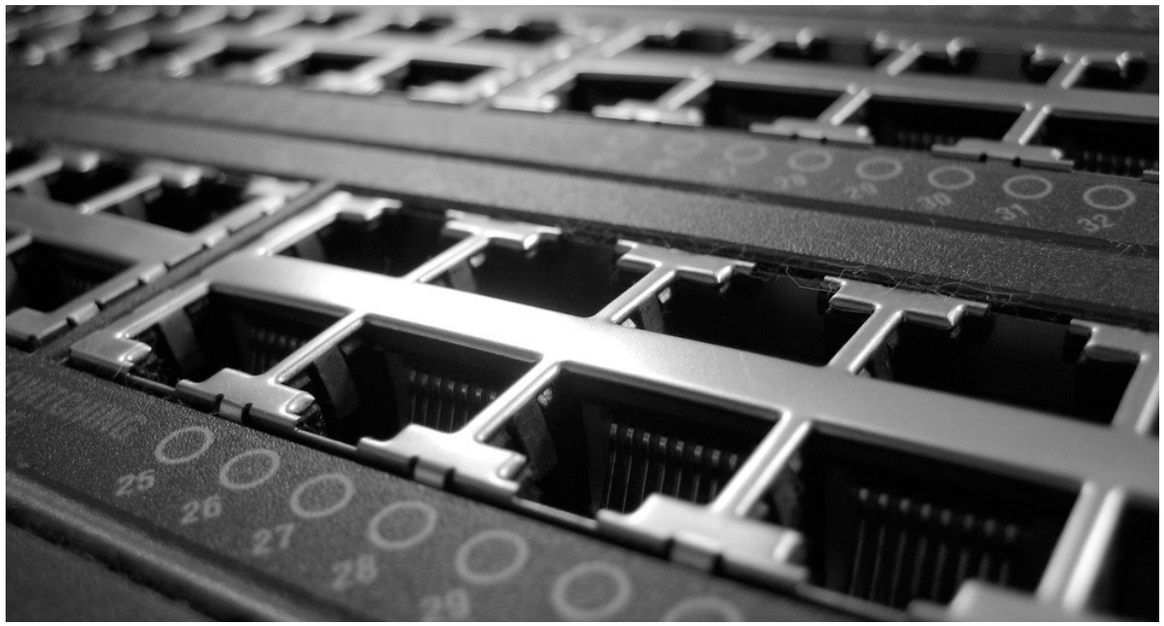


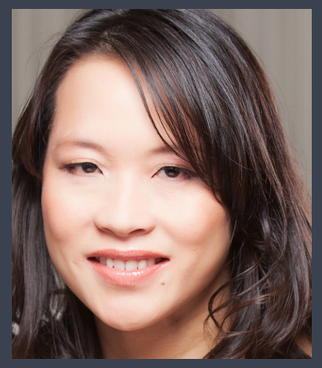
# Introduction to RNA-seq using High-Performance Computing (HPC)

