The workflowr R package: a framework for reproducible and collaborative data science

John Blischak (@jdblischak)

2018-07-11

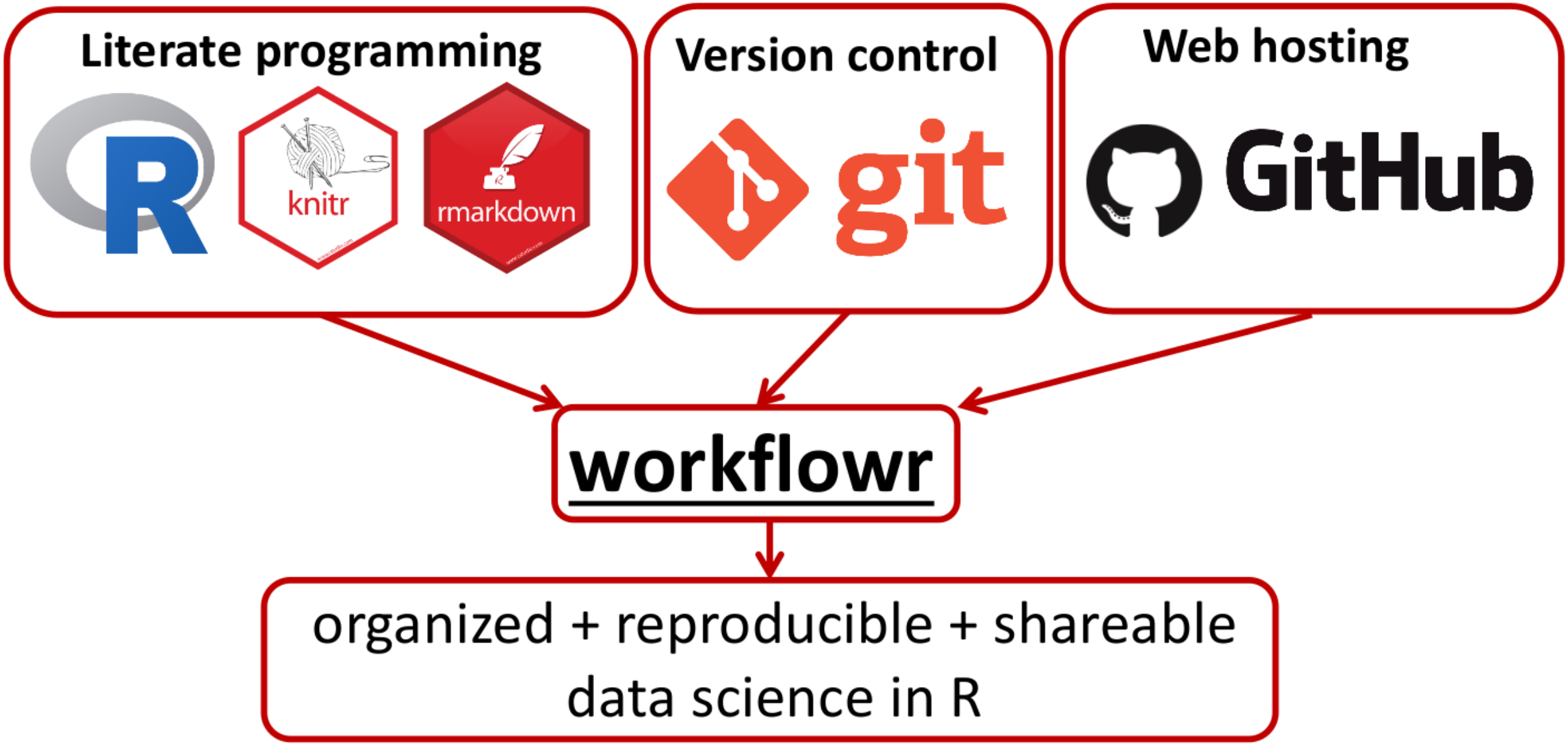
useR! 2018 Brisbane, Australia

My computational challenges

Organizing files

Tracking intermediate results

Sharing results


Literate programming
knitr
rmarkdown
Version control
git
Web hosting
GitHub
workflowr
organized + reproducible + shareable
data science in R

