

Welcome to Quantitative Biology



QB Bootcamp, Day 1
Wednesday, 29 August 2022
10:00am - 10:30am

2021 QB Bootcamp Schedule

- **Day 1:** Monday, 29 August 2022, 10am - 5pm, Wendt Conference Room, Wendt 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 - 5:00pm: **Python: flow control (tutorial, Ivan)**
- **Day 2:** Tuesday, 30 August 2022, 10am - 6pm, James Library, James Bldg.
10:00am - 10:30am: **Overview of high-performance computing (lecture, Justin)**
10:30am - 12:00pm: **Read mapping using Elzar (tutorial, Justin)**
12:00pm - 1:00pm: *Lunch (provided)*
1:00pm - 2:00pm: **Python: flow control cont. (tutorial, Ivan)**
2:00pm - 2:15pm: **Introduction to Pandas (lecture, Justin)**
2:15pm - 2:30pm: *Break*
2:30pm - 4:15pm: **Pandas I, TF analysis (tutorial, Justin)**
4:15pm - 4:30pm: *Break*
4:30pm - 6:00pm: **Pandas II, Replication origin analysis (tutorial, Ivan)**
- **Day 3:** Wednesday, 31 August 2022, 2pm - 6pm, Hershey East Conference Room, Hershey Bldg.
2:00pm - 2:30pm: **Introduction to Data Visualization (lecture, Justin)**
2:30pm - 4:00pm: **Matplotlib (tutorial, Ivan)**
4:00pm - 4:30pm: *Break*
4:30pm - 6:00pm: **Advanced visualization (tutorial, Justin)**

jbkinney (Justin B. Kinney) x +

https://github.com/jbkinney

Toggl SmartSheet Papers My NCBI Evernote Github Overleaf 21_hiring Requisiti... Kinney lab > Log In eRA Commons

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

Overview Repositories 51 Projects Packages Stars

Pinned

Customize your pins

logomaker Public Software for the visualization of sequence-function relationships
Jupyter Notebook 136 ⚡ 32

mavenn Public MAVE-NN: genotype-phenotype maps from multiplex assays of variant effect
Jupyter Notebook 17 ⚡ 3

22e_urp Public Forked from bharis12/URP_2021_Programming_Course
2021 URP Python course at CSHL
Jupyter Notebook

22e_qbbootcamp Public Materials for the 2022 QB Bootcamp at Cold Spring Harbor Laboratory
Jupyter Notebook

Justin B. Kinney
jbkinney

Edit profile

26 followers · 1 following

Cold Spring Harbor Laboratory
United States
jkinney@cshl.edu
http://kinneylab.labsites.cshl.edu/

Achievements

x2

Contribution activity

August 2022

Less More

Contribution settings ▾

2022

2021

Learn how we count contributions

jbkinney/22e_qbbootcamp: Ma +

https://github.com/jbkinney/22e_qbbootcamp

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

jbkinney / 22e_qbbootcamp Public

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags

jbkinney Update README.md

- bash Initial commit
- lectures Initial commit
- python Initial commit
- .gitattributes Initial commit
- .gitignore Initial commit
- 22_qbcourse_syllabus.pdf Added 2022 syllabi
- 22e_qbbootcamp_syllabus.pdf Modified syllabus
- LICENSE Initial commit
- README.md Update README.md
- elzar_exercise.tar.gz Initial commit

Clone

HTTPS SSH GitHub CLI

https://github.com/jbkinney/22e_qbboot

Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

About

Materials for the 2022 QB Bootcamp at Cold Spring Harbor Laboratory

- Readme
- MIT license
- 0 stars
- 1 watching
- 0 forks

Releases

No releases published [Create a new release](#)

