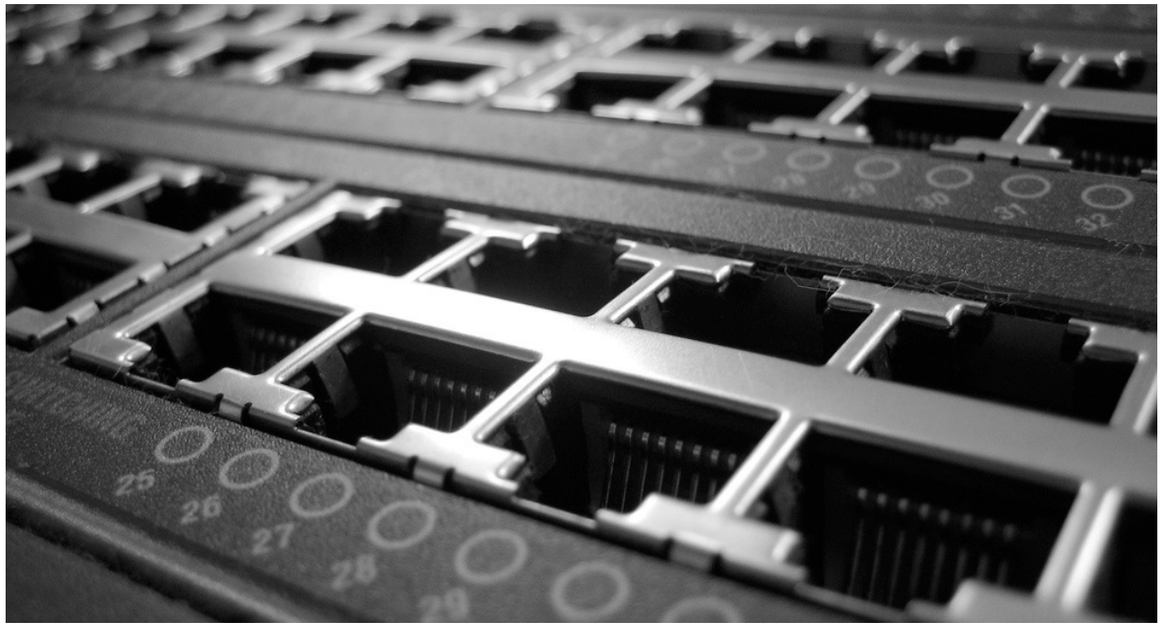


Introduction to bulk RNA-seq (Part I)

