

# Introduction to bulk RNA-seq (Part I)

Harvard Chan Bioinformatics Core

in collaboration with

FAS Research Computing

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





Shannan Ho Sui  
*Director*



John Hutchinson  
*Associate Director*



Victor Barrera



Zhu Zhuo



Preetida Bhetariya



Radhika Khetani  
*Training Director*



Meeta Mistry



Mary Piper  
*Assoc. Training Director*



Jihe Liu



Will Gammerdinger



Maria Simoneau



James Billingsley



Sergey Naumenko



Peter Kraft  
*Faculty Advisor*





Shannan Ho Sui  
*Director*



John Hutchinson  
*Associate Director*



Victor Barrera



Zhu Zhuo



Preetida Bhetariya



Radhika Khetani  
*Training Director*



Meeta Mistry



Mary Piper  
*Assoc. Training Director*



Jihe Liu



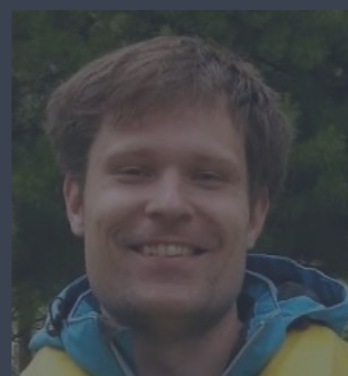
Will Gammerdinger



Maria Simoneau



James Billingsley



Sergey Naumenko



Peter Kraft  
*Faculty Advisor*

# Consulting

- 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



THE HARVARD CLINICAL  
AND TRANSLATIONAL  
SCIENCE CENTER



**HARVARD**  
MEDICAL SCHOOL



# Training

We have divided our short workshops into 2 categories:

1. Basic Data Skills - No prior programming knowledge needed (no prerequisites)
2. Advanced Topics: Analysis of high-throughput sequencing (NGS) data - Certain “Basic” workshops required as prerequisites.

*Any participants wanting to take an advanced workshop will have to have taken the appropriate basic workshop(s) within the past 6 months.*

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

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



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

**DF/HCC**

DANA-FARBER / HARVARD CANCER CENTER

**HSCI**  
HARVARD STEM CELL  
INSTITUTE



THE HARVARD CLINICAL  
AND TRANSLATIONAL  
SCIENCE CENTER



**HARVARD**  
MEDICAL SCHOOL

# Training

We have divided our short workshops into 2 categories:

1. Basic Data Skills - No prior programming knowledge needed (no prerequisites)
2. Advanced Topics: Analysis of high-throughput sequencing (NGS) data - Certain “Basic” workshops required as prerequisites.

*Any participants wanting to take an advanced workshop will have to have taken the appropriate basic workshop(s) within the past 6 months.*