Literate programming

Source code

```
---
title: "Untitled"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

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

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

```{r cars}
summary(cars)
```

Including Plots

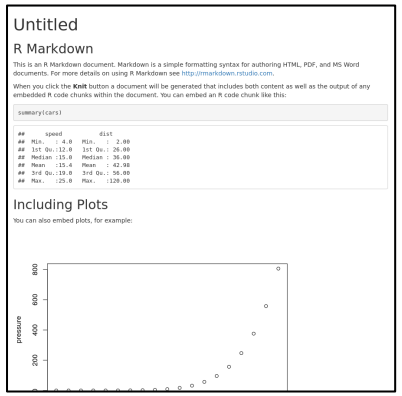
You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```

`file.Rmd`

Results



Untitled

R Markdown

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

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

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

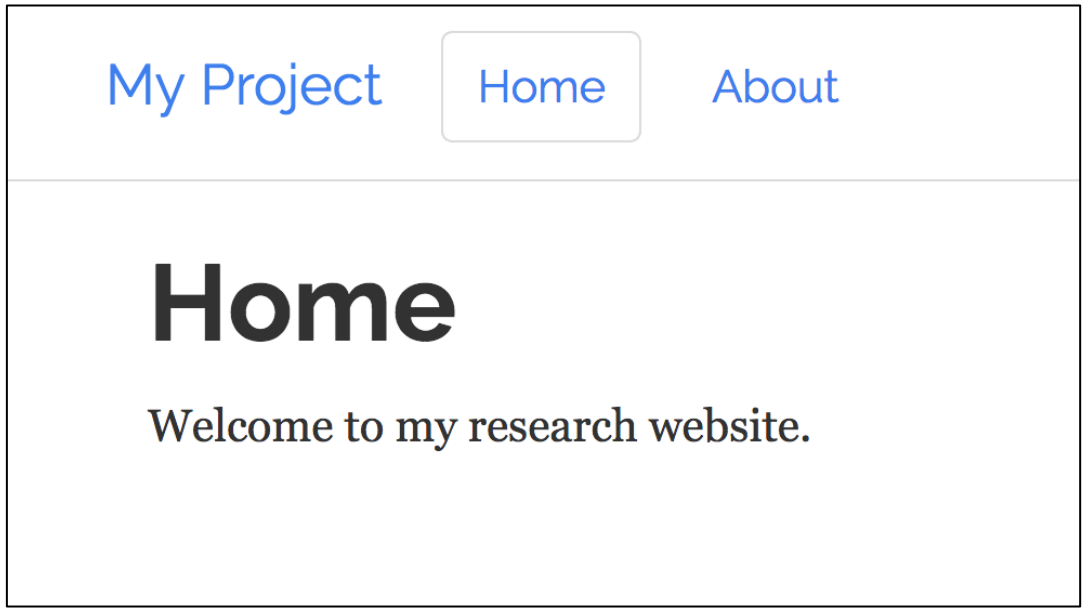
You can also embed plots, for example:

`file.html`

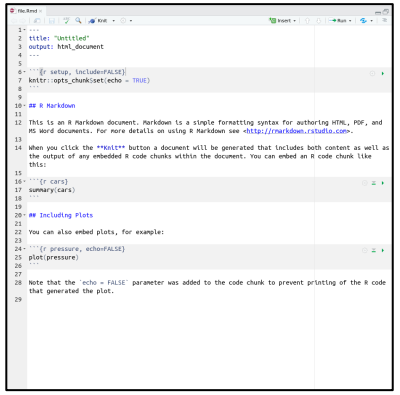
R Markdown websites

_site.yml | index.Rmd

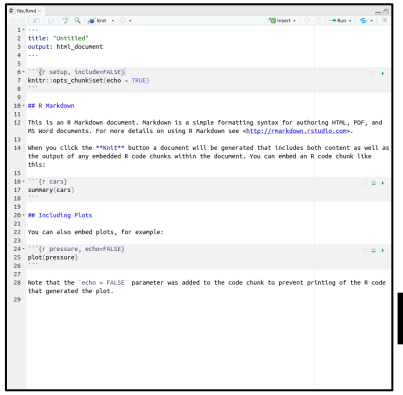
```
1  name: "My project"
2  output_dir: "results"
3  navbar:
4    title: "My Project"
5    left:
6      - text: "Home"
7        href: index.html
8      - text: "About"
9        href: about.html
10 output:
11   html_document:
12     theme: readable
13
```



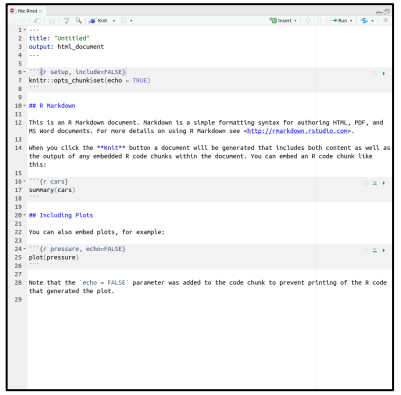
Version control



