

# Introduction to RNA-seq using High-Performance Computing (HPC)

Harvard Chan Bioinformatics Core  
in collaboration with  
HMS Research Computing

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



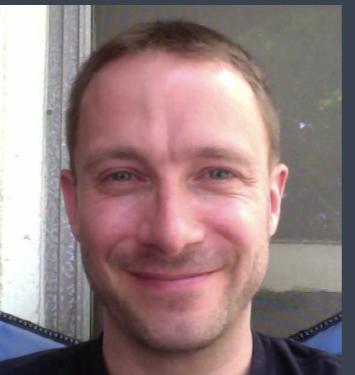
Shannan Ho Sui  
*Director*



John Hutchinson  
*Associate Director*



Victor Barrera



Rory Kirchner



Zhu Zhuo



Preetida Bhetariya



Meeta Mistry



Mary Piper  
*Assoc. Training Director*



Jihe Liu



Radhika Khetani  
*Training Director*



Maria Simoneau



James Billingsley



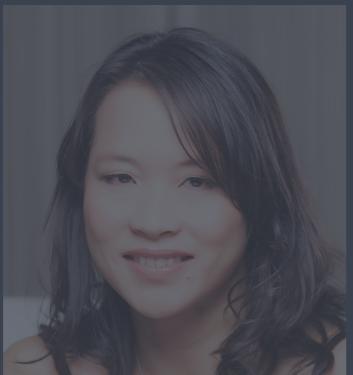
Sergey Naumenko



Joon Yoon



Peter Kraft  
*Faculty Advisor*



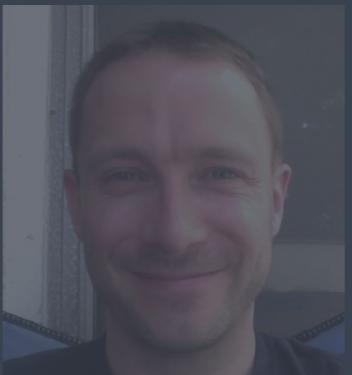
Shannan Ho Sui  
*Director*



John Hutchinson  
*Associate Director*



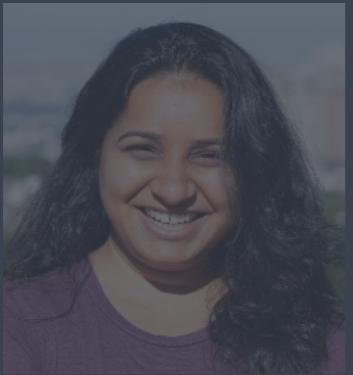
Victor Barrera



Rory Kirchner



Zhu Zhuo



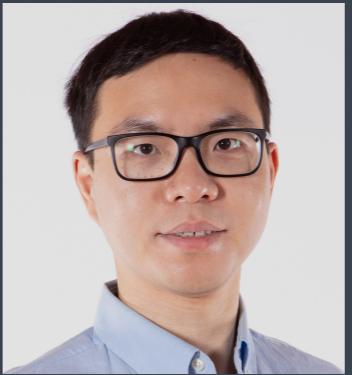
Preetida Bhetariya



Meeta Mistry



Mary Piper



Jihe Liu



Radhika Khetani  
*Training Director*



Maria Simoneau



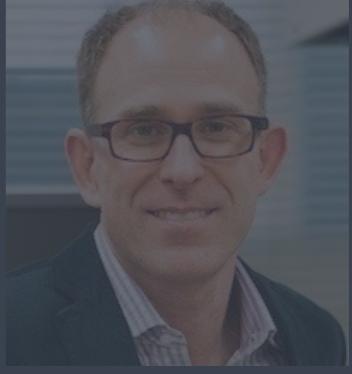
James Billingsley



Sergey Naumenko



Joon Yoon



Peter Kraft  
*Faculty Advisor*

# Consulting

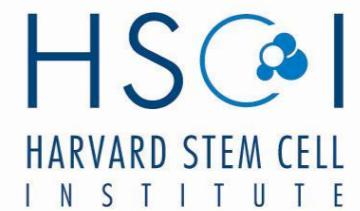
- RNA-seq analysis: bulk, single cell, small RNA
- ChIP-seq and ATAC-seq analysis
- Genome-wide methylation
- WGS, resequencing, exome-seq and CNV studies
- QC & analysis of gene expression arrays
- Functional enrichment analysis
- Grant support

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



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

NIEHS



# 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/>



# Training

We have divided our short workshops into 2 categories:

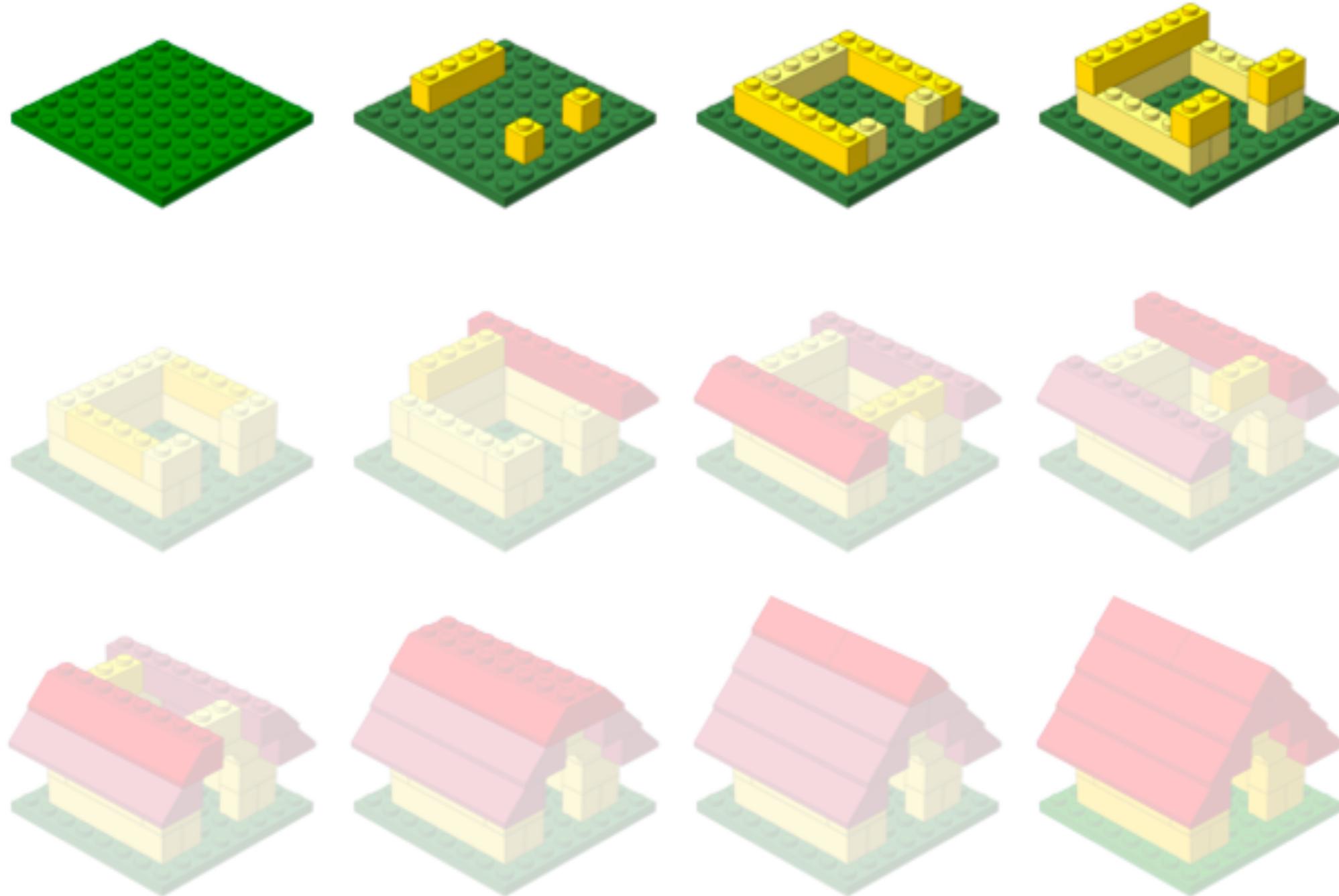
1. Basic Data Skills - No prior programming knowledge needed (no prerequisites)
2. Advanced Topics: Analysis of high-throughput sequencing (NGS) data - Certain “Basic” workshops required as prerequisites.

