

Welcome to Quantitative Biology



QB Bootcamp, Day 1
Wednesday, 1 September 2021
10:00am - 10:30am

2021 QB Bootcamp Schedule

Day 1: Wednesday, September 1, Hershey East (Hershey), 10am - 5pm

10:00am - 10:30am: **Overview of Quantitative Biology (lecture)**

10:30am - 12:00pm: **The Unix command line (tutorial)**

12:00pm - 1:00pm: *Lunch (provided)*

1:00pm - 1:30pm: **Introduction to Python and Jupyter Notebooks (tutorial)**

1:30pm - 3:00pm: **Python: data types (tutorial)**

3:00pm - 3:30pm: *Break*

3:30pm - 5:00pm: **Python: flow control (tutorial)**

Day 2: Thursday, September 2, Hershey East (Hershey), 10am - 5pm

10:00am - 10:30am: **Overview of High-Performance Computing (lecture)**

10:30am - 12:00pm: **Read mapping using Elzar (tutorial)**

12:00pm - 1:00pm: *Lunch (provided)*

1:00pm - 1:30pm: **Introduction to Pandas (lecture)**

1:30pm - 3:00pm: **Pandas I, TF analysis (tutorial)**

3:00pm - 3:30pm: *Break*

3:30pm - 5:00pm: **Pandas II, Replication origin analysis (tutorial)**

Day 3: Friday, September 3, Hershey East (Hershey), 2pm - 6pm

2:00pm - 2:30pm: **Introduction to Data Visualization (tutorial)**

2:30pm - 4:00pm: **Matplotlib (tutorial)**

4:00pm - 4:30pm: *Break*

4:30pm - 6:00pm: **Seaborn (tutorial)**

A screenshot of a web browser showing the GitHub repository [jbkinney/21_qbbootcamp](https://github.com/jbkinney/21_qbbootcamp). The repository page is displayed, featuring a list of files and a detailed sidebar on the right.

The sidebar on the right contains the following sections:

