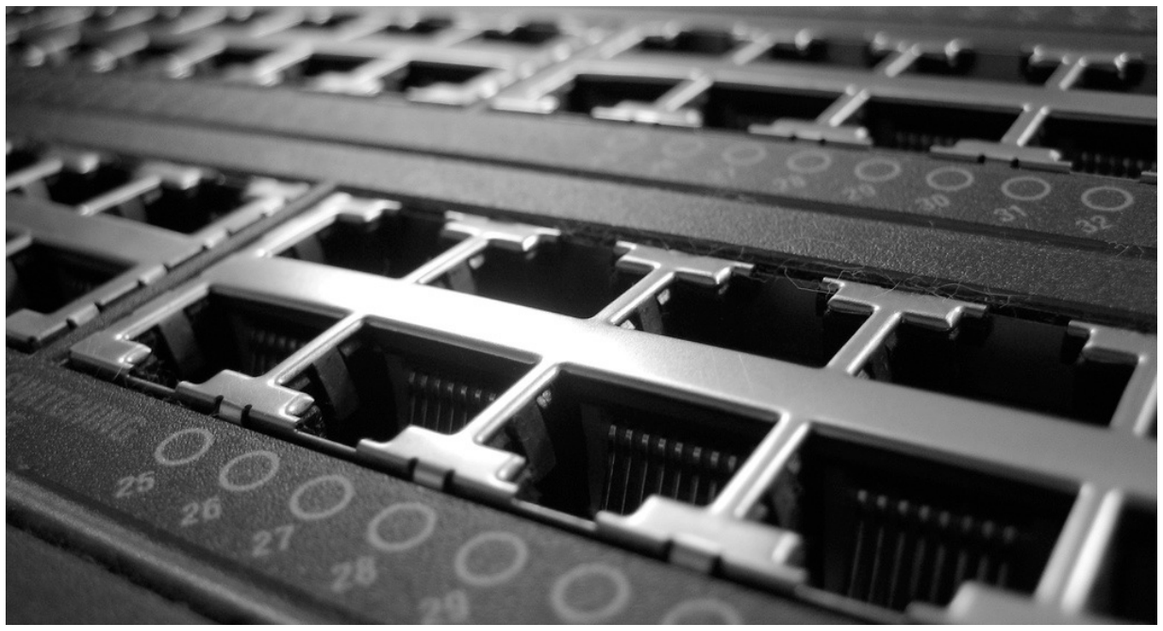


Introduction to the command-line interface

(shell)

Harvard Chan Bioinformatics Core

in collaboration with

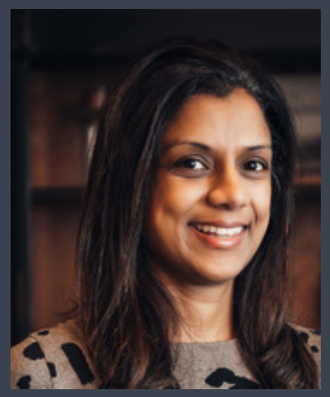
HMS Research Computing

Warm-up Question:
Why are you interested in taking this workshop?

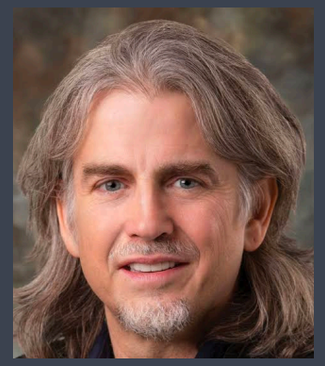
https://tinyurl.com/hbc-shell-online

Introductions!

Shannan Ho Sui
Director



Meeta Mistry
Associate Director



John Quackenbush
Faculty Advisor

Emma Berdan

Heather Wick

Will Gammerdinger

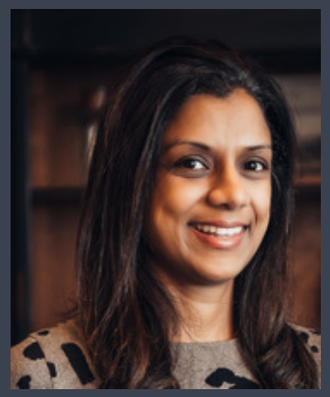
Noor Sohail

James Billingsley

Zhu Zhuo

Maria Simoneau

Shannan Ho Sui
Director



Meeta Mistry
Associate Director

John Quackenbush
Faculty Advisor

Emma Berdan

Heather Wick

Will Gammerdinger

Noor Sohail

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

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

CFAR
HARVARD UNIVERSITY
CENTER FOR AIDS RESEARCH

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NIEHS

HARVARD
CATALYST
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.*

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

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

Training

A key component of the H
researchers at Harvard b

HBC's training team is m
training and community b
research projects to ensu

Our hands-on workshops
an emphasis on **experim
for **wet-lab biologists** ai
data.

We offer three types of w

1. Short, 3-hour monthly
2. Basic Data Skills**
3. Advanced Topics: Ana

**The basic data skills

Our dedicated training team holds workshop to help
or NGS data.

o devote substantial time to material development,
raining team also participate in consultations on
best practices in NGS analysis.

ysis of high-throughput sequencing data, with
and **reproducibility**. Our workshops are designed
cing-based experiments and analysing the resulting

matics)

NGS) data**

r the advanced workshops.

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

DF/HCC
DANA-FARBER / HARVARD CANCER CENTER

HARVARD
CATALYST
THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER

HARVARD
MEDICAL SCHOOL

CFAR
HARVARD UNIVERSITY
CENTER FOR AIDS RESEARCH

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

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

Training

A key component of the H
researchers at Harvard b

HBC's training team is m
training and community b
research projects to ensu

Our hands-on workshops
an emphasis on **experim
for **wet-lab biologists** ai
data.

