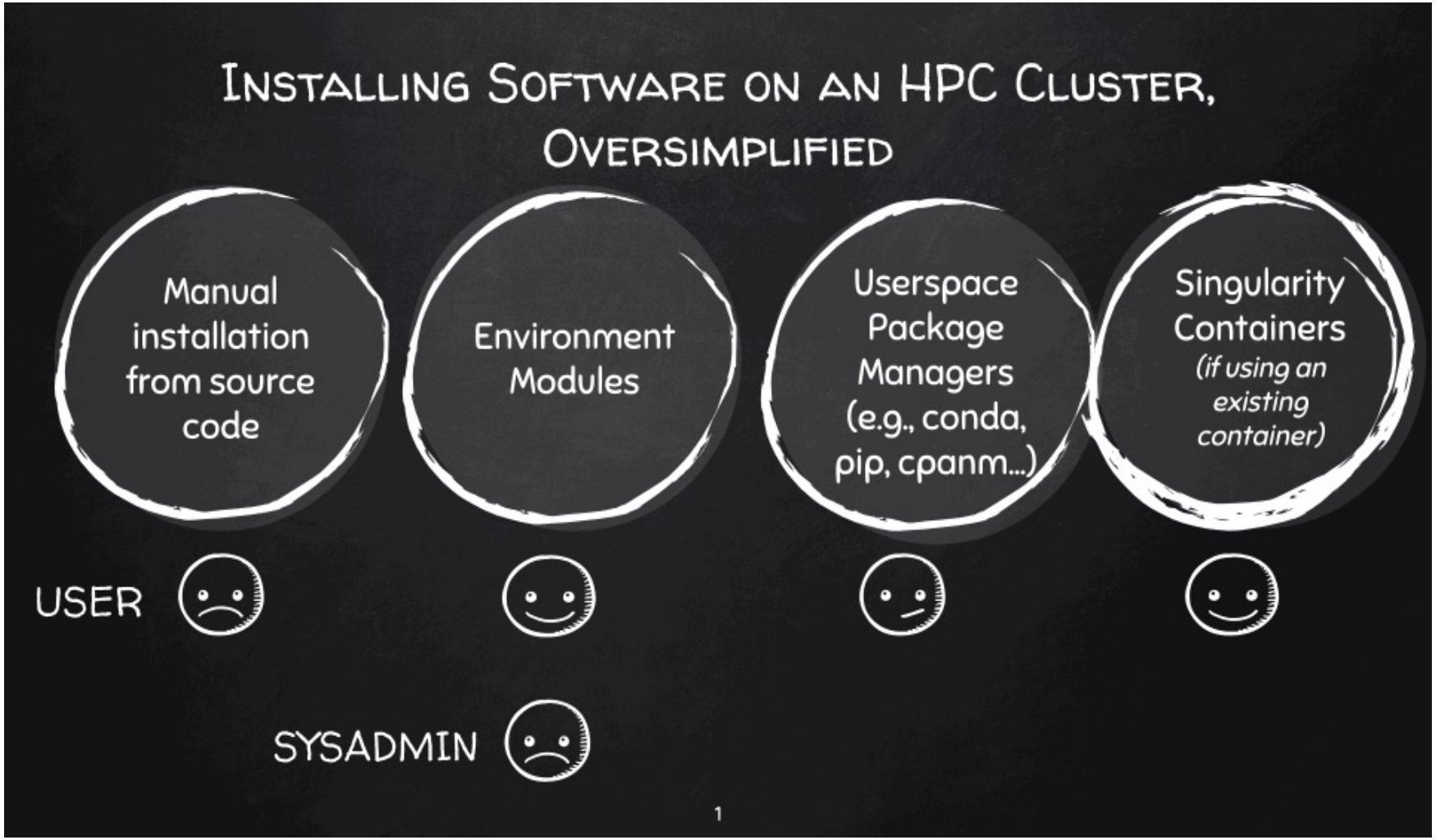
RUNNING BIOINFORMATICS ANALYSIS USING HPC

Yucheng Zhang, Senior Life Science Scientist

PURDUE UNIVERSITY | Rosen Center for Advanced Computing

Package installation



source: https://github.com/harvardinformatics/bioinformatics-coffee-hour/blob/master/singularity/images/installing_software.gif

What to expect from this bioinformatics workshop?

Objectives

- Introduce current bioinformatics resources on clusters.
- Learn how to install packages using different approaches.
- Become familiar with different ways of running bioinformatics analysis on clusters.

Running Bioinformatics Analysis using HPC

Bioinformatics resources

Bioinfo

- Maintained by Purdue Bioinformatics Core.
- > 700 tools (As of March, 2023).
- Compiled and installed from source codes on old operation systems.
- Available on Gilbreth, Brown, Bell, Scholar, and Workbench.
- Unavailable on future clusters since Negishi, due to compatibility issues.
- Send emails to bioinformatics@purdue.edu for issues.

Biocontainers

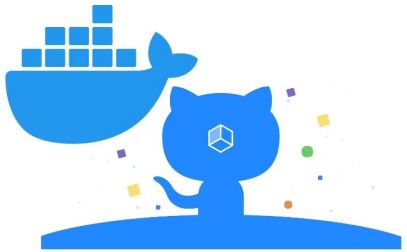
- Maintained by Purdue RCAC.
- ~600 tools (As of March, 2023).
- Deployed based on singularity containers.
- Available on all RCAC clusters including ACCESS Anvil.
- Send emails to rcac-help@purdue.edu for issues.

Containers

A **container** is an abstraction for a set of technologies that aim to solve the problem of how to get software to run reliably when moved from one computing environment to another.

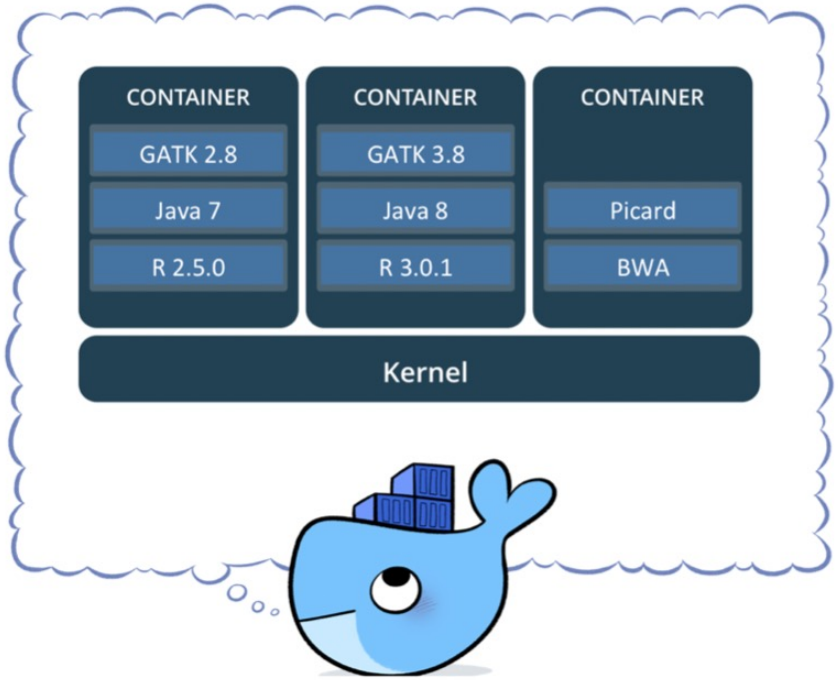
A container **image** is simply a file (or collection of files) saved on disk that stores everything you need to run a target application or applications.

Registry: a place to store (and share) container images.

