HBC
Harvard Chan Bioinformatics Core

# Understanding chromatin biology using high throughput sequencing methods

https://tinyurl.com/hbc-chipseq

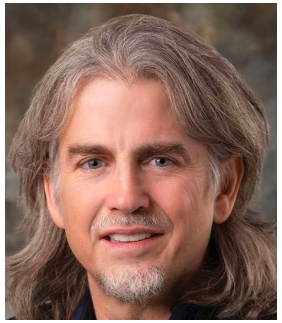
Harvard Chan Bioinformatics Core
in collaboration with
HMS Research Computing

# Introductions!

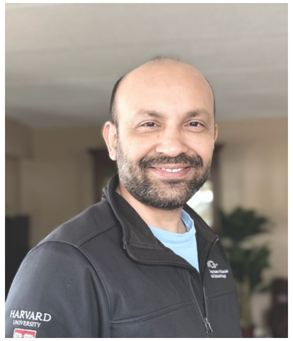
Shannan Ho Sui
*Director*

Meeta Mistry
*Associate Director*

Lorena Pantano
*Director of Bioinformatics Platform*



John Quackenbush
*Faculty Advisor*



Upen Bhattarai

Heather Wick

Will Gammerdinger

Noor Sohail

Elizabeth Partan

Alex Bartlett

Emma Berdan

James Billingsley

Zhu Zhuo

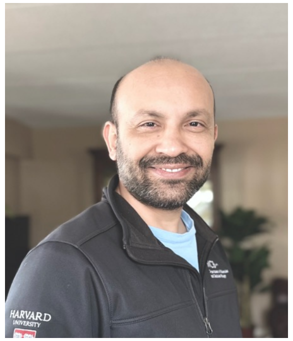
Maria Simoneau

Shannan Ho Sui
*Director*

Meeta Mistry
*Associate Director*

Lorena Pantano
*Director of Bioinformatics Platform*

John Quackenbush
*Faculty Advisor*



Upen Bhattarai

Heather Wick

Will Gammerdinger

Noor Sohail

Elizabeth Partan

Alex Bartlett

Emma Berdan

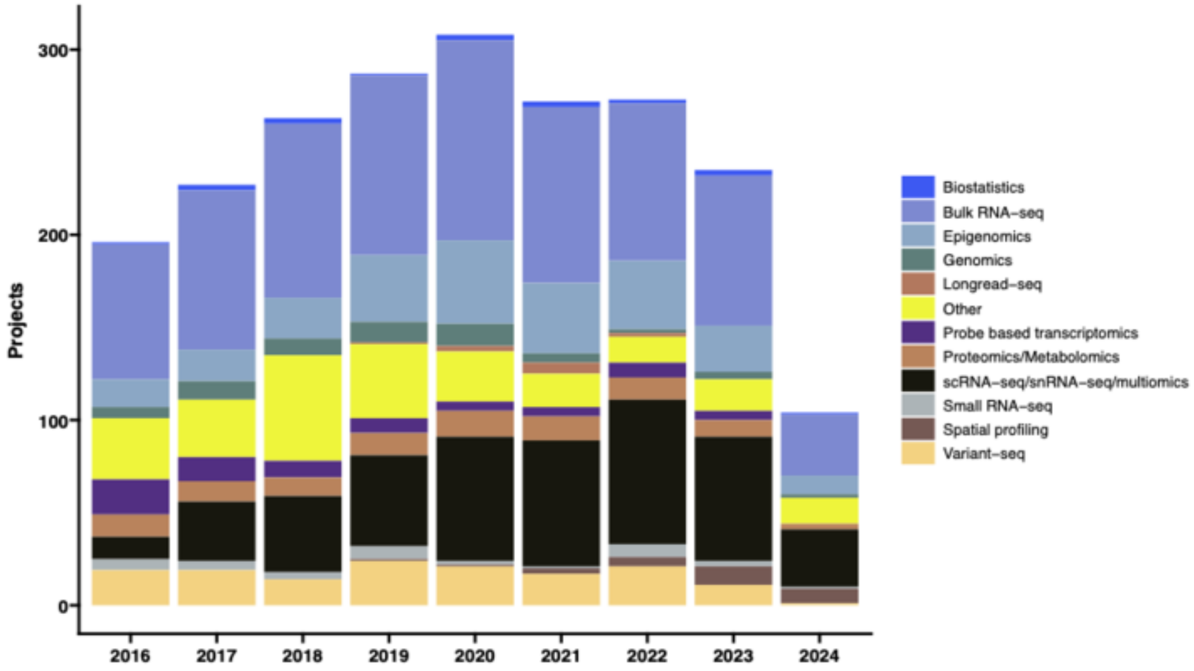
James Billingsley

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



https://bioinformatics.sph.harvard.edu/services

# 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

CFAR
HARVARD UNIVERSITY
CENTER FOR AIDS RESEARCH

VE RI TAS
HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NIEHS

HARVARD
CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

VE RI TAS
HARVARD
MEDICAL SCHOOL

# Training

- Hands-on workshops design to reflect best practices, reproducibility and an emphasis on experimental design
  - Basic Data Skills
    - Shell
    - R
  - Advanced Topics: Analysis of high-throughput sequencing data
    - Chromatin Biology
    - Bulk RNA-seq
    - Differential Gene Expression
    - scRNA-seq
    - Variant Calling
  - Current Topics in Bioinformatics

**https://bioinformatics.sph.harvard.edu/training**

# Training

- Hands-on workshops design to reflect best practices, reproducibility and an emphasis on experimental design
  - Basic Data Skills
    - Shell
    - R
  - Advanced Topics: Analysis of high-throughput sequencing data
    - Chromatin Biology
    - Bulk RNA-seq
    - Differential Gene Expression
    - scRNA-seq
    - Variant Calling
  - Current Topics in Bioinformatics

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

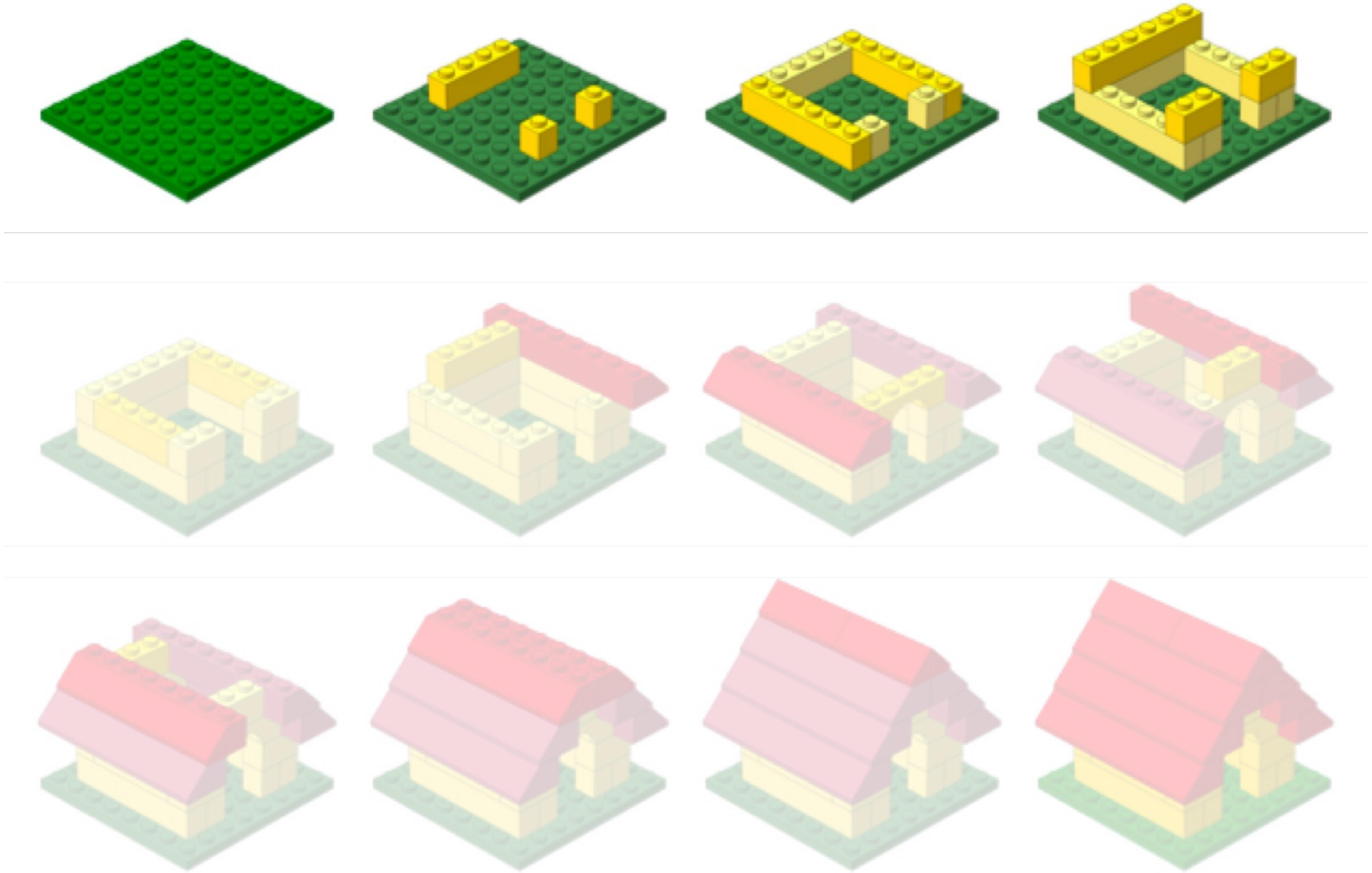
CFAR
HARVARD UNIVERSITY
CENTER FOR AIDS RESEARCH

