

HTCondor, Docker, iRODS and HPC for every scientist with the CyVerse Discovery Environment

Ian McEwen – Software Engineer University of Arizona

mian@cyverse.org









formerly

(just so you know)









Main, high-level goals:

locker

- Make storing and analyzing data easy (help scientists focus on science, not IT)
- Use prior work, serve as glue between others' excellent work:

	scovery Environm × 🕂		
	https://de.iplantcollaborative.org/de/	C Search	☆ 自 🔸 會 🤎 🕫
l analyzing data easy 💦 🔤 📷	scovery Environment		🌲 L
ocus on science, not IT)	Data: VirSorter		
	Upload • File • Edit • Download • Share • 🍣 Refresh	💎 Trash 🗸	
erve as glue between 👘 🗖 🗖	Navigation VirSorter	Details	
	Viewing: /iplant/home/shared/imicrobe/VirSorter	Last Modified: 2016 Feb 3	
WOIK:	Community Data Name Last Modified Size	Date Submitted: 2016 Feb 3	
A	PPS ▷ ☐ aegis □ ₩ Denchmark_datasets 2015 Aug 28 15:41:25 @ ♥ S	Permissions: read	
	Construction of the second sec	Share: -	
	📶 🕨 🦸 csb472 👘 🚆 🛄 Data_package 2015 Sep 8 15:42:46 📾 🖓 🧬	Size: 3.66 GB	
- HCONDOR E	News	Type: application/gzip	
	In the second	Info-Type: -	
And the second s	▷ ○ earthenv_dem_data	md5 Checksum: d063a8f	
	↓ annotations	Send to: -	
		Tags:	
	Anns v Workflow v Share v Search Anns	Search	
YOUR SCIENCE-AS-A-SERVICE API PLATFORM	Molecular Structure (0) Name Integrated by Rating	Analyses	
GWT	Motif Identification (2) 📵 👺 BAM to SAM Roger Barth + + + + + (1)	Analyses 🗸 Edit 👻 🍣 Refresh 🛛 Share 🗸	All V Search by Name
GWI	MGS (125) MGS (125)	Name Owner App	Start Date End Date Status
	Aligners (15)	DE_Word_Coun mian@iplant ᡒ DE Wo	rd 2016 Mar 29 2016 Mar 29 Comp
	Assemblers (4)	DE_Word_Coun mian@iplant 욹 DE Wo	rd 2016 Mar 22 2016 Mar 22 Comp
	Bisulfte (0)	🔲 compress_the mian@iplant 🍃 compr	es 2016 Feb 4 1 2016 Feb 4 1 Comp
	ChiPseq (3)	compress_the mian@iplant 😓 compress	es 2016 Feb 4 1 2016 Feb 4 1 Failed
	🗧 HTProcess Pipeline (11) 🕕 🐺 SAM to sorted B Roger Barth ★ ★ ★ 🖈 (4) 🔗	Displaying 1 - 4 of 22	0 item(s)
	■ NGSEP (13) ③ 畳 Sort BAM file Roger Barth 会会会会会 (0)		
	QC and Processing (25)		
	SAM Tools (7)		
			E Feedback
	Jata: VirSonter 🖉 Apps 🥰 Analiyses		
(and more)			



RODS



RODS

Integrated Rule-oriented Database System



- Data storage & organization for large amounts of data
- Known and used by scientists already
- Where mostly-all the data comes from & goes to in our usage
- Gateway into High-Performance Computing capabilities
 through HTTP APIs
- Used for jobs which need HPC capabilities, but through same interfaces for end users within the DE



- Software container system, registry, etc.
- Known and used by scientists already
- Used for encapsulation & reproducibility of execution, distribution of tools within cluster, ease of use





A brief tour of the interfaces around jobs





App interface

(for final users who just want to do some science)

≠ Trimmomatic 0.35.0	8 🖨 🖸
Analysis Name:Trimmomatic_0.35.0_analysis1	
Inputs	
Parameters	
* Phred Quality Score:	0
PHRED+33 (more common)	¥
* Unpaired Output 1:	()
unpaired_output1.fastq	
Palindrome Clip Threshold:	0
30	~
Window Size:	0
20	~
Minimum Read Length:	0
50	\$
' Log file name:	0
trimmomatic_trim.log	
Paired End or Single End Mode:	0
Paired End	×
Output File 2:	0
output2.fastq	
Simple Clip Threshold:	0
10	^
Leading Quality:	0
3	Ŷ
Read Crop Length:	0
0	^
Output File 1:	0
output.fastq	
Seed Mismatches:	0
2	A V
' Unpaired Output 2:	0
unpaired_output2.fastq	
Window Size:	0
•	



App editor interface (for intermediate users or internal people, creating interfaces around tools)

