Bioinformatics for RNA sequencing analysis

Wenwen Huo, Postdoctoral Scholar, School of Medicine, Isberg lab
Rebecca Batorsky, Senior Bioinformatics Specialist
Albert Tai, Research Assistant Professor of Immunology

June 2nd 2020

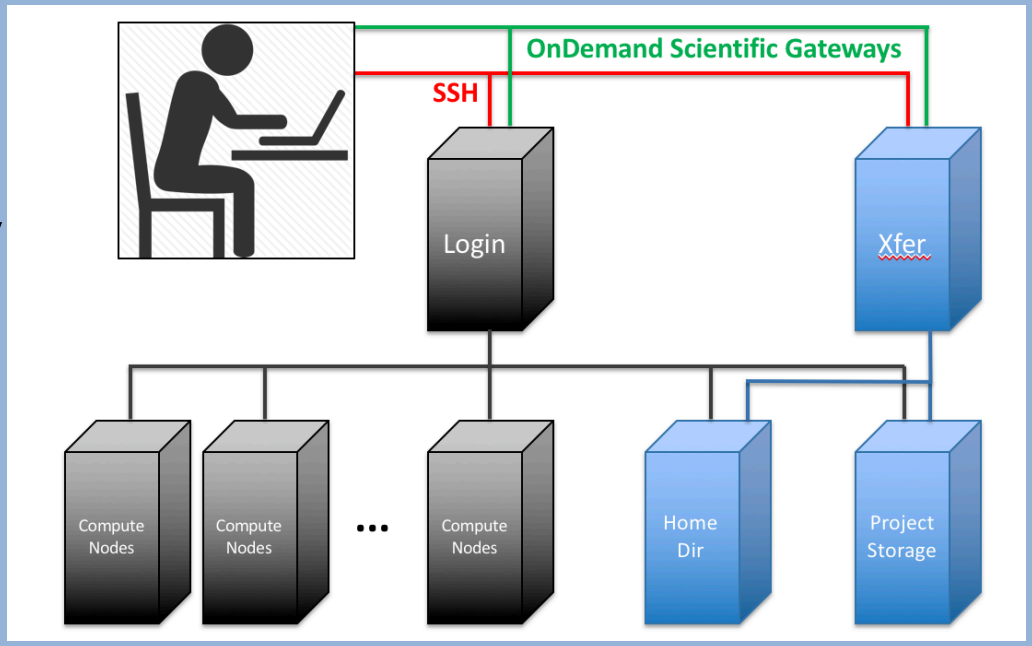
Using command line and R via OnDemand

- Structure of Tufts HPC cluster
- Command Line Basics
- R Basics

Tufts High Performance Compute Cluster

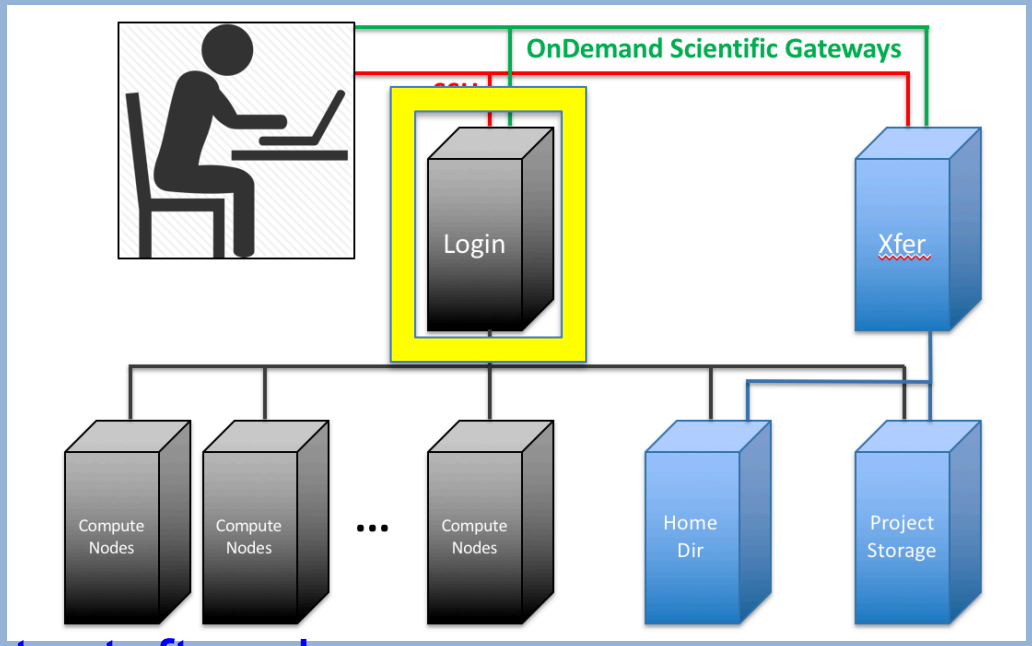
- **Cluster**

  Cluster computing is the result of connecting many local computers (nodes) together via a high speed connection to provide a single shared resource.

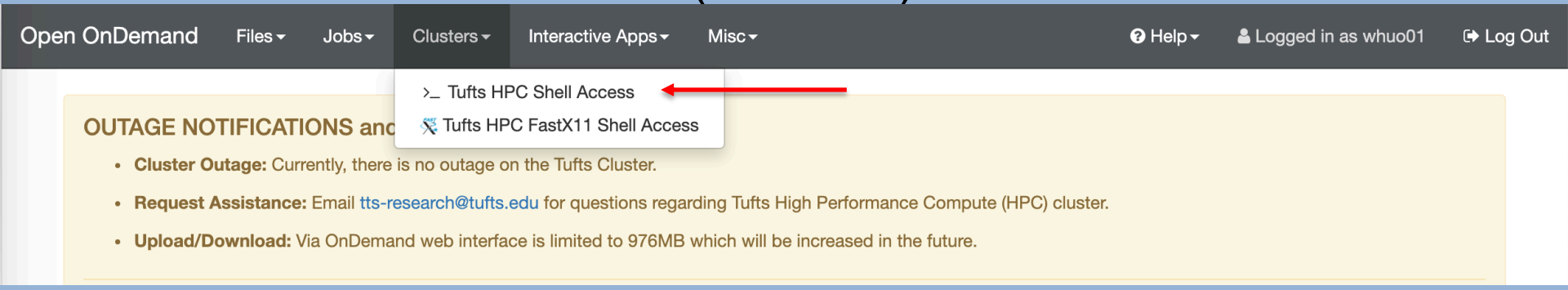


Structure of Tufts HPC Cluster

- **Login Node**

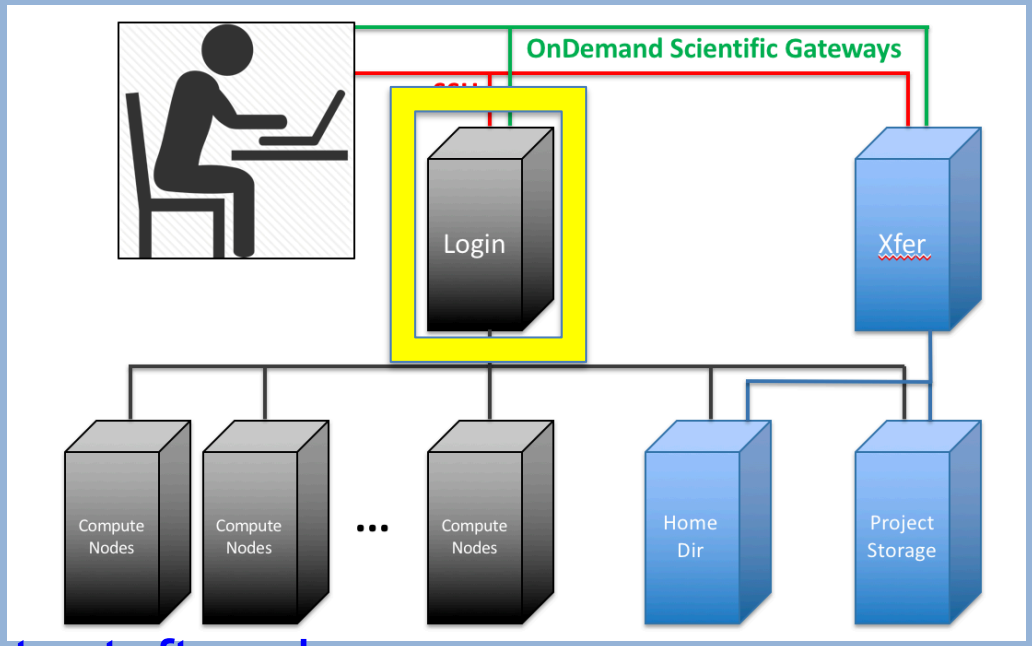


- OnDemand login
  - ➢ Go to: https://ondemand.cluster.tufts.edu
  - ➢ login with your UTLN and password
  - ➢ Clusters/Tufts HPC (FastX11) Shell Access



Structure of Tufts HPC Cluster

- **Login Node**



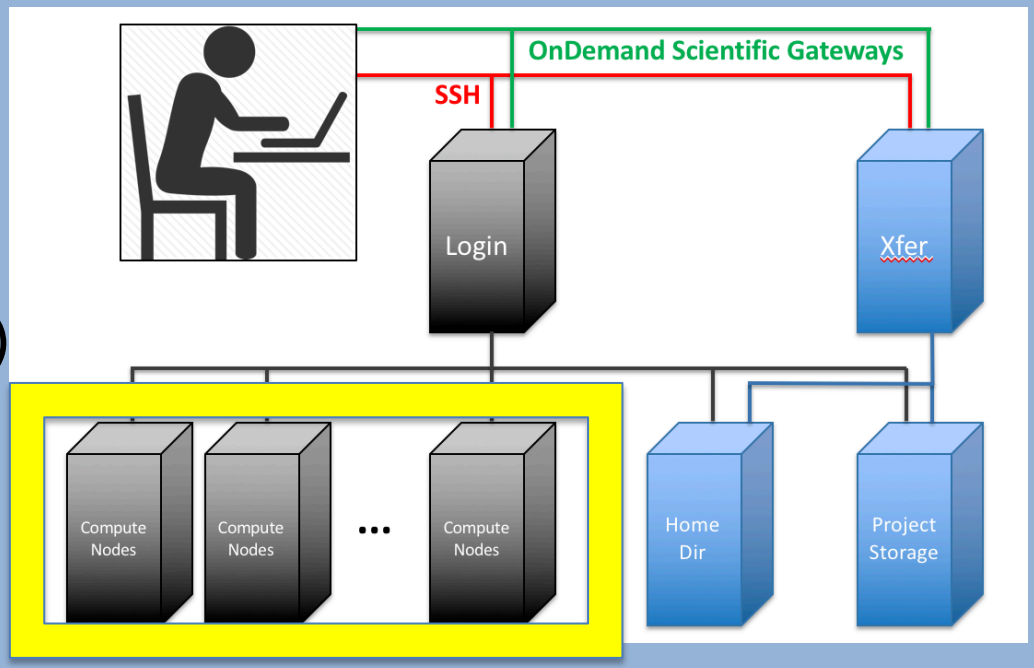
- OnDemand login
  - Go to: https://ondemand.cluster.tufts.edu
  - login with your UTLN and password
  - Clusters/Tufts HPC (FastX11) Shell Access

```
(base) [whuo01@login001 ~]$
```

Structure of Tufts HPC Cluster

- **Compute Nodes**

  Different partitions:
  - batch (3days)
  - gpu (3days)
  - interactive (4hours)
  - largemem (7days)
  - mpi (7days)
  - m4 (7days)**
  - long (21days)***



