

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

Harvard Chan Bioinformatics Core

in collaboration with

HMS Research Computing

<https://tinyurl.com/intro-to-rnaseq-adv>



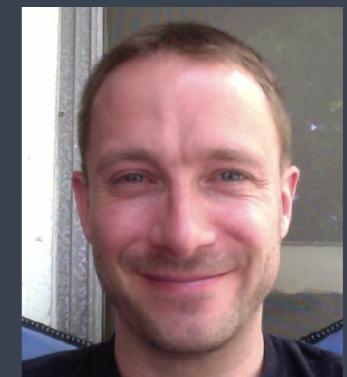
Shannan Ho Sui



John Hutchinson



Brad Chapman



Rory Kirchner



Meeta Mistry



Radhika Khetani



Mary Piper



Victor Barrera



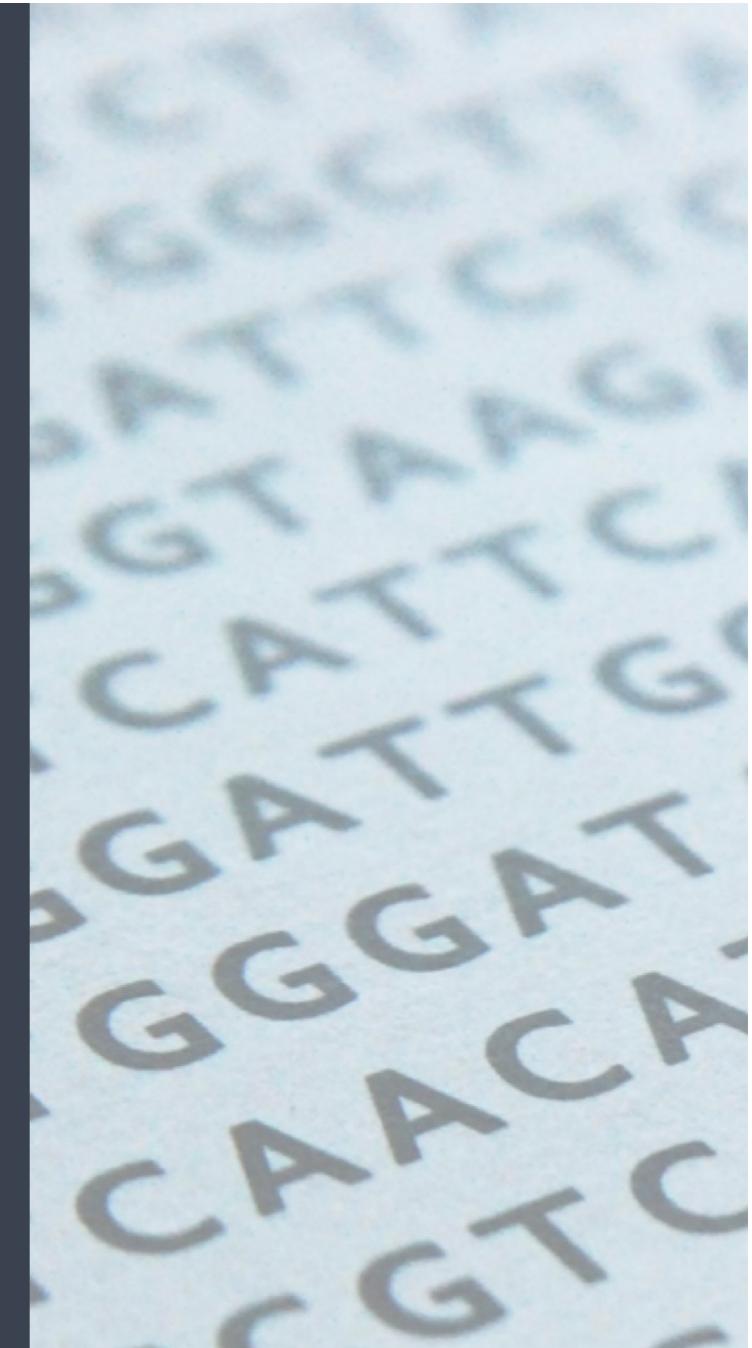
Lorena Pantano



Peter Kraft

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





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

Training

We have divided our short workshops into 2 categories:

