CFAR *Center For AIDS Research*

DUKE UNIVERSITY & MEDICAL CENTER

CFAR Core E ~~Chalk~~Slide Talk

Minimum requirements for DIY *collaborative* data science for HIV/AIDS data analysis

Cliburn Chan

CFAR Core E ~~Chalk~~Slide Talk

Why should a biologist learn data science?

- You want to put me out of my job
- You suspect we are ignoring your emails
- You like mucking with computers
- You are drowning in data
- You suspect there must be a better way than 1 billion row Excel documents

Requirements

- Basic statistics
- Basic programming skills in R
- Basic familiarity with use of BioConductor
- Practice and feedback
- Reminder: We are *not* your competition - collaborate with us!

1. Basic statistics

- Modern software makes it possible to *use statistics* as a black box of magic tricks. If you are going to summon demons, at least know what you're doing.
- *Concepts* are more important than mathematical sophistication. You do not need to know how to derive mathematical results, only how to properly use them.
- Learn to *think* like a statistician - this consists of knowing the appropriate questions to ask to address your research problem.

2. Basic programming skills in R

- Learn to *think* like a computer scientist - this consists of breaking big problems into smaller ones until you know how to use a *recipe* to solve it
- Learn how to manipulate data using modern R.
- Learn how to generate visualizations and plots
- Learn how to write *literate* programs for reproducible analysis in RStudio.

3. Basic familiarity with use of BioConductor

- BioConducotor contains hundreds of packages for working with biological data sets
- Learn to use specialized packages for your own analysis needs - e.g. gene chips, sequencing, proteomics, flow cytometry
- Builds on your basic knowledge of R

4. Practice and feedback

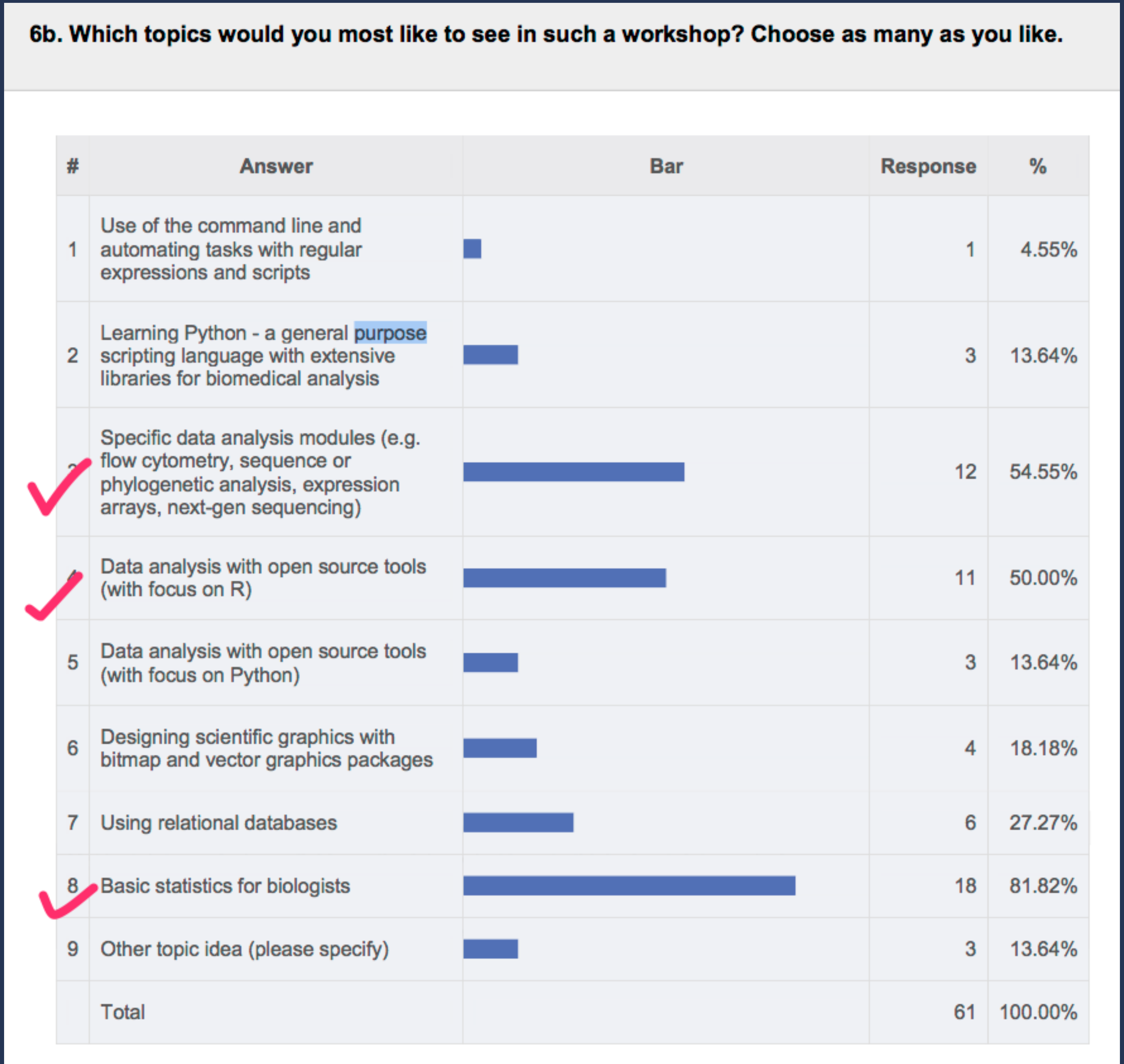
- Find a partner (or two) from your lab to practice together
- Practice regularly
- Start a data science users group with me!