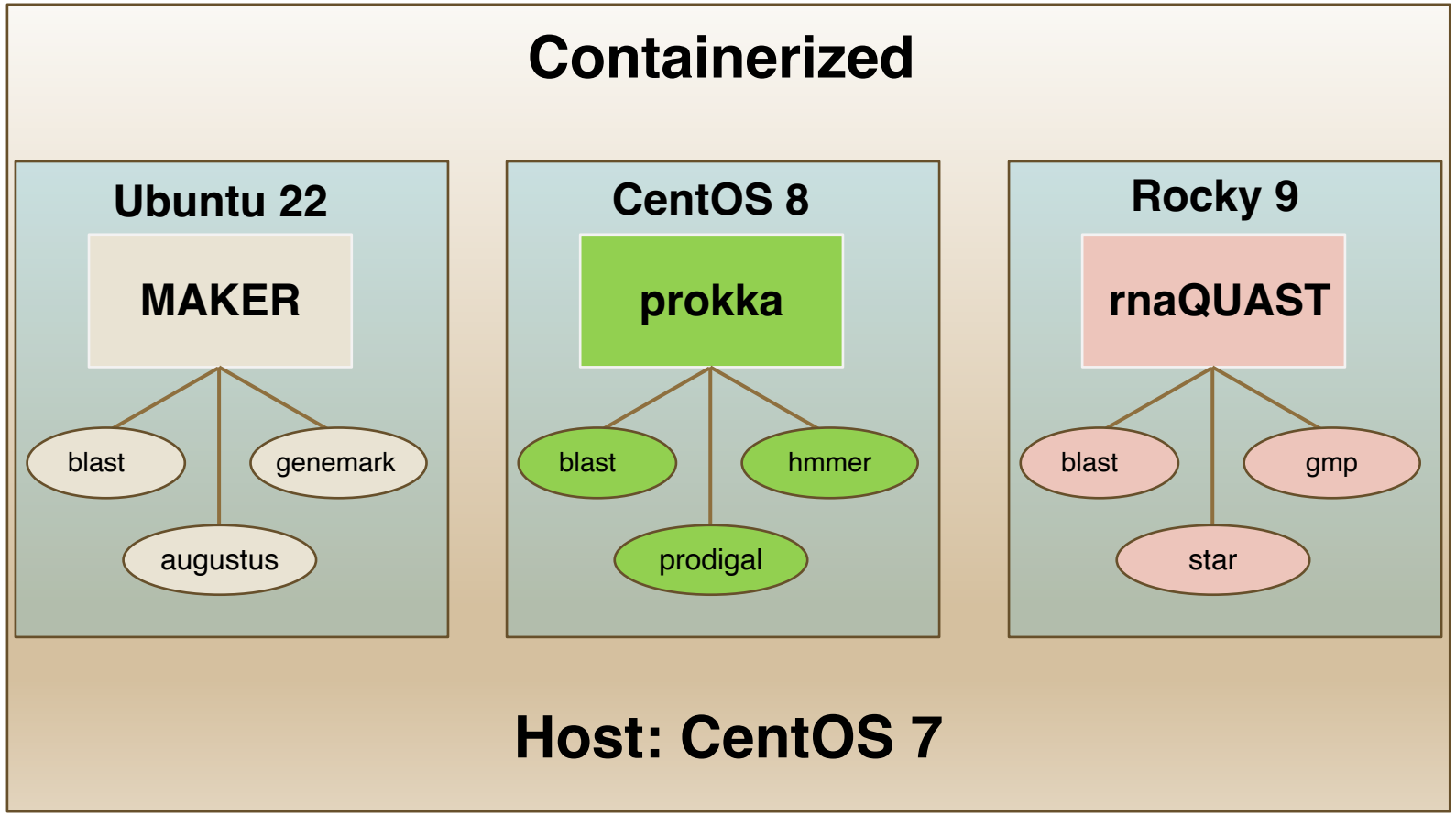


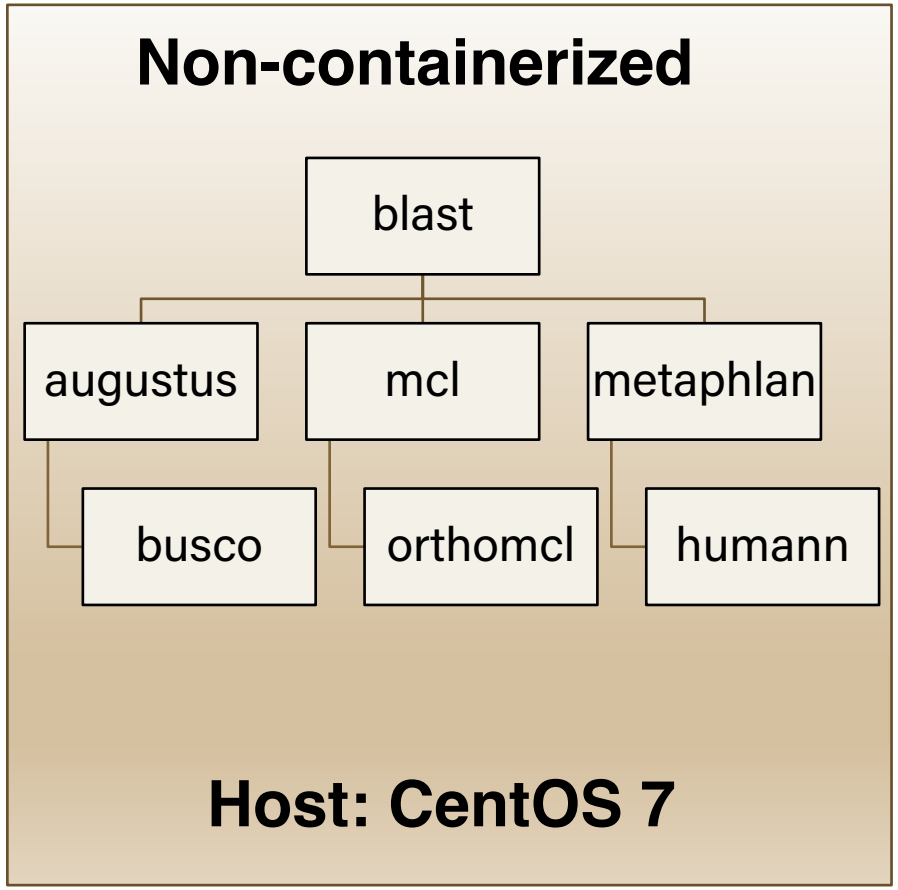
GitHub Container Registry

BioContainers



Containerized vs. non-containerized applications





Want to know more about RCAC biocontainers?

Workshops

- Containers 101
- Biocontainers 101
- **Containerization Bioinformatics Applications for HPC: March 23**

Publications

- BioContainers on Purdue Clusters

  Y Zhang, L Gorenstein. 2022. Practice and Experience in Advanced Research Computing, 1-2.

- Containerized Bioinformatics Ecosystem for HPC

  Y Zhang, L Gorenstein, P Bhutra, RT DeRue. 2022 IEEE/ACM International Workshop on HPC User Support Tools (HUST), 1-10.

