

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

Harvard Chan Bioinformatics Core

in collaboration with

HMS Research Computing

<https://tinyurl.com/hbc-intro-to-chipseq>



Shannan Ho Sui



John Hutchinson



Brad Chapman



Rory Kirchner



Meeta Mistry



Radhika Khetani



Mary Piper



Lorena Pantano



Michael Steinbaugh



Victor Barrera



Kayleigh Rutherford



Peter Kraft



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

NIEHS / CFAR  
Bioinformatics  
Core

**HSCI**  
HARVARD STEM CELL  
INSTITUTE

Center for Stem  
Cell  
Bioinformatics

 **HARVARD CATALYST**  
THE HARVARD CLINICAL  
AND TRANSLATIONAL  
SCIENCE CENTER

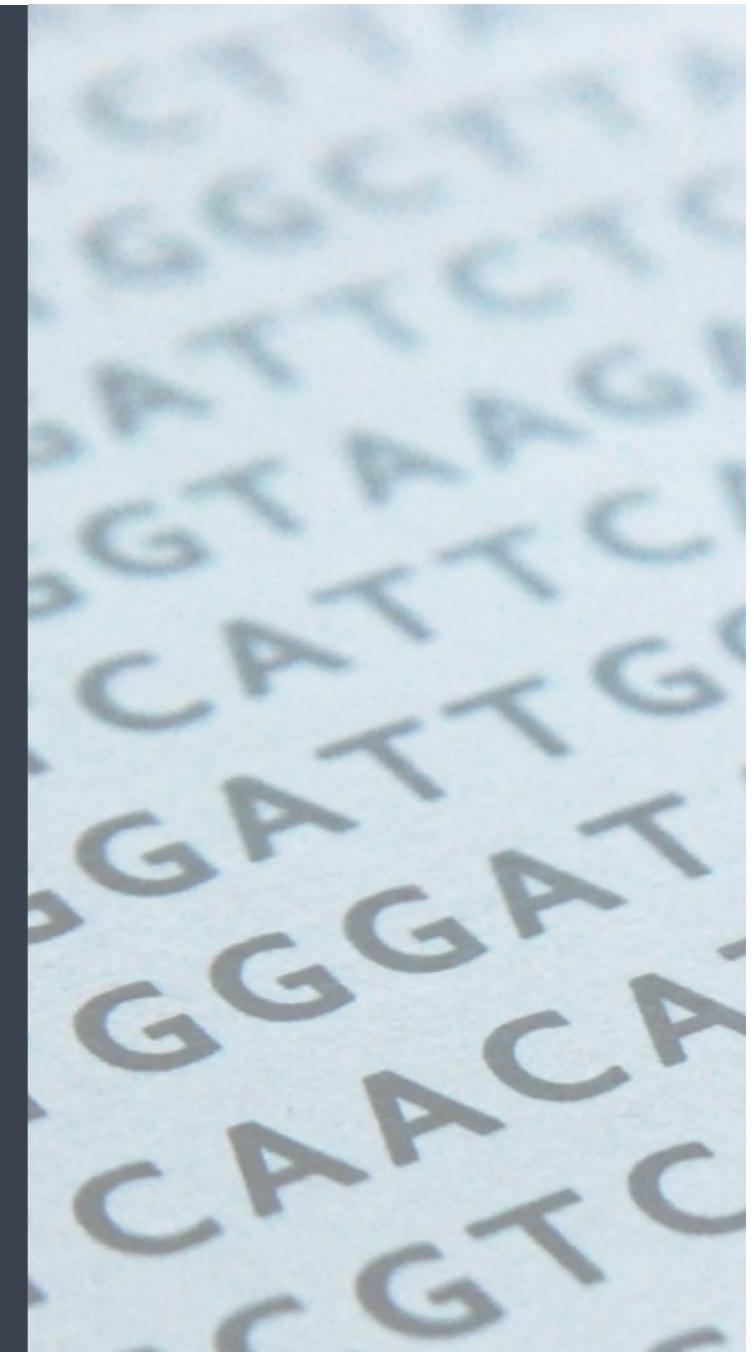
