Introduction_to_R

Johanna Jantzen

September 30, 2020

Welcome to “Understanding taxonomy,” a module in the “Using phylogenies to study trait evolution” series.

In this module, we will learn how to download a list of taxa and their corresponding Taxon IDs from an online database: the Open Tree of Life.

In future modules, we will use these taxa to create phylogenies and study how morphological or ecological traits have evolved for these taxa.

Before completing this module, it is best if you have completed the “XXXXX” module which gives you an introduction to R Studio, R Projects, running scripts, and troubleshooting.

This course has several streams and several difficulty levels. Please choose the appropriate datafiles and scripts according to your goals.

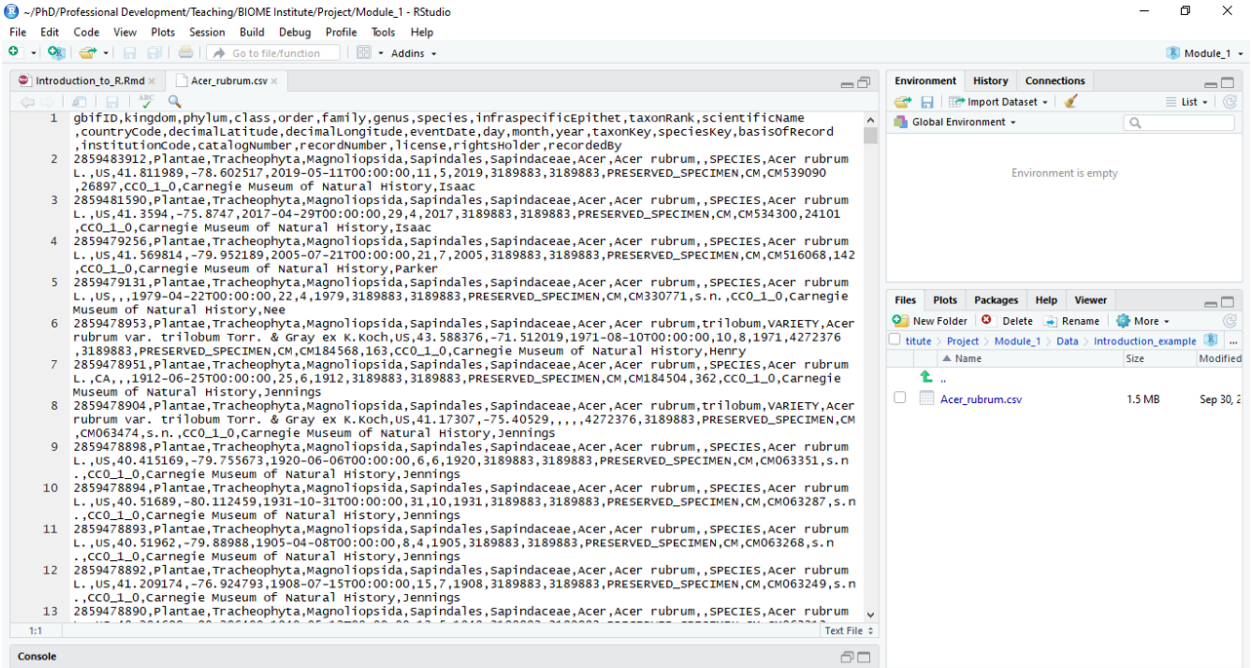
Data for each module are generally included as **csv** files. These files are able to be opened manually or using an R script. Each **csv** file contains columns separated by **commas** (,) and rows separated by **new line characters**.

Navigate to the “Data” folder in the “Files” tab (bottom right corner of the screen) and within that, to the “Introduction_example” folder. Open the file “Acer_rubrum.csv” **manually** by clicking on it and selecting “View File” and check it out. It should look like this:

Now, open the same file using **R**. To run the code below, select the row(s) of text and hit the “Run” button above and “Run Selected Lines”.

```r
example_file <- read.csv("../Data/Introduction_example/Acer_rubrum.csv",
    stringsAsFactors = FALSE, header = TRUE)
head(example_file)
```

```
##       gbifID kingdom       phylum         class      order      family
    genus
## 1 2859483912 Plantae Tracheophyta Magnoliopsida Sapindales Sapindaceae
    Acer
## 2 2859481590 Plantae Tracheophyta Magnoliopsida Sapindales Sapindaceae
    Acer
## 3 2859479256 Plantae Tracheophyta Magnoliopsida Sapindales Sapindaceae
    Acer
## 4 2859479131 Plantae Tracheophyta Magnoliopsida Sapindales Sapindaceae
    Acer
## 5 2859478953 Plantae Tracheophyta Magnoliopsida Sapindales Sapindaceae
    Acer
## 6 2859478951 Plantae Tracheophyta Magnoliopsida Sapindales Sapindaceae
    Acer
##       species infraspecificEpithet taxonRank
## 1 Acer rubrum                        SPECIES
## 2 Acer rubrum                        SPECIES
```


~/PhD/Professional Development/Teaching/BIOME Institute/Project/Module_1 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Module_1
Introduction_to_R.Rmd
Acer_rubrum.csv
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder
Delete
Rename
More
titute > Project > Module_1 > Data > Introduction_example
Name
Size
Modified
Acer_rubrum.csv
1.5 MB
Sep 30, 2
1:1
Text File
Console

```
gbifID,kingdom,phylum,class,order,family,genus,species,infraspecificEpithet,taxonRank,scientificName,countryCode,decimalLatitude,decimalLongitude,eventDate,day,month,year,taxonKey,speciesKey,basisOfRecord,institutionCode,catalogNumber,recordNumber,license,rightsHolder,recordedBy
2859483912,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum L.,US,41.811989,-78.602517,2019-05-11T00:00:00,11,5,2019,3189883,3189883,PRESERVED_SPECIMEN,CM,CM539090,26897,CC0_1_0,Carnegie Museum of Natural History,Isaac
2859481590,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum L.,US,41.3594,-75.8747,2017-04-29T00:00:00,29,4,2017,3189883,3189883,PRESERVED_SPECIMEN,CM,CM534300,24101,CC0_1_0,Carnegie Museum of Natural History,Isaac
2859479256,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum L.,US,41.569814,-79.952189,2005-07-21T00:00:00,21,7,2005,3189883,3189883,PRESERVED_SPECIMEN,CM,CM516068,142,CC0_1_0,Carnegie Museum of Natural History,Parker
2859479131,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum L.,US,,,1979-04-22T00:00:00,22,4,1979,3189883,3189883,PRESERVED_SPECIMEN,CM,CM330771,s.n.,CC0_1_0,Carnegie Museum of Natural History,Nee
2859478953,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,trilobum,VARIETY,Acer rubrum var. trilobum Torr. & Gray ex K.Koch,US,43.588376,-71.512019,1971-08-10T00:00:00,10,8,1971,4272376,3189883,PRESERVED_SPECIMEN,CM,CM184568,163,CC0_1_0,Carnegie Museum of Natural History,Henry
2859478951,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum L.,CA,,,1912-06-25T00:00:00,25,6,1912,3189883,3189883,PRESERVED_SPECIMEN,CM,CM184504,362,CC0_1_0,Carnegie Museum of Natural History,Jennings
2859478904,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,trilobum,VARIETY,Acer rubrum var. trilobum Torr. & Gray ex K.Koch,US,41.17307,-75.40529,,,,,4272376,3189883,PRESERVED_SPECIMEN,CM,CM063474,s.n.,CC0_1_0,Carnegie Museum of Natural History,Jennings
2859478898,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum L.,US,40.415169,-79.755673,1920-06-06T00:00:00,6,6,1920,3189883,3189883,PRESERVED_SPECIMEN,CM,CM063351,s.n.,CC0_1_0,Carnegie Museum of Natural History,Jennings
2859478894,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum L.,US,40.51689,-80.112459,1931-10-31T00:00:00,31,10,1931,3189883,3189883,PRESERVED_SPECIMEN,CM,CM063287,s.n.,CC0_1_0,Carnegie Museum of Natural History,Jennings
2859478893,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum L.,US,40.51962,-79.88988,1905-04-08T00:00:00,8,4,1905,3189883,3189883,PRESERVED_SPECIMEN,CM,CM063268,s.n.,CC0_1_0,Carnegie Museum of Natural History,Jennings
2859478892,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum L.,US,41.209174,-76.924793,1908-07-15T00:00:00,15,7,1908,3189883,3189883,PRESERVED_SPECIMEN,CM,CM063249,s.n.,CC0_1_0,Carnegie Museum of Natural History,Jennings
2859478890,Plantae,Tracheophyta,Magnoliopsida,Sapindales,Sapindaceae,Acer,Acer rubrum,,SPECIES,Acer rubrum
```