Bioinformatics Core at the

Harvard T.H. Chan School of Public Health

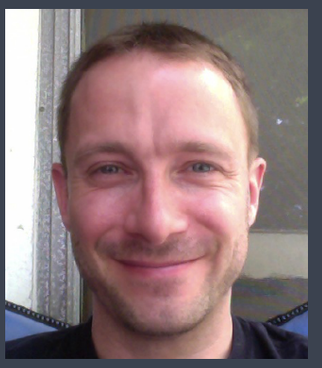
April 23-24, 2018

https://tinyurl.com/intro-rnaseq-gt


Shannan Ho Sui

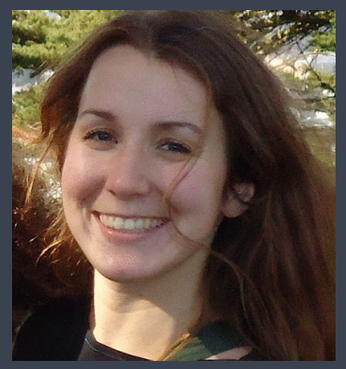
John Hutchinson

Brad Chapman


Rory Kirchner

Meeta Mistry

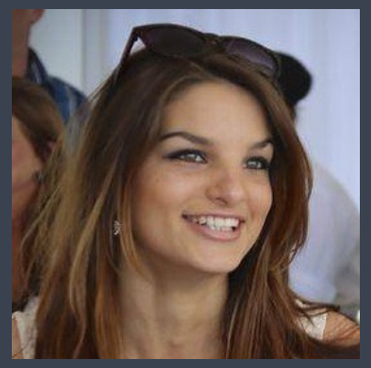
Radhika Khetani


Mary Piper

Lorena Pantano

Michael Steinbaugh

Victor Barrera


Kayleigh Rutherford

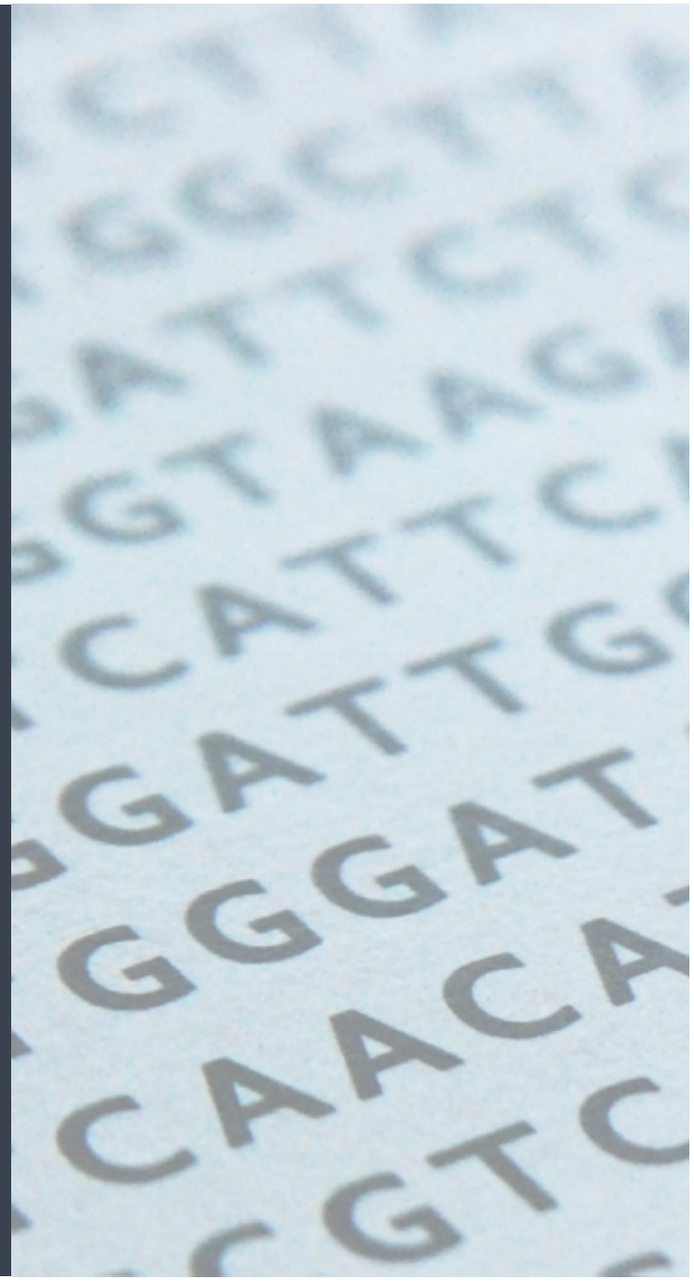
Peter Kraft

