High Performance Research Computing

A Resource for Research and Discovery

TEXAS A&M UNIVERSITY

Introduction to R

Noushin Ghaffari, PhD

Bioinformatics Scientist, Genomics and Bioinformatics, Texas A&M AgriLife Research

Research Scientist, Texas A&M High Performance Research Computing

DIVISION OF RESEARCH

TEXAS A&M UNIVERSITY

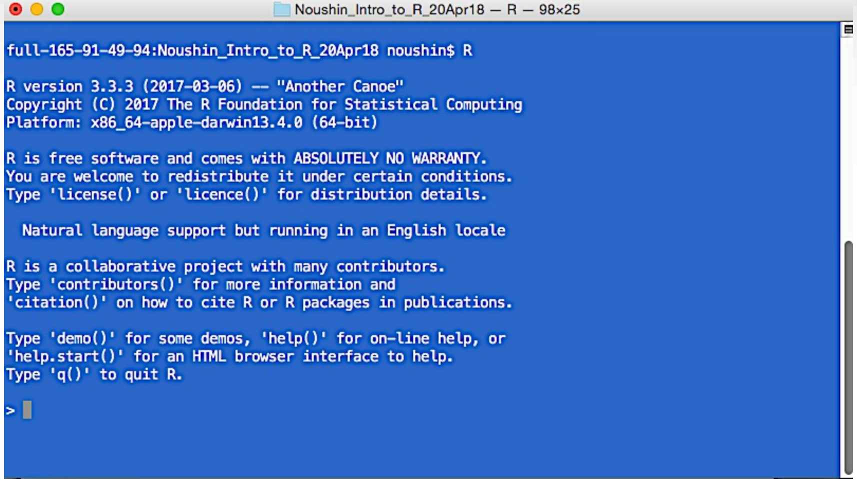
What is R?

- Open source programming language and software environment for statistical computing and graphics
- Supported by the R Foundation for Statistical Computing
- Supports multiple platforms and can be easily extended.

- The “Comprehensive R Archive Network” (CRAN) is a collection of sites which carry identical material, consisting of the R distribution(s), the contributed extensions, documentation for R, and binaries.
- The CRAN master site at WU (Wirtschaftsuniversität Wien) in Austria can be found at the URL https://CRAN.R-project.org/
    - Mirrored daily to many sites around the worldhttps://CRAN.R-project.org/mirrors.html

Why R?

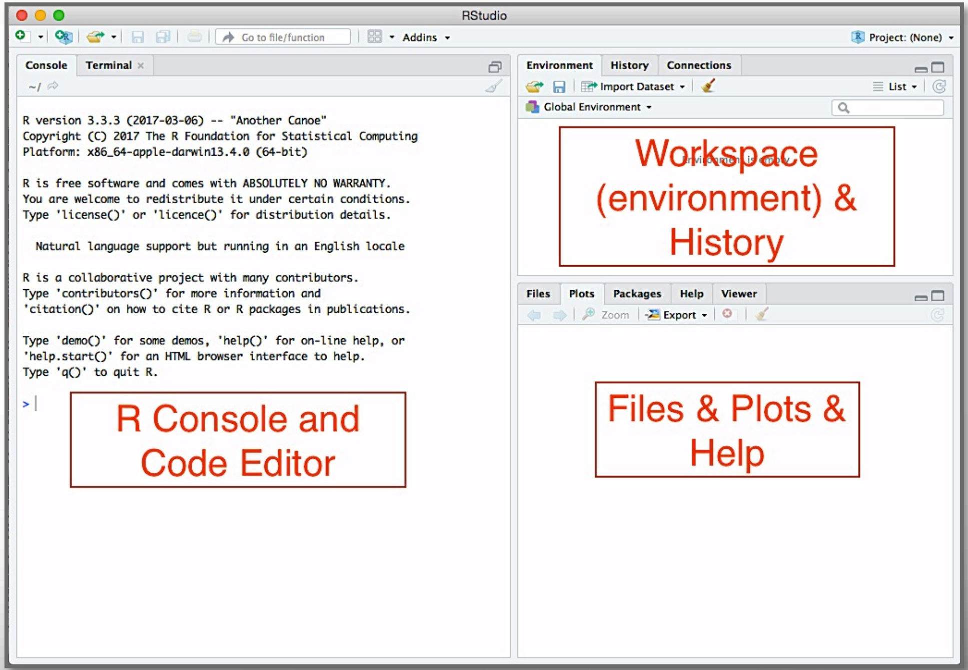
- Open source
- Comprehensive collection of statistical functions
  - Linear modeling, classification, clustering, genomics analysis, economic and financial analysis, etc.
- Large user community – support
- Collection of “packages”
  - A package is a shared code, documentation/Vignettes, data (occasionally)
- Multi platform, but not too sensitive about the source platform
- Command-line or graphical user interface (recently)
- Anyone can contribute



```
full-165-91-49-94:Noushin_Intro_to_R_20Apr18 noushin$ R

R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

What is RStudio?

- An integrated development environment (IDE) for R
- Graphical Interface – user friendly environment
- Embedded Text Editor
- Auto completion on names (functions)
- Simplified plot and output view