- Please **ALWAYS** run your programs/jobs on compute nodes!

time: hr:min:sec

Memory requested per node

nodes requested

```
(base) [whuo01@login001 ~]$ srun -t 3:00:00 --mem 16G -N 1 -n 4 -p preempt --reservation bioworkshop --pty bash
```

pseudo terminal runs bash shell

# of tasks

partition

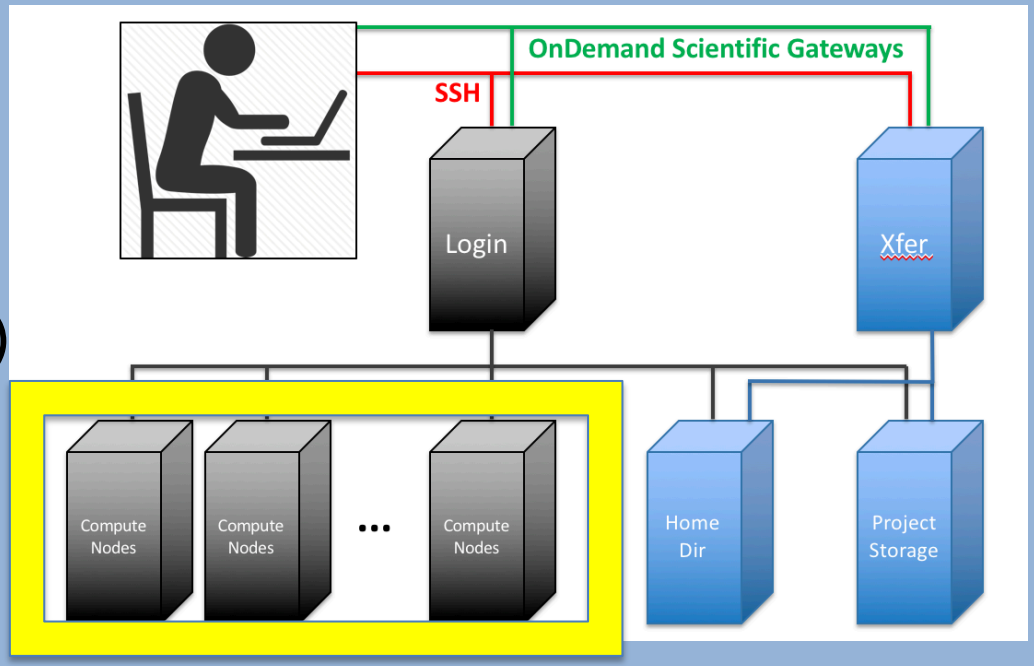
reserved spot

Structure of Tufts HPC Cluster

- **Compute Nodes**

  Different partitions:

  - batch (3days)
  - gpu (3days)
  - interactive (4hours)
  - largemem (7days)
  - mpi (7days)
  - m4 (7days)**
  - long (21days)***

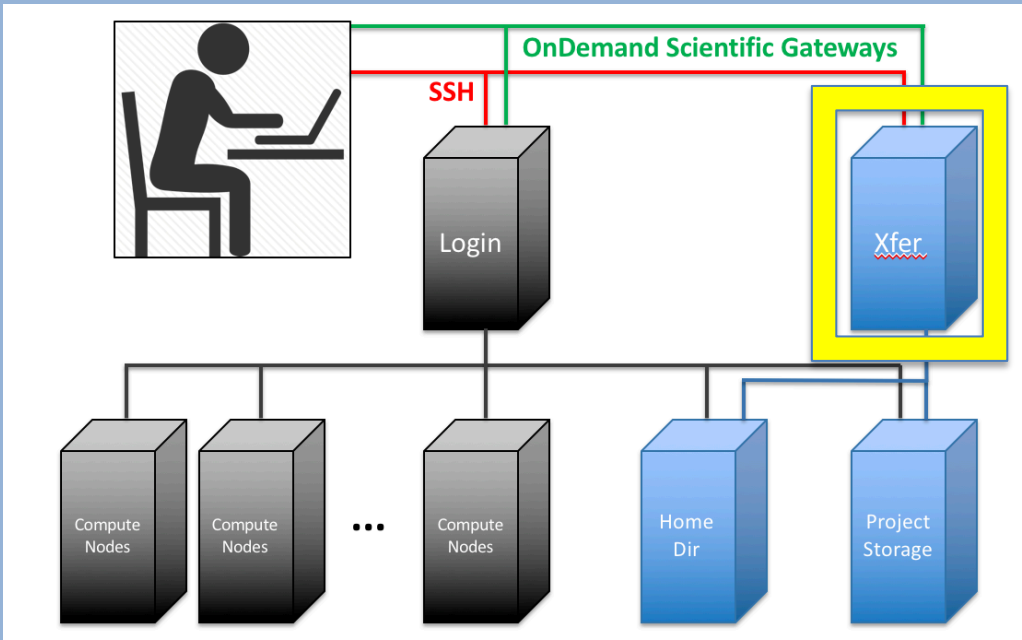


- Please **ALWAYS** run your programs/jobs on compute nodes!

```
(base) [whuo01@login001 ~]$ srun -t 3:00:00 --mem 16G -N 1 -n 4 -p preempt --reservation bioworkshop --pty bash
(base) [whuo01@pcomp41 ~]$
```

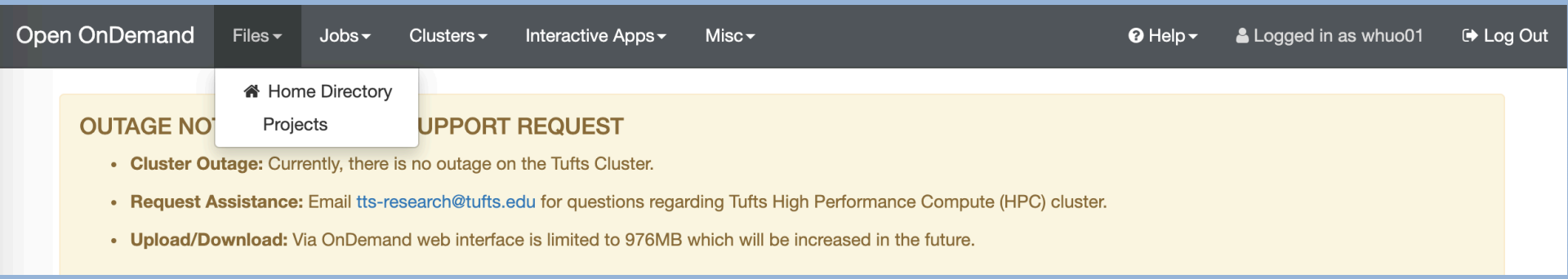
Structure of Tufts HPC Cluster

- **Xfer Node**
  - ➢ data transfer node
  - ➢ upload/download



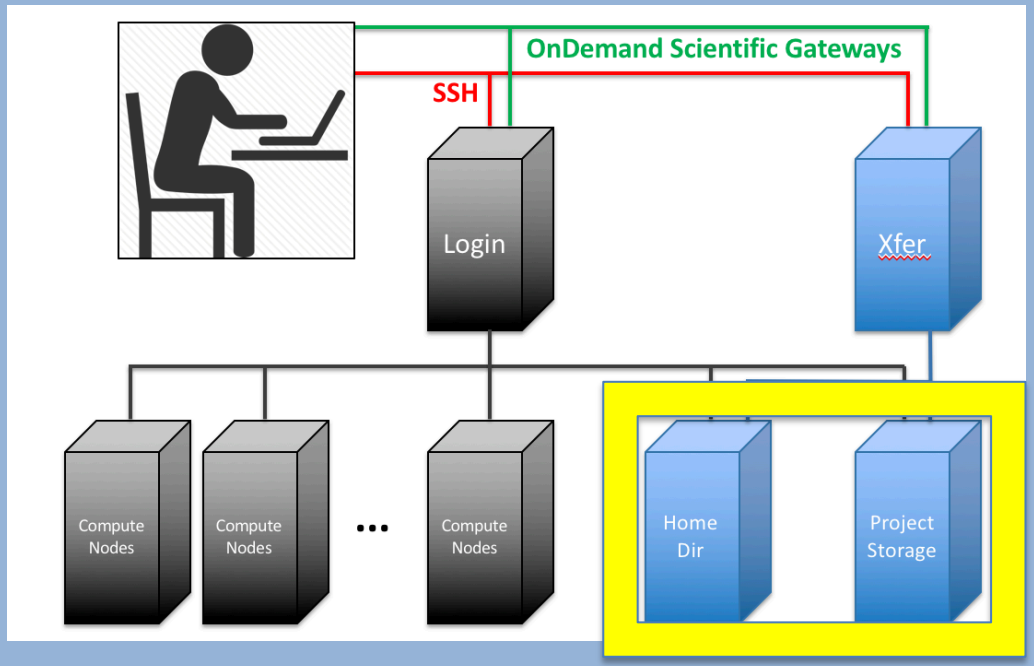
- **OnDemand**
  - ➢ Go to:

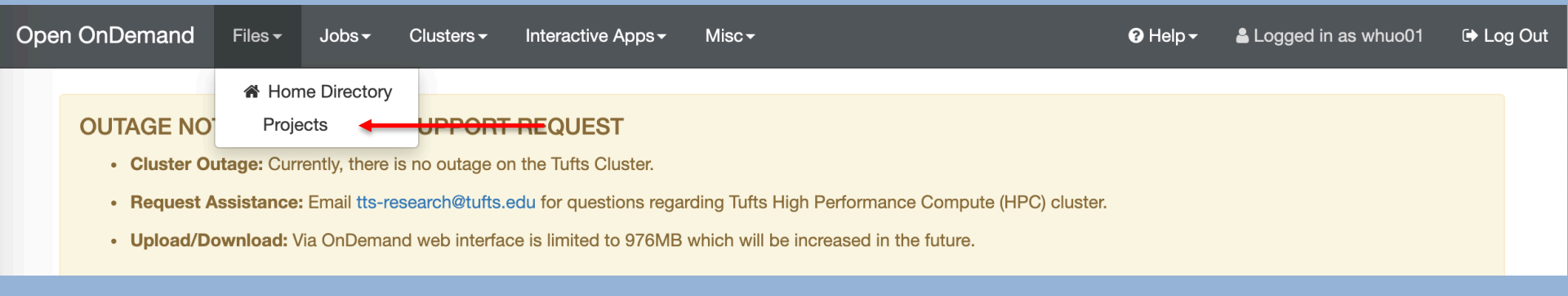
    https://ondemand.cluster.tufts.edu
  - ➢ Files (Upload/Download/View/Edit ...)



