

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



Version control



Web hosting



workflowr

organized + reproducible + shareable
data science in R

Literate programming



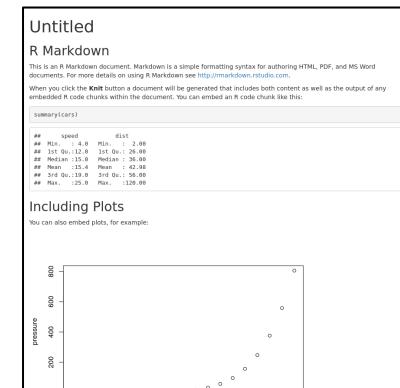
Source code

```
1<--># This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and
2<--># MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.
3<--># When you click the **Knit** button a document will be generated that includes both content as well as
4<--># any output of any embedded R code chunks within the document. You can embed an R code chunk like this:
5<--># 
6<-->#<code>summary(cars)</code>
7<-->#<code>plot(pressure)</code>
8<-->#<code># Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code
9<--># that generated the plot.
10<-->#<code>#</code>
```

file.Rmd



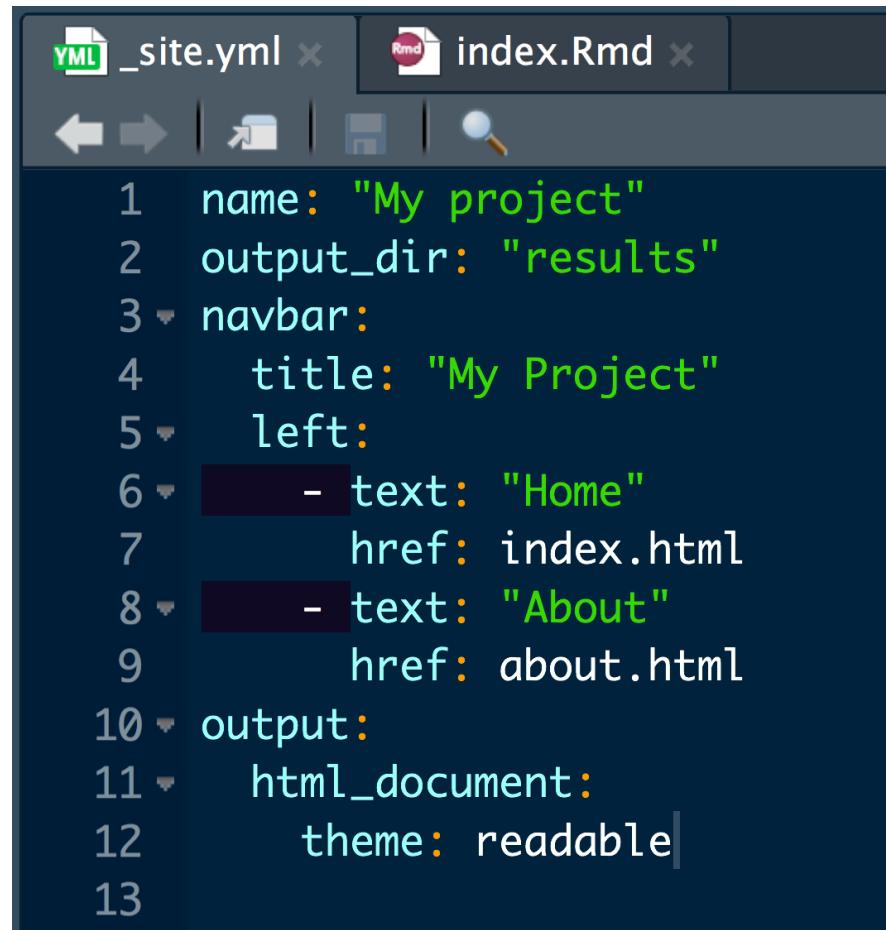
Results



file.html

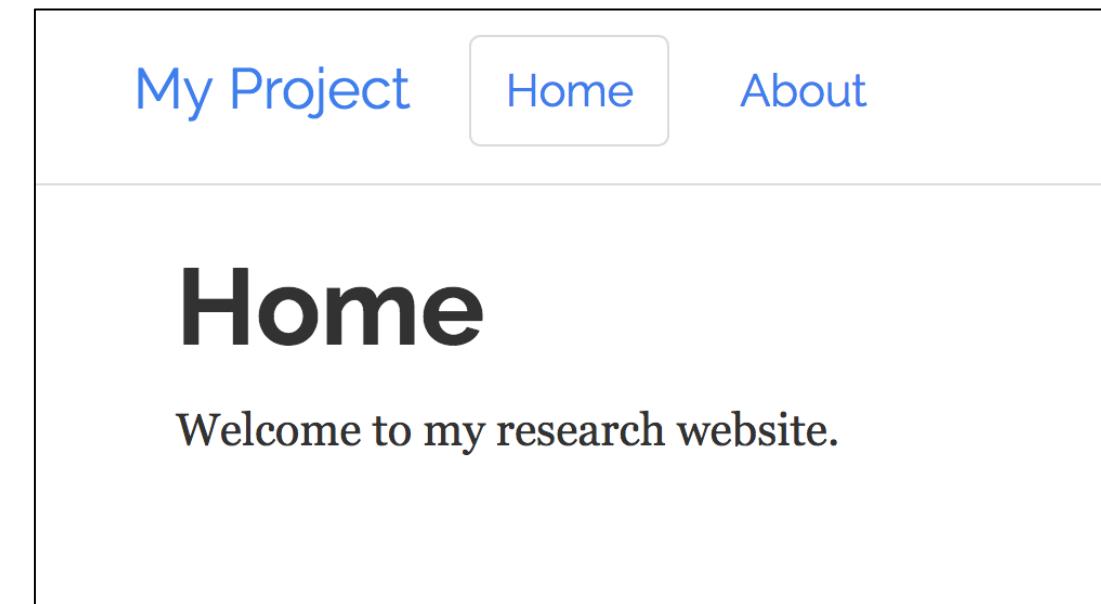


R Markdown websites



A screenshot of a code editor showing two files: `_site.yml` and `index.Rmd`. The `_site.yml` file contains configuration for a static website, including the project name, output directory, navigation bar with links to 'Home' and 'About', and the theme 'readable'. The `index.Rmd` file is partially visible at the top.

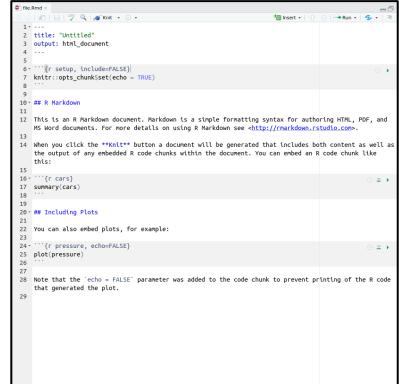
```
YML _site.yml x Rmd index.Rmd x
< > | ↻ | ⌂ | 🔎
1 name: "My project"
2 output_dir: "results"
3 navbar:
4   title: "My Project"
5 left:
6   - text: "Home"
7     href: index.html
8   - text: "About"
9     href: about.html
10 output:
11   html_document:
12     theme: readable
```



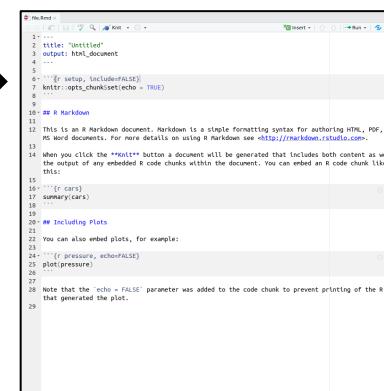
A screenshot of a generated R Markdown website. The header includes the project title 'My Project' and a navigation bar with 'Home' (which is currently active) and 'About'. The main content area features a large heading 'Home' and the text 'Welcome to my research website.'

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

Source

Your GitHub Pages site is currently being built from the /docs folder in the master branch. [Learn more.](#)

master branch /docs folder ▾