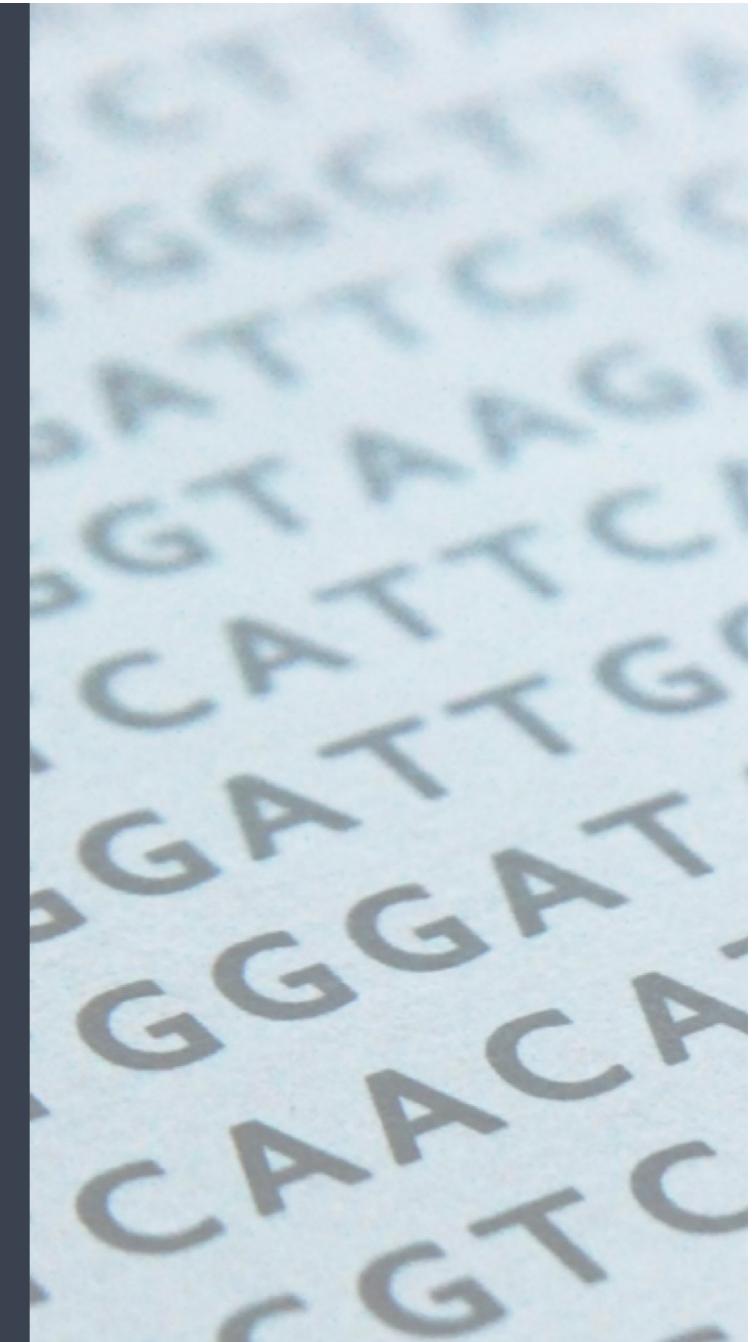
1. **Basic Data Skills** - No prior programming knowledge needed (no prerequisites)
2. **Advanced Topics: Analysis of high-throughput sequencing (NGS) data** - Certain “Basic” workshops required as prerequisites.