See DE Word Count (multi-file)		
💾 Save 🔍 Preview 👻 📴 Argum	nent order	
App Items	DE Word Count (multi-file)	Details: count lines
	Tool used:	Checkbox label:
Section	wc 0.0.1	count lines
	* App name:	
	DE Word Count (multi-file)	Checked: -I Value when checked
	* App description:	Argument option when NOT checked
Files/Folders	Counts the number of words in a set of files	0
Multiple Input Files		Check item by default
Add Delete		Do not display this item in the app.
Files	* Parameters	Tool tip text:
	count lines	Enter tool tip here
	Count words	
Input File	count characters	
Browse	Count bytes	
	input Hies:	
Input Folder		
Folder Browse	🕂 Add 🗙 Delete	
Text/Numerical Input	Name	
	Select multiple input files.	
え Info Text	Tip: You can also drag and drop files from the Data window.	
Single-line Text		
I Text		
		4
Multi-line Text		
£ Text		
Command line view		
wc -w, -m, -c,		





Workflow editor (a bit rudimentary though...)

Automate Workflow							
Bave Save				🕒 Swi	itch View		
	🕂 Ad	d 🗙 Remove 👔 🏦 Move Up 🗍 Move dow	n				
Vorkflow Information	Step	Name	Description				
	1	Compress with bzip2 (for pipeline)	Bzip2 file compress	sion utility. Make it smaller! Writes to a f	file 'co		
Select & Order Apps	2	DE Word Count (multi-file)	Counts the number	of words in a set of files			
Match Outputs to Inputs							
Quick Tips							
				Se Automate Workflow			
apps to this workflow.							
				E Save	- 🔺 Step 1: Co	ompress with bzip2 (for pipeline)	
				Workflow Information	Input file:	User Provided 🗸	
					Output(s):	compressed.bz2	
				Salast & Order Apps			
				Select & Order Apps	Input Files	Stop 1: compressed bz2	
					Output(c):	Step 1. compresseu.bzz	
				Match Outputs to Inputs	Output(s).		
				Quick Tips			
				At least one input for			
				from an output of a			
				previous app.			
C. Park							





Switch View

Tool interface

Tool Information

(for software authors or internal people, who build images that get sent to internal Docker registry)

Request New Tool Installation	Category
Discovery Environment tools are executables or binaries built.	from which apps are
Before a new app can be created, the tool that app is base available in the DE.	ed on must be made
This form contains the required information needed for ou new tool.	r team to install your
Once your request is submitted, you will receive status up installation progresses.	dates as the

