

Tools for Reproducible Research

<https://tinyurl.com/hbc-rr>



Harvard Chan Bioinformatics Core



Introductions!





Shannan Ho Sui
Director



Meeta Mistry
Associate Director



Lorena Pantano
*Director of Bioinformatics
Platform*



John Quackenbush
Faculty Advisor



Open Bhattarai



Heather Wick



Will Gammerdinger



Noor Sohail



Elizabeth
Partan



Alex Bartlett



Emma Berdan



James Billingsley



Zhu Zhuo



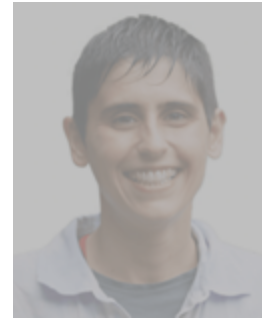
Maria Simoneau



Shannan Ho Sui
Director



Meeta Mistry
Associate Director



Lorena Pantano
*Director of Bioinformatics
Platform*



John Quackenbush
Faculty Advisor



Open Bhattarai



Heather Wick



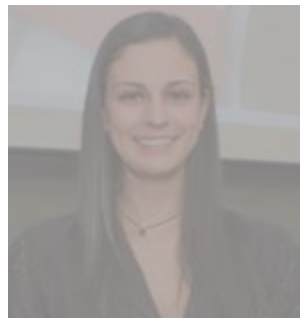
Will Gammerdinger



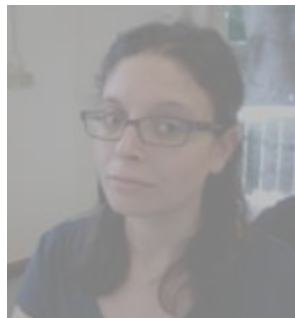
Noor Sohail



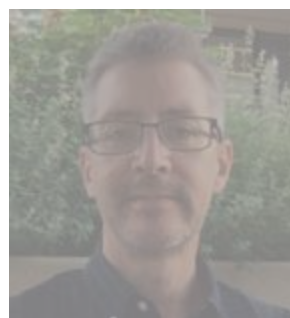
Elizabeth
Partan



Alex Bartlett



Emma Berdan



James Billingsley



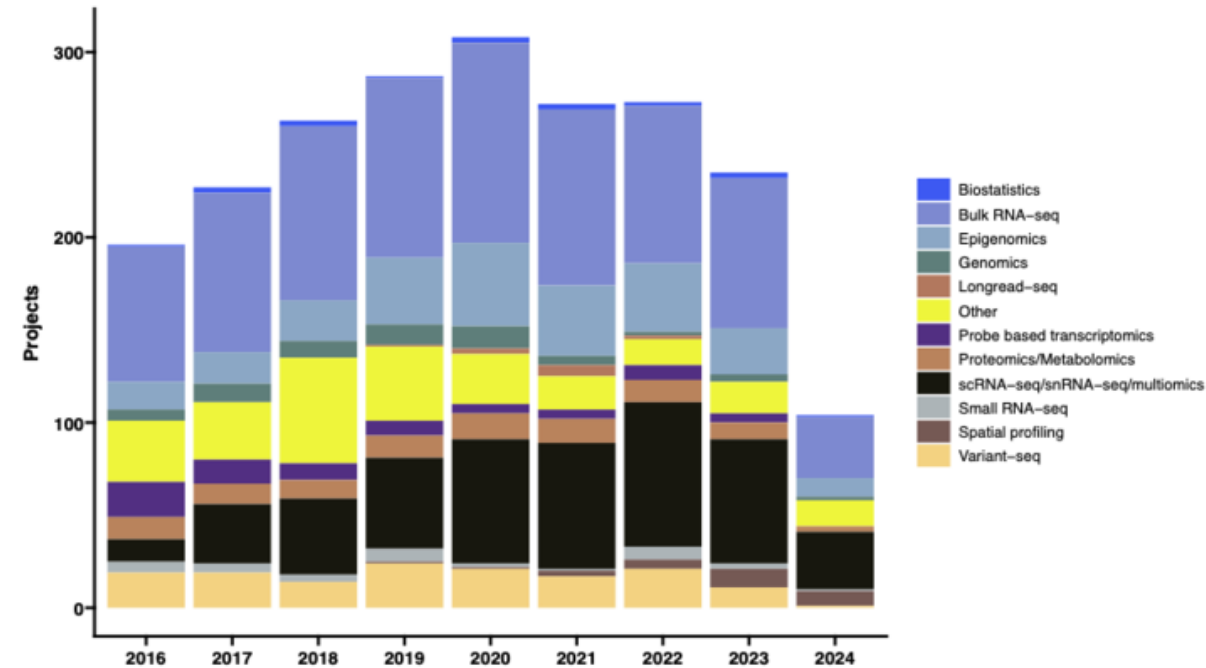
Zhu Zhuo



Maria Simoneau

Consulting

- ❖ Transcriptomics: Bulk, single cell, small RNA
- ❖ Epigenomics: ChIP-seq, CUT&RUN, ATAC-seq, DNA methylation
- ❖ Variant discovery: WGS, resequencing, exome-seq and CNV
- ❖ Multiomics integration
- ❖ Spatial biology
- ❖ Experimental design and grant support



