Lecture 1: Introduction and R tutorial

Jing Ma, Statistics, TAMU

August 26 2019

Outline

Course information

- Stat 312: *Statistics for Biology*
- Instructor: Jing Ma (jingma@tamu.edu)
- Textbook: *Modern Statistics for Modern Biology* by Holmes and Huber, available at http://web.stanford.edu/class/bios221/book/

Course overview

- First time offered
- An introduction to statistical learning methods for biology.
- Topics to be covered include clustering, dimension reduction, hypothesis testing, classification, regression, experimental design, etc.
- Learn how to use R to visualize and analyze biological data.

Grading: Homework 50%, midterm 20%, final 30%.

Course site

Everything will be on eCampus and my website.

- Syllabus
- Lectures
- Homework
- Exams

Outline

Challenges with modern biological data

- Genetic data are discrete: counts, transitions, states.
- Independence is not the norm.
- Contingency tables (chi-square or not).
- Large heterogeneous data sets.
- Need to interface statistics programs with databases, ontologies, etc.
- Non-standard parameters are common: trees, graphs, etc.
- Complex plotting procedures.
- Reproducibility of all research (diaries, write-ups, documentation).

General principles do apply

- Generative probability models (Poisson, binomial).
- Statistical methods (maximum likelihood).
- High quality graphics at three different levels.
- Data transformations.
- Removing unwanted variation.
- Experimental design.

Goal of the course

Learn the useful probabilistic tools specific to gene expression, protein, metabolic, immunological or microbiome data.

- Modeling discrete random variables (binomial, multinomial, Poisson, Dirichlet).
- Monte Carlo simulation (bootstrap, nonparametric testing, power).
- Filtering, de-noising, modeling, transforming (when to use $\log$).
- Markov chains (transitions, dependencies).
- Mixture models (latent variables, EM).
- Expectation, conditional probability, variance (we only need the basics).

Goal of the course

Learn the statistical machine learning tools for analyzing large data sets.

- Multivariate analyses (PCA, SVD, CA, MDS).
- Maximum likelihood estimation, Bayesian methods, EM.
- Clustering, mixture models
- Multiple testing, gene set enrichment analysis
- Data preprocessing, variance stabilization.
- Supervised methods: classification and regression.

Goal of the course

Learn to design your experiments and analyses.

- Power computations, randomization, sensitivity, robustness.
- Reproducible research.

Goal of the course

Learn to use R to run statistical analyses of biological data.

- Bioconductor suite of bioinformatics packages.
- R for sequence analyses and interfacing with databases (`Biostrings`, `DEseq2`, `Bayeseq`, `edgeR`).
- R for phylogenetics (`ape`, `phangorn`, `distory`, `phyloseq`).
- R for high quality visualizations (`ggplot2`, `cowplot`).
- R for multivariate multi-table analyses (`ade4`, `vegan`, `phyloseq`).
- R for input and normalization of data from modern technologies (`ShortRead`, `DEseq2`, `edgeR`).
- R for network and dynamic plotting (`igraph`, `animation`, `statnet`).
- R for convenient MC simulation (`MCMCpack`, `bootstrap`).
- RStudio and `knitr`.

Worlds of variability

Worlds of variability

- Biology cannot be easily summarized into simple principles because it is a world of complex variation.
- Variation = differences between organisms
- It is variability that has enabled evolution, and it is variability that ensures the robustness of complex biological systems. This is the rule rather than the exception in biological systems.
- Statistics and probability provide many tools for decomposing the signals in medical, genetic and ecological data.

Particularities of genomic data

Genetic sequence data are often discrete: either binary, or categorical (A,C,G,T). Most of the data come in the form of counts, or frequency tables, which we call these contingency tables.

Example:

	followed by			
first	A	C	G	T
A	34	09	14	33
C	12	13	10	10
G	11	15	17	09
T	13	22	20	23

Phenotypic data

Eyes	Black	Brunette	Red	Blonde
Brown	68	20	15	5
Blue	119	84	54	29
Hazel	26	17	14	14
Green	7	94	10	16

Non-standard parameters

- The parameters that we will be interested in are non-standard. They could be trees.
- Family trees of genes and of species are called phylogenetic trees, and are very important in the study of molecular evolution.

