

# 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



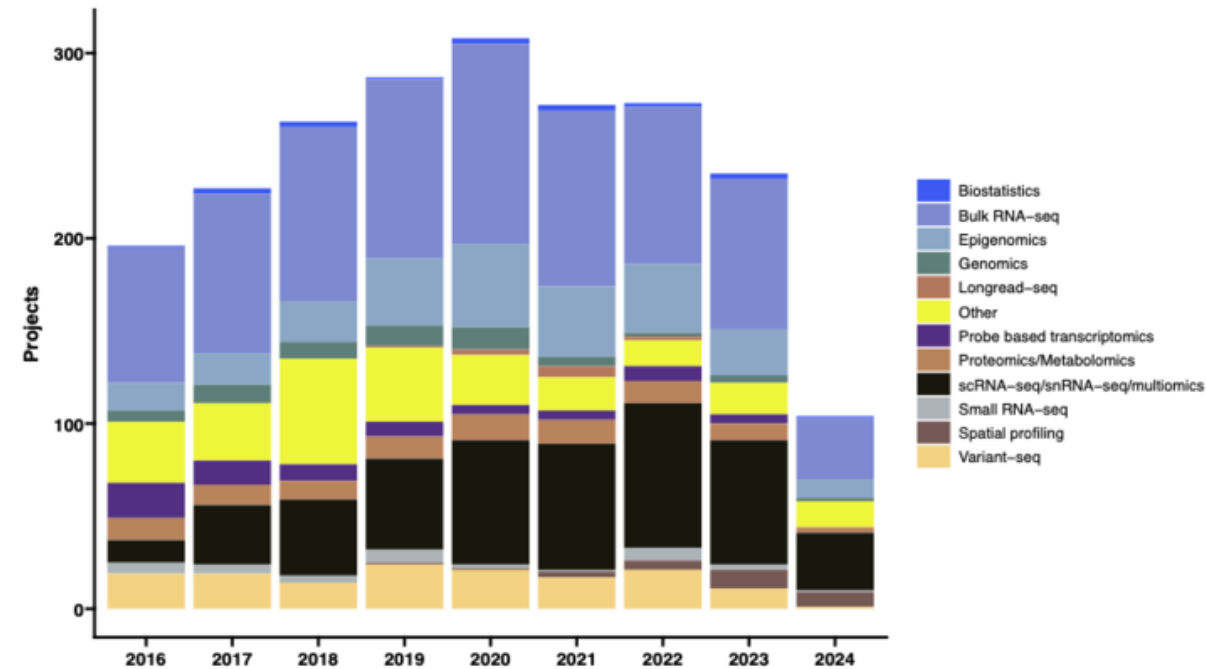
Zhu Zhuo



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



# 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



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

NIEHS

---



THE HARVARD CLINICAL  
AND TRANSLATIONAL  
SCIENCE CENTER

---



**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

# 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>



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

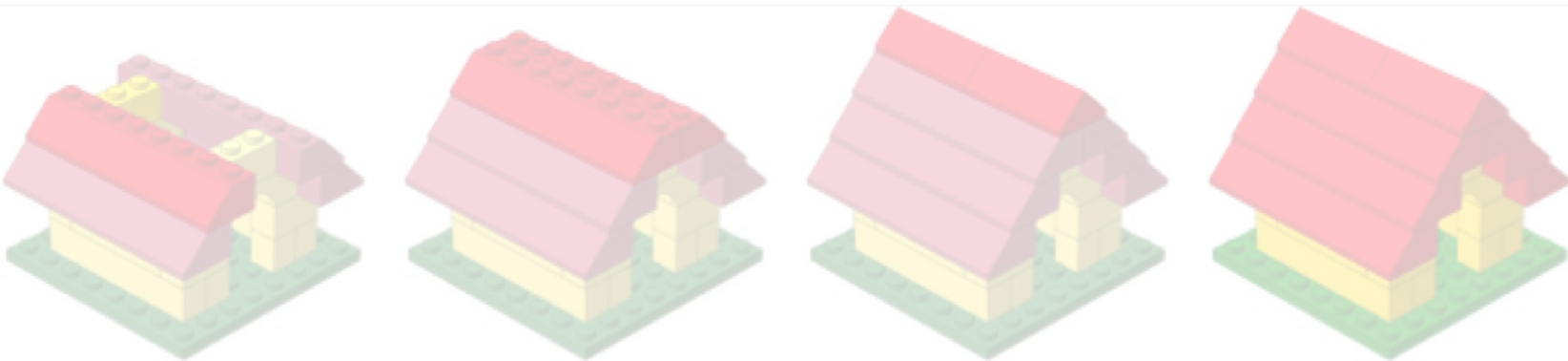
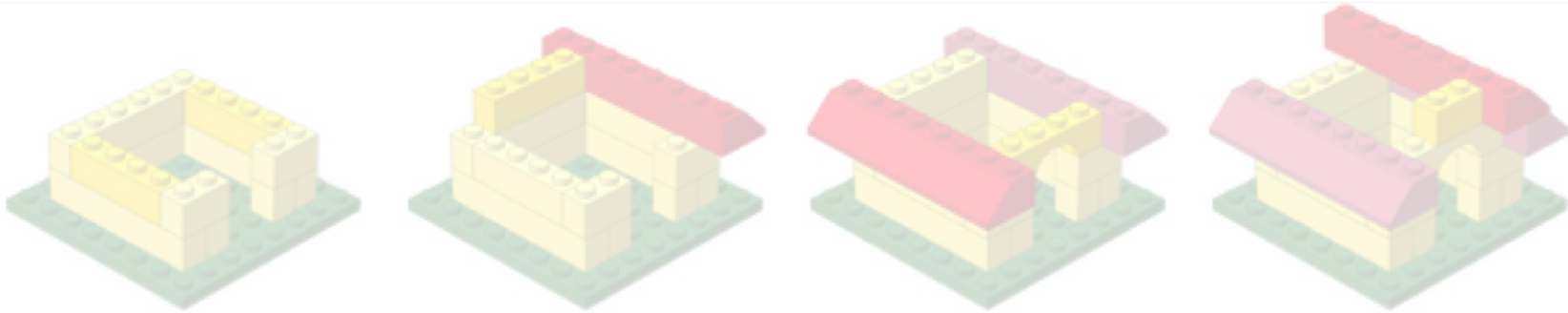
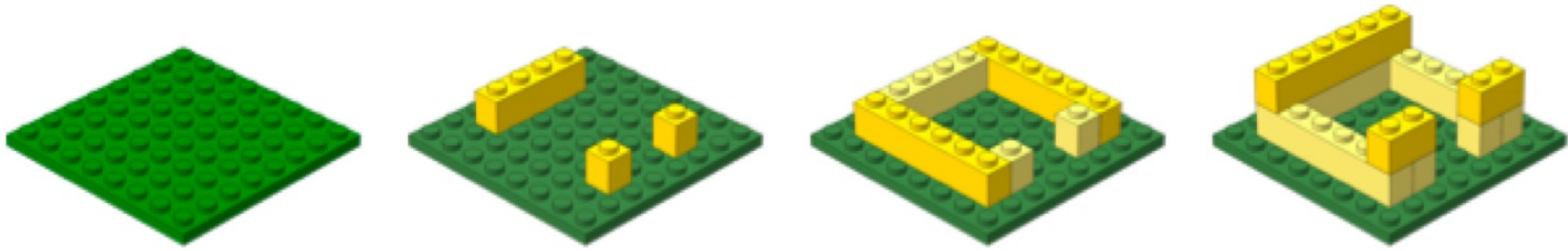
**DF/HCC**  
DANA-FARBER / HARVARD CANCER CENTER



