



```
dds <- DESeqDataSetFromMatrix(countData = cts,
                                colData = coldata,
                                design= ~ batch + condition)

dds <- DESeq(dds)
resultsNames(dds) # lists the coefficients
res <- results(dds, name="condition_trt_vs_untrt")
# or to shrink log fold changes association with condition:
res <- lfcShrink(dds, coef="condition_trt_vs_untrt", type="apeglm")
```

Bulk RNA-seq Analysis Part II

Differential Gene Expression

Harvard Chan Bioinformatics Core

<https://tinyurl.com/hbc-dge-online>



Shannan Ho Sui
Director



Victor Barrera



Amelie Jule



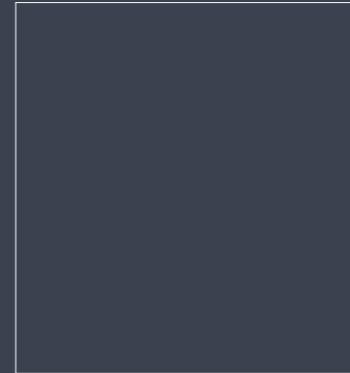
Zhu Zhuo



Radhika Khetani
Director of Education



Meeta Mistry



Heather Wick
Starts on June 19th



Will Gammerdinger



Emma Berdan



Sergey Naumenko



Maria Simoneau



Noor Sohail



James Billingsley

Consulting

- 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 HBC's mission is its training initiative. Our dedicated training team holds workshop to help researchers at Harvard better understand analytical methods for NGS data.

HBC's training team is made up of four PhD-level scientists who devote substantial time to material development, training and community building/outreach. All members of the training team also participate in consultations on research projects to ensure they remain up-to-date on current best practices in NGS analysis.

Our hands-on workshops focus on **basic data skills** and **analysis of high-throughput sequencing data**, with an emphasis on **experimental design**, current **best practices** and **reproducibility**. Our workshops are designed for **wet-lab biologists** aiming to independently design sequencing-based experiments and analysing the resulting data.

We offer three types of workshops:

1. Short, 3-hour monthly workshops (*Current topics in bioinformatics*)
2. Basic Data Skills**
3. Advanced Topics: Analysis of high-throughput sequencing (NGS) data**

***The basic data skills workshops serve as the foundation for the advanced workshops.*

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

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

Training

A key component of the HBC's mission is to provide training for researchers at Harvard and beyond.

HBC's training team is made up of experts in training and community based research projects to ensure that our trainees are well prepared for their future careers.

Our hands-on workshops focus on practical skills, with an emphasis on **experimental design** and **bioinformatics**, designed for **wet-lab biologists** and **bioinformaticians** who work with genomic data.

We offer three types of workshops:

1. Short, 3-hour monthly workshops
2. Basic Data Skills**
3. Advanced Topics: Analysis of high-throughput sequencing (NGS) data

**The basic data skills workshop is currently available online.



**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 workshops to help researchers learn how to analyze genomic and NGS data.

In addition to devote substantial time to material development, 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 **experimental design**, **bioinformatics**, and **reproducibility**. Our workshops are designed to help researchers design experiments and analyse the resulting data.

bioinformatics)

analysis of NGS) data**

and **bioinformatics** for the advanced workshops.

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

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

Training

A key component of the HBC's mission is to support researchers at Harvard by providing training.

HBC's training team is made up of scientists who provide training and community building for research projects to ensure the quality of our work.

Our hands-on workshops focus on **bioinformatics**, with an emphasis on **experimental design** and **data analysis**. We also provide training for **wet-lab biologists** aiming to understand their data.

We offer three types of workshops:

1. Short, 3-hour monthly workshops
2. Basic Data Skills**
3. Advanced Topics: Analysis of high-throughput sequencing data

**The basic data skills workshop is designed for researchers who have no prior experience with bioinformatics.



**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 workshops to help researchers learn how to analyze high-throughput sequencing (NGS) data.