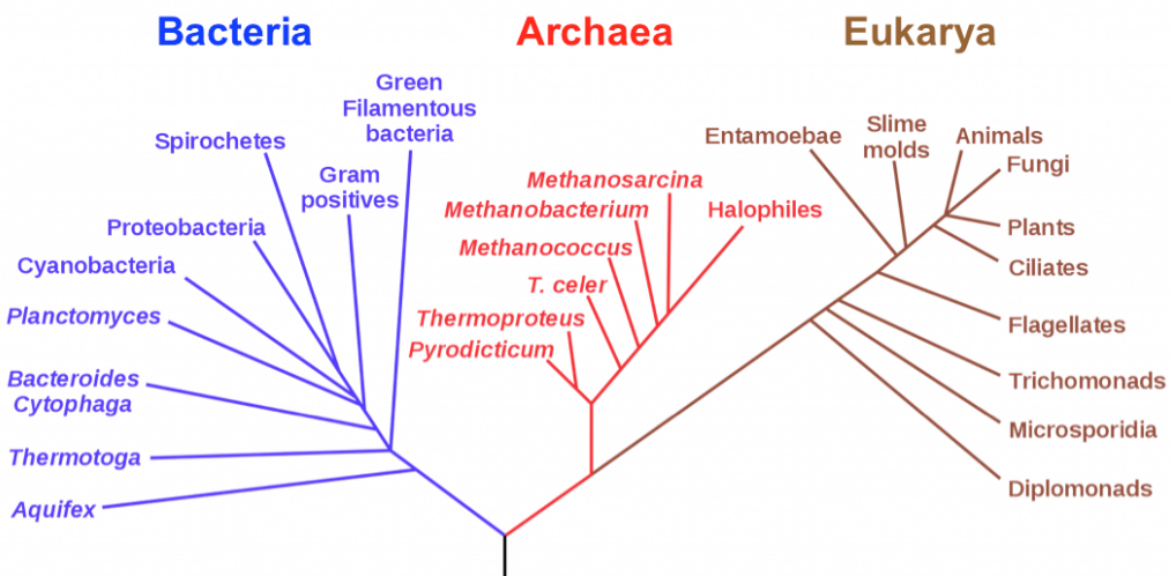
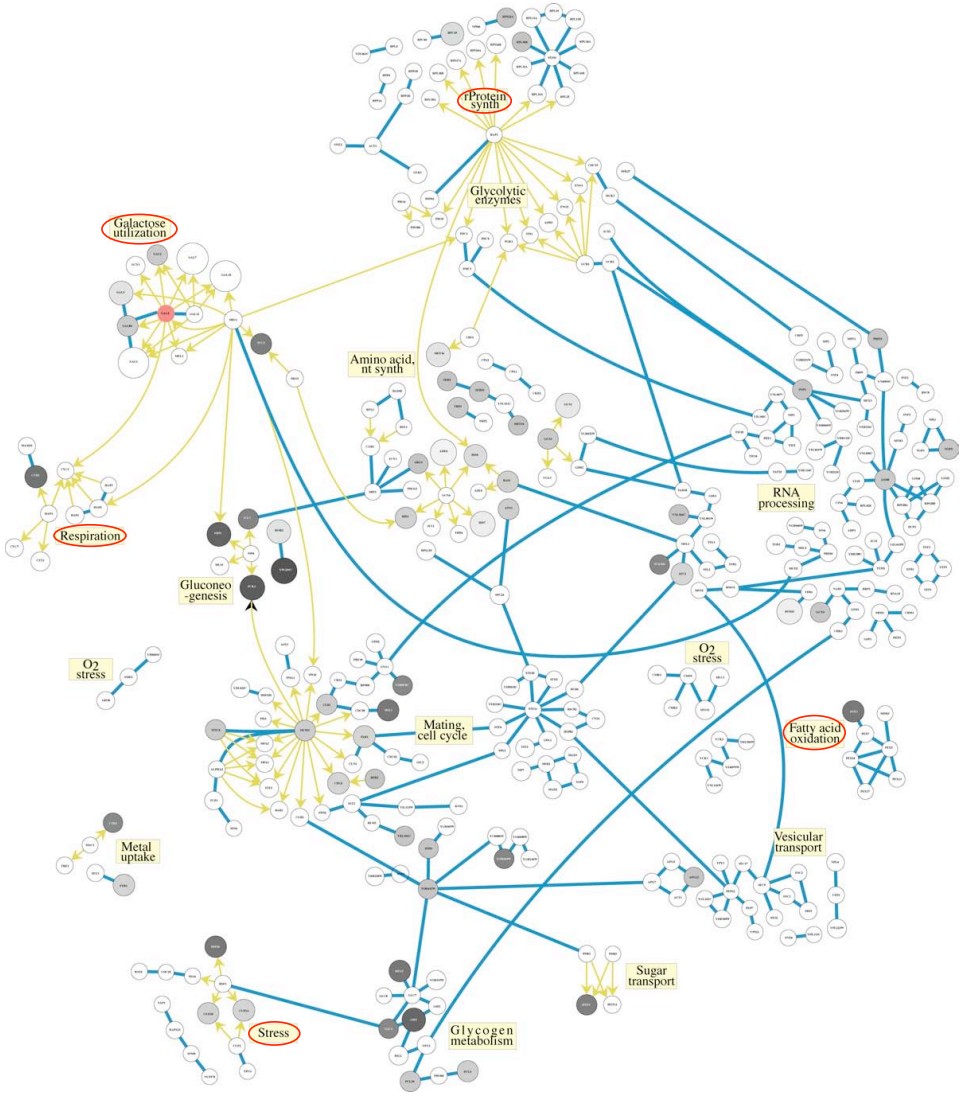