Consulting

- ❖ Transcriptomics: Bulk, single cell, small RNA
- ❖ Epigenomics: ChIP-seq, CUT&RUN, ATAC-seq, DNA methylation
- ❖ Variant discovery: WGS, resequencing, exome-seq and CNV
- ❖ Multiomics integration
- ❖ Spatial biology
- ❖ Experimental design and grant support



HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NIEHS



THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER



HARVARD
MEDICAL SCHOOL

Training

- ❖ Hands-on workshops design to reflect best practices, reproducibility and an emphasis on experimental design
 - ❖ Basic Data Skills
 - ❖ Shell
 - ❖ R
 - ❖ Advanced Topics: Analysis of high-throughput sequencing data
 - ❖ Chromatin Biology
 - ❖ Bulk RNA-seq
 - ❖ Differential Gene Expression
 - ❖ scRNA-seq
 - ❖ Variant Calling
 - ❖ Current Topics in Bioinformatics

Training

- ❖ Hands-on workshops design to reflect best practices, reproducibility and an emphasis on experimental design
 - ❖ Basic Data Skills
 - ❖ Shell
 - ❖ R
 - ❖ Advanced Topics: Analysis of high-throughput sequencing data
 - ❖ Chromatin Biology
 - ❖ Bulk RNA-seq
 - ❖ Differential Gene Expression
 - ❖ scRNA-seq
 - ❖ Variant Calling
 - ❖ Current Topics in Bioinformatics

<https://bioinformatics.sph.harvard.edu/training>



HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

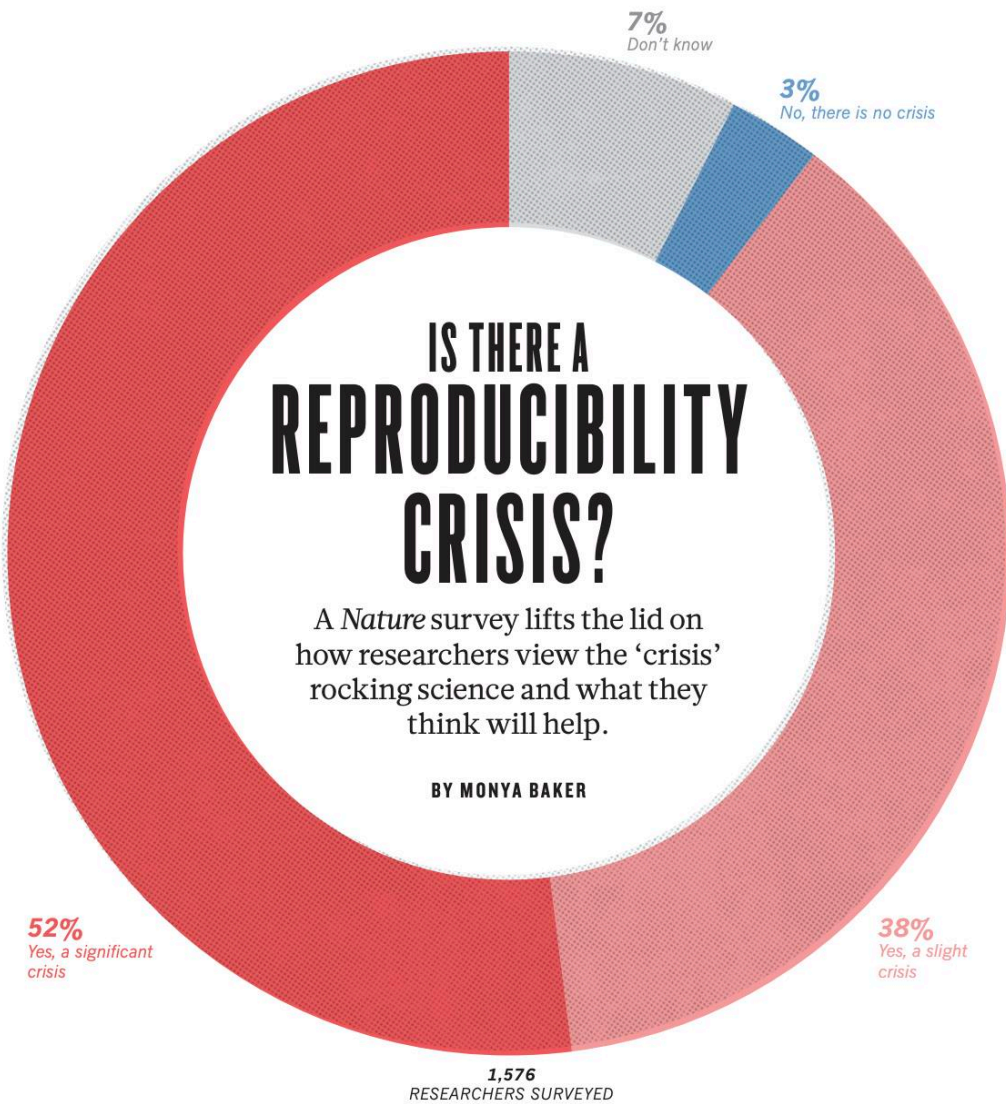
DF/HCC
DANA-FARBER / HARVARD CANCER CENTER



THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER



Workshop scope



a replication study. When work does not reproduce, researchers often assume there is a perfectly valid (and probably boring) reason. What's more, incentives to publish positive replications are low and journals can be reluctant to publish negative findings. In fact, several respondents who had published a failed replication said that editors and reviewers demanded that they play down comparisons with the original study.

Nevertheless, 24% said that they had been able to publish a successful replication and 13% had published a failed replication. Acceptance was more common than persistent rejection: only 12% reported being unable to publish successful attempts to reproduce others' work; 10% reported being unable to publish unsuccessful attempts.

Survey respondent Abraham Al-Ahmad at the Texas Tech University Health Sciences Center in Amarillo expected a "cold and dry rejection" when he submitted a manuscript explaining why a stem-cell technique had stopped working in his hands. He was pleasantly surprised when the paper was accepted³. The reason, he thinks, is because it offered a workaround for the problem.

Others place the ability to publish replication attempts down to a combination of luck, persistence and editors' inclinations. Survey respondent Michael Adams, a drug-development consultant, says that work showing severe flaws in an animal model of diabetes has been rejected six times, in part because it does not reveal a new drug target. By contrast, he says, work refuting the efficacy of a compound to treat Chagas disease was quickly accepted⁴.

THE CORRECTIVE MEASURES

One-third of respondents said that their labs had taken concrete steps to improve reproducibility within the past five years. Rates ranged from a high of 41% in medicine to a low of 24% in physics and engineering. Free-text responses suggested that redoing the work or asking someone else within a lab to repeat the work is the most common practice. Also common are efforts to beef up the documentation and standardization of experimental methods.

Any of these can be a major undertaking. A biochemistry graduate student in the United Kingdom, who asked not to be named, says that efforts to reproduce work for her lab's projects doubles the time and materials used — in addition to the time taken to troubleshoot when some things invariably don't work. Although replication does boost confidence in results, she says, the costs mean that she performs checks only for innovative projects or unexpected results.

Consolidating methods is a project unto itself, says Laura Shankman, a postdoc studying smooth muscle cells at the University of Virginia, Charlottesville. After several postdocs and graduate students left her lab within a short time, remaining members had trouble getting consistent results in their experiments. The lab decided to take some time off from new questions to repeat published work, and this revealed that lab protocols had gradually diverged. She thinks that the lab saved money overall by getting synchronized instead of troubleshooting failed experiments, but that it was a long-term investment.

people mentioned this strategy. One who did was Hanne Watkins, a graduate student studying moral decision-making at the University of Melbourne in Australia. Going back to her original questions after collecting data, she says, kept her from going down a rabbit hole. And the process, although time consuming, was no more arduous than getting ethical approval or formatting survey questions. "If it's built in right from the start," she says, "it's just part of the routine of doing a study."

THE CAUSE

The survey asked scientists what led to problems in reproducibility. More than 60% of respondents said that each of two factors — pressure to publish and selective reporting — always or often contributed. More than half pointed to insufficient replication in the lab, poor oversight or low statistical power. A smaller proportion pointed to obstacles such as variability in reagents or the use of specialized techniques that are difficult to repeat.

But all these factors are exacerbated by common forces, says Judith Kimble, a developmental biologist at the University of Wisconsin–Madison: competition for grants and positions, and a growing burden of bureaucracy that takes away from time spent doing and designing research. "Everyone is stretched thinner these days," she says. And the cost extends beyond any particular research project. If graduate students train in labs where senior members have little time for their juniors, they may go on to establish their own labs without having a model of how training and mentoring should work. "They will go off and make it worse," Kimble says.

