

# Welcome to Quantitative Biology

---



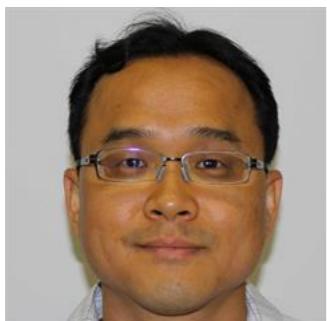
QB Bootcamp, Day 1  
Wednesday, 28 August 2019  
10:00am - 10:30am

## QB Bootcamp team

---



**Justin B. Kinney**  
(Assistant Professor)



**Wei-Chia Chen**  
Postdoc, Kinney Lab



**Ben Harris**  
WSBS, Gillis Lab



**Hussein Hijazi**  
Postdoc, Siepel Lab



**Shaina Lu**  
WSBS, Gillis /  
Zador Labs

# 2019 QB Bootcamp Schedule

---

- **Wednesday, August 28, 10a-5p**
  - 10:00a - 10:45a: Welcome to quantitative biology (**Justin**)
  - 10:45a - 12:00p: The UNIX command line (**Justin**)
  - *12:00p - 1:00p: Lunch*
  - 1:00p - 1:30p: Introduction to Python and Jupyter Notebooks (**Justin**)
  - 1:30p - 3:00p: Python: data types (**Ben**)
  - *3:00p - 3:30p: Break*
  - 3:30p - 5:00p: Python: flow control (**Ben**)
- **Thursday, August 29, 10a-5p**
  - 10:00a - 10:30a: Overview of high-performance computing (**Justin**)
  - 10:30a - 12:00p: BlackNBlue, read mapping (**Justin**)
  - *12:00p - 1:00p: Lunch*
  - 1:00p - 1:30p: Introduction to dataframes (**Justin**)
  - 1:30p - 3:00p: Pandas I, binding site analysis (**Hussein**)
  - *3:00p - 3:30p: Break*
  - 3:30p - 5:00p: Pandas II, replication origin analysis (**Hussein**)
- **Friday, August 30, 2p-6p**
  - 2:00p - 2:30p: Overview of data visualization (**Justin**)
  - 2:30p - 4:00p: Matplotlib (**Shaina**)
  - *4:00p - 4:30p: Break*
  - 4:30p - 6:00p: Seaborn (**Shaina**)

[https://github.com/jbkinney/19\\_qbbootcamp](https://github.com/jbkinney/19_qbbootcamp)

## Download this repository

The screenshot shows a GitHub repository page for 'jbkinney/19\_qbbootcamp'. The page includes a header with navigation links like Pull requests, Issues, Marketplace, and Explore. Below the header, there's a main content area with sections for Code, Issues (0), Pull requests (0), Projects (0), Wiki, Security, Insights, and Settings. A summary bar shows 12 commits, 1 branch, 0 releases, and 2 contributors. A green box highlights the 'Clone or download' button, which has a dropdown menu showing 'Clone with HTTPS' and 'Use SSH' options, along with a URL field containing 'https://github.com/jbkinney/19\_qbbootcamp'. Below the summary bar is a list of files: bash, cheatsheets, python, README.md, bnb\_exercise.tar.gz, mac\_install.sh, and qb\_syllabus.pdf. At the bottom, there's a section titled '2019 Quantitative Biology Bootcamp' with a welcome message and a 'Summary' link.

jbkinney/19\_qbbootcamp: GitHub

GitHub, Inc. [US] | github.com/jbkinney/19\_qbbootcamp

Search or jump to... / Pull requests Issues Marketplace Explore

jbkinney / 19\_qbbootcamp

Unwatch 2 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Security Insights Settings

GitHub repository for the 2019 QB Bootcamp

Manage topics

12 commits 1 branch 0 releases 2 contributors

Branch: master New pull request Create new file Upload files Find File Clone or download

jbkinney Merge branch 'master' of https://github.com/jbkinney/19\_qbbootcamp

bash	modified bash/* contents
cheatsheets	revised 7_seaborn.ipynb and 6_matplotlib.ipynb, as well as e
python	revised 7_seaborn.ipynb and 6_matplotlib.ipynb, as well as e
README.md	Add step to mac install
bnn_exercise.tar.gz	Added bnn_exercise.tar.gz
mac_install.sh	initial commit
qb_syllabus.pdf	revised syllabus

