3000788 Intro to Comp Molec Biol

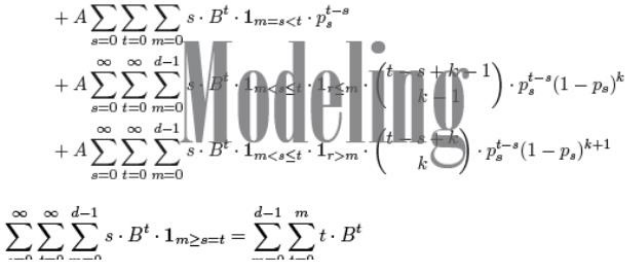
Week 1: Course introduction and logistics

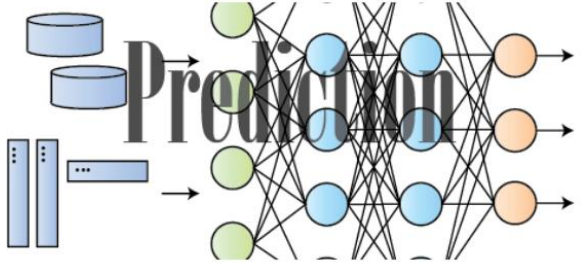
Fall 2024

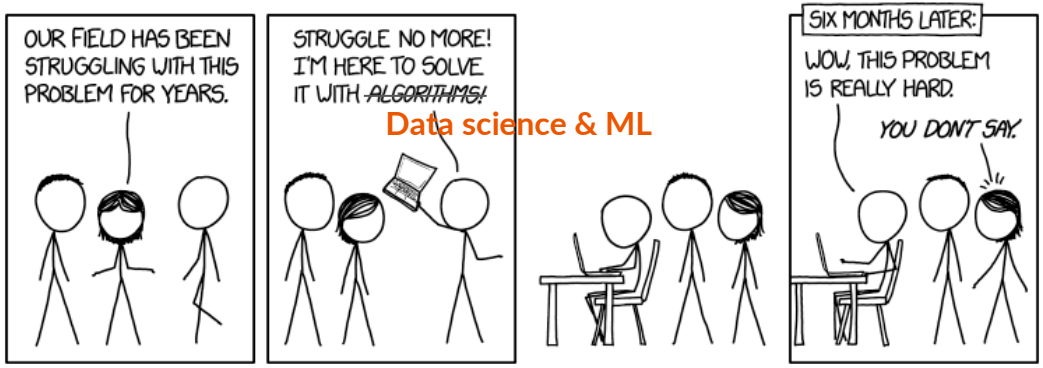
Sira Sriswasdi, PhD

- Research Affairs, Faculty of Medicine, Chulalongkorn University
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

About instructor







Credit: "Here to Help" from xkcd comic, reprinted under Creative Commons License

Data science & ML

- BS in Mathematics
- PhD in Computational Biology

Keywords

- Molecular Evolution
- Proteomics / Transcriptomics
- Machine Learning

About you

Please introduce yourself

- Name & nickname
- Graduate program & year
- Undergraduate background
- Research interest
- Thesis advisor & topic (if you already picked)

Fill out the pre-course questionnaire:
https://docs.google.com/forms/d/e/1FAIpQLSfXPgzsG5_Z7HvGTpsKw_mnwetwrfNkrzZ3-iC7b40R9XD2hw/viewform

About this course

- Survey broad topics in computational molecular biology
- Provide foundation for understanding deeper courses
  - Always ask “Why?”

- Mixing discussion and practice
- Exposure to basic bioinformatics workflows

- Some taste of
  - Python programming
  - Machine learning & data science

- Go to https://github.com/cmb-chula/comp-biol-3000788/

Course schedule

Module	Date	Topics	Assignment
1	August	DNA sequencing & applications	Problem set 1-2
2	September	Transcriptomics: bulk and single-cell	Problem set 3-5
3	October	Other omics - Proteomics - Chromatin organization - Biological networks	Problem set 6-8
4	November	Programming & machine learning	Problem set 9-10
		Post-course evaluation	

Course schedule (Thai section)

Session	Date	Time	Topics	Videos	Assignment
1	08/14/24	1-3pm	Course introduction	-	Pre-course evaluation
2	08/21/24	1-3pm	DNA sequencing and data processing	L4 / L5	PS1
3	08/28/24	1-3pm	Sequence alignment / Phylogenetics	L6 / L7	PS2
4	09/04/24	1-3pm	Metagenomics / Microarray	L8 / L9	PS3
5	09/11/24	1-3pm	RNA-seq / Expression data analysis	L11 / L13	PS4
6	09/18/24	1-3pm	RNA-seq analysis demo	-	PS5
7	09/25/24	1-3pm	Single-cell technique and demo	L14	-
8	10/02/24	1-3pm	Proteomics and demo	L16	PS6
9	10/09/24	1-3pm	Biological networks / Chromatin	L18 / L20	PS7
10	10/16/24	1-3pm	Dynamics simulation / Online resources	L21 / L22	PS8
-	10/23/24	-	-	-	-
11	10/30/24	1-3pm	Python programming session 1	L23 / L24	-
12	11/06/24	1-3pm	Python programming session 2	L25 / L26	PS9
13	11/13/24	1-3pm	Principles of machine learning	L27 / L28	-
14	11/20/24	1-3pm	Python programming session 3	L29	PS10
15	11/27/24	1-3pm	Deep learning in life sciences	L30	Post-course evaluation

Course schedule (English section)

Session	Date	Time	Topics	Videos	Assignment
1	08/16/24	1:30-3:30pm	Course introduction	-	Pre-course evaluation
2	08/23/24	1:30-3:30pm	DNA sequencing and data processing	L3 / L4	PS1
3	08/30/24	1:30-3:30pm	Sequence alignment / Phylogenetics	L5 / L6	PS2
4	09/06/24	1:30-3:30pm	Metagenomics / Microarray	L7 / L8	PS3
5	09/13/24	1:30-3:30pm	RNA-seq / Expression data analysis	L10 / L12	PS4
6	09/20/24	1:30-3:30pm	RNA-seq analysis demo	-	PS5
7	09/27/24	1:30-3:30pm	Single-cell technique and demo	L13	-
8	10/04/24	1:30-3:30pm	Proteomics and demo	L15	PS6
9	10/11/24	1:30-3:30pm	Biological networks / Chromatin	L17 / L18	PS7
10	10/18/24	1:30-3:30pm	Dynamics simulation / Online resources	L20	PS8
11	10/25/24	1:30-3:30pm	Python programming session 1	L21 / L22	-
12	11/01/24	1:30-3:30pm	Python programming session 2	L23 / L24	PS9
13	11/08/24	1:30-3:30pm	Principles of machine learning	L25 / L27	-
14	11/15/24	1:30-3:30pm	Python programming session 3	L26 / L28	PS10
15	11/22/24	1:30-3:30pm	Deep learning in life sciences	L29	Post-course evaluation

Learning strategy

- Before class
  - **Watch assigned videos** (1-2 per week, 1.5hr each)
  - **Submit your pre-class query**: https://docs.google.com/forms/d/e/1FAIpQLSdkL8Kx9Z1L41ZJeLwu0jDXOOZ7wZ1Gc8NbvQH2nQ87ytV1Yw/viewform
  - **Download data and install software**

- In class
  - **Lecture-style recap** of key topics
  - **Q & A** of your pre-class query
  - **Discussion**

Grading criteria

- Problem set [9% x 10 problem sets = 90%]
  - Can work with each other
  - But write your own answer
  - You may use ChatGPT but also report how you used it

- In-class activity [10%]
  - Pre-class query
  - Hands-on practice
  - Discussion
  - Attendance

Companion courses from MIT (7.36)

FOUNDATIONS OF COMPUTATIONAL AND SYSTEMS BIOLOGY

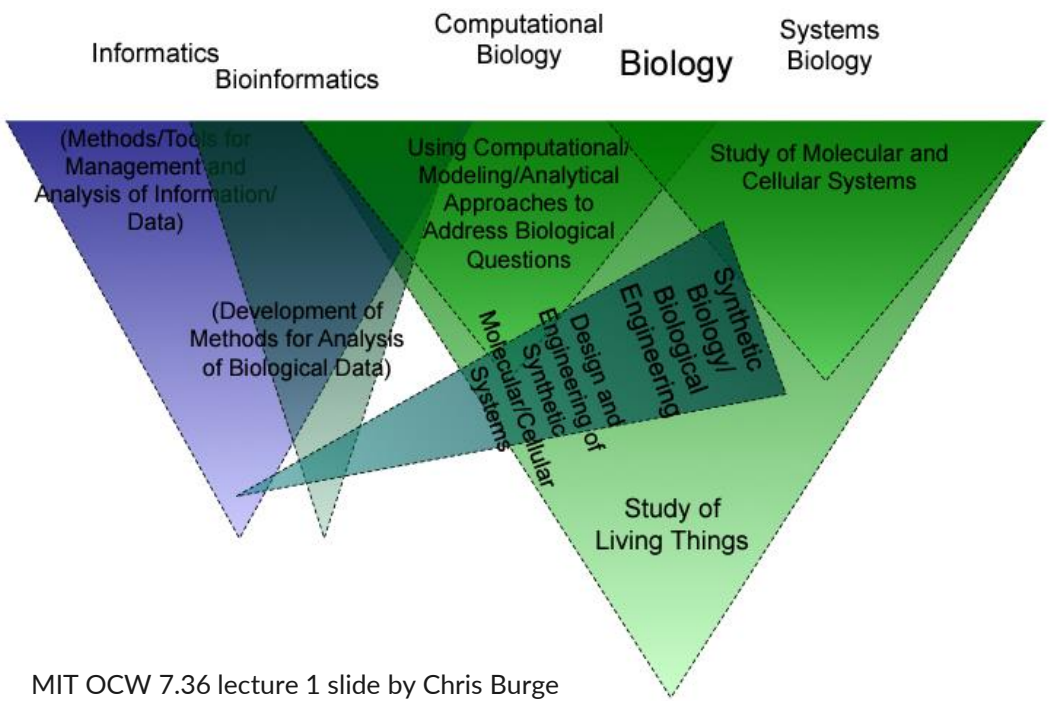
Lecture 1: Introduction to Computational and Systems Biology

Lecture 2: Local Alignment (BLAST) and Statistics

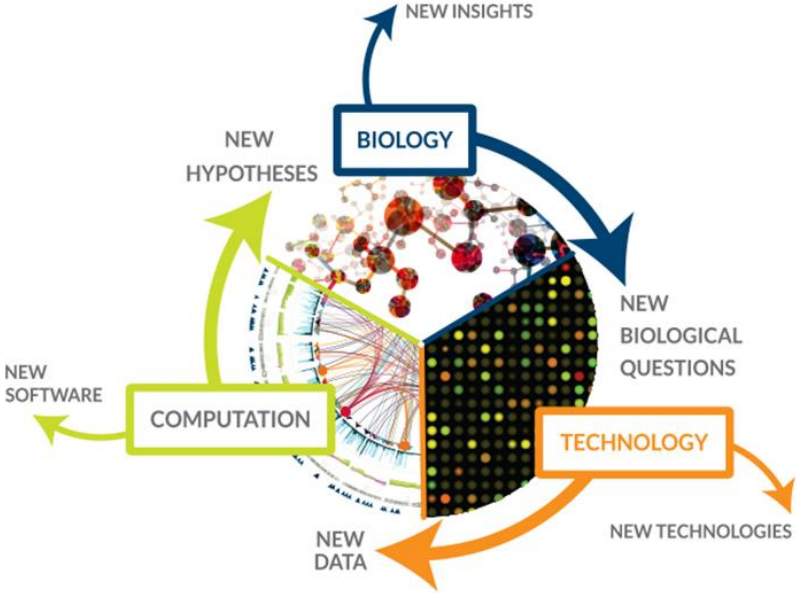
Lecture 3: Global Alignment of Protein Sequences (NW, SW, PAM, BLOSUM)

Lecture 4: Comparative Genomic Analysis of Gene Regulation

What is computational biology?



MIT OCW 7.36 lecture 1 slide by Chris Burge



www.systemsbiology.org/about/what-is-systems-biology

Biology-inspired motif analysis

Castellanos *et al*. Nat Comm 2020

How to identify the motif?

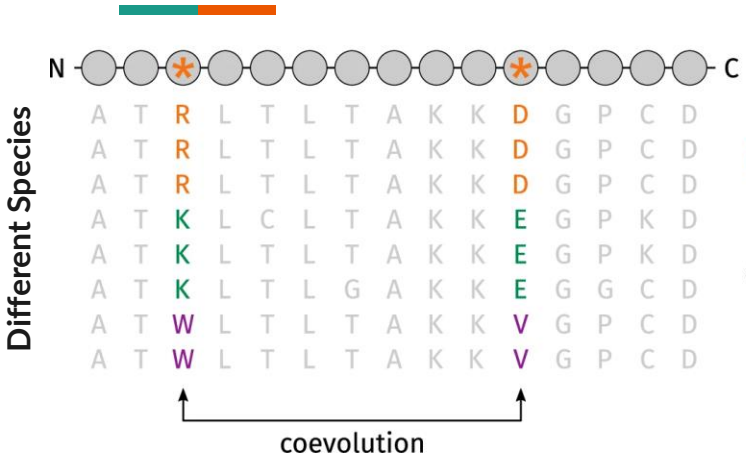
Experiment can isolate DNA sites with bound protein

Gibbs Sampling to identify similar patterns

CGGGGCTATcCAgCTGGGTCGTCACATTCCCCTT
TTTGAGGGTGCCCAATAAggGCAACTCCAAAGCGGACAAA
GGATGgAtCTGATGCCGTTTGACGACCTA
AAGGAaGCAACcCCAGGAGCGCCTTTGCTGG
AGATTATAATGTCGGTCCtTGgAACTTC
CAACTGAGATCATGCTGCATGCcAtTTTCAAC
TACATGATCTTTTGATGgcACTTGGATGAGGGAATGATGC

An Introduction to Bioinformatics Algorithm by Jones and Pevzner

Capturing co-evolution with mutual information

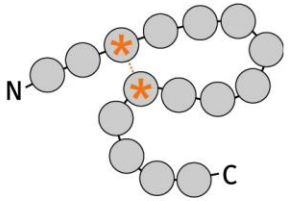


$$I(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_{(X,Y)}(x,y) \log \left(\frac{P_{(X,Y)}(x,y)}{P_X(x)\, P_Y(y)} \right)$$

Mutual Information

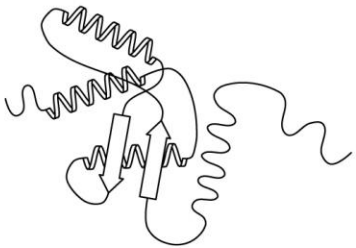
inference

constraint

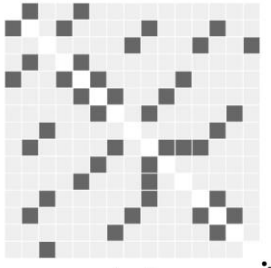


contact in 3D

Positions in protein(s) co-evolve because they interact in some fashions – usually physical bindings



reconstruct

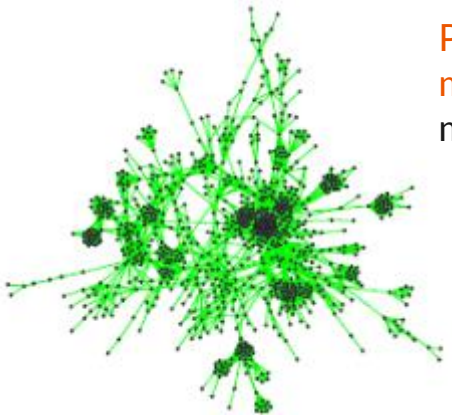


contact map

Co-evolution can be quantified and used to guide biological inferences

Bittrich *et al.* Sci Rep 2019

Using graph theory to study biological networks

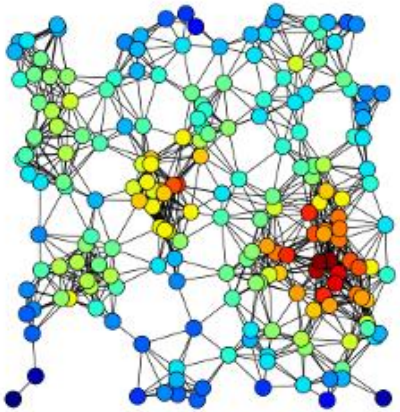


Proteins involved in many interactions might be important

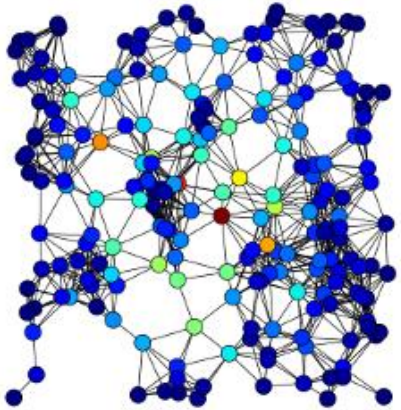
Proteins that connect other proteins might be important