Load and use biocontainers

Load biocontainers

module –-force purge # optional but highly recommended
module load biocontainers

Check available applications

module avail

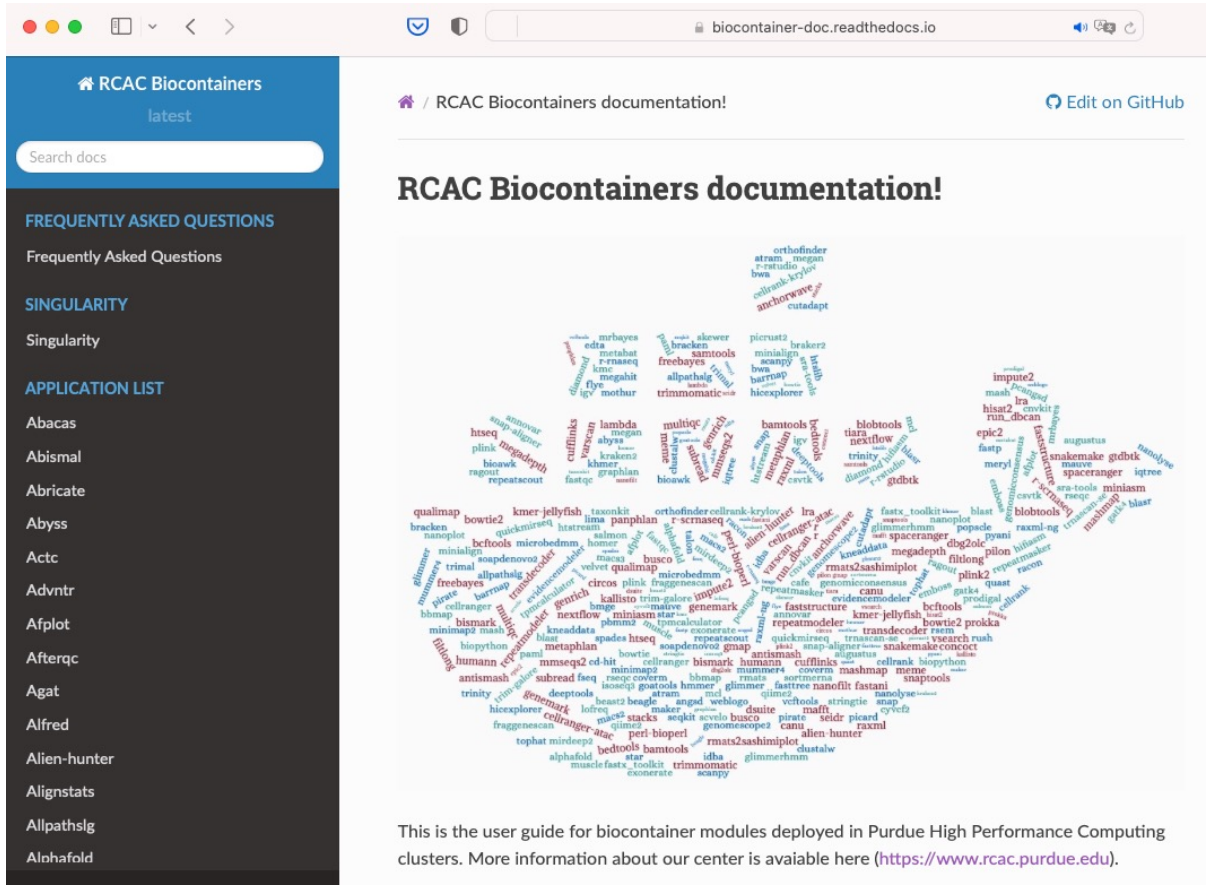
Load and run specific tools

module load samtools/1.16 # specify version can guarantee reproducibility

samtools idxstats input.bam

Biocontainers documentation

$ module load biocontainers
User guides for each biocontainer module can be found in https://biocontainer-doc.readthedocs.io/en/latest



Example job using GPU

Warning

Using `#!/bin/sh -l` as shebang in the slurm job script will cause the failure of some biocontainer modules. Please use `#!/bin/bash` instead.

Note

Notice that since version 2.2.0, the parameter `--use_gpu_relax=True` is required.

To run alphafold using GPU:

```
#!/bin/bash
#SBATCH -A myallocation     # Allocation name
#SBATCH -t 20:00:00
#SBATCH -N 1
#SBATCH -n 24
#SBATCH --gres=gpu:1
#SBATCH --job-name=alphafold
#SBATCH --mail-type=FAIL,BEGIN,END
#SBATCH --error=%x-%J-%u.err
#SBATCH --output=%x-%J-%u.out

module --force purge
ml biocontainers alphafold

run_alphafold.sh --flagfile=full_db_20221014.ff \
    --fasta_paths=sample.fasta --max_template_date=2022-02-01 \
    --output_dir=af2_full_out --model_preset=monomer \
    --use_gpu_relax=True
```

Running Bioinformatics Analysis using HPC

Installing packages from source

Make

GNU Make is a program often used for compiling software. It uses a plain text file named **makefile** or **Makefile**.

Steps

1. Unpack the source code archive.
2. ***Configure*** the package. ## Some packages do not have the ***configure*** file
3. Run ***make*** to build the programs.
4. Run ***make install*** to install the package. # Optional
   ❌ Do not run ~~**sudo make install**~~

- In default, ***make install*** will install applications into **/usr/local**, but regular users do not have permission to write into **/usr/local**.
- The best way is to install applications into your home directory or /depot by passing the option **--prefix=TargetDirName** to **./configure**.

Make example1: bwa

```
mkdir -p ~/bioinformatics
cd ~/bioinformatics
# Load gcc compiler
module load gcc ## Recommend to load the newest version of gcc

# Download source code archive from https://github.com/lh3/bwa/releases/tag/v0.7.17
wget -O bwa_0.7.17.tar.gz https://github.com/lh3/bwa/archive/refs/tags/v0.7.17.tar.gz
# Unpack source code archive
tar -xvf bwa_0.7.17.tar.gz
cd bwa-0.7.17

# compile the code
make
# Add the bwa directory to $PATH
export PATH=$PATH:$HOME/bioinformatics/bwa-0.7.17    ## This can be added to .bashrc

# Run bwa
bwa
```

Make example2: hmmer

```
# Load gcc compiler
module load gcc ## Recommend to load the newest version of gcc

# Download source code
wget http://eddylab.org/software/hmmer/hmmer.tar.gz
# Unpack source code archive
tar -xvf hmmer.tar.gz
cd hmmer-3.3.2/

# compile the code
./configure --prefix=$HOME/myapps
make
make install

# Add the myapps directory to $PATH
export PATH=$PATH:$HOME/myapps/bin
```

```
# Run hmmer
hmmsearch -h
```

Make example3: RegTools

```
mkdir -p ~/bioinformatics
cd ~/bioinformatics
# Load gcc compiler and cmake
module load gcc cmake ## Recommend to load the newest version of gcc and cmake

# Download source code archive from https://github.com/griffithlab/regtools/releases/tag/1.0.0
wget -O regtools_1.0.0.tar.gz https://github.com/griffithlab/regtools/archive/refs/tags/1.0.0.tar.gz

# Unpack source code archive
tar -xvf regtools_1.0.0.tar.gz

cd regtools-1.0.0/

# compile the source code
mkdir build
cd build/
cmake ..
make

# Add the bwa directory to $PATH
export PATH=$PATH:$HOME/bioinformatics/regtools-1.0.0/build

# Run RegTools
regtools
```