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

DF/HCC
DANA-FARBER / HARVARD CANCER CENTER

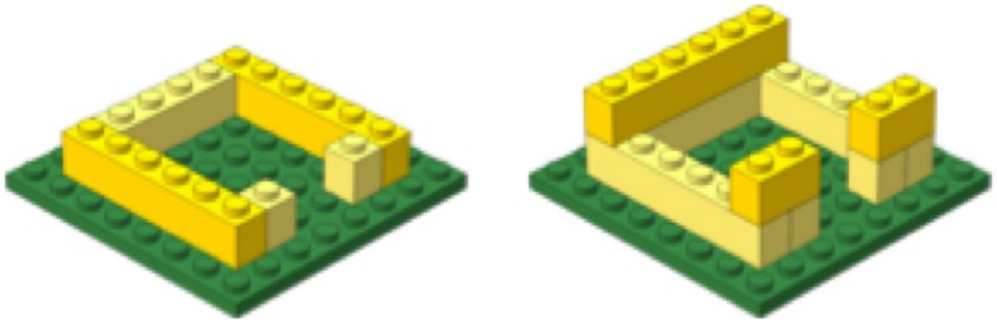
HARVARD
CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

HARVARD
MEDICAL SCHOOL



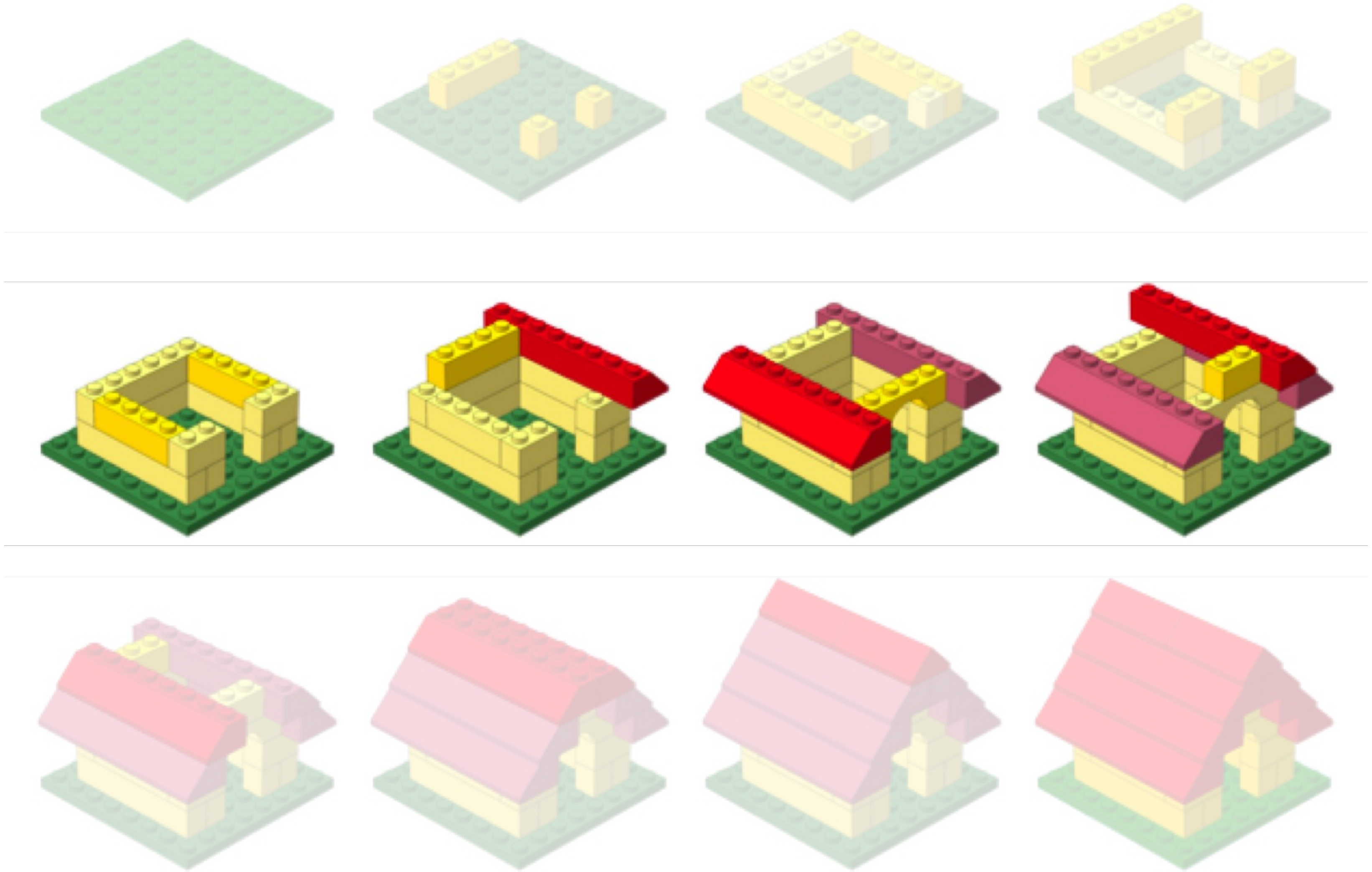
Setting up to perform Bioinformatics analysis

# 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
- Introduction to R
  - Parsing and working with smaller BED files
  - Statistical analysis, e.g. differential binding analysis
  - Generating figures from complex data