THE HARVARD CLINICAL  
AND TRANSLATIONAL  
SCIENCE CENTER

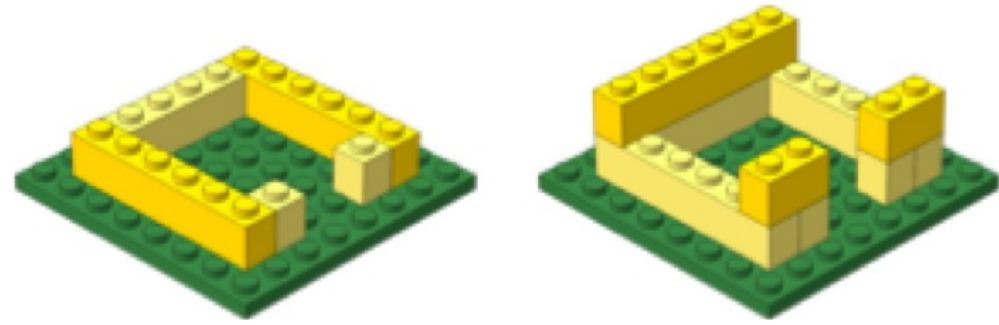






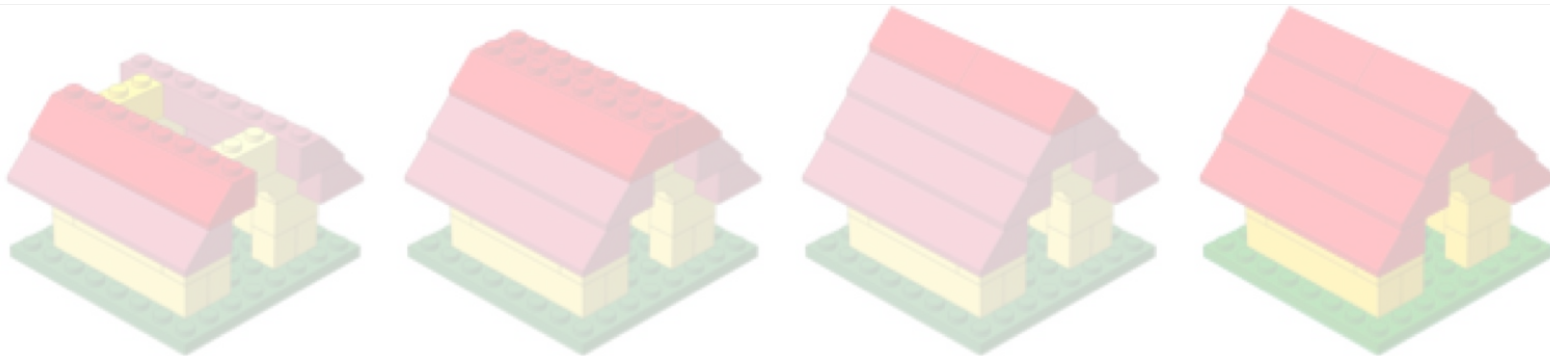
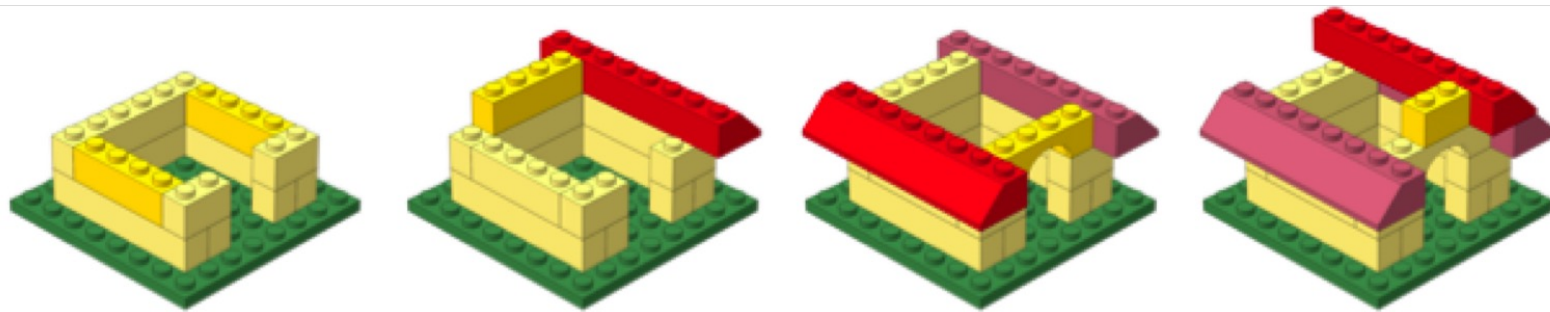
Setting up to perform Bioinformatics analysis