Figure: A speculatively rooted tree for rRNA genes, showing the three life domains: bacteria, archaea, and eukaryota. (wiki)

Graphs and networks

- Genes work together and it is important to understand how they interact in gene regulatory networks. (Right: Ideker et al. 2001. Science)



Metabolic networks and pathways

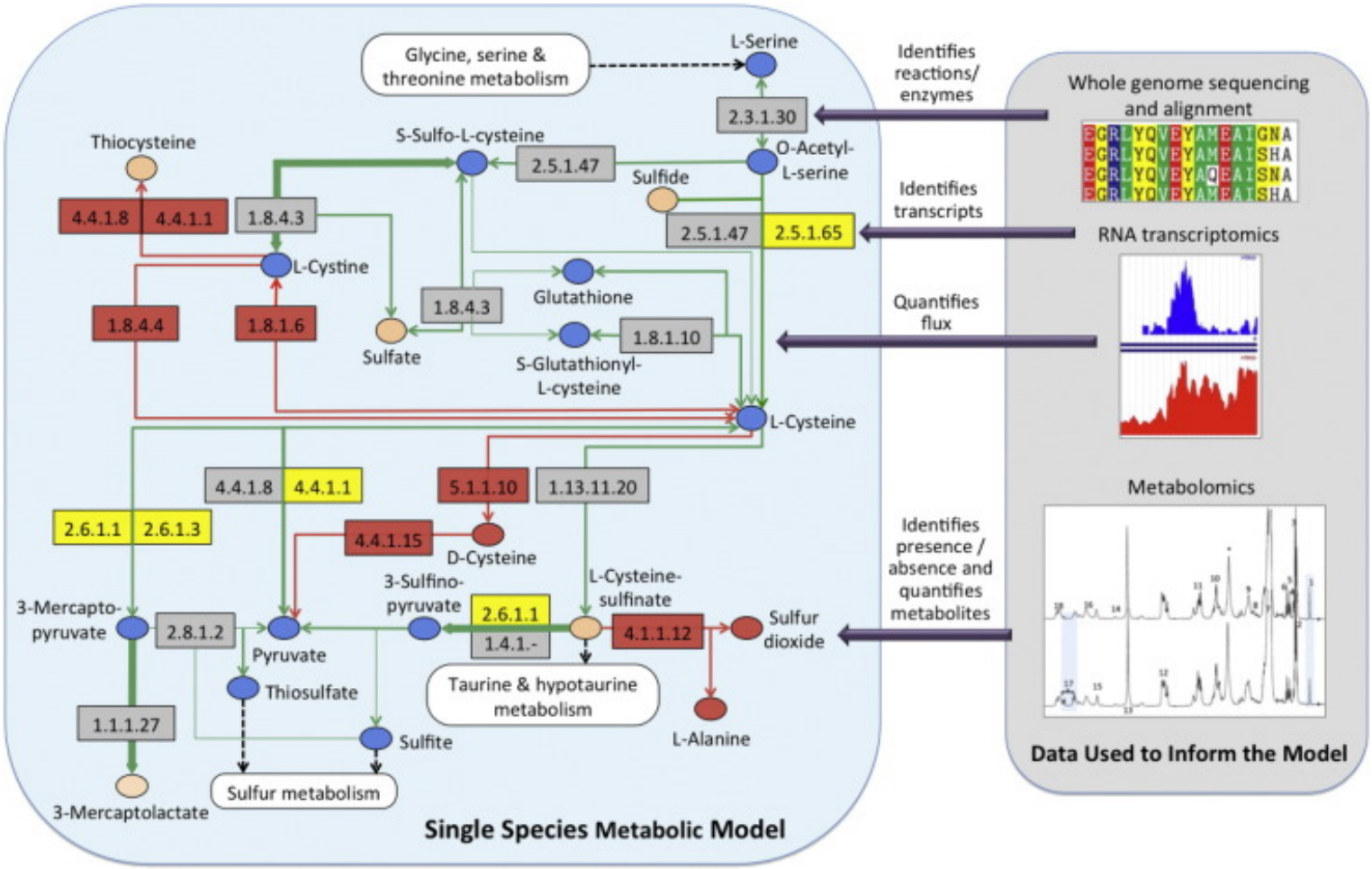


Figure: Subset of a microbial metabolic network with integrated genome, metabolomics, and RNA data. (Sung et al. 2016. Appl Transl Genom.)

Non-standard parameters

- What one actually estimates in genetics, immunology or microbiology can be very different from classical statistics.
- In classical statistics, we estimate what we don't know and very often we denote it by a greek letter called a parameter. In most cases, the parameter is a real number. We might have an estimate on its own or we might get a lower and a higher estimates which constitute a confidence interval.
- We will see that in biology the parameters are much more complicated.

Challenges that we will focus on

- Heterogeneity.
- Structured high-dimensionality.
- Graph or tree integration.
