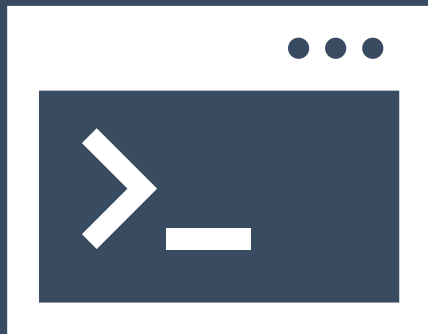


Intro to Bulk RNA-seq (Part I)

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



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



Alex Bartlett



Elizabeth
Partan



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



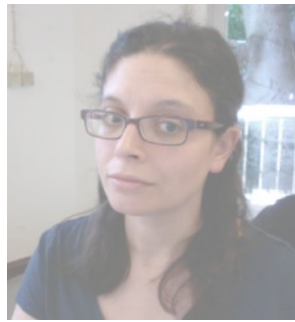
Noor Sohail



Alex Bartlett



Elizabeth
Partan



Emma Berdan



James Billingsley



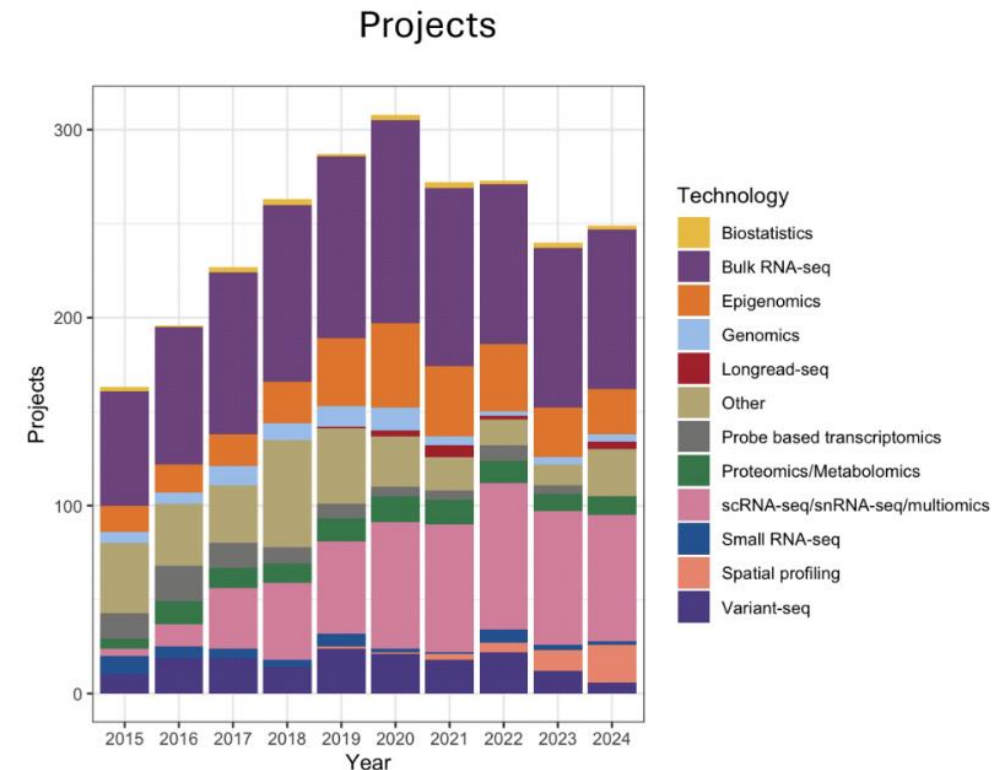
Zhu Zhuo



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

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.*

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

Training

A key component of the HBC's mission is its training initiative. Our dedicated training team holds workshops for 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 consulting 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** 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 sequencing 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.*

