

# Differential gene expression analysis workshop

---



**Owen M. Wilkins, PhD**

Bioinformatics scientist

Center for Quantitative Biology, Geisel School of Medicine at Dartmouth

**Email:** [DataAnalyticsCore@groups.dartmouth.edu](mailto:DataAnalyticsCore@groups.dartmouth.edu)

**Website:** (<https://sites.dartmouth.edu/cqb/projects-and-cores/data-analytics-core/>)

---

08/23/21



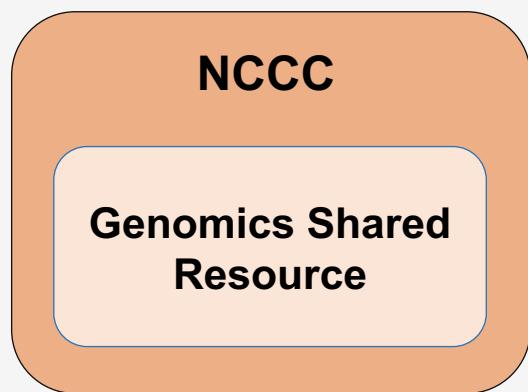
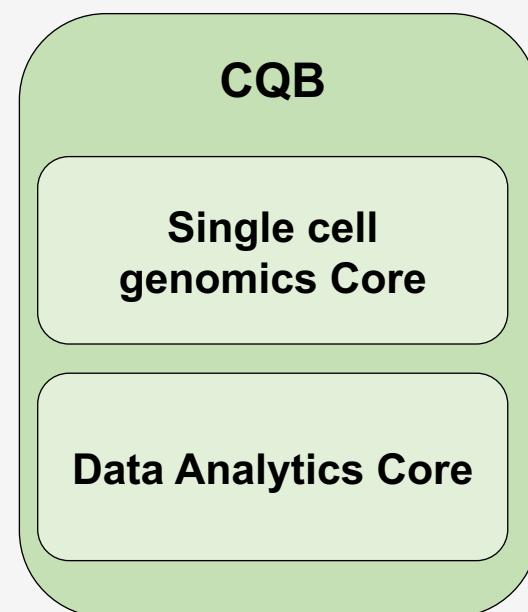
# Center for Quantitative Biology at Dartmouth

**Mission:** Support and enhance quantitative biological research at Dartmouth and to facilitate its integration with experimental biology

## Activities:

- Invest in instruments and infrastructure
- Improved sample management
- Support for new method development
- Pilot grants for new users or novel projects
- Dedicated data analysis resources

