Using BioGrids for RNA-Seq on AWS and Your Laptop

James Vincent

BioGrids Consortium
Harvard Medical School

TMEC 304
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
[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

- faithful laptop
- laboratory workstation
- HMS O2 compute cluster
- BCH compute cluster
BioGrids Benefits

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

You

<table>
<thead>
<tr>
<th>compile software</th>
<th>learn to use software</th>
</tr>
</thead>
<tbody>
<tr>
<td>compile libraries</td>
<td>optimize workflow</td>
</tr>
<tr>
<td>manage dependencies</td>
<td>get science done</td>
</tr>
<tr>
<td>manage versions</td>
<td></td>
</tr>
<tr>
<td>manage paths</td>
<td></td>
</tr>
<tr>
<td>change versions</td>
<td></td>
</tr>
</tbody>
</table>

Cover of Nature
RNA-Seq Overview

Harvard Chan Bioinformatics Core (HBC)

http://bioinformatics.sph.harvard.edu/training
RNA-Seq Overview

- Biological samples / Library prep
- sequence reads
- quality check
- adapter/quality trimming
- splice aware mapping to genome
- count reads associated with genes
- statistical analysis
  identify differentially expressed genes

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

1. Input RNA
2. Fragmentation
3. Fragmented RNA
4. Convert to cDNA and add sequencing adapters
5. DNA Library
Sequencing

1. Fragments
2. Add adaptors
3. Attach to flowcell
4. Bind to primer
5. PCR extension
6. Dissociation
7. Cluster formation
8. Sequencing
9. Signal scanning
RNA-Seq Overview

Biological samples / Library prep

sequence reads

quality check

adapter/quality trimming

splice aware mapping to genome

count reads associated with genes

statistical analysis
identify differentially expressed genes

BioGrids Apps

FastQC

(trimmomatic)

STAR

subRead

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

EC2 Dashboard

Resources

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

- 0 Running Instances
- 0 Dedicated Hosts
- 1 Volumes
- 5 Key Pairs
- 0 Placement Groups
- 0 Elastic IPs
- 3 Snapshots
- 0 Load Balancers
- 15 Security 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
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://rnaseq.uoregon.edu