Welcome to Quantitative Biology

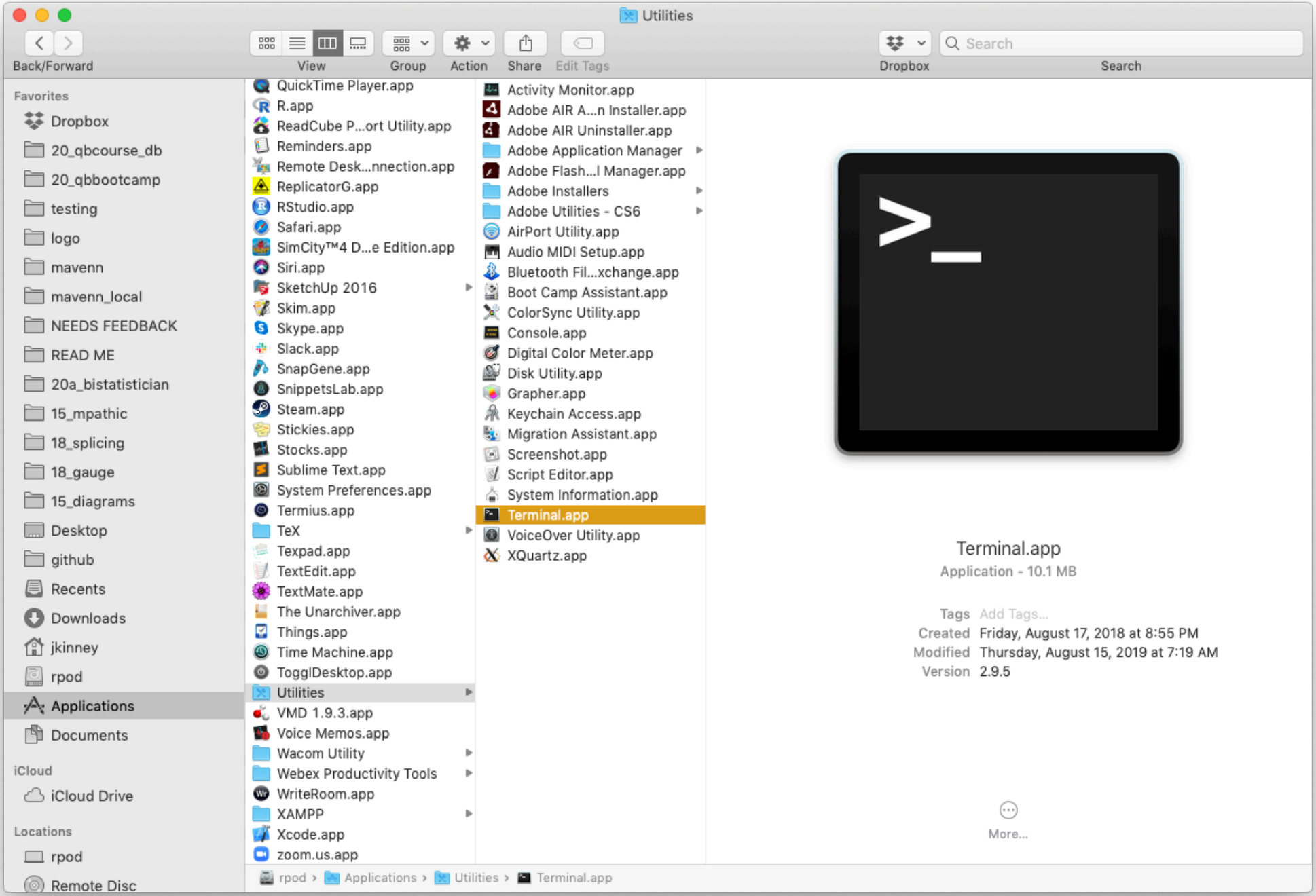
QB Bootcamp, Day 1
Wednesday, 9 September 2020
10:00am - 10:30am