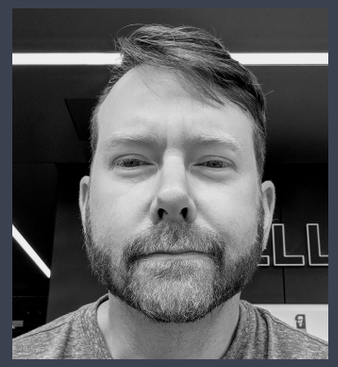
Harvard Chan Bioinformatics Core

in collaboration with

FAS Research Computing

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

Shannan Ho Sui
Director



John Hutchinson
Associate Director

Victor Barrera

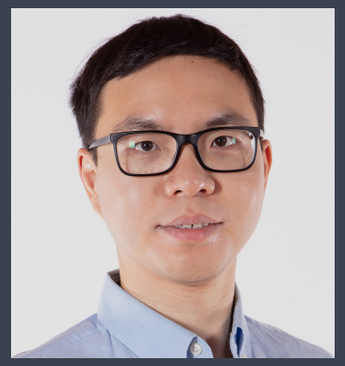
Zhu Zhuo

Preetida Bhetariya

Radhika Khetani
Training Director

Meeta Mistry

Mary Piper
Assoc. Training Director



Jihe Liu

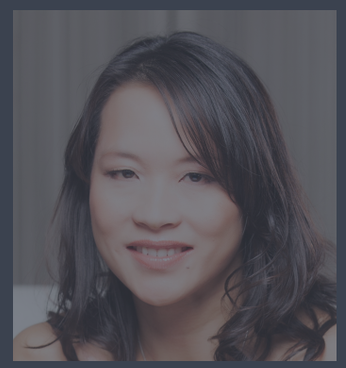
Will Gammerdinger

Maria Simoneau

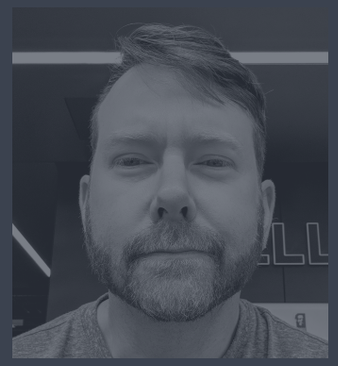
James Billingsley

Sergey Naumenko

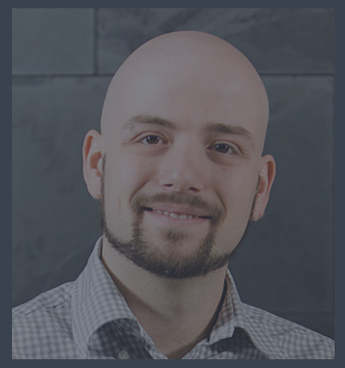
Peter Kraft
Faculty Advisor



Shannan Ho Sui
Director

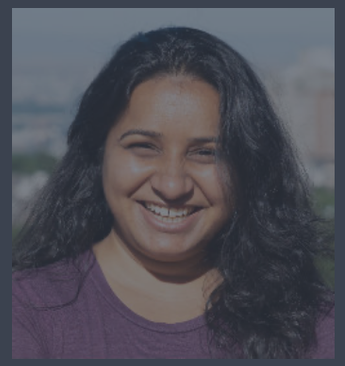


John Hutchinson
Associate Director



Victor Barrera

Zhu Zhuo

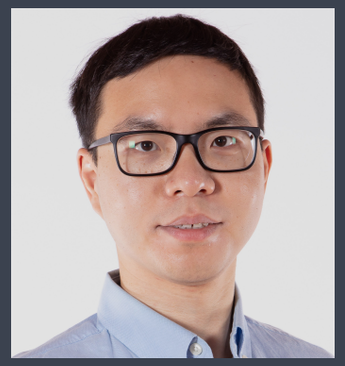


Preetida Bhetariya

Radhika Khetani
Training Director

Meeta Mistry

Mary Piper
Assoc. Training Director



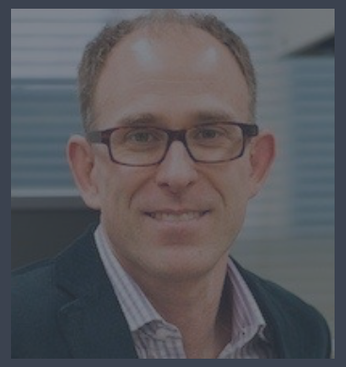
Jihe Liu

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/

CFAR
HARVARD UNIVERSITY
CENTER FOR AIDS RESEARCH

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NIEHS

HSCI
HARVARD STEM CELL
INSTITUTE

HARVARD
CATALYST
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

HARVARD CATALYST 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/

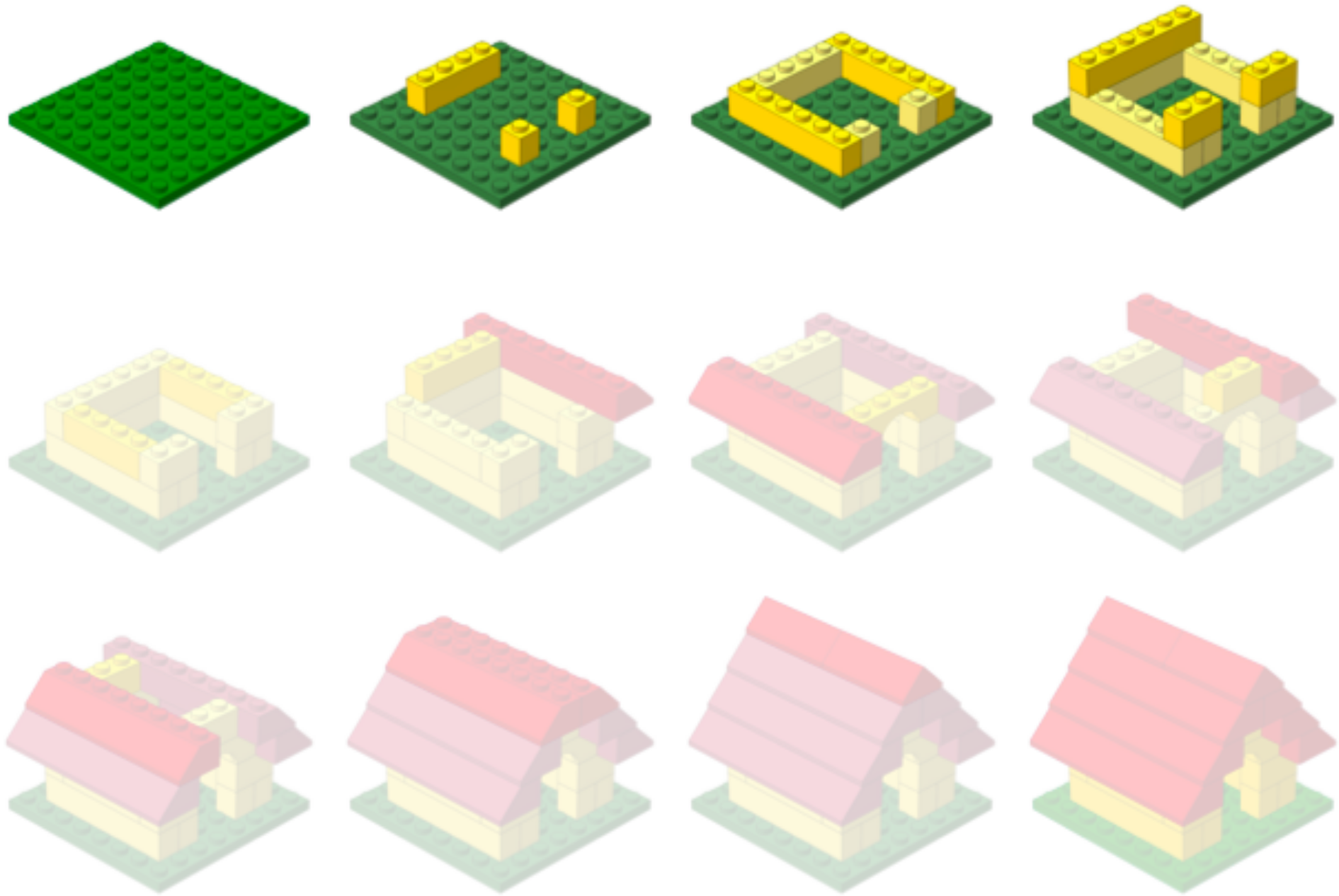
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

