



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

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

December 2021



Dartmouth
GEISEL SCHOOL OF MEDICINE

The Data Analytics Core



Personnel:



James O'Malley, PhD
Faculty Director



Owen Wilkins, PhD
Bioinformatics Research Scientist



Shannon Soucy, PhD
Senior Research Scientist



Tim Sullivan, BA
Bioinformatics Research Scientist

What we do:

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

Services:

Data Analysis

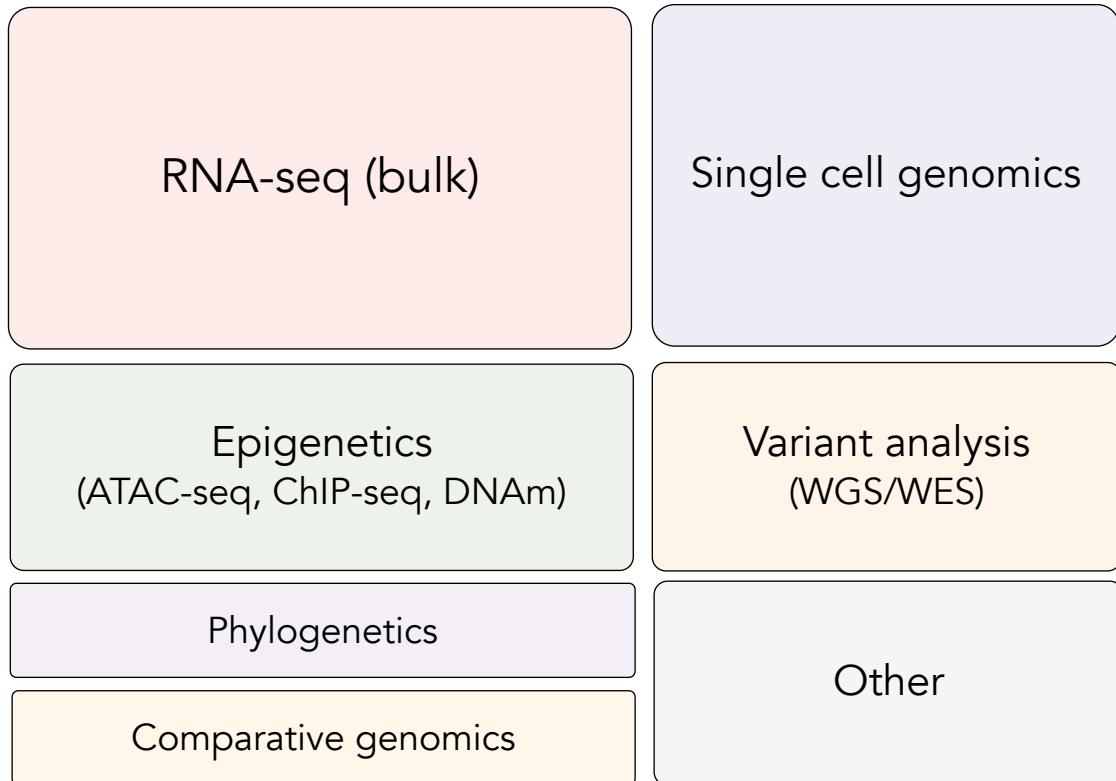
Publication
& grant
support

Bioinformatics training
(workshops)

Bioinformatics analysis services



Main services:



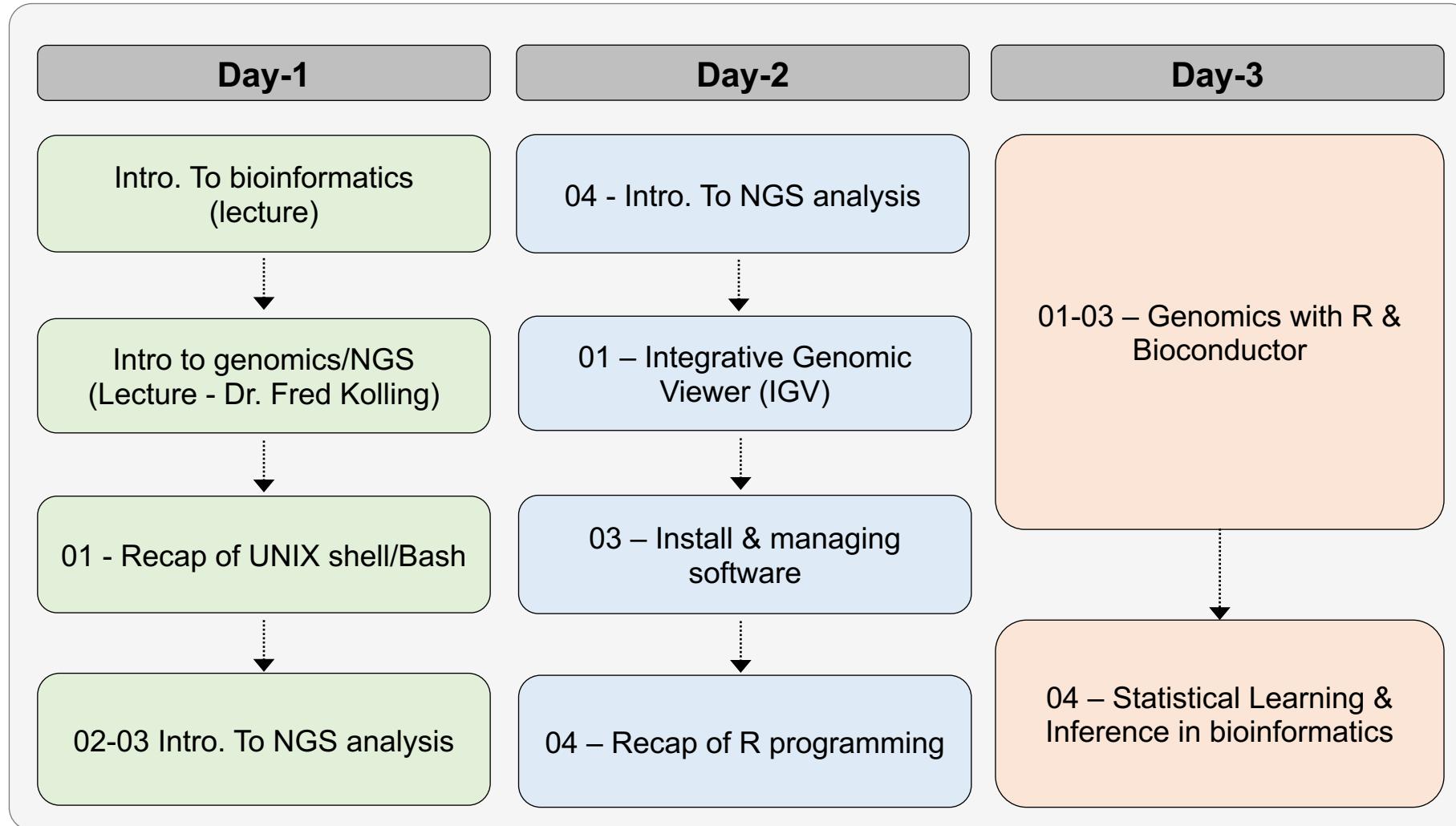
Core users:

Department	% of users
Molecular & Systems Biology	40%
Microbiology & Immunology	27%
Biochemistry & Cell Biology	7%
Epidemiology	6%
Surgery/Medicine	7%
Neurology	3%
External	10%