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



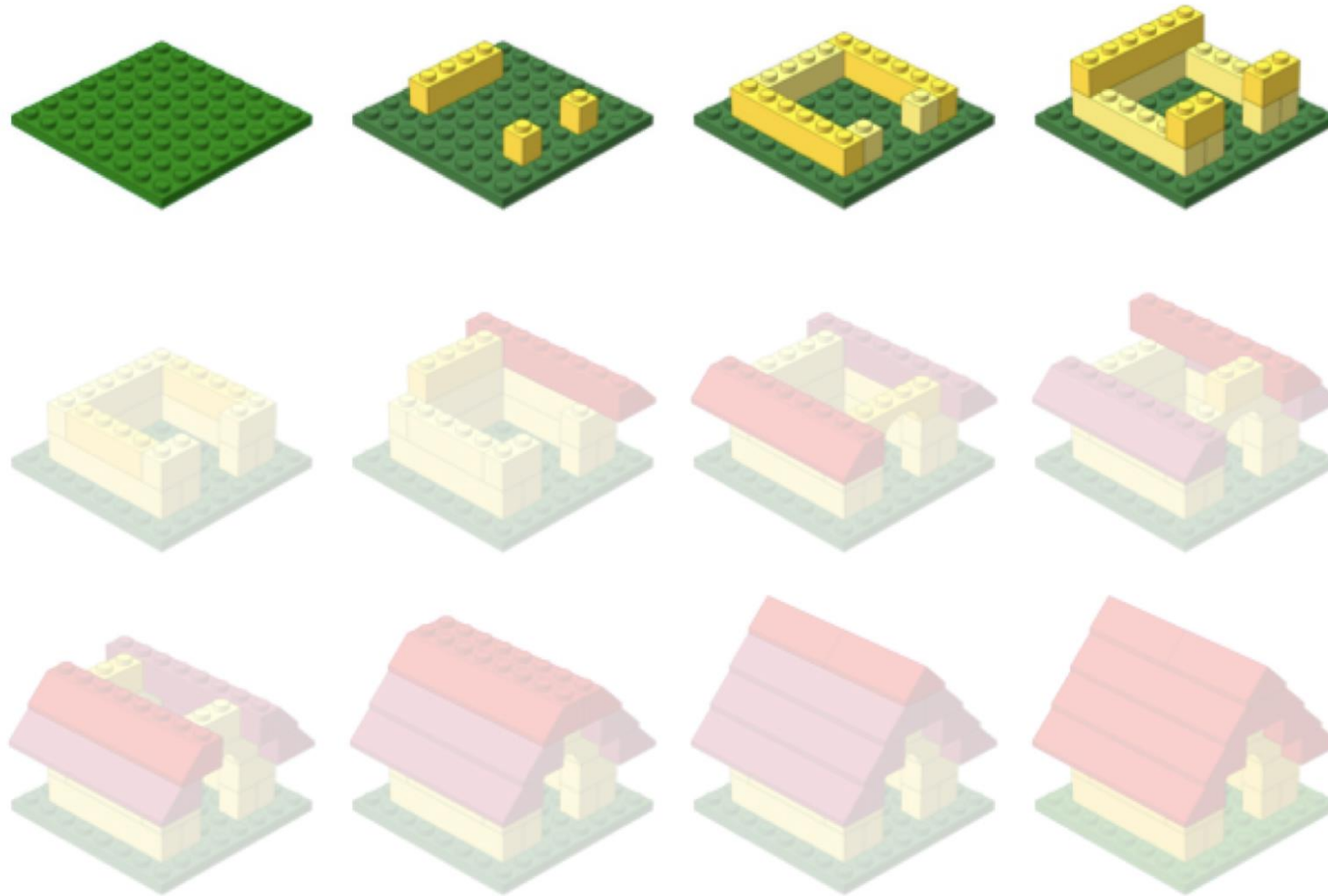
HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