We offer three types of w

1. Short, 3-hour monthly
2. Basic Data Skills**
3. Advanced Topics: Ana

**The basic data skills

Our dedicated training team holds workshop to help
or NGS data.

o devote substantial time to material development,
raining team also participate in consultations on
best practices in NGS analysis.

ysis of high-throughput sequencing data, with
and **reproducibility**. Our workshops are designed
cing-based experiments and analysing the resulting

matics)

NGS) data**

r the advanced workshops.

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

DF/HCC DANA-FARBER / HARVARD CANCER CENTER

HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

HARVARD MEDICAL SCHOOL

CFAR HARVARD UNIVERSITY CENTER FOR AIDS RESEARCH

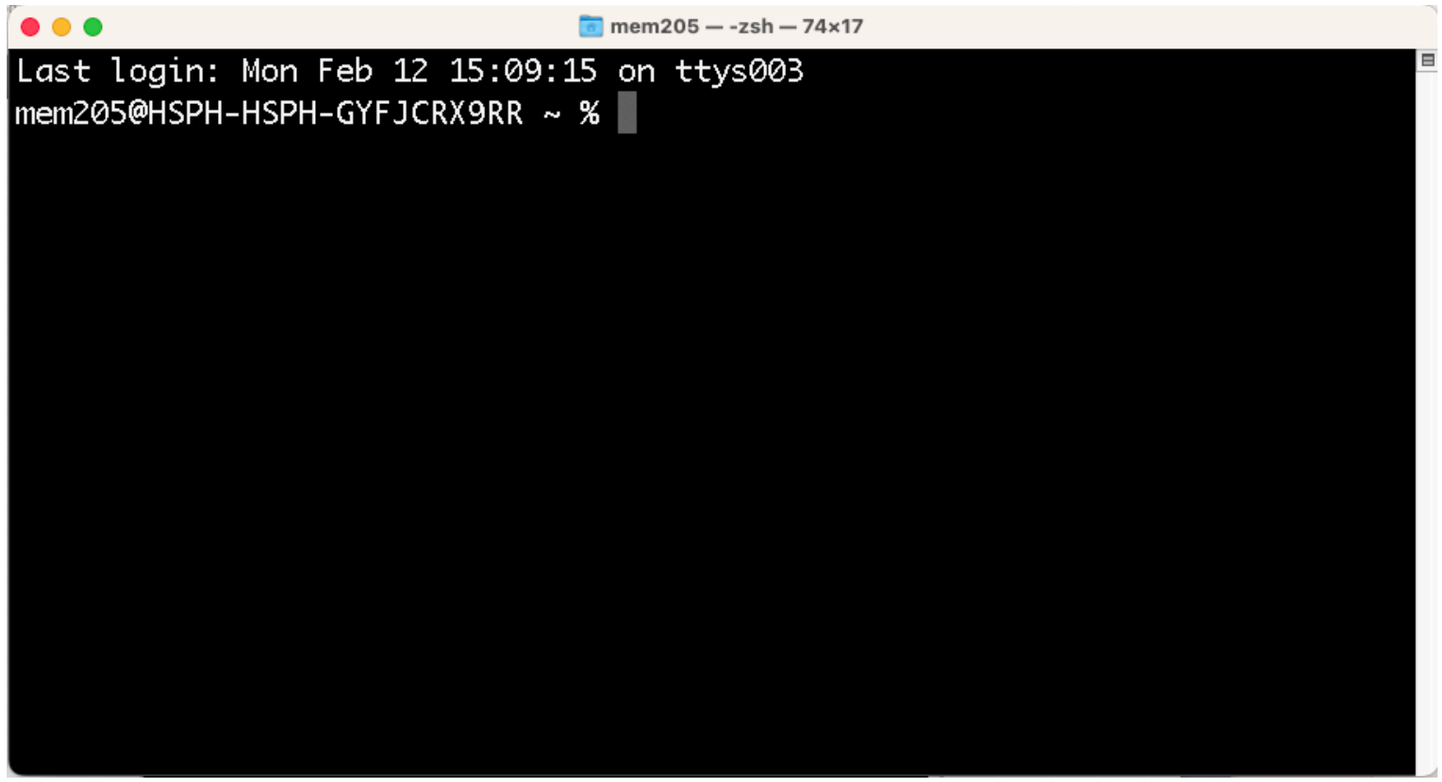
http://bioinformatics.sph.harvard.edu/training/

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

Workshop scope

What is Shell?

Shell - a program that allows users to control Unix/Linux OS with text commands



Unix /Linux - The operating systems of High performance computers

Unix /Linux - The operating systems of High performance computers

Shell - a program that allows users to control Unix/Linux OS with text commands

Unix /Linux - The operating systems of High performance computers

Shell - a program that allows users to control Unix/Linux OS with text commands

Bash - the most prevalent kind of shell

The bottom line

If you plan to process raw high throughput sequencing data yourself, you will need to learn shell.

1. You need more resources than what is available on your laptop

- Sequence data files are LARGE
- Processing these data require increased CPU and memory
- High performance compute clusters have the necessary resources!

