

NGS Data Analysis Course

Harvard Chan Bioinformatics Core

September - October 2018

<http://tinyurl.com/hbc/ngs-course-2018>



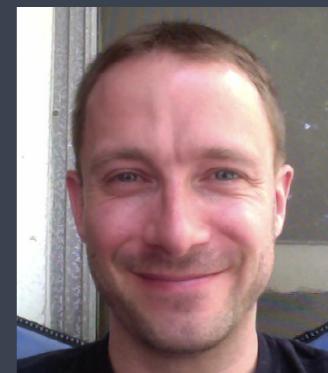
Shannan Ho Sui



John Hutchinson



Brad Chapman



Rory Kirchner



Meeta Mistry



Radhika Khetani



Mary Piper



Lorena Pantano



Kayleigh Rutherford



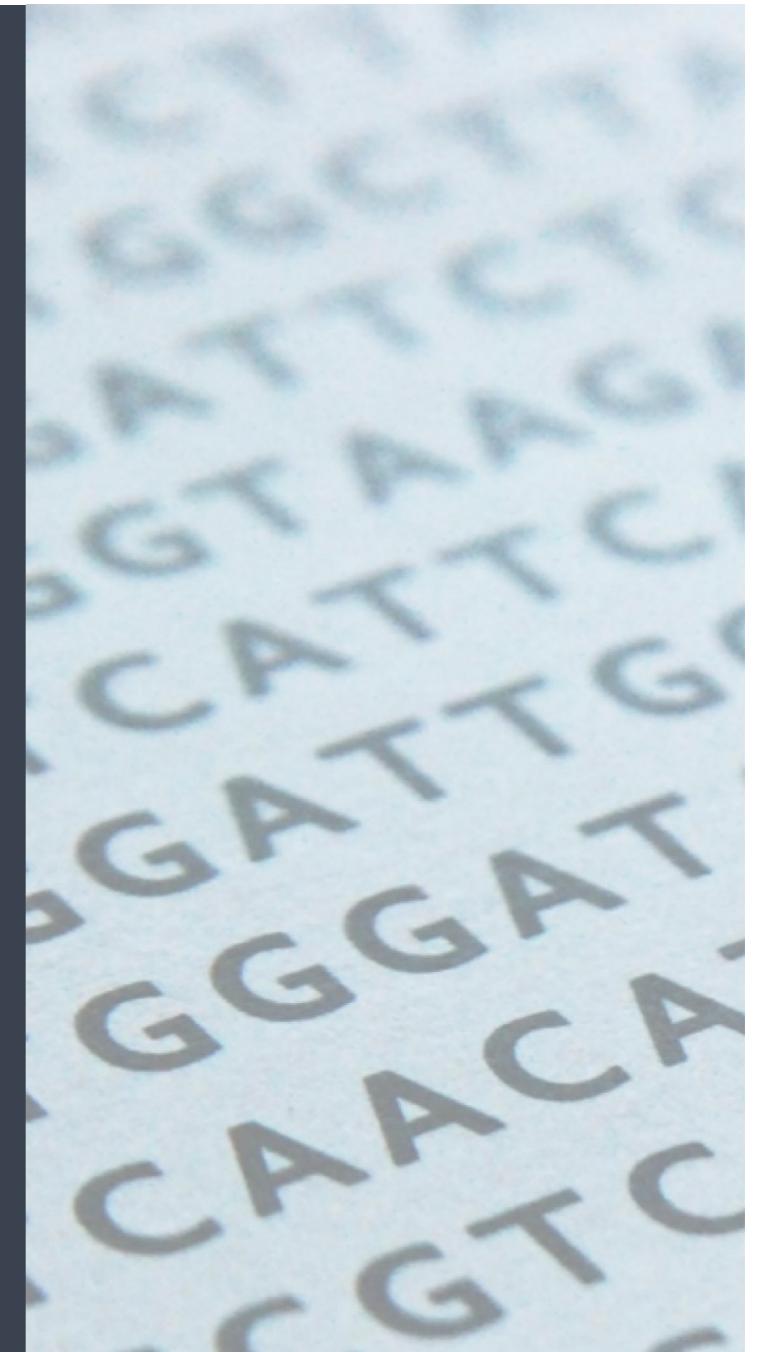
Victor Barrera



Peter Kraft

Consulting

- RNA-seq, small RNA-seq and ChIP-seq analysis
- Genome-wide methylation
- WGS, resequencing, exome-seq and CNV studies
- Quality assurance and analysis of gene expression arrays
- Functional enrichment analysis
- Grant support