Node = Protein
Edge = Protein-protein interaction

Degree

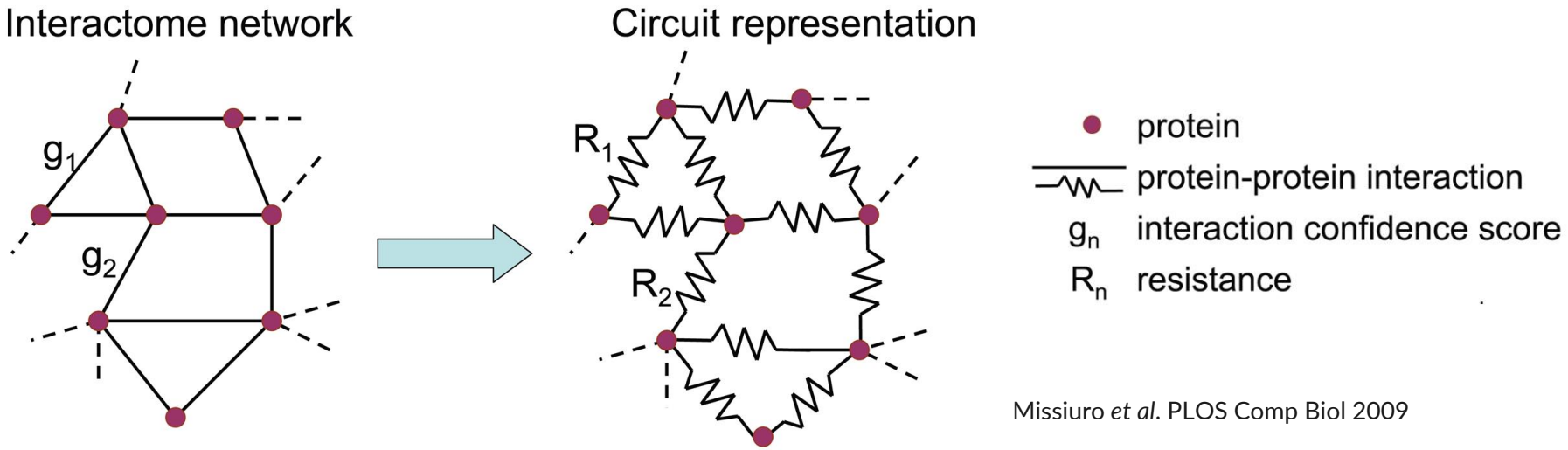


Betweenness



Images from wikipedia

Electrical circuit model for biological signal flows

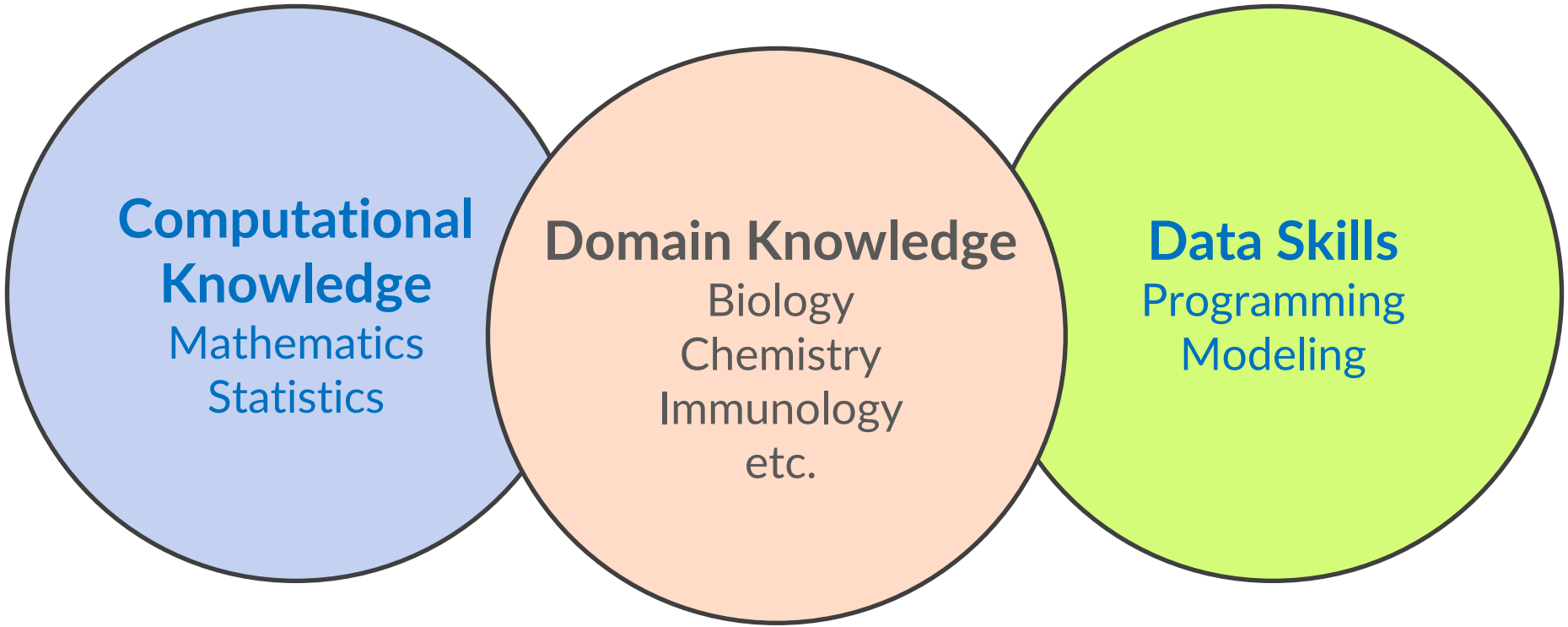


Missiuro *et al*. PLOS Comp Biol 2009

To determine how information flows between two nodes

Apply Kirchoff's laws to calculate the current through the circuit

Course objectives



Knowledge enables communication

Module 0: Statistics and computational thinking

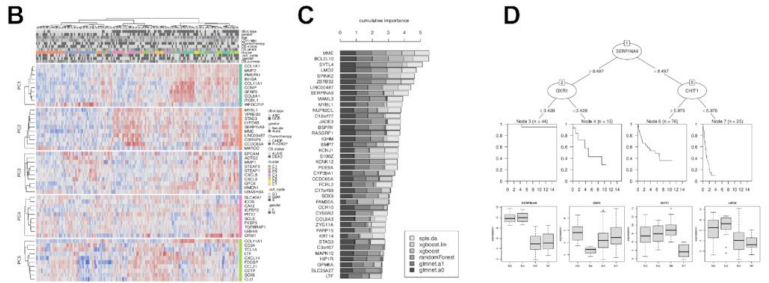
- P-values help distinguish biological pattern from random chance
  - Differential test of gene expression
  - Enrichment of biological function terms among differentially expressed genes

- How were they calculated?
  - Null hypothesis
  - Correction for multiple testing

- What can you do with computational thinking?
  - Data analysis
  - Modeling and simulation
  - Understand algorithm / workflow

What can you do with computational thinking?

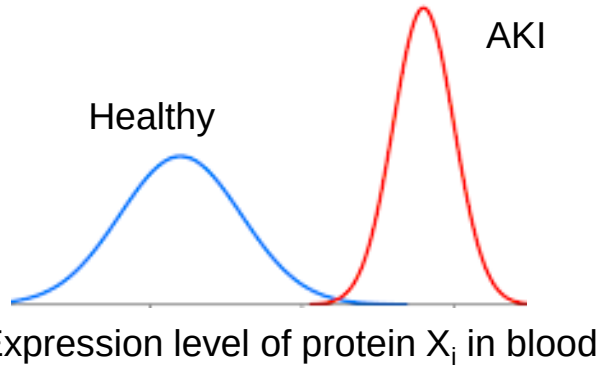
- **Data analysis**
  - Identify calculations that will support or disprove your hypothesis
  - Be aware of assumptions and limitations of each calculation
- Modeling and simulation
- Algorithm



Akhmedov, M. et al. NAR Genom and Bioinfor, 2(1):lqz019 (2019)

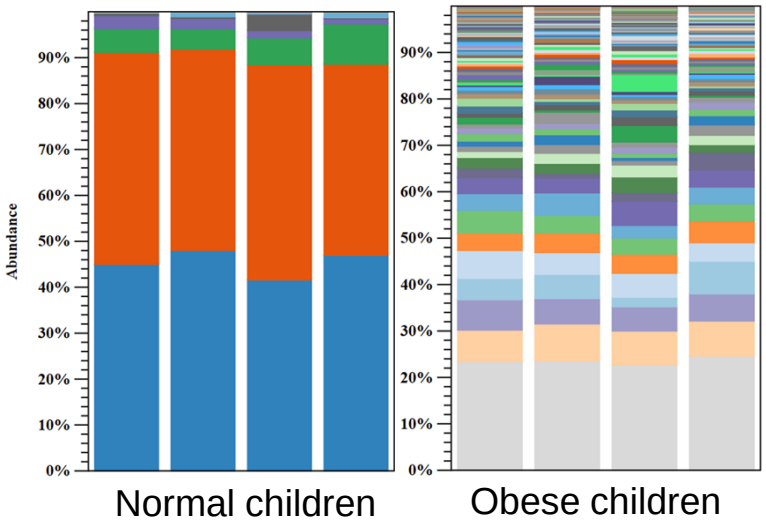
Gene ID	P61_2_C	P62_2_C	P63_2_C	P64_2_C	P68_2_C
ENSG00000000003.14	4.637576	6.183992	5.237635	2.372719	5.665966
ENSG00000000005.5	0	0	0	0	0
ENSG00000000419.12	11.22781	4.813792	2.99782	10.99452	10.7482
ENSG00000000457.13	7.656414	5.082675	7.710682	9.014404	8.488388
ENSG00000000460.16	3.172546	2.245954	5.974815	3.501081	4.162024
ENSG00000000938.12	0	0	0	0.042488	0
ENSG00000000971.15	6.626259	8.19511	5.904925	11.7748	2.050394
ENSG00000001036.13	1.790445	0.76823	3.670635	0.68115	1.894823
ENSG00000001084.11	19.53907	25.08378	11.04872	5.815902	20.23763
ENSG00000001167.14	15.34717	20.00867	17.10001	25.31168	27.41216
ENSG00000001460.17	0.889852	3.090642	0.744581	3.439525	2.417934
ENSG00000001461.16	3.771195	3.12468	1.385353	2.767444	2.973217
ENSG00000001497.16	16.75059	9.662455	15.4965	14.34071	10.62035
ENSG00000001617.11	2.998366	3.712208	3.885852	17.50663	3.019686

Quantifying the “goodness” of biomarkers



- Identify significantly differentially expressed proteins, X_1, X_2, …, X_{200} with *t*-test
- What if we want to identify the **best biomarkers**?
  - How to quantify the ability of a protein to distinguish healthy and AKI?
  - Independent two-sample *t*-test’s statistics $= \frac{\text{Mean}_1 - \text{Mean}_2}{\sqrt{\frac{1}{n}(\text{Variance}_1 + \text{Variance}_2)}}$

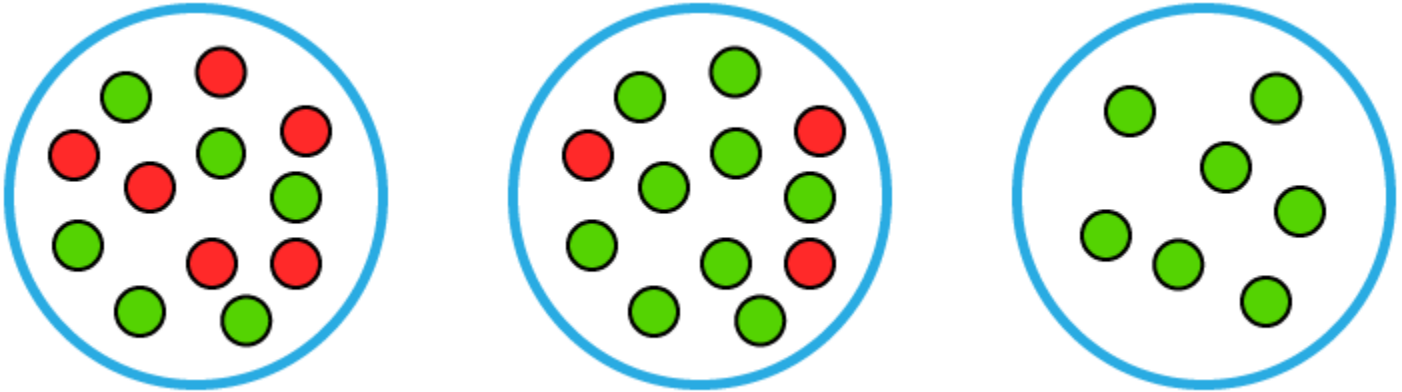
Quantifying diversity of gut microbiomes



Colors indicate different microbial taxa

- Visually, microbial taxa distributions in obese children are clearly more diverse
- Can we quantify this pattern?
  - Number of taxa
  - **Entropy** = $-\text{Frequency}_1 \log_2(\text{Frequency}_1) - \cdots - \text{Frequency}_n \log_2(\text{Frequency}_n)$

Entropy quantifies purity of a mixture

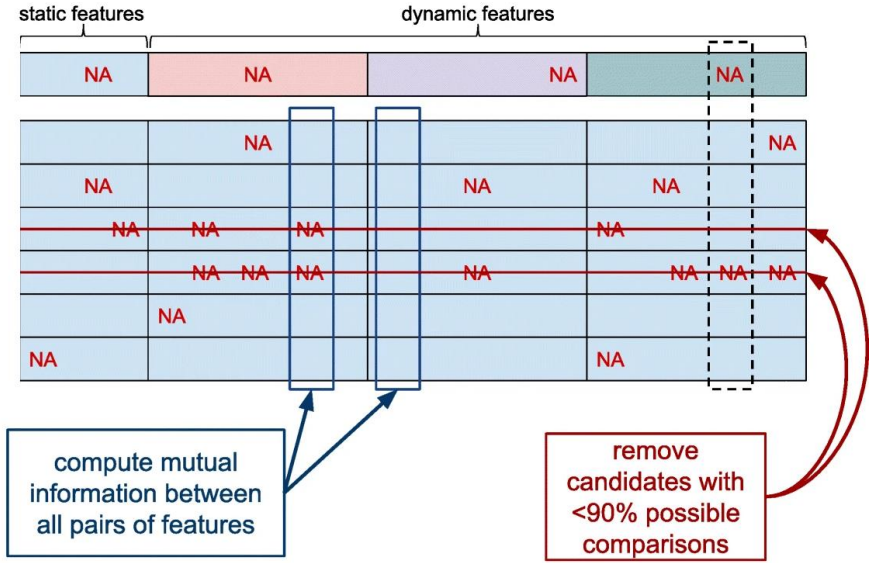


https://www.javatpoint.com/entropy-in-machine-learning

- **Entropy** = $-\text{Freq}_1 \log_2(\text{Freq}_1) - \cdots - \text{Freq}_n \log_2(\text{Freq}_n)$
  - Left sample: $-\frac{6}{12}\log_2\left(\frac{6}{12}\right) - \frac{6}{12}\log_2\left(\frac{6}{12}\right) = \log_2(2) = 1$
  - Middle sample: $-\frac{3}{12}\log_2\left(\frac{3}{12}\right) - \frac{9}{12}\log_2\left(\frac{9}{12}\right) = 0.811278$
  - Right sample: $-\frac{0}{12}\log_2\left(\frac{0}{12}\right) - \frac{12}{12}\log_2\left(\frac{12}{12}\right) = 0$

Mutual Information ~ correlation for categorical data

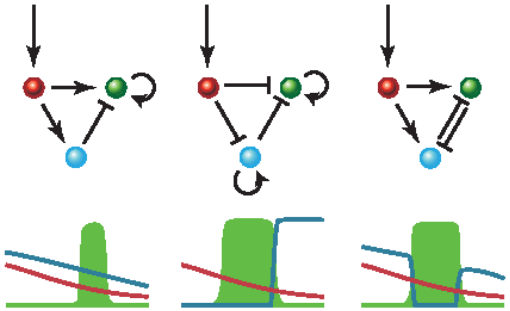
- MI = difference between P(X, Y) and P(X) P(Y)
- If X and Y are statistically independent
  - P(X, Y) = P(X) P(Y)
  - MI = 0
- $\mathrm{MI}(X;Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$

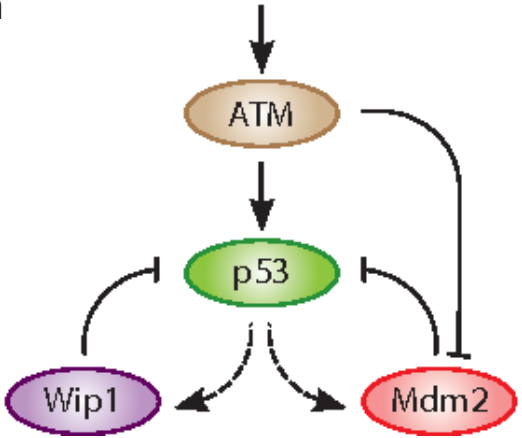


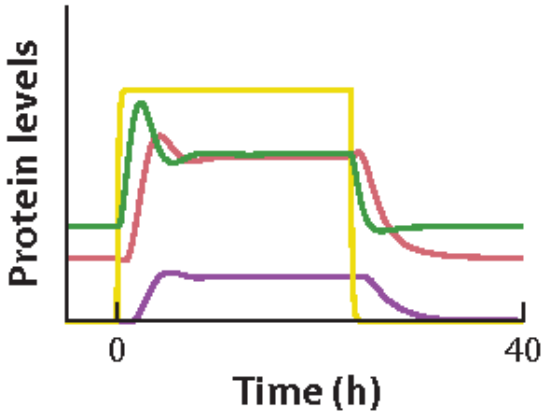
Tavazzi, E. et al. BMC Medical Informatics and Decision Making 2020

What can you do with computational thinking?