- High quality graphics.
- Reproducibility.
- Validation and confirmatory analysis.

Heterogeneity of data

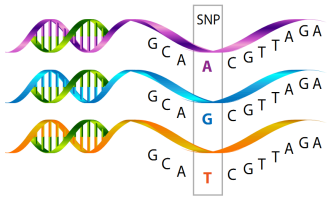
- Status: response or explanatory.
- Hidden (latent) or measured (observed).
- Types:
  - Continuous.
  - Discrete, categorical.
  - Graphs, trees.
  - Images.
  - Maps, spatial information.
- Amounts of dependency: independent, time series, spatial.
- Different technologies used (16S-rRNA, illumina, Minion, Mass-Cyto, Mass Spec, RNA-seq).

A systematic approach to data integration

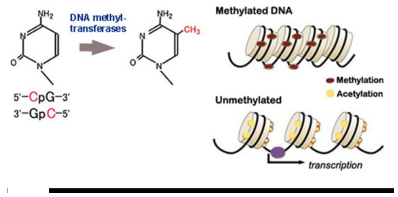
TCGA

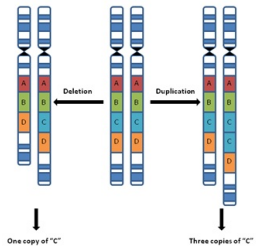
Understanding genomics to improve cancer care

Mutation

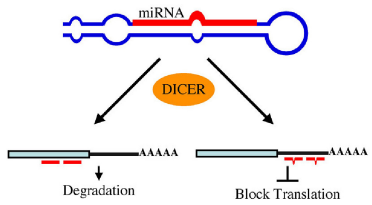


Methylation

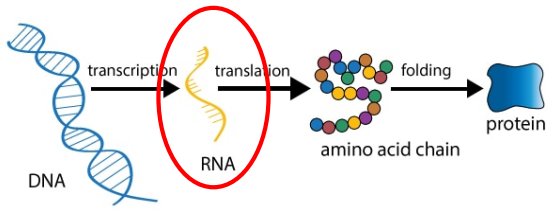




Copy Number Variation



microRNA Expression



Gene Expression

What is Probability?

- Allows one to go from a hypothetical model, usually parametric, to the probability of an event.
- Probability is the theory of randomness, which we use to model uncertainty, noise and its consequences on what would be observed under certain probabilistic models.

What is Statistics?

