

Introduction to bulk RNA-seq (Part I)

Harvard Chan Bioinformatics Core

in collaboration with

HMS Research Computing

<https://tinyurl.com/hbc-rnaseq>



Shannan Ho Sui
Director



Meeta Mistry
Associate Director



John Quackenbush
Faculty Advisor



Emma Berdan



Heather Wick



Will Gammerdinger



Noor Sohail



Upendra Bhattarai



James Billingsley



Zhu Zhuo



Maria Simoneau



Shannan Ho Sui
Director



Meeta Mistry
Associate Director



John Quackenbush
Faculty Advisor



Emma Berdan



Heather Wick



Will Gammerdinger



Noor Sohail



Upendra Bhattarai



James Billingsley



Zhu Zhuo



Maria Simoneau

Consulting

- Experimental design help
- RNA-seq analysis: bulk, single cell, small RNA
- ChIP-seq and ATAC-seq analysis
- Genome-wide methylation
- WGS, resequencing, exome-seq and CNV studies
- QC & analysis of gene expression arrays
- Functional enrichment analysis
- Grant support

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



HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NIEHS



THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER



HARVARD
MEDICAL SCHOOL

Training

A key component of the training program is for researchers at Harvard to

[HBC's training team](#) is made up of experts in training and community based research projects to ensure

Our hands-on workshops place an emphasis on **experimentation** for **wet-lab biologists** and **bioinformatics** data.

We offer three types of workshops:

1. [Short, 3-hour monthly](#)
2. [Basic Data Skills](#)**
3. [Advanced Topics: Analyzing](#)

**The basic data skills



HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

DF/HCC
DANA-FARBER / HARVARD CANCER CENTER



THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER



Our dedicated training team holds workshop to help researchers analyze and interpret NGS data.

Participants are encouraged to devote substantial time to material development, and the training team also participate in consultations on best practices in NGS analysis.

Workshops focus on the **analysis of high-throughput sequencing data**, with an emphasis on **accuracy** and **reproducibility**. Our workshops are designed to help researchers design sequencing-based experiments and analysing the resulting

(bioinformatics)

(NGS) data**

for the advanced workshops.

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

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

Training

A key component of the [unclear] researchers at Harvard b

[HBC's training team](#) is m training and community b research projects to ensu

Our hands-on workshops an emphasis on **experim** for **wet-lab biologists** ai data.

We offer three types of w

1. [Short, 3-hour monthly](#)
2. [Basic Data Skills](#)**
3. [Advanced Topics: Ana](#)

**The basic data skills



HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

DF/HCC
DANA-FARBER / HARVARD CANCER CENTER



THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER



Our dedicated training team holds workshop to help or NGS data.

to devote substantial time to material development, training team also participate in consultations on best practices in NGS analysis.

ysis of high-throughput sequencing data, with and **reproducibility**. Our workshops are designed cing-based experiments and analysing the resulting