*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/>

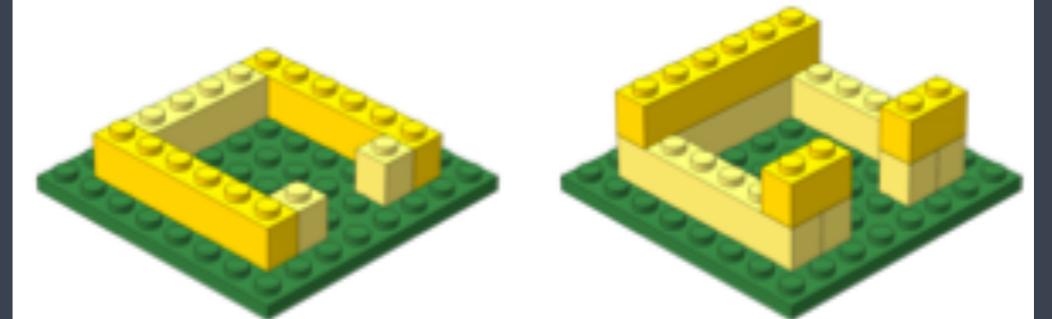




<http://anoved.net/tag/lego/page/3/>

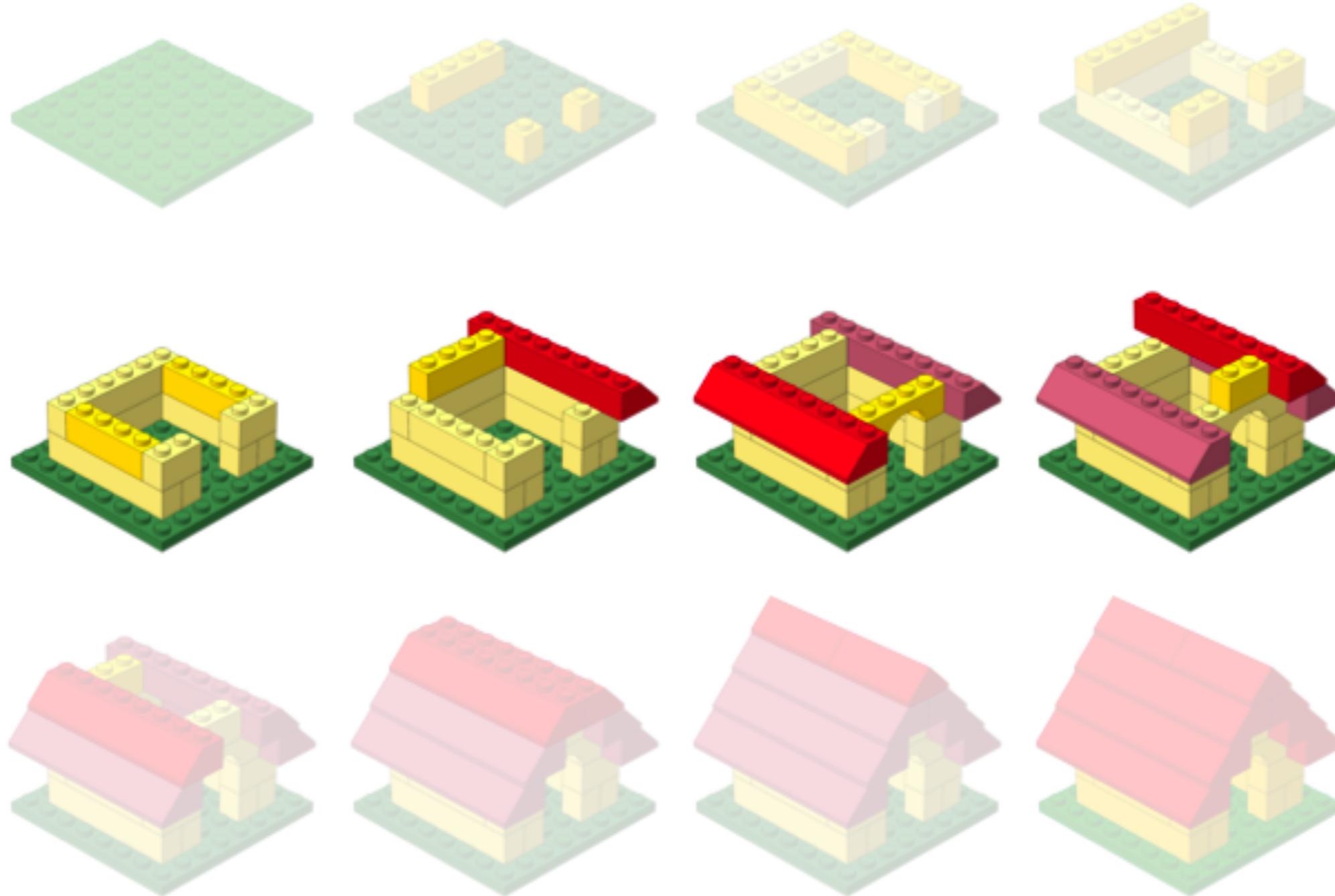
Setting up to perform Bioinformatics analysis