- Reasoning backwards from data to a potential explanation for what we see.
- Statistics is a separate subject from probability. Although the latter serves as the mathematical basis for making inferences, statistics does not exist without data, and in the case of contemporary genetics, large data sets are common.
- We need to know quite a few things about the probabilistic models that might have generated the data, in order to go back from the data to guess at the best possible model from which we think the data come.

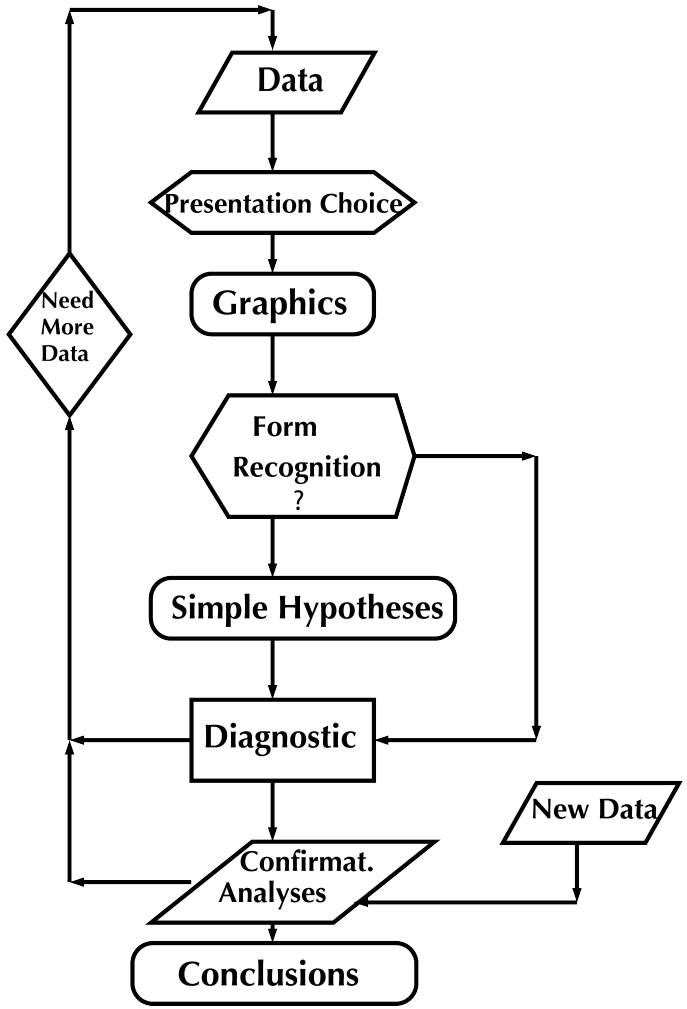
Hardest part: curse of dimensionality

- We will often encounter the high dimensionality nature of the biological phenomena.
  - Genes work together so their expression patterns can be highly correlated and thus we cannot look at variables one at a time.
- We need multivariate techniques such as principal component analysis, multidimensional scaling and cluster analysis.
- How to understand the data when one doesn't have a model? Methods such as dimension reduction and visualization are part of what we call exploratory data analyses.

Statistics is not only p-values

- Statistics is often caricatured as a method for obtaining p-values; this is far from what statisticians spend most of their time doing. It is true that before the computer age, statisticians mostly used probability theory to make statements about their inferences (such as hypothesis testing).

Modern statistical analysis is iterative



Outline

R tutorial

Prerequisites.

- You need to know the basics of probability theory, e.g. expectation, variance, binomial distribution, normal distribution.
- You need to know what are vectors, matrices and the associated operations (addition, multiplication and transposition).

If you have never learned anything about the above topics, try your best to pick up some basics. Here are two recommended books.

- A First Course in Probability (Sheldon Ross).
- Introduction to Probability (Grinstead and Snell).

R tutorial

- You can either install R on your personal laptop or use some web-based R coding environment.
- For the latter, we recommend RStudio (a free IDE for R).
- We will use RStudio to write and run R scripts.
- Please bring your laptop.
- Laptop is also needed for the final exam.
- Please contact me immediately if you need to borrow a laptop.