# Workshop scope



Bioinformatic Data Analysis

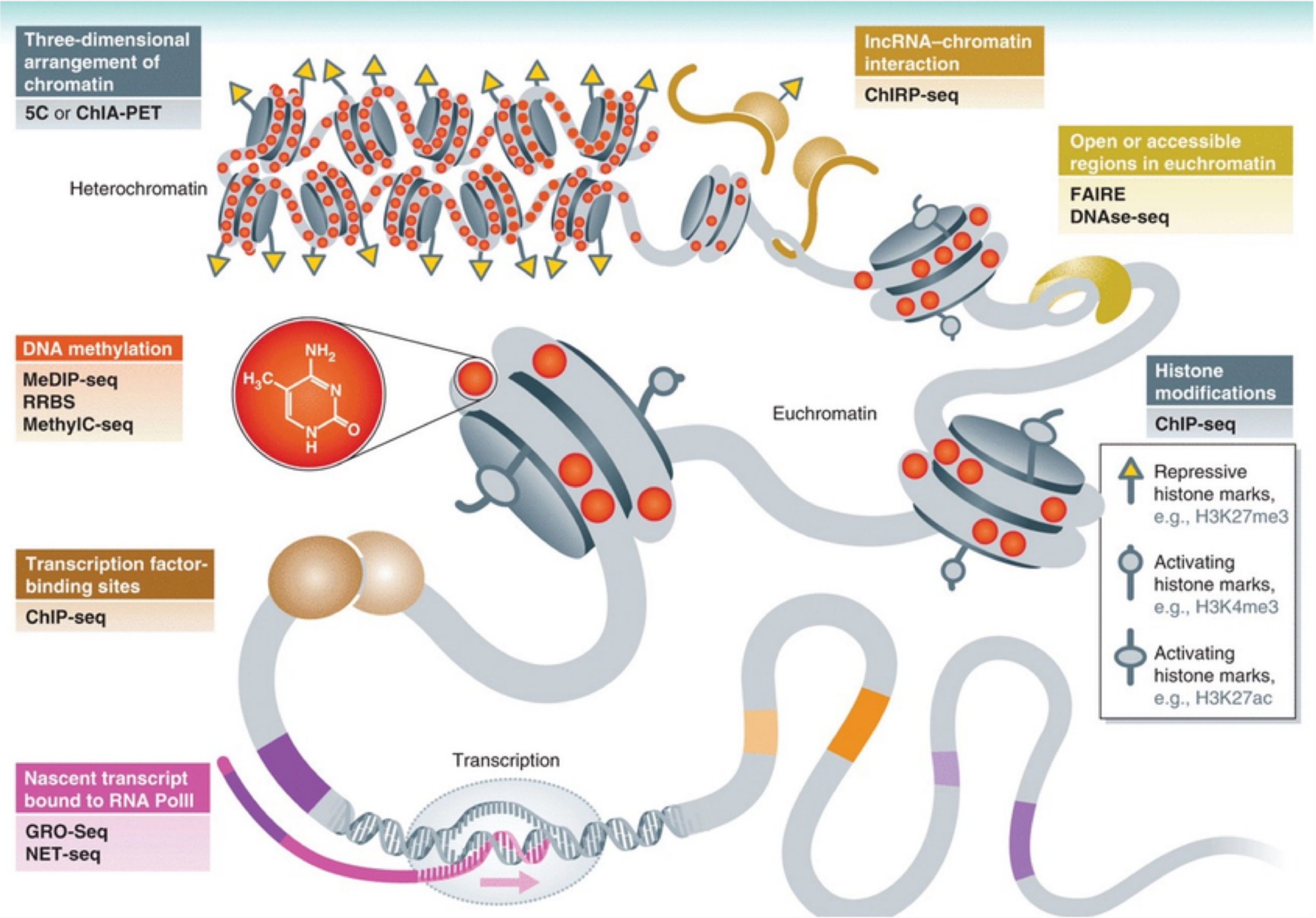
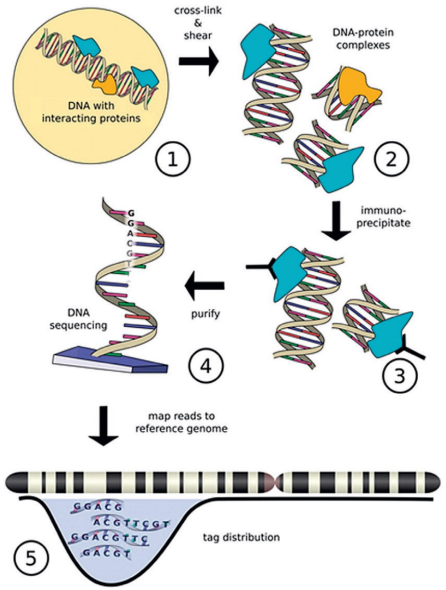

Three-dimensional arrangement of chromatin
5C or ChIA-PET
Heterochromatin
lncRNA–chromatin interaction
ChIRP-seq
Open or accessible regions in euchromatin
FAIRE
DNAse-seq
DNA methylation
MeDIP-seq
RRBS
MethylC-seq
NH2
H3C
N
O
N
H
Euchromatin
Histone modifications
ChIP-seq
Repressive histone marks, e.g., H3K27me3
Activating histone marks, e.g., H3K4me3
Activating histone marks, e.g., H3K27ac
Transcription factor-binding sites
ChIP-seq
Transcription
Nascent transcript bound to RNA PolII
GRO-Seq
NET-seq

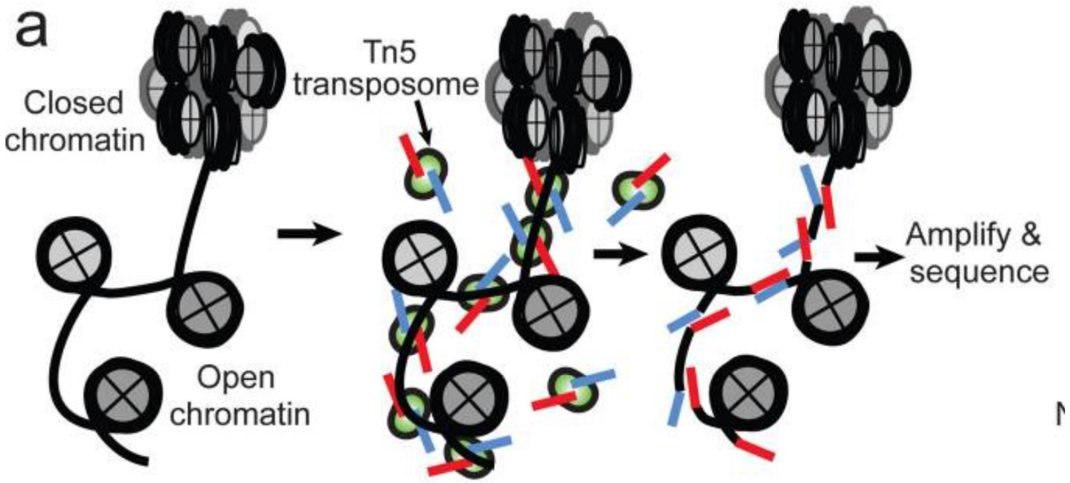
Figure adapted from Soon WW, Hariharan M, Snyder MP, “High throughput sequencing for biology and medicine”. Molecular Systems Biology 9:640 2013

# Genomic Methods for Profiling Chromatin

## ChIP-seq



## ATAC-seq



## CUT&RUN

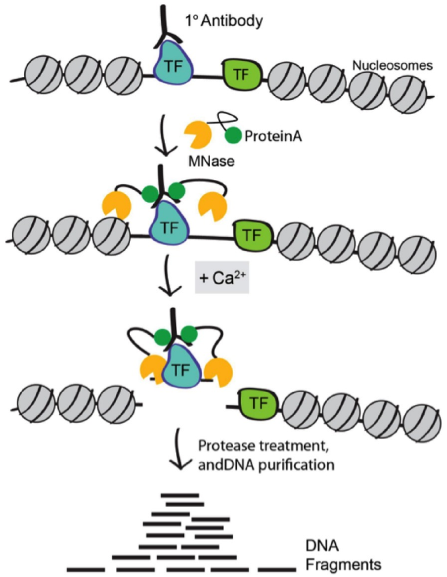
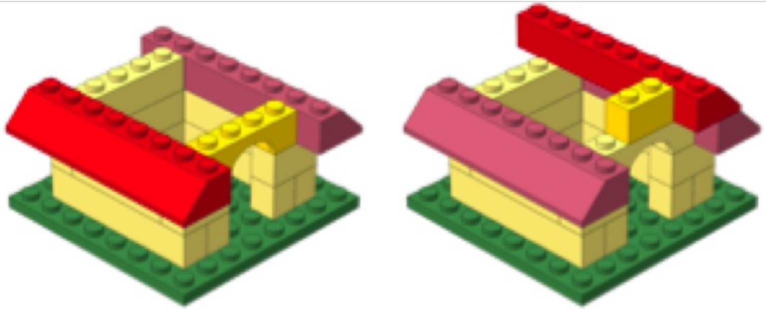


