

Introduction to R

Harvard Chan Bioinformatics Core

<https://tinyurl.com/hbc-r-online>

Sponsored by DF/HCC and HMS Foundry



Shannan Ho Sui
Director



Victor Barrera



Amelie Jule



Zhu Zhuo



Radhika Khetani
Director of Education



Meeta Mistry



Will Gammerdinger



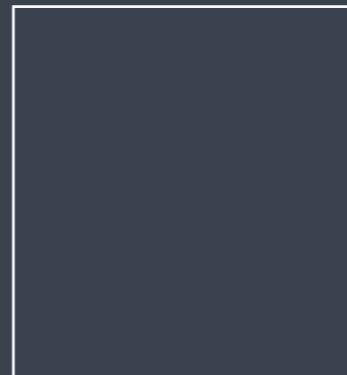
Emma Berdan



Sergey Naumenko



Maria Simoneau



Noor Sohail



James Billingsley

Consulting

- Experimental design help
- RNA-seq analysis: bulk, single cell, small RNA
- ChIP-seq and ATAC-seq analysis
- Genome-wide methylation
- WGS, resequencing, exome-seq and CNV studies
- QC & analysis of gene expression arrays
- Functional enrichment analysis
- Grant support

<http://bioinformatics.sph.harvard.edu/>



**HARVARD
T.H. CHAN**
SCHOOL OF PUBLIC HEALTH

NIEHS

CFAIR
HARVARD UNIVERSITY
CENTER FOR AIDS RESEARCH

 **HARVARD
CATALYST**
THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER

 **HARVARD**
MEDICAL SCHOOL

Training

A key component of the HBC's mission is its training initiative. Our dedicated training team holds workshop to help researchers at Harvard better understand analytical methods for NGS data.

HBC's training team is made up of four PhD-level scientists who devote substantial time to material development, training and community building/outreach. All members of the training team also participate in consultations on research projects to ensure they remain up-to-date on current best practices in NGS analysis.

Our hands-on workshops focus on **basic data skills** and **analysis of high-throughput sequencing data**, with an emphasis on **experimental design**, current **best practices** and **reproducibility**. Our workshops are designed for **wet-lab biologists** aiming to independently design sequencing-based experiments and analysing the resulting data.

We offer three types of workshops:

1. Short, 3-hour monthly workshops (*Current topics in bioinformatics*)
2. Basic Data Skills**
3. Advanced Topics: Analysis of high-throughput sequencing (NGS) data**

***The basic data skills workshops serve as the foundation for the advanced workshops.*

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

<https://hbctraining.github.io/main/>

Training

A key component of the HBC's mission is to provide training for researchers at Harvard and beyond.

HBC's training team is made up of experts in training and community building who work on research projects to ensure the quality of our training.

Our hands-on workshops focus on **bioinformatics**, with an emphasis on **experimental design** and **data analysis**. We also offer **wet-lab biologists** and **computational biologists** training in working with data.

We offer three types of workshops:

1. Short, 3-hour monthly workshops
2. Basic Data Skills**
3. Advanced Topics: Analysis of high-throughput sequencing (HTS) data

**The basic data skills workshop is designed for researchers with no prior experience in bioinformatics.



**HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH**

DF/HCC
DANA-FARBER / HARVARD CANCER CENTER



THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER



Our dedicated training team holds workshops to help researchers learn how to analyze high-throughput sequencing (HTS) or NGS data.

The training team also devote substantial time to material development, consulting, and outreach. Our training team also participate in consultations on best practices in NGS analysis.

Workshops focus on the analysis of high-throughput sequencing data, with an emphasis on **experimental design**, **data quality**, and **reproducibility**. Our workshops are designed to help researchers understand the principles of sequencing-based experiments and analysing the resulting data.

bioinformatics)

basic data skills (e.g., NGS) data**

and **advanced topics** (e.g., for the advanced workshops).

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

<https://hbctraining.github.io/main/>

Training

A key component of the HBC's mission is to support researchers at Harvard by providing training.

HBC's training team is made up of scientists who provide training and community building for research projects to ensure the quality of our work.