2019 QB Bootcamp Schedule

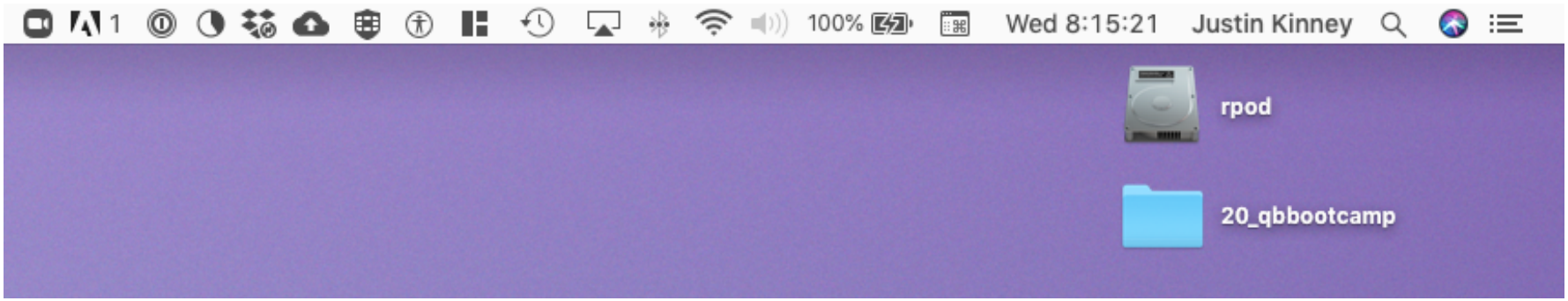
Day 1: Wednesday, September 9, Hershey East (Hershey), 10am - 5pm
10:00am - 10:30am: **Overview of Quantitative Biology (lecture)**
10:30am - 12:00pm: **The Unix command line (tutorial)**
12:00pm - 1:00pm: *Lunch (provided)*
1:00pm - 1:30pm: **Introduction to Python and Jupyter Notebooks (tutorial)**
1:30pm - 3:00pm: **Python: data types (tutorial)**
3:00pm - 3:30pm: *Break*
3:30pm - 5:00pm: **Python: flow control (tutorial)**