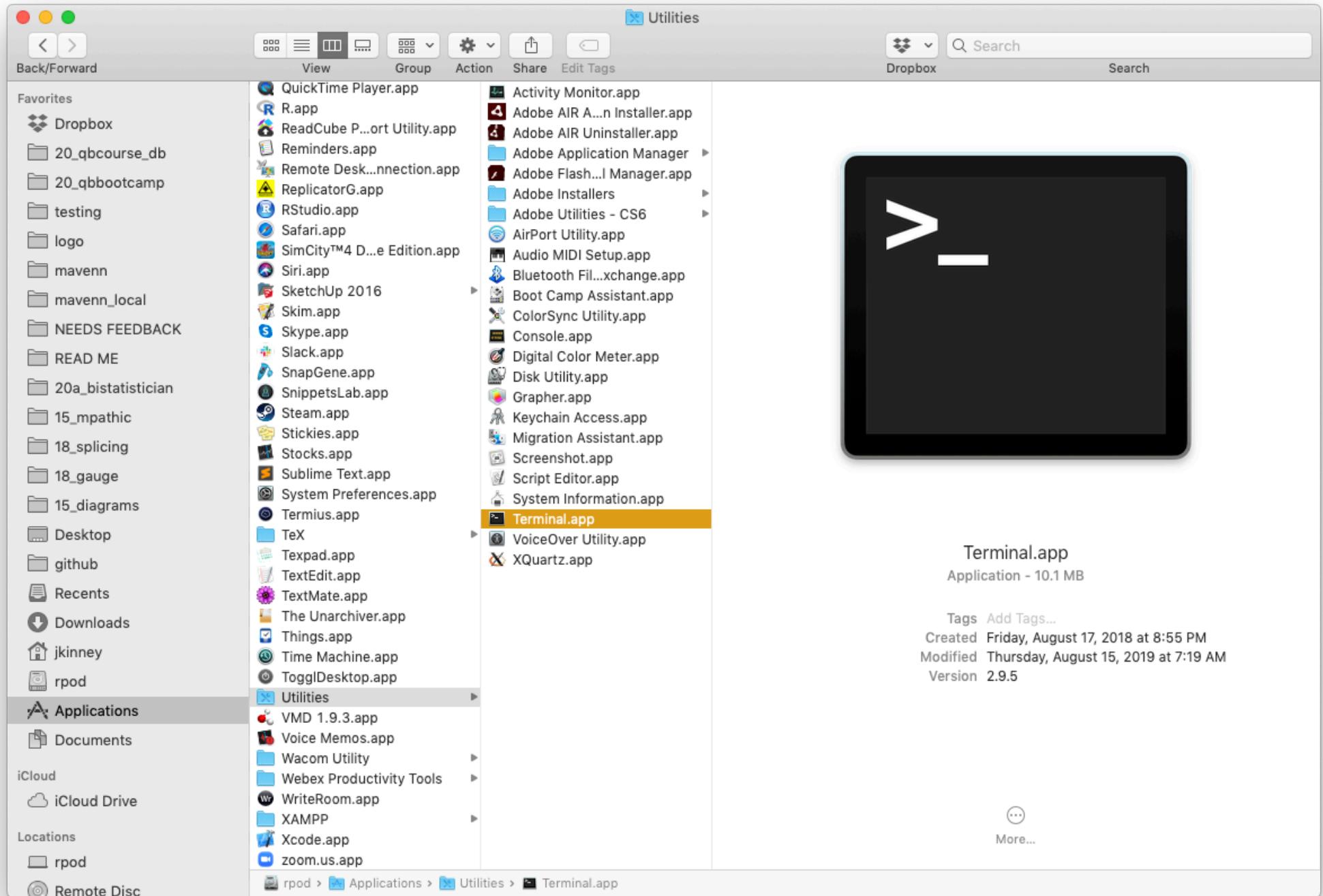
Packages

No packages published [Publish your first package](#)