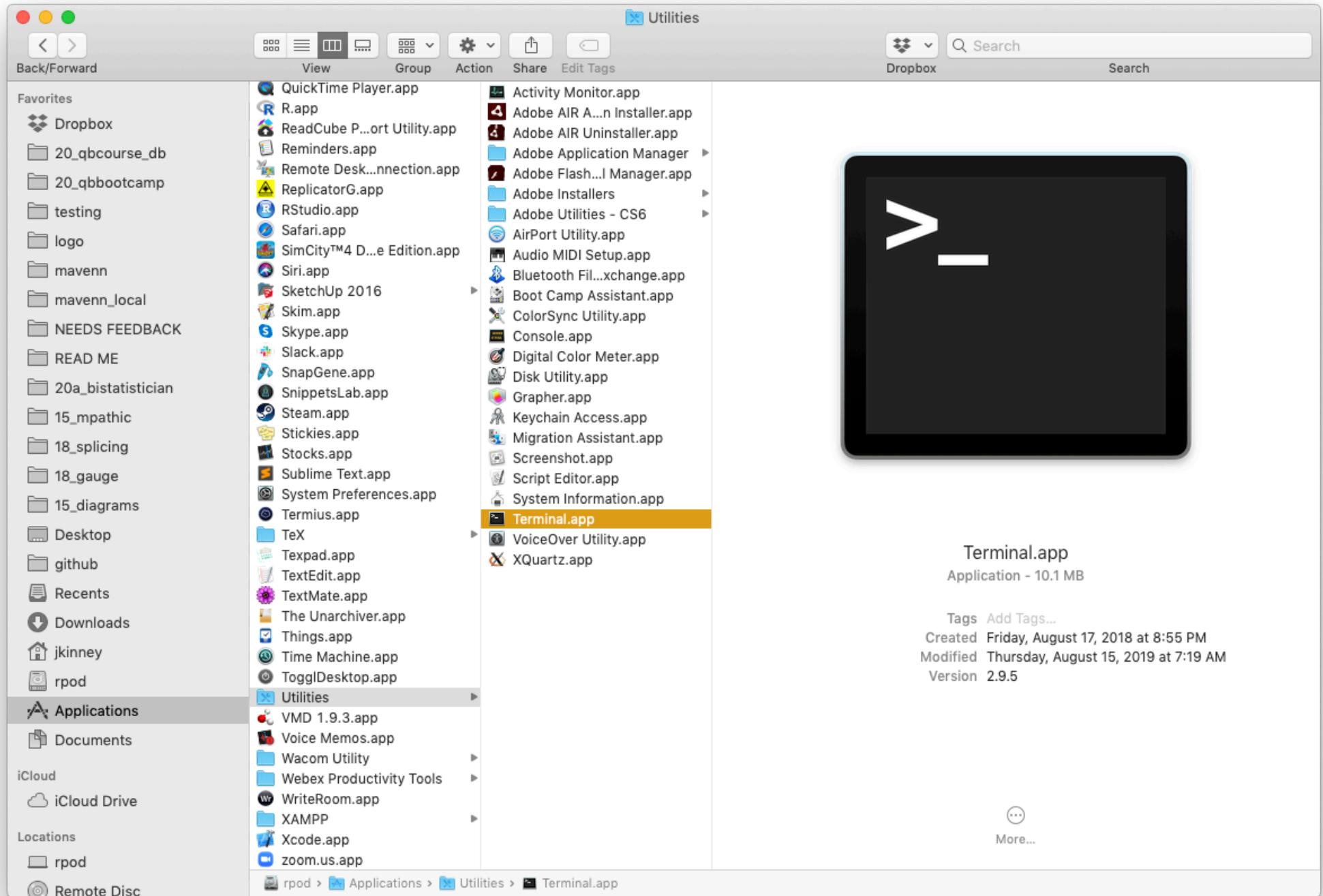
- About**: No description, website, or topics provided.
- Readme**
- Releases**: No releases published. [Create a new release](#)
- Packages**: No packages published. [Publish your first package](#)

The main content area shows the repository structure and commit history:

- Code** tab selected.
- Branches**: main (selected), 1 branch, 0 tags.
- Commits** (list):
 - jbkinney readme file is modified (initial commit)
 - bash initial commit
 - python initial commit
 - .gitignore changed name of ex...
 - 21_qbbootcamp_syllabus.pdf cleaned up dir
 - 21_qbcourse_syllabus.pdf cleaned up dir
 - README.md Update README.md
 - elzar_exercise.tar.gz readme file is modified 16 hours ago
- README.md** file preview.

