Introduction and course overview

Neo Christopher Chung

Lecture 1, 1000-719bMSB Modeling of complex biological systems

Course overview

Challenges and solutions in modeling biological systems

Large scale measurement techniques, microarrays, RNA-seq, imaging

Old and new computational methods, from statistics to deep learning

Understand fundamental organizational principles and functionalities of biosystems

Course website: https://cbml.science/cbs

Email: nchchung@gmail.com   n.chung@uw.edu.pl

Add [1000-719bMSB] in the subject

Prerequisites

Statistics and Data Analysis

Programming in R and Python

Bioinformatics and Genomics

Learning materials

An Introduction to Statistical Learning: https://www.statlearning.com/

Data Science Specialization: https://www.coursera.org/specializations/jhu-data-science

R for Data Science (book): https://r4ds.had.co.nz/

Dive into Deep Learning (Pytorch): http://d2l.ai

PyTorch Tutorials: https://pytorch.org/tutorials/index.html

Code Academy: https://www.codecademy.com/

DataCamp: https://www.datacamp.com/

Grade

Course participation: 20%

Class Note: 20%

Lab homework: 20%

Final presentation: 20%

Final report: 20%

>50% required to pass the course

Class notes, before the class

Each student write a summary for one week’s topic (>2 page; 11 pt size).

Use our reading materials and textbooks.

Must be in your own words, no copy & paste, no plagiarism, etc.

These notes will be shared with all students.

Homework

Given during the computer lab

Embedded in the course notebooks (R Markdown or Jupyter notebook)

Upload your codes and outputs (PDF files) to your Github account

Due by the following Sunday night 23:00

Final Project

Study a specific biological system and a biomedical question

Be inspired by biological functions, diseases, modeling approaches

Use the modern research practices (GitHub, Markdown/Jupyter, etc)

> Have a specific hypothesis or an exploratory goal
>
> Replicate an interesting research
>
> Experiment with how an analysis is done
>
> Improve methods and algorithms

Final Project

Choose a topic that is interesting to you.

Consider innovating and experimenting with how an analysis is done.

At the very minimum, try replicating a study that is interesting to you.

Required sections: title, abstract, introduction, methods and materials, results, discussion, references

Length: minimum 6 pages excluding figures and references

Format: single-spaced, 11 font size, Times New Roman

Reference: Nature citation style

Major methods & applications

1. statistical tests & false discovery rates
2. dimension reduction & latent variable models
3. unwanted variation & batch effects
4. single cell RNA-seq & analysis
5. cellular identities & trajectories
6. neural networks and convolution
7. natural images & spatial transcriptomics
8. interpretability of neural networks
9. analysis of medical images

Modern research practice

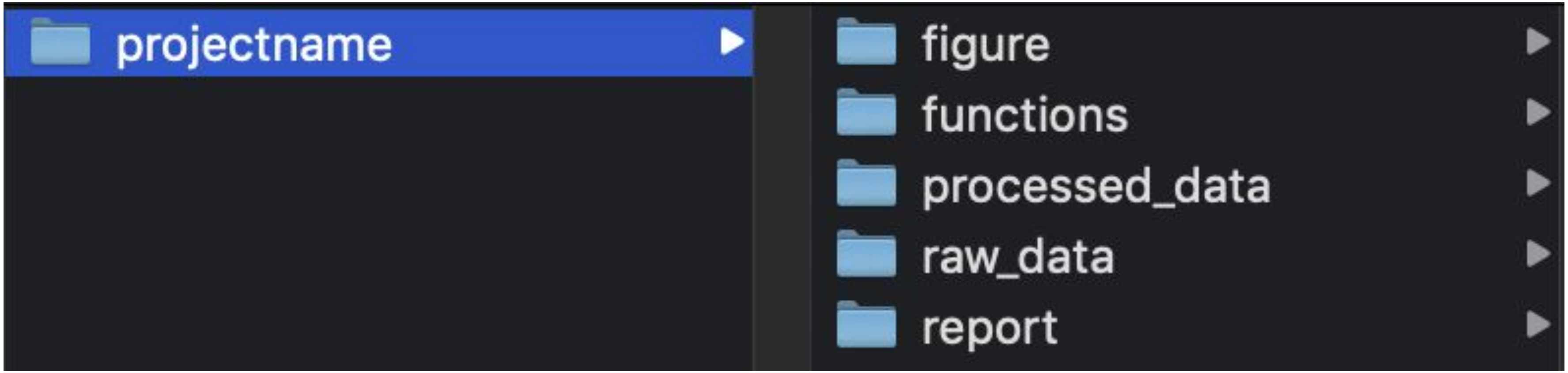
Open a Github account, invite https://github.com/ncchung as a collaborator

Make one repository for this course, create a separate directory for a homework, etc

Create a reproducible analysis and keep your R/Python scripts

Write in R Markdown or Jupyter notebooks

Organize your project



projectname

figure

functions

processed_data

raw_data

report

Models and Modeling

Conceptual model: concepts, rules, and representations

Physical model: interdependencies in physical systems

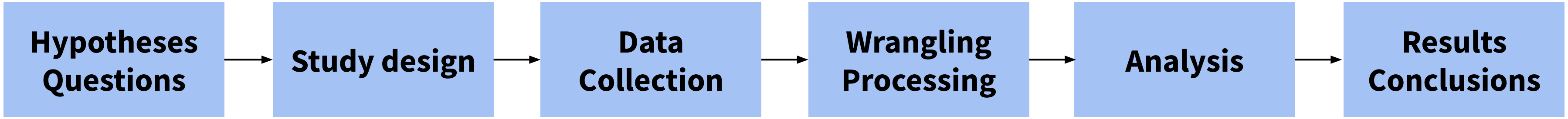
Computer model: simulate and reproduce behaviors of a systems

Mathematical model: describe a system using mathematical concepts and language

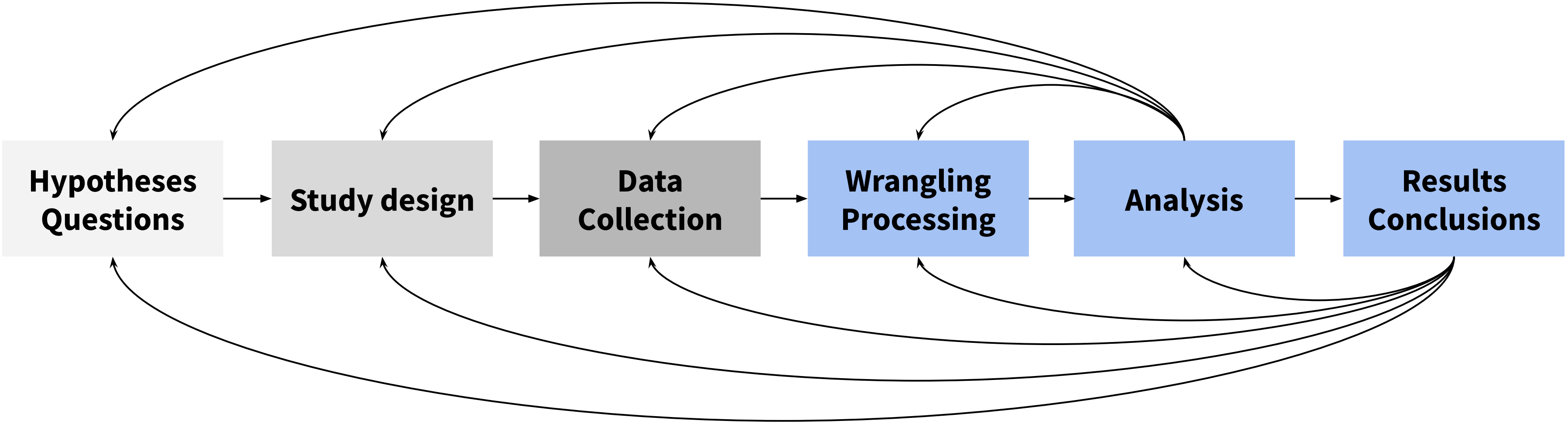
Statistical model: represent the data-generating process

→ what, how, and why **our observations** are realized

What does it mean to model



What does it mean to model



Not a linear process – revisiting earlier steps would be critical in practice
Our course focuses on the later steps – but tries to hint at earlier aspects

