Genomic Data Analysis Course Exercises

Carson Stacy

2023-10-26

Contents

Chapter 1

About

This is a compilation of exercises created for a graduate level course in Genomic Data Analysis at the University of Arkansas.

1.1 Usage

Each **bookdown** chapter is an .Rmd file, and each .Rmd file can contain one (and only one) chapter. A chapter *must* start with a first-level heading: `# A good chapter`, and can contain one (and only one) first-level heading.

Use second-level and higher headings within chapters like: `## A short section` or `### An even shorter section`.

The `index.Rmd` file is required, and is also your first book chapter. It will be the homepage when you render the book.

1.2 Render book

You can render the HTML version of this example book without changing anything:

1. Find the **Build** pane in the RStudio IDE, and
2. Click on **Build Book**, then select your output format, or select "All formats" if you'd like to use multiple formats from the same book source files.

Or build the book from the R console:

```
bookdown::render_book()
```

To render this example to PDF as a `bookdown::pdf_book`, you'll need to install XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

1.3 Preview book

As you work, you may start a local server to live preview this HTML book. This preview will update as you edit the book when you save individual .Rmd files. You can start the server in a work session by using the RStudio add-in "Preview book", or from the R console:

```
bookdown::serve_book()
```

Chapter 2

Getting Started in R

last updated: 2023-10-26

2.1 Installing Packages

First things first: Click the "Visual" button in the top-left corner of the code box. This makes the code look more like a word processor. You can always switch back to Source anytime you prefer.

The following code installs a set of R packages used in this document – if not already installed – and then loads the packages into R. Note that we utilize the US CRAN repository, but other repositories may be more convenient according to geographic location.

```
if (!require("pacman")) install.packages("pacman"); library(pacman)

# the p_load function
#    A) installs the package if not installed (like install.packages("package_name")),
#    B) loads the package (equivalent of library(package_name))

p_load("tidyverse", # An ecosystem of packages for making life in R easier
       "here", # For locating files easily
       "knitr", # For generating ("knitting") html or pdf files from .Rmd file
       "readr", # For faster and easier reading in files to R
       "pander", # For session info at the end of the document
       "BiocManager", # For installing Bioconductor R packages
       "dplyr" # A key part of the tidyverse ecosystem, has useful functions
       )
```

2.2 Exercise Description

This activity is intended to familiarize you with using RStudio and the R ecosystem to analyze genomic data

2.3 Learning outcomes

At the end of this exercise, you should be able to:

- open, modify, and knit an Rmd file to a pdf/html output
- relate Rmarkdown to a traditional lab notebook
- run commands in an Rmarkdown file

2.4 Using R and RStudio

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
# print a statement
print("R code in a .Rmd chunk works just like a script")
```

```
## [1] "R code in a .Rmd chunk works just like a script"
```

```
# preform basic calculations
2+2
```

```
## [1] 4
```

R is a useful tool for analyzing data. Let's download a data file from GitHub to work with. First, we will download the file manually and open it. Later, we will download the same file directly from the url.

- Click here to open the file in GitHub and click the download icon to download it to your computer.
- Use the "Import Dataset" in the Environment panel of RStudio to open the file browser and select the downloaded file

- You'll want to use the "From text (readr)..." option
- Adjust settings to make sure the file loads in properly.
- Copy the code that the Import Dataset feature provides for reading in the file and paste it in the code chunk below

```
# insert here the code used to load the file in from your computer
```

2.5 Load data directly from the URL

Rather than downloading the file manually and then loading it in from where we downloaded it to, we can just load it directly from the URL, as shown below. A word of caution, this won't work with any URL and you can't guarantee the URL will always work in the future.

```
# assign url to a variable
DE_data_url <- "https://raw.githubusercontent.com/clstacy/GenomicDataAnalysis_Fa23/main/data/etha

# download the data from the web
DE_results_msn24_EtOH <-
  read_tsv(file=DE_data_url)
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 5756 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (3): Gene ID, Common Name, Annotation
## dbl (15): logFC: YPS606 (WT) EtOH response, Pvalue: YPS606 (WT) EtOH respons...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Do remember that this function uses the package readr (a part of the tidyverse package we loaded above). If you don't have that package (1) installed and (2) loaded into your script, it won't work. Thankfully, the p_load function takes care of both of these simultaneously.