```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Workspace (environment) & History

R Console and Code Editor

Files & Plots & Help

Running R for this Course

- Your own laptop - You need to install R on your laptop using R download page: https://cran.r-project.org/mirrors.html
- TAMU Open Access Lab (OAL) computers have R and RStudio installed
- TAMU NetID - You can use the online Jupyter notebook for the class
  - Off campus needs VPN
- HPRC Portal to run Rstudio: http://portal.hprc.tamu.edu
- TAMU HPRC Ada System Account – You can use R_Tamu on Ada or choose Jupyter Notebook
  - Off campus needs VPN

Using Jupyter Notebook

Jupyter is a web application that allows you to easily connect to a remote server and run applications. We will be using **Jupyter** to connect to **titan.tamu.edu** and use R.

To access the **Jupyter Notebook** you will need:

-An Internet-connected device (desktop, laptop, or large tablet is preferred)

-Internet access (tamulink-wpa Wi-Fi, OAL Ethernet)

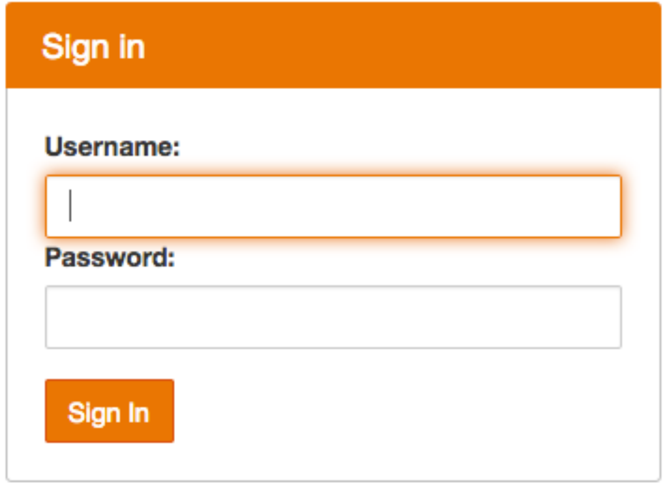
-An Internet browser

Jupyter Notebook Access

(1) Open your Internet browser.

(2) Go to one of the following links:
https://titan.tamu.edu:8000/

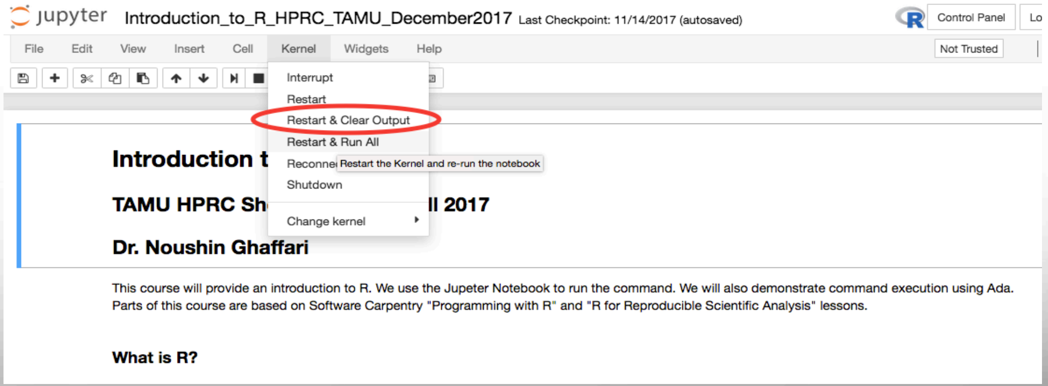
(3) You should then see a login window.
Use your NetID credentials to log in.
This is the same username and password as Howdy and Ada.



