A Practical Guide for Reproducible Papers

Aurora Blucher, PhD
Postdoc, Mills Lab, Knight Cancer Institute

Ted Laderas, PhD
Assistant Professor, DMICE

Head and Neck Project Repository
https://github.com/biodev/HNSCC_Notebook

Reproducible Paper Repository
https://github.com/ablucher/Workshop_ReproduciblePaper

Workshop Overview

- **Creating a Strategy/ Project Management Good Practices**
- **Literate Programming with R Markdown Notebooks**
- **Research Compendia with Binder / Hands-On Binder Demo**
- **Github Project Management Good Practices**
- **Bonus Round: Sub-analyses and annotation files**

Glossary

- **Software Environment**: what your code needs to run, such as operating system, programs, databases, etc.
- **Research Compendium:** data, code, and documentation, often goes along with a scientific publication
- **Literate Programming**: combining code and human-readable explanations of your code
- **Repository:** a folder for your project
- **Docker**: a program that lets you manipulate multiple operating systems on your computer

Preparing a Manuscript for PLOS Call for Papers

Our perspective for today's workshop

-ongoing project of a research group
-analysis of TCGA head and neck cancer pathways
-existing code base
-several sub-analyses
-draft manuscript

PLOS ONE

RESEARCH ARTICLE

Illuminating biological pathways for drug targeting in head and neck squamous cell carcinoma

Gabrielle Choonoo[1,2☯¤a], Aurora S. Blucher[1,3☯*], Samuel Higgins[2¤b], Mitzi Boardman[2], Sophia Jeng[1,4], Christina Zheng[1,2], James Jacobs[1,2,5], Ashley Anderson[6], Steven Chamberlin[2], Nathaniel Evans[2], Myles Vigoda[3,6], Benjamin Cordier[2], Jeffrey W. Tyner[1,3,7], Molly Kulesz-Martin[3,6], Shannon K. McWeeney[1,2,4], Ted Laderas[1,2]

1 Knight Cancer Institute, Oregon Health & Science University, Portland, Oregon, United States of America, **2** Division of Bioinformatics and Computational Biology, Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, United States of America, **3** Department of Cell, Developmental & Cancer Biology, Oregon Health & Science University, Portland, Oregon, United States of America, **4** Oregon Clinical and Translational Research Institute, Oregon Health & Science University, Portland, Oregon, United States of America, **5** Pediatric Hematology and Oncology, OHSU Doernbecher Children's Hospital, Portland, Oregon, United States of America, **6** Department of Dermatology, Oregon Health & Science University, Portland, Oregon, United States of America, **7** Division of Hematology and Medical Oncology, Oregon Health & Science University, Portland, Oregon, United States of America

☯ These authors contributed equally to this work.
¤a Current address: Regeneron Pharmaceuticals, Tarrytown, New York, United States of America
¤b Current address: Roche Sequencing Solutions, Santa Clara, California, United States of America
* blucher@ohsu.edu

Check for updates

OPEN ACCESS

Citation: Choonoo G, Blucher AS, Higgins S, Boardman M, Jeng S, Zheng C, et al. (2019) Illuminating biological pathways for drug targeting in head and neck squamous cell carcinoma. PLoS ONE 14(10): e0223639. https://doi.org/10.1371/journal.pone.0223639

Abstract

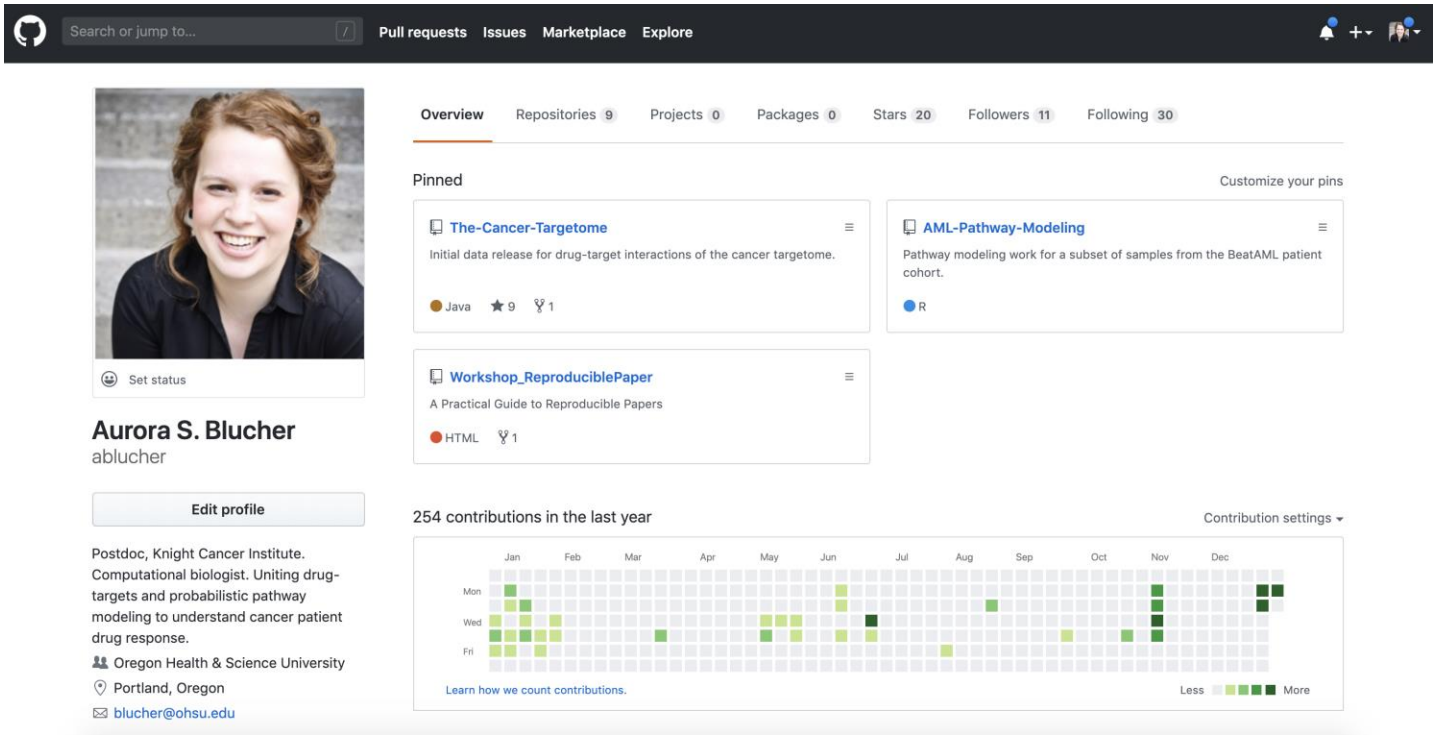
Strategy

- Where does your project live?
- Creating a roadmap for your work
- Identifying your inputs/ analysis steps/ output
    - Separate out any sub-analyses
- Re-creating your Results (Figures and Tables)
    - (Don’t forget your) Supplemental Figures/Tables
- Code Reproducibility

Where Does Your Project Live?