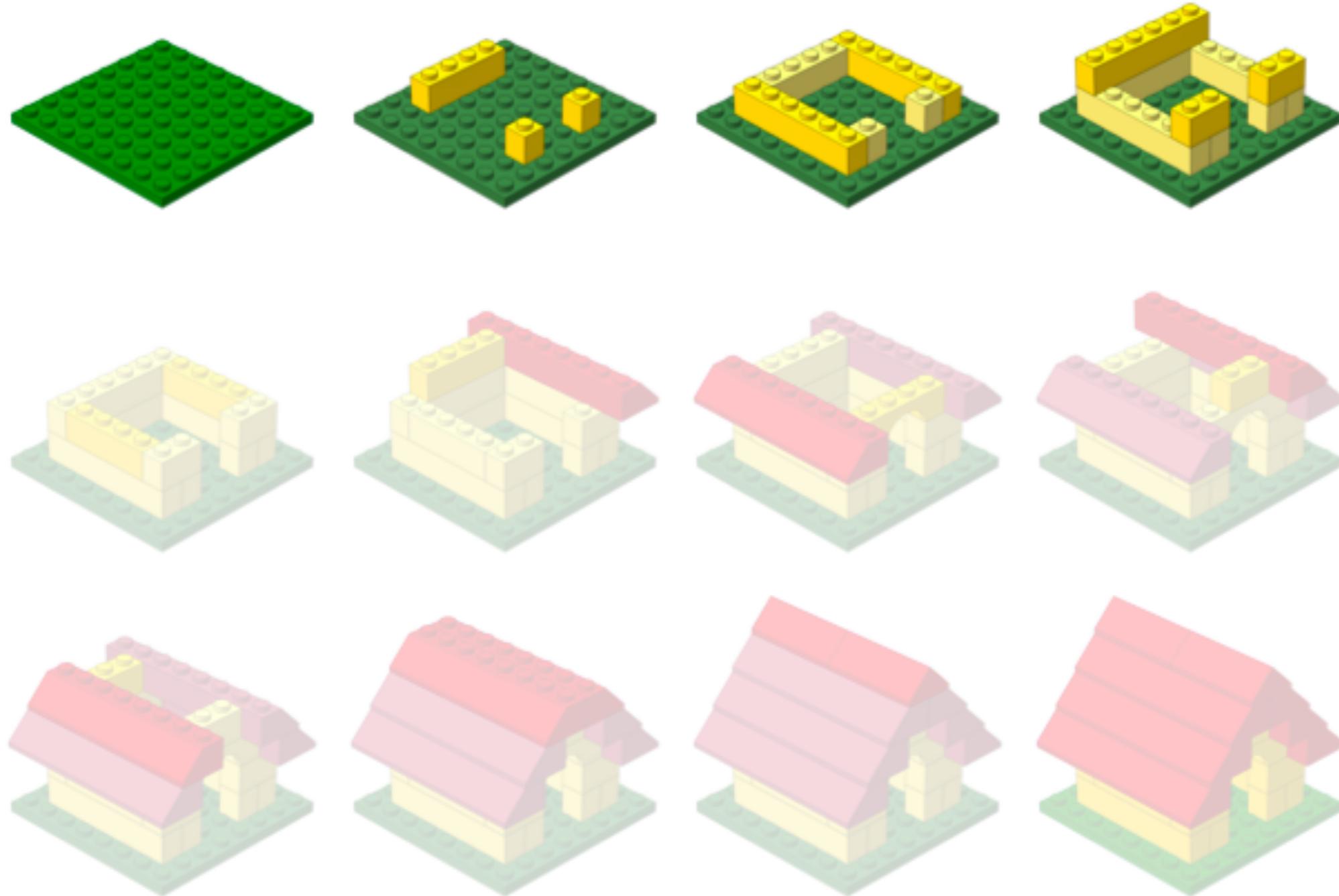
Any participants wanting to take an advanced workshop will have to have taken the appropriate basic workshop(s) within the past 6 months.

