

Using BioGrids for RNA-Seq on AWS and Your Laptop

James Vincent

*BioGrids Consortium
Harvard Medical School*

TMEC 304

July 25 2019



BioGrids
by Harvard Medical School

Today we will

Install software with BioGrids

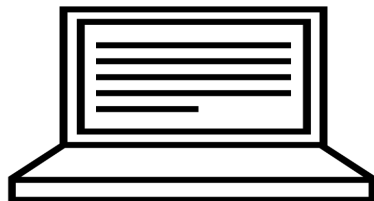
Run an RNA-Seq workflow

Replicate above on AWS and laptop



BioGrids on AWS

faithful laptop



AWS EC2 Instance



RNA-Seq Workflow

open a terminal

open a browser: biogrids.org/wiki/workshops



Subtle Things

- Capsule Environment
- `.bashrc` / `.profile` not changed
- binary installs



Avoid Time Sinks

<https://www.biostars.org/p/189261/>:

This seems to be a bug when installing fastqc using apt-get install fastqc

STARmanual.pdf

....which creates problems for STAR compilation. One option to avoid this problem is to install gcc

<http://github.gersteinlab.org/exceRpt/>

Manual Installation:

....generally not recommended ... <snip> ... instructions on how to install exceRpt and its various dependencies will [one day] be listed toward the bottom of this page.



Reproducible Research

Self Documenting

```
$ STAR --sbapp:d
```

```
Capsule:STAR using star version 2.5.3a
```

```
Version information for: /programs/i386-mac/star
```

```
Default version:                2.5.3a
In-use version:                  2.5.3a
Other available versions:        none
Overrides use this shell variable: STAR_M
```

Include in workflow:

```
STAR      --sbapp:d
samtools  --sbapp:d
```



BioGrids

by Harvard Medical School

biogrids.org help@biogrids.org

Config File

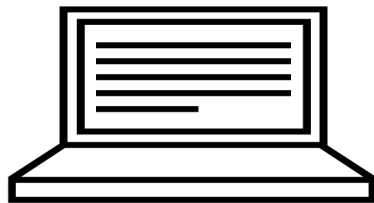
```
[installer]
site = biogrid-production
key = 70rYFBTDnmCr93VUklfbf1s3M4jdyC9bFVYHew==
user = jvincent1
```

```
[packages]
star@2.5.3a = i386-mac
samtools@1.5 = i386-mac
igv@2.4.10 = i386-mac
```



