Procedures and Guidelines

Clerissa Copeland, Jason Pither, and Mathew Vis-Dunbar

2023-02-10

Contents

Welcome

The procedures and guidelines articulated in this document represent accepted standards for the conduct and presentation of student works in the Biology undergraduate curriculum at UBC Okanagan.

These guidelines are modeled on best practices in the life sciences and where necessary, adapted specifically for the biological sciences and student engagement in learning and research.

For Students These are guidelines only. You may be asked to adhere to them directly as part of your coursework or you may be asked to work with a specific implementation of what is suggested here.

For Instructors Any concerns or omissions from these procedures and guidelines can be forwarded to Jason Pither (jason.pither@ubc.ca) or Mathew Vis-Dunbar (mathew.vis-dunbar@ubc.ca)

This is a living document. Expect that content will be added over time and adapted as needs and circumstances change.

Copyright

Please use the following for citing this document

Copeland, C., Pither, J., Vis-Dunbar, M. (2021). *Procedures and Guidelines*. https://ubco-biology.github.io/Procedures-and-Guidelines/

All source files are available at https://github.com/ubco-biology/Procedures-and-Guidelines.

UBCO Biology open materials

This resource is part of a larger project to host UBCO Biology lab materials in an open, accessible format.

All BIOL open materials can be found at https://ubco-biology.github.io/

Why Procedures and Guidelines?

When we talk about procedures and guidelines, we're very much talking about standards and conventions. Standards and conventions allow the products of research to be easily consumed, interpreted, adapted and re-used. For example, the metric system did wonders for standardizing how we measure distances and weight. The chaos that would ensue if every entomologist took specimen measurements with their own system - or determined wingspan from different points of origin across the same species!

Standards and conventions allow us to explore our data and outputs to greater extents than historically possible by helping us leverage computers; computers rely on standards to parse and merge data. For instance, without the standardization of how markers of climate change are recorded, we would be unable to pool the massive amounts of globally collected data that is used to monitor and fight climate change.

These qualities of standards and conventions - easy consumption, interpretation, adaptation, and re-use - are integral to robust, transparent, reproducible research; these qualities underpin the development of a strong evidence base on which to conduct further research and inform practices and policy.

Structure

This book is divided into 4 sections.

- File and Data Management
- Data Presentation
- Writing and Citing
- Glossary

File and data management covers content related to how to organize, name, and store research data. Broadly speaking, this covers all aspects of research data management.

Data presentation covers how data should be presented when summarized or analyzed.

Writing and citing touches on tools and approaches to support properly formatted and cited open publications.

Glossary is a set of standard definitions for concepts that one will encounter throughout the Biology undergraduate curriculum. When possible, course materials will link directly here.

File and Data Management

Chapter 1

File and Data Management

Last updated 2023-02-10

Well-organized data is critical to transparency, reproducibility, and generally maintaining one's sanity when conducting research. When we talk about file and data management, we may be referring to one of many aspects of making our data understandable to others, to a computer, or to our future selves that have succumb to memory lapses. Making data comprehensible is really about well-structured and communicated metadata that is, whenever possible, implemented according to conventions or standards.

So, when we talk about file and data management, broadly speaking, we're talking about

- File naming and file naming conventions
- Directory structures
- Organizing and formatting data at the variable level

Directory structures, being more complicated, bring with them the need to add additional documentation, such as a description of the directory structure and what we might expect to find where. It will also often include more detailed documentation about how to interpret what is inside specific files; one example of this is the data dictionary that describes each of the variables collected for the study.

Organizing and formatting data is a discussion about the best way to sort and parse our data into columns and rows so that we can effectively produce summaries, statistical calculations, and visualizations; a core concept that you will be introduced to in this guidelines document is that of "tidy data".

Chapter 2

File Naming

Last updated 2023-02-10

File naming isn't exactly fun, but it is crucial for organizing, describing, and managing all kinds of work, especially research. So, let's talk about the file naming conventions that you will be expected to use while at UBCO.

We'll start with the rules, we'll then break these rules down and explain the processes behind them.

2.1 Quick Reference

- File names should only contain letters in the English alphabet, numbers from 1-9, dashes -, and underscores _.
- Do not use spaces or special characters, including # % & < > : " / | ? * { } $! ' @ + ` =
- File names should be broken down into components that are separated by underscores _.
- If more than one word is needed in each component, these are separated by dashes -.
- All files should start with your last name and all other components should be meaningful (read on for what it means for a file name to be meaningful!)

There are four variations on how these guidelines are implemented depending on what your file contains.

2.1.1 Lab reports and manuscripts

Format LastName_Project_File-contents_Version.file-type

Example Pither_BIOL116RProject_Manuscript_V0.docx

2.1.2 Figures and plots

Format LastName_Project_Figure-title_Version.file-type

Example Pither_BIOL116RProject_Figure-freq-plot_V1.png

2.1.3 Analysis

Format LastName_Project_Analysis_Version.file-type

Example Pither_BIOL116RProject_Analysis_V0.xlsx

2.1.4 Data

Format LastName_Date_Project_Data-file-description.file-type

Example Pither_20210921_BIOL116RProject_Data.csv

2.2 What's in a name

File names need to achieve two primary goals, they need to make sense to a human reading them and they need to be constructed in a way that allows a computer to parse or process them. That is, file names should be **human interpretable** and **machine readable**. How do we achieve this?

Human interpretable

To be human interpretable, a file name needs to be meaningful. To do this, it needs to convey some basic information to a person reading it. We do this by integrating metadata into the file name. The metadata elements we include are:

- Who created the file
- The date on which it was created
- The project to which it is connected
- The nature of the contents of the file
- If it's been modified
- The type or format of the file

That is, we should be able to look at a file and tell, *who* created it, *when* it was created, *what* it is related to, *what* is inside of it, if it has been *updated*, and what *application* I should expect to be able to open it with. As we'll see shortly, we don't always include a date, and we don't always include information about modifying a file.

All said, that's a fair bit of information to hold in a file name!

Machine readable

What does it mean for a file name to be machine readable or machine interpretable? It means building our file names in such a way that we can easily organize them so that they can be sorted by an application and in a way that makes sense to us. It also means building our names according to set patterns, which can then be parsed along known delimiters. Lastly, it means building our names in such a way that if we move them from one computer to another, from one application to another, or from one operating system to another, the files remain interpretable in exactly the same way.

How do we do this? We avoid special characters and follow conventions.

Special characters

Special characters are all characters **except**:

- Any character that is a part of the English alphabet
- Numbers from `0 - 9`
- Dashes `-`
- Underscores `_`

This means that a space " " is a special character, which means that your file names should not have spaces.

When operating in a multi-lingual or non-English environment, this can prove problematic, but it is an unfortunate legacy of the development of computer standards that has yet to be fully resolved.

Conventions

Convention has file naming proceed in the following order, with each element separated by an underscore `_`, and words within an element joined with a dash `-`. The file type is generally added with a period `.` and is usually automatically generated when an application creates a file.

Element-1	_Element-2	_Element-3	_Element-4	_Element-5	.Element-5
Last-Name	_Date	_Project	_File-Contents	_Version	.File-type

Dates

Dates should be written in the following format `yyyymmdd`. They should contain no spaces, no dashes, no words, just 8 numbers. Months and Days that are from 1 - 9 should be led by a 0. For example, January 23, 2020, should be written **20200123**. When written this way, your computer will always sort your files from the earliest date to the latest date.

Keeping track of dates is especially important for data because the date on which your data was collected has direct relevance. Dates are less important for things like figures because they are derived from previously dated data.

Versions

Version tracking is achieved in file naming by adding _Vn where n is the version number. With each major change, we increase n by 1. So version 1 would read _V1, and when updated, it would read _V2.

Versions are very important for things like manuscripts and interpretations of data, such as figures and other visualizations, which we will continue to change and modify throughout a project. Data, however, while it has a collection date, should not be modified, and should not then be versioned.

2.3 An example

So what does this look like?

Say you're in BIOL 116 and you're working on your research project. Your research project involves:

- Preparing the beginning of a manuscript that states your research question, hypothesis, and proposed methods.
- Conducting your experiment and recording your data. This process might span more than one day.
- Updating your manuscript, describing any changes that were made to your methods.
- Organizing the results of your experiment and interpreting and visualizing your data.

- Updating your manuscript to include your results and your interpretation of these results, including a visual interpretation.
- Completing your manuscript by discussing the importance of and / or limitations of the experiment, and finally producing a conclusion.

In this scenario, we have **1 project**, **1 manuscript**, **1 dataset**, and at least **1 figure**. In addition, our dataset is constructed from data collected over several days, and our manuscript is revised 3 times before final submission.

So, first we will come up with a project name, and then we will **date our data** and **version our figures and manuscript**. And we'll see how this evolves over the course of several days.

Day 1

We create the following file:

Pither_BIOL116RProject_Lab-report_V0.docx

This is our manuscript, so it will get a version, but no date. Looking at it, we quickly see that this is a lab report (Lab-report), authored by someone with the last name Pither (Pither) associated with a BIOL 116 Research Project (BIOL116RProject), that it has only just been created (V0), and that I should expect it to open in Microsoft Word (docx).

Let's imagine that I will put my research question, hypothesis, and methods in this document and submit it.

Day 2

Today, I conducted the first part of our experiment and collected some data. Now we have the following files:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Lab-report_V0.docx
```

We have not changed our manuscript, so there's no change to the name. However, we have collected some data. We can easily see who collected this data (Pither), when it was collected (September 21, 2021), that it's connected to the BIOL 116 Research Project (BIOL116RProject), and that it's data related to PH exposure. Lastly, it is formatted as comma separated values (csv), which can be opened by any spreadsheet program or text editor.

Day 3-5

I continue to collect data over the next several days, and here is what my files now look like:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Lab-report_V0.docx
```

Again, we have not changed our manuscript, so there's no change to the name. However, we have collected some more data related to PH. We have one file for each day, organized from the earliest day of collection to the most recent.

Day 6

Today, I did two things. I have no more data to collect, so I updated my manuscript to include any modifications made to my original methods section, I then submitted this. I also started to analyze my data; to do this, I merged all my data into a single file for analysis. Now my files look like this:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Analysis_V0.xlsx
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
```

At this stage, I have my data collated into a document where I can work on it without impacting the original data. We can see that I have done this in Excel (xlsx), and that I should expect to be able to open this file in Excel. I also now have a V1 of my manuscript, as I have now added a new section to it; when submitting it, my TA knows that the file with V1 should have this updated section.

Day 7

Today, I built two visualizations using the data in my analysis document, one linear regression and one bar plot of frequency counts; I save these as images to insert into my manuscript. I then updated my manuscript to include my results and these two figures and submitted V2 of my manuscript. Now my files look like this:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Analysis_V0.xlsx
Pither_BIOL116RProject_Figure-freq-plot_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V0.png
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
Pither_BIOL116RProject_Lab-report_V2.docx
```

We can start to see the advantage here of naming conventions. I can easily see which files are which, what they contain, and what their timeline of development is. Also, my computer easily sorts these into meaningful categories - my data is grouped together, sorted by date. My analyses, figures, and manuscripts are all respectively grouped and sorted by version.

Day 8

I got feedback that my linear regression model had an error in it. So I fixed this today, added the new figure into my manuscript, and wrote the discussion and conclusion sections. I'm now ready to submit. Here is what my files look like now (I will be submitting V3 of my manuscript):

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Analysis_V0.xlsx
Pither_BIOL116RProject_Figure-freq-plot_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V1.png
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
Pither_BIOL116RProject_Lab-report_V2.docx
Pither_BIOL116RProject_Lab-report_V3.docx
```

Chapter 3

Directories

Last updated 2023-02-10

Directories are just folders; we'll use the terms interchangeably. All of the files on your computer are organized around folders. In some respects, this means that your computer is synonymous with a filing cabinet; you open it up and there are a bunch of folders holding files and sub-folders. Following this analogy, if you were to open your computer–and you were on a Mac–you'd see 16 folders, including the following:

```
Applications/
Users/
home/
Library/
Volumes/
System/
bin/
usr/
```

There's a very good chance you've never seen any of these folders. We call this the root of your file system; these folders don't sit inside of any other folders, they only hold other folders and files.

The folder called `Users` holds all of the files that you create on your computer. In fact, there is a folder in there named after the user name that you use to log into your computer with. And within that a series of folders that you should be fairly familiar with including a `Downloads/` and `Documents/` folder. It also includes a `Desktop/` folder–your desktop is just another folder containing files, but one that has special status in terms of how those files are shown to you–ie, on your desktop when you start up your computer.

When writing about directories, directory names are frequently followed by a slash–`/`–to differentiate them from files.

If we were to represent this graphically–as a hierarchy–we'd have something like this:

```
Users/
    yourUserName/
        Downloads/
        Documents/
        Desktop/
```

3.1 Working directory

Your **working directory** is the folder you're working in or the folder that holds the file that you have open.

Let's say you create a folder on your `Desktop/` called `BIOL-125/` to hold all of your coursework for this class. And in it you have another folder for your research project–`Research-Project/`–in which you have a Word file for keeping notes–`biol-125_research-project_notes.docx`.

A directory map would now look something like the following:

```
Users/
    yourUserName/
        Downloads/
        Documents/
        Desktop/
     |      BIOL-125
     |   |      Research-Project/
     |   |   |        biol-125_research-project_notes.docx
```

When you open your `Research_Project/` folder, we call this your working directory–the directory that holds the files that you are currently working with, or that you currently have access to. If you clicked on your `Desktop/`, your working directory would switch to your `Desktop/`, since we know that your `Desktop/` is just another folder.

This has implications for how the applications that you work with access the files on your computer. If you opened your file `biol-125_research-project_notes.docx`, `Research-Project` would not only be your working directory, it would be Microsoft Word's working directory for that file. With `biol-125_research-project_notes.docx` open, if you went to `File > Open...` in Word, it would prompt you to open a file in the directory `Research-Project`, because this is where the application is looking for working files.

Generally, in day to day life, we don't need to worry about things like working directories. When we conduct research though, and we use tools like `R`, working directories–and knowing where directories and files are in relation to the directory that you are currently in–have significant implications for how things work and for computational reproducibility. In the next section, we'll look at common directory structures used to organize research. For the moment, we'll assume that you have a research project and that associated with that research project you have data and that you're analyzing that data, and that you're using two different folders to hold the files associated with these activities. So a map something like the following:

```
Users/
    yourUserName/
        Downloads/
        Documents/
        Desktop/
     |      BIOL-125
     |   |      Research-Project/
     |   |   |        Data/
     |   |   |        Analysis/
     |   |   |        biol-125_research-project_notes.docx
```

If you have a file in your `Analysis/` folder open, `Analysis/` is your working directory. If this file needs to access a file with data in your `Data/` directory, you need to be explicit about where this file lives–it lives up one level in the hierarchy in another folder called `Data/`.

3.2 Relative and Absolute Paths

When talking about file locations, a path tells us where a file lives within a directory structure. Paths can be absolute or relative. In general, relative paths are preferrable for reproducibility.

Absolute paths

Absolute paths tell us the location of a file relative to the whole directory structure. Using the previous example, the absolute path of `biol-125_research-project_notes.docx` is `/Users/yourUserName/Desktop/BIOL-125/Research-Project/`.

Relative paths

Relative paths on the other hand tell us the location of a file relative a working directory. So, if you're working directory is your Desktop, the relative path of

`biol-125_research-project_notes.docx` is `BIOL-125/Research-Project`.

3.3 Set a Working Directory in RStudio

Preferred Method: Creating a Project

Let's say you're working on an experiment and you'd like to write a reproducible lab report using Markdown in RStudio. The first step is to create a new R project–which creates a file ending in `.Rproj`. A `.Rproj` file is a file that sits in the root directory of your project. When you open your RStudio session via this project file, it automatically tells `R` to use that root folder as your working directory. In other words, it tells R to look to this root folder for any files or folders you refer to within your R Markdown script.

The biggest benefit to this approach for setting a working directory is that if you decide–or need–to move your project's root folder to another location on your computer or transfer your project's root folder to a different computer, all of the file paths specified in your `R` script will still be valid and work, because they will all be relative to your project's root folder.

There are two options to create a new R project:

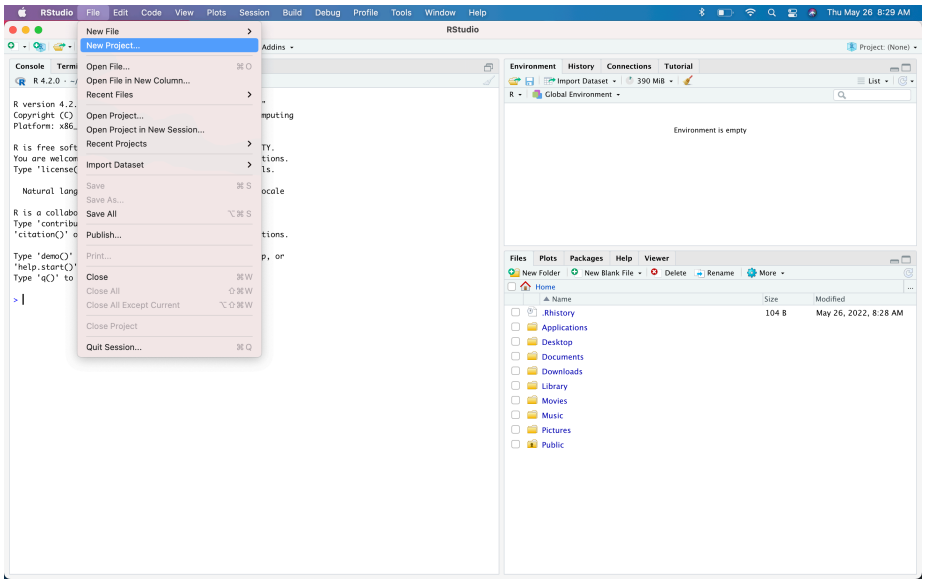
- Create a new project file within a *brand new folder*.
- Create a new project file within an *existing folder*.

So if you have a bunch of files related to your experiment and these are already saved in a folder on your computer, it makes more sense to create a new project within that existing folder. Alternatively, if you are starting from scratch you might prefer to create a new project within a brand new folder.

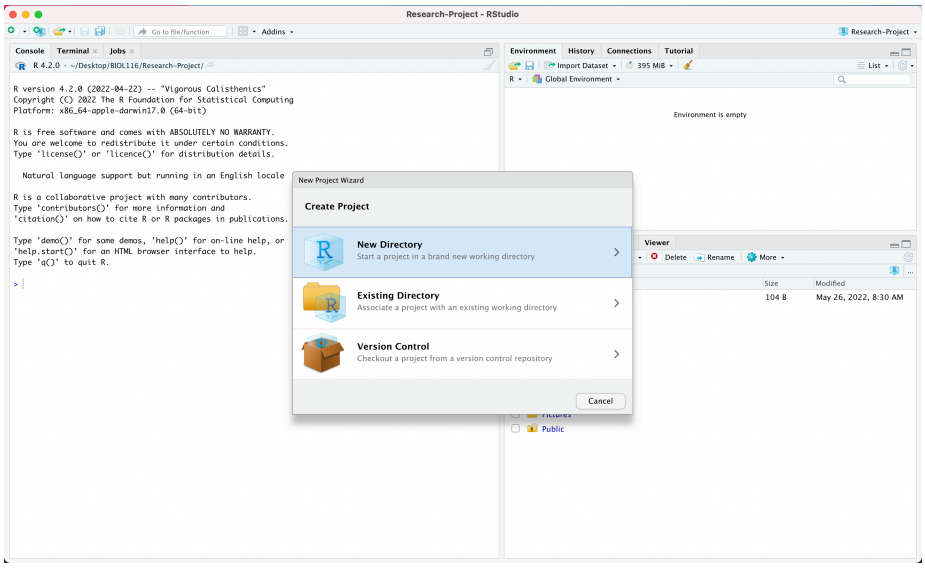
Let's go through an example of how to create a new project with each method.

Create a New Project in a New Folder

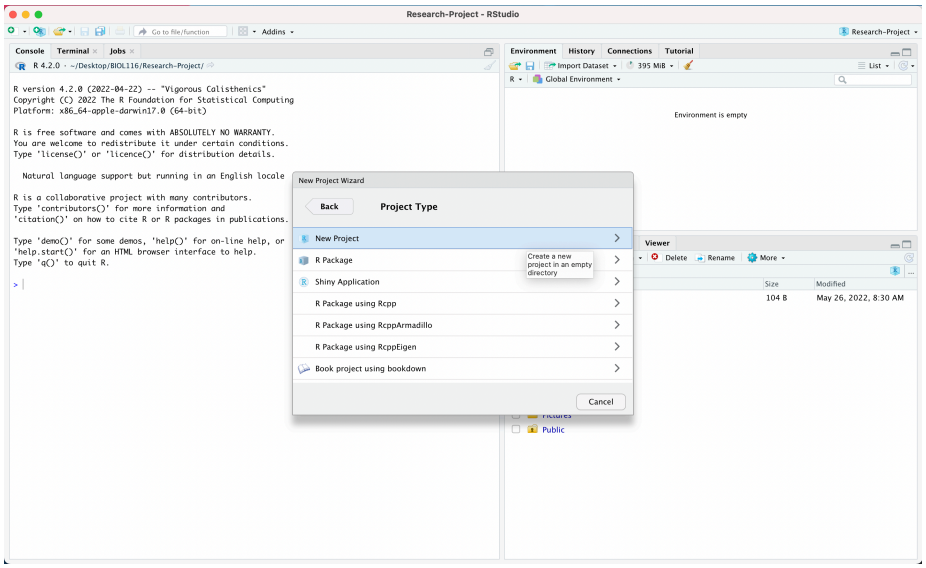
The first step is to open RStudio and select 'New Project' from the 'File' drop down menu.



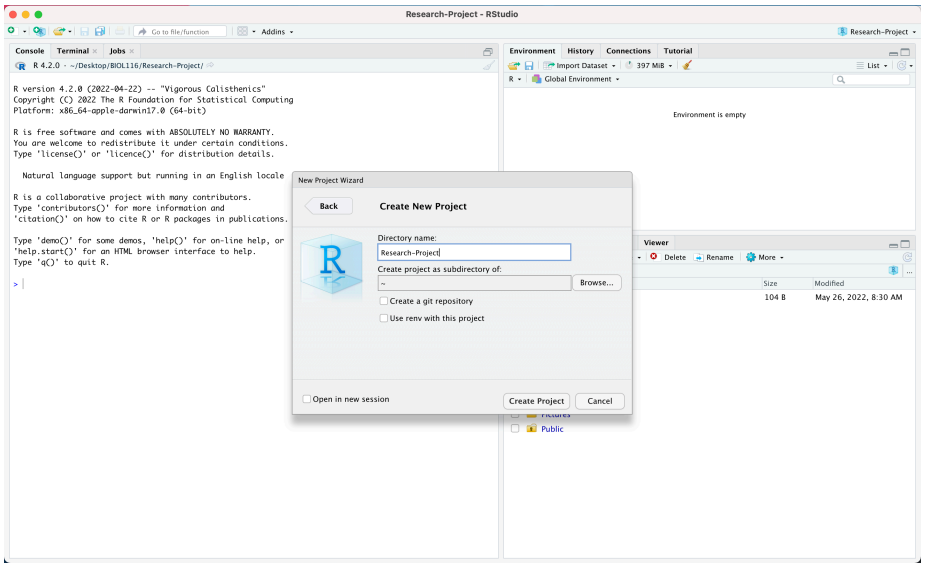
You will then be asked to specify if you'd like to create a 'New Directory' or use an 'Existing Directory'. In this case select 'New Directory'.



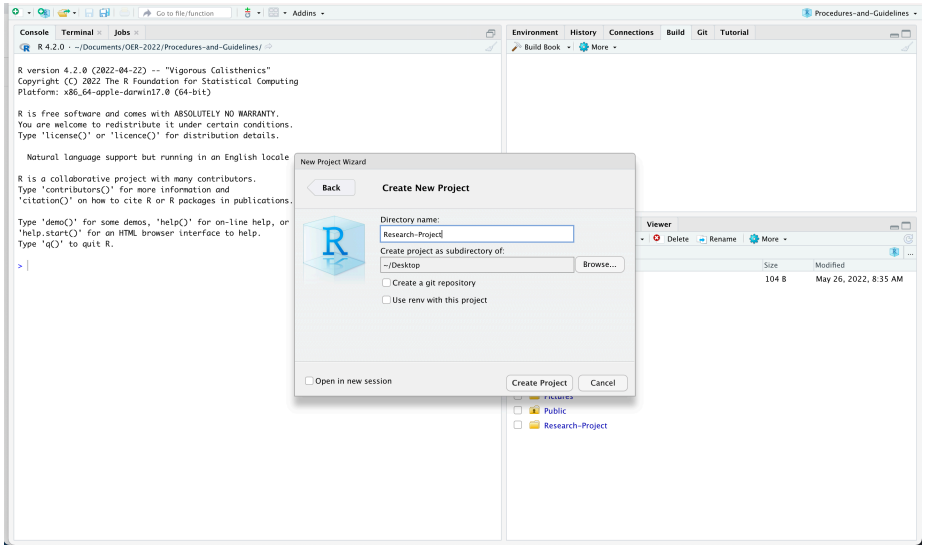
You will then be prompted to choose a project type. Choose 'New Project'.



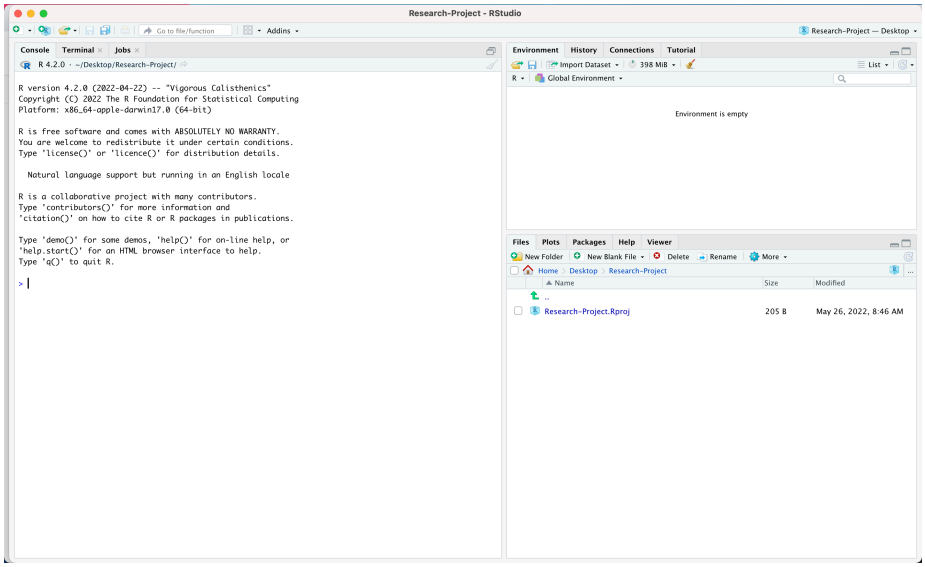
Now name what will be your working directory/root folder according to appropriate naming conventions. Click 'Browse' and select a location to save this folder to on your computer.



Then click 'Create Project'.

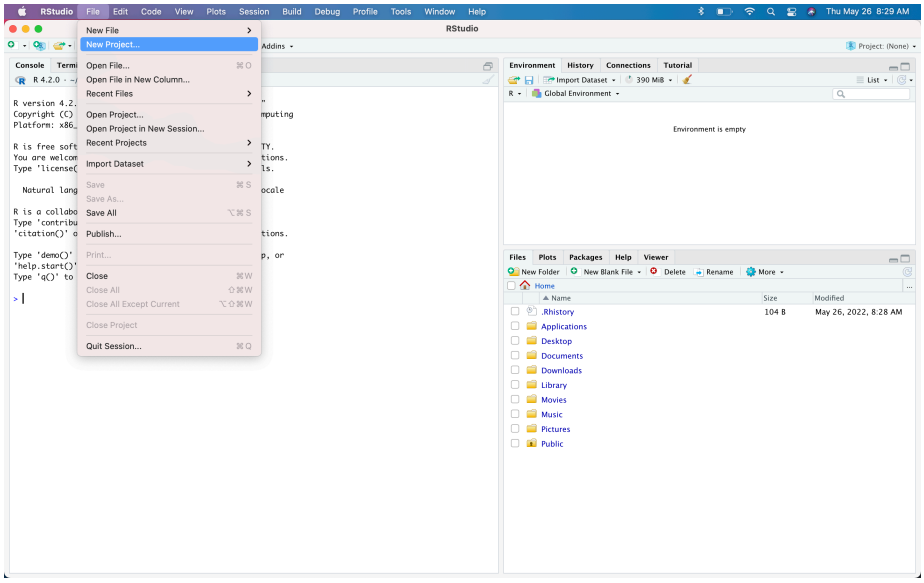


Finally, R will bring you back to the main screen and under the 'Files' region (bottom right panel) you will see your new project file. Now R is automatically using your new folder containing the project file as the working directory. As you create new files associated with your experiment, make sure you save them within the root folder that contains this project file.

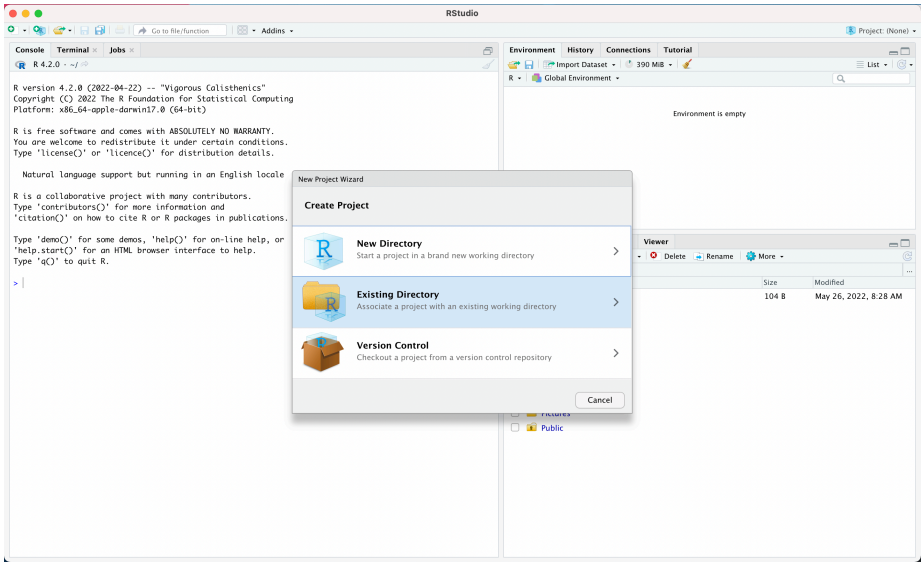


Create a New Project in an Existing Folder

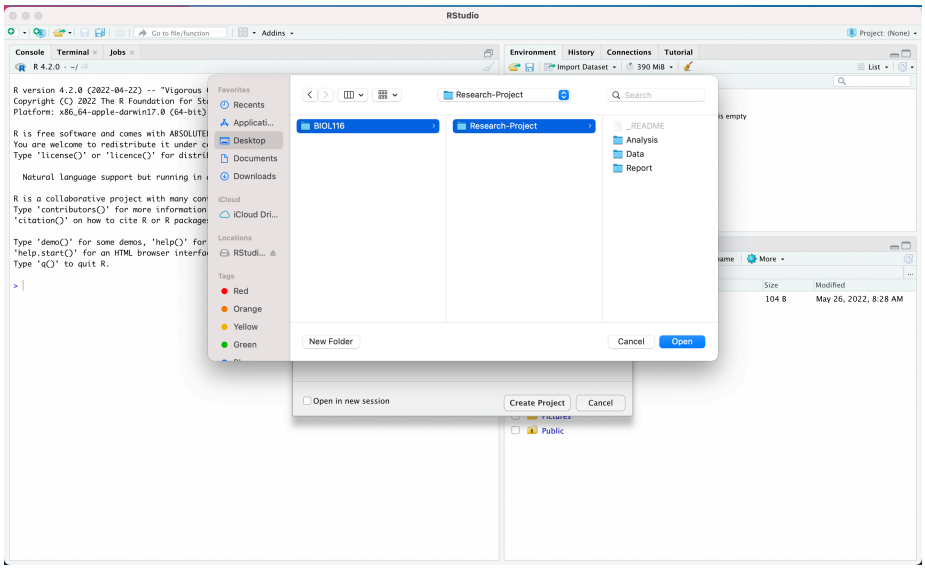
Recall, that the first step is to open RStudio and select 'New Project' from the 'File' drop down menu.



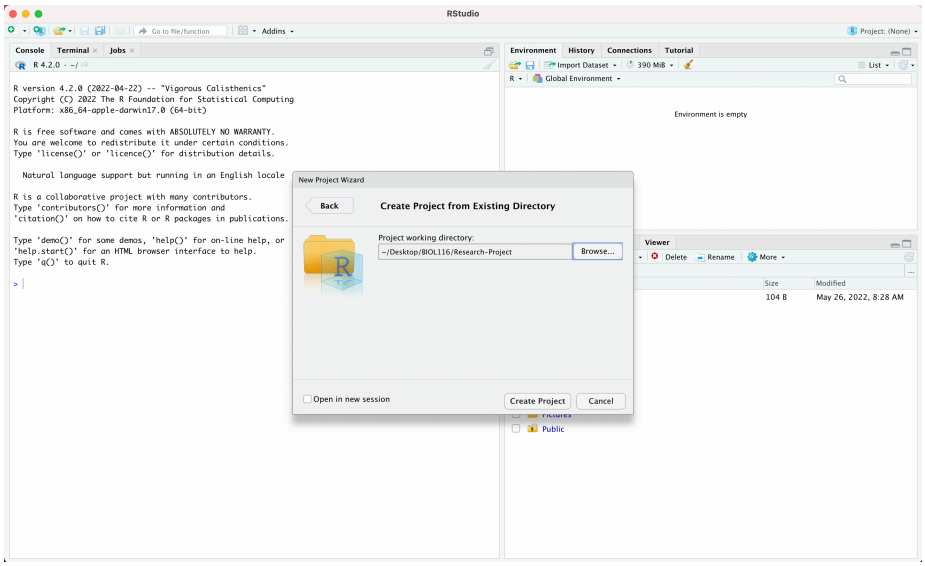
You will then be asked to specify if you'd like to create a 'New Directory' or use an 'Existing Directory'. In this case select 'Existing Directory'.



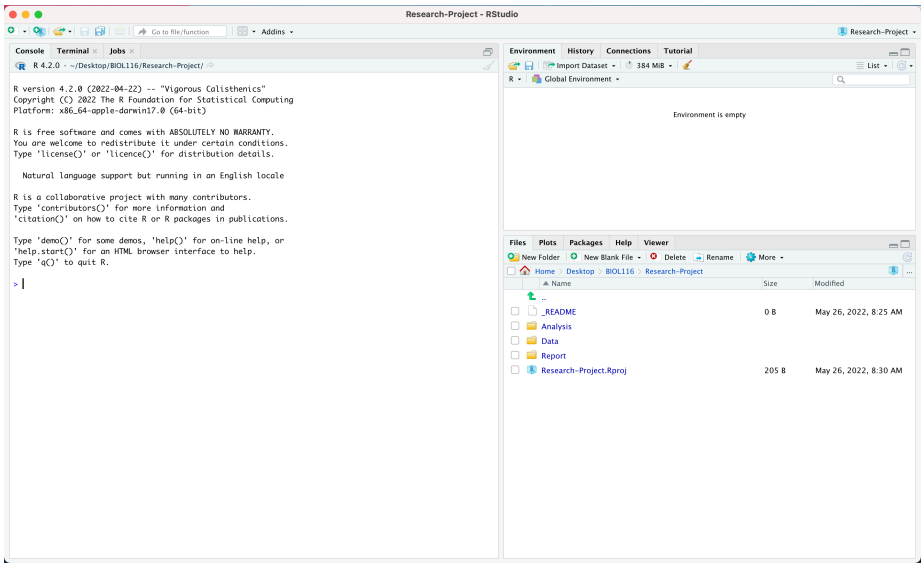
Select the folder from your computer that you'd like to use as your working directory for the project.



Then click 'Create Project'.



Finally, `R` will bring you back to the main screen and under the 'Files' region (bottom right panel) you will see your new project file. Now `R` is automatically using your new folder containing the project file as the working directory. Remember, as you create new files associated with your experiment, make sure you save them within the root folder that contains this project file.



If you close RStudio and want to re-open the project you'll need to tell R that you want to use the root folder with the project file in it as your working directory. To do this, you'll need to locate the project file on your computer and open it, or open RStudio and select 'Open Project' under the 'File' drop down menu.

You may have heard of other methods that can be used to set a working directory in `R`. In fact, there's a built in command–`setwd()`–to set a working directory other than the one currently being used. If you chose this method over using an R Project file, you would need to set your working directory every time you wanted to work on your project, or, alternatively, have the first line of code in your script set your working directory using an absolute path. The downside of using this method is that it relies on extra steps or absolute paths to access your working directory. Consequently, if you move the project's root folder–containing all the files related to your experiment–you'll have to edit your `R` script and change the absolute path. This impacts computational reproducibility. If you send your work to a lab partner or colleague they will have to edit the script, updating the absolute path to match where they've placed the project folder on their computer. An R project negates this problem by leveraging relative paths.

Chapter 4

Directory Structures

Last updated 2023-02-10

Now that we've covered naming conventions and have a sense of how our systems are organized around folder structures, let's pretend that you store everything on your computer in a single folder–some of us are probably known to use our desktops for this. Imagine how long it would take you to find data you collected on a specific day a few years ago. Instead of keeping every document in a single place, we often organize our files using directory or folder structures. This helps us save precious time and improve our productivity.

One major aspect of Open Science is ensuring transparency in the research process. This includes sharing files from all the steps of the research lifecycle (i.e. a priori hypothesis, study design, data, analysis etc.) with others so that our research can be understood and replicated more easily.

The way we currently organize folders and files on our computer may make sense to us, but a problem arises when we need to share those folders and files with other people. A folder name that is meaningful for us may make no sense to another person. So let's cover some conventions to help you organize the files on your computer in a way that is meaningful to both you and others.

4.1 Directory Hierarchies

First, let's talk about how to properly structure a folder hierarchy.

The highest-ranking folder is generally called the **root directory**, or sometimes the top-level folder. We'll call it the root directory here. This folder will contain all of the subfolders and files related to a particular project, including its data, analysis, lab reports etc. It will also contain what is called a readme file.

The structure should look something like the following:

```
  Project-Folder/
|     Experiment-Data/
|   |     File-1
|   |     File-2
|     Experiment-Analysis/
|   |     File-1
|     Experiment-Report/
|   |     File-1
|   |     File-2
```

Here, our root folder is called Project-Folder and it contains three subdirectories, one for data, Experiment-Data, one for analyses, Experiment-Analysis, and one for a report Experiment-Report. Each subdirectory then contains one or many files.

4.2 Directory Naming

Key file naming conventions, such as avoiding special characters, are equally as important for directories as they are for naming files. Remember, we consider special characters to be anything other than letters in the English alphabet, numbers from 1-9, dashes "-", and underscores "_".

In case there is any doubt, here are some examples of what are considered special characters - characters you want to avoid!

% & < > : " / | ? * { } $! ' @ + ` =

Remember: spaces, " ", qualify as special characters when it comes to file naming.

Remember: when we name the root folder we want to clearly communicate what the project is. And similarly, within this root folder, we want clearly labeled subfolders for each relevant aspect of the project. Common discrete subdirectories include ones for figures, data, manuscript etc.

4.3 readme files and data dictionaries

When naming files we embed metadata into our file naming conventions to encode relevant information for the reader. But we can only store so much information in a file name. So we also include three additional files:

- A _README file which resides in our root folder and elaborates on the contents of our folder structure.

- A _README file that resides in our data directory and discusses some of the particulars of the how, where, and who did the actual data collection.
- A _DATA-DICTIONARY file that also resides in our data directory and elaborates on how our data is stored and organized.

These files - containing a brief description of the major folder contents, naming conventions, and data structure - are critical for transparency and reproducibility because they allow others to easily understand the contents of your directory and data without needing to ask you. This is especially helpful when working with a group or sharing directories with others.

General rules

A readme file and data dictionary should:

- Exist in at least two locations, the root directory and the data directory.
- Be prepended with an underscore "_". This will push these files to the top of the directory for easy access.
- _README and _DATA-DICTIONARY files should be in all caps, so they really stand out; this should be the first thing you look at when looking at any directory or folder, as this is your guide to its contents.

File formats

Readme files and data dictionaries should be written in plain text, for this will ensure that the files describing your project can be opened on any computer. You will often see readme files called _README.txt or _DATA-DICTIONARY.txt.

Our example readme and data dictionaries use a plain text format called markdown.

markdown

We recommend that you create your readme files as markdown, a way of formatting plain text files, allowing us to provide additional meaning to our content. For example, in plain text, if we want to emphasize content, we don't really have a way of doing this. In markdown, we can use italics and bolding if needed. We can also create lists and tables.

Learn the basic syntax of markdown here.

4.4 Root folder readme

To create a root folder readme file, use any markdown or text editor (e.g. VS Code, Typora, notepad etc.), open a new file and save it to the root folder for your project, ensuring the file type is .md.

Name your readme file _README.md.

Next, we want to add some content to our _README.md. The purpose of this document is to describe the directory structure of our project. To adequately describe our directory structure we should include:

- A brief description of the project or purpose of the root folder
- Date when the root folder was created and who created it
- Date when the readme file was last updated and who updated it
- A brief description of the contents of each major folder within the root folder
- A brief description of file naming conventions used within the directory

To see an example root directory readme file click here.

4.5 Data directory readme

Next, we want to create another readme file, but this file will be placed within the subfolder that contains our project's data. To do this, open any markdown or text editor (e.g. VS Code, Typora, notepad etc.), open a new file and save it to the data subfolder for your project, ensuring the file type is .md. Name your readme file _README.md.

The purpose of this readme file is to provide a description of data collection methods. We will include:

- Date when the data directory was created and who created it
- Date when the readme file was updated and who updated it
- A brief description of each data that was collected, the methods used for collection, and the date range for when each dataset was collected
- A description of who was involved in data collection
- A brief description of where the data was collected

To see an example data directory readme file click here.

4.6 Data dictionary

Lastly, we need to create a data dictionary which elaborates on how our data is stored and organized. To do this, open any markdown or text editor (e.g. VS Code, Typora, notepad etc.), open a new file and save it to the data subfolder for your project. This time we will save the file as _DATA-DICTIONARY.md.

A data dictionary helps others understand the meaning of each element in your datasets within the broader context of the project. Typically you will have an individual readme file for each dataset. This file should include:

- Date when the data dictionary was created and who created it
- Date when the data dictionary file was updated and who updated it
- A description of the raw data file
- A description of each variable for all datasets including data type, units, number of levels if categorical, and a description of variable levels where relevant
- When describing variables you need to provide the full names and definitions of each variable because often variables are abbreviated in datasets

To see an example data dictionary click here.

4.7 Example BIOL 116

In our previous BIOL 116 example, we used a flat folder structure to hold all of our files. With this example, no readme files were created, as we made the assumption that the project was "simple" enough in its structure to not warrant a readme file. On reflection, this was an oversight, and we probably should have created a readme file describing what was in each file. Neither did we create a data dictionary. We'll do better on future assignments now that we know about the value of both forms of documentation. Anyway, in that example, we ended up with one folder of files that looked like the following before submitting our final assignment:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Analysis_V0.xlsx
Pither_BIOL116RProject_Figure-freq-plot_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V1.png
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
```

```
Pither_BIOL116RProject_Lab-report_V2.docx
Pither_BIOL116RProject_Lab-report_V3.docx
```

We can see that this might start to get unwieldy if we have a few more files joining the party. So let's break this apart into folders...

Top Level folder

```
BIOL116RProject/
```

Inside of BIOL116RProject we have one file and four subdirectories:

```
 _README.md
Data/
Analysis/
Figures/
Report/
```

Note that we created a _README.md file to describe our directory structure. We'll now distribute our files across these folders...

Data Folder

Creating a _README.md and a _DATA-DICTIONARY.md to describe our data files and their contents...

```
_DATA-DICTIONARY.md
_README.md
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
```

Analysis Folder

```
Pither_BIOL116RProject_Analysis_V0.xlsx
```

Figures Folder

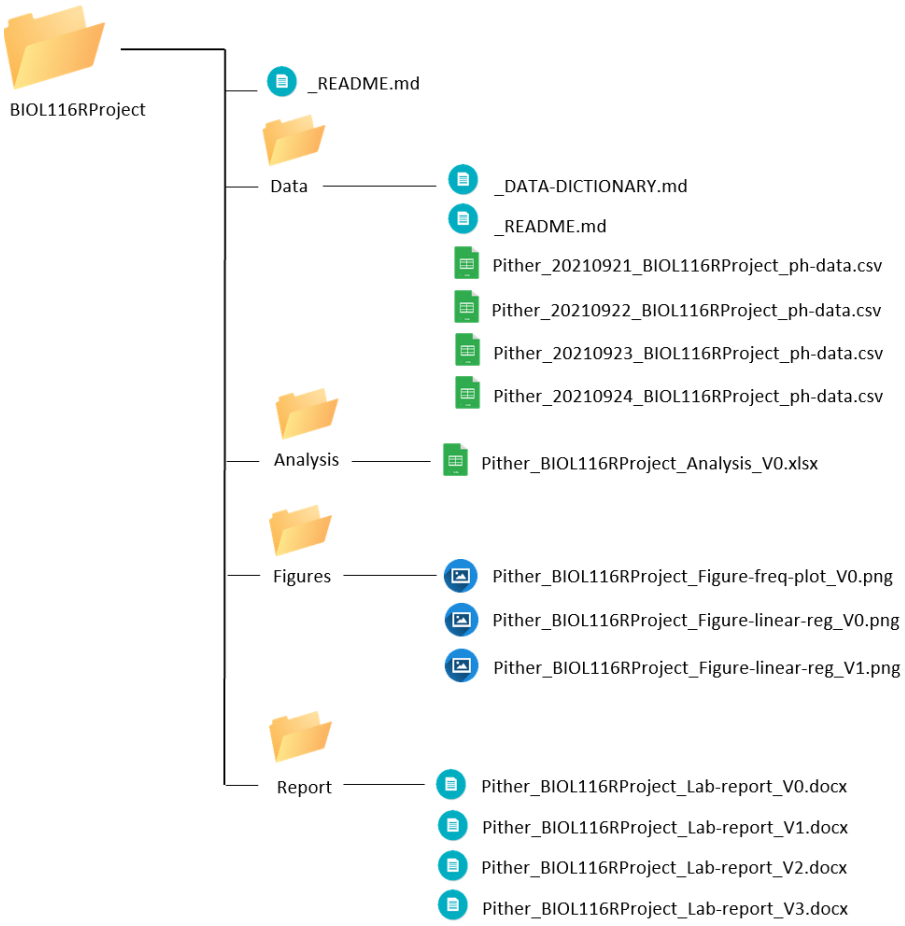
```
Pither_BIOL116RProject_Figure-freq-plot_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V1.png
```

Report Folder

```
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
Pither_BIOL116RProject_Lab-report_V2.docx
Pither_BIOL116RProject_Lab-report_V3.docx
```

Screenshot

And finally a screenshot from our desktop file manager...



4.8 Example BIOL 125

Let's work through another example where we start our project off using both appropriate directory structure and file naming conventions. Say you're a student in BIOL 125 working on a research project testing mealworm food preferences…

Day 1

On day one of our research project, we are asked to prepare the beginning of a lab report that states our research question, hypothesis, and proposed methods. First, we need to create the root folder for our project:

```
BIOL125MealwormProject/
```

Within our root folder, we create a _README.md file to describe our directory structure. Let's add our project name, date the folder was created and who created it, a short description of the project, group member names, file structure (major folders and their proposed content), and naming conventions to this readme file. Your file should look something like this:

_README.md

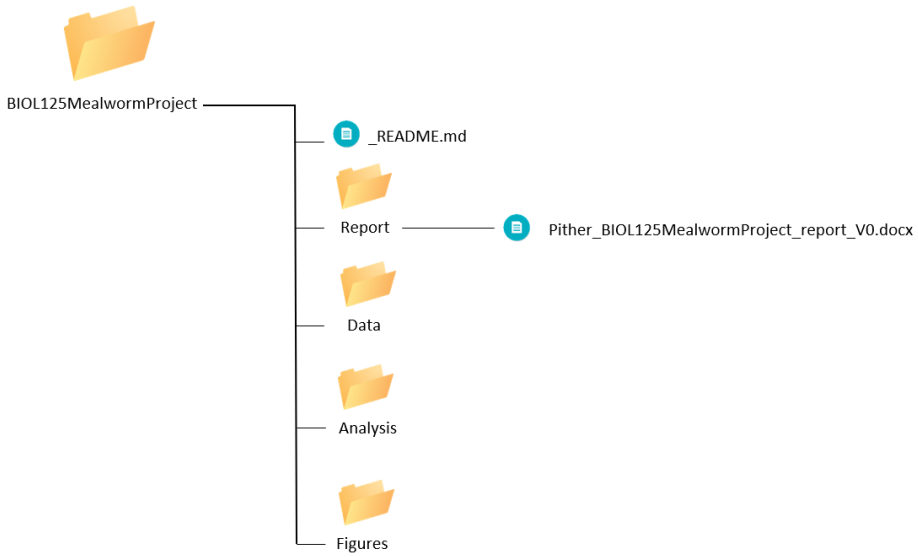
Since we have just started our project, there won't be files in most subfolders we create. However, it's good to have the skeleton of what we want our directory to look like so everyone in the group places new files in the correct location. Later, if needed, we can modify our directory structure and update our readme file to reflect those changes.

Since we will be using the naming conventions outlined in Chapter 1, we can list those naming conventions here. It may seem strange to outline the naming conventions for documents that haven't been created yet, but having a strategy for naming files from the beginning of the project is very important. It ensures everyone is following the same set of rules when they add or edit files in the project, which helps everyone stay on the same page when working in a shared directory.

Now that our root folder and root readme files are set up, we need to create the subfolders within BIOL125MealwormProject/. Since we outlined the major subfolders in our readme file as Report, Data, Analysis, and Figures, we'll use these same names when we create the subfolders. Finally, we can open up a new Word document for our lab report and save it to the Report folder using the appropriate naming conventions.

```
Pither_BIOL125MealwormProject_report_V0.docx
```

Here is what our project directory looks like so far:



Day 2

Today, we completed a pilot experiment and collected some data. We saved this data file into our project's corresponding Data folder using appropriate naming conventions.

New files:

`Pither_20210915_BIOL125MealwormProject_food-preference-data.csv`

Since we have added our first dataset into our project folder, we need to create a corresponding _README.md and _DATA-DICTIONMARY.md.

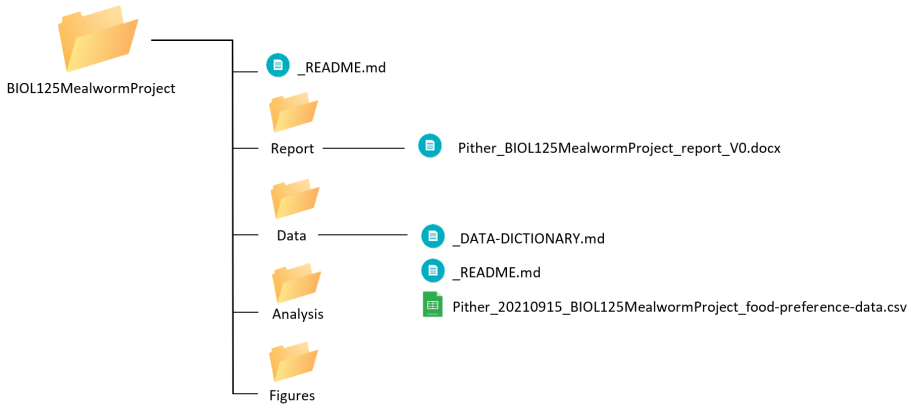
Let's start by creating the data directory readme and provide a description of our data set, collection methods, who collected the data, and where it was collected. It should look something like this:

_README.md

Next, let's create a data dictionary for our new dataset. It should look something like this:

_DATA-DICTIONARY.md

Now our project directory looks like this:



Day 3

Now that our group has completed its pilot project, we decided to expand our data collection and start recording how far mealworms are willing to travel to get food. In addition to this new distance data, we continued to collect data on food preferences.

New files:

```
Pither_20210916_BIOL125MealwormProject_food-preference-data.csv
Pither_20210916_BIOL125MealwormProject_distance-data.csv
```

Since we have a new dataset, we'll have to update our data directory readme file with a description of the new dataset. Remember to note the date it was updated and who it was updated by. Our updated data directory _README.md file should look something like this:

_README.md

Now our updated data directory readme file includes descriptions for both datasets.

In the interest of organization, let's keep our food preferences and distance data in separate subfolders. So, we'll create two new subfolders within the Data folder. We'll call one Food-preferences/ and the other Distance/. This way, we can organize csv files into specific folders for each corresponding dataset. After doing this, we need to update the _README.md in our root folder since we've modified our directory structure. This readme should now look something like this:

_README.md

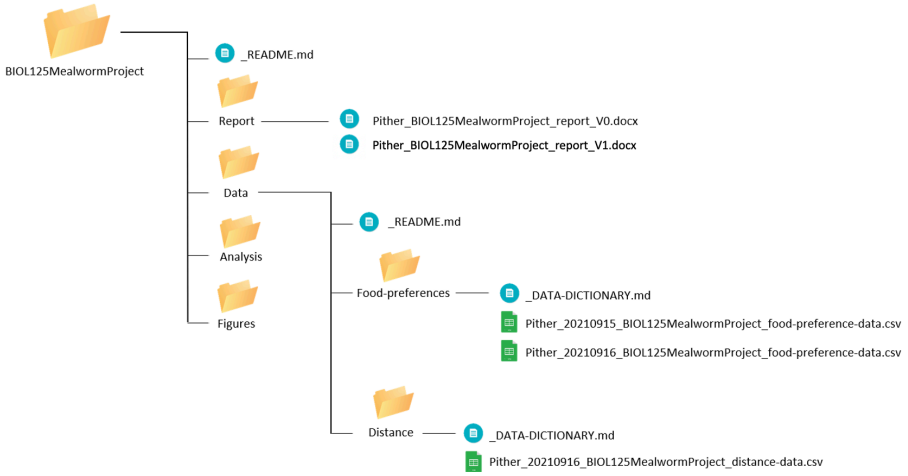
Next, let's also create a data dictionary for our new dataset within our distance subfolder. Remember to note the date it was updated and who it was updated by. It should look something like this:

_DATA-DICTIONARY.md

Since we made some updates to our project design and methods, I'll also go ahead and update our lab report to reflect those changes alongside justification for the changes. Then, I'll be sure to save my updated lab report using the appropriate naming conventions.

```
Pither_BIOL125MealwormProject_report_V1.docx
```

Now our project directory looks like:



Day 4-5

Over these days, we collected our last rounds of data, created some figures, and analyzed the data. So we have a bunch of new files that we need to make sure are placed correctly within our project directory.