Save

workflowr

Organized

Reproducible

Shareable

Version-controlled websites

John Blischak - github.com/jdblischak/workflowr

The image displays a grid of nine screenshots from various workflowr-powered websites, illustrating the tool's application in different fields:

- dropEstAnalysis**: A website for Bayesian Ancestral Gene Expression Reconstruction (BAGER). It features a "Table of content" section with links to dataset annotations, figures, and other resources.
- BAGER**: The main BAGER website, which includes a "Home" page with a brief description of the method and a "Steps" section detailing the analysis pipeline.
- rccSims**: A website for Rank Conditional Coverage Simulations. It shows a "Rank Conditional Coverage Simulations" page with a "Workflow checks" section and a "Welcome" message.
- Data in Motion**: A website for "Data in Motion", described as a "living, interactive view of real-life data". It includes a "Data in Motion" section with a heatmap visualization and a "More Info" section with links to GitHub repositories.
- sc-daparkinsons**: A website for Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease. It features a "Welcome" page with a heatmap and a "Raw sequencing data" section.
- RepoDecisions**: A website for Reproductive Decisions, a portal for systematic literature review on reproducibility in ecology and conservation. It includes a "Home" page and a "Welcome" section.
- epi_stats**: A website for Epidemiology assignments in epidemiology and statistics. It contains sections for "Thoughts and assignments in epidemiology and statistics", "Blog posts", "Worked-out assignments", "Classical Methods in Data Analysis", and "Modern Methods in Data Analysis".
- THREE PRIME SEQUENCING IN HUMAN LCL**: A website for Three Prime Sequencing in Human LCL. It includes a "Home" page and a "Project Overview" section with a graph showing glucose spikes over time.
- predicting Glucose Spikes**: A website for predicting blood sugar spikes from personal diet. It features a "Project Overview" page with a graph showing glucose levels throughout the day.