# 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

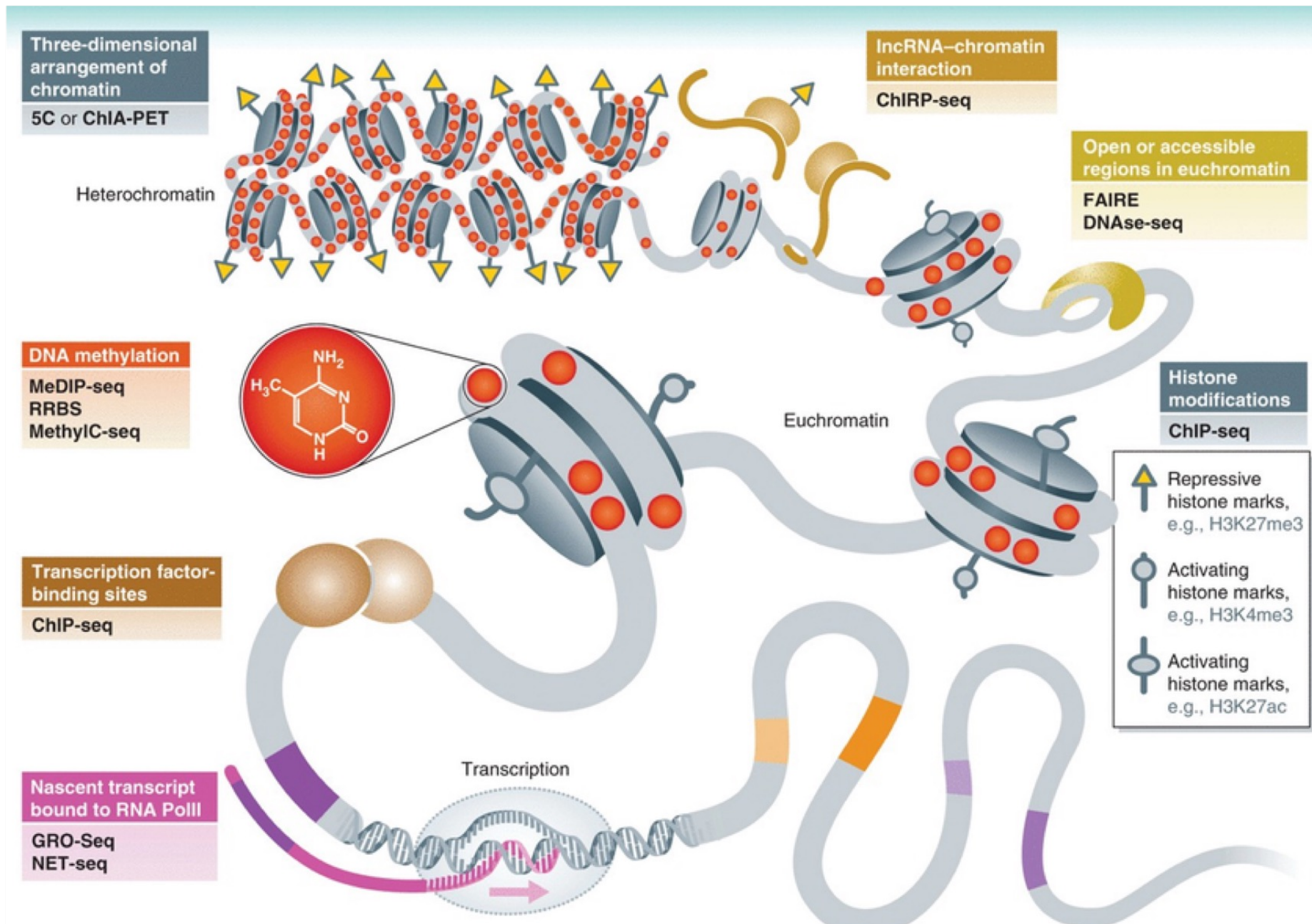
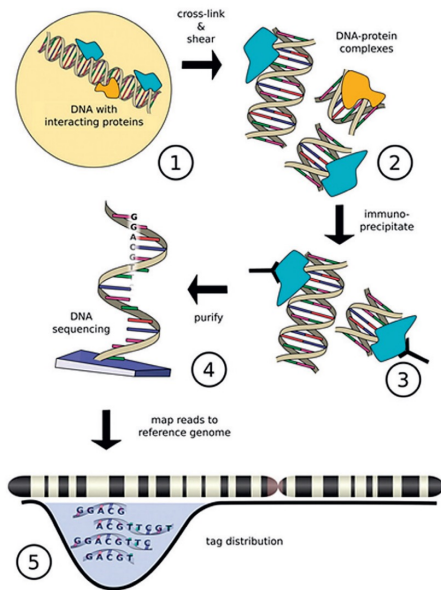