# 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

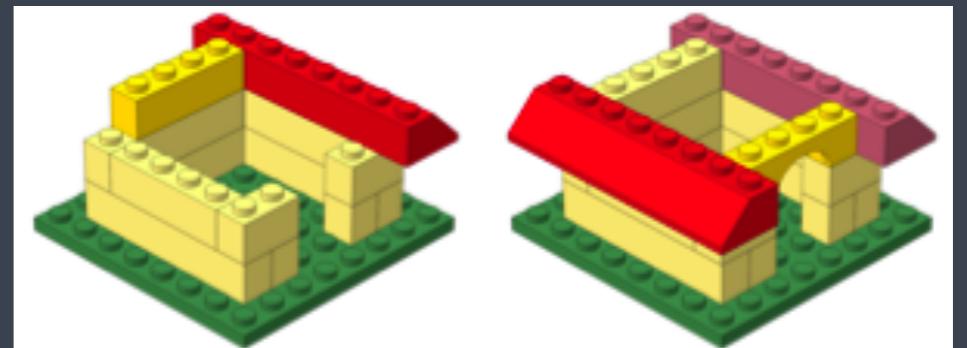
# Workshop scope



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

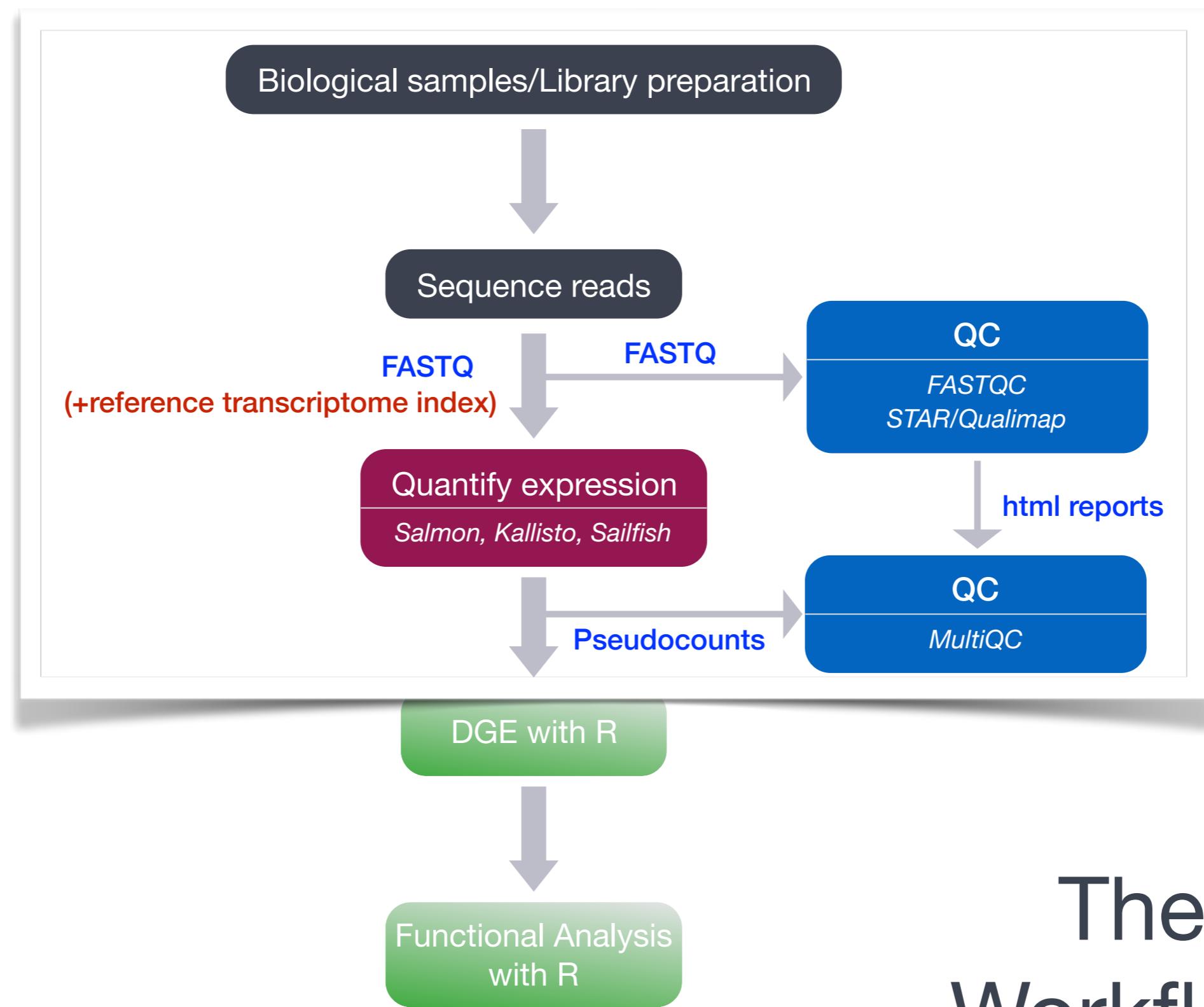
Bioinformatics data analysis

# 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>

# 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

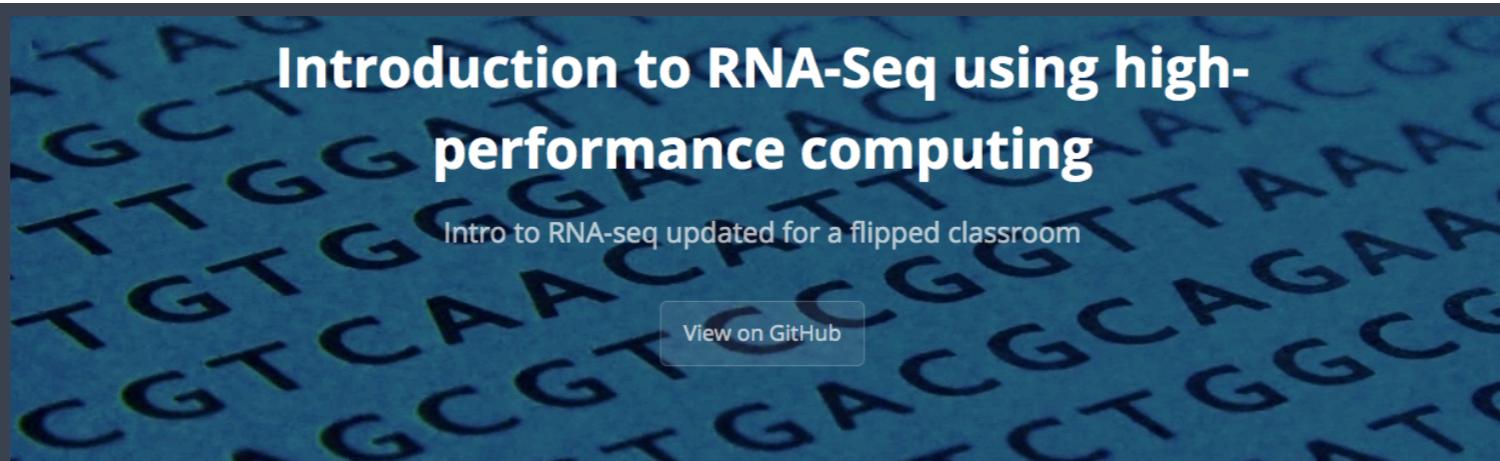
Time	Topic	Instructor
09:30 - 09:45	Workshop introduction	Radhika
09:45 - 10:25	Working in an HPC environment	Radhika
10:25 - 11:05	Project Organization and Best Practices in Data Management	Meeta
11:05 - 11:45	Quality Control of Sequence Data: Running FASTQC	Jihe
11:45 - 12:00	Overview of self-learning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



