

Welcome to Quantitative Biology



QB Bootcamp, Day 1
Tuesday, 3 September 2024
10:00am - 10:30am

2024 QB Bootcamp Syllabus

2024 Quantitative Biology Bootcamp, CSHL School of Biological Sciences

Instructors:

- Justin Kinney, jkinney@cshl.edu
- Ivan Iossifov, iossifov@cshl.edu

Teaching assistants:

- Ari Benjamin, benjami@cshl.edu
- Kaiser Loell, loell@cshl.edu

GitHub Repository: https://github.com/jbkinney/24e_qbbootcamp/

Day 1: Tuesday, 3 September 2024, 10am - 6pm, Plimpton Conference Room, Beckman 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 Break*
- 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, 4 September 2024, 1pm - 6pm, Plimpton Conference Room, Beckman Bldg.

- 1:00pm - 1:30pm: **NextGen sequencing and high-performance computing (lecture, Justin)**
- 1:30am - 3:00pm: **Read mapping using Elzar (tutorial, Justin)**
- 3:00pm - 3:15pm: *Break*
- 3:15pm - 3:30pm: **Introduction to dataframes (lecture, Justin)**
- 3:30pm - 5:00pm: **Pandas I, TF analysis (tutorial, Ivan)**

Day 3: Thursday, 5 September 2024, 10am - 6pm, Plimpton Conference Room, Beckman Bldg.

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

jbkinney (Justin B. Kinney) x +

https://github.com/jbkinney

Toggl Overleaf SmartSheet Papers Github ChatGPT CSHL: Lawson Supergroup Sched... Mimecast eRA eRA Commons Concur

jbkinney Type ⌂ to search

Overview Repositories 59 Projects Packages Stars 1

Pinned Order updated Customize your pins

24e_qbootcamp Public Repository for the 2024 QB Bootcamp Jupyter Notebook

24e_urp Public Code for the 2024 CSHL URP Python Programming Course Jupyter Notebook 1 3

mavenn Public MAVEN-NN: genotype-phenotype maps from multiplex assays of variant effect Jupyter Notebook 24 5

logomaker Public Software for the visualization of sequence-function relationships Jupyter Notebook 178 35

Justin B. Kinney jbkinney

Edit profile

36 followers · 1 following

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

61 contributions in the last year Contribution settings ▾ 2024

Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul 2023

Mon Wed Fri 2022

Learn how we count contributions Less More 2021

2020

2019

2018

Contribution activity

jbkinney/24e_qbbootcamp: R x +

https://github.com/jbkinney/24e_qbbootcamp

New Chrome available :

Toggl Overleaf SmartSheet Papers Github ChatGPT CSHL: Lawson Supergroup Sched... Mimecast eRA eRA Commons Concur

jbkinney / 24e_qbbootcamp Type ⌈ to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

24e_qbbootcamp Public Unpin Unwatch 1 Fork 0 Star 0

main 1 Branch 0 Tags

jbkinney Update README.md

- bash
- lectures
- python
- .gitignore
- 23e_qbbootcamp_syllabus.pdf
- LICENSE
- README.md

Clone

Local Codespaces

HTTPS SSH GitHub CLI

https://github.com/jbkinney/24e_qbbootcamp.

Clone using the web URL.

Open with GitHub Desktop

Download ZIP

15 minutes ago

About

Repository for the 2024 QB Bootcamp

- 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

2024 Quantitative Biology Bootcamp

Utilities

Back/Forward View Group Action Share Edit Tags

Dropbox Search

Favorites

- Dropbox
- 20_qbcourse_db
- 20_qbbootcamp
- testing
- logo
- mavenn
- mavenn_local
- NEEDS FEEDBACK
- READ ME
- 20a_bistatistician
- 15_mpathic
- 18_splicing
- 18_gauge
- 15_diagrams
- Desktop
- github
- Recents
- Downloads
- jkinney
- rpod