Jupyter Notebook Access - 2

(4) Upon successful login you will see some files and directories. Choose on the “HPRC_R_Courses" directory, then "Introduction_to_R_HPRC_ResearchComputingSymposium_Jun18", and click on "Introduction_to_R_12Jun18.ipynb”

(5) Before class begins, please restart your kernel and clear all outputs. To do this, click on “Kernel -> Restart & Clear Output”



Introduction t

TAMU HPRC Sh ll 2017

Dr. Noushin Ghaffari

This course will provide an introduction to R. We use the Jupeter Notebook to run the command. We will also demonstrate command execution using Ada. Parts of this course are based on Software Carpentry "Programming with R" and "R for Reproducible Scientific Analysis" lessons.

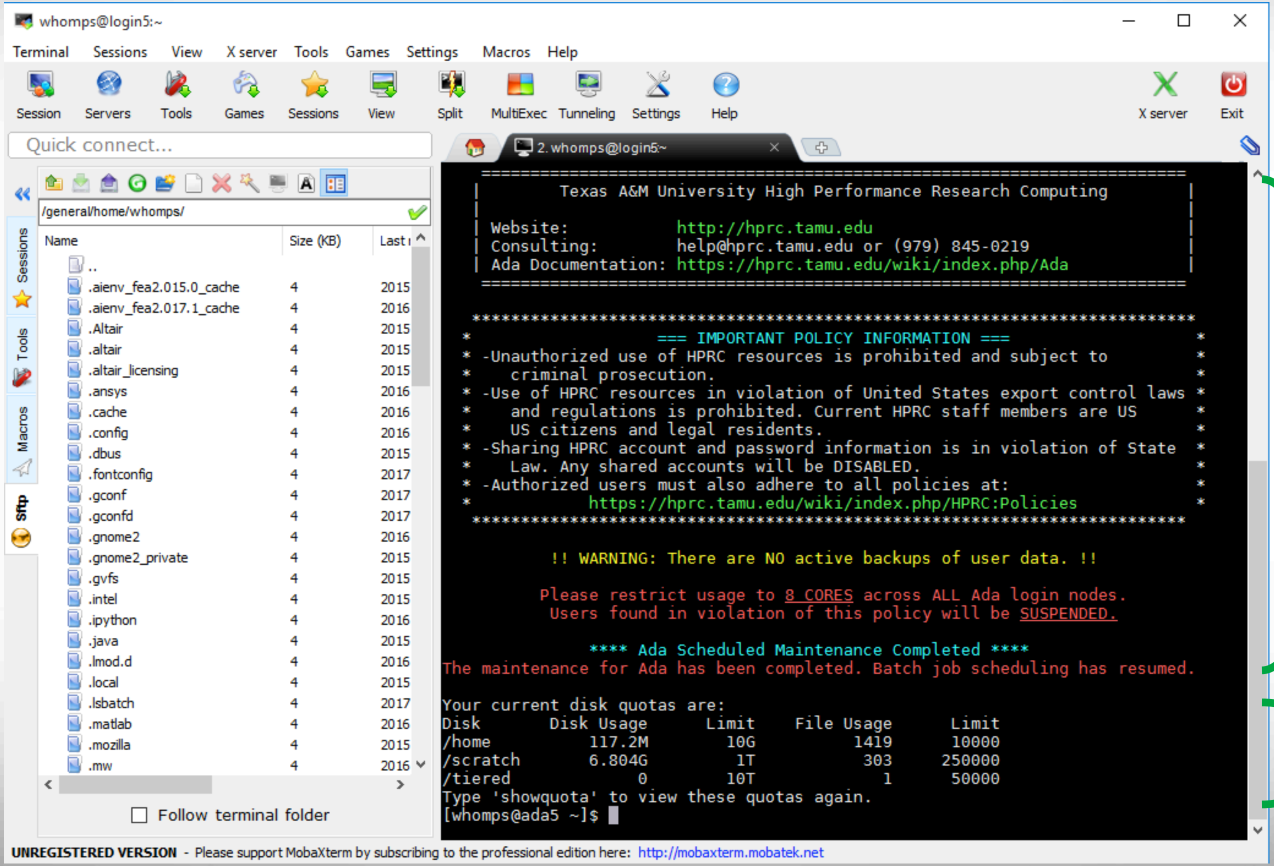
What is R?