Advantages of R

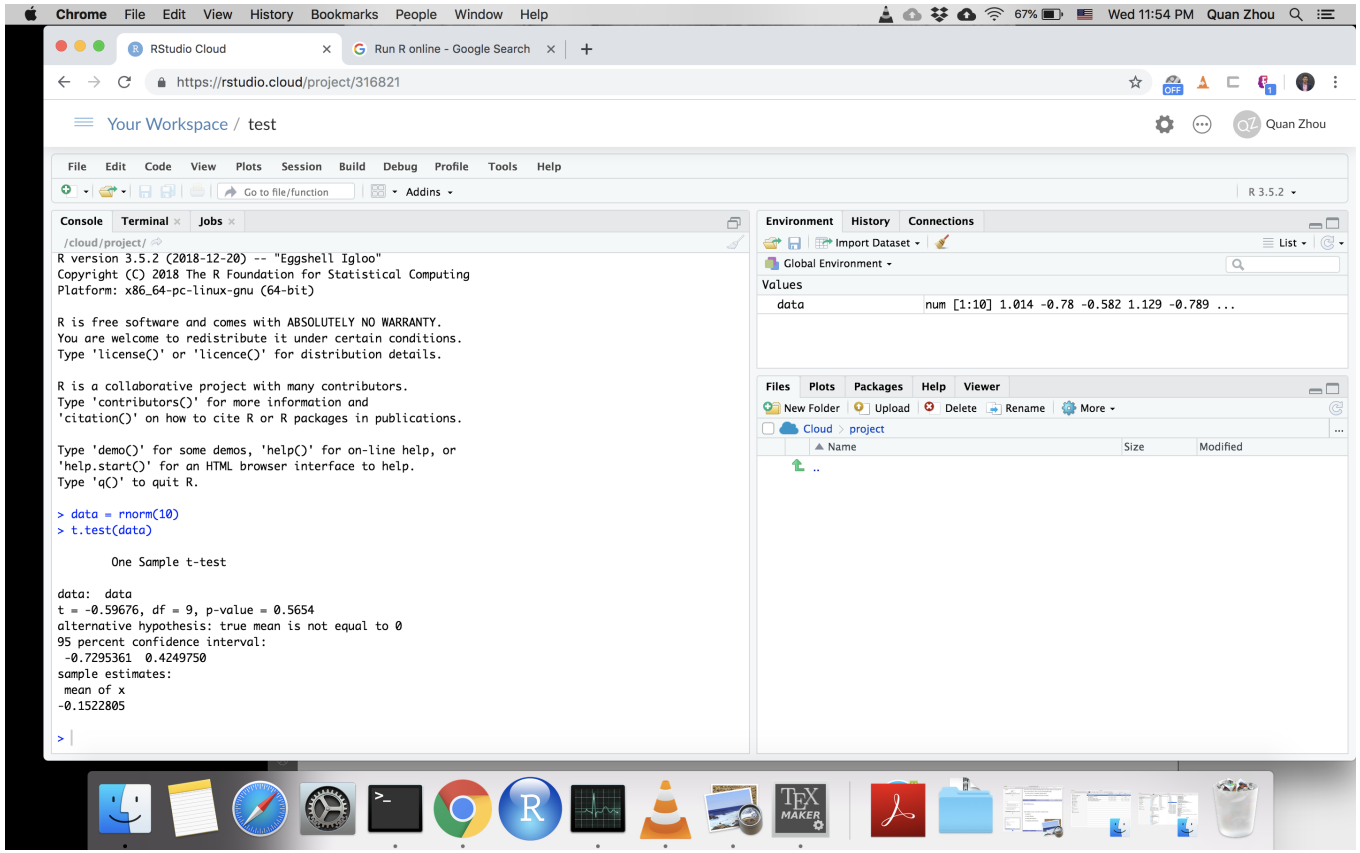
- Availability of $> 10,000$ packages, all statistical/machine learning methods available.
- High quality graphics: `ggplot2`.
- Reproducibilty: R Markdown.
- Free and open source.
- Seamless interaction with standard databases such as GenBank, Gene Ontology Consortium, UCSC Human Genome Project, KEGG.

