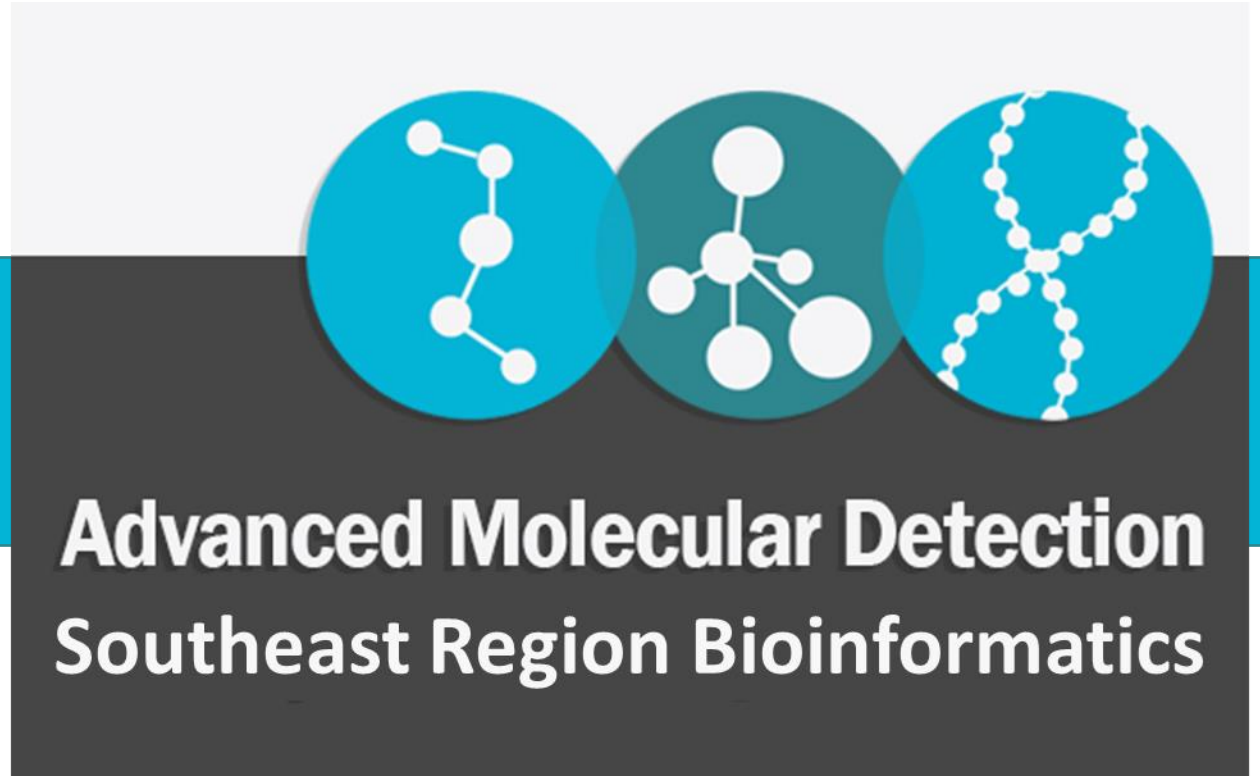


Introduction to R

04/29/2024

Outline

Agenda

Notes

Introduction

Features and Installation of R

Questions

Agenda

May 13 – R Data Types and Structures Part - 2
May 27 – Importing Data & Iterations in R Part - 3

Future Trainings

- ONT & FL's Flisochar pipeline
- StaPH-B Toolkit Programs/Pipelines
- GISAID flagged SARS-CoV-2
- R Training Series
- Dryad pipeline
- …and more

Advanced Molecular Detection
Southeast Region Bioinformatics

Updates

- Molly has changed roles to the Bioinformatics Supervisor for Florida Department of Health
- Thank you everyone for joining the BRR Quarterly meeting, please reach out to us if you have any questions or requests.

Advanced Molecular Detection
Southeast Region Bioinformatics

What is R?

- R is a programming language and free software developed by Ross Ihaka and Robert Gentleman in 1993
- An integrated suite of software features for data manipulation, calculation, and graphical display
- R is used among data miners, bioinformaticians and statisticians for data analysis and developing statistical software
- R software environment is an open-source free software environment within the GNU package, available under the GNU General Public License

Advanced Molecular Detection
Southeast Region Bioinformatics

Uses of R

R has

- an effective data handling & storage facility
- a suite of operators for calculations on arrays, in particular matrices
- a large, coherent, integrated collection of intermediate tools for data analysis
- graphical facilities for data analysis & display either directly at the computer or on hard-copy
- a well developed, simple & effective programming language which includes conditionals, loops, user defined recursive functions, and input and output facilities

Features of R

features of R images - Bing images

Advanced Molecular Detection
Southeast Region Bioinformatics

Data Processing

R's data structures include vectors, lists, arrays, and data frames

Vectors:

- Most basic R data objects
- Six types of vectors namely logical, integer, double, complex, character, & raw