Open in Desktop Download ZIP

21 hours ago 16 hours ago

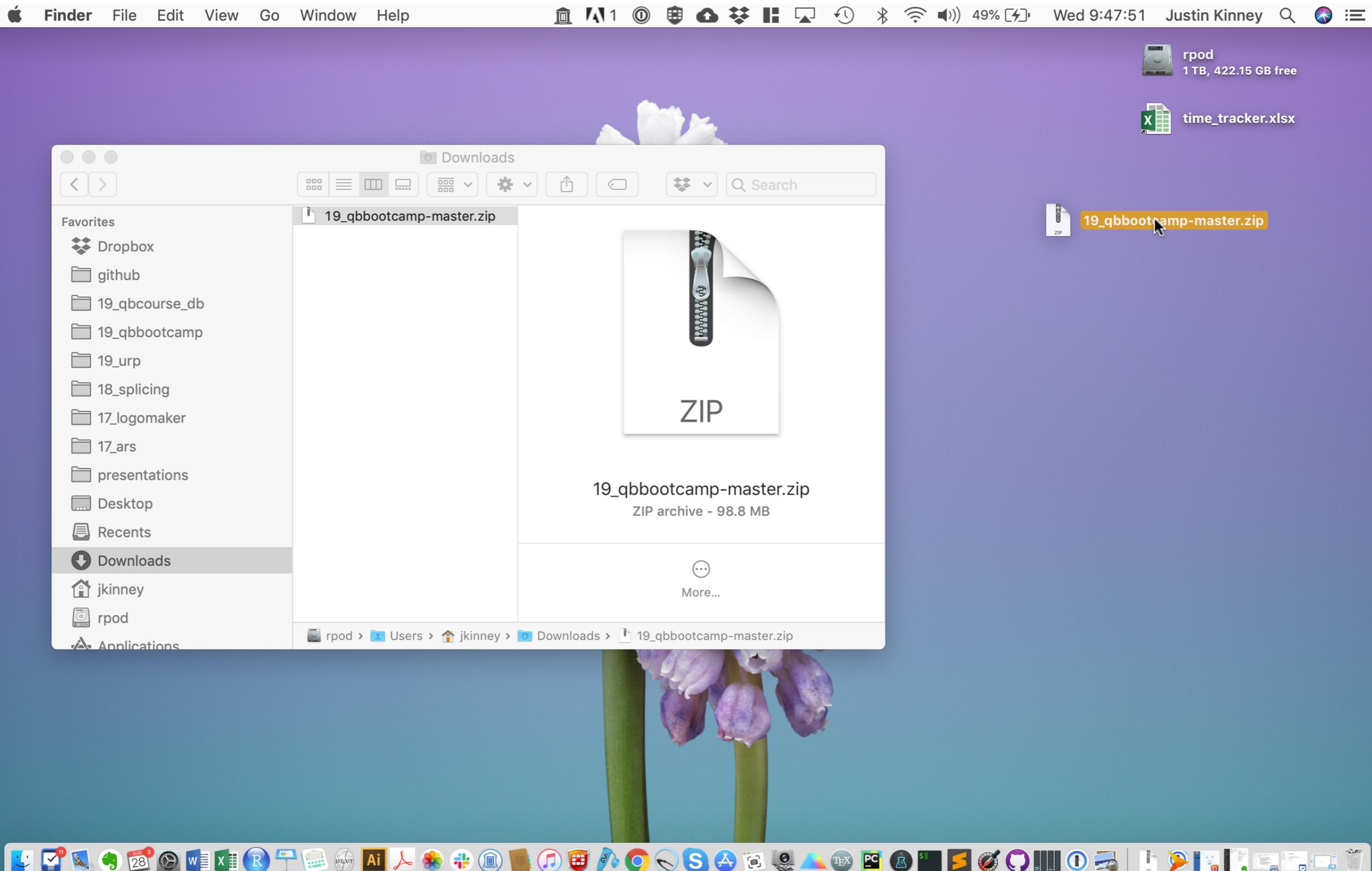
README.md

## 2019 Quantitative Biology Bootcamp

Welcome to the 2019 QB Bootcamp in the Watson School for Biological Sciences at Cold Spring Harbor Laboratory! This Github repository contains the Jupyter notebooks, shell scripts, and data sets that we will work through in this bootcamp.