Running Bioinformatics Analysis using HPC

Installing packages using conda

Conda

Conda is an open-source package manager and virtual environment manager for installing packages.

```
module spider anaconda
module load  anaconda/XXXX          #The latest version is recommended
conda create --name MyEnv python=3.9 1stPackage 2ndPackage
conda env list        #Confirm the conda environment is created
conda activate MyEnv
conda install 3rdPackage
conda deactivate
```

conda channels

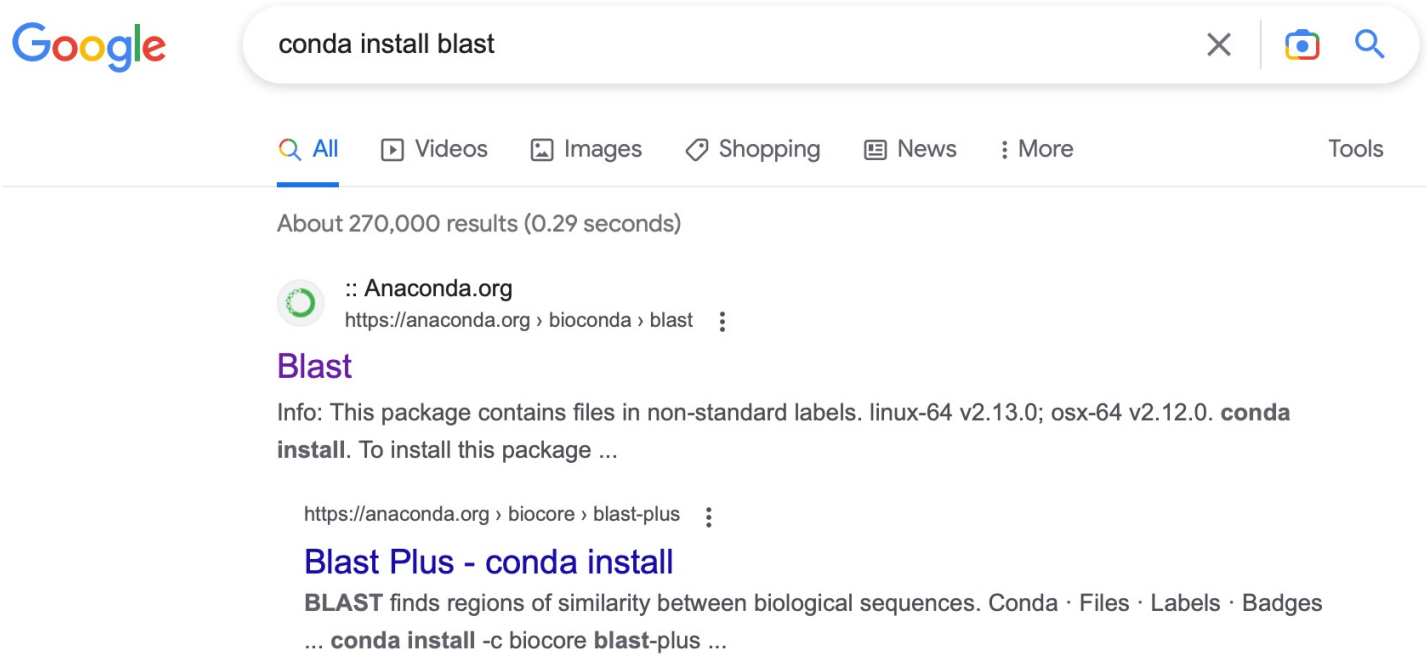
- Conda channels are the locations where packages are stored.
- Conda search or download packages from channels.
- The default set of channels is called **defaults**.
- To install a package that is not in **defaults**, you need to tell conda which channel contains the package.

```
conda install -c pytorch pytorch
conda install -c conda-forge r-base
```

bioconda

Bioconda is the channel for bioinformatics applications.
As of today, bioconda includes over 10 thousands bioinformatics applications.

```
conda install -c bioconda blast
conda install -c bioconda samtools
```



To use conda to install packages, Google
conda install packageName

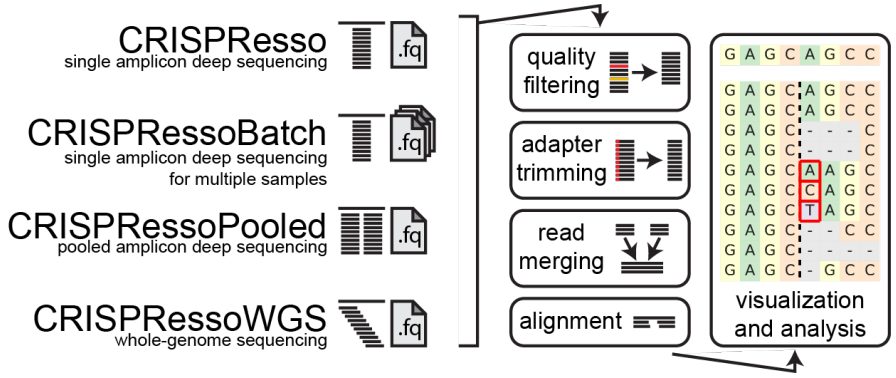
BIOCONDA®

Running Bioinformatics Analysis using HPC

Containerized applications

Docker containerization

https://github.com/pinellolab/CRISPResso2



Docker

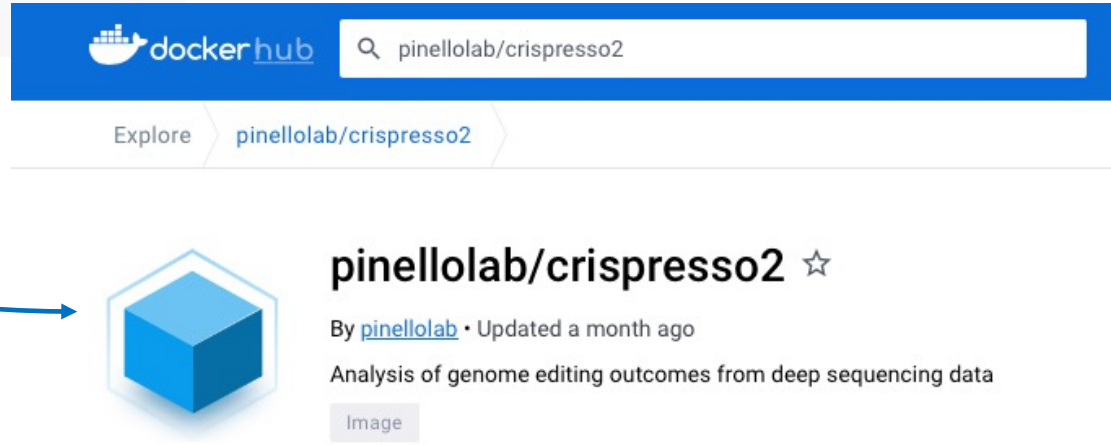
CRISPResso2 can be used via the Docker containerization system. This system allows CRISPResso2 to run on your system without configuring and installing additional packages. To run CRISPResso2, first download and install docker: https://docs.docker.com/engine/installation/

Next, Docker must be configured to access your hard drive and to run with sufficient memory. These parameters can be found in the Docker settings menu. To allow Docker to access your hard drive, select 'Shared Drives' and make sure your drive name is selected. To adjust the memory allocation, select the 'Advanced' tab and allocate at least 4G of memory.

To run CRISPResso2, make sure Docker is running, then open a command prompt (Mac) or Powershell (Windows). Change directories to the location where your data is, and run the following command:

```
docker run -v ${PWD}:/DATA -w /DATA -i pinellolab/crispresso2 CRISPResso -h
```

https://hub.docker.com/r/pinellolab/crispresso2



PURDUE UNIVERSITY | Rosen Center for Advanced Computing

Singularity workflow on HPC

1 (Optional). **Build** singularity containers on a computer system where you have root or sudo privilege, e.g., your personal computer with singularity installed.