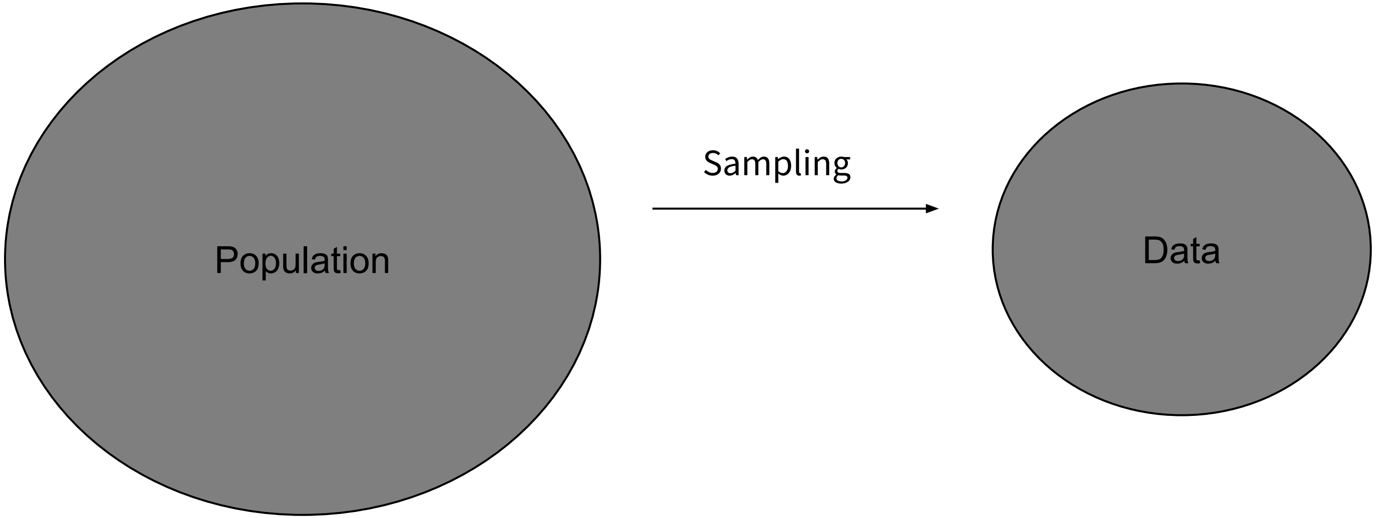
How we will achieve these challenges

1. Identify the questions → study recent papers and trends in computational biology
2. Design the study → come up with a new hypothesis, an improved methods, etc
3. Collect the data → run an experiment, observe, or identify a dataset
4. Process the data → clean up, transform, and normalize
5. Analyze the data → apply statistical and machine learning methods
6. Disseminate the results → write a final report and give a final presentation

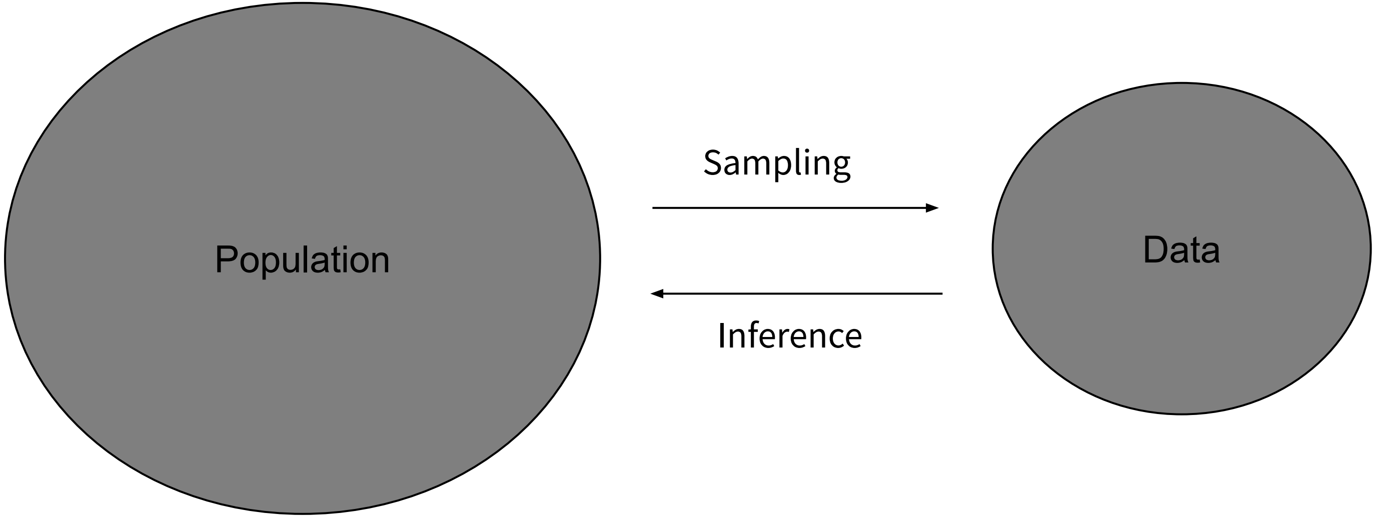
Data

1. Gene expression: abundance of RNAs
    a. Bulk cell
    b. Single cell
2. Abundance of proteins (e.g., proteomics) and metabolites (e.g., metabolomics)
3. Genetic variation; single-nucleotide polymorphisms
4. Spatial transcriptomics
5. Natural images
6. Medical images

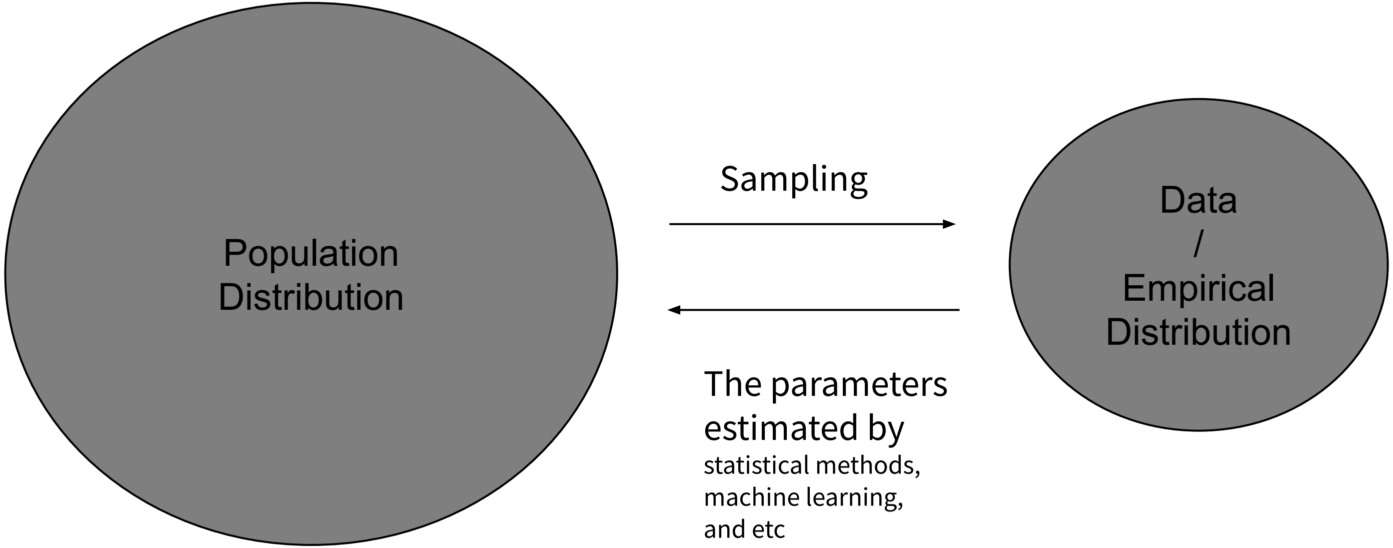
What is statistics?



What is statistics?