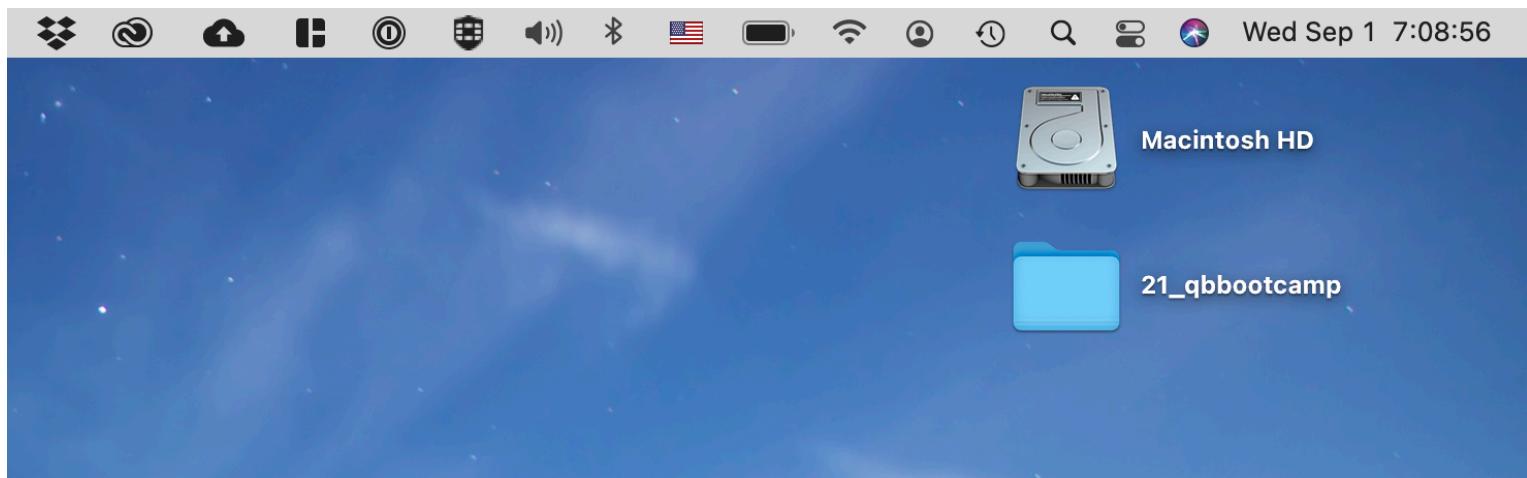
A green box highlights the "Code" dropdown menu in the sidebar, which includes options for "Go to file", "Add file", and "Code". Below this, the "Clone" section is also highlighted.

```
graph TD; A[Code] --> B[Go to file]; A --> C[Add file]; A --> D[Code]; D --> E[Clone]; E --> F[HTTPS]; E --> G[SSH]; E --> H[GitHub CLI]; E --> I[Link]; E --> J[Open with GitHub Desktop]; E --> K[Download ZIP]
```



Desktop — bash — 91x14

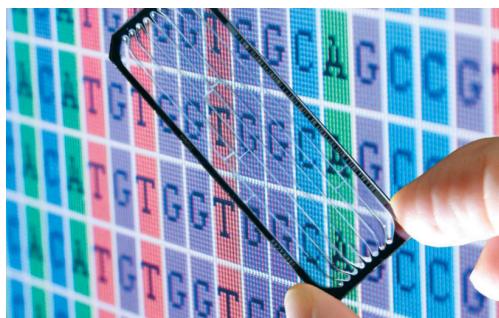
```
[base) jkinney@u1:~/Desktop/
(base) jkinney@u1:~/Desktop$ git clone https://github.com/jbkinney/21_qbbootcamp.git
objc[1732]: Class AMSupportURLConnectionDelegate is implemented in both ?? (0x1f6e378f0) and ?? (0x11661c2b8). One of the two will be used. Which one is undefined.
objc[1732]: Class AMSupportURLSession is implemented in both ?? (0x1f6e37940) and ?? (0x11661c308). One of the two will be used. Which one is undefined.
Cloning into '21_qbbootcamp'...
remote: Enumerating objects: 70, done.
remote: Counting objects: 100% (70/70), done.
remote: Compressing objects: 100% (55/55), done.
remote: Total 70 (delta 22), reused 59 (delta 14), pack-reused 0
Receiving objects: 100% (70/70), 295.84 MiB | 6.84 MiB/s, done.
Resolving deltas: 100% (22/22), done.
(base) jkinney@u1:~/Desktop$
```



What is Quantitative Biology?

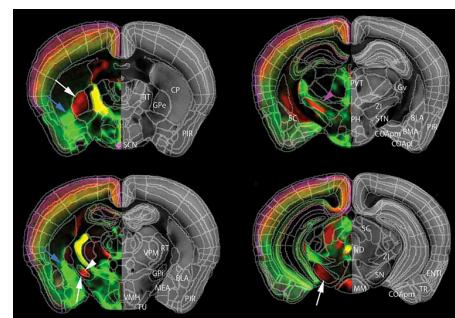
Quantitative biology is a vast field

Genomics



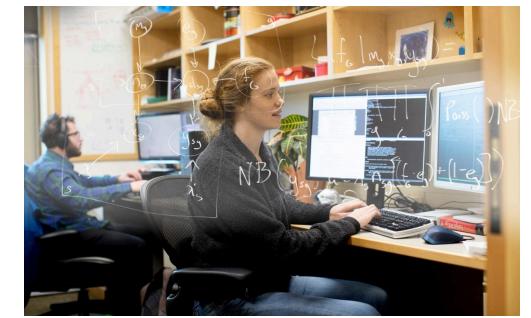
Functional genomics
Evolutionary genomics
Genome dynamics
Technology development