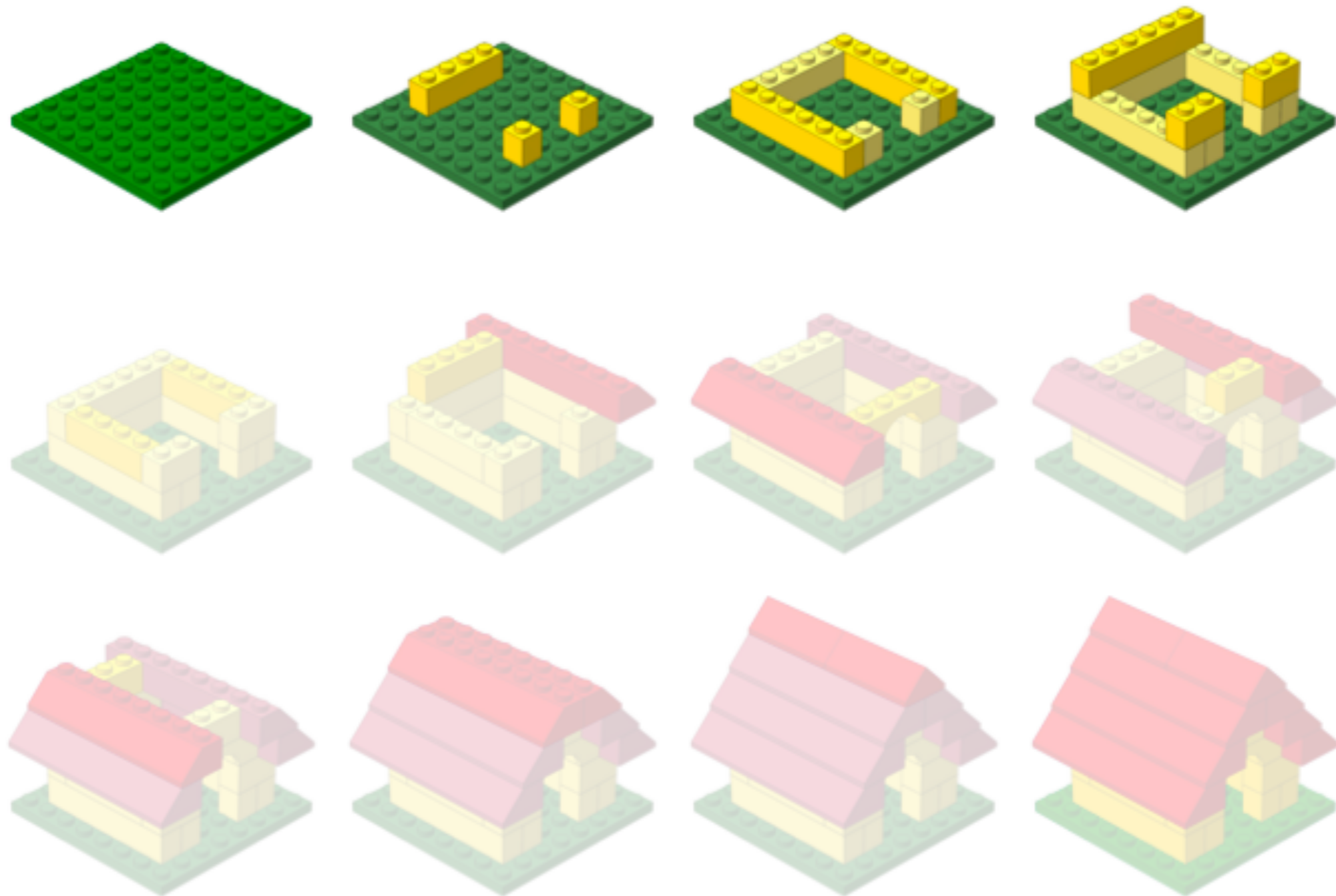
matics)

[NGS\) data](#)**

or the advanced workshops.