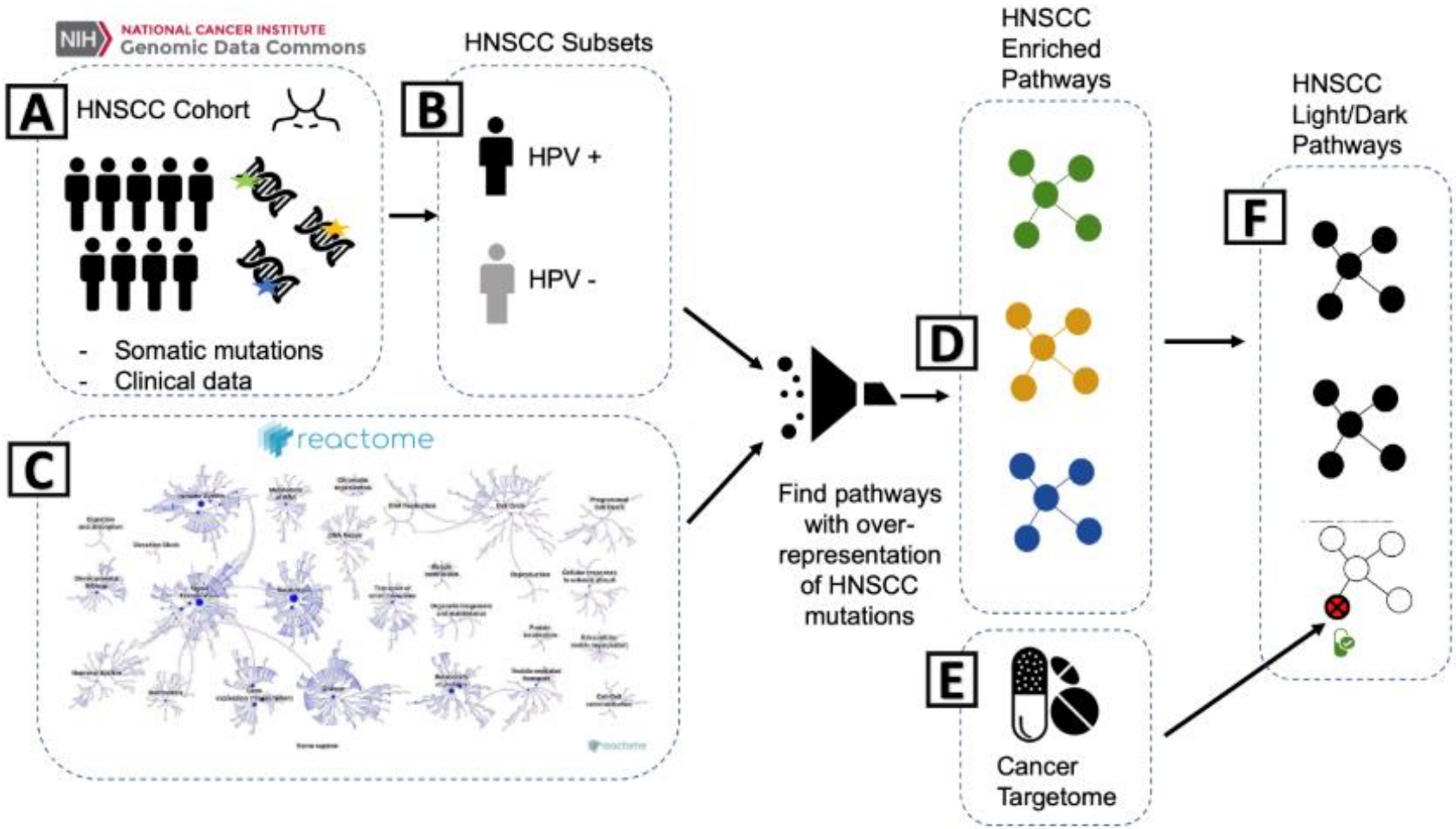
Give Projects a Home with GitHub Repositories

- Great for project management!
- Open (private/public options)
- Not necessarily tied to an institution/group
- Add collaborators with more privileges
- Part of your research portfolio



Creating a roadmap for your work

Your Overview Figure



…and prepare to delegate

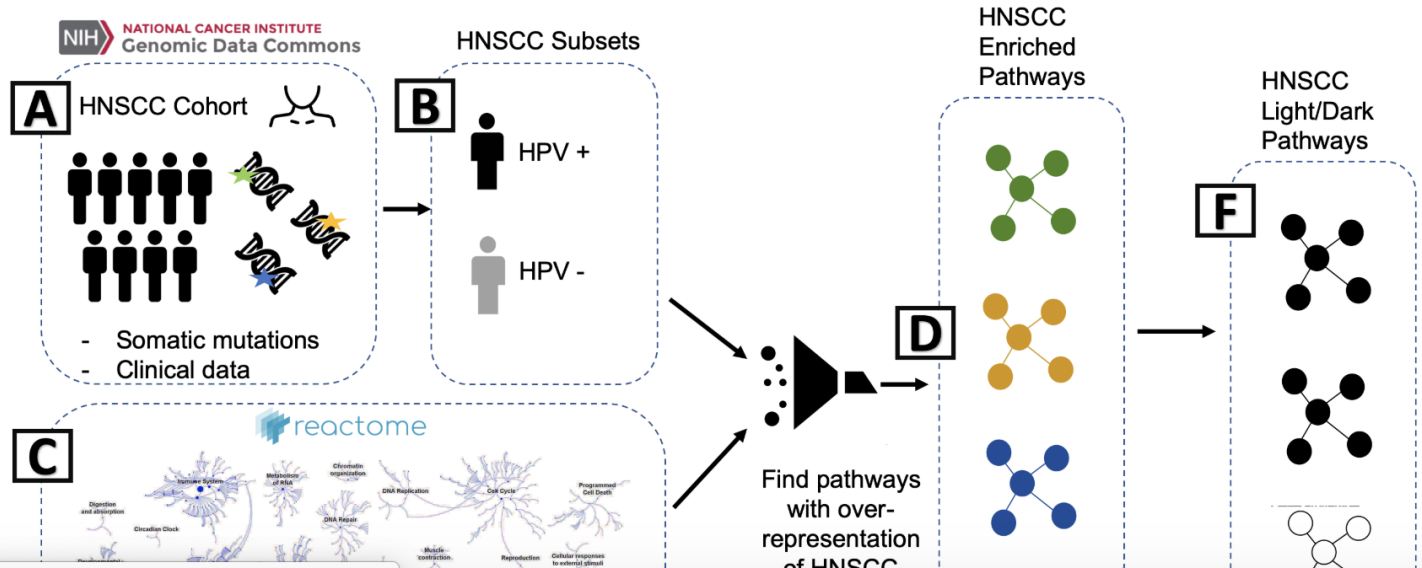
Use your GitHub README.md as a Project Overview