Summary

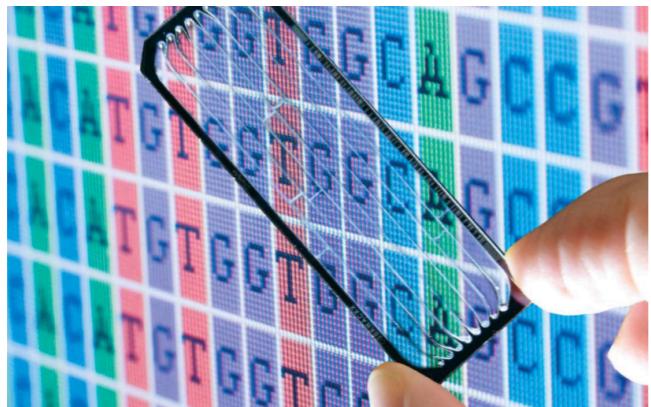
**Move this repository from Downloads to your Desktop  
and decompress by double-clicking**



## **What is Quantitative Biology?**

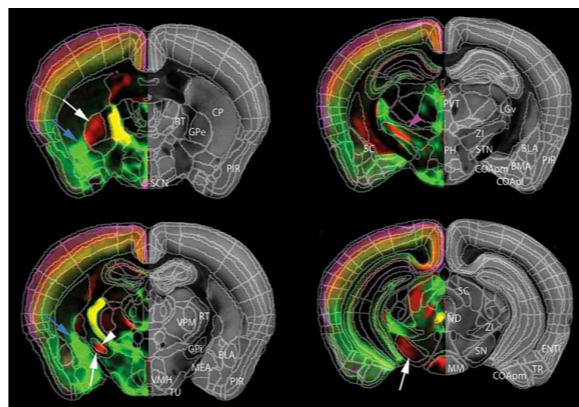
# Quantitative biology is a vast field

## Genomics



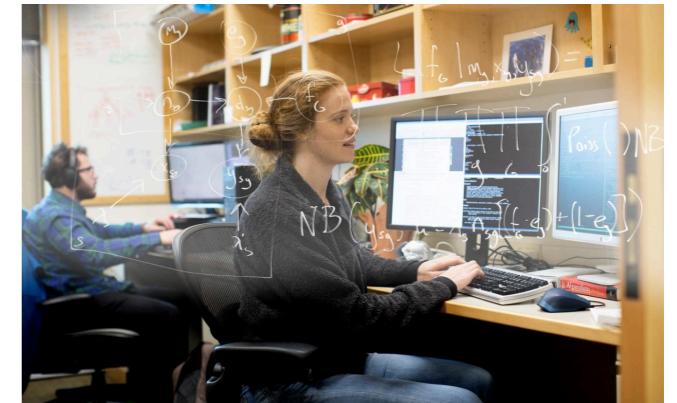
Functional genomics  
Evolutionary genomics  
Genome dynamics  
Technology development

## Neuroscience



Data analysis  
Modeling neural systems  
Behavioral modeling

## Other



Biophysics  
Machine learning  
Software development

## **Who does Quantitative Biology at CSHL?**

## Core QB program

---



**Molly Gale  
Hammell**



**Dan  
Levy**



**Saket  
Navlakah**



**Ivan  
Iossofov**



**David  
McCandlish**



**Justin  
Kinney**



**Hannah  
Meyer**



**Peter  
Koo**



**Alexander  
Krasnitz**



**Adam  
Siepel**

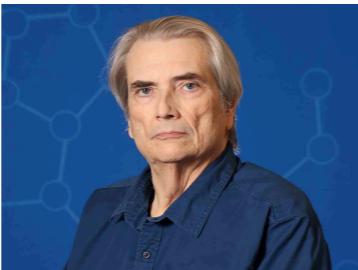
## QB Associated Faculty

---

### Genomics



Alexander  
Dobin



Richard  
McCombie

### Neuroscience



Tatiana  
Engel



Jesse  
Gillis



Doreen  
Ware



Alexei  
Koulakov



Partha  
Mitra

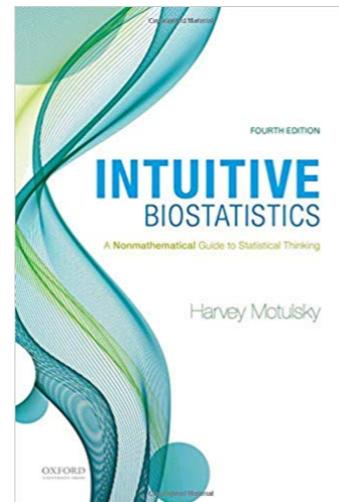
**What QB skills should all biology researchers have?**