Arrays:

- Data objects that store data in more than two dimensions
- Stores the values having only a similar kind of data types

data processing images - Bing images

Advanced Molecular Detection
Southeast Region Bioinformatics

Data Processing

Lists:

- R objects which contain elements of different types like numbers, strings, vectors, & another list inside it
- Collection of data which is ordered and changeable

Data frames:

- Two-dimensional data structure which can store data in tabular format
- Data frames have rows and columns, and each column can be a different vector. Also, different vectors can be of different data types

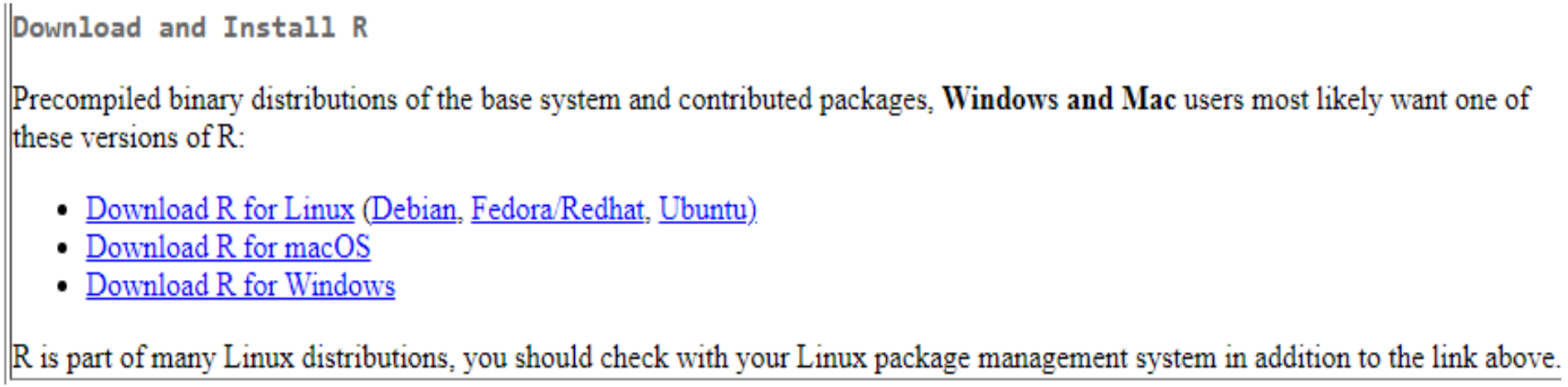
Advanced Molecular Detection
Southeast Region Bioinformatics

Programming

- R is an interpreted language, users can access it through a command-line interpreter
- R supports procedural programming with functions and, for some functions, object oriented programming with generic functions
- R is highly extensible through the use of packages for specific functions and specific applications

How to install R & R Studio?

To download R, go to this website and choose the download link that corresponds to your computer



Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

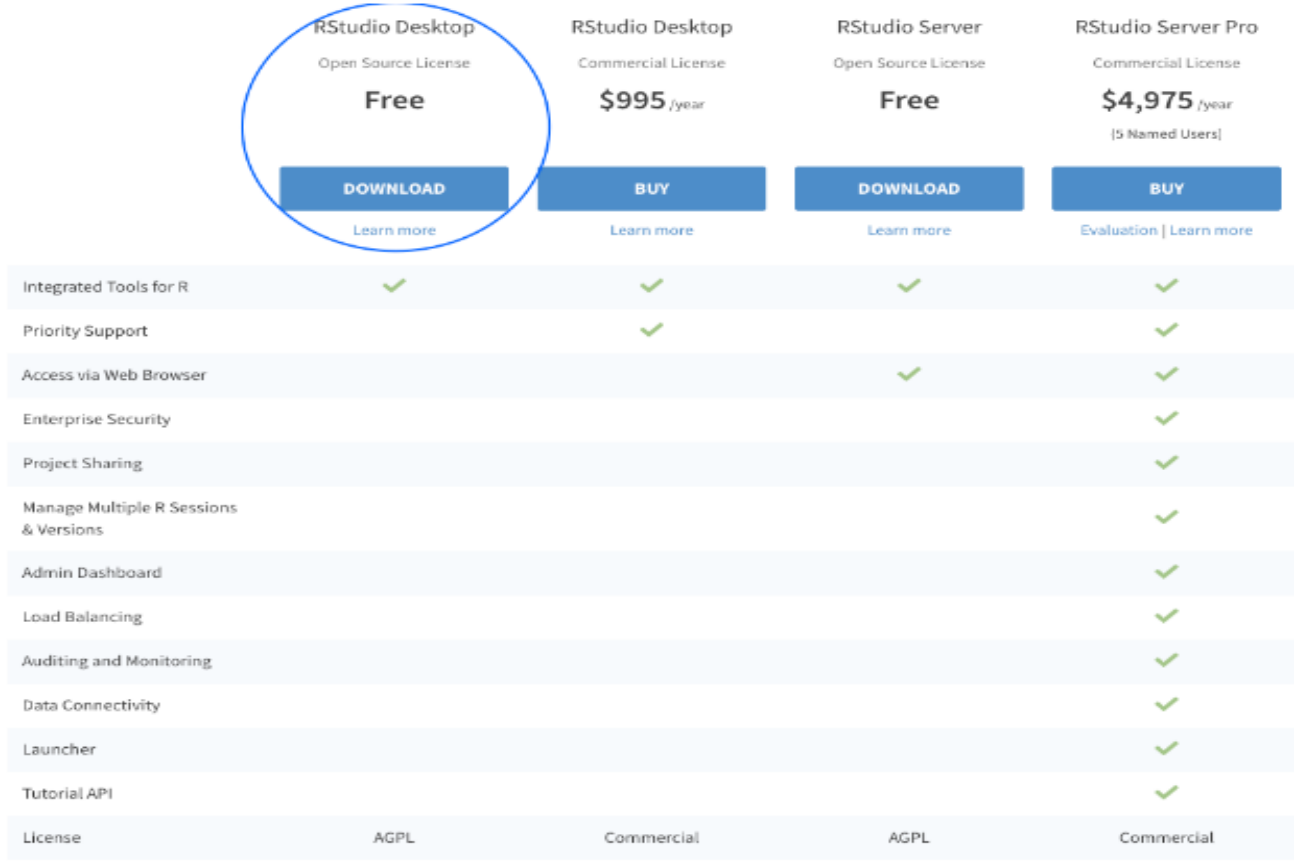
- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Advanced Molecular Detection
Southeast Region Bioinformatics