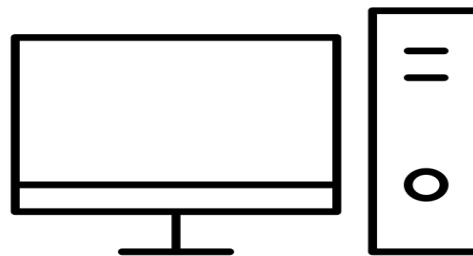
This Is Handy

faithful laptop



biogrids save mysetup.txt

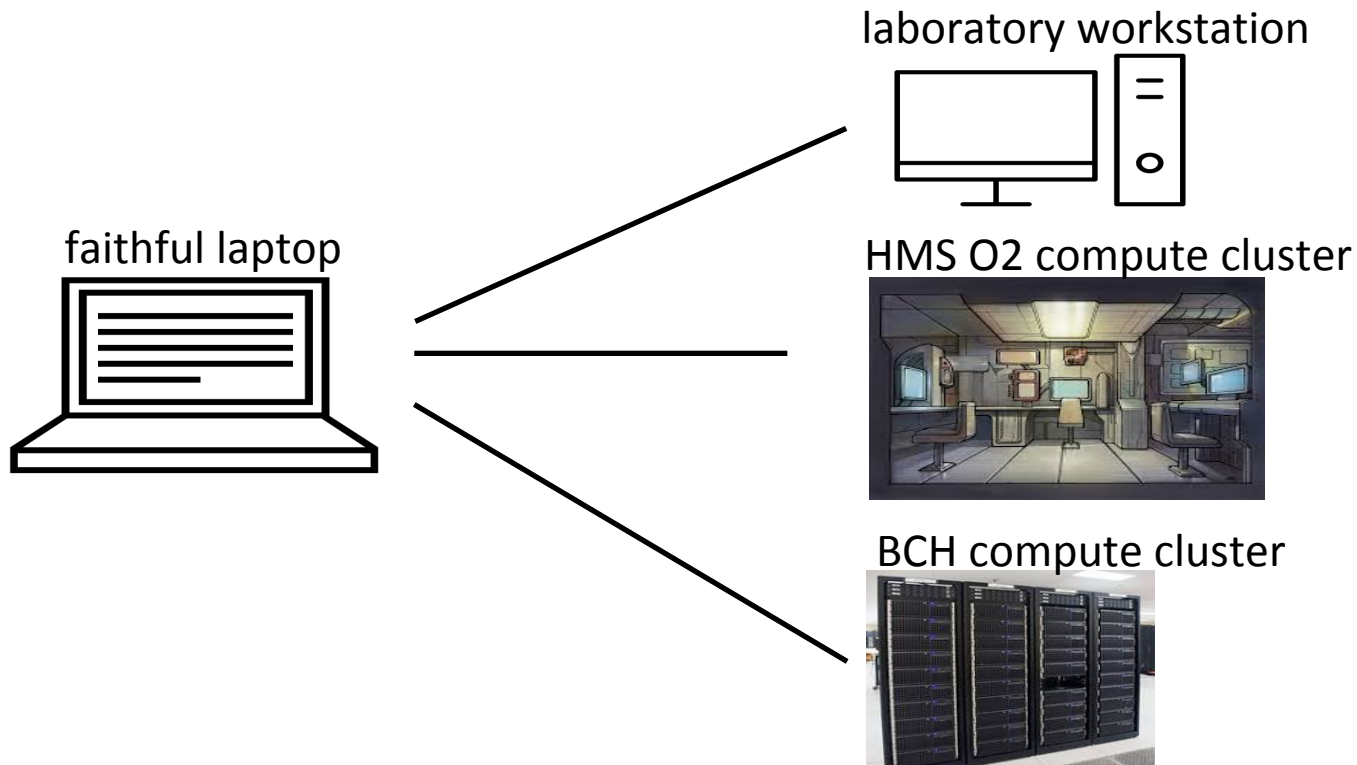
new workstation



biogrids reactivate mysetup.txt



BioGrids is Portable

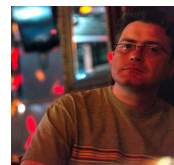


BioGrids Benefits

save time - reduce headaches
scale and share workflows
part of reproducible research



BioGrids Consortium



Personnel



SBGrid
BioGrids

Compute Infrastructure



Funding

HMS Tools and Technologies Committee



BioGrids
by Harvard Medical School

biogrids.org help@biogrids.org

Why BioGrids?

You



Cover of Nature

compile software
compile libraries
manage dependencies
manage versions
manage paths
change versions
....

learn to use software
optimize workflow
get science done



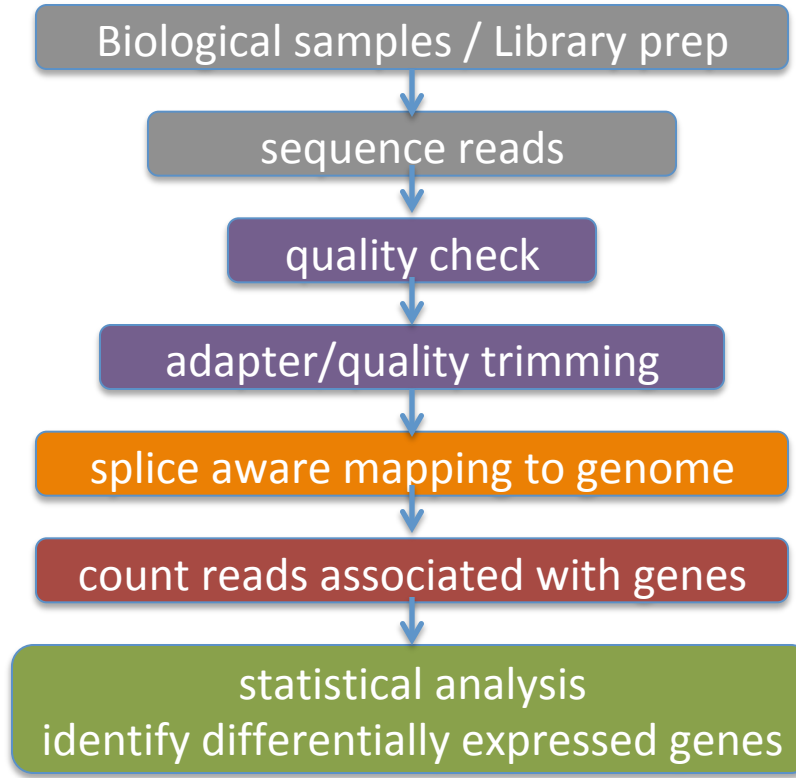
RNA-Seq Overview

Harvard Chan Bioinformatics Core (HBC)