# Learn to interpret standard statistics

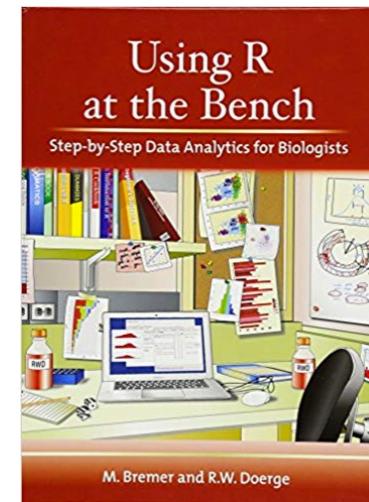
---

## Key statistical concepts:

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

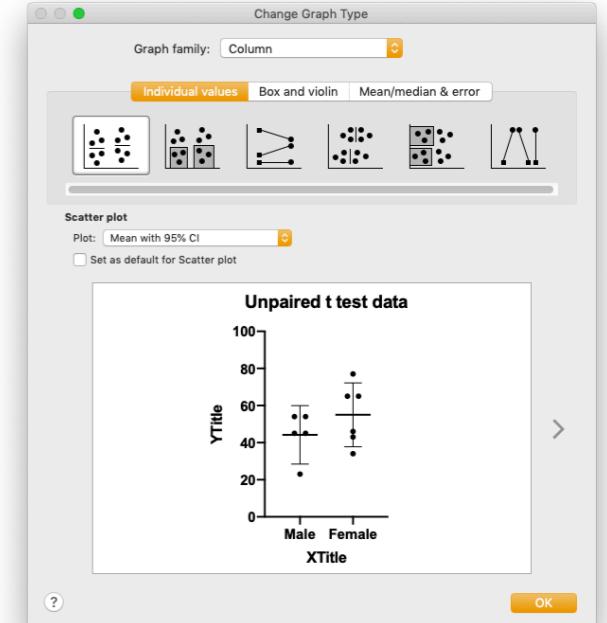
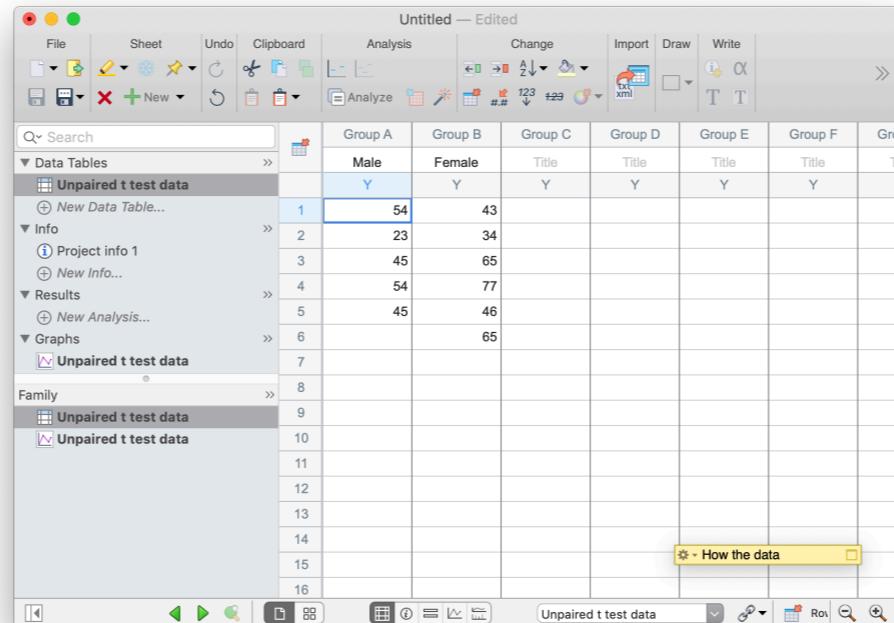
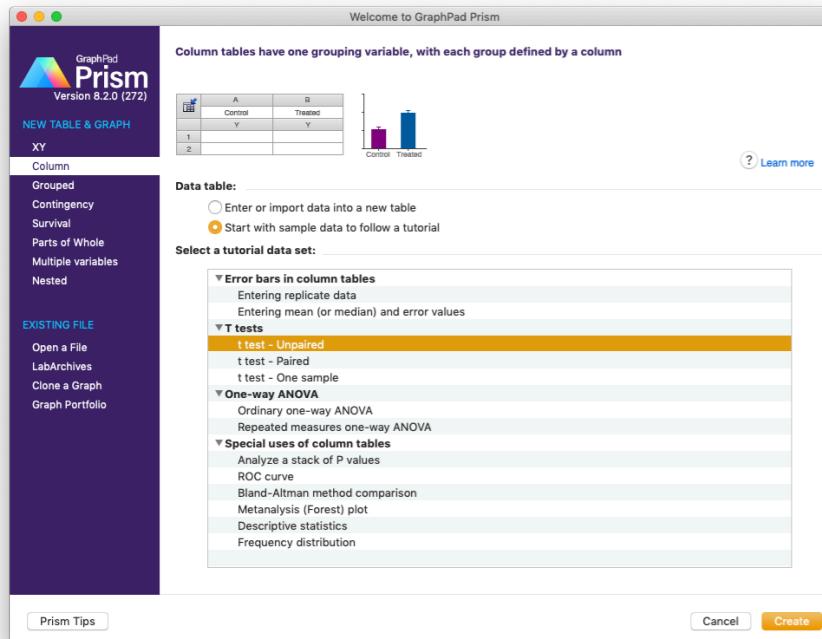


