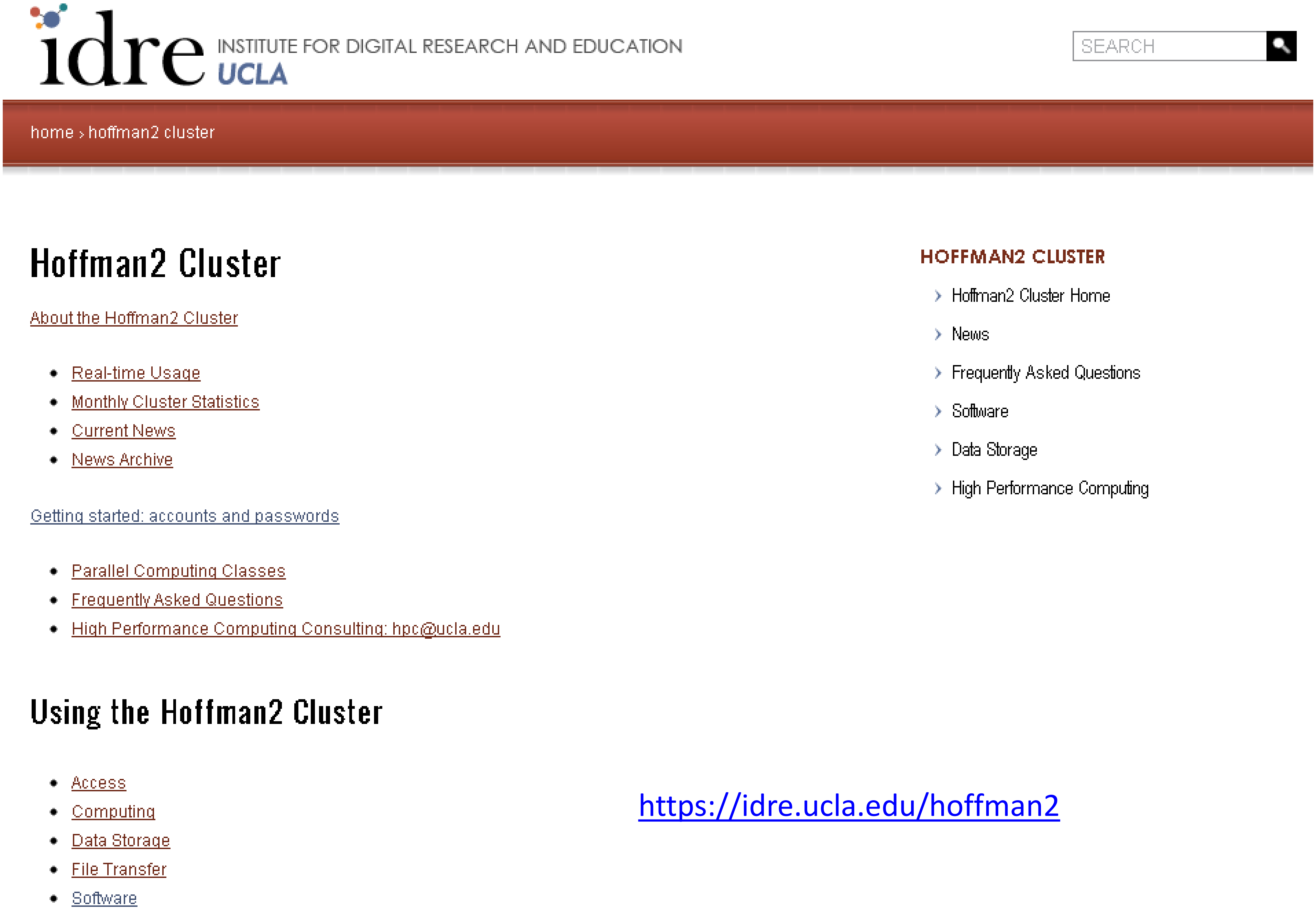
Introduction to RNAseq on Hoffman2

Richard T Wang, PhD

UCLA Collaboratory

Outline

- Background on hoffman2 cluster
- Connecting to hoffman2
- Transferring data
- Software availability
- File formats and tools for NGS data
  - raw reads, fastq, SAM/BAM
- Running an analysis using DESeq and cufflinks/tophat



idre INSTITUTE FOR DIGITAL RESEARCH AND EDUCATION UCLA

SEARCH

home › hoffman2 cluster

Hoffman2 Cluster

About the Hoffman2 Cluster

- Real-time Usage
- Monthly Cluster Statistics
- Current News
- News Archive

Getting started: accounts and passwords

- Parallel Computing Classes
- Frequently Asked Questions
- High Performance Computing Consulting: hpc@ucla.edu

