



Fundamentals of Bioinformatics workshop

Shannon M. Soucy, PhD

Bioinformatics scientist

Genomic Data Science Core

Center for Quantitative Biology, Geisel School of Medicine at Dartmouth

Email: GDSC@groups.dartmouth.edu

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

March 2024



Dartmouth
GEISEL SCHOOL OF MEDICINE

Center for Quantitative Biology (CQB) at Dartmouth

- COBRE funded initiative to enhance NIH-funded quantitative research (PI: Michael Whitfield)
- Facilitates integration of quantitative & experimental biological research
- Supports multiple core facilities

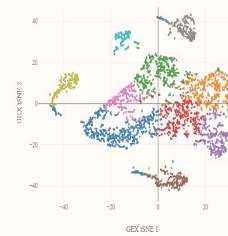


Single cell genomics Core

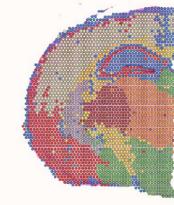
- Implement state of the art single cell workflows
- Maintain & develop existing technologies
- Acquire & operate novel instrumentation



10X Multiome
(scRNA- & scATAC-seq)



10X Visium
Spatial genomics

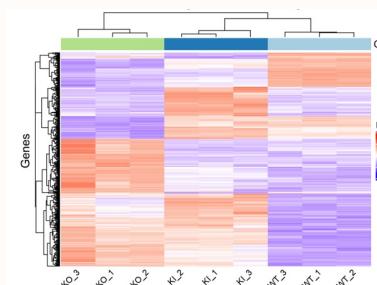
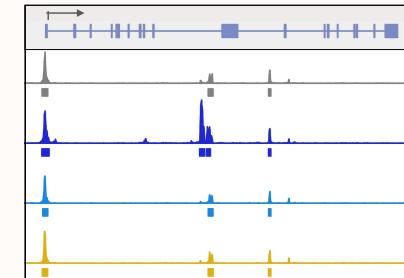


Director: Fred Kolling

Genomic Data Science Core

- Genomic and bioinformatic analysis services
- Publication & grant support
- Bioinformatics workshops

GRX2



The Genomic Data Science Core



Personnel:



Owen Wilkins, PhD
GDSC Co-director



Shannon Soucy, PhD
GDSC Co-director



Tim Sullivan, BA
Bioinformatics Research Scientist



Noelle Kosarek, PhD
Bioinformatics Research Scientist

What we do:

Genomic & bioinformatic data analysis to the CQB
& Dartmouth research community

Services:

Data Analysis

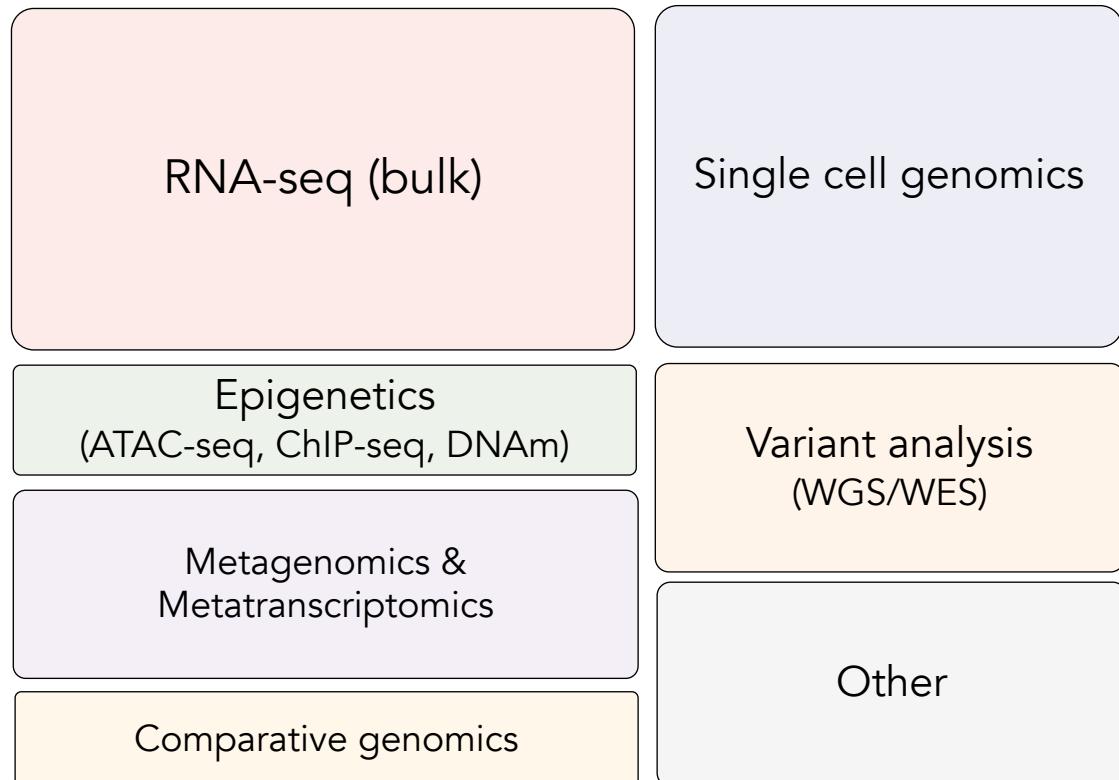
Publication
& grant
support

Bioinformatics training