Applications

- Documents

iCloud

- iCloud Drive

Locations

- rpod

Remote Disc

QuickTime Player.app R.app ReadCube Port Utility.app Reminders.app Remote Desk...nnection.app ReplicatorG.app RStudio.app Safari.app SimCity™4 D...e Edition.app Siri.app SketchUp 2016 Skim.app Skype.app Slack.app SnapGene.app SnippetsLab.app Steam.app Stickies.app Stocks.app Sublime Text.app System Preferences.app Termius.app TeX Texpad.appTextEdit.app TextMate.app The Unarchiver.app Things.app Time Machine.app TogglDesktop.app Utilities

- VMD 1.9.3.app
- Voice Memos.app
- Wacom Utility
- Webex Productivity Tools
- WriteRoom.app
- XAMPP
- Xcode.app
- zoom.us.app

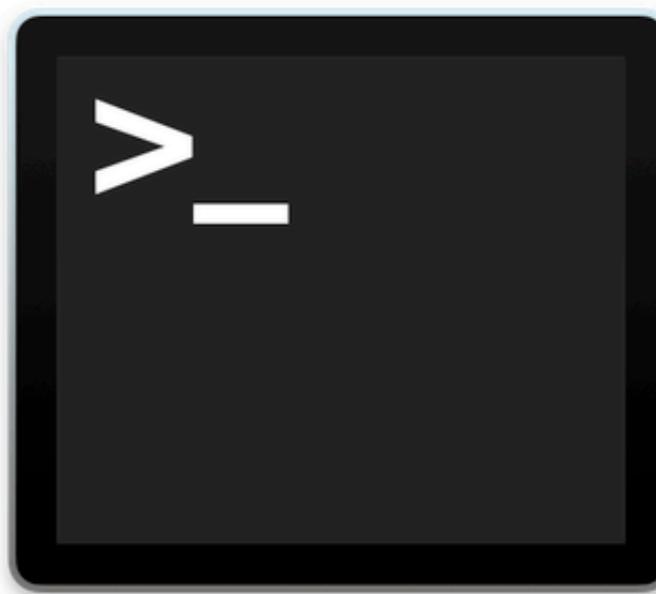
Activity Monitor.app Adobe AIR A...n Installer.app Adobe AIR Uninstaller.app Adobe Application Manager ► Adobe Flash...l Manager.app Adobe Installers ► Adobe Utilities - CS6 ► AirPort Utility.app Audio MIDI Setup.app Bluetooth Fil...xchange.app Boot Camp Assistant.app ColorSync Utility.app Console.app Digital Color Meter.app Disk Utility.app Grapher.app Keychain Access.app Migration Assistant.app Screenshot.app Script Editor.app System Information.app Terminal.app VoiceOver Utility.app XQuartz.app

Terminal.app

Application - 10.1 MB

Tags Add Tags...
Created Friday, August 17, 2018 at 8:55 PM
Modified Thursday, August 15, 2019 at 7:19 AM
Version 2.9.5