Using the Hoffman2 Cluster

- Access
- Computing
- Data Storage
- File Transfer
- Software

HOFFMAN2 CLUSTER

- Hoffman2 Cluster Home
- News
- Frequently Asked Questions
- Software
- Data Storage
- High Performance Computing

https://idre.ucla.edu/hoffman2

Hoffman2 Overview

- 1300 active users
- 11,000 processors
- 3 data centers
- 20Gb home directory
- If your lab contributes nodes, you have additional storage space and different time limits

Hoffman2 Overview

- Every UCLA student has
  - 20Gb home directory
  - 24 run time
  - 2TB scratch space (7 day limit)
- If your lab contributes nodes or bought storage
  - priority on your nodes
  - additional storage space
  - different time limits (eg 14 day runs, all cores, etc)

Hoffman2

- Grid portal (not discused)
- Login node
  - There are 4 login nodes
  - Connect via SSH, putty
- hoffman2.idre.ucla.edu
- Requires UCLA sponsor
  - Can use Weihong Yan from UCLA Collaboratory
  - Limits on resources available w/o sponsor

Connecting via SSH

- To connect to hoffman2 using ssh

  ```
  ssh –l username hoffman2.idre.ucla.edu
  ```

- To use X11 forwarding
  - On a MAC
    ssh -Y login_id hoffman2.idre.ucla.edu
  - Everyone else
    ssh -X *login_id* hoffman2.idre.ucla.edu

- There are 4 login nodes, do NOT do analysis there
- Upon login, you are in your home directory which has 20Gb of space

Interactive shell

- To do analysis, you can request an interactive shell.
- This is like logging into one node of the cluster and using the machine for whatever you want.
- Default 2 hours, 1GB of RAM

  -
    ```
    qrsh
    ```
- Default 24 hours, 1GB of RAM, interactive

  -
    ```
    qrsh -l i,time=24:00:00
    ```
- Requests for large resources are unlikely to be granted
- You can also do light work on a login node and submit jobs through the login node

Transfer data onto hoffman2

- Globus online
  - For transferring large file sets
  - Eg. Your data is on a portable hard drive and you want to transfer to hoffman2
- http://www.ucgrid.org/go/go.html
- http://hpc.ucla.edu/hoffman2/file-transfer/gol.php
- SFTP
  - Secure FTP
    - Cyberduck (mac), Filezilla (mac), Bitwise SSH client(win)
- Rsync
- http://hpc.ucla.edu/hoffman2/file-transfer/file-transfer.php

Data sizes

- One lane of 100x100 PE run, 100M reads = 20Gb raw data
- If your lab is a cluster member, you will have a lab-specific cluster storage area
  - /u/home/labname/username

Learn a text editor

- For working on the command line, you need to write scripts, etc
- Learn a text editor to make your life easier
  - nano is the easiest
  - vi, vim, emacs are much more powerful but challenging to learn

Available Software

- List of software tools available
- https://idre.ucla.edu/hoffman2/software
- Many tools/versions available, but not all loaded into your environment. To load a specific tool so that you can run it
  - `module load <software name>`
  - E.g. `module load samtools`
- To view all available modules
  - `module available`

