Setup: Differential expression + functional enrichment analysis

J. Oberstaller & J. Donglasan

11/04/2020

Materials

We'll begin by downloading our exercise materials and setting up a new R-Project for our analysis:

1. Follow this **link** and click 'download' in the upper right corner.

2. Extract the downloaded zip file to where you want it.

3. Open RStudio and click on 'New Project' in the 'File' tab.

4. Create the new project by choosing 'Existing Directory'

5. Browse to the directory where you extracted the zip file and make sure 'R' is the last folder in your project directory file path.

   You should see the folders (Ranalysis, Rdata, etc..) when you open the R folder.

6. Click "create project".

For more info behind the logic of creating RStudio projects and adhering to an organizational directory-structure as you build your data-analysis skills, see **this post on reproducible scientific data-analyses** from Software Carpentry. We don't use exactly the same structure they do, but the concepts are the same: structured analyses make sharing and reproducing analyses much easier!

Pipeline-overview

Goal: The purpose of this analysis is to identify differentially expressed genes from the raw counts-data we generated with featureCounts from an experiment evaluating the effects of drugs on malaria parasites (Justin's RNAseq exercise a few weeks ago).

Input data: We will start with the featureCounts output file–a table of counts-data containing read-counts for all genes across all 4 samples (2 control samples, 2 drug-treated samples).

Before we begin

Before we begin, let's take a moment to get organized. The importance of documentation and good record-keeping are essential to producing high-quality and reproducible computational analyses, just as they are at the bench!

We recommend you keep your analyses organized by project (just as we organized this example). Looking around:

- **Rdata**: this folder contains our input featureCounts data file, and the genome annotation file for P. falciparum (which we'll need to perform our GO enrichment analysis).
- **Ranalysis**: this folder contains any scripts we create to analyze our data. You'll be using the "dge.analysis.R" script.
- **Routput**: we will direct any output data-files from our analyses to this folder.
- **Rfigs**: we will direct any figures we generate from our analyses to this folder.
- **Rsource**: this folder contains any R source-scripts we create to set up our environment for our analyses–custom functions, which packages to load, etc. etc. Don't worry about this for now.

Begin analyses

Now back to RStudio. Open the "Rdata/dge.analysis.R" script and complete the exercise.