 **HARVARD**  
MEDICAL SCHOOL

Harvard  
Catalyst  
Bioinformatics  
Consulting

HMS  
Tools &  
Technology  
  
Harvard  
NeuroDiscovery  
Center

# Consulting

- RNA-seq, small RNA-seq and ChIP-seq analysis
- Genome-wide methylation
- WGS, resequencing, exome-seq and CNV studies
- Quality assurance and analysis of gene expression arrays
- Functional enrichment analysis
- Grant support

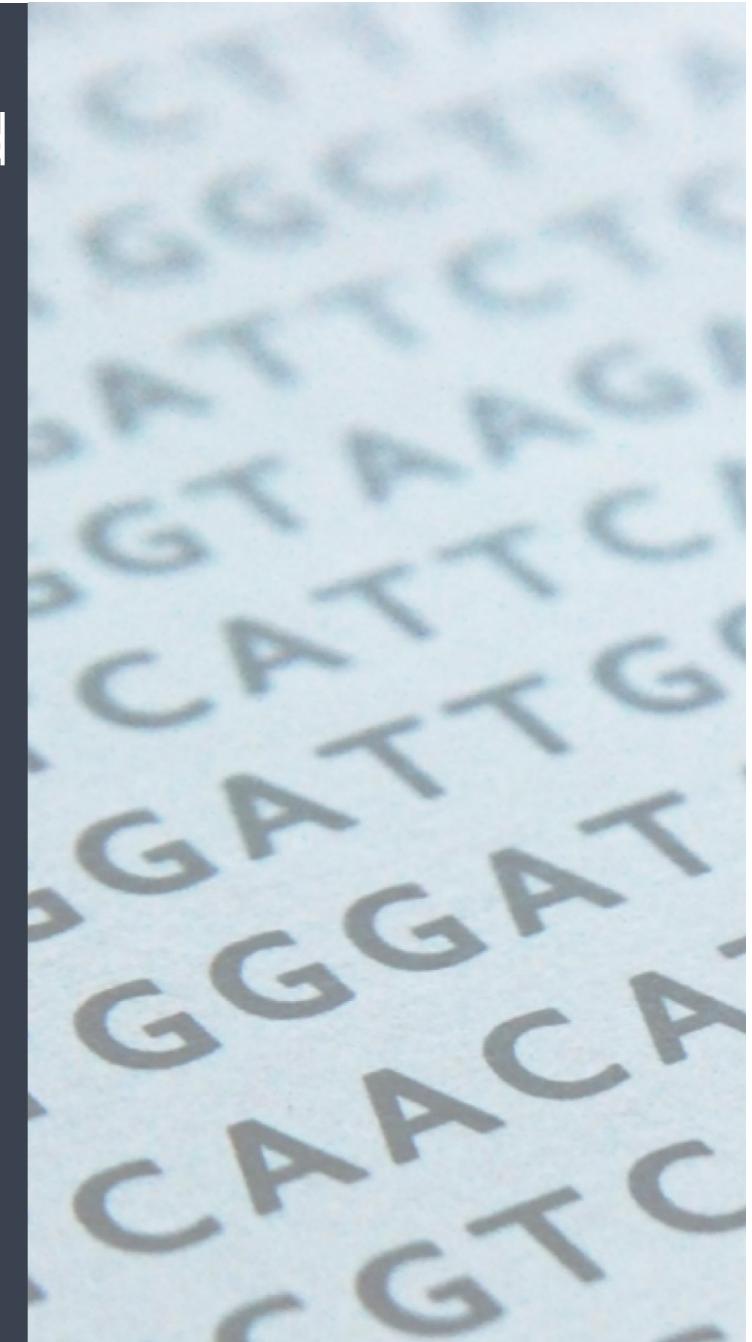


# Training

- Short workshops on introductory, intermediate and advanced topics related to NGS data analysis
- Monthly, 2-3 hour, hands-on and free workshops on “Current Topics in Bioinformatics”
- In-depth courses (8- or 12-day formats) [Fall 2018]