```
version: 2rko6xn
message: Start new…
```



```
version: d1zyskv
message: Update parameters…
```



```
version: z6o3b97
message: Label axes…
```

Version control terminology

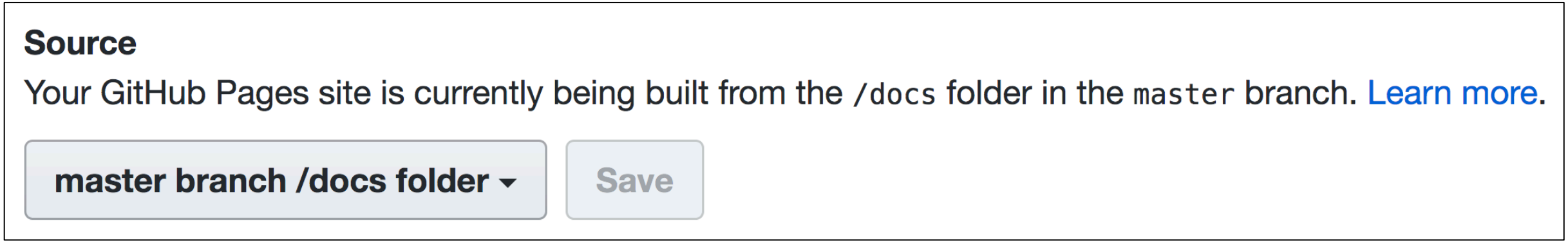
repository – the tracked files and their revision history

commit – a snapshot of the current state of the files

Web hosting

GitHub

GitHub Pages – hosts one website per code repository

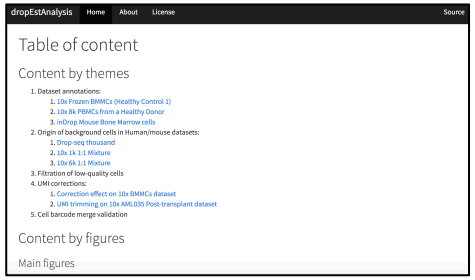


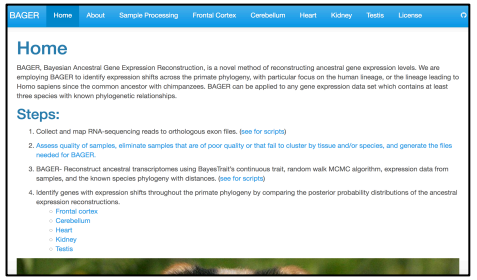