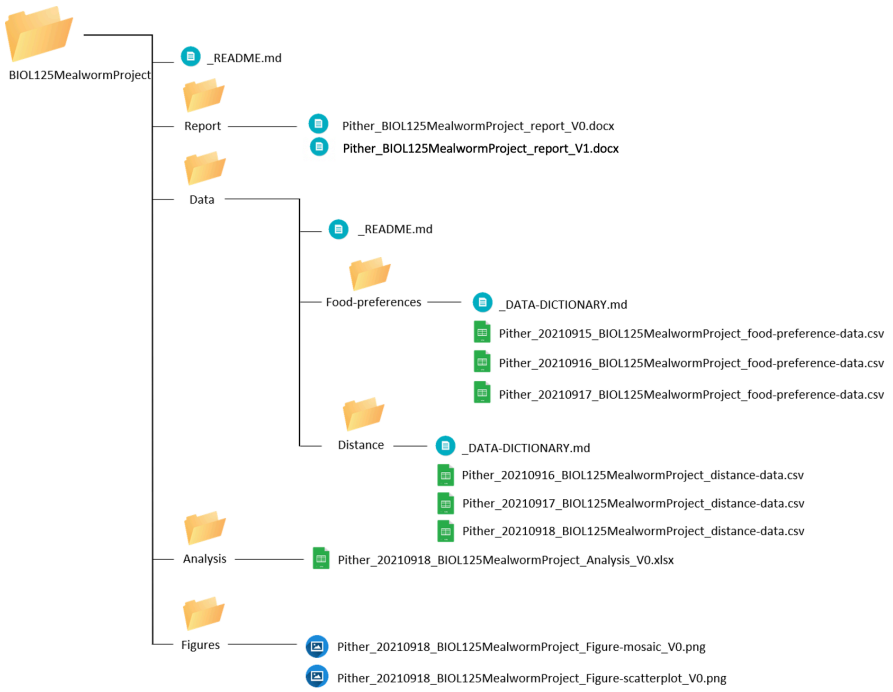
New files:

```
Pither_20210917_BIOL125MealwormProject_food-preference-data.csv
Pither_20210917_BIOL125MealwormProject_distance-data.csv
Pither_20210918_BIOL125MealwormProject_distance-data.csv
Pither_20210918_BIOL125MealwormProject_Analysis_V0.xlsx
Pither_20210918_BIOL125MealwormProject_Figure-mosaic_V0.png
Pither_20210918_BIOL125MealwormProject_Figure-scatterplot_V0.png
```

We'll save all 3 new data files into the Data/ subfolder of our directory. Since we've already described these datasets in the data directory _README.md file and have a corresponding _DATA-DICTIONARY.md for both, there are no more updates needed.

Next, we'll save our analysis into the Analysis/ subfolder and the figures into the Figure/ subfolder.

Now our project directory looks like:



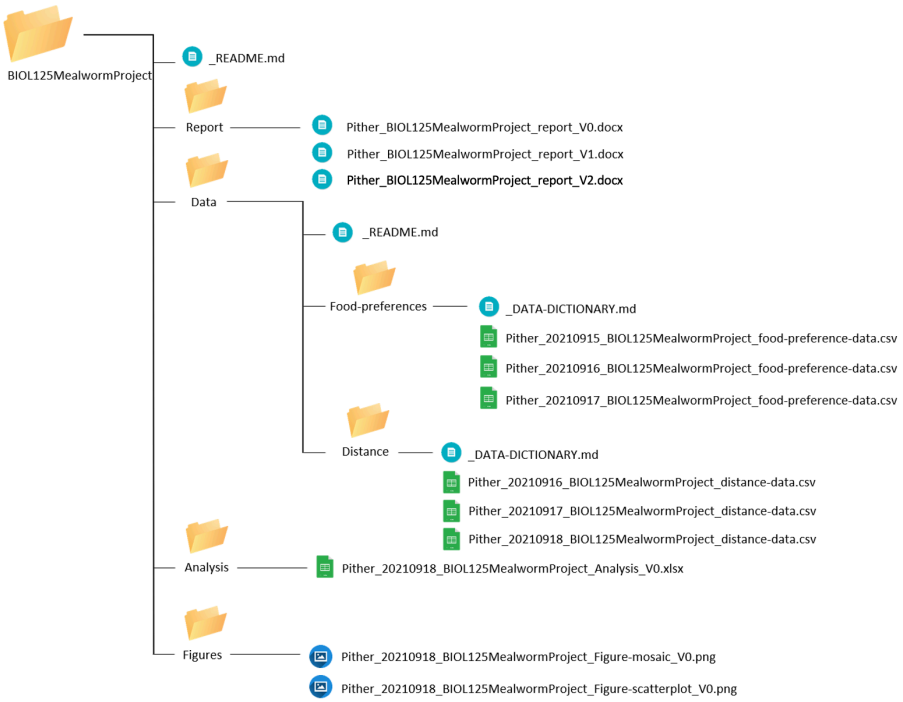
Day 6

Today is the last day of our project and we completed the final copy of our report. We will make sure this is saved into the Report folder using the appropriate naming conventions.

New files:

```
Pither_BIOL125MealwormProject_report_V2.docx
```

Our final directory looks like this:



We can start to see that if you were to share your entire project directory with another person, it would be relatively easy for them to locate files and understand the meaning behind each document in our project. They would also know when changes were made and who made these changes, so if they had any questions, they'd know exactly who to ask!

Chapter 5

Tidy data

Last updated 2023-02-10

We've talked about file naming, directory structures, and documentation to ensure accessible, interpretable, and transparent data. Now it's time to talk about organizing individual variables within a given file. When properly organized, data values can be effectively analyzed, summarized, and visualized. When not, they can be onerous to work with and risk misinterpretation.

In general, your data files should adhere to the principles of "tidy data". Tidy data is governed by the following 3 rules[1]:

- Each variable must have its own column.
- Each observation must have its own row.
- Each value must have its own cell.

It's easy to veer from these rules, as it's often easier to collect data using data collection tools that violate these rules. When this is the case, we need to know how to re-organize our data to make it "tidy".

5.1 Wide Data

Non-tidy data - sometimes called "wide" data as it tends to use more columns, but fewer rows - tends to lump observations together in cells. It is often easier to collect data in this way.

So, say we collected data about the number of trout caught at local lakes across several days. We might end up with the following data tables if we used a separate table, sheet of paper etc. to record our findings.

[1]See: Wickham, H. & Grolemund, G. (2017). Tidy Data. *In R for Data Science*.

Site	Trout_Caught_Day_1
Mabel-lake	1
Postill-lake	3
Ellison-lake	0

Site	Trout_Caught_Day_2
Mabel-lake	3
Postill-lake	4
Ellison-lake	5

Site	Trout_Caught_Day_3
Mabel-lake	3
Postill-lake	5
Ellison-lake	1

It is also very likely that we set up an Excel sheet where we recorded the site as the first column and our days and fish caught combined in subsequent columns, one column for each day. Even if we hadn't collected our data this way, we might be tempted to group our above data together for analysis or assignment submission in this way.

Doing this, we'd end up with a table something like the following:

Site	Trout_Caught_Day_1	Trout_Caught_Day_2	Trout_Caught_Day_3
Mabel-lake	1	3	3
Postill-lake	3	4	5
Ellison-Lake	0	5	1

But for analysis - for "tidy" data - we want one column per variable. In this case, we have three variables:

- site
- day
- quantity caught

So let's get this cleaned up…

5.2 Tidy Data

In the previous example, our data were organized where day and quantity caught shared common columns. That is, not every variable had a dedicated column and consequently, not every variable had a value in every given cell - day did not have any cell values.

Tidy data breaks this down and reserves one column per variable and one row per observation. Remember, we have three variables: site, day, and quantity caught. So let's transform this…

First, working with a collection tool where we have one table per day:

Site	Day	Trout_Caught
Mabel-lake	1	1
Postill-lake	1	3
Ellison-lake	1	0

Site	Day	Trout_Caught
Mabel-lake	2	3
Postill-lake	2	4
Ellison-lake	2	5

Site	Day	Trout_Caught
Mabel-lake	3	3
Postill-lake	3	5
Ellison-lake	3	1

And second, gathering this data into a single dataset, sorted by site:

Site	Day	Trout_Caught
Mabel-lake	1	1
Mabel-lake	2	3
Mabel-lake	3	3
Postill-lake	1	3
Postill-lake	2	4
Postill-lake	3	5
Ellison-lake	1	0
Ellison-lake	2	5
Ellison-lake	3	1

Now that's tidy data!

5.3 Side by Side Comparison

Wide Data

Site	Trout_Caught_Day_1	Trout_Caught_Day_2	Trout_Caught_Day_3
Mabel-lake	1	3	3
Postill-lake	3	4	5
Ellison-Lake	0	5	1

Tidy Data

Site	Day	Trout_Caught
Mabel-lake	1	1
Mabel-lake	2	3
Mabel-lake	3	3
Postill-lake	1	3
Postill-lake	2	4
Postill-lake	3	5
Ellison-lake	1	0
Ellison-lake	2	5
Ellison-lake	3	1

Data Presentation

Chapter 6

Figures & Tables

Last updated 2023-02-10

These guidelines are based on current "best practices" in Biology. You may encounter small differences when working with data or reading the results of research from other disciplines. Our aim is to achieve consistency among faculty, instructors, and students in how data are summarized and presented within lab reports and research papers.

6.1 Tables

When presenting data in a table keep in mind the following:

- The heading is placed above the table.
- The table should be interpretable as a stand alone object using an informative heading and judicious footnotes.
- Sample sizes and units are always included.
- Use horizontal lines only; these are often placed above and below headings, and at the bottom of tables.

Example

The following is an example of a properly formatted table.

Table 1. Summary of trait measurements made on individuals of *Solidago* ssp. collected within shaded and open habitats in the vicinity of Portland, Oregon. "sd" denotes standard deviation.

Trait

Habitat: Shaded (n = 20)

Habitat: Open (n = 18*)

Mean (sd)

95% confidence interval

Mean (sd)

95% confidence interval

Leaf area (cm2)

4.59 (0.974)

4.14 - 5.05

4.54 (0.972)

4.24 - 5.15

Leaf mass (mg)

2.52 (0.765)

2.15 - 2.89

2.62 (0.705)

2.25 - 2.99

Root mass (mg)

9.97 (2.754)

8.67 - 11.26

9.90 (2.454)

8.37 - 11.16

* data for two individuals misplaced

6.2 Descriptive & Summary Statistics

Here are some general guidelines to follow when displaying descriptive or summary statistics:

- Round numbers to one more digit for measures of centre (e.g. mean), and 2 more digits for measures of spread (e.g. sd) than was used in measuring the data
  - For detailed guidelines about significant digits, consult the following webpage: https://www.physics.uoguelph.ca/significant-digits-tutorial

- Units are preceded by a space within text passages:
  - e.g. "Average height was 34.2 cm (± 3.43 SEM)."

Describing Numerical Variables

- Report mean with standard deviation, and additionally median with interquartile range for variables that exhibit a non-normal frequency distribution (e.g. is skewed) or that includes outliers
- Parameter estimates (i.e. mean) should be accompanied by measures of uncertainty, i.e. the *standard error of the mean* `SEM` or confidence *interval* (notation: lower limit – upper limit); confidence *limits*: (lower limit, upper limit)
- Confidence intervals are strongly encouraged because they inform about *effect size*
- Measures of uncertainty for an estimate, such as SEM, can be preceded by a `±` sign; do not make the common mistake of reporting a `±` sign with a standard deviation, as it is not a measure of uncertainty in an estimate

Describing Categorical Variables

- Report a frequency table (raw data) or a summary table with proportions for categories (the main descriptive statistic of interest), along with the confidence interval for the proportion if appropriate.

6.3 Results of Statistical Tests

Here are some guidelines for reporting the results of statistical tests:

- Your *Methods* section should clearly state the significance level (α), and this should be decided prior to the study
- Test statistics (e.g. Student's `t` or an `F` from ANOVA) should be rounded to 2 decimal places, and associated `P-values` should report 3 decimal places, or if smaller than 0.001, then <0.001. `P-values` do not indicate `effect size`, so reporting $P = 10^{-6}$ is not more impressive than P <0.001
- Concluding statements should, in the absence of a table, include the test, test statistic value, degrees of freedom `df` or sample sizes, `P-value`, and `confidence interval` (if appropriate) in parentheses.
- For example: "On average, hair loss was significantly greater among fathers compared to childless men (Welch's 2-sample t-test; $t = 4.23$; n F = 18, n C = 20; $P = 0.018$; 95% CI for difference: 9.34 – 18.22%)."

- Regression and ANOVA results should be shown in a standard ANOVA table format.

Example of ANOVA table format

Table 3.

	SS	df	MS	F	P
Treatment	7.224	2	3.6122	7.29	0.004
Error	9.415	19	0.4955		
Total	16.640	21	4.1078		

6.4 Figures

When displaying data using a figure, follow these guidelines:

- The figure heading should be placed below the graph and should provide sufficient information so the figure can be interpreted on its own
- The heading can include statistical statements (Fig. 1) or simply describe what is being shown (Fig. 2)
- Sample size(s) must be reported
- The first time a particular type of graph is shown (e.g. boxplot), an explanation of the graph features must be provided. Subsequent figures of the same type can refer to the first for details. See Fig. 2 for an example
- Use hollow symbols so that overlapping points can be seen (Fig. 3)
- Orient all text horizontal (except y-axis label), including all tick labels
- Place axis tick marks outside of the figure border to avoid overlapping with observations
- Data points should not touch the axes
- Fitted lines (e.g. least-squares regression) included in figures should be fully explained in the heading,
  - e.g. "Line represents a least-squares linear regression line, $y = 0.3 + 4.5x$ ($F = 5.65$, df = 36; P = 0.021)".
- For more complex statistics (e.g. lines associated with mixed effects models) refer the reader to the text for details
- Bar plots should **only** be used to visualize categorical data (e.g. proportion of students with brown or blue eyes) or counts (number of flies on scat)
- When comparing numerical data among categories or groups (Figs. 1,2) use stripcharts (Fig. 1) when sample sizes are small (i.e. <20) and boxplots otherwise (Fig. 2).
- Note that the stripchart has the advantage of showing all the data (i.e. each observation is represented by a point), whereas the boxplot summarizes

the data visually. When sample sizes per group are very large, it would get very messy if you tried to show all the data. However, there are graphs like "violin plots" that offer a nice compromise.

Examples

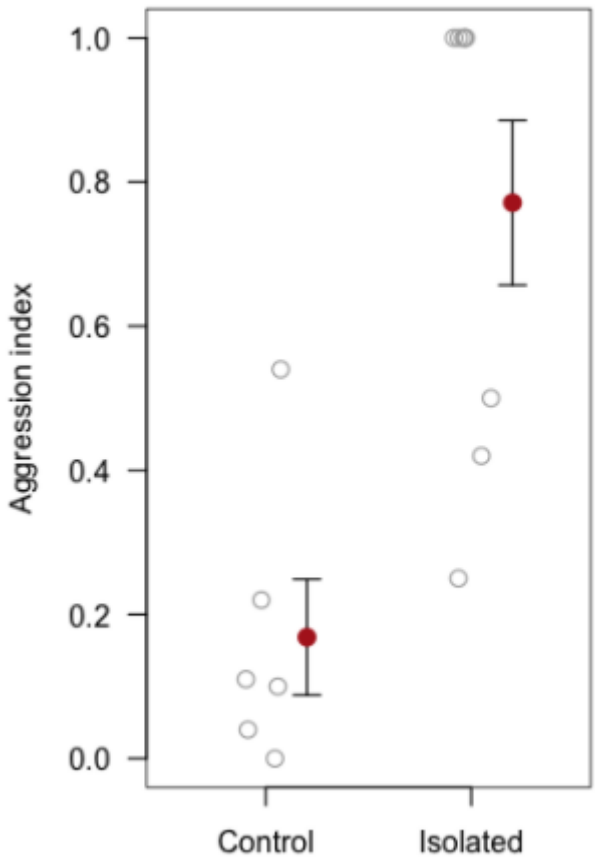


Figure 1. Aggression was significantly higher among isolated ants ($n = 8$) compared to the control group ($n = 6$) (see text for details). Shown are individual observations (grey circles), group means (solid circles) with +/- 1 SEM.

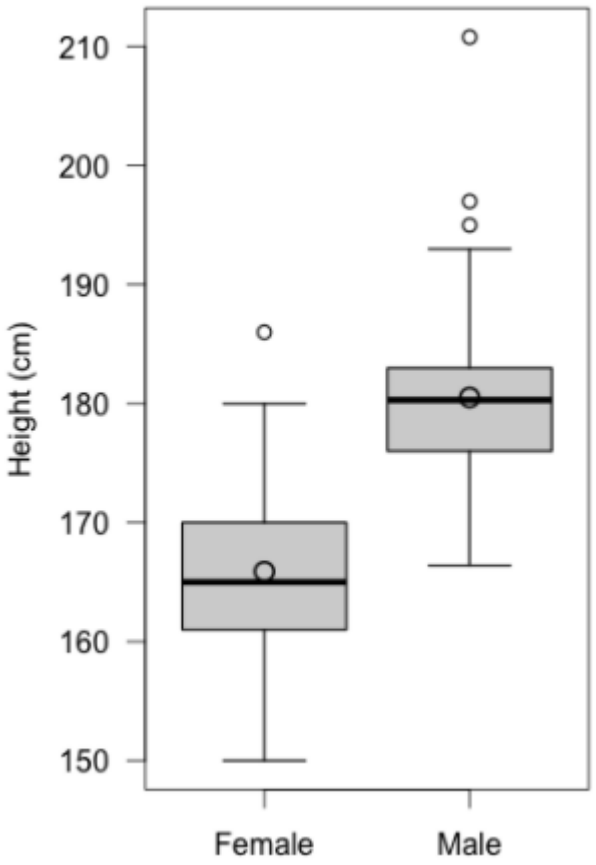


Figure 2. Height of male (n = 64) and female (n = 90) students within BIOL202. Thick horizontal lines represent group medians, large circles represent group means, boxes delimit 1st to 3rd quartiles, whiskers extend to 1.5 x IQR, and small circles represent extreme observations.

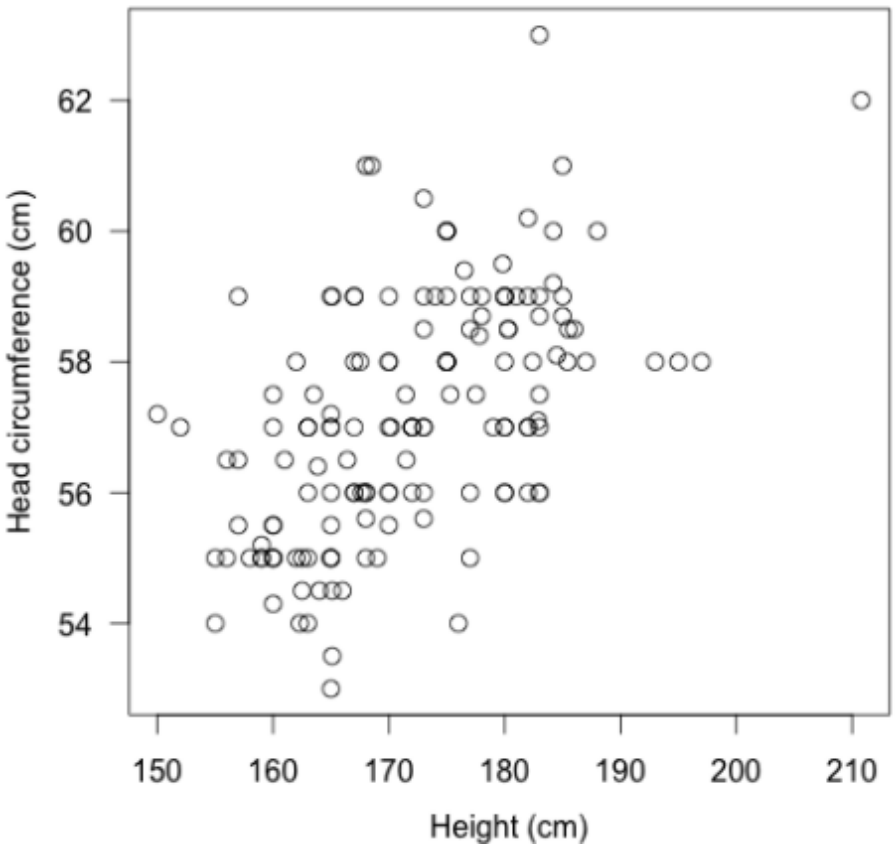


Figure 3. Head circumference versus height for $n = 150$ BIOL202 students. The positive association is highly significant (Pearson $r = 0.82$; $P < 0.001$).

Chapter 7

Sketches & Drawings

Last updated 2023-02-10

We've heard it before, a picture is worth a thousand words. While a common place phrase, it echoes very true in biological research. Describing in text the morphological features of an organism is no easy task. Trying to form a mental image of an organism described in text still more difficult! As a simple example, imagine trying to draw a representation of *Fritillaria pudica* from the botanical description in the Flora of North America. No easy task. So let's talk about sketches…

7.1 The Role of Sketches

Detailed notetaking is an essential aspect of the research process. There are countless benefits to taking detailed notes throughout as you work in the lab or on research projects. Some of these benefits include:

- Increased transparency and openness in your research and thought processes.
- Enhanced reproducibility of your experiments and lab work.
- Your notes may be able to help you explain variability or unexpected results.

While detailed written notes can be extremely helpful, their combination with sketches is powerful. Some things you might sketch include:

- Experiment plan or design.
- Any apparatus's used. This is especially handy if you are using a novel method or apparatus.

- Organisms or structures viewed under the microscope.
- Organisms or structures viewed during a dissection.

Providing a quick simple sketch alongside your written notes provides additional context to help both yourself and your reader understand what you observed or your train of thought.

7.2 Sketching Guidelines

As we've seen elsewhere, convention and standards help us communicate in unambiguous terms - they allow others in our field to easily interpret what we're trying to say, and they allow us to process our information in computational environments in known and reproducible ways.

The same goes for how we construct sketches in biological research. These sketches are generally simple in their articulation and are embedded within your notebook.

General Guidelines

- Do **not** put a title at the top of your drawing. Instead, place a caption at the bottom left. Captions can be several sentences long and should adequately describe what the sketch is.
- Use a minimum number of fine lines and use dotted lines to show depth. Do not shade.
- Place the drawing so that the labels can be put in a column on one side. When possible, labels should be to the right of the figure.
- Label lines should be straight and should not intersect.

Organisms and Structures

Some additional guideline to ensure your sketches are clear and readable when drawing organisms or their structures include:

- Orient your drawing so that the anterior or oral aspect of the organism is at the top of the page.
- Labelling should be as complete as possible. If any structures have been removed or displaced you should indicate this on your drawing. If the manual or text you are using mentions certain structures that you cannot locate, make a note of this on your drawing. For example, you may say "nuclei not seen" in your caption.
- Genus and species names should be

  - Underlined
    or *Italicized*
  - The genus should be Capitalized
  - The specific epithet should be all lower case

Microscopes

When drawing organisms or structures viewed under the microscope:

- Do not draw a circle around your drawing.
- Put the scale (or magnification) of your drawing at the bottom right of your drawing. Instructions for determining scale/magnification can be found here.

7.3 Good Example Sketches

Example 1

This sketch is properly labeled and label lines do not intersect. It contains a descriptive caption below the sketch with the species identified with proper formatting and capitalization and scale noted.

Example 2

This figure is properly labeled and label lines do not intersect. It contains a descriptive caption below the sketch for easy interpretation, and depth is indicated with a dashed line.

7.4 Poor Example Sketch

This figure is not directed to have the anterior region at the top, the labelling, while potentially accurate, is not all directionally to the right of the drawing itself. It employs shading and it's wrapped in a circle. Did we miss anything? Oh, the caption is above, not below the sketch and the genus and specific epithet are not properly formatted.

Caption Vorticella Campanula viewed using a microscope.

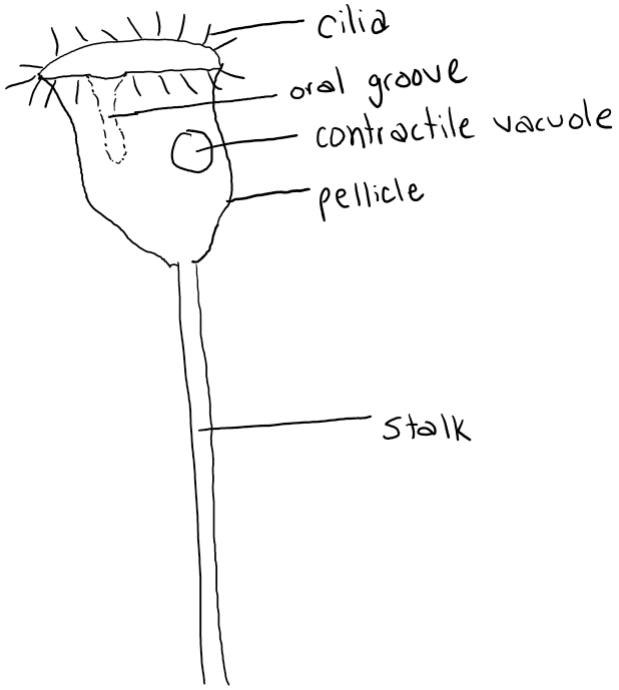


Figure 7.1: *Caption* Whole (wet) mount of *Vorticella campanula* viewed at 400x total magnification using a brightfield compound microscope. Macronucleus not seen.Image produced by Clerissa Copeland licensed under CC BY-NC-SA 4.0

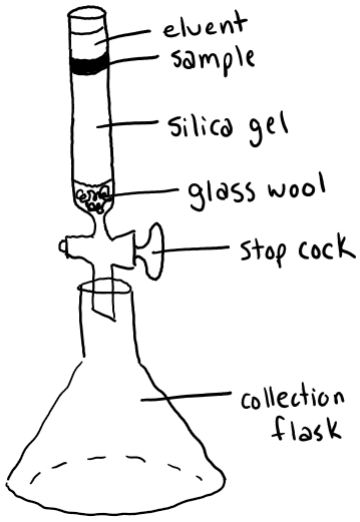


Figure 7.2: *Caption* Column chromatography apparatus set-up used to purify the experiment sample. Silica gel represents the stationary phase whereas the eluent represents the mobile phase. As the components of the sample separated, the stopcock was closed and the collection flask swapped such that only the compound of interest was collected.Image prodced by Clerissa Copeland licensed under CC BY-NC-SA 4.0

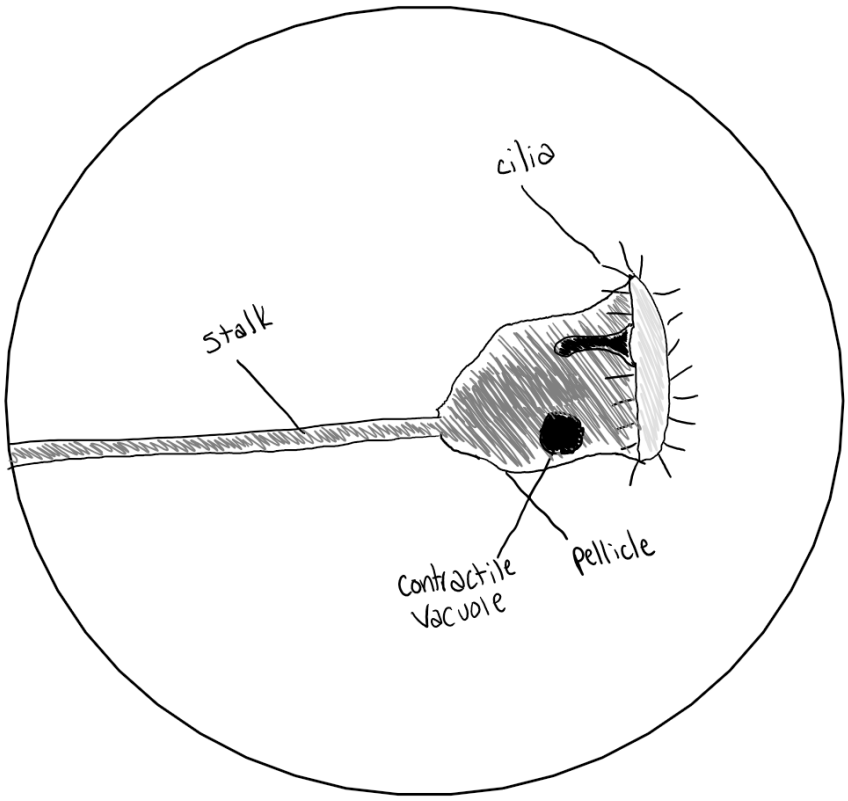


Figure 7.3: Image prodced by Clerissa Copeland licensed under CC BY-NC-SA 4.0

Writing and Citing

Chapter 8

Markdown

Last updated 2023-02-10

Markdown is a markup language used to format plain text files to help us provide additional meaning to our content. Markup languages are ideal authoring tools because they work on the principle of separating content from formatting. Markup languages are ideal authoring tools for the sciences because they rely on plain text, so any computer anywhere at any time will be able to open them and consequently, we will be able to read them.