- Data analysis
- **Modeling and simulation**
  - Identify mechanisms that underlie the phenomenon or system of interest
  - Develop models
  - Study the behavior of the models / systems
  - Synthesize new data
- Algorithm

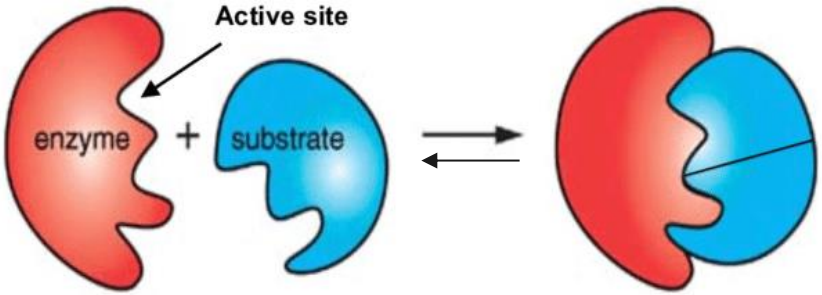






Rue and Garcia-Ojalvo. Annu Rev Biophys 42:605-627 (2013)

Enzyme substrate binding



Graham Hutchings. “Development of new highly active nano gold catalysts for selective oxidation reactions” (2014)

- $K_{dissociation}$ = [E][S] / [E-S] at equilibrium

- Association = P(E meeting S) x $P_{binding}$ = k_1 [E][S]
- Dissociation = Number of E-S molecules x $P_{dissociating}$ = k_2 [E-S]
- At equilibrium, Association = Dissociation
  - k_1 [E][S] = k_2 [E-S] → $K_{dissociation}$ = k_2 / k_1

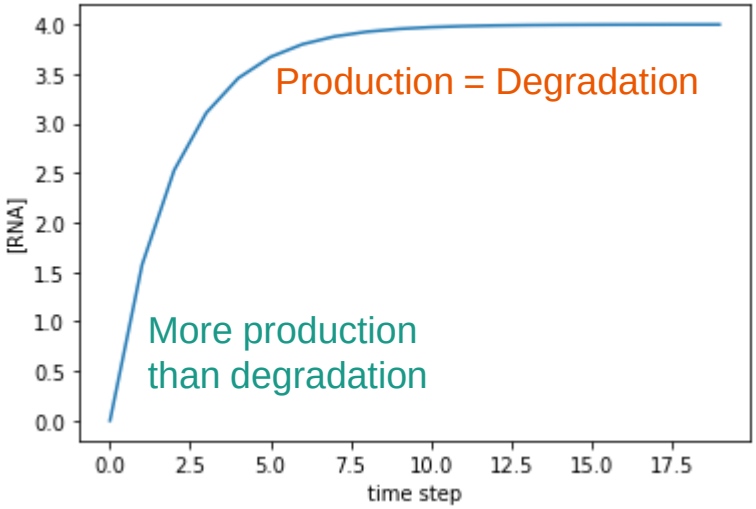
A simple gene expression model

$k_{transcription}$

$$DNA \rightarrow DNA + RNA$$

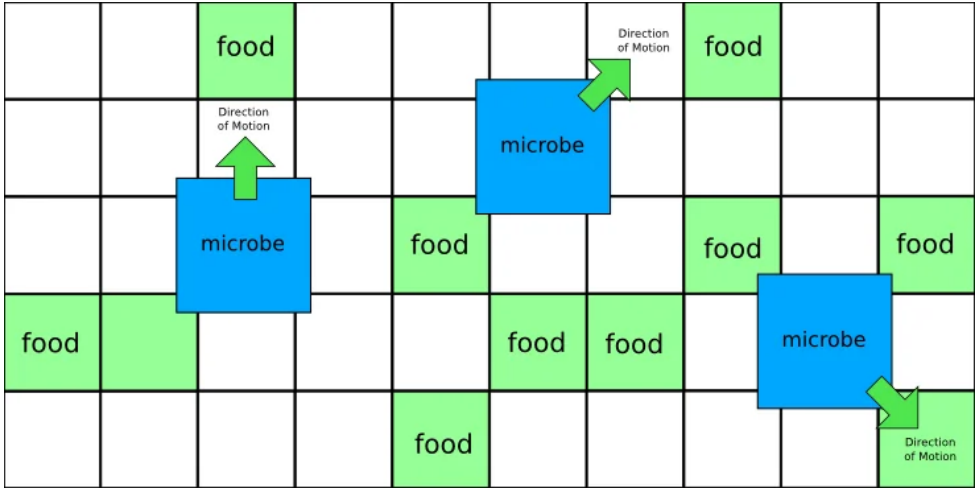
$$RNA \rightarrow \emptyset$$

$k_{degradation}$

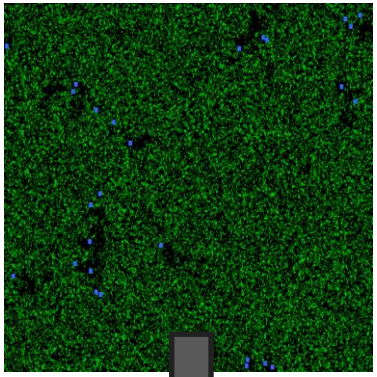


- $\frac{d[RNA]}{dt} = k_{transcription} - k_{degradation}[RNA]$
- This is called an **ordinary differential equation**

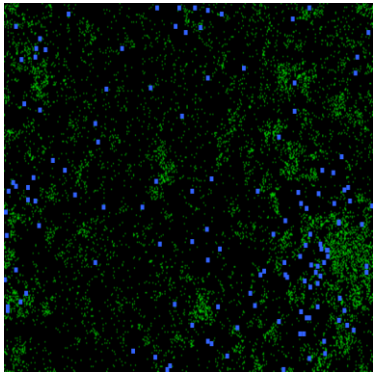
Microbial growth



Time = 0

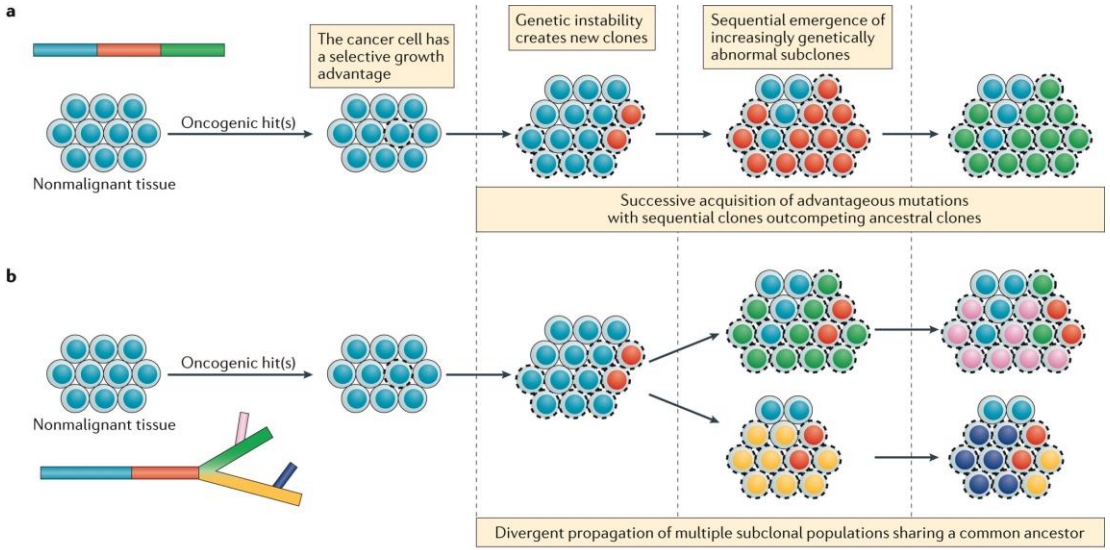


Time = 100



- Model different movement behaviors (probability of moving in a certain direction) with different genes
- See more at **https://beltoforion.de/en/simulated_evolution/**

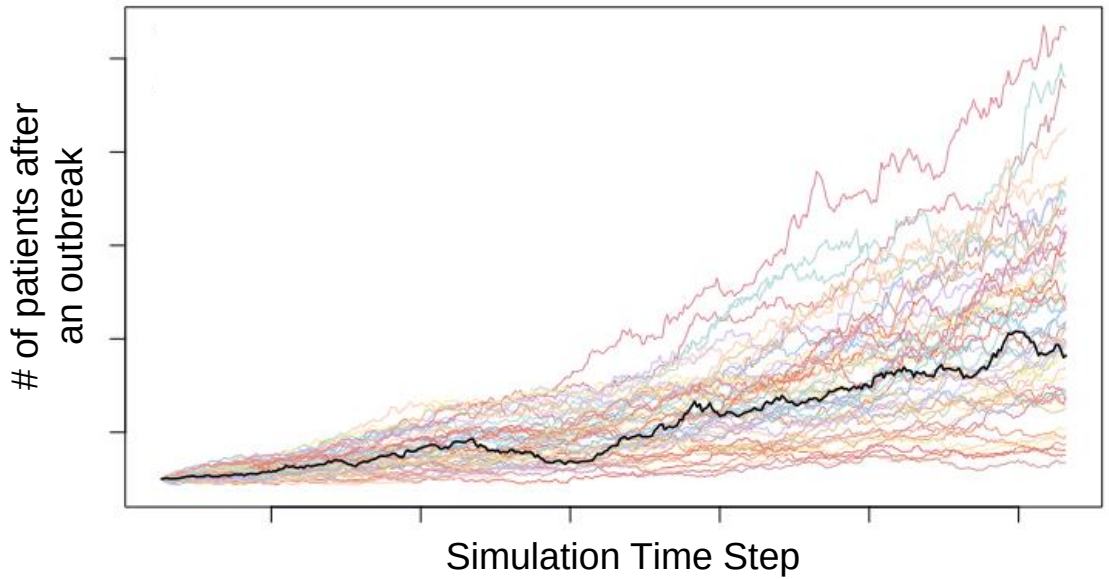
Tumor growth



Dagogo-Jack and Shaw, Nat Rev Clin Oncol (2017)

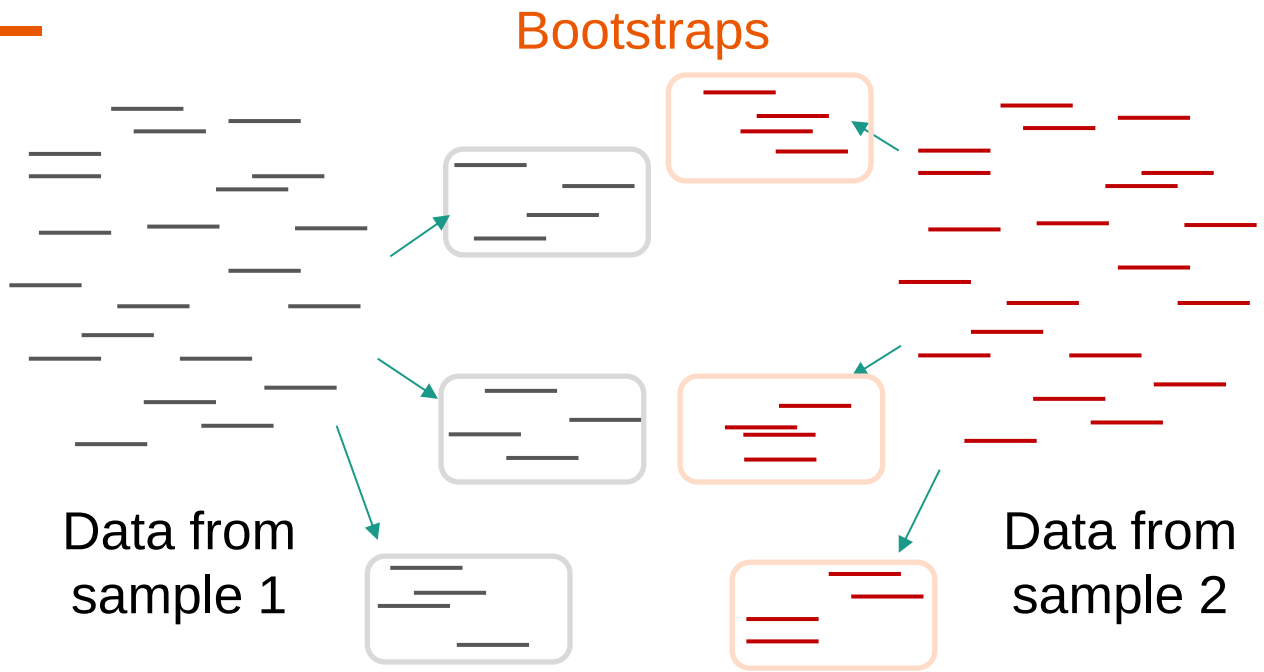
- Cell's actions: Gain mutation, Divide, Die
  - What are the parameters influencing these actions?

Monte Carlo technique



- Perform repeated sampling to explore broad parameter scenarios
- Provide estimates for the probability of different scenarios and outcomes
- May require millions of repetition (can utilize multiple CPUs)

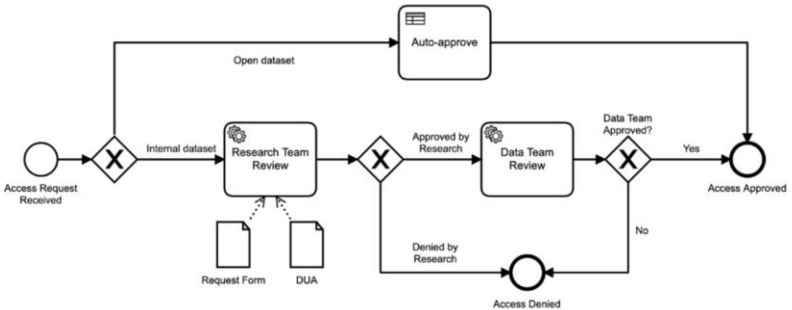
Bootstrapping



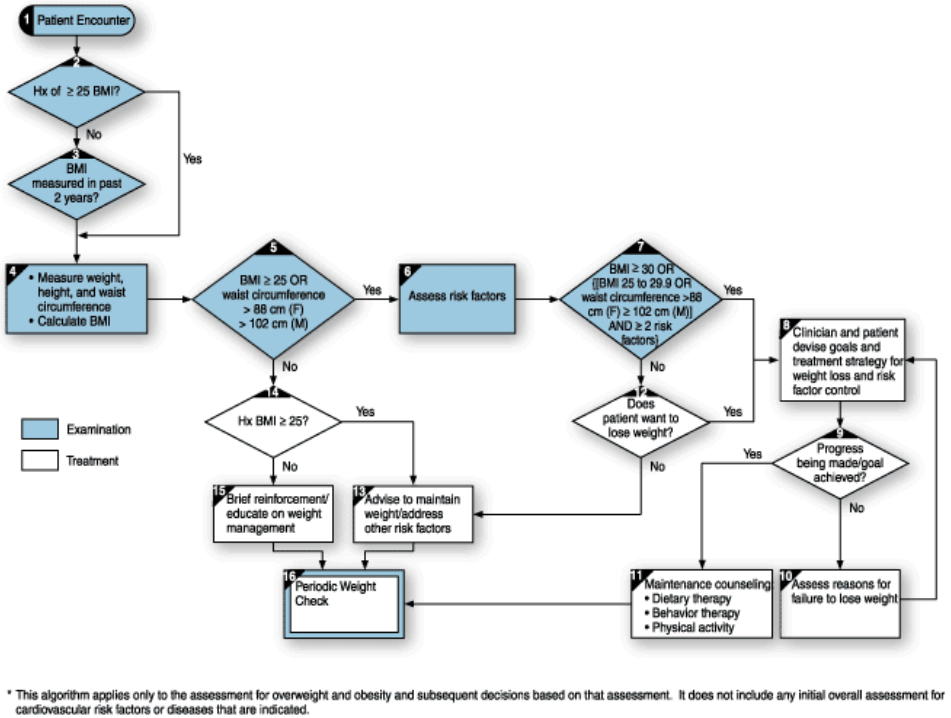
- Instead of using the whole data to perform one calculation
- Sampling from data to perform multiple calculations → Estimate variance

What can you do with computational thinking?

- Data analysis
- Modeling and simulation
- **Algorithm**
  - Formulating step-by-step instructions
  - Identifying weak points

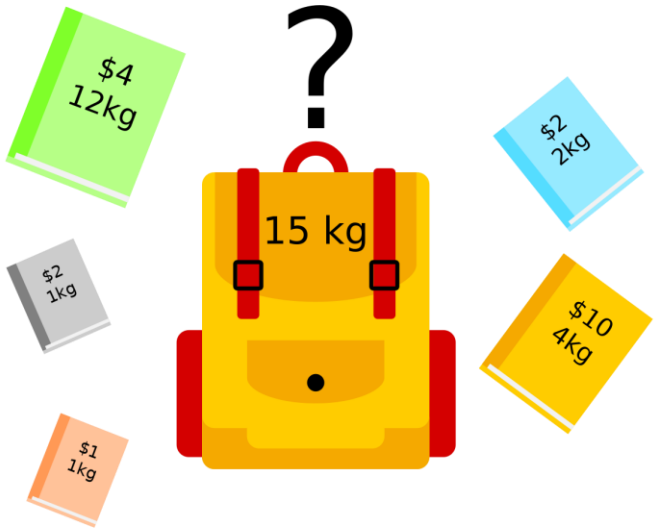


https://www.aridhia.com/blog/accessibility/



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

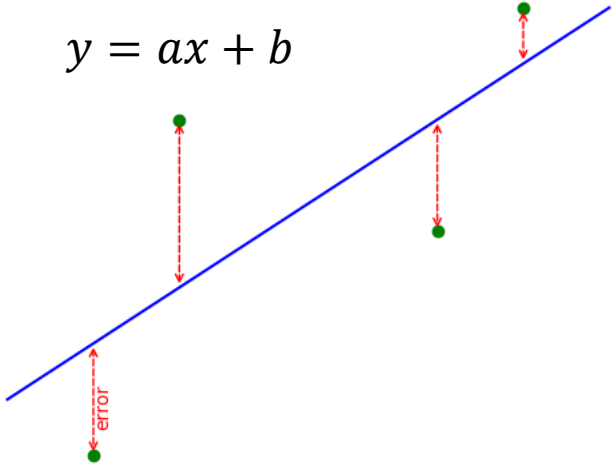
Knapsack problem



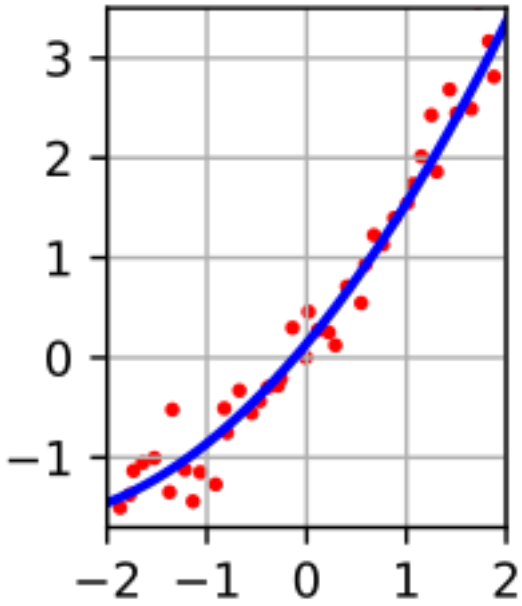
https://en.wikipedia.org/wiki/Knapsack_problem

- Limited capacity
- Want to pick as much value as possible while not exceeding the capacity
- General optimization setting
  - **Objective** = Total value
  - **Constraint** = Total weight ≤ W
- Paying **X** baht with the minimal number of coins and bank notes

Curve fitting with least square



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



$y = ax^2 + bx + c$

- Finding the **best** *a*, *b*, and *c* that make the curve fit to the observations
- Minimize least square error $\frac{1}{n}\left((\hat{y}_1 - y_1)^2 + \cdots + (\hat{y}_n - y_n)^2\right)$

Newton-Raphson method

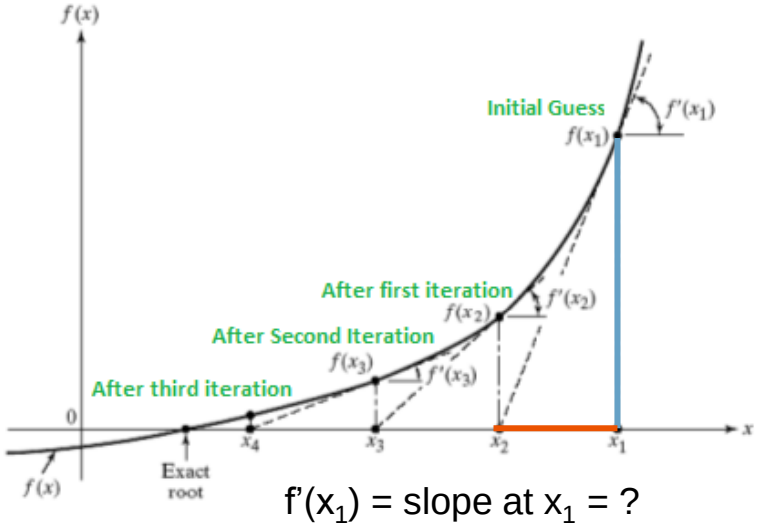
- Want to maximize L(x) by solving

$$f(x) = \frac{dL(x)}{dx} = 0$$

- Start with an initial guess x_1
- Calculate $f(x)$ and $f'(x)$ at x_1
- Define $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$