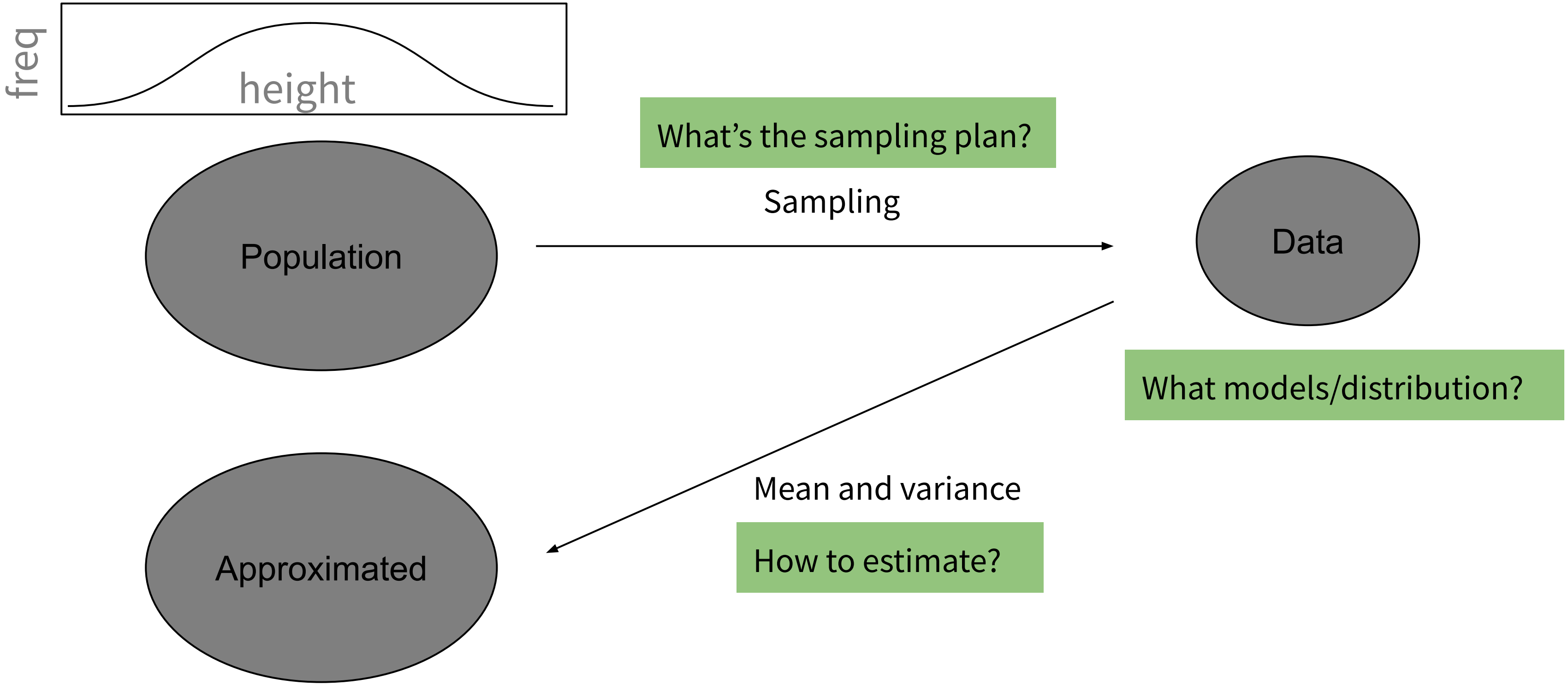


Statistical inference



The parameters estimated by statistical methods, machine learning, and etc

Height example



Population

What's the sampling plan?

Sampling

Data

What models/distribution?

Mean and variance

How to estimate?

Approximated

Limitations and Obstacles

Deriving a model is inherently approximation.

"All Models Are Wrong, Some Are Useful" - George Box

Sampling might be **NOT** random, balanced, etc.

Differences in estimation, inference, and predictions.

Best to check the data, to go back to experiment, and to test your data-driven conclusion.

Be open to new new approaches and emerging disciplines →

Emerging and interconnected disciplines

Statistics: the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data

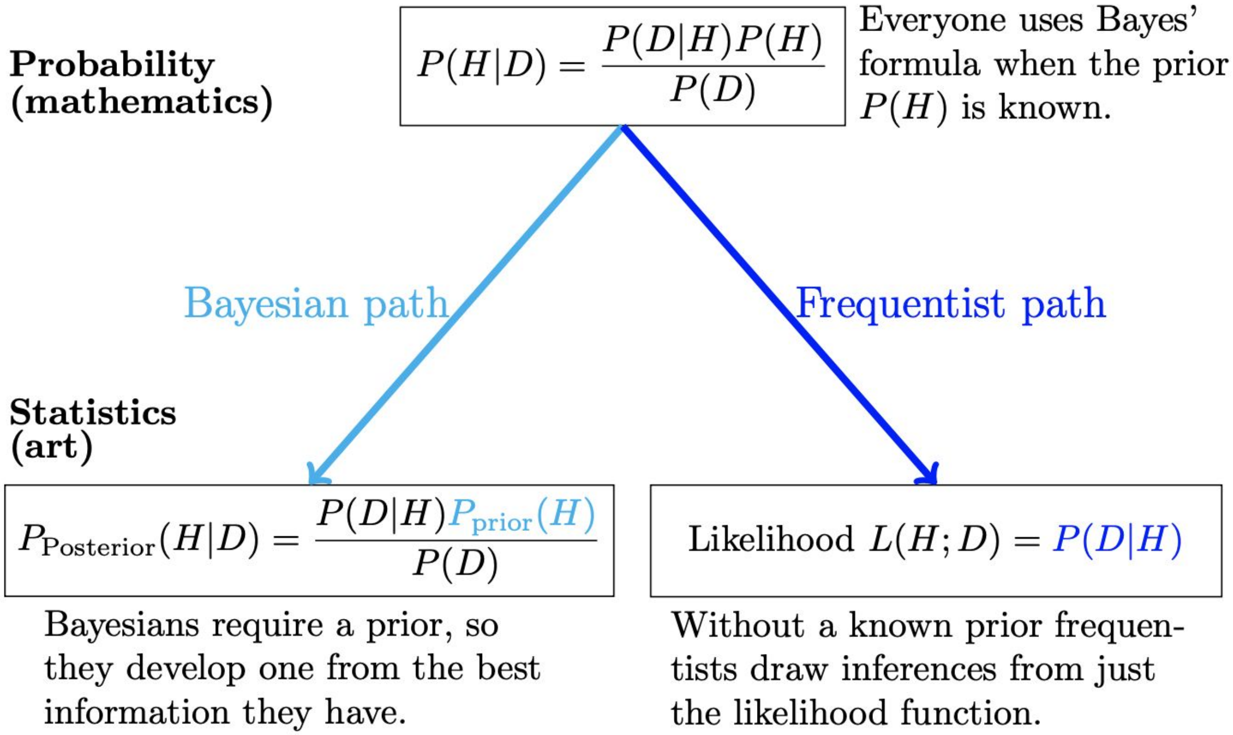
Data science: a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

Machine learning: the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead

… artificial intelligence, data analytics, statistical programming ...

From Wikipedia

Frequentist vs. Bayesian statistics



Probability (mathematics)

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Everyone uses Bayes' formula when the prior $P(H)$ is known.

Bayesian path

Frequentist path

Statistics (art)

$$P_{\text{Posterior}}(H|D) = \frac{P(D|H)P_{\text{prior}}(H)}{P(D)}$$

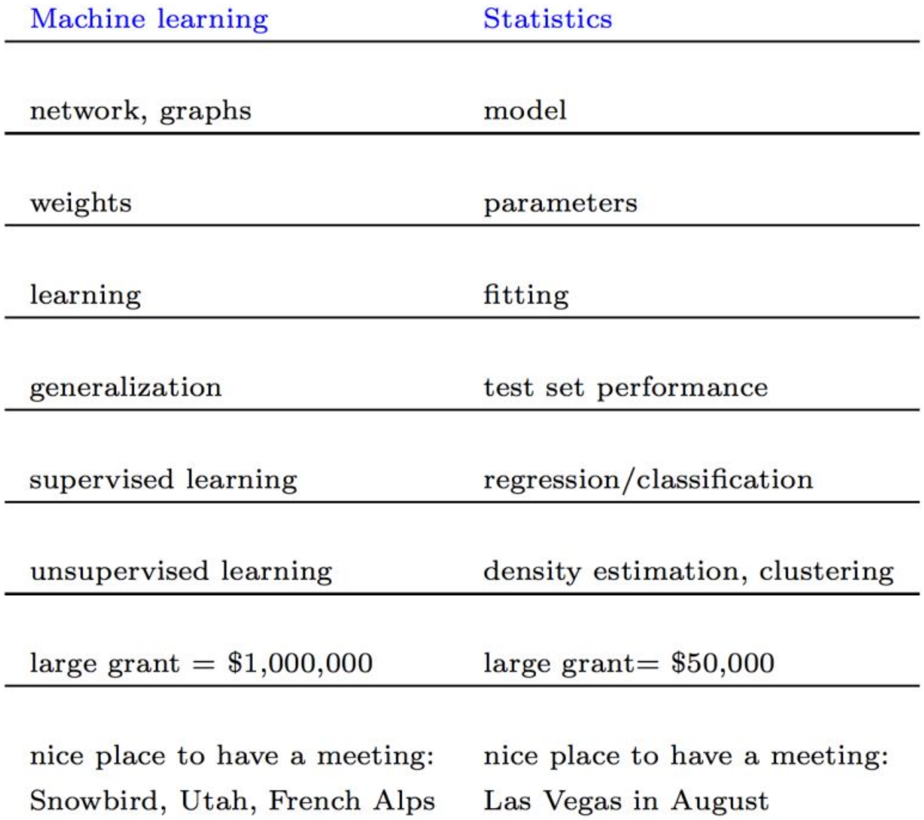
Bayesians require a prior, so they develop one from the best information they have.

$$\text{Likelihood } L(H;D) = P(D|H)$$

Without a known prior frequentists draw inferences from just the likelihood function.