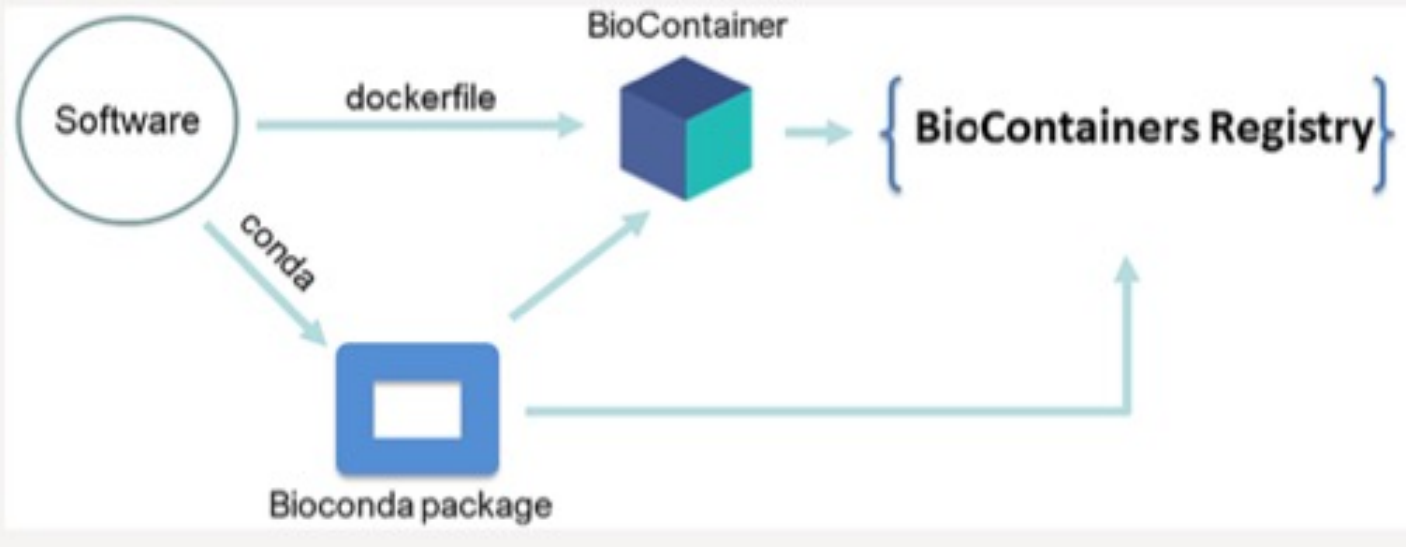
2. **Pull** the public containers or **transfer** your own containers to HPC.

3. **Run** singularity containers on the HPC system.

source: https://github.com/harvardinformatics/bioinformatics-coffee-hour/blob/master/singularity/images/installing_software.gif

BIOCONDA®

- BioContainers is integrated with Bioconda, which is the conda channel for bioinformatics applications.
- BioContainers registry is the largest registry for bioinformatics applications.
- As of today, BioContainers provides containers for over 10 thousand bioinformatics applications.



J. Proteome Res. 2021, 20, 4, 2056–2061

You can find almost all bioinformatics applications from here:https://bioconda.github.io/conda-package_index.html

Singularity demo1: samtools

https://bioconda.github.io/recipes/samtools/README.html#package-samtools

Installation

With an activated Bioconda channel (see set-up-channels), install with:

```
conda install samtools
```

and update with:

```
conda update samtools
```

or use the docker container:

```
docker pull quay.io/biocontainers/samtools:<tag>
```

(see samtools/tags for valid values for <tag>)

Click to get tags/versions

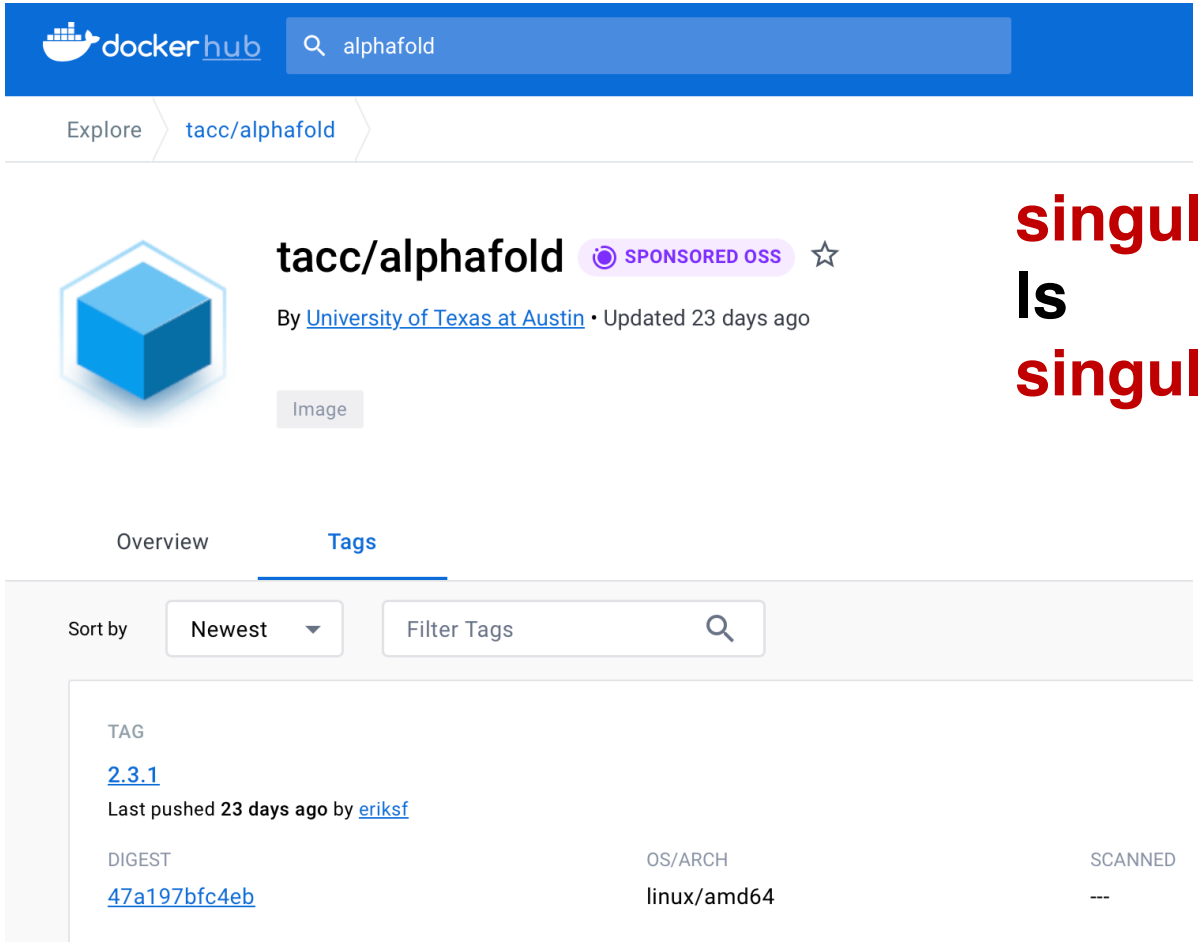
singularity pull docker://quay.io/biocontainers/samtools:1.16.1--h6899075_1

ls

singularity exec samtools_1.16.1--h6899075_1.sif **samtools**

Singularity demo2: Alphafold

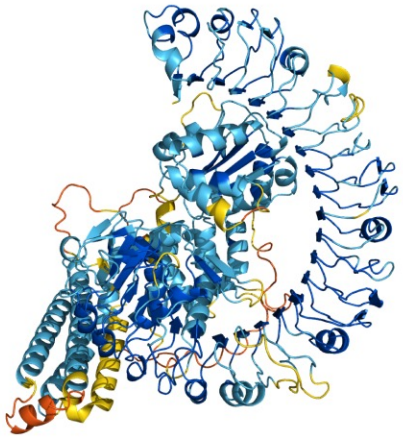
https://hub.docker.com/r/tacc/alphafold/tags



singularity pull docker://tacc/alphafold:2.3.1

ls

singularity run alphafold_2.3.1.sif --helpfull



PURDUE UNIVERSITY | Rosen Center for Advanced Computing

Running Bioinformatics Analysis using HPC

R

.Rprofile

- Each cluster has multiple versions of R and packages installed with one version of R may not work with another version of R.
- Libraries for each R version must be installed in a separate directory.
- Define the directory where your R packages will be installed using the environment variable **R_LIBS_USER**.

