

# Welcome to Quantitative Biology

---



QB Bootcamp, Day 1  
Tuesday, 5 September 2023  
10:00am - 10:30am

# 2023 QB Bootcamp Syllabus

---

## Instructors:

- Justin Kinney, [jkinney@cshl.edu](mailto:jkinney@cshl.edu)
- Ivan Iossifov, [iossifov@cshl.edu](mailto:iossifov@cshl.edu)

## Teaching assistants:

- Chandana Rajesh, [rajesh@cshl.edu](mailto:rajesh@cshl.edu)
- Michael Passalacqua, [passala@cshl.edu](mailto:passala@cshl.edu)

**GitHub Repository:** [https://github.com/jbkinney/23e\\_qbbootcamp](https://github.com/jbkinney/23e_qbbootcamp)

**Day 1:** Tuesday, 5 September 2023, 10am - 6pm, Hershey East Room, Hershey Bldg.

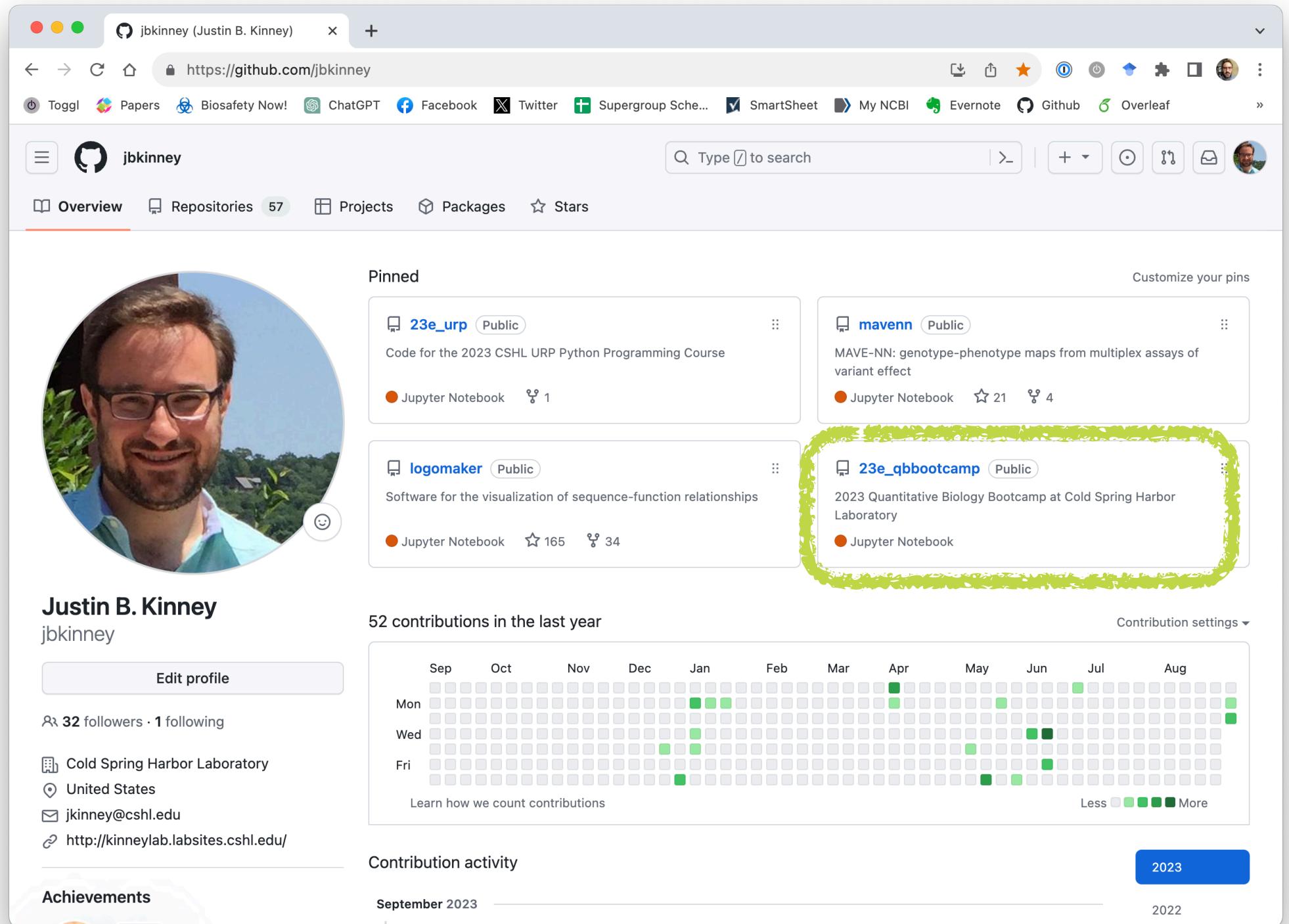
- 10:00am - 10:30am: **Overview of Quantitative Biology (lecture, Justin)**
- 10:30am - 12:00pm: **The Unix command line (tutorial, Ivan)**
- 12:00pm - 1:00pm: *Lunch (provided)*
- 1:00pm - 1:30pm: **Introduction to Python and Jupyter Notebooks (tutorial, Justin)**
- 1:30pm - 3:30pm: **Python: data types (tutorial, Justin)**
- 3:30pm - 4:00pm: *Break*
- 4:00pm - 6:00pm: **Python: flow control (tutorial, Ivan)**