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

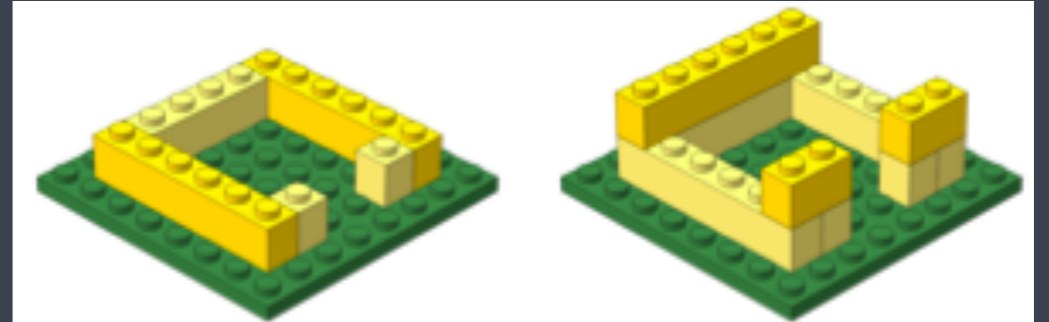
<https://hbctraining.github.io/main/>



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

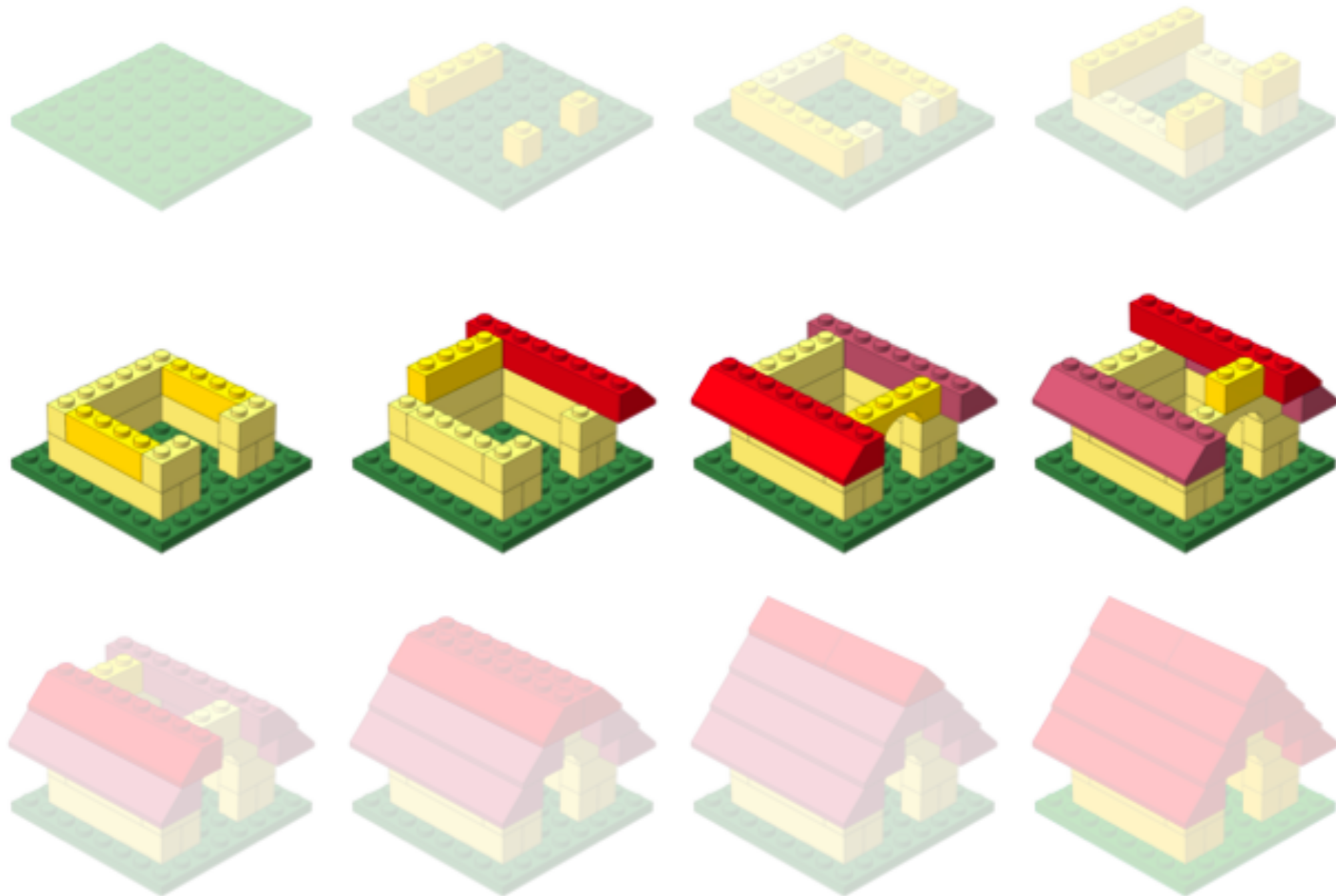
Setting up to perform Bioinformatics analysis

