



# RNA-seq data 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/>)

---

07/08/20



**Dartmouth**  
GEISEL SCHOOL OF MEDICINE

# Who are we? 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



**Yue (Frank) Wang, PhD**  
Bioinformatics Research Scientist



**Carol Ringelberg**  
Data analyst



**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
- **Submission of raw & processed data to public databases**
- **Training**
  - One-on-one analysis
  - Group workshops

# Goals of the workshop



- Build an appreciation of key concepts for experimental design and generation of a typical RNA-seq dataset
- Develop understanding of pre-processing and QC steps for preparing a gene expression matrix for a typical RNA-seq dataset
- Gain familiarity working with RNA-seq data at the command-line (Bash)
- Learn the fundamentals of differential expression analysis and build foundational skills for performing differential expression in your own datasets (R/Bioconductor)

# Schedule

- **Can be found at:** [https://github.com/Dartmouth-Data-Analytics-Core/RNA-seq workshop July2020/blob/master/schedule.md](https://github.com/Dartmouth-Data-Analytics-Core/RNA-seq_workshop_July2020/blob/master/schedule.md)
- **9am-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\\_workshop\\_July2020](https://github.com/Dartmouth-Data-Analytics-Core/RNA-seq_workshop_July2020)

- You should have the repo downloaded locally

- Day 1 we will be copying Bash code from markdowns (.md) into the terminal

- Day 2 we will be running R markdowns (.Rmd)

The screenshot shows a GitHub repository page for 'Dartmouth-Data-Analytics-Core / RNA-seq\_workshop\_July2020'. The repository is private and has 164 commits, 4 branches, and 0 tags. The commits are listed in reverse chronological order, showing updates to various files like '05-quantification.md', 'experimental\_description.md', and 'README.md'. Below the commits is the contents of the 'README.md' file, which provides details about the RNA-seq data analysis workshop. At the bottom is the logo for the Center for Quantitative Biology (CQB).

Branch: master

owenwilkins committed b4a0d5c 1 hour ago

164 commits 4 branches 0 tags

File	Description	Time Ago
Day-1	Update 05-quantification.md	2 hours ago
Day-2	Added experimental description	yesterday
QC_reports	Add files via upload	23 hours ago
figures	Add files via upload	4 days ago
misc	Add files via upload	27 days ago
README.md	Update README.md	5 days ago
cheat-sheets.md	Update cheat-sheets.md	last month
schedule.md	Update schedule.md	last month
useful_links.md	Rename Useful_links.md to useful_links.md	last month
welcome-&-setup.md	Update welcome-&-setup.md	1 hour ago

README.md

**RNA-seq data analysis workshop, July 2020**

This workshop will be delivered on July 8th-9th 2020 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)

The logo for the Center for Quantitative Biology (CQB) features a stylized green DNA helix icon next to the letters 'CQB' in a bold, green, sans-serif font. Below the letters, the full name 'Center for Quantitative Biology' is written in a smaller, green, sans-serif font.

# Logistics II

## ➤ For day 1:

- Multiple tabs open
- Terminal window (ssh to discovery7)
- Web browser to GitHub repo
- If you finish: edit the code, try different options, generate scripts
- Use the ***cheat sheets!***

```
d41294d@discovery7:/dartfs-hpc/rc/lab/G/GMBSR_bioinfo
Last login: Tue Jun 30 15:30:30 on ttys002
-bash: /anaconda3/etc/profile.d/conda.sh: No such file or directory

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
[Owenl@vpn-two-factor-general-230-140-214 ~] $ sh ssh_to_dartfs_discovery.sh
d41294d@discovery7.dartmouth.edu's password:
Last login: Tue Jun 30 15:30:46 2020 from vpn-two-factor-general-230-140-214.dartmouth.edu

[cheduled maintenance is complete]

[Announcements]

Questions? Email Research.Computing@dartmouth.edu
===== Research Computing Notices =====
[05/22/20] Passwordless logging to Research Computing servers with PuTTY
[05/21/20] Quarterly reboot of Research Computing systems scheduled 06/16

Run 'notice' for more details of any announcements above.
[d41294d@discovery7 ~] $ cd bioin
[d41294d@discovery7 ~] $ ls
genomic_references Labs misc pipelines workshops
[d41294d@discovery7 ~] $
```

```
# count how many reads are NOT a primary alignment (FLAG=256)
samtools view -c -F 256 SRR1039508Aligned.out.sorted.REV.bam
```

### Viewing Alignments in IGV

The Integrative Genomics Viewer (IGV) from the Broad Institute is an extremely useful tool for visualization of alignment files (as well as other genomic file formats). Viewing your alignments in this way can be used to explore your data, troubleshoot issues you are having downstream of alignment, and inspect coverage for and quality of reads in specific regions of interest (e.g. in variant calling). I strongly encourage you to download the IGV for your computer from their [website](#) and play around with some BAM file to get familiar with all its various features.

Here, we will create a small subset of a BAM file, download it onto our local machines, and view it using the IGV web app (for speed). You can open the IGV web app in your browser [here](#).