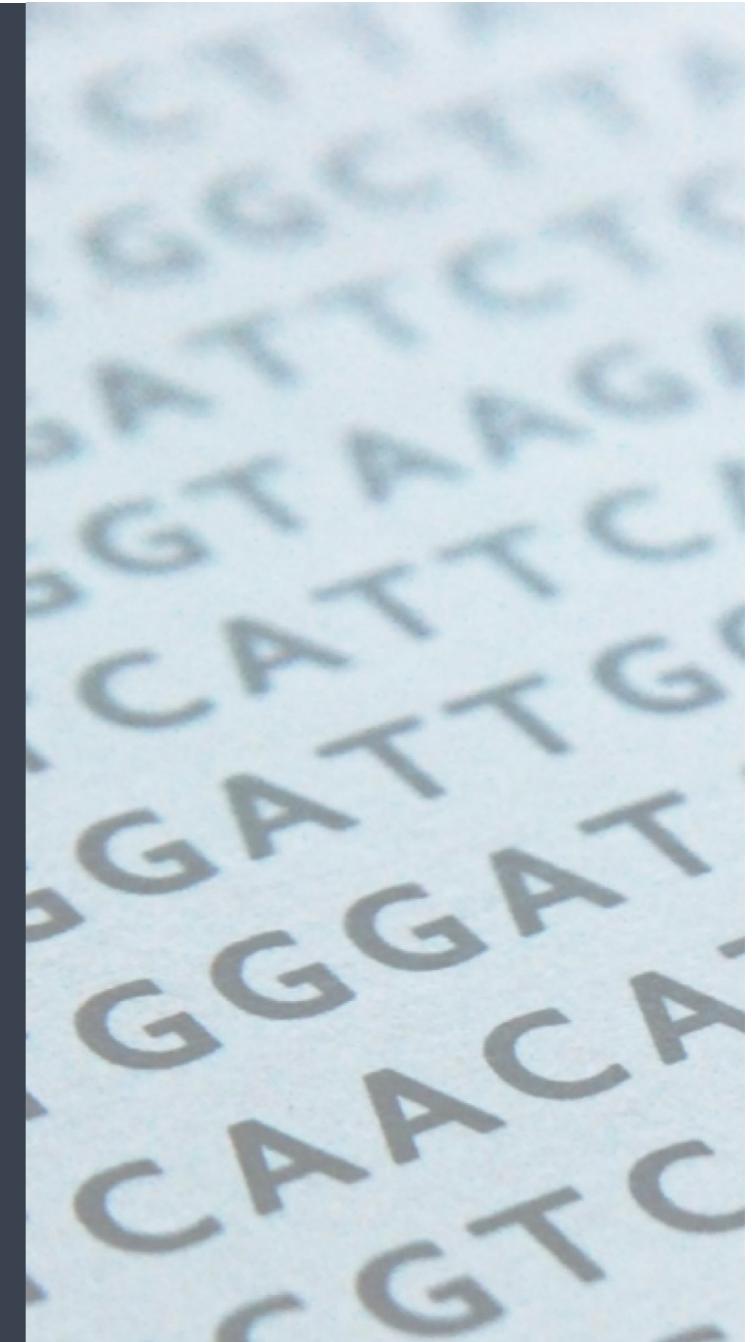


Training

- Short workshops on introductory, intermediate and advanced topics related to NGS data analysis
 - Monthly, 2-3 hour, hands-on and free workshops on “Current Topics in Bioinformatics”
 - In-depth courses (8- or 12-day formats)

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

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





HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NIEHS / CFAR
Bioinformatics
Core

HSCI
HARVARD STEM CELL
INSTITUTE

Center for Stem
Cell
Bioinformatics

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

 **HARVARD**
MEDICAL SCHOOL

Harvard
Catalyst
Bioinformatics
Consulting

HMS
Tools &
Technology



HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NIEHS / CFAR
Bioinformatics
Core

HSCI
HARVARD STEM CELL
INSTITUTE

Center for Stem
Cell
Bioinformatics

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

 **HARVARD**
MEDICAL SCHOOL

Harvard
Catalyst
Bioinformatics
Consulting

HMS
Tools &
Technology

Want more information?