(workshops)

Bioinformatics analysis services



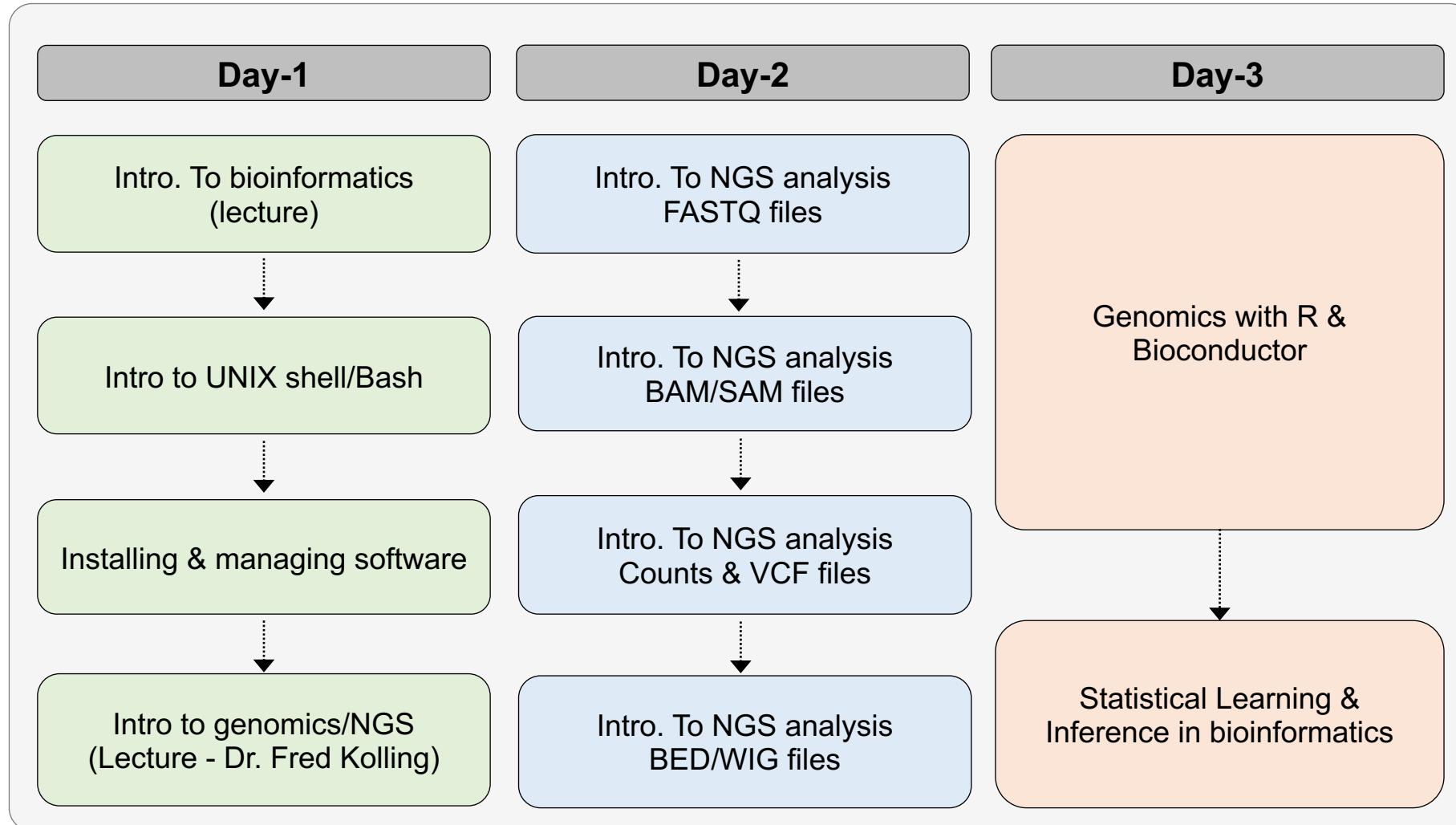
Main services:



Other services:

- Tunneling to leverage Rstudio on discovery
- Host reference data on DartFS for Human, Mouse, & Zebrafish
 - FASTA
 - Gene annotation (Ensembl & GENCODE)
 - Genome indices (STAR, Bowtie, HISAT, RSEM)
 - RefFlat & ribosomal intervals (Picard Tools)
- Custom workflow development
- One on One training (consultation hours)

Workshop outline



Schedule

- Can be found at: [https://github.com/Dartmouth-Data-Analytics-Core/Bioinformatics workshop-2024/blob/main/schedule.md](https://github.com/Dartmouth-Data-Analytics-Core/Bioinformatics_workshop-2024/blob/main/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/Bioinformatics_workshop-2024/
- You should have the repo downloaded locally
- Day 1 we will be copying Bash code from markdowns (.md) into the terminal, or into R Studio
- This is deliberate, and will help you learn how to organize code



Screenshot of a GitHub repository page for Dartmouth-Data-Analytics-Core / Bioinformatics_workshop (Private). The repository has 218 commits, 4 branches, and 0 tags. The README.md file contains the following content:

```
Fundamentals of Bioinformatics, December 2020

This workshop will be delivered on December 14, 15, & 17 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

Workshop goals:

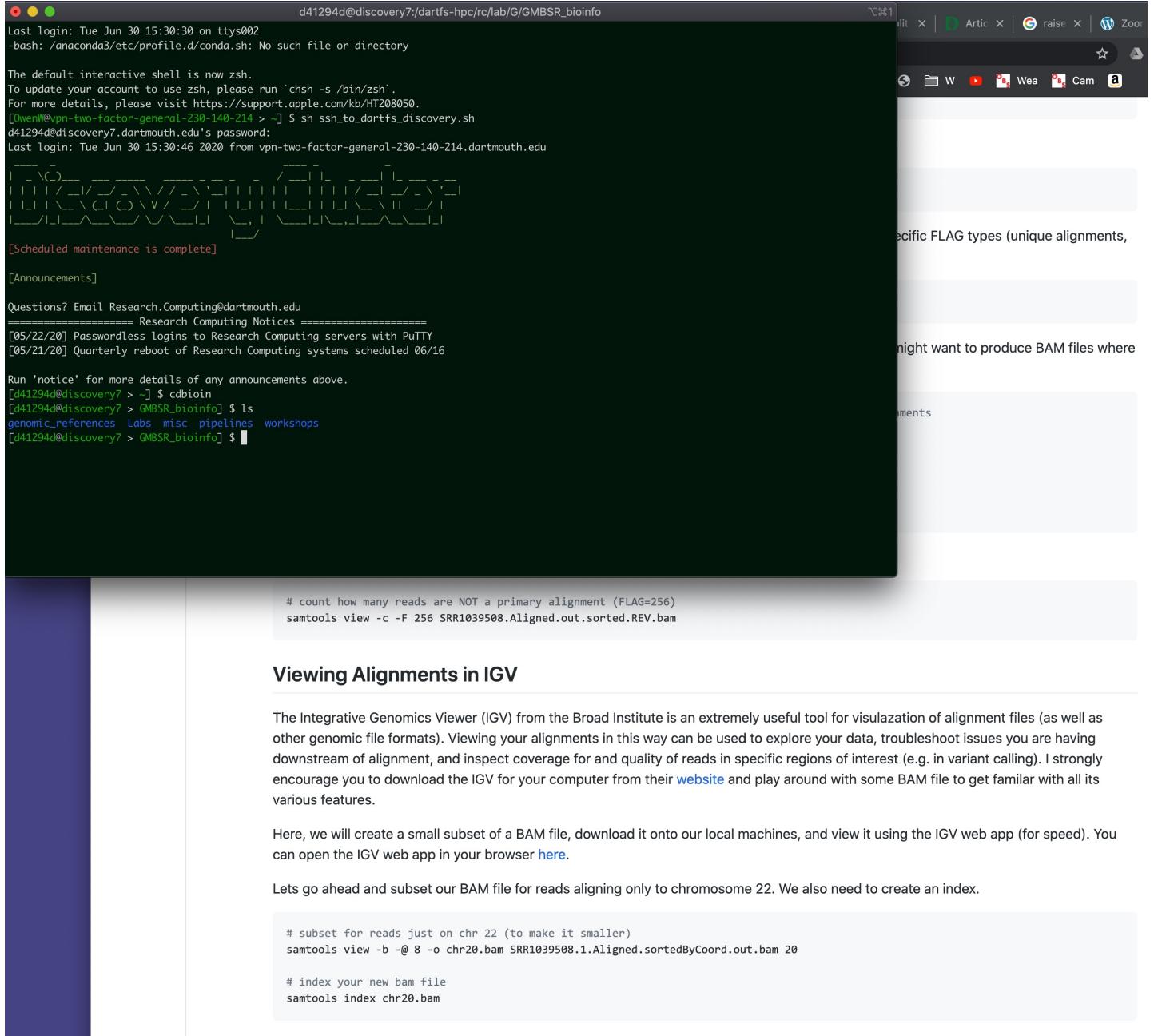
- Develop a working understanding what Bioinformatic data analysis involves, how it is done, and what skills it requires
- Gain an appreciation for how next-generation sequencing data is generated (NGS) and how the information generated is stored
- Learn the major file-types used in bioinformatic data analysis and how to manipulate them
- Learn how to install standard bioinformatic software using Conda
- Understand the concepts of reference genomes and genome annotations and where to find them
- Learn how to leverage the Integrative Genomics Viewer (IGV) for exploring genomics data
- Gain a working knowledge of basic programming in R and how it can be used for Bioinformatics

```

Logistics II

➤ When working in the terminal:

- Multiple windows 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.
[OwenW@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

[Scheduled 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 bioinfo
[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

➤ When working in R/Rstudio:

- Multiple windows open (including GitHub repo)

- Copy and paste R code into scripting window

- Save the scripts you generate as .R files

The screenshot shows a desktop environment with several windows open. In the center is a large RStudio window displaying an R script titled 'Untitled2.R'. The script contains code for reading ChIP-seq data from BED files and combining them into GrangesList objects. A red arrow points from the text 'TO DO: UPDATE THIS FIGURE TO' in the R script towards a 'Feature Distribution' plot in the bottom right corner.

The RStudio interface includes a top navigation bar with tabs like 'Environment', 'History', and 'Connections'. On the left, there's a 'Data' sidebar listing various objects: annolist, dat, dat2, dat3, df, df2, fr, fr_h3k27ac, fr_h3k27ac_anno, fr_h3k9ac, gene.df, ht, ht_h3k27ac, ht_h3k9ac, lm_2, lm1, and lm2. Below the Data sidebar is a 'Plots' tab containing the 'Feature Distribution' chart.