Structure of Tufts HPC Cluster

- **Home Directory**
  - /cluster/home/utln
  - 5GB (small!)
  - Visible only to you
- **Project Storage**
  - /cluster/tufts/lab_n
  - 50GB+
  - Collaboration friendly – visible to all lab members

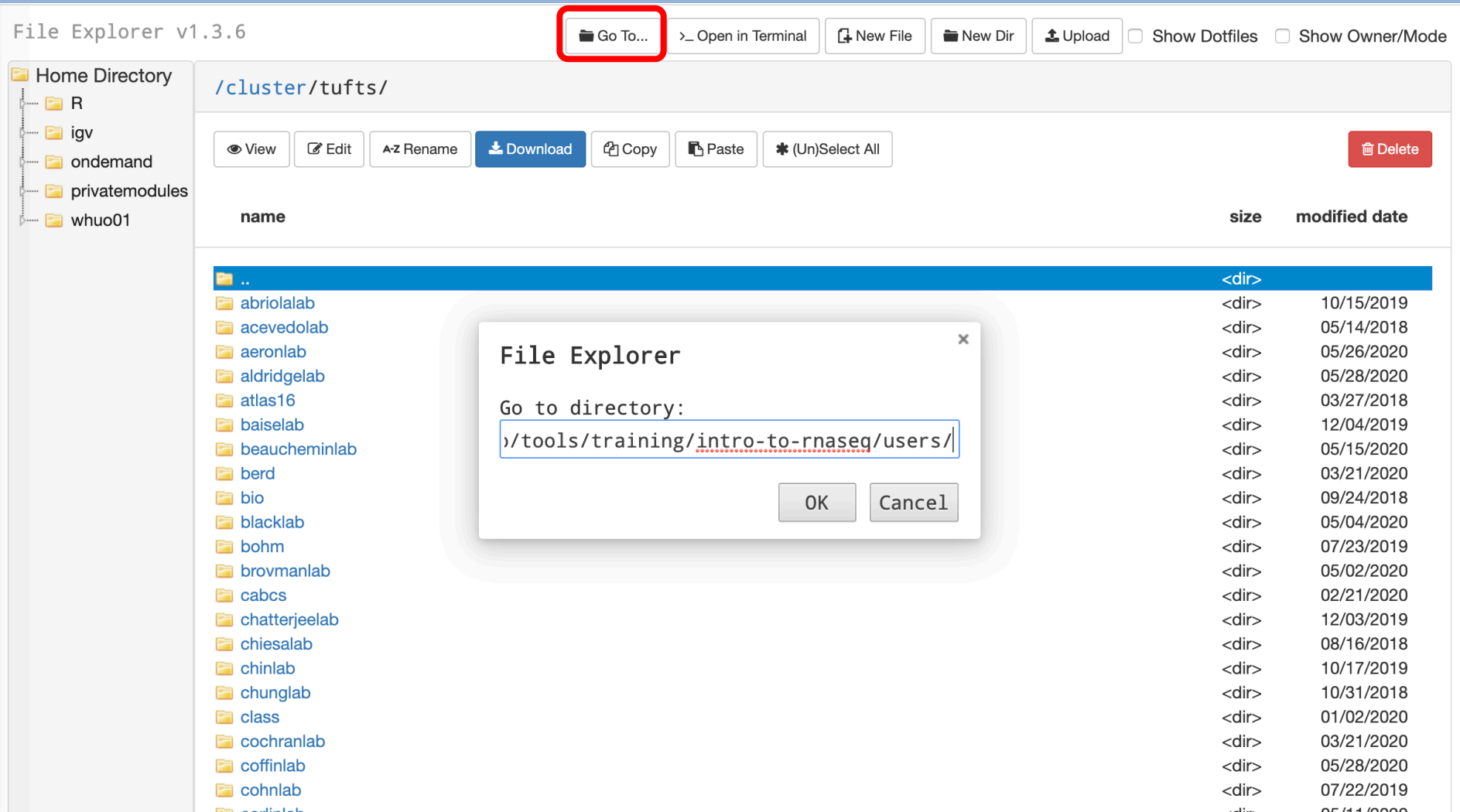




Structure of Tufts HPC Cluster

- **Project Storage**

  /cluster/tufts/bio/tools/training/intro-to-rnaseq/users/



Using command line and R via OnDemand

Command Line Basics

R Basics

1. Path: which folder you are in and how to change folder

pwd: print the current working directory
ls : list content in the folder
cd : change directory
cd .. : go back one directory
cd ../.. : go back two directories

```
(base) [whuo01@login001 ~]$ pwd
/cluster/home/whuo01
(base) [whuo01@login001 ~]$ cd /cluster/tufts/bio/tools/training/intro-to-rnaseq/users/
(base) [whuo01@login001 users]$ mkdir whuo01
(base) [whuo01@login001 users]$ cd whuo01
(base) [whuo01@login001 whuo01]$ cp /cluster/tufts/bio/tools/training/intro-to-rnaseq/intro-to-RNA-seq-May-2020.tar.gz .
(base) [whuo01@login001 whuo01]$ tar -xvzf intro-to-RNA-seq-May-2020.tar.gz
intro-to-RNA-seq/
intro-to-RNA-seq/ERP004763_info.txt
intro-to-RNA-seq/raw_data/ intro-to-RNA-seq/raw_data/sample_info.txt
intro-to-RNA-seq/raw_data/WT/
intro-to-RNA-seq/raw_data/WT/ERR458499.fastq.gz
intro-to-RNA-seq/raw_data/WT/ERR458498.fastq.gz
…
…
(base) [whuo01@login001 whuo01]$ cd intro-to-RNA-seq
(base) [whuo01@login001 intro-to-RNA-seq]$ ls
ERP004763_info.txt          raw_data       scripts
(base) [whuo01@login001 intro-to-RNA-seq]$
```

Using command line and R via OnDemand

Command Line Basics

R Basics

1. Path: which folder you are in and how to change folder

pwd: get the current directory
ls : show content in the folder
ls -ltr : show content in list format with timestamp
cd : go to a specific directory
cd .. : go back one directory
tree : list all contents in a folder

```
(base) [whuo01@login001 ~]$ cd /cluster/tufts/bio/tools/training/intro-to-rnaseq/users/whuo01/intro-to-RNA-seq
(base) [whuo01@login001 intro-to-RNA-seq]$ tree
.
├── ERP004763_info.txt
├── raw_data
│   ├── sample_info.txt
│   ├── SNF2
│   │   ├── ERR458500.fastq.gz
│   │   ├── ERR458501.fastq.gz
│   │   ├── ERR458502.fastq.gz
│   │   ├── ERR458503.fastq.gz
│   │   ├── ERR458504.fastq.gz
│   │   ├── ERR458505.fastq.gz
│   │   └── ERR458506.fastq.gz
│   └── WT
│       ├── ERR458493.fastq.gz
│       ├── ERR458494.fastq.gz
│       ├── ERR458495.fastq.gz
│       ├── ERR458496.fastq.gz
│       ├── ERR458497.fastq.gz
│       ├── ERR458498.fastq.gz
│       └── ERR458499.fastq.gz
└── scripts
    ├── fastqc.sh
    ├── featurecounts.sh
    ├── featurecount_stat.R
    ├── intro.R
    ├── mapping_percentage.R
    ├── star_align_individual.sh
    └── star_align_practice.sh
4 directories, 23 files
(base) [whuo01@login001 intro-to-RNA-seq]$
```

Using command line and R via OnDemand

Command Line Basics

R Basics

1. Path: which folder you are in and how to change folder

pwd: get the current directory
ls : show content in the folder
ls -ltr : show content in list format with timestamp
cd : go to a specific directory
cd .. : go back one directory
tree : list all contents in a folder

2. Run scripts:

sh test.sh # run a script named test.sh
sh ./test.sh # run test.sh in the current folder
sh ./subfolder/test.sh # run test.sh in the subfolder

```
(base) [whuo01@login001 intro-to-RNA-seq]$ tree
.
├── ERP004763_info.txt
├── raw_data
│   ├── sample_info.txt
│   ├── SNF2
│   │   ├── ERR458500.fastq.gz
│   │   ├── ERR458501.fastq.gz
│   │   ├── ERR458502.fastq.gz
│   │   ├── ERR458503.fastq.gz
│   │   ├── ERR458504.fastq.gz
│   │   ├── ERR458505.fastq.gz
│   │   └── ERR458506.fastq.gz
│   └── WT
│       ├── ERR458493.fastq.gz
│       ├── ERR458494.fastq.gz
│       ├── ERR458495.fastq.gz
│       ├── ERR458496.fastq.gz
│       ├── ERR458497.fastq.gz
│       ├── ERR458498.fastq.gz
│       └── ERR458499.fastq.gz
└── scripts
    ├── fastqc.sh
    ├── featurecounts.sh
    ├── featurecount_stat.R
    ├── intro.R
    ├── mapping_percentage.R
    ├── star_align_individual.sh
    └── star_align_practice.sh
4 directories, 23 files
(base) [whuo01@login001 intro-to-RNA-seq]$ sh fastqc.sh
sh: fastqc.sh: No such file or directory
```

```
(base) [whuo01@login001 intro-to-RNA-seq]$ tree
.
├── ERP004763_info.txt
├── raw_data
│   ├── sample_info.txt
│   ├── SNF2
│   │   ├── ERR458500.fastq.gz
│   │   ├── ERR458501.fastq.gz
│   │   ├── ERR458502.fastq.gz
│   │   ├── ERR458503.fastq.gz
│   │   ├── ERR458504.fastq.gz
│   │   ├── ERR458505.fastq.gz
│   │   └── ERR458506.fastq.gz
│   └── WT
│       ├── ERR458493.fastq.gz
│       ├── ERR458494.fastq.gz
│       ├── ERR458495.fastq.gz
│       ├── ERR458496.fastq.gz
│       ├── ERR458497.fastq.gz
│       ├── ERR458498.fastq.gz
│       └── ERR458499.fastq.gz
└── scripts
    ├── fastqc.sh
    ├── featurecounts.sh
    ├── featurecount_stat.R
    ├── intro.R
    ├── mapping_percentage.R
    ├── star_align_individual.sh
    └── star_align_practice.sh
4 directories, 23 files
(base) [whuo01@login001 intro-to-RNA-seq]$ sh fastqc.sh
sh: fastqc.sh: No such file or directory
(base) [whuo01@login001 intro-to-RNA-seq]$ sh scripts/fastqc.sh
Started analysis of ERR458493.fastq.gz
Approx 5% complete for ERR458493.fastq.gz
Approx 10% complete for ERR458493.fastq.gz
Approx 15% complete for ERR458493.fastq.gz
```

Using command line and R via OnDemand

Command Line Basics

R Basics

1. Path: which folder you are in and how to change folder

pwd: get the current directory
ls : show content in the folder
ls -ltr : show content in list format with timestamp
cd : go to a specific directory
cd .. : go back one directory

2. Run scripts:

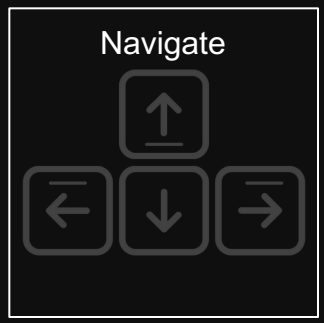
sh test.sh # run a script named test.sh
sh ./test.sh # run test.sh in the current folder
sh ./subfolder/test.sh # run test.sh in the subfolder

3. Edit scripts or other files using nano:

nano filename: enters the file named filename and creates one if filename does not exist
nano ./subfolder/filename: similar with above

```
(base) [whuo01@login001 intro-to-RNA-seq]$ cp ./scripts/star_align_practice.sh ./scripts/star_align_ERR458500.sh
(base) [whuo01@login001 intro-to-RNA-seq]$ nano ./scripts/star_align_ERR458500.sh
```

```
 GNU nano 2.0.9                File: scripts/star_align_ERR458500.sh

## Use STAR aligner to align fastq files
module load STAR/2.6.1d
mkdir -p STAR

## Fastq files to align, separated by commas for multiple lanes of a single sample
FASTQ="raw_data/WT/ERR458493.fastq.gz"

## Name the output file
OUT="WT_ERR458493"

## Defing reference genome directory
REF_DIR="/cluster/tufts/bio/data/genomes/Saccharomyces_cerevisiae/UCSC/sacCer3"

# execute STAR in the runMode "alignReads"
STAR --genomeDir ${REF_DIR}/Sequence/STAR \
--readFilesIn ${FASTQ} \
--readFilesCommand zcat \
--outFileNamePrefix STAR/${OUT}_ \
--outFilterMultimapNmax 1 \
--outSAMtype BAM SortedByCoordinate \
--runThreadN 4 \
--alignIntronMin 1 \
--alignIntronMax 2500 \
--sjdbGTFfile ${REF_DIR}/Annotation/Genes/sacCer3.gtf \
--sjdbOverhang 49

                        [ Read 28 lines ]
^G Get Help   ^O WriteOut   ^R Read File   ^Y Prev Page   ^K Cut Text   ^C Cur Pos
^X Exit       ^J Justify    ^W Where Is    ^V Next Page   ^U UnCut Text ^T To Spell
```