Benefits of Markdown:

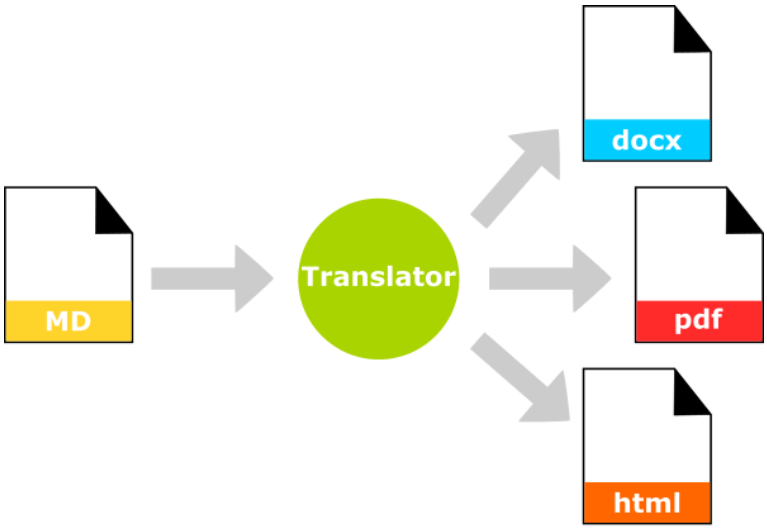
- Simple and easy to learn
- Can be used to generate many different output formats (i.e. pdf, html, docx etc.)
- Can be read on any device with any operating system
- Many different web-applications and websites (e.g. Reddit, GitHub) use markdown

In fact, when we think about the best practices in science, Markdown is an essential authoring tool because it allows for reproducibility and interpretability (i.e. is independent of specific operating systems or programs).

To learn more about Markdown see Matt Cone's Markdown Guide here.

8.1 How Markdown Works

Markdown is a two step process. You write, with markup, in a plain text file. Another application then formats your document based on your markup.



This makes markdown very useful for producing many different types of documents from the same piece of prose, whether that be pdf, html, docx, you name it. It also means that you have one working, editable document, which is plain text, and a series of distributable copies in other formats. More importantly, formatted in a way to address the needs of that particular audience.

All of the content that you're reading right now was authored in Markdown. Note at the top of the page, you can download this content as a pdf or as an epub. But we only had to write it once.

8.2 What You Need to Get Started

You'll need a text editor or a dedicated Markdown editor! What's the difference?

A text editor will not format your markdown, it'll just display the plain text. When you open the _README.md files noted throughout this book (either in **TextEdit** on a Mac or **Notepad** on Windows), you see the raw, plain text output.

TextEdit on your mac is not actually a plain text editor; while TextEdit can read and open a Markdown file, it cannot create a Markdown file. MacOS does not come with a plain text editor like Notepad on Windows. So if you're on a Mac, you'll need to download a text editor.

A dedicated Markdown editor will format your Markdown as you type, so you get an idea as to how it will render if you were to save it as a pdf, html etc. There are a lot of Markdown editors available, many of them free and open source, and many for very specific use cases. When you have some free time,

you may choose to check out the following comprehensive list https://www.markdownguide.org/tools/.

In between a plain text editor and a dedicated Markdown editor are those text editors with add-ons to help you navigate your markup. These options provide a lot of flexibility and finding a text editor of this ilk that you like will be of benefit going forward in your degree. An excellent free and open source, cross platform option is VS Code, which has built in support for Markdown rendering.

VS Code

VS Code is a source code editor. This handy program has many features, one of which is functioning as a text editor. We can use it to create Markdown files by following the instructions below.

Download VS Code at https://code.visualstudio.com/. Once downloaded, launch the program and you'll be presented with a `Welcome` page. Select `New File` and you will be prompted to select a file type; choose `Text-File`.

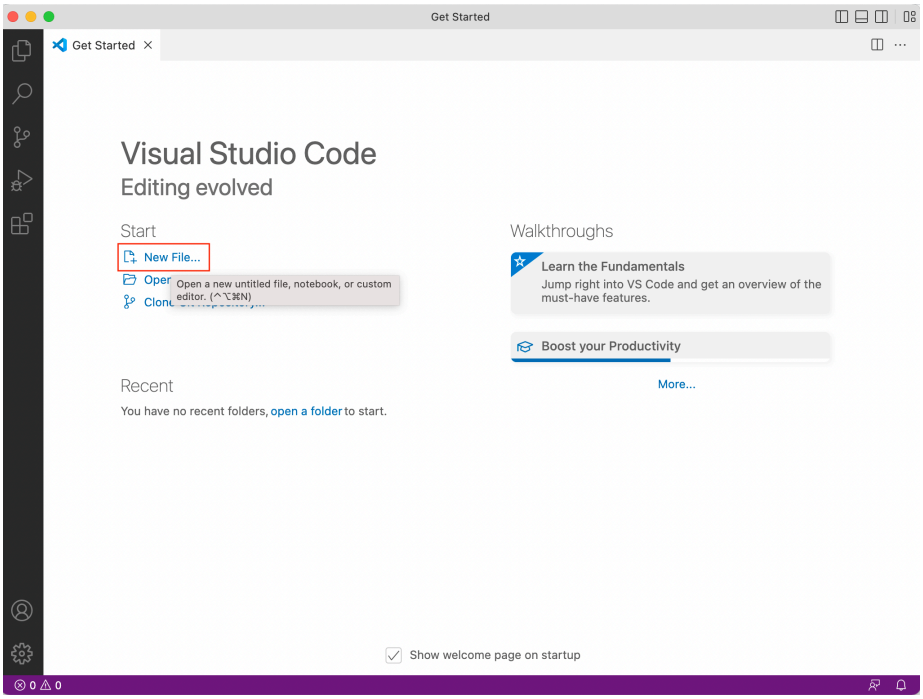


Figure 8.1: VS Code Welcome Page.

Next you will prompted to `Choose a language`. In this case we want a Markdown file. Simply click on `Choose a language` and type `Markdown` into the search bar.

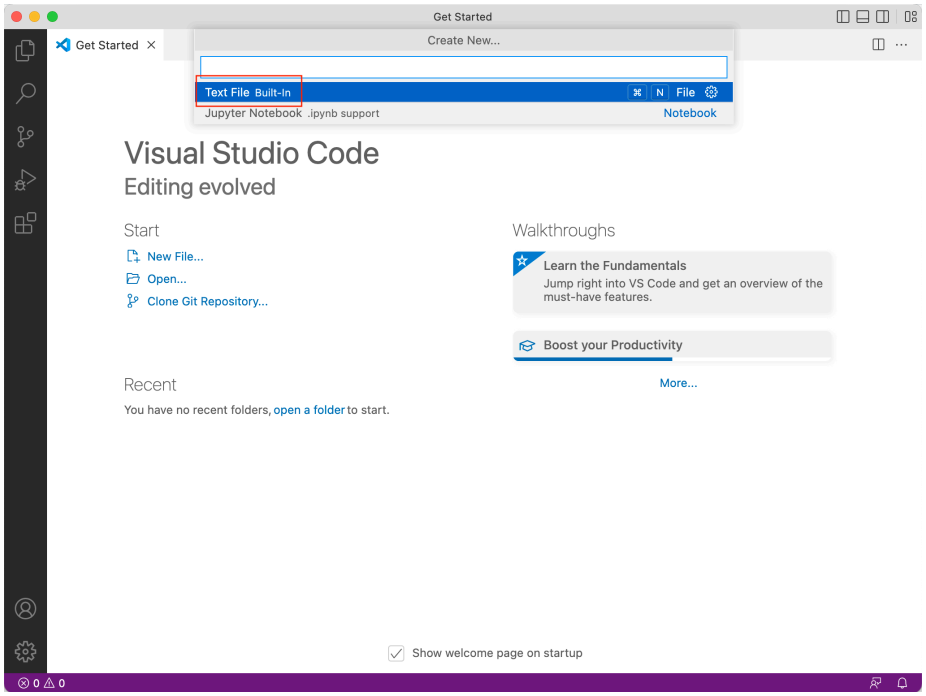


Figure 8.2: VS Code Select a File Type.

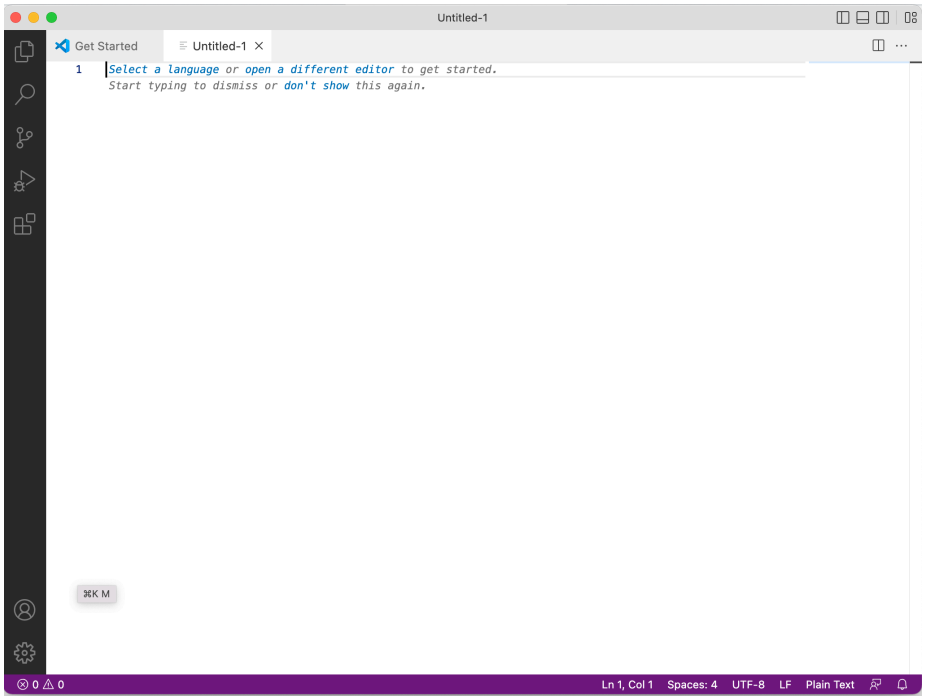


Figure 8.3: VS Code Choose a Language.

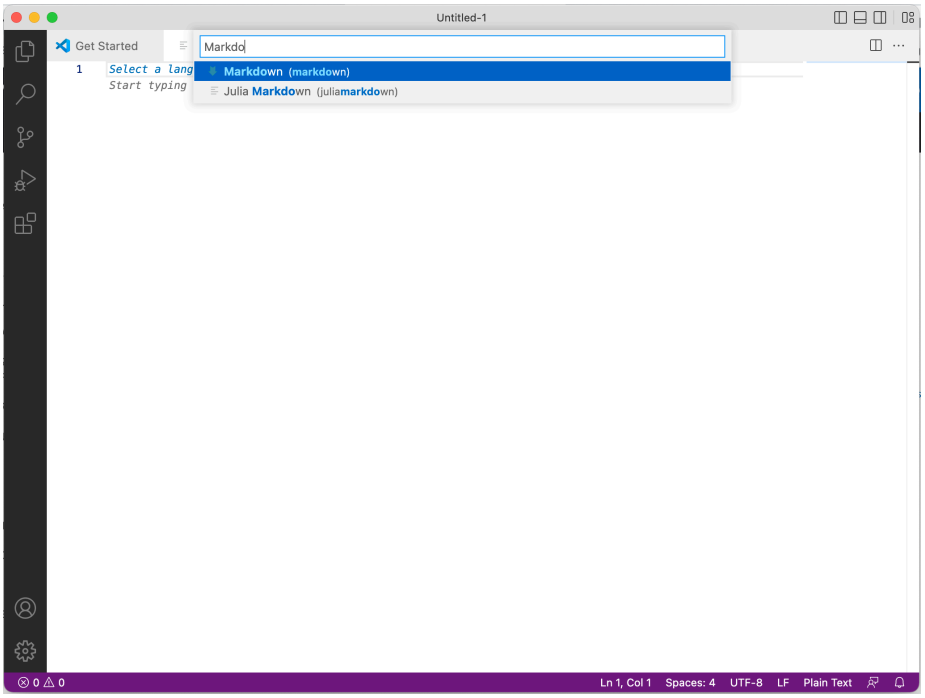


Figure 8.4: VS Code Select Markdown as a Language.

Now you're ready to start writing your plain text file using Markdown syntax. Don't forget to name and save the file to your computer. If you want to preview your document, simply click the `Preview` icon in the top right corner of the window.

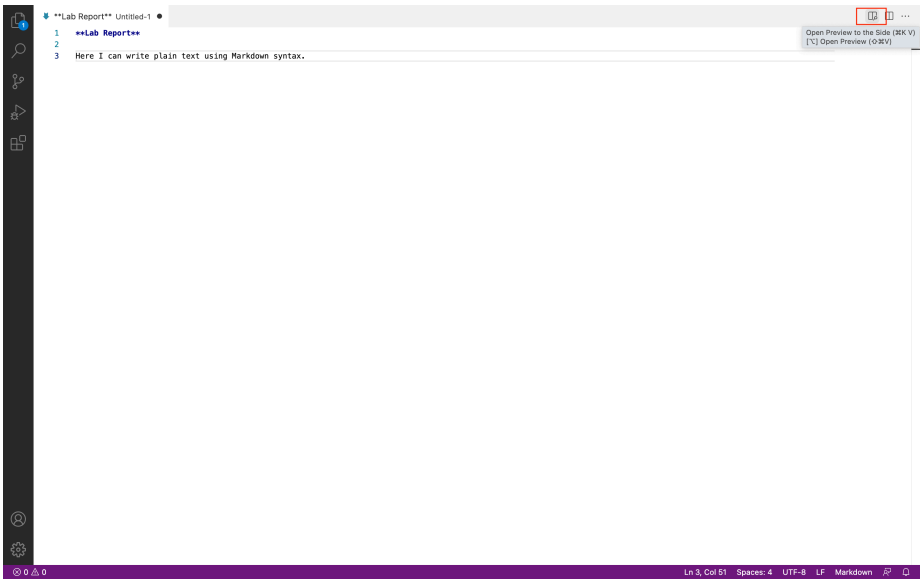


Figure 8.5: VS Code Markdown Preview Icon.

8.3 Prose

You can just start writing if you've opened up a new document in MacDown, Markdown Edit, or whatever text or Markdown editor you happen to be using.

However, there's a couple of things you'll note right away:

- Whether you use one space, " ", or many spaces in between your words, it will only render (display with formatting) as if there was one space.
- Just hitting 'Enter' once doesn't put you on a new line. You need to hit 'Enter' twice.

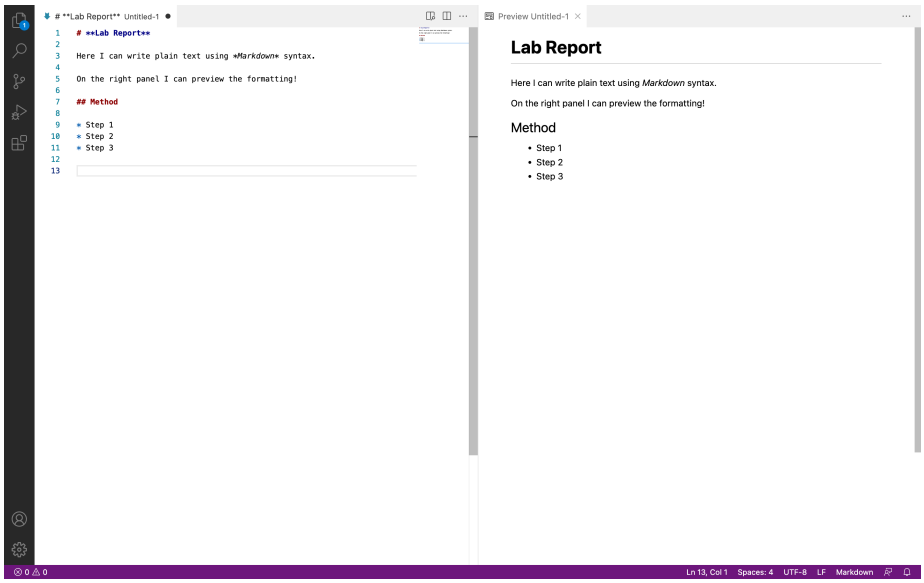


Figure 8.6: VS Code Markdown Preview.

8.4 Structure

Headings

You can add structure to your text document by adding headers with different hierarchies. To do this, you add a number sign `#` before the text. The number of `#` symbols indicates the hierarchy of the header.

Make sure you include a space " " between the `#` and your header

For example, the following...

```
# This is a first-tier header
## This is a second-tier header
### This is a third-tier header
#### This is a fourth-tier header
```

...would render as

This is a first-tier header

This is a second-tier header

This is a third-tier header

This is a fourth-tier header

8.5 Emphasis and Style

There are a number of different Markdown syntaxes that we can use to add style and emphasis to specific parts of our document. Below are a few examples.

Italics

You can make text italicized by encasing the text in a single asterisk `*` or underscore `_`.

Input

```
*This is italicized*

_This is also italicized_
```

Output

This is italicized

This is also italicized

Bold

You can make text bold by encasing the text in two asterisks `**` or underscores `__`.

Input

```
**This is bold**

__This is also bold__
```

Output

This is bold

This is also bold

Bold and Italics

If you want to make text both bold and italicized you can encase the text with three asterisks `***` or underscores `___`.

Input

```
***This is bold and italicized***

___This is also bold and italicized___
```

Output

This is bold and italicized

This is also bold and italicized

NOTE When italicizing or bolding characters within a string of text, it's better to use an asterisk `*` rather than an underscore `_`. For example: Biology *is* awesome NOT Biology _is_ awesome

Strikethrough

You can strike through text by encasing it with two tildes, `~~`.

Input

```
~~This text has a strike through it~~
```

Output

~~This text has a strike through it~~

8.6 Code

You can show text as code by encasing it with three backticks, ```` ``` ````.

Input

````
```
This text looks like code
```
````

Output

```
This text looks like code
```

8.7 Blockquotes

To create a blockquote using markdown, you need to place a greater-than, `>` sign in front of the text.

Input

```
> This text is placed within a block quote
```

Output

> This text is placed within a block quote

Leave a blank line both before the blockquote and after it.

8.8 Lists

Ordered lists

To create an ordered list in Markdown, you'll need to place the item number and a period in front of the text.

Input

```
1. First item
2. Second item
3. Third item
```

Output:

1. First item
2. Second item
3. Third item

Unordered lists

You can create an unordered list by placing an asterisk `*`, dash `-`, or plus sign `+` in front of the text.

Input

```
* first item
* second item
* third item

- first item
- second item
- third item

+ first item
+ second item
+ third item
```

Output

- first item
- second item
- third item

8.9 Tables

When creating tables using Markdown, pipes `|` separate columns while line breaks are used to separate rows. The column header is separated by three or more hyphens `---` between each column's pipe `|`. A colon `:` can be added to the left of the hyphens to left align the column, to the right to right align, and on both ends to centre. The following example is left aligned.

Input

```
| Column 1 Header | Column 2 Header | Column 3 Header |
| :--- | :--- | :--- |
| Column 1 item | Column 2 item | Column 3 item |
| Column 1 item | Column 2 item | Column 3 item |
```

Output

Column 1 Header	Column 2 Header	Column 3 Header
Column 1 item	Column 2 item	Column 3 item
Column 1 item	Column 2 item	Column 3 item

NOTE

- The number of hyphens, `-`, used can make the cell width look incorrect. However, as long as there are three or more hyphens the rendered output will be the same.
- Put a space, " ", between each pipe `|` and the following word or dash `-`

8.10 Links

To create a link to a url or another document, encase the text with square brackets, `[ ]`, and follow the text immediately with the link encased in parentheses, `( )`.

Input

```
To visit the UBC Okanagan Faculty of Biology website click [here](https://biology.ok.ubc.ca/).
```

Output

To visit the UBC Okanagan Faculty of Biology website click here.

8.11 Images

Creating a link to an image follows a similar format to that of links, but the square brackets encasing the text are preceded with an exclamation mark `!`. You can place either the url link or path to the image on your computer in the parentheses. If using a path on your computer, use a relative path.

Input

```
![A magnificent caterpillar. Photo by Erik Karits on Unsplash](images/caterpillar)
![A magnificent caterpillar. Photo by Erik Karits on Unsplash](https://unsplash.com/photos/F7SRtQ
```

Output

8.12 Markdown Flavours

Everything here is basic, or core, Markdown and will be supported by any markdown editor. Since Markdown is simply encoding document structure through markup, several different implementations have expanded on this core set and allow you to do other things, including footnotes, references etc.

We'll keep it simple for the moment; this is all you really need for your _README.md files! Later, you'll be introduced to RMarkdown, which, when used in conjunction with R, will allow you to render statistical analyses within your Markdown document and build reference lists, among other things.

Figure 8.7: A magnificent caterpillar. Photo by Erik Karits on Unsplash

Chapter 9

APA Citations

Last updated 2023-02-10

This chapter is a brief overview of the 7th edition of APA. For a more in depth review, please consult the library's APA Citation Guide. For the full APA manual, please consult the Publication Manual of the American Psychological Association, available through the library.

Additional resources you may wish to consult include:

- The APA Style Blog - great for searching for examples not listed in the 7th edition
- Purdue OWL's website for still more examples

9.1 In-text Citations

In-text citations always appear right after the content you are summarizing, paraphrasing, or quoting. The format is as follows:

Summarizing or Paraphrasing

- (Author, YYYY)

Quoting

- (Author, YYYY, p. #) 1 page
- (Author, YYYY, pp. ##-##) Multiple pages

Narrative vs Parenthetical

If you mention the author or authors in text, you do not need to include this information in the brackets, "()". This is called a narrative citation.

Narrative in-text citation: Raimi (2018) outlines the risks and benefits of fracking through economic analysis and energy security benefits.

The alternative is called a parenthetical citation.

Parenthetical in-text citation: Several benefits and risks can be identified in the implementation of fracking for oil extraction. Considerations include regulation, water pollution, tremors etc. (Raimi, 2018).

Quick Format Guide

# of Authors	Narrative Example	Parenthetical Example
1	Bradley (2017)	(Bradley, 2017)
2	Janmaat and Rahimova (2018)	(Janmaat & Rahimova, 2018)
3 or more	Mei et al. (2018)	(Mei et al., 2018)

9.2 Reference List

NOTE

- Every source used in your in-text citations needs to be listed as part of your reference list, in alphabetical order by author(s)' last names.
- The word **References** should appear at the top of your reference list, and it should be centred and bolded on the page
- Titles should be written in sentence case, that is, capitalize the first word and only subsequent proper nouns. If the title is broken up by a colon (:), capitalize the first word after the colon.
- List all authors in the order that they appear in the source.

Journal article with a DOI (1-2 authors)

List all authors in the reference list and in-text citations.

Janmaat, J., & Rahimova, N. (2018). Managing drought risk in the Okanagan: A roll for dry-year option contracts? *Canadian Public Policy*, *44*(2), 112-125. https://doi.org/10.3138/cpp.2017-003

Journal article with a DOI (3-20 authors)

List all authors in the reference list and only the first author in the in-text citations.

Mei, Y., Yu, K., Lo, J. C. Y., Takeuchi, L. E., Hadjesfandiari, N., Yazdani-Ahmadabadi, H., Brooks, D. E., Lange, D., & Kizhakkedathu, J. N. (2018). Polymer-nanoparticle interaction as a design principle in the development of a durable ultrathin universal binary antibiofilm coating with long-term activity. *ACS Nano*, *12*(12), 11881-11891. https://doi.org/10.1021/acsnano.8b05512

Chapter 10

Types of Sources

Last updated 2023-02-10

As you conduct research and throughout your studies you will encounter many different types of sources. This chapter will briefly introduce and provide examples of three commonly utilized sources: primary, secondary, and review sources.

For more in-depth information about types of sources visit the UBC Library Research Skills for Biologists Guide and the UBC Literature Review Guide.

Primary Sources

Primary sources are scientific papers describing the first hand accounts of events or theories. In other words, an original publication where the author(s) discuss work performed by them. Primary sources can describe either exploratory or confirmatory research. For a refresher on these types of research see The Burden of Proof section of the Open Science Primer. Examples of primary sources include:

- conference proceedings
- journal articles

The link below provides you with an example of a primary paper found through the UBC Okanagan Library. You will need your cwl to access the paper.

Hay, T. N., Phillips, L. A., Nicholson, B. A., & Jones, M. D. (2015). Ectomycorrhizal community structure and function in interior spruce forests of british columbia under long term fertilization. Forest Ecology and Management, 350, 87-95. https://doi.org/10.1016/j.foreco.2015.04.023.

In contrast, the link here provides you with an example of a non-primary sourced paper https://en.wikipedia.org/wiki/Spruce%E2%80%93fir_forests.

Secondary Sources

Secondary sources are scientific papers written by a person who did not participate in the events first hand. The author(s) of a secondary source typically generalizes, interprets, or analyzes primary sources. Examples of secondary sources include:

- review articles
- edited volumes
- books
- abstracts and indexes

To help you distinguish between a primary and secondary source, ask yourself "Did the researchers collect data to answer a question or are they reporting on someone else's data?" If the answer is the former, the source is primary. However, if the answer is reporting on someone else's data the source is secondary.

Review Sources

Review sources are a specific type of secondary source where authors summarize the existing literature for a topic. These reviews are extremely helpful for identifying

- research already conducted
- current knowledge and theories
- significant gaps in the literature
- areas for future research

Four common types of review sources are literature reviews, scoping reviews, systematic reviews, and meta-analyses.

Literature Reviews

Literature reviews, also known as traditional reviews, provide a comprehensive, critical, and objective analysis of a topic. These reviews do not typically follow a structured methodological protocol and the articles included vary and may not necessarily be exhaustive. Additionally, author(s) of a literature review will offer a qualitative synthesis of the included sources along with a conclusion or discussion about the findings.

Scoping Reviews

Scoping reviews follow a specific **pre-defined** methodological study protocol including a detailed search strategy aimed at gathering a broad and exhaustive list of sources to review. The purpose of following a pre-defined study protocol is to reduce bias and enhance reproducibility of the review. Moreover, a scoping

review offers a qualitative synthesis of the evidence based on the included sources along with an evidence based discussion of the findings.

For more information on scoping reviews see the Joanna Briggs Institute Guidelines or PRISMA checklist for conducting a scoping review.

Systematic Reviews

Systematic reviews are similar to scoping reviews in that they follow a specific **pre-defined** methodological study protocol. However, the research question addressed in a systematic review is more specific. Additionally, this type of review will include a critical appraisal of all sources of evidence rather than all sources being considered equally valid. The outcomes of systematic reviews are often used to inform clinical guidelines.

Meta-Analyses

Meta-analyses further expand on systematic reviews by including a quantitative statistical analysis of the results. Typically this involves combining the effect sizes of similar studies to determine an overall effect size related to the research question.

For more information on evidence synthesis, systematic reviews, and meta-analyses see the following links:

- Joanna Briggs Institute Guidelines
- PRISMA checklist
- Collaboration for Environmental Evidence
- O'Dea et al. (2021). *Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: A PRISMA extension. Biological Reviews of the Cambridge Philosophical Society*, *96*(5), 1695-1722. https://doi.org/10.1111/brv.12721

Chapter 11

Finding & Evaluating Published Evidence

Last updated 2023-02-10

11.1 Types of Evidence

Published works in journals are generally divided into three categories: original research, secondary research, and opinion or commentary.

11.1.1 Original Research

Original research – also called primary research – directly reports on the findings from a study. This work is generally considered to be novel, and even with peer review is subject to error that requires close scrutiny.