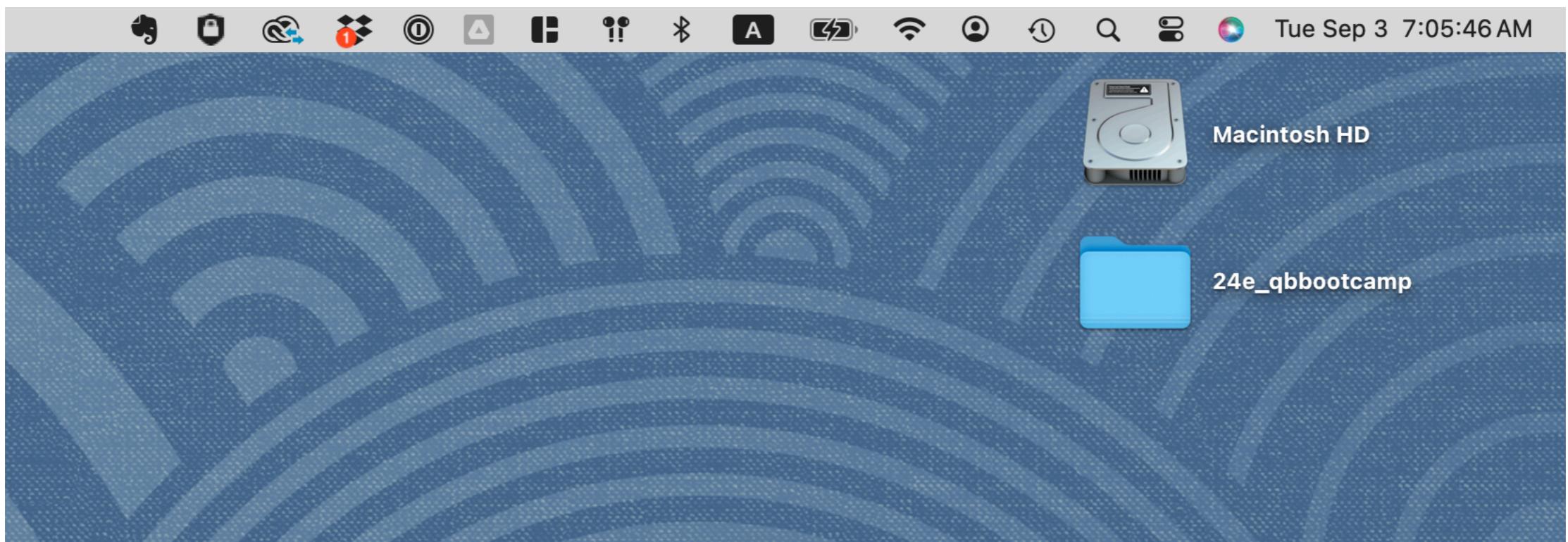
More...