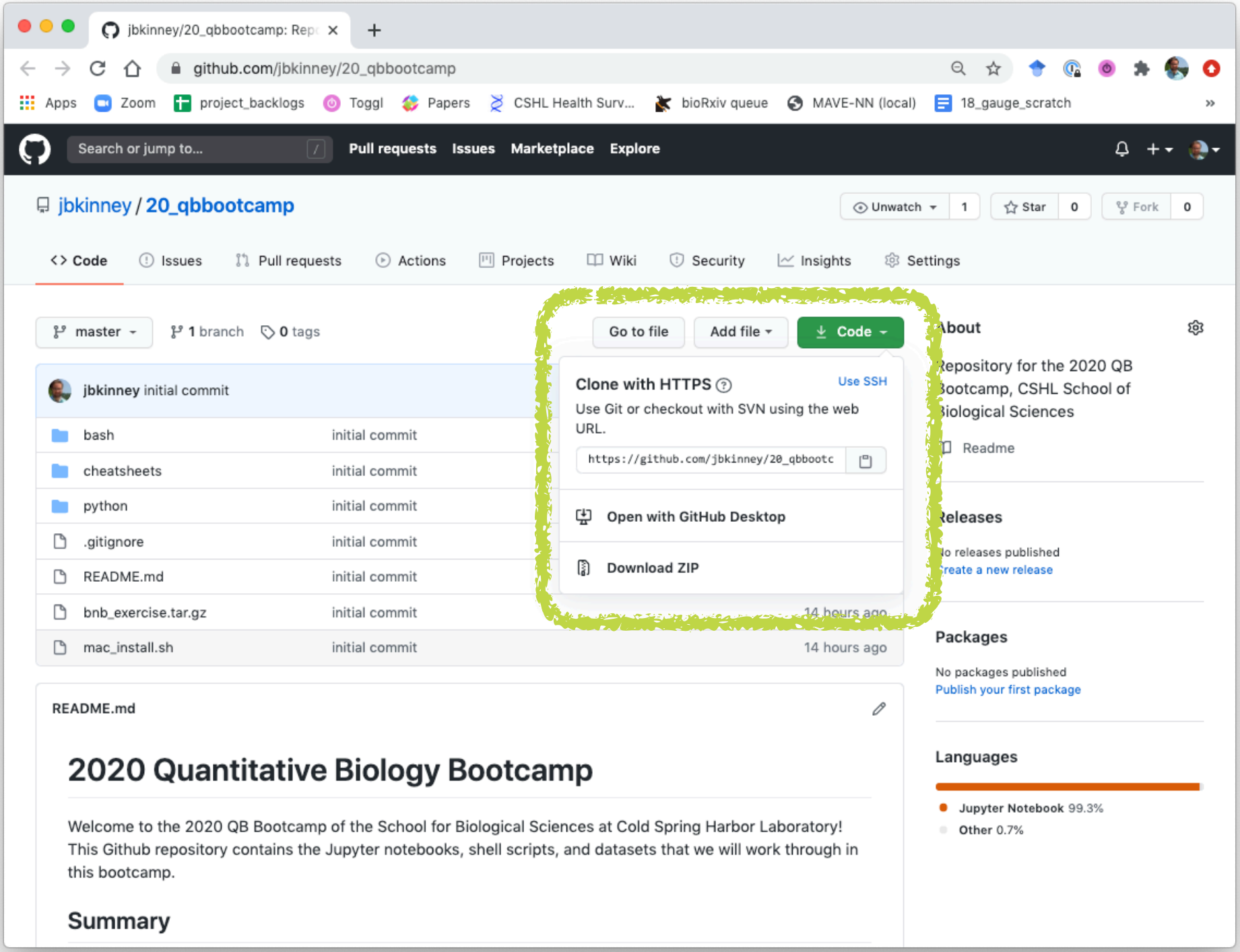
Day 2: Thursday, September 10, Hershey East (Hershey), 10am - 5pm
10:00am - 10:30am: **Overview of High-Performance Computing (lecture)**
10:30am - 12:00pm: **Read mapping using BlackNBlue (tutorial)**
12:00pm - 1:00pm: *Lunch (provided)*
1:00pm - 1:30pm: **Introduction to Pandas (lecture)**
1:30pm - 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 11, 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)**


Utilities
Back/Forward
View
Group
Action
Share
Edit Tags
Dropbox
Search
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 P...ort 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.app
TextEdit.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

```
(base) jkinney@rpod:~$ cd Desktop
(base) jkinney@rpod:~/Desktop$ git clone https://github.com/jbkinney/20_qbbootcamp.git
Cloning into '20_qbbootcamp'...
remote: Enumerating objects: 11, done.
remote: Counting objects: 100% (11/11), done.
remote: Compressing objects: 100% (11/11), done.
remote: Total 62 (delta 0), reused 7 (delta 0), pack-reused 51
Unpacking objects: 100% (62/62), done.
Checking out files: 100% (53/53), done.
(base) jkinney@rpod:~/Desktop$
```