README.md

HNSCC_Notebook

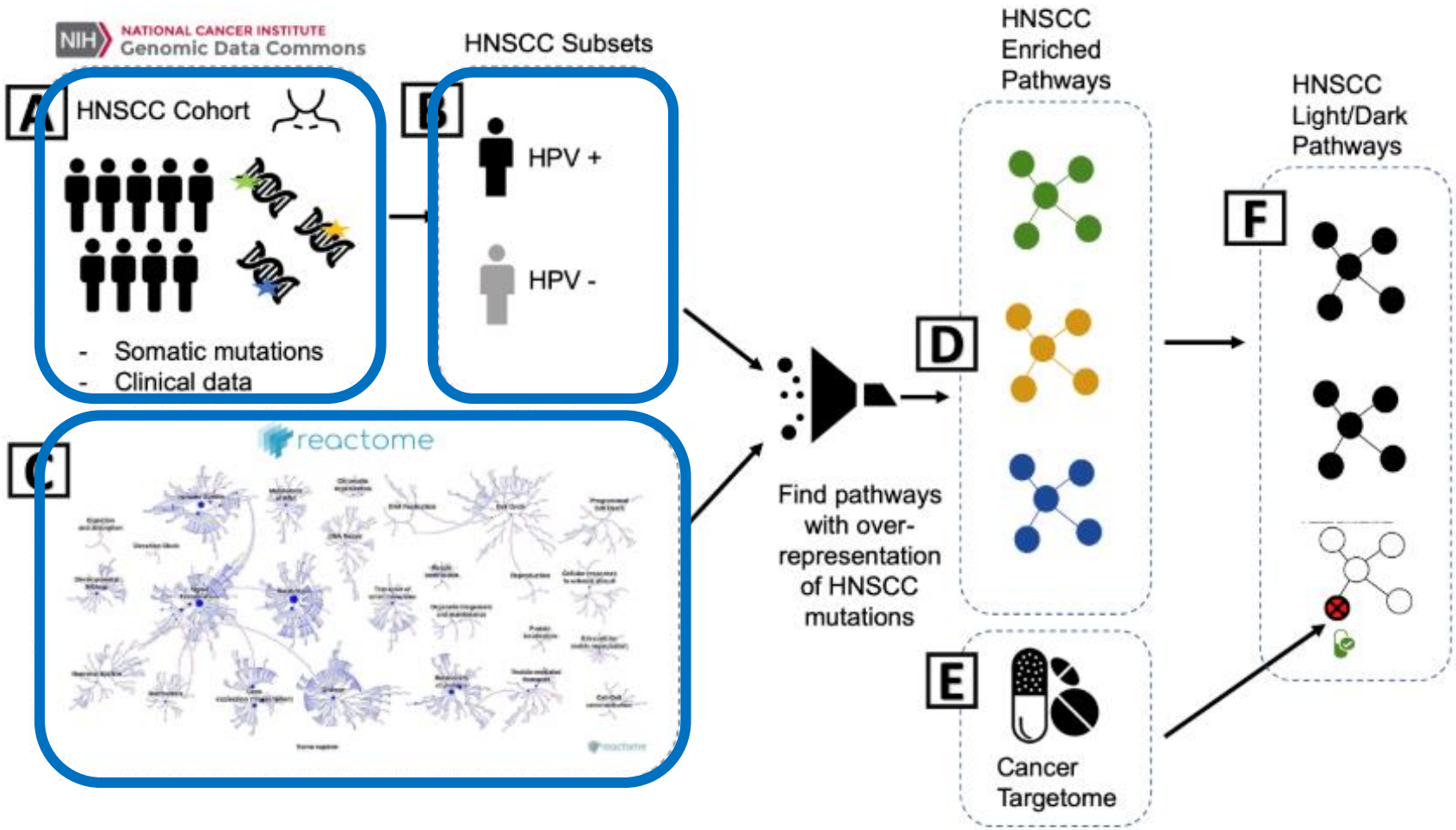
launch binder

This repository contains the workflow for light and dark pathway analysis of head and neck squamous cell carcinoma (HNSCC). For the TCGA cohort, we first assessed biological pathways that are significantly enriched for gene mutations in HNSCC patient tumors and then evaluated whether these pathways contained drug targets of FDA-approved cancer drugs. Enriched pathways containing drug targets are "light" to drugs and therefore of interest for targeting with the current set of approved drugs. Enriched pathways containing no drug targets are "dark" to drugs and of interest for future therapeutics development.



Identifying your inputs/ analysis steps/ output

Identify key inputs-data files, pathway databases, annotation files



Identify key inputs-data files, pathway databases, annotation file

"Good Enough Practices in Scientific Computing"
Greg Wilson & Jennifer Bryan. 2017.

Data set

- TCGA Head and Neck Squamous Cell Carcinoma Cohort
  - Mutation Data
  - Copy Number Data
  - Cohort/clinical annotation
- Best: include the open source, non PHI data files with your project
- Next best: link to the public repository where data can be downloaded

Resources

- Reactome pathway database
  - File of pathway IDs, names, and gene members
- HPV status annotation file
  - Additional cohort annotation file
- Cancer Targetome drug-target interactions file
- Include versions/access dates, and any modifications or clean-up you've done