What is the r	name of the too	01?	
Please brief	ly describe the	tool:	
			0.1-1

Cancel

* Tool Aname:		
Description: Trimmomatic: A floxible filmmer for flumina Sequence Data. * Type: executable Altibudor: Dr. R. Veera Manikandan Version: 0.33 tockion: Thyp: Sequence Data. * Topie: 0.33 tockion: Thyp: Sequence Data. * Topie: 0.43 * Topie: oreara * Topie: orear	* Tool Name:	trimmomatic-0.33
 Type: executable Type: executable The Versa Manikandan Vestor: 0.33 Location: https://pantoollaborative.org Tool Implementation Implementation Implem	Description:	Trimmomatic: A flexible trimmer for Illumina Sequence Data.
* Type: executable Type: executable This //panticollaborative org To Ir. R. Vera Manikandan Uersion: 0.33 Costainer Intips/Inpanticollaborative org To Iruppementor: driveera * Intips/Inpanticollaborative org To Iruppementor: driveera * Intips/Inpanticollaborative org To Iruppementor: driveera * Intips/Inpanticollaborative org To Iruppementor: * Intips/Inpanticollaborative org To Iruppementor: * Intips/Inpanticollaborative org * Iruppementor: * Iruppe		
* Type: Conclusion Attribution: Dr. R. Veera Manikandan Version: Q 33 * Location: https://gantcollaborative.org * Location: Intps://gantcollaborative.org * Location: drevera * Implementor: drevera * Implementor: drevera * Implementor: drevera * Tool Implementor: drevera * Ass Deate Fee Name(s) Implementor: Sample Output Files: Implementor: * Ass Deate Fee Name(s) Implementor: Container Image Southimmomatic Tag: v0 33 URL: https:// Southimmomatic Tag: v0 33 URL: https:// Southimmomatic Maree: Maree: Working Direcony: Vexentor is contain any risky		
* Type:: executable Type:: Order Manikandan Version:: O.3 Location: https://plantcollaborative.org Tool Implementation * Implementation regioneration: divera@cmcvellore.ac.in main: materia: Add ★ Decke Fe Name(s) Sample https:// Sample h		
* Type: executable Attribution: Dr. R. Veera Manikandan Version: 0.33 * Location: https://plantcollaborative.org * Tool: Insplementation * Implementation * Imp		
 *Type: executable Attribution: Dr. R. Veeta Manikandan Version: 0.23 Cocation: https://plantcollaborative.org Tool Implementation *Implementor: dirveera@cmovellore.ac.in Sample Input Files: Add X Detele Fe Name(s) Container Image *Name: 5000/trimmomatic Tage: v0.33 URL: https://j		
* Type: pxecutable Attitution: Dr. R. Vetra Manikandan Version: D.3.3 Location: https://ipintcollaborative.org * Tool Implementor drveera * Implementor drveera@cmcvellore.ac.in * Implementor drveera@cmcvellore.ac.in * Implementor drveera@cmcvellore.ac.in * Implementor drveera@cmcvellore.ac.in * Add X Detele Fee Name(s) Sample Output Files: * * Add X Detele Fee Name(s) Sample Output Files: * * Add X Detele Fee Name(s) * Name: 50000trimmomatic Tag: v0.33 URL: https:// Amme: S000/trimmomatic * Ocntainer Details Name: Working Directory: Working Directory: WARNING: Do not add a bol without an Entry Point setting if its Docker image also does not have a default "ENTRYPOINT", if a bol like this is required, then its Network Kode setting should be setting none' to contain any risky scripts run by this tool. Entry Point		
 * Type: executable Attributon: Dr. R. Veera Manikandan Version: D. 3.3 * Location: https://ipintoollaborative.org * Location: https://ipintoollaborative.org * Indigementor driveera drivera@cmovellore.ac.in Email: Sample Input Files: 		
Attribution: Dr. R. Vera Manikandan Version: 0.33 Location: https://plant.oollaborative.org Tool Implementation implementor: Implementor: drveera@cmcvellore.ac.in Email: Sample Input Files: Implementor: drveera@cmcvellore.ac.in Feh Name(s) Sample Output Files: Implementor: add X Delete File Name(s) Sample Output Files: Implementor: Soutput Files:	* Туре:	executable
Version: 0.33 * Locator: https://jplant.collaborative.org Tool Implementor: drveera@cmcvellore.ac.in * Implementor: drveera@cmcvellore.ac.in * Add X Delete File Name(s) Sample Dutput Files: * * Add X Delete File Name(s) File Name(s) * Container Image * * Name: \$000/trimmomatic Tag: vo 33 URL: https://jwlinkocollaborative.org * Add to all bol without an Entry Point setting if its Docker image allso doces not have a default "ENTRYPOINT". If a tool like this is required, then its Network Mode setting should be set to "none" to contain any risky scripts run by this tool. Entry Point:	Attribution:	Dr. R. Veera Manikandan
Location: https://pintcollaborative.org Tool Implementation Implementor: drveera Implementor: drveera Implementor: drveera@cmcvellore.ac.in Email: Sample Output Files: Add X Debte File Name(s) Container Image Name: S000/trimmomatic Tag: v0.33 URL: https://	Version:	0.33
Tool Implementation I	* Location:	https://iplantcollaborative.org
Implementor: drvera Implementor: drvera@cmcvellore.ac.in Email: Sample Input Files: Implementor: drvera@cmcvellore.ac.in File Name(s) Sample Output Files: Sample Output Files: Implementor Implementor: drvera@cmcvellore.ac.in File Name(s) Sample Output Files: Implementor: Implementor Implementor: Implementor Implementor: Implementor Implementor: Implementor Sample Output Files: Implementor Implementor: Implementor Implementor: Implementor Implementor: Implementor Sample Output Files: Implementor Implementor: Implementor Implementor: <td>- Tool Implemente</td> <td>tion</td>	- Tool Implemente	tion
Implementor driveera@cmcvellore.ac.in Email: Greena@cmcvellore.ac.in Sample Input Files: Greena@cmcvellore.ac.in Implementor Greena@cmcvellore.ac.in	* Implementor:	drveera
Sample Input Files:	* Implementor Email:	drveera@cmcvellore.ac.in
Add X Delete Fie Name(s) Sample Output Files: Add X Delete Fie Name(s) Container Image Name: S000/trimmomatic Tag: V0.33 URL: https://5000/trimmomatic Container Details Name: Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker Image also does not have a default 'ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to 'nome' to contain any risky scripts run by this tool. Entry Point Memory Limit:	Sample Input F	les:
File Name(s) Sample Output Files:		
Sample Output Files: Add X Dekte File Name(s) Container Image Name: S000/trimmomatic Tag: V0.33 URL: https://	📫 Add 🗙 🛛	Delete
Sample Output Files:	File Name(s)	Delete
Add X Delete File Name(s) Container Image Name: S000/trimmomatic Tag: V0.33 URL: https://	File Name(s)	Jelete
File Name(s) Container Image Name: 123 v0.33 URL: https:// 5000/trimmomatic Container Details Name: Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default "ENTRYPOINT". If a tool like this is required, then its Network Mode setting should be set to "none" to contain any risky scripts run by this tool. Entry Point: Memory Limit:	He Add X	Files:
Container Image Name: Tag: V0.33 URL: https://5000/trimmomatic Container Details Name: Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default 'ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to 'none' to contain any risky scripts run by this tool. Entry Point: Memory Limit:	File Name(s)	Files:
Container Image Name: S000/trimmomatic Tag: V0.33 URL: https://5000/trimmomatic Container Details Name: Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default 'ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to 'none' to contain any risky scripts run by this tool. Entry Point: Memory Limit:	Add X File Name(s) Sample Output Add X File Name(s)	Files:
Container Image Name: S000/trimmomatic Tag: V0.33 URL: https://	Add X File Name(s) Sample Output Add X File Name(s) File Name(s)	Files: Delete
Tag: v0.33 URL: https:// bttps:// 5000/trimmomatic Container Details Name: Working Directory: Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default 'ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to 'none' to contain any risky scripts run by this tool. Entry Point:	Add X File Name(s) Sample Output Add X File Name(s)	Files: Delete
Intg. volsa URL: https://5000/trimmomatic ▲ Container Details Name: Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default `ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to `none` to contain any risky scripts run by this tool. Entry Point: Memory Limit:	Add X File Name(s) Sample Output Add X File Name(s) File Name(s) Container Image Name:	Files:
Container Details Name: Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default 'ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to 'none' to contain any risky scripts run by this tool. Entry Point: Memory Limit:	Add X File Name(s) Sample Output Add X File Name(s) File Name(s) Container Image Name: Tag:	Files: Delete
Container Details Name: Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default `ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to `none` to contain any risky scripts run by this tool. Entry Point: Memory Limit:	Add X File Name(s) Sample Output Add X File Name(s) File Name(s) Container Image Name: Tag: UEL:	Files: Delete
Name: Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default 'ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to 'none' to contain any risky scripts run by this tool. Entry Point: Memory Limit:	Add X File Name(s) Sample Output Add X File Name(s) Container Image Name: Tag: URL:	Files: Delete 5000/trimmomatic v0.33 https://5000/trimmomatic
Working Directory: WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default 'ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to 'none' to contain any risky scripts run by this tool. Entry Point: Memory Limit:	Add X File Name(s) File Name(s) Sample Output Add X File Name(s) File Name(s) Container Image Name: Tag: URL: Container D	Files: Delete 5000/trimmomatic v0.33 https://5000/trimmomatic tails
WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default `ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to `none' to contain any risky scripts run by this tool. Entry Point: Memory Limit:	Add X File Name(s) File Name(s) Sample Output Add X File Name(s) File Name(s) Container Image Name: Tag: URL: Container Do Name:	Files: Delete
Entry Point: Memory Limit:	Add X I File Name(s) Sample Output Add X I File Name(s) Container Image Name: Tag: URL: Container D Name: Working Directo	Files: Delete Files: Delete 5000/trimmomatic V0.33 https://5000/trimmomatic tails
Memory Limit:	Add X I File Name(s) Sample Output Add X I File Name(s) Container Image Name: Tag: URL: Container D Name: Working Directo WARNING: Do n Mode setting sh	Files:
	Add X I File Name(s) Sample Output Add X I File Name(s) Container Image Name: Tag: URL: Container Dr Name: Working Directo WARNING: Do n Mode setting sh Entry Point:	Files: Delete Files: Delete 5000/trimmomatic V0.33 https:// 5000/trimmomatic tails tails tails tails tails
	Add X File Name(s) Sample Output Add X File Name(s) Container Image Name: Tag: URL: Container Du Name: Working Directo WarNING: Do a Mode setting sh Entry Point: Memory Limit:	Files: Delete Files: Delete 5000/trimmomatic V0.33 https://5000/trimmomatic ttails ttails ttails tot add a tool without an Entry Point setting if its Docker image also does not have a default 'ENTRYPOINT'. If a tool like this is required, then its Network puid be set to 'none' to contain any risky scripts run by this tool.



Now let's talk about just the HTCondor part, behind the scenes







(maybe I should focus this a bit)













Future (potential) Work & Ideas

- Time limitation (e.g., 48h per job, then request more)
- Other computational limits (number of processes (no forkbombs!), memory, CPU, network)
- Interactive jobs (iPython etc., or even just for exploration)
- More complex (non-linear, more flexible) workflows
- Public APIs and command-line tools (expand into more technical users)
- "connect your own compute" spin up machines with appropriate software, put keys/other credentials into DE and have your own jobs go to non-limited, no-waiting compute you control

(question time!)

https://de.iplantcollaborative.org/ (try it out yourself) more to say/ask?

Ian McEwen University of Arizona mian@cyverse.org