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

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





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

NIEHS / CFAR  
Bioinformatics  
Core

**HSCI**  
HARVARD STEM CELL  
INSTITUTE

Center for Stem  
Cell  
Bioinformatics

 **HARVARD CATALYST**  
THE HARVARD CLINICAL  
AND TRANSLATIONAL  
SCIENCE CENTER

 **HARVARD**  
MEDICAL SCHOOL

Harvard  
Catalyst  
Bioinformatics  
Consulting

HMS  
Tools &  
Technology

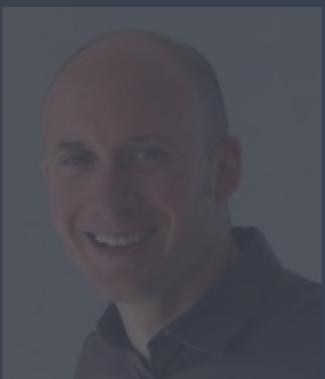
# Introductions!



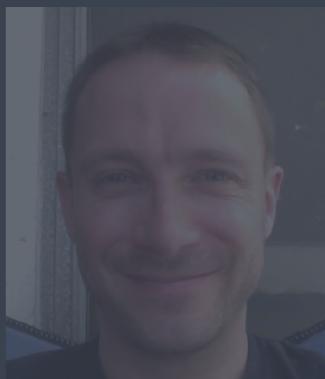
Shannan Ho Sui



John Hutchinson



Brad Chapman



Rory Kirchner



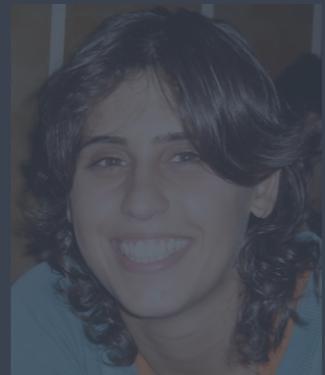
Meeta Mistry



Radhika Khetani



Mary Piper