DF/HCC
DANA-FARBER / HARVARD CANCER CENTER



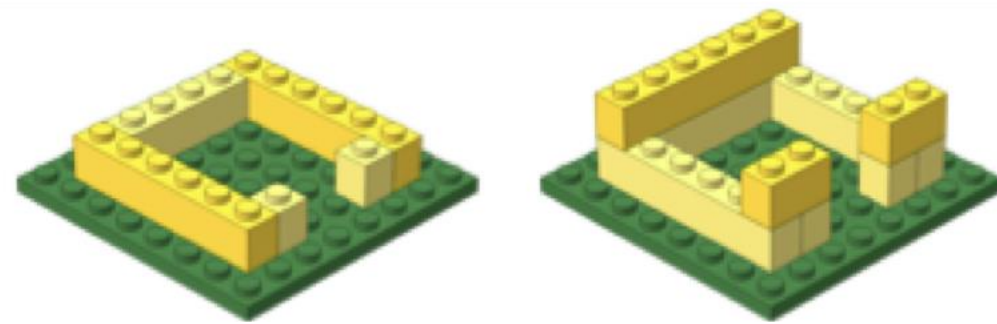
THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER





Learning Bioinformatics

Setting up



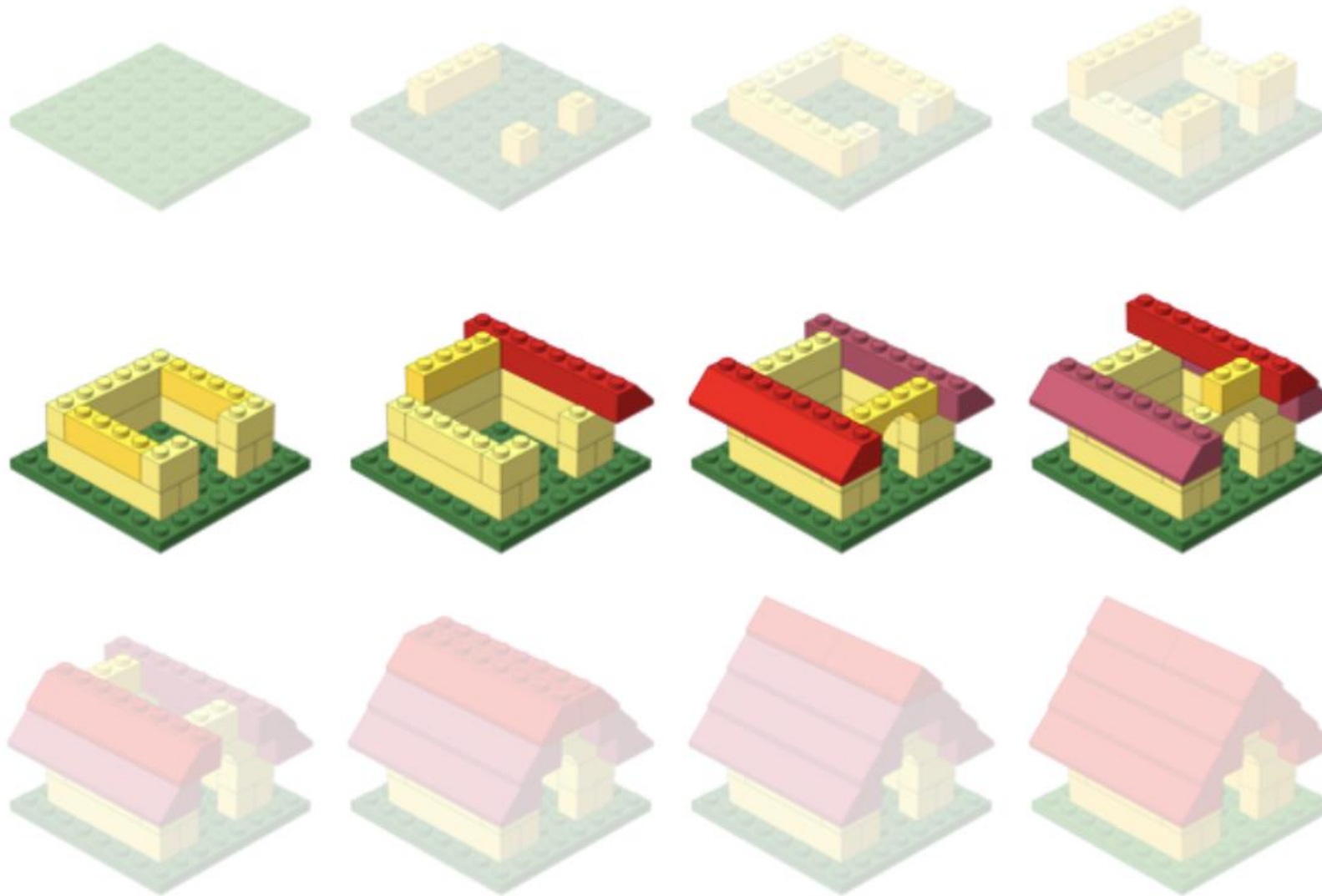
❖ Shell for Bioinformatics

- ❖ Dealing with large data files