The quality of this evidence is governed by a number of factors, which include:

- Study type (experimental, observational)
- Study design (blinding, use of controls etc)
- Systematic bias (sampling distribution, sampling size, etc)
- Other bias (financial conflicts, cognitive biases, etc)
- Level of reporting (access to protocols or registrations, data availabability, code availability, etc)
- Appropriate choice of statistical analysis

11.1.2 Secondary Research

Secondary research – also called synthesis research – cumulatively reports on the findings of original research. This kind of reporting attempts to establish the extent of knowledge on a specific topic (systematic review), identify if sufficient evidence exists on a specific topic to suggest the evidence is conclusive (meta analysis), scope the degree or level of evidence available in a specific field of inquiry (scoping review), or offer high level commentray backed by some evidence on a specific topic (narrative review).

The quality of this evidence is governed by a number of factors, which include:

- Study type (meta-analysis, systematic review, scoping review, narrative review)
- Systematic bias (reasonable attempts to find *all* published literature, attempts to find unpublished results, etc)
- Other bias (financial conflicts, cognitive biases, etc)
- Level of reporting (access to protocols or registrations, data availabability, code availability, etc)
- Appropriate choice of statistical analysis (for meta analysis)

11.1.3 Opinion & Commentary

Commentaries offer 'expert' opinion on topics, sometimes suggesting conclusions, sometime suggesting future directions. While an important part of science communication, these neither report comprehensively on a given study nor systematically evaluate existing published research. These should not be treatee as sources of evidence.

11.2 Sources of Evidence

Biology is a big field, with diverse areas of research. The following databases are the primary databases available through UBC to support research in a breadth of topics in Biology.

11.2.1 General

11.2.1.1 Web of Science Core Collection

Link

Description

A rich collection of citation indexes representing the citation connections between scholarly research articles found in most globally significant journals, books, and proceedings in the sciences, social sciences and art & humanities.

Topic coverage

Life sciences, biomedical sicences, engineering, social sciences, arts & humanities. Strongest coverage of natural sciences, health sciences, engineering, computer science, materials sciences.

Resources

Coming.

11.2.1.2 Scopus

Link

Description

Citations and abstracts for journal articles, conference proceedings, and other resources in the sciences, social sciences, arts, and humanities.

Topic coverage

Very broad. Similar to adding Web of Science and PubMed in one portal.

Resources

Coming.

11.2.1.3 CAB Abstracts

Link

Description

Provides international literature in the applied life sciences including all areas of agriculture, forestry, global health, nutrition, and conservation, leisure and tourism.

Topic coverage

Community and Regional Planning, Fisheries, Forestry, Landscape Architecture, Agricultural Sciences, Environmental Design, Food, Nutrition and Health, Human Nutrition

Resources

Coming.

11.2.2 Health & Biological Processes

11.2.2.1 PubMed

Link

Description

The U.S. National Library of Medicine's free, web searchable database. The premier international index to biomedical research covering nearly 6000 scholarly journals and indexing over 30 million citations from 1946 to present.

Topic coverage

All areas of health reasearch, including clinical research, nursing and allied health and other paramedical professions, as well biochemistry.

Resources

Coming.

11.2.3 Niche

There are many niche databases available. These include sources for taxanomic research – Zoological Record – agricultural research – PubAg – aquatic research – ASFA – chemical substances – SciFinder etc.

11.3 Google Scholar and AI Citation Searching

The list of databases under Sources of Evidence are called abstract and indexing databases; they pull or are provided with content from journals and journal publishers; specifically, they are provided with a list of abstracts when a new issue is published. So, when you search, you're not generally searching an article's full text, just it's summary, which is often beneficial. They employ the equivalent to editorial boards to **review the quality of journals** before including them in their index. Their content is relatively **static**; run a search today and run the same search 6 months from now and the only difference will be content that has been added to the index since you last ran your search. The results returned to you are **not dependent on who you are**; while you can re-order a result set by relevance, citation count etc, the pool of results returned to you and your friend(s) will be identical. While they may index millions of articles, they rarely return millions of hits; they are built to **cater to specific disciplines** and as a result, the result sets are usually manageable. You also get access to the full set of results.

Google Scholar is pretty amazing, but it works quite differently from the databases in Sources of Evidence. The search platform does work with publishers, but it also crawls sites known to host academic content. It frequently searches the full text of articles, not just the abstracts, so it risks bringing in less relevant content. It ingests more or less whatever it can find; there is no editorial review on the content. What it returns in your results list is dependent on who you are and when you're running your search; don't anticipate that running a search at different times, from different machines, or as two different people will ever return the same result set. The database is huge; result sets are consequently also generally overwhelmingly large, but you're also only provided access to the first 1000 hits. All in, there's a fair bit happening behind the scenes that you don't control.

AI informed citation searching services are also pretty amazing. These discovery tools generally rely on network analyses, using things like citations, to draw connections between papers and other algorithms to topically cluster papers. The indexes that they have to draw on don't generally come from publishers directly or by crawling the web like Google Scholar does, but rather by leveraging the meta data available through the organizations that, for example, issue DOIs for journal articles.

11.3.1 When to Use Which

These three kinds of services suggest three different approaches to the discovery of published evidence. Which you use will be determined by what your're trying to acheive with a review of the published evidence.

The first – those listed under Sources of Evidence – are curated, stable, and reproducible. When being systematic in your approach, or when you need a confined set of literature that is generally accepted within academia, these should be your primary source of evidence. Any bias introduced here is through publication bias and the curatorial work done on selecting journals to index.

The second – using a service like Google Scholar – are great when you are already familiar with a subject area, and need a quick, topical citation. Don't expect to get a full evidence summary here, but do expect to find complimentary evidence to what you've already found, and to find it quickly. The page ranking algorithms used introduce bias. This is somewhat offset by the fact that these systems are rather indescriminate in what they'll include.

The third – clustering and networking services – are great compliments to the above two sources, especially for serindipedous discovery. No search will ever return all relevant results; citation tracking and thematic clustering of abstracts can be hugely beneficial to trying to understand the scope of published evidence available. For the purposes of evidence synthesis, these tools on their own are insufficient. For the purposed of large evidence synthesis and exploratory efforts,

these tools are invaluable. Bias in these systems will largely result from the clustering algorithms used.

11.4 Grey Literature

Grey literature is a somewhat ambiguous topic. The simplest definition is anything that has not been formally published in an indexed publication - that is, anything that you wouldn't find by searching in a database subscribed to by your library. Grey literature tends to characterized by being difficult to find, hard to cite, and lacking the checks and balances of traditional publication, such as peer review and copy editing. Examples of grey literature include posters, conference abstracts, government reports, and blogs. This may also include pre-prints, especially if those pre-prints never materialize into a formal publication. Gery literature is not inherntly less valuable than formally published literature; however it may require more judicious evaluation before being trusted as quality evidence.

Grey literature is generally found either through a search engine such as Google or DuckDuckGo, on conference websites, on the personal websites of academics, in pre-print repositories, or through government or NGO portals. For example, for Canadian government publications, there is the Federal Science Libraries Network, a search portal for seven science-based departments and agencies of the Canadian government. And while many pre-print servers exist, a common portal and hosting service is OSF Preprints.

11.5 Searching Basics

There are four key concepts that can help with effective searching: breaking up a search into its conceptual parts, using boolean logic to tie concepts together, using stemming and wildcards to account for variations in how terms are written or articulated, and ensuring that concepts that are comprised of more than one term are grouped together.

11.5.1 Concepts

Research questions are generally comprised of at least 3 concepts: the organism or population of interest, the intervention or thing being studied in relation to that organism or population, and the measure of interest or the outcome that we're interested in. This is often formalized as a PIO framework - Population, Intervention, Outcome. There are many variations on this, including a PEO – Population, Exposure, Outcome – PICO where the C covers a comparison – PICOS, where the S is for Study type etc. These frameworks can be useful for

articulating a research question in terms of it's conceptual components; pick the simplest one that resonates with you and your question.

Putting this into an example, we might ask about the relationship between smoking and cancer in men. We have a population – men – an intervention, or perhaps more appropriate an exposure – smoking – and an outcome – cancer. A very simple search for this question might then be:

men AND smoking AND cancer

Similarlily, we might ask if an increase in water salinity increases mortality in zebra fish. Again, we have a population – zebra fish – an intervetion or exposure (if an experimental study it would be an intervention, if an observational study it would be an exposure) – salinity – and an outcome to measure – mortality. A very simple search for this question might then be:

"zebra fish" AND salinity AND mortality

11.5.2 Boolean logic

Boolean logical operators are comprised of AND OR and NOT. Generally when searching the literature, we don't use NOT, but there are circumstances where it may be warranted.

AND is an intersect, OR is a union. AND means both elements must be in the set, OR means either element could be in the set. AND is thus used to find the intersection between concepts while OR is used to find the union among synonyms. In the example

men AND smoking AND cancer

AND is used to find results that have all three terms. Often times, there is more than one way to articualte a given concept, which is when we use OR. For example

men AND smoking AND (cancer OR neoplasm)

We always want to use a reasonable number of synonyms to ensure we are capturing different approaches to describing any given entity or phenomenon.

11.5.3 Stemming & Wildcards

Words can have alternate spellings. There are two ways to work around this. We could use OR, for example

color OR colour

Alternatively, we could use a wildcard

colo#r

where the # represents 1 or 0 characters.

Wildcards will differe between databases. Read the documentation to know which wildcards are available and how to use them.

Stemming stems a word. For example, we might want to find the term smoke, smoker, smoking, smoked etc. Again, we could use OR

smoke OR smoker OR smoking OR smoked

Alternatively, we could use a stem

smok*

which will look for the letters *smok* and any combination of letters thereafter, capturing all the variations (and more potentially) of interest.

:::note

The stemming wildcard * is generally pretty universal across databases. However, neither stemming nor wildcards more generally operate in Google or Google Scholar in the same way as they do in the more structured databases listed in the Sources of Evidence section.

11.5.4 Phrase searching

When a concept is comprised of more than one term, we need to explicity account for that, and we do that with the use of quotations. For example, a search for

zebra fish

is equivalent to a search for

zebra AND fish

which is significantly broader than a search for

“zebra fish”

where, in the former, both terms need to be present, and in the latter they need to be not only present, but also directly adjacent to each other.

11.6 Evaluating the Literature

Evaluating published evidence comes in three flavours. These cascade down from conventional markers of quality, to how a study is reported on, and finally to the actual study design.

The first is via proxy measures, which is usually the first thing we learn about. Examples of proxy measures include peer review, author affiliations, society vs

commercial publishers etc These are proxy measures because the literature is not being evaluated directly.

The second is via reporting; how much information does a given publication provide on how they conducted their study? This might include things like publication of a protocol, availability of data and scripts etc. These kinds of reporting allow the reader to benchmark bias (protocol) and verify the reproducibility of findings (data and code). There are an increasingly large number of reporting frameworks available. In the biological sciences, two common frameworks include the Materials Design Analysis Reporting (MDAR) Framework for primary research and the Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology, an extension of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Guidelines originally developped for health research.

The third is with explicit evaluation of the study design and potentially each reported outcome. These tools generally ask questions related to appropriate study design and analysis as a way of evaluating study bias, and usually include questions like, was the study design appropritate for the research question, was the data collected in the most appropriate way, were the statistical tests applied appropriate for the data, etc. There are a lot of tools to choose from out there to engage in this kind of evaluation, not least because what is appropriate for an experimental study will be different from an observational study, let alone whether that experimental study used randomization or not.

These tools are generally used for large knowledge synthesis activities (systematic reviews and meta analyses), but can also be useful guides for systematically evaluating individual studies, especially as a means of building critical literacies.

The National Health and Medical Research Council of Australia, hosts a great list of tools for a variety of study types.

An evaluation of a variety of tools used in review protocols is available in Farrah, K., Young, K., Tunis, M.C. et al. Risk of bias tools in systematic reviews of health interventions: an analysis of PROSPERO-registered protocols. Syst Rev 8, 280 (2019). https://doi.org/10.1186/s13643-019-1172-8.

Chapter 12

APA Citations

Last updated 2023-02-10

This chapter is a brief overview of the 7th edition of APA. For a more in depth review, please consult the library's APA Citation Guide. For the full APA manual, please consult the Publication Manual of the American Psychological Association, available through the library.

Additional resources you may wish to consult include:

- The APA Style Blog - great for searching for examples not listed in the 7th edition
- Purdue OWL's website for still more examples

12.1 In-text Citations

In-text citations always appear right after the content you are summarizing, paraphrasing, or quoting. The format is as follows:

Summarizing or Paraphrasing

- (Author, YYYY)

Quoting

- (Author, YYYY, p. #) 1 page
- (Author, YYYY, pp. ##-##) Multiple pages

Narrative vs Parenthetical

If you mention the author or authors in text, you do not need to include this information in the brackets, "()". This is called a narrative citation.

Narrative in-text citation: Raimi (2018) outlines the risks and benefits of fracking through economic analysis and energy security benefits.

The alternative is called a parenthetical citation.

Parenthetical in-text citation: Several benefits and risks can be identified in the implementation of fracking for oil extraction. Considerations include regulation, water pollution, tremors etc. (Raimi, 2018).

Quick Format Guide

# of Authors	Narrative Example	Parenthetical Example
1	Bradley (2017)	(Bradley, 2017)
2	Janmaat and Rahimova (2018)	(Janmaat & Rahimova, 2018)
3 or more	Mei et al. (2018)	(Mei et al., 2018)

12.2 Reference List

NOTE

- Every source used in your in-text citations needs to be listed as part of your reference list, in alphabetical order by author(s)' last names.
- The word **References** should appear at the top of your reference list, and it should be centred and bolded on the page
- Titles should be written in sentence case, that is, capitalize the first word and only subsequent proper nouns. If the title is broken up by a colon (:), capitalize the first word after the colon.
- List all authors in the order that they appear in the source.

Journal article with a DOI (1-2 authors)

List all authors in the reference list and in-text citations.

Janmaat, J., & Rahimova, N. (2018). Managing drought risk in the Okanagan: A roll for dry-year option contracts? *Canadian Public Policy*, *44*(2), 112-125. https://doi.org/10.3138/cpp.2017-003

Journal article with a DOI (3-20 authors)

List all authors in the reference list and only the first author in the in-text citations.

Mei, Y., Yu, K., Lo, J. C. Y., Takeuchi, L. E., Hadjesfandiari, N., Yazdani-Ahmadabadi, H., Brooks, D. E., Lange, D., & Kizhakkedathu, J. N. (2018). Polymer-nanoparticle interaction as a design principle in the development of a durable ultrathin universal binary antibiofilm coating with long-term activity. *ACS Nano*, *12*(12), 11881-11891. https://doi.org/10.1021/acsnano.8b05512

Chapter 13

Academic Integrity

Last updated 2023-02-10

Academic integrity is the act of performing honest, responsible scholarship, much like scientific integrity is honest, responsible science. Academic integrity is very much about how we conduct ourselves in the pursuit of our studies, respecting those we learn from and work with and the contributions they make to our scholarly and scientific endeavours.

Learn more about academic integrity from the UBCO "Academic Integrity Resources for Students".

Chapter 14

Reference Management: Zotero

Last updated 2023-02-10

Imperative to managing our education and research is managing our readings and references. We can do this manually or we can take advantage of a tool - citation management software - specifically designed to keep your references organized, shareable, and easily citable.

There are several options out there. Some are great for inter-institutional collaborations, some for review projects that have you processing 10s of thousands of citations, and others are great for managing your day to day needs.

Here we're going to get you set up with a tool well suited for the latter - Zotero.

This overview will get you set up with an account, and importing your first reference.

Zotero is comprised of 4 components

1. A **desktop application**.
2. A **cloud application** that syncs with your desktop application.
3. A **web browser plugin** that gathers bibliographic information when you're looking at an article online.
4. A **word processor plugin** that facilitates creating citations and reference lists while writing.

14.1 Installation & first launch

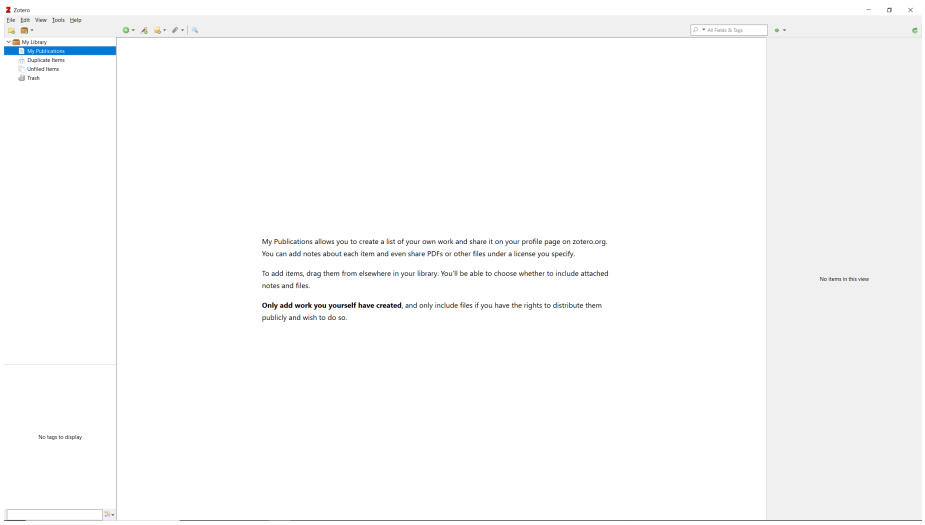
First, get your environment set up. **Quite your word processor** and **make sure that you are in a position to quite your web browser**.

Now, the **desktop application and word processor plugin**.

Head to https://www.zotero.org/download/, download Zotero and Run the installer.

First launch

Once installation is complete, if you launch the program, you should be greeted with the following:



That's step one and four covered, as the word processor plugin is installed automatically.

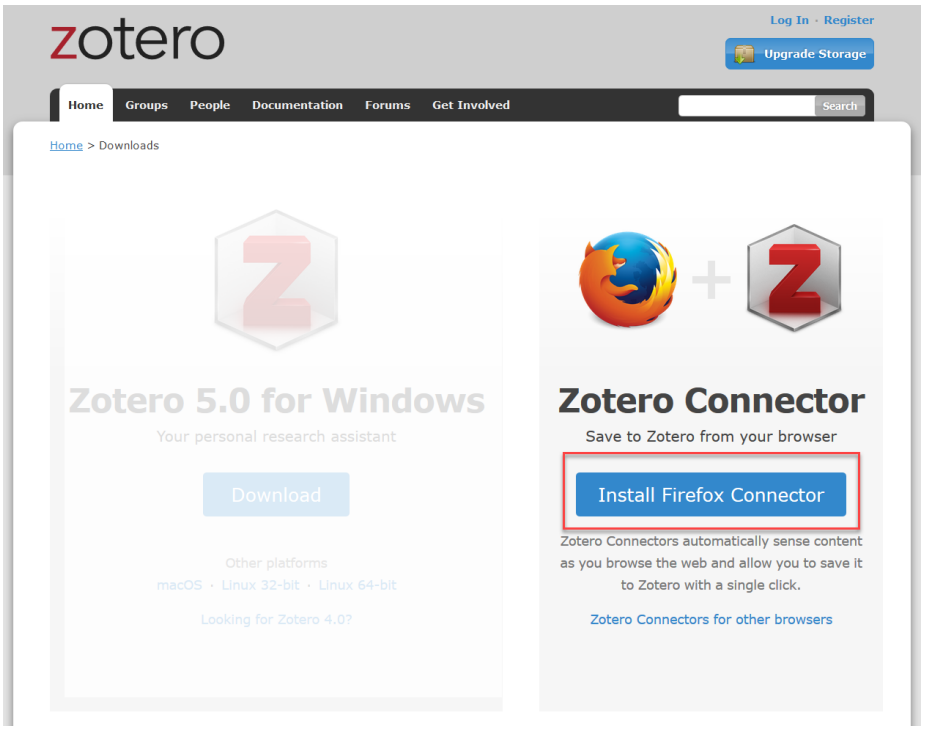
The plugin works with Microsoft Office, LibreOffice, and Google Docs. More on this later.

14.2 Browser plugin

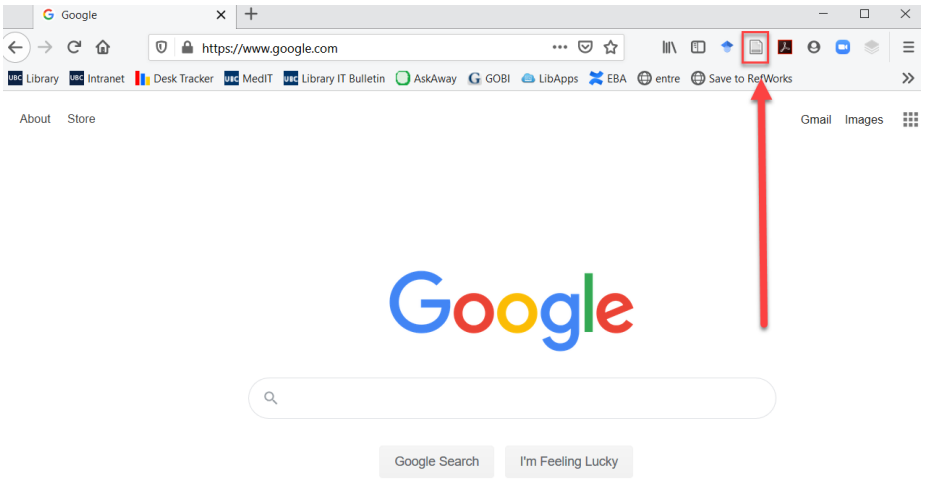
At this stage, Zotero will have launched a new window in your web browser, and you should see the opportunity to install the web browser plugin.

If this doesn't happen, simply return to https://www.zotero.org/download/ and install the plugin from there.

Zotero will automatically detect your browser. We recommend using either Chrome or Firefox; the plugin for Safari is currently in beta development and a bit more complicated to get configured.



Now, to make sure everything is lined up, quite and then relaunch your web browser. You should then see a small icon in the upper right hand corner. If you're in Chrome, you'll have a puzzle piece in this same spot, click that and you'll see you're extensions, including the Zotero extension.



14.3 Accounts & sync setup

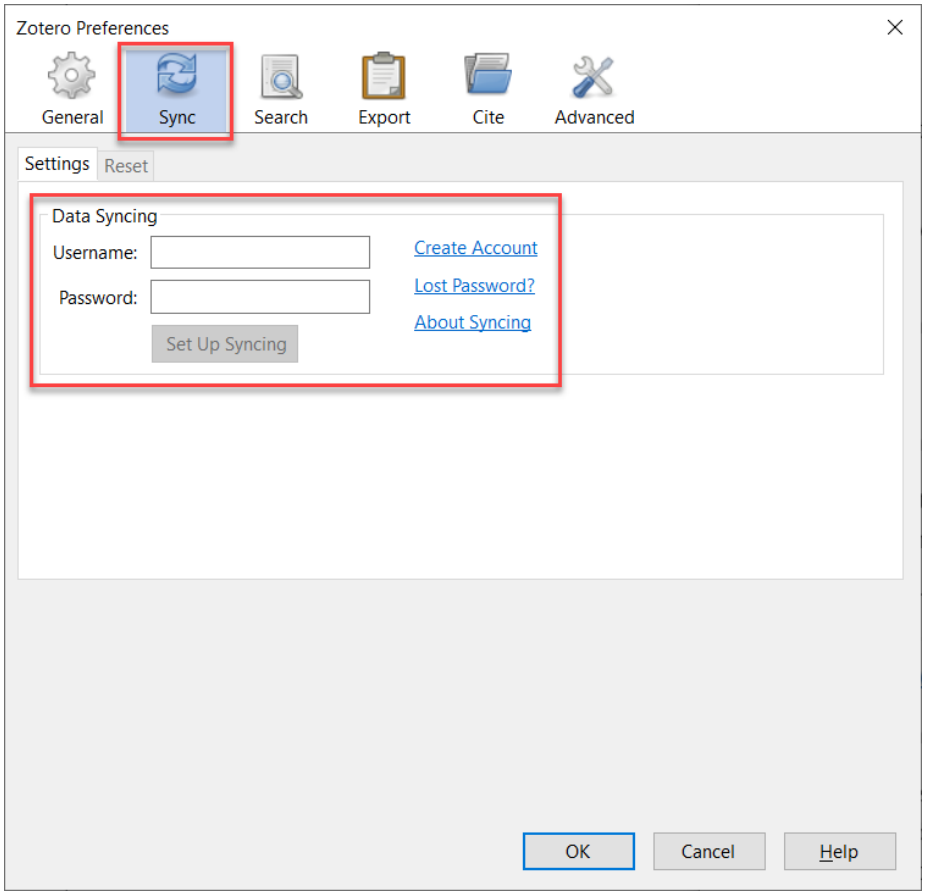
An account is required for cloud syncing and sharing folders.

Fill in the necessary credentials here https://www.zotero.org/user/register/ and we're almost done.

Connecting Zotero Desktop with the Cloud

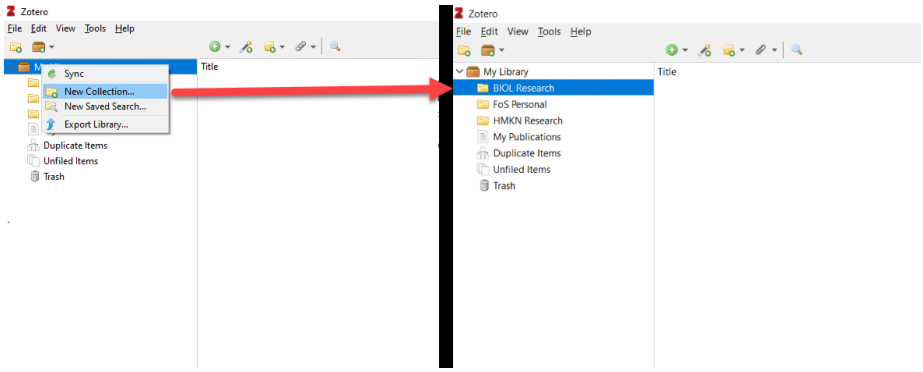
Now that you've created an account, let's get your Desktop application connected with you account.

In Zotero, go to `Edit > Preferences` and in the preferences tab go to `Sync`, fill in your credentials and click `Set Up Syncing`.



14.4 Adding citations

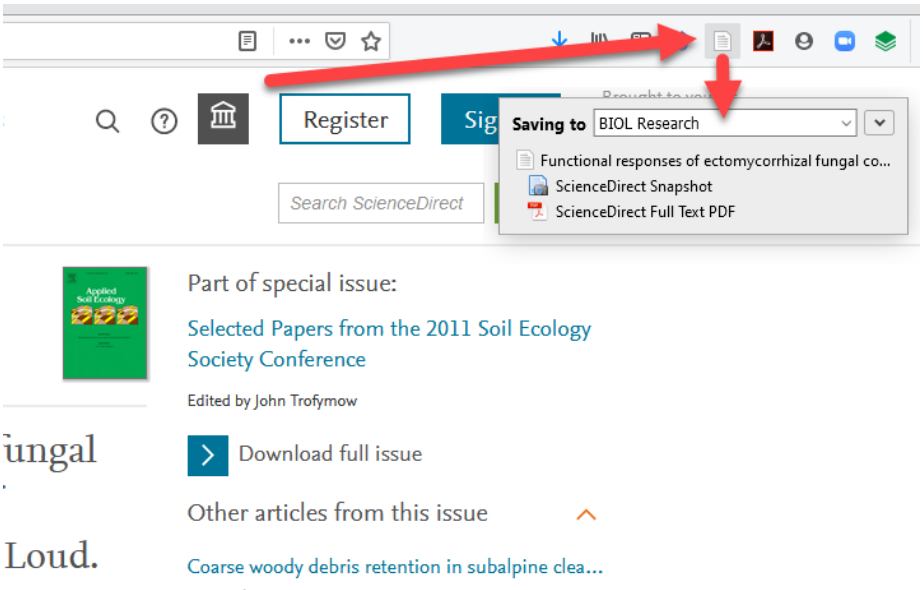
First things first - because we want to be organized - in Zotero, create a folder for your project by right clicking on `My Library > New Collections...` and give it a name, like `BIOL Research` because that's informative.