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

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



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

DF/HCC

DANA-FARBER / HARVARD CANCER CENTER

HSCI  
HARVARD STEM CELL  
INSTITUTE

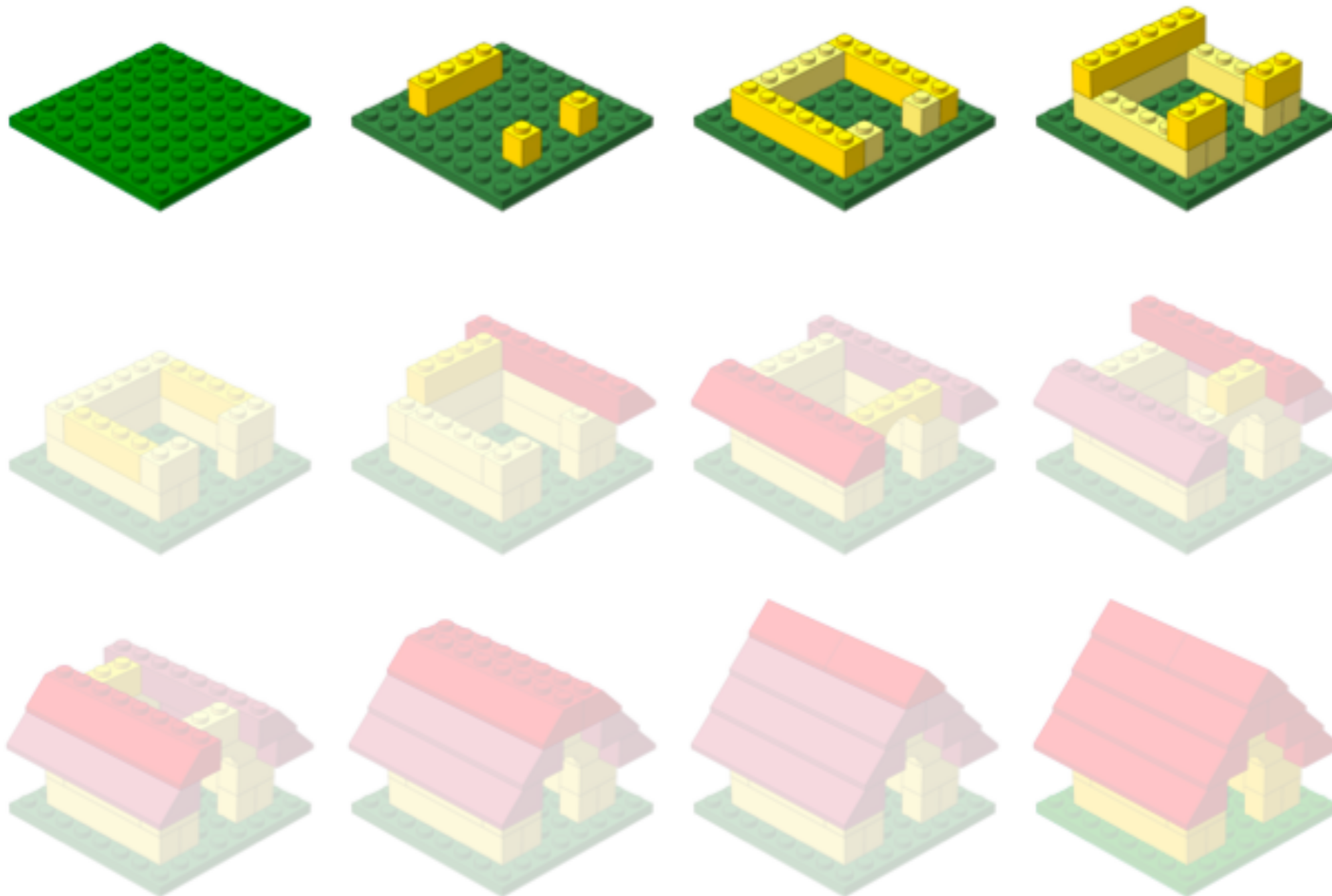


THE HARVARD CLINICAL  
AND TRANSLATIONAL  
SCIENCE CENTER



**HARVARD**  
MEDICAL SCHOOL





<http://anoved.net/tag/lego/page/3/>

Setting up to perform Bioinformatics analysis