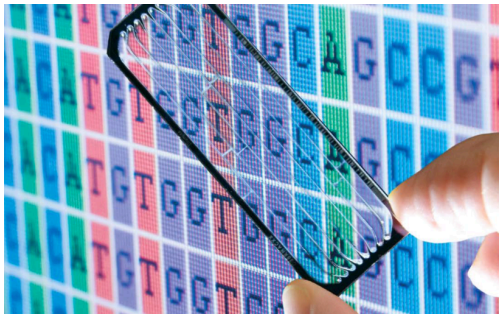



jbkinney/20_qbbootcamp: Rep
github.com/jbkinney/20_qbbootcamp
Apps
Zoom
project_backlogs
Toggl
Papers
CSHL Health Surv...
bioRxiv queue
MAVE-NN (local)
18_gauge_scratch
Search or jump to...
Pull requests
Issues
Marketplace
Explore
jbkinney / 20_qbbootcamp
Unwatch 1
Star 0
Fork 0
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
master
1 branch
0 tags
Go to file
Add file
Code
Clone with HTTPS
Use SSH
Use Git or checkout with SVN using the web URL.
https://github.com/jbkinney/20_qbbootc
Open with GitHub Desktop
Download ZIP
jbkinney initial commit
bash initial commit
cheatsheets initial commit
python initial commit
.gitignore initial commit
README.md initial commit
bnb_exercise.tar.gz initial commit 14 hours ago
mac_install.sh initial commit 14 hours ago
About
Repository for the 2020 QB Bootcamp, CSHL School of Biological Sciences
Readme
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
Languages
Jupyter Notebook 99.3%
Other 0.7%
README.md
2020 Quantitative Biology Bootcamp
Welcome to the 2020 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.
Summary

What is Quantitative Biology?

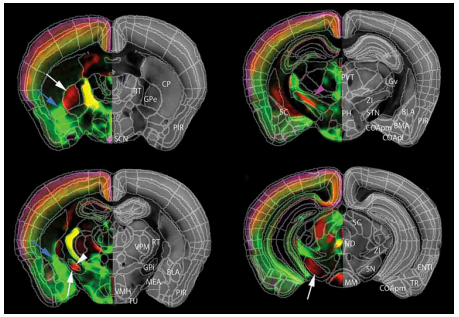
Quantitative biology is a vast field

Genomics



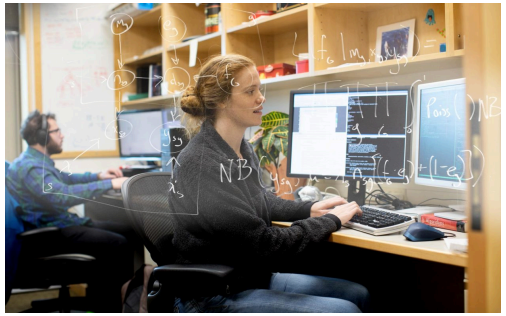
Functional genomics
Evolutionary genomics
Genome dynamics
Technology development

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

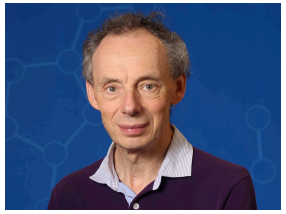
Ivan Iossofov

David McCandlish

Justin Kinney

Hannah Meyer

Peter Koo



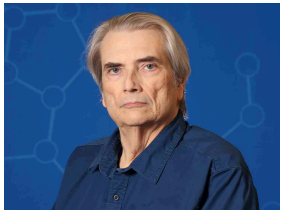
Alexander Krasnitz

Adam Siepel

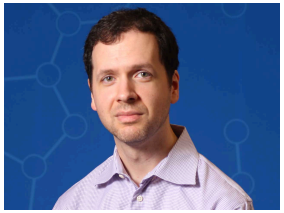
QB Associated Faculty

Genomics

Alexander Dobin



Richard McCombie



Jesse Gillis

Doreen Ware

Neuroscience

Tatiana Engel

Alexei Koulakov

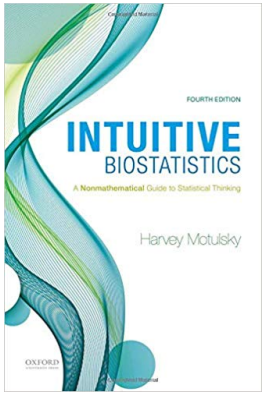
Partha Mitra

What QB skills should all biology researchers have?