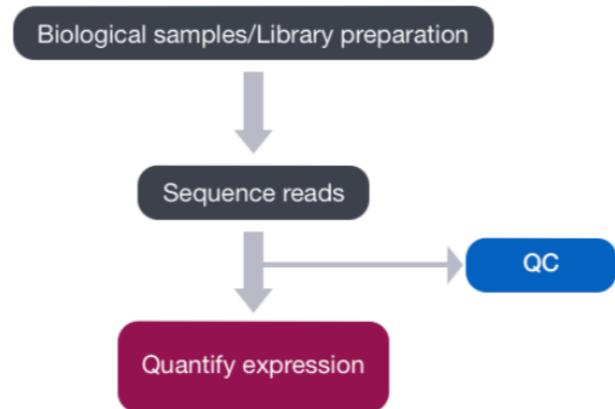
The background of the slide features a blue-tinted grid of DNA sequence data, specifically FASTQ reads, with letters A, T, C, G visible.

### Learning Objectives:

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

### Quality Control of FASTQ files

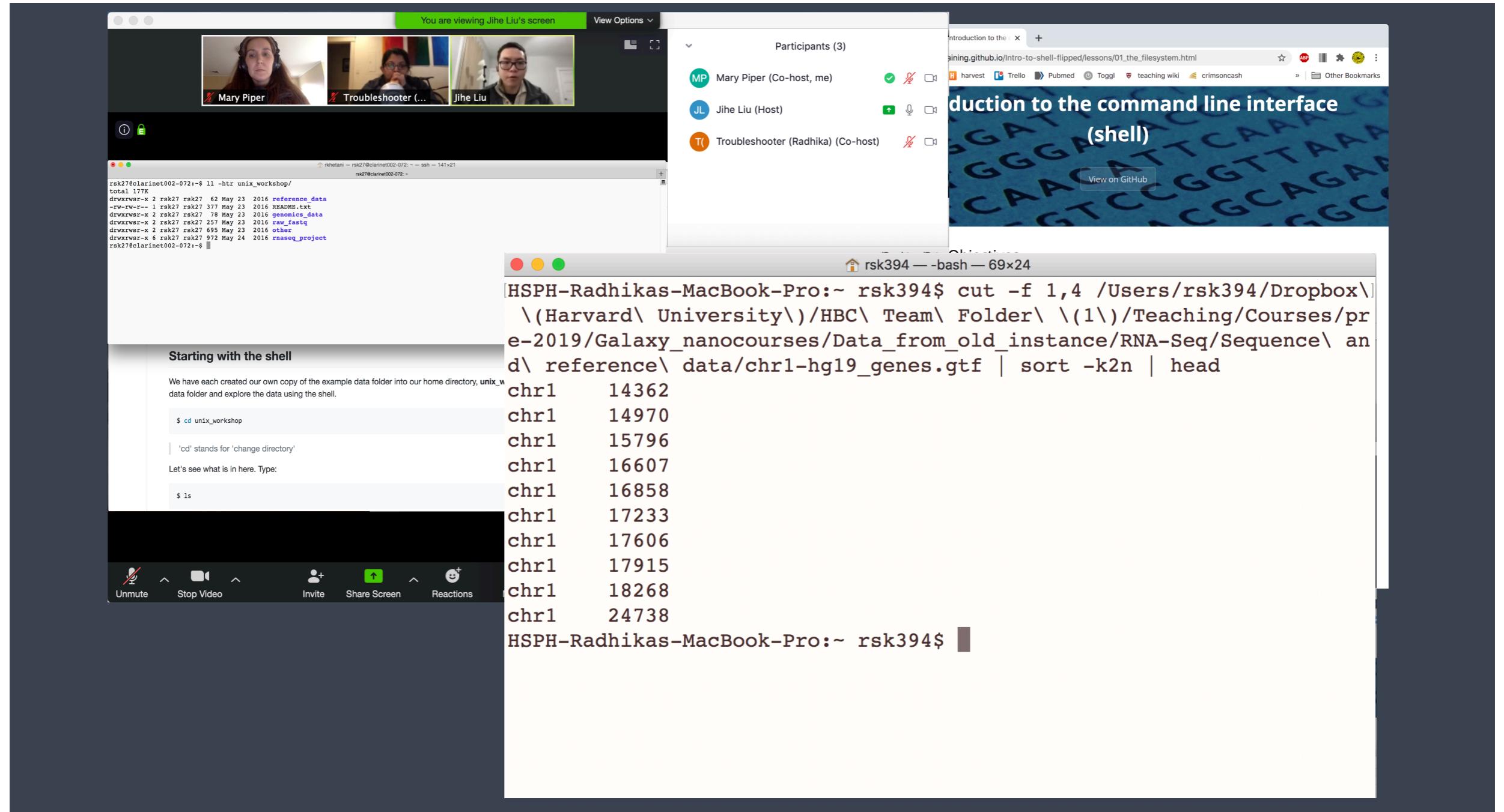
The first step in the RNA-Seq workflow is to take the FASTQ files received from the sequencing facility and assess the quality of the sequence reads.