- Repeat the process for x_3, x_4, …
- Stop when x_i converges to a value

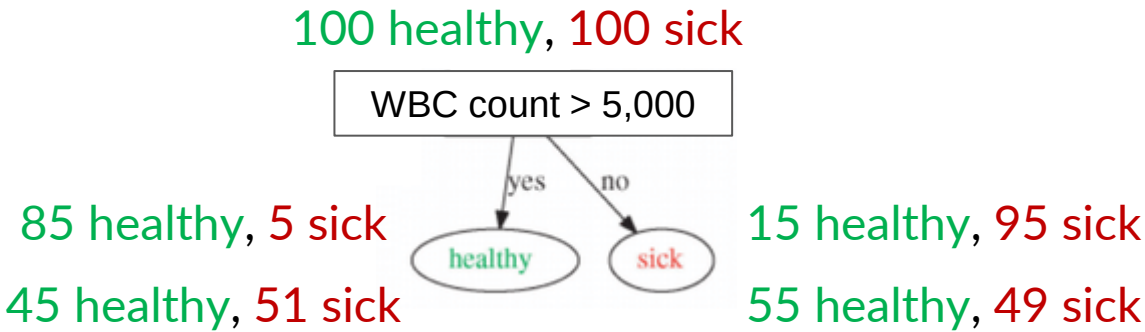
f(x) = the first derivative of the objective L(x)



f'(x_1) = slope at x_1 = ?

https://predictivehacks.com/newton-raphson-method-in-python/

Constructing a decision point



- Gini impurity: $\sum p(1-p)$
- Entropy: $-\sum p \ln(p)$
  - Minimal at p = 0 or 1 → Perfect split
  - Maximal at p = 0.5 → 50-50 split
- **Search for variable and cutoff** that strongly reduce impurity

(Almost) all algorithms involve optimization

- **Curve fitting** = minimize square error
- **Drug-protein docking** = minimize energy
- **Hospital queuing** = minimize wait time
- **Diagnosis** = maximize accuracy
- **Triaging** = maximize number of severe patients at the top of the list

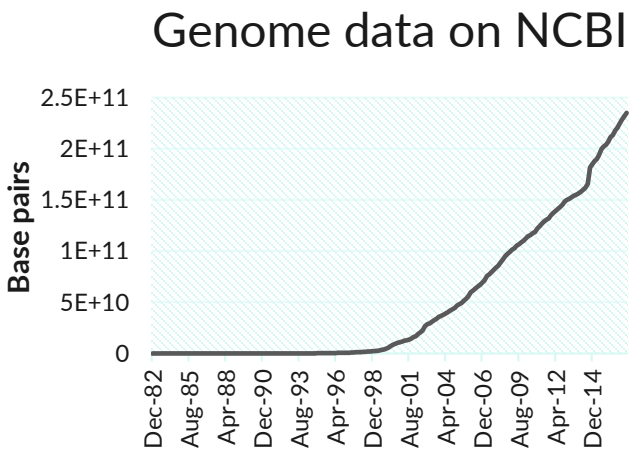
Module 1: DNA sequencing & applications

- What do you think kickstarted modern-day computational biology?

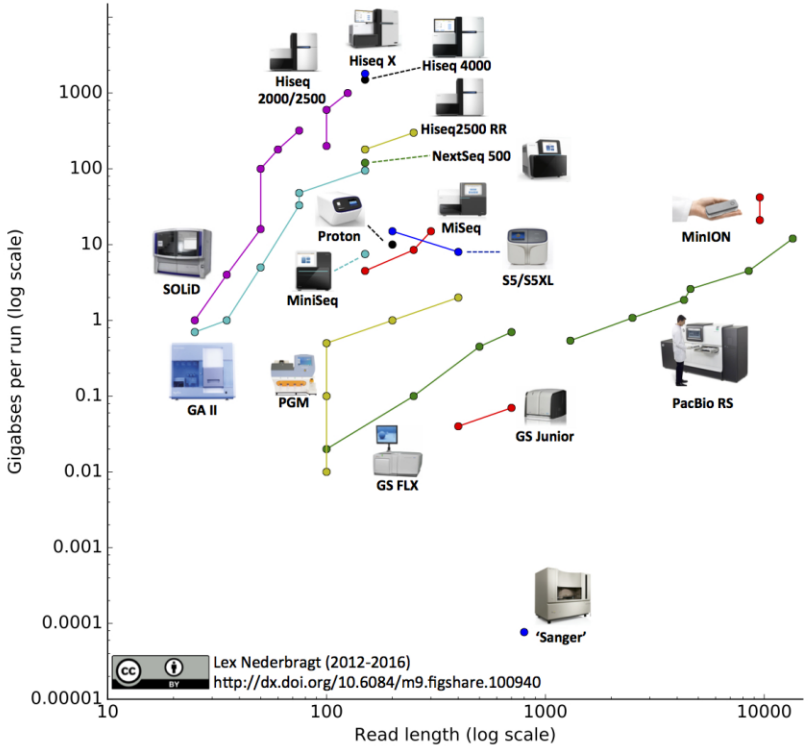
- First draft of human genome, 26 June 2000
  - BLAST sequence alignment

- Gene structure annotation
  - Exome sequencing
  - Oligo nucleotide microarray

Improvements in DNA sequencing

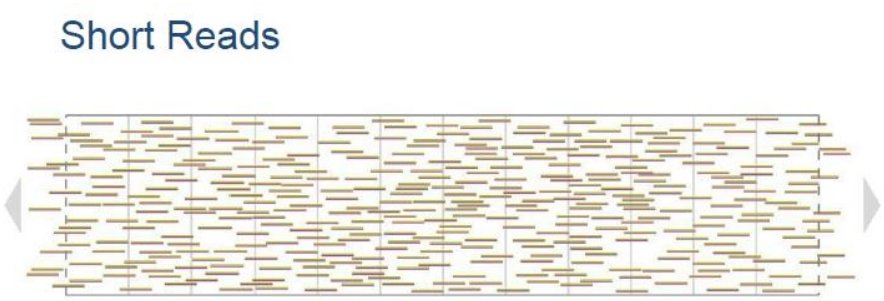


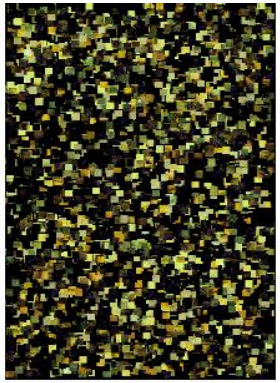
Plotted with data from
https://www.ncbi.nlm.nih.gov/genbank/statistics/



flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/

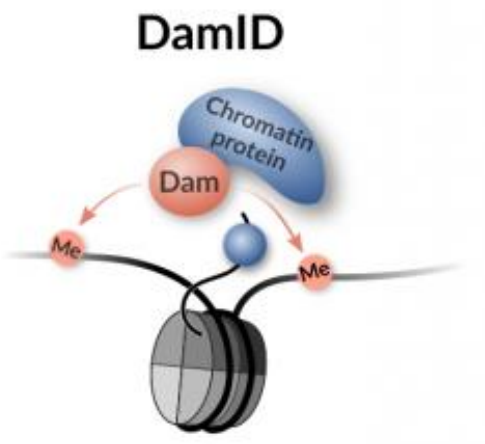
Combining short read and long read data



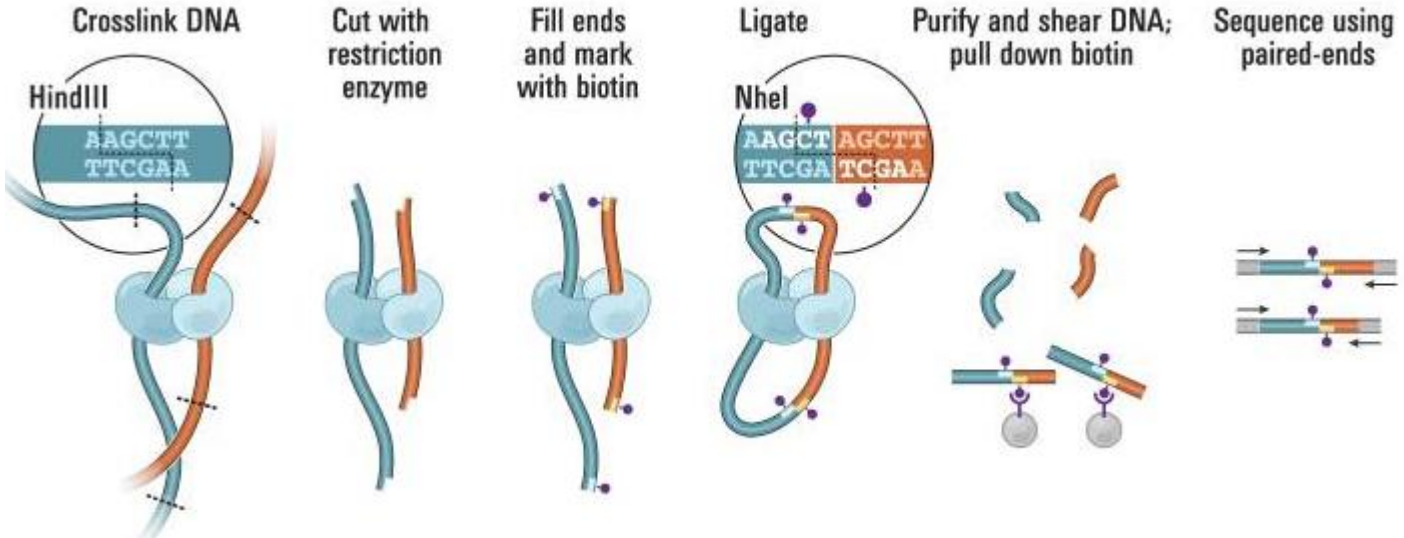


DNA sequencing applications

- Bisulfite-seq
  - DNA methylation
- ChIP-seq, DamID-seq
  - Histone modification, DNA-binding protein
- DNAse-seq, MNase-seq, ATAC-seq
  - DNA accessibility
- 3C, 4C, 5C, Hi-C, ChIA-PET
  - Chromatin folding structure