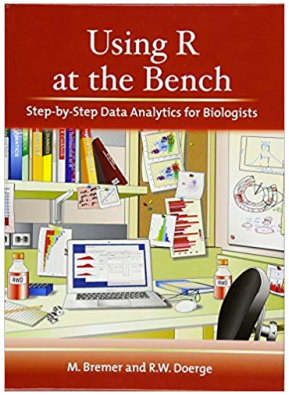
Learn to interpret standard statistics

Key statistical concepts:

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



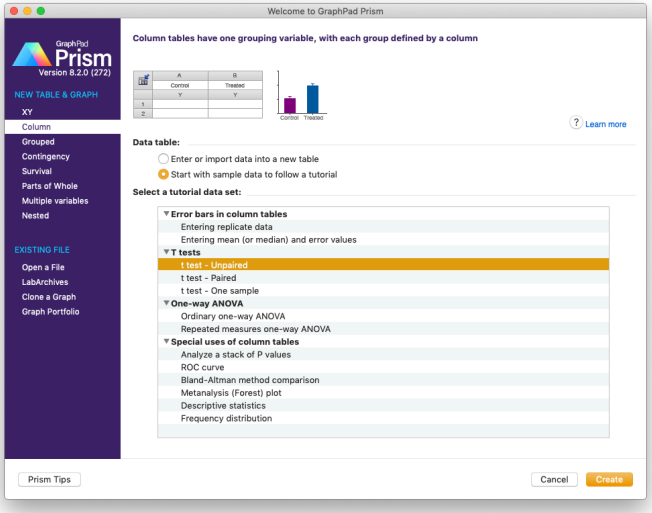
Intuitive Biostatistics, 4th ed
Motulsky (2018)
[CEO, GraphPad Software]

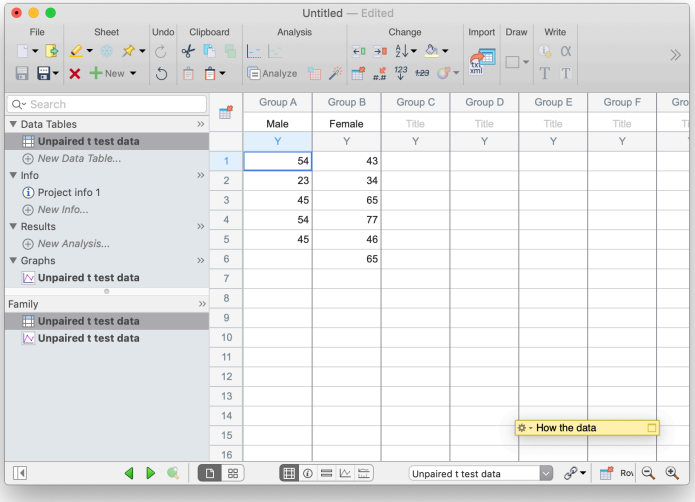


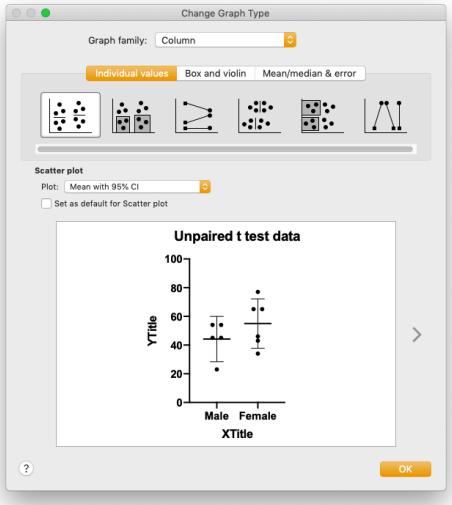
Using R at the Bench
Bremmer & Doerge (2015)