DF/HCC DANA-FARBER / HARVARD CANCER CENTER

HSCI HARVARD STEM CELL INSTITUTE

HARVARD CATALYST THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

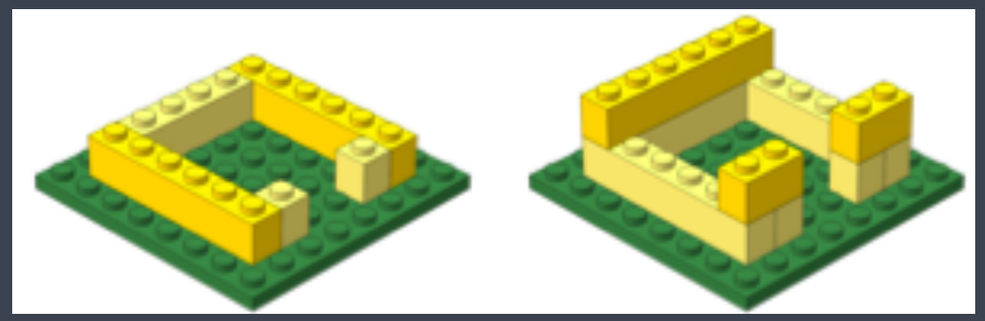
HARVARD MEDICAL SCHOOL



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

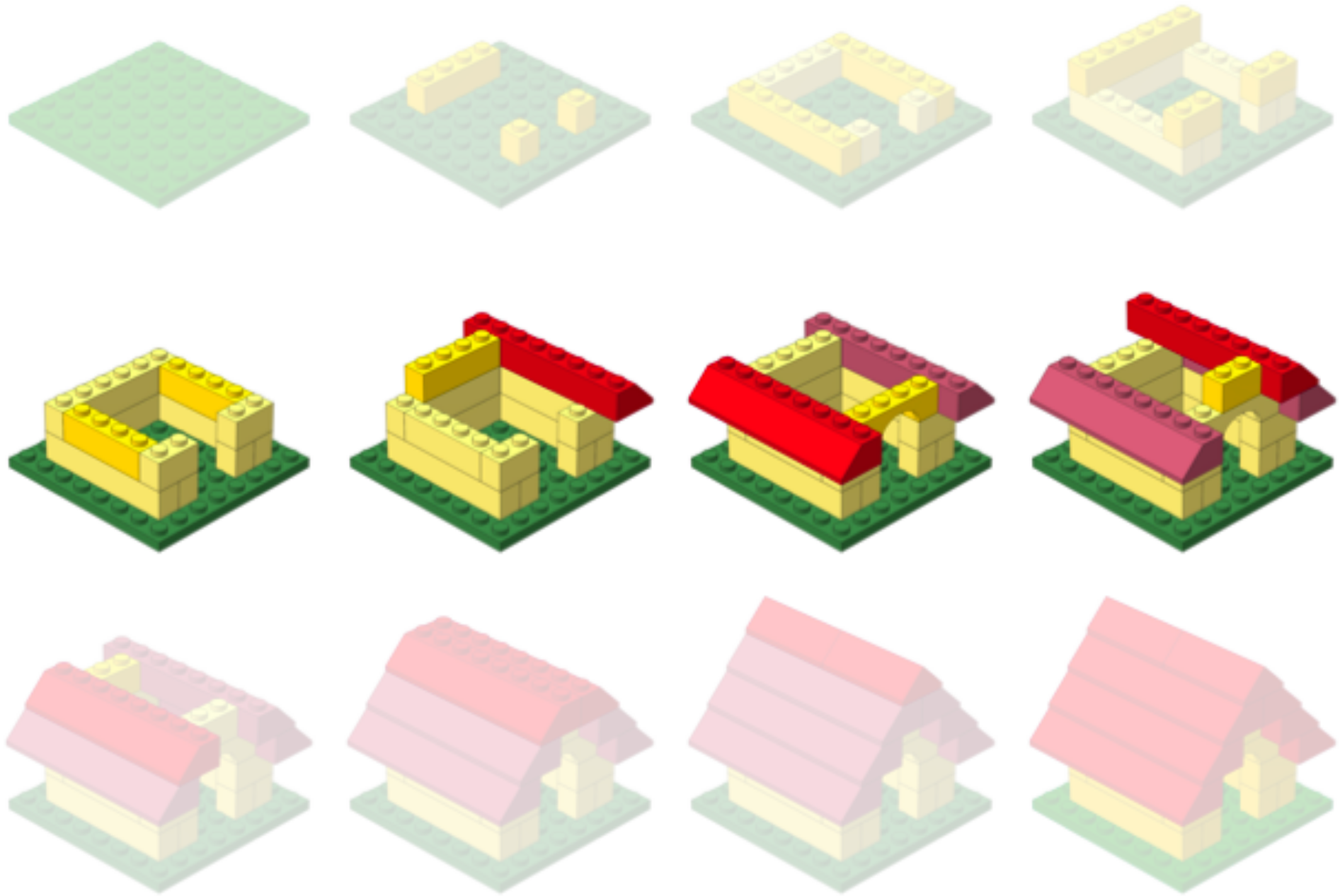
Setting up to perform Bioinformatics analysis