→ **GitHub Repository**

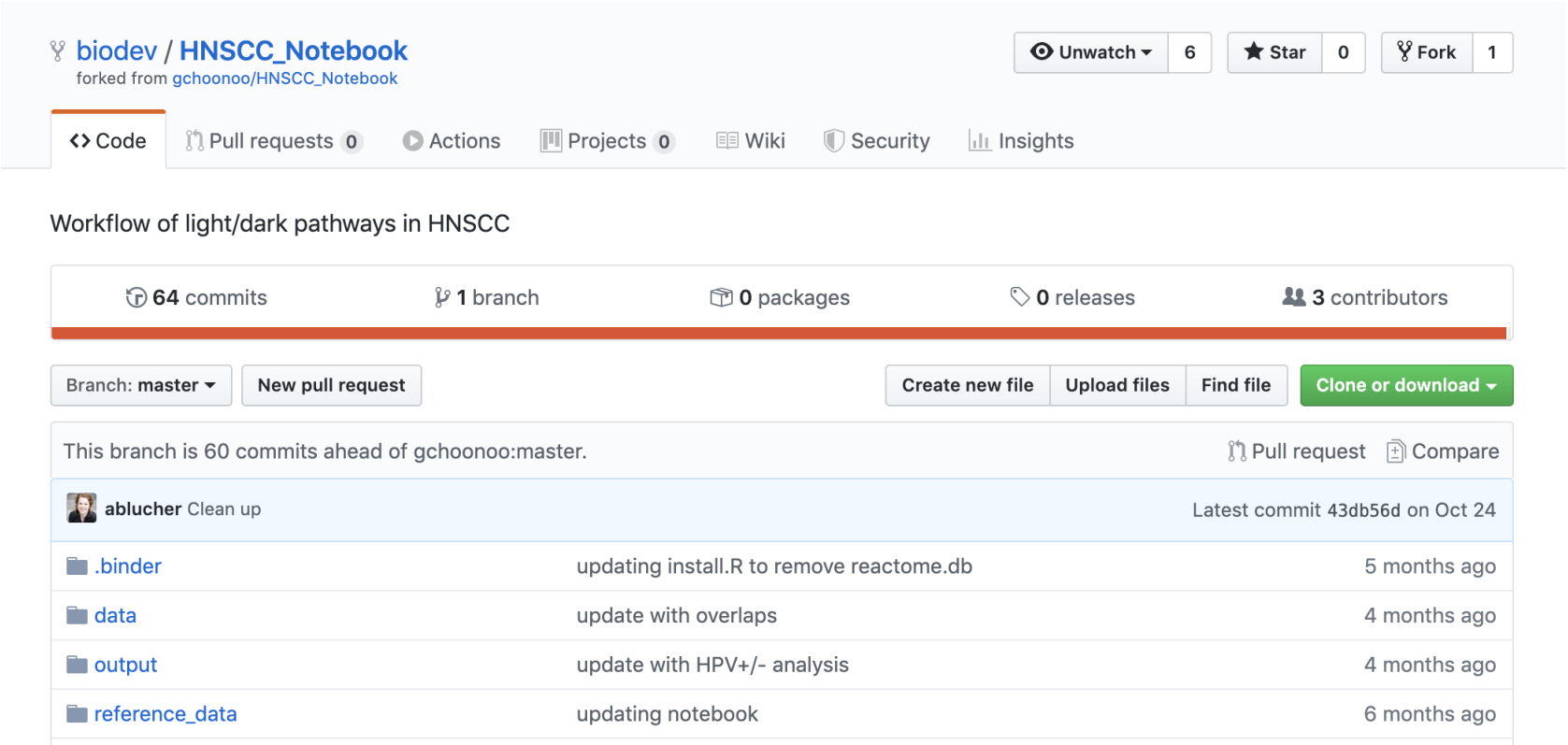
Good Practices in Project Organization

MyProject_Folder

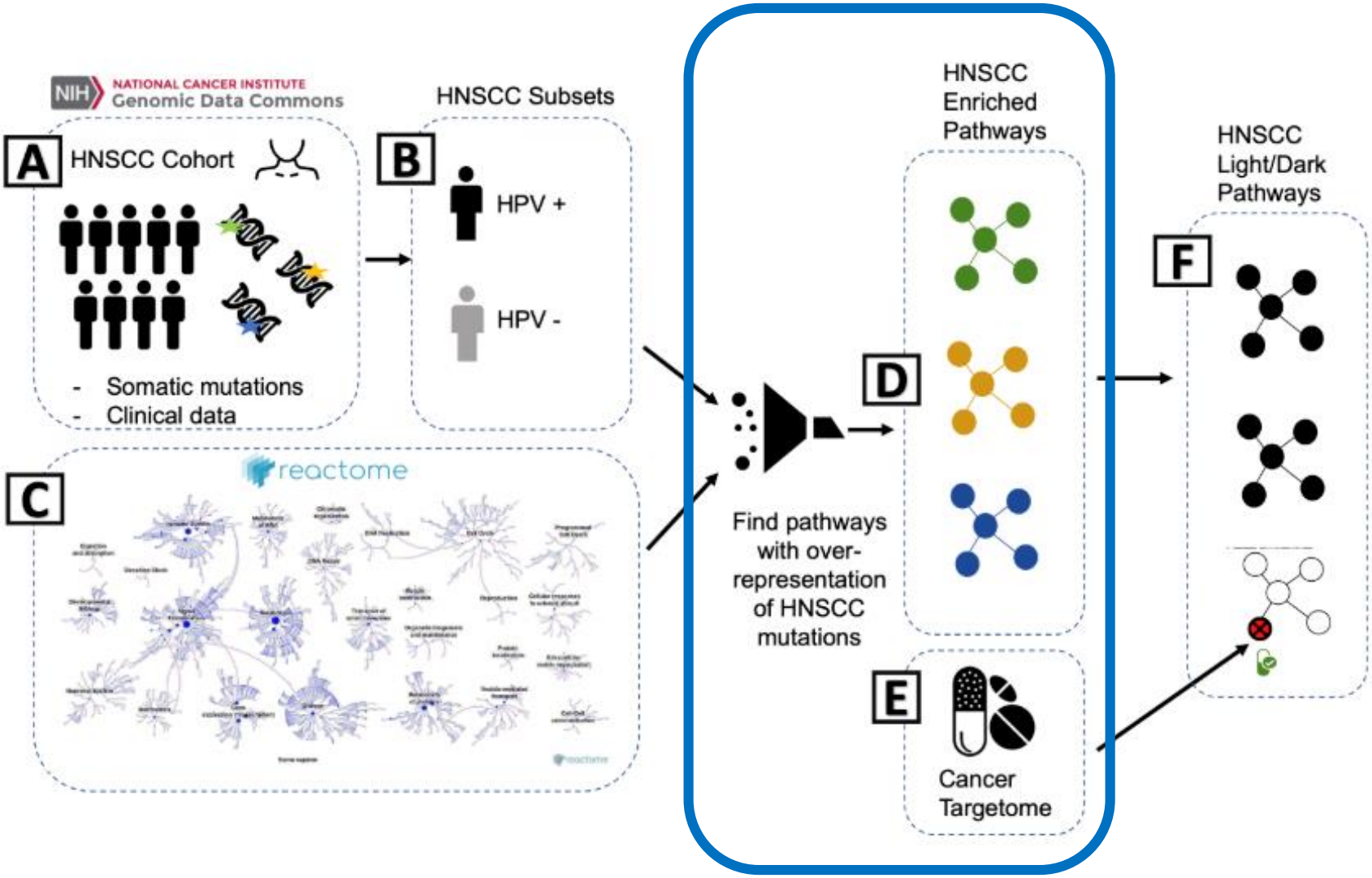
>data ← original_data, cleaned_data, resources

>R ← .R scripts, markdown files, notebooks

>output ← figures, tables, etc.



Identify key analysis steps



Good Practices in Project Organization

- **What are the main scripts used for analysis?**
  - **versus exploratory/one-off scripts**
- **Do they run?**
- **Are input files and output files clearly described?**
- **Packages/dependencies at top of scripts**
- **Helpful commenting**

→ **GitHub Repository**

Great stage for a code review/ coding buddy

http://ropensci.org <- open code reviewers for scientific R packages

here() package in R

MyProject_Folder

>data

>R

>output

looks for .Rproj file
here() makes this your root directory
all file paths now relative to root

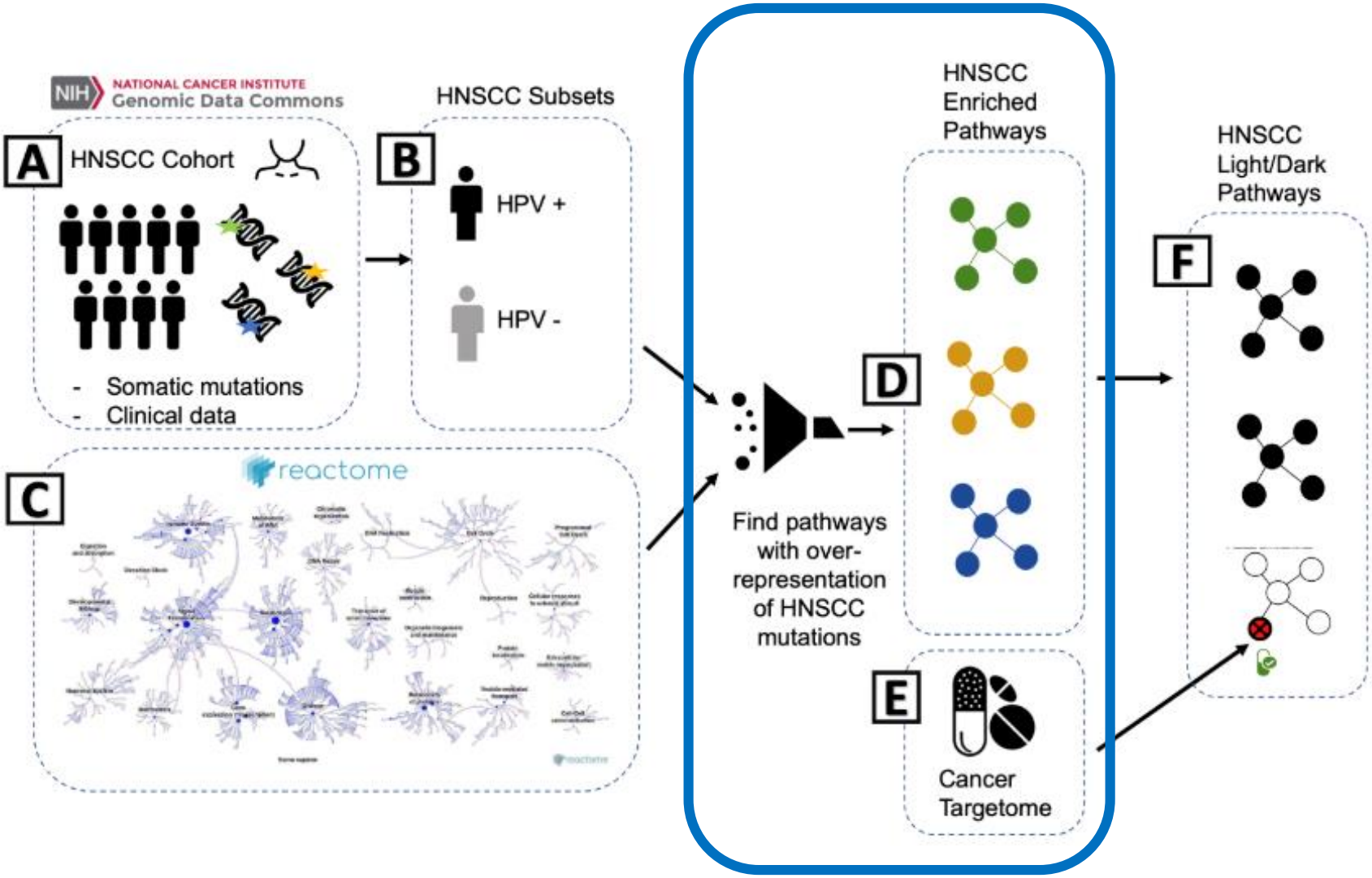
```
>library(here) #attach library
>here() #show me my root directory
>myfile<-read_csv(here("data",  "myfile.csv")) #read in file
```

cross-platform compatible file paths
Can move an Rmarkdown report anywhere
in project and will still execute