Exit
Ctrl + x

Using command line and R via OnDemand

Command Line Basics

R Basics

1. Path: which folder you are in and how to change folder
   - **pwd**: get the current directory
   - **ls** : show content in the folder
   - **ls -ltr** : show content in list format with timestamp
   - **cd** : go to a specific directory
   - **cd ..** : go back one directory

2. Run scripts:
   - **sh test.sh** # run a script named test.sh
   - **sh ./test.sh** # run test.sh in the current folder
   - **sh ./subfolder/test.sh** # run test.sh in the subfolder

3. Edit scripts or other files using nano:
   - **nano filename**: enters the file named filename and creates one if filename does not exist
   - **nano ./subfolder/filename**: similar with above

4. Quick view a file:
   - **head filename**: show the top 10 lines of a file
   - **cat filename**: print all the content in filename to screen
   - **cat filename | head**: chaining commands using |

```
(base) [whuo01@login001 intro-to-RNA-seq]$ cp ./scripts/star_align_practice.sh ./scripts/star_align_ERR458500.sh
(base) [whuo01@login001 intro-to-RNA-seq]$ nano ./scripts/star_align_ERR458500.sh
(base) [whuo01@login001 intro-to-RNA-seq]$ cat ./scripts/star_align_ERR458500.sh
```

```
(base) [whuo01@login001 intro-to-RNA-seq]$ cp ./scripts/star_align_practice.sh ./scripts/star_align_ERR458500.sh
(base) [whuo01@login001 intro-to-RNA-seq]$ nano ./scripts/star_align_ERR458500.sh
(base) [whuo01@login001 intro-to-RNA-seq]$ cat ./scripts/star_align_ERR458500.sh
## Use STAR aligner to align fastq files
module load STAR/2.6.1d
mkdir -p STAR

## Fastq files to align, separated by commas for multiple lanes of a single sample
FASTQ="raw_data/SNF2/ERR458500.fastq.gz"

## Name the output file
OUT="SNF2_ERR458500"

## Defing reference genome directory
REF_DIR="/cluster/tufts/bio/data/genomes/Saccharomyces_cerevisiae/UCSC/sacCer3"

# execute STAR in the runMode "alignReads"
STAR --genomeDir ${REF_DIR}/Sequence/STAR \
--readFilesIn ${FASTQ} \
--readFilesCommand zcat \
--outFileNamePrefix STAR/${OUT}_ \
--outFilterMultimapNmax 1 \
--outSAMtype BAM SortedByCoordinate \
--runThreadN 4 \
--alignIntronMin 1 \
--alignIntronMax 2500 \
--sjdbGTFfile ${REF_DIR}/Annotation/Genes/sacCer3.gtf \
--sjdbOverhang 49



(base) [whuo01@login001 intro-to-RNA-seq]$ sh ./scripts/star_align_ERR458500.sh
```

comments start with

Oriniginal: WT/ERR458493.fastq.gz

Oriniginal: WT_ERR458493

Takes ~ 5 to 10 min to finish

Using command line and R via OnDemand

Command Line Basics

R Basics

1. Run R script without using Rstudio:
   module load R/3.6.3
   Rscript sometest.R # run script that's written in R
   Rscript ./subfolder/sometest.R # run script that's in subfolder

Using command line and R via OnDemand

Command Line Basics

R Basics

1. Run R script without using Rstudio:

module load R/3.6.3

Rscript sometest.R # run script that's written in R

Rscript ./subfolder/sometest.R # run script that's in subfolder

2. Working inside Rstudio (version **3.5.0**):

a). know your working directory

getwd(): get the current working directory

setwd("/cluster/tufts/bio/tools/training/intro-to-rnaseq/users/whuo01/intro-to-RNA-seq/") : set working directory

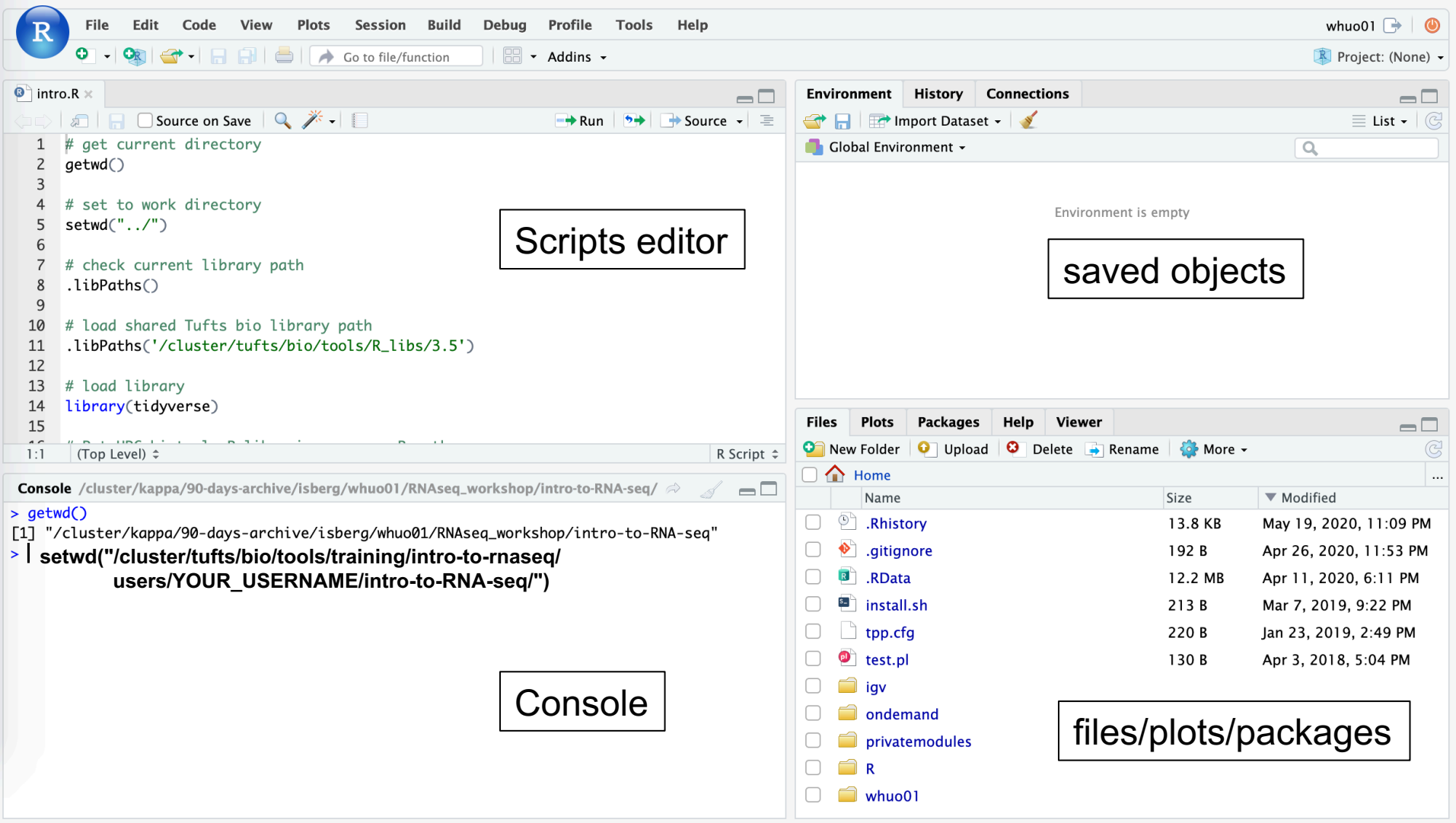
b). load library

.libPaths('/cluster/tufts/bio/tools/R_libs/3.5')

library(DESeq2)

c). assign variable

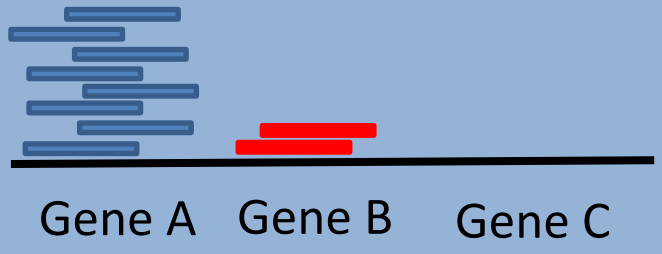
meta <- read.table("raw_data/sample_info.txt")


File Edit Code View Plots Session Build Debug Profile Tools Help
whuo01
Go to file/function
Addins
Project: (None)
intro.R
Source on Save
Run
Source
1 # get current directory
2 getwd()
3
4 # set to work directory
5 setwd("../")
6
7 # check current library path
8 .libPaths()
9
10 # load shared Tufts bio library path
11 .libPaths('/cluster/tufts/bio/tools/R_libs/3.5')
12
13 # load library
14 library(tidyverse)
15
Scripts editor
1:1 (Top Level)
R Script
Console /cluster/kappa/90-days-archive/isberg/whuo01/RNAseq_workshop/intro-to-RNA-seq/
> getwd()
[1] "/cluster/kappa/90-days-archive/isberg/whuo01/RNAseq_workshop/intro-to-RNA-seq"
> setwd("/cluster/tufts/bio/tools/training/intro-to-rnaseq/
users/YOUR_USERNAME/intro-to-RNA-seq/")
Console
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
saved objects
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Home
Name Size Modified
.Rhistory 13.8 KB May 19, 2020, 11:09 PM
.gitignore 192 B Apr 26, 2020, 11:53 PM
.RData 12.2 MB Apr 11, 2020, 6:11 PM
install.sh 213 B Mar 7, 2019, 9:22 PM
tpp.cfg 220 B Jan 23, 2019, 2:49 PM
test.pl 130 B Apr 3, 2018, 5:04 PM
igv
ondemand
privatemodules
R
whuo01
files/plots/packages

Analysis pipeline

Yeast Wild Type Phenotype

Yeast Mutant Phenotype



Process reads	FastQC
Read Alignment	STAR
Gene Quantification	featureCounts
Test for Differential Expression	R: DESeq2
Test for Functional Enrichment	R:ClusterProfiler

```
(base) [whuo01@login001 ~]$ cd /cluster/tufts/bio/tools/training/intro-to-rnaseq/users/whuo01/intro-to-RNA-seq
(base) [whuo01@login001 intro-to-RNA-seq]$ tree
.
├── ERP004763_info.txt
├── raw_data
│   ├── sample_info.txt
│   ├── SNF2
│   │   ├── ERR458500.fastq.gz
│   │   ├── ERR458501.fastq.gz
│   │   ├── ERR458502.fastq.gz
│   │   ├── ERR458503.fastq.gz
│   │   ├── ERR458504.fastq.gz
│   │   ├── ERR458505.fastq.gz
│   │   └── ERR458506.fastq.gz
│   └── WT
│       ├── ERR458493.fastq.gz
│       ├── ERR458494.fastq.gz
│       ├── ERR458495.fastq.gz
│       ├── ERR458496.fastq.gz
│       ├── ERR458497.fastq.gz
│       ├── ERR458498.fastq.gz
│       └── ERR458499.fastq.gz
└── scripts
    ├── fastqc.sh
    ├── featurecounts.sh
    ├── featurecount_stat.R
    ├── intro.R
    ├── mapping_percentage.R
    ├── star_align_individual.sh
    └── star_align_practice.sh
4 directories, 23 files
(base) [whuo01@login001 intro-to-RNA-seq]$
```