Say we're searching Web of Science Core Collection, we find an article we're interested in and we follow the link to read the article. And let's say that takes us here (que to follow the link): https://www.sciencedirect.com/science/article/pii/S0929139312000224

We'll see that when we're looking at an article, our web browser plugin changes to reflect this by now looking like a document.

Click on it, and Zotero will harvest both the requisite metadata and the pdf for you.

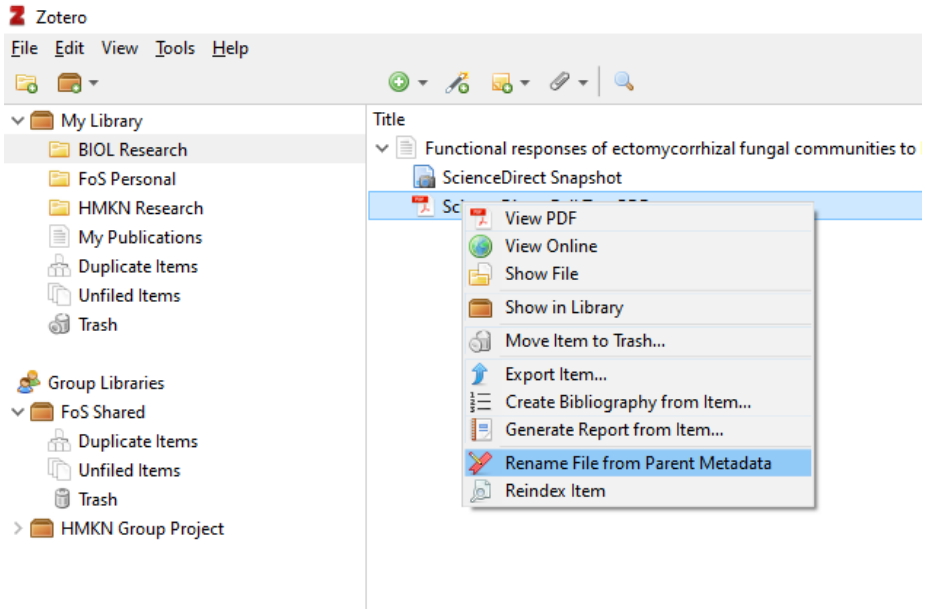


14.5 Renaming PDFs

Head back into Zotero and you'll see your reference.

Select the arrow adjacent to the title and you'll see your PDF.

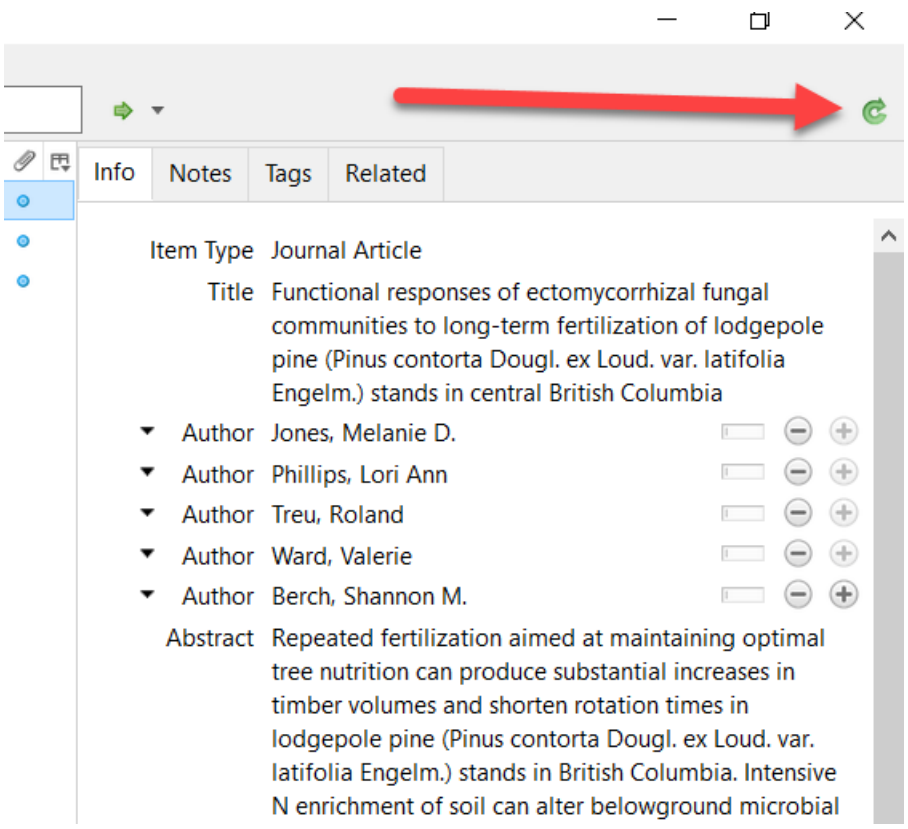
Right click the PDF and select `Rename File from Parent Metadata`. There! You've got a nicely titled article. Double click the PDF and it opens in your default PDF reader.



14.6 Syncing to the cloud

Click the sync button and send your citation and PDF to the cloud.

Sync happens automatically on every launch. And notes or highlights that you make on your PDF will also be synced and updated.



Zotero Cloud

Head online and log in to your online account at https://www.zotero.org/user/login/

You'll see your recently synced citation and associated attachments.

Chapter 15

Copyright

Last updated 2023-02-10

Copyright is the legal ownership of a work, like a book, image, graph, journal article. Copyright law governs how and when we are allowed to redistribute what someone else has produced and how others can redistribute what we've produced.

Learn more about copyright at UBC from the UBC Learning Commons.

R

Chapter 16

ggplot

ggplot is based on a grammar graphics - a way of approaching describing the construction of a graphic from common building blocks. Building a graphic with ggplot then follows some common patterns of construction.

At its most basic, we supply ggplot with a dataset and some aesthetics–that is, the variables we wish to display on the plot and how they should appear. Laslty, we define a plot type.

Step by step this looks like

1. call ggplot()
2. provide ggplot with a data set
3. provide ggplot with the variables of interest and their aesthetic properties
4. define a plot type with geom_*plotType*()

This is a brief introduction. For more in depth examples and solutions, check out ggplot2: Elegant Graphics for Data Analysis by Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen.

16.1 The Basic Graph

The following assumes you're using R and RMarkdown.

Install the libraries if needed

```
install.packages("ggplot2")
install.packages("palmerpenguins")
```

Load them

```
library(ggplot2) # for graphics
library(palmerpenguins) # penguins data set
```

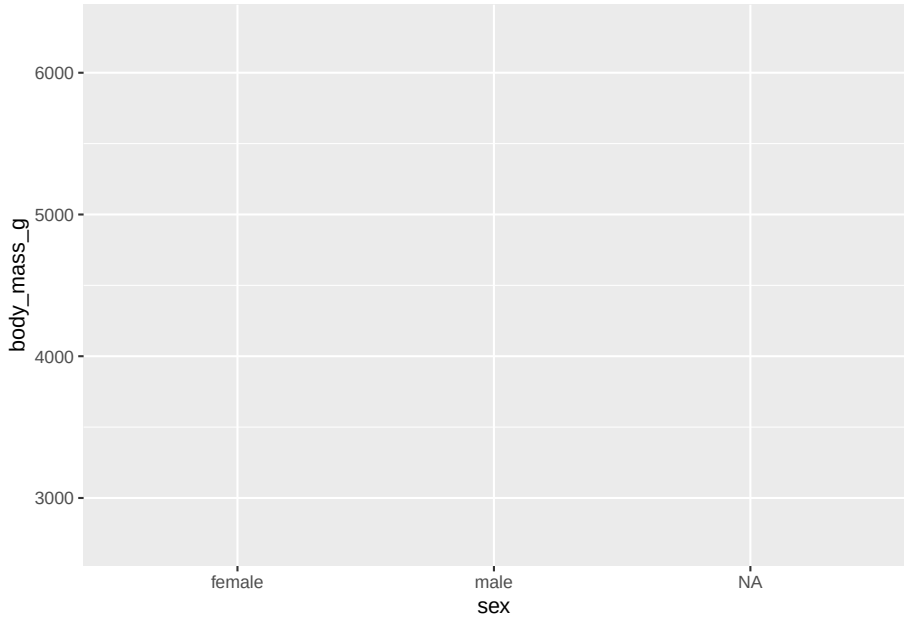
First, make sure we know a bit about our data set

```
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_l~1 body_~2 sex    year
##   <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
## 1 Adelie  Torgersen           39.1          18.7         181    3750 male   2007
## 2 Adelie  Torgersen           39.5          17.4         186    3800 fema~  2007
## 3 Adelie  Torgersen           40.3          18           195    3250 fema~  2007
## 4 Adelie  Torgersen           NA            NA            NA      NA <NA>   2007
## 5 Adelie  Torgersen           36.7          19.3         193    3450 fema~  2007
## 6 Adelie  Torgersen           39.3          20.6         190    3650 male   2007
## # ... with abbreviated variable names 1: flipper_length_mm, 2: body_mass_g
```

Next, we call ggplot, define our data set, and then the variables to plot on the x and y axes:

```
ggplot(data = penguins, aes(x = sex, y = body_mass_g))
```

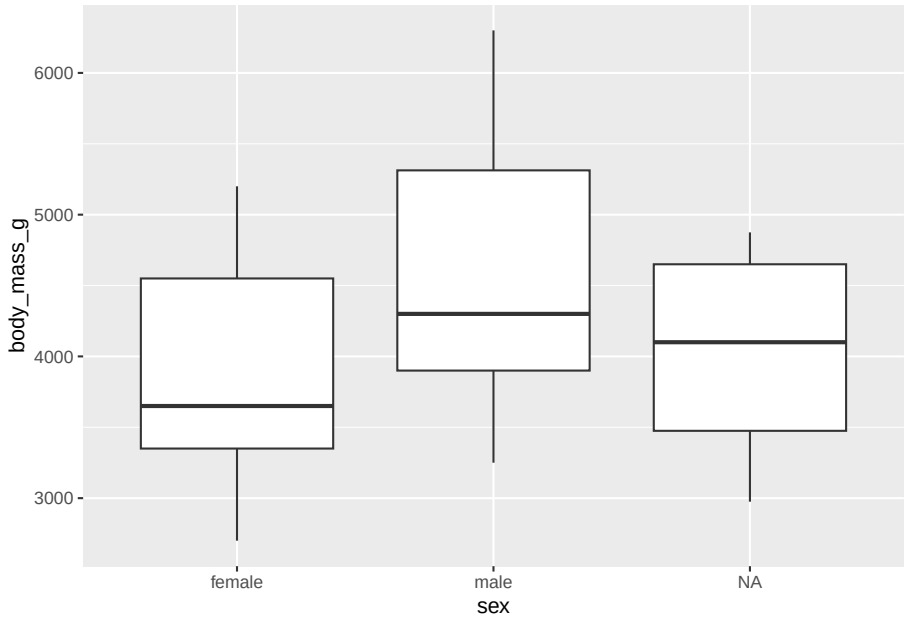


This sets us up with a blank grid with x and y axis ticks corresponding to your variable values and x and y axis labels corresponding to your variables.

Now we call a plot type to represent our data, in this case, a box plot

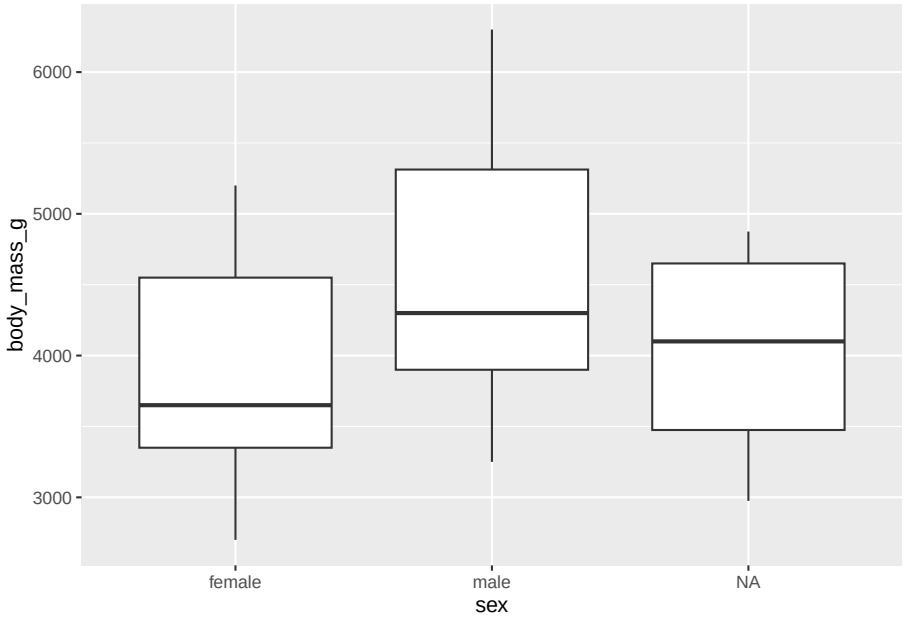
```
ggplot(data = penguins, aes(x = sex, y = body_mass_g)) +
  geom_boxplot()
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



We have some `NA` values in our data set. We won't worry about cleaning those up here. But we will suppress the error message with a code chunk option.

```
```{r, warning = FALSE}
ggplot(data = penguins, aes(x = sex, y = body_mass_g)) +
  geom_boxplot()
```
```

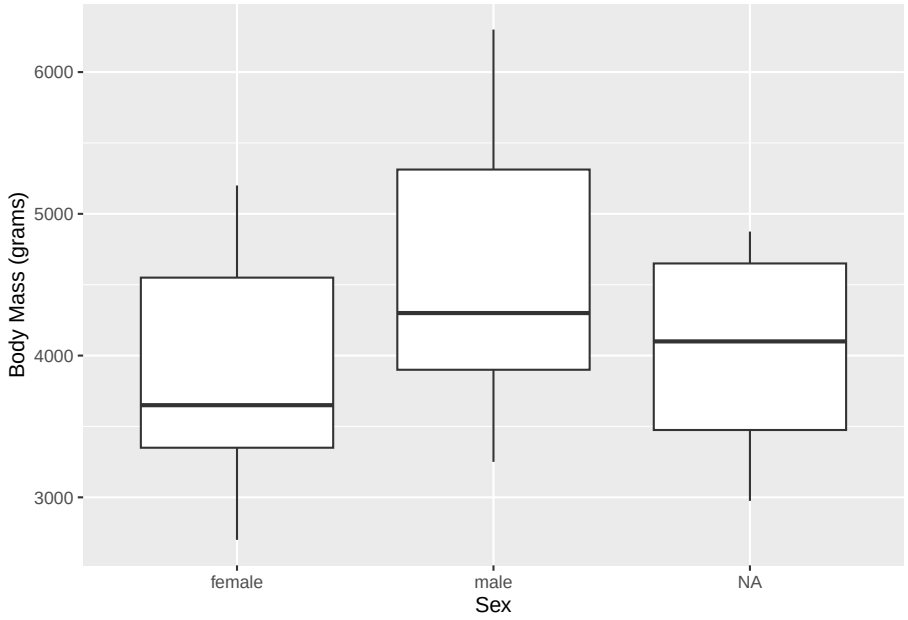


Chunk options are independent of ggplot itself and impact the knitting process of your RMarkdown document. For a more detailed overview of the code chunks options available to you, check out Xie Yihuis page on Knitr chunk options.

16.2 Labeling and captions

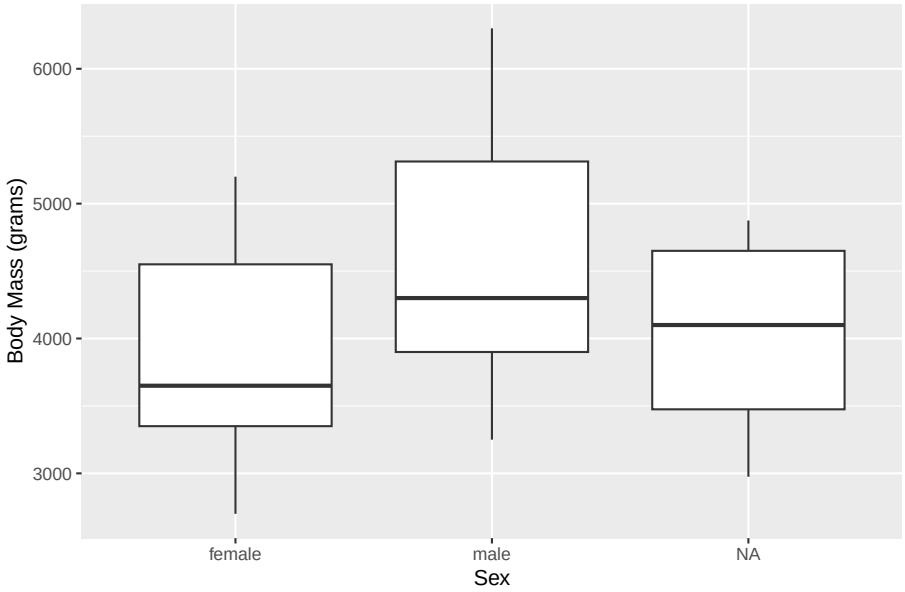
Labels default to our variable names, which may not be what we want on our graph. We can override this with `labs()`.

```
ggplot(data = penguins, aes(x = sex, y = body_mass_g)) +
  geom_boxplot() +
  labs(
    x = "Sex",
    y = "Body Mass (grams)"
  )
```



Captions are also important. There are two ways that we can add this information. The first uses `labs()`:

```
ggplot(data = penguins, aes(x = sex, y = body_mass_g)) +
  geom_boxplot() +
  labs(
    x = "Sex",
    y = "Body Mass (grams)",
    caption = "Palmer Penguins. Body mass grouped by sex. n = 344."
  )
```



Palmer Penguins. Body mass grouped by sex. n = 344.

The second is better if we're using RMarkdown. Note also that we can include markdown syntax, italicizing the *n*.

```
```{r, fig.cap = "Palmer Penguins. Body mass grouped by sex. *n* = 344."}
ggplot(data = penguins, aes(x = sex, y = body_mass_g)) +
  geom_boxplot(na.rm = TRUE) +
  labs(
    x = "Sex",
    y = "Body Mass (grams)"
  )
```
```

If you're knitting your report to html instead of pdf and you want to take advantage of automatic figure numbering–knitting to pdf will take care of automatic figure numbering by default–use the output option `bookdown::html_document2` in your YAML.

```
---
title: My Report
output: bookdown::html_document2
---
```

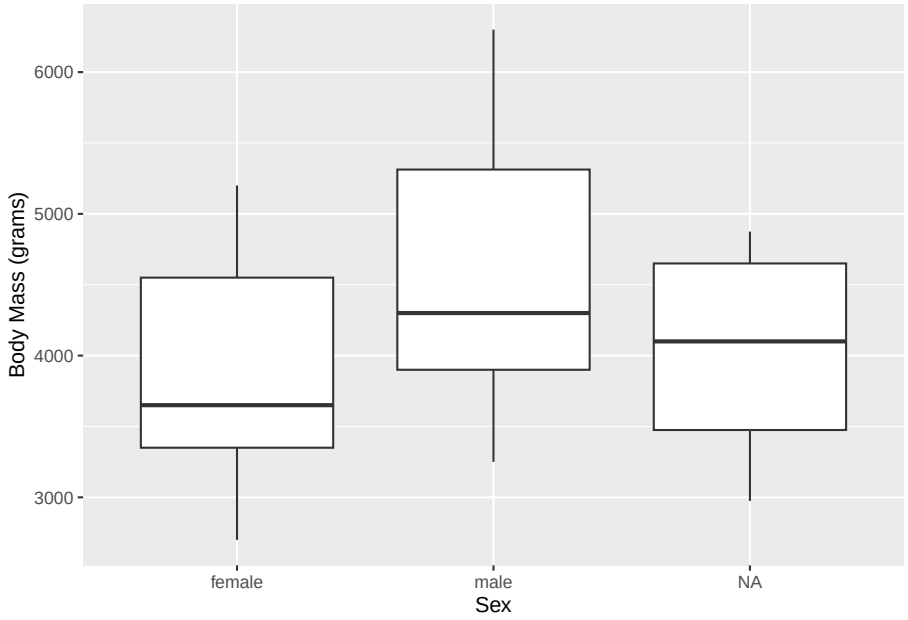


Figure 16.1: Palmer Penguins. Body mass grouped by sex. $n = 344$.

16.3 Size, shape & colour

Aesthetics such as size, shape, colour, and opacity are powerful ways of visually highlighting aspects of our data. These aesthetics can be mapped to individual variables, which is a great way to increase the number of dimensions–variables–we can plot. They can also be mapped to all data points. When used to map to a variable, we include this within the `aes()` argument. When used to map to all data points associated with a particular geom, we include this within the geom_*plotType*() argument.

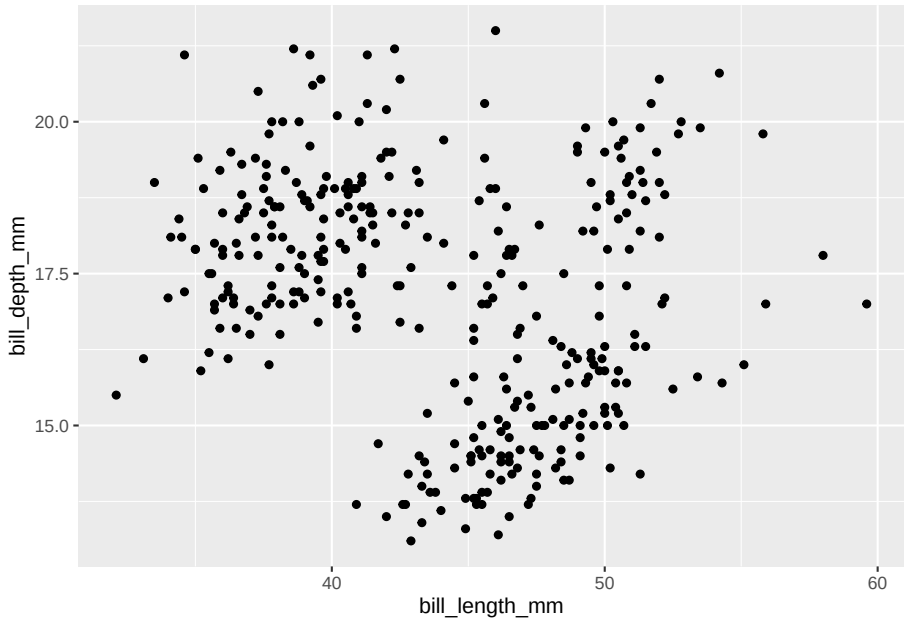
Options include:

- size
- colour
- fill
- shape
- alpha

Things like bars have both colour (the outside line) and fill (the inside body) properties. Things like lines and points have colour, but not fill.

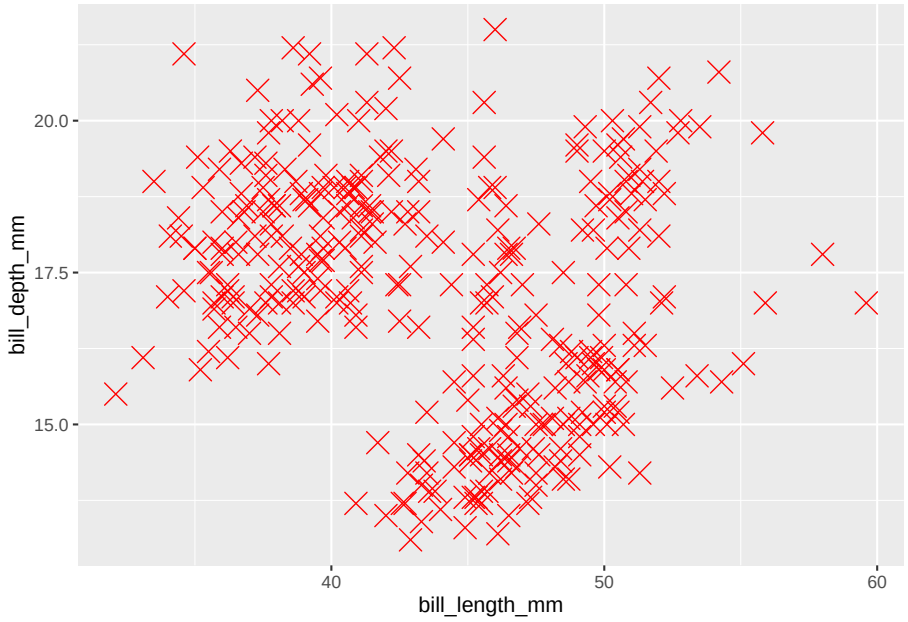
Plotting bill length against bill depth:

```
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm)) +
  geom_point()
```



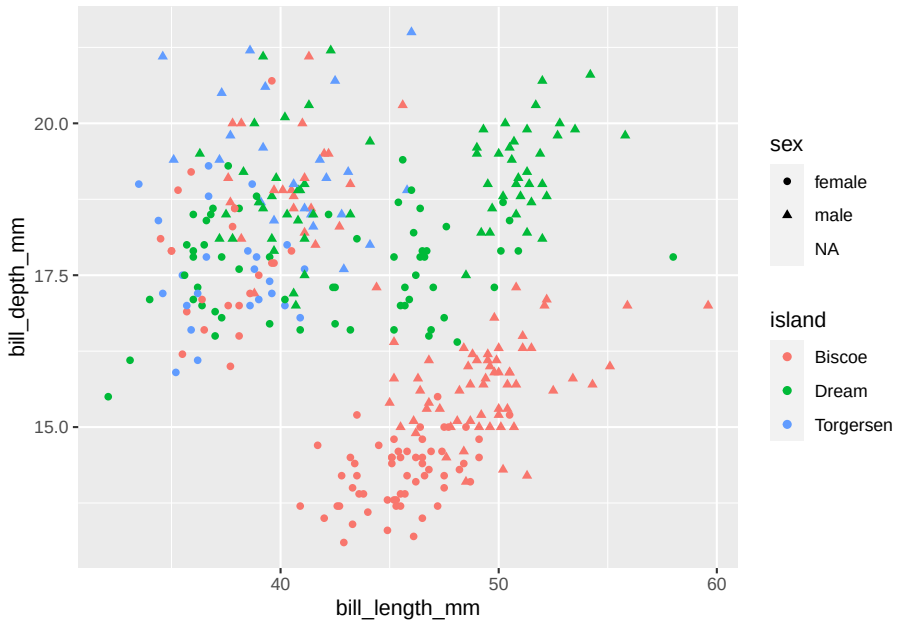
Adjusting the aesthetics of all data points

```
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm)) +
  geom_point(colour = 'red', size = 5, shape = 4)
```



Increasing the number of variables we're plotting:

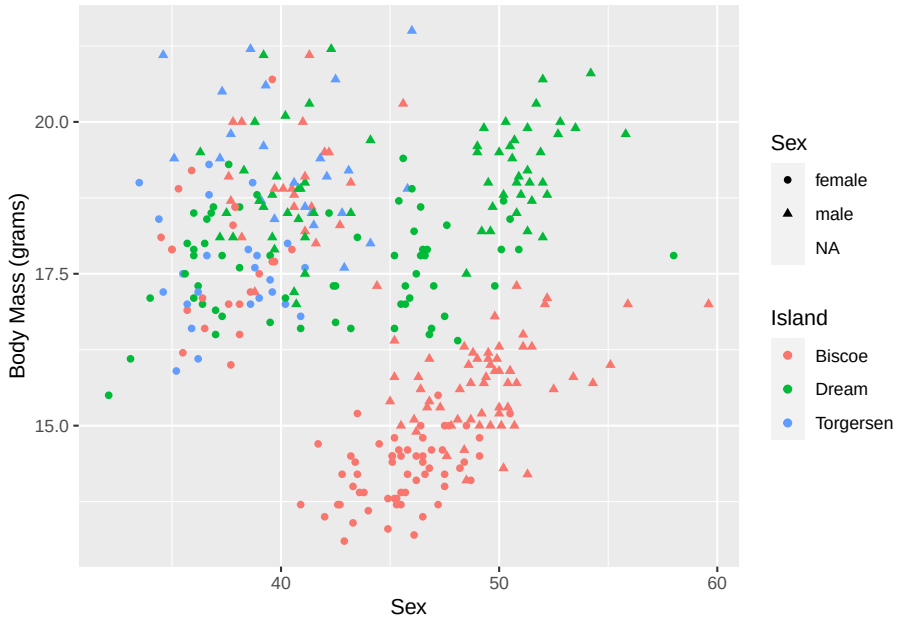
```
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm, colour = island, shape = sex))
  geom_point()
```



Adding more dimensions does not always increase the clarity of your graph, as the above example demonstrates!

And with some proper labeling:

```
ggplot(data = penguins, aes(x = bill_length_mm, y = bill_depth_mm, colour = island, sha
  geom_point() +
  labs(
    x = "Sex",
    y = "Body Mass (grams)",
    colour = "Island",
    shape = "Sex"
  )
```

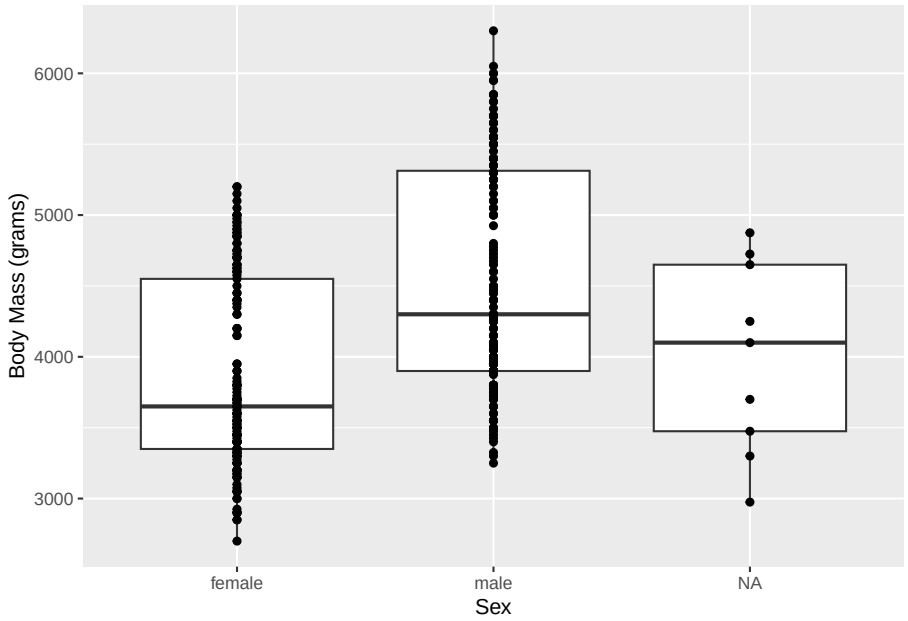


16.4 More than one geom

It can be handy occasionally to have more than one geom per plot, like including both lines and dots. To do this, we feed our data set and aesthetic mappings into `ggplot()`, and then call multiple geoms. Using the box plot example from earlier and adding individual data points:

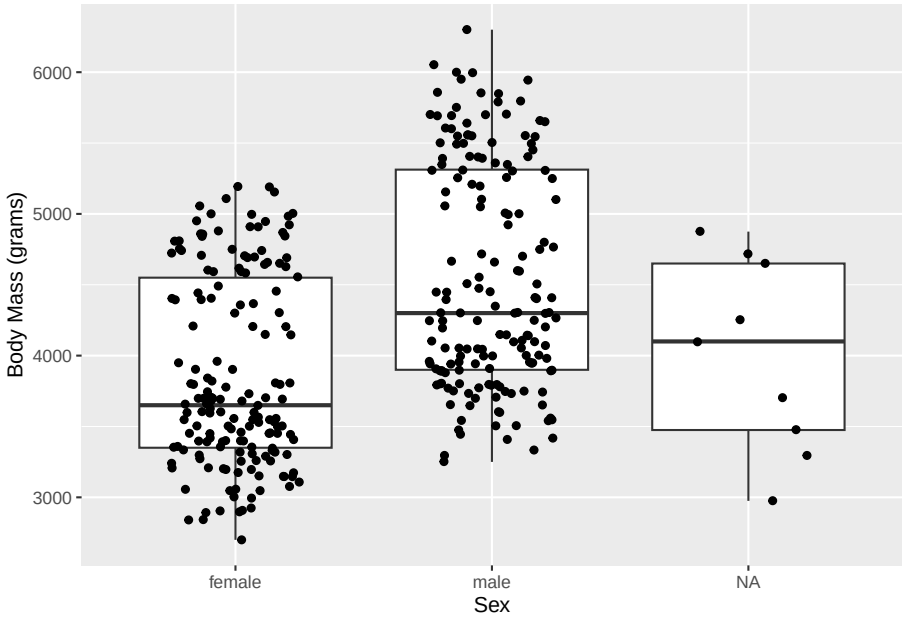
```
ggplot(data = penguins, aes(x = sex, y = body_mass_g)) +
  geom_boxplot() +
```

```
  geom_point() +
  labs(
    x = "Sex",
    y = "Body Mass (grams)"
  )
```



When there are a lot of data points, using 'jitter' to create lateral space between points can be useful.

```
ggplot(data = penguins, aes(x = sex, y = body_mass_g)) +
  geom_boxplot() +
  geom_jitter(width = 0.25) +
  labs(
    x = "Sex",
    y = "Body Mass (grams)"
  )
```



If you start typing `geom_` you'll see a full list of available plots to you with ggplot. If you'd like more in depth coverage of geoms built into ggplot, see the reference page section on geoms.

16.5 More than one plot

There are several ways to place more than one plot side by side. One of the easiest is to use patchwork.

Install

```
install.packages("patchwork")
```

Load

```
library(patchwork)
```

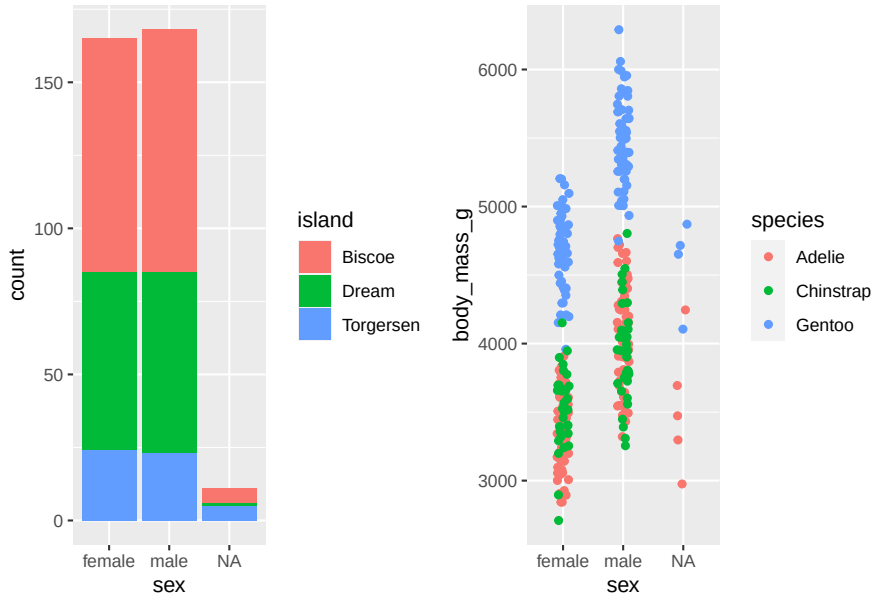
This time we store our plots as variables

```
barGraph <- ggplot(data = penguins, aes(x = sex, fill = island)) +
  geom_bar()

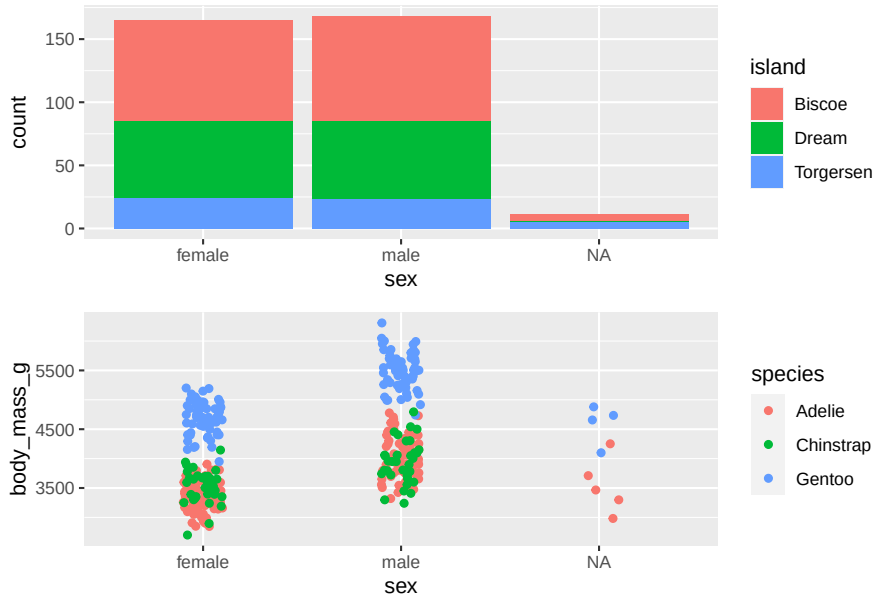
dotPlot <- ggplot(data = penguins, aes(x = sex, y = body_mass_g, colour = species)) +
  geom_jitter(width = 0.1)
```

Then patchwork will arrange them

```
barGraph + dotPlot
```




```
barGraph / dotPlot
```

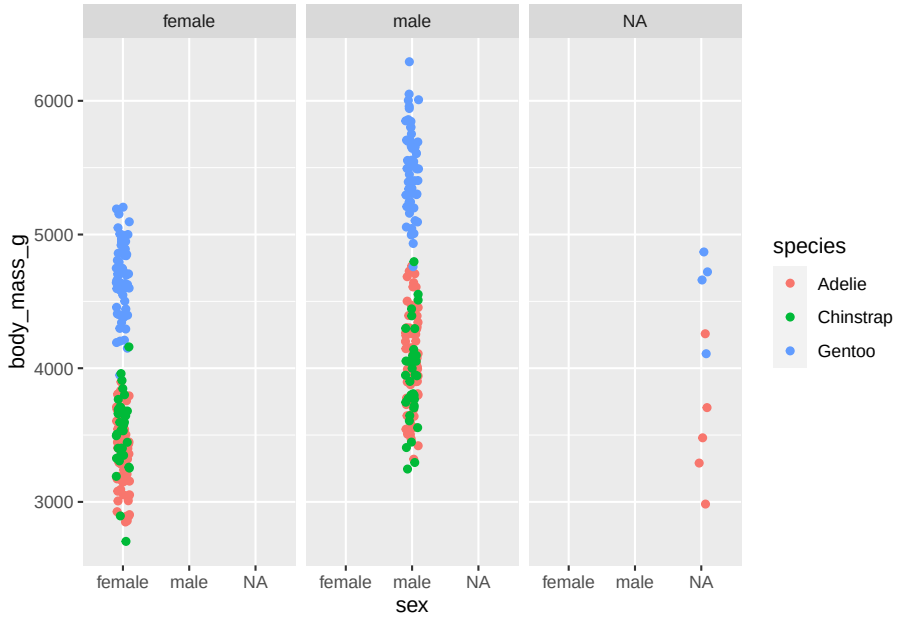


There are many ways in which patchwork can arrange plots. See the chapter Arranging Plots in *ggplot2: Elegant Graphics for Data Analysis* for more complex examples.

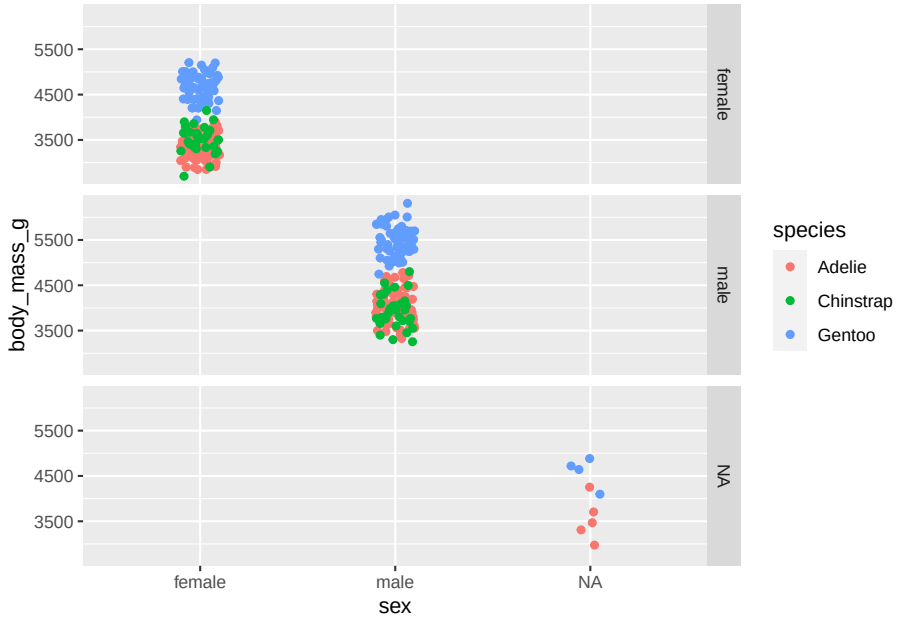
16.6 Faceting a Plot

We can also facet a plot with a call to `facet_grid()`.

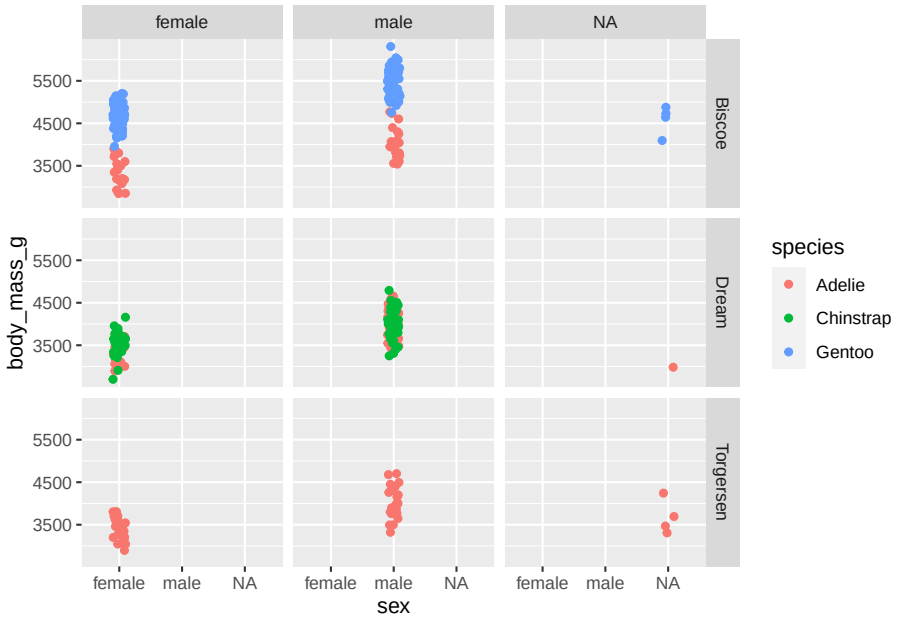
```
dotPlot +
  facet_grid(cols = vars(sex))
```




```
dotPlot +
  facet_grid(rows = vars(sex))
```

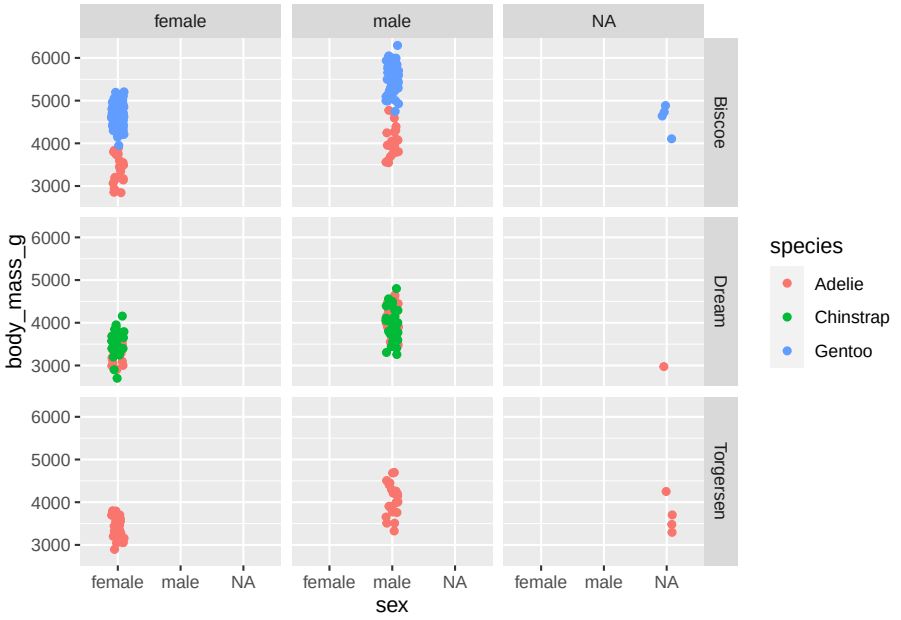


```
dotPlot +
  facet_grid(cols = vars(sex), rows = vars(island))
```



A slightly different notation is also valid to express the above graph

```
dotPlot +
  facet_grid(island ~ sex)
```



See additional facet options on the ggplot facet_grid() reference page.

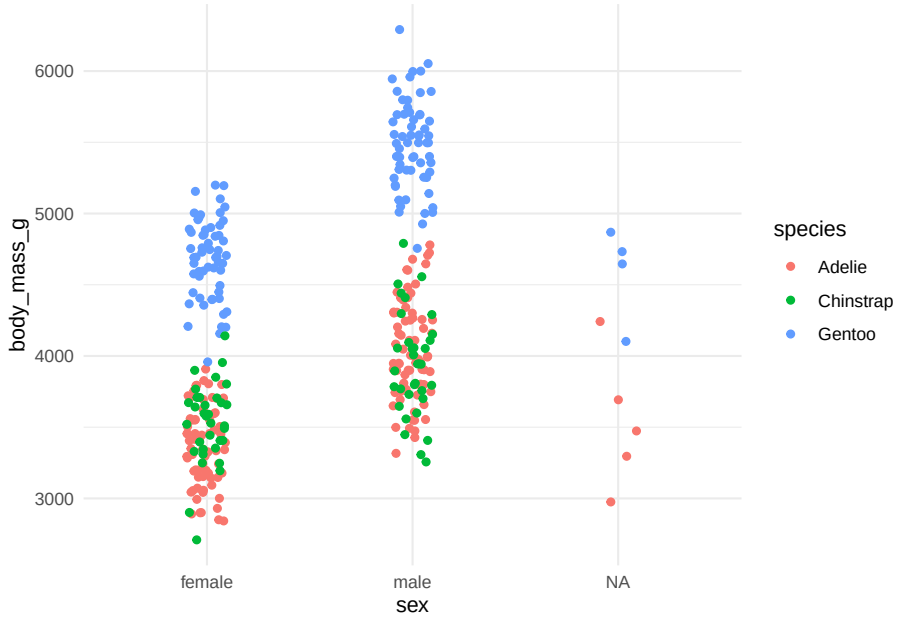
16.7 Cusomizing Look and Feel

Many visual aspects of your graph can be customized. Most of these are controlled within themes.

Themes

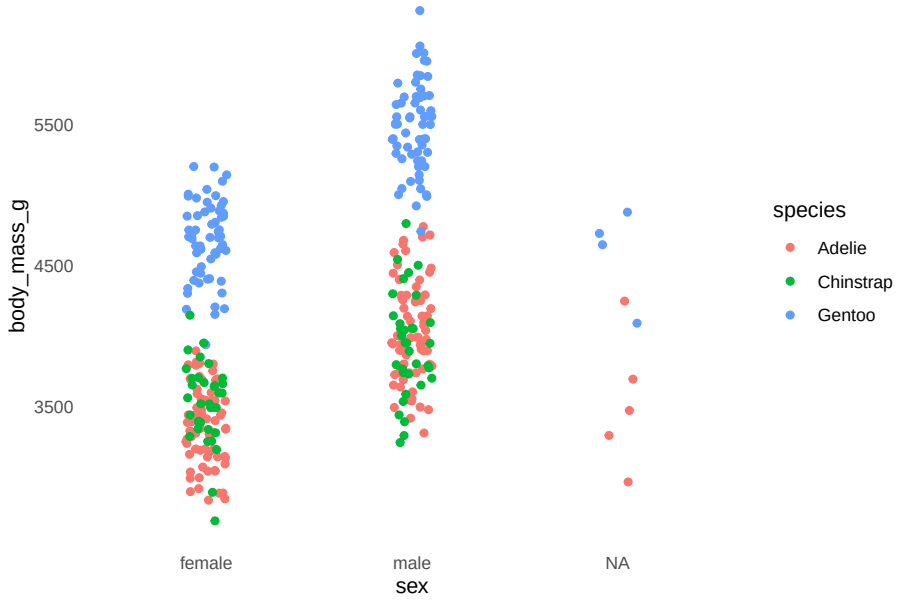
There are many built in themes, for example, minimal. If you start to type `theme_` RStudio will prompt you with a list of built in themes to choose from.

```
dotPlot +
  theme_minimal()
```



Within a theme, we can start to customize other elements. Things that we can customize included axes elements, legend elements, panel elements, and plot elements. For example, we can build on the theme minimal and remove the panel grids above, we do this with a separate, additional call to theme():

```
dotPlot +
  theme_minimal() +
  theme(
    panel.grid = element_blank()
  )
```



A full list of theme options are available on the ggplot theme reference page.

Colours

There are several ways of customizing the colours used in our plots, including using a custom colour palette.

It's critical to remember to use appropriate combinations of colour depending on if your data is divergent, continuous, or qualitative in nature.

Examples of each of these include:

Sequential - for ordered data

Diverging - for data with a central location from which other values diverge

Qualitative- for categorical data with not natural order

Instead of having to generate your own custom colour palettes, a good alternative are the palettes produced by ColourBrewer that already have due consideration to things like contrast, colour blind audiences, benig print friendly etc.

Install

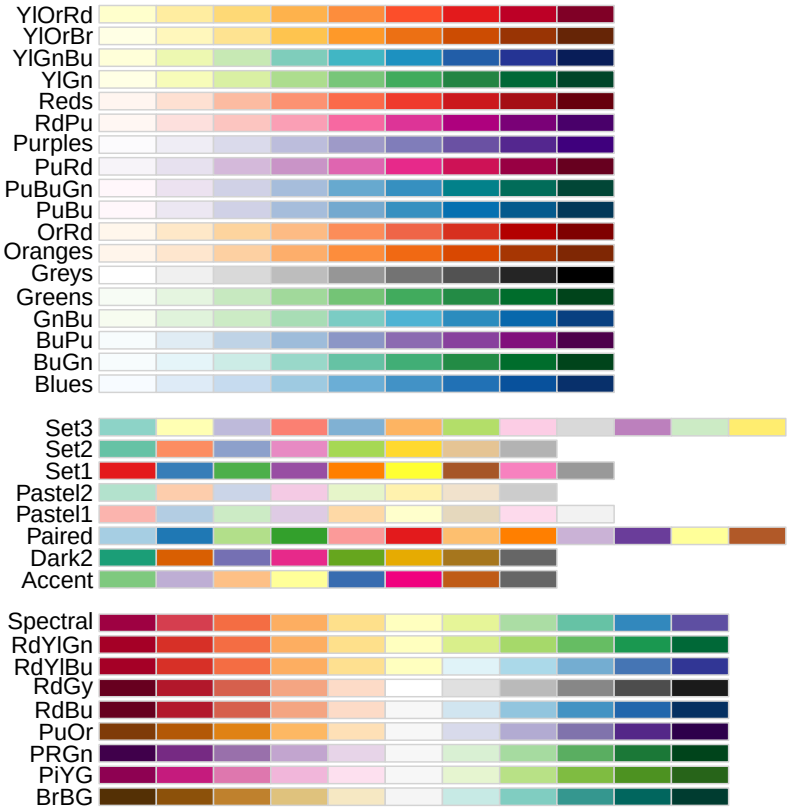
```
install.packages("RColourBrewer")
```

Load

```
library(RColorBrewer)
```

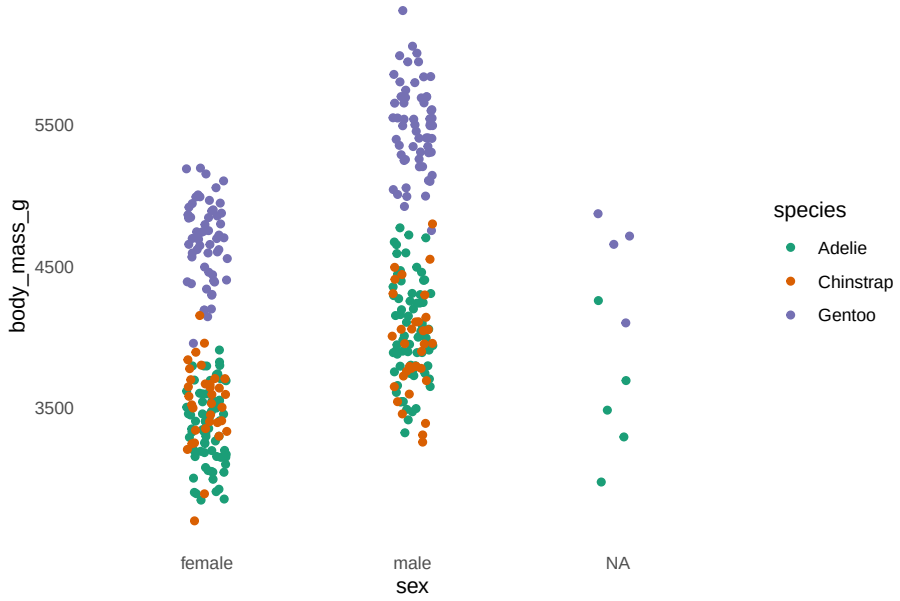
View the palettes available to us, noting it's grouped by sequential, qualitative, and diverging palettes.

```
display.brewer.all()
```

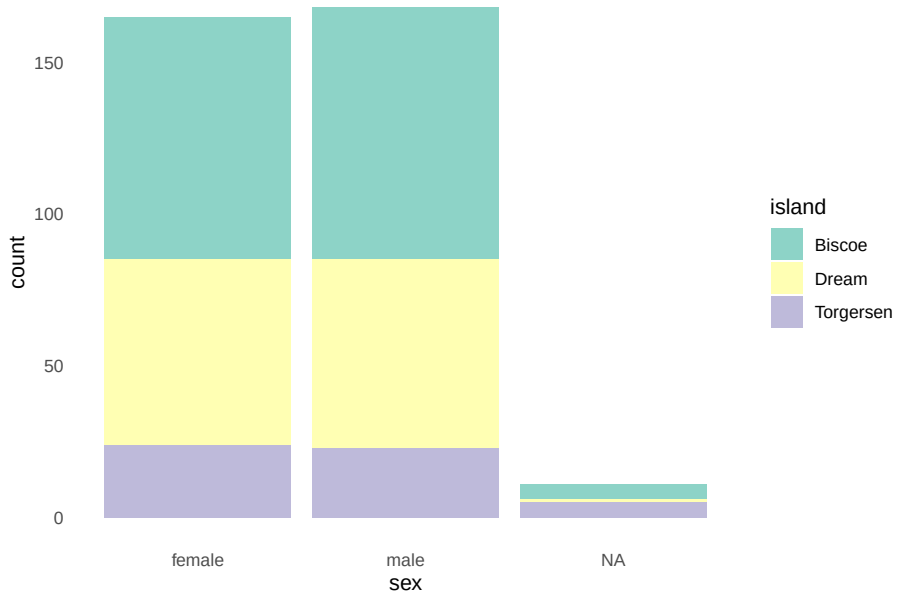


Use a palette, by making an additional call to either `scale_color_brewer()` or `scale_fill_brewer()` depending on you graph.