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

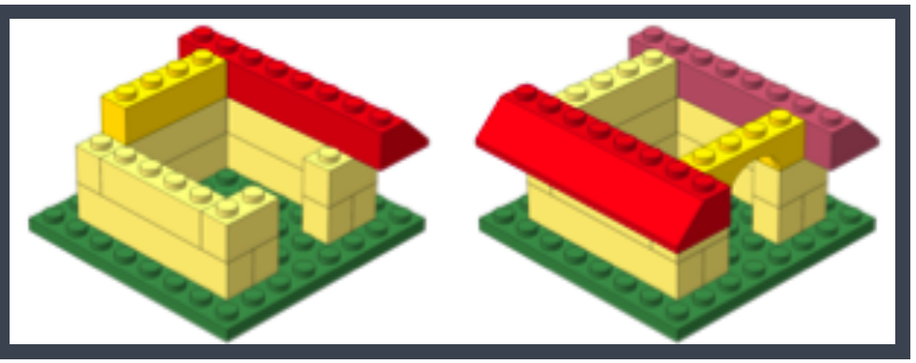
Workshop scope



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

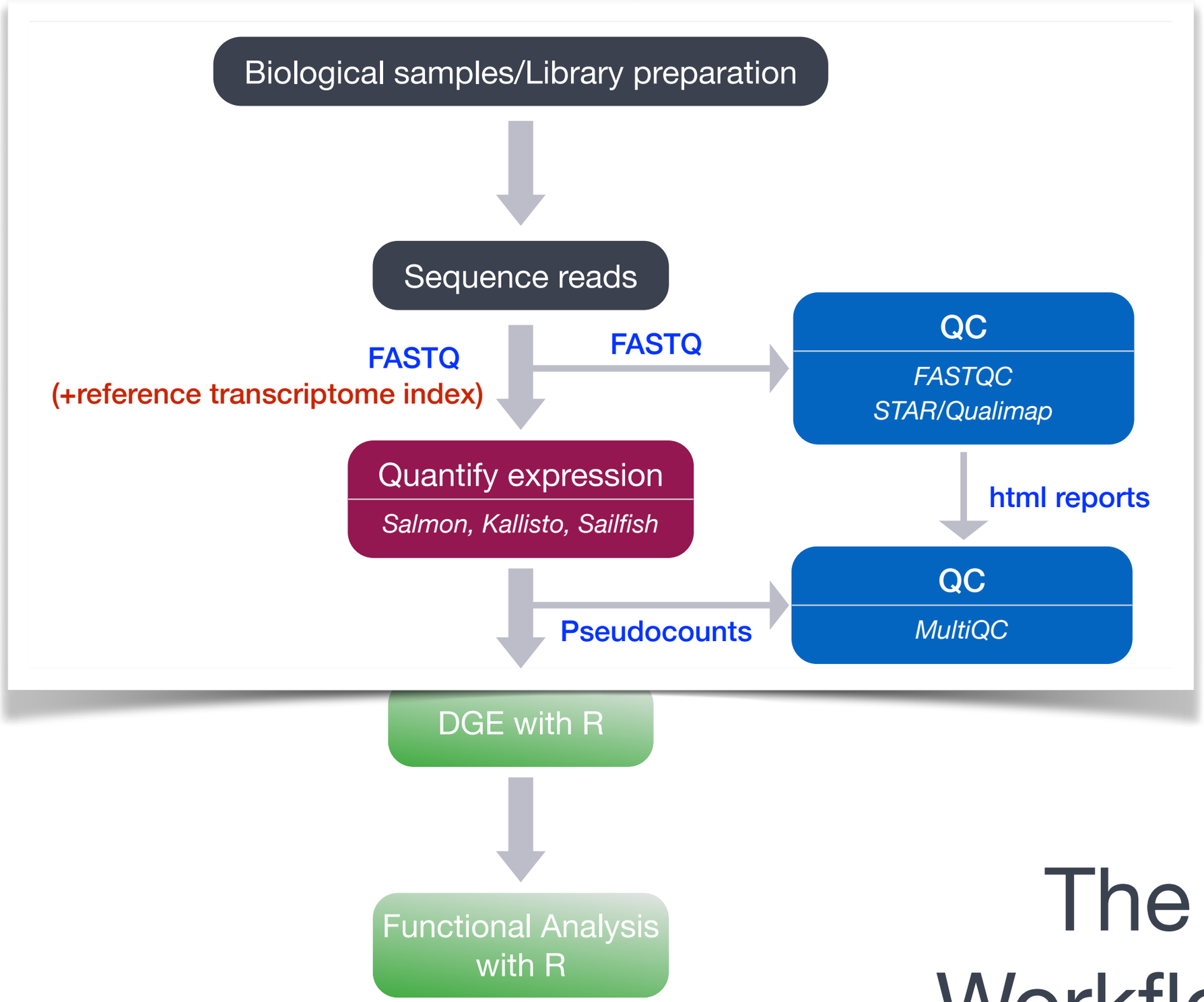
Bioinformatics data analysis

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

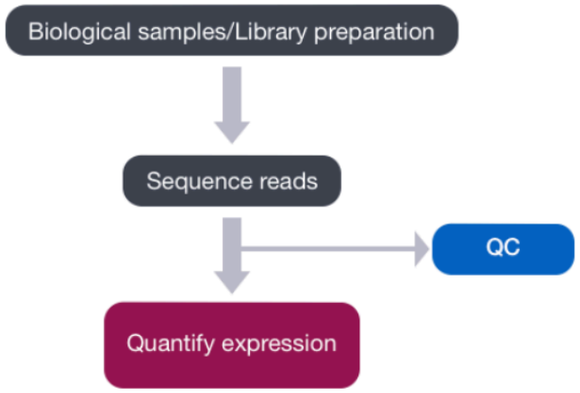
Course materials online

Learning Objectives:

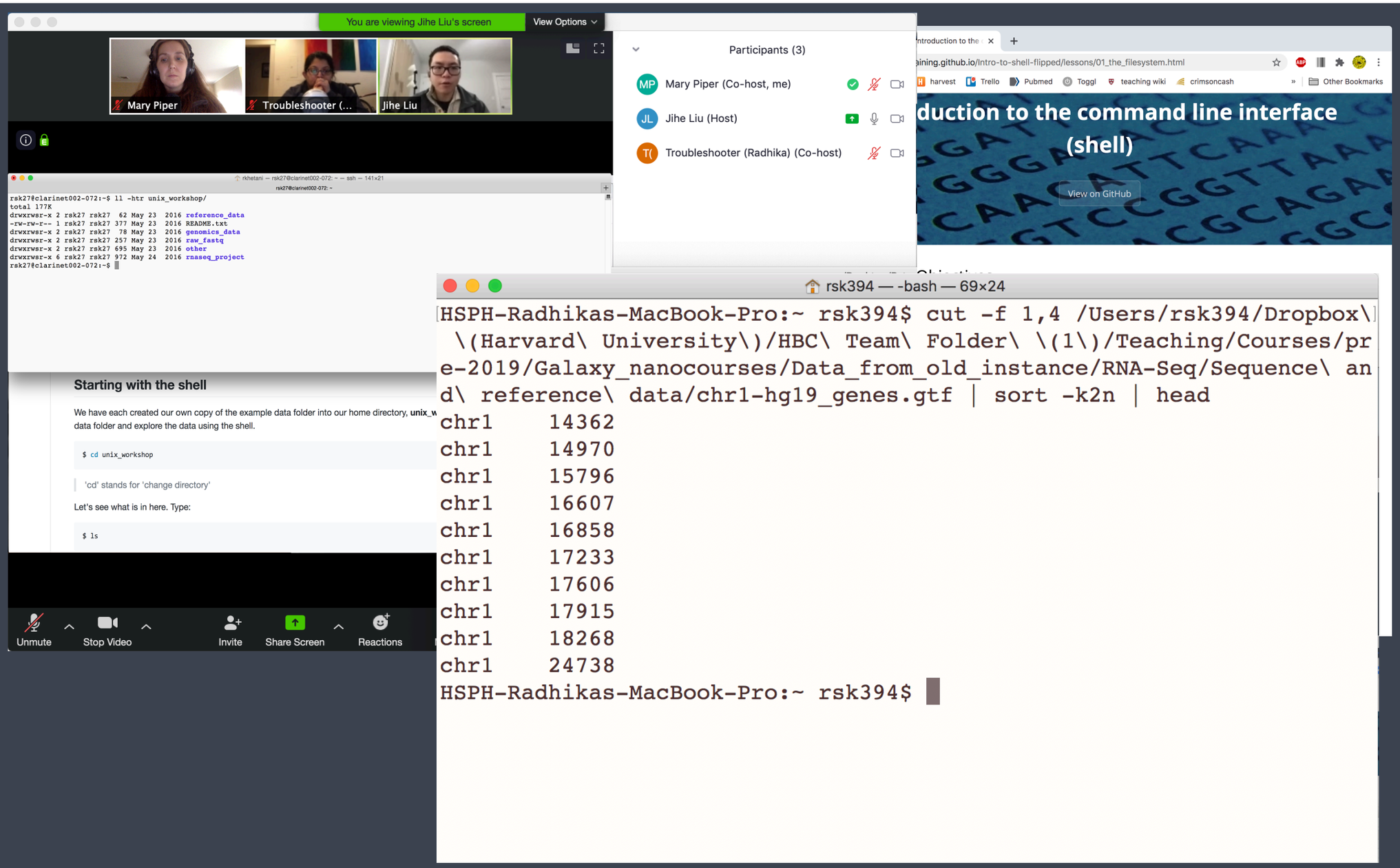
- Understand the quality values in a FASTQ file
- Create a quality report using FASTQC