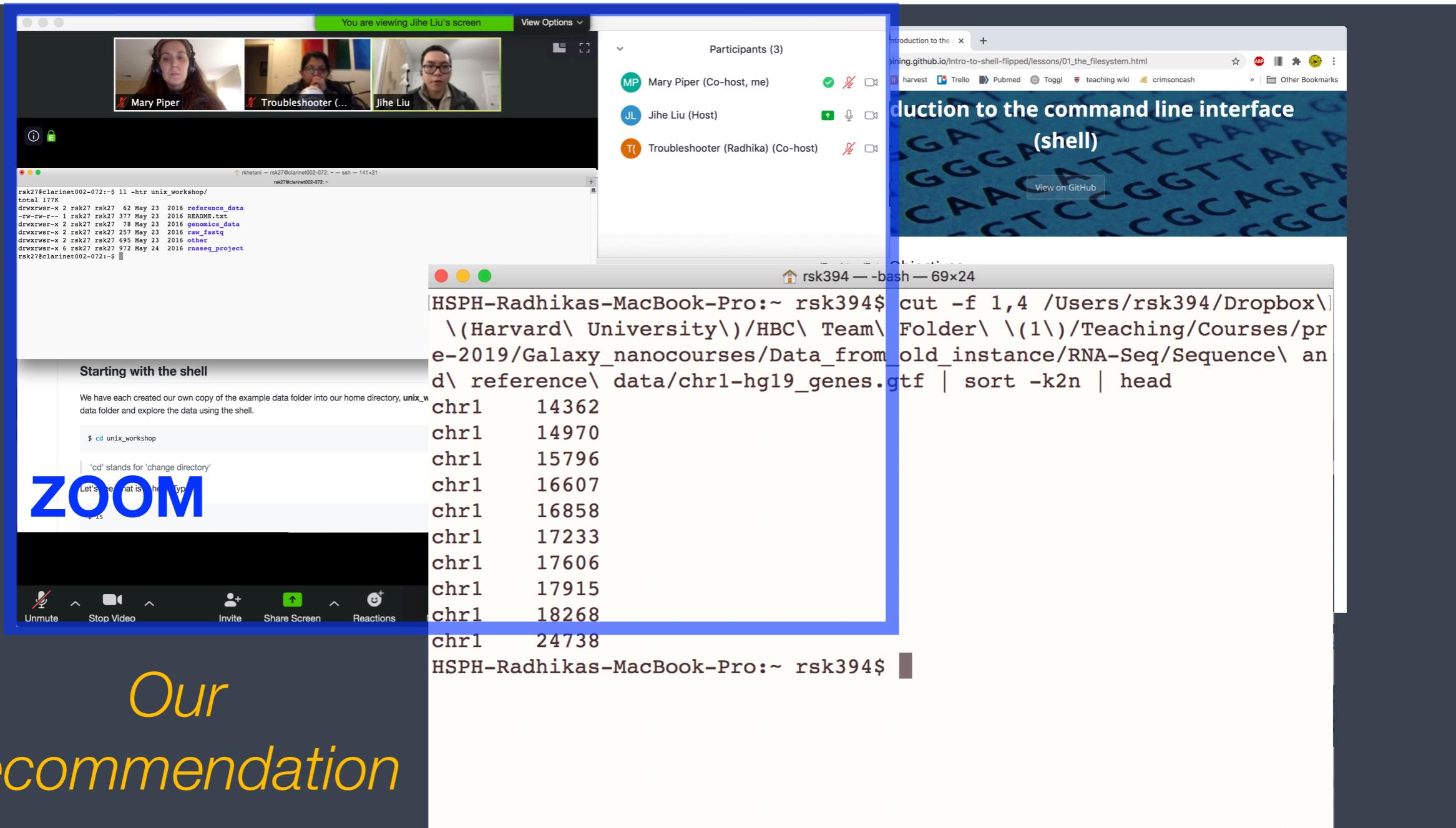
```
graph TD; A[Biological samples/Library preparation] --> B[Sequence reads]; B --> C[QC]; C --> D[Quantify expression]
```

The diagram illustrates the RNA-Seq workflow. It starts with 'Biological samples/Library preparation', which leads to 'Sequence reads'. From 'Sequence reads', the process branches into two paths: one leading to 'QC' (Quality Control) and another leading to 'Quantify expression'.