Identifying your inputs/ analysis steps/ output

Separate out any sub-analyses

Identify key analysis steps

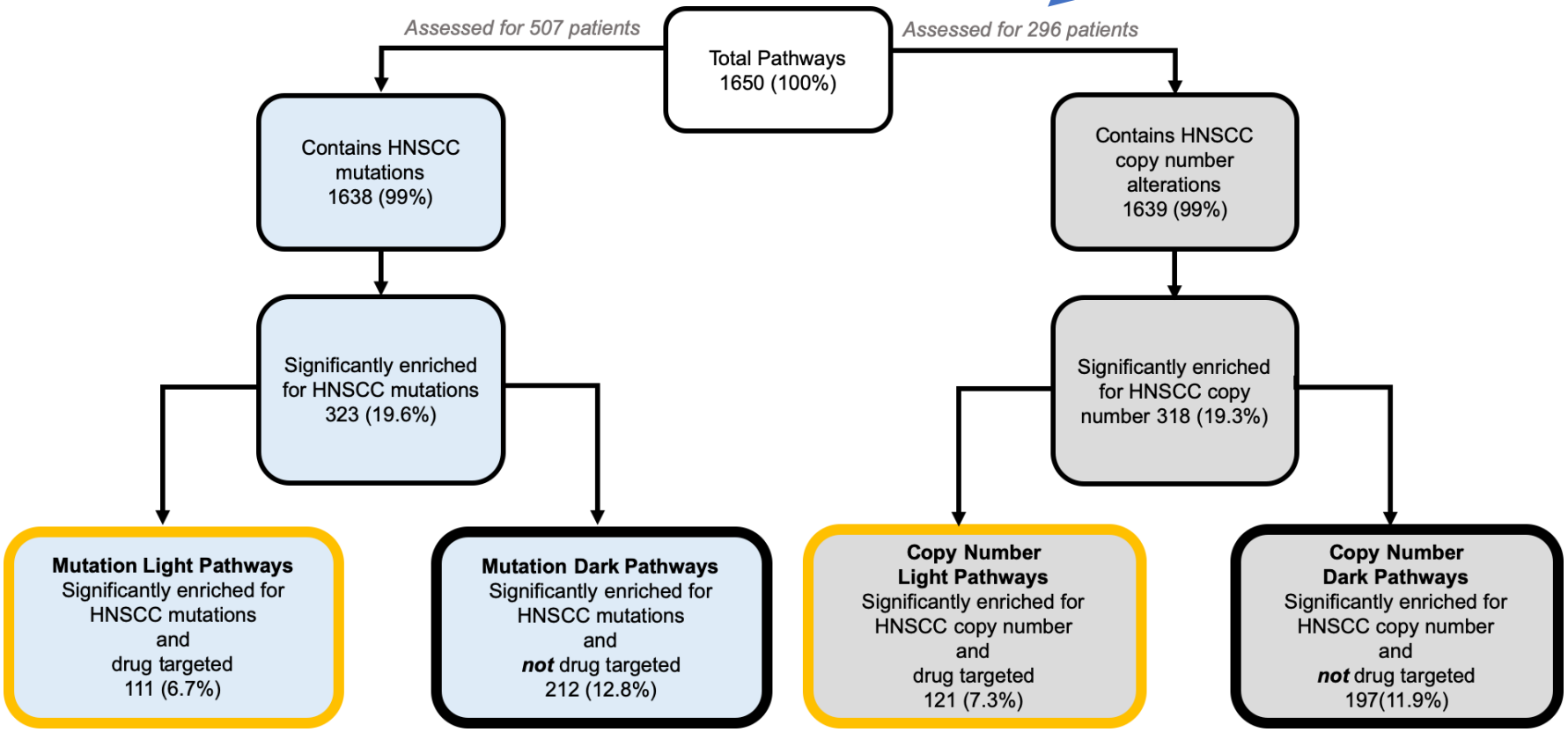


Do you have similar sub-analyses?