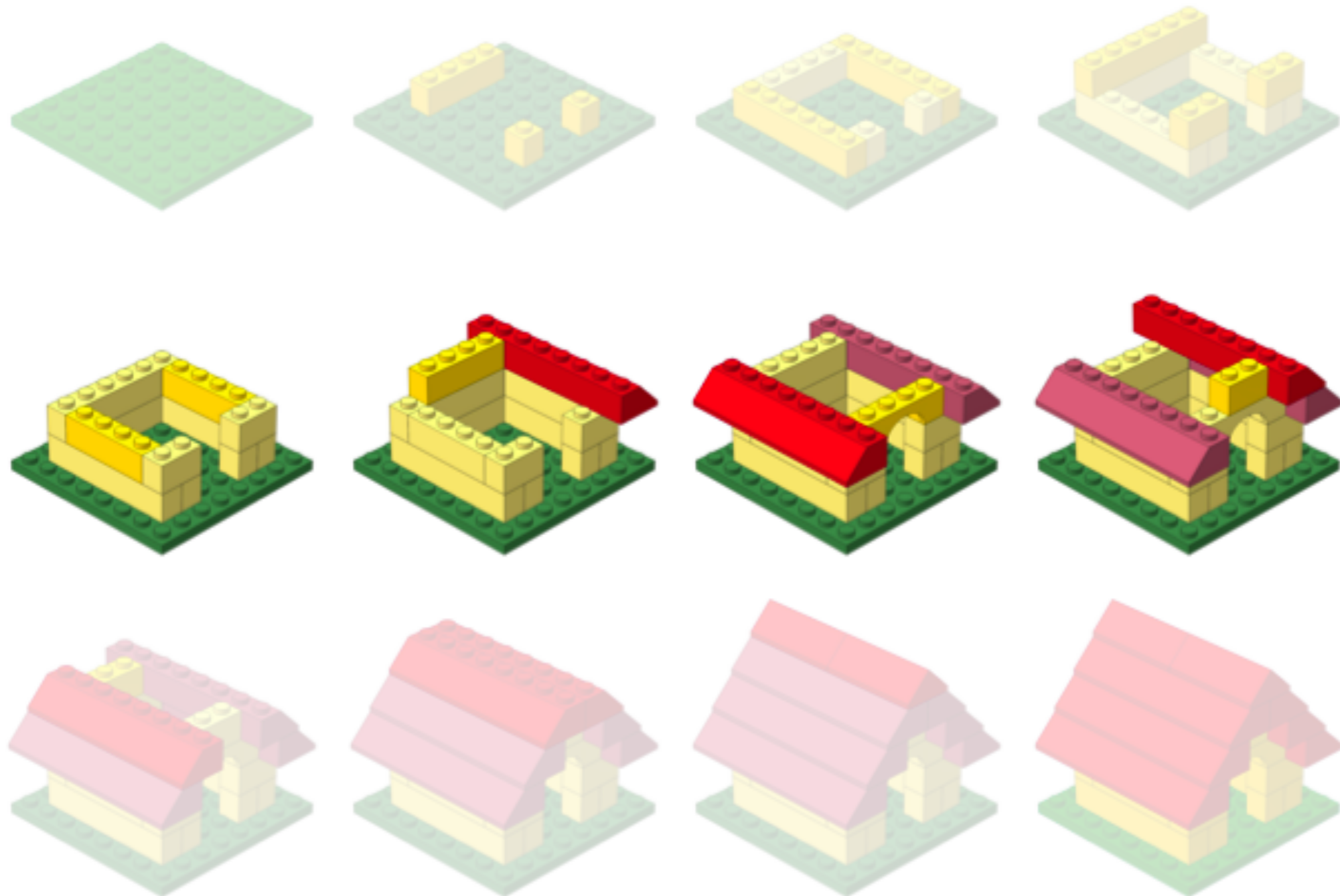
# Setting up...



- ✓ Introduction to the command-line interface (shell, Unix, Linux)
  - Dealing with large data files
  - Performing bioinformatics analysis
    - Using tools
    - Accessing and using compute clusters
- ✓ R
  - Parsing and working with smaller results text files
  - Statistical analysis, e.g. differential expression analysis
  - Generating figures from complex data