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

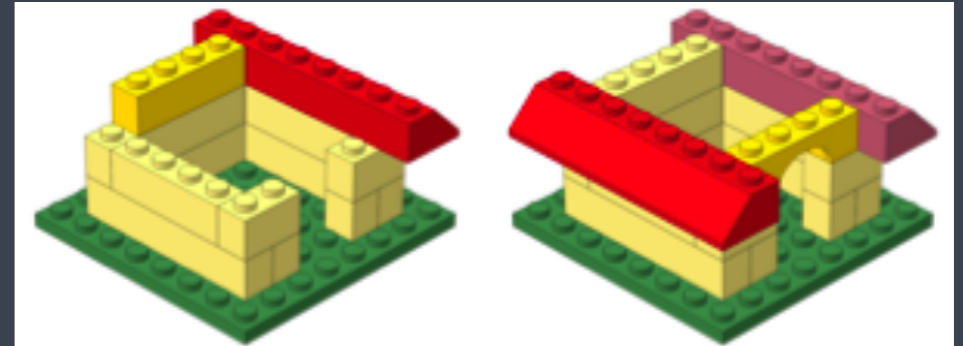
Workshop scope



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

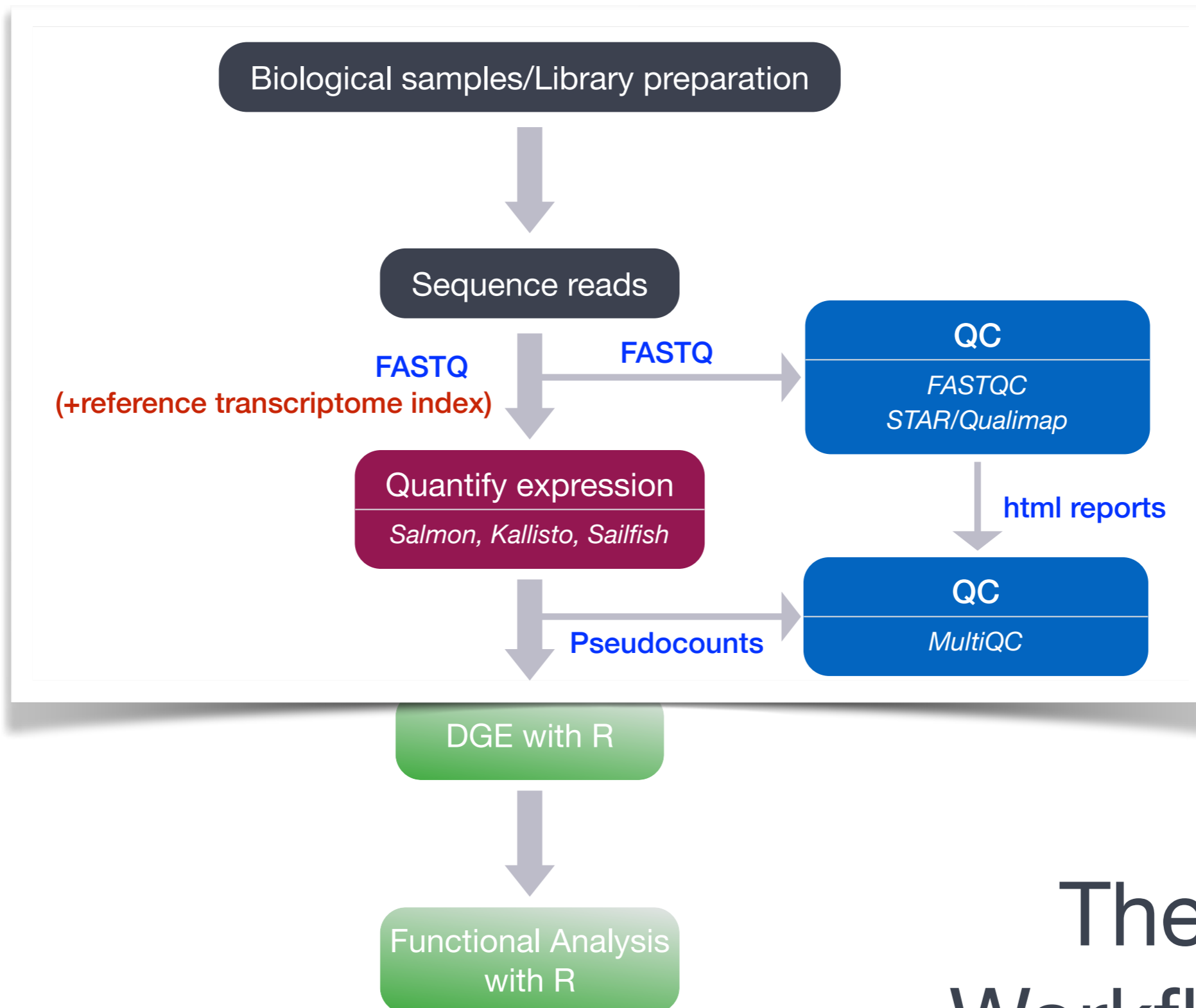
Bioinformatics data analysis

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.



The Workflow

Logistics

Course webpage

<https://tinyurl.com/hbc-rnaseq>

Course schedule online

Workshop Schedule

NOTE: The *Basic Data Skills* Introduction to the command-line interface workshop is a prerequisite.