Neuroscience



Data analysis
Modeling neural systems
Behavioral modeling

Other



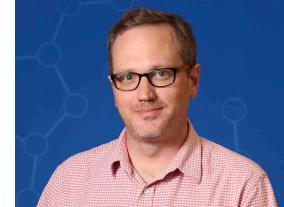
Biophysics
Machine learning
Software development

Who does Quantitative Biology at CSHL?

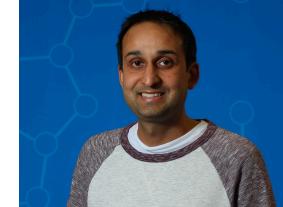
Core QB program



**Molly Gale
Hammell**



**Dan
Levy**



**Saket
Navlakah**



**Ivan
Iossofov**



**David
McCandlish**



**Peter
Koo**



**Justin
Kinney**



**Hannah
Meyer**



**Alexander
Krasnitz**



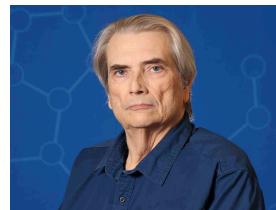
**Adam
Siepel**

QB Associated Faculty

Genomics



Alexander
Dobin



Richard
McCombie

Neuroscience



Tatiana
Engel



Jesse
Gillis



Doreen
Ware



Alexei
Koulakov



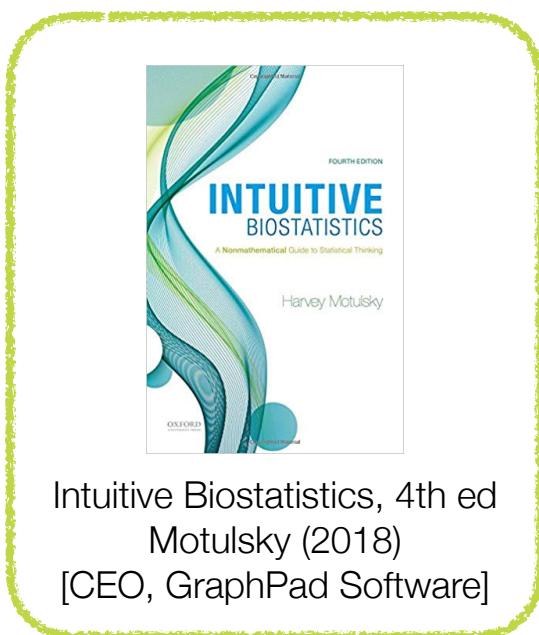
Partha
Mitra

What QB skills should all biology researchers have?

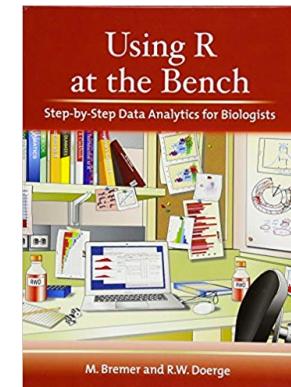
Learn to interpret standard statistics

Key statistical concepts:

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

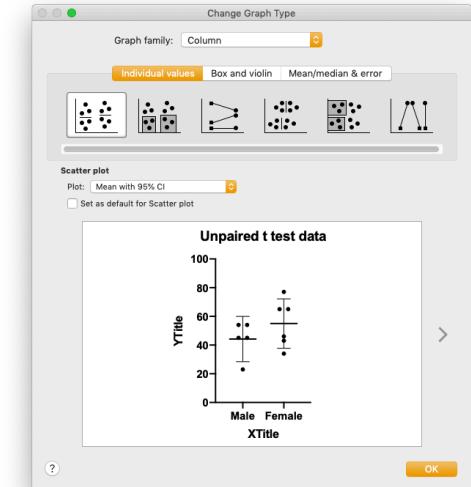
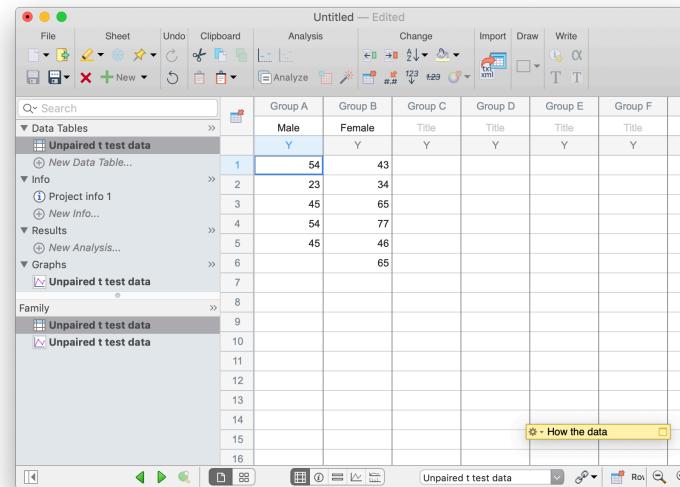
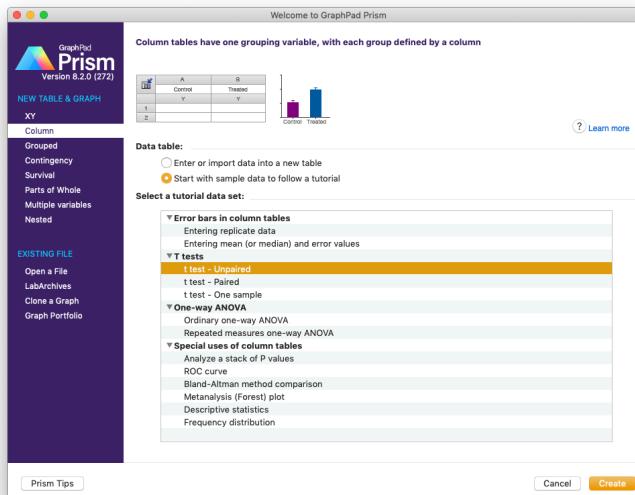


Intuitive Biostatistics, 4th ed
Motulsky (2018)
[CEO, GraphPad Software]



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 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



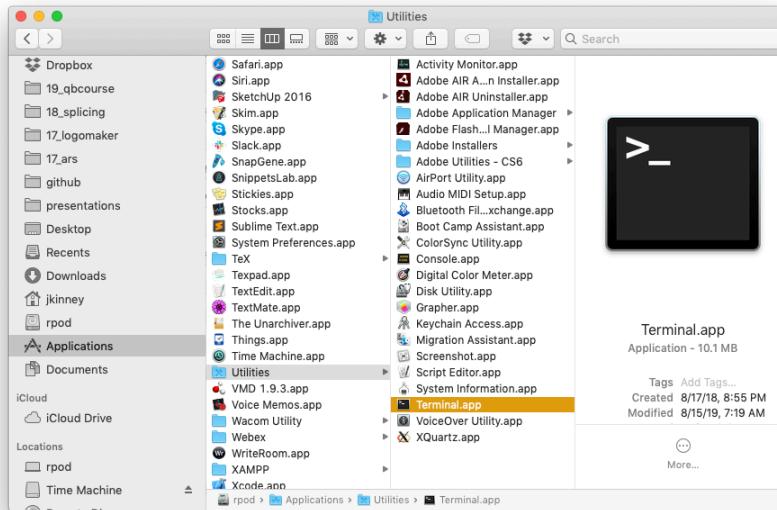
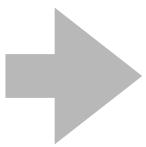
Databases