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

# 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

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

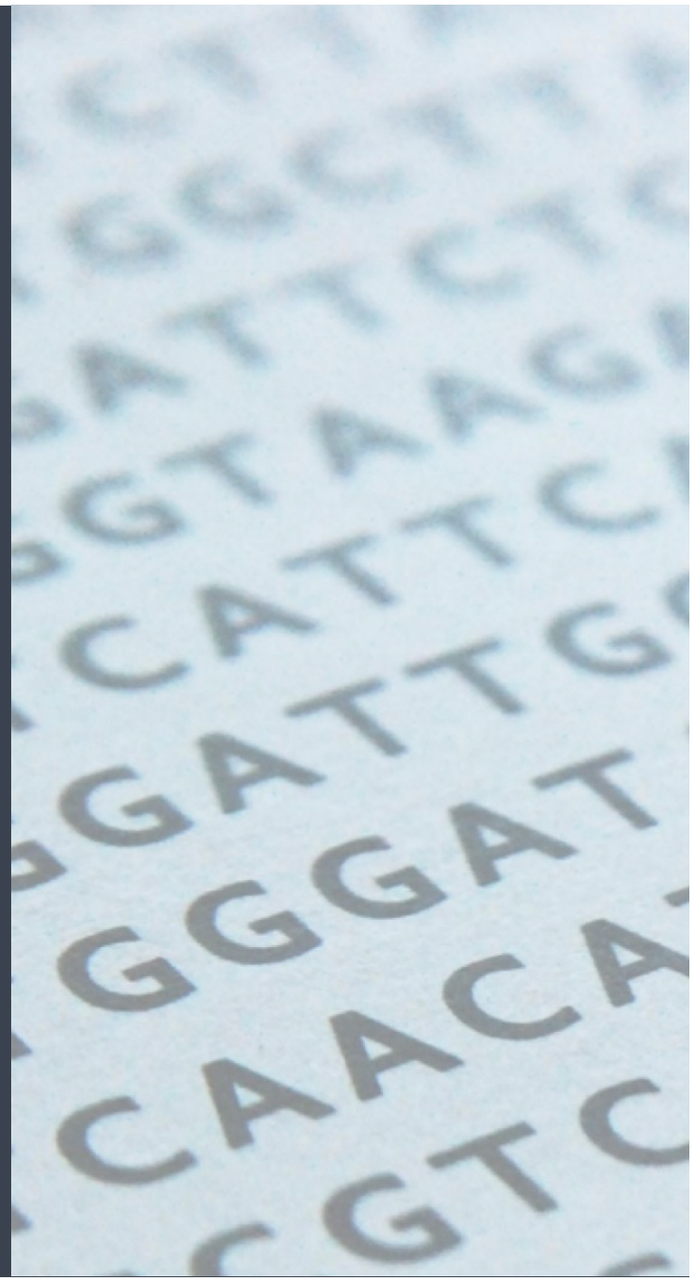


# Training

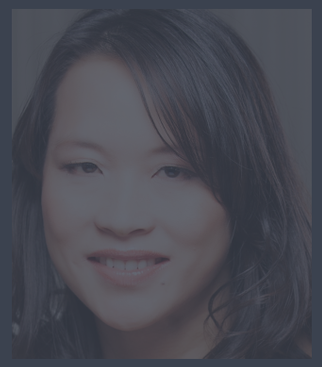
- Short workshops on introductory, intermediate and advanced topics related to NGS data analysis
- Monthly, 2-3 hour, hands-on and free workshops on “Current Topics in Bioinformatics”
- In-depth courses (8- or 12-day formats)