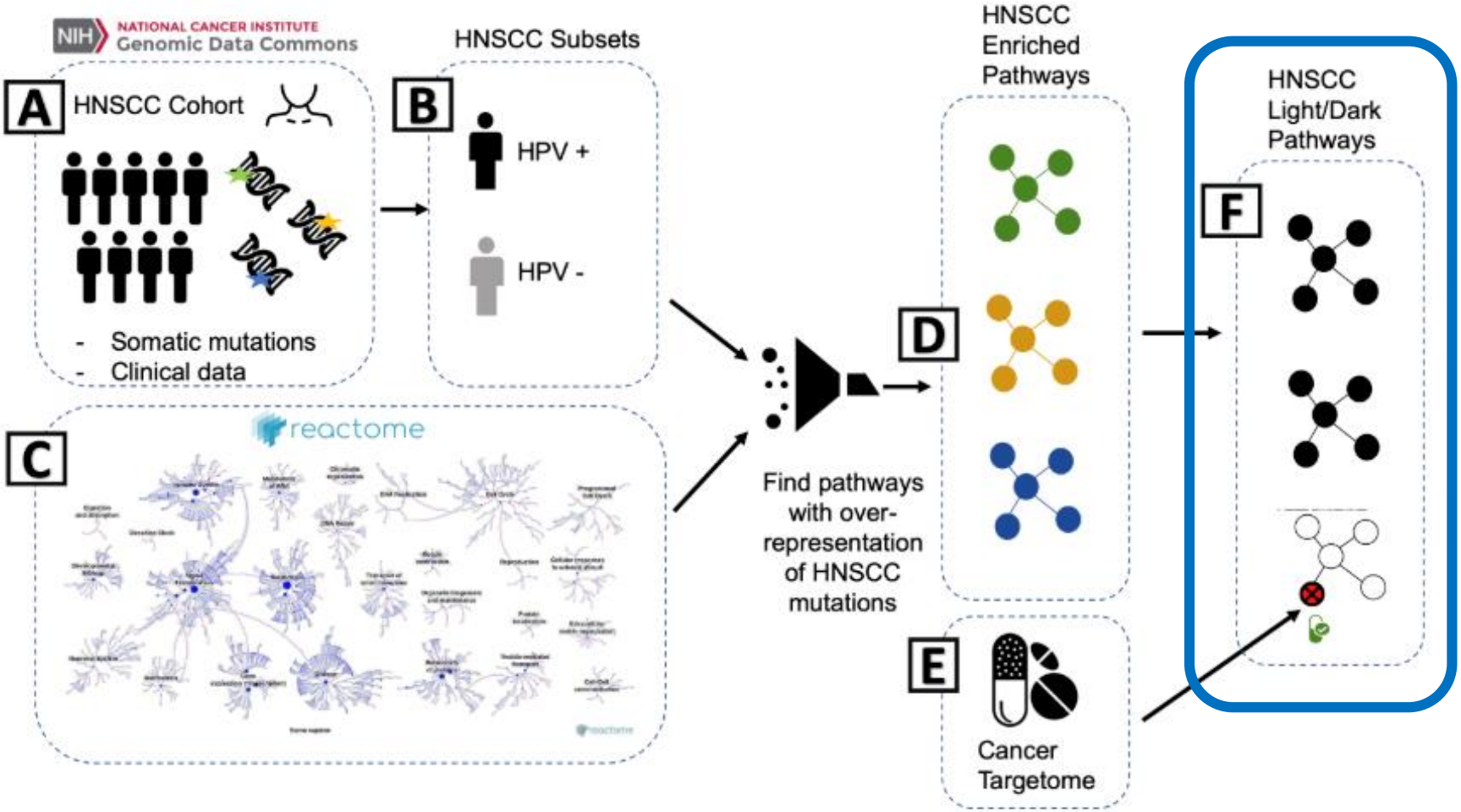
Consider adding workflow figures

Differentiate between sequential versus parallel tasks

Sample sizes, coverage, serve as reproducibility landmarks



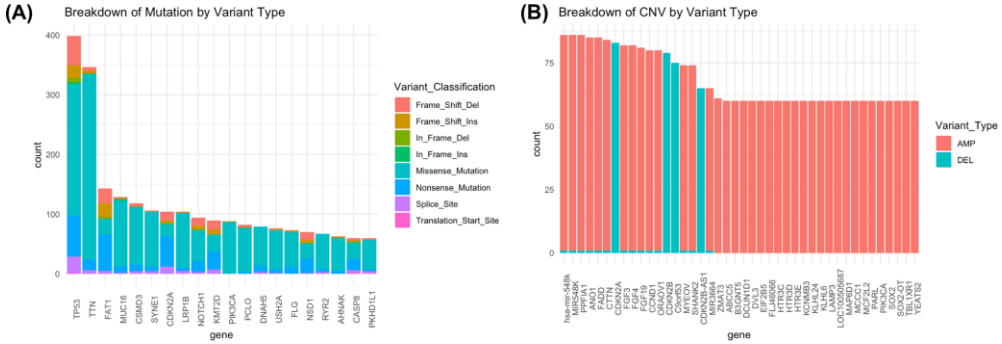
Identify key outputs



Recreating Your Results

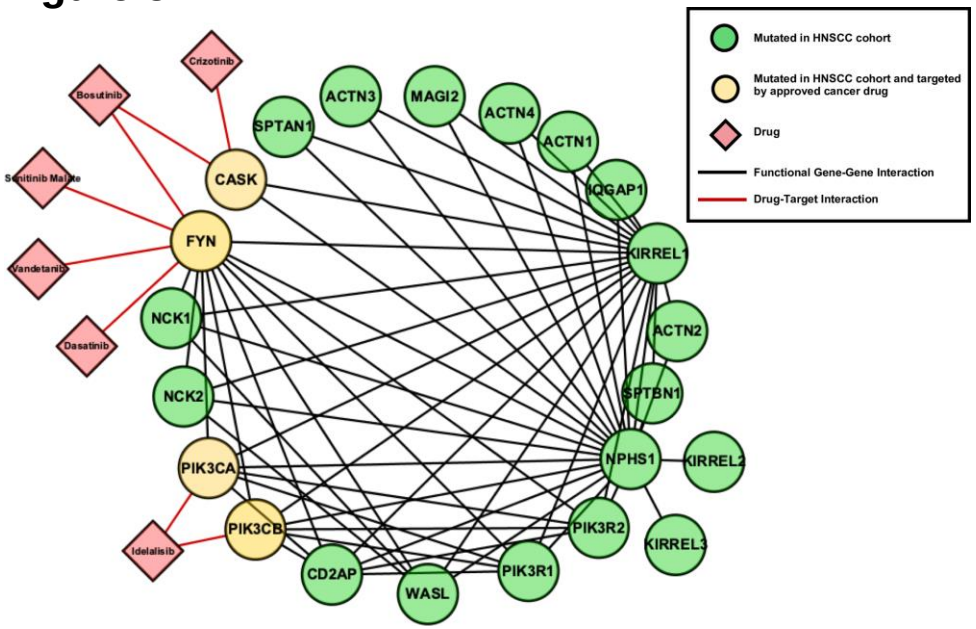
Where do all my figures and tables come from?

Figure 2. A and B.



Created within R scripts

Figure 5.



Created in another software application (Cytoscape/ ReactomeFIVIz)

Recreating Your Results

Don’t forget your supplemental!

Good Practices in Project Organization

Make a clear path to your outputs

Imagine you are guiding a friend who is excited about your research!

Add links to key outputs directly in your README.md

Output

Mutation Enriched Light Pathways are pathways found to be mutationally enriched in the cohort and also drug-targeted.

Mutation Enriched Dark Pathways are pathways found to be mutationall enriched in the cohort and are not currently drug-targeted.

Copy Number Enriched Light Pathways are pathways found to be copy number enriched in the cohort and also drug-targeted.

Copy Number Enriched Dark Pathways are pathways found to be copy number enriched in the cohort and are not currently drug-targeted.

Additional output for the sub-analysis of HPV cohorts:

HPV-Positive Cohort Light Pathways

HPV-Positive Cohort Dark Pathways

HPV-Negative Cohort Light Pathways

HPV-Negative Cohort Dark Pathways

Code Reproducibility

Literate programming/ R markdown notebooks

- Walk-through R markdown notebook

Reproducible Software Environment

- Best Practice is to reproduce the entire software environment used in analysis
- Many tools for this that are language specific: R: renv and Python: virtualenv
- Docker: lets you reproduce the entire software environment (analysis software versions, software dependencies and software packages needed) in a OS independent manner
- Need to specify packages and versions (use tags to specify releases)
- Don't get too dependent on any one install of software – ensure that your analysis can be run across OSes and versions

binder

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

Creating a “Binder”

=

Creating a “Binder-Ready” Repository (e.g. Git Repo)