http://www-math.mit.edu/~dav/05.dir/class17-slides-all.pdf

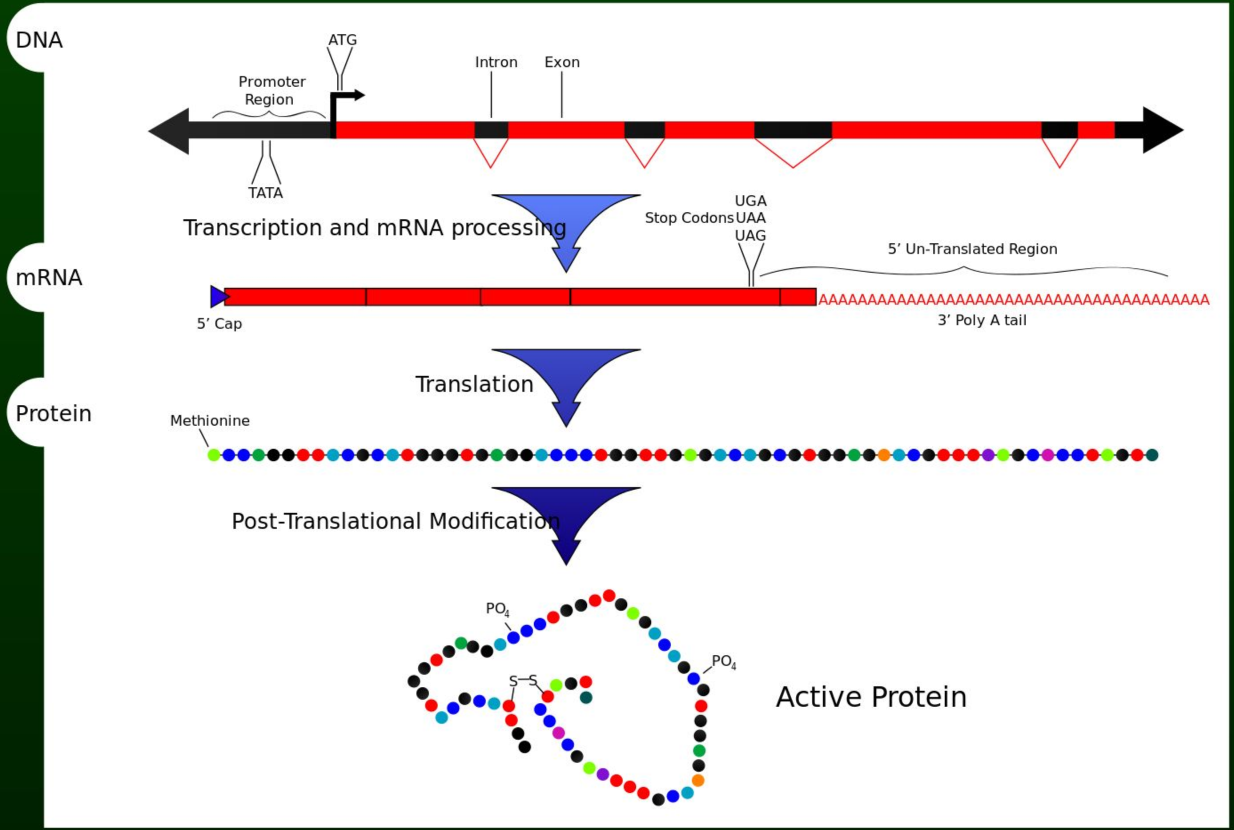
Statistics vs. Machine Learning



Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant= \$50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Robert Tibshiriani

Central Dogma of Molecular Biology : Eukaryotic Mode



https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

High-throughput molecular biology

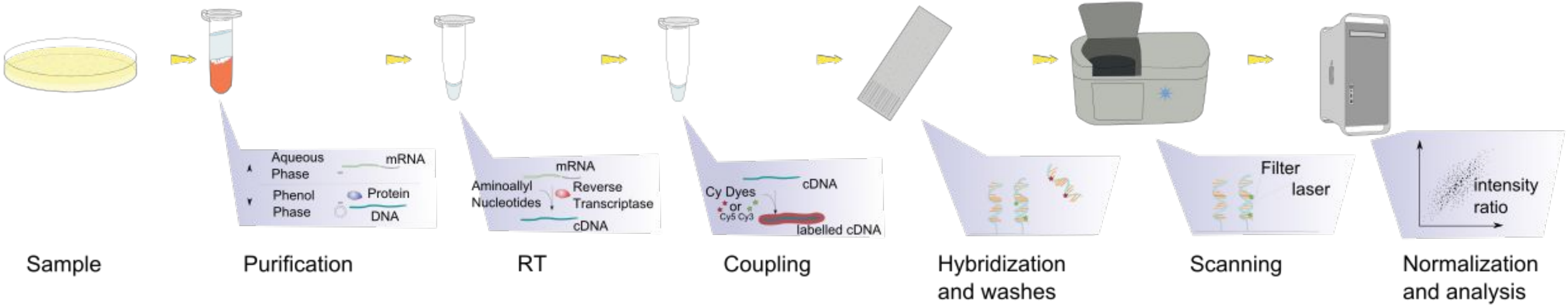
DNA sequences: DNA sequencing, tiling microarray

Amounts of RNAs: sequencing (via reverse transcription), microarray, RT-PCRs

Amounts of proteins: mass spec, immunocytochemistry, protein microarray

Available in a given organism, at a specific time and environment

Microarrays



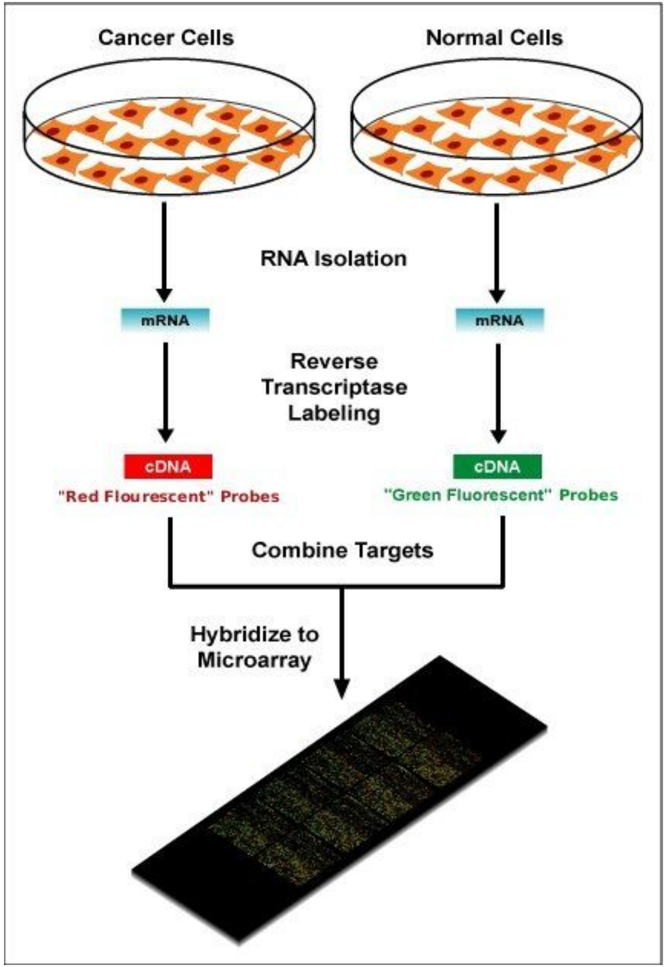
One-channel vs. two-channel

Single-channel microarrays

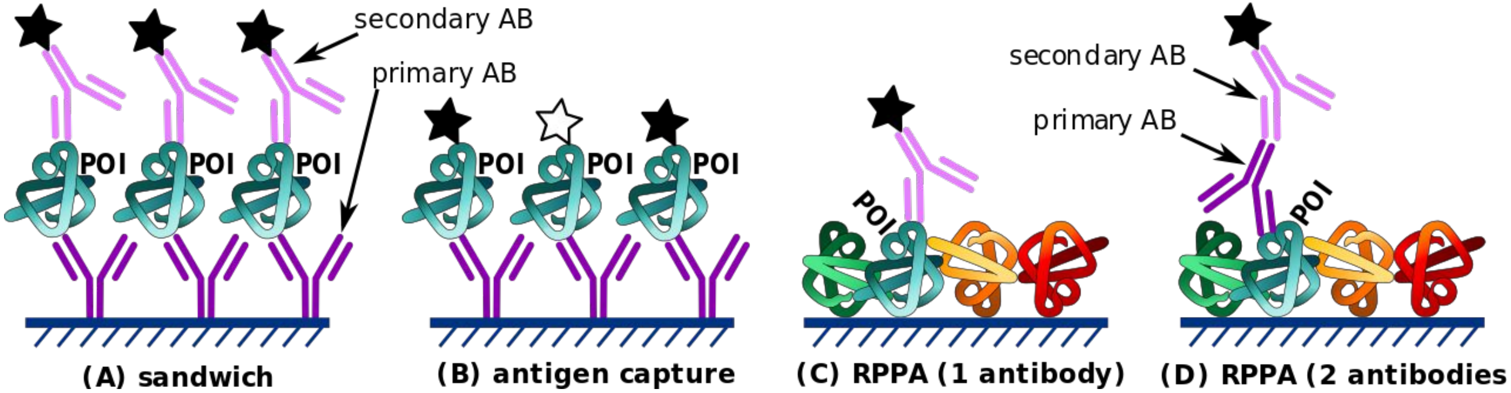
- Use one sample per microarray
- Provide intensity/abundance data
- Numeric values are relative to other probes in that experiment