WHAT CAN BE DONE?

Respondents were asked to rate 11 different approaches to improving reproducibility in science, and all got ringing endorsements. Nearly 90% — more than 1,000 people — ticked "More robust experimental design" "better statistics" and "better mentorship". Those ranked higher than the option of providing incentives (such as funding or credit towards tenure) for reproducibility-enhancing practices. But even the lowest-ranked item — journal checklists — won a whopping 69% endorsement.

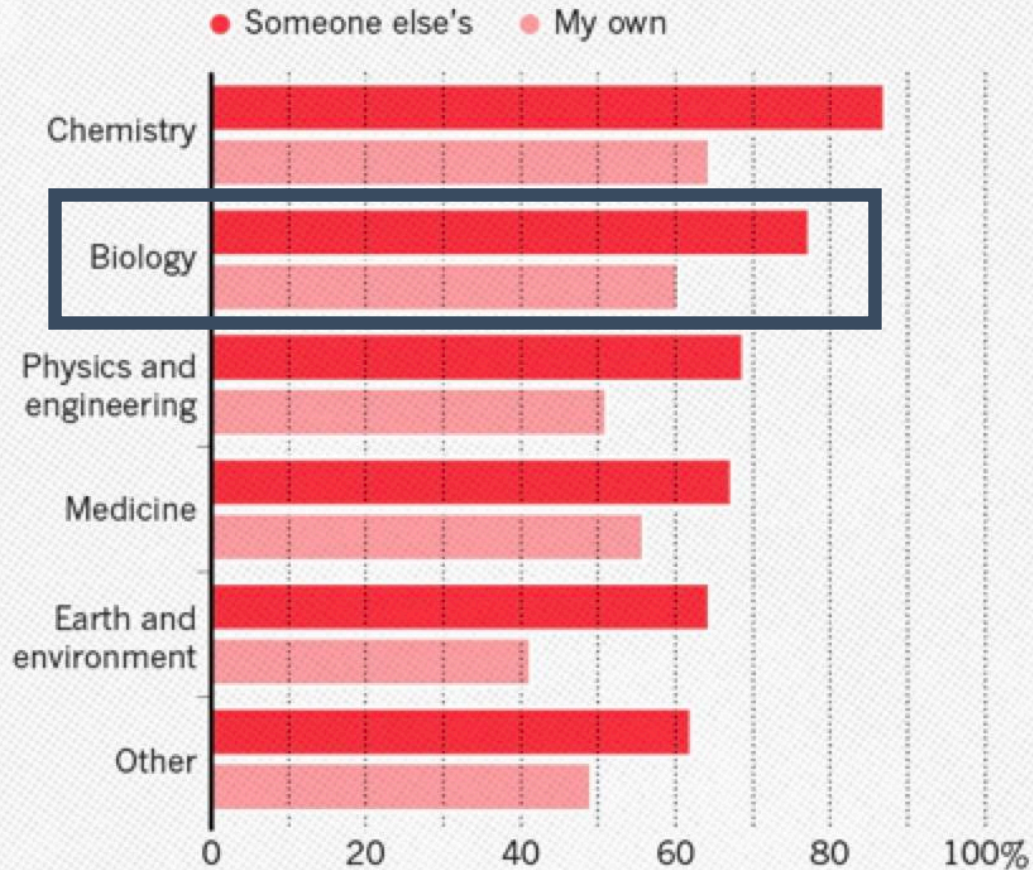
The survey — which was e-mailed to *Nature* readers and advertised on affiliated websites and social-media outlets as being 'about reproducibility' — probably selected for respondents who are more receptive to and aware of concerns about reproducibility. Nevertheless, the results suggest that journals, funders and research institutions that advance policies to address the issue would probably find cooperation, says John Ioannidis, who studies scientific robustness at Stanford University in California. "People would probably welcome such initiatives." About 80% of respondents thought that funders and publishers should do more to improve reproducibility.

"It's healthy that people are aware of the issues and open to a range of straightforward ways to improve them," says Munafò. And given that these ideas are being widely discussed, even in mainstream media, take

"REPRODUCIBILITY IS LIKE BRUSHING YOUR TEETH. ONCE YOU LEARN IT, IT BECOMES A HABIT."

HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

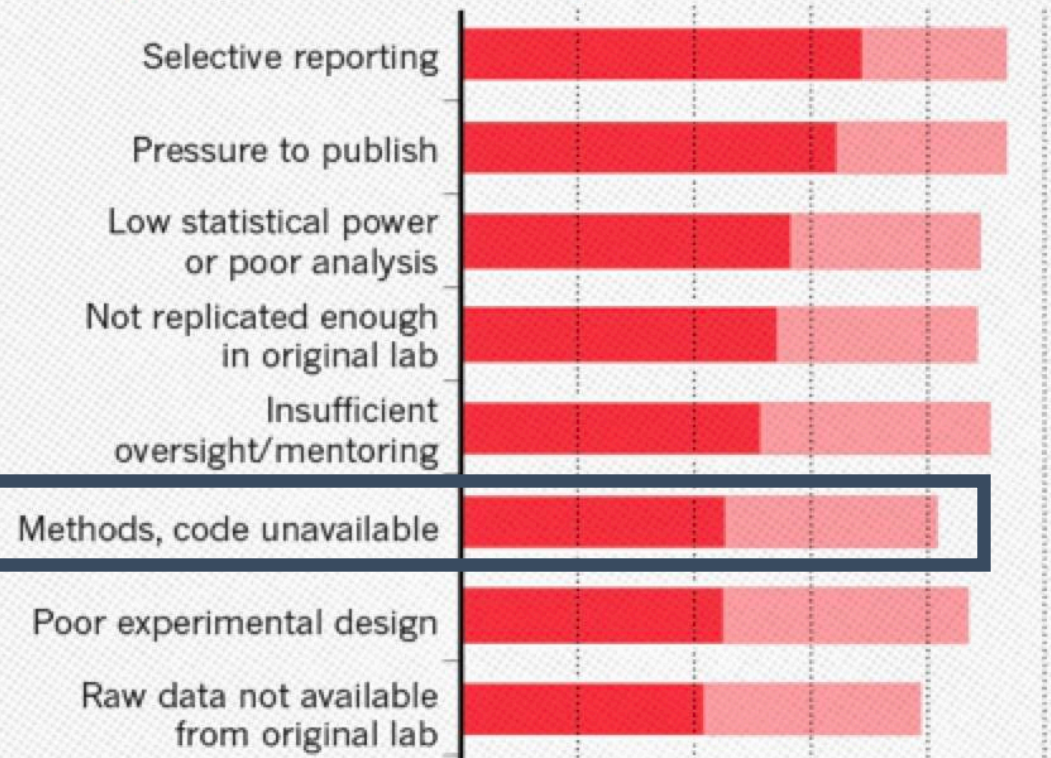
Most scientists have experienced failure to reproduce results.



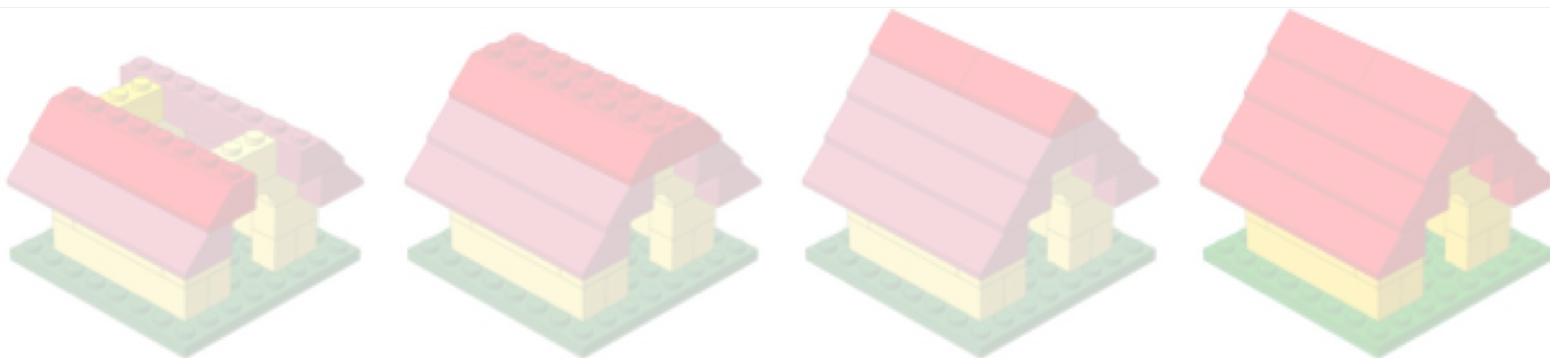
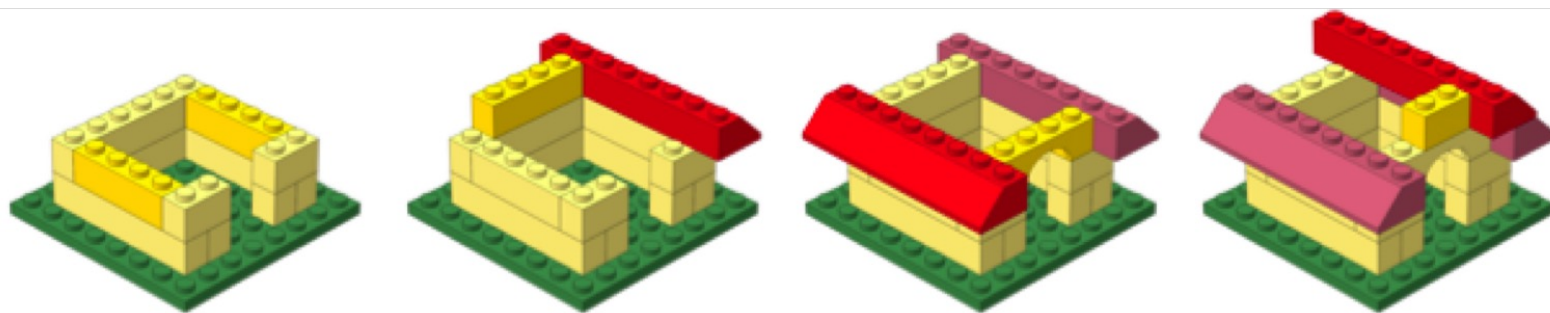
WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