=

Your Repository + Code + Configuration Files

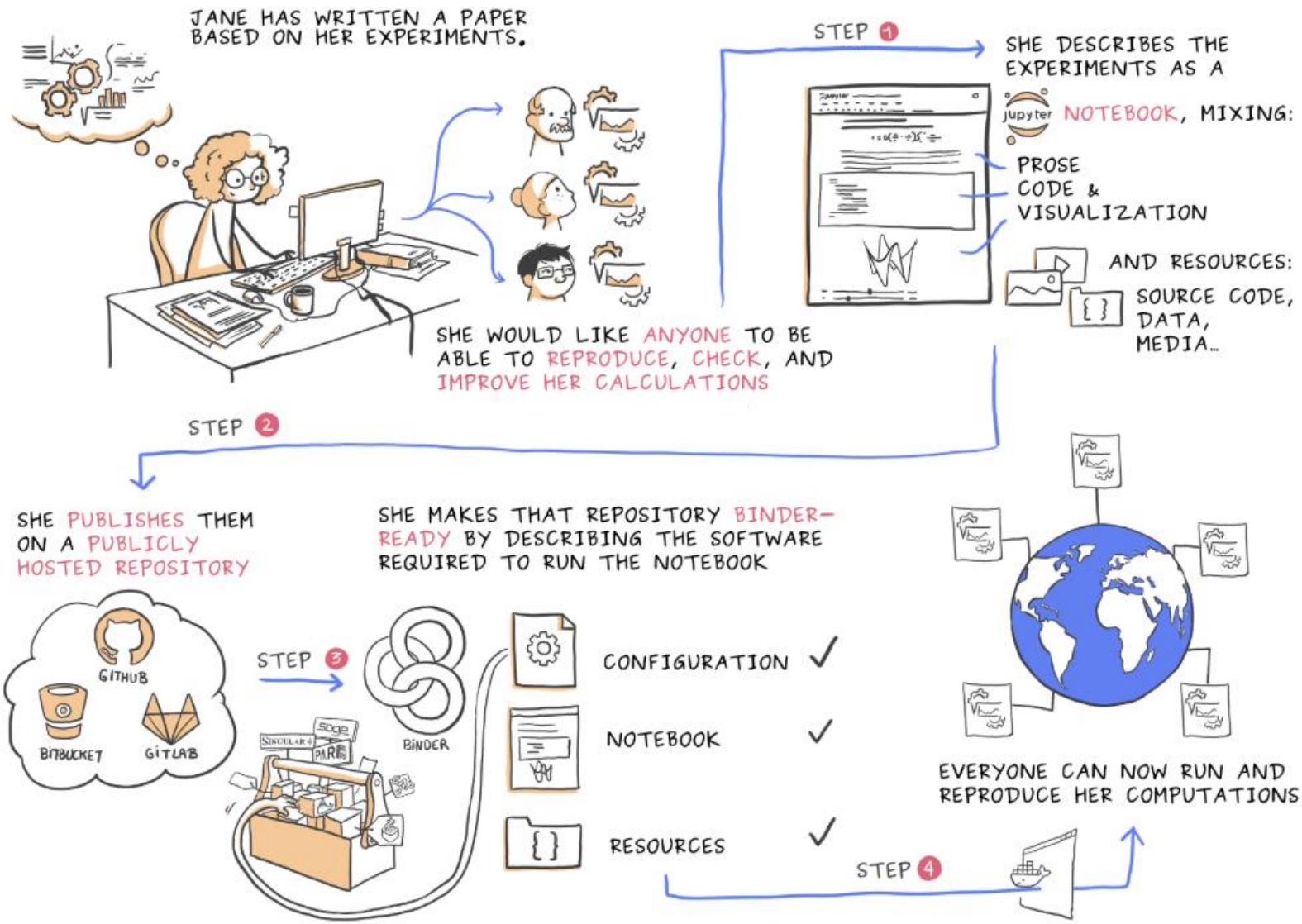


Figure credit: Juliette Taka, Logilab and the OpenDreamKit project

Hands On - Setting up a Github Repository/Compendia for Binder

Github repository (public)

R markdown notebook

Configuration for Binder

Option 1. install.R and runtime.txt
install.R #R script that with install.packages() calls
runtime.txt #specify R version here

Demo for today

Option 2. Docker file set up
binder/ Dockerfile

Alternate option

More info: Research Compendium: https://research-compendium.science/

Holepunch Package for Binder: https://github.com/karthik/holepunch

http://bit.ly/bdc_binder

mybinder.org

Build and launch a repository

GitHub repository name or URL

GitHub ▾ https://github.com/ablucher/Workshop_ReproduciblePaper

Git branch, tag, or commit

Git branch, tag, or commit

URL to open (optional)

rstudio URL ▾

launch

Copy the URL below and share your Binder with others:

```
https://mybinder.org/v2/gh/ablucher/Workshop_ReproduciblePaper/master?urlpath=rstudio
```

Copy the text below, then paste into your README to show a binder badge: launch binder

m

```
[![Binder](https://mybinder.org/badge_logo.svg)](https://mybinder.org/v2/gh/ablucher/Workshop_Repro
```

.rst

```
.. image:: https://mybinder.org/badge_logo.svg
 :target: https://mybinder.org/v2/gh/ablucher/Workshop_ReproduciblePaper/master?urlpath=rstudio
```

This will take a while the first time you build your binder!

How Docker Operates Behind the Scenes (repo2docker)

- **Docker = a program to let you run multiple operating systems on your computer**

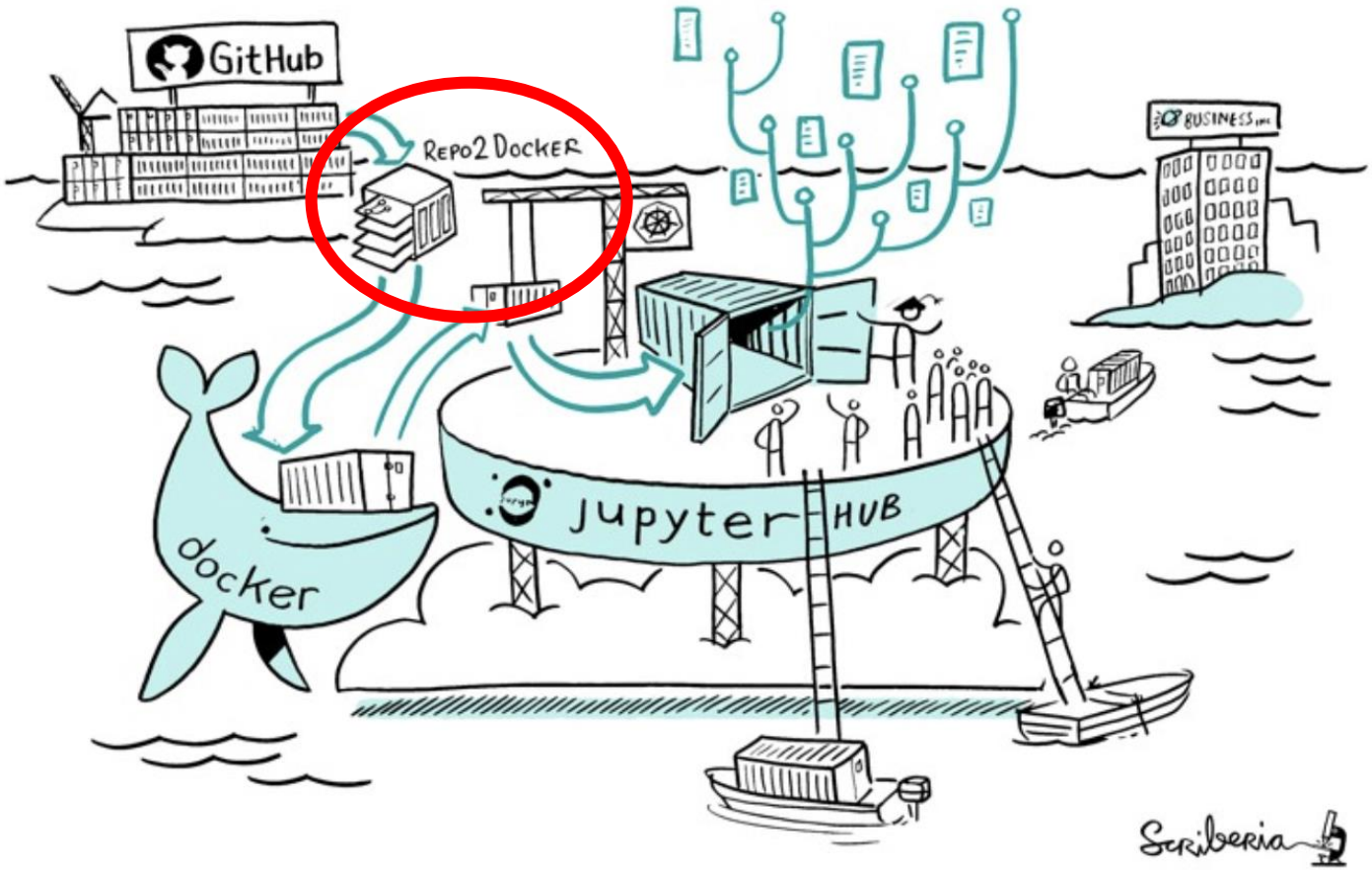
- **We use Docker to specify our software environment as an image and run it as a container**