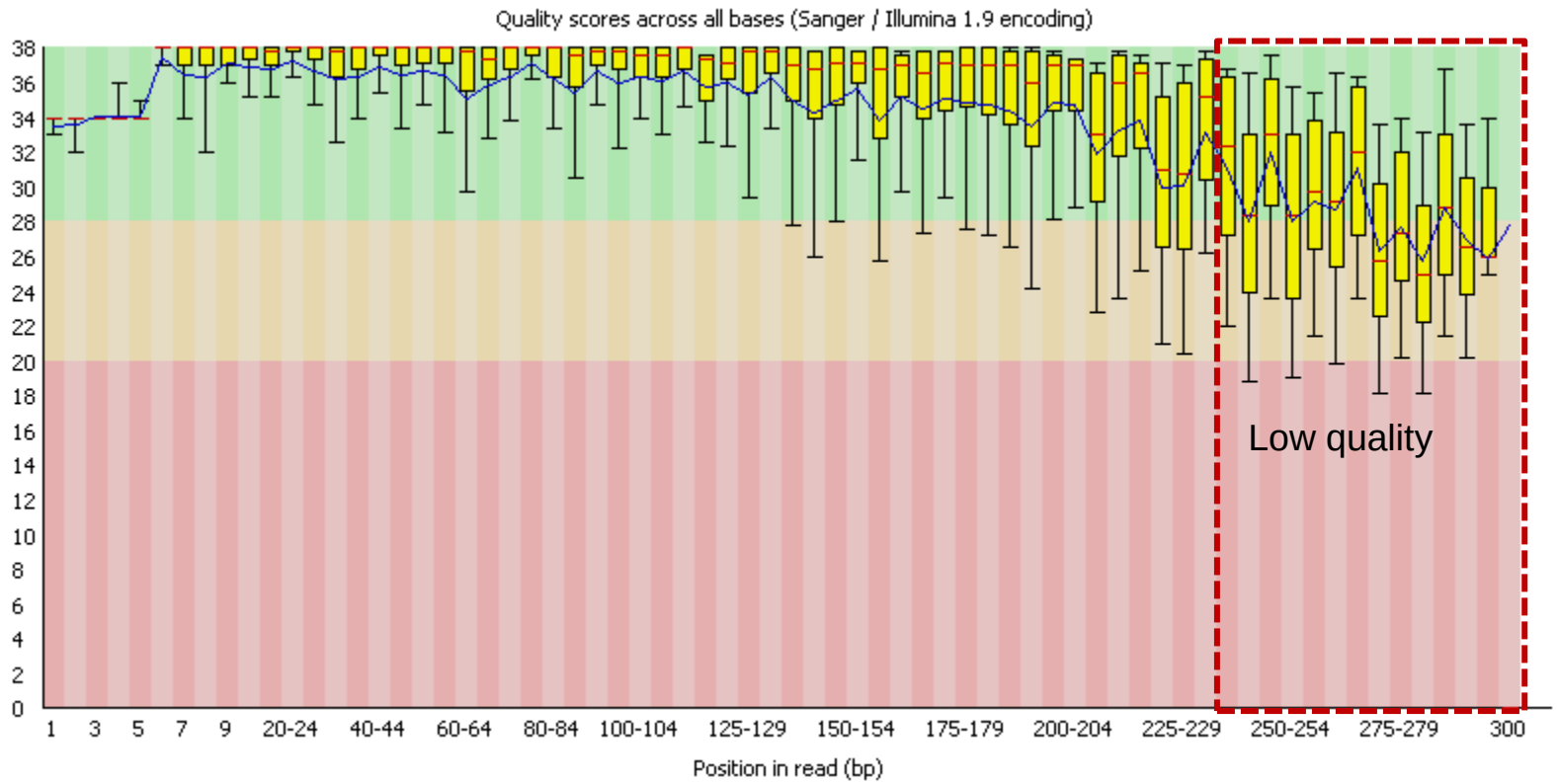


Hi-C: Chromatin folding structure



Lieberman-Aiden *et al*. Science 2009

DNA sequencing QC



Important file formats

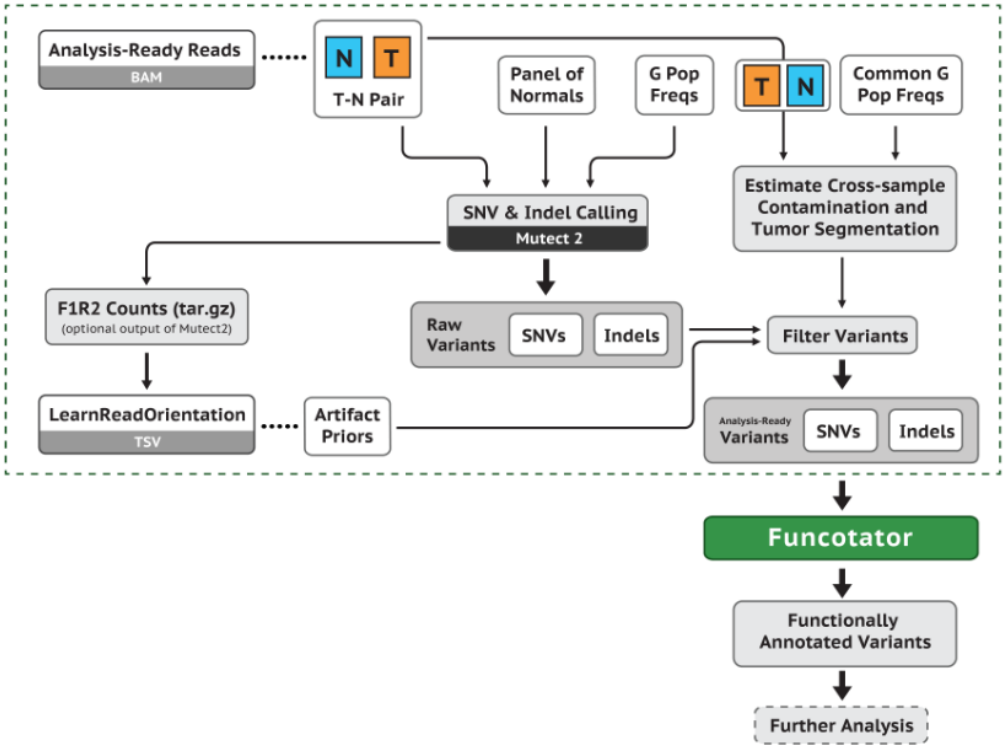
```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                       ATAGCT..............TCAGC
-r003                             ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

The corresponding SAM format is:[1]

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

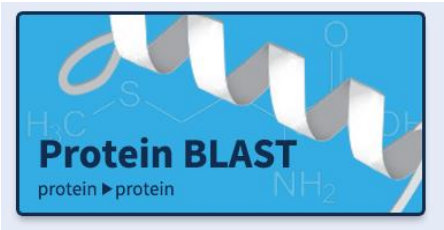
- SAM, BAM, BED, GTF, GFF, FASTA, FASTQ, VCF, MAF

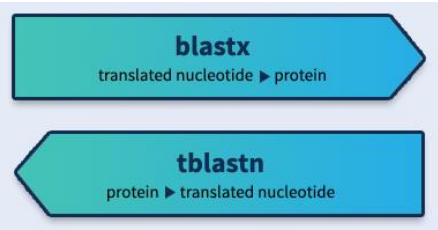
GATK variant calling pipelines

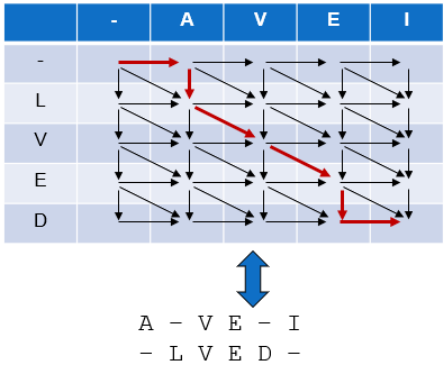


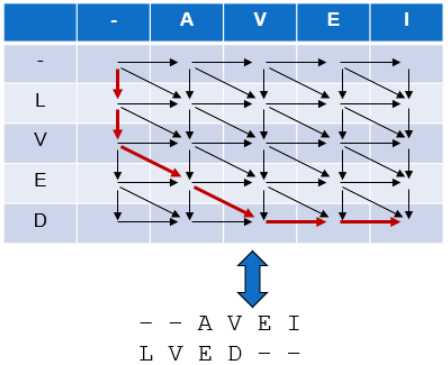
Basic Local Alignment Search Tool (BLAST)

- How does it work?
  - How should we adjust its parameters?
- What questions can it answer?
  - Functional annotation
  - Taxonomy annotation
  - A start for evolutionary analysis
- Variants of BLAST
  - MEGABLAST
  - PSI-BLAST



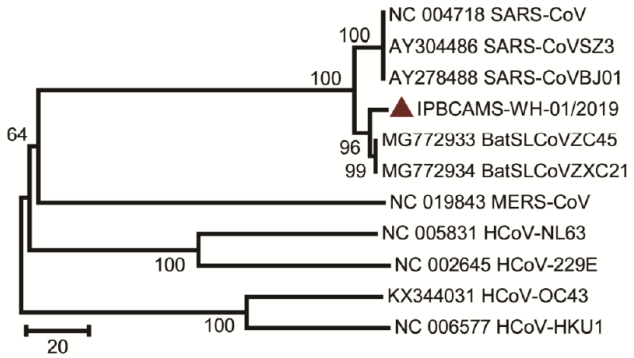






Phylogenetics

- Evolution is informative
- So many parameters and models
  - How to properly analyze?
  - Which tools to use?



Guo *et al*. Clin Infect Dis (2020)

MX: Analysis Preferences

Phylogeny Reconstruction

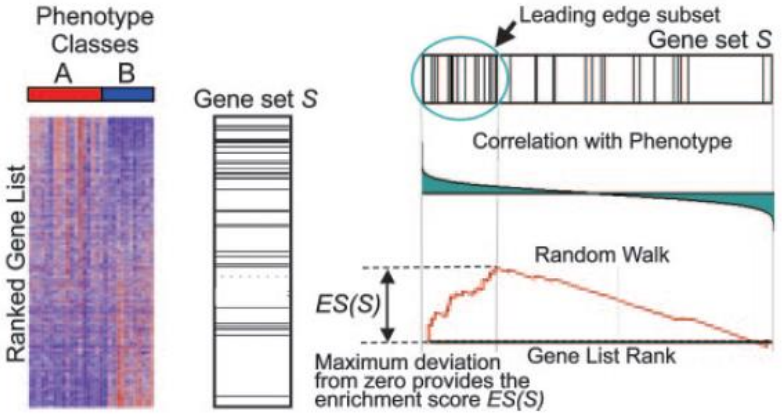
Option	Setting
ANALYSIS	
Statistical Method →	*Maximum Likelihood*
PHYLOGENY TEST	
Test of Phylogeny →	None
No. of Bootstrap Replications →	None / Bootstrap method
SUBSTITUTION MODEL	
Substitutions Type →	*Nucleotide*
Model/Method →	*Tamura-Nei model*
RATES AND PATTERNS	
Rates among Sites →	*Uniform Rates*
No of Discrete Gamma Categories →	Not Applicable
DATA SUBSET TO USE	
Gaps/Missing Data Treatment →	*Use all sites*
Site Coverage Cutoff (%) →	Not Applicable
TREE INFERENCE OPTIONS	
ML Heuristic Method →	*Nearest-Neighbor-Interchange (NNI)*
Initial Tree for ML →	*Make initial tree automatically (Default - NJ/BioNJ)*
Initial Tree File →	Not Applicable
Branch Swap Filter →	*None*
SYSTEM RESOURCE USAGE	
Number of Threads →	*7*

Module 2: Transcriptomics

- Microarray vs RNA-seq vs Nanostring
  - Different data models and computational analyses
- Pseudoalignment of RNA-seq
  - kallisto, salmon
- StringTie's hybrid alignment and *de novo* assembly pipeline
- Pairing RNA-seq processing and differential expression analysis tools
  - HTSeq-count with DESeq2
  - salmon with sleuth

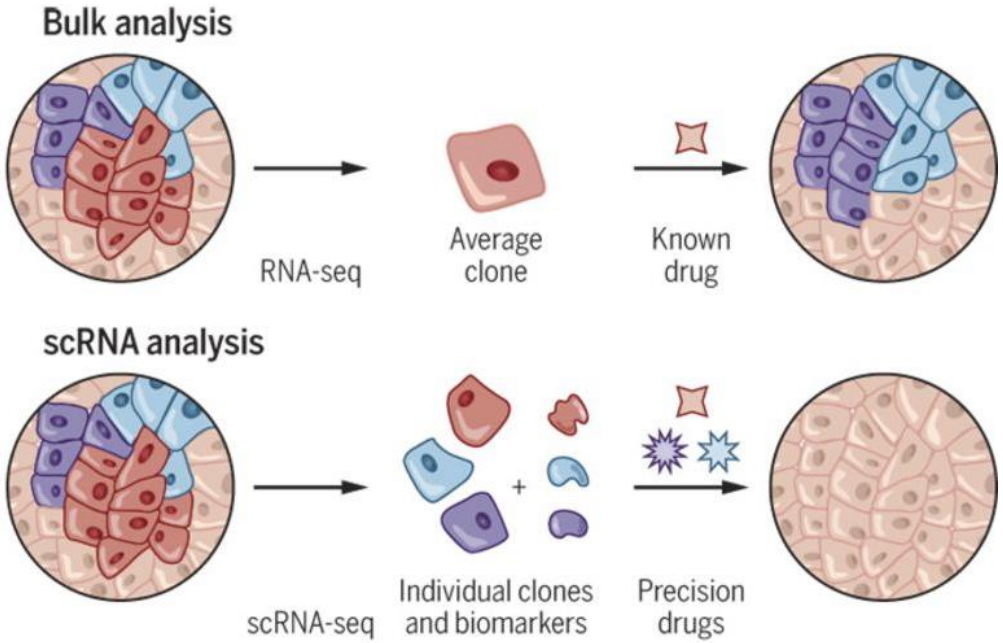
Functional enrichment analysis

- Overrepresentation analysis
  - Frequent functional terms
  - Hypergeometric distribution
- Gene set enrichment analysis (GSEA)
  - Random walk model
- Gene-gene network topology
  - Frequent functional terms
  - Nearby gene on network

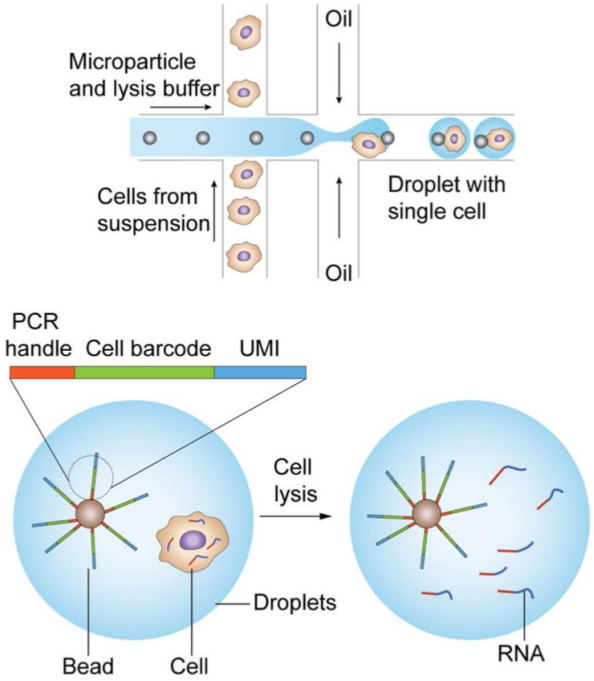


Subramanian *et al*. PNAS. 102:15545-15550 (2005)

Toward single-cell technology

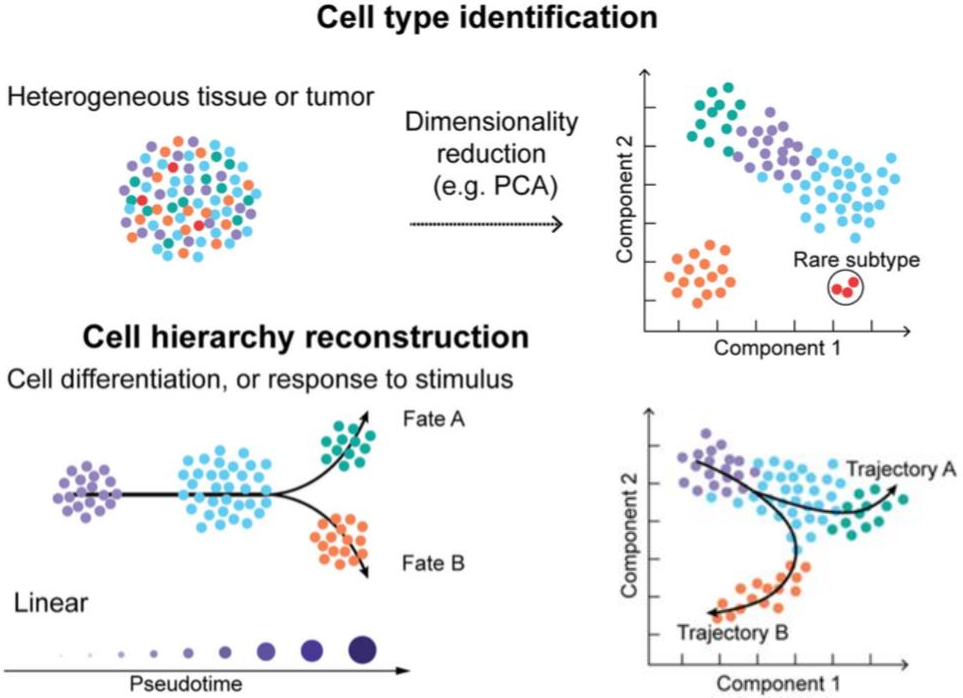


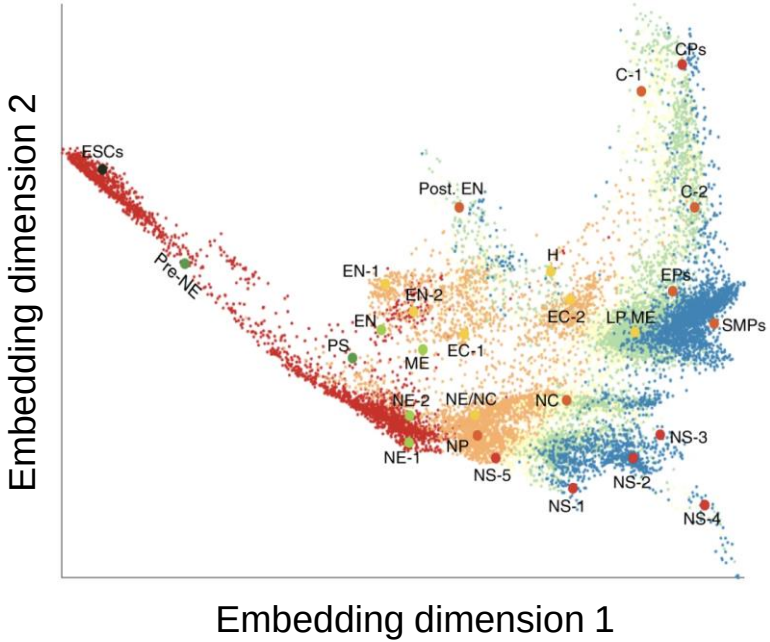
Shalek and Benson. Science Trans Med. 9:eaan4730 (2017)



Hwang *et al.* Exp & Mol Med 50:96 (2018)

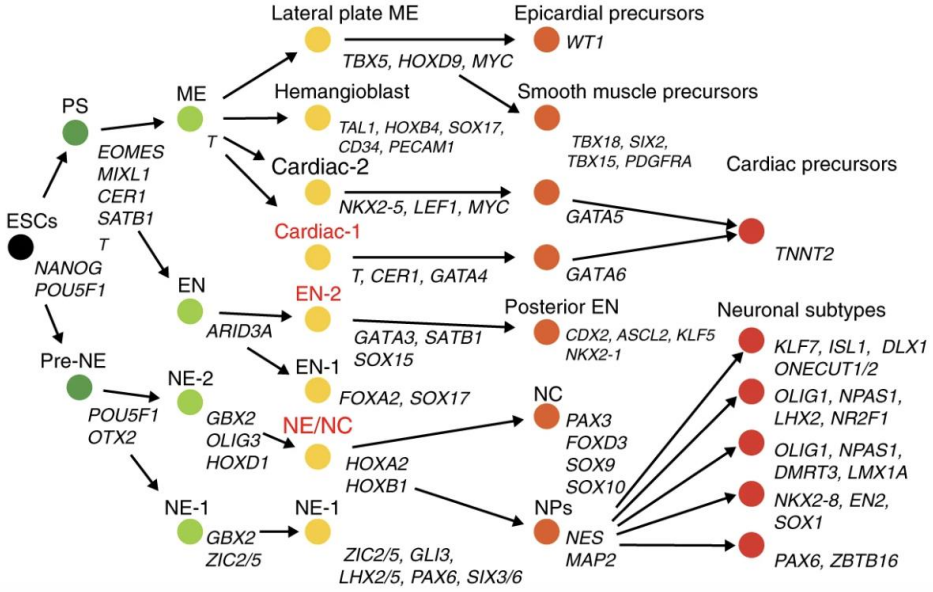
Power of single-cell data

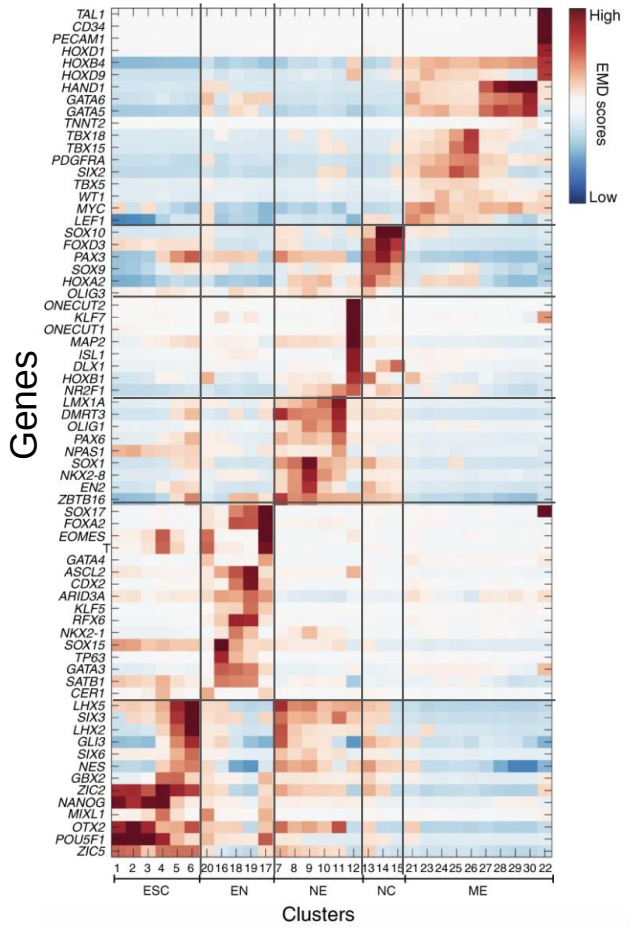




Hwang *et al.* Exp & Mol Med 50:96 (2018)

Power of single-cell data

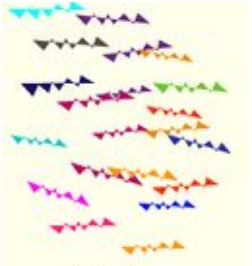




Source: Moon et al. Nature Biotechnology 37:1482-92 (2019)

Module 3: Other omics

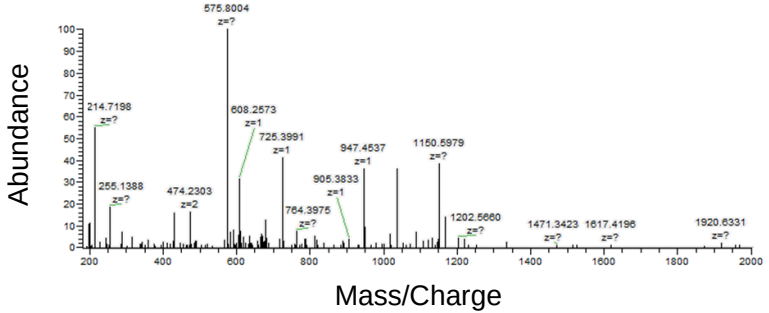
Peptides mixture



Mass spectrometer

Image from planetorbitrap.com

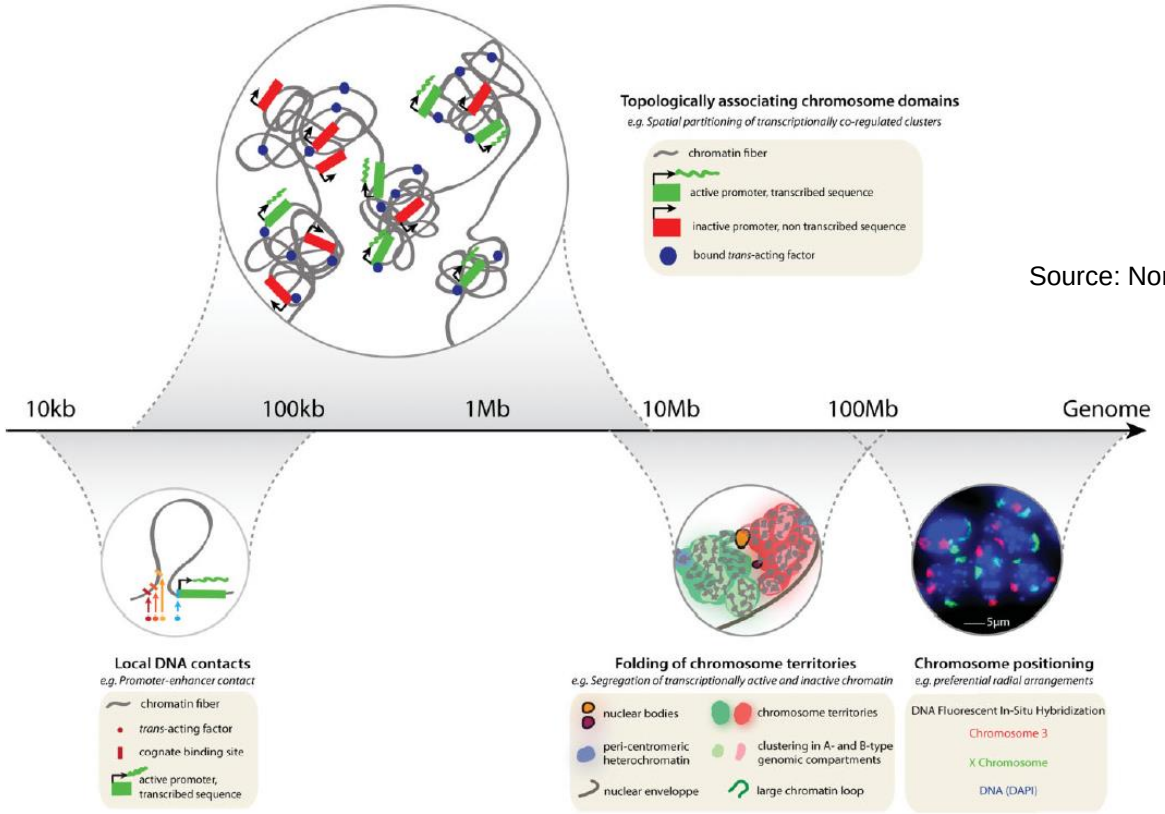
Peptide mass spectrum



SAPYSAMIADQVAQR

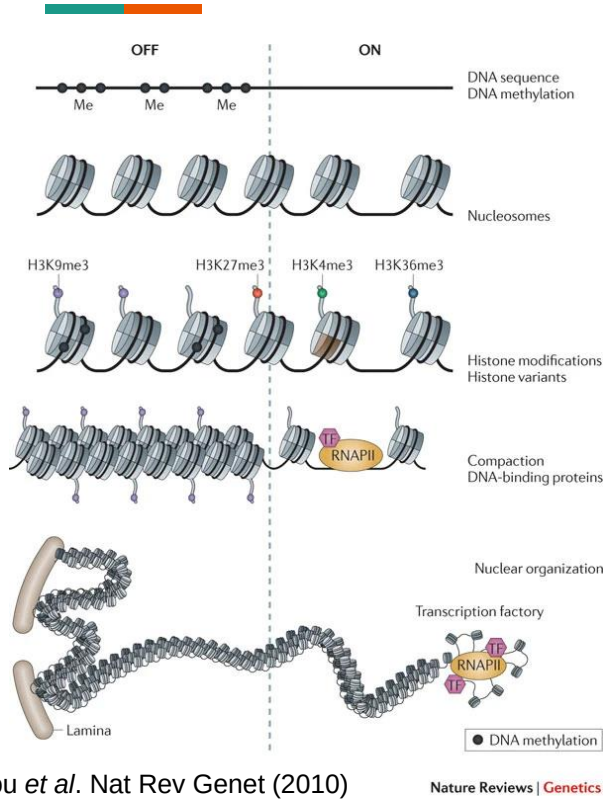
Bassani-Sternberg *et al*. Mol Cell Protomics 14(3), 658-673 (2015)

Chromatin organization

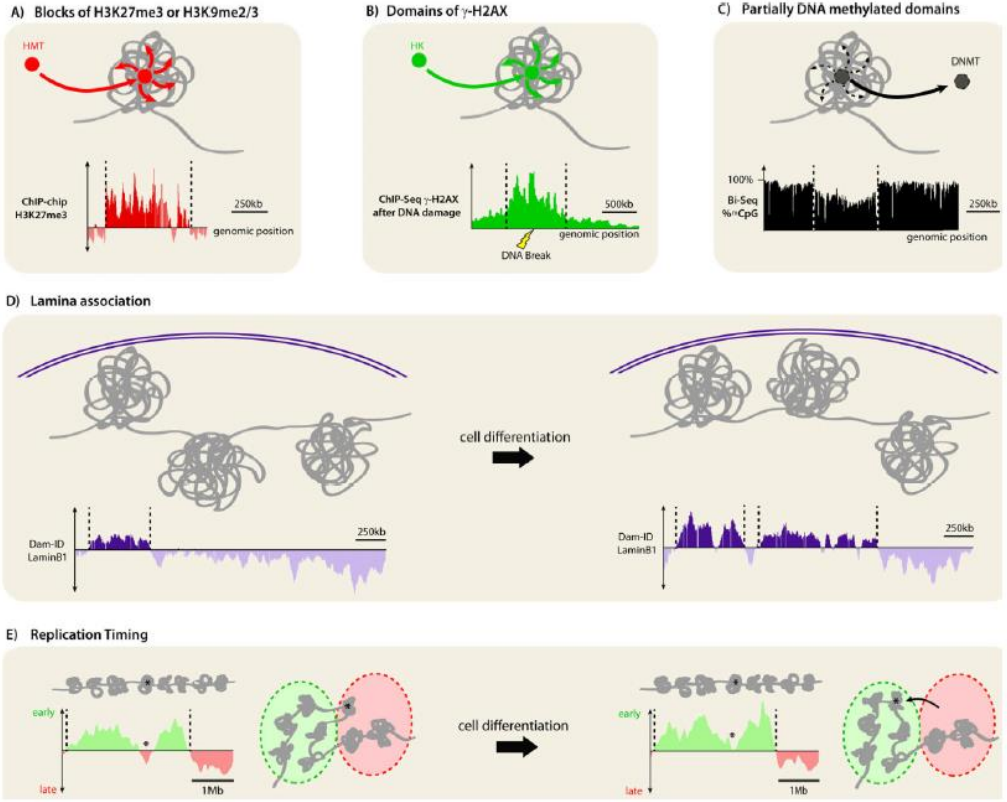


Source: Nora *et al*. Bioessay (2013)

Epigenetics and gene expression regulation

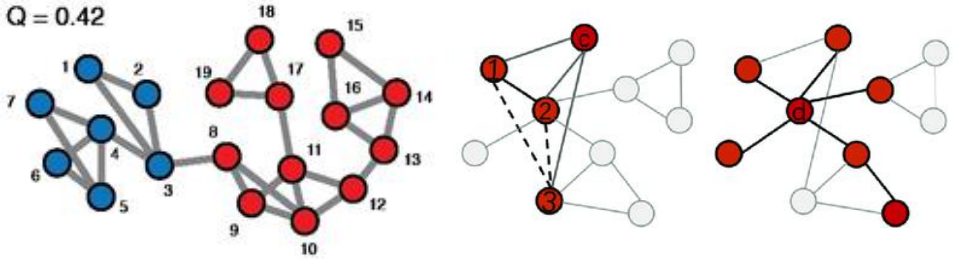


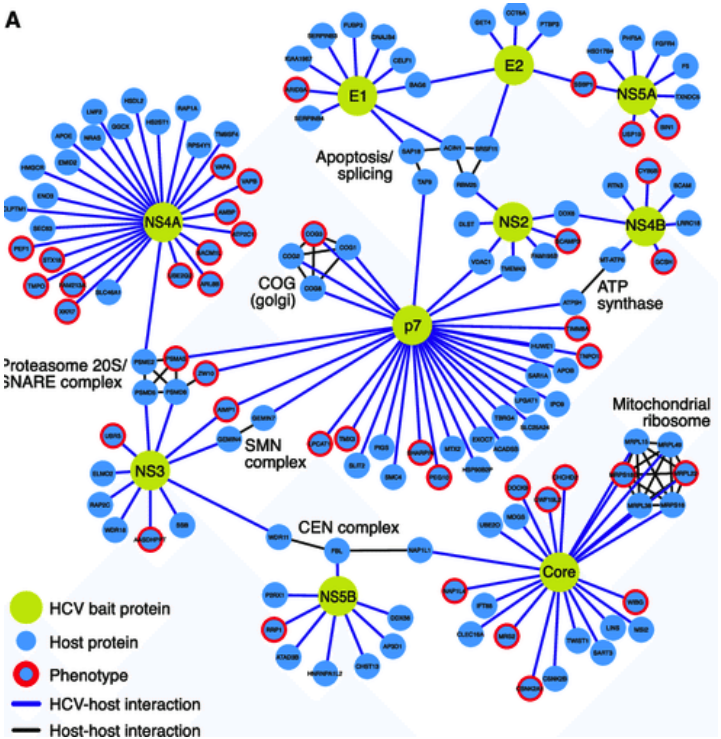