workflowr

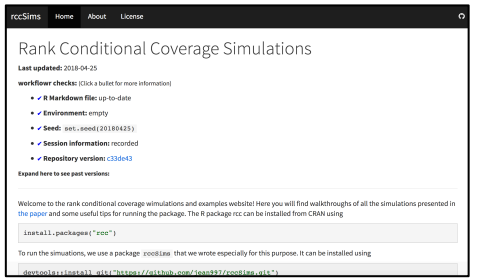
Organized

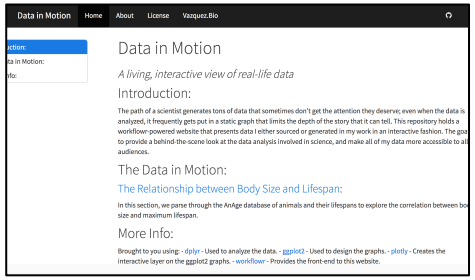
Reproducible

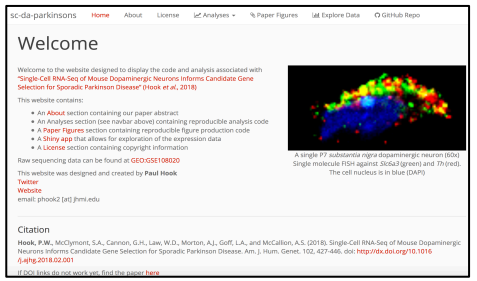
Shareable

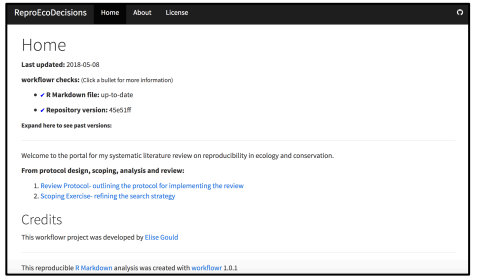


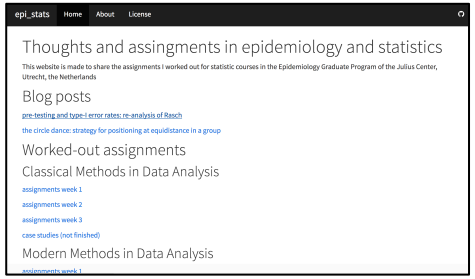


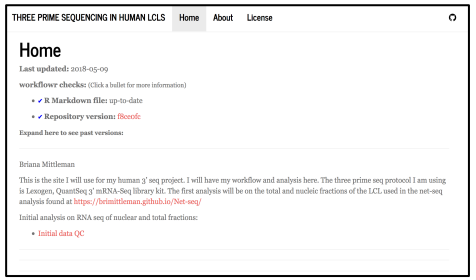


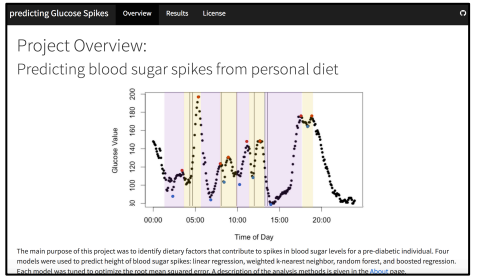












Version-controlled websites

Organized

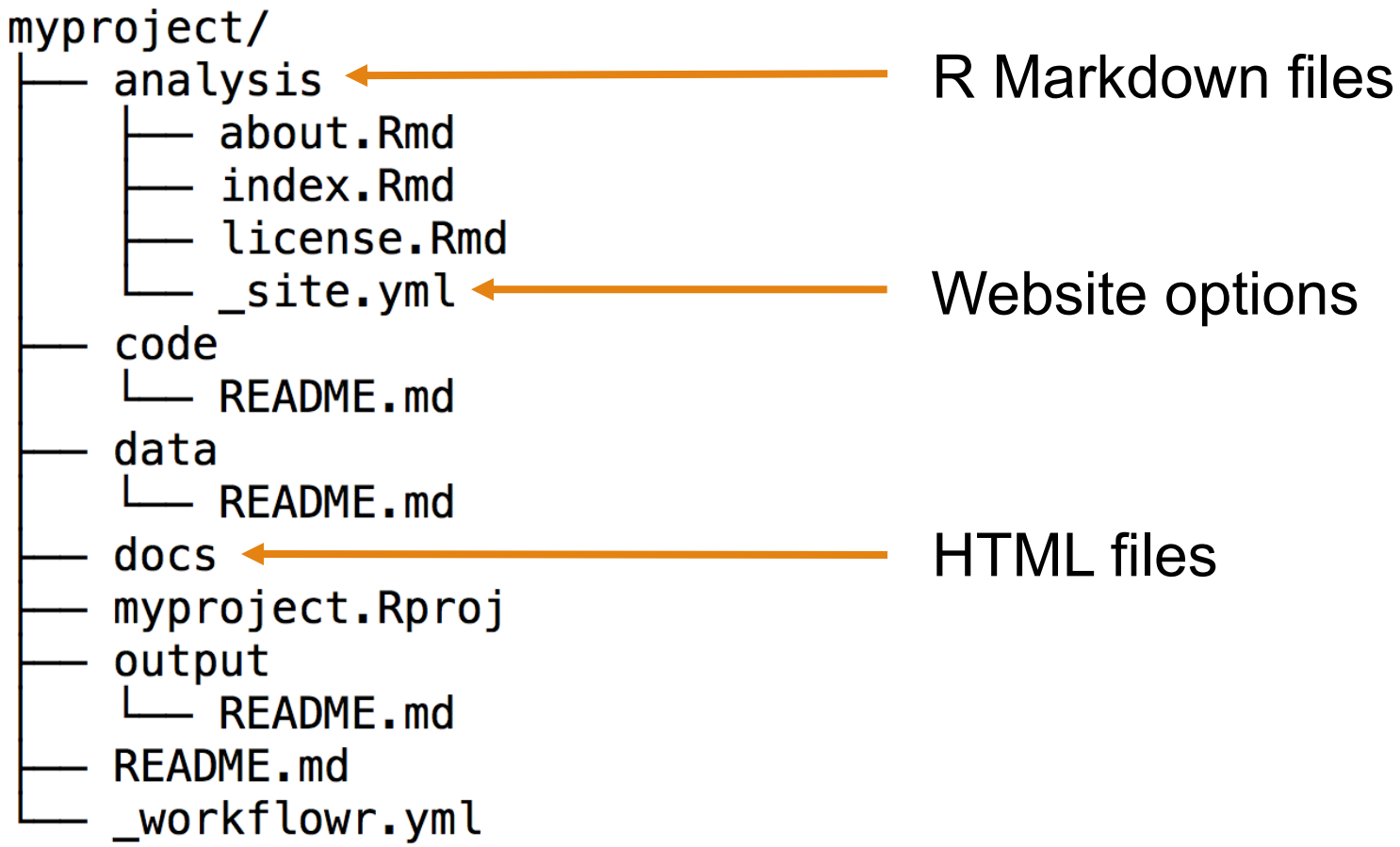
Start a new project