Our hands-on workshops focus on **bioinformatics**, with an emphasis on **experimental design** and **data analysis**. We also provide training for **wet-lab biologists** aiming to understand their data.

We offer three types of workshops:

1. Short, 3-hour monthly workshops
2. Basic Data Skills**
3. Advanced Topics: Analysis of high-throughput sequencing data

**The basic data skills workshop is designed for researchers with no prior experience in bioinformatics.



**HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH**

DF/HCC
DANA-FARBER / HARVARD CANCER CENTER



THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER



Our dedicated training team holds workshops to help researchers learn how to analyze **high-throughput sequencing (NGS) data**.

The training team also devote substantial time to material development, and our training team also participate in consultations on best practices in NGS analysis.

Workshops focus on the analysis of high-throughput sequencing data, with an emphasis on **experimental design**, **data quality**, and **reproducibility**. Our workshops are designed to help researchers understand the principles of sequencing-based experiments and analysing the resulting data.

bioinformatics)

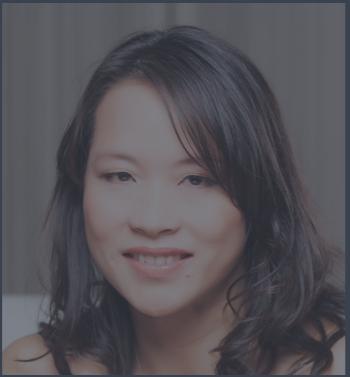
basic data skills (e.g., NGS) data**

and **advanced topics** (e.g., for the advanced workshops).

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

<https://hbctraining.github.io/main/>

Introductions!