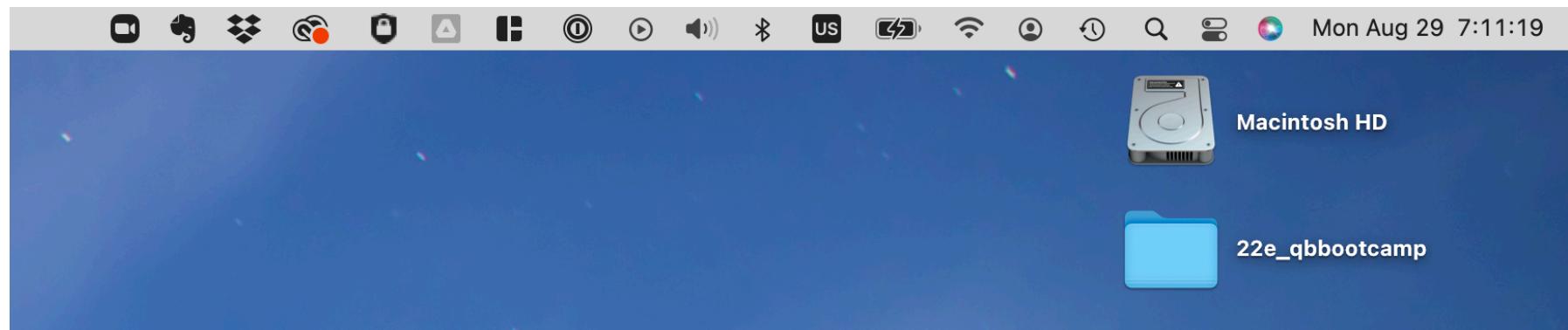
Contributors 2

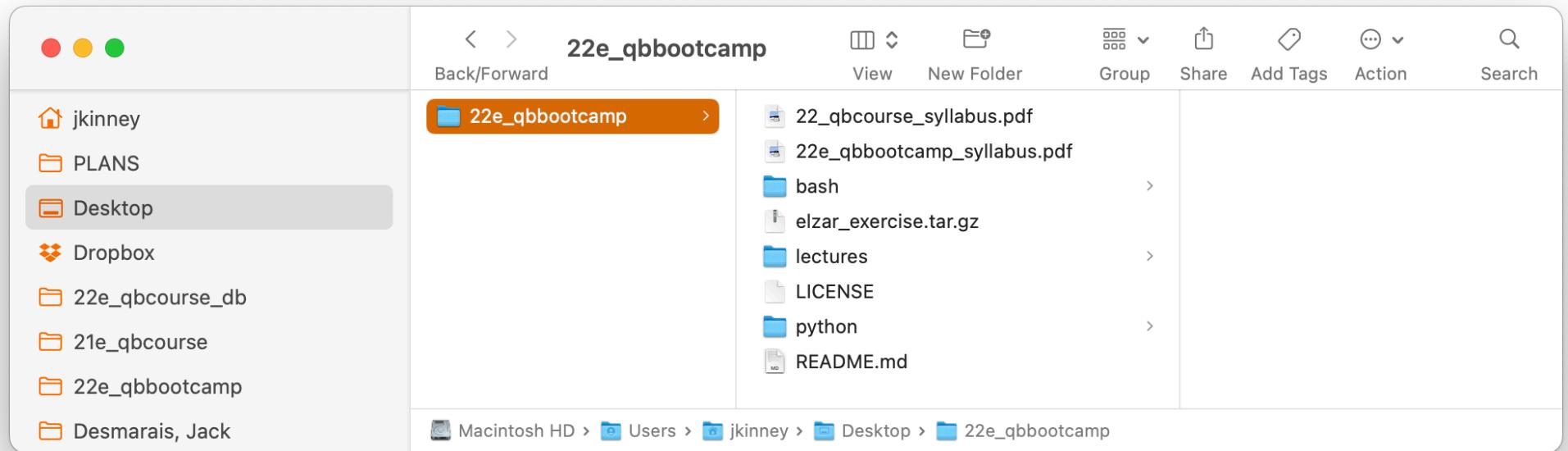
jbkinney Justin B. Kinney

2022 Quantitative Biology Bootcamp



```
Desktop — bash — 96x14
[(base) jkinney@presslaptop20:~/Desktop$ git clone https://github.com/jbkinney/22e_qbbootcamp.git]
Cloning into '22e_qbbootcamp'...
remote: Enumerating objects: 66, done.
remote: Counting objects: 100% (32/32), done.
remote: Compressing objects: 100% (28/28), done.
remote: Total 66 (delta 17), reused 10 (delta 4), pack-reused 34
Receiving objects: 100% (66/66), 137.37 MiB | 6.17 MiB/s, done.
Resolving deltas: 100% (26/26), done.
(base) jkinney@presslaptop20:~/Desktop$
```