- Many labs closely associated with Norris Cotton Cancer Center



Workshop outline



Schedule

- Can be found at: [https://github.com/Dartmouth-Data-Analytics-Core/Bioinformatics workshop-Dec-2021/blob/master/schedule.md](https://github.com/Dartmouth-Data-Analytics-Core/Bioinformatics_workshop-Dec-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/Bioinformatics_workshop-Dec-2021
- 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
```

The page also features the CQB logo (Center for Quantitative Biology) and a list of workshop goals.

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!***

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>.
[OwenII@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 ~] \$ cdbioin
[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 tabs open (including GitHub repo)

- Copy and paste R code into scripting window

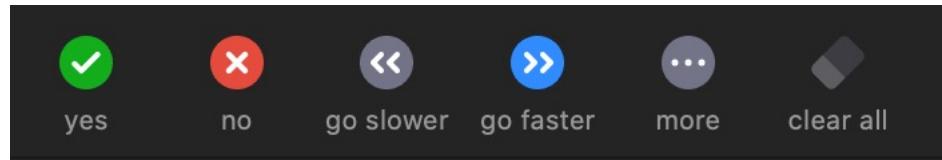
- Save the scripts you generate as .R files

The screenshot shows a desktop environment with several windows open, illustrating a workflow for genomic data analysis:

- GitHub Repository Window:** Shows 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:** The main workspace shows an R script titled "Untitled2.R" with code for reading ChIP-seq data from BED files. The code includes imports for `rtracklayer`, `forebrain`, and `granges` packages. It defines variables like `fr_h3k27ac` and `ht_h3k27ac` for heart and forebrain H3K27ac peaks respectively.
- RStudio Console:** Displays the command-line history of the R session, showing the execution of the script and its output.
- Feature Distribution Plot:** A stacked bar chart titled "Feature Distribution" showing the percentage of genomic features across different categories for four samples: Forebrain_H3K27ac, Heart_H3K27ac, Forebrain_H3K9ac, and Heart_H3K9ac. The categories include Promoter (<=1kb), Promoter (1-2kb), 5' UTR, 3' UTR, 1st Exon, Other Exon, 1st Intron, Other Intron, Downstream (<=300), and Distal Intergenic.
- Diagram of NGS Data Analysis Pipeline:** A flowchart titled "Typical NGS data analysis pipeline" showing the process from "Aligned reads (.bam)" through "Genome annotation (.GTF)" and "Reference genome (.fa)" to "Research questions?".
- Textual Notes:** A red dashed arrow points from the GitHub repository window to the RStudio console, indicating the flow of data from the repository into the R session. A note "TO DO: UPDATE THIS FIGURE TO THE NEW ONE" is present in the RStudio console area.
- Footnote:** At the bottom of the RStudio console, there is a note about the use of Bioconductor for downstream analysis.

Logistics IV

- Use buttons in Participants tab in zoom



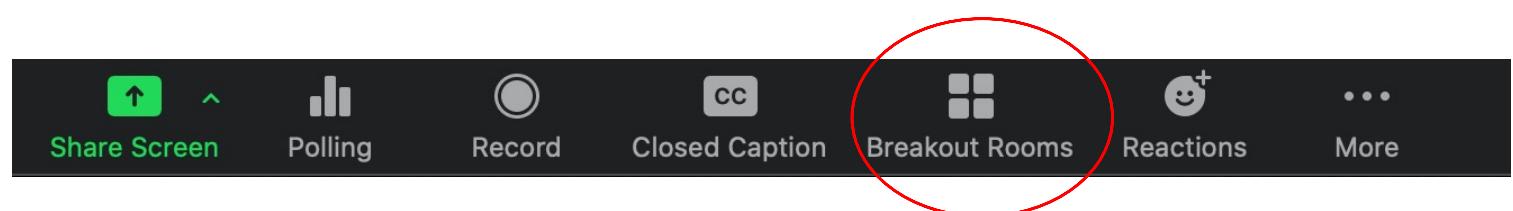
- You'll be muted, but if you want to ask a question, just raise your hand



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
- Shannon Soucy (Shannon.Margaret.Soucy@Dartmouth.edu)
- Tim Sullivan (Timothy.J.Sullivan@dartmouth.edu)
- Owen Wilkins (omw@Dartmouth.edu)

Questions?

...then.. Introductions!

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