```
dotPlot +
  theme_minimal() +
  theme(
    panel.grid = element_blank()
  ) +
  scale_color_brewer(palette = "Dark2")
```




```
barGraph +
  theme_minimal() +
  theme(
    panel.grid = element_blank()
  ) +
  scale_fill_brewer(palette = "Set3")
```


150
100
50
0
count
female
male
NA
sex
island
Biscoe
Dream
Torgersen

Glossary

Chapter 17

Glossary

Last updated 2023-02-10

A priori hypothesis: Hypothesis that is generated before the research study takes place. Presenting the hypothesis before the study takes place helps in avoiding replacing the hypothesis later with one that fits the data better aka hypothesizing after the fact (HARKing).

Alternative hypothesis: In contrast to the null hypothesis, the alternative hypothesis suggests there is a relationship between phenomena, variables, or populations. In other words, any differences are not the result of random chance.

Analysis of variance (ANOVA): A statistical test used to compare the mean of a numeric variable in relation to a single categorical variable that has more than two groups.

Assumptions: There are often assumptions associated with statistical tests. This means that for the test to provide reliable results, the data must meet specific criteria or conditions. These assumptions need to be checked prior to conducting any analyses.

Bias: Error is introduced and false conclusions might be drawn because our sample doesn't meet established standards for faithful representation of our population of interest.

Binomial distribution: A discrete probability distribution of the number of successes where there are exactly two possible outcomes (e.g., success and failure).

Binomial test: A statistical test that determines the probability of getting a particular proportion when there are exactly two possible outcomes (e.g., success and failure).

Burden of proof: The obligation that when a causal link is suggested that evidence to support this link must be presented. This can be accomplished

through independent replication of studies where if they demonstrate the same conclusions it reinforces the validity of the causal link between those variables.

Chi-square (χ^2) contingency test: A statistical test used to assess whether there is an association between categorical variables. This test is used on contingency tables that are larger than 2 x 2.

Chi-square (χ^2) goodness of fit test: A statistical test used to test how well an observed discrete frequency (or probability) distribution fits some specified expectation.

Choice experiment: A type of scientific experiment where a categorical response variable is measured in relation to a manipulated independent variable.

Citizen science: When members of the public engage in the research process. Often involving collaboration with researchers.

Clinical trials: Experiments that involve i) human participants who are assigned in advance to a group that receives a particular treatment designed to produce a biomedical or behavioural result and ii) evaluation of the effect of the treatments

Coefficient of determination (R2): The proportion of the variance in the response variable that can be explained or predicted by the independent variable. It describes the strength of the relationship between two variables.

Coefficient of variation (CV): A relative measure of variability that indicates the size of a standard deviation in relation to its mean.

Comma-separated values (CSV) file: A plain text file where each line of the file represents a record and each field (column) entry for that record is separated by a comma. This file format is frequently used by researchers to store data.

Confidence interval: An estimated range of values that has an associated probability. The probability describes the likelihood that this range of values will contain the true value of a parameter (ie. mean). For example, a 95% confidence interval suggests we can be 95% confident that the true parameter lies within that range of values. Or in other words, on average we can expect the true parameter to lie in this range, 95% of the time.

Confirmatory research: Researchers used a well designed experiment to test the validity of predetermined hypotheses that can be disproved.

Continuous quantitative data: Numeric data that lies on a continuum. So there are infinite possible values between integers and our data collection tool or convention determines to what decimal point we record the values to. Some examples include temperature, height, and distance.

Critical analysis: Careful examination and evaluation of all parts of a research article including consideration of the study's strengths and weaknesses as they relate to study design, implementation, data collection, data analysis, and interpretation.

Data transformation: A process where the format of the values within a dataset are changed. For example, all of the values within a dataset might be log transformed. This is often done when the original data does not meet the assumptions of a particular statistical test. After the data transformation researchers will re-assess those assumptions to see if they can perform the test on the newly transformed data.

Delimiters: One or more characters used to separate strings of text. Some commonly used delimiters include commas (,), colons(:), semicolons(;) and pipes(|).

Descriptive statistics: A number used to summarize or describe a given data set or sample. Examples include mean, median, mode, standard deviation, and interquartile range.

Discrete quantitative data: Numeric data that encompasses only whole integers and no fractions in-between. For example number of people or number or petals on a flower.

Diversity: The practice or quality of having individuals who vary in terms of social class, ethnic background, sexual orientation, gender, religion, ability, etc.

Effect size: A measure of the degree of association between one variable and another, or in experimental contexts, of the impact of one variable on another.

Equity: The practice of treating all segments of society in such a manner that everyone has a similar chance of achieving a given outcome. Some individuals and groups may need more or different support than others to achieve that outcome. "Equality" , in contrast, refers to the practice of providing identical support and opportunities to all.

Exploratory research: Research that is performed to gain a better understanding of an existing problem. For example, it might give rise to hypotheses that can then be tested through confirmatory research.

File and data management: Refers to practices used to collect, generate, and store data and files throughout the research process. Researchers should document what type of data they have collected, the methods used, and any relevant context. Files and data should be stored such that they are organized, accessible, and interpretable by both the researcher and others.

Fisher's exact test: A statistical test used to assess whether there is an association between categorical variables. This test is used on contingency tables that have exactly 2 x 2 dimensions.

HARKing: A form of questionable research practices where the researcher changes their hypothesis after the study is conducted so that the hypothesis better fits the data. In other words, the researcher suggests that this after the fact hypothesis was formed a priori. This has a number of implications including harming the progress of science by preventing the research community from identifying already falsified hypotheses, contributes to the replication crisis, and

it increases the probability that the findings are not reproducible or generalizable in the population of interest.

Hypothesis: A proposed explanation for an observed phenomenon. Often structured in an "If… then… because…" format. Hypotheses must be present a priori, be falsifiable, and measurable.

Hypothesis testing: Typically involves setting a null and alternative hypothesis and performing an appropriate statistical analysis to test those hypotheses. Often used in confirmatory research.

Inclusion: The philosophy or practice of considering individuals from diverse backgrounds in relation to the community, organization, or society, and ensuring that they feel that they belong, supporting them in giving their best efforts, and giving them equal opportunities to advance and participate in decision-making.

Interquartile range: A descriptive statistic that measures the variation within the middle section of a set of values. Specifically, it describes the range between the first and third quartiles of a set of values.

Linear regression: A statistical method used to model the linear relationship between independent and dependent variables.

Literate programming: A coding paradigm where natural language is written alongside or between lines of code to provide an explanation for the code's logic. This practice helps enhance reproducibility and understanding by guiding readers through the programmers thought process.

Literature review: A review of scholarly sources related to a specific research question or topic. Involves recording a list of research studies consulted, how they were found, and the strengths, limitations, and weaknesses of each.

Long format data: A method for organizing data where all of one subject's observations are represented by distinct rows.

Markdown: Markdown is a markup language that is used to format plain text files to help us provide additional meaning to our content. For example, using Markdown you can use bold, italics, and create tables. Markup languages are ideal authoring tools because they work on a principle of separating out content from formatting.

Mean: A commonly used descriptive statistic that measures the central tendency of a numeric variable. Specifically, the mean is the arithmetic average of a group of values.

Measured response experiment: A type of scientific experiment where a quantitative response variable is measured in relation to a manipulated independent variable.

Median: A descriptive statistic measuring the central tendency of a numeric variable. The median is the value separating the upper and lower halves of the variable. In other words the middle value of a group of numbers.

Metadata: Data that provides information about other data.

Mode: A descriptive statistic for a either a numeric or categorical variable. The mode is the value that appears most frequently.

Nominal categorical data: In contrast to ordinal categorical data, this data has two or more discrete categories that have no natural order. For example, hair colour or blood type.

Null hypothesis (Ho): Used alongside the alternative hypothesis in hypothesis testing. The null hypothesis states that there is no significant effect or relationship between phenomena, variables, or populations. Rather any differences observed are the result of random chance.

Odds ratio: Measure of the relative odds of the occurrence of a specific event (ie. cancer) given the exposure to a variable of interest (ie. smoking). This ratio is often used to determine the odds of health related outcomes.

One-sample t-test: A statistical test used to compare a numeric response variable (ie. mean) to an expected value.

Open notebooks: Involves i) publishing or linking to data on an online platform before results are published in a peer-reviewed journal; ii) making information about the methodology and equipment used in a study publicly available; iii) openly discussing both positive and negative results in real time, as they are obtained.

Open science: A movement and set of practices intended to combat the replication crisis, QRPs, and style trumping substance by making all parts of the scientific research process transparent and accessible, allowing for a critical review of how a study was conducted, ultimately enabling that study to be independently replicated. It also involves changing scientific culture to reward not just novel findings, but also the many other aspects of conducting good scientific research.

Ordinal categorical data: In contrast to nominal categorical data, this data has two or more discrete categories have a natural order but there is no clearly defined interval between each category. For example, storm severity is often classified by stages - Stage 1, Stage 2 … Stage 5 - where Stage 1 is less severe than stage 5. However, we don't know how much more severe one stage is than the next.

P value (p): The probability of getting a result that is the same or more extreme than what was observed. If the probability of getting that result due to random chance is sufficiently low, then it could be interpreted that there is a significant relationship. In contrast, a high p value indicates a larger likelihood that the result was due to random chance and therefore there may be no significant relationship. The p value required to establish significance is set by the researchers in advance of the study and is known as the significance level (α).

Paired t-test: A statistical test used to compare the means of two samples where an observation in one sample can be paired with an observation in the other sample. For example, observations might be linked because they were before and after observations on the same subject or in the same place,.

Participatory research: Turns the relationship between researcher and subject into a partnership, where both contribute to the research question, methods, and outcomes.

Pearson correlation: A statistical test that measures the linear correlation between two numeric variables.

Peer review: Peers of the author critically review the author's study. Traditionally peer review focused on the evaluation of studies prior to publication, however open science practices suggest additional peer review at the study design stage prior to implementation. This ensures the study design meets accepted quality standards before it is conducted.

Plain text: Simple text that is human readable. It can includes letters, numbers, symbols, and spaces but does not have any special formatting and is not computationally tagged.

Post-hoc test: If a significant result is found when performing a statistical test, post hoc tests can be done to provide more details about where those significant differences are arising from. They are another form of statistical tests.

Prediction: The expected results of an experiment based on a specific hypothesis.

Probability: Describes the likelihood of an event occurring. For example, when a fair coin is flipped the probability of getting tails is 0.5.

Proportion: A number between 0 and 1 that represents the fraction of the total population with a certain attribute. For example, if 10 students have red hair in a school with 100 students, then the proportion of students with red hair is 10/100 or 0.1.

Questionable research practices: A grey area of scientific practice in which researchers do not engage in outright misconduct such as fraud or plagiarism, but may unwittingly break rules of acceptable scientific practice in the pursuit of novel and promising results.

R: a programming language and free software for statistical computing. When used throughout the research process, it allows for openness in the research workflow and computational reproducibility.

Raw data: Data that has been collected but not yet processed in any way.

Relative path: In contrast to an absolute path, a relative path is a URL that only contains a portion of the path. It is relative to the root of the document and thus should start the path with the directory name that contains the document. For example, if you are writing a Markdown document and would like

to include an image of a mealworm, place the mealworm image into the same directory (folder) as the Markdown document and the relative path may look like /BIOL116/project/mealworm.png. Whereas the absolute path might look like C::/Documents/BIOL116/project/mealworm.png.

Random assignment: Assigning participants of a research study to each condition using a method of randomization. This ensures that each participant has an equal chance of being placed in each condition and helps to minimize bias.

Range: A descriptive statistic or measure of variation for a set of values. Specifically, it measures the difference between the highest and lowest values.

Registered report: A publishing format where peer review of the research question and methods is conducted prior to data collection. High quality protocols are then provisionally accepted for publication if the authors follow through with the registered methods.

Replication: Thorough repetition of a study, using the same methods but different data.

Replication crisis: Many studies cannot be competently analyzed or replicated. This is because critical information about them—design, data, methods, lab notes, analyses and code—may not be made available, or may be poorly communicated. This problem is escalated further because new and original findings are considered more exciting than re-testing or replicating previously conducted studies.

Reproducibility: Obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis.

Research transparency: The quality or practice of revealing all inputs and outputs of the research process clearly, as well as making evident the exact reasoning and process used in coming to a decision or taking actions in research, in such a way that the study can be replicated. As well, transparency means taking care to disclose important information in a respectful and responsible fashion.

Research lifecycle: The traditional research cycle involves five stages, 1) develop idea, 2) design study, 3) collect and analyze data, 4) write report, and 5) publish report. Traditionally, peer review has been conducted after writing the report and prior to publication. However, open science proposes revising the research life cycle by introducing an additional peer review after the study design stage.

Scientific method: An empirical method for acquiring knowledge which includes making an observation, asking a question, forming a hypothesis, making a prediction based on the hypothesis, and testing the prediction.

Scientific integrity: Adhering to professional values and practices when conducting, reporting, and applying the results of scientific activities. Adherence to

these values ensures objectivity, clarity, and reproducibility, and provides insulation from bias, fabrication, falsification, plagiarism, inappropriate influence, political interference, censorship, and inadequate procedural and information security.

Significance level (α): The probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.05 means there is a 5% chance that researchers might conclude a significant relationship or difference exists when there is no true relationship or difference. This is also known as the Type I error rate. The significance level is set by researchers before conducting a study and the p value result is compared to the α to determine if there is a significant relationship or difference.

Spearman rank correlation: A nonparametric statistical test that measures the statistical dependence of ranking between two numeric variables.

Standard deviation: A type of descriptive statistic that is used to quantify the amount of variation within a set of values. A set of values with a large standard deviation exhibits high variability whereas a low standard deviation indicates the values are close together.

Standard error: The standard deviation of the sampling distribution for a specific parameter. For example, if the parameter of interest is the mean, the standard error of the mean would be the standard deviation of the sampling distribution of the mean.

Statistical analysis: Involves collecting, organizing, exploring, interpreting, and presenting data to uncover patterns or trends in the data. Involves using statistics to describe the study sample and use that sample to make inferences about the population of interest.

Statistical significance: If a result is determined to have statistical significance it means that the result from the study is not likely to have occurred randomly or by chance. In other words, the result is likely to be caused by something other than chance. The significance level ( ) is set by the researcher in advance of the study being performed. Often is set to 0.05, which indicates a 5% chance of making the wrong decision and determining that the null hypothesis is false when it is in fact true.

Study power (aka statistical power): The probability that a random sample taken from a population will lead to rejection of the study's null hypothesis if that null hypothesis is in fact false. That is, power is a measure of how reliable a study is as a test for its hypothesis; power is positively influenced by things like large sample sizes and relationships characterized by large effect sizes.

Two-sample t-test: A statistical test used to compare the means of two independent samples.

Variance: A descriptive statistic that measures how far a set of numbers is spread out from their average value. In other words, a high variance indicates

the values are spread further from the mean whereas a low variance indicates they are close to the mean.

Version control: Saving changes to files while retaining the changes on all previous versions of the file. This practice contributes to transparency and openness in science.

Wide format data: A method for organizing data where each row represents an individual subject and each column represents an observation for that subject.

Appendix

A1: Magnification of A Drawing

To determine the scale of a sketch or drawing, we must know the size of the object we're drawing and the size of the image we're representing this object with. Mathematically, this is expressed as

$$m = \frac{d}{o}$$

where

- m = magnification (or scale) of drawing
- d = size of drawing
- o = size of object

17.1 Object Size

Two different methods are described here to calculate an object's size under the microscope. The first uses estimation whereas the second, more precise measure, involves calibration of an ocular micrometer. It is very likely that you frequently won't have access to an ocular micrometer, in which case the estimation method should suffice.

17.2 Method 1: Estimation

Field of View Diameter

First, determine the field of view diameter of your microscope for each objective setting.

These values might be provided to you in your lab manual. If so, proceed to the next section, estimating object size.

To calculate field of view diameter, we first need to know the field number (FN). When you look at your microscope eyepiece, the FN is displayed after the ocular lens magnification. For example, you might see 10x/16. In this case, 10x is the ocular lens magnification and 16 mm is the FN.

To calculate the field of view diameter use the following formula:

$$d = \frac{f}{o}$$

where

- d = field of view diameter
- f = field number
- o = objective magnification

Assuming the FN on our microscope is 16 mm, let's calculate the field of view diameter for the 4x objective.

$$d = \frac{16mm}{4} = 4mm$$

Repeat this process for the remainder of the objectives on your microscope. Your final result should look something like this:

Ocular-magnification	Objective-magnification	Total-magnification	Field-of-view-diameter
10x	4x	40x	4 mm
10x	10x	100x	1.6 mm
10x	40x	400x	0.4 mm
10x	100x	1000x	0.16 mm

Estimating Object Size

Next, estimate the size of the object you are looking at.

Since you know the diameter of the field of view for each lens, you can easily estimate the length or size of objects you are looking at by roughly comparing the object's size to the diameter of the circle. The formula for doing this is:

$$o = fr \times d$$

where

- o = object size
- fr = fraction of the field of view taken up by the object
- d = field of view diameter

For example, suppose you are looking at a protozoan that extends 3/4 of the way across the field under 400x total magnification. Using the table above for the microscope in this example, we can see that at 400x total magnification, the objective magnification is 40x and the diameter of the field of view is 0.4 mm. Plugging these numbers into the formula above, we can estimate the length of the object at 0.3 mm.

$$o = \frac{3}{4} \times 0.4mm = 0.3mm$$

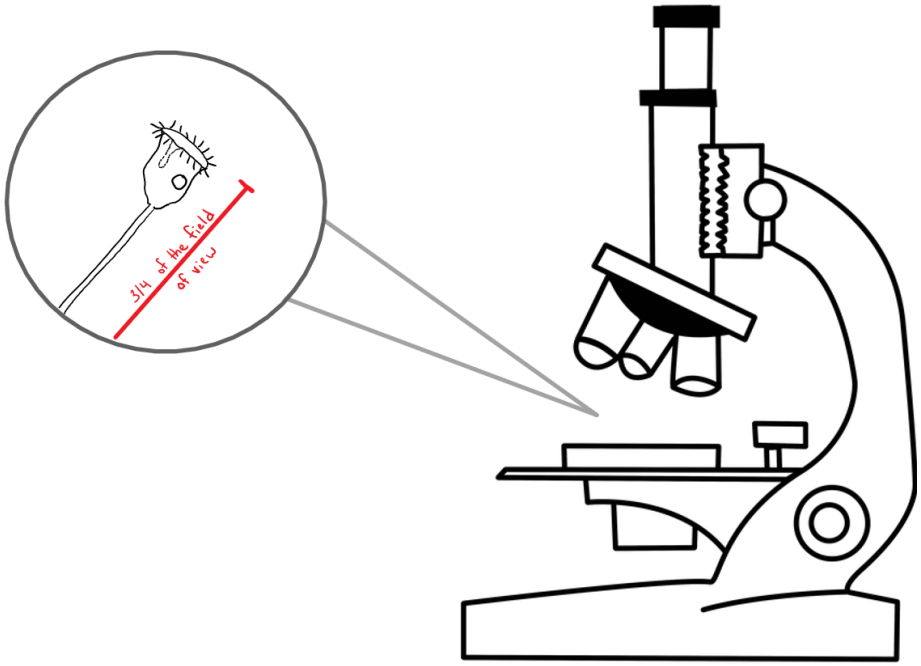


Figure 17.1: *Caption* Estimating object size.Image produced by Clerissa Copeland licensed under CC BY-NC-SA 4.0

17.3 Method 2: Ocular Micrometer

The ocular micrometer is inserted into one of the eyepieces of the microscope and looks like a ruler.

Figure 17.2: *Caption* The Ocular Micrometer.Image produced by Mathew Vis-Dunbar licensed under CC BY-NC-SA 4.0

The distance between the markings on the micrometer are used to measure objects in the field of view and represent different sizes of increments at different magnifications because the diameter of the field of view decreases as the magnification increases.

In order to determine the size represented by the increments at different magnifications, the ocular micrometer must be calibrated against a stage micrometer (a special slide containing a scale marked off in 10μm increments).

Calibrating the Ocular Micrometer

To calibrate the ocular micrometer, follow these steps:

1. Look through the microscope eyepieces and note the ocular micrometer.
2. Obtain a stage micrometer and bring the scale into focus using the 40x objective (remember to focus it first using the 4x and 10x objectives).
3. Rotate the ocular lens tube until the markings of the ocular micrometer are lined up parallel to those of the stage micrometer and superimposed on them.
4. Use the mechanical stage adjustment knob to move the stage micrometer so that the lines of both micrometers coincide at the left edge of the field.

5. Starting with the lowest magnification objective, carefully observe and count the number of increments on the ocular micrometer that correspond to a single 10μm division (the smallest division on the stage micrometer) of the stage micrometer. Repeat this for each objective lens.
6. Calculate the size of one ocular micrometer unit using the following formula:

$$o = \frac{10\mu m}{u}$$

where

- o = size of 1 ocular micrometer unit (in μm)

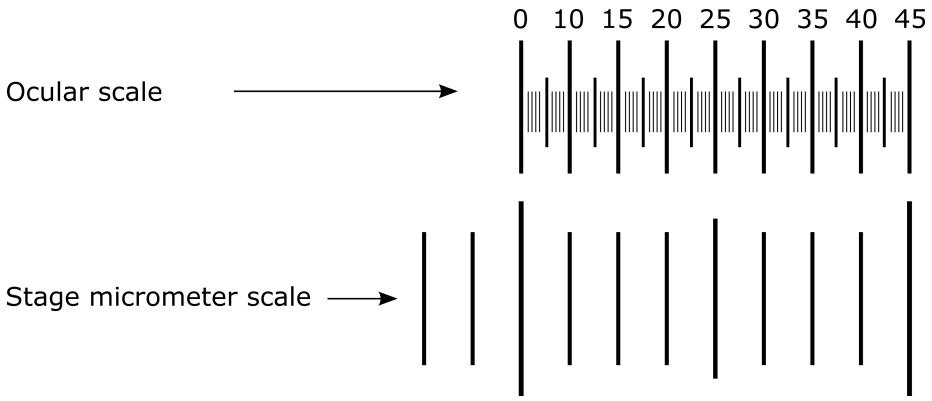


Figure 17.3: *Caption* Calibration of the ocular micrometer. Note that this diagram is only an example and will not necessarily reflect the alignment you will see under your own microscope.Image produced by Mathew Vis-Dunbar licensed under [CC BY-NC-SA 4.0]

- $10\mu m$ is the width of 1 stage micrometer unit
- u = # of ocular micrometer units in 1 stage micrometer unit

Example: Using a particular objective lens, you find that there are 5 ocular units in one stage micrometer unit. By using the formula above, you can then determine that each ocular micrometer unit for that objective is equal to 2 μm. See the calculations below:

$$o = \frac{10\mu m}{5 ocular micrometer units} = 2\mu m$$

7. Now you can calculate the size of a specimen by counting the number of ocular micrometer increments spanned by the specimen and multiplying it by the width of the ocular micrometer unit at that specific magnification. If your specimen is longer than it is wide, do this for both the width and the length of your specimen.

Make sure to measure more than one specimen and report the range of sizes for each dimension to account for the natural variation that occurs in the population. For example, a bacillus-shaped bacterium that ranges between 1 to 5 μm in width and 4 to 10 μm in length would be expressed as 1-5μm x 4-10 μm.

17.4 Drawing Size

Next, you must measure your drawing.

Now that you have estimated the size of the object, you must measure your drawing using a ruler. Let's suppose that in the above example that the protozoan drawing is 6 cm (or 60 mm) long.

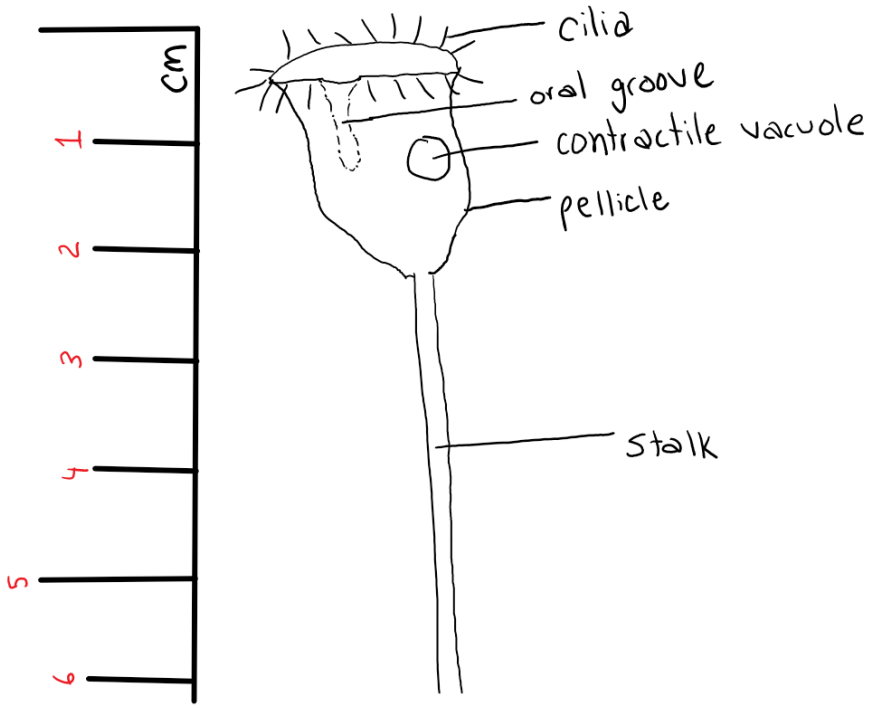


Figure 17.4: *Caption* Measuring drawing size.Image produced by Clerissa Copeland licensed under CC BY-NC-SA 4.0

17.5 Calculating Scale

Lastly, you can determine the drawing magnification using the following formula:

$$m = \frac{d}{o}$$

where

- m = magnification (or scale) of drawing
- d = size of drawing
- o = size of object

Now that you know both the size of the drawing and the size of the object, you can plug that information into the above formula.

In the previous example (Under Estimating object size using field of view diameter), the protozoan drawing is 60 mm and the size of the protozoan was estimated to be 0.3 mm. Plugging these values into the formula, we get a magnification of 200x.

$$m = \frac{60mm}{0.3mm} = 200$$

Alternatively, you can write this as a ratio with drawing size: real life size. In this case the scale would be 200:1.

Place the scale or magnification on the bottom right of your drawing.

Units must match and then they cancel out, leaving the ”x” to indicate magnification. You may also use the word ”times” instead of ”x”.

Here is the final image.

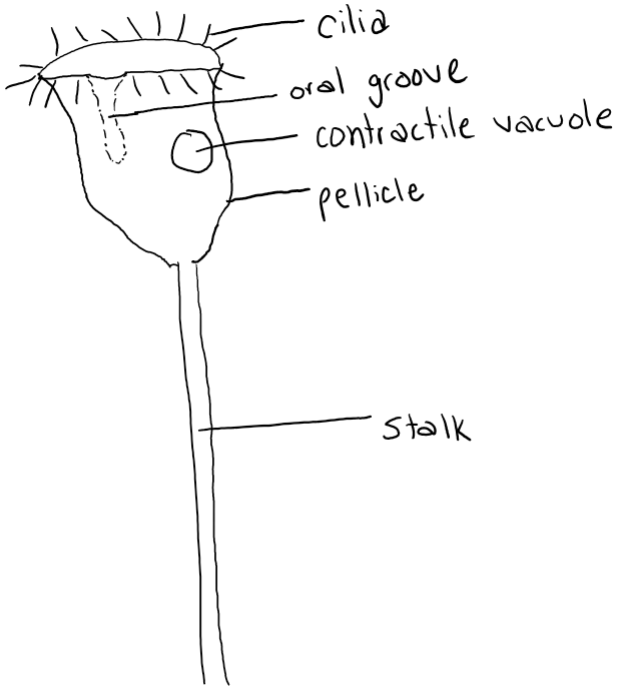


Figure 17.5: Whole (wet) mount of *Vorticella campanula* viewed at 400x total magnification using a brightfield compound microscope. Macronucleus not seen.Image produced by Clerissa Copeland licensed under CC BY-NC-SA 4.0