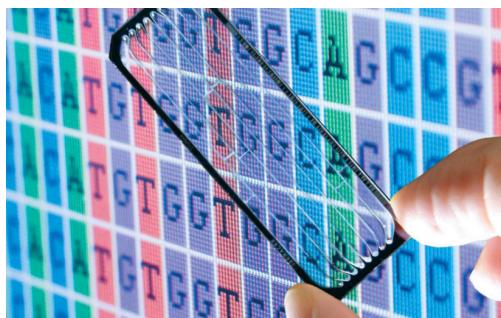




What is Quantitative Biology?

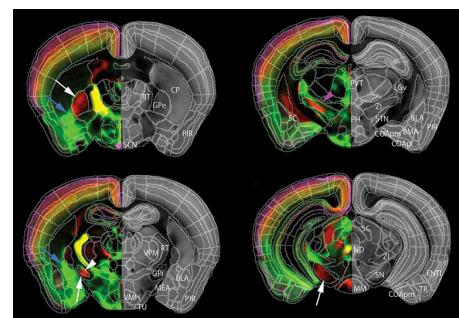
Quantitative biology is a vast field

Genomics



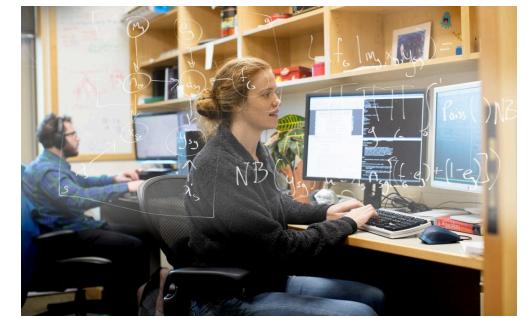
Functional genomics
Evolutionary genomics
Genome dynamics
Technology development

Neuroscience



Data analysis
Modeling neural systems
Behavioral modeling

Other



Biophysics
Machine learning
Software development

Who does Quantitative Biology at CSHL?