<http://bioinformatics.sph.harvard.edu/training>



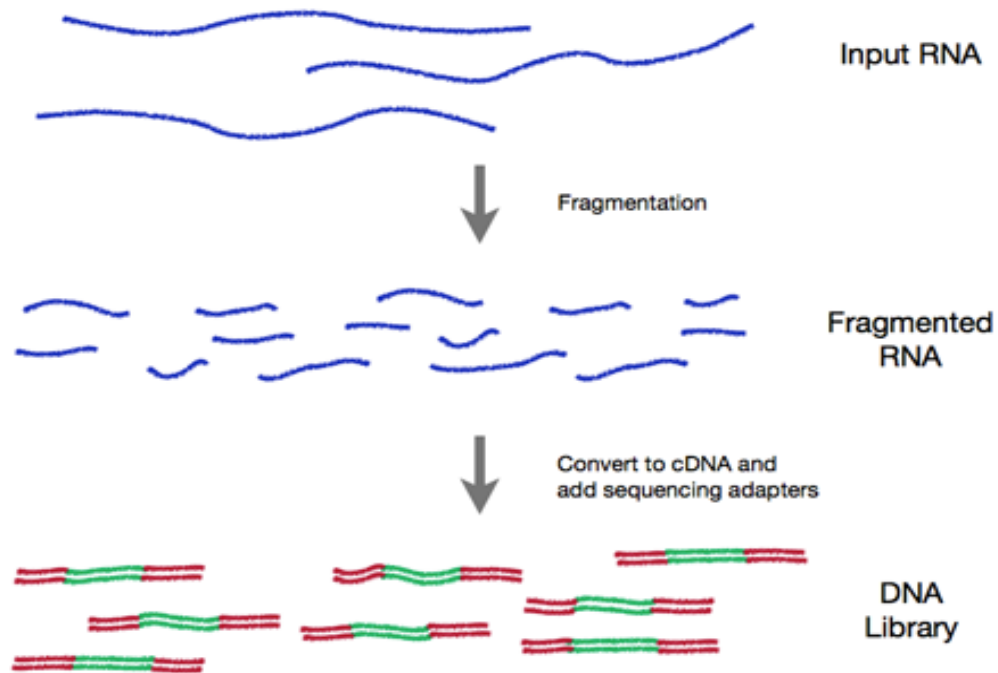
RNA-Seq Overview



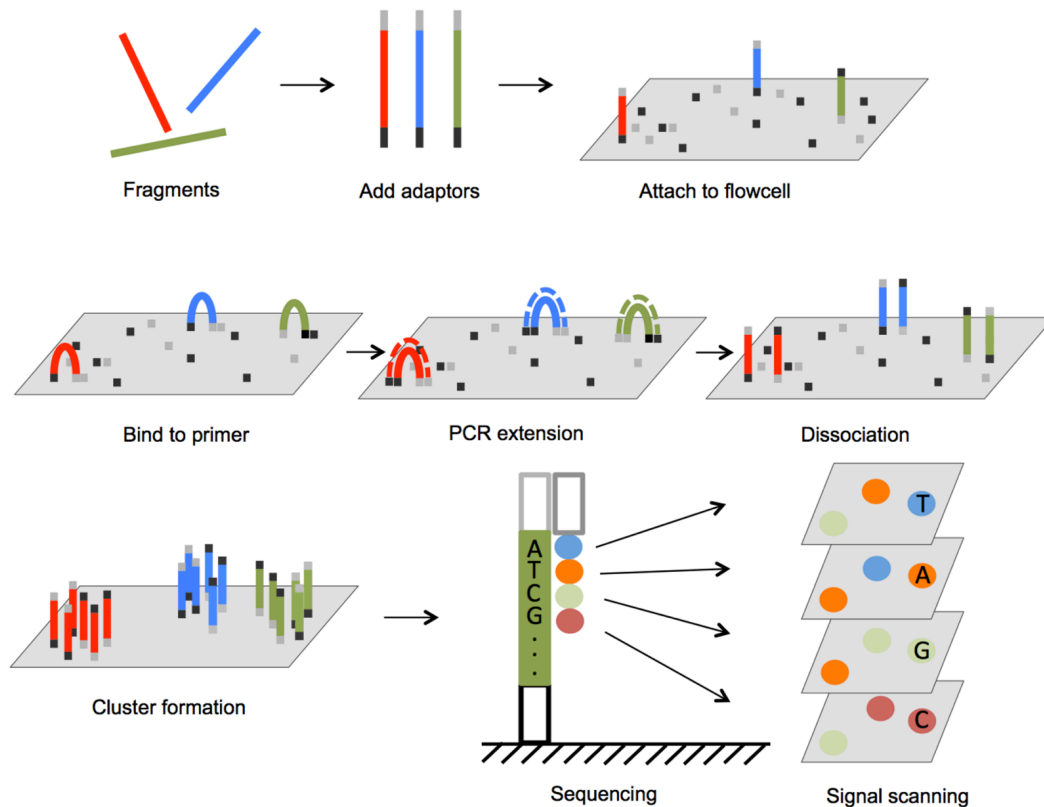
hbctraining.github.io/Intro-to-rnaseq-hpc-02