The training team also devote substantial time to material development, and our 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 **experimental design**, **data quality**, and **reproducibility**. Our workshops are designed to help researchers understand the principles of sequencing-based experiments and analysing the resulting data.

bioinformatics)

basic data skills (e.g., NGS) data**

and **advanced topics** (e.g., for the advanced workshops).

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

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

Introductions!



Shannan Ho Sui
Director



Victor Barrera



Amelie Jule



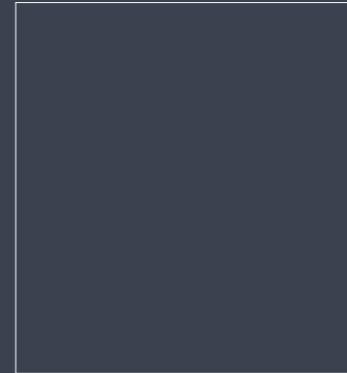
Zhu Zhuo



Radhika Khetani
Director of Education



Meeta Mistry



Heather Wick
Starts on June 19th



Will Gammerdinger



Emma Berdan



Sergey Naumenko



Maria Simoneau



Noor Sohail



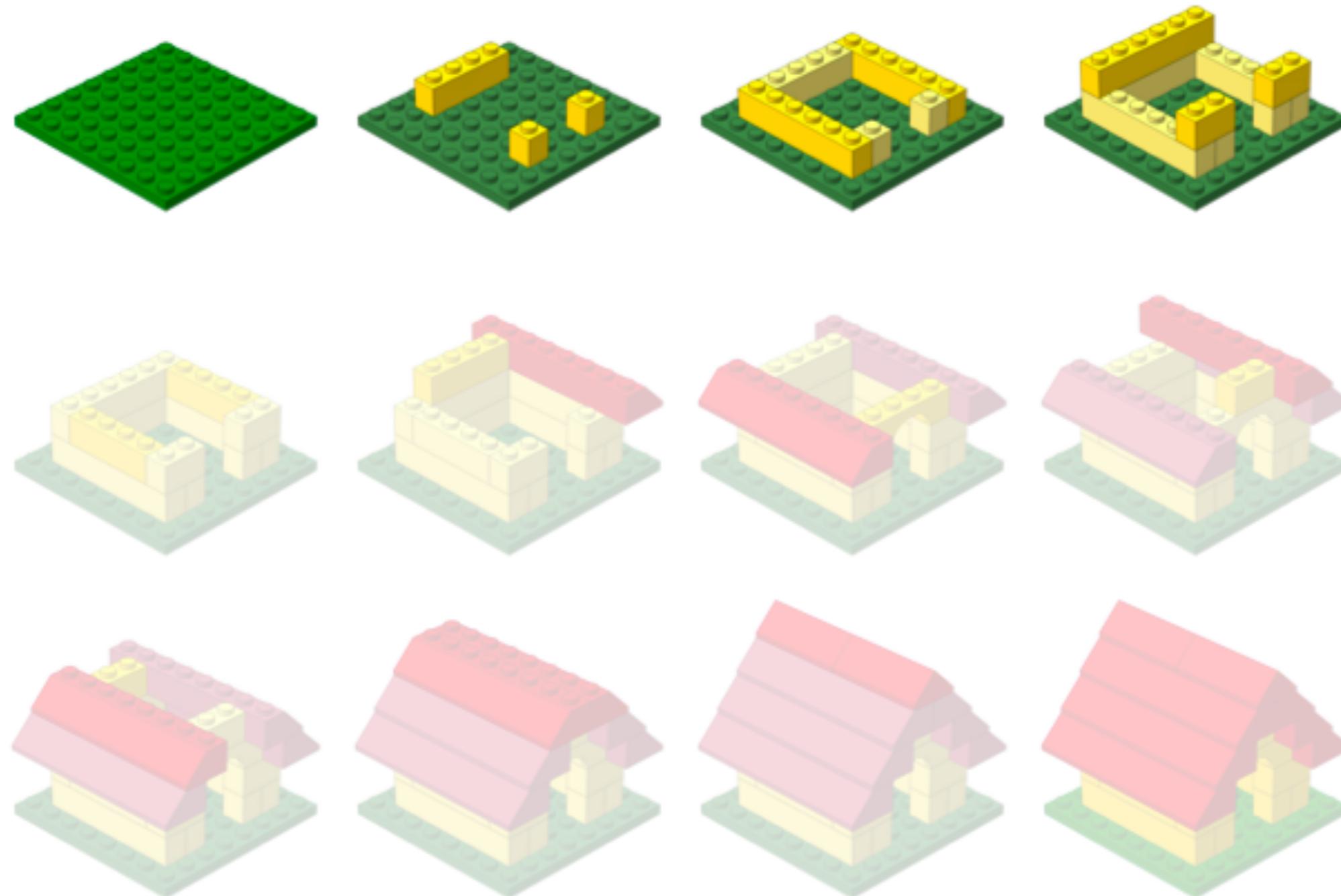
James Billingsley

Introductions!

How do you pronounce your name?

How do you plan to use differential gene expression?

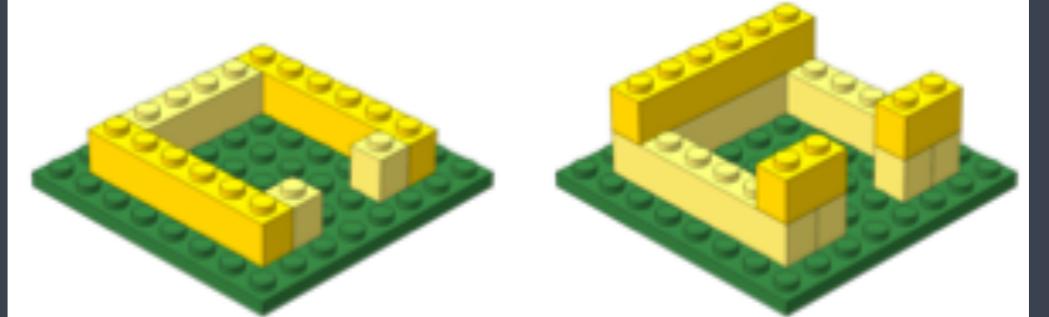
Workshop Scope...