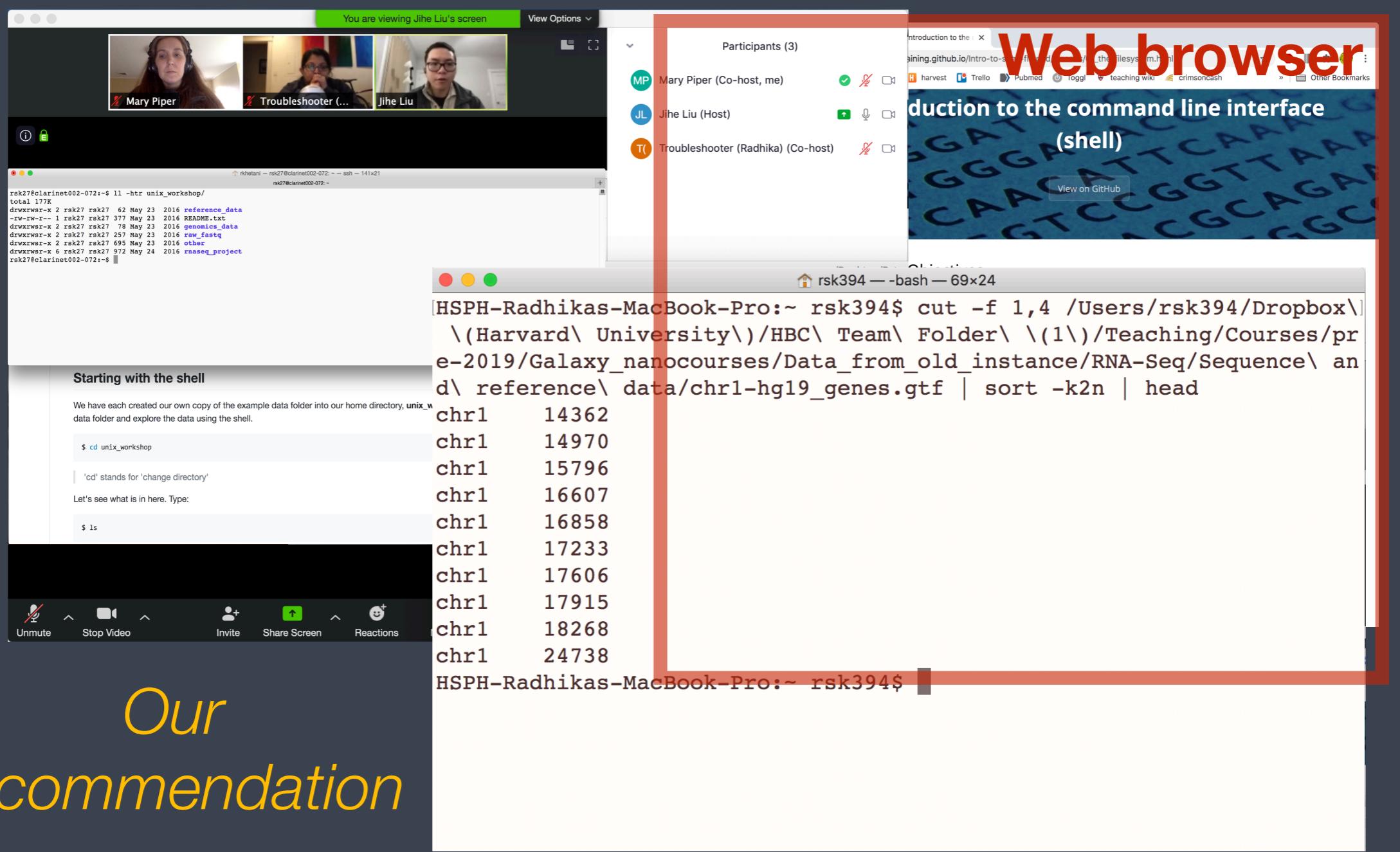
# Single screen & 3 windows?



# Single screen & 3 windows?



# Single screen & 3 windows?



# Single screen & 3 windows?

The image shows a video conference interface with three main windows:

- Video Feed:** Shows three participants: Mary Piper, Troubleshooter (Radhika), and Jihe Liu.
- Participants List:** Shows three participants: Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host).
- Terminal Session:** A green-bordered window titled "rsk394 — bash — 69x24" displays a command-line interface. The command run is:

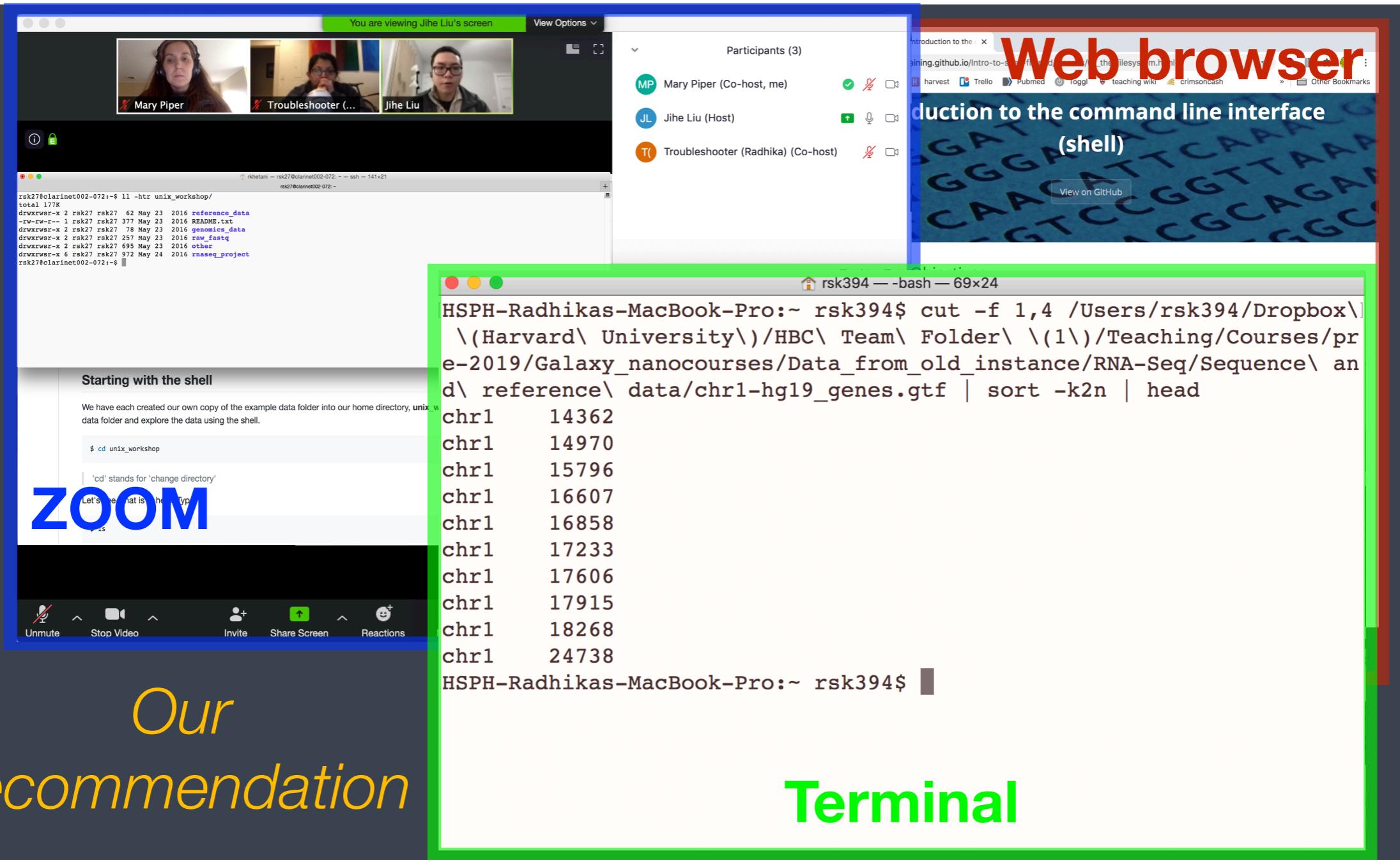
```
HSPH-Radhikas-MacBook-Pro:~ rsk394$ cut -f 1,4 /Users/rsk394/Dropbox\\(Harvard\\ University\\)/HBC\\ Team\\ Folder\\ \\(1\\)/Teaching/Courses/pre-2019/Galaxy_nanocourses/Data_from_old_instance/RNA-Seq/Sequence\\ and\\ reference\\ data/chrl-hg19_genes.gtf | sort -k2n | head
```

Output:

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

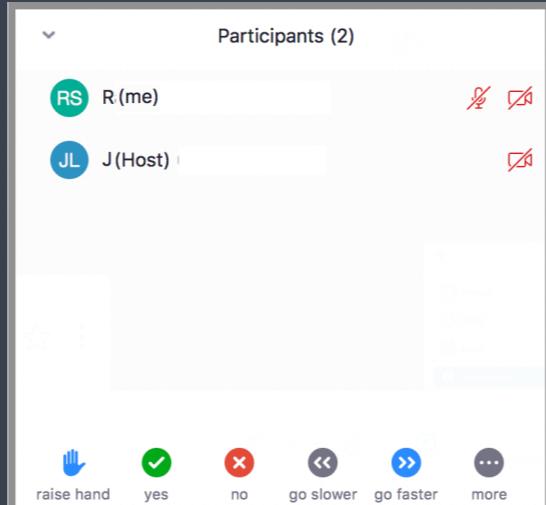
*Our recommendation* **Terminal**

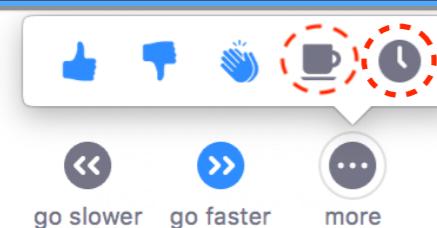
# Single screen & 3 windows?



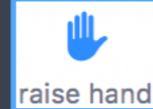
# Odds and Ends (1/2)

- ❖ Quit/minimize all applications that are not required for class
- ❖ Click on “Participants” to open that panel in Zoom



- ▶  = "agree", "I'm all set" (equivalent to a **green post-it**)
- ▶  = "disagree", "I need help" (equivalent to a **red post-it**)
- ▶  If you are away from the computer use the coffee cup or clock icon

# Odds and Ends (2/2)

- ❖ Questions for the presenter?
  - Post the question in the Chat window OR
  - Raise your hand  when the presenter asks for questions
- ❖ Technical difficulties with logging in or running commands?
  - Start a private chat with the *Troubleshooter* with a description of the problem.

# Thanks!

- Andy Bergman & Kathleen Keating 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](mailto:hbctraining@hsph.harvard.edu)

*O2 (HMS-RC):* [rchelp@hms.harvard.edu](mailto:rchelp@hms.harvard.edu)

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

## Twitter

*HBC:* @bioinfocore

*HMS-RC:* @hms\_rc