Zhou *et al*. Nat Rev Genet (2010)



Biological networks

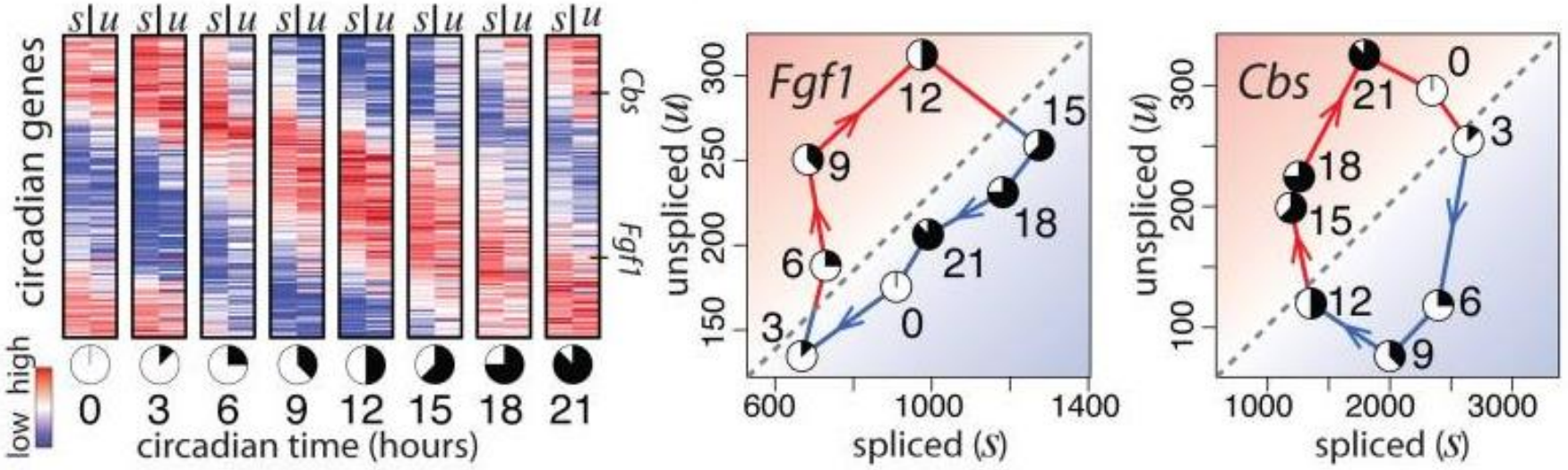
- Topological properties of networks
  - Relationship to biology
- Applications in biomedicine
- Visualization and analysis with CytoScape





Source: Ramage et al. Mol Cell (2015)

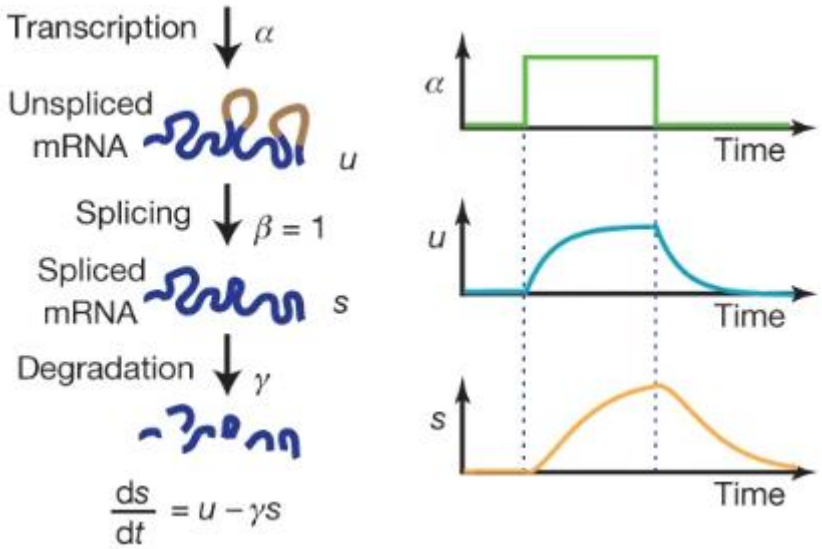
Dynamics modeling in Systems Biology



La Manno *et al.* Nature 2018

- Gene upregulation first produce unspliced RNA
- Followed by processing into mature RNA and proteins

RNA velocity model



La Manno *et al.* Nature 2018

- Model RNA synthesis as differential equations
- Simulation can be performed to analyze the dynamics of the system
  - Try various parameter values

Human Protein Atlas: Protein localization

ACE2

SUMMARY | TISSUE | RNA BRAIN | RNA SINGLE CELL | RNA TISSUE CELL | PATHOLOGY | RNA IMMUNE | MS BLOOD | RNA SUBCELL | RNA CELL LINE | METABOLIC

PROTEIN SUMMARY

RNA DATA

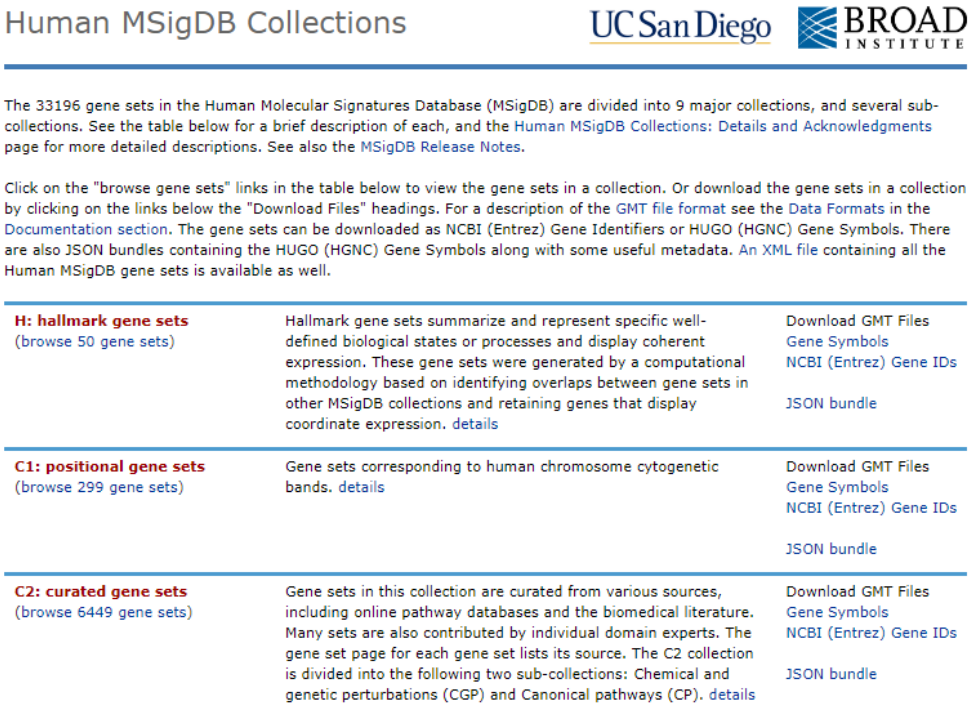
GENE/PROTEIN

ANTIBODIES AND VALIDATION

HUMAN PROTEIN ATLAS SUMMARY[i]

Protein[i]	Angiotensin I converting enzyme 2
Gene name[i]	ACE2
Tissue specificity[i]	Tissue enhanced (gallbladder, intestine, kidney)
Tissue expression cluster[i]	Intestine & Kidney - Transmembrane transport (mainly)
Single cell type specificity[i]	Cell type enriched (Proximal enterocytes)
Single cell type expression cluster[i]	Enterocytes - Digestion (mainly)
Immune cell specificity[i]	Not detected in immune cells
Brain specificity[i]	Not detected in human brain
Cancer prognostic summary	Prognostic marker in renal cancer (favorable) and liver cancer (favorable)
Predicted location[i]	Membrane, Secreted (different isoforms)
Extracellular location[i]	Secreted to blood

MSigDB: Curated gene sets



Human MSigDB Collections

UC San Diego BROAD INSTITUTE

The 33196 gene sets in the Human Molecular Signatures Database (MSigDB) are divided into 9 major collections, and several sub-collections. See the table below for a brief description of each, and the Human MSigDB Collections: Details and Acknowledgments page for more detailed descriptions. See also the MSigDB Release Notes.

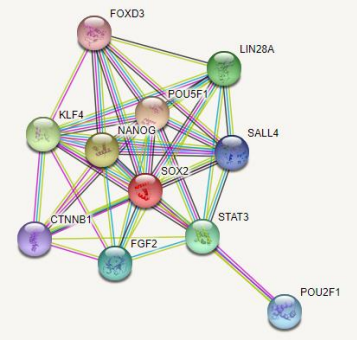
Click on the "browse gene sets" links in the table below to view the gene sets in a collection. Or download the gene sets in a collection by clicking on the links below the "Download Files" headings. For a description of the GMT file format see the Data Formats in the Documentation section. The gene sets can be downloaded as NCBI (Entrez) Gene Identifiers or HUGO (HGNC) Gene Symbols. There are also JSON bundles containing the HUGO (HGNC) Gene Symbols along with some useful metadata. An XML file containing all the Human MSigDB gene sets is available as well.

H: hallmark gene sets (browse 50 gene sets)	Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression. details	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
C1: positional gene sets (browse 299 gene sets)	Gene sets corresponding to human chromosome cytogenetic bands. details	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
C2: curated gene sets (browse 6449 gene sets)	Gene sets in this collection are curated from various sources, including online pathway databases and the biomedical literature. Many sets are also contributed by individual domain experts. The gene set page for each gene set lists its source. The C2 collection is divided into the following two sub-collections: Chemical and genetic perturbations (CGP) and Canonical pathways (CP). details	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle

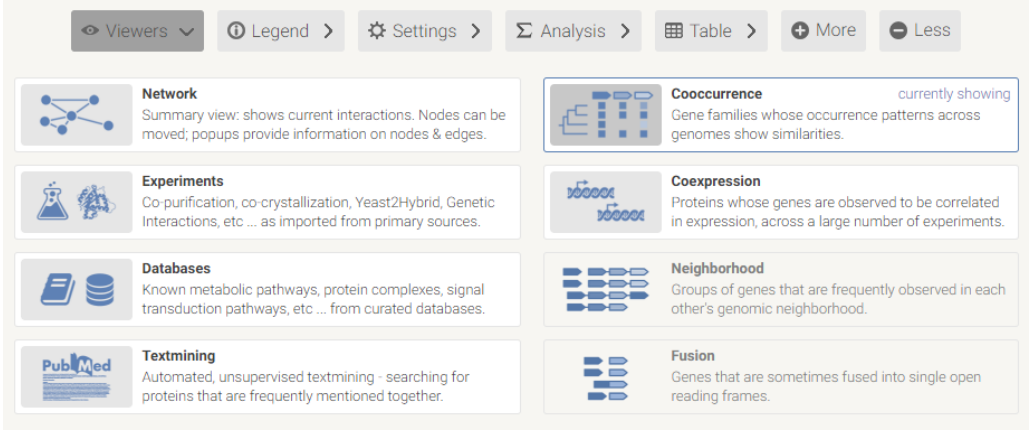
- Function-based / Disease-based / Publication-based / Genomic location-based

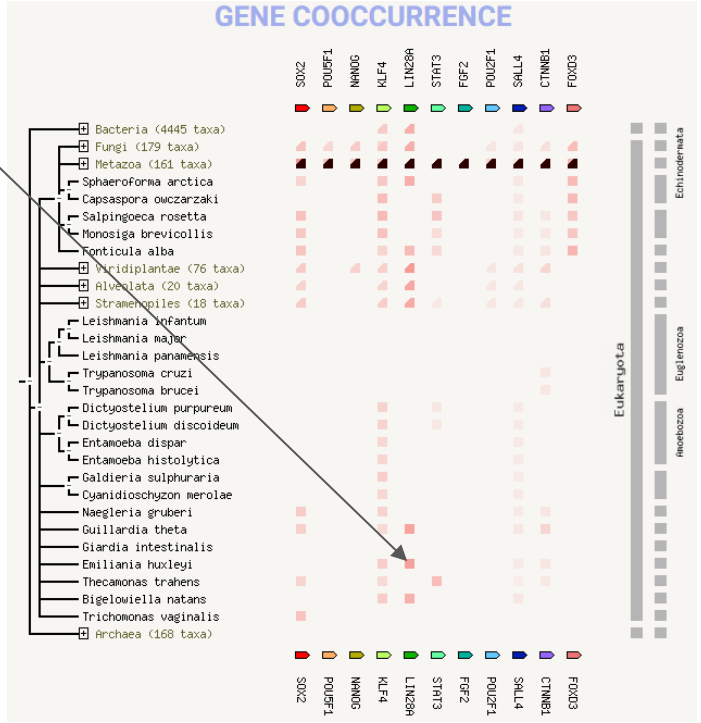
- For gene panel, enrichment test, etc.

STRING: Protein-protein interaction

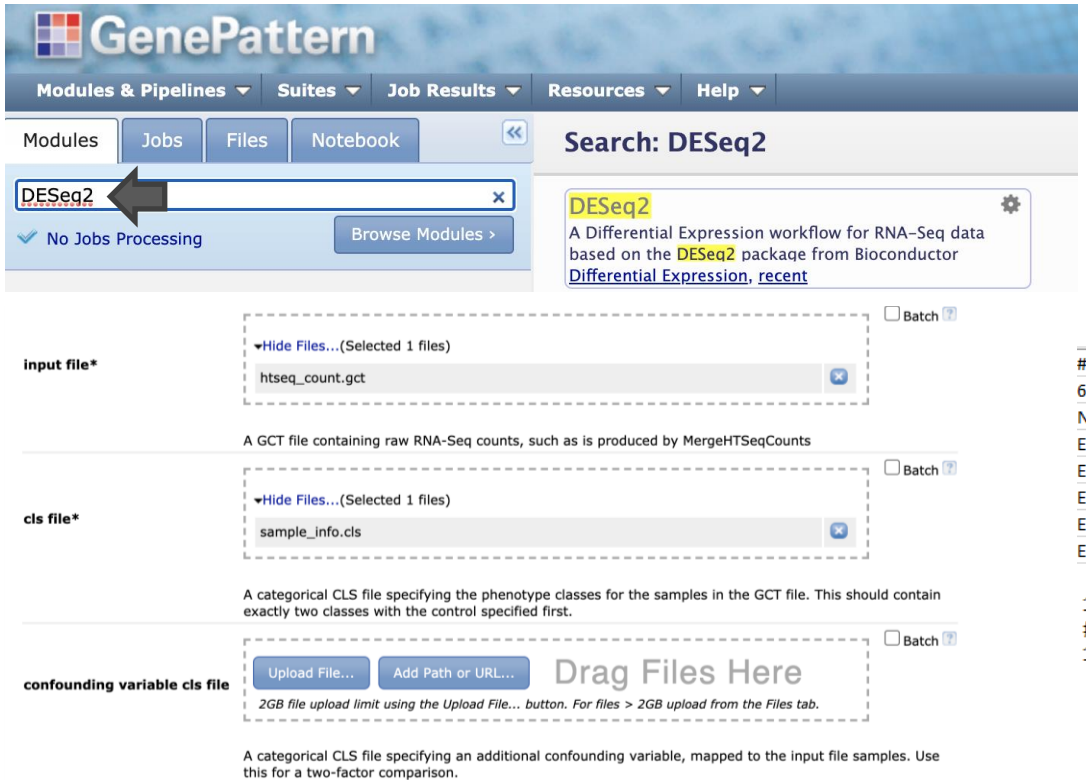