# Workshop scope

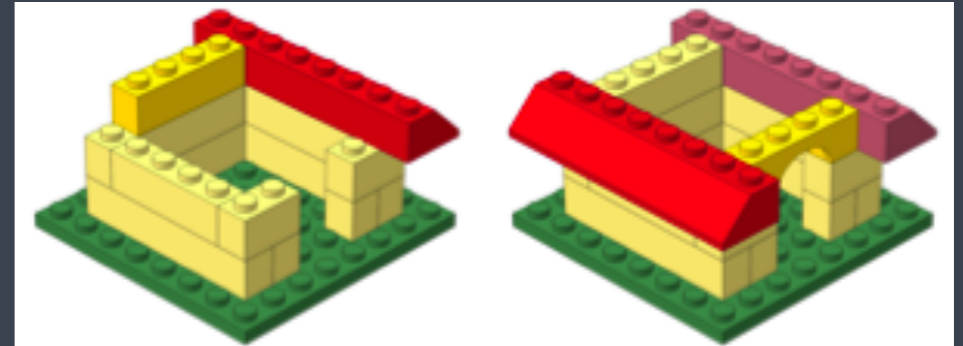


<http://anoved.net/tag/lego/page/3/>

Bioinformatics data analysis

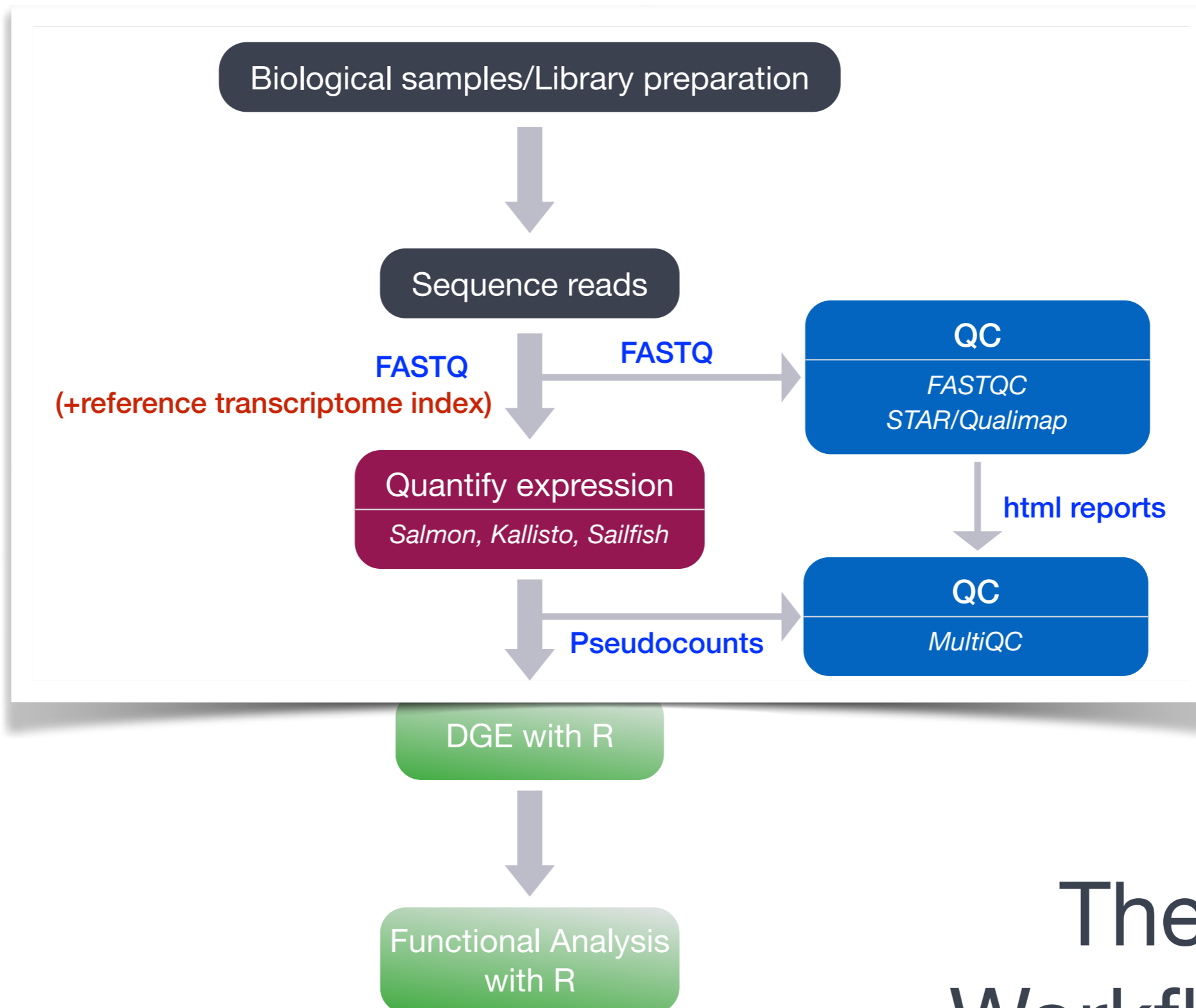


# Learning Objectives



- ✓ 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-fasrc>

# Course schedule online

## Workshop Schedule

**NOTE:** The *Basic Data Skills* Introduction to the command-line interface workshop is a prerequisite.

### Pre-reading

- [Shell basics review](#)
- [Introduction to RNA-seq](#)