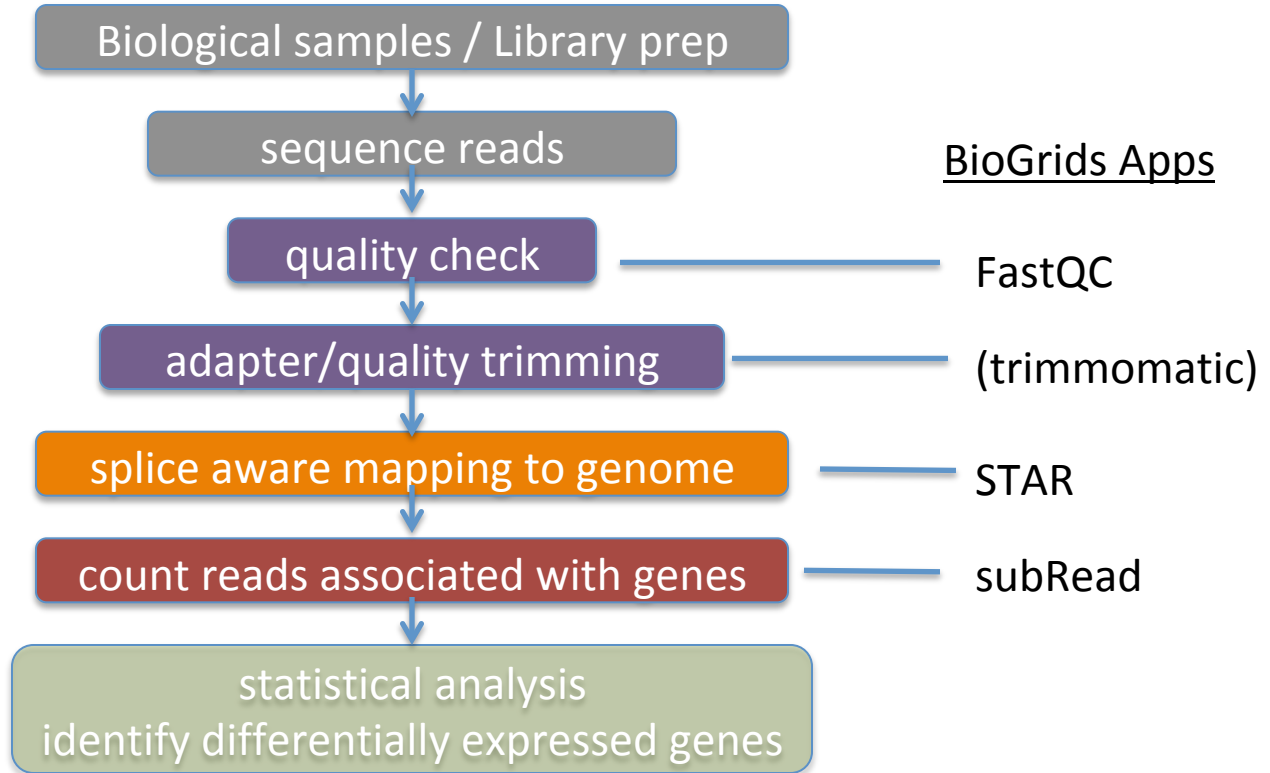
RNA Prep



Sequencing



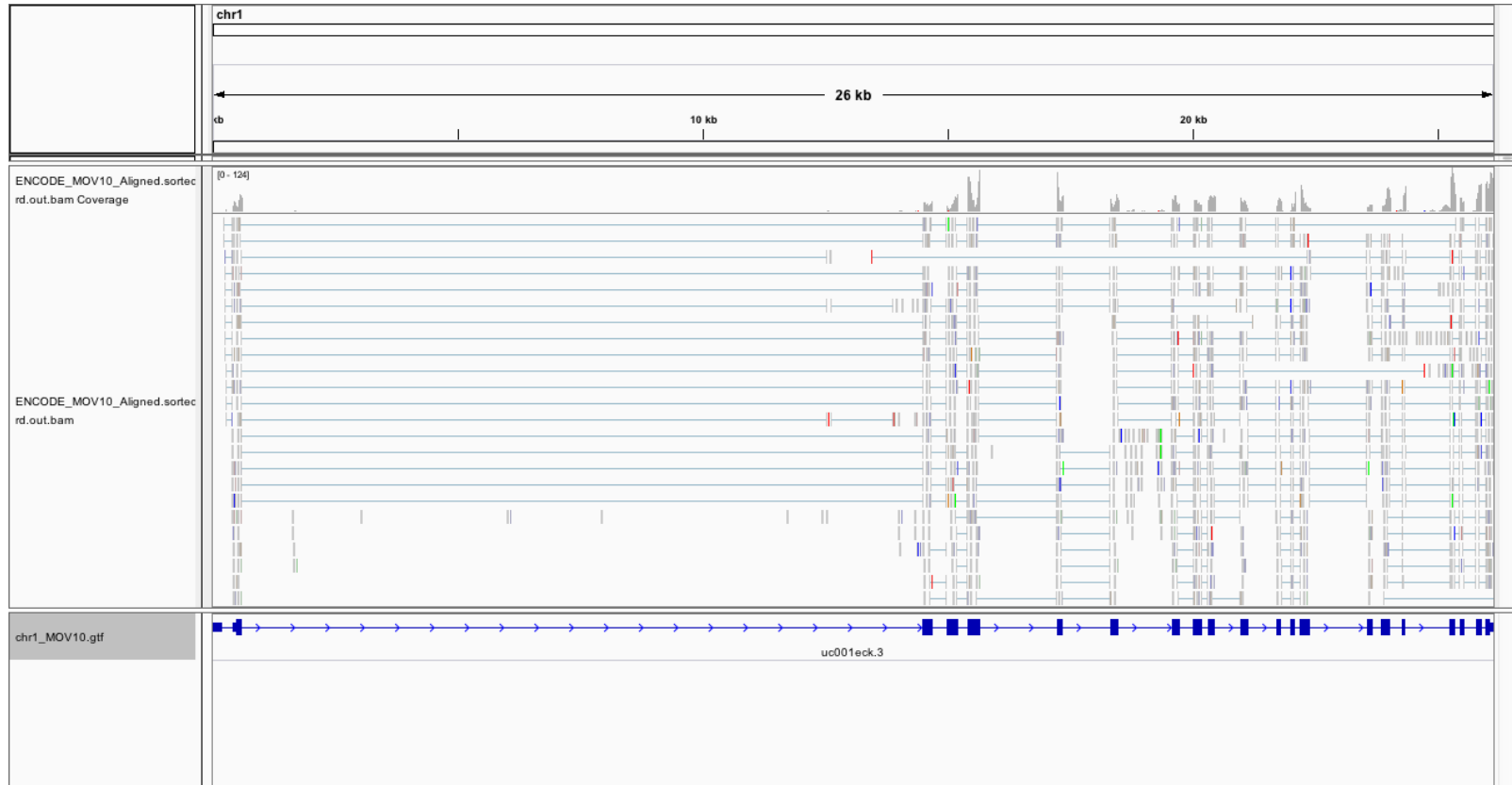
RNA-Seq Overview



hbctraining.github.io/Intro-to-rnaseq-hpc-02



Mapped Reads



Check Results

IGV

- 1: Genomes / Load Genome from File... (chr1_MOV10.fa)
- 2: File / Load from file... (.gtf file)
- 3: File / Load from file... (.bam file)