● Always/often contribute ● Sometimes contribute

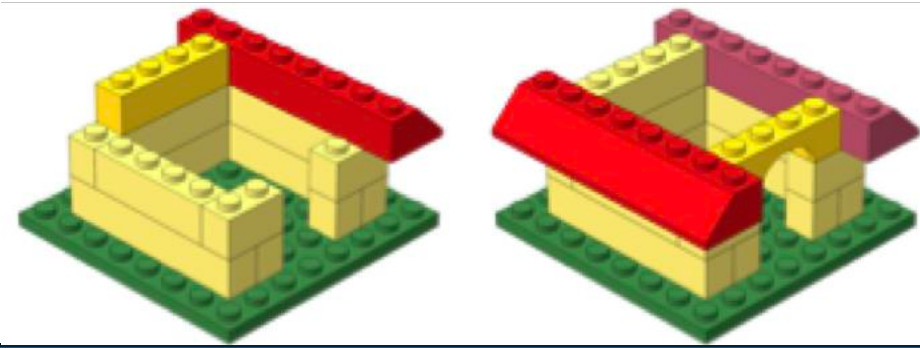


“Reproducibility is a minimum necessary condition for a finding to be believable and informative.”



Bioinformatic Data Analysis

Workshop Scope



- ❖ Generate reports for your analyses using RMarkdown
- ❖ Track changes as you work on files using a version control system called Git (GitHub Desktop tool)
- ❖ Collaborate effectively, and disseminate code & other documents using Github

Logistics



Course schedule

Day 1

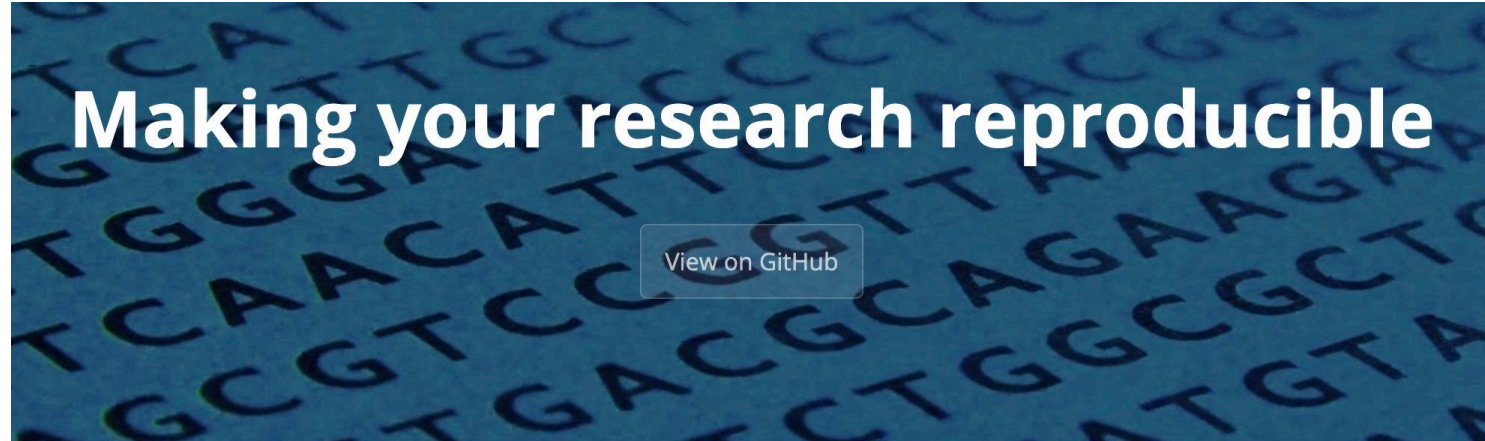
Time	Topic	Instructor
09:30 - 9:45	Workshop Introduction	Will
09:45 - 10:30	Making your data analysis reproducible	Julie Goldman
10:30 - 10:35	Break	
10:35 - 11:10	RMarkdown Basics	Heather
11:10 - 11:55	RMarkdown Intermediate	Will
11:55 - 12:00	Assignment review	Will