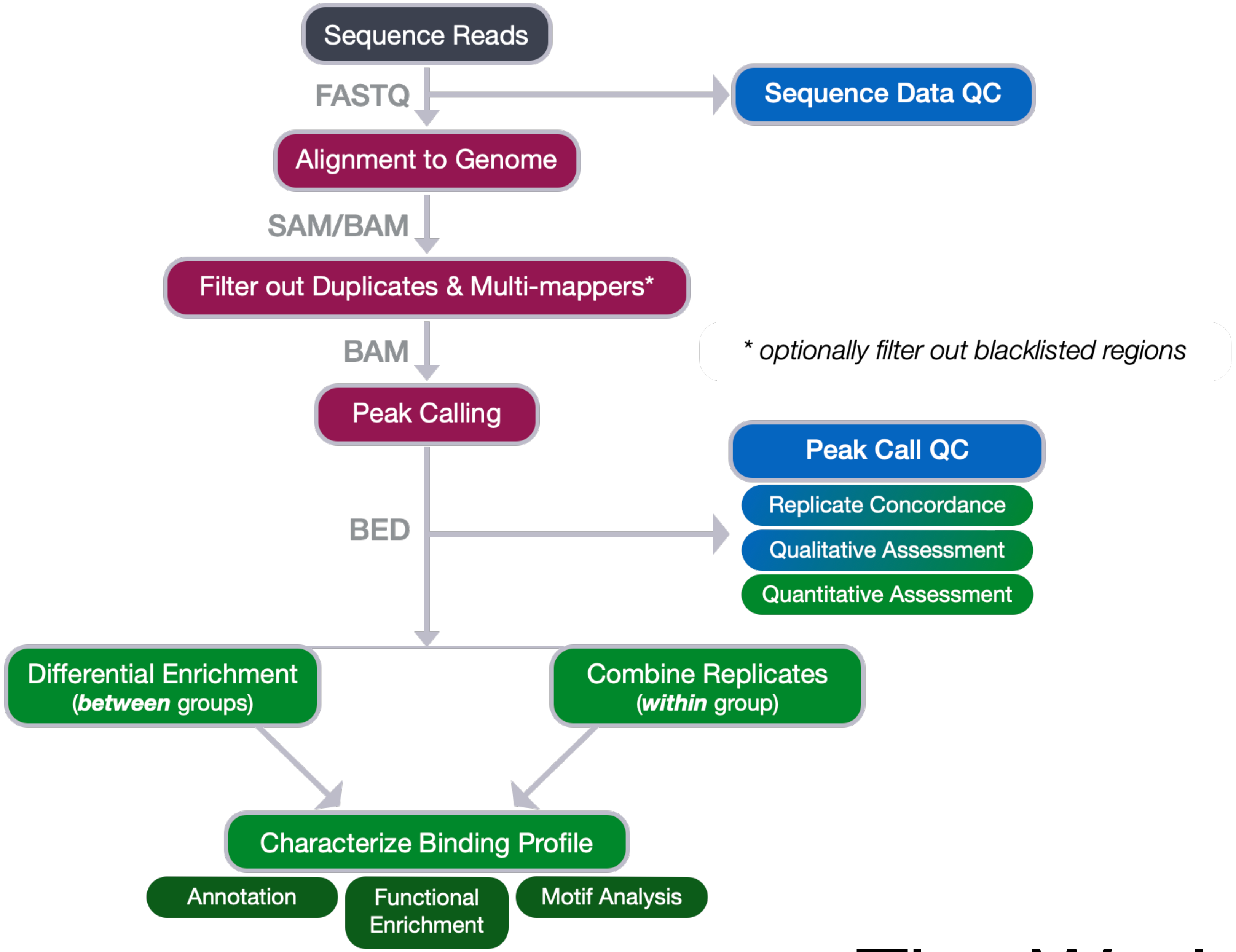
*Figure 1. CUT&RUN schematic (see text for details).*

# Learning Objectives



- Describe important considerations for setting up a successful ChIP-seq, CUT&RUN or ATAC-seq experiment
- Describe the steps in an ChIP-seq analysis workflow (from sequence data to peak calls) and contrast any differences for CUT&RUN and ATAC-seq analyses
- Learn how to handle various file formats encountered when analyzing ChIP-seq and related data
- Implement shell scripts on a high-performance compute cluster to perform the above steps

# The Workflow



Sequence Reads

FASTQ

Sequence Data QC

Alignment to Genome

SAM/BAM

Filter out Duplicates & Multi-mappers*

BAM

* *optionally filter out blacklisted regions*

Peak Calling

BED

Peak Call QC

Replicate Concordance

Qualitative Assessment

Quantitative Assessment

Differential Enrichment
(***between*** groups)

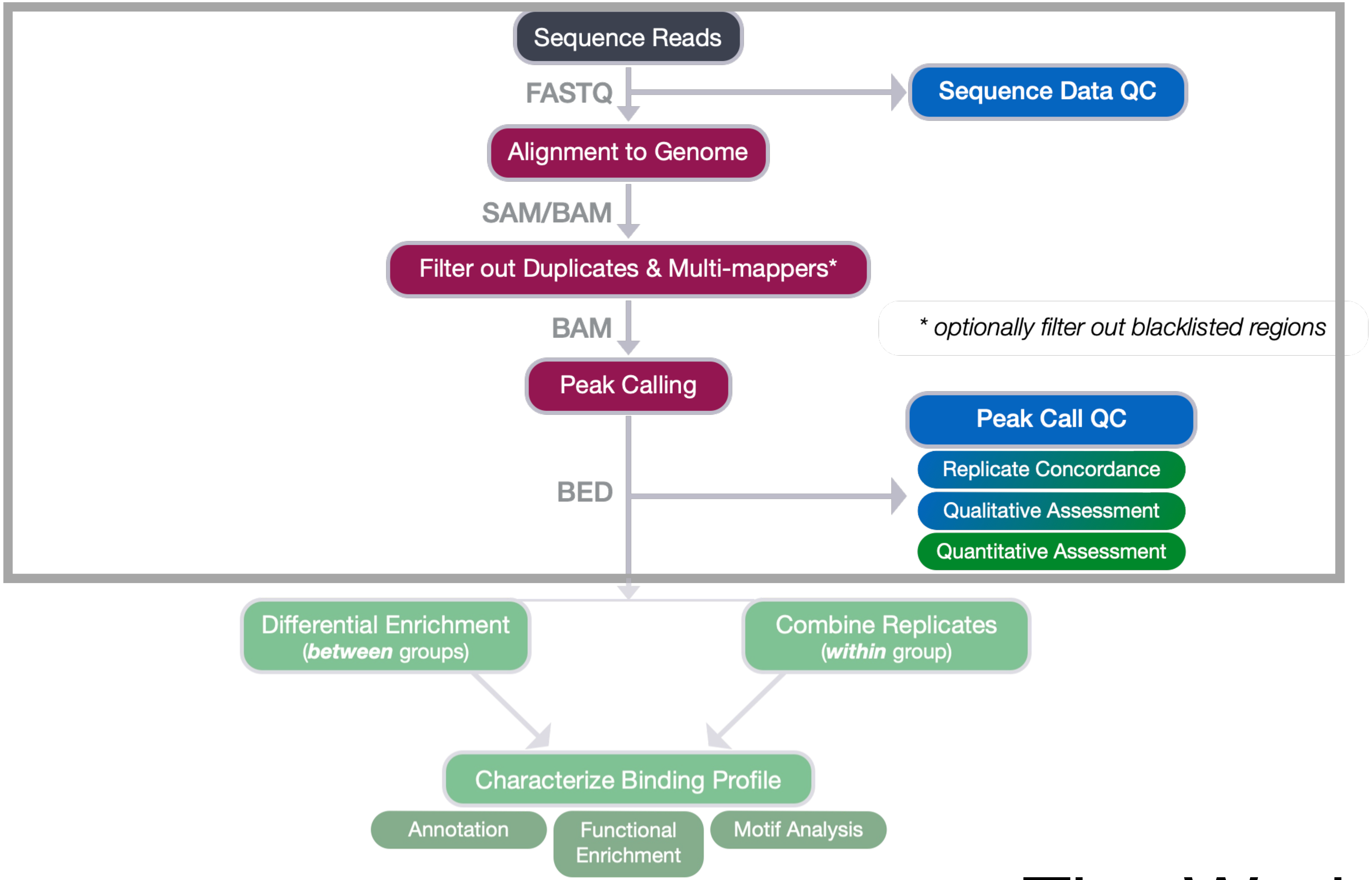
Combine Replicates
(***within*** group)