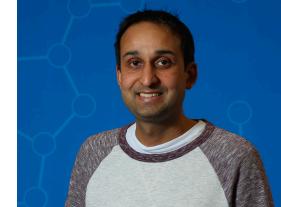
Core QB program



**Molly Gale
Hammell**



**Dan
Levy**



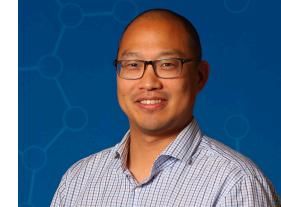
**Saket
Navlakah**



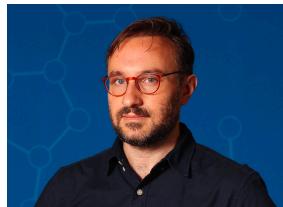
**Ivan
Iossofov**



**David
McCandlish**



**Peter
Koo**



**Justin
Kinney**



**Hannah
Meyer**



**Alexander
Krasnitz**



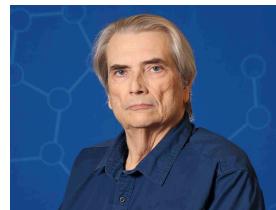
**Adam
Siepel**

QB Associated Faculty

Genomics



Alexander
Dobin



Richard
McCombie

Neuroscience



Tatiana
Engel



Doreen
Ware



Alexei
Koulakov



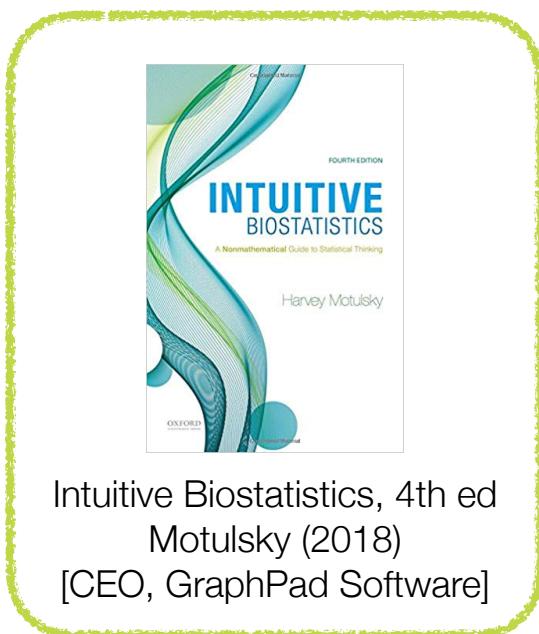
Partha
Mitra

What QB skills should all biology researchers have?

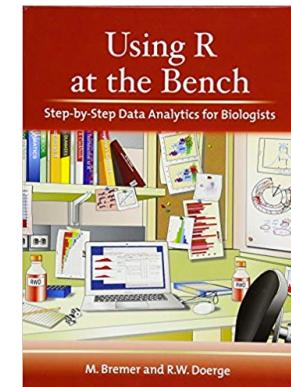
Learn to interpret standard statistics

Key statistical concepts:

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

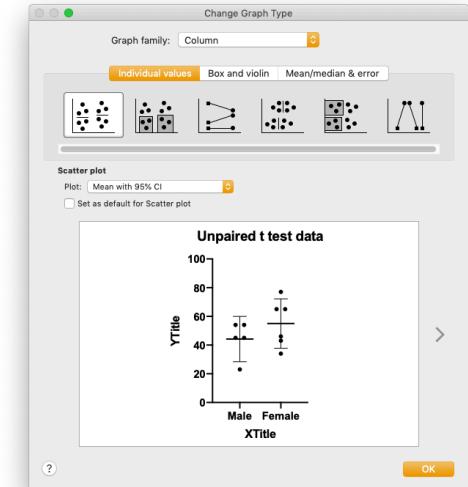
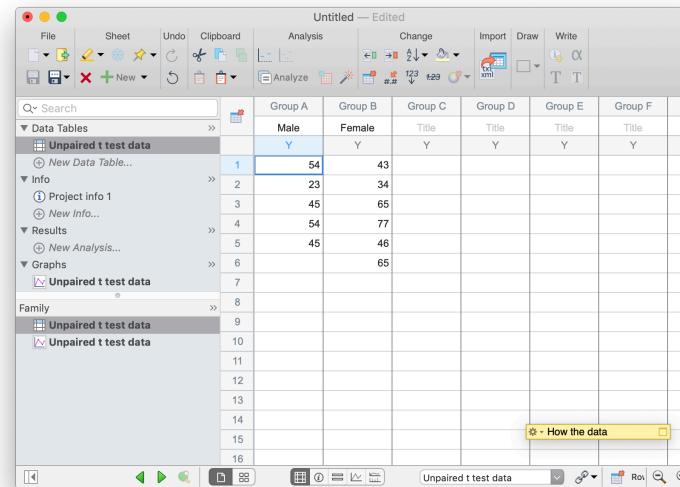
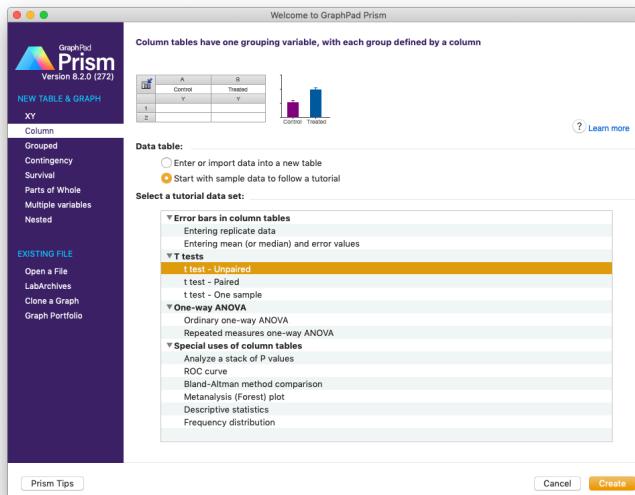