Figure adapted from Soon WW, Hariharan M, Snyder MP, "High throughput sequencing for biology and medicine". Molecular Systems Biology 9:640 2013

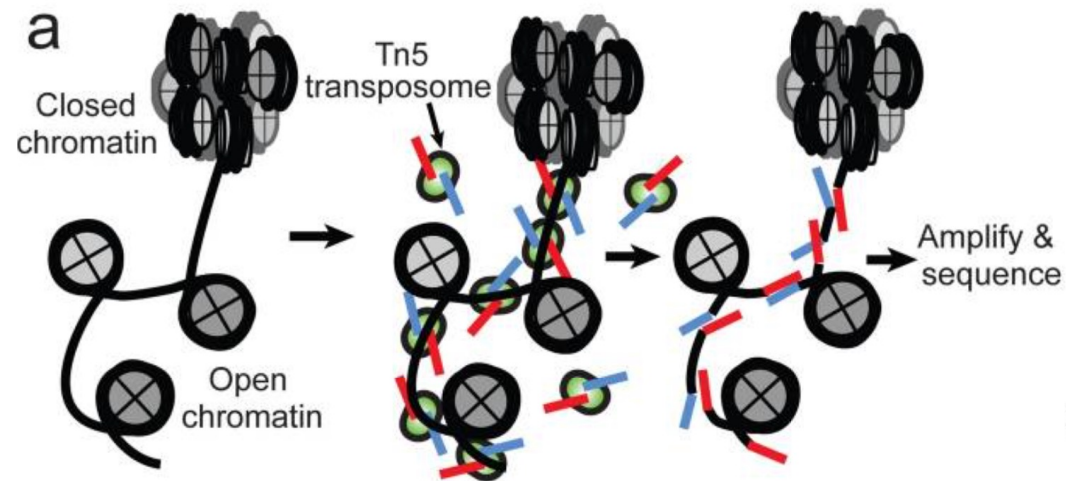


# Genomic Methods for Profiling Chromatin

## ChIP-seq



## ATAC-seq



## CUT&RUN

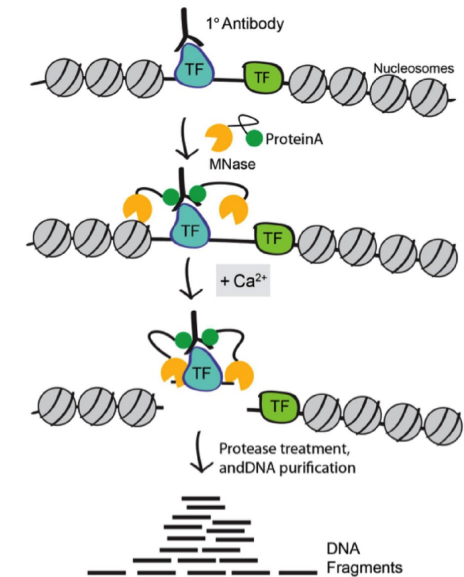
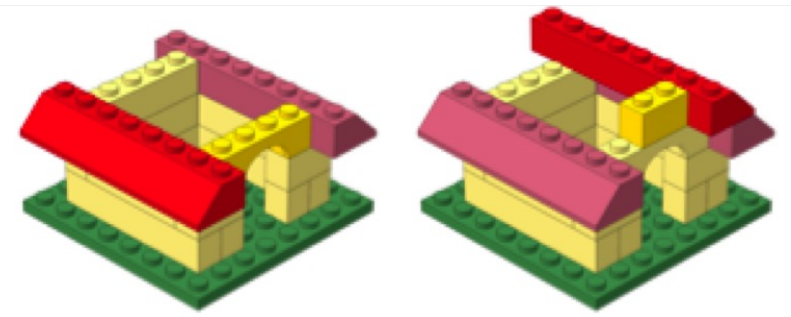
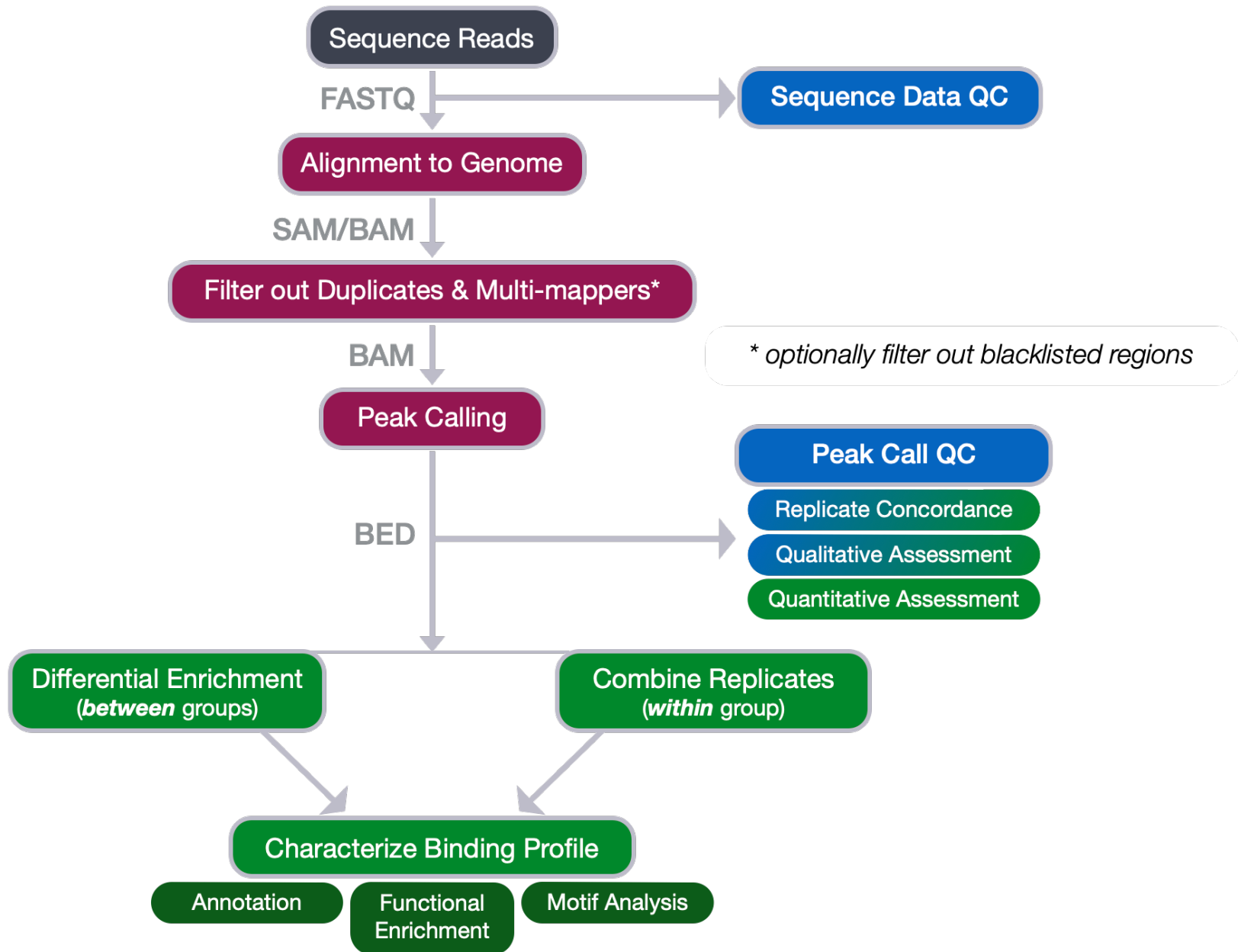


Figure 1. CUT&RUN schematic (see text for details).