R Studio

- Now that R is downloaded, let's download R Studio. Navigate to this link

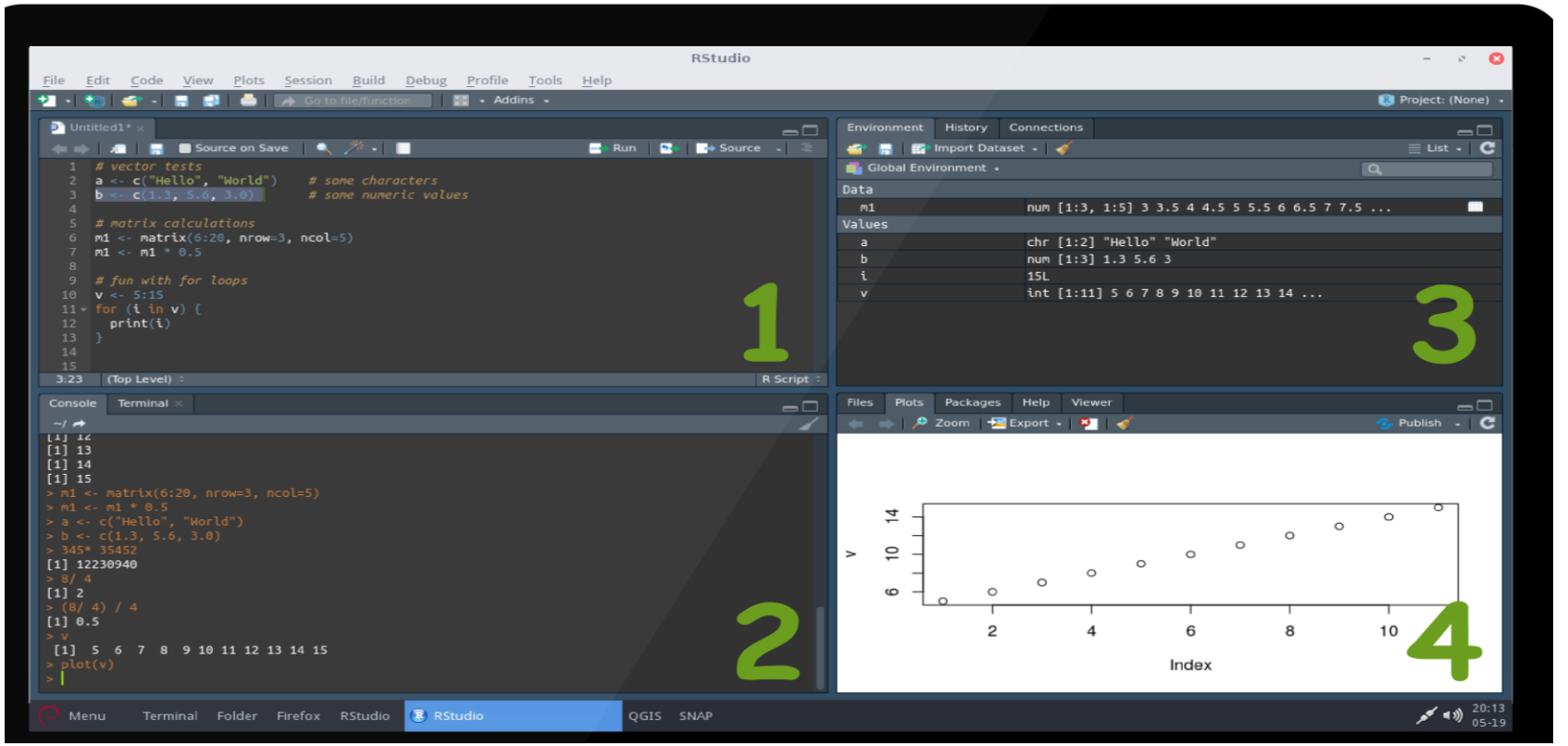


	RStudio Desktop Open Source License Free DOWNLOAD Learn more	RStudio Desktop Commercial License $995 /year BUY Learn more	RStudio Server Open Source License Free DOWNLOAD Learn more	RStudio Server Pro Commercial License $4,975 /year [5 Named Users] BUY Evaluation | Learn more
Integrated Tools for R	✓	✓	✓	✓
Priority Support		✓		✓
Access via Web Browser			✓	✓
Enterprise Security				✓
Project Sharing				✓
Manage Multiple R Sessions & Versions				✓
Admin Dashboard				✓
Load Balancing				✓
Auditing and Monitoring				✓
Data Connectivity				✓
Launcher				✓
Tutorial API				✓