Lorena Pantano



Michael Steinbaugh



Victor Barrera

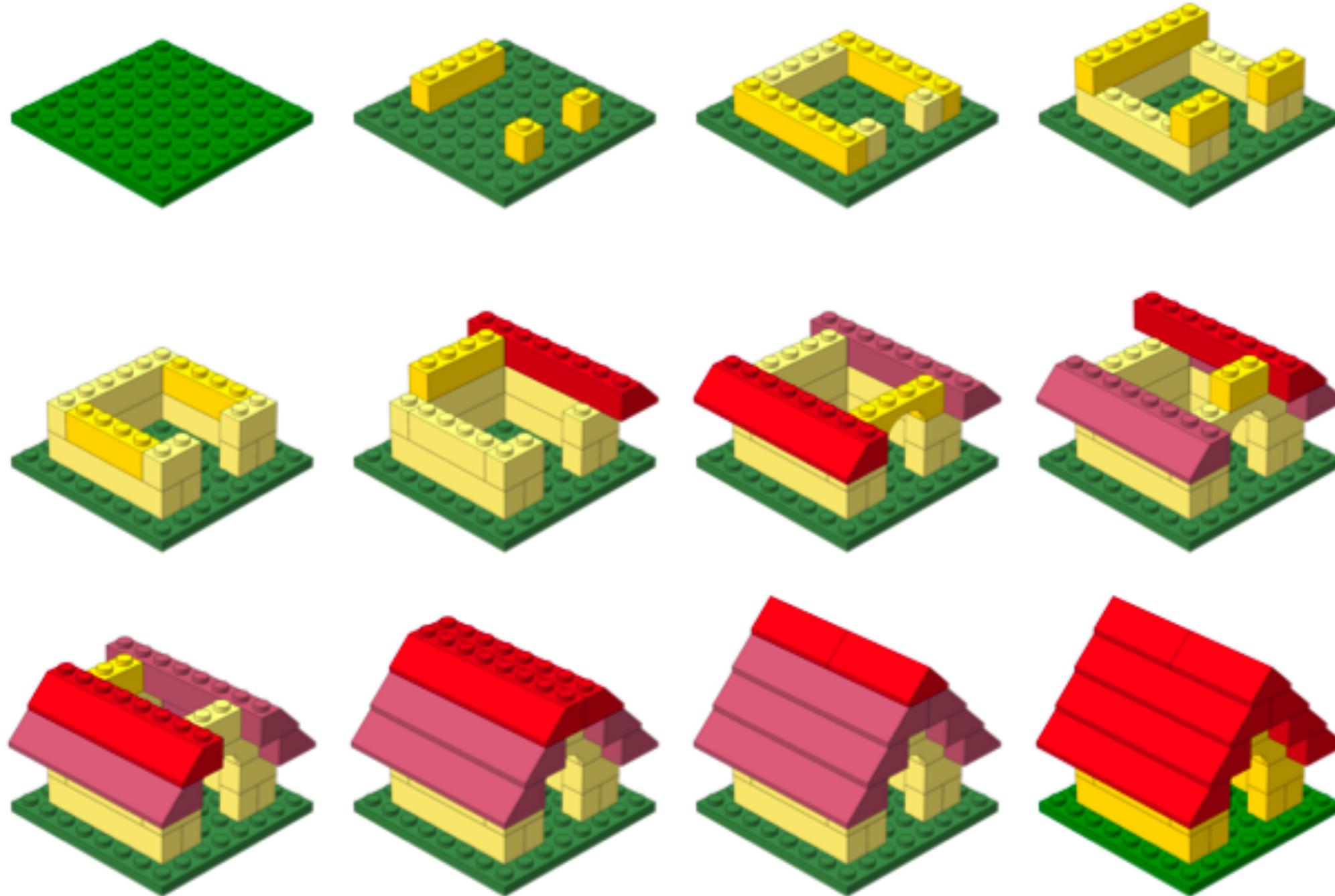


Kayleigh Rutherford



Peter Kraft

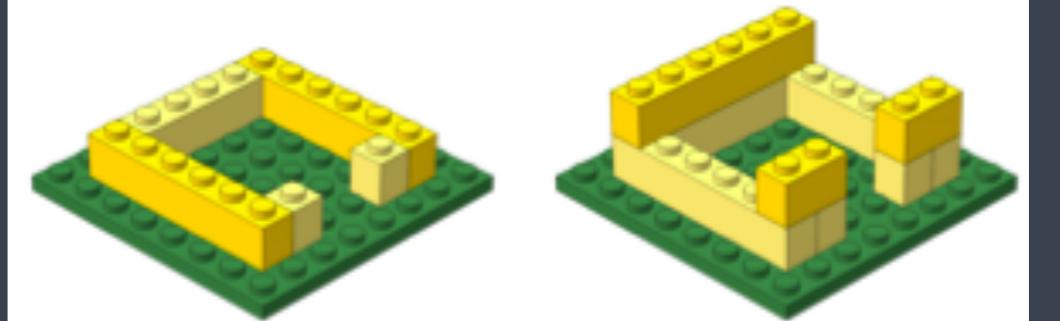
# Workshop scope



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

Bioinformatics data analysis

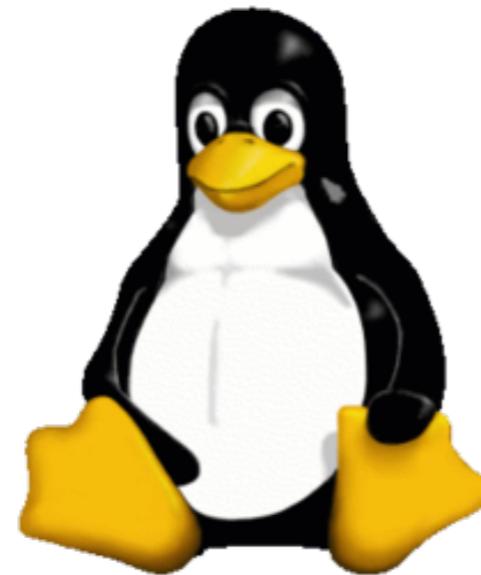
# Base components



- ✓ Introduction to the UNIX shell
  - Dealing with large data files
  - Using bioinformatics tools
  - Accessing and using compute clusters
- ✓ R (*outside the scope of today's workshop*)
  - Parsing and working with smaller results text files
  - Statistical analysis, e.g. differential expression analysis
  - Generating figures from complex data



```
rkhetani — rsk27@clarinet002-072: ~ — ssh — 75x51
