

# Package as Analysis

*Joshua H. Cook*

*2019-03-11*



# Contents

<b>Welcome</b>	<b>5</b>
About . . . . .	6
Resources . . . . .	6
License . . . . .	6
<b>1 Framework</b>	<b>7</b>
1.1 Introduction . . . . .	7
<b>2 Getting Started</b>	<b>9</b>
2.1 R Package Set Up . . . . .	9
2.2 Remote . . . . .	10
2.3 Extras . . . . .	12
<b>3 Workflow</b>	<b>15</b>
<b>4 Example: Allele-specific <i>KRAS</i> CNA</b>	<b>17</b>



# Welcome

This is a manual on how to use the standard R package framework for data analysis. Though potentially more work, especially at the start, the purpose of using the R package framework is to maintain a clear and reproducible analysis.

**[This book is currently in progress, though any feedback is welcome.]**

## Why

Why bother with maintaining a package framework while also doing data analysis? It is a great question, especially when one considers the complexity and fluidity of an analysis and the rigidity of the R package framework. The answer is that some rigidity is needed - but just the basics. That is what the R package framework provides. There is a place for everything, though sometimes getting it to work (ie. build and pass checks) requires a few extra steps.

During my own analyses, I found things were getting much too disorganized and decentralized. What would start out as exploratory would morph into a subdirectory graph more complex than the original parent analysis. Perhaps for more organized people, this is unnecessary, but for those of us who want order and aren't sure how to get it, this framework offers a great place to start.

The final (more abstract) reason for using the R package framework is to battle the current issue of reproducibility. Reproducibility is the cornerstone of science - if a finding is true, anyone should be able to replicate it. However, the scientific community has been dealing with an astounding amount of irreproducibility, most famously documented by the Open Science Collaboration. If the analysis is organized as an R package, though, an analysis can be re-run entirely by anyone else familiar with R. Thus, whether they are collaborators and competitors, anyone should be able to follow the analysis a scientist publishes.

## Advantages

There are many advantages to using this framework. Here are just a few, though I am sure you will find there are many others:

- Because this is a standard framework, others will be able to navigate the directories and files adeptly.
- The implementation of tests on functions and subroutines will make bugs easier to find and increase overall confidence in the validity of the analysis
- This is a seamless mixture of scripts and markdown files for the separation of functions and analysis
- Complete documentation of functions makes returning to code later much easier!
- The analysis can take advantage of normal R package tools such as Travis-CI and Codecov integration, pkgdown, and devtools (build checks, documentation, etc.).

## Examples

[coming soon] Allele-specific *KRAS* copy number alteration

## About

### About this Book

[TODO]

### About the Author

I am a classically-trained biologist-turned computational biologist. I graduated with degrees in Molecular Biology and Biochemistry, and Chemistry from the University of California, Irvine in 2017. My research focused on investigating the patterns and mechanisms of dissemination by which *Toxoplasma gondii*, an obligate, intracellular parasite, infects a human host Cook *et al.*, 2018. I started my graduate studies at Harvard Medical School in 2018, and after rotating in a chemical biology lab and a *Vibrio cholerae* lab, I finally decided to study cancer using computational biology. Since then, and continuing still today, I have been learning computer programming and statistics, trying to catch up to my peers. Consequently, I have fallen in love with R, especially because of the *Tydyverse* and tidy data.

## Resources

The best resource for making R packages is R Packages by Hadley Wickham.

There are some useful R packages you will want to have installed:

- `devtools` - will do most of the development building and checking
- `roxygen2` - makes all of the documentation
- `usethis` - for preparing all of the pieces and tools you want to include
- `testthat` - for running tests
- `kintr` - for compiling all of the Rmarkdown files

These can be installed using the following code.

```
install.packages(c("devtools", "roxygen2", "usethis", "testthat", "knitr"))
```

## License

This work is under a Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0)

# Chapter 1

## Framework

### 1.1 Introduction

I have decided to introduce the R package framework before going through the set-up process because, quite frankly, I think this is more important. There are many sites that outline the steps of creating an R package (they likely do it better than I will next chapter), but the main point of this book is how to make the R package into a data analysis project. Therefore, understanding the use of the framework should be the reader's focus, here.

I try to adhere to the following process for each piece of the framework:

1. what is it and what is its main role?
2. when does it need to be used or adjusted?
3. what is unique about its use for a data analysis project?