Characterize Binding Profile

Annotation

Functional Enrichment

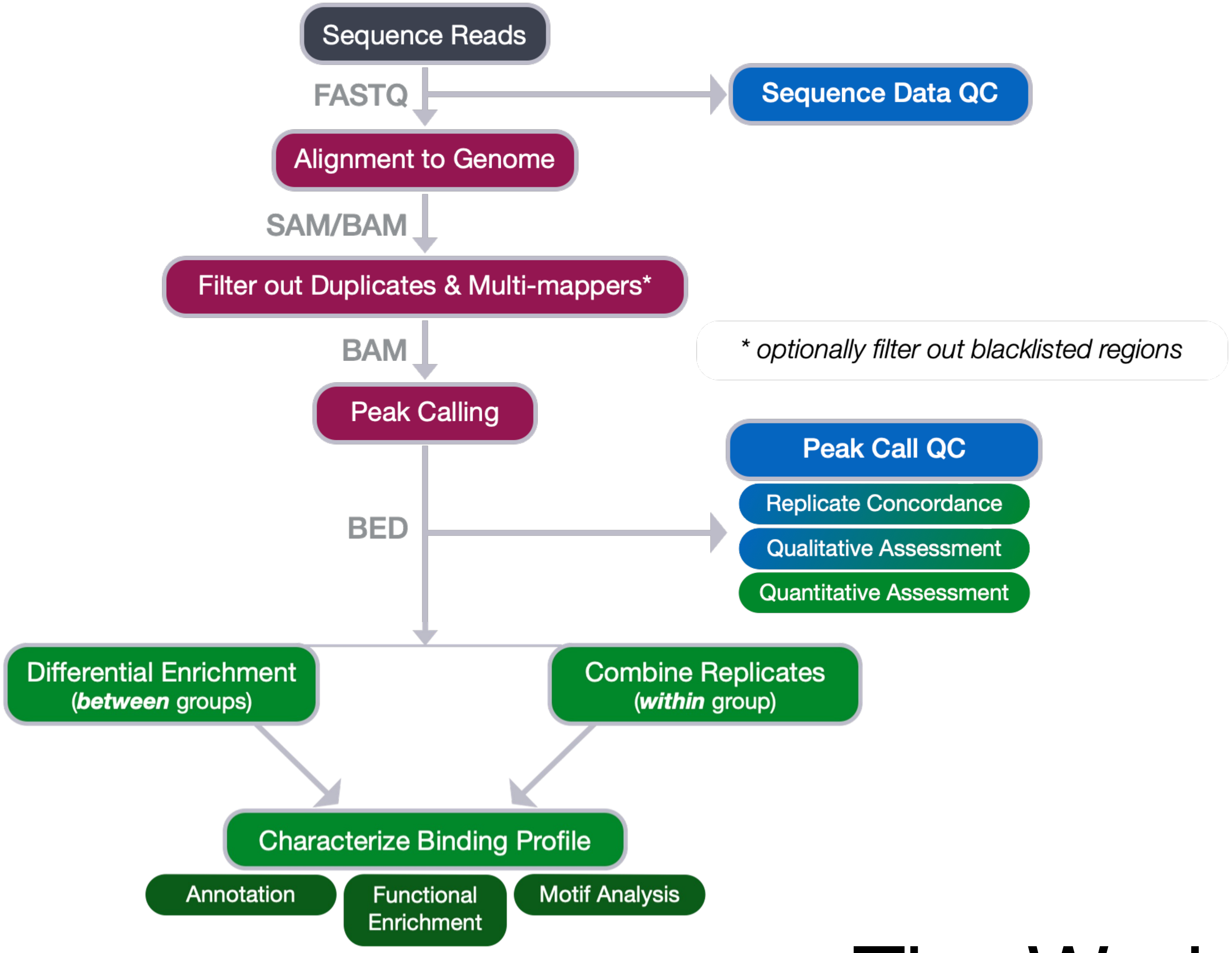
Motif Analysis



# The Workflow

*Boxes in green represent parts of the workflow that will not be covered in this workshop*

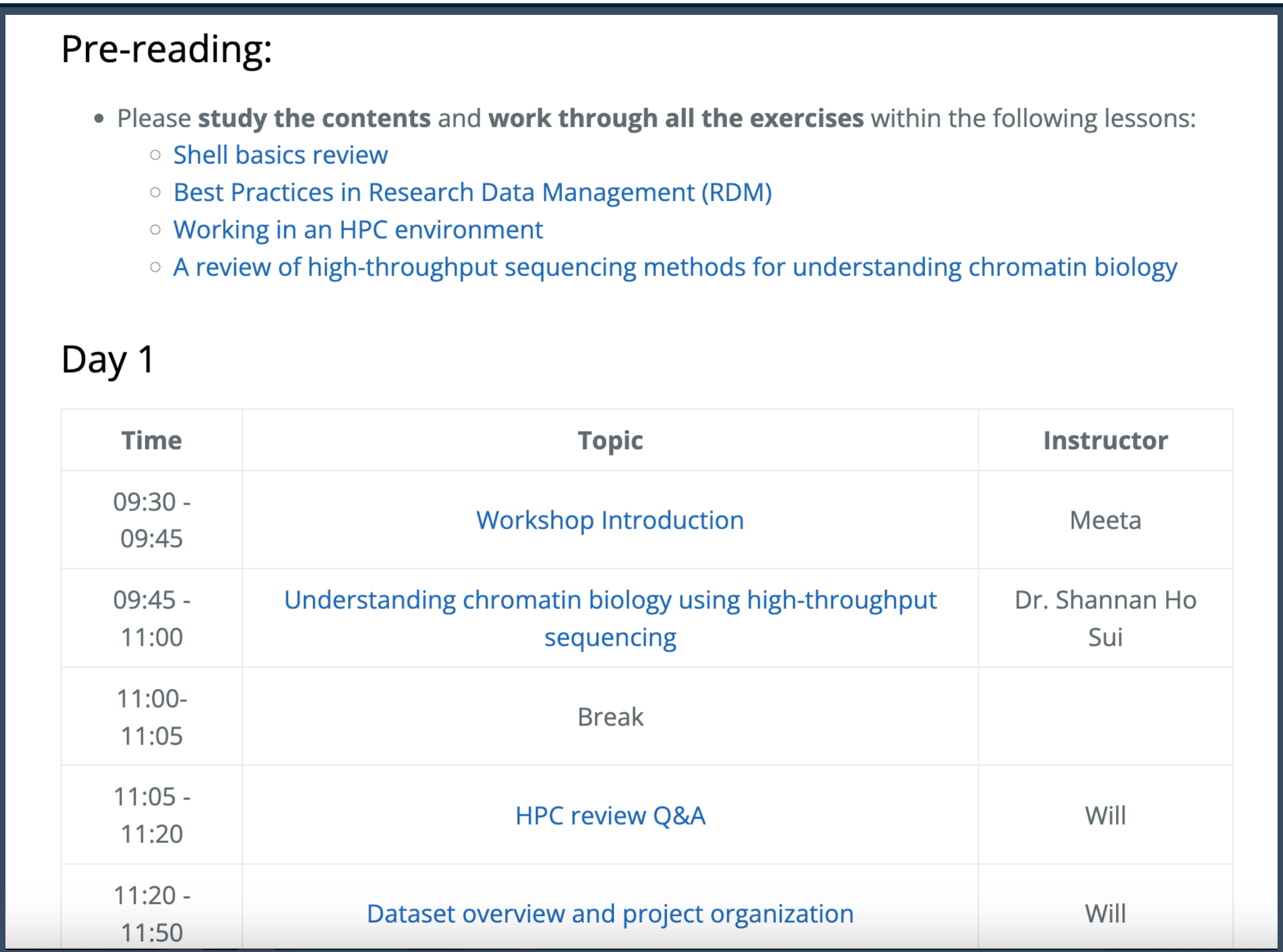
# The Workflow



*Boxes in green represent parts of the workflow that will not be covered in this workshop*

# Logistics

# Course schedule



## Pre-reading:

- Please **study the contents** and **work through all the exercises** within the following lessons:
  - Shell basics review
  - Best Practices in Research Data Management (RDM)
  - Working in an HPC environment
  - A review of high-throughput sequencing methods for understanding chromatin biology