Lets go ahead and subset our BAM file for reads aligning only to chromosome 22. We also need to create an index.

```
# subset for reads just on chr 22 (to make it smaller)
samtools view -b -@ 8 -o chr20.bam SRR1039508.1.Aligned.sortedByCoord.out.bam 20

# index your new bam file
samtools index chr20.bam
```

# Logistics III

## ➤ For day 2:

- Run Rmd files locally
- .Rmd files are in the repo you downloaded

The screenshot shows the RStudio interface with the following details:

- RStudio Environment:** Project: (None), Global Environment: Environment is empty.
- R Markdown Document:** RNAseq\_Workshop-day2-PART2.Rmd, containing R code and a figure.
- Figure:** A scatter plot titled "Dispersion estimation by shrinkage" from Figure 1 of the DESeq2 paper. The y-axis is "dispersion estimate" on a log scale from 0.001 to 100. The x-axis is "mean of normalized counts" on a log scale from 1 to 10000. The plot shows black dots representing individual genes, a red fitted curve labeled "prior mean", and blue vertical error bars representing MLE and MAP estimates. A legend indicates: MLE = Maximum-likelihood estimate, MAP = Maximum a posteriori estimate, and "Each black dot represent 1 gene". The plot is cited as Love et al., 2014, Genom. Biol.
- R Code:**

```
139 2. how far the initial dispersion is from the prior mean
140
141 !DESeq2 dispersion estimation
```

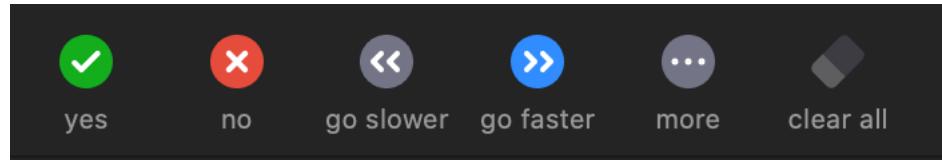
```
142
143 This Figure taken frm the DESeq2 paper demonstrates the process of shrinkage, where the initial dispersion esimates for each gene (estimated by maximum-likelihood) are shrunk towards the prior mean (based on the fitted curve in red) to a final MAP estimate. For dispersion estimates further away from the line, you can see that their estimates are shrunken more than those are originally closer to the line.
144
145 This curve also shows the expected trend for dispersion estimates over a range of expression levels. Importantly, the dispersion tends to decrease as the mean increases. Through inspecting disperion estimates for our own data, we can determine if the model is a good fit for our data, and therefore if it can be used to accurately test DE.
146
147 We can plot the dispersion estimates for our own data using:
148 ``{r}
149 plotDispEsts(dds)
150 ```
151
152 This is an example of a well calibrated set of dispersion estimates: the final MAP estimates are well scattered around the fitted line, and the dispersion trend decreases with increasing mean expression.
153
154 If there was a lot more structure in these plots, we would be concerned that the model is not estimating dispersions well for our data, indicating something may be wrong with the dataset, e.g. outlier samples, a batch effect, low quality samples/data, potential contamination etc.
155
156 *** To implement the confidence interval for dispersion estimation, one will need to add a confidence interval to the dispersion estimation function. See the vignette for more information.
```

```
248:32 [green] Chunk 15 [blue] R Markdown [green]
```

- R Documentation:** The right pane shows the documentation for the `grep` function. It includes the description, usage examples, and arguments.

# Logistics IV

- Use buttons in Participants tab in zoom



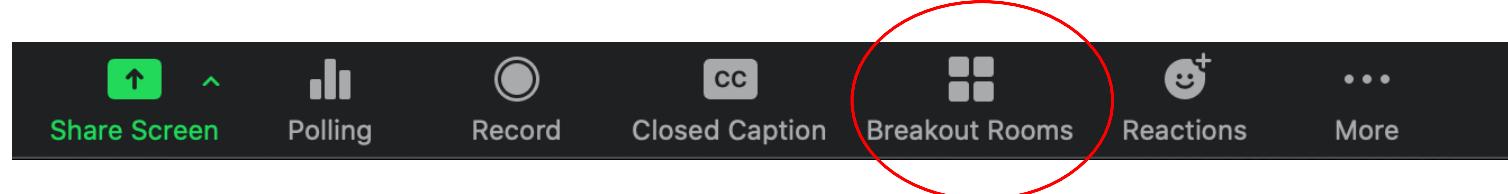
- You'll be muted, but if you want to ask a question, just raise your hand



Raise Hand

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

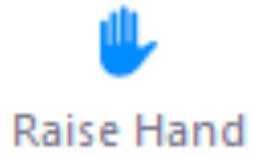
- We will use these when we split up to run code independently
- We've tried to pair everyone based on experience
- If you're stuck, message us, and we will come help you in your (BR)
- When we are going to move on, breakout rooms will close



- Please be courteous on zoom..

# How to get help?

- **Raise 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))
- Owen Wilkins ([omw@Dartmouth.edu](mailto:omw@Dartmouth.edu))
- Yue Wang ([yue.wang@Dartmouth.edu](mailto:yue.wang@Dartmouth.edu))

# Questions?

**Questions later?**

**Bioinformatics office hours** - Fridays 1-2pm. <https://dartmouth.zoom.us/s/96998379866>

Also, give us feedback about this workshop & other workshops you would like to see!

Bring questions for discussion at end of Day 2

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

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