License	AGPL	Commercial	AGPL	Commercial

Dataquest : Tutorial: Getting Started with R and RStudio – Dataquest

Advanced Molecular Detection
Southeast Region Bioinformatics

First look at R Studio



r studio pictures - Bing images

Advanced Molecular Detection
Southeast Region Bioinformatics

RStudio Panes

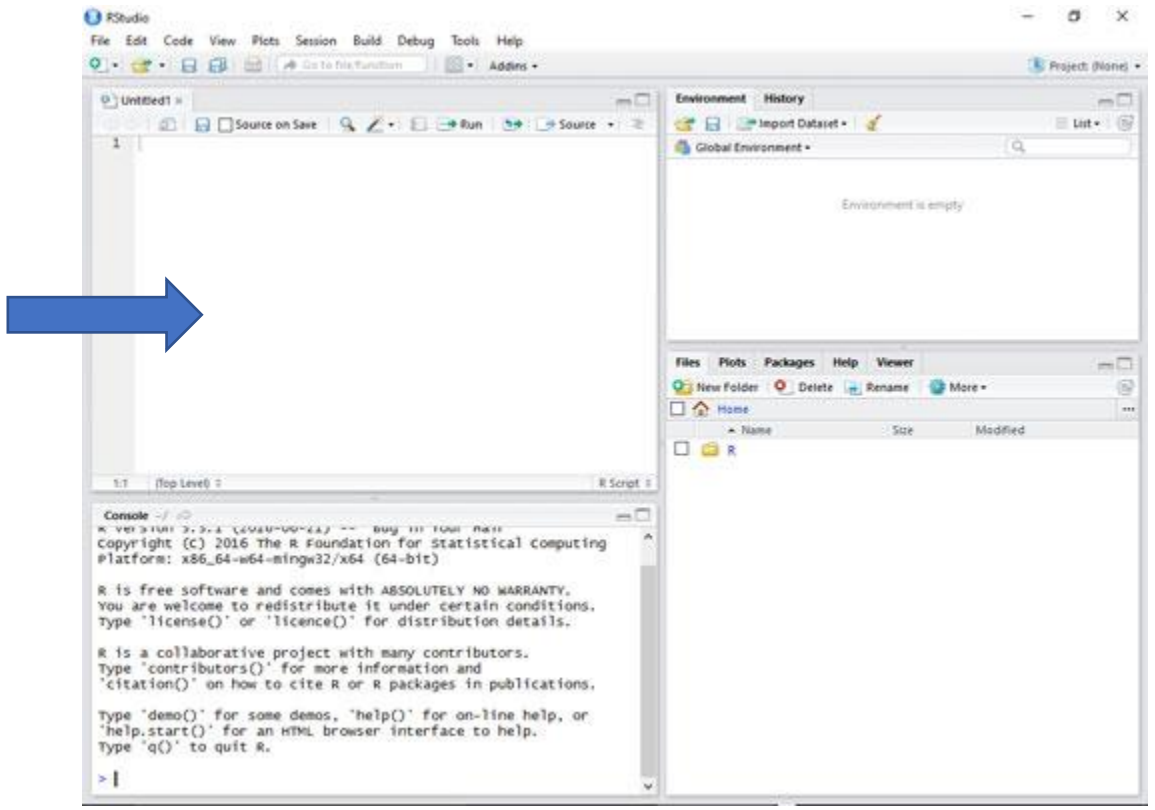
RStudio has four main panes each in a quadrant of your screen which include:

1. Source Editor
2. Console
3. Workspace Browser (and History)
4. Plots (and Files, Packages, Help)

Advanced Molecular Detection
Southeast Region Bioinformatics

R Source Editor

- This is the pane on the top left of your screen.
- Helps you open, edit, and execute the programs