Command Line tools

R (Rstudio)

Process reads: QC

Read Alignment

Gene Quantification using Feature Count

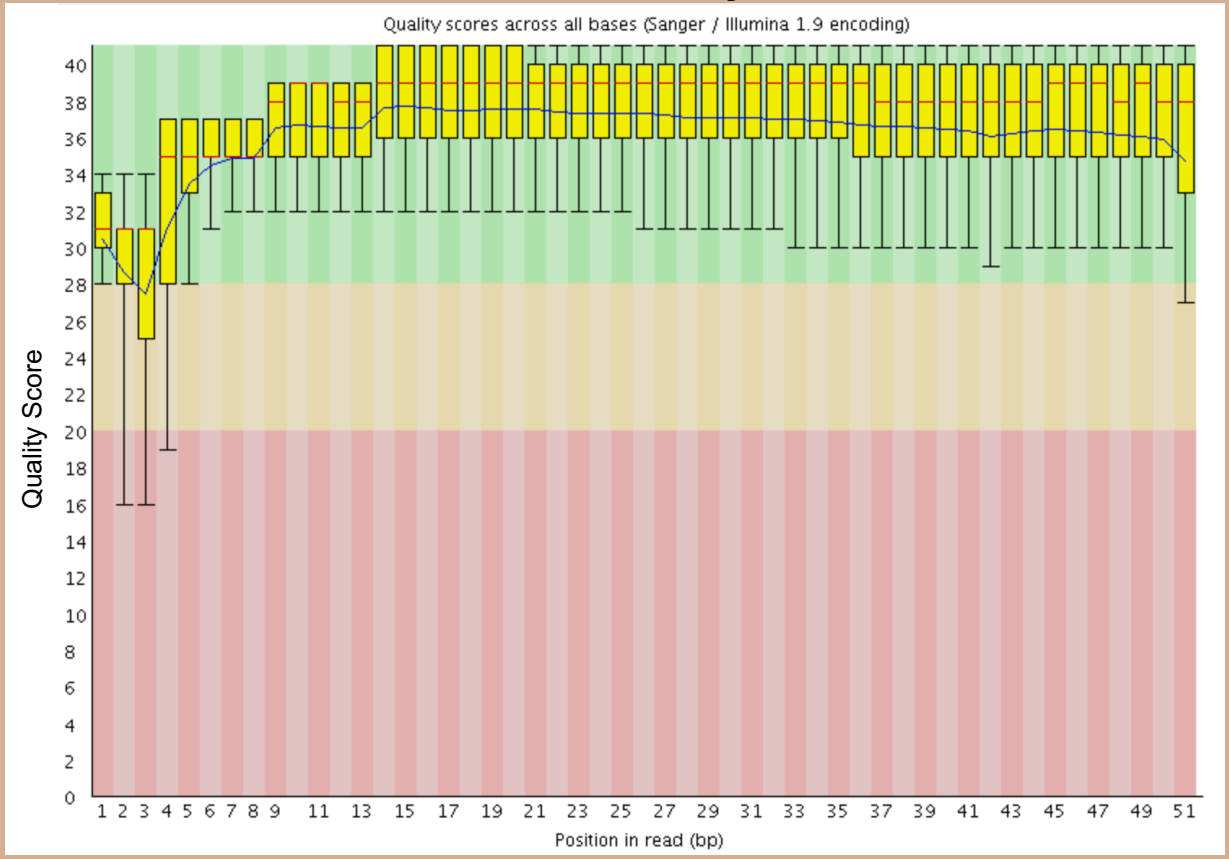
Test for Differential Expression

Test for Functional Enrichment

```
(base) [whuo01@omega001 intro-to-RNA-seq]$ module load fastqc
(base) [whuo01@omega001 intro-to-RNA-seq]$ mkdir fastqc
(base) [whuo01@omega001 intro-to-RNA-seq]$ fastqc raw_data/WT/*.fastq.gz -o fastqc --extract
```

- Process reads
- Read Alignment
- Gene Quantification
- Test for Differential Expression
- Test for Functional Enrichment

Reads Quality Control



```
(base) [whuo01@login001 ~]$ cd /cluster/tufts/bio/tools/training/intro-to-rnaseq/users/whuo01/intro-to-RNA-seq
(base) [whuo01@login001 intro-to-RNA-seq]$ tree
.
├── ERP004763_info.txt
├── raw_data
│   ├── sample_info.txt
│   ├── SNF2
│   │   ├── ERR458500.fastq.gz
│   │   ├── ERR458501.fastq.gz
│   │   ├── ERR458502.fastq.gz
│   │   ├── ERR458503.fastq.gz
│   │   ├── ERR458504.fastq.gz
│   │   ├── ERR458505.fastq.gz
│   │   └── ERR458506.fastq.gz
│   └── WT
│       ├── ERR458493.fastq.gz
│       ├── ERR458494.fastq.gz
│       ├── ERR458495.fastq.gz
│       ├── ERR458496.fastq.gz
│       ├── ERR458497.fastq.gz
│       ├── ERR458498.fastq.gz
│       └── ERR458499.fastq.gz
└── scripts
    ├── fastqc.sh
    ├── featurecounts.sh
    ├── featurecount_stat.R
    ├── intro.R
    ├── mapping_percentage.R
    ├── star_align_individual.sh
    └── star_align_practice.sh
4 directories, 23 files
(base) [whuo01@login001 intro-to-RNA-seq]$
```

Command Line tools

R (Rstudio)

Process reads: QC

Read Alignment

Gene Quantification using Feature Count

Test for Differential Expression

Test for Functional Enrichment

Practice: Process the rest of reads before moving on to the next step

```
(base) [whuo01@login001 ~]$ cd /cluster/tufts/bio/tools/training/intro-to-rnaseq/users/whuo01/intro-to-RNA-seq
(base) [whuo01@login001 intro-to-RNA-seq]$ tree
.
├── ERP004763_info.txt
├── raw_data
│   ├── sample_info.txt
│   ├── SNF2
│   │   ├── ERR458500.fastq.gz
│   │   ├── ERR458501.fastq.gz
│   │   ├── ERR458502.fastq.gz
│   │   ├── ERR458503.fastq.gz
│   │   ├── ERR458504.fastq.gz
│   │   ├── ERR458505.fastq.gz
│   │   └── ERR458506.fastq.gz
│   └── WT_1
│       ├── ERR458493.fastq.gz
│       ├── ERR458494.fastq.gz
│       ├── ERR458495.fastq.gz
│       ├── ERR458496.fastq.gz
│       ├── ERR458497.fastq.gz
│       ├── ERR458498.fastq.gz
│       └── ERR458499.fastq.gz
└── scripts
    ├── fastqc.sh
    ├── featurecounts.sh
    ├── featurecount_stat.R
    ├── intro.R
    ├── mapping_percentage.R
    ├── star_align_individual.sh
    └── star_align_practice.sh
4 directories, 23 files
(base) [whuo01@login001 intro-to-RNA-seq]$ sh ./scripts/star_align_individual.sh
```

Command Line tools

R (Rstudio)

Process reads: QC

Read Alignment

Gene Quantification using Feature Count

Test for Differential Expression

Test for Functional Enrichment

Practice: Process the rest of reads before moving on to the next step

```
(base) [whuo01@pcomp41 intro-to-RNA-seq]$ sh scripts/featurecounts.sh
(base) [whuo01@pcomp41 intro-to-RNA-seq]$ head featurecounts/featurecounts_results.txt
```

- Process reads
- Read Alignment
- Gene Quantification
- Test for Differential Expression
- Test for Functional Enrichment

Feature counts table

```
(base) [whuo01@login001 featurecounts]$ head featurecounts_results.txt
# Program:featureCounts v1.6.3; Command:"featureCounts" "-a" "/cluster/tufts/bio/data/genomes/Saccharomyces_cerevisiae/UCSC/sacCer3/Annot
ation/Genes/sacCer3.gtf" "-o" "featurecounts/featurecounts_results.txt" "STAR/SNF2_ERR458500_Aligned.sortedByCoord.out.bam" "STAR/SNF2_ER
R458501_Aligned.sortedByCoord.out.bam" "STAR/SNF2_ERR458502_Aligned.sortedByCoord.out.bam" "STAR/SNF2_ERR458503_Aligned.sortedByCoord.out
.bam" "STAR/SNF2_ERR458504_Aligned.sortedByCoord.out.bam" "STAR/SNF2_ERR458505_Aligned.sortedByCoord.out.bam" "STAR/SNF2_ERR458506_Aligne
d.sortedByCoord.out.bam" "STAR/WT_1_ERR458493_Aligned.sortedByCoord.out.bam" "STAR/WT_1_ERR458494_Aligned.sortedByCoord.out.bam" "STAR/WT
_1_ERR458495_Aligned.sortedByCoord.out.bam" "STAR/WT_1_ERR458496_Aligned.sortedByCoord.out.bam" "STAR/WT_1_ERR458497_Aligned.sortedByCoor
d.out.bam" "STAR/WT_1_ERR458498_Aligned.sortedByCoord.out.bam" "STAR/WT_1_ERR458499_Aligned.sortedByCoord.out.bam"
Geneid  Chr     Start   End     Strand  Length  STAR/SNF2_ERR458500_Aligned.sortedByCoord.out.bam       STAR/SNF2_ERR458501_Aligned.sorte
dByCoord.out.bam        STAR/SNF2_ERR458502_Aligned.sortedByCoord.out.bam       STAR/SNF2_ERR458503_Aligned.sortedByCoord.out.bam       S
TAR/SNF2_ERR458504_Aligned.sortedByCoord.out.bam        STAR/SNF2_ERR458505_Aligned.sortedByCoord.out.bam       STAR/SNF2_ERR458506_Align
ed.sortedByCoord.out.bam        STAR/WT_1_ERR458493_Aligned.sortedByCoord.out.bam       STAR/WT_1_ERR458494_Aligned.sortedByCoord.out.bam
STAR/WT_1_ERR458495_Aligned.sortedByCoord.out.bam       STAR/WT_1_ERR458496_Aligned.sortedByCoord.out.bam       STAR/WT_1_ERR458497_Align
ed.sortedByCoord.out.bam        STAR/WT_1_ERR458498_Aligned.sortedByCoord.out.bam       STAR/WT_1_ERR458499_Aligned.sortedByCoord.out.bam
YAL069W chrI    335     649     +       315     0       0       0       0       0       0       0       0       0       0       0       0
0       0
YAL068W-A       chrI    538     792     +       255     0       0       0       0       0       0       0       0       0       0       0
0       0       0
YAL068C chrI    1807    2169    -       363     0       0       0       1       0       0       1       0       0       0       0       0
0       0
YAL067W-A       chrI    2480    2707    +       228     0       0       0       0       0       0       0       0       0       0       0
0       0       0
YAL067C chrI    7235    9016    -       1782    18      21      10      12      12      11      19      1       3       1       2       1
0       4
YAL066W chrI    10091   10399   +       309     0       1       0       0       0       1       0       0       0       0       0       0
0       0
YAL065C chrI    11565   11951   -       387     2       0       0       0       2       0       1       1       2       0       4       0
0       3
YAL064W-B       chrI    12046   12426   +       381     3       3       1       1       0       3       2       1       1       0       2
1       3       1
```