https://hbctraining.github.io/main/training_spring2019.html

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

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

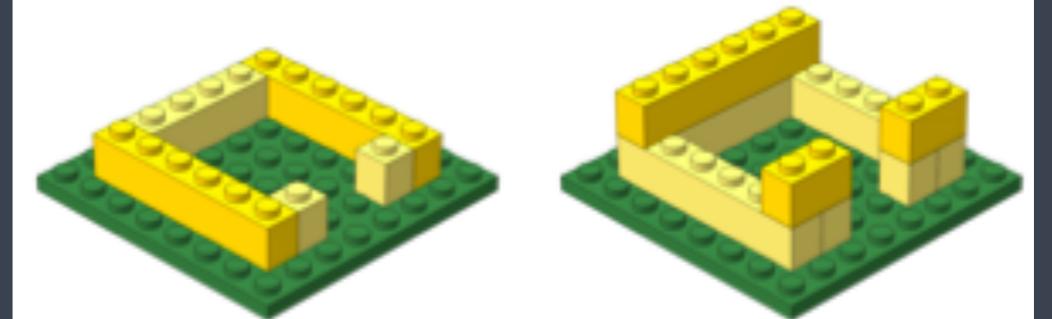




<http://anoved.net/tag/lego/page/3/>

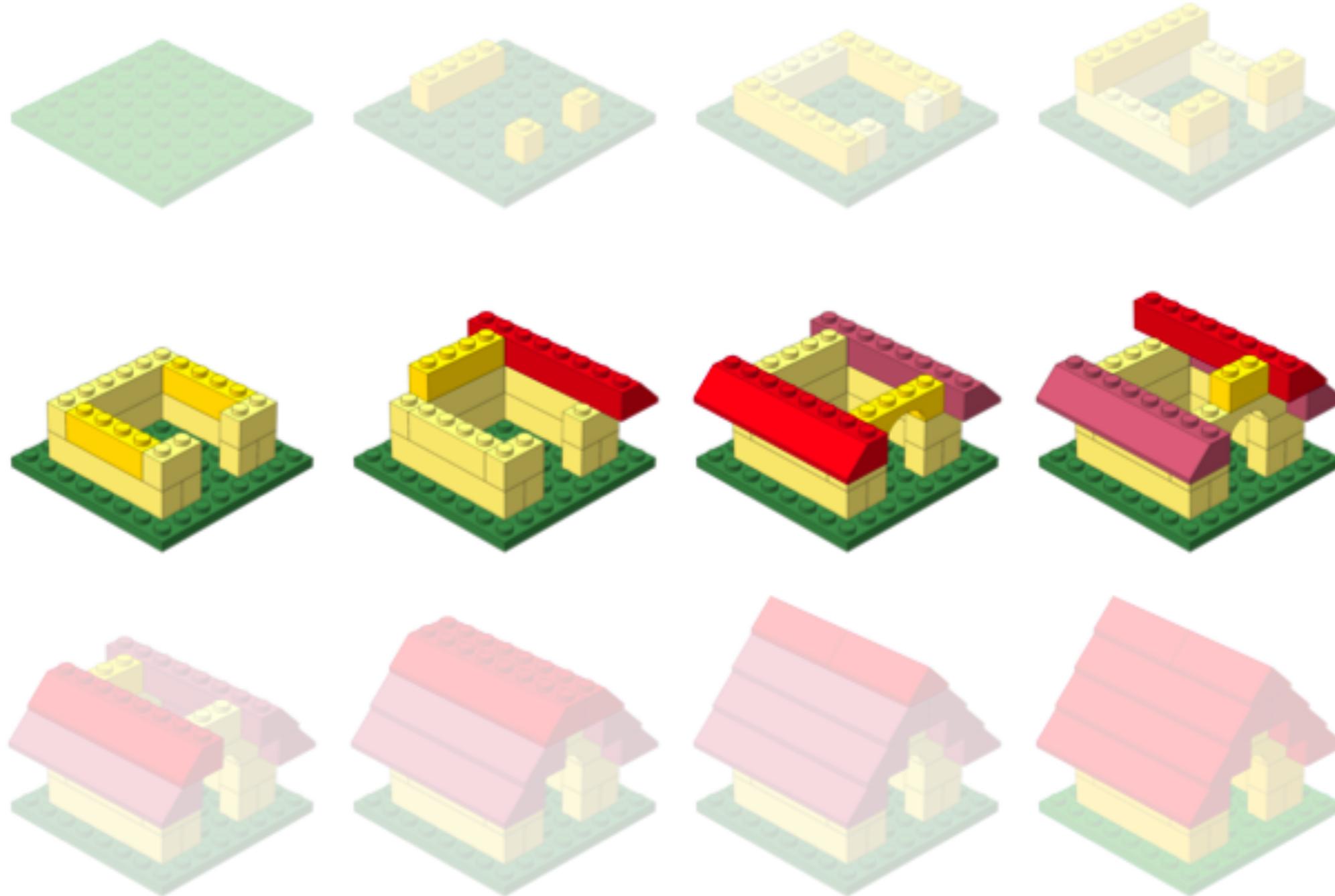
Setting up to perform Bioinformatics analysis