Connecting to HPRC to Use R

- SSH (secure shell)
  - The only program allowed for remote access; encrypted communication; freely available for Linux/Unix and Mac OS X hosts;
- For Microsoft Windows PCs, use *MobaXterm*
    - https://hprc.tamu.edu/wiki/HPRC:MobaXterm
      - You are able to view images and use GUI applications with MobaXterm
  - or *Putty*
    - https://hprc.tamu.edu/wiki/HPRC:Access#Using_PuTTY
      - You can not view images or use GUI applications with PuTTY

- Both state of Texas law and TAMU regulations prohibit the sharing and/or illegal use of computer passwords and accounts
- Don’t write down passwords
- Don’t choose easy to guess/crack passwords
- Change passwords frequently

Using SSH - MobaXterm (on Windows)



Name	Size (KB)	Last
..		
.aienv_fea2.015.0_cache	4	2015
.aienv_fea2.017.1_cache	4	2016
.Altair	4	2015
.altair	4	2015
.altair_licensing	4	2015
.ansys	4	2016
.cache	4	2016
.config	4	2016
.dbus	4	2015
.fontconfig	4	2017
.gconf	4	2017
.gconfd	4	2017
.gnome2	4	2016
.gnome2_private	4	2015
.gvfs	4	2015
.intel	4	2015
.ipython	4	2016
.java	4	2015
.lmod.d	4	2016
.local	4	2015
.lsbatch	4	2017
.matlab	4	2016
.mozilla	4	2015
.mw	4	2016

```
  ==========================================================================
 |        Texas A&M University High Performance Research Computing         |
 |                                                                          |
 | Website:          http://hprc.tamu.edu                                   |
 | Consulting:       help@hprc.tamu.edu or (979) 845-0219                   |
 | Ada Documentation: https://hprc.tamu.edu/wiki/index.php/Ada              |
  ==========================================================================

 ****************************************************************************
 *                 === IMPORTANT POLICY INFORMATION ===                      *
 * -Unauthorized use of HPRC resources is prohibited and subject to         *
 *    criminal prosecution.                                                 *
 * -Use of HPRC resources in violation of United States export control laws *
 *    and regulations is prohibited. Current HPRC staff members are US      *
 *    US citizens and legal residents.                                      *
 * -Sharing HPRC account and password information is in violation of State  *
 *    Law. Any shared accounts will be DISABLED.                            *
 * -Authorized users must also adhere to all policies at:                   *
 *           https://hprc.tamu.edu/wiki/index.php/HPRC:Policies             *
 ****************************************************************************

           !! WARNING: There are NO active backups of user data. !!

         Please restrict usage to 8 CORES across ALL Ada login nodes.
          Users found in violation of this policy will be SUSPENDED.

                **** Ada Scheduled Maintenance Completed ****
The maintenance for Ada has been completed. Batch job scheduling has resumed.

Your current disk quotas are:
Disk         Disk Usage      Limit    File Usage      Limit
/home            117.2M        10G          1419      10000
/scratch         6.804G         1T           303     250000
/tiered               0        10T             1      50000
Type 'showquota' to view these quotas again.
[whomps@ada5 ~]$
```

message of the day

your quotas

Using SSH to Access Ada

```
ssh –X user_NetID@ada.tamu.edu
```

https://hprc.tamu.edu/wiki/Ada:Access

You may see something like the following the first time you connect to the remote machine from your local machine:

```
Host key not found from the list of known hosts.
Are you sure you want to continue connecting (yes/no)?
```