```
LIN28A      25      PEDAARAAD-EPQLLHGAGICKWFNVRMGFGFLSMTARAGVALDPPVDV
                    P  A R AD EP     +G CKWF+V+ GFGF+ +        +   D+
EOD22827    19      PNTATRVADPEPA---PSGKCKWFDVQKGFGFIDVE-------NQEQDL

LIN28A      73      FVHQSKLHMEGFRSLKEGEAVEFTFKKSAKG--LESIRVTGPGGVFCIG
                    FVHQ+ +  +GFRSL EGEA+EF   + AK   L++I VTGPGG F  G
EOD22827    58      FVHQTDIKAKGFRSLAEGEALEFKVSRDAKTNKLKAIEVTGPGGDFVEG

LIN28A      120     SERRP
                    + R P
EOD22827    107     APREP
```





DESeq2 on GenePattern



- Input expression data (RNA-seq read count) and sample label
- .gct and .cls are text files

#1.2

60488	12					
Name	Description	sample1	sample2	sample3	sample4	s
ENSG00000242268.2	ENSG00000242268.2	2	10	0	2	
ENSG00000270112.3	ENSG00000270112.3	8	6	0	0	
ENSG00000167578.15	ENSG00000167578.15	102	633	468	1200	
ENSG00000273842.1	ENSG00000273842.1	0	0	0	0	
ENSG00000078237.5	ENSG00000078237.5	370	364	1220	692	

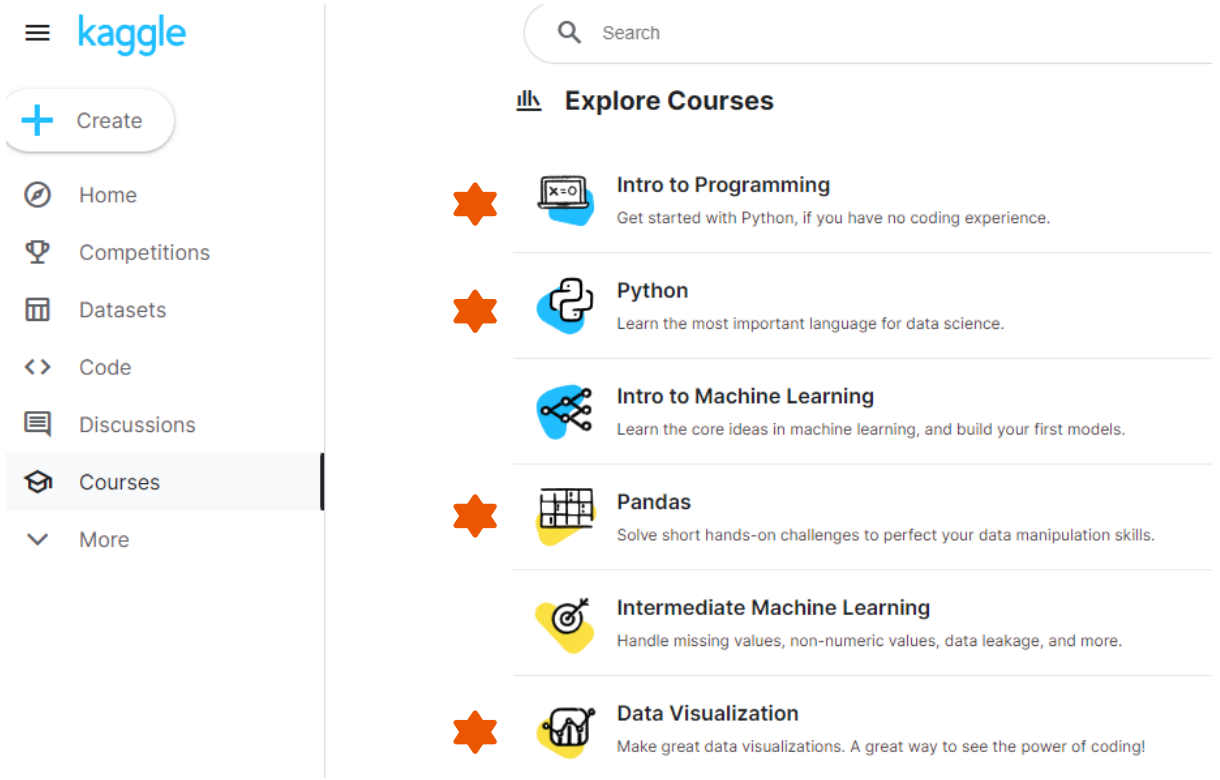
```
12 2 1
# alive dead
1 1 0 0 1 0 0 0 1 0 1 1
```

Module 4: Programming and machine learning

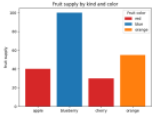
- Kaggle’s programming courses

- In-class practice & problem sets
  - Handling of tabular data
  - Statistical analyses
  - Data exploration
  - Visualization

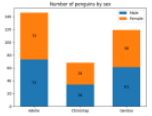
Kaggle's programming



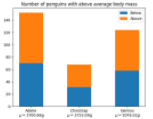
Matplotlib (https://matplotlib.org/stable/)



Bar color demo

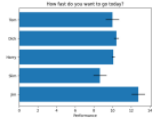


Bar Label Demo

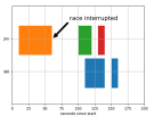


Stacked bar chart

Grouped bar chart with labels

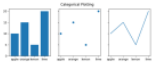


Horizontal bar chart

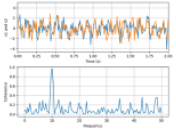


Broken Barh

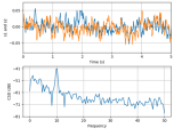
CapStyle



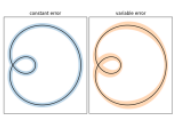
Plotting categorical variables



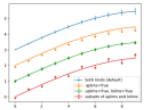
Plotting the coherence of two signals



Cross spectral density (CSD)



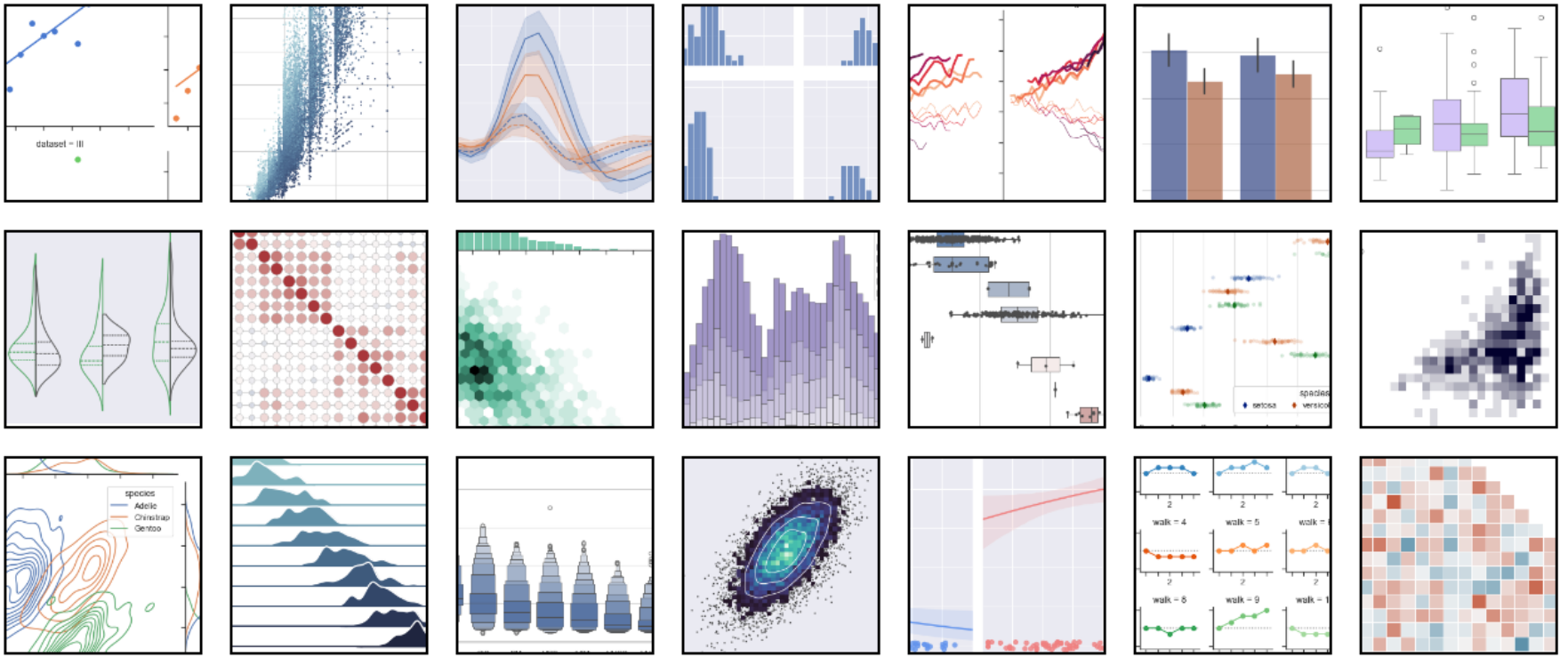
Curve with error band



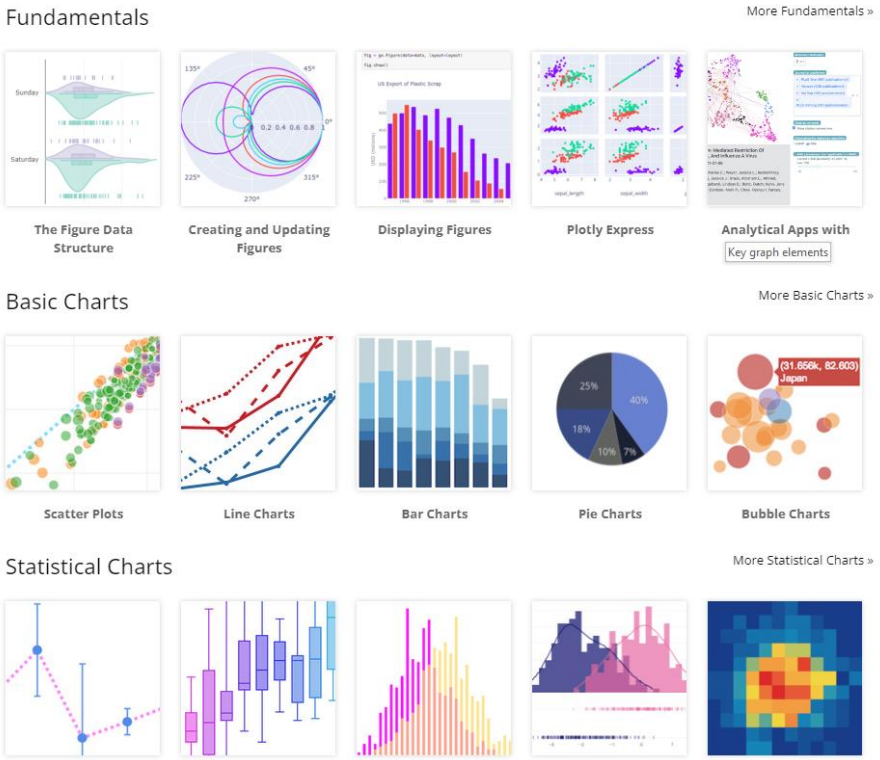
Errorbar limit selection

```python
import matplotlib.pyplot as plt
import numpy as np

# data from https://allisonhorst.github.io/palmerpenguins/

species = (
    "Adelie\n $\\mu=$3700.66g",
    "Chinstrap\n $\\mu=$3733.09g",
    "Gentoo\n $\\mu=5076.02g$",
)
weight_counts = {
    "Below": np.array([70, 31, 58]),
    "Above": np.array([82, 37, 66]),
}
width = 0.5

fig, ax = plt.subplots()
bottom = np.zeros(3)

for boolean, weight_count in weight_counts.items():
    p = ax.bar(species, weight_count, width, label=boolean, bottom=bottom)
    bottom += weight_count

ax.set_title("Number of penguins with above average body mass")
ax.legend(loc="upper right")

plt.show()
```

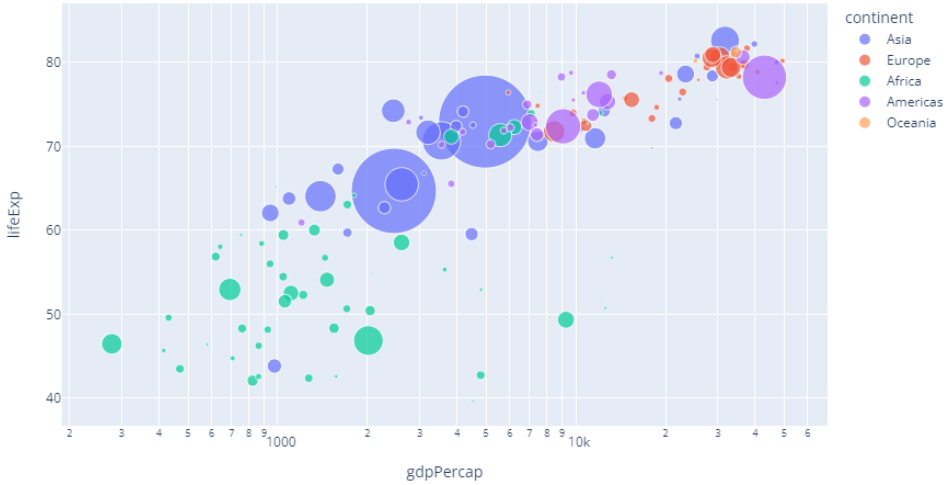
Seaborn (https://seaborn.pydata.org)



Plotly (https://plotly.com/python/)