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
- ✓ R
 - Parsing and working with smaller results text files
 - Statistical analysis, e.g. differential expression analysis
 - Generating figures from complex data

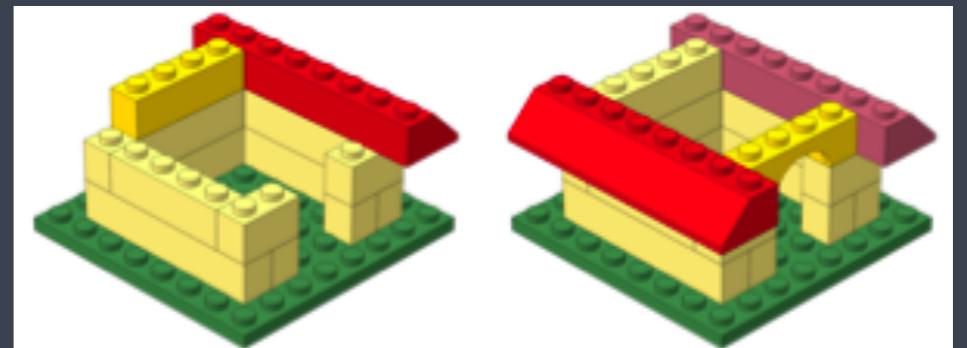
Workshop scope



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

Bioinformatics data analysis

Learning Objectives



- ✓ Describe best practices for designing a bulk RNA-seq experiment
- ✓ Describe steps in an RNA-seq analysis workflow (from sequence data to expression quantification).
- ✓ Implement shell scripts on a high-performance compute cluster to perform the above steps.

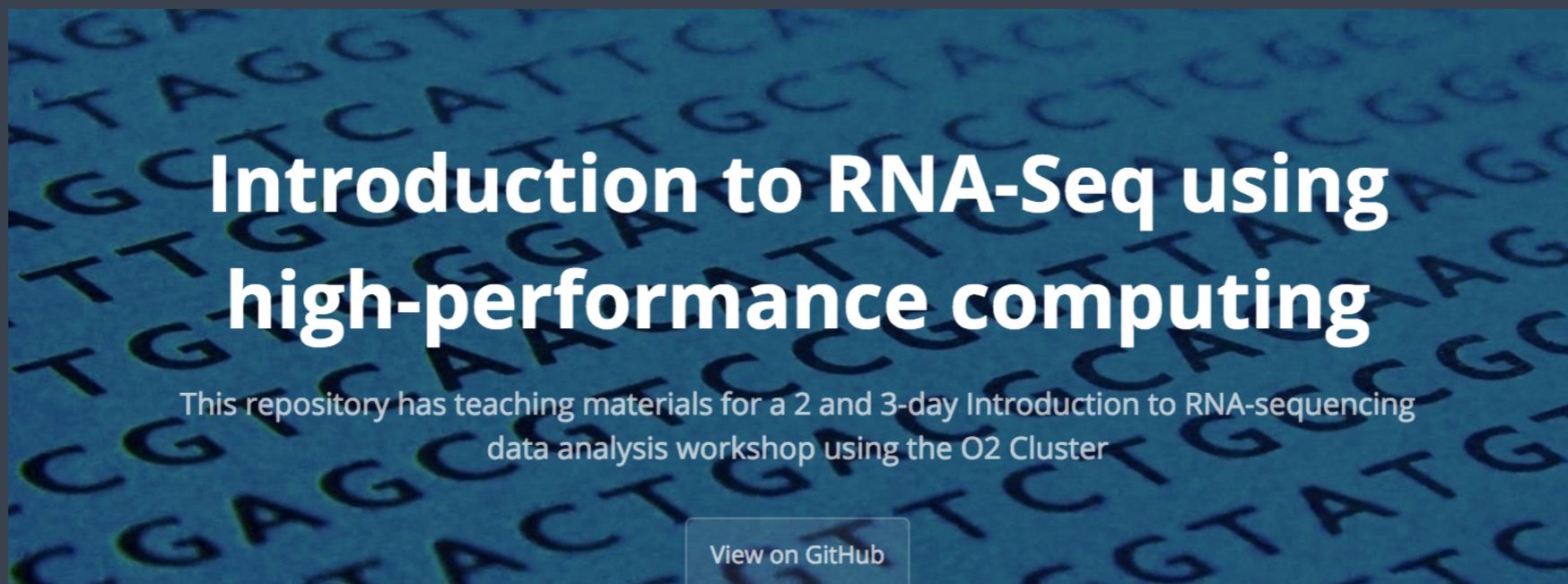
We won't be covering how to perform differential gene expression (DGE) analysis on count data in this workshop. A DGE workshop will be held on April 1st/2nd and the pre-requisite for it is a working knowledge of R (March 12th/13th).