A. (Nice to have) Basic familiarity with command line

- This is always nice to have since skill with the command line and Unix Shell can greatly increase your productivity
- Some forms of genomic analysis require working with the Unix Shell- If you will be working remotely (e.g. using the Duke computer cluster or Amazon cloud)

B. (Nice to have) Basic familiarity with SQL and relational databases

- Most large and well-annotated data sets live in relational databases that can be queried efficiently with SQL



6b. Which topics would you most like to see in such a workshop? Choose as many as you like.

#	Answer	Bar	Response	%
1	Use of the command line and automating tasks with regular expressions and scripts		1	4.55%
2	Learning Python - a general purpose scripting language with extensive libraries for biomedical analysis		3	13.64%
✓ 3	Specific data analysis modules (e.g. flow cytometry, sequence or phylogenetic analysis, expression arrays, next-gen sequencing)		12	54.55%
✓ 4	Data analysis with open source tools (with focus on R)		11	50.00%
5	Data analysis with open source tools (with focus on Python)		3	13.64%
6	Designing scientific graphics with bitmap and vector graphics packages		4	18.18%
7	Using relational databases		6	27.27%
✓ 8	Basic statistics for biologists		18	81.82%
9	Other topic idea (please specify)		3	13.64%
	Total		61	100.00%

Coming Opportunities Offered by Core E

We are finalizing the planning for an extended series of educational opportunities for CFAR investigators. Would love feedback on the stuff we are planning.

Basic Statistics for Biologists

- Series of 14 lectures planned
- Based on statistics refresher course for clinicians
- Supplemented with R code so you don't just learn the concepts but can also do the computations
- Dates?

Data analysis with R

- Two-day workshop planned for August 14 and 15 – (9am-5pm) in the CRTP classroom (2nd floor Hock Plaza, 2424 Erwin Road).
- Day 1 will introduce the use of RStudio and cover basic usage of R and the `tidyverse` package for data manipulation - loading, cleaning, filtering, transforming, sorting, summarization etc
- Day 2 will cover using the `ggplot2` package to create publication-quality plots in R
- Makes use of HIV/AIDS data sets
    - UNC HIV clinic demographic data (generously shared by UNC CFAR Biostatistics Core)
    - Peptide array data analysis and visualization (generously shared by Tomaras lab)

Using BioConductor for Assay Data Analysis

- Series of hands-on practice sessions with mini-lectures
- Session 1: Introduction to BioConductor
- Session 2: Pipeline for analysis of gene chip data
- Session 3: Pipeline for analysis of sequencing data
- Session 4: Pipeline for analysis of flow cytometry data
- Requests?

HTS Course (intensive)

- 6 week full-time (4 days per week) summer course to be held in 2017, 2018 and 2019
- Covers statistics, computing, bioinformatics and hands-on lab practice preparing material for RNA-Seq
- Sponsored by NIH BD2K program
- Limited to 24 participants - FREE
- 2017 course is FULL but consider applying next year if interested
- Course website

User Group

If there is interest, I will consider setting up regular user group session to provide mentoring and sharing of knowledge. Venue, frequency etc to be determined.