Clean sample name and import to R

Sample name

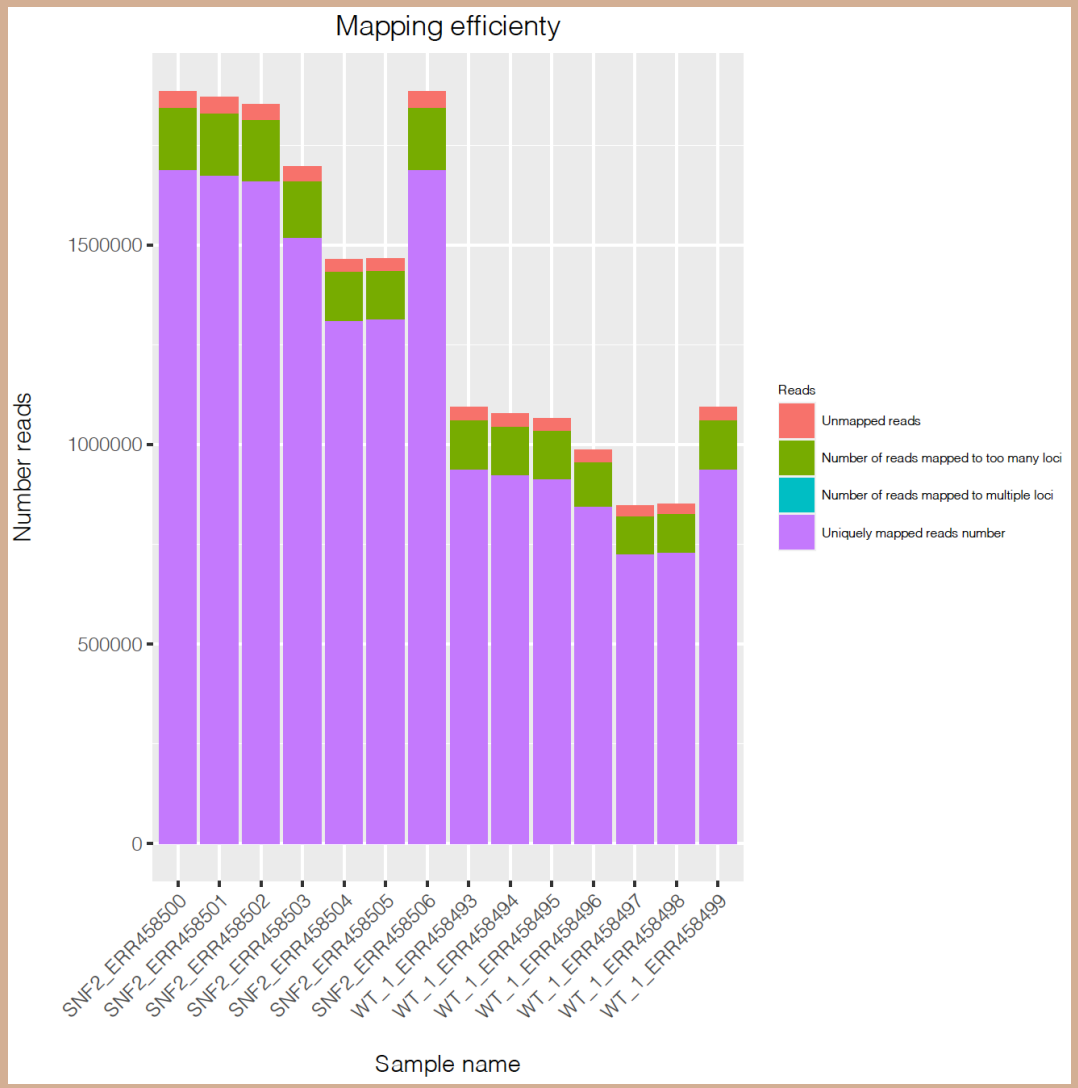
Transcript name

	SNF2_ERR458500	SNF2_ERR458501	SNF2_ERR458502
YAL069W	0	0	0
YAL068W-A	0	0	0
YAL068C	0	0	0
YAL067W-A	0	0	0
YAL067C	18	21	10
YAL066W	0	1	0
YAL065C	2	0	0
YAL064W-B	3	3	1
YAL064C-A	12	7	8
YAL064W	3	3	2
YAL063C-A	2	0	3
YAL063C	35	38	46
YAL062W	127	136	113
YAL061W	124	100	125
YAL060W	300	309	323

Optional: Read alignment QC

- Process reads
- Read Alignment
- Gene Quantification
- Test for Differential Expression
- Test for Functional Enrichment

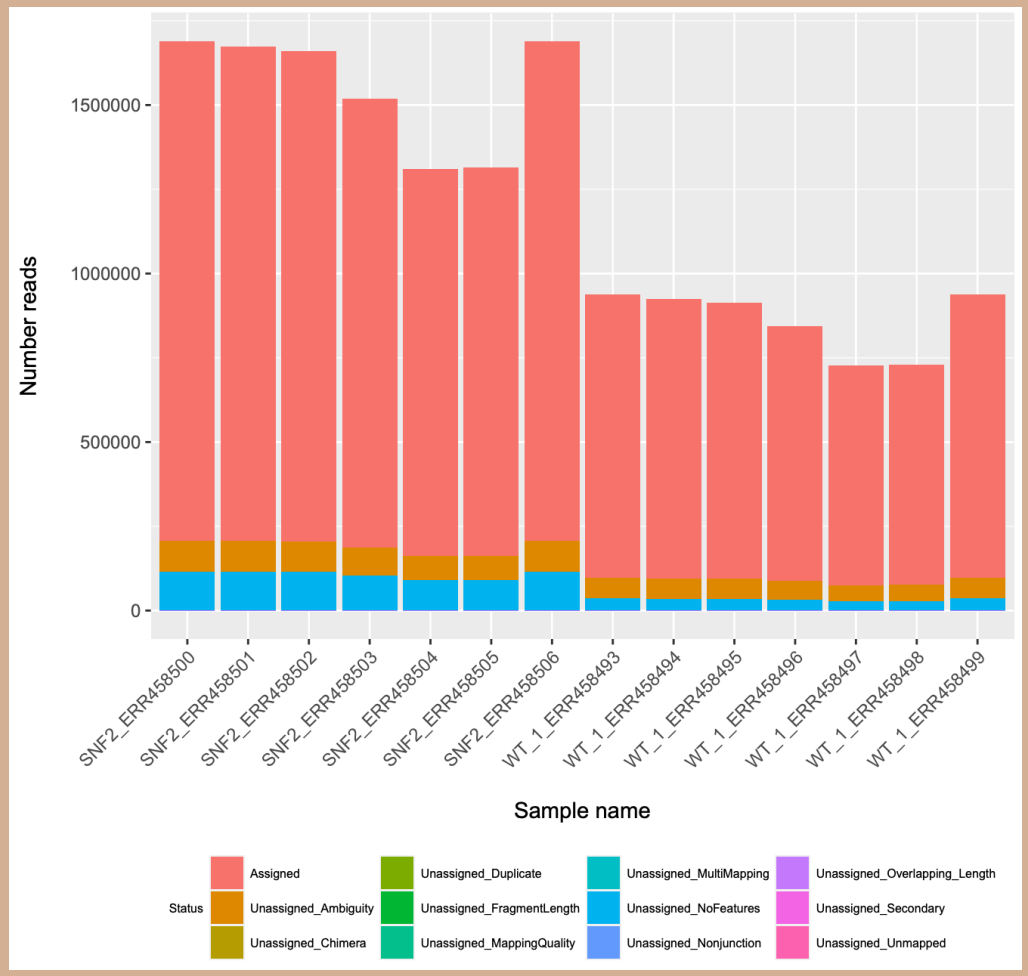
Mapping Statistics




```
(base) [whuo01@pcomp41 intro-to-RNA-seq]$ Rscript scripts/mapping_percentage.R
```

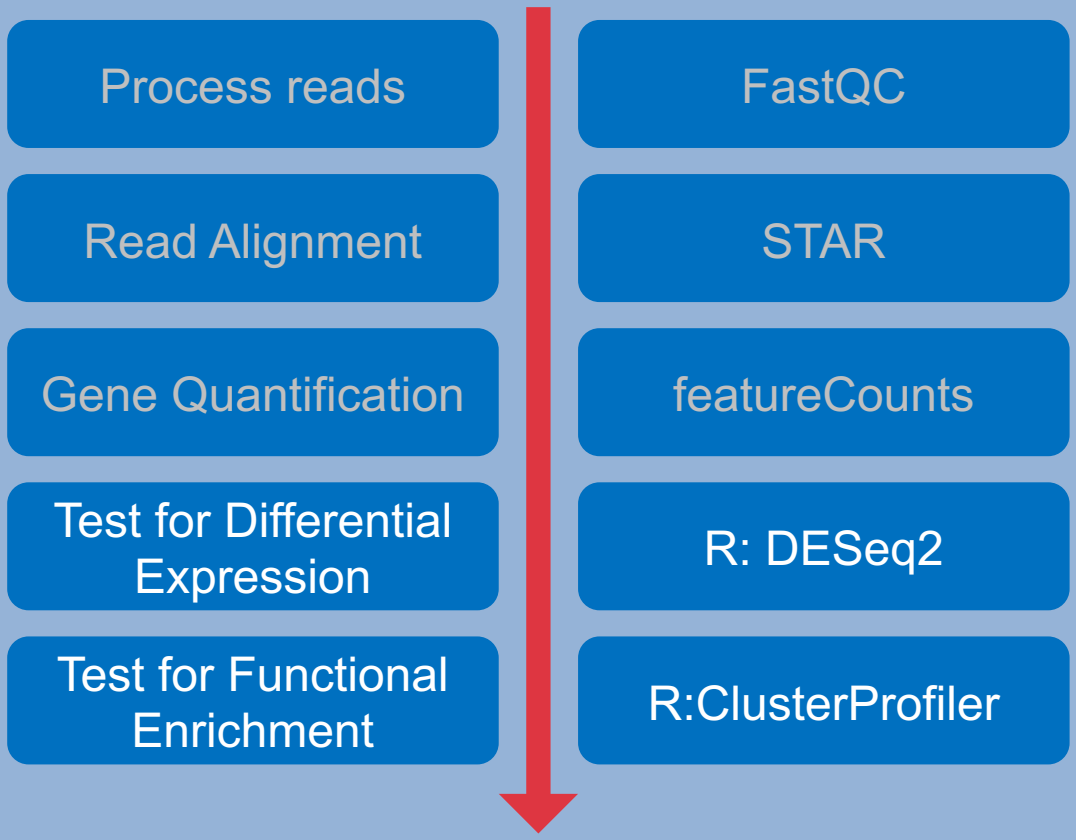
Optional: Feature count QC

- Process reads
- Read Alignment
- Gene Quantification
- Test for Differential Expression
- Test for Functional Enrichment




```
(base) [whuo01@pcomp41 intro-to-RNA-seq]$ Rscript scripts/featurecount_stat.R
```

Analysis pipeline



```
> # Input files: data, meta
> # DESeq2 transforms featurecount to log2FoldChange
> dds <- DESeqDataSetFromMatrix(countData = data, colData = meta, design = ~ condition)
```

- Process reads
- Read Alignment
- Gene Quantification
- Test for Differential Expression
- Test for Functional Enrichment

Input data files

data

Sample

Gene

	SNF2_ERR458500	SNF2_ERR458501	SNF2_ERR458502
YAL069W	0	0	0
YAL068W-A	0	0	0
YAL068C	0	0	0
YAL067W-A	0	0	0
YAL067C	18	21	10
YAL066W	0	1	0
YAL065C	2	0	0
YAL064W-B	3	3	1
YAL064C-A	12	7	8
YAL064W	3	3	2
YAL063C-A	2	0	3
YAL063C	35	38	46
YAL062W	127	136	113
YAL061W	124	100	125
YAL060W	300	309	323

meta

Sample

	condition
SNF2_ERR458500	SNF2
SNF2_ERR458501	SNF2
SNF2_ERR458502	SNF2
SNF2_ERR458503	SNF2
SNF2_ERR458504	SNF2
SNF2_ERR458505	SNF2
SNF2_ERR458506	SNF2
WT_ERR458493	WT
WT_ERR458494	WT
WT_ERR458495	WT
WT_ERR458496	WT
WT_ERR458497	WT
WT_ERR458498	WT
WT_ERR458499	WT

Love MI, Huber W, Anders S (2014). *Genome Biology*, **15**, 550. doi: 10.1186/s13059-014-0550-8.