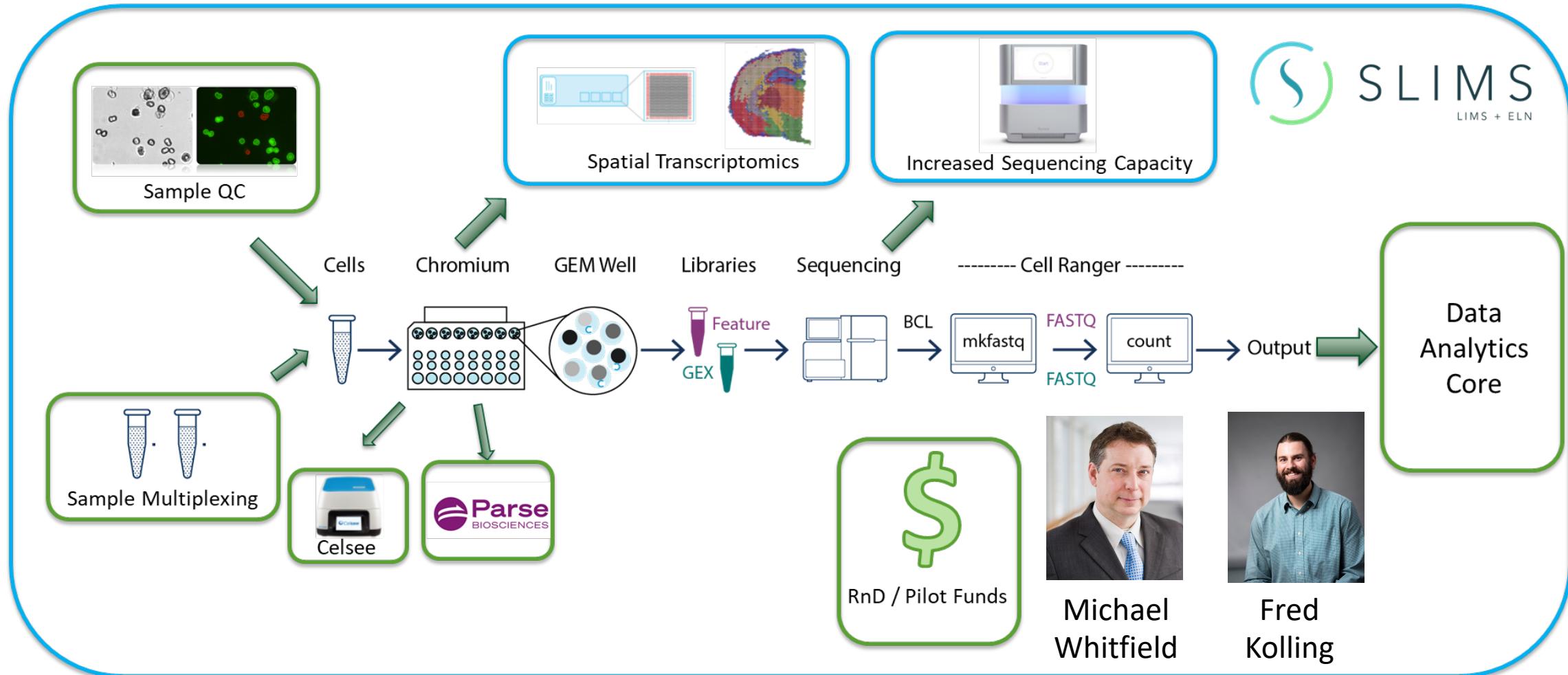
**Website:** <https://sites.dartmouth.edu/cqb/>



# The Single cell Genomics Core (SCG)



- The SCG Core provides end-to-end single cell services



# The Data Analytics Core



## Mission statement

Facilitate advanced genomic & bioinformatic data analysis solutions to CQB faculty & the Dartmouth research community



**James O'Malley, PhD**

Director



**Shannon Soucy, PhD**

Senior Research Scientist



**Tim Sullivan, BA**

Bioinformatics Research Scientist



**Owen Wilkins, PhD**

Bioinformatics Research Scientist

### ➤ Genomic data analysis

- Analytical support for Dartmouth researchers
- Pipeline development & maintenance

### ➤ Bioinformatics consulting

### ➤ Publication & grant support

- Writing for methods & results sections
- Letters of support

### ➤ Training

- Group workshops

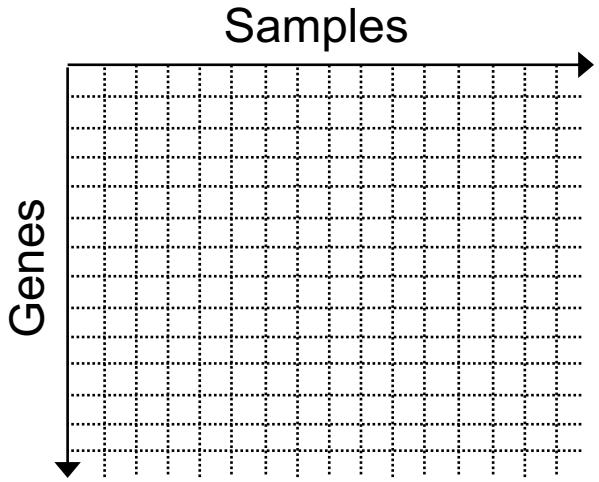


# Goals of the workshop

- Understand the basic principles of a differential expression analysis using RNA-seq data
- Develop a working understanding of the fundamental statistics behind a typical differential expression analysis using R/Bioconductor packages
- Perform a differential expression analysis using R/Bioconductor packages
- Learn how to explore the results and make robust insights from your data