Learn to compute standard statistics

Prism 8







Alternatively:

python™

R

MATLAB

Excel

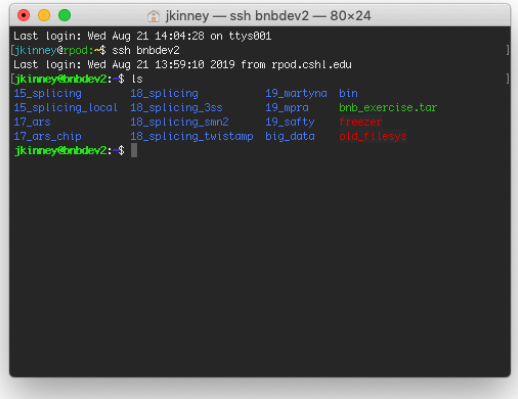
Learn to navigate UNIX systems

Sequencer

Microscope

Databases

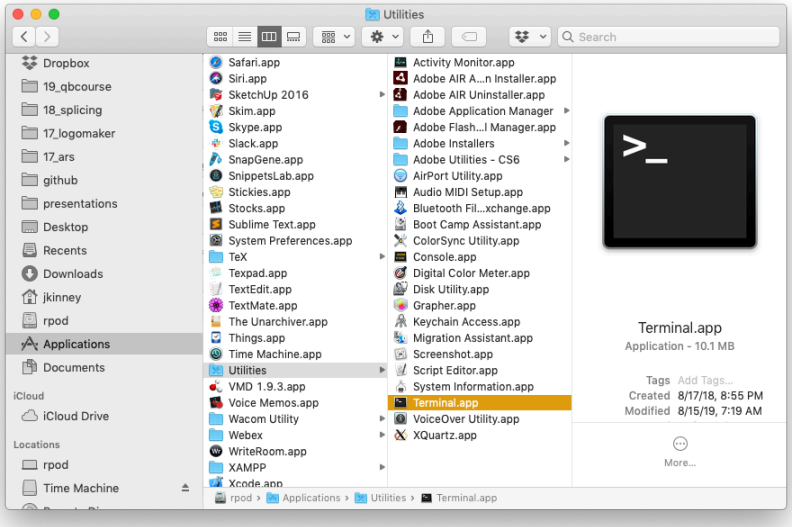
High Performance
Computer Cluster

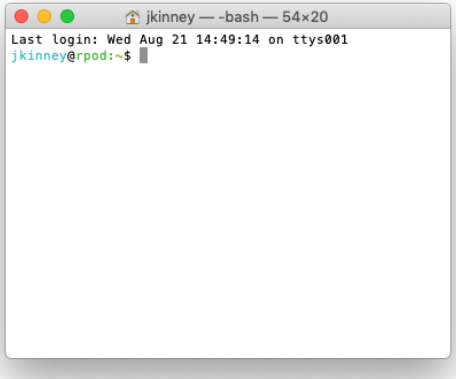


UNIX command line

Mac OS X is based on UNIX

Finder



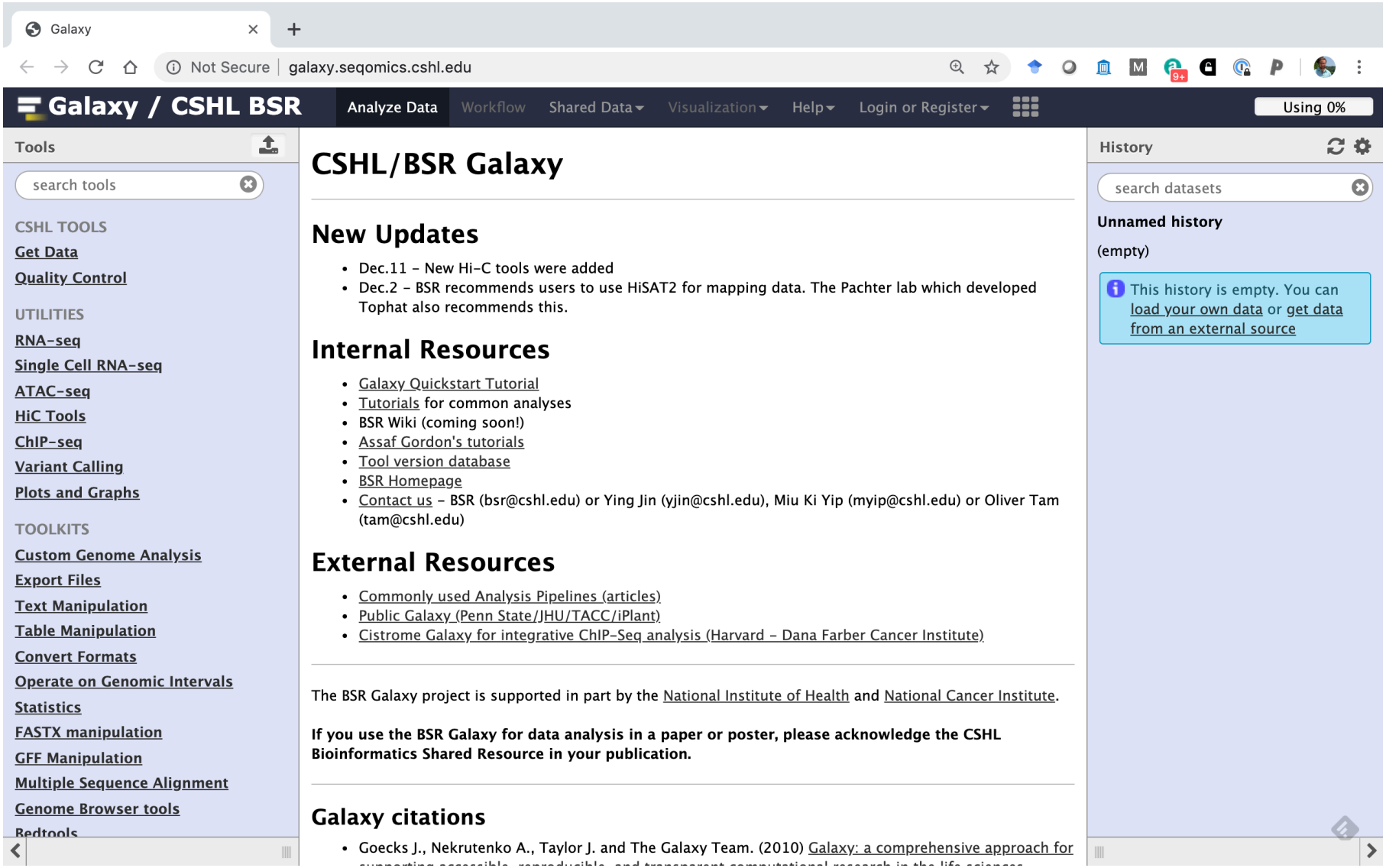


Applications/Utilities/Terminal.app

Become familiar with at least one programming language

language	strengths	weaknesses
	- elegant language - easy to learn - flexibility: use for large pipelines or local data analysis - highly valued skill in industry - **primary language for deep learning**	- clunky dataframes - clunky statistics - clunky graphics
	- streamlined for statistics - highly developed for genomics - great graphics	- strange language - not great for building pipelines
	- used heavily in neuroscience and by old people	- proprietary - poorly supported - bad graphics - bad for strings

Learn to analyze your own sequencing data



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

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)

PCWorld FROM IDG

BEST ONLINE COURSES FOR PYTHON AT A GLANCE

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

- 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]*

https://www.pcworld.com/article/3287981/best-python-courses.html

Learn to use LaTeX

19_mclb.tex

Typeset | MANUAL | pdfLaTeX + Bibliography | Configuration | Share | Outline | View | Editor | New Tab