### Day 1

Time	Topic	Instructor
09:30 - 09:45	Workshop introduction	Radhika
09:45 - 10:25	Working in an HPC environment	Radhika
10:25 - 11:05	Project Organization and Best Practices in Data Management	Meeta
11:05 - 11:45	Quality Control of Sequence Data: Running FASTQC	Jihe
11:45 - 12:00	Overview of self-learning materials and homework submission	Jihe/Meeta

# Course materials online

## Introduction to RNA-Seq using high-performance computing

Intro to RNA-seq updated for a flipped classroom

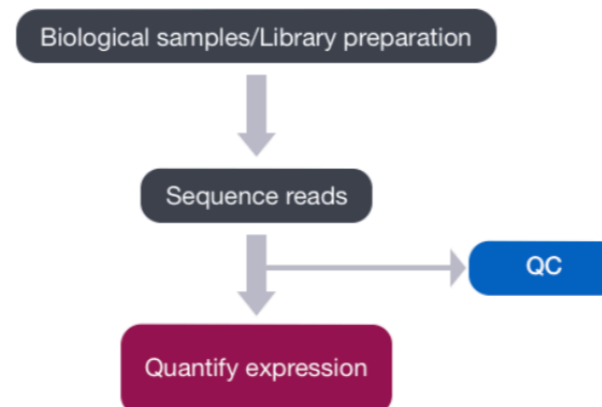
[View on GitHub](#)

### Learning Objectives:

- Understand the quality values in a FASTQ file
- Create a quality report using FASTQC

### Quality Control of FASTQ files

The first step in the RNA-Seq workflow is to take the FASTQ files received from the sequencing facility and assess the quality of the sequence reads.





# Single screen & 3 windows?

The image shows a Zoom meeting interface with three participants: Mary Piper, Troubleshooter (Radhika), and Jihe Liu. A terminal window is open, displaying the output of the `ls -ltr` command in a directory named `unix_workshop`. The output lists files with their permissions, owners, dates, and names. A second terminal window is also open, showing the execution of a complex `cut` command to extract specific columns from a file. The output of this command is a list of chromosome 1 coordinates.

```
rsk27@clarinet002-072:~$ ll -ltr 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:~$
```

```
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 screenshot 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 list of participants: Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host). The main content area is a terminal window titled 'rsk394 -- bash -- 69x24'. The terminal shows 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$
```

Overlaid on the terminal is a presentation slide titled 'Introduction to the command line interface (shell)'. The slide features a blue background with white text and a 'View on GitHub' button. The slide content includes the text 'Introduction to the command line interface (shell)' and a 'View on GitHub' button. A large blue 'ZOOM' watermark is visible on the left side of the terminal window. At the bottom of the slide, there is a navigation bar with icons for Unmute, Stop Video, Invite, Share Screen, and Reactions.

*Our  
recommendation*

# Single screen & 3 windows?

The screenshot 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 titled 'rsk27@clarinet002-072:~\$ ll -ltr unix\_workshop/'. The terminal output lists files and their details:

```
rsk27@clarinet002-072:~$ ll -ltr 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:~$
```

Below the terminal window, there is a text box titled 'Starting with the shell' with instructions and code snippets:

```
$ cd unix_workshop

'cd' stands for 'change directory'

Let's see what is in here. Type:

$ ls
```

Overlaid on the right side of the Zoom interface is a red-bordered box containing a terminal window titled 'rsk394 -- bash -- 69x24'. The terminal shows a command and its 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$
```

Overlaid on the top right of the red-bordered box is a text box with the text 'Web browser' in large red font, and below it, 'Introduction to the command line interface (shell)' in white font on a blue background with a DNA sequence pattern. A 'View on GitHub' button is visible below the text.

*Our  
recommendation*



# Single screen & 3 windows?

The screenshot 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 'Participants (3)' list. In the center, a terminal window is open, displaying a list of files and a command being executed. To the right, a browser window is open, showing a page titled 'Introduction to the command line interface (shell)'. The terminal window is highlighted with a green border.