Process reads

Read Alignment

Gene Quantification

Test for Differential Expression

Test for Functional Enrichment

```
> # DESeq2 transforms featurecount to log2FoldChange
> dds <- DESeqDataSetFromMatrix(countData = data, colData = meta, design = ~ condition)
> dds <- DESeq(dds)
> # Tell DESeq2 which conditions you would like to compare
> contrast <- c("condition", "SNF2", "WT")
> res_unshrunken <- results(dds, contrast=contrast)
> results <- lfcShrink(dds, contrast=contrast, res=res_unshrunken)
> rld <- rlog(dds, blind=TRUE)
> rld_counts <- assay(rld)
> # output tables: results and rld_counts
```

output files

results: log2FoldChange with pvalue and p-adjusted (FDR)

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
YAL069W	0.000000e+00	NA	NA	NA	NA	NA
YAL068W-A	0.000000e+00	NA	NA	NA	NA	NA
YAL068C	9.021961e-02	-0.043634441	0.35298047	-0.12361413	9.016208e-01	NA
YAL067W-A	0.000000e+00	NA	NA	NA	NA	NA
YAL067C	5.977662e+00	1.632313960	0.37520919	4.12031328	3.783576e-05	1.234224e-04
YAL066W	9.812029e-02	-0.046943453	0.36423350	-0.12887995	8.974526e-01	NA
YAL065C	1.257050e+00	-1.311117160	0.53025320	-2.40606774	1.612527e-02	3.116017e-02
YAL064W-B	1.607497e+00	-0.478036753	0.49638637	-0.97070341	3.316960e-01	4.231887e-01

rld_counts: regularized log transformed counts. y=log2(n + n_0)

	SNF2_ERR458500	SNF2_ERR458501	SNF2_ERR458502	SNF2_ERR458503	SNF2_ERR458504	SNF
YAL069W	0.00000000	0.00000000	0.000000000	0.00000000	0.000000e+00	
YAL068W-A	0.00000000	0.00000000	0.000000000	0.00000000	0.000000e+00	
YAL068C	-2.05284510	-2.05277066	-2.052678613	-2.03518019	-2.050732e+00	
YAL067W-A	0.00000000	0.00000000	0.000000000	0.00000000	0.000000e+00	
YAL067C	2.85518301	2.92415585	2.665424931	2.74363842	2.787779e+00	
YAL066W	-2.05330217	-2.03529421	-2.053124206	-2.05234800	-2.051044e+00	
YAL065C	0.10262823	0.02271448	0.023084190	0.02600336	1.213956e-01	
YAL064W-B	0.58300398	0.58389721	0.509902895	0.51530992	4.804996e-01	
YAL064C-A	2.49883637	2.35760285	2.390449413	2.28097988	2.344918e+00	

http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Process reads

Read Alignment

Gene Quantification

Test for Differential Expression

Test for Functional Enrichment

```
> # DESeq2 transforms featurecount to log2FoldChange
> dds <- DESeqDataSetFromMatrix(countData = data, colData = meta, design = ~ condition)
> dds <- DESeq(dds)
> # Tell DESeq2 which conditions you would like to compare
> contrast <- c("condition", "SNF2", "WT")
> res_unshrunken <- results(dds, contrast=contrast)
> results <- lfcShrink(dds, contrast=contrast, res=res_unshrunken)
> rld <- rlog(dds, blind=TRUE)
> rld_counts <- assay(rld)
> # output tables: results and rld_counts
```

output files

results: log2FoldChange with pvalue and p-adjusted (FDR)

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
YAL069W	0.000000e+00	NA	NA	NA	NA	NA
YAL068W-A	0.000000e+00	NA	NA	NA	NA	NA
YAL068C	9.021961e-02	-0.043634441	0.35298047	-0.12361413	9.016208e-01	NA
YAL067W-A	0.000000e+00	NA	NA	NA	NA	NA
YAL067C	5.977662e+00	1.632313960	0.37520919	4.12031328	3.783576e-05	1.234224e-04
YAL066W	9.812029e-02	-0.046943453	0.36423350	-0.12887995	8.974526e-01	NA
YAL065C	1.257050e+00	-1.311117160	0.53025320	-2.40606774	1.612527e-02	3.116017e-02
YAL064W-B	1.607497e+00	-0.478036753	0.49638637	-0.97070341	3.316960e-01	4.231887e-01

rld_counts: regularized log transformed counts. y=log2(n + n_0)

	SNF2_ERR458500	SNF2_ERR458501	SNF2_ERR458502	SNF2_ERR458503	SNF2_ERR458504	SNF
YAL069W	0.00000000	0.00000000	0.000000000	0.00000000	0.000000e+00	
YAL068W-A	0.00000000	0.00000000	0.000000000	0.00000000	0.000000e+00	
YAL068C	-2.05284510	-2.05277066	-2.052678613	-2.03518019	-2.050732e+00	
YAL067W-A	0.00000000	0.00000000	0.000000000	0.00000000	0.000000e+00	
YAL067C	2.85518301	2.92415585	2.665424931	2.74363842	2.787779e+00	
YAL066W	-2.05330217	-2.03529421	-2.053124206	-2.05234800	-2.051044e+00	
YAL065C	0.10262823	0.02271448	0.023084190	0.02600336	1.213956e-01	
YAL064W-B	0.58300398	0.58389721	0.509902895	0.51530992	4.804996e-01	
YAL064C-A	2.49883637	2.35760285	2.390449413	2.28097988	2.344918e+00	

http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Process reads

Read Alignment

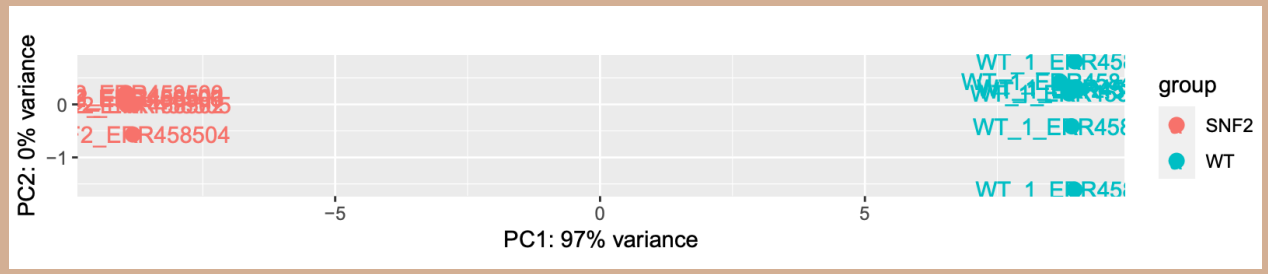
Gene Quantification

Test for Differential Expression

Test for Functional Enrichment

```
> # DESeq2 transforms featurecount to log2FoldChange
> dds <- DESeqDataSetFromMatrix(countData = data, colData = meta, design = ~ condition)
> dds <- DESeq(dds)
> # Tell DESeq2 which conditions you would like to compare
> contrast <- c("condition", "SNF2", "WT")
> res_unshrunken <- results(dds, contrast=contrast)
> results <- lfcShrink(dds, contrast=contrast, res=res_unshrunken)
> rld <- rlog(dds, blind=TRUE)
> rld_counts <- assay(rld)
> # output tables: results and rld_counts
> # plot PCA
> plotPCA(rld, intgroup="condition") + geom_text(aes(label=name))
```

How well are the replicates?

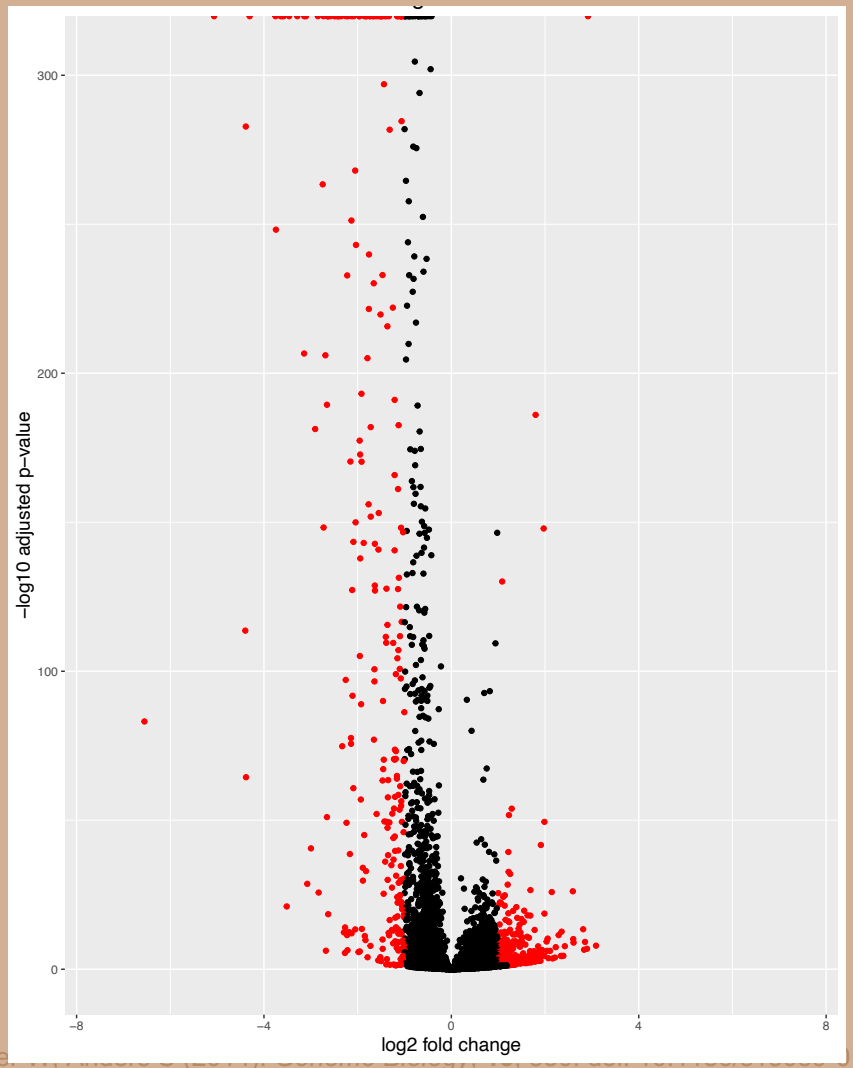


Love MI, Huber W, Anders S (2014). *Genome Biology*, **15**, 550. doi: 10.1186/s13059-014-0550-8.

```
> # output table: results
> # load necessary library ggplot2
> library(ggplot2)
> # red dots: threshold of padj<0.05 and absolute value of log2foldchange >=1
> res_table <- results %>% data.frame() %>% rownames_to_column(var="gene") %>% as_tibble()
> res_table <- res_table %>% mutate(threshold_OE = padj < 0.05 & abs(log2FoldChange) >= 1)
> # make volcano plot
> ggplot(res_table) + …
```