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

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

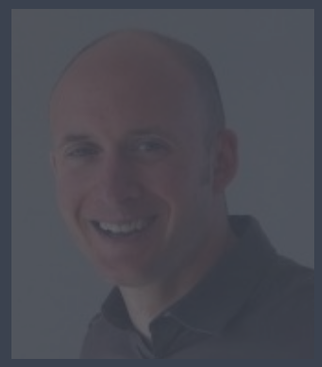


# Introductions!

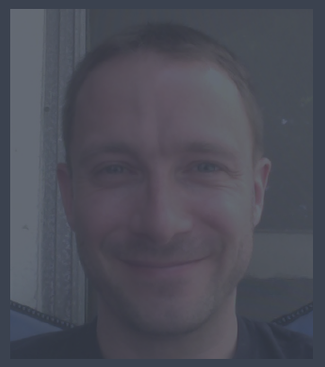


Shannan Ho Sui

John Hutchinson



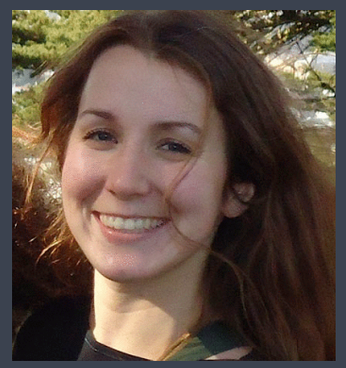
Brad Chapman



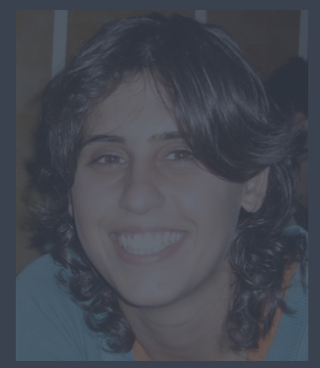
Rory Kirchner

Meeta Mistry

Radhika Khetani

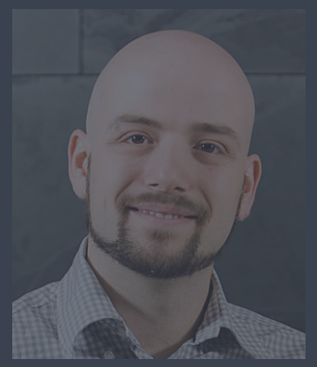


Mary Piper

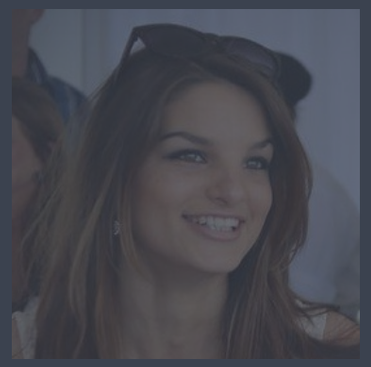


Lorena Pantano

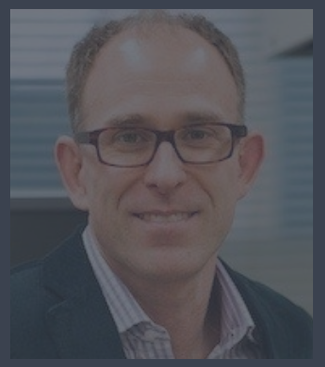
Michael Steinbaugh



Victor Barrera

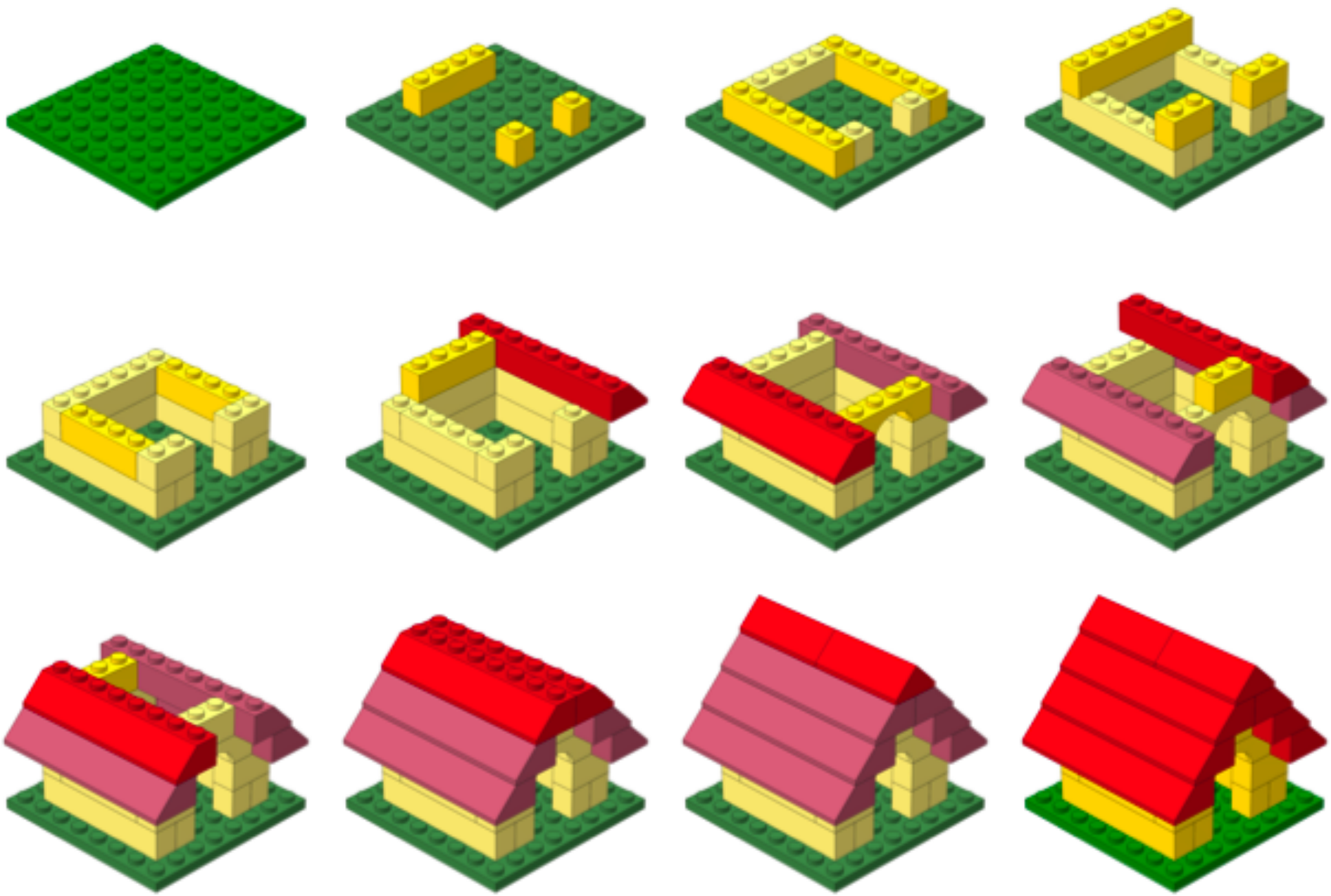


Kayleigh Rutherford



Peter Kraft