Module command

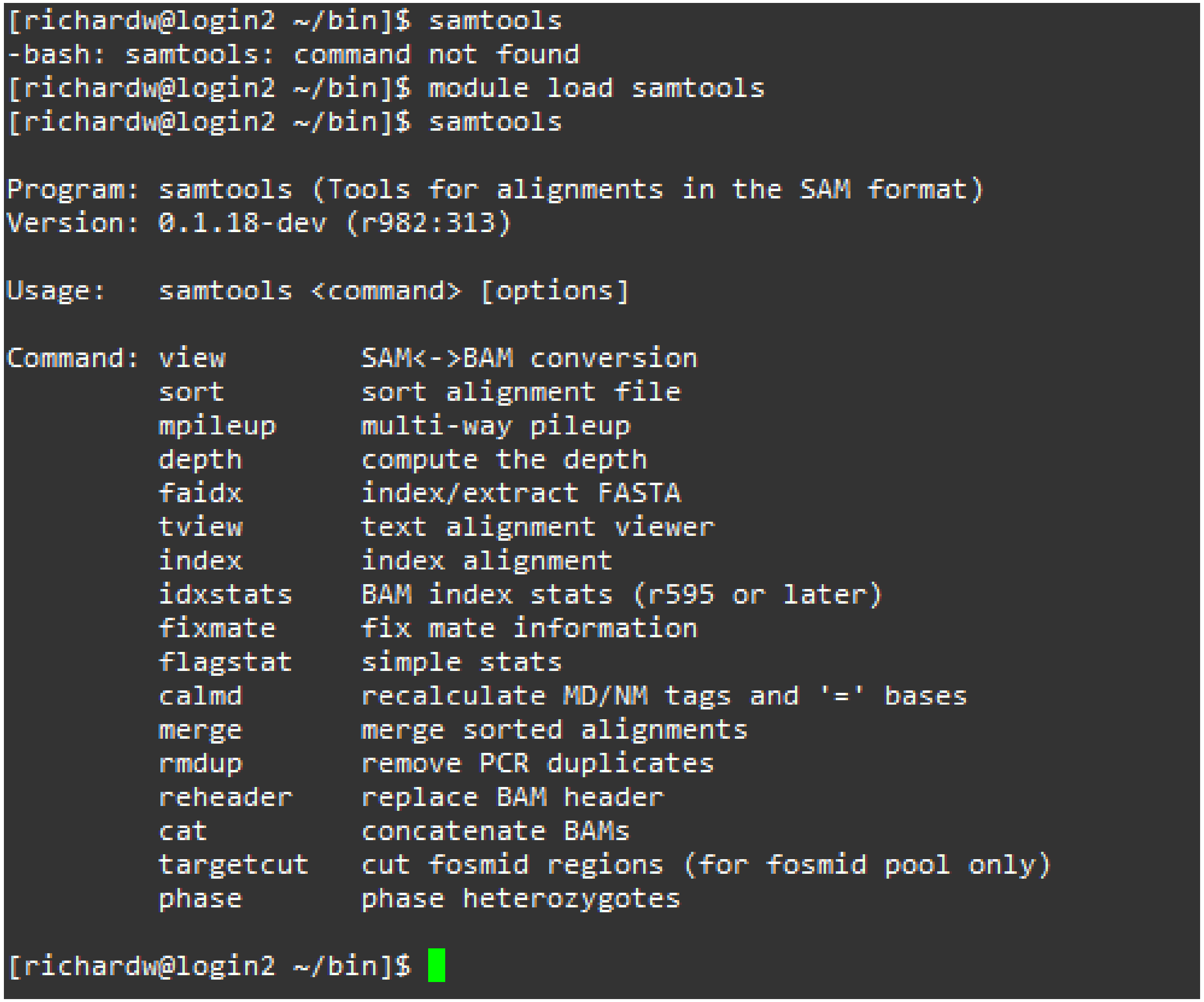
- Type
  `module load R`
  loads version 2.12.2
- If you wanted a specific version
  `module load R/2.15.1`
- To unload
  `module unload R/2.15.1`
- Example

```
module list
module load R/2.15.1
module list
module unload R.15.1
module list
```

- To run RNAseq analysis, you will need to load the correct modules before running the analysis
- Example prior to Tophat alignment

```
module load tophat/2.0.4
module load bowtie/0.12.8
module load samtools
```

Samtools



```
[richardw@login2 ~/bin]$ samtools
-bash: samtools: command not found
[richardw@login2 ~/bin]$ module load samtools
[richardw@login2 ~/bin]$ samtools

Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.18-dev (r982:313)

Usage:   samtools <command> [options]

Command: view        SAM<->BAM conversion
         sort        sort alignment file
         mpileup     multi-way pileup
         depth       compute the depth
         faidx       index/extract FASTA
         tview       text alignment viewer
         index       index alignment
         idxstats    BAM index stats (r595 or later)
         fixmate     fix mate information
         flagstat    simple stats
         calmd       recalculate MD/NM tags and '=' bases
         merge       merge sorted alignments
         rmdup       remove PCR duplicates
         reheader    replace BAM header
         cat         concatenate BAMs
         targetcut   cut fosmid regions (for fosmid pool only)
         phase       phase heterozygotes

[richardw@login2 ~/bin]$
```

Samtools

- Allows you to look at a SAM/BAM file
- More useful for BAM files which are the output of alignment in Tophat, BWA, etc

```
samtools view bamfile.bam | less -S
```

Noninstalled software

- If it’s not installed
  - Ask staff to consider installing for all users
  - Install to local user directory
    - Eg install to ~/bin directory
    - Eg. You need a particular version of software
  - You can set $PATH to include a directory so you don’t have to type
  - `/u/home/<sponsor>/<username>/<directory>`