Quality Control of FASTQ files

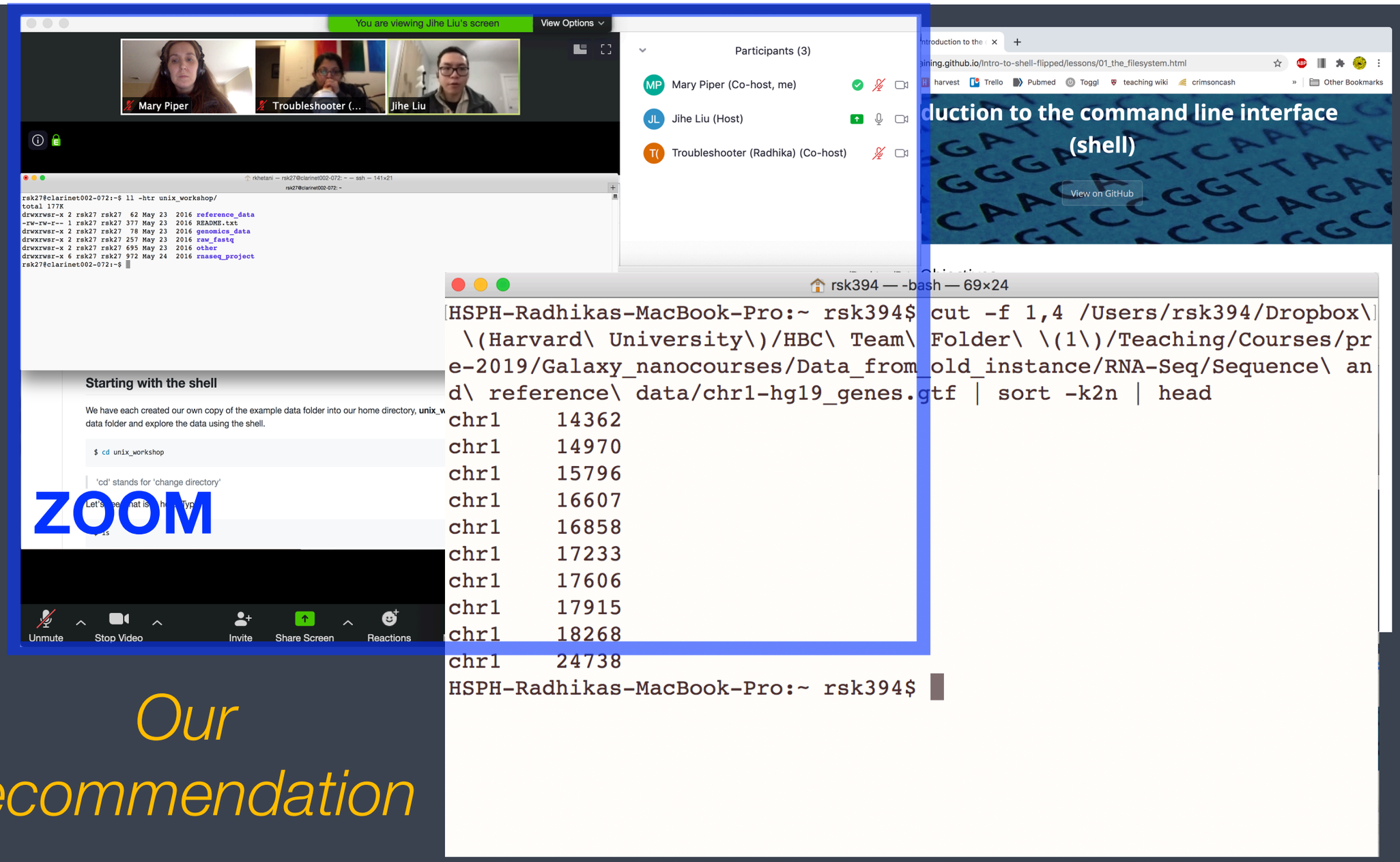
The first step in the RNA-Seq workflow is to take the FASTQ files received from the sequencing facility and assess the quality of the sequence reads.



Single screen & 3 windows?

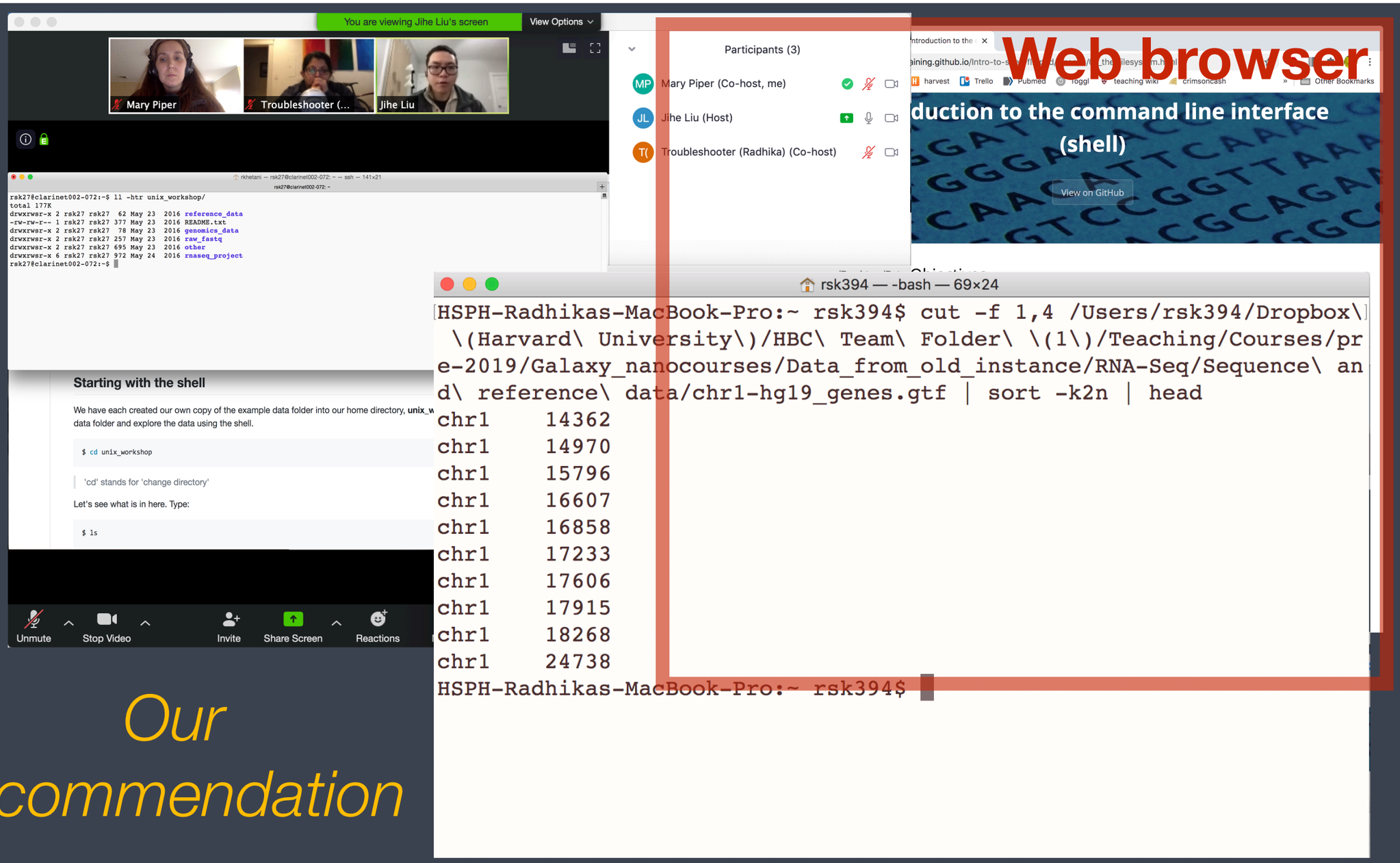


Single screen & 3 windows?