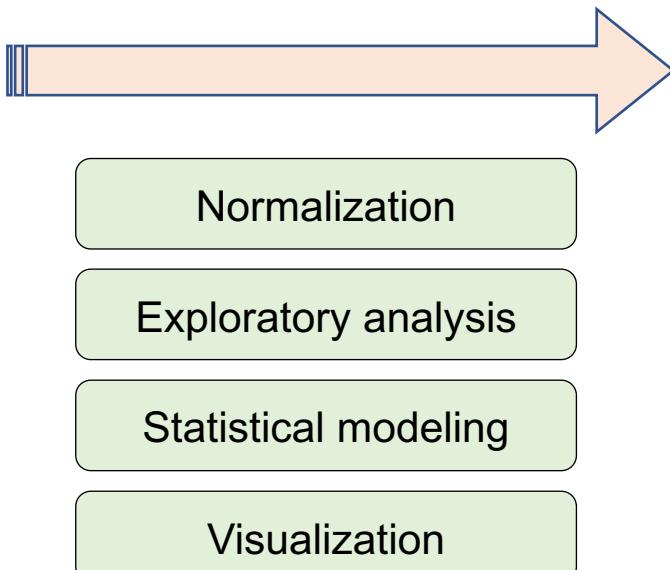
# Workshop outline

**Gene expression matrix**  
(read counts)



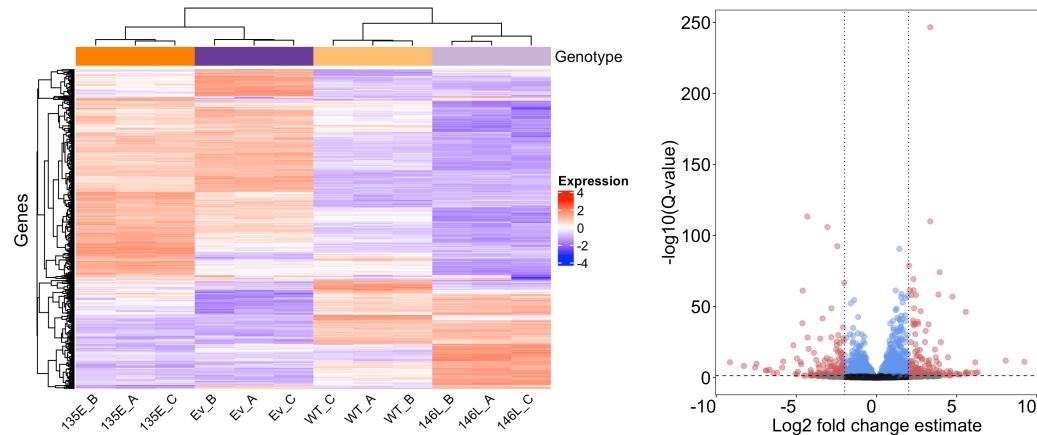
+

**Sample metadata**

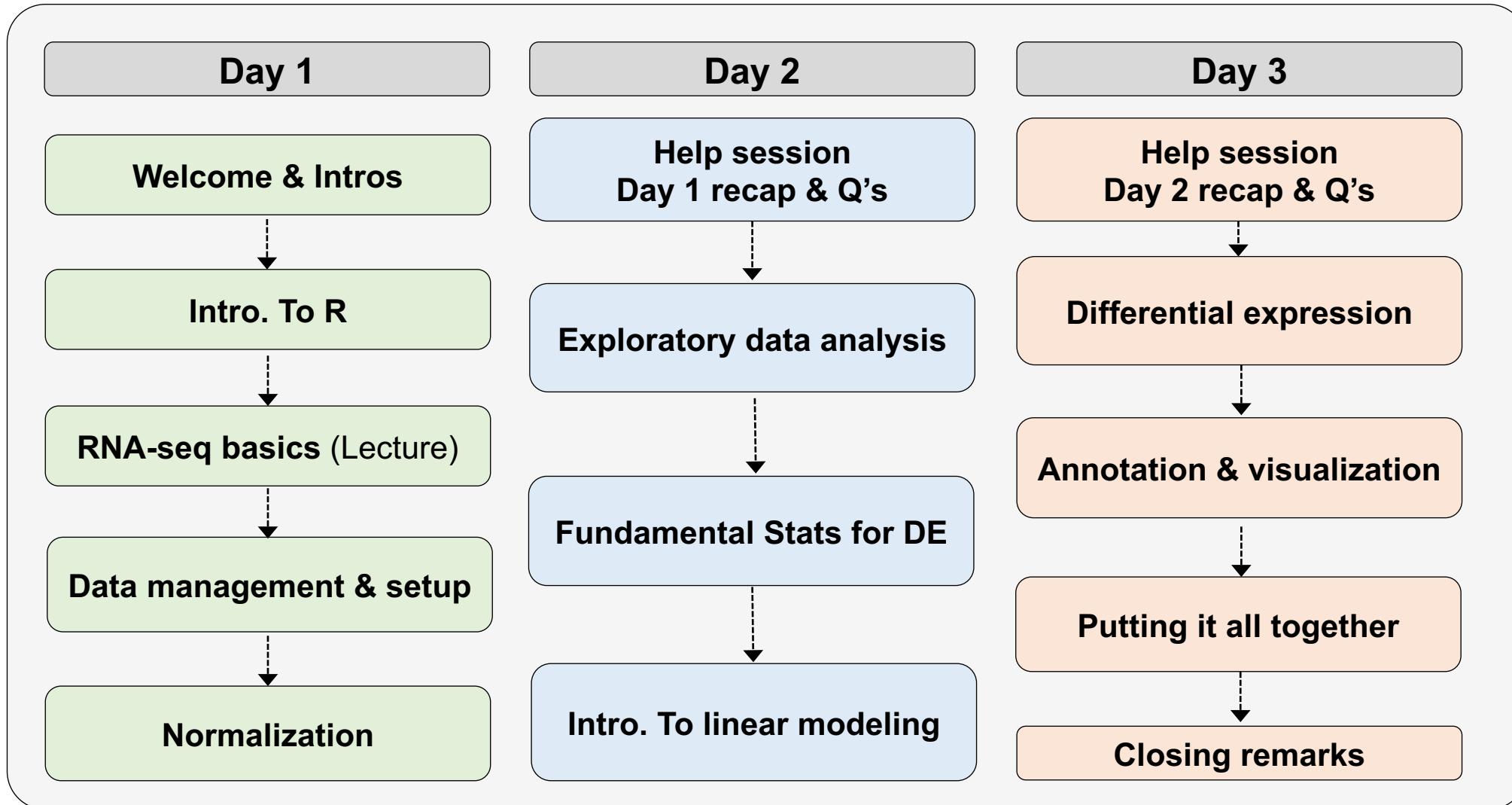


**Differential expression results**

ID	Ifc	IfcSE	stat	pvalue	padj	Gene
ENSG032219	2.89	0.12	-25.4	1.6E-251	1.7E-249	ARID4A
ENSG012951	-3.44	0.15	-24.5	1.2E-132	1.4E-122	KLK10
ENSG016754	-3.14	0.12	-24.1	2.9E-123	1.4E-118	KLK5
ENSG080097	-2.78	0.11	-20.7	5.7E-115	2.4E-109	LGALS1
ENSG006277	5.92	0.28	19.8	1.9E-80	3.5E-92	UCHL1
.....	.....	.....	.....	.....	.....	.....



