Introduction to

RAMA-SOMPALLAE@uiowa.edu
Iowa Institute of Human genetics
University of Iowa

05/20/2015
What we will learn

• Basic functionalities in R
• Import and export of data
• Data transformation
• Simple visualization in R
What is R?

- R is a software environment for data analysis
- Created by Robert Gentleman & Ross Ihaka, Dept of Statistics, University of Auckland, New Zealand.
- It was developed to share the statistical methods.
- It is mainly an implementation of S language developed at Bell labs
- R’s development began in early 1990s
- R is made public in 1993 and current version R3.2.0
- R core Development team formed in 1997
Why R (is great)?

- R is freely available and licensed under open source
- it contains suite of functions for basic calculations and data analysis
- is extensible with community developed packages
- Over 4000 packages
- operators to work with arrays and large data matrices
- Also has a graphical system
- Good documentation
Drawbacks of R

• A bit of learning curve
• Can be very slow depending on the objects in workspace
• Update R versions to meet Packages compatibility
Where is it used

• Revolution Analytics
• Google
• Facebook uses for analyzing user behavior.
• National Weather Service: Flood forecasts.
• Orbitz: Statistical analysis to suggest best hotel.
• Trulia: Statistical modeling for real estate.
Lab- 1
Installing R and R studio

R - http://cran.r-project.org/

R Studio - http://www.rstudio.com/
R Studio

Editor

Workspace, History

Console

Plots, Files, Packages, Help

5/20/15

IIHG Bioinformatics course
R programing

R code can be executed on the command line directly or run as a script

Commands are separated either by ‘;’ or by a newline.

R is case sensitive.

Help files for R functions are accessed by preceding the name of the function with ? (e.g. ?require).
Go to R script
Data Types

- Supports any type of data
- Vectors (character, numeric, logical)
  - c('GeneA', 'GeneB', 'GeneC', 'GeneD', 'GeneE')
  - c(1,2,3)
- Matrix (2 dimensional data of single data type)
- Array (can have more than 2 dimensions)
- Data frames (matrix with multiple data types)
- Lists (collection of objects)
- Factors

```
exp1  exp2
[1,]  1  6
[2,]  2  7
```
Data frames

• Once read in, datasets in R are typically stored as data frames, which have a matrix structure. Observations are arranged as rows and variables, either numerical or categorical, are arranged as columns.

• Individual rows, columns, and cells in a data frame can be accessed through many methods of indexing.

• We most commonly use object[row,column] notation.
Basic operators

Comparison operators
- equal: ==
- not equal: !=
- greater/less than: > <
- greater/less than or equal: >= <=

Logical operators
- AND: &
- OR: |
- NOT: !
Installing R packages

• If you have sufficient privileges, you can install and update most R packages.

• `install.packages("packagename")`
• To update packages, you can update them all at once using:

• `update.packages()`
R environment

• R creates hidden files .RData, .Rhistory and .Rprofile (optional) to control R environment.
• History - save all commands you type
• Rdata saves everything in memory (can be large - be careful)
• Helpful to rename
  • save.image(file="my_project.RData")
  • save(myVec, file="my_project_object.RData")
Lab session -2
RNA Seq
RNA Seq pipeline

1. Raw Data (FastQ)
2. QC (fastqc)
3. Remove Duplicates
4. Align to Ref (60 aligners)

- Differential Expression
- Novel Isoforms
- Variants
Bioconductor

• Software for the analysis of genomic data.
• Open source and uses R
• Release coincides with R
• Install from website

source("http://www.bioconductor.org/biocLite.R")
biocLite("package")
Practice tutorial


How to get the speed.

- Try to use it in your daily research and
- Hangout with coders
Resources

• R bloggers
• Package documentation
• Books – R in Nutshell (Joseph Adler)
• Online courses Coursera, Harvard edX, code school etc
• [http://r-dir.com/learn/courses.html](http://r-dir.com/learn/courses.html)
Plotting

```r
plot(mycooldata[,1], mycooldata[,2])
plot(mycooldata[,1], mycooldata[,2], xlab="height", ylab="weight")
plot(mycooldata[,1], mycooldata[,2], xlab="height", ylab="weight", main = "Height Vs Weight")

# save to file
pdf(myplot.pdf)
plot(mycooldata[,1], mycooldata[,2], xlab="height", ylab="weight", main = "Height Vs Weight")
dev.off()
```
Functions

```r
a <- 2

mysqrt <- function(n) {
  sqrt(n)
}

print(mysqrt(a))
```
Loop

- for (i in 1:50) {
  print i
}

- while (i<5) {
  print i
}
Steps in data analysis

- Read data (input)
- Review data
- Load or install required packages
- Analyze
- Save output to files
- Save commands to a script `myscript.R`
  ```
  ## In R
  source("myscript.R")
  ## Or from the command line
  R CMD BATCH myscript.R
  ```
- Save scripts, `projectNameDate.R`,