[TODO]

Include a subsection on each main piece of package:

- what is it for
- things to be wary of
- specific use for data analysis





## Chapter 2

# Getting Started

Setting up the package is mostly automated and is well documented in *R Packages* by Hadley Wickham. If you are running the analysis on your local machine, I would recommend using RStudio (which you likely already do), but this is possible to do on a remote computing cluster (which is how I work). I begin by going through the steps of setting up the basic package framework, which is the same for local or remote work. Following that, there is a section for how I work remotely. This can be skipped if you only work locally or if you already have a system you enjoy (though I highly recommend the system I currently use). I finish off with a few extras that I recommend using, but are not necessary.

### 2.1 R Package Set Up

The set up process is rather simple. If using RStudio, you can start a new R project as a package. Otherwise, the following command will get the basic framework started. There is a lot of overlap between the devtools and usethis package. I believe that RStudio is trying to fade out devtools and instead have people use the various packages that were split from it, including usethis.

The advantage to using the usethis functions for seemingly simple tasks (such as making the “data-raw” directory) is that it will also add the necessary lines to “.RBuildignore,” the “DESCRIPTION,” and “NAMESPACE” if needed.

It’s easy to create the package from the R console.

```
usethis::create("path/to/package/pkgname")
```

You can then add a license as shown below. I generally use a GPL-3, though you can get a lot of information on the common licenses at [choosealicense.com](http://choosealicense.com).

```
usethis::use_gpl3_license(name = "Your Name")
```

Prepare the project to use roxygen for documentation.

```
usethis::use_roxygen_md()
```

Create a README file. You can also opt to use a normal Markdown file with `usethis::use_readme_md()`, though I would recommend to just go with an RMarkdown file.

```
usethis::use_readme_rmd()
```

Create a “NEWS” file for announcing major changes to the project.

```
usethis::use_news_md()
```

Create a “data-raw” directory.

```
usethis::use_data_raw()
```

Finally, set up the use of testthat package for testing.

```
usethis::use_testthat()
```

**Note:** If you are working remotely (ie. sshing to the computer running the code), many of the usethis functions will open the file that you just asked them to create (eg. open “NEWS.md” after using `use_news_md()`) in vim. To suppress this, just pass the parameter `open = FALSE`. Otherwise, it is set to `interactive()`.

### 2.1.1 Git and GitHub

If you are programming, you should be using git. This is especially important in the sciences because git logs can be used to resolve legal conflicts and issues of data falsification. GitHub is not essential, though I would highly recommend you use it because it makes managing files and collaboration much easier. It is also essential for taking advantage of some of the best parts of an R package such as build checks and `pkgdown` (see Extras below)

To get started with git, there are *tons* of resources available, so I will not describe it here. If you are new to git and GitHub, here are a few good resources to get you started:

- An introduction to Git: what it is, and how to use it
- How To Use Git: A Reference Guide
- GitHub Guides: Hello World

There is a usethis function to initiate git (`usethis::use_git()`) though I always prefer to set up myself.

## 2.2 Remote

If you conduct work remotely, I’m going to assume that you have ssh set up and running. Otherwise, there are plenty of resources available, and you should review the material available by your system admin.

Though I prefer Rstudio for normal package development, I spare my computer the pain of performing complex and heavy computation, opting instead to off-load it to the Harvard Medical School Research Computing Cluster. Therefore, I use SublimeText3 as my text editor and send code to the remote computing node over ssh using iTerm2 as my terminal. Finally, I use SSH File System (SSHFS) to “mount” the remote directory to my local directory.

### 2.2.1 SublimeText3 Set-Up

Here are the handful of SublimeText3 (ST3) packages I use for R coding, followed by any particular notes on their use:

- LSP - “Gives Sublime Text 3 rich editing features for languages with Language Server Protocol support”
- MarkdownEditing - “Markdown plugin for Sublime Text. Provides... more robust syntax highlighting and useful Markdown editing features for Sublime Text.”
- R-IDE - “[A]iming to utilize the use of language server + better support R Markdown + better support of R packaging + ...”
- SendCode - “Send code and text to macOS and Linux Terminals, iTerm, ConEmu, Cmdr, Tmux, Terminus; R (RStudio), Julia, IPython.”

LSP and R-IDE handle syntax and completion in ST3. It isn't a great system, so if you know of a better set-up in ST3, please let me know. MarkdownEditing and R-IDE combine to make RMarkdown feasible. The SendCode package essentially copies, pastes, and runs my code written in ST3 to the terminal when I press `command + return`. This way, I can type in ST3 and run in the terminal without using the mouse.