R packages are tailored to various formats

- Maps
- Trees, networks and graphs
- Microarrays, mass spectroscopy
- Text
- Images
- DNA, AA, sequencing data, HTS, RNA-seq
- QIIME, mothur formats

Web-based RStudio

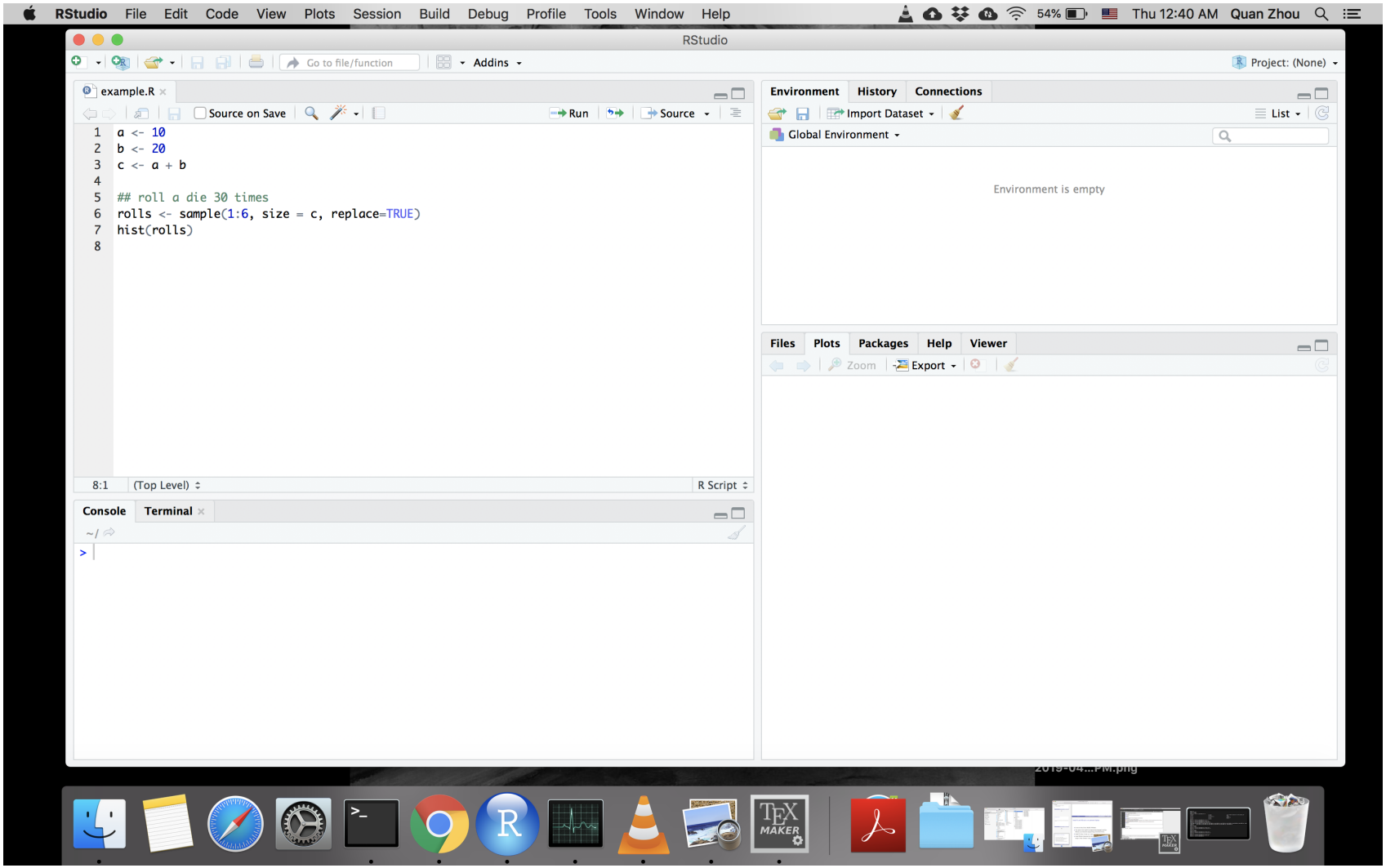
https://rstudio.cloud is free and easy to use. You only need to register with your email.