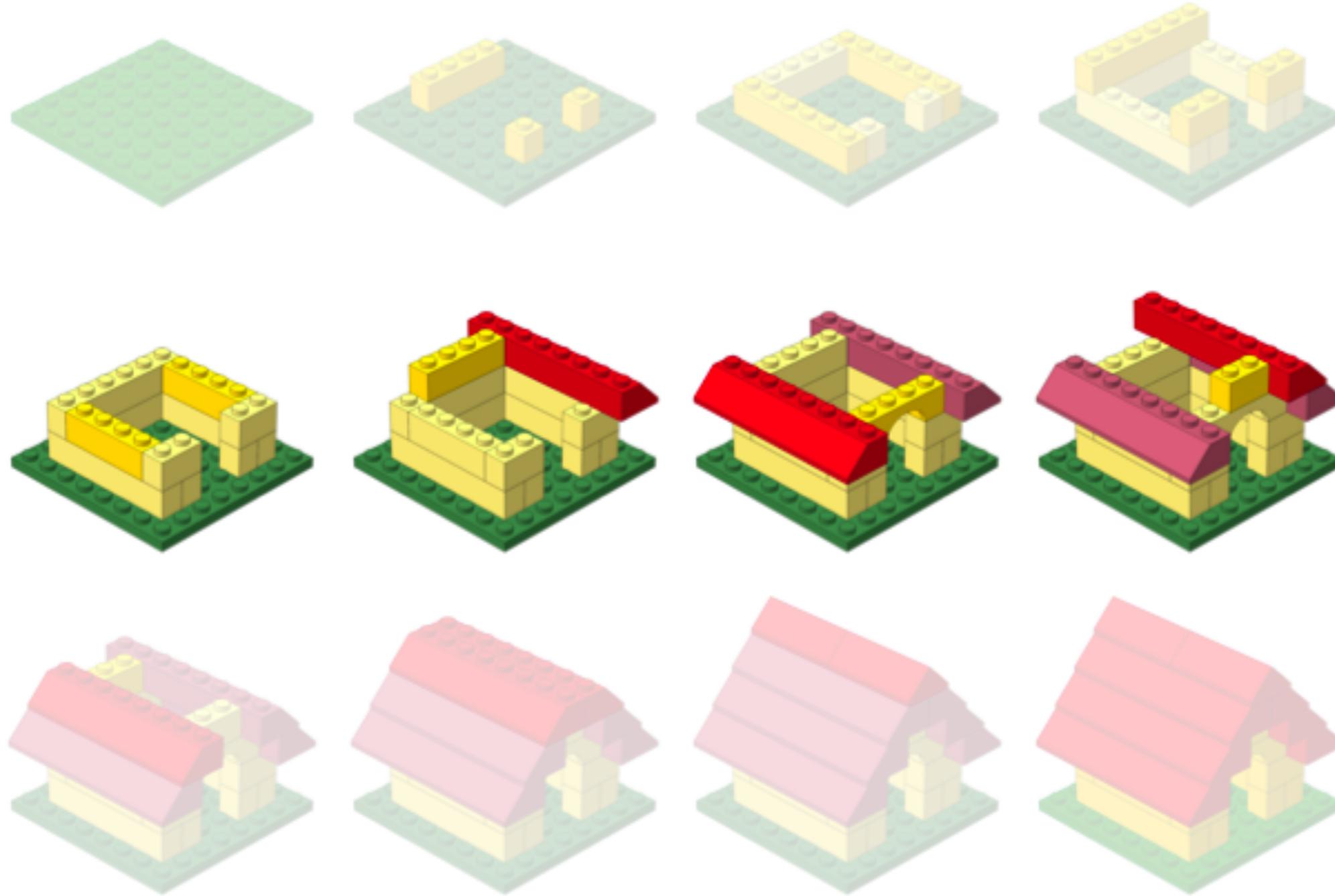
<http://anoved.net/tag/lego/page/3/>

Setting up to perform Bioinformatics analysis

Setting up...



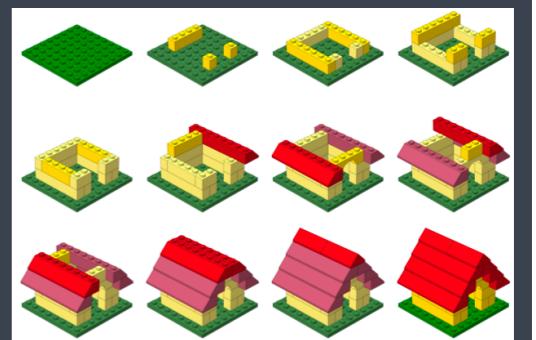
- ✓ Introduction to the command-line interface (shell, Unix, Linux)
 - Dealing with large data files
 - Performing bioinformatics analysis
 - Using tools
 - Accessing and using compute clusters
- ✓ R
 - Parsing and working with smaller results text files
 - Statistical analysis, e.g. differential expression analysis
 - Generating figures from complex data



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

Bioinformatics data analysis

Workshop Scope



Differential Gene Expression analysis

- ✓ Understand the considerations for performing statistical analysis on RNA-seq data
- ✓ Start with gene counts (after alignment and counting)
- ✓ Perform QC on count data
- ✓ Use DESeq2 to perform differential expression analysis on the count data and obtain a list of significantly different genes
- ✓ Visualize results of the analysis
- ✓ Perform functional analysis on the lists of differentially expressed genes

Logistics

Course webpage

<https://tinyurl.com/hbc-dge-online>

Course schedule online

Workshop Schedule

Pre-reading

1. [Workflow \(raw data to counts\)](#)
2. [Experimental design considerations](#)