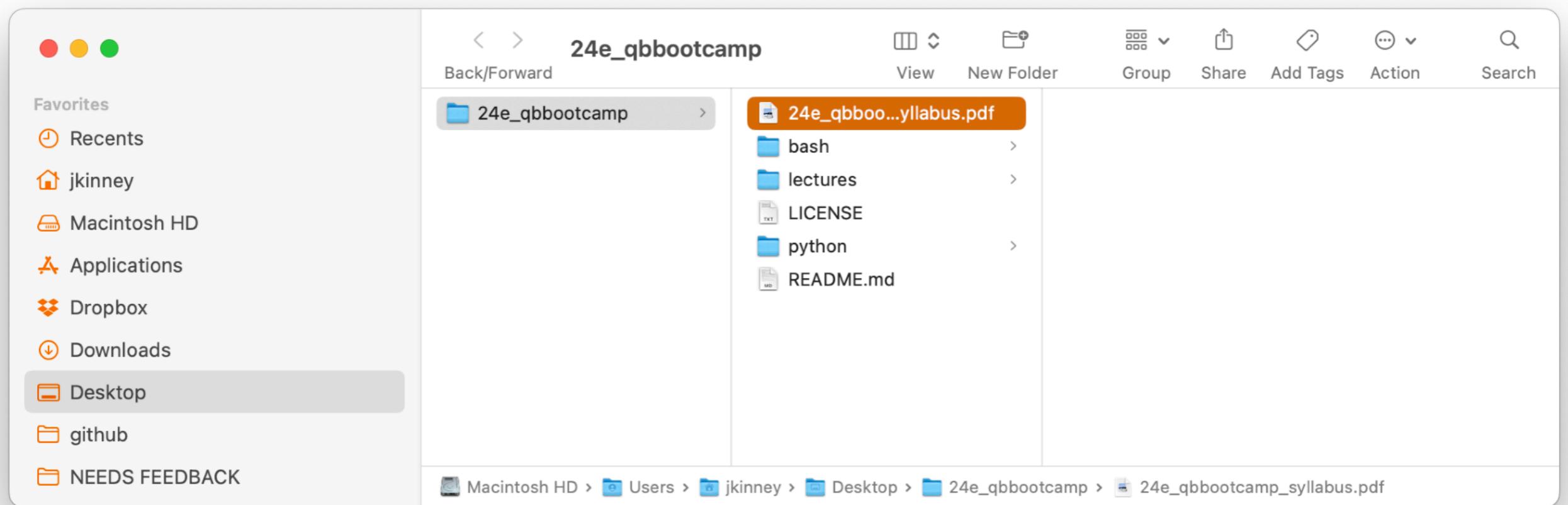


rpod > Applications > Utilities > Terminal.app

Desktop — jkinney@bamdev1:/grid/kinney/home — -bash — 99x15

```
[base) jkinney@MJM-2023:~/Desktop$ git clone https://github.com/jbkinney/24e_qbbootcamp.git
Cloning into '24e_qbbootcamp'...
remote: Enumerating objects: 49, done.
remote: Counting objects: 100% (12/12), done.
remote: Compressing objects: 100% (11/11), done.
remote: Total 49 (delta 5), reused 2 (delta 0), pack-reused 37 (from 1)
Receiving objects: 100% (49/49), 75.97 MiB | 14.24 MiB/s, done.
Resolving deltas: 100% (17/17), done.
(base) jkinney@MJM-2023:~/Desktop$
```





What is Quantitative Biology?

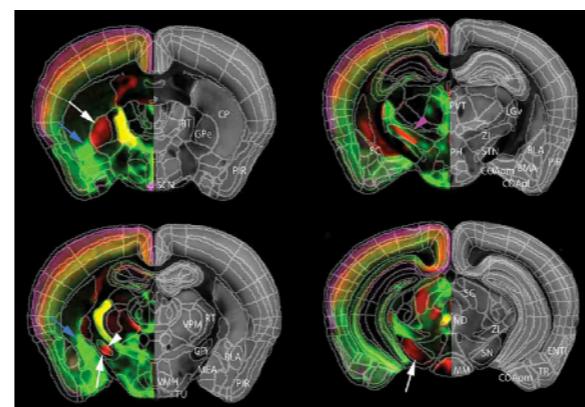
Quantitative biology is a vast field

Genomics



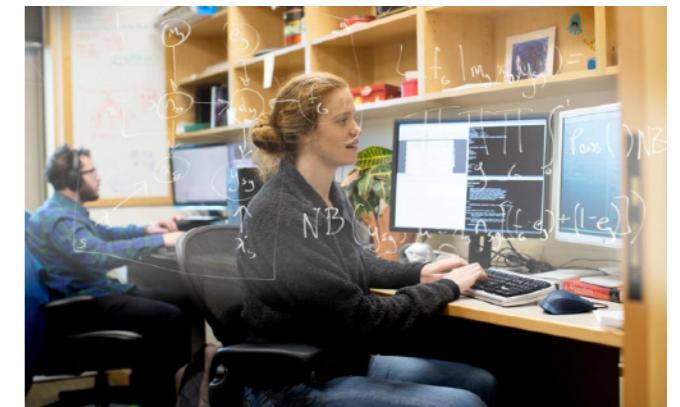
Functional genomics
Evolutionary genomics
Genome dynamics
Technology development