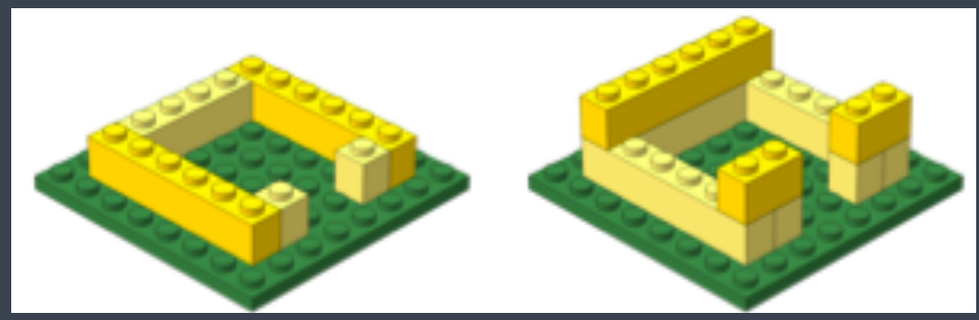
# Workshop scope



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

# Bioinformatics data analysis

# Base components



- ✓ Introduction to the 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 — 75×51
rsk27@clarinet002-072: ~
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.”*

http://en.wikipedia.org/wiki/Unix, http://en.wikipedia.org/wiki/Tux

# 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

# 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

***Bioinformatics:***

- A lot of NGS-analysis tools are created for UNIX
- Computational resources that can handle large datasets require a working knowledge of UNIX