Type yes, hit enter and you will then see the following:

```
Host 'ada.tamu.edu' added to the list of known hosts.
user_NetID@ada.tamu.edu's password:
```

Login and Set up

- Login to Ada using SSH or MobaXterm
- Let’s take a look at the path and create appropriate directories

```
echo $SCRATCH
cd $SCRATCH
Pwd
mkdir Intro_to_R_Jun18
```

Transferring the Code

The R code and its submission script should be copied to users’ local directory

```
cd $SCRATCH/Intro_to_R_Jun18

cp /scratch/training/Intro_to_R/Scripts/* .
```

Data_Generator.R - 1

```
#randomly generates 10000 number in a matix[100,100]
rand_data <- matrix(ncol=100,nrow=100,data=runif(10000,1,100))

#getting log2 of the data and transposing the results
rand_data_t_log2 <- t(log2(rand_data))

#calculating the min, max and average for rows and columns
mean_cols <- apply(rand_data,2,mean)
mean_rows <- apply(rand_data,1,mean)
min_rows <- apply(rand_data,1,min)
min_cols <- apply(rand_data,2,min)
max_rows <- apply(rand_data,1,max)
max_cols <- apply(rand_data,2,max)
```

Data_Generator.R - 2

```
#making plot and saving results in a pdf file
pdf("Rand_Data_QC.pdf")
g_range <- range(0,min_rows,max_rows)
plot(mean_rows, type="l", col="blue", ylim=g_range, ann=FALSE)
lines(mean_cols, type="l", col="red")
lines(max_rows, type="l", col="green")
lines(max_cols, type="l", col="purple")
lines(min_rows, type="l", col="royalblue")
lines(min_cols, type="l", col="coral")
legend(80,40,c("Mean Rows","Mean Cols","Max Rows","Max Cols","Min Rows","Min Cols"),col=c("blue","red","green","purple","coral","royalblue"),lty=1,cex=0.55)
dev.off()
```

Data_Generator.R - 3

```
#saving data files
write.csv(rand_data,"rand_data.csv")
write.csv(rand_data_t_log2,"rand_data_t_log2.csv")
```

R_Script_Submit.sh

```
#BSUB -J Testing_R_Script    # sets the job name to Testing_R_Script.
#BSUB -L /bin/bash           # uses the bash login shell to initialize the job's execution environment.
#BSUB -W 1:00                # sets to 5 hours the job's runtime wall-clock limit.
#BSUB -n 1                   # assigns 1 core for execution.
#BSUB -R "span[ptile=1]"     # assigns 1 core per node.
#BSUB -R "rusage[mem=5000]"  # reserves ~5GB per process/CPU for the job
#BSUB -M 5000                # sets to ~5GB the per process enforceable memory limit.
#BSUB -o stdout.%J           # directs the job's standard output to stdout.jobid

## Load the necessary modules
module purge
module load R_tamu/3.4.2-iomkl-2017A-Python-2.7.12-default-mt

## Launch R with proper parameters
Rscript Data_Generator.R
```

Submitting the RScript to Ada

```
bsub < $SCRATCH/Intro_to_R_Jun18/R_Script_Submit.sh

Bjobs

ls –l
```

Now, let’s copy the run’s output to your local computer:

```
scp NetID@ada.tamu.edu:/scratch/user/NetID/ \

Intro_to_R_Jun18/* path_to_a_directory_on_local_machine
```

Course Content

Parts of this course are based on Software Carpentry "Programming with R" and "R for Reproducible Scientific Analysis" lessons.

> "*Software Carpentry* is a volunteer non-profit organization dedicated to teaching basic computing skills to researchers."

- https://software-carpentry.org/lessons/
- Reference page for R lessons (cheat sheet)
  - http://swcarpentry.github.io/r-novice-inflammation/reference/
  - http://swcarpentry.github.io/r-novice-gapminder/reference

Any question?

nghaffari@tamu.edu