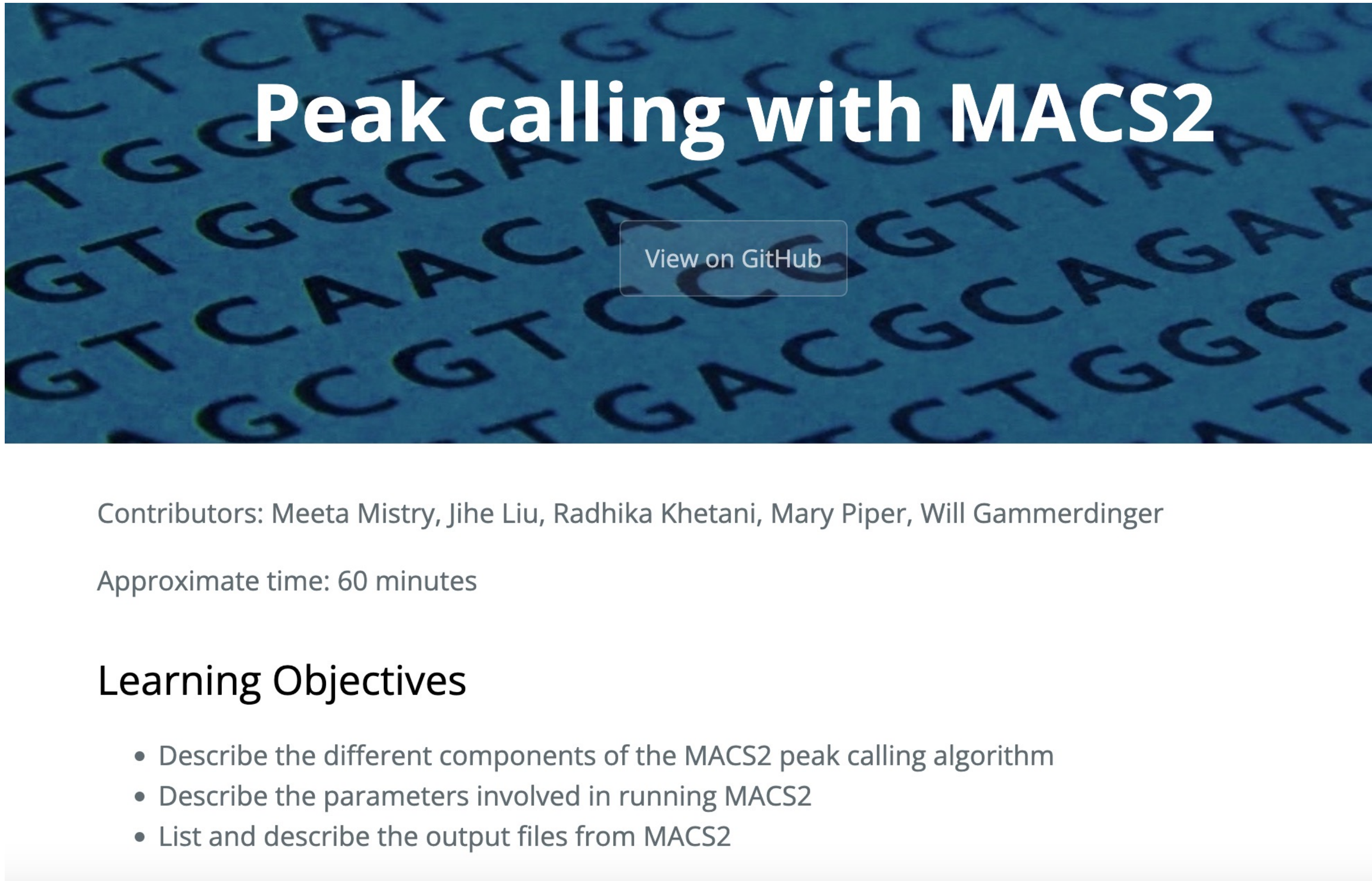
## Day 1

| Time | Topic | Instructor |
|---|---|---|
| 09:30 - 09:45 | Workshop Introduction | Meeta |
| 09:45 - 11:00 | Understanding chromatin biology using high-throughput sequencing | Dr. Shannan Ho Sui |
| 11:00- 11:05 | Break | |
| 11:05 - 11:20 | HPC review Q&A | Will |
| 11:20 - 11:50 | Dataset overview and project organization | Will |

https://tinyurl.com/hbc-chipseq

# Course materials

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



# Peak calling with MACS2

Contributors: Meeta Mistry, Jihe Liu, Radhika Khetani, Mary Piper, Will Gammerdinger

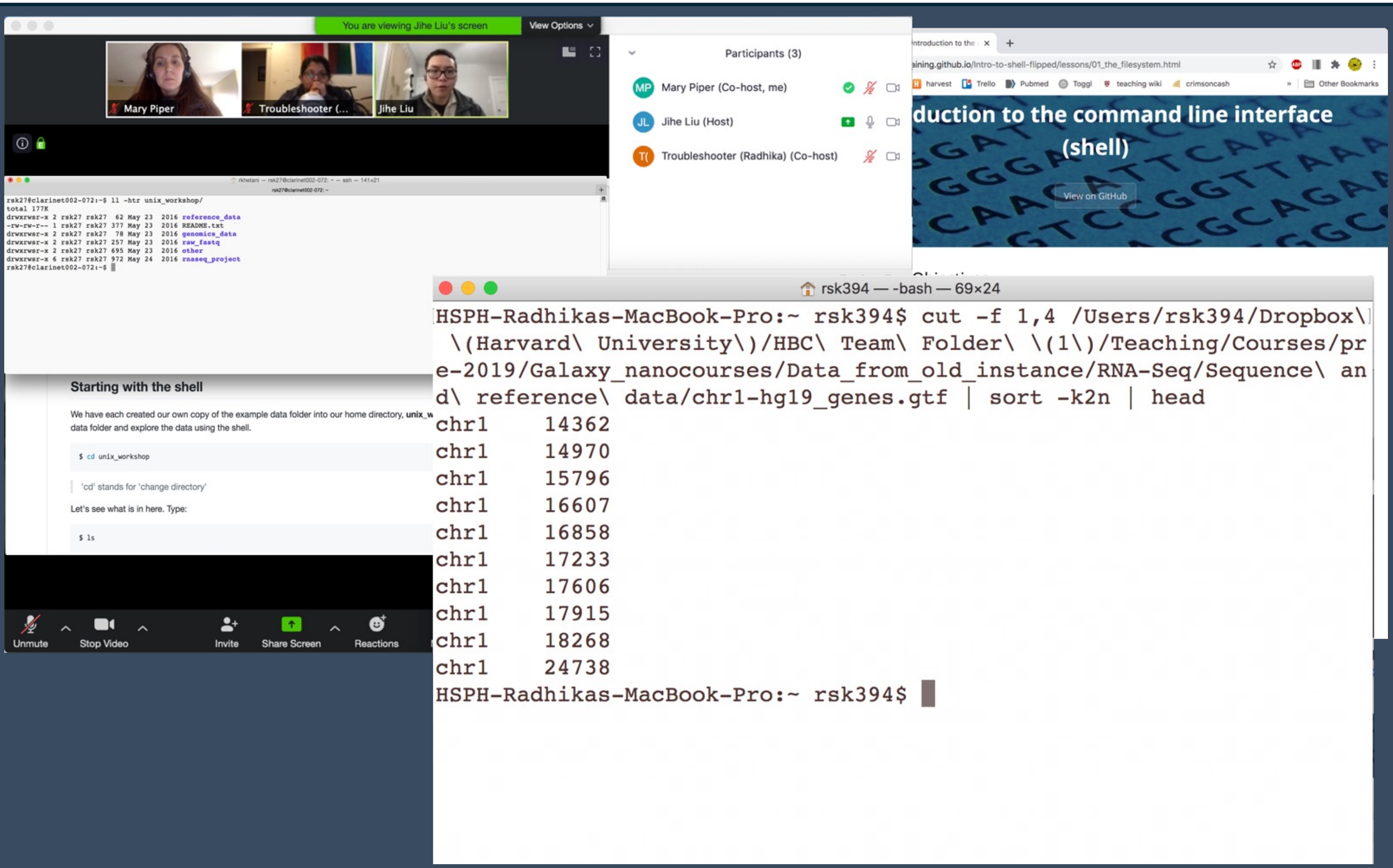
Approximate time: 60 minutes

## Learning Objectives

- Describe the different components of the MACS2 peak calling algorithm
- Describe the parameters involved in running MACS2
- List and describe the output files from MACS2

https://tinyurl.com/hbc-chipseq

# Single Screen & 3 Windows



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

### 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
```

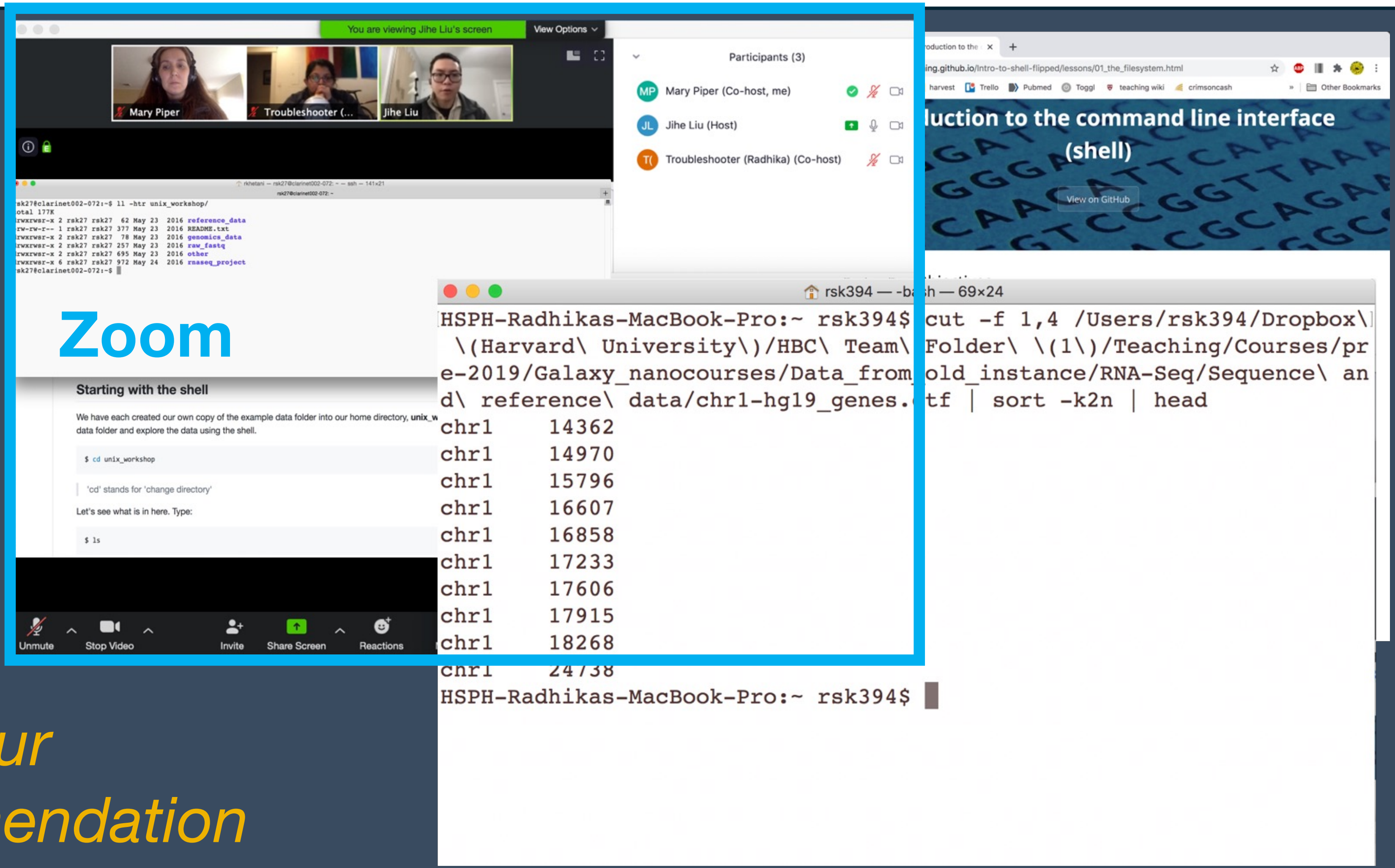
> 'cd' stands for 'change directory'

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.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



