## R and Bioconductor

Bioconductor (http://bioconductor.org/) is a set of R packages for doing bioinformatics.

Bioconductor has lots of other packages. You can browse them via http://bioconductor.org/packages/release/BiocViews.html.

In this assignment we will use the same GFF datafile that you used in Assignment 1. Bioconductor has a package that can parse GFF files.

Next, look at the next page and find and R script for your class. The script is complicated because it will give you an opportunity to learn several scripting techniques and utilities. In Rstudio, create a new R script and type the contents from page two into the script. Note that you will learn better if you go through the effort to type the script instead of copying-and-pasting.

1. Add a comment describing what each line does. You will likely need to consult help pages and reverse engineer the script (delete/change lines to see how it effects the output) to determine what it is doing.
2. Using comment lines at the beginning of the script write a detailed explanation of what the script is does and how it works over all.

## Submitting



Work independently but feel free to discuss your answers with each other and ask for help. There is a lot of resources on Blackboard and in your textbooks. Google is also very helpful to understand what commands do.



When you are done upload your annotated script to Canvas. Remember to include the information/references of any resources you used to complete the assignment (apart from manual pages or help messages).

```r
# This script uses the genomeIntevals package from the Bioconductor
# Project. This package is not installed on the cluster and will need
# to be installed. The following code does this.
#
# NOTE: If you are not using the cluster, you may need to follow the
# following instructions to install Bioconductor on your local machine.
# https://www.bioconductor.org/install/
if (!requireNamespace("genomeIntervals", quietly = TRUE)) {
  BiocManager::install("genomeIntervals")
}

library("genomeIntervals")

url <- paste0("https://downloads.yeastgenome.org/curation/",
              "chromosomal_feature/saccharomyces_cerevisiae.gff")
gff <- basename(url)

if (!file.exists(gff)) {
  download.file(url, destfile = gff, quiet = TRUE)
}

gff_data <- readGff3(gff, quiet=TRUE)
annotation_data <- annotation(gff_data)
gene_data <- subset(annotation_data,type=="gene")
chr_counts <- table(gene_data$seq_name)
chr_data <- as.data.frame(chr_counts)
names(chr_data) <- c("Chr", "Genes")
chr_sorted <- chr_data[rev(order(chr_data$Genes)),]
print(chr_sorted)
```