Shannan Ho Sui
Director



Victor Barrera



Amelie Jule



Zhu Zhuo



Radhika Khetani
Director of Education



Meeta Mistry



Will Gammerdinger



Emma Berdan



Sergey Naumenko



Maria Simoneau



Noor Sohail

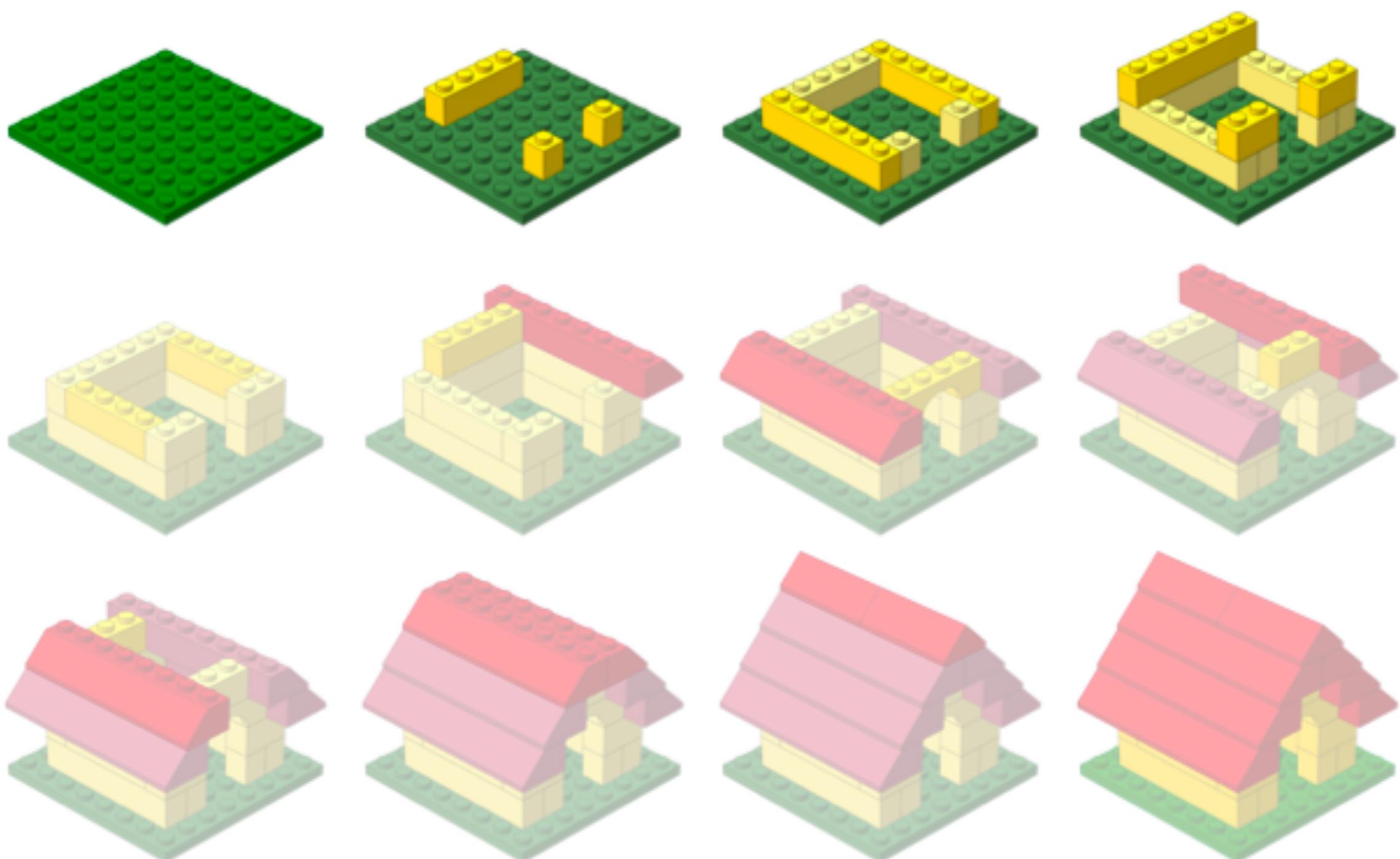


James Billingsley

Tell us a bit about yourselves!

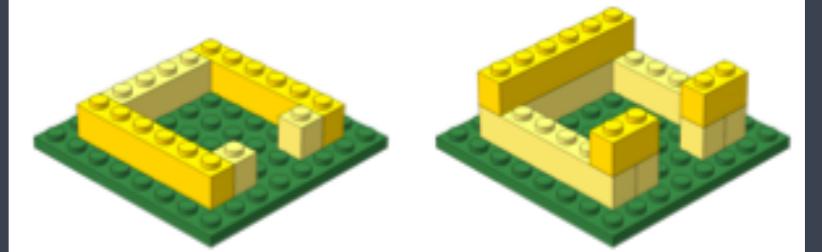
- How do you pronounce your name?
- How do you plan to use R?

Workshop Scope...



Learning R

Workshop Scope



- ✓ Comfortably use RStudio (a graphical interface for R)
- ✓ Fluently interact with R using RStudio
- ✓ Become familiar with R syntax
- ✓ Understand data structures in R
- ✓ Inspect and manipulate data structures
- ✓ Install packages and use functions in R

CRAN

(Comprehensive R Archive Network)



Available CRAN Packages By Name

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

A3	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
abc	Tools for Approximate Bayesian Computation (ABC)
ABCanalysis	Computed ABC Analysis
abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
abcdeFBA	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
ABCOptim	Implementation of Artificial Bee Colony (ABC) Optimization
ABCp2	Approximate Bayesian Computational Model for Estimating P2
abcrf	Approximate Bayesian Computation via Random Forests

*CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal*

- The main repository for R packages
- Easy to install

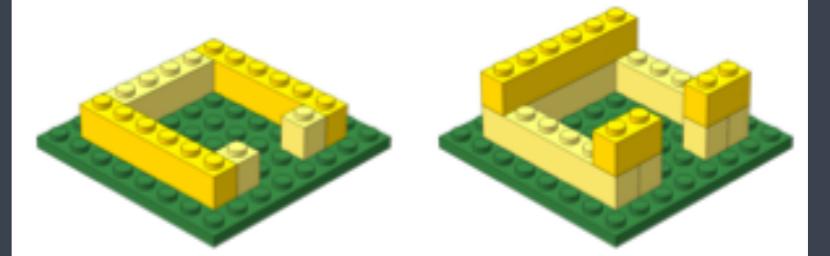
<https://cran.r-project.org/>



- An alternative package repository; “..provides tools for the analysis and comprehension of *high-throughput genomic data*.”
- Includes (but is not limited to) tools for:
 - performing statistical analysis
 - accessing public datasets
- Open source and open development
- Free