2.6 Working with data in R

To get a quick summary of our data and how it looks

```
# take a quick look at how the data is structured
glimpse(DE_results_msn24_EtOH)
```

```
## Rows: 5,756
## Columns: 18
## $ `Gene ID`                                    <chr> "YMR105C", "YML100W", "YER053~
## $ `Common Name`                                <chr> "PGM2", "TSL1", "PIC2", "NCE1~
## $ Annotation                                   <chr> "Phosphoglucomutase", "Large ~
## $ `logFC: YPS606 (WT) EtOH response`           <dbl> 7.5999973, 7.7618280, 6.69400~
## $ `Pvalue: YPS606 (WT) EtOH response`          <dbl> 9.40e-38, 1.04e-35, 3.03e-39,~
## $ `FDR: YPS606 (WT) EtOH response`             <dbl> 3.26e-35, 1.54e-33, 2.07e-36,~
## $ `logFC: YPS606 msn2/4ΔΔ EtOH response`       <dbl> 0.78481798, 0.60949852, 1.735~
## $ `Pvalue: YPS606 msn2/4ΔΔ  EtOH response`     <dbl> 3.430000e-06, 8.401730e-04, 4~
## $ `FDR: YPS606 msn2/4ΔΔ  EtOH response`        <dbl> 7.420000e-06, 1.398507e-03, 2~
## $ `logFC: WT v msn2/4ΔΔ: EtOH response`        <dbl> -6.815179, -7.152329, -4.9580~
## $ `Pvalue: WT v msn2/4ΔΔ: EtOH response`       <dbl> 6.34e-32, 2.53e-30, 1.35e-27,~
## $ `FDR: WT v msn2/4ΔΔ: EtOH response`          <dbl> 3.65e-28, 7.28e-27, 2.59e-24,~
## $ `logFC: WT v msn2/4ΔΔ: unstressed`           <dbl> -0.144061475, -0.365016862, -~
## $ `Pvalue: WT v msn2/4ΔΔ: unstressed`          <dbl> 0.350436027, 0.041423492, 0.4~
## $ `FDR: WT v msn2/4ΔΔ:unstressed`              <dbl> 0.998531082, 0.998531082, 0.9~
## $ `logFC: WT v msn2/4ΔΔ: EtOH absolute`        <dbl> -6.959241, -7.517346, -5.0845~
## $ `Pvalue: WT v msn2/4ΔΔ: EtOH absolute`       <dbl> 8.55e-37, 2.04e-35, 3.06e-36,~
## $ `FDR: WT v msn2/4ΔΔ: EtOH absolute`          <dbl> 1.64e-33, 1.96e-32, 3.52e-33,~
```

We see in the output there are 5756 rows and 18 columns in the data. The same information should be available in the environment panel of RStudio

Chapter 3

Seeing Data in RStudio

If we want to take a closer look at the data, we have a few options. To see just the first few lines we can run the following command:

```
head(DE_results_msn24_EtOH)
```

```
## # A tibble: 6 x 18
##   `Gene ID` `Common Name` Annotation                      logFC: YPS606 (WT) E~1
##   <chr>     <chr>         <chr>                                           <dbl>
## 1 YMR105C   PGM2          Phosphoglucomutase                               7.60
## 2 YML100W   TSL1          Large subunit of trehalose 6-p~                  7.76
## 3 YER053C   PIC2          Mitochondrial copper and phosp~                  6.69
## 4 YPR149W   NCE102        Protein involved in regulation~                  0.714
## 5 YKL035W   UGP1          UDP-glucose pyrophosphorylase ~                  4.42
## 6 YLR258W   GSY2          Glycogen synthase                                7.52
## # i abbreviated name: 1: `logFC: YPS606 (WT) EtOH response`
## # i 14 more variables: `Pvalue: YPS606 (WT) EtOH response` <dbl>,
## #   `FDR: YPS606 (WT) EtOH response` <dbl>,
## #   `logFC: YPS606 msn2/4ΔΔ EtOH response` <dbl>,
## #   `Pvalue: YPS606 msn2/4ΔΔ  EtOH response` <dbl>,
## #   `FDR: YPS606 msn2/4ΔΔ  EtOH response` <dbl>,
## #   `logFC: WT v msn2/4ΔΔ: EtOH response` <dbl>, ...
```

This can be difficult to look at. For looking at data similar to an Excel file, RStudio allows this by clicking on the name of the data.frame in the top right corner of the IDE. We can also view a file by typing `View(filename)`. To open the data in a new window, click the "pop out" button next to "filter" just above the opened dataset.