Two-channel microarrays

- Use cDNA prepared from two samples per microarray
- e.g. cancer vs. normal tissue
- Typically uses Cy3 (~green) and Cy5 (~red) fluorophores
- Intensities of each fluorophore ~ ratios of two samples
- Identify up- and down-regulated genes w.r.t. the reference sample

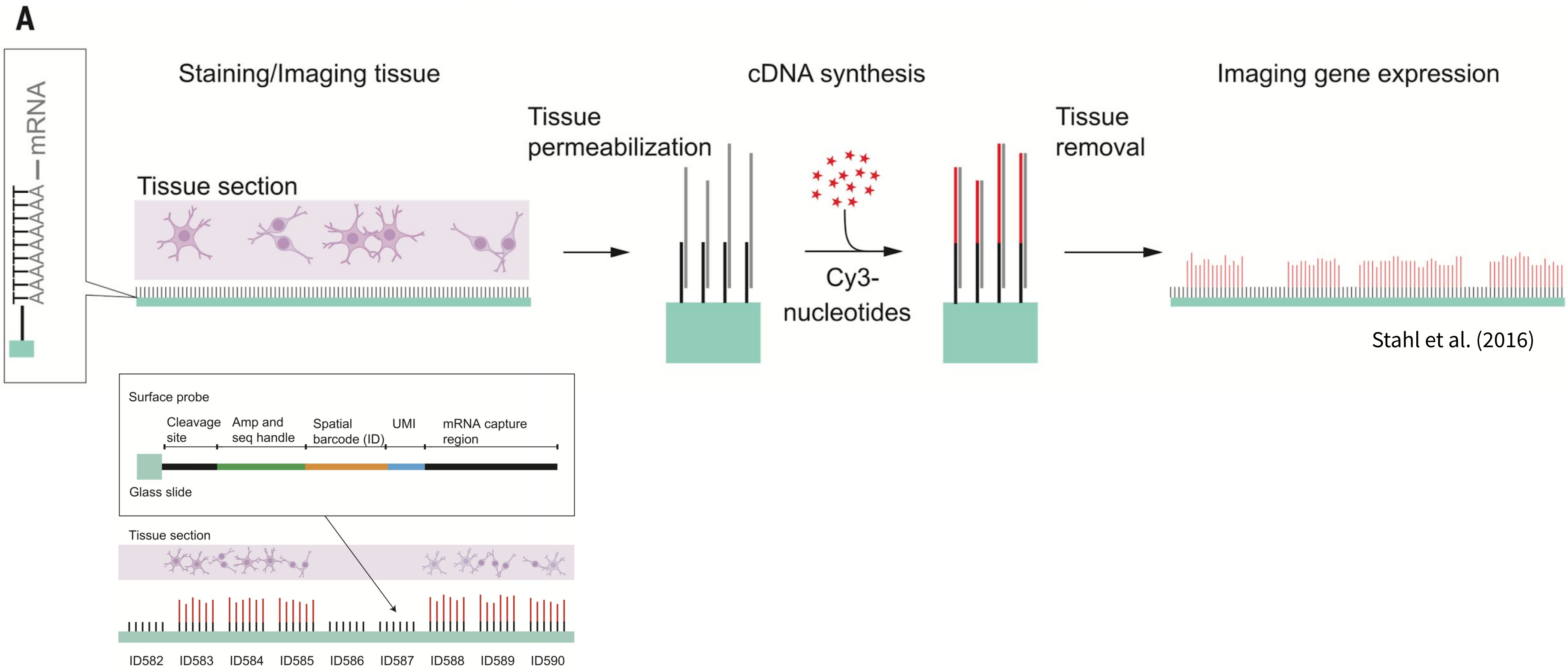


Protein microarrays



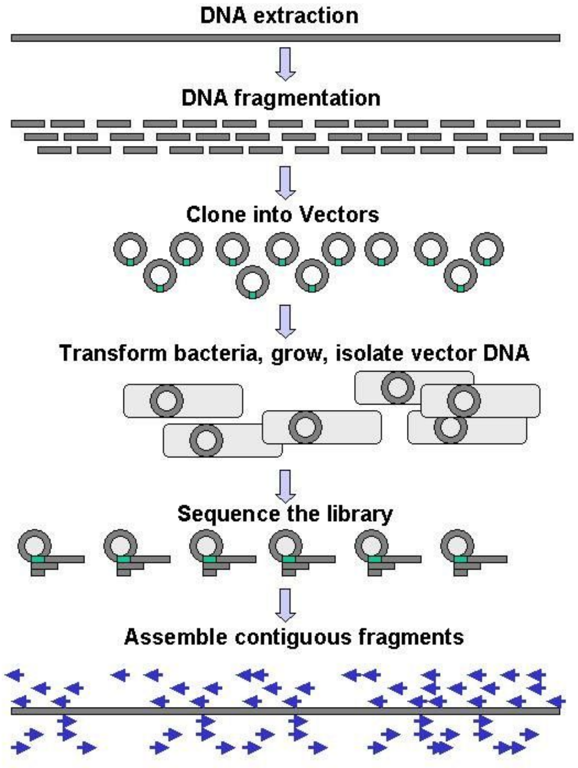
The lysate is arrayed onto the microarray and probed with antibodies against the target protein of interest. These antibodies are typically detected with chemiluminescent, fluorescent or colorimetric assays.

Microarray for a tissue section

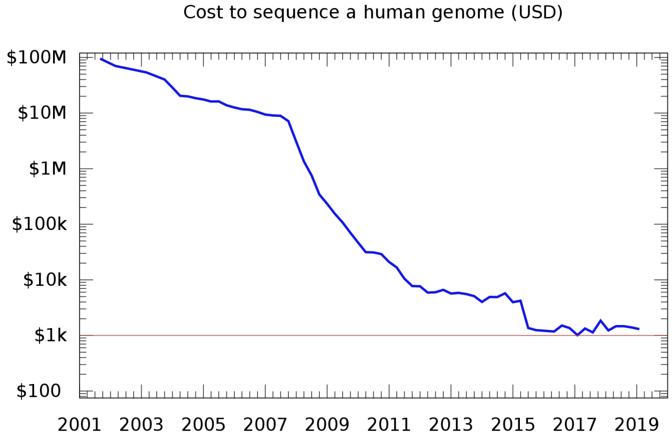


Stahl et al. (2016)

High throughput sequencing

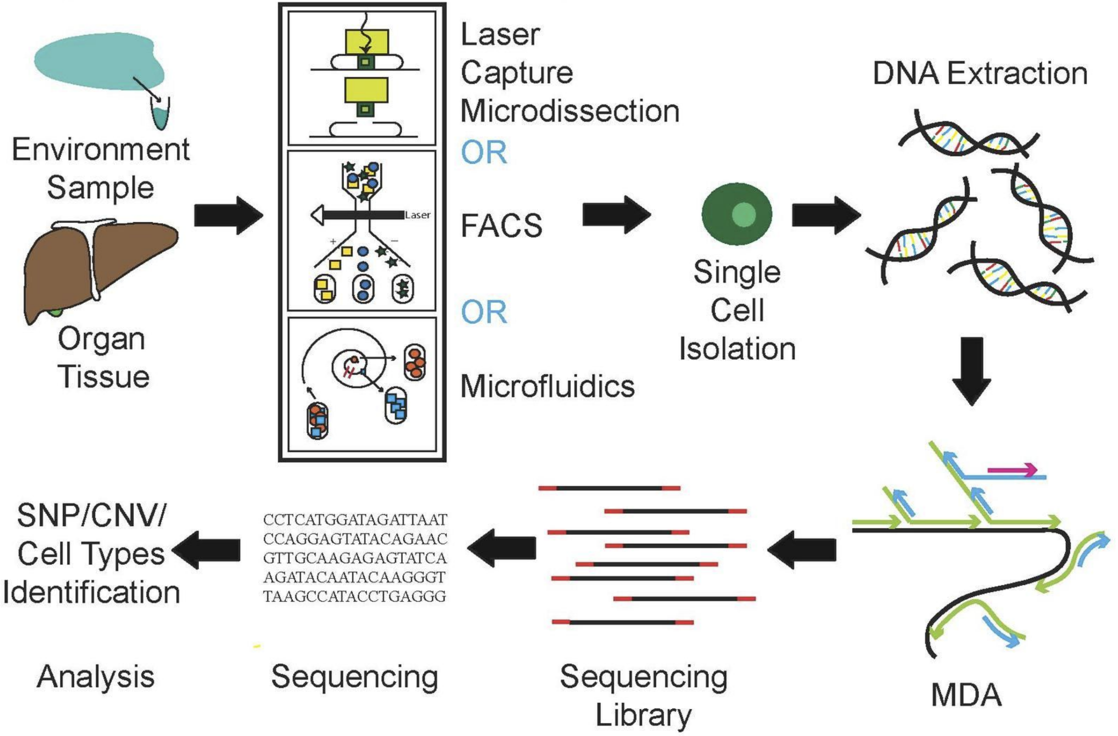

DNA extraction
DNA fragmentation
Clone into Vectors
Transform bacteria, grow, isolate vector DNA
Sequence the library
Assemble contiguous fragments