www.bioconductor.org

Workshop Scope



- Comfortably use RStudio (a graphical interface for R)
 - Fluently interact with R using RStudio
 - Become familiar with R syntax
 - Understand data structures in R
 - Inspect and manipulate data structures
 - Install packages and use functions in R
- ✓ Visualize data using *ggplot2*
- ✓ Utilize pipes, tibbles and functions from the Tidyverse package suite

Logistics

Course webpage

<https://tinyurl.com/hbc-r-online>

Course schedule online

Workshop Schedule

Day 1

Time	Topic	Instructor
10:00 - 10:30	Workshop Introduction	Jihe
10:30 - 11:45	Introduction to R and RStudio	Radhika
11:45 - 12:00	Overview of self-learning materials and homework submission	Mary

Before the next class:

1. Please **study the contents** and **work through all the code** within the following lessons:
 - o [R Syntax and Data Structure](#)
 - o [Functions and Arguments](#)
 - o [Reading in and inspecting data](#)
2. **Complete the exercises:**
 - o Each lesson above contain exercises; please go through each of them.
 - o **Copy over** your code from the exercises into a text file.
 - o **Upload the saved text file** to [Dropbox](#) the **day before the next class**.

Course materials online



Introduction to R

[View on GitHub](#)

Approximate time: 70 min

Learning Objectives

- Employ variables in R.
- Describe the various data types used in R.
- Construct data structures to store data.

The R syntax

Now that we know how to talk with R via the script editor or the console, we want to use R for something more than adding numbers. To do this, we need to know more about the R syntax.

Below is an example script highlighting the many different “parts of speech” for R (syntax):

- the **comments** `#` and how they are used to document function and its content
- **variables and functions**
- the **assignment operator** `<-`

The 2 Window problem...

The screenshot shows the RStudio interface. The top bar displays the path: ~/Dropbox (HBC)/HBC Team Folder (1)/Teaching/Intro-to-R - RStudio. The left pane contains an R script named "Intro-to-R.R" with the following code:

```
351
352 animals[4, 2] <- "Gray"
353
354 animals$color <- factor(animals$color)
355 animals$new2 <- c(1,2,3)
356
357 vector1 <- c(6:11)
358 data.frame(animals[, 1:2], vector1, animals[, 3:4])
359
360
362:1 (Top Level) ▾
```

The console window below shows the R environment:

```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/Dropbox (HBC)/HBC Team Folder (1)/Teaching/Intro-to-R/.RData]

>
```

The right pane shows the Global Environment and a file browser:

Name	Size	Modified
..		
.RData	5.8 MB	May 3, 2018, 1:40
.Rhistory	17.4 KB	Nov 15, 2018, 1:2
data		
de_sleuth.R	2.6 KB	Oct 10, 2018, 10:0
figures		
Intro-to-R.R	11.9 KB	May 1, 2018, 3:31

A callout box highlights the code in the script editor:

```
rownames(metadata)

metadata[c("sample10", "sample12"),]
```

The text "Selecting using indices with logical operators" is displayed below the highlighted code.

The explanatory text below states:

With dataframes, similar to vectors, we can use logical vectors for specific columns in the dataframe to select only the rows in a dataframe with TRUE values at the same position or index as in the logical vector. We can then use the logical vector to return all of the rows in a dataframe where those values are TRUE.

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!



Homework and Expectations

- ❖ 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

Odds and Ends

- ❖ Name tags
- ❖ Post-its
 - green - I am all set
 - red - I need time/help
- ❖ Quit/minimize all applications that are not required for class
- ❖ Phones on vibrate/silent
- ❖ Bathrooms

Contact us!

HBC training team: hbctraining@hsph.harvard.edu

HBC consulting: bioinformatics@hsph.harvard.edu