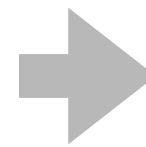
Mac OS X is based on UNIX



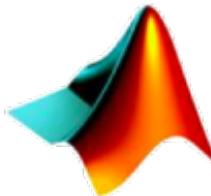
Finder



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 statistics- clunky graphics
	<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 and by old people	<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 Galaxy web interface with the following details:

- Header:** Galaxy / CSHL BSR. Navigation links include Analyze Data, Workflow, Shared Data, Visualization, Help, Login or Register, and a grid icon.
- Left Sidebar (Tools):**
 - Search bar: search tools
 - CSHL TOOLS:**
 - [Get Data](#)
 - [Quality Control](#)
 - UTILITIES:**
 - [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](#)
- Main Content Area:**

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

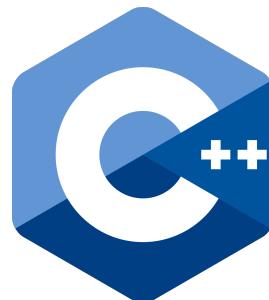
 - Goecks J., Nekrutenko A., Taylor J. and The Galaxy Team. (2010) [Galaxy: a comprehensive approach for supporting reproducible computational workflows](#).
- Right Sidebar (History):**
 - Search bar: search datasets
 - Unnamed history** (empty)
 - This history is empty. You can [load your own data](#) 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 file `19_mclb.tex` open. The left pane displays the LaTeX code, and the right pane shows the rendered document.

Left Pane (Code View):

```
19_mclb.tex
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

```

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 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
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**

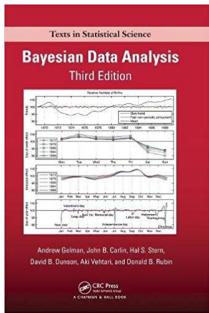
**Master at least one
of these topics**

**Learn selected
topics as needed**

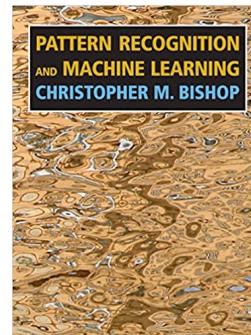
Learn to work through technical books systematically and independently

Mentored independent study in QB:

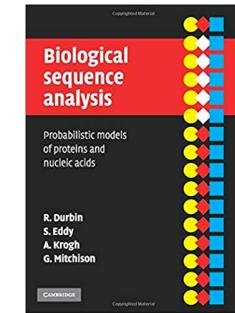
Faculty will help interested students pursue directed reading of graduate-level material.
Email me <jkinney@cshl.edu> if interested.



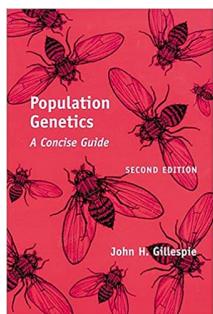
Bayesian Data Analysis, 3rd ed
Gelman et al., 2013



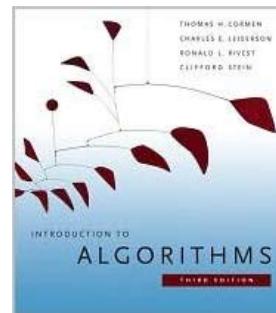
**Pattern Recognition and
Machine Learning**
Bishop, 2006



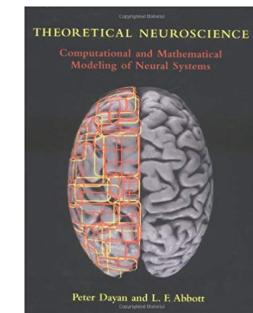
Biological Sequence Analysis
Durbin et al., 1998



**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 Peter Koo to get on mailing list.