- ❖ Performing Bioinformatic Analyses
 - ❖ Using tools
 - ❖ Accessing and using computer clusters

❖ Introduction to R

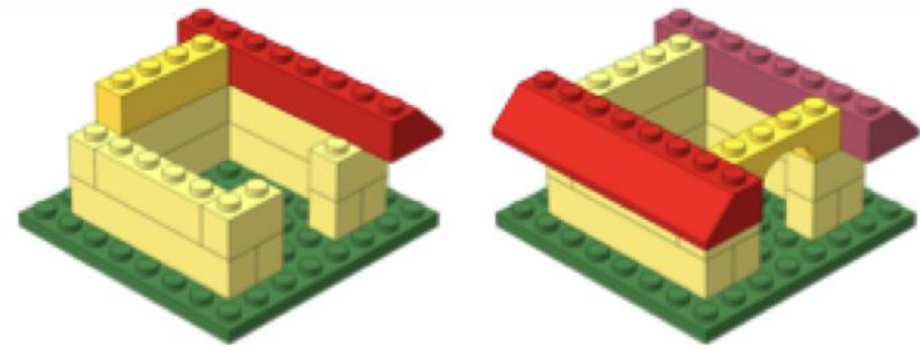
- ❖ Parsing and working with smaller results text files
- ❖ Statistical analyses, e.g. differential expression analysis
- ❖ Generating figures from complex data