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:~$
```



*“UNIX is user-friendly.*

*It's just very selective about who its friends are.”*

# Why UNIX?

- ◆ UNIX is a **stable**, **efficient** and **powerful** operating system
- ◆ It can easily coordinate the use and sharing of a computer's (or a system's) resources, i.e. built to allow multi-user functionality
- ◆ Can easily handle complex and repetitive tasks easily on large and small datasets

# Linux

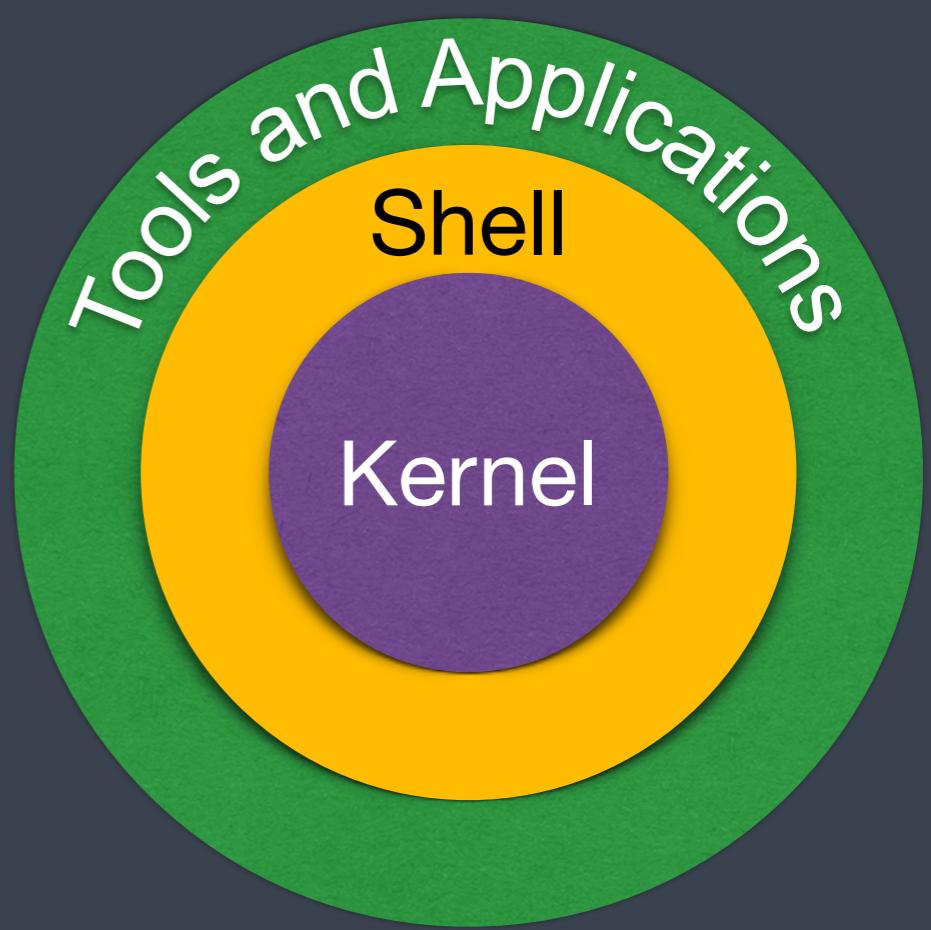
- ♦ Linux is a free, open-source operating system based on UNIX
- ♦ It has the same components as the original, but the open source community is involved in active development of various distinct distributions of Linux