**Day 2:** Wednesday, 6 September 2023, 1pm - 6pm, Samet Conference Room, Koch Bldg.

- 1:00pm - 1:30pm: **Overview of high-performance computing (lecture, Justin)**
- 1:30am - 3:00pm: **Read mapping using Elzar (tutorial, Justin)**
- 3:00pm - 3:30pm: *Break*
- 3:30pm - 3:00pm: **Introduction to Pandas (lecture, Justin)**
- 4:00pm - 5:30pm: **Pandas I, TF analysis (tutorial, Justin)**

**Day 3:** Thursday, 7 September 2023, 10am - 6pm, Plimpton Conference Room, Beckman Bldg.

- 10:00am - 11:30am: **Pandas II, Replication origin analysis (tutorial, Ivan)**
- 11:30pm - 12:00pm: **Introduction to Data Visualization (lecture, Justin)**
- 12:00pm - 1:00pm: *Lunch (provided)*
- 1:00pm - 2:30pm: **Matplotlib (tutorial, Ivan)**
- 2:30pm - 3:00pm: *Break*
- 3:00pm - 4:30pm: **Advanced visualization (tutorial, Justin)**



jbkinney/23e\_qbbootcamp: 20 · +

[https://github.com/jbkinney/23e\\_qbbootcamp](https://github.com/jbkinney/23e_qbbootcamp)

Toggl Papers Biosafety Now! ChatGPT Facebook Twitter Supergroup Sche... SmartSheet My NCBI Evernote Github Overleaf

jbkinney / 23e\_qbbootcamp Type / to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

23e\_qbbootcamp Public Unpin Unwatch 1 Fork 0 Star 0

main 1 branch 0 tags

jbkinney Merge branch 'main' of https://github.com/jbkinney/23e\_qbbootcamp

bash Initial commit

python Initial commit

.gitignore added syllabus and elzar\_exercise.tar.gz

23e\_qbbootcamp\_syllabus.pdf added syllabus and elzar\_exercise.tar.gz

LICENSE Initial commit

README.md Update README.md

elzar\_exercise.tar.gz added syllabus and elzar\_exercise.tar.gz

README.md

**2023 Quantitative Biology Bootcamp**

Welcome to the 2023 QB Bootcamp of the School for Biological Sciences at Cold Spring Harbor Laboratory! This Github repository contains the Jupyter notebooks, shell scripts, and datasets that we will work through in this bootcamp.

Go to file Add file ▾ < > Code ▾

Local Codespaces

Clone

HTTPS SSH GitHub CLI

[https://github.com/jbkinney/23e\\_qbbootcamp](https://github.com/jbkinney/23e_qbbootcamp)

Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

About

2023 Quantitative Biology Bootcamp at Cold Spring Harbor Laboratory

Readme MIT license Activity 0 stars 1 watching 0 forks

Releases

No releases published Create a new release

Packages

No packages published Publish your first package

Languages

jbkinney / 23e\_qbbootcamp

Type / to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

23e\_qbbootcamp Public

Unpin Unwatch 1 Fork 0 Star 0

main 1 branch 0 tags

jbkinney Merge branch 'main' of https://github.com/jbkinney/23e\_qbbootcamp

bash Initial commit

python Initial commit

.gitignore added syllabus and elzar\_exercise.tar.gz

23e\_qbbootcamp\_syllabus.pdf added syllabus and elzar\_exercise.tar.gz

LICENSE Initial commit

README.md Update README.md

elzar\_exercise.tar.gz added syllabus and elzar\_exercise.tar.gz

README.md

**2023 Quantitative Biology Bootcamp**

Welcome to the 2023 QB Bootcamp of the School for Biological Sciences at Cold Spring Harbor Laboratory! This Github repository contains the Jupyter notebooks, shell scripts, and datasets that we will work through in this bootcamp.

Go to file Add file ▾ < > Code ▾

Local Codespaces

Clone

HTTPS SSH GitHub CLI

[https://github.com/jbkinney/23e\\_qbbootcamp](https://github.com/jbkinney/23e_qbbootcamp)

Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

About

2023 Quantitative Biology Bootcamp at Cold Spring Harbor Laboratory

Readme MIT license Activity 0 stars 1 watching 0 forks

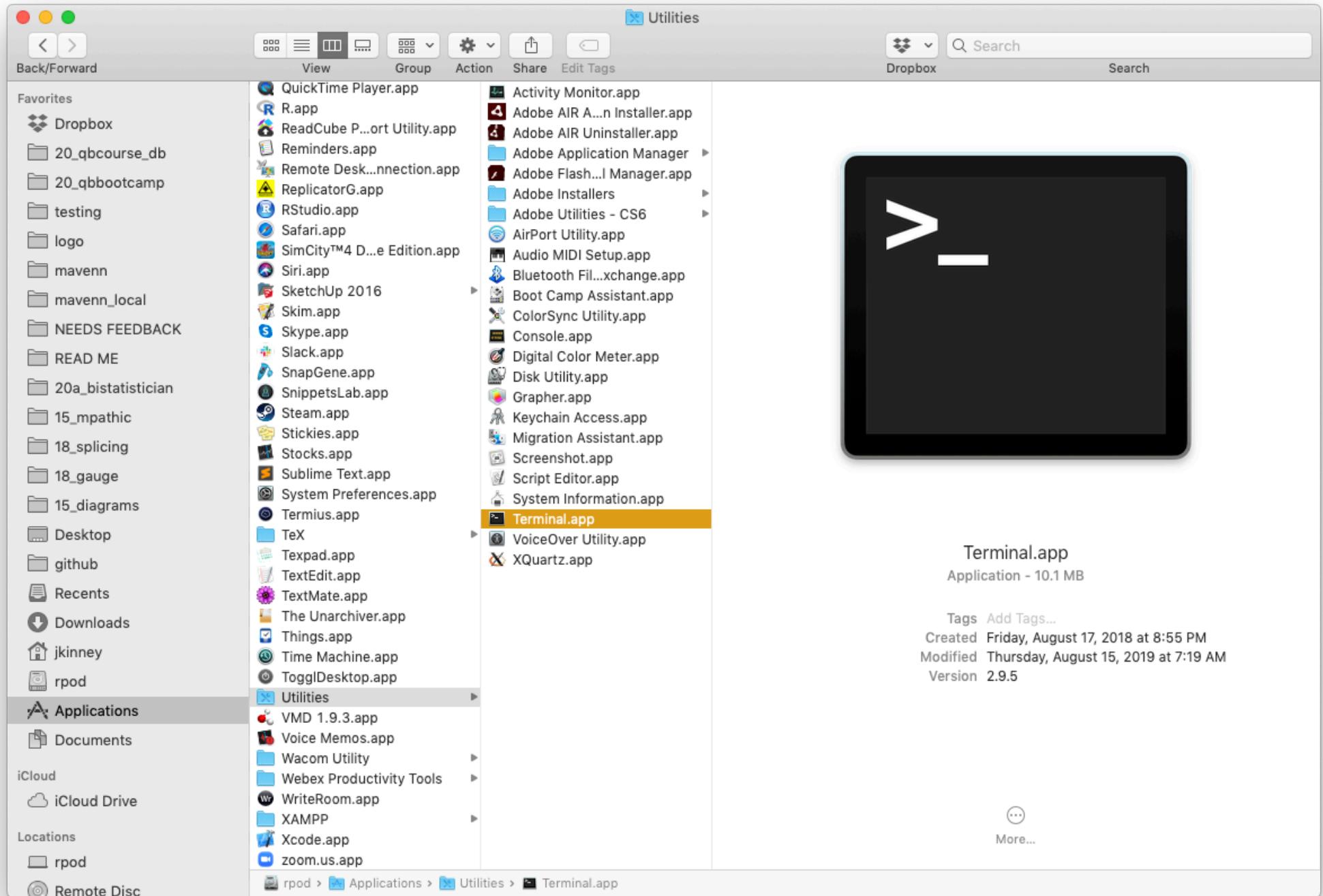
Releases

No releases published Create a new release

Packages

No packages published Publish your first package

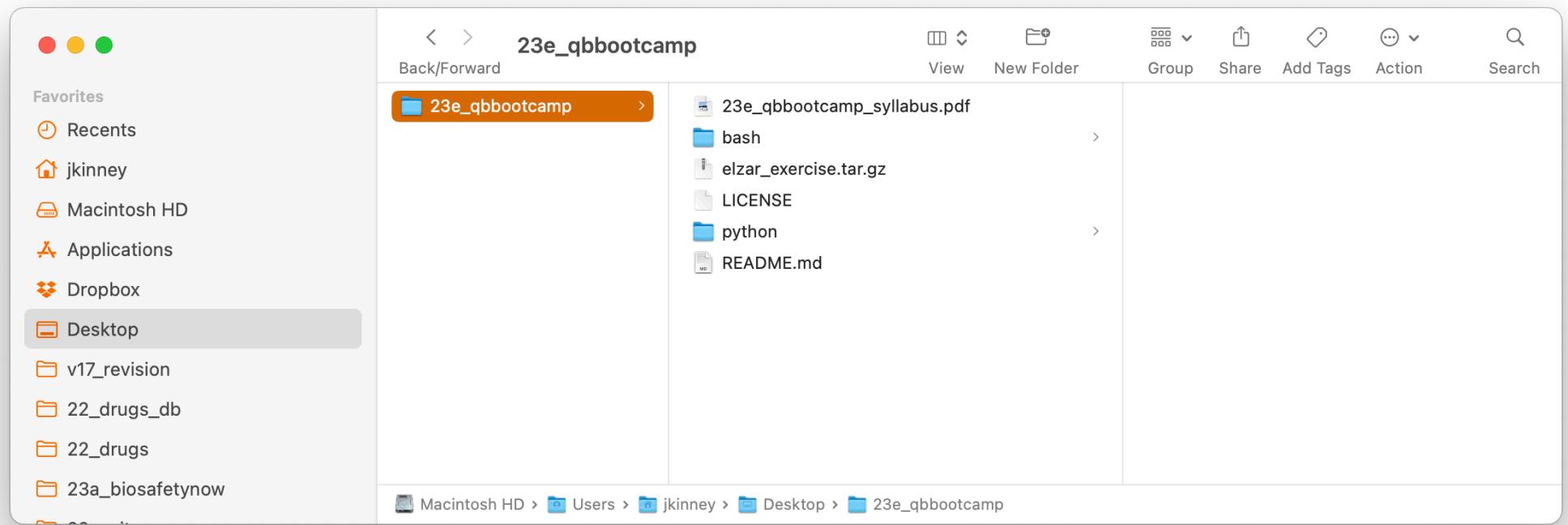
Languages



Desktop — bash — 104x17

```
[base) jkinney@dobbs:~/Desktop$ git clone https://github.com/jbkinney/23e_qbbootcamp.git
Cloning into '23e_qbbootcamp'...
remote: Enumerating objects: 34, done.