DevOps with BioGrids

workflow

bioinformatics

software
stack

BioGrids

compute
resources

*laptop
HMS O2
AWS*



AWS Hands On

<https://sbgrid.signin.aws.amazon.com/console>

username: **workshop21**

password: **Biogrids_Workshop1**



AWS - Amazon Web Services

The screenshot displays the AWS Management Console interface. At the top, the navigation bar includes the AWS logo, 'Services' with a dropdown arrow, 'Resource Groups' with a dropdown arrow, a star icon, a notification bell, and the user 'jvincent @ sbgri'. The left-hand navigation pane is titled 'EC2 Dashboard' and lists various options: 'Events', 'Tags', 'Reports', 'Limits', '+ INSTANCES', '+ IMAGES', '+ ELASTIC BLOCK STORE', '+ NETWORK & SECURITY', '+ LOAD BALANCING', and '+ AUTO SCALING'. The main content area is titled 'Resources' and shows a summary of EC2 resources in the 'US East (N. Virginia)' region. The resources are listed in two columns: Running Instances (0), Elastic IPs (0), Dedicated Hosts (0), Snapshots (3), Volumes (1), Load Balancers (0), Key Pairs (5), Security Groups (15), and Placement Groups (0). Below this list is a blue banner with the text 'Learn more about the latest in AWS Compute from AWS re:Invent by viewing the [EC2 Videos](#).' and a close button. At the bottom of the main content area, there is a section titled 'Create Instance' with the text 'To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.' and a blue button labeled 'Launch Instance' with a dropdown arrow.

aws Services ▾ Resource Groups ▾ ⭐ 🔔 jvincent @ sbgri

EC2 Dashboard

- Events
- Tags
- Reports
- Limits
- + INSTANCES
- + IMAGES
- + ELASTIC BLOCK STORE
- + NETWORK & SECURITY
- + LOAD BALANCING
- + AUTO SCALING

Resources

You are using the following Amazon EC2 resources in the US East (N. Virginia) region:

0 Running Instances	0 Elastic IPs
0 Dedicated Hosts	3 Snapshots
1 Volumes	0 Load Balancers
5 Key Pairs	15 Security Groups
0 Placement Groups	

Learn more about the latest in AWS Compute from AWS re:Invent by viewing the [EC2 Videos](#).

Create Instance

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

[Launch Instance](#) ▾



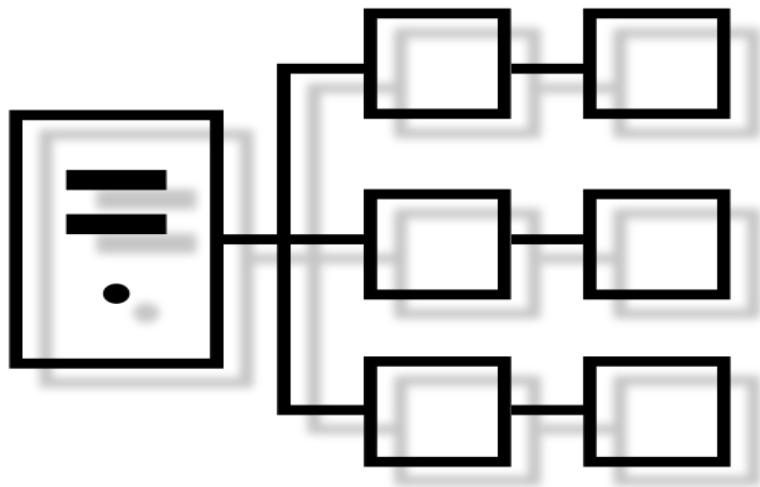
BioGrids

by Harvard Medical School

biogrids.org help@biogrids.org

AWS Parallel Cluster

aws-parallelcluster.readthedocs.io



scalable HPC cluster



help@biogrids.org

**BioGrids is funded by the
Harvard Medical School
Tools and Technologies Committee**



Additional Resources

ENCODE data files can be found here for CalTech RNA-Seq :

<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/>

Use this bam file: `wgEncodeCaltechRnaSeqK562R1x75dAlignsRep1V2`

Region of MOV10 gene: `chr1:113,214,934-113,243,900`

How to download whole genome:

- UCSC ftp site: `hgdownload.cse.ucsc.edu`
- UCSC web site: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/>
- UCSC recommends using an ftp client for large file downloads
- chr1 is only 70M



References

TRAINING

hbctraining.github.io/Intro-to-rnaseq-hpc-O2

AWS

<https://aws.amazon.com/ec2/getting-started>

ENCODE

<https://www.encodeproject.org>

IMAGES

<https://www.diagenode.com/en/categories/Library-preparation-for-RNA-seq>

<https://rnaseq.uoregon.edu>