```
sk27@clarinet002-072:~$ ll -htr unix_workshop/
otal 177K
rwxrwsr-x 2 rsk27 rsk27  62 May 23  2016 reference_data
rw-rw-r-- 1 rsk27 rsk27 377 May 23  2016 README.txt
rwxrwsr-x 2 rsk27 rsk27  78 May 23  2016 genomics_data
rwxrwsr-x 2 rsk27 rsk27 257 May 23  2016 raw_fastq
rwxrwsr-x 2 rsk27 rsk27 695 May 23  2016 other
rwxrwsr-x 6 rsk27 rsk27 972 May 24  2016 rnaseq_project
sk27@clarinet002-072:~$
```

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'

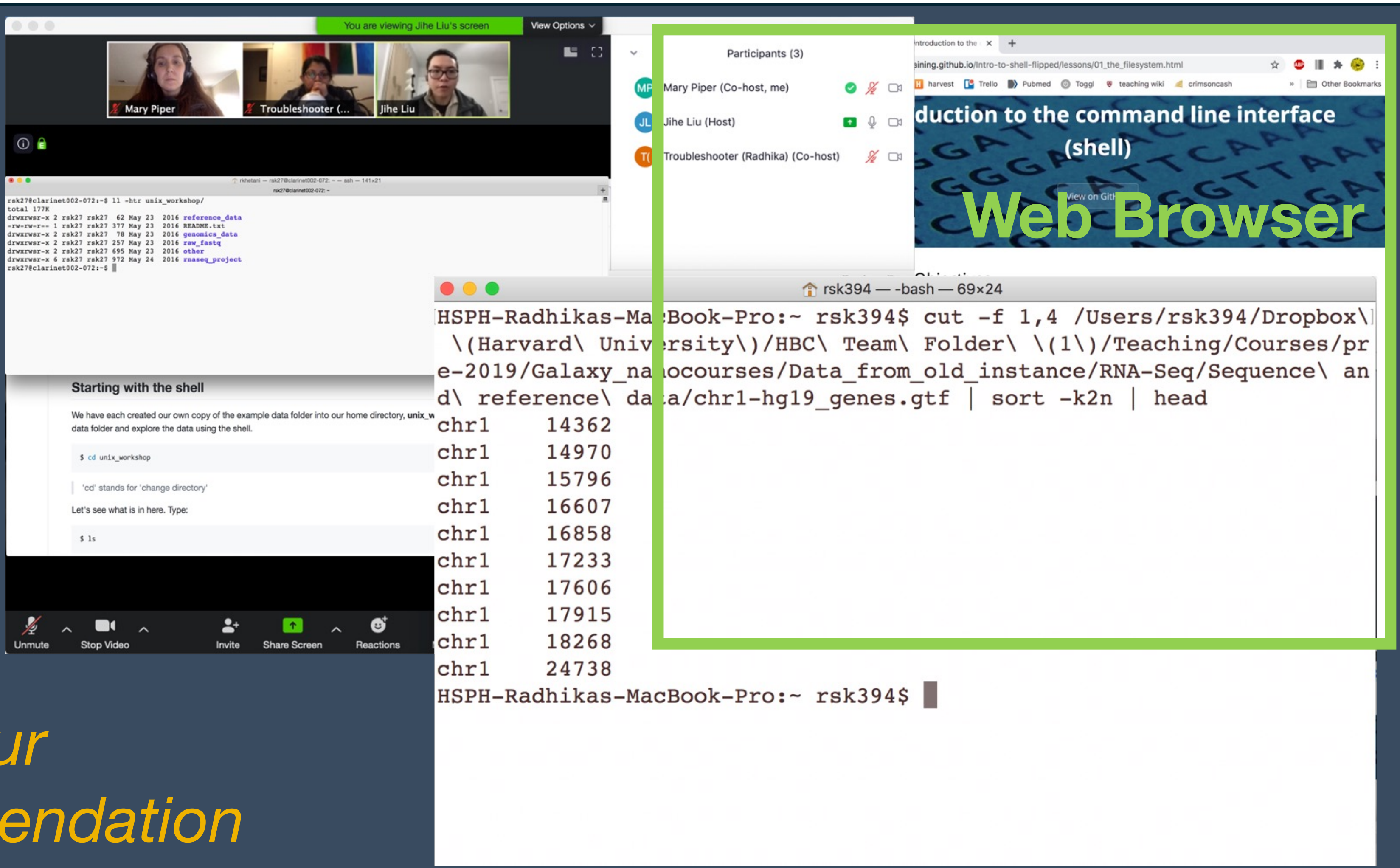
Let's see what is in here. Type:

```
$ 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.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*

# Single Screen & 3 Windows



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

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'