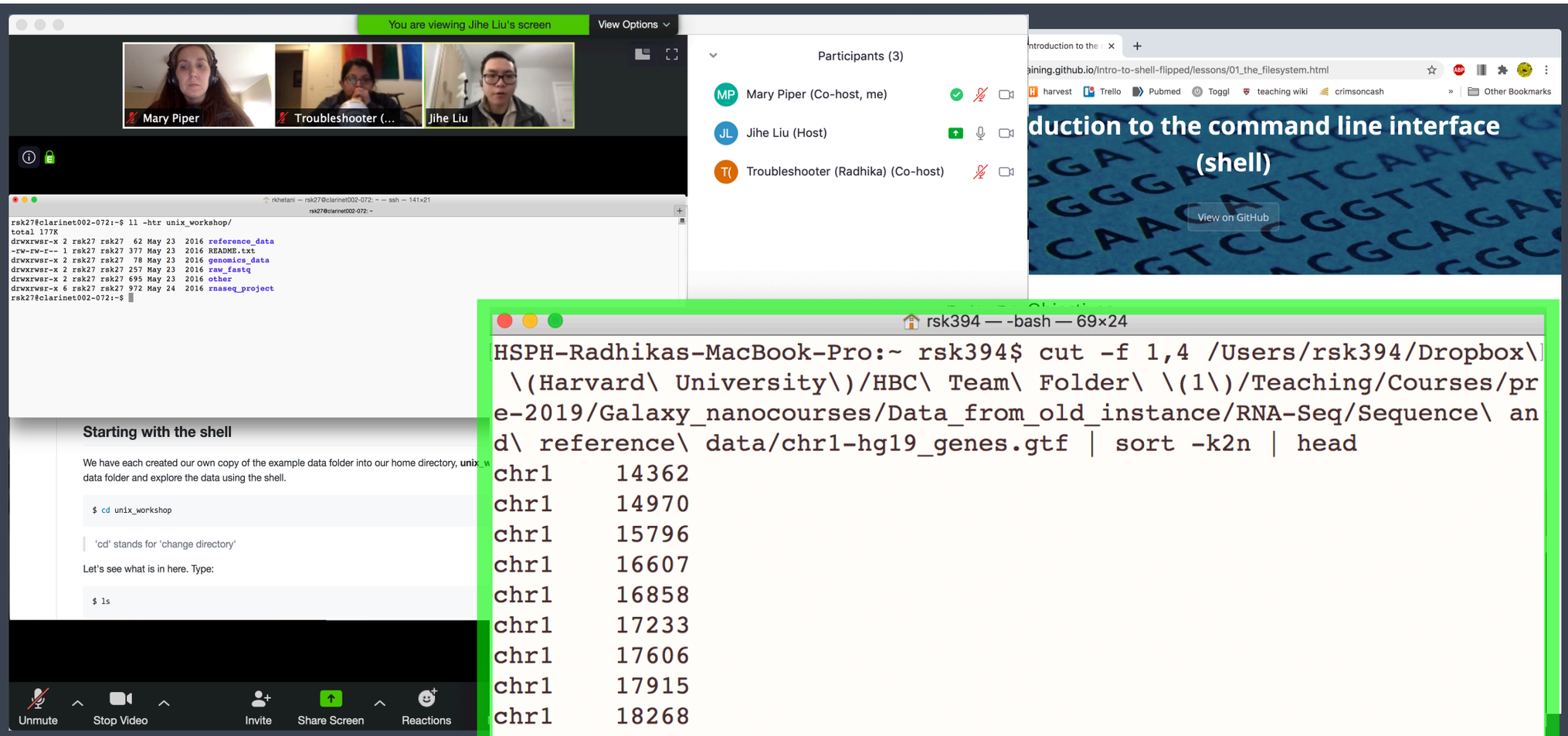
Our recommendation

Single screen & 3 windows?