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



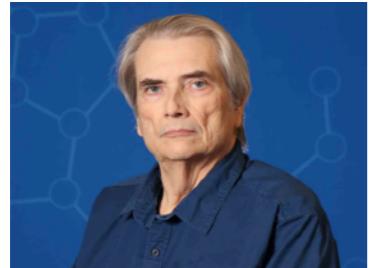
Adam
Siepel



Michael
Wigler

QB Associated Faculty

Genomics



**Richard
McCombie**



**Tom
Gingeras**



**Doreen
Ware**

Neuroscience



**Ben
Cowley**



**David
Klindt**



**Alexei
Koulakov**



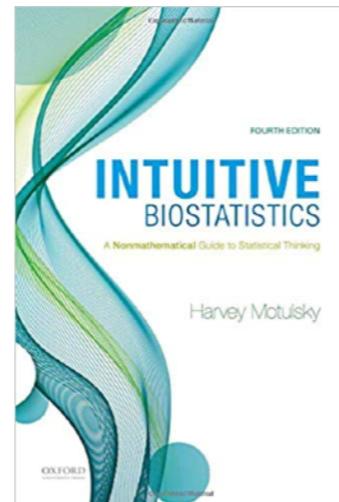
**Partha
Mitra**

What QB skills should all biology researchers have?

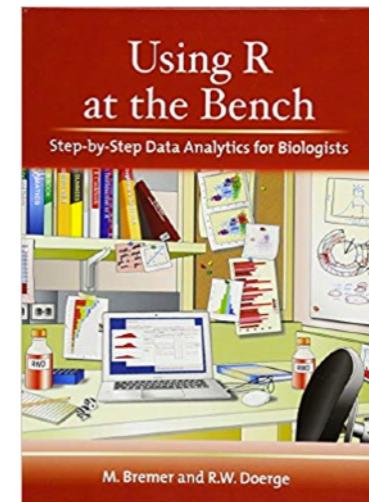
Learn to interpret standard statistics

Key statistical concepts:

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

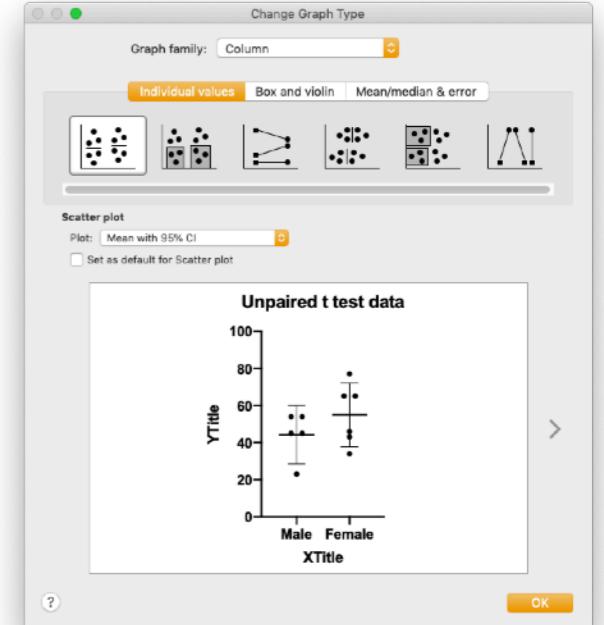
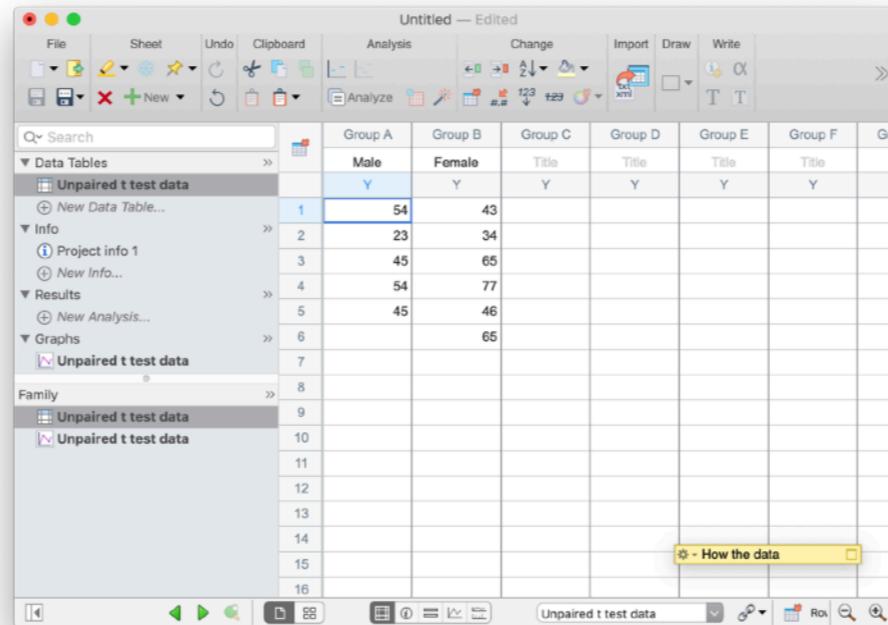
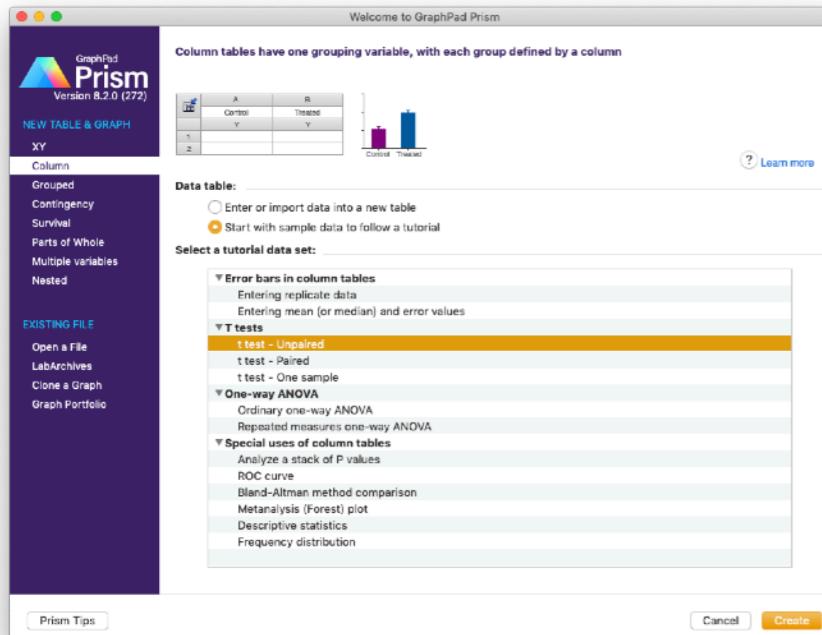


Intuitive Biostatistics, 4th ed
Motulsky (2018)
[CEO, GraphPad Software]



Using R at the Bench
Bremmer & Doerge (2015)

Learn to compute standard statistics



Alternatively:



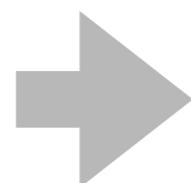
Learn to navigate UNIX systems



Sequencer



Microscope



High Performance
Computer Cluster

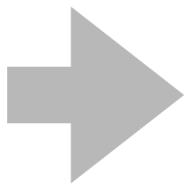
A screenshot of a macOS terminal window showing a UNIX command line session. The user 'jkinney' is logged in via SSH to a host named 'bnbdev2'. The terminal shows the output of the 'ls' command, listing various directory names and files.