Figure 1: Image of csv file as opened manually

```
## 3 Acer rubrum                        SPECIES
## 4 Acer rubrum                        SPECIES
## 5 Acer rubrum             trilobum   VARIETY
## 6 Acer rubrum                        SPECIES
##                                     scientificName countryCode
    decimalLatitude
## 1                                    Acer rubrum L.          US
    41.81199
## 2                                    Acer rubrum L.          US
    41.35940
## 3                                    Acer rubrum L.          US
    41.56981
## 4                                    Acer rubrum L.          US
                NA
## 5 Acer rubrum var. trilobum Torr. & Gray ex K.Koch          US
    43.58838
## 6                                    Acer rubrum L.          CA
                NA
##   decimalLongitude           eventDate day month year taxonKey speciesKey
## 1        -78.60252 2019-05-11T00:00:00  11     5 2019  3189883    3189883
## 2        -75.87470 2017-04-29T00:00:00  29     4 2017  3189883    3189883
## 3        -79.95219 2005-07-21T00:00:00  21     7 2005  3189883    3189883
## 4               NA 1979-04-22T00:00:00  22     4 1979  3189883    3189883
## 5        -71.51202 1971-08-10T00:00:00  10     8 1971  4272376    3189883
## 6               NA 1912-06-25T00:00:00  25     6 1912  3189883    3189883
##        basisOfRecord institutionCode catalogNumber recordNumber license
## 1 PRESERVED_SPECIMEN              CM      CM539090        26897 CC0_1_0
## 2 PRESERVED_SPECIMEN              CM      CM534300        24101 CC0_1_0
## 3 PRESERVED_SPECIMEN              CM      CM516068          142 CC0_1_0
## 4 PRESERVED_SPECIMEN                 CM       CM330771          s.n. CC0_1_0
## 5 PRESERVED_SPECIMEN                 CM       CM184568           163 CC0_1_0
## 6 PRESERVED_SPECIMEN                 CM       CM184504           362 CC0_1_0
##                         rightsHolder recordedBy
## 1 Carnegie Museum of Natural History      Isaac
## 2 Carnegie Museum of Natural History      Isaac
## 3 Carnegie Museum of Natural History     Parker
## 4 Carnegie Museum of Natural History        Nee
## 5 Carnegie Museum of Natural History      Henry
## 6 Carnegie Museum of Natural History   Jennings
```

We can use several functions to learn more about our dataset. For example, we can use *nrow()* to count how many rows there are in the dataset. Try it now. How many rows are there? What about columns?

```
nrow(example_file)
```

```
## [1] 6688
```

```
ncol(example_file)
```

```
## [1] 27
```

You will notice that we are using the **object** *example_file* to refer to the dataset. **Objects** are easy ways to assign large amounts of information to a single string of letters and/or numbers.

Let's assign a single individual row to a new object: *row1*

```
row1 <- example_file[1,]
```

In this example, we are assigning the **object** *row1* the value of *example_file[1,]* which represents the first row and all columns of the **object** *example_file* using the **operator** <- . Let's see what that looks like. Here, rather than asking to see a subset of the data, as we did above with the *head()* **function**, we can run the **object** itself and see the whole thing.

```
row1
```

```
##       gbifID kingdom       phylum         class      order      family
   genus
## 1 2859483912 Plantae Tracheophyta Magnoliopsida Sapindales Sapindaceae
   Acer
##       species infraspecificEpithet taxonRank scientificName countryCode
## 1 Acer rubrum                        SPECIES  Acer rubrum L.          US
##   decimalLatitude decimalLongitude           eventDate day month year
   taxonKey
## 1        41.81199        -78.60252 2019-05-11T00:00:00  11     5 2019
   3189883
##   speciesKey      basisOfRecord institutionCode catalogNumber recordNumber
## 1    3189883 PRESERVED_SPECIMEN              CM      CM539090        26897
##   license                       rightsHolder recordedBy
## 1 CC0_1_0 Carnegie Museum of Natural History      Isaac
```

