



# Fundamentals of Bioinformatics 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/>

---

December 2020



**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



**Tim Sullivan, BS**  
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

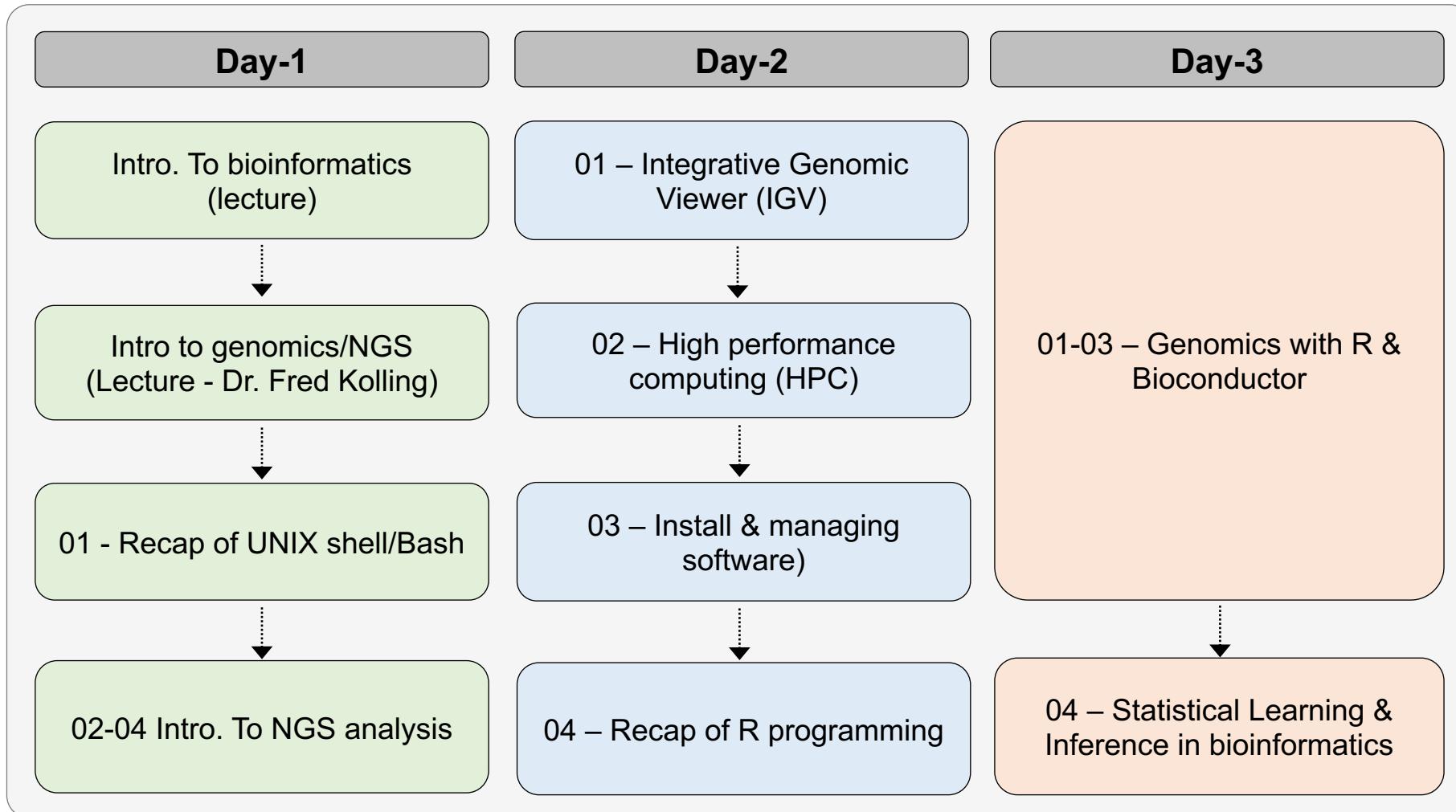
### ➤ Publication & grant support

- Writing for methods & results sections
- Letters of support

### ➤ Training

- One-on-one analysis
- Group workshops

# Workshop outline



# Schedule

- Can be found at: [https://github.com/Dartmouth-Data-Analytics-Core/Bioinformatics workshop/blob/master/schedule.md](https://github.com/Dartmouth-Data-Analytics-Core/Bioinformatics_workshop/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/Bioinformatics\\_workshop](https://github.com/Dartmouth-Data-Analytics-Core/Bioinformatics_workshop)
- 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 the GitHub repository page for Dartmouth-Data-Analytics-Core/Bioinformatics\_workshop.