Intuitive Biostatistics, 4th ed  
Motulsky (2018)  
[CEO, GraphPad Software]



Using R at the Bench  
Bremmer & Doerge (2015)

# Learn to compute standard statistics



Alternatively:



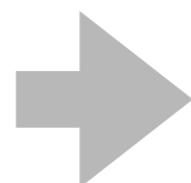
# Learn to navigate UNIX systems



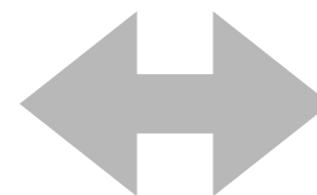
Sequencer



Microscope



High Performance  
Computer Cluster

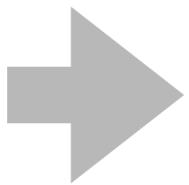
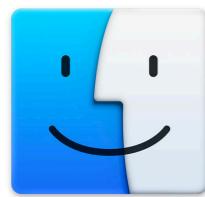
A screenshot of a Mac OS X terminal window showing a UNIX command line session. The user 'jkinney' is connected via SSH to a host named 'bnbdev2'. The terminal displays the output of the 'ls' command, listing various directories and files related to splicing and genome analysis.

UNIX command line

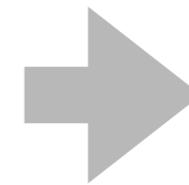
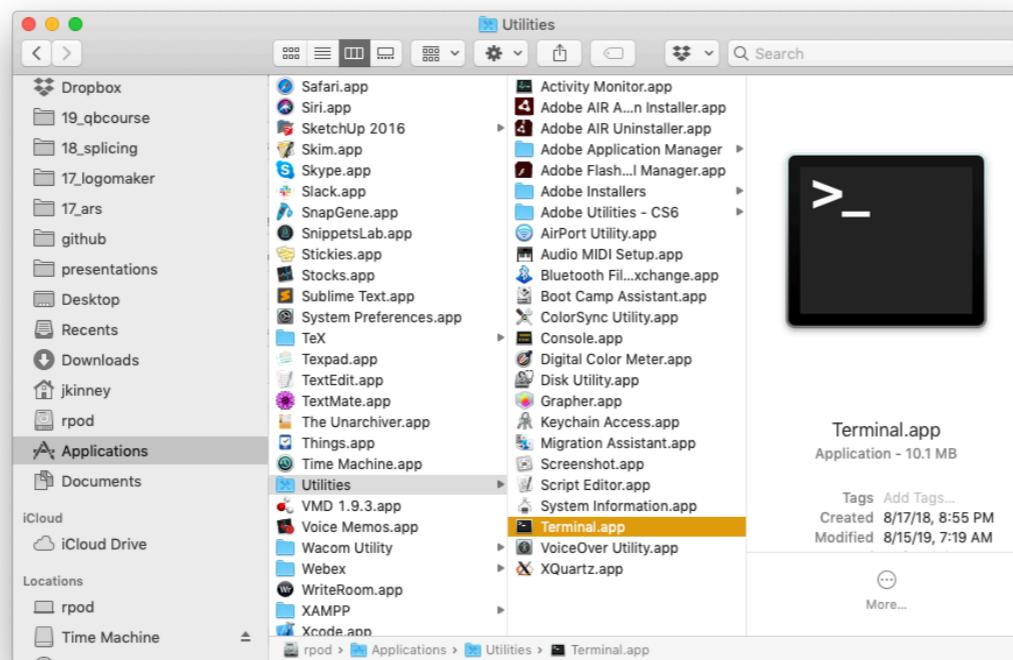