# Components

The UNIX/Linux system is functionally organized at 3 levels:

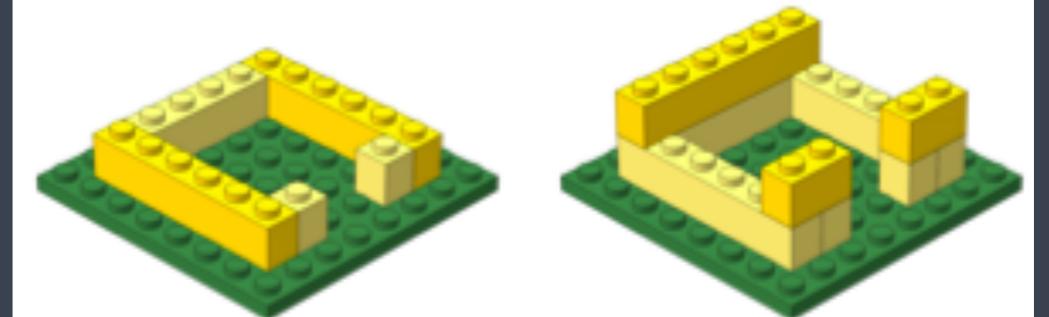
- ◆ **The kernel**, which schedules tasks and manages storage: *the brain of the system*
- ◆ **The shell**, *an interpreter* that helps interprets our input for the kernel
- ◆ **Utilities, tools and applications**, which use the shell to communicate with the kernel



# The “shell”

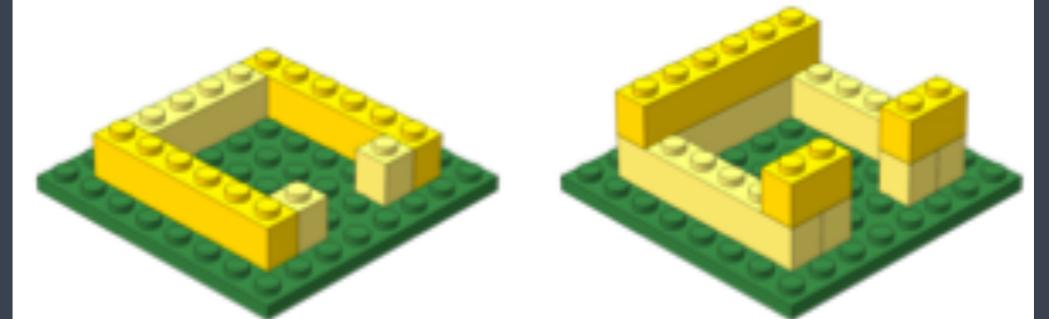
- ◆ The shell is **an interpreter**
- ◆ It is independent of the operating system
- ◆ Dozens of shells have been developed throughout UNIX history, and a lot of them are still in use
- ◆ The most commonly used shell is **bash**