File formats in NGS

- qseq or fastq (Illumina)
  - these are raw reads/basecalls from the sequencer
  - qseq is being phased out in favor of fastq
  - fastq 4 lines per read
    - identifier
    - read bases
    - ?
    - quality scores
  - For paired end data, you have 3 sets of files
    - read1 = first end
    - read2 = barcodes
    - read3 = second end
- BAM (aligned reads)
  - output of tophat is accepted_hits.bam
  - You can read the BAM file spec http://samtools.sourceforge.net/

For R/Bioconductor

- Need R and Bioconductor software
- Module load R
- library()
- Unfortunately, DESeq is not installed
- Set these variables. Installed packages will go here
  - R_LIBS = /u/home/eeskin/richardw/Rlibs
  - R_LIBS_USER = /u/home/eeskin/richardw/Rlibs

Linux enviromental variables

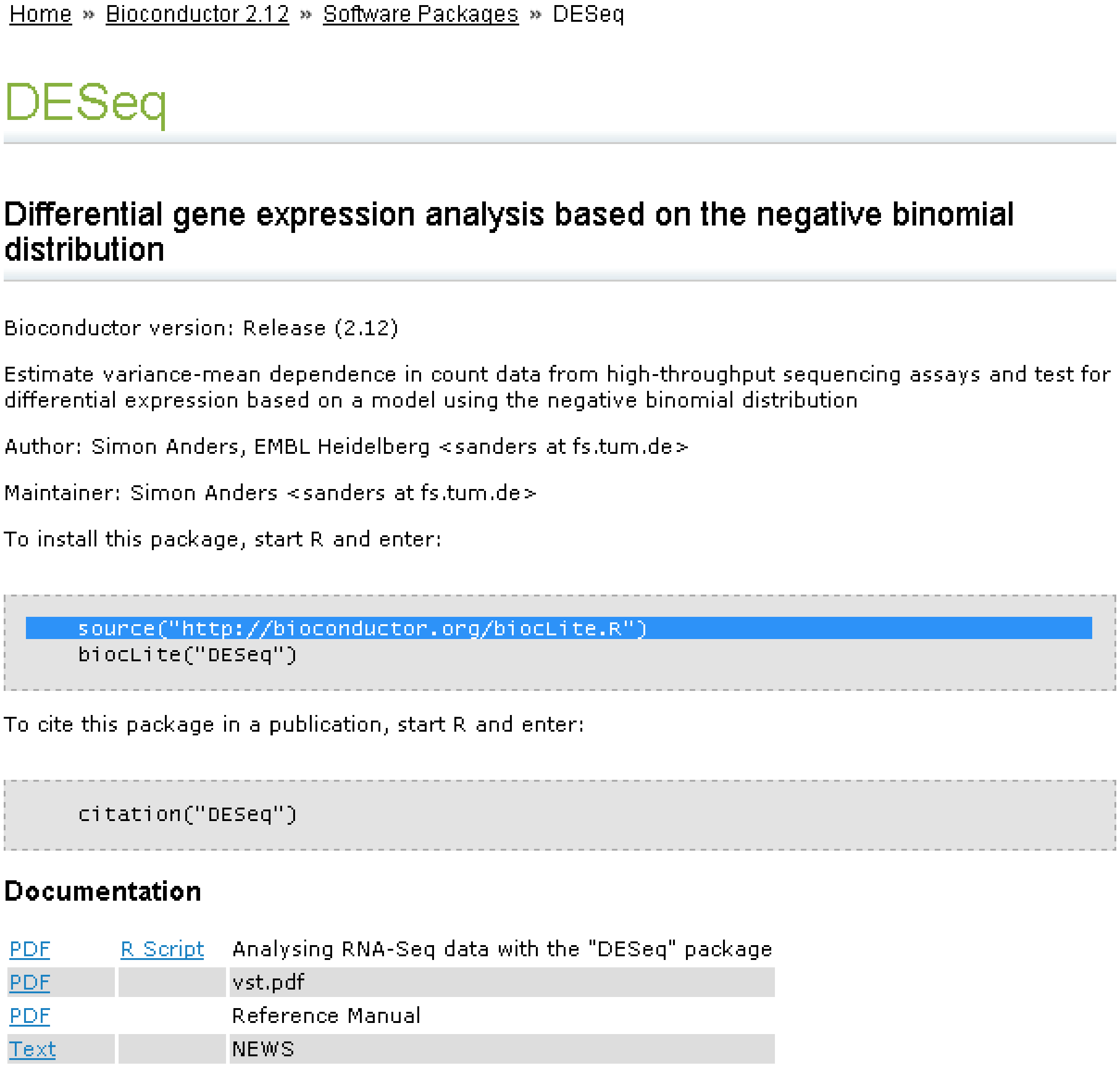
- To view variable, type
  - `env`
  - `echo $variable` (**eg** `echo $PATH`)
- To set a variable like R_LIBS, pick a location such as your home directory
  - `export R_LIBS =/u/home/eeskin/richardw/Rlibs`
  - `export R_LIBS_USER = /u/home/eeskin/richardw/Rlibs`