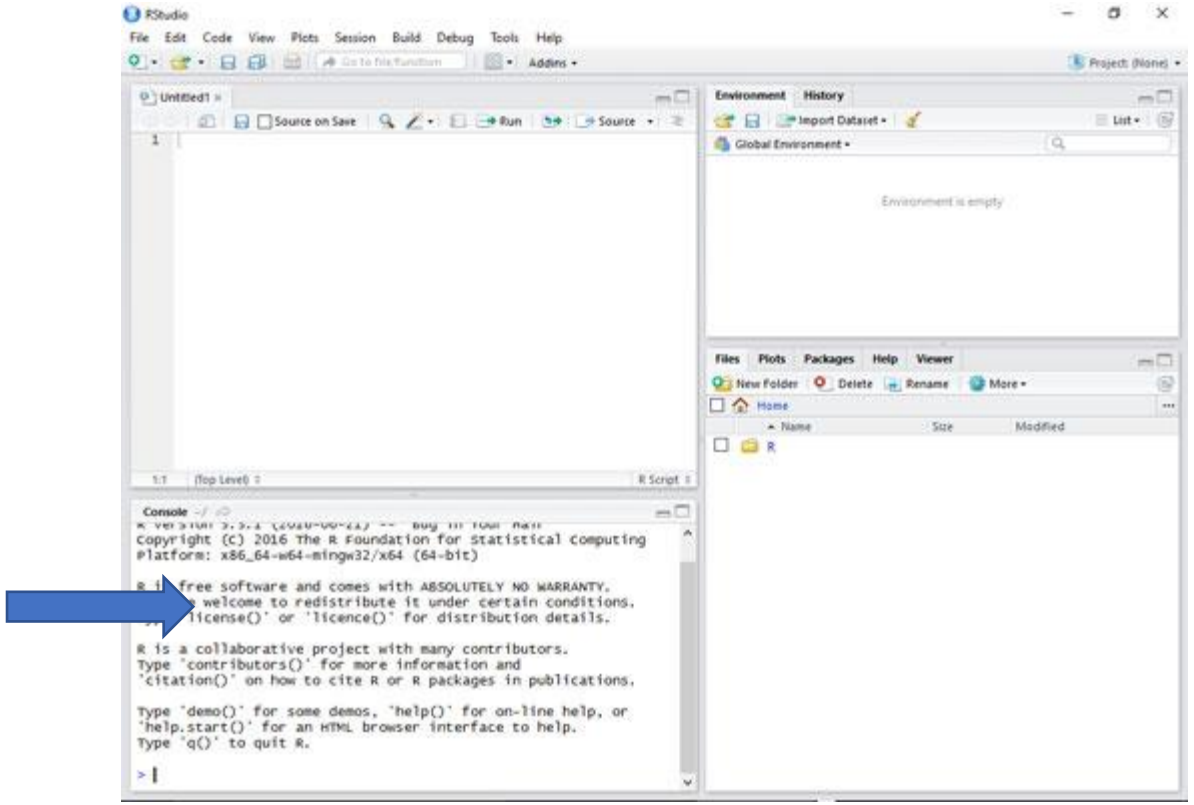


console pane in rstudio - Bing images

Advanced Molecular Detection
Southeast Region Bioinformatics

R Console

- This is the pane on the bottom left of your screen
- Console is where you can type code that executes immediately