- **Images versus containers**
  - **Images are the definition for the operating systems**
  - **Containers are the actual running instance**

- **Option #2 is using Dockerfile to build our image**
  - **Dockerfile = configuration file**

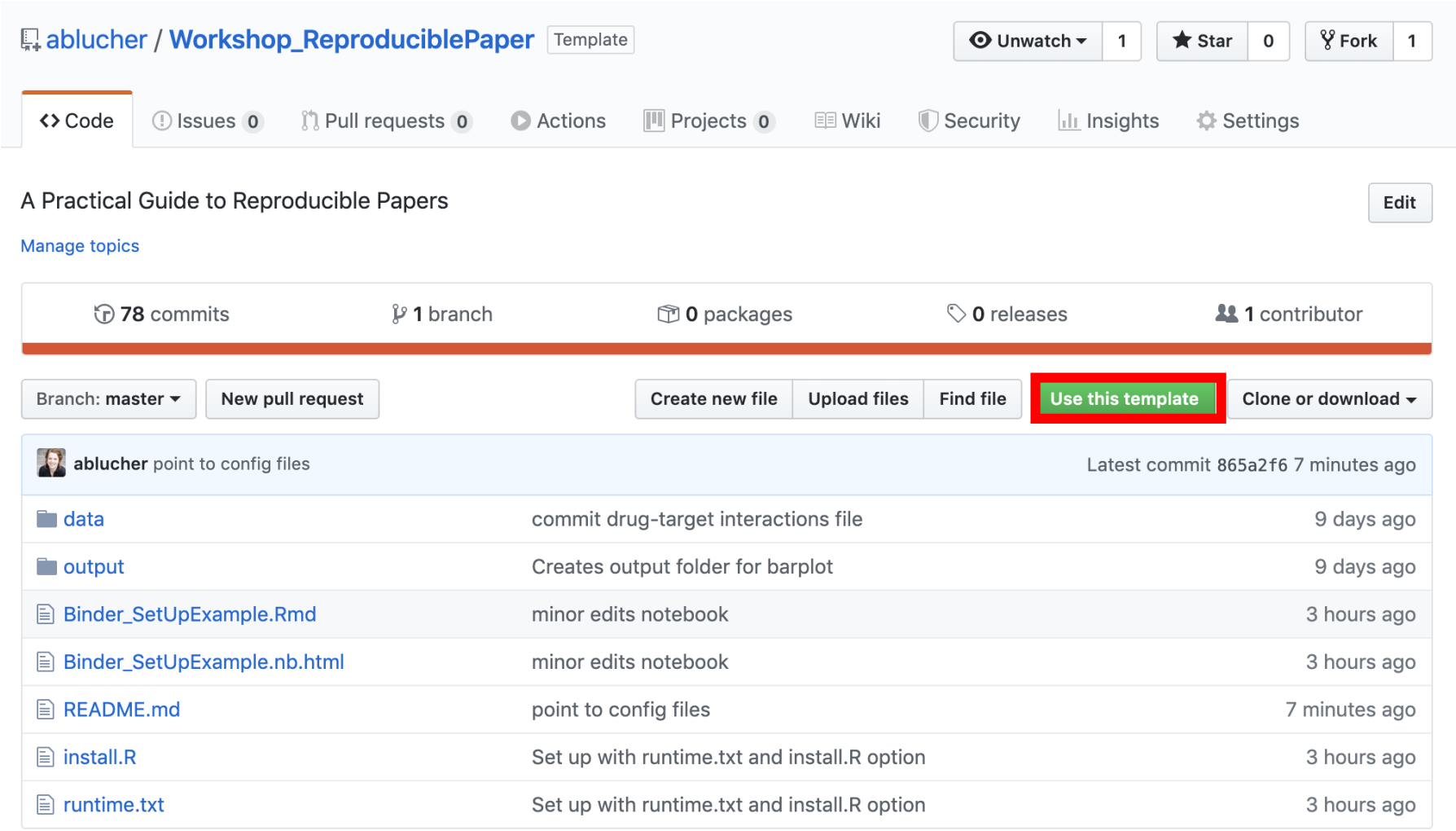
Docker and R Reproducibility
https://colinfay.me/docker-r-reproducibility/

What's going on behind the scenes?



A representation of the BinderHub architecture. This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence.

Using This Workshop as a Template



Good Practices for GitHub Project Management

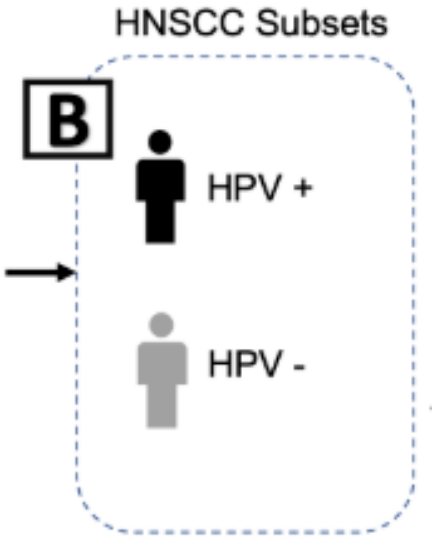
Making Version Control Work For You

- **Make sure all your files are in the repository**
- **Add numbering to your figures and tables to match manuscript drafts**
- **Clean up duplicate files**
  - **Remove outdated versions (version control means you have a history!)**
- **A Quick Guide to Organizing Computational Biology Projects:**
  - https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424

Bonus Round: Sub-analyses and annotation files

Sub-analyses and annotation files

- Adding annotation
  - Flag columns, date columns, curator columns
  - Add a README (can be tab in spreadsheet)
  - Explain to someone else <-> have a buddy
  - Don't be afraid of manual annotation steps – they often are information rich and incredibly valuable!!
    - But you need to leave a paper trail



HPV status annotated from 3 primary sources
-methods write-up
-citations for original papers
-README that explains the annotation file

Final Thoughts

- **Protocol/ Methods Documentation**

- **Iterative Process**

  Have a tester! Partner up with some for code review!

- **Time/effort commitment for reproducibility is non-trivial**

Acknowledgements

Ted Laderas | Pierrette Lo

Biodata Club

Head and Neck Squamous Cell Carcinoma Precision Medicine Group

Shannon McWeeney & Molly Kulesz-Martin

Gabrielle Choonoo | Mitzi Boardman | James Jacobs | Christina Zheng | Samuel Higgins | Sophia Jeng | Steve Chamberlin | Nate Evans | Miles Vigoda | Chase Mathieson | Ben Cordier | Ashley Anderson

Additional/ Backup Slides

How Docker Operates Behind the Scenes (repo2docker)

- **Docker = a program to let you run multiple operating systems on your computer**

- **We use Docker to specify our software environment as an image and run it as a container**

- **Images versus containers**
  - **Images are the definition for the operating systems**
  - **Containers are the actual running instance**

- **Option #2 is using Dockerfile to build our image**
  - **Dockerfile = configuration file**

Docker and R Reproducibility
https://colinfay.me/docker-r-reproducibility/

What's in our Docker file? Example docker file from Ted

```
FROM rocker/binder:3.5.3

USER root
RUN apt-get update -qq && apt-get -y --no-install-recommends install \
  libxml2-dev \
  libcairo2-dev \
  libsqlite3-dev \
  libmariadbd-dev \
  libmariadb-client-lgpl-dev \
  libpq-dev \
  libssh2-1-dev \
  unixodbc-dev \
  libsasl2-dev \
  && install2.r --error \
    --deps TRUE \
    dplyr \
    ggplot2 \
    here
COPY . ${HOME}
RUN chown -R ${NB_USER} ${HOME}

USER ${NB_USER}
```