2. Many bioinformatics tools are only available as command-line tools

3. Many HTS filetypes are binary.

- Binary files are not human readable
- Binary files need an interpreter

4. There are many useful commands that can help work with enormous data files

- Commands for easily viewing files: less, cat, head, tail

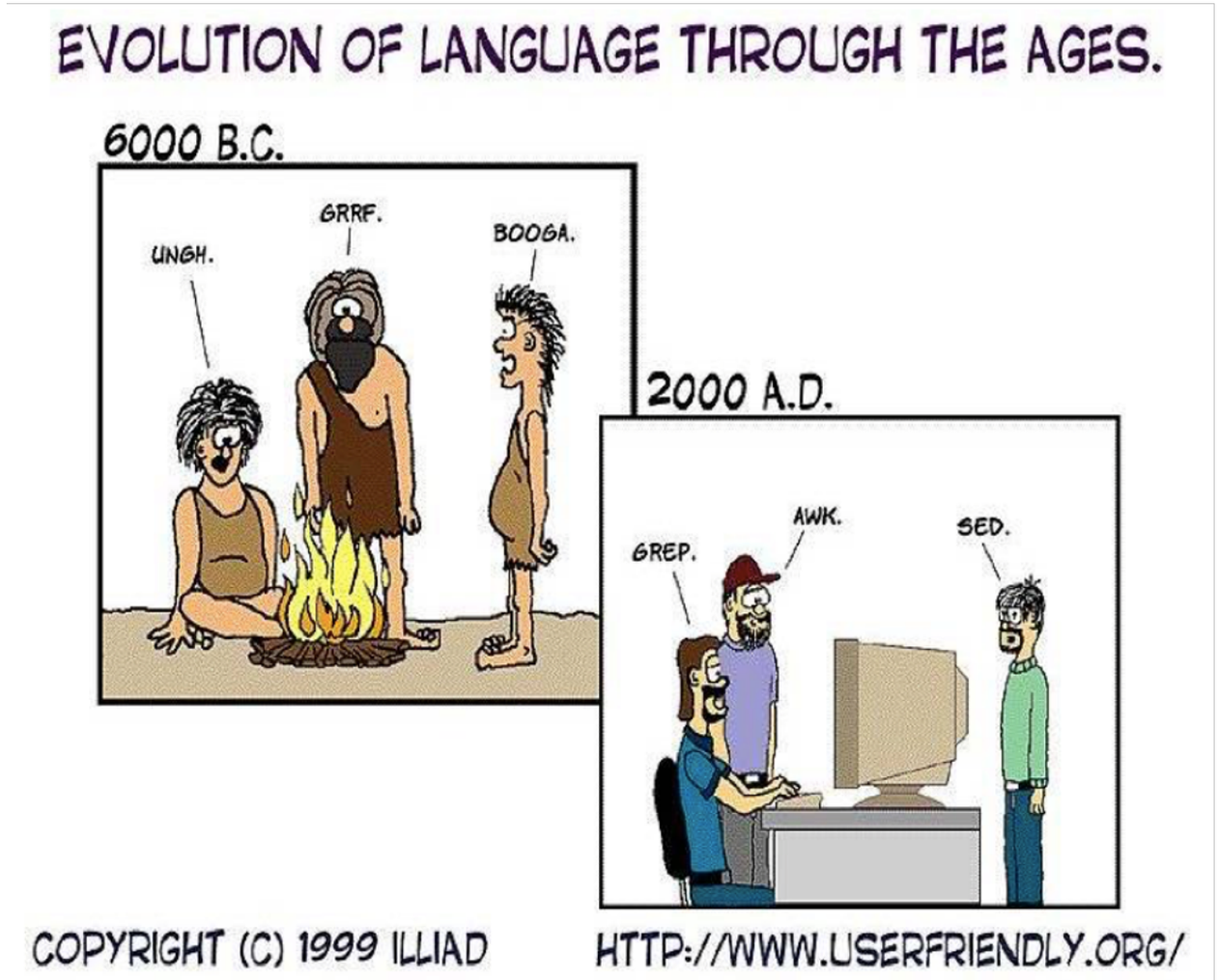
```
0  ##gff-version 3.2.1
1  ##sequence-region ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16 ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19 ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21 ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22 ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23 ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24 ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