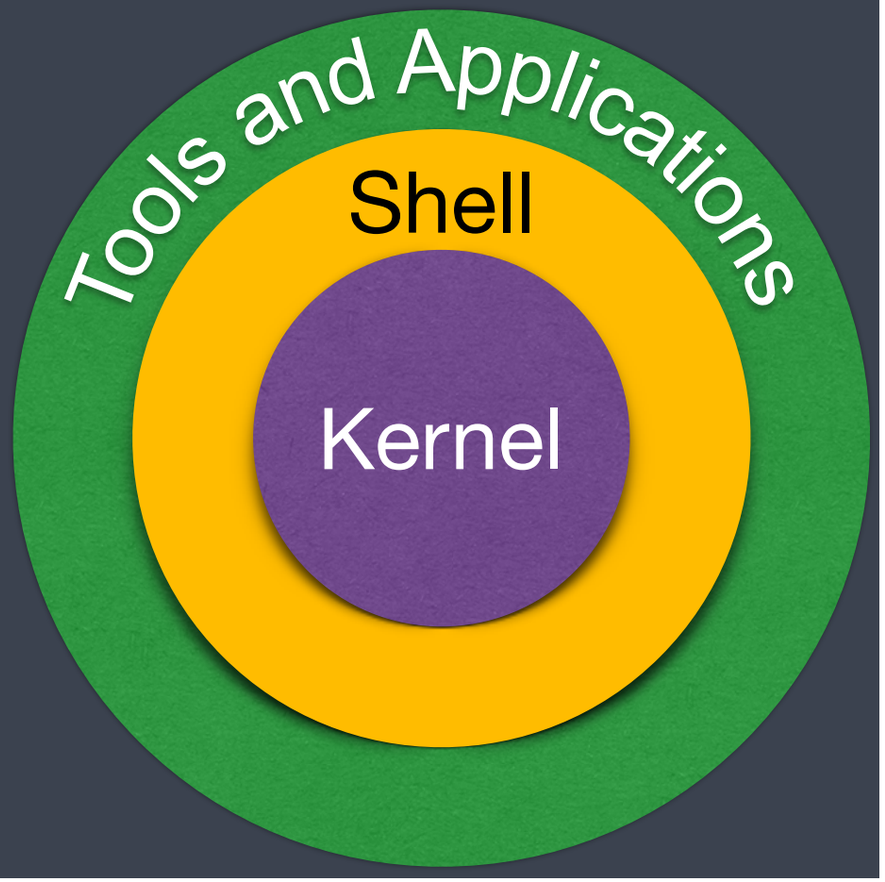
# Linux

- 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:

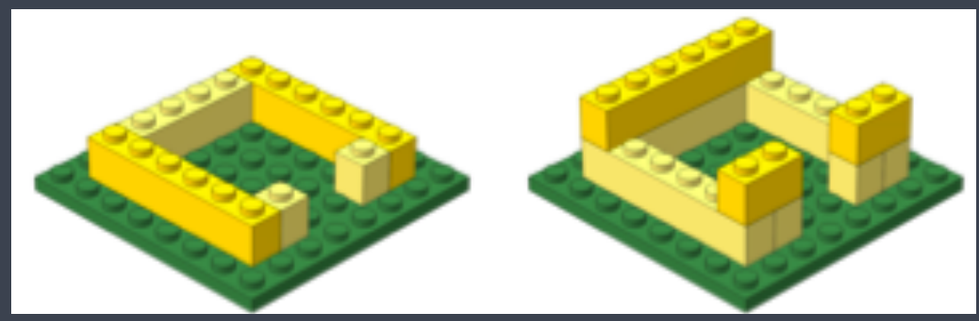
- The kernel, which schedules tasks and manages storage: ***the brain of the system***
- The shell, ***an interpreter*** that helps interprets our input for the kernel
- Utilities, tools and applications, which use the shell to communicate with the kernel



# The “shell”

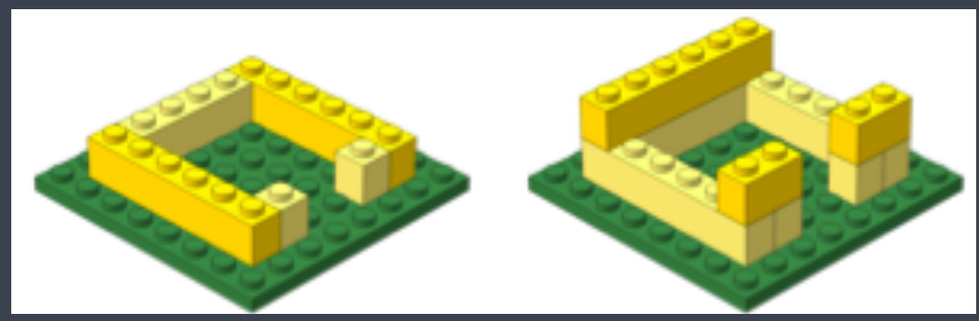
- The shell is **an interpreter**
- It is independent of the operating system
- Dozens of shells have been developed throughout UNIX history, and a lot of them are still in use
- The most commonly used shell is **bash**

# Learning Objectives



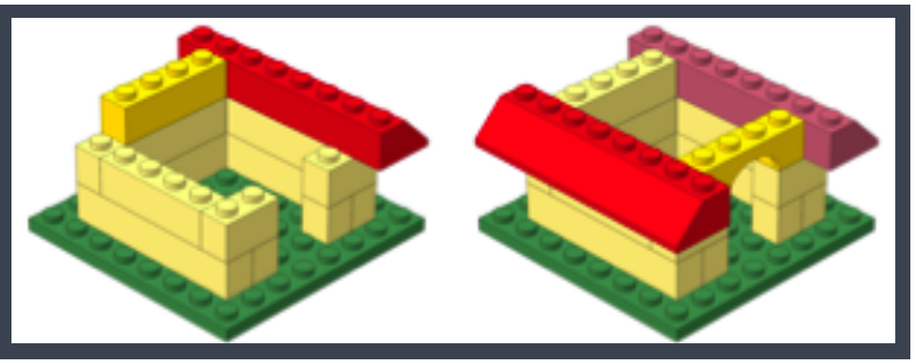
- ✓ Learn what a “shell” is and become comfortable with the command line interface
  - Find your way around a UNIX filesystem
  - Work with small and large data files
  - Become more efficient when performing repetitive tasks

# Learning Objectives



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

# Learning Objectives



- ✓ Describe best practices for designing an RNA-seq experiment
- ✓ Describe steps in an RNA-seq analysis workflow
- ✓ Use the local compute cluster to efficiently run the RNA-seq workflow from sequence files to count matrices.