Databases

# Mac OS X is based on UNIX

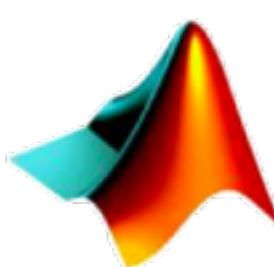


Finder

A screenshot of a Terminal window. The title bar reads "jkinney — bash — 54x20". The command line shows the user's last login information: "Last login: Wed Aug 21 14:49:14 on ttys001" and the current command prompt "jkinney@rpod:~\$".

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></ul>	<ul style="list-style-type: none"><li>- clunky dataframes</li><li>- clunky statistics</li><li>- clunky graphics</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 and by old people</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 top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Help, Login or Register, and a search bar. The main content area features sections for New Updates, Internal Resources, and External Resources. The left sidebar contains a 'Tools' section with a search bar and categories like CSHL TOOLS (Get Data, Quality Control), UTILITIES (RNA-seq, Single Cell RNA-seq, ATAC-seq, HiC Tools, ChIP-seq, Variant Calling, Plots and Graphs), and 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). The right sidebar shows an 'History' section titled 'Unnamed history (empty)' with a note about loading data.

**CSHL/BSR Galaxy**

**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 \(Harvard – Dana Farber Cancer Institute\)](#)

The BSR Galaxy project is supported in part by the [National Institute of Health](#) and [National Cancer Institute](#).

If you use the BSR Galaxy for data analysis in a paper or poster, please acknowledge the CSHL Bioinformatics Shared Resource in your publication.

**Galaxy citations**

- Goecks J., Nekrutenko A., Taylor J. and The Galaxy Team. (2010) [Galaxy: a comprehensive approach for supporting reproducible computational workflows](#) in the *Life Sciences Software Tools and Applications* journal.

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](#)]
- [Learning Python with PyCharm](#) [[lynda.com](#)]
- [DataCamp](#) [[datacamp.com](#)]
- [Introduction to Python: Absolute Beginner](#) [[edx.com](#)]
- [Introduction to Computer Science and Programming Using Python](#) [[edx.com](#)]
- [Python and Django Full Stack Web Developer Bootcamp](#) [[udemy.com](#)]
- [AI Programming with Python](#) [[udacity.com](#)]
- [Introduction to Computing in Python](#) [[edx.com](#)]
- [Python I: Essentials](#) [[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 document class declarations, package imports, and author information.
- Right Panel (Preview):** Shows the rendered document page titled "Biophysical models of cis-regulation as interpretable neural networks".
- Header Bar:** Includes tabs for "MANUAL", "pdfTeX + Bibliography", "Configuration", "Typeset", "View", "Editor", and "New Tab".
- Toolbar:** Features icons for "Share", "Outline", "View", "Editor", and "New Tab".
- Page Header:** "19\_mclb.tex" and "Page 1 of 6".
- Page Content:**
  - Title:** Biophysical models of cis-regulation as interpretable neural networks
  - Authors:** Ammar Tareen and Justin B. Kinney, both from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory, NY 11724. Their emails are listed as `tareen@cshl.edu` and `jkinney@cshl.edu`.
  - Abstract:** A brief summary of the research, mentioning the use of deep learning frameworks to infer biophysical models from MPRA data.
  - Introduction:** A section discussing the three main types of biophysical models for cis-regulation: thermodynamic, kinetic, and stochastic. It highlights the use of deep neural networks for training biophysically interpretable models.
  - Thermodynamic models as deep neural networks:** A section explaining how thermodynamic models are specified by molecular complexes and how they relate to deep neural networks.