Logistics

Course webpage (wiki)

<https://tinyurl.com/intro-to-rnaseq-adv>

Course materials online

A blue-tinted background image showing a dense grid of DNA sequence data, specifically ACGT nucleotide bases, arranged in a repeating pattern.

Introduction to RNA-Seq using high-performance computing

This repository has teaching materials for a 2 and 3-day Introduction to RNA-sequencing data analysis workshop using the O2 Cluster

[View on GitHub](#)

Learning Objectives

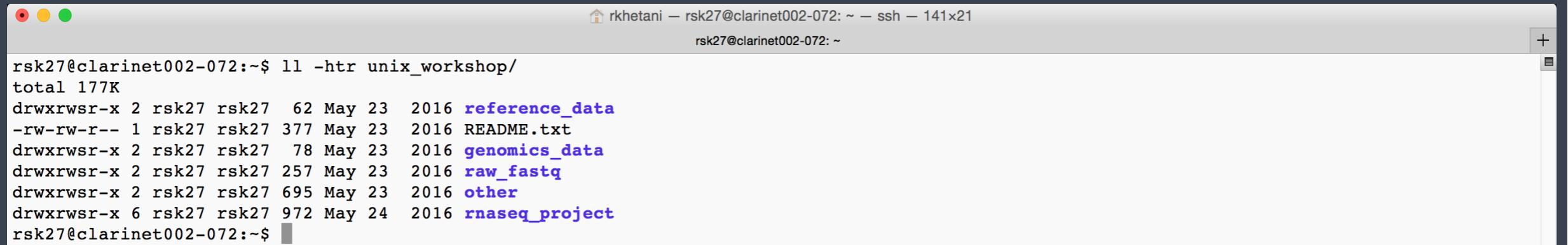
- How do you access the shell?
- How do you use it?
 - Getting around the Unix file system
 - looking at files
 - manipulating files
 - automating tasks
- What is it good for?

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...



A screenshot of a Mac OS X terminal window. The title bar shows the user is rkhetani on a host named rsk27@clarinet002-072, connected via ssh, with a window size of 141x21. The terminal prompt is rsk27@clarinet002-072:~\$. The user has run the command 'ls -ltr unix_workshop/' which lists the contents of the 'unix_workshop' directory. The output shows several files and directories with their permissions, modification times, and names. The names of the files ('reference_data', 'README.txt', 'genomics_data', 'raw_fastq', 'other', 'rnaseq_project') are highlighted in blue, indicating they are links or specific files of interest.

```
rsk27@clarinet002-072:~$ ls -ltr 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
- ❖ Phones on vibrate/silent!

Thanks!

- 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

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

HBC consulting: bioinformatics@hsph.harvard.edu

Twitter

HBC: @bioinfocore

HMS-RC: @hms_rc