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

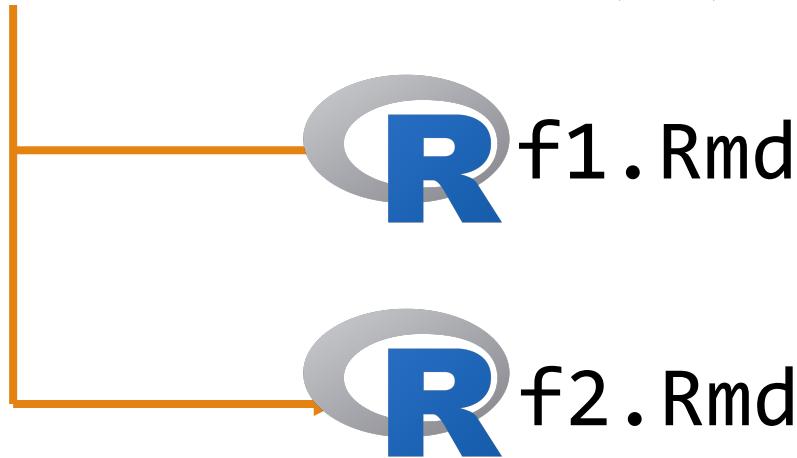
Organized directory structure

```
myproject/
  └── analysis ← R Markdown files
      ├── about.Rmd
      ├── index.Rmd
      ├── license.Rmd
      └── _site.yml ← Website options
  └── code
      └── README.md
  └── data
      └── README.md
  └── docs ← HTML files
  └── myproject.Rproj
  └── output
      └── README.md
  └── README.md
  └── _workflowr.yml
```

Reproducible

Run code in clean environment

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



Tracking intermediate results

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

Performs 3-steps:

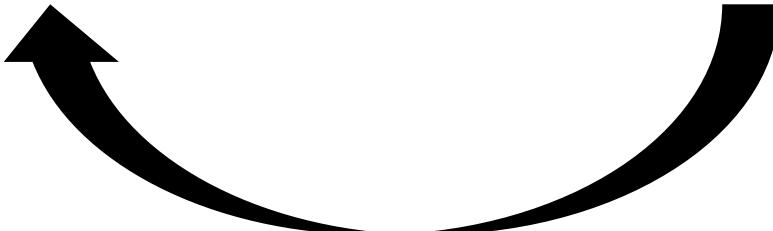
1. Commits `analysis/file.Rmd`
2. Builds `analysis/file.Rmd`
3. Commits `docs/file.html` and figure files

Combining rmarkdown and Git

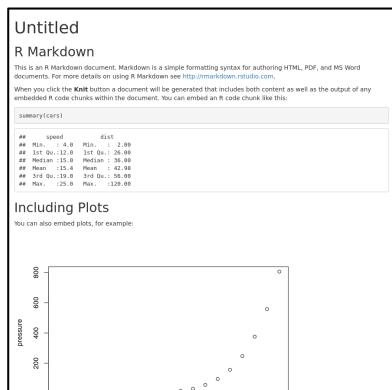
Source code

```
knitr.Rmd
1 title: "Untitled"
2 output: html_document
3 ...
4 ...
5 ...
6 # [r echo = TRUE, include=FALSE]
7 knitr::opts_chunk$set(echo = TRUE)
8 ...
9 ...
10 ## R Markdown
11 ...
12 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.
13 ...
14 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:
15 ...
16 ...
17 (r cars)
18 summary(cars)
19 ...
20 ...
21 ## Including Plots
22 ...
23 You can also embed plots, for example:
24 ...
25 (r pressure, echo=FALSE)
26 plot(pressure)
27 ...
28 Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code
29 that generated the plot.
```

long9jt



Results



ln412fy

Source code

```
File: R_Handout.Rmd
1: #> R_Handout
2: title: "Untitled"
3: output: html_document
4:
5:
6: ```{r setup, include=FALSE}
7: knitr::opts_chunk$set(echo = TRUE)
8: ...
9:
10: ## # R Markdown
11:
12: 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.
13:
14: 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:
15:
16: ```{r cars}
17: summary(cars)
18:
19:
20: ## Including Plots
21:
22: You can also embed plots, for example:
23:
24: ```{r pressure, echo=FALSE}
25: plot(pressure)
26: ...
27:
28: Note that the "echo = FALSE" parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```

wr1q7bk



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  Max. : 25.0
## #> 1st Qu.: 10.0  Median : 16.0
## #> 3rd Qu.: 22.0  Max. : 29.0
## #>
## #> Mean : 15.4  Std. Dev.: 12.0
## #> Coef. of variation: 0.791
## #> Min. :  5.0  1st Qu.: 10.0
## #> Median : 15.0  3rd Qu.: 26.0
## #> Max. : 36.0  128.0
```