The repository has 4 branches and 0 tags. The master branch contains 218 commits by owenwilkins, merged from the master branch of https://github.com/Dartmouth-Data-Analytics-Core/Bioinformatics\_workshop. The commits are listed below:

File	Description	Time Ago
Day-1	Update Day1b-basic_bash_scripting.md	2 hours ago
Day-2	Merge branch 'master' into shannonmargaret-patch-2	33 minutes ago
Day-3	u	1 hour ago
figures	u	5 hours ago
.DS_Store	igv update based on pull req.	34 minutes ago
README.md	Update README.md	23 minutes ago
cheat-sheets.md	u	1 hour ago
closing_remarks.md	update	10 minutes ago
schedule.md	Update schedule.md	22 minutes ago
welcome-&-setup.md	u	1 hour ago

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](mailto:DataAnalyticsCore@groups.dartmouth.edu)

**CQCB**  
Center for Quantitative Biology

**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 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:~$ cd bioinfo
d41294d@discovery7:~/bioinfo$ ls
genomic_references  Labs  misc  pipelines  workshops
d41294d@discovery7:~/bioinfo$
```

# 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 tabs open (including GitHub repo)

- Copy and paste R code into scripting window

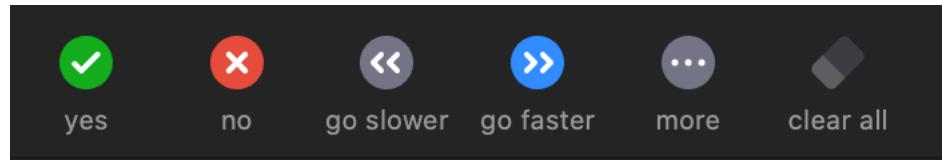
- Save the scripts you generate as .R files

The screenshot shows a desktop environment with several windows open:

- GitHub Repository:** A browser window showing the `Bioinformatics_workshop/01-genome-annotation` repository on GitHub. It displays 373 lines of code (265 sloc) and a size of 24.1 KB.
- RStudio Environment:** An RStudio interface with the following panes:
  - Code:** An R script titled "Untitled2.R" containing code for reading ChIP-seq data from BED files and creating GrangesList objects.
  - Data:** A list of objects in the global environment, including `annolist`, `dat`, `dat2`, `dat3`, `df`, `df2`, `fr`, `fr_h3k27ac`, `fr_h3k27ac_anno`, `fr_h3k9ac`, `gene.df`, `ht`, `ht_h3k27ac`, `ht_h3k9ac`, `lm_2`, and `lm1`.
  - Plots:** A "Feature Distribution" plot showing the percentage of genomic features for various categories across four samples: Forebrain\_H3K27ac, Heart\_H3K27ac, Forebrain\_H3K9ac, and Heart\_H3K9ac. The categories are color-coded: Promoter (<=1kb) in light blue, Promoter (1-2kb) in medium blue, 5' UTR in green, 3' UTR in purple, 1st Exon in red, Other Exon in orange, 1st Intron in yellow, Other Intron in light purple, Downstream (<=300) in light green, and Distal Intergenic in brown.
- Bioinformatics Workshop Guide:** A PDF titled "Part 1 - Working with NGS Data" which includes a "TO DO: UPDATE THIS FIGURE TO SHOW THE NEW ANALYSIS" section.
- Console:** The R console showing the command history for the analysis.

# Logistics IV

- Use buttons in Participants tab in zoom



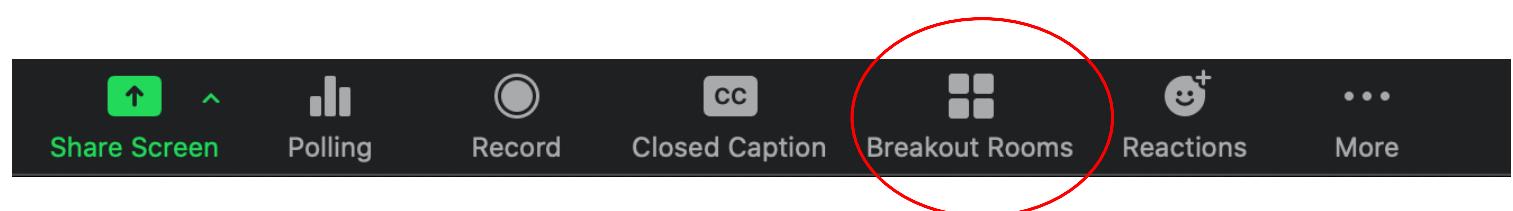
- You'll be muted, but if you want to ask a question, just raise your hand



Raise Hand

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

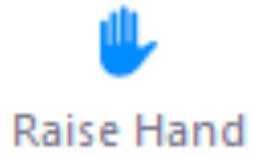
- We will use these when we split up to run code independently
- We've tried to pair everyone based on experience
- If you're stuck, message us, and we will come help you in your (BR)
- When we are going to move on, breakout rooms will close



- 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 or the TAs**

- 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?

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

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