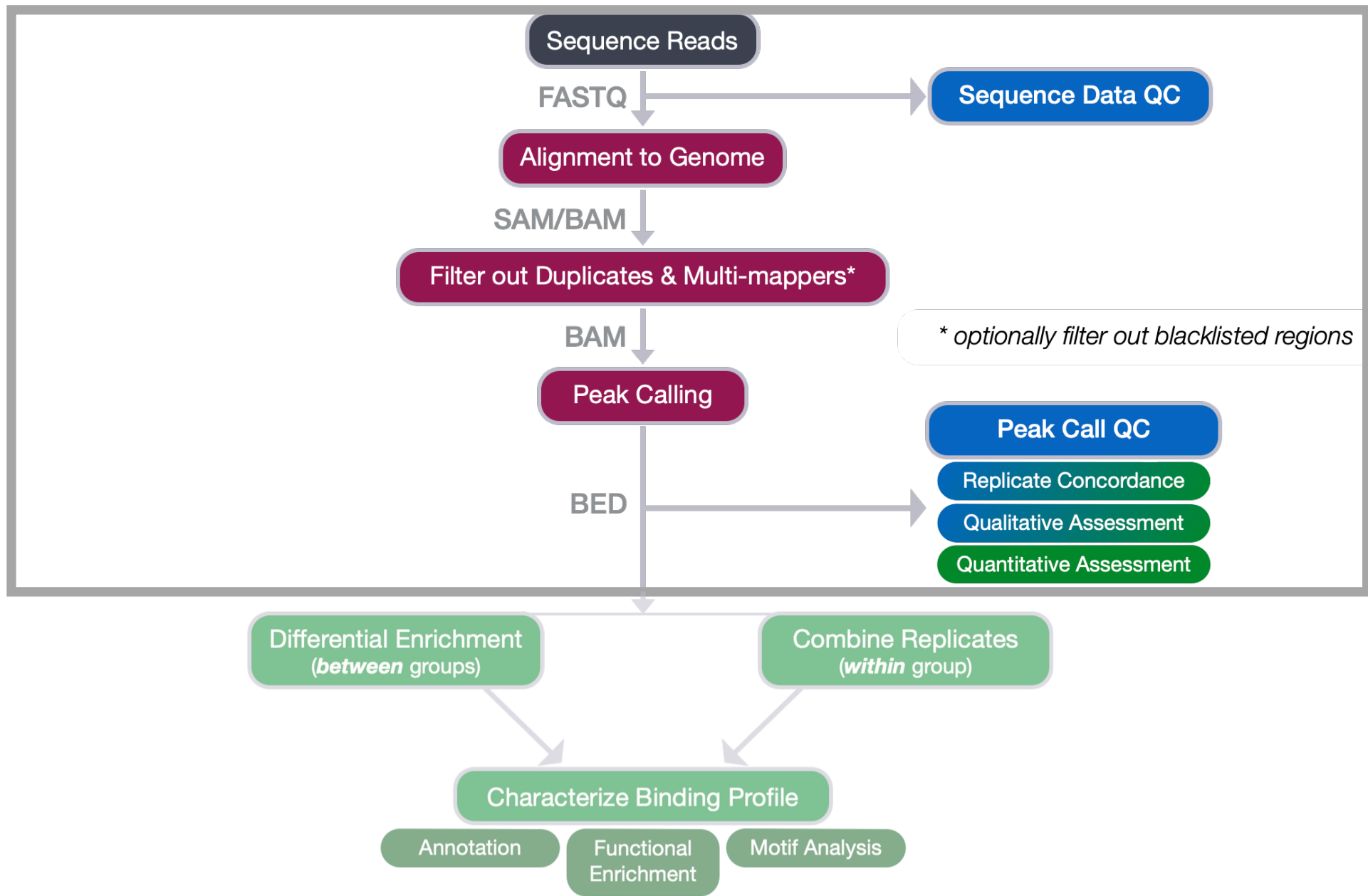
# Learning Objectives



- ❖ 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

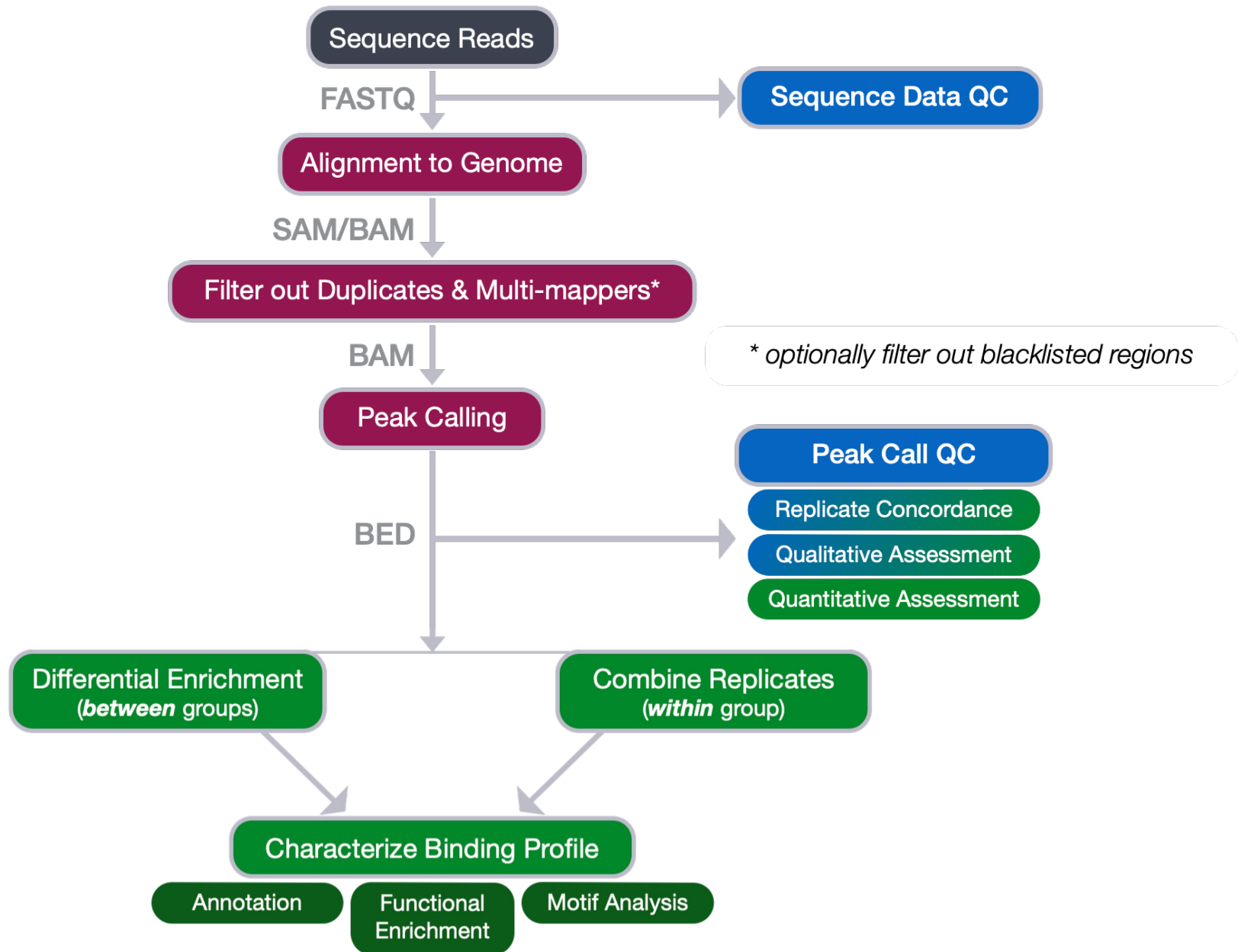


# The Workflow



# 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	<a href="#">Understanding chromatin biology using high-throughput sequencing</a>	Dr. Shannan Ho Sui
11:00- 11:05	Break	
11:05 - 11:20	<a href="#">HPC review Q&amp;A</a>	Will
11:20 - 11:50	<a href="#">Dataset overview and project organization</a>	Will