5. Automation is the name of the game

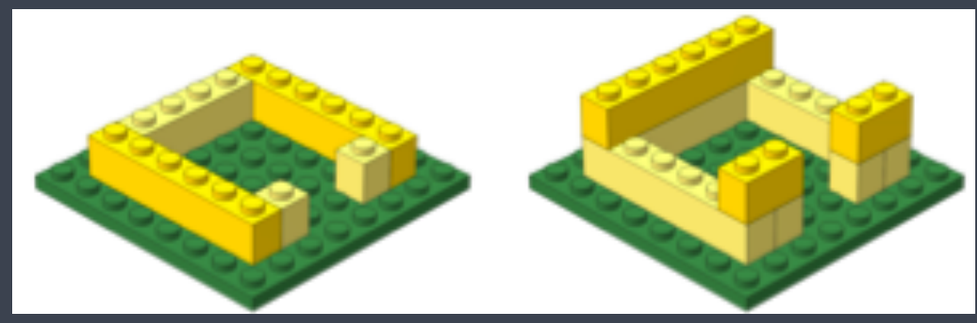
- Launch many jobs with one command
- Code is used an reused to iterate tasks over multiple files
- Parallelization to complete tasks using multiple cores and increase speed!

This could be you watching your analysis run.

6. Bonus! Maybe understand some coding jokes?



Learning Objectives



- ✓ Learn what a “shell” is and become comfortable with the command-line interface
  - Find your way around a filesystem using written commands
  - Work with small and large data files
  - Become more efficient when performing repetitive tasks
- ✓ Understand what a computational cluster is and why we need it

Logistics

Course webpage

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