Assignment #1

- [Practice with RMarkdown](#)
- Upload the files requested in the above exercise to [Dropbox](#) **day before the next class**.
- [Email us](#) about questions that you need answered to work through the exercise.
- [Answer key](#)

Course materials

- ❖ We continuously update our materials to reflect changes in the field/software



Learning Objectives

- Describe the need for reproducible research
- Create RMarkdown reports for sharing analysis methods, code and results

Making your research reproducible

We have already made a case about reproducibility in the introduction to this workshop. In this lesson we will focus on one of the tools to enable and empower you to perform analysis reproducibly.

When you do lab work, you use lab notebooks to organize your methods, results, and conclusions for future retrieval and reproduction. The information in these notebooks is converted into a more concise experimental description for the Methods section when publishing the results.

Single Screen & 3 Windows

The image illustrates a single-screen setup with three overlapping windows:

- Zoom Meeting:** Shows a video call with three participants: Mary Piper (Co-host, me), Troubleshooter (Radhika) (Co-host), and Jihe Liu (Host). The interface includes a 'You are viewing Jihe Liu's screen' notification and a 'View Options' dropdown.
- RStudio IDE:** Displays R code in the editor and the console. The code includes:

```
483  
484  
485 getwd()  
486  
487 # square root function  
488 sqrt(81)  
489  
490 # round function  
491 round(3.14159)  
492 ?round  
493  
494  
495
```

The console shows the output of these commands:

```
> # round function  
> round(3.14159)  
[1] 3  
> ?round  
>
```

The Environment pane shows the values of variables:

```
Values  
number 15  
x      3  
y      10
```
- Web Browser:** Shows a page with a 'View on GitHub' button and a background image of DNA sequence letters (A, T, C, G).

Single Screen & 3 Windows

Zoom

Our Recommendation

```
# Assignment operator
x <- 3

# Functions
getwd()
sqrt(81)
round(3.14159)
?round
```

```
> x <- 3
> # Functions
> getwd()
[1] "/Users/mariyaper/Desktop/R-testing"
> sqrt(81)
[1] 9
> round(3.14159)
[1] 3
> ?round
```

Rounding of Numbers

Description

ceiling takes a single numeric argument x and returns a numeric vector containing the smallest integers not less than the corresponding elements of x.

floor takes a single numeric argument x and returns a numeric vector containing the largest integers not greater than the corresponding elements of x.

trunc takes a single numeric argument x and returns a numeric vector containing the integers formed by truncating the values in x toward 0.

round rounds the values in its first argument to the specified number of decimal places (default 0). See 'Details' about "round to even" when rounding off a 5.

signif rounds the values in its first argument to the specified number of significant digits.

Usage

```
ceiling(x)
floor(x)
trunc(x, ...)
```

Single Screen & 3 Windows

Participants (3)

- Mary Piper (Co-host, me)
- Ji He Liu (Host)
- Troubleshooter (Radhika) (Co-host)

Web Browser

```
1 # Assignment operator
2 x <- 3
3
4 # Functions
5 getwd()
6
7 sqrt(81)
8
9 round(3.14159)
10 ?round
11
```

Environment History Connections

Global Environment

Values

x	3
---	---

Files Plots Packages Help Viewer

R: Rounding of Numbers

Round (base)

Rounding of Numbers

Description

integers not less than the corresponding elements of x.

floor: takes a single numeric argument x and returns a numeric vector containing the largest integers not greater than the corresponding elements of x.

trunc: takes a single numeric argument x and returns a numeric vector containing the integers formed by truncating the values in x toward 0.

round rounds the values in its first argument to the specified number of decimal places (default 0). See 'Details' about "round to even" when rounding off a 5.

signif rounds the values in its first argument to the specified number of significant digits.