*obtain RNA abundances via reverse transcription



Single cell sequencing

Single Cell Genome Sequencing Workflow



Environment Sample

Organ Tissue

Laser Capture Microdissection

OR

FACS

OR

Microfluidics

Single Cell Isolation

DNA Extraction

MDA

Sequencing Library

CCTCATGGATAGATTAAT
CCAGGAGTATACAGAAC
GTTGCAAGAGAGTATCA
AGATACAATACAAGGGT
TAAGCCATACCTGAGGG

Sequencing

SNP/CNV/ Cell Types Identification

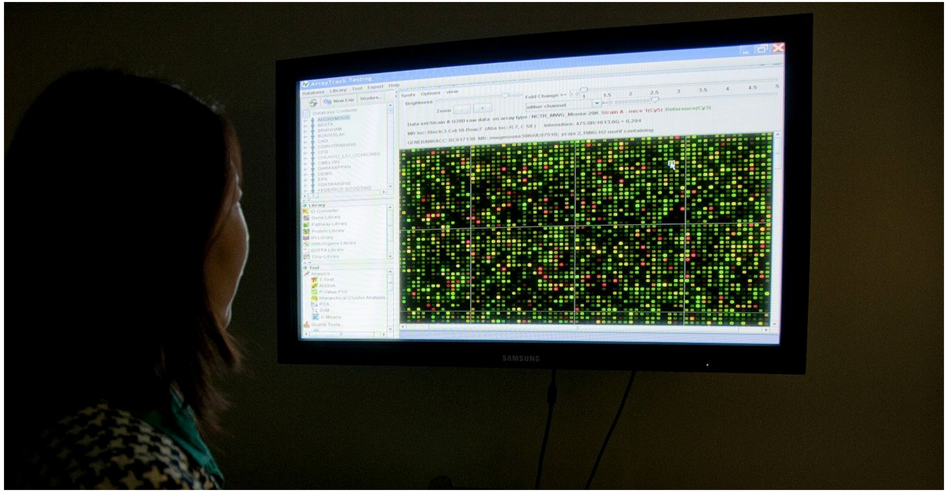
Analysis

Data processing

→

→ Images,
Signals,
Spectra

→ FASTA: sequence records
FASTQ: with quality scores
SAM: with mapping info



Data matrix

Observations

	Sample 1	Sample 2	…	Sample *n*
Variable 1	16.4	0.2	10.1	1.5
Variable 2	4.2	6.1	10.5	33.1
Variable 3	0.5	10.4	98.3	1.8
Variable 4	4.6	61.4	1.2	0.1
Variable 5	1.5	3.5	11.2	4.1
Variable 6	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
Variable *m*	…	…	…	…

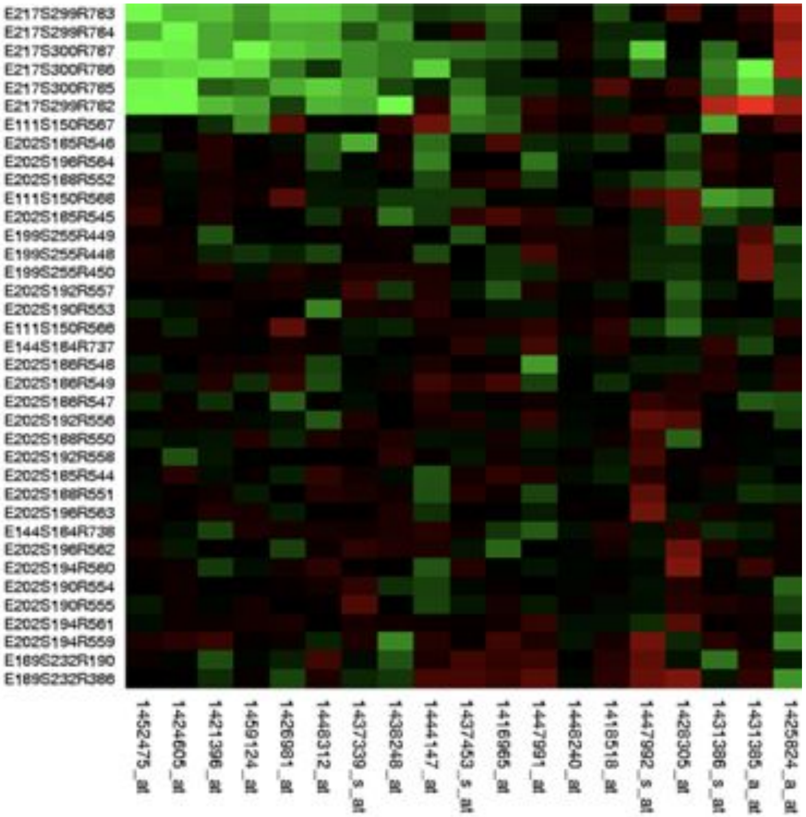
Example

Gene Expression related to Cancer vs. Normal?

	Normal Samples		Cancer Samples	
	Cell 1	Cell 2	Cell 3	Cell 4
Gene 1	16.4	0.2	10.1	1.5
Gene 2	4.2	6.1	10.5	33.1
Gene 3	0.5	10.4	98.3	1.8
Gene 4	4.6	61.4	1.2	0.1
Gene 5	1.5	3.5	11.2	4.1
Gene 6	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
Gene *m*	…	…	…	…

Visualized

Large values: Green
Middle values: Black
Low values: Red



Importance of data structures

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data

The principles of tidy data provide a standard way to organize data values within a dataset.

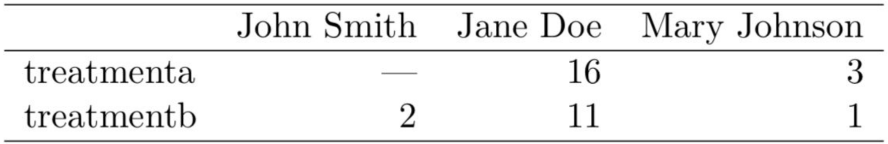
The principles of tidy data are closely tied to those of relational databases

Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).

Tidy Data, Hadley Wickham. The Journal of Statistical Software, (59) 2014

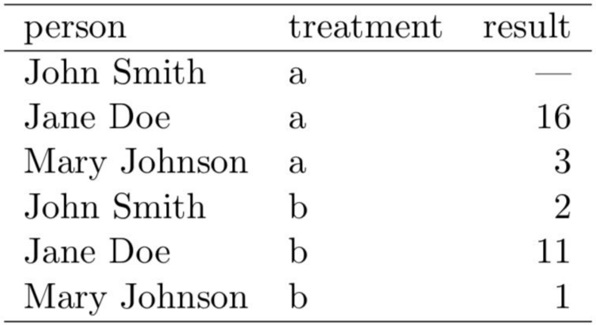
Tidy data

“each variable is a column, each observation is a row, and each type of observational unit is a table.”