```
> wflow_start("myproject")
```

1. Creates directory with template files
2. Changes working directory
3. Initiates Git repository and commits files

Also available as RStudio Project Template

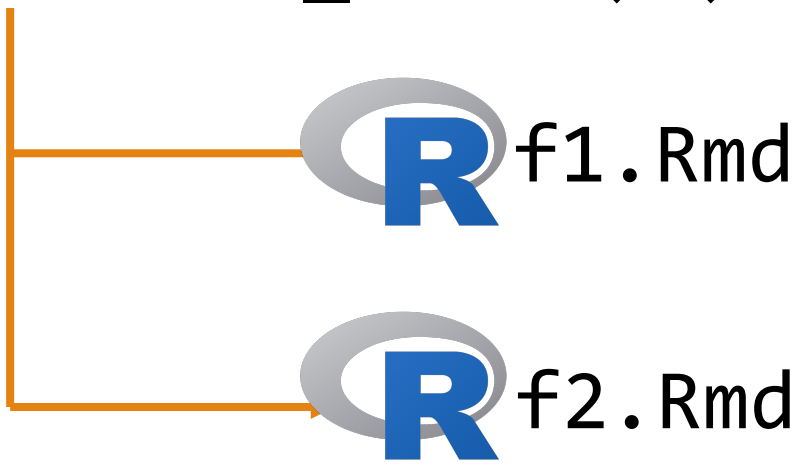
Organized directory structure



Reproducible

Run code in clean environment