DESEQ ANALYSIS

Required Bioconductor packages for DESeq

- Rsamtools
- DESeq
- GenomicFeatures
- TxDb.Hsapiens.UCSC.hg19.knownGene

- Install with biocLite(<package>)



Home » Bioconductor 2.12 » Software Packages » DESeq

DESeq

Differential gene expression analysis based on the negative binomial distribution

Bioconductor version: Release (2.12)

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution

Author: Simon Anders, EMBL Heidelberg <sanders at fs.tum.de>

Maintainer: Simon Anders <sanders at fs.tum.de>

To install this package, start R and enter:

```
    source("http://bioconductor.org/biocLite.R")
    biocLite("DESeq")
```

To cite this package in a publication, start R and enter:

```
    citation("DESeq")
```

Documentation

PDF	R Script	Analysing RNA-Seq data with the "DESeq" package
PDF		vst.pdf
PDF		Reference Manual
Text		NEWS

DESeq Workflow

1. Generate reads (Illumina: 1 lane, 100M reads x 100bp = 10Gb)
2. Quality assessment
    1. ShortRead package
    2. RNASeQC (not bioconductor)
3. Read adjustments: trimming reads for adapter contamination, remove chimeric reads, ...
    1. ShortRead package
    2. Biostrings package
4. Align reads (Tophat, BWA, Bowtie, ...)
5. Importing annotation (eg gene locations)
    1. GenomicFeatures package
    2. TxDb.<species> data package
6. Count overlaps between reads and annotations (e.g., how many reads land in a gene?)
    1. GenomicRanges package
7. differential expression analysis
    1. DESeq package (DEXseq for exons)
    2. edgeR package
8. Gene set enrichment
    1. goseq package

Howto

- Organize your files by directory
  - 1-reads, 2-align, 3-qc
- Align reads using bowtie
  - Sample alignment w/ known reference

Analysis

- My directory
  - /u/home/eeskin/richardw/collaboratory/workshop3
- 1-reads: directory of raw reads (paired end)
- 2-align: directory for aligning reads into BAM files
- 3-deseq: directory to run DESeq
- gtf: contains a human GTF file
- Indexes: contains the bowtie hg19 index

Reads

- Reads will come from the sequencers as either QSEQ or FASTQ files. If QSEQ, you can convert to FASTQ
- Paired end reads come in pairs
- FASTQ
  - 4 lines per read (ID, bases, <forgot>, quality scores)

Align reads

- Use tophat to align
- On hoffman, we need to module load tophat and its underlying tools
  - `module load tophat/1.3.3`
  - `module load bowtie/0.12.8`
  - `moduel load samtools`
- We also need a
  - genome reference (hg19)
  - GTF file

Align reads

- I provided a script that runs tophat/bowtie
  - run_tophat.sh
- To submit to cluster, run qsub:
  - `qsub --cwd --V run_tophat.sh`
- Takes time to align!
- Will result in accepted_hits.bam file

Qsub commands

```
# to submit a script using qsub
qsub <script>
# list qsub jobs
qstat
# kill a job
qdel <jobid>
```

DESeq

- The default version of R is 2.12, but we will want a more recent version
  - `module load R/2.15.1`
- You can run R interactively, but for a real dataset, you’ll want to write a script and run it batchmode
  - deseq_demo.R

Bioconductor install notes

- On hoffman2, you cannot write to system files, so package install will go to your home directory
- biocLite() works for most things
- Found error with locfit (an R package) when using DESeq
  - To resolve, install/update the R package for locfit
  - install.packages(locfit)

CUFFLINKS

Tophat/cufflinks

- Analysis proceeds the same way as DESeq
- Take raw reads and align
- Perform differential expression using Cufflinks

References

- Qsub parameters
  - http://hpc.ucla.edu/hoffman2/computing/sge_qrsh.php
- Hoffman2
- R/Bioconductor