How to set Up R Preferences with .Rprofile

```
curl -#LO https://www.rcac.purdue.edu/files/knowledge/run/examples/apps/r/Rprofile_example
mv -ib Rprofile_example ~/.Rprofile
```

R package installation: load required modules

```
if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager")

BiocManager::install("Cardinal")
```

```
fftwtools.c:28:18: fatal error: fftw3.h: No such file or directory
 #include<fftw3.h>
                  ^
compilation terminated.
make: *** [fftwtools.o] Error 1
ERROR: compilation failed for package 'fftwtools'
```

```
module load r/4.2.2

module load fftw/3.3.7
```

```
* DONE (Cardinal)

The downloaded source packages are in
        '/tmp/RtmpRSSU7U/downloaded_packages'
```

r-rnanseq and r-scrnaseq

R-RNAseq

Customized R container for RNAseq analysis.

- ComplexHeatmap
- DESeq2
- DEXSeq
- edgeR
- ggrepel
- Limma
- pheatmap
- tidyverse

https://biocontainer-doc.readthedocs.io/en/latest/source/r-rnaseq/r-rnaseq.html

R-scRNAseq

Customized R container for scRNAseq analysis.

- CellChat
- CoGAPS
- DESeq2
- doSNOW
- DropletUtils
- edgeR
- Limma
- miQC
- monocle
- monocle3
- Nebulosa
- ProjecTILs
- rliger
- scCATCH
- scDblFinder
- SCHNAPPs
- scMappR
- seurat
- seurat-wrappers
- SingleR
- SnapATAC
- SoupX
- tidyverse
- tricycle
- velocyto.R

And more

https://biocontainer-doc.readthedocs.io/en/latest/source/r-scrnaseq/r-scrnaseq.html

Running Bioinformatics Analysis using HPC

Python

conda-env-mod

To facilitate the process of creating and using Conda environments, we support a script (*conda-env-mod*) that generates a module file for an environment, as well as an optional Jupyter kernel to use this environment in a JupyterHub notebook.

You must load one of the anaconda modules in order to use this script.

Create a conda environment

```
$ module load anaconda/2020.11-py38
$ coda-env-mod create -n mypackages --local-python
```

Load the conda environment

```
$ module load use.own
$ module load conda-env/mypackages-py3.8.5  ## py3.8.5 is the python version in the anaconda module
```

Install packages

```
$ conda install PackageName
```

Create conda environment for Jupyter

```
$ conda-env-mod create -n mypackages --jupyter #--jupyter always implies '--local-python'
```

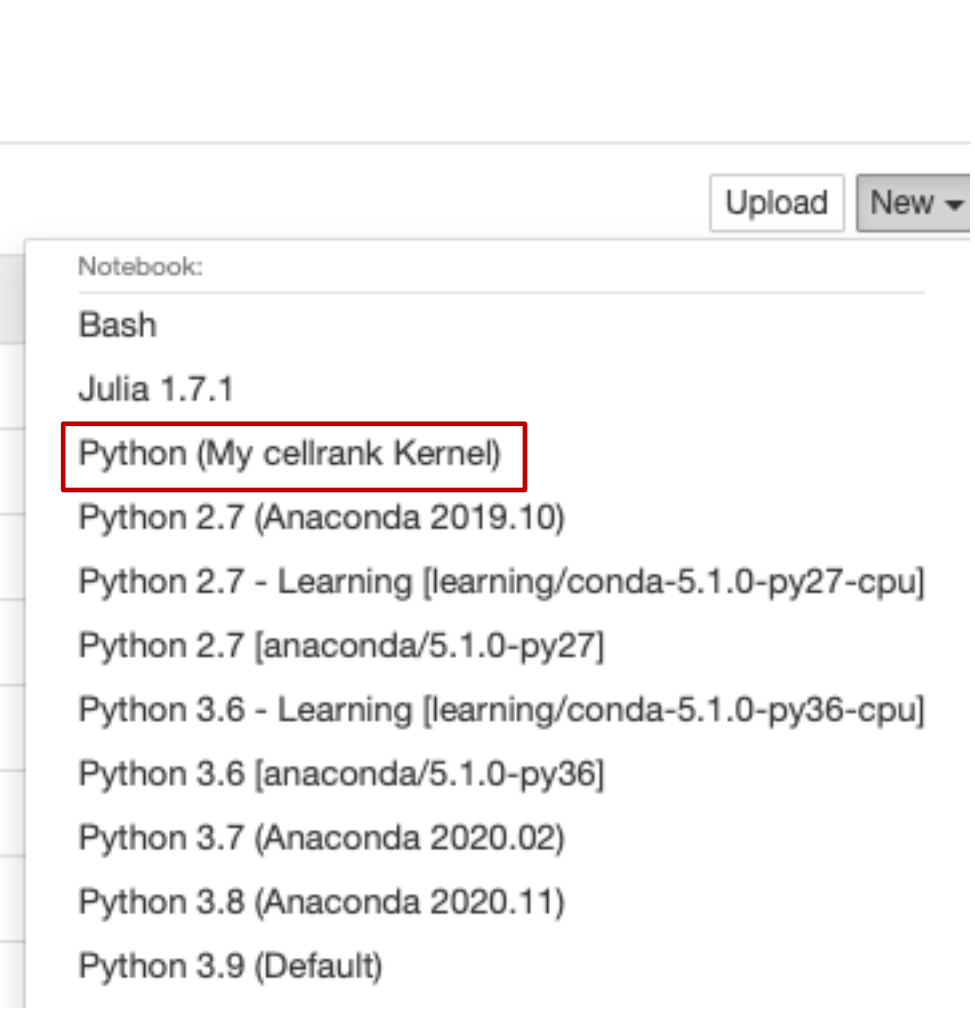
conda-env-mod is very powerful. Detailed usage can be found here
https://www.rcac.purdue.edu/knowledge/negishi/run/examples/apps/python/packages