Usage

```
ceiling(x)
floor(x)
trunc(x, ...)
```

Our Recommendation

Single Screen & 3 Windows

The image illustrates a single-screen setup for a Zoom meeting. The Zoom window at the top shows three participants: Mary Piper, Troubleshooter (Radhika), and Jihe Liu. Below the Zoom window, the RStudio interface is shown, displaying R code in the editor and the console output. The console output shows the execution of the following R code:

```
483  
484  
485 getwd()  
486  
487 # square root function  
488 sqrt(81)  
489  
490 # round function  
491 round(3.14159)  
492 ?round  
493  
494  
495
```

The console output shows the following results:

```
> # round function  
> round(3.14159)  
[1] 3  
> ?round  
>
```

The RStudio window also shows the environment pane with the following values:

```
Values  
number 15  
x 5  
y 10
```

The browser window at the top right shows a search for 'View on GitHub'.

Our Recommendation

R Studio

Single Screen & 3 Windows

Zoom

Web Browser

R Studio

```
1 # Assignment operator
2 x <- 3
3
4 # Functions
5 getwd()
6
7 sqrt(81)
8
9 round(3.14159)
10 ?round
11
```

```
> x <- 3
> # Functions
> getwd()
[1] "/Users/mariyaper/Desktop/R-testing"
> sqrt(81)
[1] 9
> round(3.14159)
[1] 3
> ?round
```

integers not less than the corresponding elements of x.
floor: takes a single numeric argument x and returns a numeric vector containing the largest integers not greater than the corresponding elements of x.
trunc: takes a single numeric argument x and returns a numeric vector containing the integers formed by truncating the values in x toward 0.
round rounds the values in its first argument to the specified number of decimal places (default 0). See 'Details' about "round to even" when rounding off a 5.
signif rounds the values in its first argument to the specified number of significant digits.

Usage
ceiling(x)
floor(x)
trunc(x, ...)

Our Recommendation

R Studio

Course participation

- ❖ Mandatory review of self-learning lessons and assignments
- ❖ Attendance required for all classes
- ❖ Your questions and active participation drive learning
- ❖ **We look forward to all of your questions!**



Course participation

- ❖ At-home lessons and exercises after each session
- ❖ Cover material not previously discussed
- ❖ Provides us feedback to help pace the course appropriately
- ❖ 3-5 hours to complete
- ❖ Homework load is heavier in the beginning of this workshop series and tapers off

Using AI for Assignments

❖ Do

- ❖ Try to resolve error messages with it
- ❖ Test code written by AI on a dataset where you have expected results
- ❖ Take the time to review the generated code line-by-line

❖ Don't

- ❖ Implement it in replacement to learning
- ❖ Write code that you don't understand
- ❖ Assume the output from an AI process is correct

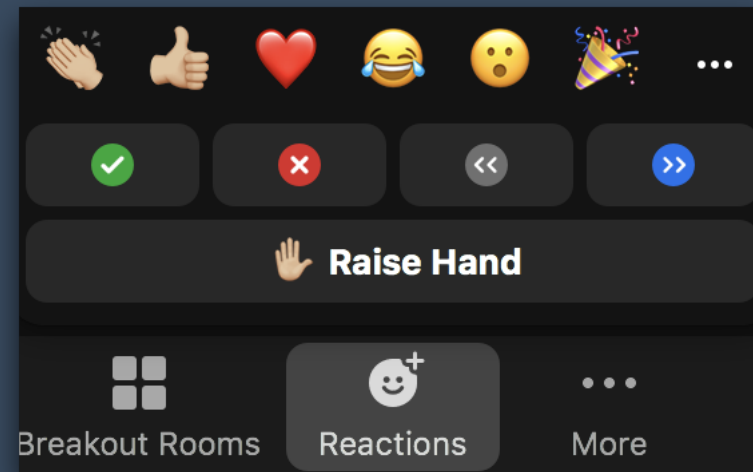
Odds & Ends

❖ Quit/minimize all applications that are not required for class

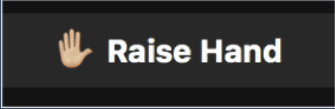
❖ Are you all set?

❖  = "agree", "I'm all set"

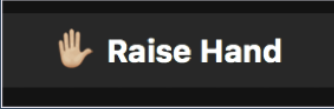
❖  = "disagree", "I need help"



Odds & Ends

- ❖ Questions for the presenter?
 - ❖ Post the question in the Chat window OR
 - ❖  when the presenter asks for questions
 - ❖ Let the Moderator know

Odds & Ends

- ❖ Questions for the presenter?
 - ❖ Post the question in the Chat window OR
 - ❖  when the presenter asks for questions
 - ❖ Let the Moderator know
- ❖ Technical difficulties with software?
 - ❖ Start a private chat with the Troubleshooter with a description of the problem

Contact Us



- ❖ *HBC training team:* hbctraining@hsph.harvard.edu
- ❖ *HBC consulting:* bioinformatics@hsph.harvard.edu