# Develop core quantitative knowledge

---

## **Fundamentals**

Calculus  
Linear Algebra  
Algorithms (basic)  
Statistics (basic)

## **Intermediate material**

Bayesian inference  
Machine learning  
Sequence analysis  
Population genetics  
Theoretical neuroscience  
Algorithms (intermediate)

## **Advanced material**

Molecular biophysics  
Stochastic processes  
Dynamical systems  
Information theory  
Deep learning  
...

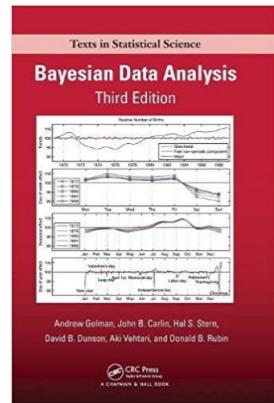
**Master all of  
these topics**

**Master at least one  
of these topics**

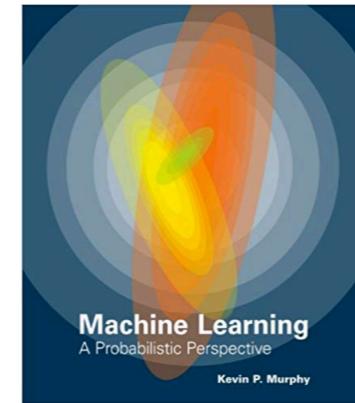
**Learn selected  
topics as needed**

# Learn to work through technical books systematically and independently

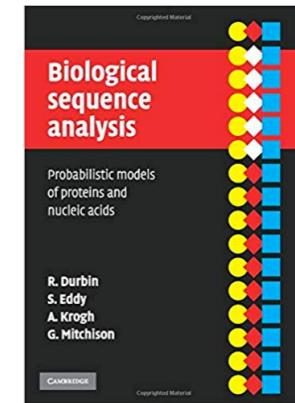
**New QB initiative:** faculty will help interested students pursue directed independent study of graduate-level material.  
Email me <[jkinney@cshl.edu](mailto:jkinney@cshl.edu)> if interested.



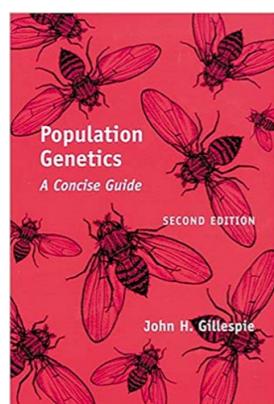
**Bayesian Data Analysis, 3rd ed**  
Gelman et al., 2013



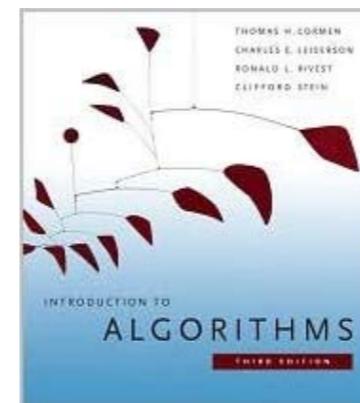
**Machine Learning:  
A Probabilistic Perspective**  
Murphy, 2012



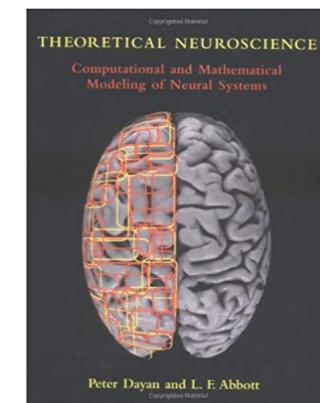
**Biological Sequence Analysis**  
Durbin et al., 1998



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



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



**Theoretical Neuroscience**  
Dayan and Abbott, 2001

## Other tips

---

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

Wednesdays at 12pm, Hawkins.

### **Attend QB Tea Time**

Wednesdays at 4pm, Samet.

Email me to get on mailing list.