Workshop scope



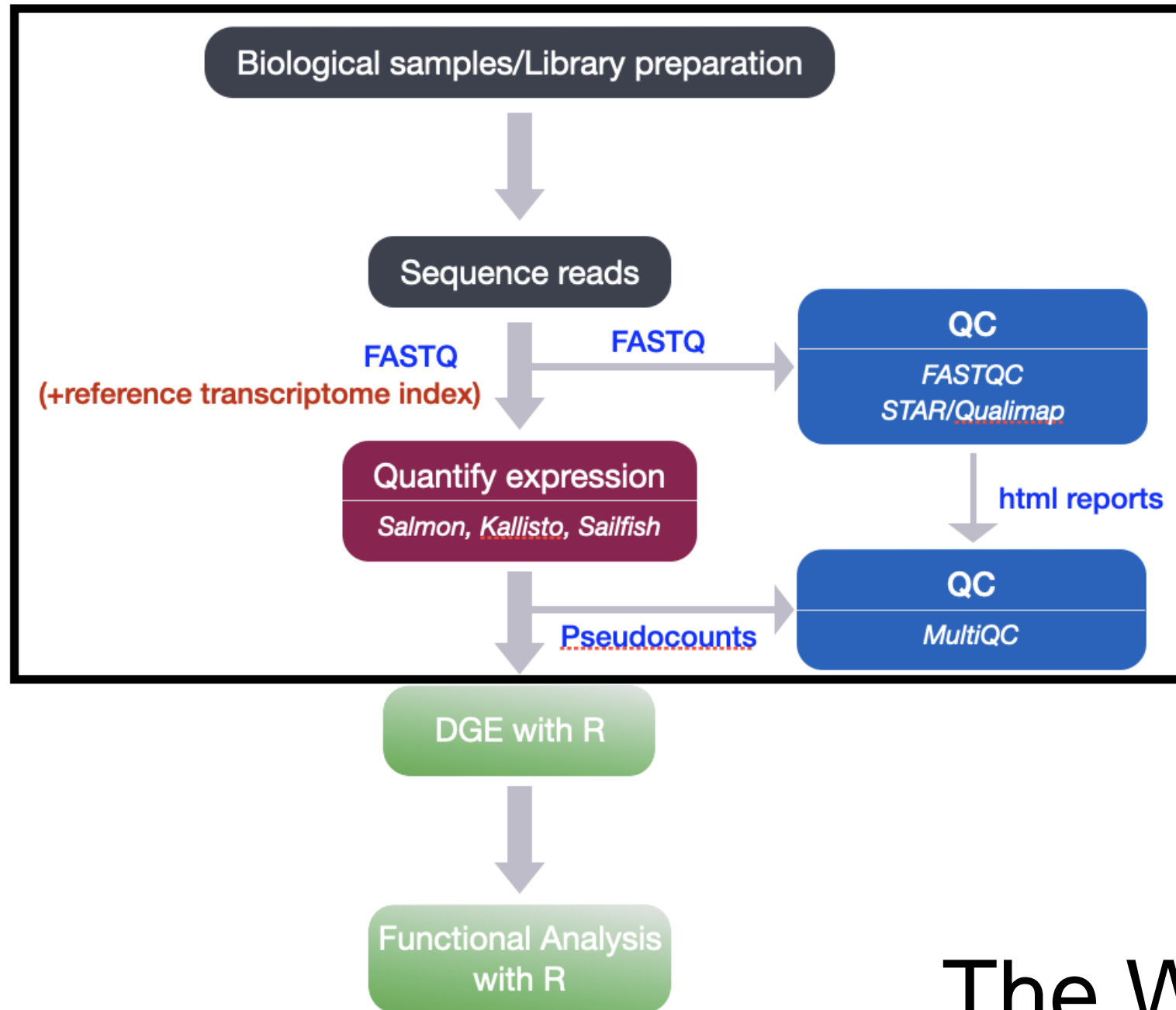
Bioinformatics Data Analysis

Learning Objectives



- ❖ Describe best practices for designing a bulk RNA-seq experiment
- ❖ Describe steps in an RNA-seq analysis workflow (from sequence data to expression quantification)
- ❖ Implement shell scripts on a high-performance compute cluster to perform the above steps

We won't be covering how to perform differential gene expression (DGE) analysis on count data in this workshop.



The Workflow

Logistics



Course Webpage

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

Course schedule

Day 1

Time	Topic	Instructor
09:30 - 09:45	Workshop Introduction	Will
09:45 - 10:25	Working in an HPC environment - Review	Open
10:25 - 11:05	Project Organization (using Data Management best practices)	Will
11:05 - 11:45	Quality Control of Sequence Data: Running FASTQC	Open
11:45 - 12:00	Overview of self-learning materials and homework submission	Will

Before the next class:

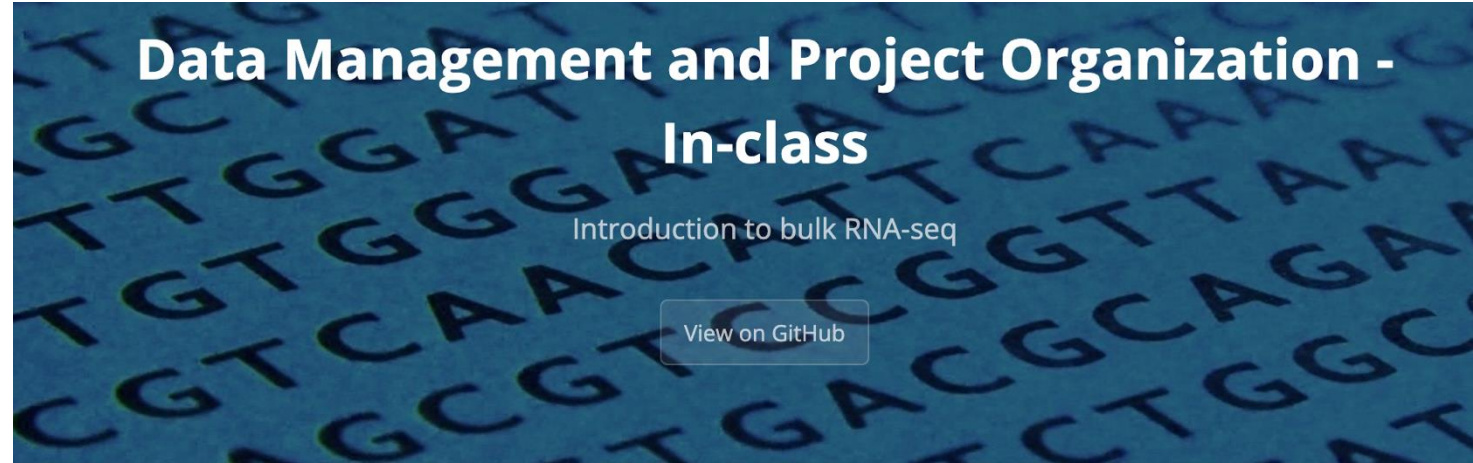
1. Please **study the contents** and **work through all the code** within the following lessons:

- Experimental design considerations
- Quality Control of Sequence Data: Running FASTQC on multiple samples
- Quality Control of Sequence Data: Evaluating FASTQC reports

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

Course materials

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



Learning Objectives

- Describe the example RNA-seq experiment and its objectives.
- Demonstrate strategies for good data management and project organization.

The Dataset

The dataset we are using for this workshop is part of a larger study described in [Kenny PJ et al., *Cell Rep* 2014](#). The authors are investigating interactions between various genes involved in Fragile X syndrome, a disease of aberrant protein production, which results in cognitive impairment and autistic-like features. **The authors sought to show that RNA helicase MOV10 regulates the translation of RNAs involved**

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

Single Screen & 3 Windows

The screenshot displays a Zoom meeting interface with a single screen showing three overlapping windows:

- Video Window:** Shows three participants: Mary Piper, Troubleshooter (...), and Jihe Liu. The status bar at the bottom includes controls for Unmute, Stop Video, Invite, Share Screen, and Reactions.
- Participants Window:** Lists three participants: Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host).
- Terminal Window:** Shows a shell session on a MacBook Pro. The terminal output includes a directory listing of files and a command execution:

```
rsk27@clarinet002-0721:~$ ll -ltr unix_workshop/
total 177K
drwxrwxr-x 2 rsk27 rsk27 62 May 23 2016 reference_data
-rw-rw-r-- 1 rsk27 rsk27 377 May 23 2016 README.txt
drwxrwxr-x 2 rsk27 rsk27 78 May 23 2016 genomics_data
drwxrwxr-x 2 rsk27 rsk27 257 May 23 2016 raw_fastq
drwxrwxr-x 2 rsk27 rsk27 495 May 23 2016 other
drwxrwxr-x 6 rsk27 rsk27 372 May 24 2016 enaseq_project
rsk27@clarinet002-0721:~$

HSPH-Radhikas-MacBook-Pro:~ rsk394$ cut -f 1,4 /Users/rsk394/Dropbox\
\ (Harvard\ University)\ /HBC\ Team\ Folder\ \ (1\)/Teaching/Courses/pr
e-2019/Galaxy_nanocourses/Data_from_old_instance/RNA-Seq/Sequence\ an
d\ reference\ data/chr1-hg19_genes.gtf | sort -k2n | head
chr1    14362
chr1    14970
chr1    15796
chr1    16607
chr1    16858
chr1    17233
chr1    17606
chr1    17915
chr1    18268
chr1    24738
HSPH-Radhikas-MacBook-Pro:~ rsk394$
```