	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Tidy data



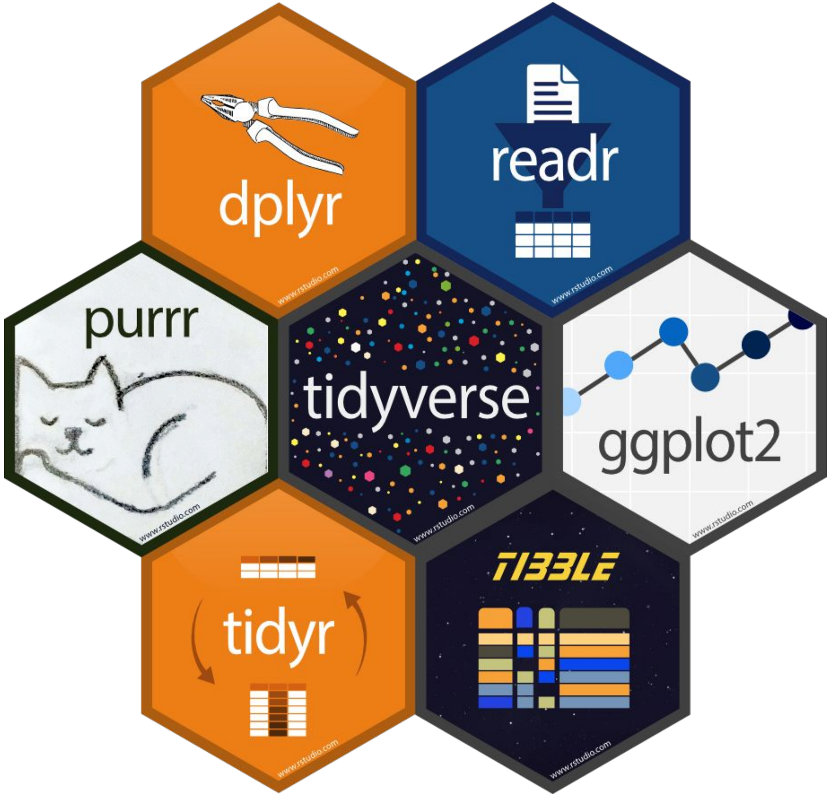
person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

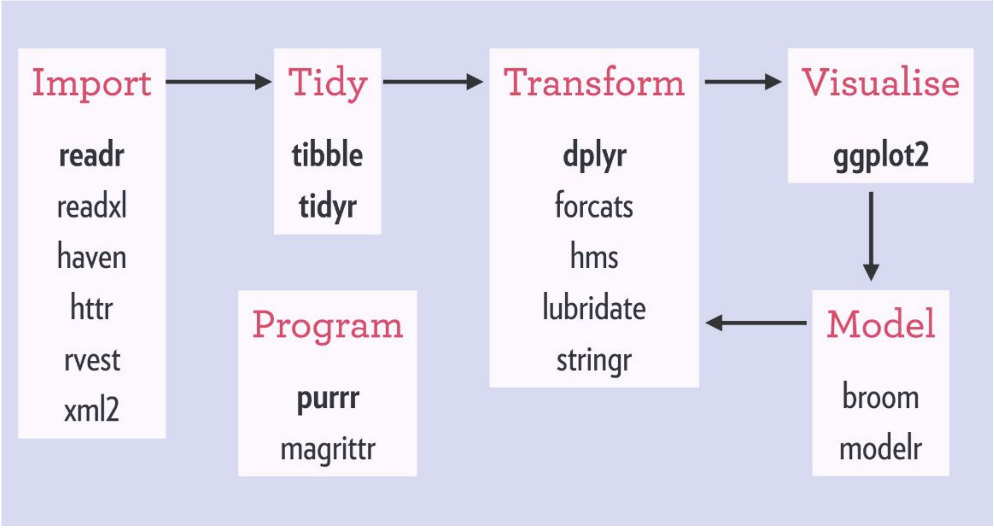
Help you use “tidyverse” packages in R; check out ggplot2, dplyr, tidyr, etc.

Use the most intuitive/effective representations for the task.

Wickham (2014)

Tidyverse & Tidymodel





Import

readr
readxl
haven
httr
rvest
xml2

Tidy

tibble
tidyr

Transform

dplyr
forcats
hms
lubridate
stringr

Visualise

ggplot2

Model

broom
modelr

Program

purrr
magrittr

https://www.tidyverse.org/

Exploratory Data Analysis

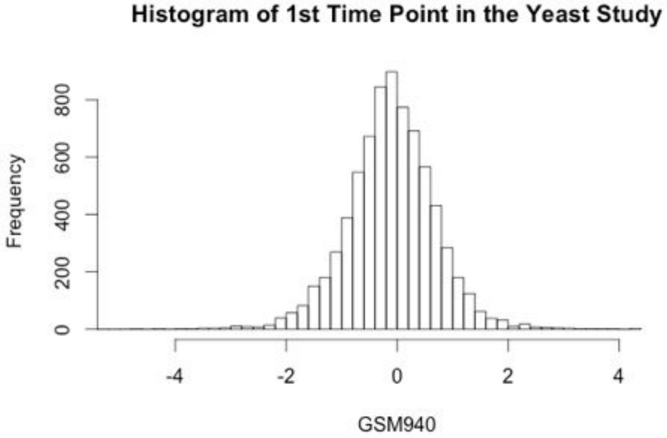
J. W. TUKEY wrote a book titled Exploratory Data Analysis, in 1977

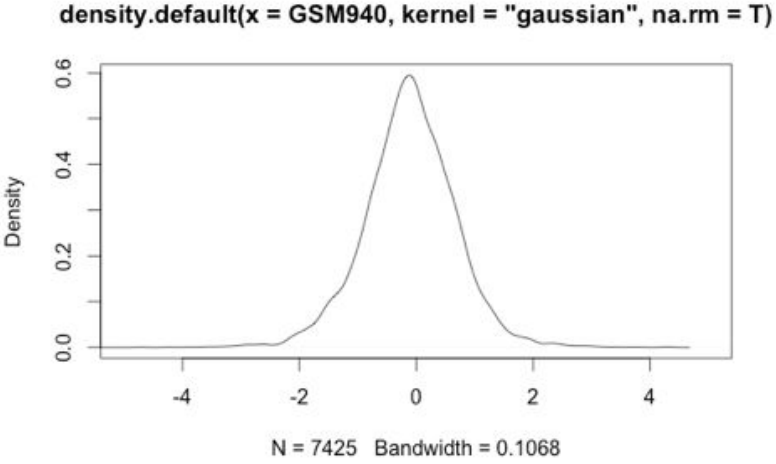
Some of Tukey’s inventions include

- Boxplot
- Jackknife estimation
- Fast Fourier Transform (FFT) algorithm
- Naming of 'bit'

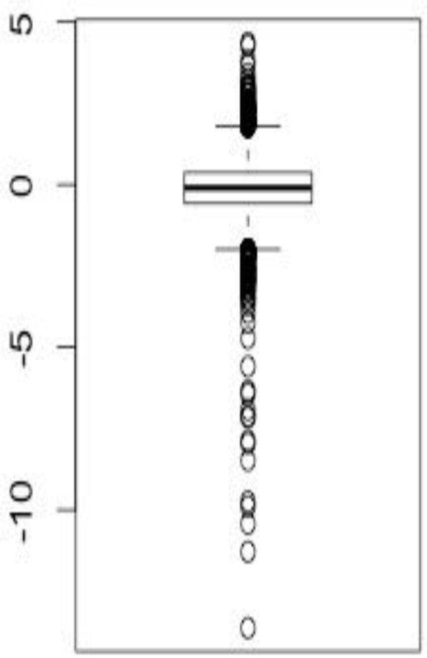
Histogram & Density

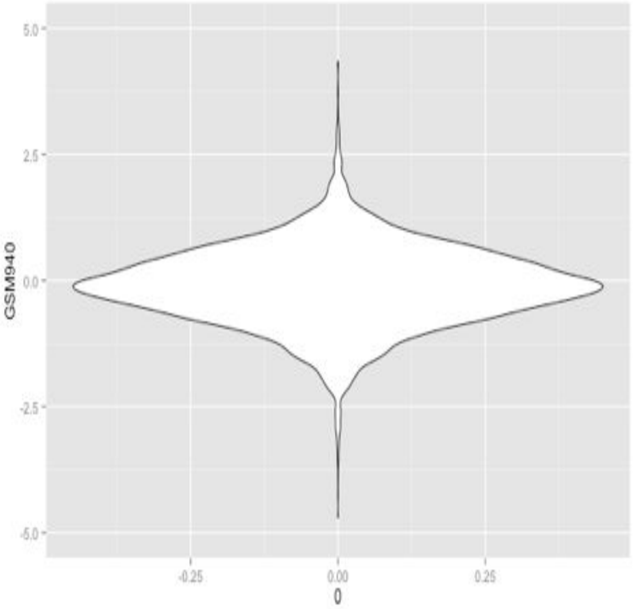
Visualizing the distribution of the data