Consulting email: bioinformatics@hsph.harvard.edu

Training email: hbctraining@hsph.harvard.edu

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

Twitter: @bioinfocore



Class introductions!





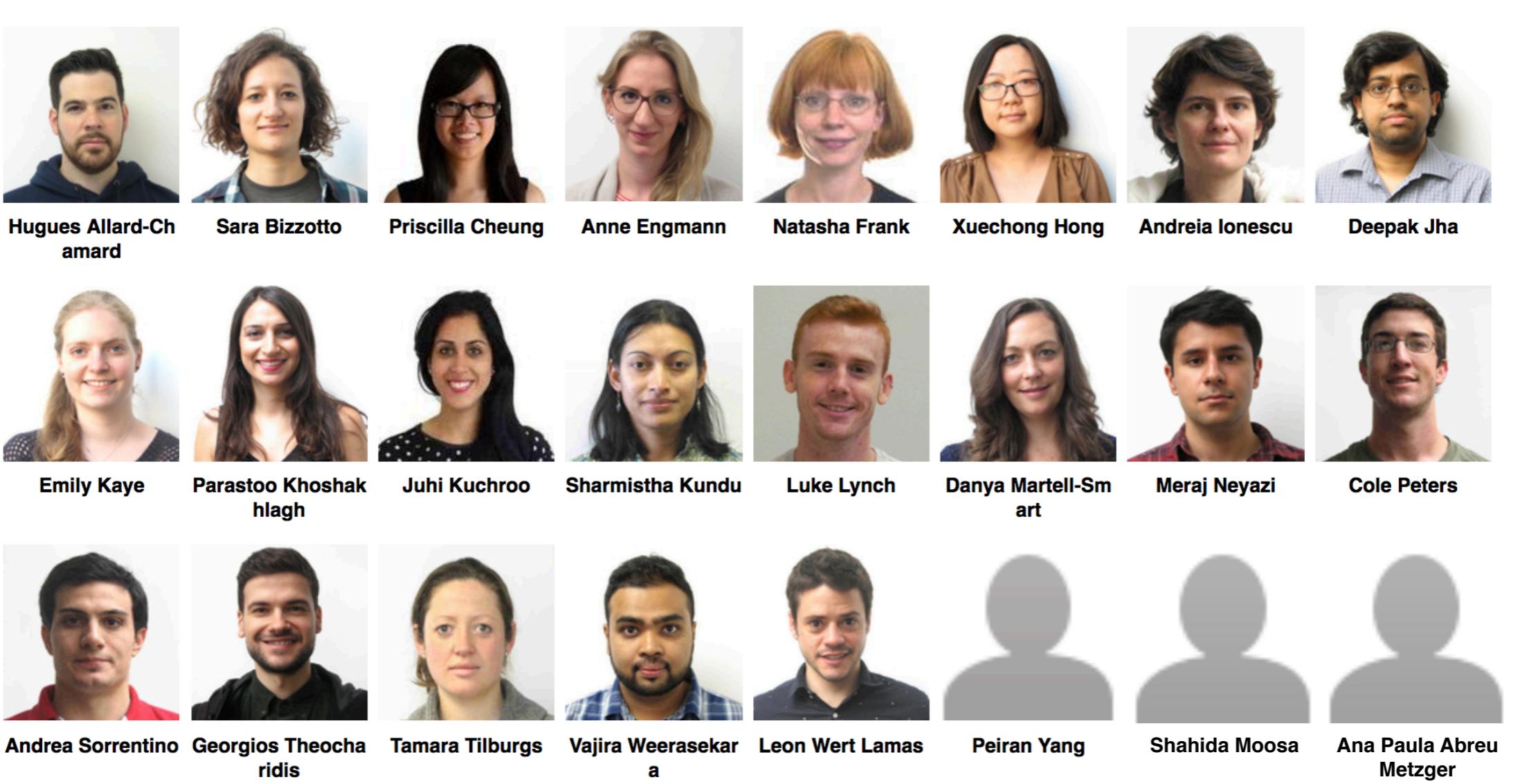
Meeta Mistry



Radhika Khetani



Mary Piper



NGS Data Analysis Course Participants

Course Scope

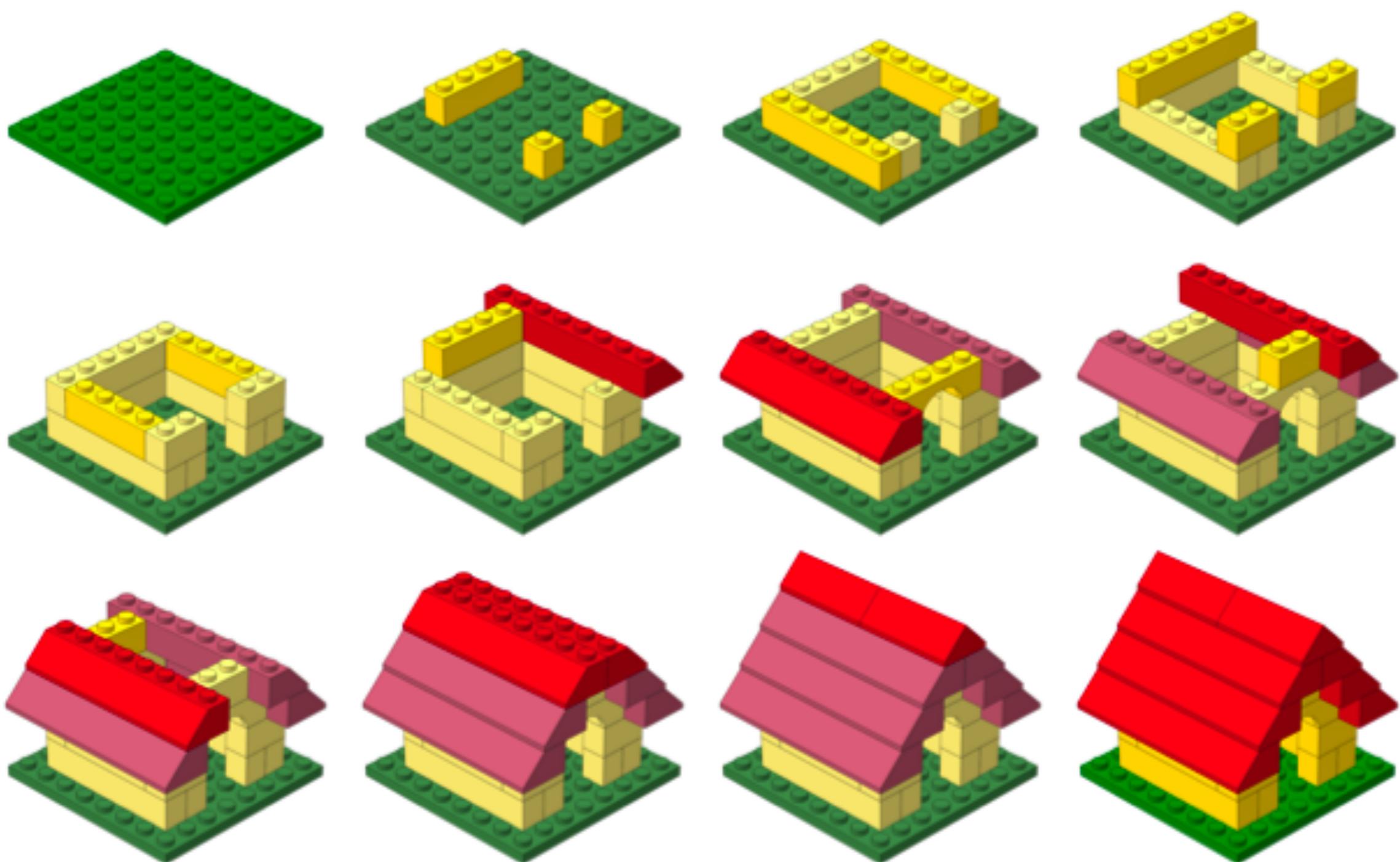


- ✓ Comprehend the nature of Next-Generation Sequencing (NGS) data
 - ◆ Multiple technologies
 - ◆ Caveats
 - ◆ Options and strategies
- ✓ Understand how tools and workflows for NGS-based analysis work
- ✓ Utilize these tools and workflows
 - ◆ Big data = Big Computational Requirements; what does this really entail?
 - ◆ UNIX command-line interface
 - ◆ R

Course Scope

- ✓ Implement best practices for the analysis of RNA-Seq (bulk and single cell), ChIP-Seq and Variant calling data
 - ◆ Experimental design
 - ◆ Quality control and Assessment
 - ◆ Reproducibility
- ✓ Become a resource for your group