```
> wflow_build(c("f1.Rmd", "f2.Rmd"))
```



Tracking intermediate results

```
> wflow_publish("analysis/file.Rmd")
```

Performs 3-steps:

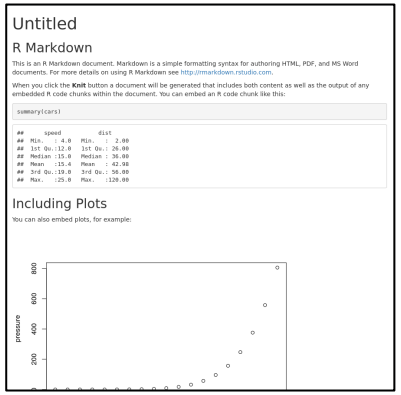
1. Commits `analysis/file.Rmd`
2. Builds `analysis/file.Rmd`
3. Commits `docs/file.html` and figure files

Combining rmarkdown and Git

Source code

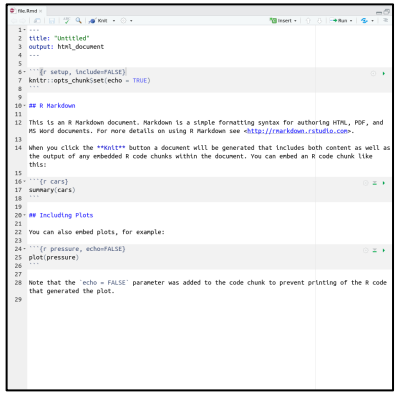
1ong9jt

Results



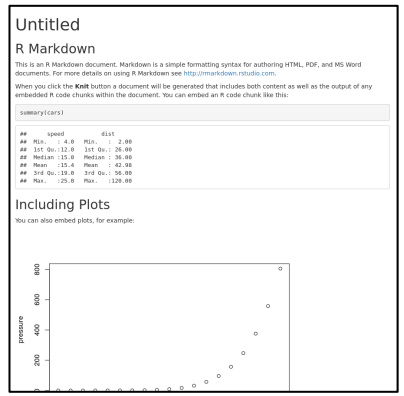
ln412fy

Source code



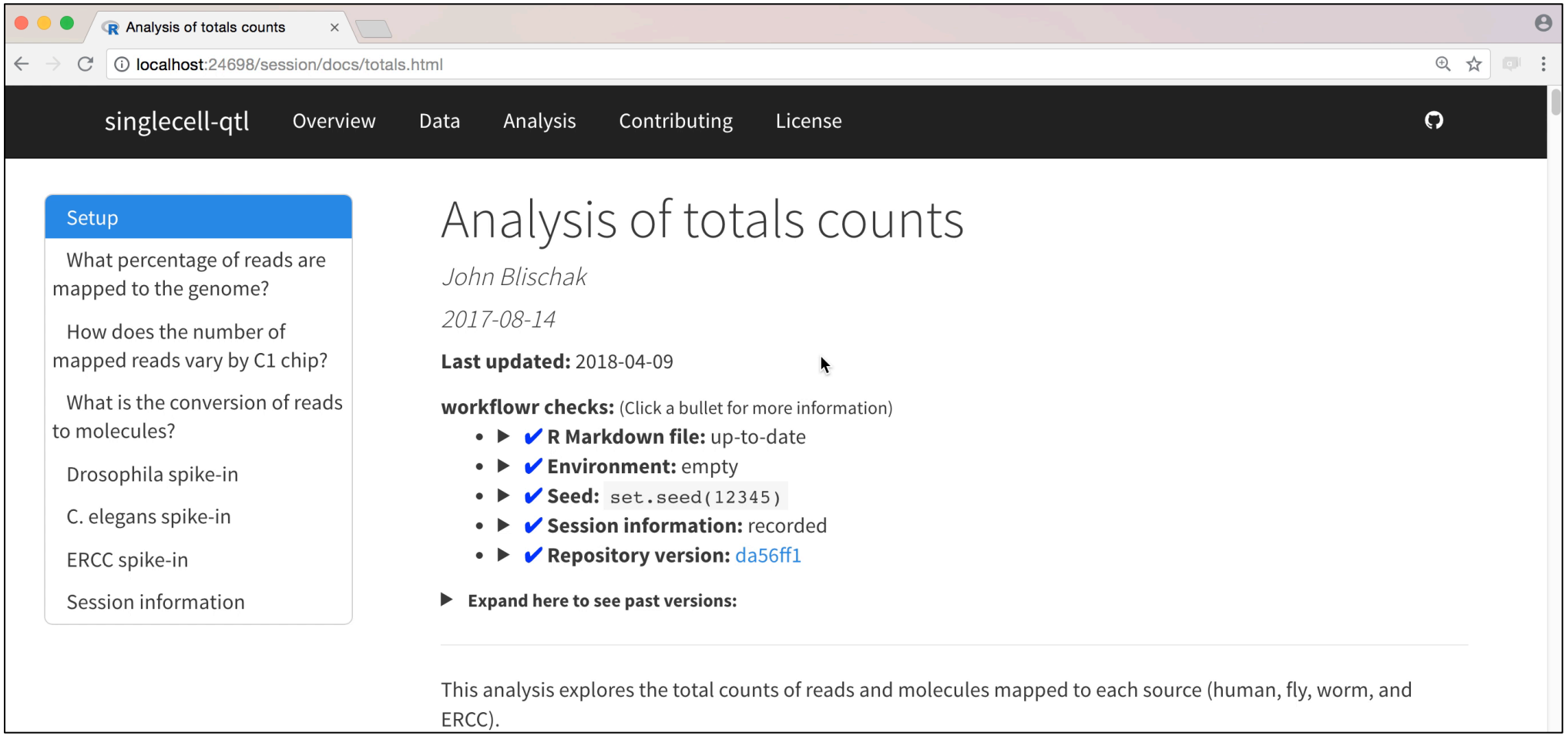
wr1q7bk

Results



3tg6lse

View past results



Other reproducibility features