3.1 Exploring the data

This dataset includes the log fold changes of gene expression in an experiment testing the ethanol stress response for the YPS606 strain of *S. cerevisiae* and an *msn2/4ΔΔ* mutant. There are also additional columns of metadata about each gene. In later classes, we will cover the details included, but we can already start answering questions.

Using RStudio, answer the following questions:

1. How many genes are included in this study?
2. Which gene has the highest log fold change in the *msn2/4ΔΔ* mutant EtOH response?
3. How many HSP genes are differentially expressed (FDR < 0.01) in unstressed conditions for the mutant?
4. Do the genes with the largest magnitude fold changes have the smallest p-values?
5. Which isoform of phosphoglucomutase is upregulated in response to ethanol stress? Do you think *msn2/4* is responsible for this difference?

Be sure to knit this file into a pdf or html file once you're finished.

System information for reproducibility:

```
pander::pander(sessionInfo())
```

R version 4.3.1 (2023-06-16)

Platform: aarch64-apple-darwin20 (64-bit)

locale: en_US.UTF-8||en_US.UTF-8||en_US.UTF-8||C||en_US.UTF-8||en_US.UTF-8

attached base packages: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *BiocManager(v.1.30.22)*, *pander(v.0.6.5)*, *knitr(v.1.44)*, *here(v.1.0.1)*, *lubridate(v.1.9.3)*, *forcats(v.1.0.0)*, *stringr(v.1.5.0)*, *dplyr(v.1.1.3)*, *purrr(v.1.0.2)*, *readr(v.2.1.4)*, *tidyr(v.1.3.0)*, *tibble(v.3.2.1)*, *ggplot2(v.3.4.4)*, *tidyverse(v.2.0.0)* and *pacman(v.0.5.1)*

loaded via a namespace (and not attached): *utf8(v.1.2.3)*, *generics(v.0.1.3)*, *stringi(v.1.7.12)*, *hms(v.1.1.3)*, *digest(v.0.6.33)*, *magrittr(v.2.0.3)*, *evaluate(v.0.22)*, *grid(v.4.3.1)*, *timechange(v.0.2.0)*, *bookdown(v.0.36)*, *fastmap(v.1.1.1)*, *rprojroot(v.2.0.3)*, *fansi(v.1.0.5)*, *scales(v.1.2.1)*, *cli(v.3.6.1)*,

rlang(v.1.1.1), *crayon(v.1.5.2)*, *bit64(v.4.0.5)*, *munsell(v.0.5.0)*, *withr(v.2.5.1)*, *yaml(v.2.3.7)*, *parallel(v.4.3.1)*, *tools(v.4.3.1)*, *tzdb(v.0.4.0)*, *colorspace(v.2.1-0)*, *curl(v.5.1.0)*, *vctrs(v.0.6.4)*, *R6(v.2.5.1)*, *lifecycle(v.1.0.3)*, *bit(v.4.0.5)*, *vroom(v.1.6.4)*, *pkgconfig(v.2.0.3)*, *pillar(v.1.9.0)*, *gtable(v.0.3.4)*, *glue(v.1.6.2)*, *Rcpp(v.1.0.11)*, *xfun(v.0.40)*, *tidyselect(v.1.2.0)*, *rstudioapi(v.0.15.0)*, *htmltools(v.0.5.6.1)*, *rmarkdown(v.2.25)* and *compiler(v.4.3.1)*

Chapter 4

Cross-references

Cross-references make it easier for your readers to find and link to elements in your book.

4.1 Chapters and sub-chapters

There are two steps to cross-reference any heading:

1. Label the heading: `# Hello world {#nice-label}`.
    - Leave the label off if you like the automated heading generated based on your heading title: for example, `# Hello world` = `# Hello world {#hello-world}`.
    - To label an un-numbered heading, use: `# Hello world {-#nice-label}` or `{# Hello world .unnumbered}`.
2. Next, reference the labeled heading anywhere in the text using `\@ref(nice-label)`; for example, please see Chapter 4.
    - If you prefer text as the link instead of a numbered reference use: any text you want can go here.

4.2 Captioned figures and tables

Figures and tables *with captions* can also be cross-referenced from elsewhere in your book using `\@ref(fig:chunk-label)` and `\@ref(tab:chunk-label)`, respectively.