# Course materials

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



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

# 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 a command to list files in a directory. The terminal output is as follows:

```
total 177K
drwxrwxr-x 2 rsk27 rsk27 62 May 23 2016 reference_data
-rw-rw-r-- 1 rsk27 rsk27 337 May 23 2016 README.txt
drwxrwxr-x 2 rsk27 rsk27 78 May 23 2016 genomics_data
drwxrwxr-x 2 rsk27 rsk27 237 May 23 2016 raw_fastq
drwxrwxr-x 2 rsk27 rsk27 495 May 23 2016 other
drwxrwxr-x 6 rsk27 rsk27 372 May 24 2016 rnaseq_project
rsk27@clariinet002-0721:~$
```

A browser window is also open, displaying a page titled "Introduction to the command line interface (shell)". The page content includes a "View on GitHub" button and a background image of a DNA sequence.

The terminal window also 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$
```



# 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 a list of files and directories, including 'reference\_data', 'RNAseq\_data', 'genomics\_data', 'raw\_fastq', 'other', and 'rnaseq\_project'. A browser window is also open, showing a page titled 'Introduction to the command line interface (shell)' with a 'View on GitHub' button. The terminal window shows the following 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.txt | 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

The screenshot shows a Zoom meeting interface with three windows open. The top window is a Zoom meeting window showing three participants: Mary Piper, Troubleshooter (Radhika), and Jihe Liu. The middle window is a participants list showing three participants: Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host). The bottom window is a terminal window showing a shell prompt and a command being executed. The terminal output shows a list of files and directories. The web browser window is titled 'Introduction to the command line interface (shell)' and shows a page with the text 'Web Browser' in large green letters.

```
rsk27@clariinet002-0721~$ ll -ltr unix_workshop/
total 177K
drwxrwxr-x 2 rsk27 rsk27 62 May 23 2016 reference_data
-rw-rw-r-- 1 rsk27 rsk27 337 May 23 2016 README.txt
drwxrwxr-x 2 rsk27 rsk27 78 May 23 2016 genomics_data
drwxrwxr-x 2 rsk27 rsk27 237 May 23 2016 raw_fastq
drwxrwxr-x 2 rsk27 rsk27 495 May 23 2016 other
drwxrwxr-x 6 rsk27 rsk27 372 May 24 2016 rnaseq_project
rsk27@clariinet002-0721~$
```

```
HSPH-Radhikas-MacBook-Pro:~ rsk394$ cut -f 1,4 /Users/rsk394/Dropbox\
\ (Harvard\ University)\ /HBC\ Team\ Folder\ \ (1\)/Teaching/Courses/pr
e-2019/Galaxy_nacourses/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

The screenshot displays a Zoom meeting interface with three windows open. The top window shows a video call with three participants: Mary Piper, Troubleshooter (Radhika), and Jihe Liu. The middle window is a 'Participants (3)' list. The bottom-left window is a terminal window showing a shell session with the following output:

```
rk276clacinet002-0721-$ ll -ltr unix_workshop/
total 177K
drwxrwxr-x 2 rsk27 rsk27 62 May 23 2016 reference_data
-rw-rw-r-- 1 rsk27 rsk27 337 May 23 2016 README.txt
drwxrwxr-x 2 rsk27 rsk27 78 May 23 2016 genomics_data
drwxrwxr-x 2 rsk27 rsk27 237 May 23 2016 raw_fastq
drwxrwxr-x 2 rsk27 rsk27 495 May 23 2016 other
drwxrwxr-x 6 rsk27 rsk27 972 May 24 2016 rnaseq_project
rk276clacinet002-0721-$
```

The bottom-right window is a terminal window titled 'rsk394' showing 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$
```

The top-right window is a browser window showing a page titled 'Introduction to the command line interface (shell)' with a 'View on GitHub' button.

*Our  
Recommendation*

**Terminal**

# Single Screen & 3 Windows

The image shows a Zoom meeting interface with three windows highlighted by colored boxes:

- Zoom (Blue box):** Shows the Zoom meeting controls and a terminal window. The terminal output is as follows:

```
rsk394@macbook-pro:~$ cd unix_workshop/
rsk394@macbook-pro:~/unix_workshop$ ls
reference_data
raw_fastq
raw_genome_data
raw_reads
reference_data/chr1-hg19_genes.txt
reference_data/chr2-hg19_genes.txt
reference_data/chr3-hg19_genes.txt
reference_data/chr4-hg19_genes.txt
reference_data/chr5-hg19_genes.txt
reference_data/chr6-hg19_genes.txt
reference_data/chr7-hg19_genes.txt
reference_data/chr8-hg19_genes.txt
reference_data/chr9-hg19_genes.txt
reference_data/chr10-hg19_genes.txt
reference_data/chr11-hg19_genes.txt
reference_data/chr12-hg19_genes.txt
reference_data/chr13-hg19_genes.txt
reference_data/chr14-hg19_genes.txt
reference_data/chr15-hg19_genes.txt
reference_data/chr16-hg19_genes.txt
reference_data/chr17-hg19_genes.txt
reference_data/chr18-hg19_genes.txt
reference_data/chr19-hg19_genes.txt
reference_data/chr20-hg19_genes.txt
reference_data/chr21-hg19_genes.txt
reference_data/chr22-hg19_genes.txt
reference_data/chr23-hg19_genes.txt
reference_data/chr24-hg19_genes.txt
reference_data/chr25-hg19_genes.txt
reference_data/chr26-hg19_genes.txt
reference_data/chr27-hg19_genes.txt
reference_data/chr28-hg19_genes.txt
reference_data/chr29-hg19_genes.txt
reference_data/chr30-hg19_genes.txt
reference_data/chr31-hg19_genes.txt
reference_data/chr32-hg19_genes.txt
reference_data/chr33-hg19_genes.txt
reference_data/chr34-hg19_genes.txt
reference_data/chr35-hg19_genes.txt
reference_data/chr36-hg19_genes.txt
reference_data/chr37-hg19_genes.txt
reference_data/chr38-hg19_genes.txt
reference_data/chr39-hg19_genes.txt
reference_data/chr40-hg19_genes.txt
reference_data/chr41-hg19_genes.txt
reference_data/chr42-hg19_genes.txt
reference_data/chr43-hg19_genes.txt
reference_data/chr44-hg19_genes.txt
reference_data/chr45-hg19_genes.txt
reference_data/chr46-hg19_genes.txt
reference_data/chr47-hg19_genes.txt
reference_data/chr48-hg19_genes.txt
reference_data/chr49-hg19_genes.txt
reference_data/chr50-hg19_genes.txt
reference_data/chr51-hg19_genes.txt
reference_data/chr52-hg19_genes.txt
reference_data/chr53-hg19_genes.txt
reference_data/chr54-hg19_genes.txt
reference_data/chr55-hg19_genes.txt
reference_data/chr56-hg19_genes.txt
reference_data/chr57-hg19_genes.txt
reference_data/chr58-hg19_genes.txt
reference_data/chr59-hg19_genes.txt
reference_data/chr60-hg19_genes.txt
reference_data/chr61-hg19_genes.txt
reference_data/chr62-hg19_genes.txt
reference_data/chr63-hg19_genes.txt
reference_data/chr64-hg19_genes.txt
reference_data/chr65-hg19_genes.txt
reference_data/chr66-hg19_genes.txt
reference_data/chr67-hg19_genes.txt
reference_data/chr68-hg19_genes.txt
reference_data/chr69-hg19_genes.txt
reference_data/chr70-hg19_genes.txt
reference_data/chr71-hg19_genes.txt
reference_data/chr72-hg19_genes.txt
reference_data/chr73-hg19_genes.txt
reference_data/chr74-hg19_genes.txt
reference_data/chr75-hg19_genes.txt
reference_data/chr76-hg19_genes.txt
reference_data/chr77-hg19_genes.txt
reference_data/chr78-hg19_genes.txt
reference_data/chr79-hg19_genes.txt
reference_data/chr80-hg19_genes.txt
reference_data/chr81-hg19_genes.txt
reference_data/chr82-hg19_genes.txt
reference_data/chr83-hg19_genes.txt
reference_data/chr84-hg19_genes.txt
reference_data/chr85-hg19_genes.txt
reference_data/chr86-hg19_genes.txt
reference_data/chr87-hg19_genes.txt
reference_data/chr88-hg19_genes.txt
reference_data/chr89-hg19_genes.txt
reference_data/chr90-hg19_genes.txt
reference_data/chr91-hg19_genes.txt
reference_data/chr92-hg19_genes.txt
reference_data/chr93-hg19_genes.txt
reference_data/chr94-hg19_genes.txt
reference_data/chr95-hg19_genes.txt
reference_data/chr96-hg19_genes.txt
reference_data/chr97-hg19_genes.txt
reference_data/chr98-hg19_genes.txt
reference_data/chr99-hg19_genes.txt
reference_data/chr100-hg19_genes.txt
```
- Web Browser (Green box):** Shows a page titled "Introduction to the command line interface (shell) Web Browser".
- Terminal (Purple box):** Shows a terminal window with 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_nacourses/Data_from_old_instance/RNA-Seq/Sequence\ an
d\ reference\ data/chr1-hg19_genes.txt | 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**

# 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

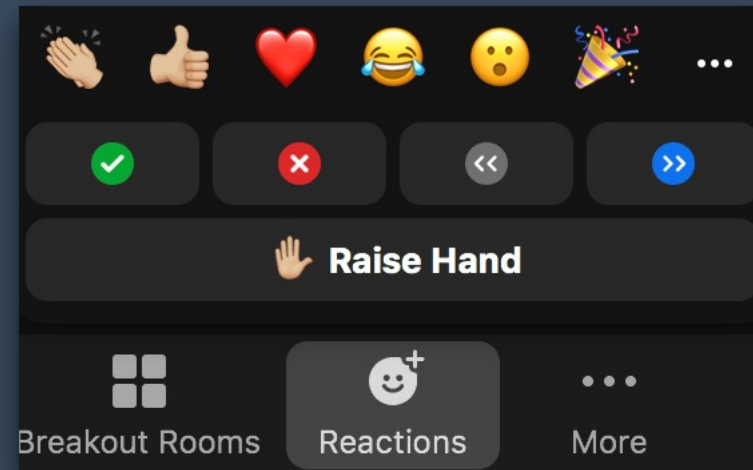
# Odds & Ends

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

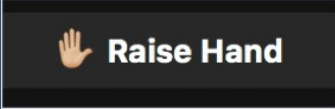
❖ Are you all set?

❖  = "agree", "I'm all set"

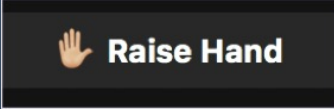
❖  = "disagree", "I need help"



# 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

---

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