remote: Counting objects: 100% (10/10), done.
remote: Compressing objects: 100% (9/9), done.
remote: Total 34 (delta 1), reused 8 (delta 1), pack-reused 24
Receiving objects: 100% (34/34), 79.19 MiB | 6.34 MiB/s, done.
Resolving deltas: 100% (7/7), done.
(base) jkinney@dobbs:~/Desktop$
```



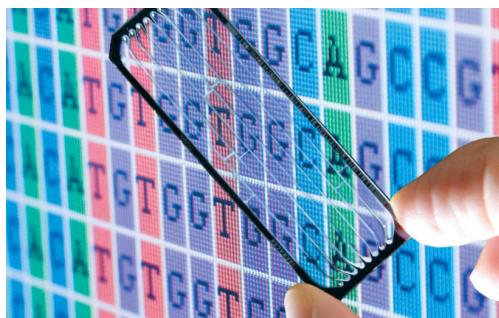


## **What is Quantitative Biology?**

# Quantitative biology is a vast field

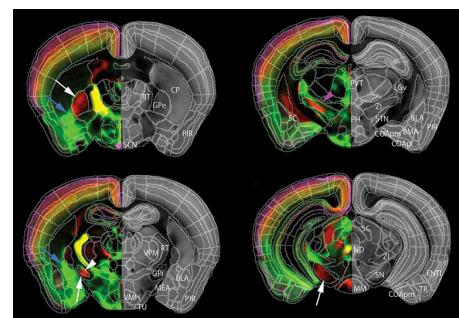
---

## Genomics



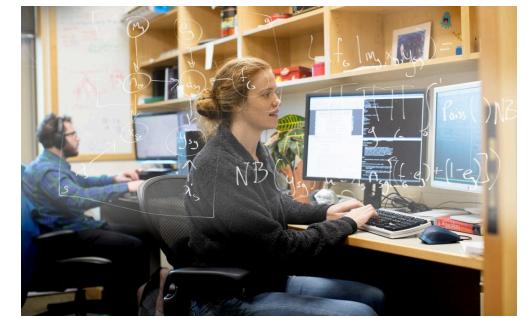
Functional genomics  
Evolutionary genomics  
Genome dynamics  
Technology development