UNIX command line

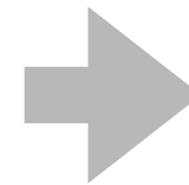
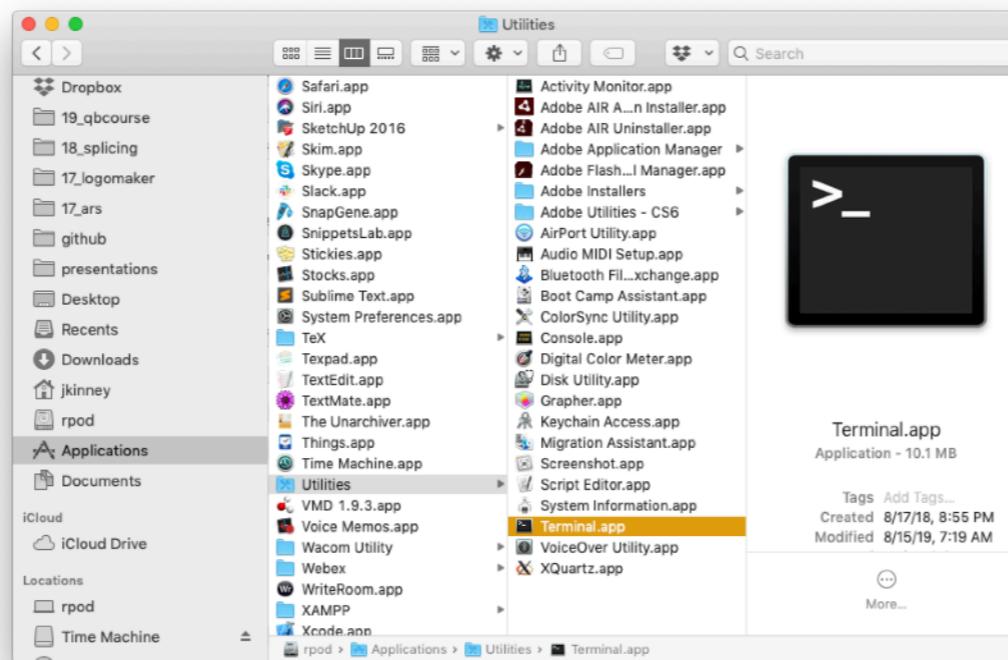


Databases

Mac OS X is based on UNIX



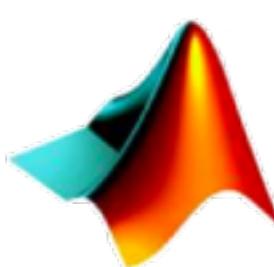
Finder

A screenshot of a Terminal window. The title bar says "jkinnery — bash — 54x20". The content shows a command line session:

```
Last login: Wed Aug 21 14:49:14 on ttys001
jkinnery@rpod:~$
```

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 '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 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 it 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.' Another note encourages users to acknowledge the CSHL Bioinformatics Shared Resource. The right sidebar shows an 'History' section titled 'Unnamed history (empty)' with 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](#)]
- [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 two main panes. The left pane displays the LaTeX source code for a document titled "Biophysical models of cis-regulation as interpretable neural networks". The right pane shows the rendered PDF output.

Left Pane (Code View):

```
19_mclb.tex
1 \usepackage{utf8}[inputenc] % allow utf-8 input
2 \usepackage[T1]{fontenc} % use 8-bit T1 fonts
3 \usepackage{hyperref} % hyperlinks
4 \usepackage{url} % simple URL typesetting
5 \usepackage{booktabs} % professional-quality tables
6 \usepackage{amsfonts} % blackboard math symbols
7 \usepackage{nicefrac} % compact symbols for 1/2, etc.
8 \usepackage{microtype} % microtypography
9 \usepackage{soul} % for \ul
10 \usepackage{graphicx} % for figures
11 \usepackage{upgreek}
12
13 \title{Biophysical models of cis-regulation as\\ interpretable neural networks}
14
15
16
17 \author{%
18   Ammar Tareen \\
19   Simons Center for Quantitative Biology \\
20   Cold Spring Harbor Laboratory \\
21   Cold Spring Harbor, NY 11724 \\
22   \texttt{tareen@cshl.edu} \\
23
24   And \\
25   Justin B. Kinney \\
26   Simons Center for Quantitative Biology \\
27   Cold Spring Harbor Laboratory \\
28   Cold Spring Harbor, NY 11724 \\
29   \texttt{jkinney@cshl.edu} \\
30
31 \begin{document}
32
33 \maketitle
34
35 \begin{abstract}
36
37 Biophysical models that describe gene regulation, as well as other cis-regulatory processes, can be
38 formulated as deep neural networks. This is true of quasi-equilibrium (a.k.a.\ thermodynamic)
39 models as well as non-equilibrium (a.k.a.\ kinetic) models. This observation suggests new ways of
40 using powerful deep learning frameworks for training biophysically interpretable neural networks
41 using data produced by massively parallel reporter assays (MPRAs). We demonstrate this
42 }
```