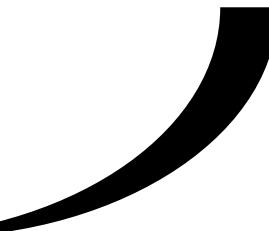
Including Plots

You can also embed plots, for example:

A scatter plot showing the relationship between speed and distance for cars. The x-axis is labeled "speed" and ranges from 4 to 36. The y-axis is labeled "dist" and ranges from 0 to 120. The plot shows a positive correlation, with points scattered generally upwards and to the right. There are several outliers at higher speeds (around 25-30) with very high distances (around 100-120).

speed	dist
4.0	2.0
10.0	4.0
15.0	6.0
22.0	8.0
25.0	12.0
29.0	14.0
36.0	18.0
5.0	4.5
10.0	7.0
15.0	9.0
20.0	11.0
25.0	15.0
28.0	19.0
30.0	22.0
32.0	25.0
35.0	29.0

3tg6lse



View past results

The screenshot shows a web browser window with the title "Analysis of totals counts". The URL in the address bar is "localhost:24698/session/docs/totals.html". The page has a dark header with navigation links: "singlecell-qt1", "Overview", "Data", "Analysis", "Contributing", and "License". A search icon is also present in the header.

The main content area features a sidebar on the left with a blue header titled "Setup". The sidebar contains the following items:

- What percentage of reads are mapped to the genome?
- How does the number of mapped reads vary by C1 chip?
- What is the conversion of reads to molecules?
- Drosophila spike-in
- C. elegans spike-in
- ERCC spike-in
- Session information

The main content area displays the analysis details:

Analysis of totals counts

John Blischak
2017-08-14

Last updated: 2018-04-09

workflowr checks: (Click a bullet for more information)

- **R Markdown file:** up-to-date
- **Environment:** empty
- **Seed:** set.seed(12345)
- **Session information:** recorded
- **Repository version:** da56ff1

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

This analysis explores the total counts of reads and molecules mapped to each source (human, fly, worm, and ERCC).

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

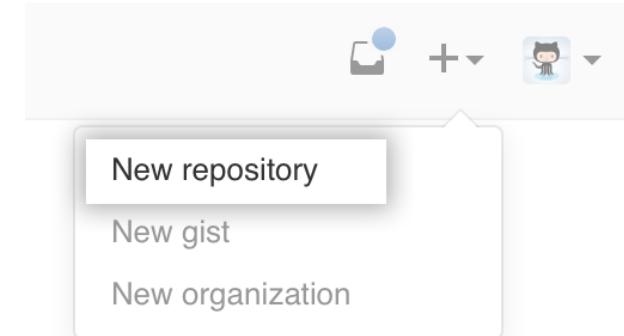
workflowr checks: (Click a bullet for more information)

- ► **✗ 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



© 2018 GitHub Inc.

> `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

John Blischak - github.com/jdblischak/workflowr

The image displays a grid of nine screenshots from various workflowr-powered websites, illustrating the tool's application in different fields:

- dropEstAnalysis**: A website for Bayesian Ancestral Gene Expression Reconstruction (BAGER). It includes a "Table of content" section with links to dataset annotations, figures, and other resources.
- BAGER**: The main BAGER website, featuring a "Home" page with a brief description of the method and a "Steps" section detailing the analysis pipeline.
- rccSims**: A website for Rank Conditional Coverage Simulations, showing a "Rank Conditional Coverage Simulations" page with simulation parameters and results.
- Data in Motion**: A website for "Data in Motion", which provides an "Interactive view of real-life data". It includes sections on "Introduction", "The Data in Motion", and "More Info".
- sc-daparkinsons**: A website for Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease. It features a "Welcome" page with a heatmap visualization and a "Paper Figures" section.
- RepoDecisions**: A website for Reproductive Decisions, designed for systematic literature reviews. It includes a "Home" page and a "Welcome" section.
- epi_stats**: A website for Epidemiology assignments in epidemiology and statistics. It contains sections for "Thoughts and assignments", "Blog posts", "Worked-out assignments", and "Modern Methods in Data Analysis".
- THREE PRIME SEQUENCING IN HUMAN LCL**: A website for three-prime sequencing in human lymphoblastoid cell lines (LCL). It includes a "Home" page and a "Project Overview" section with a graph of glucose spikes over time.
- predicting Glucose Spikes**: A website for predicting blood sugar spikes from personal diet data. It features a "Project Overview" page with a detailed graph showing glucose levels throughout a day.