Intuitive Biostatistics, 4th ed
Motulsky (2018)
[CEO, GraphPad Software]



Using R at the Bench
Bremmer & Doerge (2015)

Learn to compute standard statistics



Alternatively:



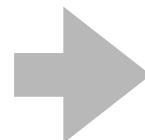
Learn to navigate UNIX systems



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



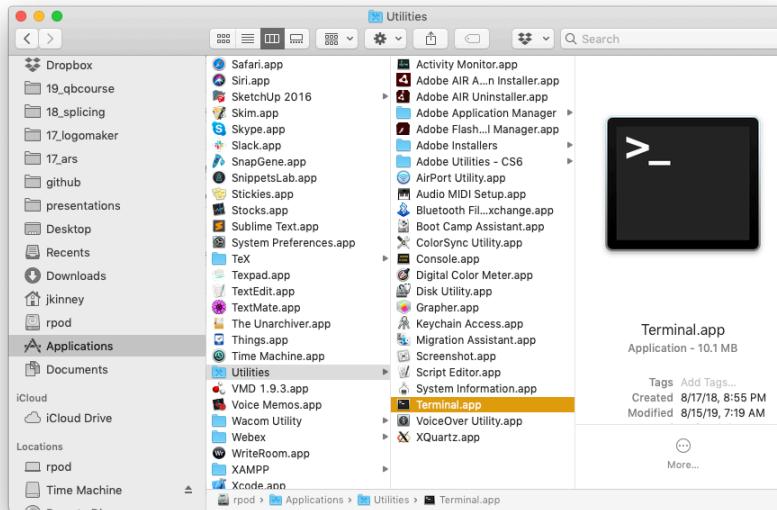
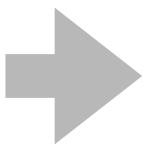
Databases



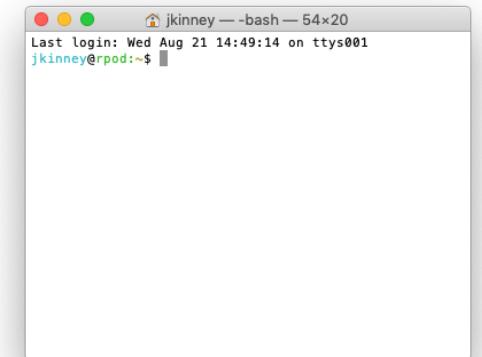
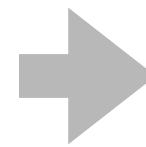
Mac OS X is based on UNIX



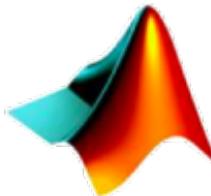
Finder



Applications/Utilities/Terminal.app



Become familiar with at least one programming language

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

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; and TOOLKITS like Custom Genome Analysis), and FASTX manipulation, GFF Manipulation, Multiple Sequence Alignment, Genome Browser tools, and Bedtools. The main content area features a "New Updates" section with two bullet points: "Dec.11 – New Hi-C tools were added" and "Dec.2 – BSR recommends users to use HiSAT2 for mapping data. The Pachter lab which developed Tophat also recommends this." Below this is an "Internal Resources" section with links to Galaxy Quickstart Tutorial, Tutorials for common analyses, BSR Wiki (coming soon!), Assaf Gordon's tutorials, Tool version database, BSR Homepage, and Contact us. The "External Resources" section includes links to Commonly used Analysis Pipelines (articles), Public Galaxy (Penn State/JHU/TACC/iPlant), and Cistrome Galaxy for integrative ChIP-Seq analysis (Harvard – Dana Farber Cancer Institute). A note at the bottom states: "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." The right sidebar shows a history panel titled "History" with an "Unnamed history" entry, which is currently empty. It includes a message: "This history is empty. You can 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, author information, and a detailed abstract.
- Right Panel (Preview):** Shows the rendered document as a PDF. The title is "Biophysical models of cis-regulation as interpretable neural networks". It features two authors: Ammar Tareen and Justin B. Kinney, both from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory. The abstract discusses the use of deep learning for biophysical models of gene regulation.
- Top Bar:** Includes standard window controls (red, yellow, green), a **MANUAL** link, a **pdfLaTeX + Bibliography** button, and other document navigation tools.
- Bottom Bar:** Shows the page number (Page 1 of 6) and various document navigation icons.