Let's see what just the first column would look like.

```
col1 <- example_file[,1]
head(col1)
```

```
## [1] 2859483912 2859481590 2859479256 2859479131 2859478953 2859478951
```

Sometimes it's hard to visualize lots of data at once, which is why the summary tools, viewing a subset, or viewing the structure of the dataset can be useful.

```
str(example_file)
```

```
## 'data.frame':    6688 obs. of  27 variables:
##  $ gbifID              : num  2.86e+09 2.86e+09 2.86e+09 2.86e+09 2.86e+09
   ...
##  $ kingdom             : chr  "Plantae" "Plantae" "Plantae" "Plantae" ...
##  $ phylum              : chr  "Tracheophyta" "Tracheophyta" "Tracheophyta"
   "Tracheophyta" ...
##  $ class               : chr  "Magnoliopsida" "Magnoliopsida" "
   Magnoliopsida" "Magnoliopsida" ...
##  $ order               : chr  "Sapindales" "Sapindales" "Sapindales" "
   Sapindales" ...
##  $ family              : chr  "Sapindaceae" "Sapindaceae" "Sapindaceae" "
   Sapindaceae" ...
##  $ genus               : chr  "Acer" "Acer" "Acer" "Acer" ...
##  $ species             : chr  "Acer rubrum" "Acer rubrum" "Acer rubrum" "
   Acer rubrum" ...
##  $ infraspecificEpithet: chr  "" "" "" "" ...
##  $ taxonRank           : chr  "SPECIES" "SPECIES" "SPECIES" "SPECIES" ...
##  $ scientificName      : chr  "Acer rubrum L." "Acer rubrum L." "Acer
   rubrum L." "Acer rubrum L." ...
##  $ countryCode         : chr  "US" "US" "US" "US" ...
##  $ decimalLatitude     : num  41.8 41.4 41.6 NA 43.6 ...
##  $ decimalLongitude    : num  -78.6 -75.9 -80 NA -71.5 ...
##  $ eventDate           : chr  "2019-05-11T00:00:00" "2017-04-29T00:00:00"
   "2005-07-21T00:00:00" "1979-04-22T00:00:00" ...
##  $ day                 : int  11 29 21 22 10 25 NA 6 31 8 ...
##  $ month               : int  5 4 7 4 8 6 NA 6 10 4 ...
##  $ year                : int  2019 2017 2005 1979 1971 1912 NA 1920 1931
   1905 ...
##  $ taxonKey            : int  3189883 3189883 3189883 3189883 4272376
   3189883 4272376 3189883 3189883 3189883 ...
##  $ speciesKey          : int  3189883 3189883 3189883 3189883 3189883
   3189883 3189883 3189883 3189883 3189883 ...
##  $ basisOfRecord       : chr  "PRESERVED_SPECIMEN" "PRESERVED_SPECIMEN" "
   PRESERVED_SPECIMEN" "PRESERVED_SPECIMEN" ...
##  $ institutionCode     : chr  "CM" "CM" "CM" "CM" ...
##  $ catalogNumber       : chr  "CM539090" "CM534300" "CM516068" "CM330771"
   ...
##  $ recordNumber        : chr  "26897" "24101" "142" "s.n." ...
##  $ license             : chr  "CC0_1_0" "CC0_1_0" "CC0_1_0" "CC0_1_0" ...
##  $ rightsHolder        : chr  "Carnegie Museum of Natural History" "
   Carnegie Museum of Natural History" "Carnegie Museum of Natural History" "
   Carnegie Museum of Natural History" ...
##  $ recordedBy          : chr  "Isaac" "Isaac" "Parker" "Nee" ...
```

Another important element is the concept of **classes** of **objects**. If the data is a number, it is often stored as a *numeric* or *integer* **class** while words or other groups of letters are stored as *character* **classes**. If you look at column *kingdom*, for example, you can see that it has the **class** *character* while the *gbifID* which is composed of numbers, is stored as *numeric.*

I think that is enough of an introduction for now so let's get into this module!