# Workshop outline

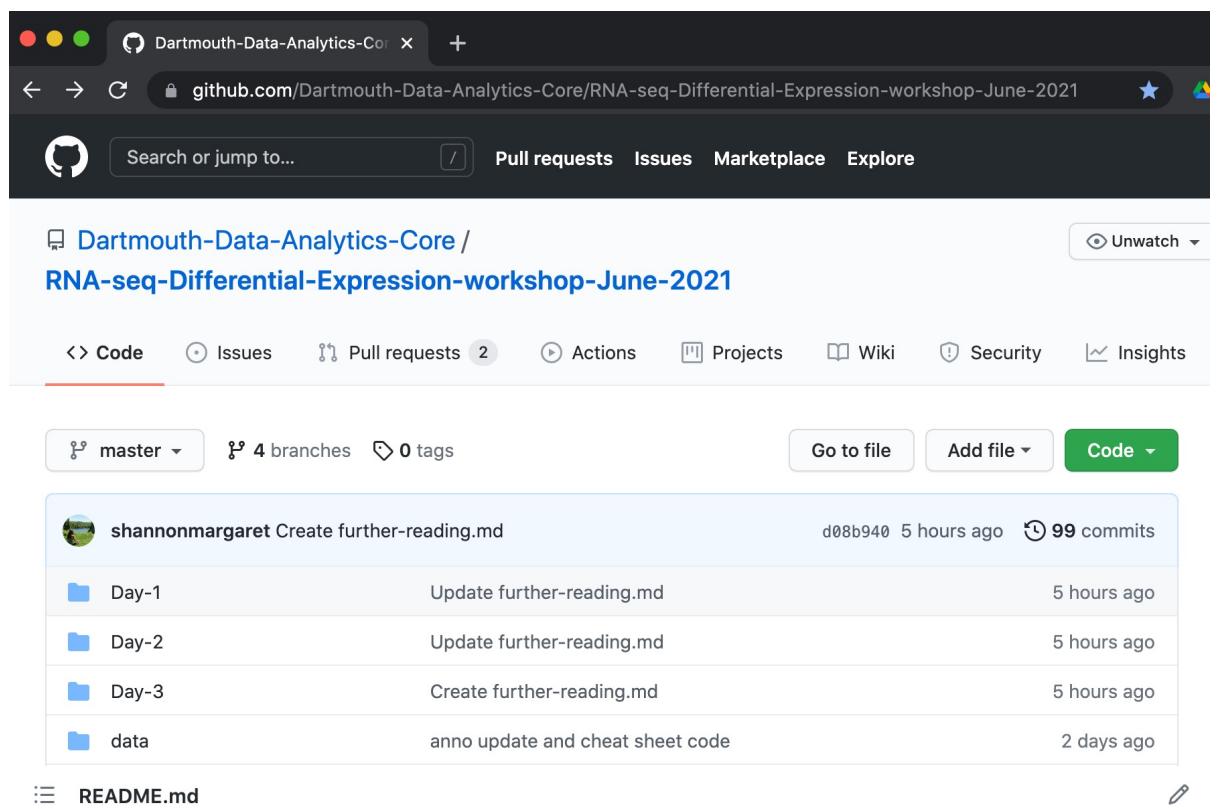


# Schedule

- Can be found at: <https://github.com/Dartmouth-Data-Analytics-Core/RNA-seq-Differential-Expression-workshop-June-2021/blob/master/schedule.md>
- 12pm-5pm each day
- Schedule is best guess, and we may deviate from it based on time
- If you will be absent for a session, just let us know

# Logistics |

- Course materials are all online (and will stay there):  
<https://github.com/Dartmouth-Data-Analytics-Core/RNA-seq-Differential-Expression-workshop-June-2021>
- You will download the materials on your local machine during the workshop
- We will be copying R code from markdowns (.md) into the RStudio