19_mclb.tex

```
\usepackage[utf8]{inputenc} % allow utf-8 input
\usepackage[T1]{fontenc}    % use 8-bit T1 fonts
\usepackage{hyperref}       % hyperlinks
\usepackage{url}            % simple URL typesetting
\usepackage{booktabs}       % professional-quality tables
\usepackage{amsfonts}       % blackboard math symbols
\usepackage{nicefrac}       % compact symbols for 1/2, etc.
\usepackage{microtype}      % microtypography
\usepackage{soul} % for \ul
\usepackage{graphicx} % for figures
\usepackage{upgreek}

\title{Biophysical models of cis-regulation as\\ interpretable neural networks}


\author{%
  Ammar Tareen \\
  Simons Center for Quantitative Biology\\
  Cold Spring Harbor Laboratory\\
  Cold Spring Harbor, NY 11724 \\
  \texttt{tareen@cshl.edu} \\
  \And
  Justin B. Kinney \\
  Simons Center for Quantitative Biology\\
  Cold Spring Harbor Laboratory\\
  Cold Spring Harbor, NY 11724 \\
  \texttt{jkinney@cshl.edu} \
}

\begin{document}

\maketitle

\begin{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
```

100% | Page 1 of 6

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 determines the probability P_s of each state occurring in thermodynamic equilibrium via the Boltzmann distribution,[1]

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

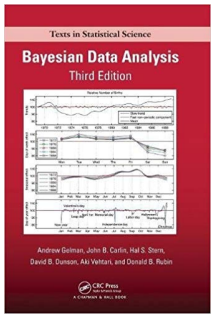
[1]To reduce notational burden, all ΔG values are assume to be in thermal units. At 37°C, one thermal unit is $1\,k_BT = 0.62$ kcal/mol, where k_B is Boltzmann's constant and T is temperature.

Develop core quantitative knowledge

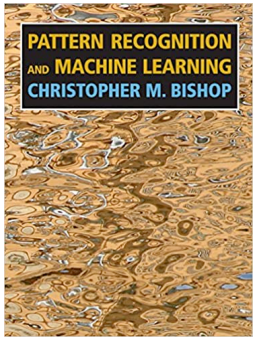
Fundamentals	**Intermediate material**	**Advanced material**
Calculus	Bayesian inference	Molecular biophysics
Linear Algebra	Machine learning	Stochastic processes
Algorithms (basic)	Sequence analysis	Dynamical systems
Statistics (basic)	Population genetics	Information theory
	Theoretical neuroscience	Deep learning
	Algorithms (intermediate)	…
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

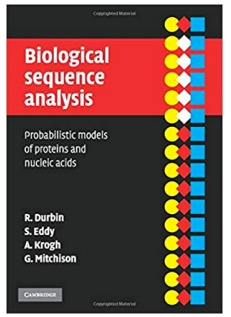
New QB initiative: faculty will help interested students pursue directed independent study of graduate-level material.
Email me <jkinney@cshl.edu> if interested.



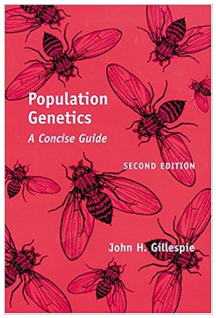
Bayesian Data Analysis, 3rd ed
Gelman et al., 2013



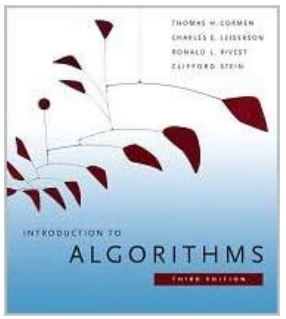
Pattern Recognition and Machine Learning
Bishop, 2006



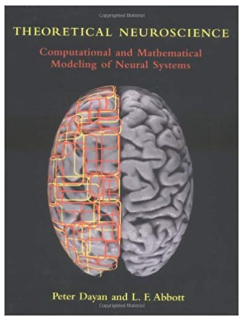
Biological Sequence Analysis
Durbin et al., 1998



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.