19_mclb.tex Content (Partial):

```
22 \usepackage[utf8]{inputenc} % allow utf-8 input
23 \usepackage[T1]{fontenc} % use 8-bit T1 fonts
24 \usepackage{hyperref} % hyperlinks
25 \usepackage{url} % simple URL typesetting
26 \usepackage{booktabs} % professional-quality tables
27 \usepackage{amsfonts} % blackboard math symbols
28 \usepackage{nicefrac} % compact symbols for 1/2, etc.
29 \usepackage{microtype} % microtypography
30 \usepackage{soul} % for \ul
31 \usepackage{graphicx} % for figures
32 \usepackage{upgreek}
33
34 \title{Biophysical models of cis-regulation as\\ interpretable neural networks}
35
36
37 \author{%
38   Ammar Tareen \\
39   Simons Center for Quantitative Biology\\
40   Cold Spring Harbor Laboratory\\
41   Cold Spring Harbor, NY 11724 \\
42   \texttt{tareen@cshl.edu} \\
43   And \\
44   Justin B. Kinney \\
45   Simons Center for Quantitative Biology\\
46   Cold Spring Harbor Laboratory\\
47   Cold Spring Harbor, NY 11724 \\
48   \texttt{jkinney@cshl.edu} \\
49 }
50
51 \begin{document}
52
53 \maketitle
54
55 \begin{abstract}
56 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

```

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 ΔG_s and an associated activity α_s . These energies determine the probability P_s of each state occurring in thermodynamic equilibrium via the Boltzmann distribution,¹

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

¹To reduce notational burden, all ΔG 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**

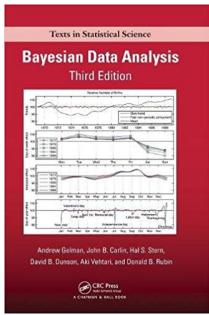
**Master at least one
of these topics**

**Learn selected
topics as needed**

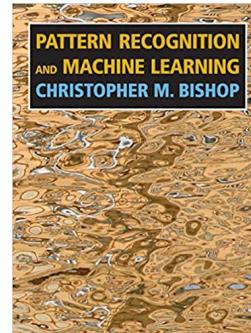
Learn to work through technical books systematically and independently

Mentored independent study in QB:

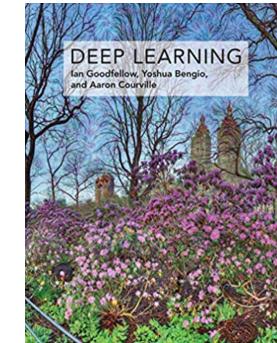
Faculty will help interested students pursue directed reading of graduate-level material.
Email me <jkinney@cshl.edu> if interested.



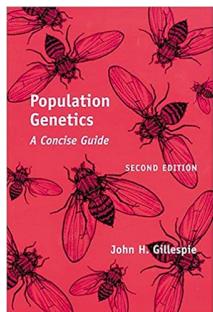
Bayesian Data Analysis, 3rd ed
Gelman et al., 2013



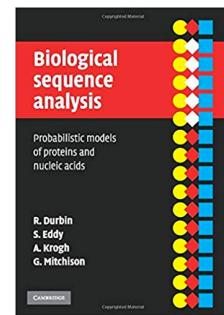
**Pattern Recognition and
Machine Learning**
Bishop, 2006



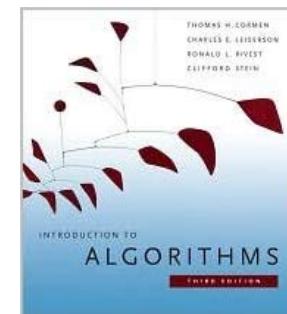
Deep Learning
Goodfellow et al., 2016



**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.