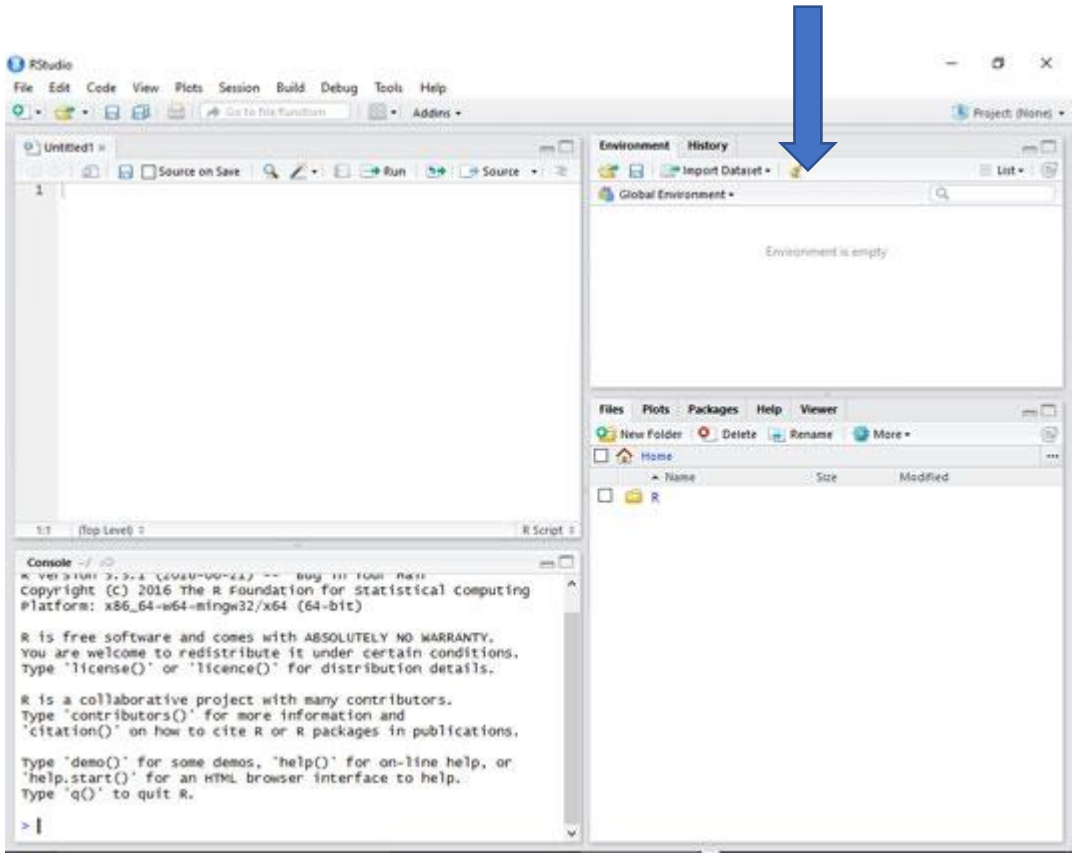


console pane in rstudio - Bing images

Advanced Molecular Detection
Southeast Region Bioinformatics

Environment Pane

- This pane is on the top right of your screen
- Environment pane exhibits what objects (i.e., dataframes, arrays, values and functions) are present in your environment (workspace)



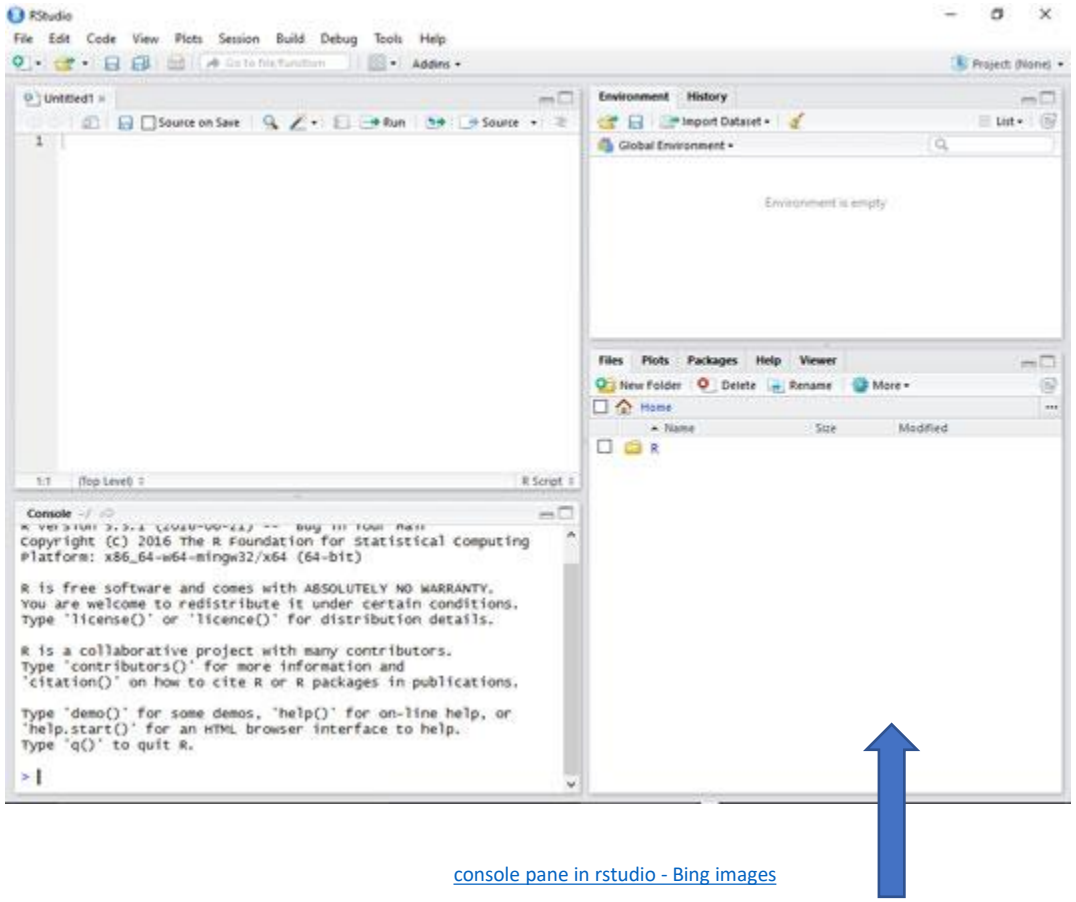
console pane in rstudio - Bing images

Advanced Molecular Detection
Southeast Region Bioinformatics

Fourth Pane

This pane has number of tabs:

- Files tab – has a navigable manager, just like file system on your operating system
- Packages tab – shows packages that are installed and those that can be installed
- Plot tab – graphics you create will appear
- Help tab – allows you to search the R documentation for help

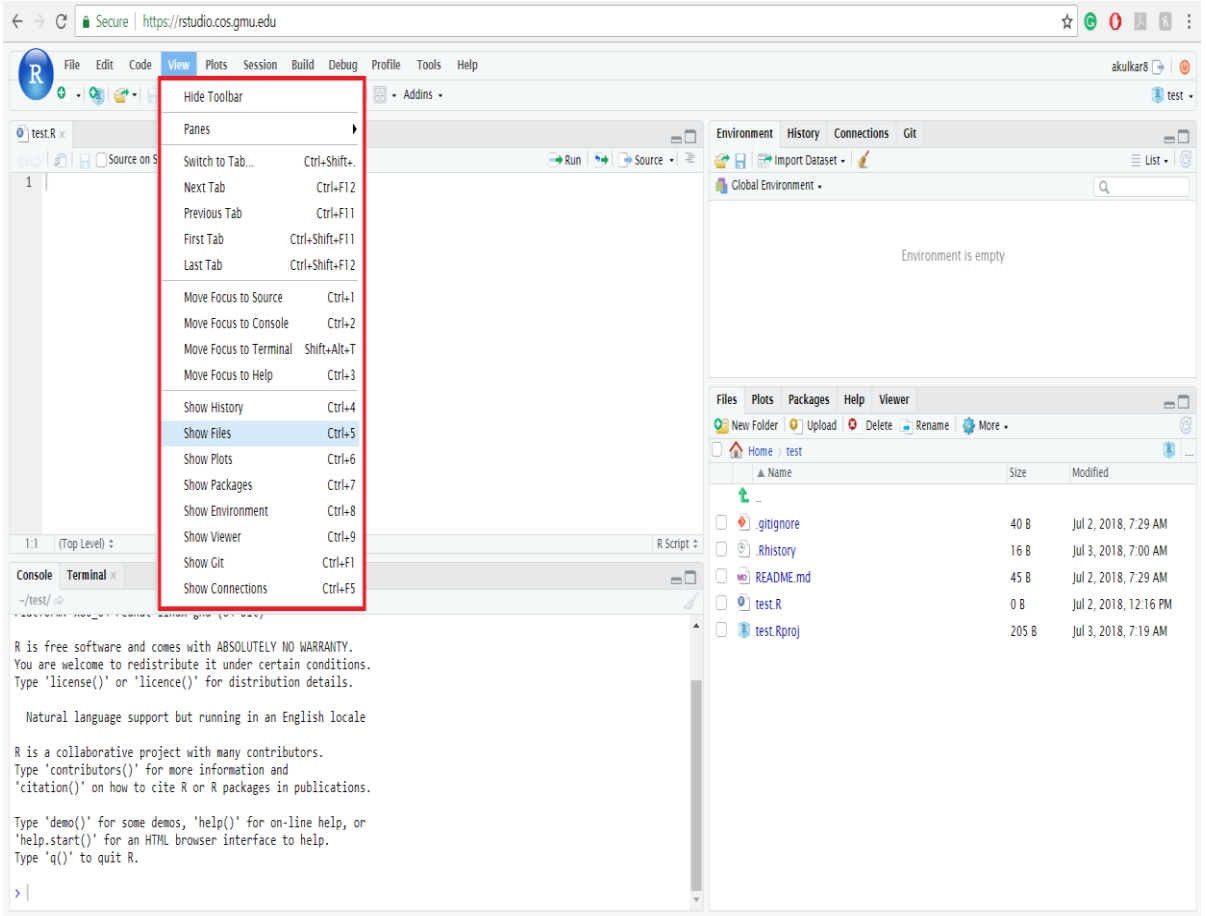


console pane in rstudio - Bing images

Advanced Molecular Detection
Southeast Region Bioinformatics

Rstudio Panes

- Rstudio panes layout can be adjusted as per your wish
- Click view and select Panes to adjust the layout



RStudio view tab images - Bing images

Advanced Molecular Detection
Southeast Region Bioinformatics

Packages in R

- R packages are a collection of R functions, compiled code and sample data
- They are stored under a directory called “library” in the R environment
- By default, R installs a set of packages during installation. More packages can be added later when needed
- Packages which are already installed have to be loaded explicitly to be used by the R
- E.g. ggplot, tidyverse, dplyr, ggtree, plotly

Advanced Molecular Detection
Southeast Region Bioinformatics

Difference Between R & Other Languages

Python:

- Python & R are interpreted, dynamically typed programming languages with duck typing that can be extended by importing packages
- Python is a general-purpose programming language while R is specifically designed for doing statistical analysis

Stata:

- Stata & R are designed to be easily extendable, outputs from both can become inputs for further analysis
- R is free software while Stata is not