```
output: workflowr::wflow_html
```

Records the session information at the end

Sets a seed prior to running code

Reproducibility report

Last updated: 2018-04-13

workflowr checks: (Click a bullet for more information)

- ▶ ✔ **R Markdown file:** up-to-date
- ▶ ✔ **Environment:** empty
- ▶ ✔ **Seed:** `set.seed(20180402)`
- ▶ ✔ **Session information:** recorded
- ▶ ✔ **Repository version:** 122e247

▶ **Expand here to see past versions:**

Last updated: 2018-04-13

workflowr checks: (Click a bullet for more information)

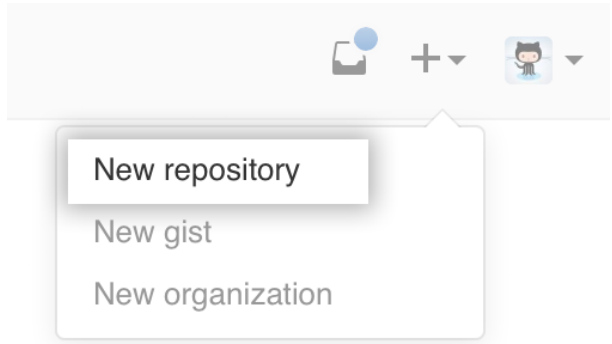
- ▶ ✖ **R Markdown file:** uncommitted changes
- ▶ ✖ **Environment:** objects present
- ▶ ✖ **Seed:** none
- ▶ ✖ **Session information:** unavailable
- ▶ ✔ **Repository version:** 17dda30

▶ **Expand here to see past versions:**

Shareable

Distribute results for sharing

Create new GitHub repository