Right Pane (Preview):

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 MPRA 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)

**Master all of
these topics**

Intermediate material

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

**Master at least one
of these topics**

Advanced material

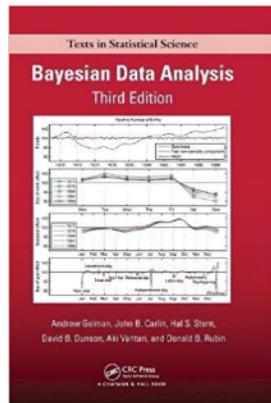
Molecular biophysics
Stochastic processes
Dynamical systems
Information theory
...

**Learn selected
topics as needed**

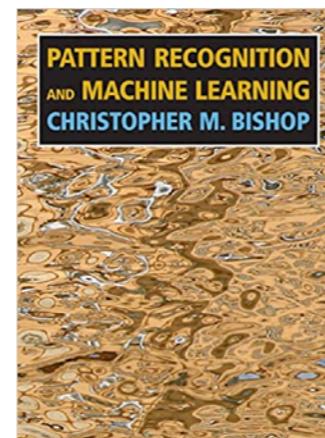
Learn to work through technical books systematically and independently

Mentored independent study in QB:

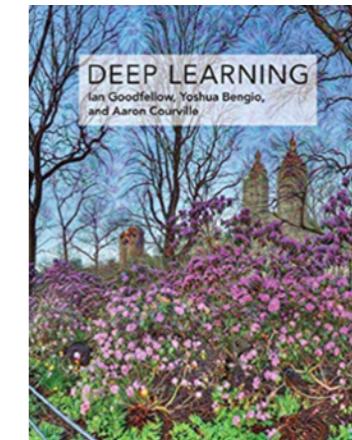
Faculty will help interested students pursue directed reading of graduate-level material.
Email me <jkinney@cshl.edu> if interested.



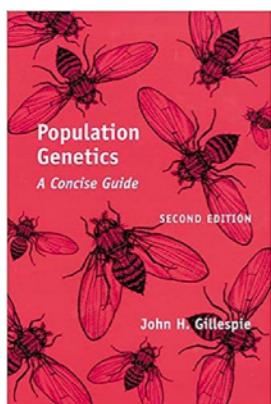
Bayesian Data Analysis, 3rd ed
Gelman et al., 2013



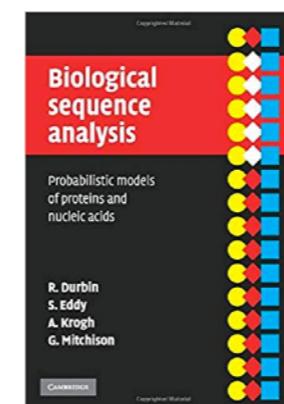
**Pattern Recognition and
Machine Learning**
Bishop, 2006



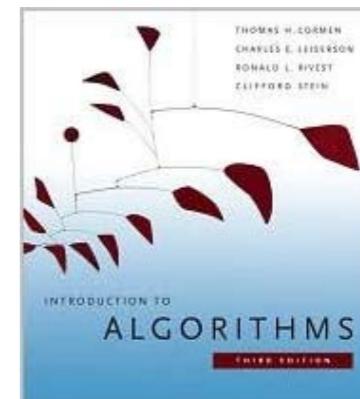
Deep Learning
Goodfellow et al., 2016



**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.