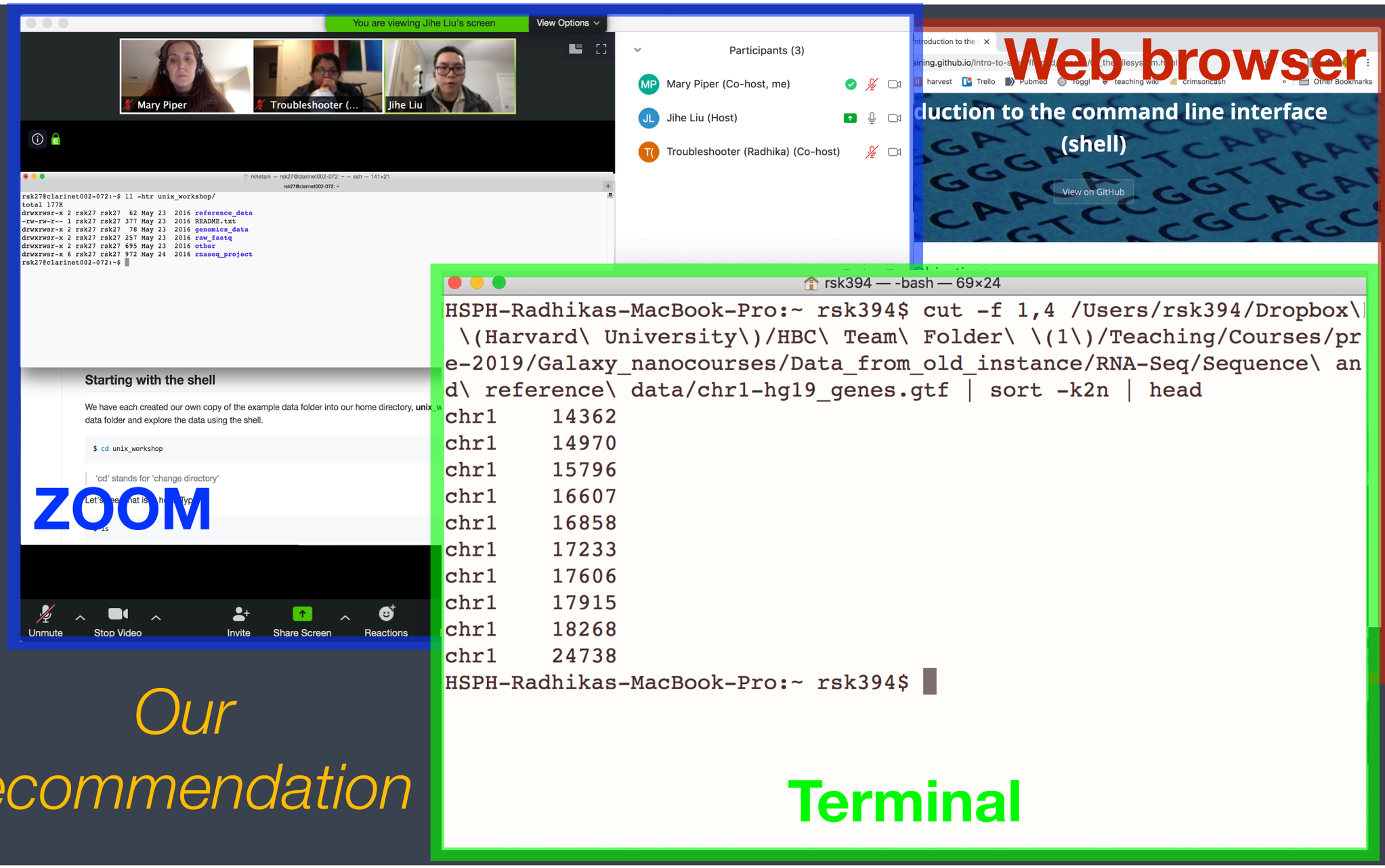
Our recommendation

Single screen & 3 windows?



Our recommendation

Single screen & 3 windows?

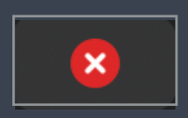


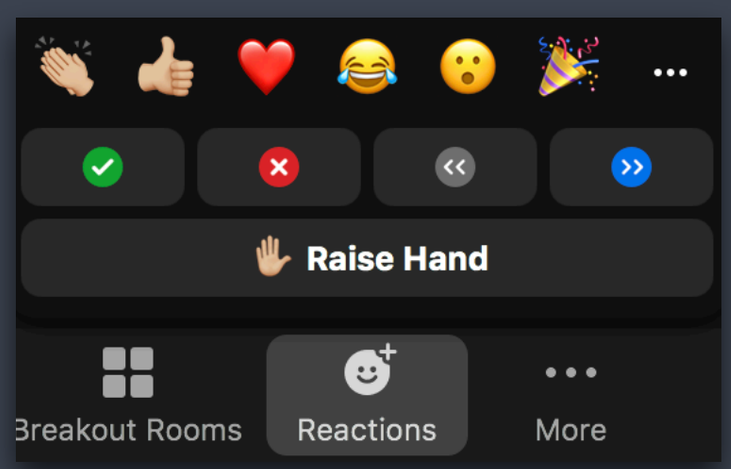
Our recommendation

Odds and Ends

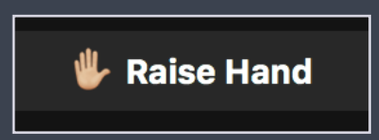
- Quit/minimize all applications that are not required for class

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)

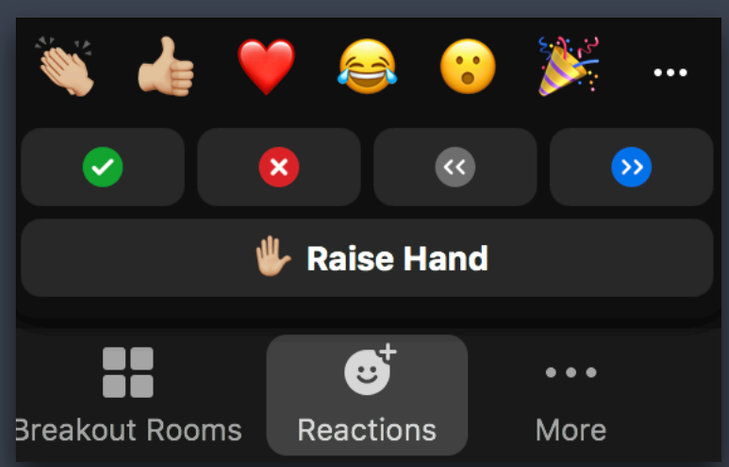


Odds and Ends (contd.)

- 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

Contact us!

HBC training team: hbctraining@hsph.harvard.edu

HBC consulting: bioinformatics@hsph.harvard.edu

FAS-RC: create a ticket

Twitter

HBC: @bioinfocore

FAS-*RC:* @fas_rc