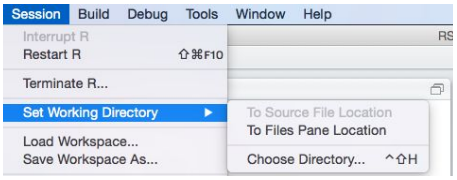
Working in RStudio

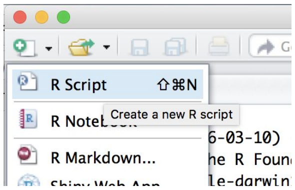
Today, we will be working in RStudio using the statistical software R. R is a program and a programming language. It can be a powerful tool for the analysis of biological data. It can also be very useful for producing informative graphics for your work. Today, you will be working with a samtools *depth* output file, which has **coverage** at every site in the genome.

1. Make a folder on your computer for today's exercise. Download the provided file "chr4.depth.out.zip" to this new folder. Unzip it locally.

2. Set the working directory in RStudio to the new folder you made in Step 1, where you downloaded the file to (see right image).

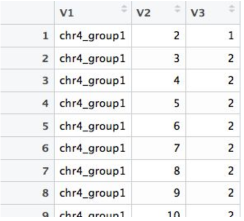


3. At top left, click to open a new R script file. Name your script coverage.R. Type everything into this area and click Run to execute it in your R console below. Save the script when you are done.



4. Load the dataset into R and start exploring it.
   a. Run: `samtools.depth=read.table(file="chr4.depth.out")`
   b. Test the following R functions: head, tail, and length on your data frame.

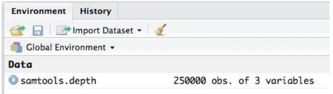
```
head(samtools.depth)
tail(samtools.depth)
length(samtools.depth)
```



	V1	V2	V3
1	chr4_group1	2	1
2	chr4_group1	3	2
3	chr4_group1	4	2
4	chr4_group1	5	2
5	chr4_group1	6	2
6	chr4_group1	7	2
7	chr4_group1	8	2
8	chr4_group1	9	2
9	chr4_group1	10	2

   c. Were you surprised by the output of length? Each column in this file is a vector (default labeled V1..V3). Try specifying a single vector within your data frame to calculate length. Write the output below.

```
length(samtools.depth$V1)
```



samtools.depth 250000 obs. of 3 variables

   d. Use the commands below to select data within this data frame and store in new variables. Explore with other rows and columns in your dataset!

```
col2=samtools.depth[,2]
row13=samtools.depth[13,]
cell2_13=samtools.depth[13,2]
```

   e. Repeat the head, tail and length functions from 4b on your new data frames.

   f. Now, we will use the R function subset to extract a single contig of the file. Confirm your subset by using head, tail, and length.

```
chr4_group5=subset(samtools.depth,V1=="chr4_group5")
```

   g. Try repeating this subset with a different contig. See link for more details.

5. Use summary statistics to explore the file further.
   a. Use the R commands mean and sd to calculate average coverage and standard deviation of coverage in V3. Use the same format as length in 4c.
   b. Write out the results of both commands below:

      Average: Standard deviation:

   c. Now using the extracted contig from 4f, calculate just the coverage of that contig and write the result below. How does this compare to the whole genome?

   d. What are some reasons you would expect coverage to be different in one chromosome or one region of a chromosome as compared to the whole genome?

6. Now let's try some more advanced features of exploring this data graphically.
   a. Make a histogram of the depth in column three. It will display in the bottom right within RStudio (*Select the Plots tab*).

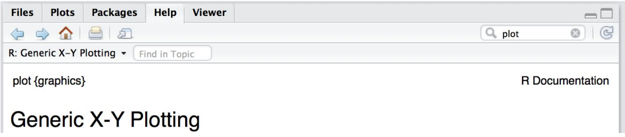
      ```
      hist(samtools.depth$V3)
      ```

   b. Now make a histogram for the contig you subsetted in 4f. How does it look?

   c. Finally let's plot the depth along the contig from 4f.

      ```
      plot(chr4_group5$V2,chr4_group5$V3)
      ```

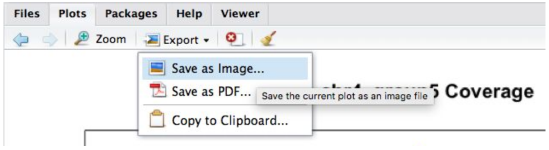
   d. Read about the command plot in help and work to make your graph look nice by adding color, a title and names for the axes that are more informative than the defaults.



plot {graphics} R Documentation

Generic X-Y Plotting

   e. Save the final graph as an image (select PNG). Save your RScript and now you have a script to reproduce this plot and others like it anytime you like!



Congrats on getting started with R and the RStudio environment.
To learn more, start with this excellent R tutorial.