```
rsk27@clarinet002-072:~$ ll -ltr 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:~$
```

```
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 screenshot 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 list of participants: Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host). The main content area is divided into three windows:

- Terminal (green border):** A terminal window titled "rsk394 -- bash -- 69x24" showing the execution of a command to sort and head a file. The output is a list of chromosome 1 coordinates.

```
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$
```
- Web browser (red border):** A browser window titled "Web browser" showing a page titled "Introduction to the command line interface (shell)". The page content includes a "View on GitHub" button and a background image of DNA sequence letters (G, A, T, C).
- Document (blue border):** A document titled "Starting with the shell" with the text: "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." Below the text is a code block: 

```
$ cd unix_workshop
```

 and a note: "'cd' stands for 'change directory'".

A large blue "ZOOM" watermark is overlaid on the document window.

*Our recommendation*



**Terminal**

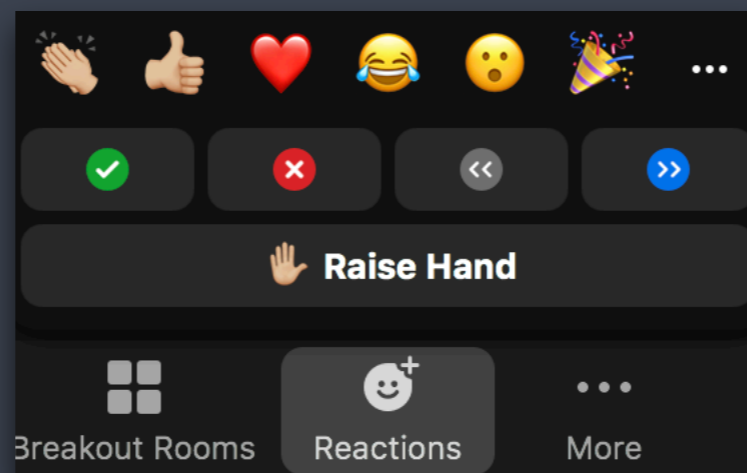
# Odds and Ends

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



# Odds and Ends

- ❖ Quit/minimize all applications that are not required for class
- ❖ Are you all set?
  - ▶  = "agree", "I'm all set" (equivalent to a **green post-it**)
  - ▶  = "disagree", "I need help" (equivalent to a **red post-it**)



# Odds and Ends (contd.)

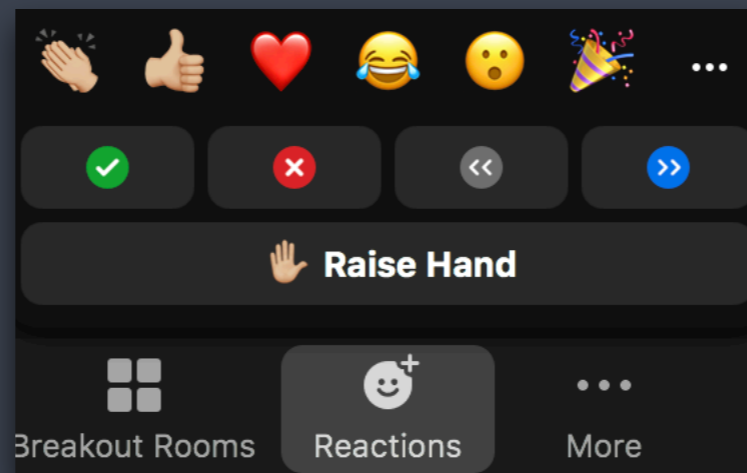
## ❖ Questions for the presenter?

- Post the question in the Chat window OR

-  when the presenter asks for questions

## ❖ Technical difficulties with software?

- Start a private chat with the *Troubleshooter* with a description of the problem.



# Thanks!

- **Daniel Caunt** and **Maggie McFee** from FAS-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](mailto:hbctraining@hsph.harvard.edu)

*HBC consulting:* [bioinformatics@hsph.harvard.edu](mailto:bioinformatics@hsph.harvard.edu)

*FAS-RC:* [create a ticket](#)

## Twitter

*HBC:* @bioinfocore

*FAS-RC:* @fas\_rc