See Figure 4.1.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

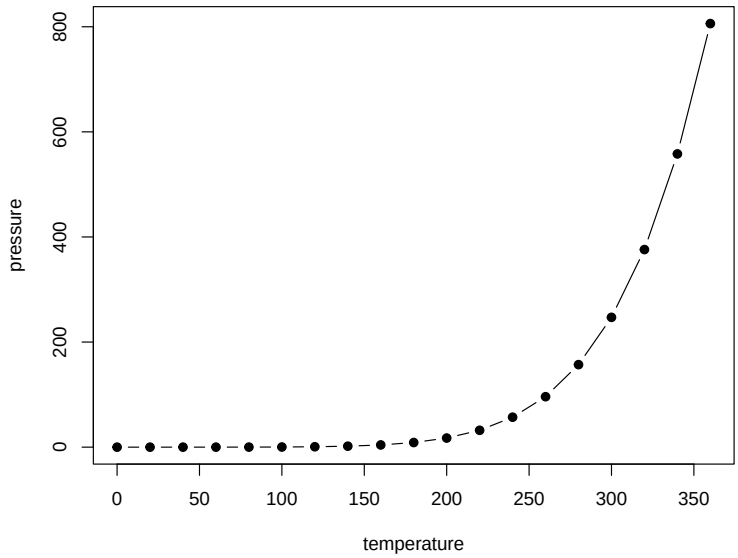

Figure 4.1: Here is a nice figure!

Don't miss Table 4.1.

```
knitr::kable(
  head(pressure, 10), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

Table 4.1: Here is a nice table!

temperature	pressure
0	0.0002
20	0.0012
40	0.0060
60	0.0300
80	0.0900
100	0.2700
120	0.7500
140	1.8500
160	4.2000
180	8.8000

Chapter 5

Parts

You can add parts to organize one or more book chapters together. Parts can be inserted at the top of an .Rmd file, before the first-level chapter heading in that same file.

Add a numbered part: `# (PART) Act one {-}` (followed by `# A chapter`)

Add an unnumbered part: `# (PART\*) Act one {-}` (followed by `# A chapter`)

Add an appendix as a special kind of un-numbered part: `# (APPENDIX) Other stuff {-}` (followed by `# A chapter`). Chapters in an appendix are prepended with letters instead of numbers.

Chapter 6

Footnotes and citations

6.1 Footnotes

Footnotes are put inside the square brackets after a caret `^[]`. Like this one [1].

6.2 Citations

Reference items in your bibliography file(s) using `@key`.

For example, we are using the **bookdown** package [Xie, 2023] (check out the last code chunk in index.Rmd to see how this citation key was added) in this sample book, which was built on top of R Markdown and **knitr** [Xie, 2015] (this citation was added manually in an external file book.bib). Note that the `.bib` files need to be listed in the index.Rmd with the YAML `bibliography` key.

The RStudio Visual Markdown Editor can also make it easier to insert citations: https://rstudio.github.io/visual-markdown-editing/#/citations

[1] This is a footnote.

Chapter 7

Blocks

7.1 Equations

Here is an equation.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{7.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (7.1).

7.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 7.1.

Theorem 7.1. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

7.3 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Chapter 8

Sharing your book

8.1 Publishing

HTML books can be published online, see: https://bookdown.org/yihui/bookdown/publishing.html

8.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

8.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `gitbook` uses the same social sharing data across all chapters in your book- all links shared will look the same.

Specify your book's source repository on GitHub using the `edit` key under the configuration options in the `_output.yml` file, which allows users to suggest an edit by linking to a chapter's source file.

Read more about the features of this output format here:

https://pkgs.rstudio.com/bookdown/reference/gitbook.html

Or use:

```
?bookdown::gitbook
```

Chapter 9

Hello bookdown

All chapters start with a first-level heading followed by your chapter title, like the line above. There should be only one first-level heading (`#`) per .Rmd file.

9.1 A section

All chapter sections start with a second-level (`##`) or higher heading followed by your section title, like the sections above and below here. You can have as many as you want within a chapter.

An unnumbered section

Chapters and sections are numbered by default. To un-number a heading, add a `{.unnumbered}` or the shorter `{-}` at the end of the heading, like in this section.

Bibliography

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL http://yihui.org/knitr/. ISBN 978-1498716963.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*, 2023. URL https://github.com/rstudio/bookdown. R package version 0.36.