Course Scope



Building Blocks

Course Logistics



Course webpage

HBC_LongCourse

2018-2019 Fall 1

In-Depth NGS Data Analysis

[View Course Stream](#)

Coming Up [View Calendar](#)
Nothing for the next week

Home Announcements Assignments Discussions People


HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH
Powerful ideas for a healthier world

HBC2018
In-depth NGS Data Analysis

Course Information Teaching Staff Getting Started

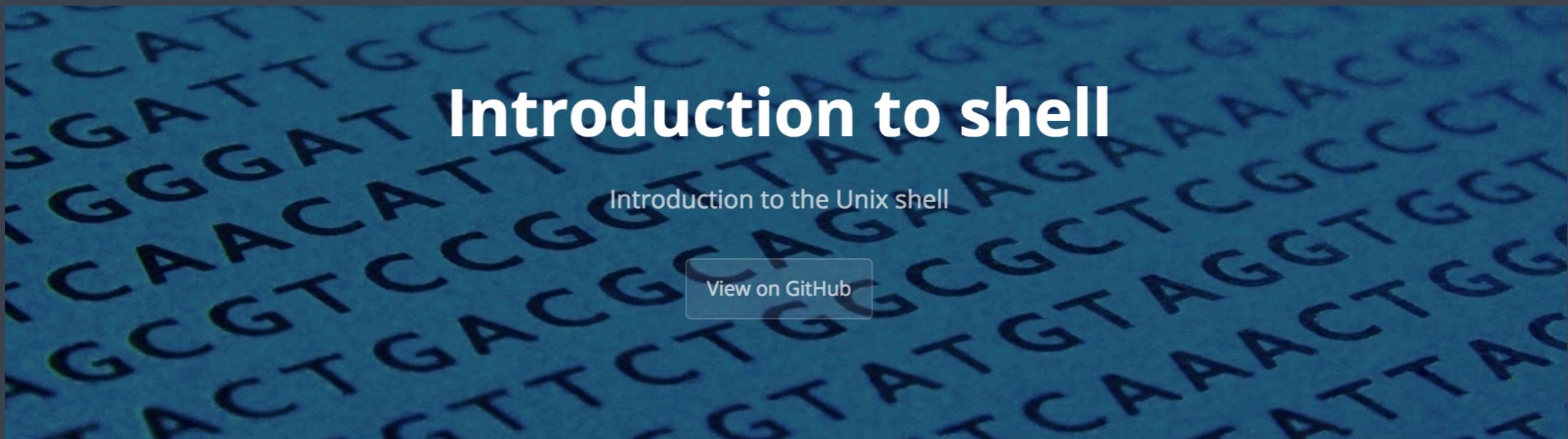
Weekly Course Material

Week 1 Week 2 Week 3 Week 4

Week 5 Week 6

<http://tinyurl.com/hbc/ngs-course-2018>

Course materials online



Introduction to shell

Introduction to the Unix shell

[View on GitHub](#)

Learning Objectives

- How do you access the shell?
- How do you use it?
 - Getting around the Unix file system
 - looking at files
 - manipulating files
 - automating tasks
- What is it good for?

Setting up

We will spend most of our time learning about the basics of the shell by exploring experimental data.

Since we are going to be working with this data on our remote server, **Orchestra 2 (O2)**, we first need to log onto the server. After we're logged on, we will each make our own copy of the example data folder.

Solution1 to the 2 Window problem

rsk27@clarinet002-072:~\$ ls -lht unix_workshop/
total 177K
drwxrwsr-x 2 rsk27 rsk27 62 May 23 2016 reference_data
-rw-rw-r-- 1 rsk27 rsk27 377 May 23 2016 README.txt
drwxrwsr-x 2 rsk27 rsk27 78 May 23 2016 genomics_data
drwxrwsr-x 2 rsk27 rsk27 257 May 23 2016 raw_fastq
drwxrwsr-x 2 rsk27 rsk27 695 May 23 2016 other
drwxrwsr-x 6 rsk27 rsk27 972 May 24 2016 rnaseq_project
rsk27@clarinet002-072:~\$

Intro-to-Unix/01_the_filesystem

GitHub, Inc. [US] https://github.com/hbctraining/Intro-t... Radhika

Starting with the shell

We have each created our own copy of the example data folder into our home directory, **unix_workshop**. Let's go into the data folder and explore the data using the shell.