Pre-reading

- [Shell basics review](#)
- [Introduction to RNA-seq](#)

Day 1

Time	Topic	Instructor
09:30 - 09:45	Workshop introduction	Radhika
09:45 - 10:25	Working in an HPC environment	Radhika
10:25 - 11:05	Project Organization and Best Practices in Data Management	Meeta
11:05 - 11:45	Quality Control of Sequence Data: Running FASTQC	Jihe
11:45 - 12:00	Overview of self-learning materials and homework submission	Jihe/Meeta

Course materials online

Introduction to RNA-Seq using high-performance computing

Intro to RNA-seq updated for a flipped classroom

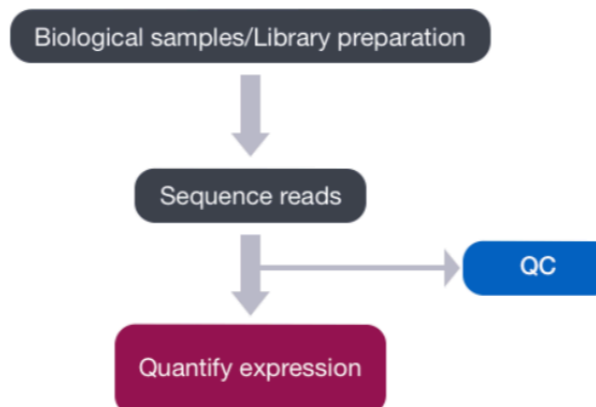
[View on GitHub](#)

Learning Objectives:

- Understand the quality values in a FASTQ file
- Create a quality report using FASTQC

Quality Control of FASTQ files

The first step in the RNA-Seq workflow is to take the FASTQ files received from the sequencing facility and assess the quality of the sequence reads.



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!



Single screen & 3 windows?

The screenshot displays a Zoom meeting interface with three windows visible:

- Terminal Window (Top Left):** Shows a shell session on a host named 'rsk27@clarinet002-072'. The user runs the command `ll -ltr unix_workshop/`, resulting in a directory listing of files and folders in the `unix_workshop` directory.
- Participants List (Top Center):** Shows three participants: Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host).
- Browser Window (Top Right):** Displays a page titled "Introduction to the command line interface (shell)" with a "View on GitHub" button. The page content includes DNA sequence motifs like "GGGATTTC" and "CAACATTCAAA".

Below the terminal window, a document titled "Starting with the shell" is visible, containing instructions and code snippets:

```
$ cd unix_workshop
```

'cd' stands for 'change directory'

Let's see what is in here. Type:

```
$ ls
```

At the bottom, a terminal window from a host named 'rsk394' is shown, displaying the execution of a complex pipeline command:

```
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 screenshot shows a Zoom meeting interface. At the top, there are three video thumbnails for participants: Mary Piper, Troubleshooter (...), and Jihe Liu. Below the thumbnails is a list of participants: Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host). The main content area is divided into three windows:

- Terminal Window (Top Left):** Shows a file listing in a terminal window:

```
rsk27@clarinet002-072:~$ ll -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:~$
```
- Terminal Window (Bottom Left):** Shows a terminal window with the command `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\ and\ reference\ data/chr1-hg19_genes.gtf | sort -k2n | head` and its output:

```
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$
```
- Web Browser Window (Right):** Shows a page titled "Introduction to the command line interface (shell)" with a "View on GitHub" button.

A large blue "ZOOM" watermark is overlaid on the terminal windows. The Zoom meeting controls are visible at the bottom of the screen.

*Our
recommendation*

Single screen & 3 windows?

The screenshot shows a Zoom meeting interface. At the top, there are three video thumbnails for participants: Mary Piper, Troubleshooter (...), and Jihe Liu. Below the thumbnails is a terminal window showing the output of a command: `ls -ltr`. The output lists files and directories in a table format. To the right of the terminal is a 'Participants (3)' list showing Mary Piper (Co-host, me), Jihe Liu (Host), and Troubleshooter (Radhika) (Co-host). Below the participants list is a web browser window displaying a page titled 'Introduction to the command line interface (shell)'. The browser window has a red border and contains the text 'Web browser' in large red letters. Below the browser window is another terminal window showing a command: `cut -f 1,4 /Users/rsk394/Dropbox/.../reference/data/chr1-hg19_genes.gtf | sort -k2n | head`. The output of this command is a list of chromosome 1 coordinates. At the bottom of the Zoom interface, there are controls for Unmute, Stop Video, Invite, Share Screen, and Reactions.

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

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\
and\
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 shows a Zoom meeting interface. At the top, there are three video thumbnails for participants: Mary Piper, Troubleshooter (...), and Jihe Liu. Below the thumbnails is a 'Participants (3)' list with icons for each participant. In the background, a browser window displays a page titled 'Introduction to the command line interface (shell)' with a 'View on GitHub' button. In the foreground, a terminal window is open, showing the output of a command. The terminal output is as follows:

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

The terminal window is highlighted with a green border. Below the terminal window, there is a text box with the following text:

```
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?

The image shows a Zoom meeting interface with three participants: Mary Piper, Troubleshooter (Radhika), and Jihe Liu. The main window displays a terminal window with the following output:

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

A second terminal window is overlaid on top, showing 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$
```

A web browser window is also visible, showing a page titled "Introduction to the command line interface (shell)" with a "View on GitHub" button. The text "Web browser" is overlaid in red on the browser window.

The word "ZOOM" is overlaid in blue on the terminal window.

*Our
recommendation*

Terminal



Odds and Ends (1/2)

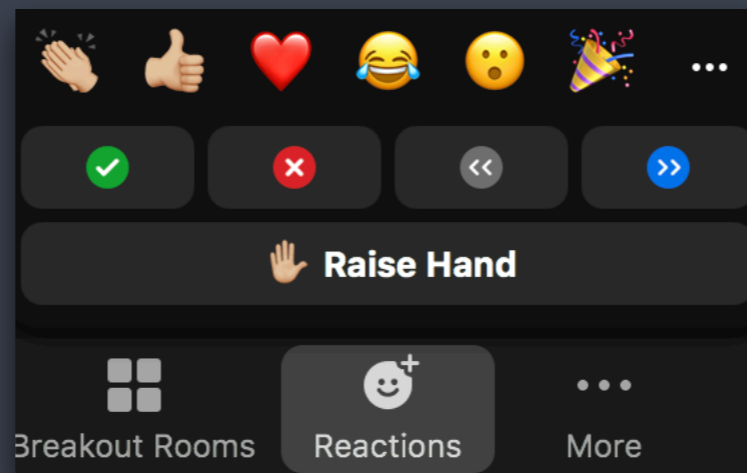
- ❖ Quit/minimize all applications that are not required for class

Odds and Ends (1/2)

- ❖ Quit/minimize all applications that are not required for class
- ❖ Captioning is available upon request

Odds and Ends (1/2)

- ❖ Quit/minimize all applications that are not required for class
- ❖ Captioning is available upon request
- ❖ Are you all set?
 - ▶  = "agree", "I'm all set" (equivalent to a **green post-it**)
 - ▶  = "disagree", "I need help" (equivalent to a **red post-it**)



Odds and Ends (2/2)

❖ Questions for the presenter?

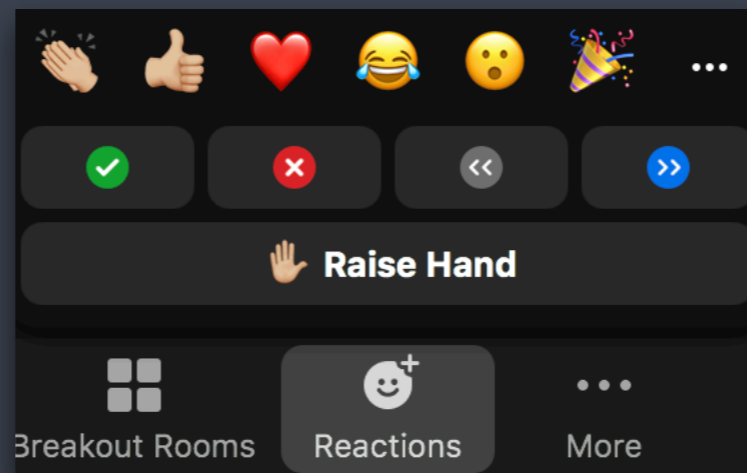
- Post the question in the Chat window OR

-  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!

- Andy Bergman & Kathleen Chappell 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