## Neuroscience



Data analysis  
Modeling neural systems  
Behavioral modeling

## Other



Biophysics  
Machine learning  
Software development

**Who does Quantitative Biology at CSHL?**

## Core QB program

---



Ivan  
Iossofov



Peter  
Koo



Hannah  
Meyer



Justin  
Kinney



Dan  
Levy



Saket  
Navlakah



Alexander  
Krasnitz



David  
McCandlish



Adam  
Siepel



Michael  
Wigler

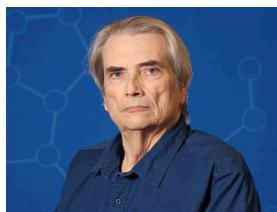
## QB Associated Faculty

---

### Genomics

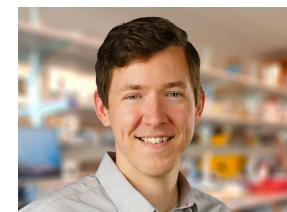


Alexander  
Dobin



Richard  
McCombie

### Neuroscience



Ben  
Cowley



Tom  
Gingeras



Doreen  
Ware



Alexei  
Koulakov



Partha  
Mitra

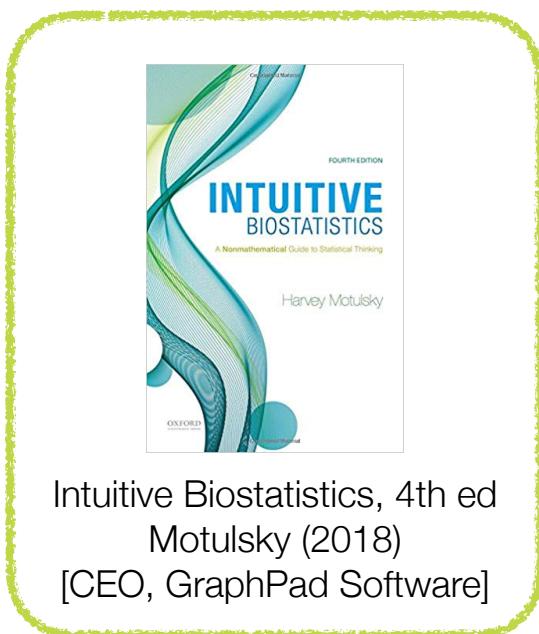
**What QB skills should all biology researchers have?**

## Learn to interpret standard statistics

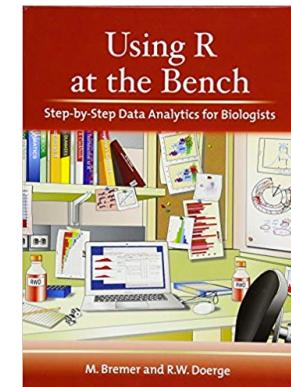
---

### Key statistical concepts:

- P-values
- Multiple hypothesis testing
- Confidence intervals
- Regression
- ANOVA
- Survival analysis

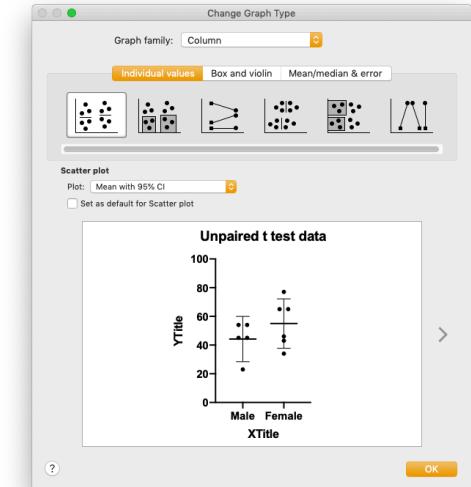
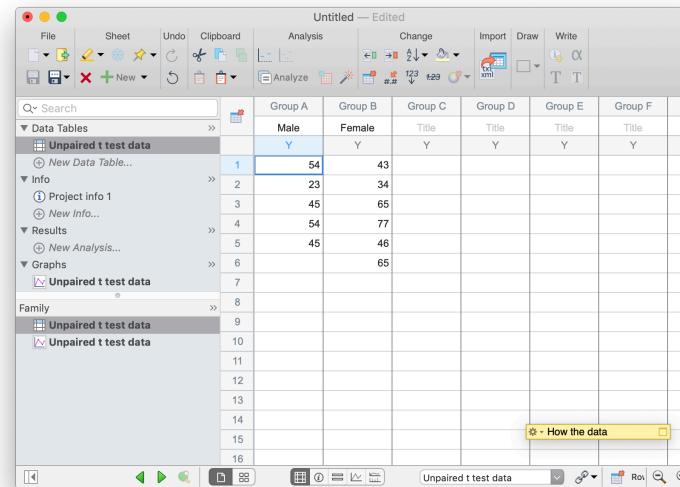
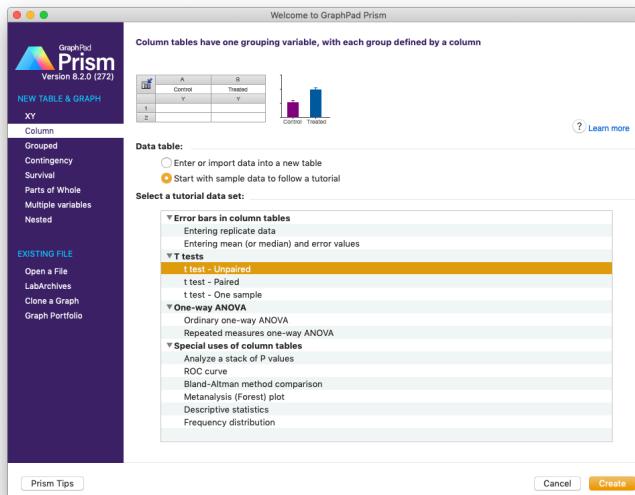


Intuitive Biostatistics, 4th ed  
Motulsky (2018)  
[CEO, GraphPad Software]



Using R at the Bench  
Bremmer & Doerge (2015)

# Learn to compute standard statistics



Alternatively:



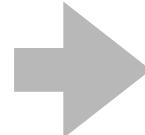
# Learn to navigate UNIX systems



Sequencer



Microscope



High Performance  
Computer Cluster

A screenshot of a UNIX command line terminal window showing a file listing. The terminal title is "jkinney — ssh bnbdev2 — 80x24". The command "ls" was run, displaying files like 15\_splicing, 18\_splicing, 19\_mrna, 19\_wpx, 18\_splicing\_local, 18\_splicing\_3ss, 19\_softy, 17\_arcs, 18\_splicing\_sim2, 19\_exercise\_tor, 17\_arcs\_chip, 18\_splicing\_twistamp, big\_data, and old\_filesys.

```
jkinney@bnbdev2:~$ ls
15_splicing          18_splicing          19_mrna      bin
15_splicing_local    18_splicing_3ss     19_wpx       bnb_exercise_tor
17_arcs              18_splicing_sim2    19_softy    freezer
17_arcs_chip         18_splicing_twistamp  big_data   old_filesys
jkinney@bnbdev2:~$
```

UNIX command line



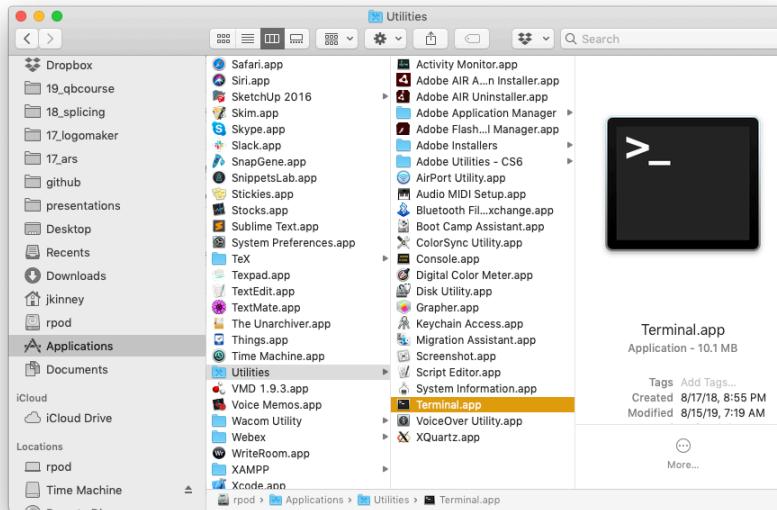
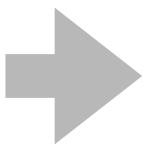
Databases



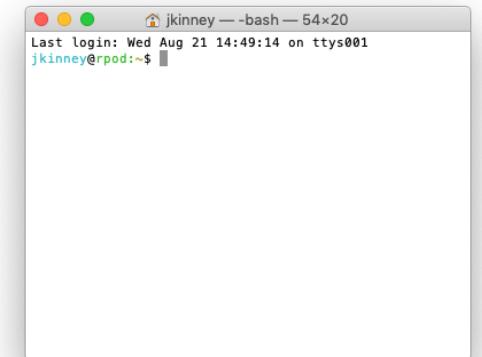
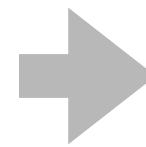
# Mac OS X is based on UNIX



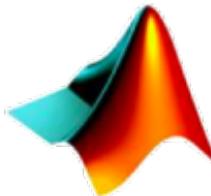
Finder



Applications/Utilities/Terminal.app



## Become familiar with at least one programming language

language	strengths	weaknesses
 python™	<ul style="list-style-type: none"><li>- elegant language</li><li>- easy to learn</li><li>- flexibility: use for large pipelines or local data analysis</li><li>- highly valued skill in industry</li><li>- <b>primary language for deep learning</b></li></ul>	<ul style="list-style-type: none"><li>- clunky dataframes</li><li>- clunky graphics</li><li>- clunky statistics</li></ul>
	<ul style="list-style-type: none"><li>- streamlined for statistics</li><li>- highly developed for genomics</li><li>- great graphics</li></ul>	<ul style="list-style-type: none"><li>- strange language</li><li>- not great for building pipelines</li></ul>
 MATLAB	<ul style="list-style-type: none"><li>- used heavily in neuroscience</li></ul>	<ul style="list-style-type: none"><li>- proprietary</li><li>- poorly supported</li><li>- bad graphics</li><li>- bad for strings</li></ul>

# Learn to analyze your own sequencing data

The screenshot shows the CSHL/BSR Galaxy homepage. The left sidebar contains a navigation menu with sections for Tools (including CSHL TOOLS like Get Data, Quality Control, UTILITIES like RNA-seq, Single Cell RNA-seq, ATAC-seq, HiC Tools, ChIP-seq, Variant Calling, Plots and Graphs), TOOLKITS (Custom Genome Analysis, Export Files, Text Manipulation, Table Manipulation, Convert Formats, Operate on Genomic Intervals, Statistics, FASTX manipulation, GFF Manipulation, Multiple Sequence Alignment, Genome Browser tools, Bedtools), and a search bar for tools.