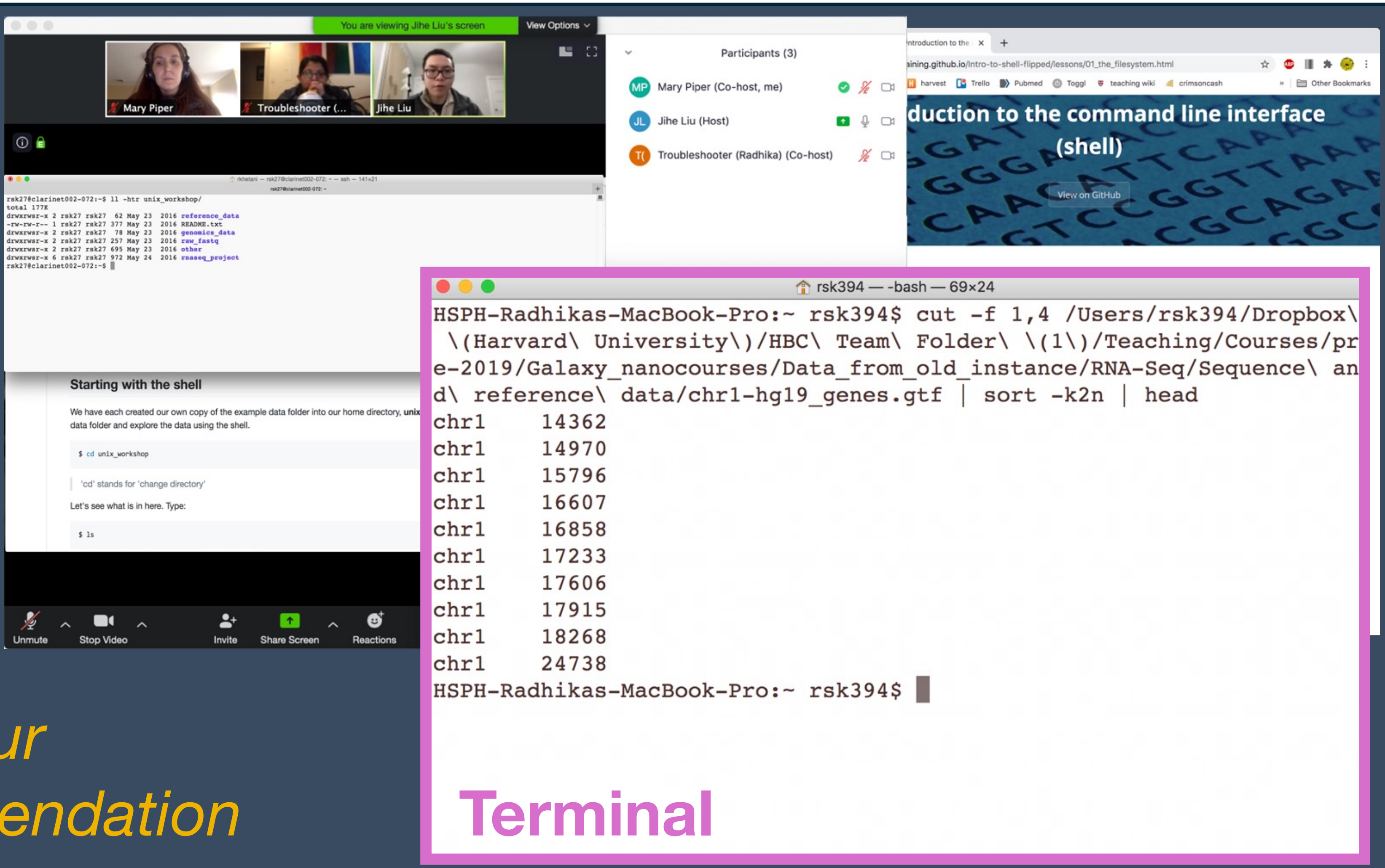
Let's see what is in here. Type:

```
$ 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.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*

# Single Screen & 3 Windows



duction to the command line interface (shell)

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

Starting with the shell

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

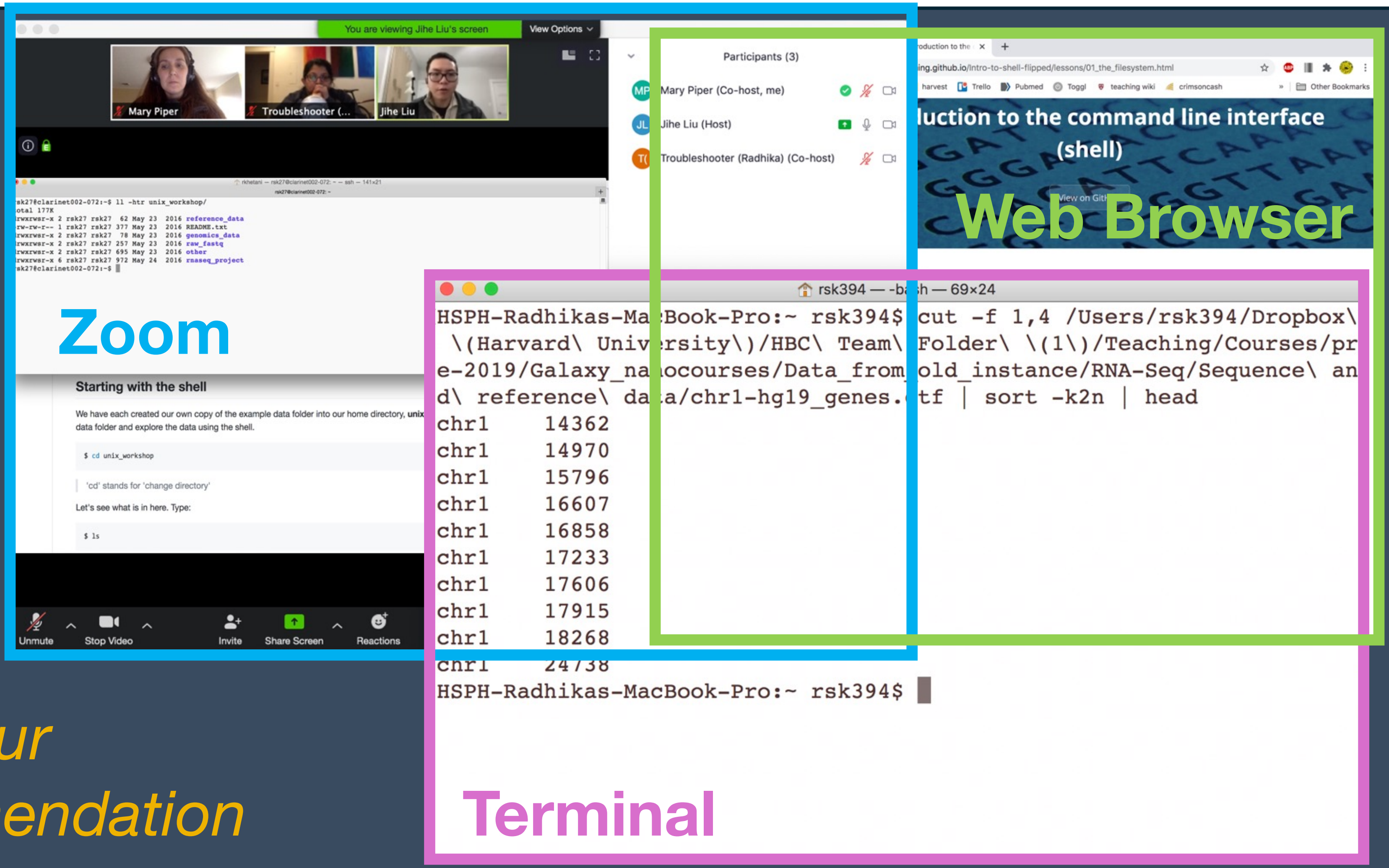
```
$ ls
```

```
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.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



```
sk27@clarinet002-072:~$ ll -htr unix_workshop/
otal 177K
rwxrwsr-x 2 rsk27 rsk27  62 May 23  2016 reference_data
rw-rw-r-- 1 rsk27 rsk27 377 May 23  2016 README.txt
rwxrwsr-x 2 rsk27 rsk27  78 May 23  2016 genomics_data
rwxrwsr-x 2 rsk27 rsk27 257 May 23  2016 raw_fastq
rwxrwsr-x 2 rsk27 rsk27 695 May 23  2016 other
rwxrwsr-x 6 rsk27 rsk27 972 May 24  2016 rnaseq_project
sk27@clarinet002-072:~$
```

Starting with the shell

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

luction to the command line interface (shell)

```
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*

# 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

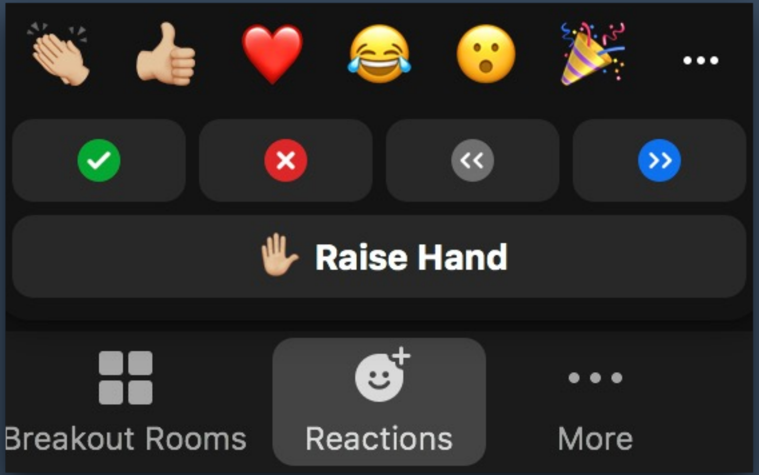
- At-home lessons and exercises after each session
- Cover material not previously discussed
- Provides us feedback to help pace the course appropriately
- 3-5 hours to complete
- Homework load is heavier in the beginning of this workshop series and tapers off

# Using AI for Assignments

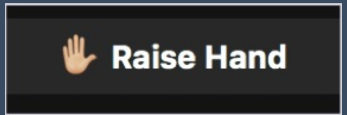
- Do
  - Try to resolve error messages with it
  - Test code written by AI on a dataset where you have expected results
  - Take the time to review the generated code line-by-line
- Don’t
  - Implement it in replacement to learning
  - Write code that you don’t understand
  - Assume the output from an 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"

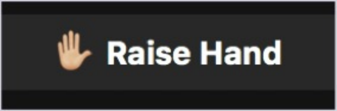


# Odds & Ends

- Questions for the presenter?
  - Post the question in the Chat window OR
  -  when the presenter asks for questions

  - Let the Moderator know

# Odds & Ends

- Questions for the presenter?
  - Post the question in the Chat window OR
  - 

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
Harvard Chan Bioinformatics Core

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