Course schedule online

Workshop Schedule

Day 1

Time	Topic	Instructor
9:30 - 10:10	Workshop introduction	Meeta
10:10 - 11:40	Introduction to Shell	Mary
11:40 - 12:00	Overview of self-learning materials and homework submission	Jihe

Before the next class:

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

- Wildcards and shortcuts in Shell
- Examining and creating files
- Searching and redirection
- Shell scripts and variables in Shell

> **NOTE:** To run through the code above, you will need to be **logged into O2** and **working on a compute node** (i.e. your command prompt should have the word `compute` in it).
>
> 1. Log in using `ssh rc_trainingXX@o2.hms.harvard.edu` and enter your password (replace the "XX" in the username with the number you were assigned in class).
> 2. Once you are on the login node, use `srun --pty -p interactive -t 0-2:30 --mem 1G`

Course materials online

Learning Objectives

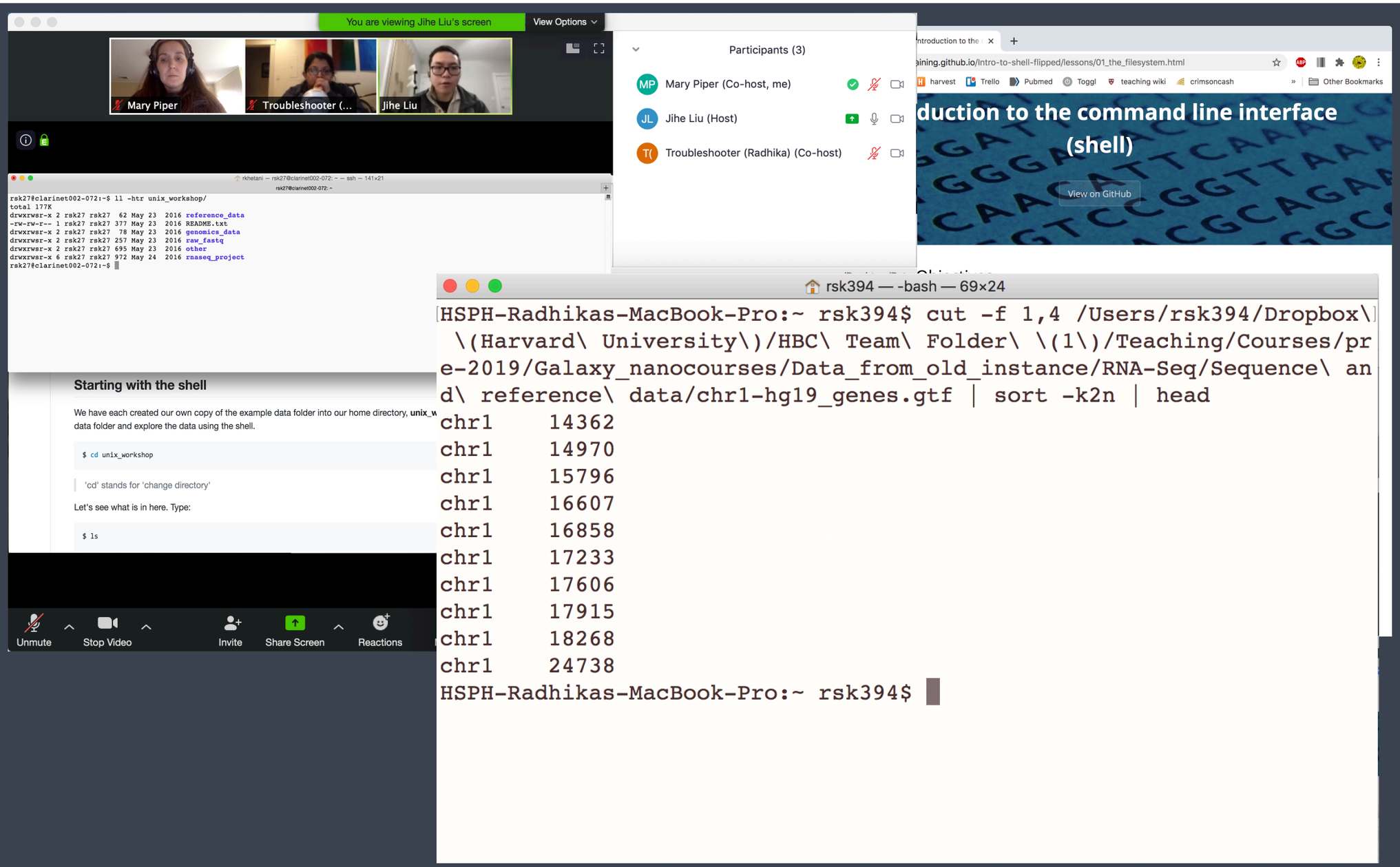
- How do you access the shell?
- How do you use it?
  - Getting around the Unix file system
  - looking at files
  - manipulating files
  - automating tasks
- What is it good for?

Setting up

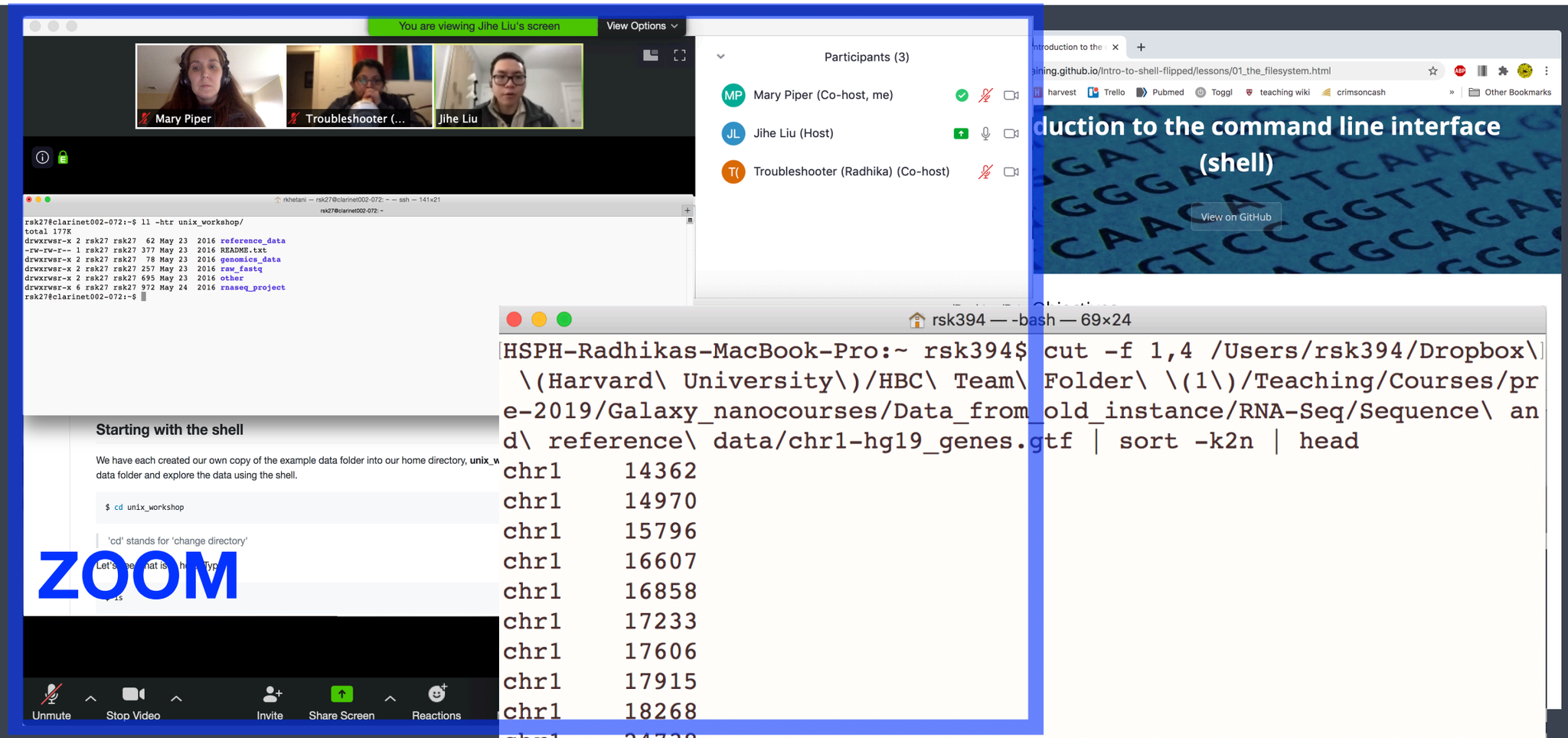
We will spend most of our time learning about the basics of the shell command-line interface (CLI) by exploring experimental data on the **O2** cluster. So, we will need to log in to this remote compute cluster first before we can start with the basics.

Let's take a quick look at the basic architecture of a cluster environment and some cluster-specific jargon prior to logging in.