# Learning Objectives



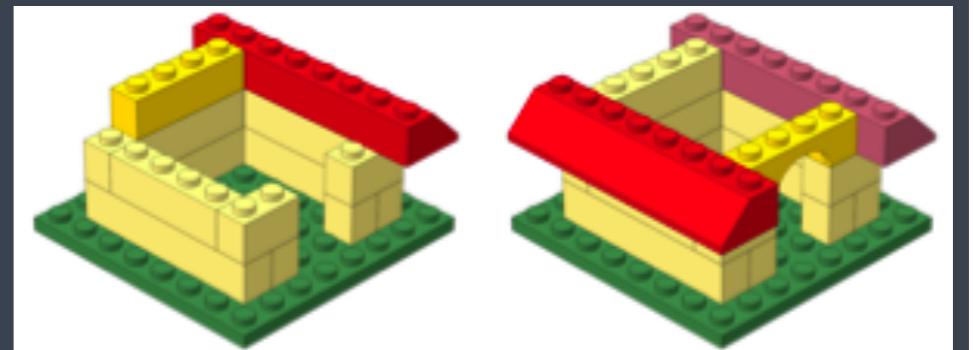
- ✓ Learn what a “shell” is and become comfortable with the command line interface
  - Find your way around a filesystem using the command line
  - Work with small and large data files
  - Become more efficient when performing repetitive tasks

# Learning Objectives



- ✓ Learn what a “shell” is and become comfortable with the command line interface
  - Find your way around a filesystem using the command line
  - 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
  - Independently access the O2 cluster
  - Perform analysis using the cluster (run programs, pipelines, etc.)

# Learning Objectives



- ✓ Describe best practices for designing an ChIP-seq experiment
- ✓ Describe steps in a typical ChIP-seq analysis workflow
- ✓ Use HMS-RC's O2 compute cluster to efficiently run the ChIP-seq workflow from sequence reads to peak calls, including QC and visualization.

Some steps in the ChIP-seq workflow require a working knowledge of R, and we won't be covering these in much detail.

# Logistics

# Course schedule

<https://tinyurl.com/hbc-intro-to-chipseq>

# Course materials online



## Introduction to ChIP-Seq using high-performance computing

Intro to ChIPseq using HPC

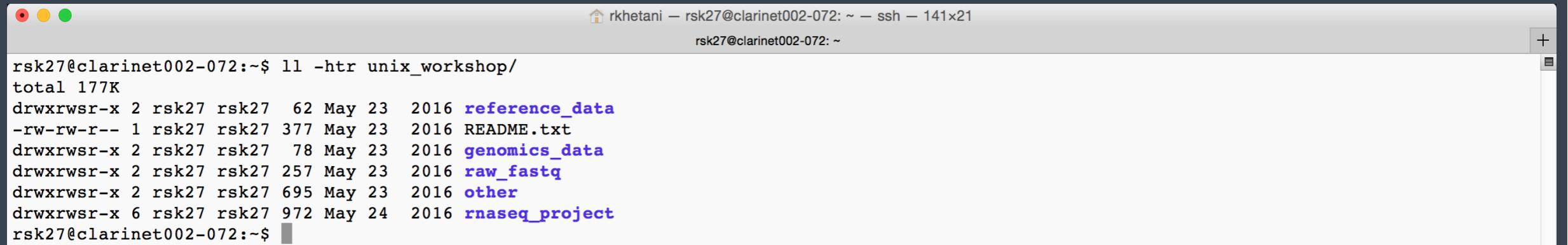
[View on GitHub](#)

## Workshop Schedule

### Day 1

Time	Topic	Instructor
9:00 - 9:40	Workshop Introduction	Radhika
9:40 - 10:30	Introduction to the Shell	Radhika
10:30 - 10:45	Break	
10:45 - 11:35	Introduction to the Shell (cont.)	Meeta

# The 2 Window problem...



```
rsk27@clarinet002-072:~$ ls -l 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\_workshop**. Let's go into the 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
```

# Odds and Ends

- ❖ Name tags: Tent Cards
- ❖ Post-its
- ❖ Wi-Fi: **HMS Public**
- ❖ Lunch locations
- ❖ Bathrooms
- ❖ Water Fountain
- ❖ Phones on vibrate/silent!

# Thanks!

- Shannan Ho Sui (HBC)
- Andy Bergman (HMS-RC)
- Kristina Holton (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)

## Twitter

*HBC:* @bioinfocore

*HMS-RC:* @hms\_rc