```
> wflow_git_push()
```

Source

Your GitHub Pages site is currently being built from the `/docs` folder in the `master` branch. Learn more.

master branch /docs folder ▾ | Save

Installation

1. Install R
   - (Recommended) Install RStudio
   - (Optional) Install pandoc
   - (Optional) Install Git
2. Install workflowr from CRAN
   - `install.packages("workflowr")`
3. Create an account on GitHub

Documentation: https://jdblischak.github.io/workflowr

In summary, using workflowr…

Enables you to start working reproducibly immediately

Allows you to focus on your analysis

Shares your results online

Acknowledgements

Co-authors: Peter Carbonetto, Matthew Stephens

Early adopters for testing and feedback

Authors and contributors to knitr, rmarkdown, git2r, callr

THE UNIVERSITY OF CHICAGO

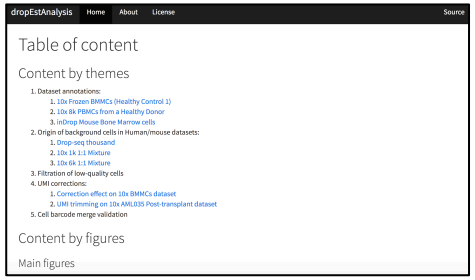
GORDON AND BETTY MOORE FOUNDATION

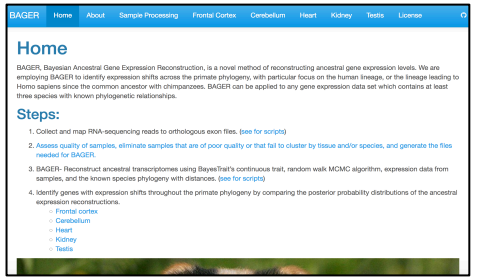
workflowr

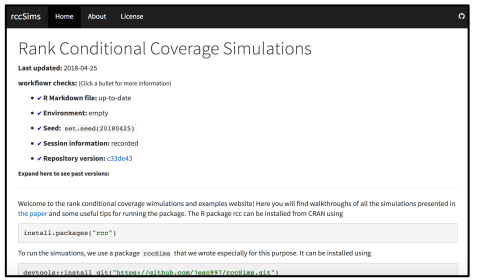
Organized

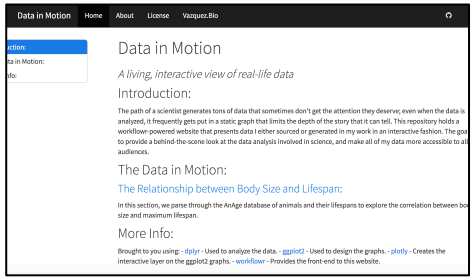
Reproducible

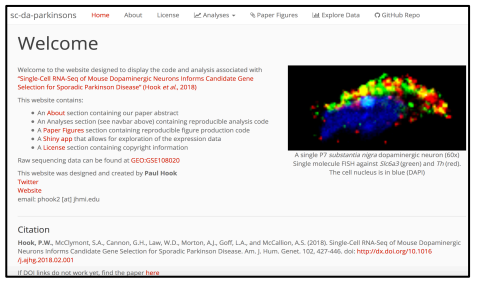
Shareable

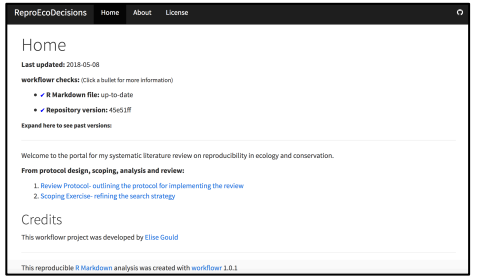


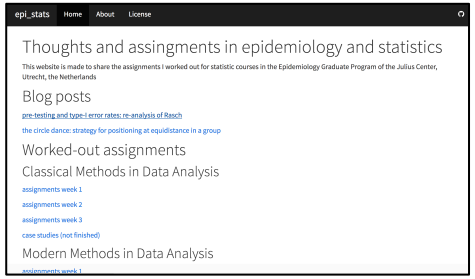


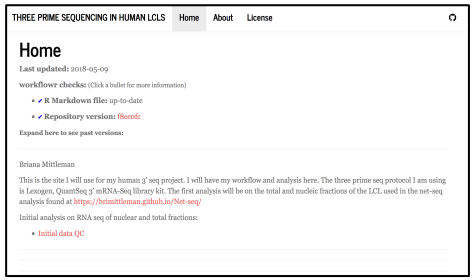


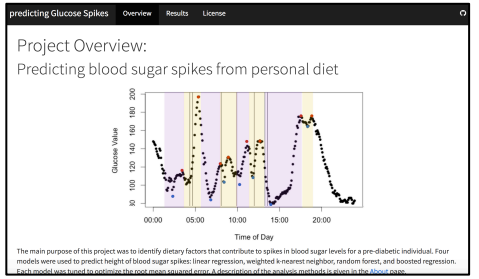












Version-controlled websites