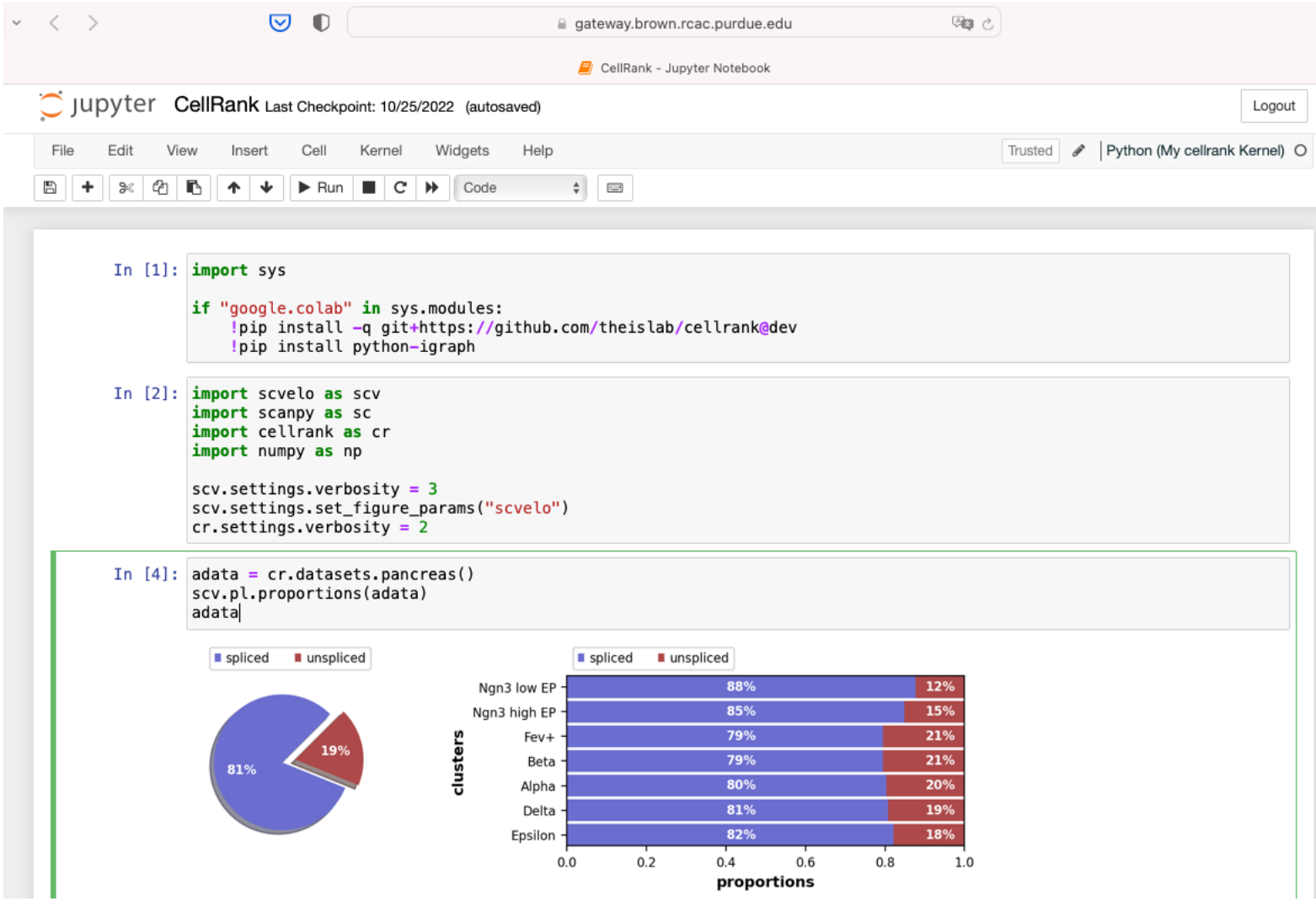
conda-env-mod demo: cellrank

```
$ module load anaconda/2020.11-py38
$ conda-env-mod create -n cellrank --jupyter
$ module load use.own
$ module load conda-env/cellrank-py3.8.5
$ conda install -c conda-forge -c bioconda cellrank

$ python
Python 3.8.5 (default, Sep  4 2020, 07:30:14)
[GCC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more
information.
>>> import cellrank
```

https://github.com/theislab/cellrank

Open OnDemand Jupyter





Running Bioinformatics Analysis using HPC

Interactive Jobs

Interactive Jobs

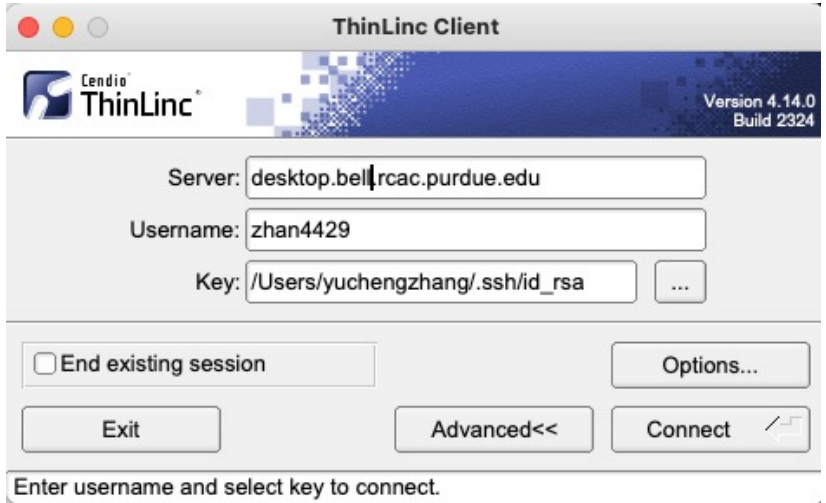
Interactive jobs are run on compute nodes, while giving you a shell to interact with. They give you the ability to type commands or use a graphical interface in the same way as if you were on a front-end login host.

1 node, 24 cores, and 10 hours walltime

```
sinteractive -N1 -n24 -t10:00:00 -A accountName
```

1 node, 12 cores, 1 GPU and 4 hours walltime

```
sinteractive -N1 -n12 --gres=gpu:1 -t4:00:00 -A accountName
```



https://www.cendio.com/thinlinc/download

Batch Jobs: CPU

Using `#!/bin/sh -l` as shebang in the slurm job script will cause the failure of some biocontainer modules. Please use `#!/bin/bash` instead.

```
#!/bin/bash

#SBATCH -A accountName     # the queue you want to use
#SBATCH -t 20:00:00
#SBATCH -N 1
#SBATCH -n 24
#SBATCH --job-name=star
#SBATCH --mail-type=FAIL,BEGIN,END
#SBATCH --error=%x-%j-%u.err
#SBATCH --output=%x-%J-%u.out

module --force purge
module load biocontainers star/2.7.10a

STAR  --runThreadN 24  --runMode genomeGenerate  \
        --genomeDir ref_genome  \
        --genomeFastaFiles ref_genome.fasta
```