Single screen & 3 windows?

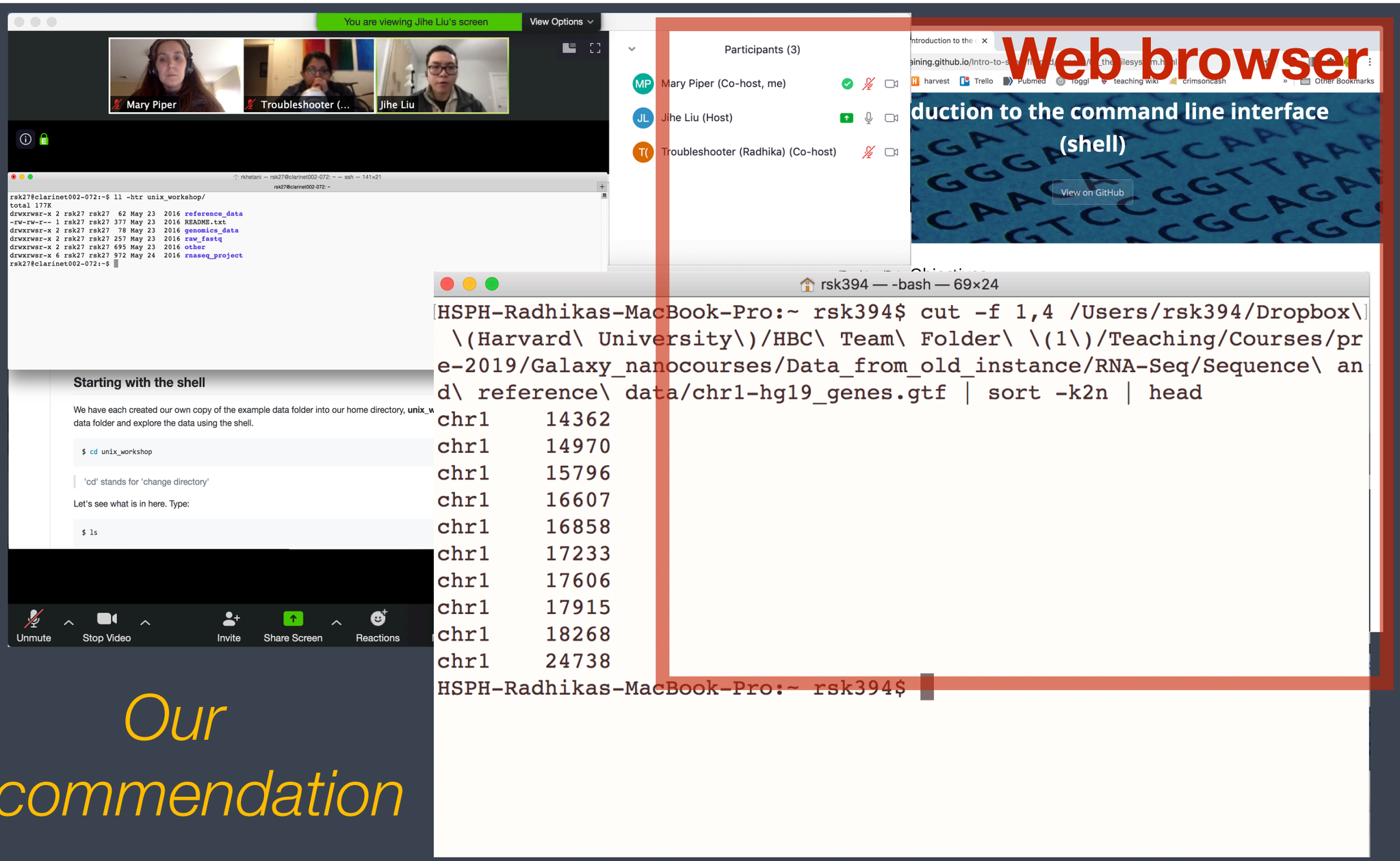


Single screen & 3 windows?



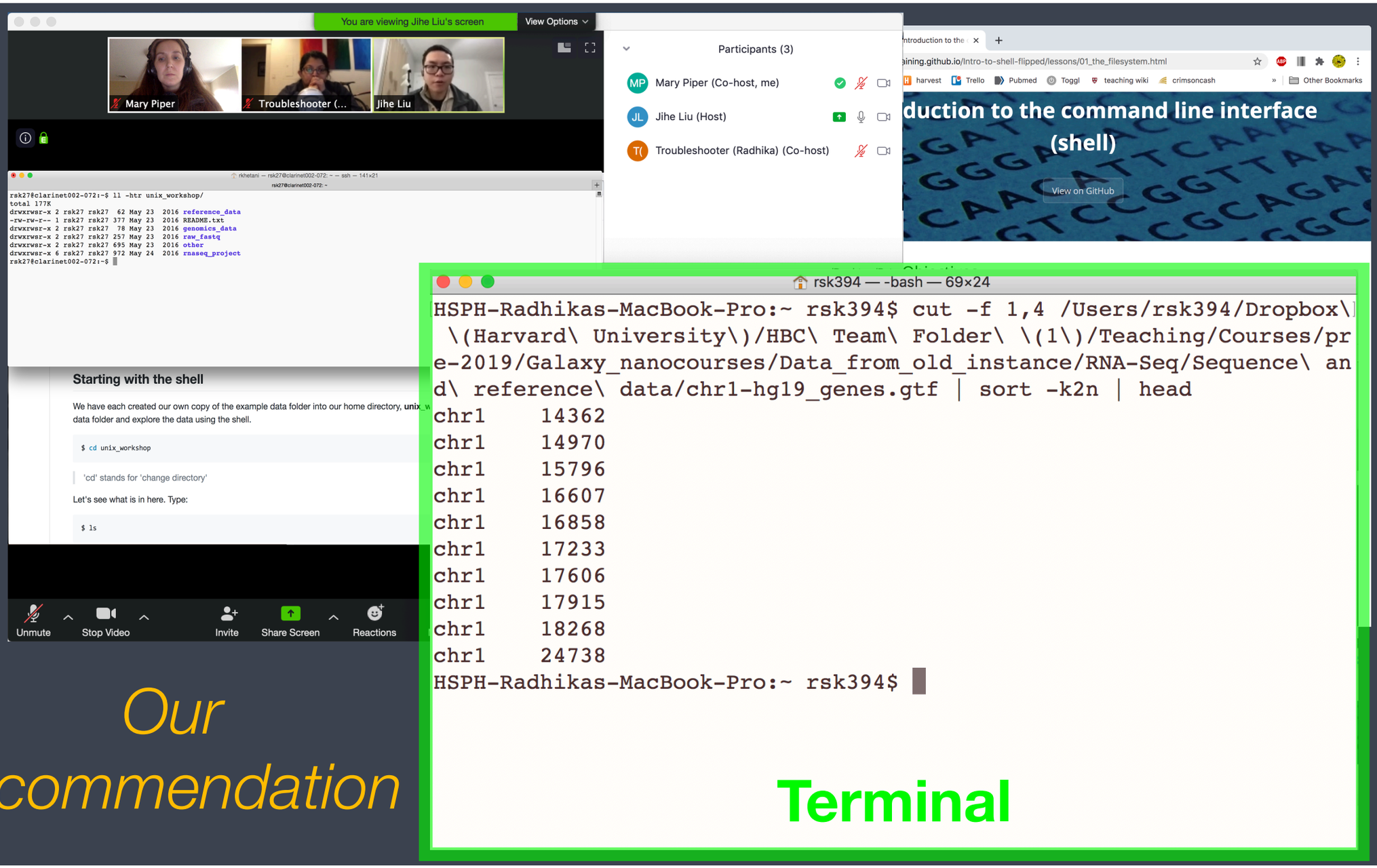
Our recommendation

Single screen & 3 windows?



Our recommendation

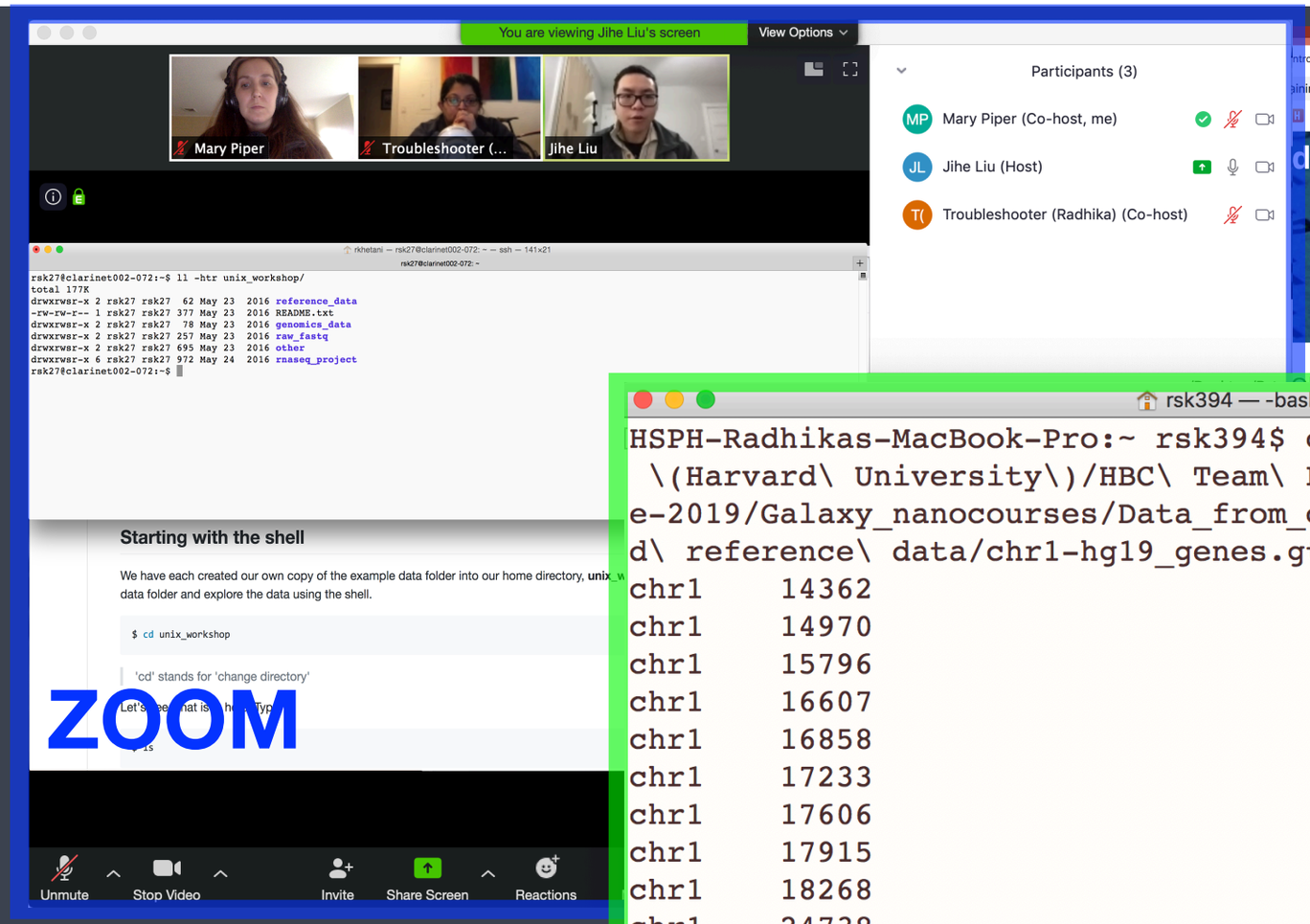
Single screen & 3 windows?

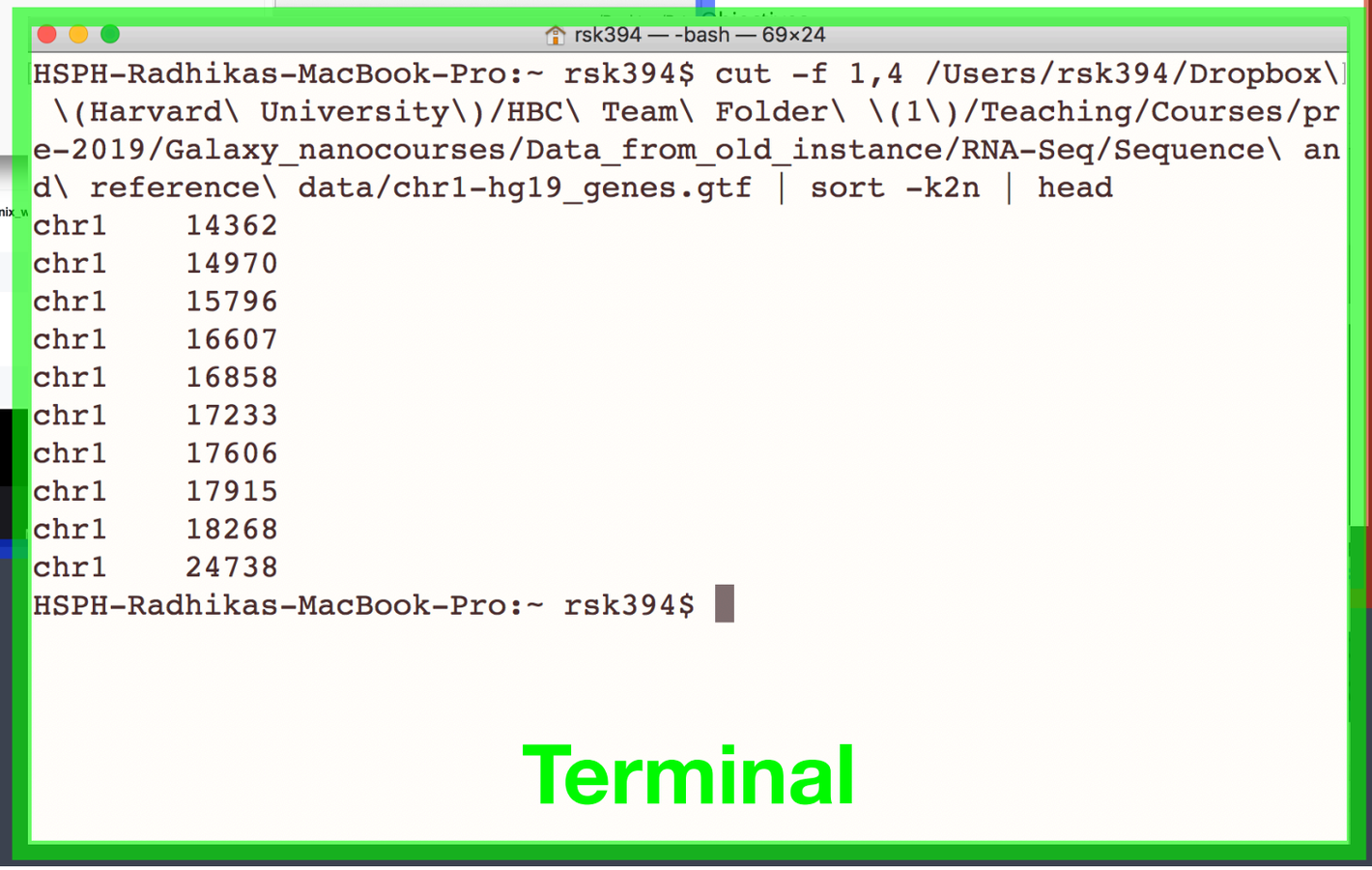


Our recommendation

Terminal

Single screen & 3 windows?





Our recommendation

Course participation

- Please keep your videos on, we would love to see your faces!
- 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!

Homework and Expectations

- 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

Odds and Ends (1/2)

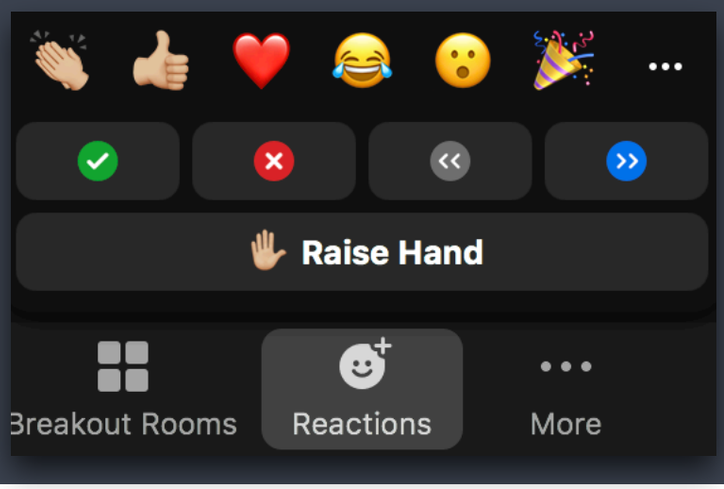
- Quit/minimize all applications that are not required for class