Day 1

Time	Topic	Instructor
10:00 - 10:30	Workshop Introduction	Jihe
10:30 - 10:45	R refresher Q & A	Radhika
10:45 - 11:15	RNA-seq pre-reading discussion	Radhika
11:15 - 12:00	Intro to DGE / setting up DGE analysis	Meeta

Before the next class:

1. Please **study the contents** and **work through all the code** within the following lessons:
 - [RNA-seq counts distribution](#)
 - [Count normalization](#)
 - [Sample-level QC \(PCA and hierarchical clustering\)](#)
2. **Complete the exercises:**
 - Each lesson above contain exercises; please go through each of them.
 - **Copy over** your code from the exercises into a text file.
 - **Upload the saved text file to Dropbox** the **day before the next class**.

Questions?

- **If you get stuck due to an error** while running code in the lesson, [email us](#)
- Post any **conceptual questions** that you would like to have **reviewed in class here**.

Course webpage

Introduction to DGE

[View on GitHub](#)

Approximate time: 60 minutes

Learning Objectives

- Explore different types of normalization methods
- Become familiar with the `DESeqDataSet` object
- Understand how to normalize counts using DESeq2

Normalization

The first step in the DE analysis workflow is count normalization, which is necessary to make accurate comparisons of gene expression between samples.

```
graph TD; A["Pseudocounts with  
Kallisto, Sailfish, Salmon"] --> B["Read counts  
associated with genes"]; B --> C["Normalization"]; C --> D["Unsupervised clustering analyses"]; C -.-> E["Quality control"]
```

The diagram illustrates the DE analysis workflow. It starts with 'Pseudocounts with Kallisto, Sailfish, Salmon', followed by 'Read counts associated with genes'. This leads to 'Normalization', which then leads to 'Unsupervised clustering analyses'. A bracket on the right side groups 'Normalization' and 'Unsupervised clustering analyses' under the heading 'Quality control'.

The 2 Window problem...

The screenshot shows the RStudio interface. The top bar displays the path: ~/Dropbox (HBC)/HBC Team Folder (1)/Teaching/Intro-to-R - RStudio. The left pane contains an R script named "Intro-to-R.R" with the following code:

```
351
352 animals[4, 2] <- "Gray"
353
354 animals$color <- factor(animals$color)
355 animals$new2 <- c(1,2,3)
356
357 vector1 <- c(6:11)
358 data.frame(animals[, 1:2], vector1, animals[, 3:4])
359
360
362:1 (Top Level) ▾
```

The console window below shows the R environment:

```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/Dropbox (HBC)/HBC Team Folder (1)/Teaching/Intro-to-R/.RData]

>
```

The right pane shows the Global Environment and a file browser:

Name	Size	Modified
..		
.RData	5.8 MB	May 3, 2018, 1:40
.Rhistory	17.4 KB	Nov 15, 2018, 1:2
data		
de_sleuth.R	2.6 KB	Oct 10, 2018, 10:0
figures		
Intro-to-R.R	11.9 KB	May 1, 2018, 3:31

A callout box highlights the code in the script editor:

```
rownames(metadata)

metadata[c("sample10", "sample12"),]
```

The text "Selecting using indices with logical operators" is displayed below the highlighted code.

The explanatory text below states:

With dataframes, similar to vectors, we can use logical vectors for specific columns in the dataframe to select only the rows in a dataframe with TRUE values at the same position or index as in the logical vector. We can then use the logical vector to return all of the rows in a dataframe where those values are TRUE.

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!



Homework and Expectations

- ❖ 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 can be heavier in the beginning of this workshop series, but it tapers off

Odds and Ends

- ❖ Name tags
- ❖ Post-its
 - green - I am all set
 - red - I need time/help
- ❖ Quit/minimize all applications that are not required for class
- ❖ Phones on vibrate/silent
- ❖ Bathrooms

Contact us!

HBC training team: hbctraining@hsph.harvard.edu

HBC consulting: bioinformatics@hsph.harvard.edu