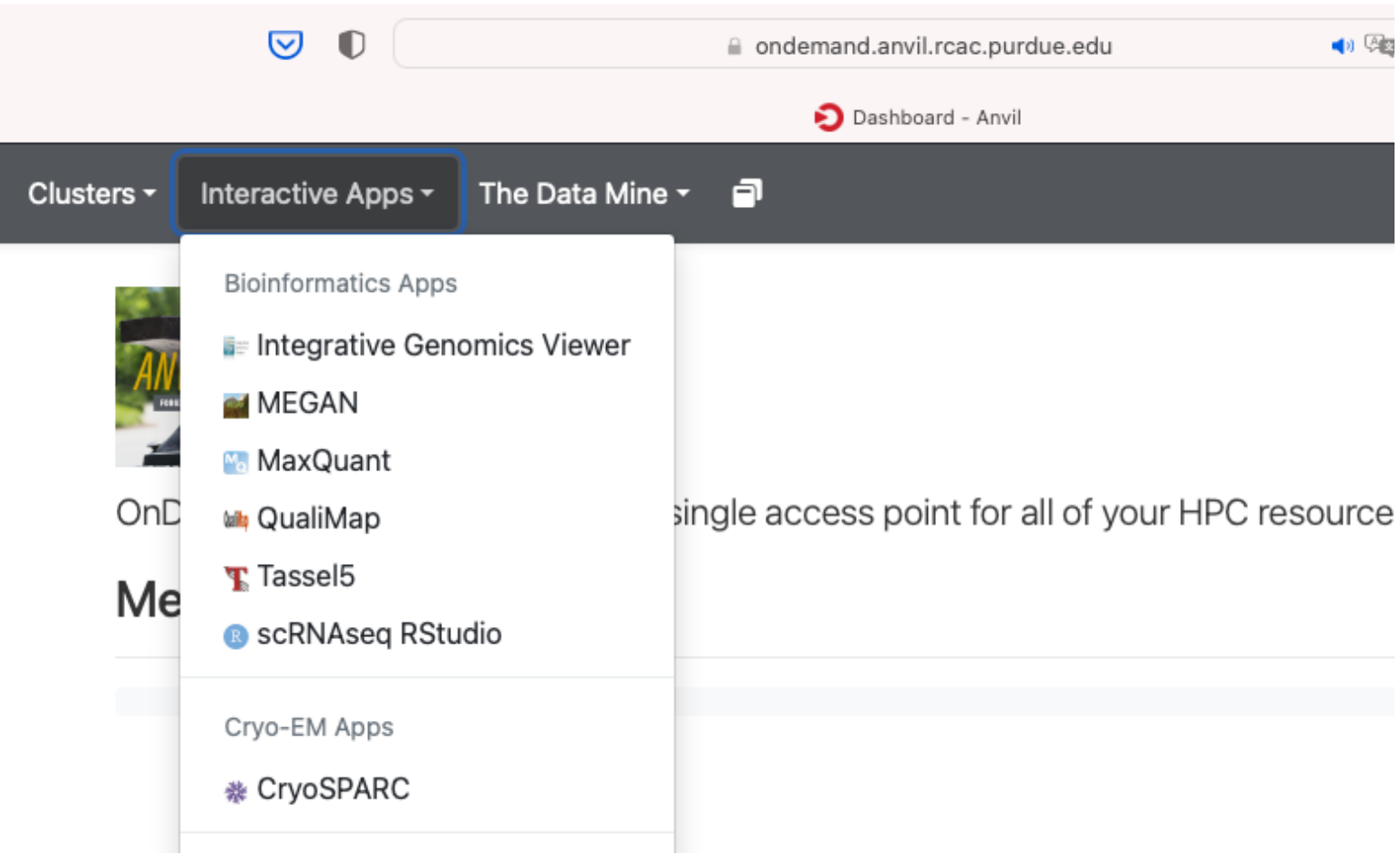
%x: job name
%j: jobid
%u: userid

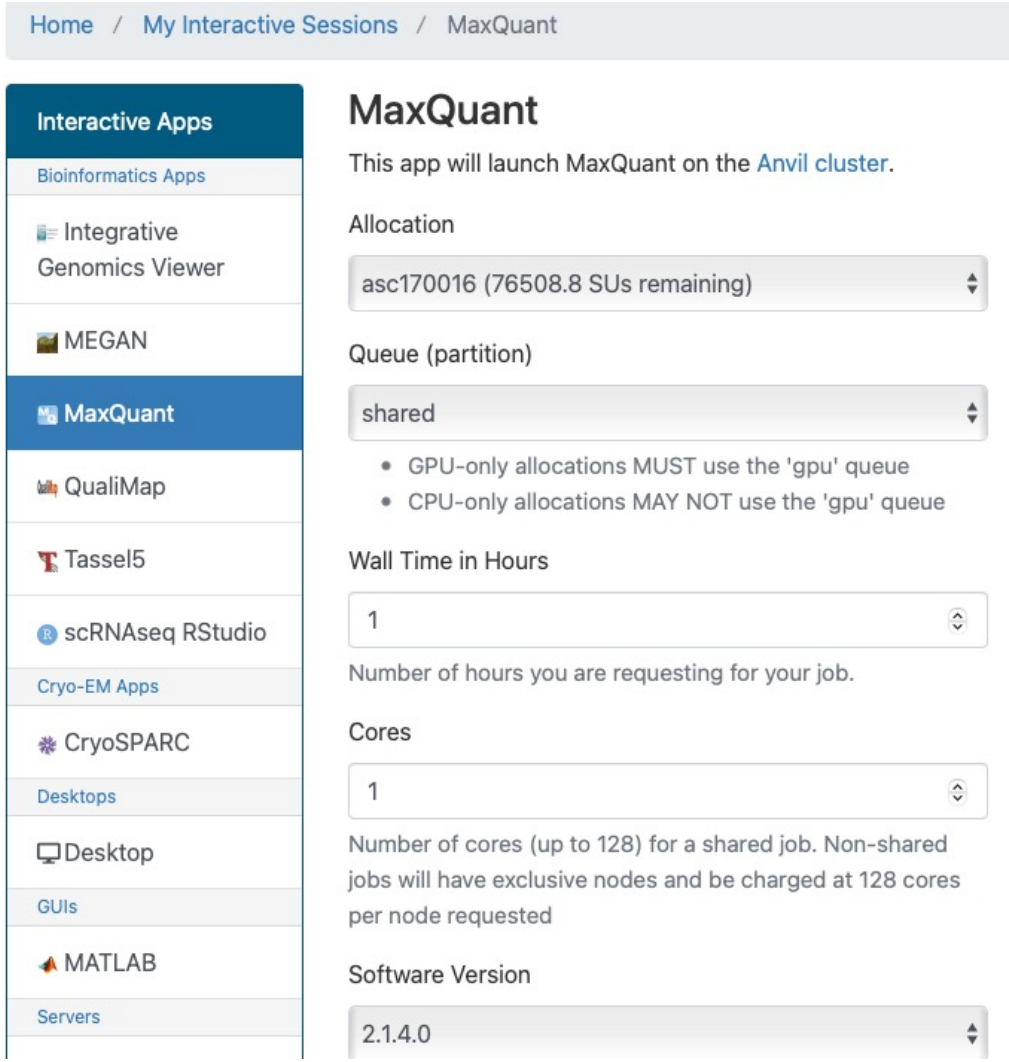
Batch Jobs: GPU

Using `#!/bin/sh -l` as shebang in the slurm job script will cause the failure of some biocontainer modules. Please use `#!/bin/bash` instead.

```
#!/bin/bash

#SBATCH -A accountName     # the queue you want to use
#SBATCH -t 20:00:00
#SBATCH -N 1
#SBATCH -n 24
#SBATCH  --gres=gpu:1
#SBATCH --job-name=parabricks
#SBATCH --mail-type=FAIL,BEGIN,END
#SBATCH --error=%x-%j-%u.err
#SBATCH --output=%x-%J-%u.out

module --force purge
module load  biocontainers parabricks

pbrun haplotypecaller \
      --ref FVZG01.1.fsa_nt \
      --in-bam output.bam \
      --out-variants variants.vcf
```

Upcoming


ondemand.anvil.rcac.purdue.edu
Dashboard - Anvil
Clusters
Interactive Apps
The Data Mine
Bioinformatics Apps
Integrative Genomics Viewer
MEGAN
MaxQuant
QualiMap
Tassel5
scRNAseq RStudio
Cryo-EM Apps
CryoSPARC
single access point for all of your HPC resource


Home / My Interactive Sessions / MaxQuant
Interactive Apps
Bioinformatics Apps
Integrative Genomics Viewer
MEGAN
MaxQuant
QualiMap
Tassel5
scRNAseq RStudio
Cryo-EM Apps
CryoSPARC
Desktops
Desktop
GUIs
MATLAB
Servers
MaxQuant
This app will launch MaxQuant on the Anvil cluster.
Allocation
asc170016 (76508.8 SUs remaining)
Queue (partition)
shared
GPU-only allocations MUST use the 'gpu' queue
CPU-only allocations MAY NOT use the 'gpu' queue
Wall Time in Hours
1
Number of hours you are requesting for your job.
Cores
1
Number of cores (up to 128) for a shared job. Non-shared jobs will have exclusive nodes and be charged at 128 cores per node requested
Software Version
2.1.4.0

Running Bioinformatics Analysis using HPC

What Comes Next?

What Comes Next?

Upcoming Seminars:

- Research Storage 101: March 10
- **Containerization Bioinformatics Applications for HPC: March 23**
- Open OnDemand 101: March 24
- Workflow Automation Tools for Many-Tasks Computing: March 30
- NLP101: March 31
- Time Series Forecasting 101: April 7

https://www.rcac.purdue.edu/news/events

THANK YOU

Feel free to reach out to zhan4429@purdue.edu with questions.

General help: rcac-help@purdue.edu

Coffee Hour: https://www.rcac.purdue.edu/coffee

Thursday -- Software compilation, Slurm workflows, Bioinformatics, Containers