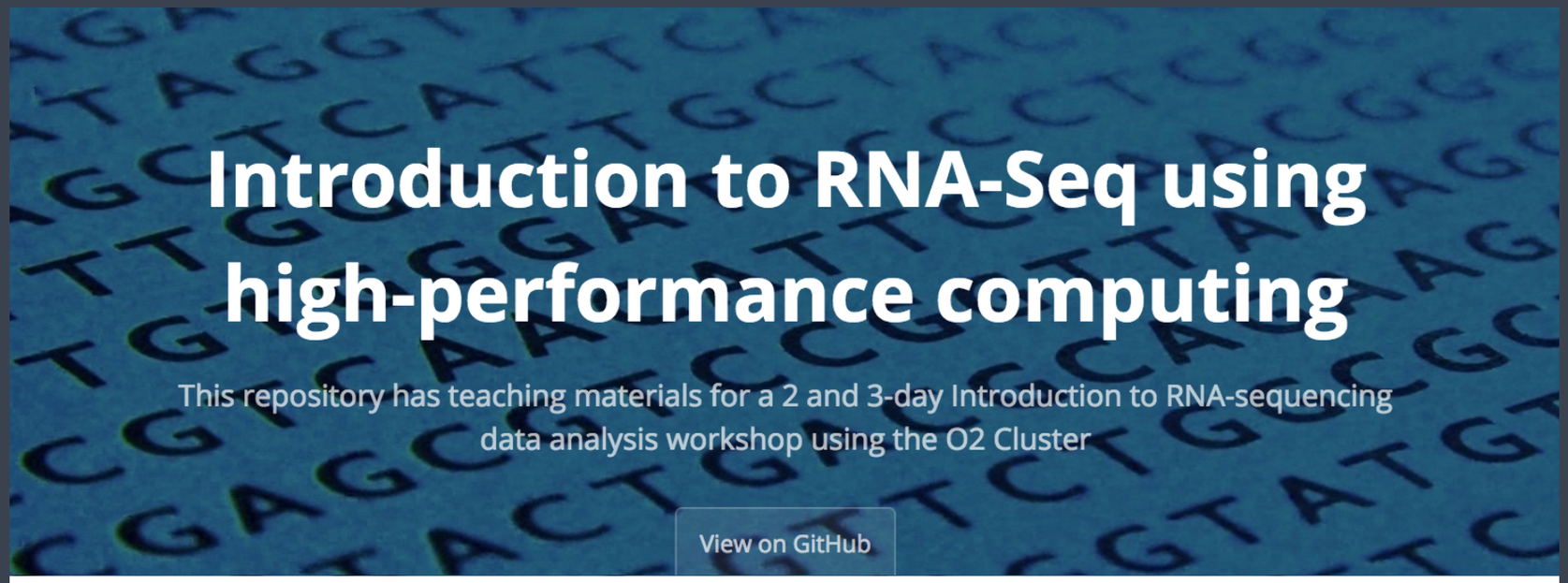
We won't be covering how to perform differential gene expression analysis on count data in this workshop, since it requires a working knowledge of R.

# Logistics

# Course webpage

https://tinyurl.com/intro-rnaseq-gt

# 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 by exploring experimental data.

Since we are going to be working with this data on our remote server, **Orchestra 2 (O2)**, we first need to log onto the server. After we're logged on, we will each make our own copy of the example data folder.

# The 2 Window problem…

rkhetani — rsk27@clarinet002-072: ~ — ssh — 141×21

rsk27@clarinet002-072: ~

```
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_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

- Post-its
- Phones on vibrate/silent!

# Contact us!

*HBC training:* hbctraining@hsph.harvard.edu

*HBC consulting:* bioinformatics@hsph.harvard.edu

## Twitter

*HBC:* @bioinfocore

*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.*