The screenshot shows a GitHub repository page. At the top, the URL is [github.com/Dartmouth-Data-Analytics-Core/RNA-seq-Differential-Expression-workshop-June-2021](https://github.com/Dartmouth-Data-Analytics-Core/RNA-seq-Differential-Expression-workshop-June-2021). The repository name is **Dartmouth-Data-Analytics-Core / RNA-seq-Differential-Expression-workshop-June-2021**. Below the repository name, there are tabs for Code (selected), Issues, Pull requests (2), Actions, Projects, Wiki, Security, and Insights. Under the Code tab, it shows the master branch (4 branches, 0 tags). A list of commits is shown, all made by user **shannonmargaret**:

- Create further-reading.md (d08b940, 5 hours ago, 99 commits)
- Update further-reading.md (5 hours ago)
- Update further-reading.md (5 hours ago)
- Create further-reading.md (5 hours ago)
- anno update and cheat sheet code (2 days ago)

## RNA-seq differential expression workshop, August 2021

This workshop will be delivered on August 23, 25, & 27 by the Data Analytics Core (DAC) of the [Center for Quantitative Biology at Dartmouth](#).

The DAC aims to facilitate advanced bioinformatic, computational, and statistical analysis of complex genomics data for the Dartmouth research community.

If you have questions about this workshop, or would like to discuss data analysis services available from the Data Analytics Core, please visit our [website](#), or email us at: [DataAnalyticsCore@groups.dartmouth.edu](mailto:DataAnalyticsCore@groups.dartmouth.edu)



# Logistics II

- Multiple tabs open:
  - Web browser
  - RStudio
- Copy & paste code from Markdowns to Rstudio
- If you finish: edit the code, try different options, generate scripts
- Use the **cheat sheets for R!**

The screenshot shows a Mac desktop with two windows open. On the left is a Chrome browser window displaying a GitHub page for RNA-seq-Differential-Expression analysis. The page contains R code for filtering results by adjusted p-value (FDR) and writing them to CSV files. It also includes a section on visualization, mentioning Volcano plots, MA plots, and Heatmaps (hierarchical clustering). On the right is an RStudio interface. The Environment pane shows various objects like colData\_sub, cts, dds, ha, ha1, ht1, mat\_scaled, mat1, p, and n2. The Global Environment pane lists formal classes and large matrices. A hierarchical clustering heatmap is displayed in the Plots pane, showing gene expression levels across samples grouped by condition (Dex vs. untreated).

RNA-seq-Differential-Expression

377 lines (279 sloc) 16.7 KB

```
# look at first few rows
head(res_ord)
```

We've now added a lot of useful information to our results that will help us interpret them in more detail. We may also wish to restrict the table to only those results that were statistically significant (at a threshold of 5%).

```
# subset @ 5% adjusted pval sig. level
res_order_FDR_05 <- res_ord[res_ord$padj<0.05,]
nrow(res_order_FDR_05)
```

Now write the table to a .csv file so that you can view it in other software (e.g. Excel) or share with others.

```
# write csv file for complete results
write.csv(as.data.frame(res_ord), file= "DE_results.csv")

# write csv for significant results
write.csv(as.data.frame(res_order_FDR_05), file="DE_results.FDR.0.05.csv")
```

**Part 2: Visualization of Differential Expression**

Several specific plot types exist that are useful for visualizing the results of a differential expression analysis, each providing insight on complimentary aspects of the results.

Below we will explore the major plot types useful for visualization of RNA-seq differential expression results, including:

- Volcano plots
- MA plots
- Heatmaps (hierarchical clustering)

**Volcano plot**

Volcano plots contrast the log<sub>2</sub> fold change (effect size) against the -log<sub>10</sub> P-value (statistical significance). The -log<sub>10</sub>(*P*) of a really small number is a very large value, therefore any gene that has a very small *P*-value will appear higher up along the y-axis. In contrast, the -log<sub>10</sub> of 1 is equal to 0, therefore genes with low statistical significance (*P*-values approaching 1) will appear lower down on the plot.

```
> draw(ht1, row_title = "Genes", column_title = "Hierarchical clustering of DEGs (padj<0.05)")
> |
```