Advanced Molecular Detection
Southeast Region Bioinformatics

Difference Between R & Other Languages

SAS:

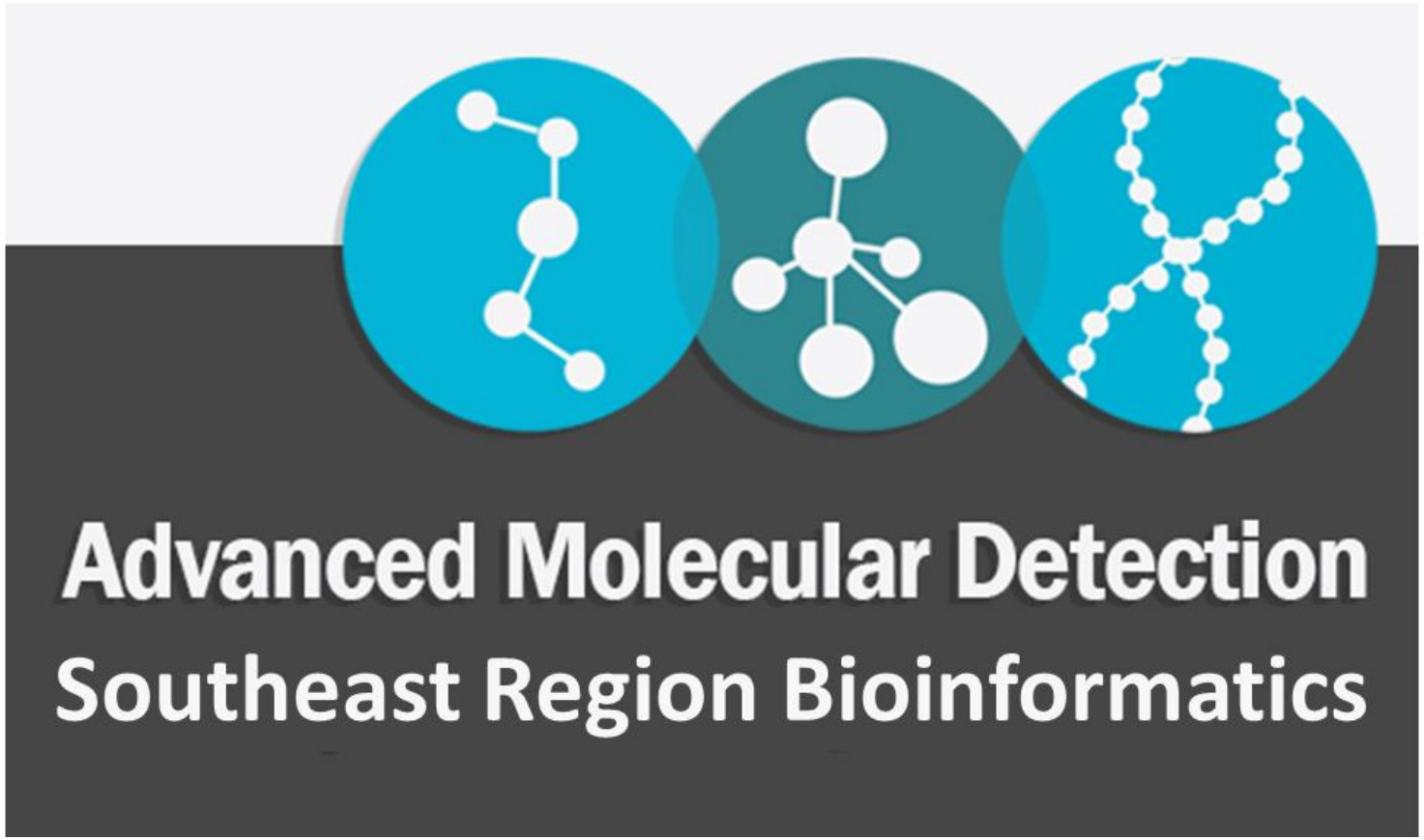
- SAS can only store data in rectangular data sets while R's more versatile data structures allow it to perform difficult analysis more flexibly
- Completely integrating functions in SAS requires a developer's kit but, in R, user defined functions are already on equal footing with provided functions
- R is more convenient for periodic reports but most of them prefer SAS for big data problems
- SAS is expensive

Advanced Molecular Detection
Southeast Region Bioinformatics

References to Learn R

- Welcome | R for Data Science (had.co.nz)
- R-intro.pdf (r-project.org)
- R for applied epidemiology and public health | The Epidemiologist R Handbook (epirhandbook.com)

Advanced Molecular Detection
Southeast Region Bioinformatics



Questions?

bphl-sebioinformatics@flhealth.gov

Molly Mitchell, PhD
Bioinformatician
Molly.Mitchell@flhealth.gov

Sam Marcellus, MPH
Bioinformatician
Samantha.Marcellus@flhealth.gov

Lakshmi Thsaliki, MS
Bioinformatician
Lakshmi.Thsaliki@flhealth.gov