Before moving on, I made this snippet to quickly add a code chunk.

```
<snippet>
  <content><![CDATA[
  ```{r ${1:chunk_name}}
  $0
  ```
  ]]></content>
  <!-- Optional: Set a tabTrigger to define how to trigger the snippet -->
  <tabTrigger>rchunk</tabTrigger>
  <!-- Optional: Set a scope to limit where the snippet will trigger -->
  <scope>text.html.markdown.multimarkdown, text.html.markdown</scope>
  <description>create a Rmd code chunk</description>
</snippet>
```

### 2.2.2 Using SSH File System

I use SSH File System (SSHFS) to “mount” my remote directory to my local directory. It is essentially SFTP and SSH combined (the details go right over my head) and it is fairly easy to get set-up. Here is a link to get everything going, and I have included the steps I used below.

To start I downloaded and installed FUSE for macOS. Then I downloaded and installed the latest build of SSHFS. Finally, I created an empty directory that will become the place that I mount the remote directory. Typically, I make the root of the package, though I could see instances where you would want the package in a subdirectory below it.

```
# on local
mkdir ~/Lab_Projects/pkgName
```

On the remote server, I make a directory with the same name

```
# on remote
/path/to/compute/directory/pkgName
```

Finally, to connect the two, I use the following command that is pretty much identical to initiating a normal ssh session.

```
# on local
sshfs username@remote.host.com:/path/to/compute/directory/pkgName ~/Lab_Projects/pkgName
```

Now, the computer will treat the mount just like a normal flash drive, and ST3 fully accepts it. The only change I made to ST3 was to map a key-binding to “Project/Refresh Folders”. This way, if new files are created remotely, a quick key-stroke and everything is visible in the ST3 sidebar.

### 2.2.3 Git and GitHub

If working remotely, I have found it much easier to handle everything git-related on the remote side. Therefore, I created ssh RSA keys and shared the public one with the GitHub repo so I could push over ssh. Setting this up is pretty simple and well outlined in Connecting to GitHub with SSH.

## 2.3 Extras

Though these next few items are not required, I *highly recommend* implementing them because they each take advantage of the fact that this project adheres to the standard R framework. Their different functions are all reasons to go through the trouble of mainting this framework.

### 2.3.1 pkgdown

Pkgdown ties a bow around your package, slaps it on the bottom, and builds a gorgeous and professional website rich with useful features. It builds the documentation for easy reference, presents the vignettes, and organizes all of the package meta-data so it is easily viewable and understandable. Here are some packages that take advantage of pkgdown:

- pkgdown (of course)
- ggplot2
- ggsci
- ggasym (a shameless plug of my own lil' package)

The use of pkgdown obviously begins with a usethis function.

```
usethis::use_pkgdown()
```

All that you have to do from there is use `pkgdown::build_site()` to build the site whenever the project is at a good stopping point for the day. (If working remotely, pass `preview = FALSE` to prevent pkgdown from searching for a browser to display in when completed.)

To show the website on GitHub, go to “Settings” in the repository, and select “master branch /docs folder” from the options in the “GitHub Pages” section. It should look something like this (another shameless plug for lil’ ole’ ggasym).

### 2.3.2 Travis-CI, Appveyor, and Codecov

GitHub integration also opens up the use of continuous integration (CI) apps. Travis-CI and Appveyor are useful for checking the build status of the package. I just use both because they each require so little effort to integrate and each provides their own suite of functions. Notably, Appveyor build the package on Linux and Windows. To get started, just use usethis.

Codecov provides an indication as to how well the package’s tests cover the code. Though not a perfect measure of test quality (nothing ever will be), I find this tool to be helpful for me to find which functions I have and have not created tests for.

```
usethis::use_travis()
usethis::use_coverage("codecov")
usethis::use_appveyor()
```

You then just follow the instructions printed out to get everything set up. If this your first time using any of the tools, then you will have to grant them access to your GitHub repositories, and they will do the rest.

The usethis command will also procude the markdown code for showing the status badges for each tools. Placing these below the package name in the README.Rmd is sstandard practice and will tell pkgdown to put them in the side bar of the site.

On top of looking good and being informative for you during the development process, these badges will also provide visitors an indication as to the quality and maintaince of the package. A few good badges will likely make visitors more trusting of your results.

### **2.3.3 Spelling**

TODO



## Chapter 3

# Workflow





## Chapter 4

### Example: Allele-specific *KRAS* CNA

[in progress]