- Process reads
- Read Alignment
- Gene Quantification
- Test for Differential Expression
- Test for Functional Enrichment

Volcano plot

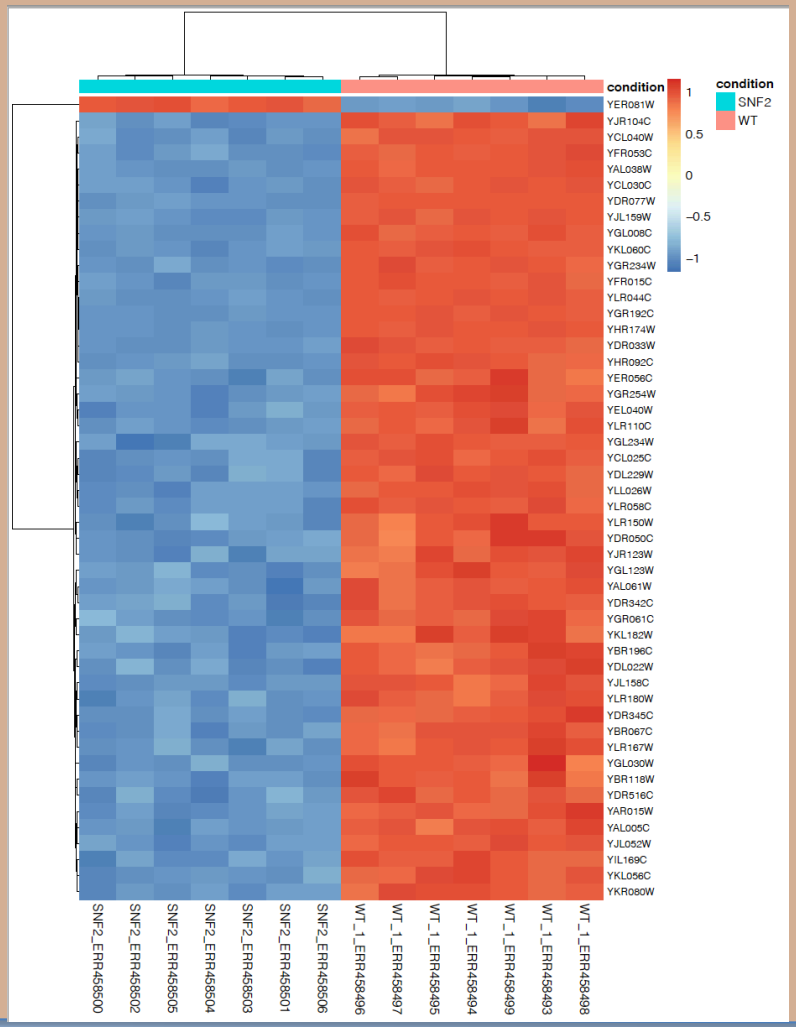


Love MI, Hube… …014-0550-8.

```
> # output file: results rld_counts
> # select a subset of genes to plot heatmap
> rld_counts_sig <- rld_counts[gene_list, ]
> # plot heatmap
> pheatmap(rld_counts_sig, cluster_rows = T, show_rownames = T, annotation = meta,
border_color = NA, fontsize = 10, scale = "row", fontsize_row = 8, height = 20)
```

- Process reads
- Read Alignment
- Gene Quantification
- Test for Differential Expression
- Test for Functional Enrichment

Heatmap and clustering analysis

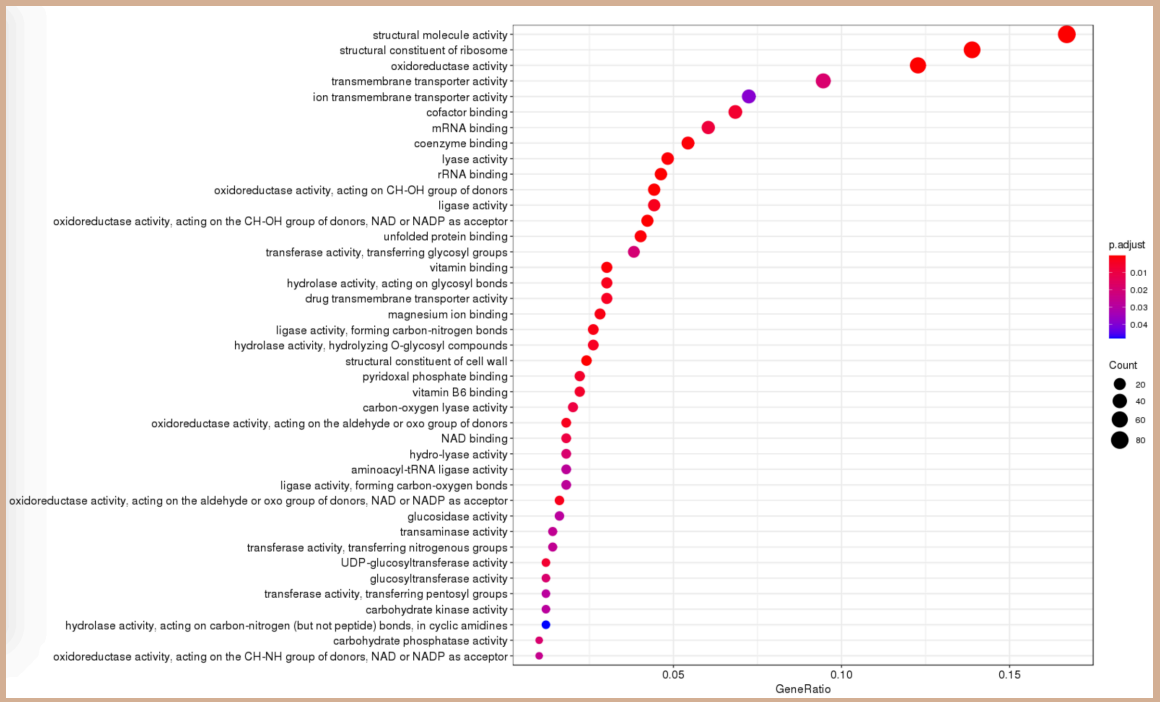


Love MI, Huber W, Anders S (2014). *Genome Biology*, **15**, 550. doi: 10.1186/s13059-014-0550-8.

```
> # output file: results and rld_counts  -> extract genes of interest
> # load necessary libraries
> library(clusterProfiler)
> library(org.Sc.sgd.db)
> # Run GO enrichment for a list of genes: gene_list
> ego <- enrichGO(gene = gene_list, keyType = "ENSEMBL", OrgDb = org.Sc.sgd.db)
> cluster_summary <- data.frame(ego)
> dotplot(ego, showCategory=50)
```

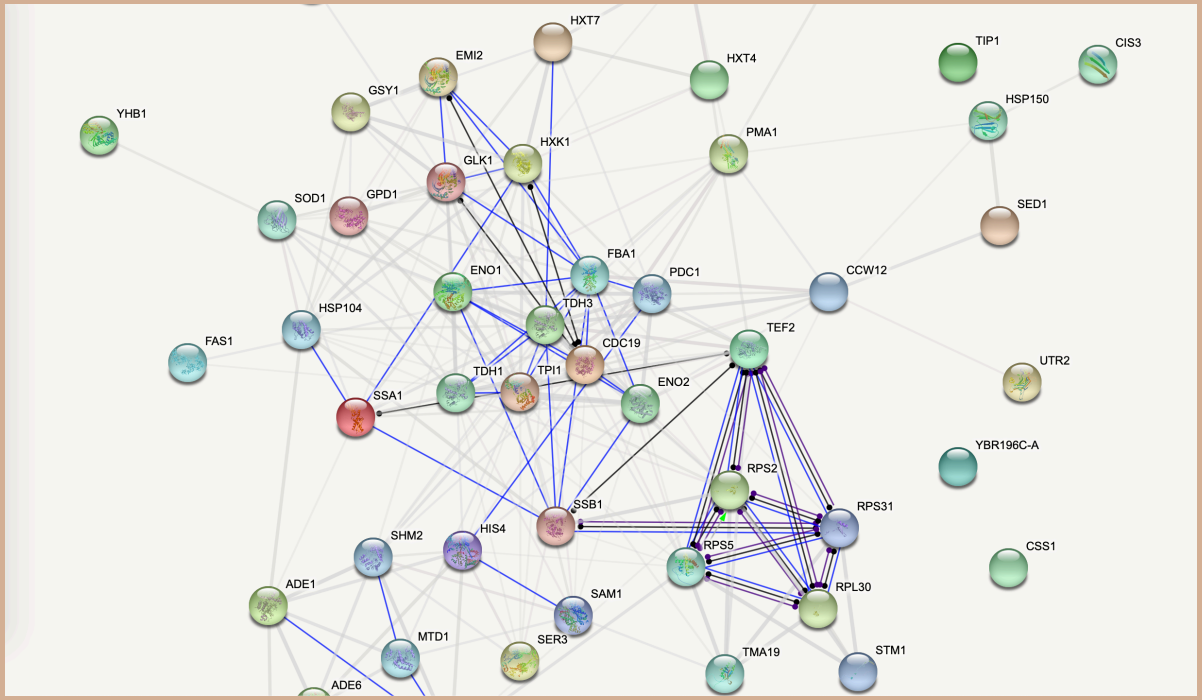
- Process reads
- Read Alignment
- Gene Quantification
- Test for Differential Expression
- Test for Functional Enrichment

dotplot: Enrichment analysis using clusterProfiler



- Process reads
- Read Alignment
- Gene Quantification
- Test for Differential Expression
- Test for Functional Enrichment

Enrichment analysis using web server: String-db.org



Molecular Function (GO)

GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0005199	structural constituent of cell wall	5 of 17	0.00011
GO:0036094	small molecule binding	21 of 931	0.00021
GO:0051287	NAD binding	5 of 30	0.00030
GO:0003824	catalytic activity	33 of 2212	0.00030
GO:0050662	coenzyme binding	8 of 139	0.00051
			(more ...)

Cellular Component (GO)

GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0071944	cell periphery	25 of 790	6.91e-09
GO:0005829	cytosol	26 of 845	6.91e-09
GO:0005886	plasma membrane	19 of 528	2.38e-07
GO:0009277	fungal-type cell wall	10 of 112	5.55e-07
GO:0005576	extracellular region	9 of 120	7.16e-06
			(more ...)

What's next?

Next step: follow self-guided tutorial and get familier with the pipeline.

What to do with your own data?
– one simply solution: move your reads.fastq files into raw_data folder and apply the same pipeline.

Questions?
– Ask questions and seek practice buddies on Piazza

Advanced reading material

An 1.5 day workshop for RNA sequencing analysis
https://hbctraining.github.io/DGE_workshop/

DESeq2 package description
http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Alternative analysis tools: Geneious Prime
https://www.geneious.com/tutorials/expression-analysis/
https://www.youtube.com/watch?v=dwG-4i85CWw&t=16s

Initiate shell access on compute node

```
srun -t 3:00:00 --mem 16G -N 1 -n 4 -p preempt --reservation bioworkshop --pty bash
cd /cluster/tufts/bio/tools/training/intro-to-rnaseq/users/
mkdir YOUR_USERNAME
cd YOUR_USERNAME
```

Setup Rstudio on OnDemand

Number of cores: 1
Amount of Memory: 32 Gb
R version: 3.5.0
Partition: Preempt
Reservation for class, training, workshop: Bioinformatics workshop