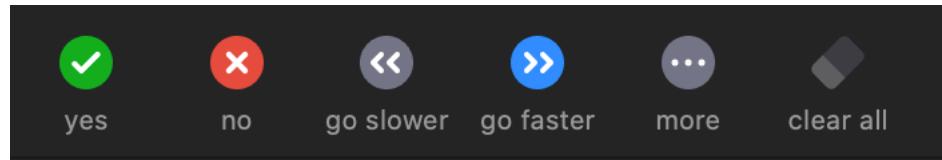
The 'Feature Distribution' chart is a stacked horizontal bar chart showing the percentage of genomic features for four samples: Forebrain_H3K27ac, Heart_H3K27ac, Forebrain_H3K9ac, and Heart_H3K9ac. The x-axis represents the percentage from 0 to 100. The y-axis lists the samples. The legend on the right identifies the genomic features by color: Promoter (<=1kb) in light blue, Promoter (1-2kb) in medium blue, 5' UTR in green, 3' UTR in orange, 1st Exon in red, Other Exon in yellow, 1st Intron in purple, Other Intron in light purple, Downstream (<=300) in light yellow, and Distal Intergenic in brown.

At the bottom of the RStudio window, there's a note: 'At this point in the analysis, we often want to perform various tasks on the reduced representation of the data, such as query overlapping peak regions between different sample groups, annotate regions based on their genomic or transcriptional context, or perform complex statistical analysis. The R statistical programming environment, and more specifically Bioconductor provides one avenue through which to perform these downstream analysis.'

Below the RStudio window, another note states: 'By far the largest advantage to using R to perform specific stages of genomic data analysis are the large number of packages available.'

Logistics IV

- Use buttons in Participants tab in zoom



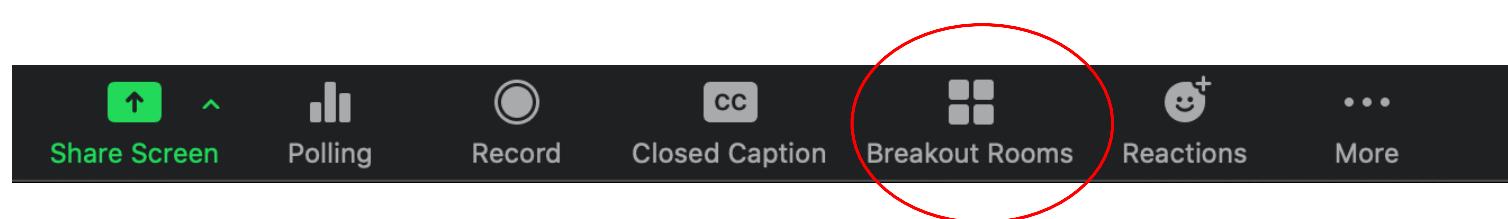
- You'll be muted, but if you want to ask a question, just raise your hand



Raise Hand

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

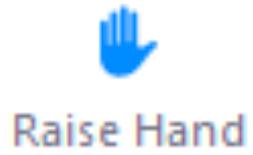
- We will use these to provide time for you to run code independently
- If you're stuck, message us on slack, and we will come help you in your (BR)
- When we are going to move on, breakout rooms will close



- 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 or the TAs**

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



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

- GDSC: GDSC@groups.dartmouth.edu
- Shannon Soucy (Shannon.Soucy@Dartmouth.edu)
- Tim Sullivan (Timothy.J.Sullivan@dartmouth.edu)
- Owen Wilkins (omw@Dartmouth.edu)
- Noelle Kosarek ([Noelle.N.Kosarek.GR@dartmouth.edu](mailto>Noelle.N.Kosarek.GR@dartmouth.edu))

Questions?

...then.. Introductions!

Use the slack channel #intros to tell us :

Your Name

Your Research interests

Why are you taking the workshop?