Single Screen & 3 Windows

The image shows a Zoom meeting interface. At the top, there are three video thumbnails for participants: Mary Piper, Troubleshooter (...), and Jihe Liu. Below the thumbnails is a terminal window with a shell prompt. The terminal shows a command to run a script, followed by a list of files and their sizes. The output of the command is a list of chromosome coordinates. The Zoom interface also shows a 'Participants (3)' list on the right side.

Zoom

Starting with the shell

We have each created our own copy of the example data folder into our home directory, `unix_w` 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
```

```
HSPH-Radhikas-MacBook-Pro:~ rsk394$ cut -f 1,4 /Users/rsk394/Dropbox/\
(Harvard\ University\)/HBC\ Team\ Folder\ \ (1\)/Teaching/Courses/pr
e-2019/Galaxy_nanocourses/Data_from_old_instance/RNA-Seq/Sequence\ an
d\ reference\ data/chr1-hg19_genes.txt | sort -k2n | head
```

```
chr1    14362
chr1    14970
chr1    15796
chr1    16607
chr1    16858
chr1    17233
chr1    17606
chr1    17915
chr1    18268
chr1    24738
```

HSPH-Radhikas-MacBook-Pro:~ rsk394\$

Introduction to the command line interface (shell)

View on GitHub

*Our
Recommendation*

Single Screen & 3 Windows

The screenshot displays a Zoom meeting interface with three windows open:

- Terminal Window:** Shows a shell prompt with a directory listing of files including `reference_data`, `README.txt`, `genomics_data`, `raw_fastq`, `other`, and `cnaseq_project`.
- Participant List:** Lists three participants: Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host).
- Web Browser Window:** Displays a slide titled "Introduction to the command line interface (shell)" with the text "Web Browser" overlaid in green.

Below the Zoom interface, a terminal window shows the execution of a command to cut and sort data:

```
HSPH-Radhikas-MacBook-Pro:~ rsk394$ cut -f 1,4 /Users/rsk394/Dropbox/\(Harvard\ University\) /HBC\ Team\ Folder\ \(1\) /Teaching/Courses/pr e-2019/Galaxy_n a courses/Data_from_old_instance/RNA-Seq/Sequence\ an d\ reference\ da a/chr1-hg19_genes.gtf | sort -k2n | head
```

```
chr1 14362
chr1 14970
chr1 15796
chr1 16607
chr1 16858
chr1 17233
chr1 17606
chr1 17915
chr1 18268
chr1 24738
```

The terminal prompt is `HSPH-Radhikas-MacBook-Pro:~ rsk394$`.

*Our
Recommendation*

Single Screen & 3 Windows

The image shows a Zoom meeting interface with three windows. The top window is a video call with three participants: Mary Piper, Troubleshooter (...), and Jihe Liu. The middle window is a participants list showing Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host). The right window is a browser window displaying a tutorial titled "Introduction to the command line interface (shell)" with a "View on GitHub" button. The bottom window is a terminal window with the following command and output:

```
HSPH-Radhikas-MacBook-Pro:~ rsk394$ cut -f 1,4 /Users/rsk394/Dropbox\
\ (Harvard\ University)\ /HBC\ Team\ Folder\ \ (1\)/Teaching/Courses/pr
e-2019/Galaxy_nanocourses/Data_from_old_instance/RNA-Seq/Sequence\ an
d\ reference\ data/chr1-hg19_genes.gtf | sort -k2n | head
chr1    14362
chr1    14970
chr1    15796
chr1    16607
chr1    16858
chr1    17233
chr1    17606
chr1    17915
chr1    18268
chr1    24738
HSPH-Radhikas-MacBook-Pro:~ rsk394$
```

*Our
Recommendation*

Terminal

Single Screen & 3 Windows

The image shows a Zoom meeting interface with three windows highlighted by colored boxes:

- Zoom (Blue box):** Shows the Zoom meeting controls and a terminal window. The terminal output is as follows:

```
mac27@clarinet002-0721:~$ ll -ltr unix_workshop/
total 177K
-rwxrwxr-x 2 rsk27 rsk27 62 May 23 2016 reference_data
-rw-rw-r-- 1 rsk27 rsk27 377 May 23 2016 README.txt
-rwxrwxr-x 2 rsk27 rsk27 78 May 23 2016 genomics_data
-rwxrwxr-x 2 rsk27 rsk27 257 May 23 2016 raw_fastq
-rwxrwxr-x 2 rsk27 rsk27 495 May 23 2016 other
-rwxrwxr-x 6 rsk27 rsk27 372 May 24 2016 enaseq_project
mac27@clarinet002-0721:~$
```
- Web Browser (Green box):** Shows a browser window with the URL `ng.github.io/intro-to-shell-flipped/lessons/01_the_filesystem.html`. The page content includes the text "Introduction to the command line interface (shell)" and "Web Browser".
- Terminal (Purple box):** Shows a terminal window with the following command and output:

```
HSPH-Radhikas-MacBook-Pro:~ rsk394$ cut -f 1,4 /Users/rsk394/Dropbox\
\ (Harvard\ University\)/HBC\ Team\ Folder\ \ (1\)/Teaching/Courses/pr
e-2019/Galaxy_nacourses/Data_from_old_instance/RNA-Seq/Sequence\ and
d\ reference\ data/chr1-hg19_genes.txt | sort -k2n | head
chr1 14362
chr1 14970
chr1 15796
chr1 16607
chr1 16858
chr1 17233
chr1 17606
chr1 17915
chr1 18268
chr1 24738
HSPH-Radhikas-MacBook-Pro:~ rsk394$
```

*Our
Recommendation*

Terminal

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

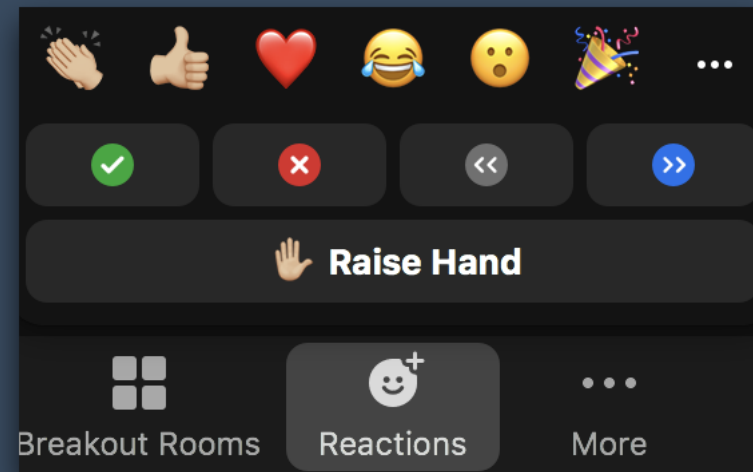
Odds & Ends

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

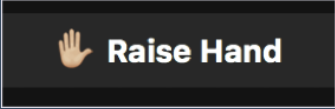
❖ Are you all set?

❖  = "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 Troubleshooter know

Odds & Ends

❖ Questions for the presenter?

❖ Post the question in the Chat window OR

❖  when the presenter asks for questions

❖ Let the Troubleshooter know

❖ Technical difficulties with software?

❖ Start a private chat with the Troubleshooter with a description of the problem

Thanks!

- ❖ Kathleen Chappell 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.

Contact Us

- ❖ *HBC training team:* hbctraining@hsph.harvard.edu
- ❖ *HBC consulting:* bioinformatics@hsph.harvard.edu
- ❖ *O2 (HMS-RC):* rchelp@hms.harvard.edu