\$ cd unix_workshop

'cd' stands for 'change directory'

Let's see what is in here. Type:

\$ ls

You will see:

genomics_data other raw_fastq README.txt reference_data

ls stands for 'list' and it lists the contents of a directory.

There are five items listed. What are they? We can use a "modifier" with `ls` to get more information; this modifier is called an argument (more below).

\$ ls -F

genomics_data/ other/ raw_fastq/ README.txt reference_data/

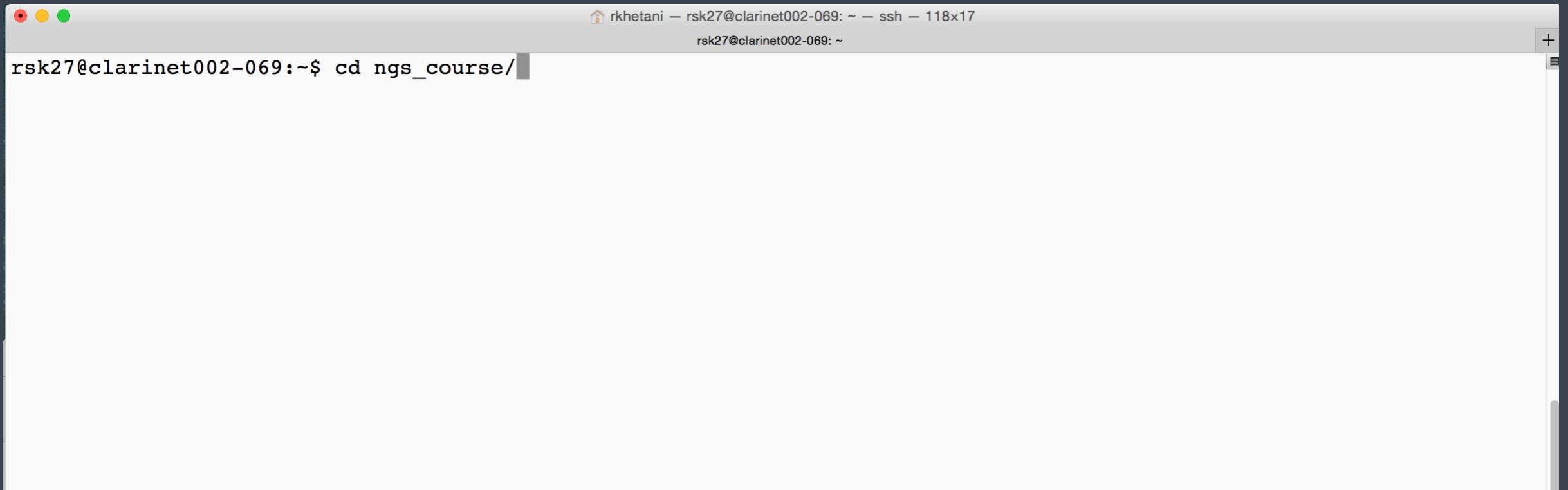
Anything with a "/" after it is a directory. Things with a "*" after them are programs. If there are no decorations after the name, it's a file.

All commands are essentially programs that are able to perform specific, commonly-used tasks.

You can also use the command:

\$ ls -l

Solution2 to the 2 Window problem

A screenshot of a macOS terminal window. The title bar shows "rkhetani — rsk27@clarinet002-069: ~ — ssh — 118x17". The main pane contains the command "rsk27@clarinet002-069:~\$ cd ngs_course/".

Let's go into the data folder and explore the data using the shell. We will do this sequentially using the `cd` command. `cd` stands for "change directory".

```
$ cd ngs_course
```

```
$ cd unix_lesson
```

Let's see what is in the `unix_lesson` directory/folder. Type:

```
$ ls
```

You will see:

```
genomics_data  raw_fastq  README.txt  reference_data
```

Odds and Ends

- ✿ Name tags: Tent Cards
 - ✿ Post-its
 - ✿ Wi-Fi: HMS Public if nothing else works
 - ✿ Lunch locations
 - ✿ Phones on vibrate/silent!
 - ✿ Bathrooms
 - ✿ Water Fountains



Thanks!

- Kristina Holton and Andy Bergman from HMS-RC
- [Data Carpentry](#)

These materials have been developed by members of the teaching team at the [Harvard Chan Bioinformatics Core \(HBC\)](#). These are open access materials distributed under the terms of the [Creative Commons Attribution license \(CC BY 4.0\)](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