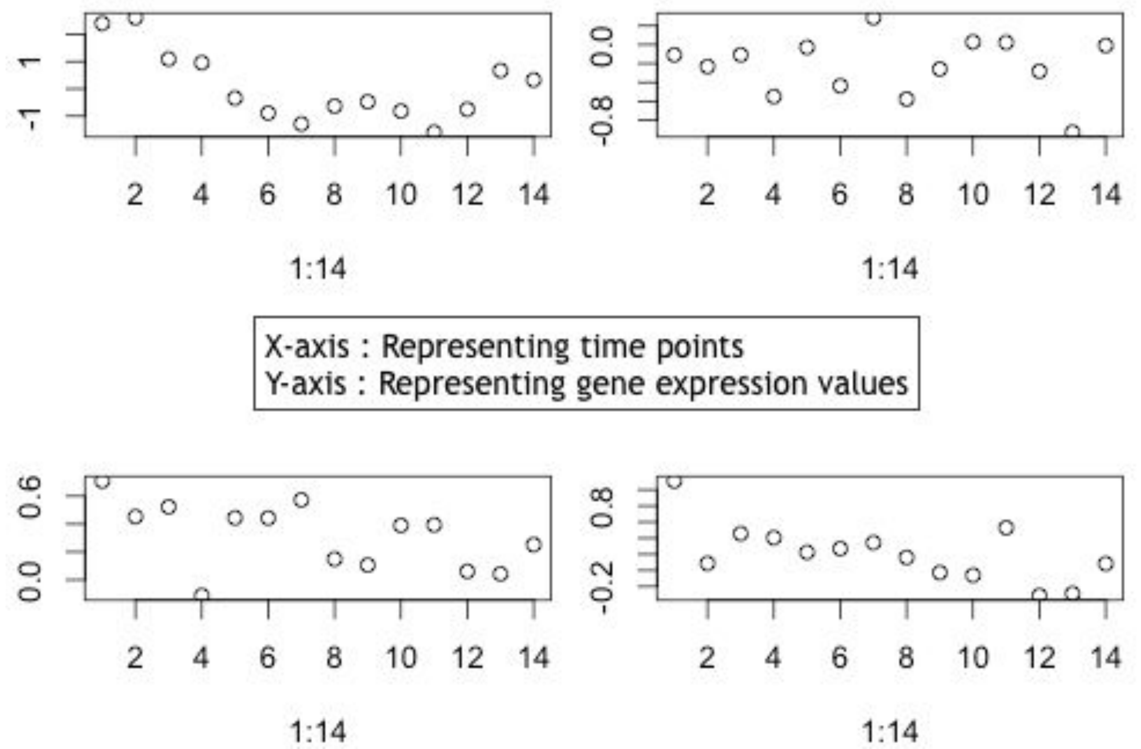
Boxplot & Violinplot





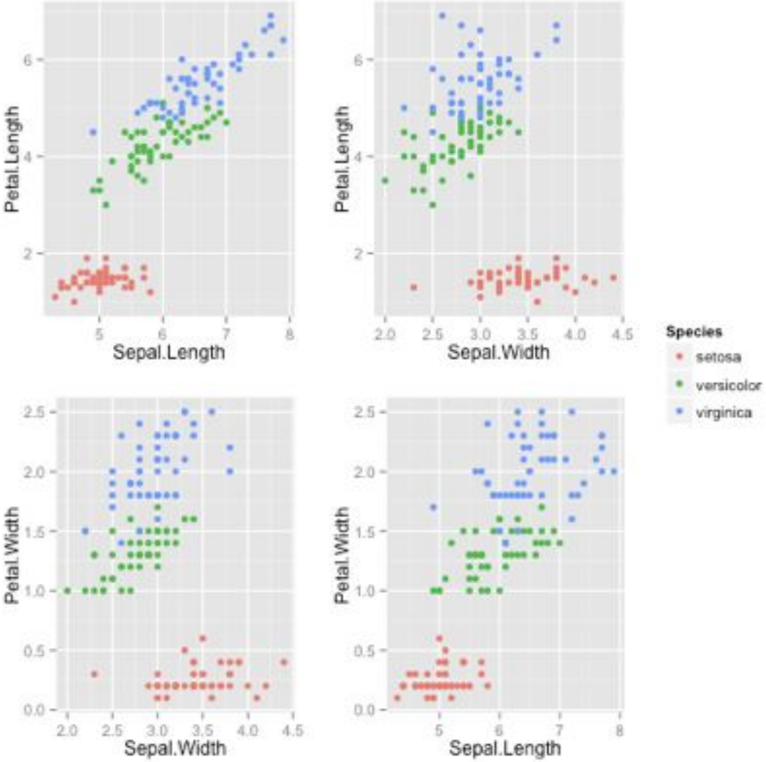
Scatterplot

Visualizing the relationship between two variables

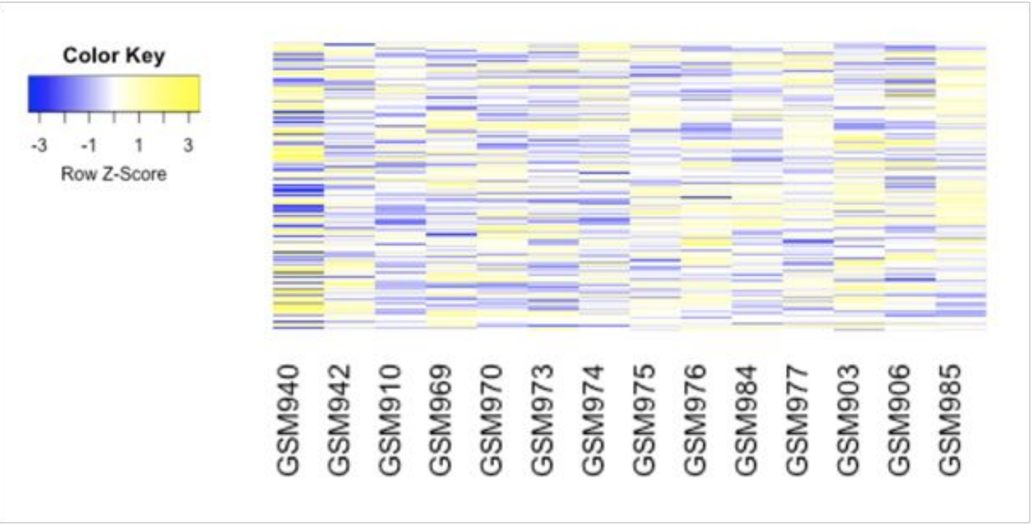


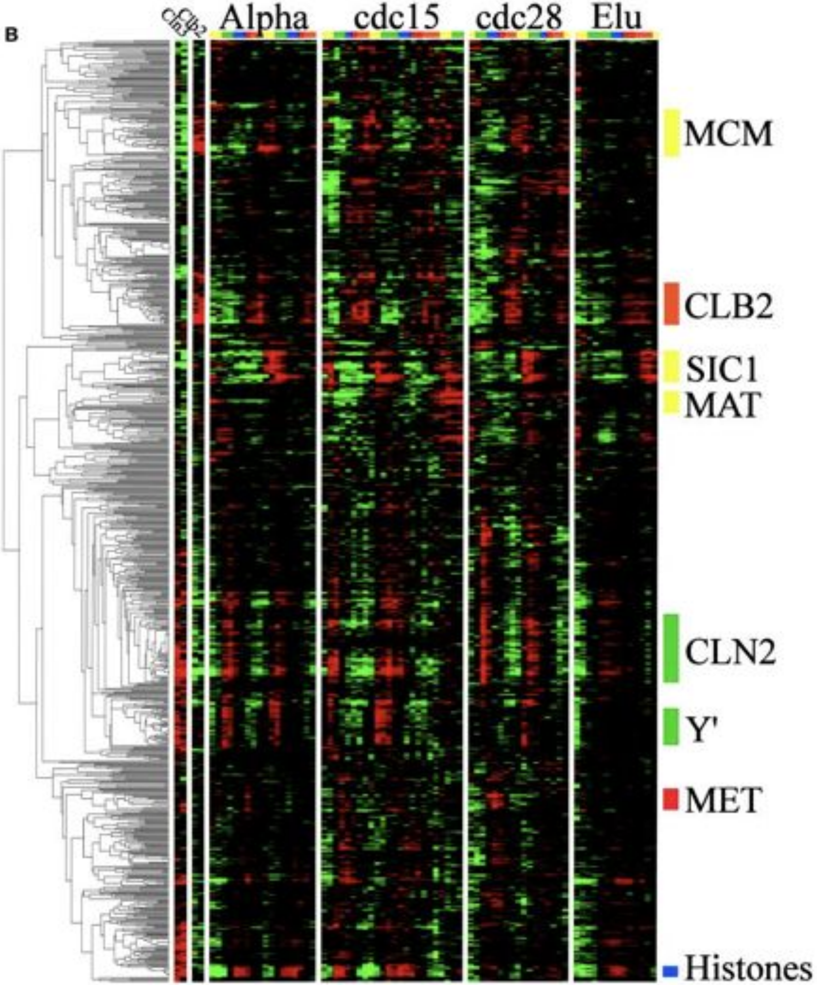
X-axis : Representing time points
Y-axis : Representing gene expression values

Scatterplot by groups



Heatmaps


Color Key
-3
-1
1
3
Row Z-Score
GSM940
GSM942
GSM910
GSM969
GSM970
GSM973
GSM974
GSM975
GSM976
GSM984
GSM977
GSM903
GSM906
GSM985


B
Cln3
Clb2
Alpha
cdc15
cdc28
Elu
MCM
CLB2
SIC1
MAT
CLN2
Y'
MET
Histones