Odds and Ends (1/2)

- Quit/minimize all applications that are not required for class
- Captioning is available upon request

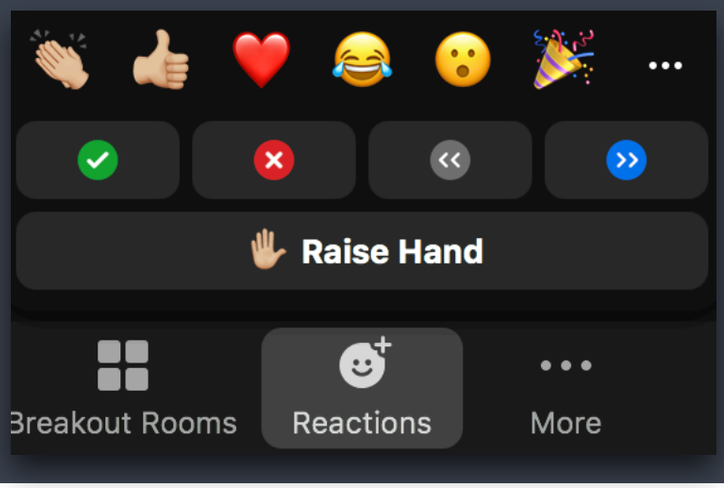
Odds and Ends (1/2)

- Quit/minimize all applications that are not required for class
- Captioning is available upon request
- 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 (2/2)

- Questions for the presenter?
  - Post the question in the Chat window OR
  - ✋ **Raise Hand** when the presenter asks for questions
  - Let the **Moderator** 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

Contact us!

HBC training team: hbctraining@hsph.harvard.edu

O2 (HMS-RC): rchelp@hms.harvard.edu

HBC consulting: bioinformatics@hsph.harvard.edu

Twitter

HBC: @bioinfocore

HMS-RC: @hms_rc