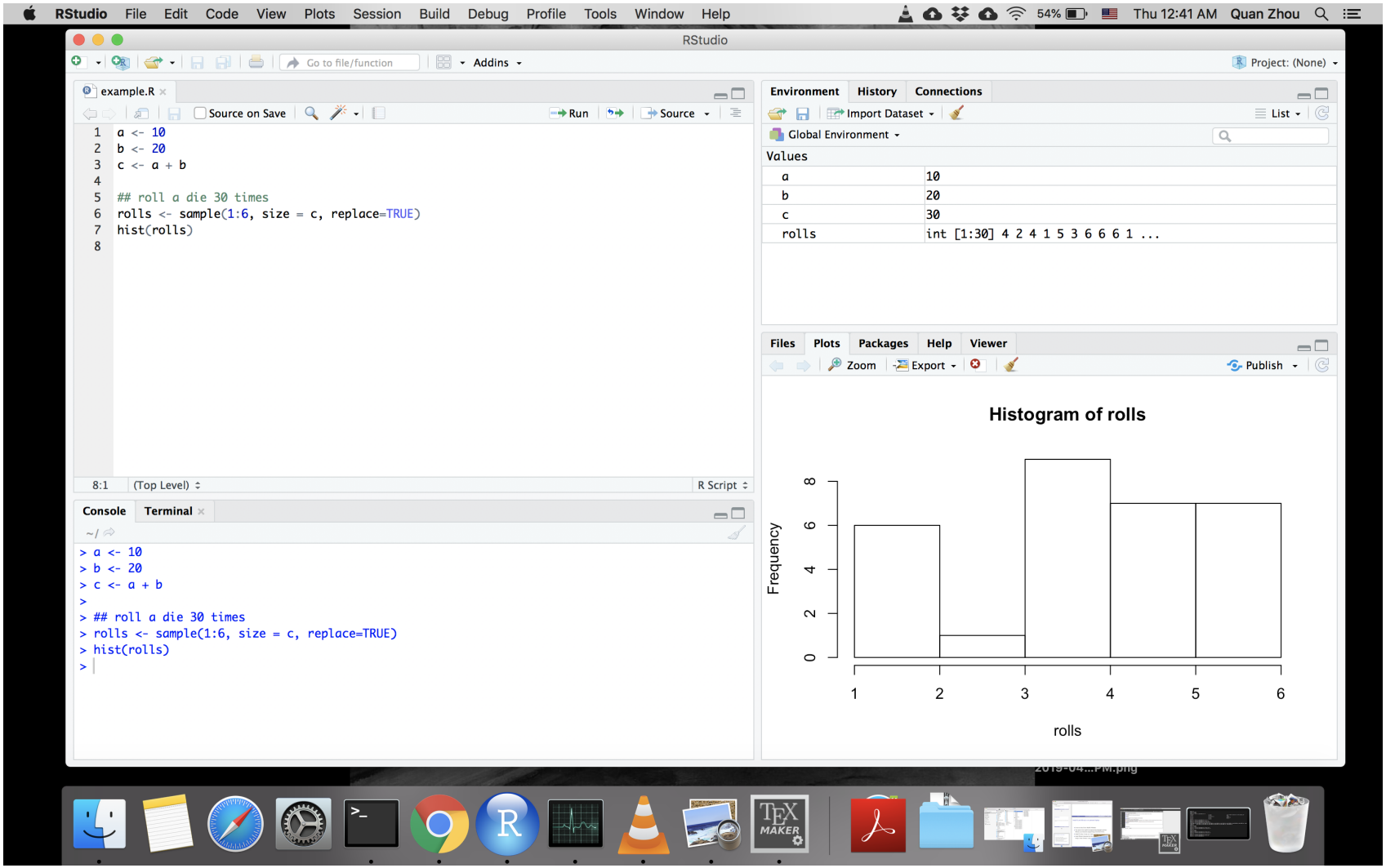
Install R and RStudio on personal laptops

- R runs on Unix/Linux, MacOS, Windows.
- You need to first install R (a programming language) and then RStudio (an interactive programming environment for R).
- Install R: http://cran.us.r-project.org
- Install RStudio (download the free version): https://www.rstudio.com/products/rstudio/download/

RStudio on your laptop



RStudio on your laptop



Homework

Complete an R tutorial course at the following site

- Lynda (Learning R): https://lynda.tamu.edu
- If time allows, challenge yourself with the intermediate-level course.

We will have a quiz next Monday (laptop not needed).

Homework

In particular, make sure you understand

- how to use RStudio
- data types and data structures in `R`
- arithmetic operations including exponentiation and modulo
- variable assignment and comparison
- how to create and use a sequence
- how to create a vector and insert, delete or access its elements
- how to create a matrix and access its rows, columns and elements
- arithmetic operations involving vectors and matrices