The main content area features:

- New Updates**:
  - Dec.11 – New Hi-C tools were added
  - Dec.2 – BSR recommends users to use HiSAT2 for mapping data. The Pachter lab which developed Tophat also recommends this.
- Internal Resources**:
  - [Galaxy Quickstart Tutorial](#)
  - [Tutorials for common analyses](#)
  - [BSR Wiki \(coming soon!\)](#)
  - [Assaf Gordon's tutorials](#)
  - [Tool version database](#)
  - [BSR Homepage](#)
  - [Contact us](#) – BSR (bsr@cshl.edu) or Ying Jin (yjin@cshl.edu), Miu Ki Yip (myip@cshl.edu) or Oliver Tam (tam@cshl.edu)
- External Resources**:
  - [Commonly used Analysis Pipelines \(articles\)](#)
  - [Public Galaxy \(Penn State/JHU/TACC/iPlant\)](#)
  - [Cistrome Galaxy for integrative ChIP-Seq analysis \(Harvard – Dana Farber Cancer Institute\)](#)
- Galaxy citations**:
  - Goecks J., Nekrutenko A., Taylor J. and The Galaxy Team. (2010) [Galaxy: a comprehensive approach for supporting reproducible computational workflows](#). *Nature Methods* 7(11): 947–948.

The right sidebar displays a history panel titled "History" with an "Unnamed history" section. It indicates that the history is empty and provides instructions to load your own data or get data from an external source.

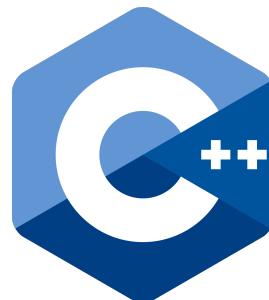
Don't be shy about asking QB labs to help you learn.

**What skills do you need to do research in Quantitative Biology?**

## Learn to program well

---

**Tip: it is better to know one language well than many languages superficially.**



# How to learn to program

---



## BEST ONLINE COURSES FOR PYTHON AT A GLANCE

Our picks for the best subscription / fee-based Python courses and tutorials

- 1. Ask for guidance**
- 2. Work on projects that require it**
- 3. Google your questions & read help threads**
- 4. Read package documentation**
- 5. Read select books**
- 6. Take online courses (don't worry about cost)**

- [Python For Everybody](#) [[coursera.com](https://www.coursera.com)]
- [Learning Python with PyCharm](#) [[lynda.com](https://www.lynda.com)]
- [DataCamp](#) [[datacamp.com](https://www.datacamp.com)]
- [Introduction to Python: Absolute Beginner](#) [[edx.com](https://www.edx.com)]
- [Introduction to Computer Science and Programming Using Python](#) [[edx.com](https://www.edx.com)]
- [Python and Django Full Stack Web Developer Bootcamp](#) [[udemy.com](https://www.udemy.com)]
- [AI Programming with Python](#) [[udacity.com](https://www.udacity.com)]
- [Introduction to Computing in Python](#) [[edx.com](https://www.edx.com)]
- [Python I: Essentials](#) [[quickstart.com](https://www.quickstart.com)]

# Learn to use LaTeX

The screenshot shows a LaTeX editor interface with the following details:

- Left Panel (Code View):** Displays the LaTeX source code for the document `19_mclb.tex`. The code includes package imports like `\usepackage{utf8}`, `\usepackage[T1]{fontenc}`, and `\usepackage{hyperref}`, and defines the title, authors, and abstract.
- Right Panel (Preview View):** Shows the rendered document as a PDF. The title page features the title "Biophysical models of cis-regulation as interpretable neural networks", the names of the authors (Ammar Tareen and Justin B. Kinney), their institutions (Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory), and their email addresses. Below the title is an "Abstract" section followed by sections 1 and 2.
- Top Bar:** Includes standard window controls (red, yellow, green buttons), a menu bar with "MANUAL", "pdfLaTeX + Bibliography", "Configuration", and "Typeset" (highlighted), and a toolbar with icons for "Share", "Outline", "View", "Editor", and "New Tab".
- Bottom Status Bar:** Shows "Page 1 of 6" and various document navigation icons.

**Document Content (Title Page):**

Biophysical models of cis-regulation as  
interpretable neural networks

Ammar Tareen  
Simons Center for Quantitative Biology  
Cold Spring Harbor Laboratory  
Cold Spring Harbor, NY 11724  
`tareen@cshl.edu`

Justin B. Kinney  
Simons Center for Quantitative Biology  
Cold Spring Harbor Laboratory  
Cold Spring Harbor, NY 11724  
`jkinney@cshl.edu`

**Abstract**

Biophysical models that describe gene regulation, as well as other cis-regulatory processes, can be formulated as deep neural networks. This is true of quasi-equilibrium (a.k.a. thermodynamic) models as well as non-equilibrium (a.k.a. kinetic) models. This observation suggests new ways of using powerful deep learning frameworks for training biophysically interpretable neural networks using data produced by massively parallel reporter assays (MPRAs). We demonstrate this capability using previously published MPRA data, and find that using deep learning frameworks to infer such biophysical models yields a dramatic improvement over previously reported model inference approaches.