RStudio

Environment History Connections

R Global Environment

colData\_sub Formal class DFrame
cts Large matrix (969952 elements, 8.2 MB)
dds Large DESeqDataSet (60622 elements, 50 M)
ha Formal class HeatmapAnnotation
ha1 Formal class HeatmapAnnotation
ht1 Large Heatmap ( 637 kB)
mat\_scaled num [1:1698, 1:8] 8.19 5.49 11.75 4.29 7...
mat1 num [1:1698, 1:8] 8.19 5.49 11.75 4.29 7...
p List of 9
n2 List of 9

Files Plots Packages Help Viewer

Hierarchical clustering of DEGs (padj<0.05)

Group

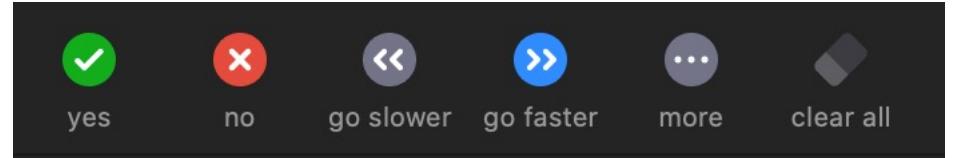
Expression Group

Genes

SRP1039521 SRP1039513 SRP1039517 SRP1039509 SRP1039520 SRP1039508 SRP1039512 SRP1039516

# Logistics III

- Use buttons in Participants tab in zoom



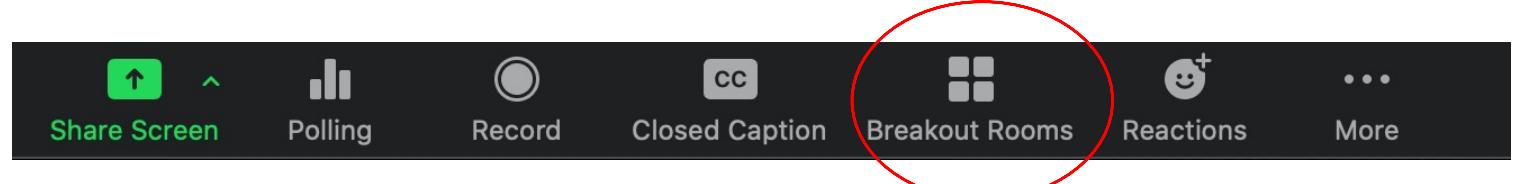
- You'll be muted, but if you want to ask a question, just raise your hand



Raise Hand

- We'll be using *breakout rooms (BRs)*

- We will use these when we split up to run code independently
- We've tried to pair everyone based on experience
- If you're stuck, message us, and we will come help you in your (BR)
- When we are going to move on, breakout rooms will close



- Please be courteous on zoom..

# How to get help?



Raise Hand

- **Raise your hand in zoom** (bottom right, participants tab)

- **Use the slack channel to message one of us**

- Use the *general* channel if it might benefit everyone
- Message us directly if its specific



- **If all else fails, email us:**

- DAC: [DataAnalyticsCore@groups.dartmouth.edu](mailto:DataAnalyticsCore@groups.dartmouth.edu)
- Shannon Soucy ([Shannon.Margaret.Soucy@Dartmouth.edu](mailto:Shannon.Margaret.Soucy@Dartmouth.edu))
- Tim Sullivan ([Timothy.J.Sullivan@dartmouth.edu](mailto:Timothy.J.Sullivan@dartmouth.edu))
- Owen Wilkins ([omw@Dartmouth.edu](mailto:omw@Dartmouth.edu))

## **Questions after the workshop?**

### **Bioinformatics office hours**

Friday 1-2pm (every other week, check calendar):  
<https://sites.dartmouth.edu/cqb/upcoming-events/calendar/>

Zoom link: <https://dartmouth.zoom.us/s/96998379866>

Passcode: bioinfo

- At the end, please give us feedback about this workshop, there will be a survey!
- And please ask lots of questions!

# Questions?

**...then.. Introductions!**

Name, department/program, research interests, why are you taking the workshop?