**1 Introduction**

There are three main types of biophysical models for cis-regulation: thermodynamic, kinetic, and stochastic. Here we focus on the first two kinds of models, both of which can be inferred (at least in principle) from the type of time-averaged data that MPRAs produce. Thermodynamic models are currently the standard way to biophysically model gene regulation [1–6]. These models assume that cis-regulatory complexes form as they would in thermodynamic equilibrium, and that this equilibrium is not greatly disturbed by the downstream kinetic processes that they regulate. By contrast, kinetic models assume that a cis-regulatory system is in steady state, but not necessarily thermal equilibrium. Thermodynamic models have proven remarkably successful at explaining the quantitative activity of a small number of bacterial promoters [7–9]. They have also been applied to a variety of regulatory contexts in yeast [10] and metazoans [11, 12]. Kinetic models have been applied less extensively, but there is a great deal of interest in them due to their ability to perform computations that thermodynamic models cannot [13–15]. However, confidently constructing either type of biophysical model for real biological systems remains a major challenge. A major stumbling block is the lack of available software. Although it was shown early on that biophysical models could be inferred from MPRA data [16], no general-purpose software for performing this type of MPRA data analysis has been described.

**2 Thermodynamic models as deep neural networks**

Thermodynamic models are specified by a set of molecular complexes, or “states”, which we index using  $s$ . Each state has both a Gibbs free energy  $\Delta G_s$  and an associated activity  $\alpha_s$ . These energies determine the probability  $P_s$  of each state occurring in thermodynamic equilibrium via the Boltzmann distribution,<sup>1</sup>

$$P_s = \frac{e^{-\Delta G_s}}{\sum_{s'} e^{-\Delta G_{s'}}}. \quad (1)$$

<sup>1</sup>To reduce notational burden, all  $\Delta G$  values are assumed to be in thermal units. At 37°C, one thermal unit is  $1 k_B T = 0.62 \text{ kcal/mol}$ , where  $k_B$  is Boltzmann’s constant and  $T$  is temperature.

## Develop core quantitative knowledge

---

### **Fundamentals**

Calculus  
Linear Algebra  
Algorithms (basic)  
Statistics (basic)

### **Intermediate material**

Bayesian inference  
Introductory machine learning  
Deep learning  
Sequence analysis  
Population genetics  
Theoretical neuroscience  
Algorithms (intermediate)

### **Advanced material**

Molecular biophysics  
Stochastic processes  
Dynamical systems  
Information theory  
...

**Master all of  
these topics**

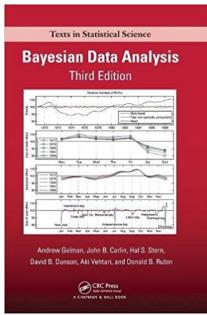
**Master at least one  
of these topics**

**Learn selected  
topics as needed**

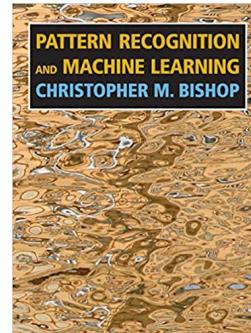
## Learn to work through technical books systematically and independently

### Mentored independent study in QB:

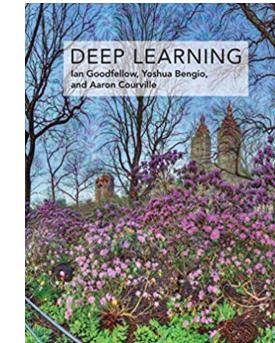
Faculty will help interested students pursue directed reading of graduate-level material.  
Email me <[jkinney@cshl.edu](mailto:jkinney@cshl.edu)> if interested.



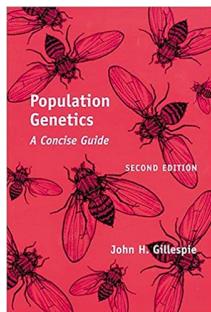
**Bayesian Data Analysis, 3rd ed**  
Gelman et al., 2013



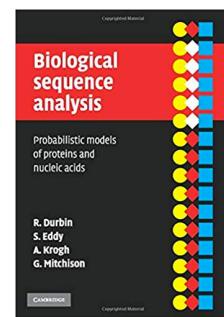
**Pattern Recognition and  
Machine Learning**  
Bishop, 2006



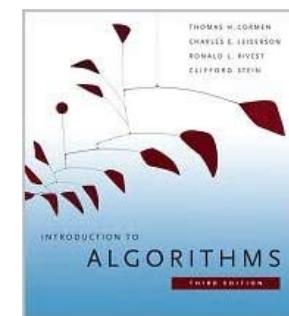
**Deep Learning**  
Goodfellow et al., 2016



**Population Genetics:  
A concise guide, 2nd ed**  
Gillespie, 2004



**Biological Sequence Analysis**  
Durbin et al., 1998



**Introduction to Algorithms**  
Cormen et al., 2009

## **Other tips**

---

### **Attend the weekly QB seminars**

Wednesdays at 12pm, Hawkins.

### **Attend QB Tea Time**

Tuesday at 3pm, Samet.

Email Peter Koo to get on mailing list.