```
import plotly.express as px
df = px.data.gapminder()

fig = px.scatter(df.query("year==2007"), x="gdpPercap", y="lifeExp",
                 size="pop", color="continent",
                 hover_name="country", log_x=True, size_max=60)
fig.show()
```



Machine learning

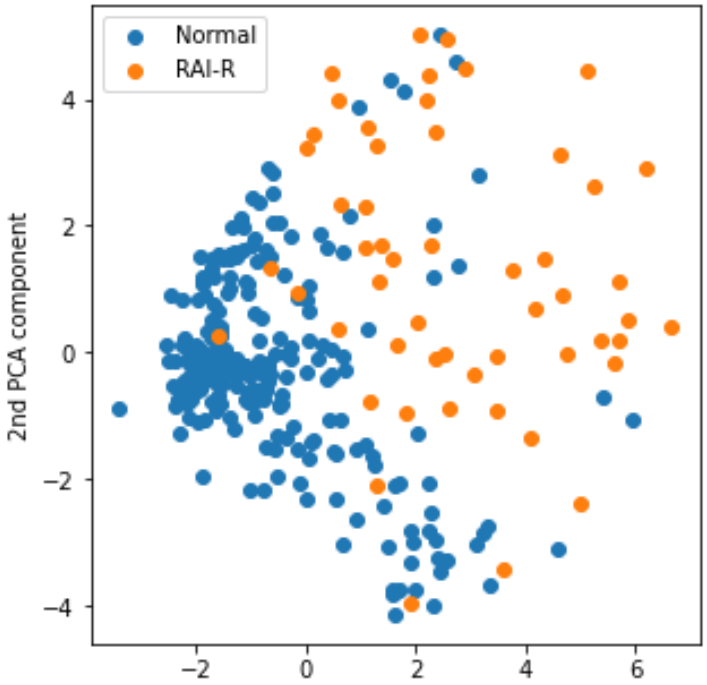
- Unsupervised learning
  - PCA, PCoA, t-SNE, UMAP
  - Clustering techniques

- Supervised learning
  - Predict cancer subtype
  - Identify potential biomarker genes

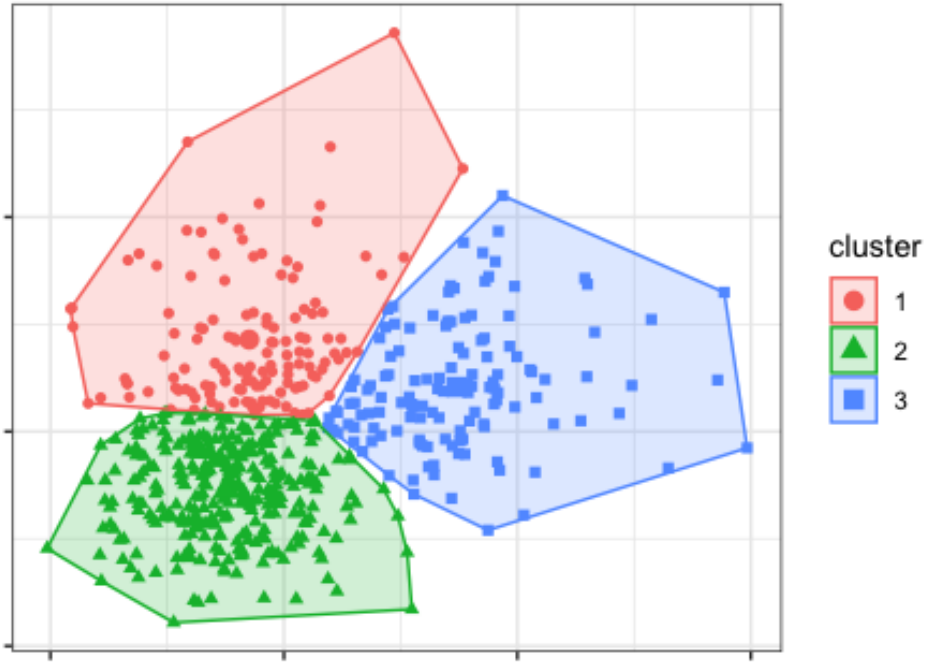
- (a touch of) Deep learning
  - AlphaFold

Two primary branches of unsupervised learning

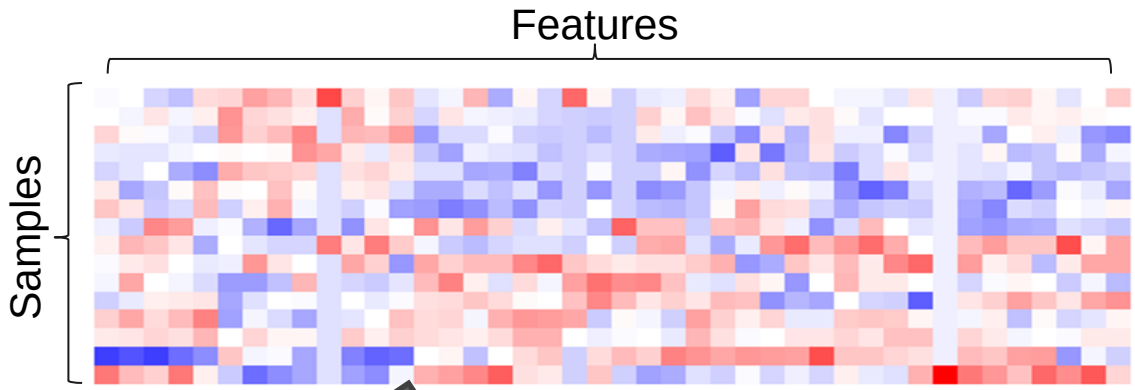
Dimensionality Reduction

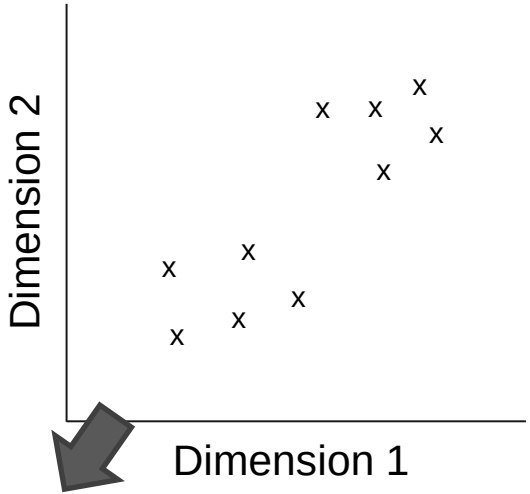


Clustering

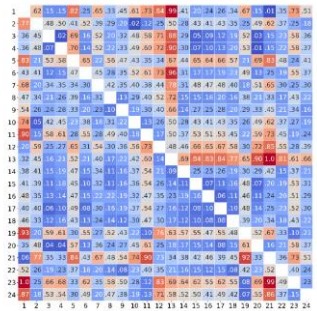


Principal Coordinate Analysis (PCoA)

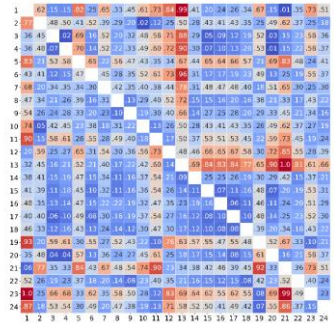




Non-Euclidean distance

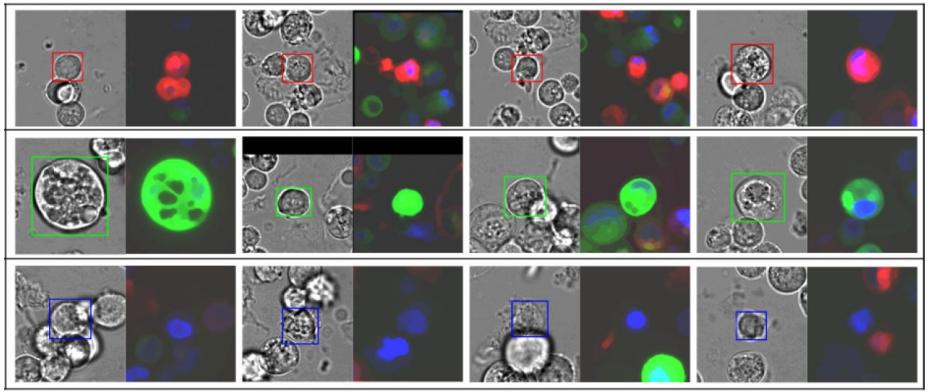


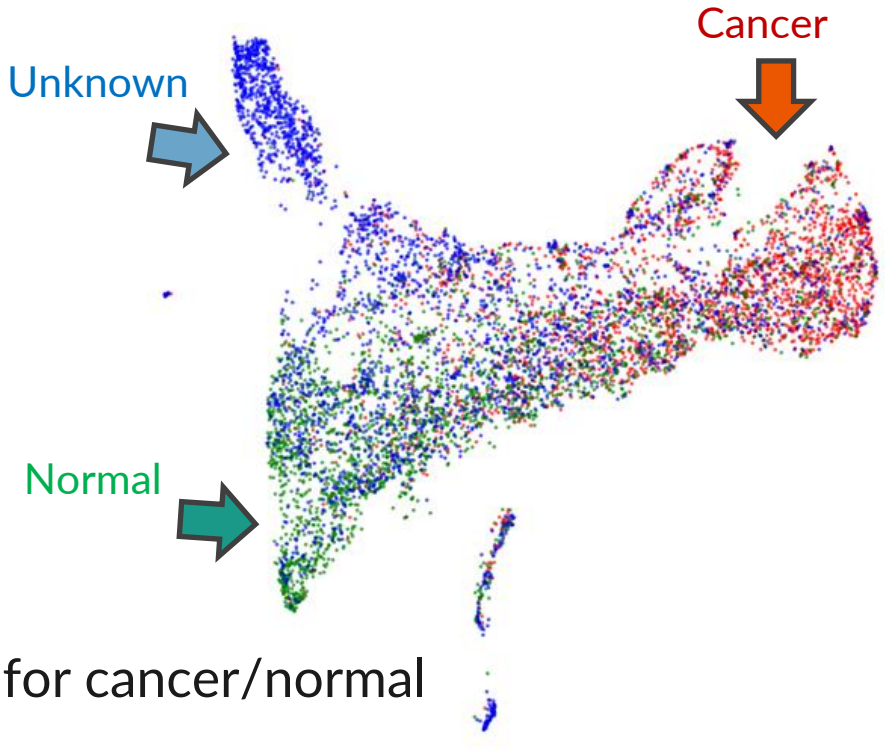
Pearson correlation



Euclidean distance

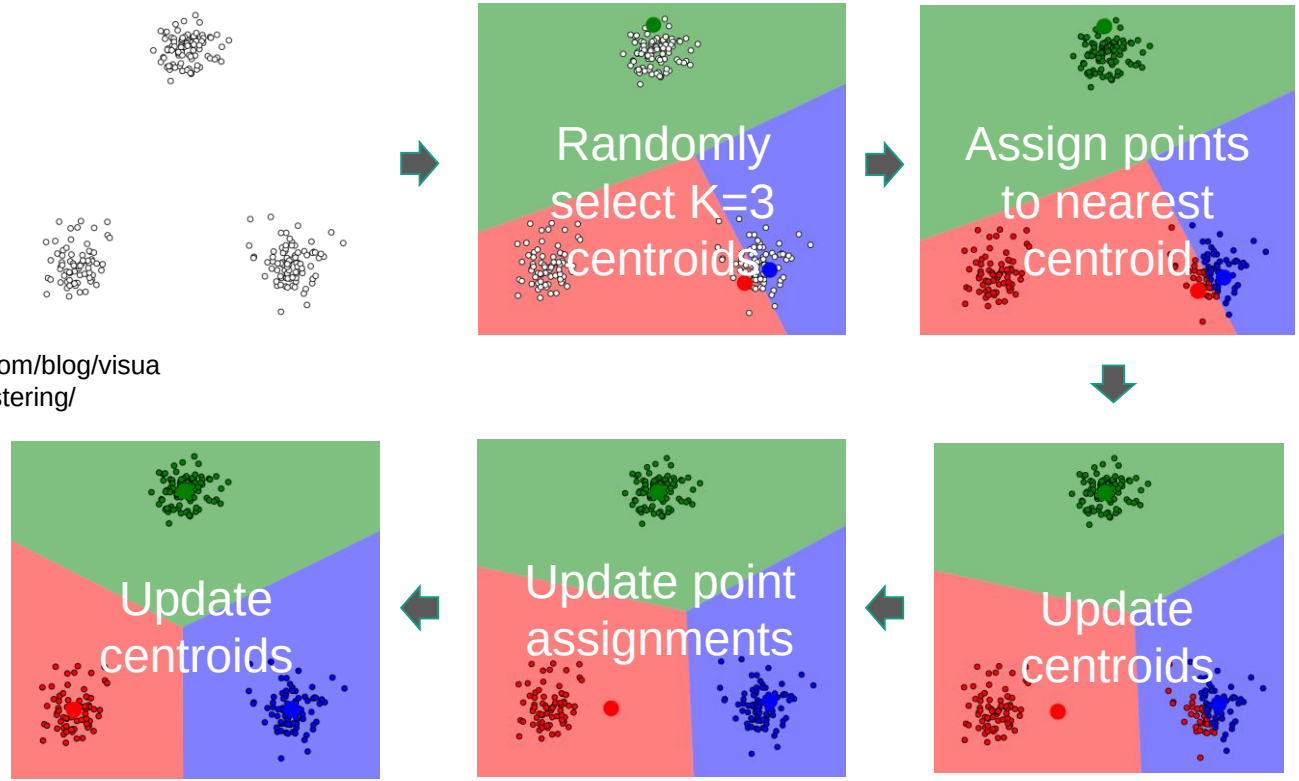
2D visualization for cell images





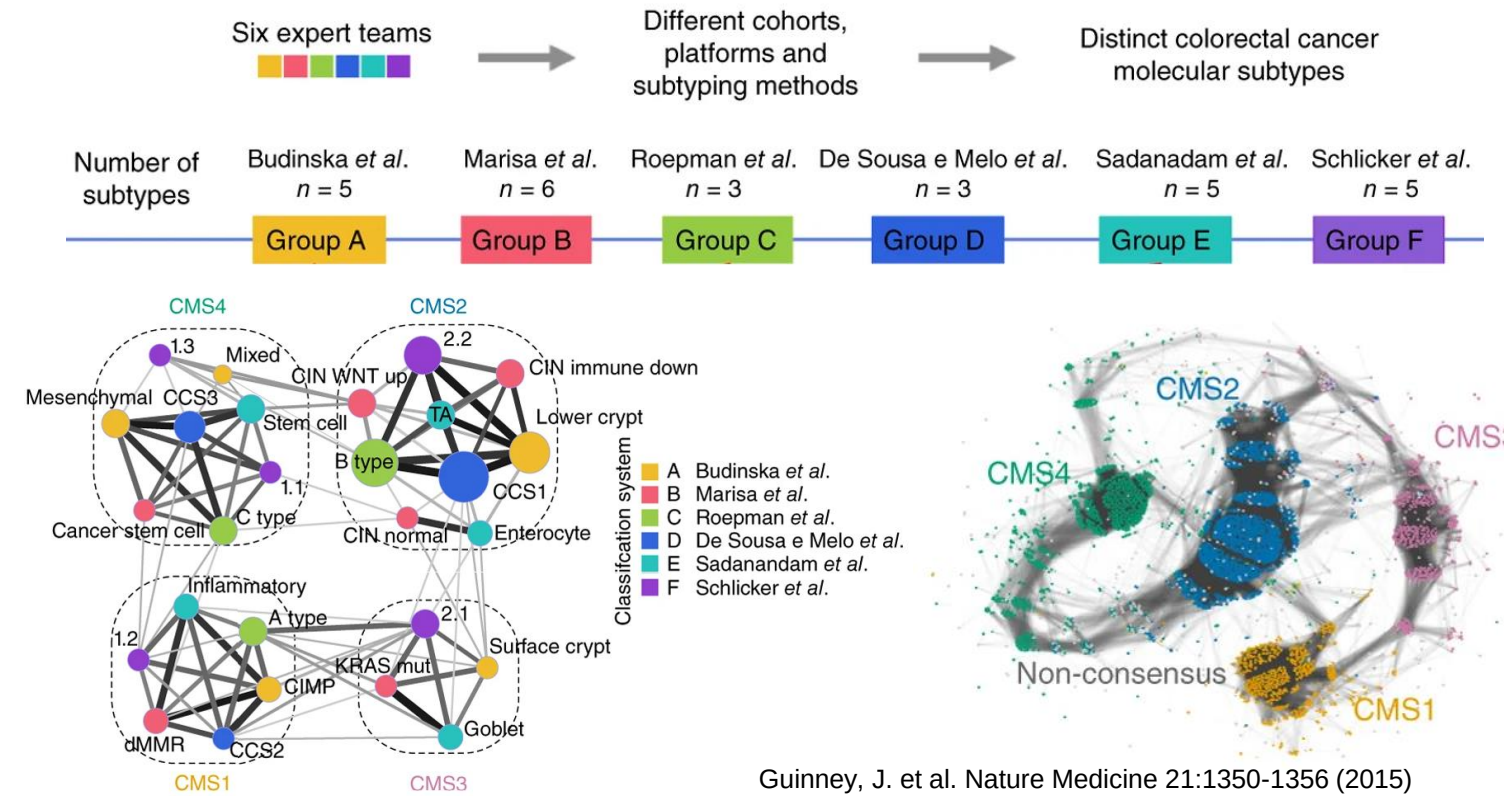
- Extracted from deep learning model for cancer/normal cell type classification

k-means algorithm



www.naftaliharris.com/blog/visualizing-k-means-clustering/

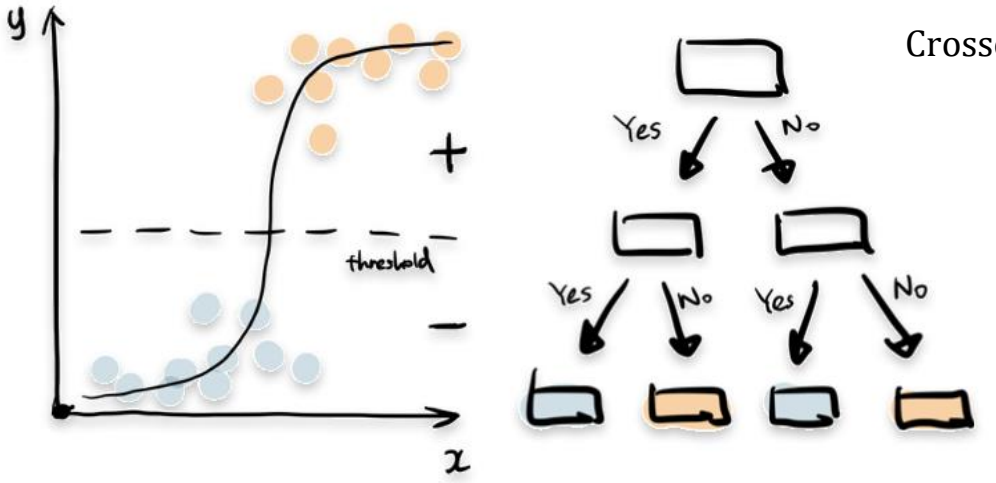
Consensus colorectal cancer subtyping



Guinney, J. et al. Nature Medicine 21:1350-1356 (2015)

The cores of supervised learning

Model



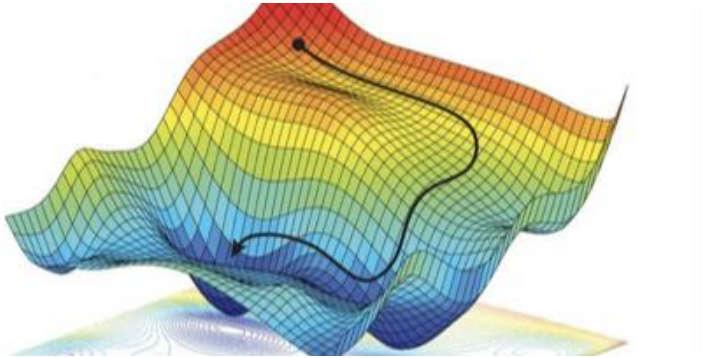
https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

Objective / Loss Function

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100$$

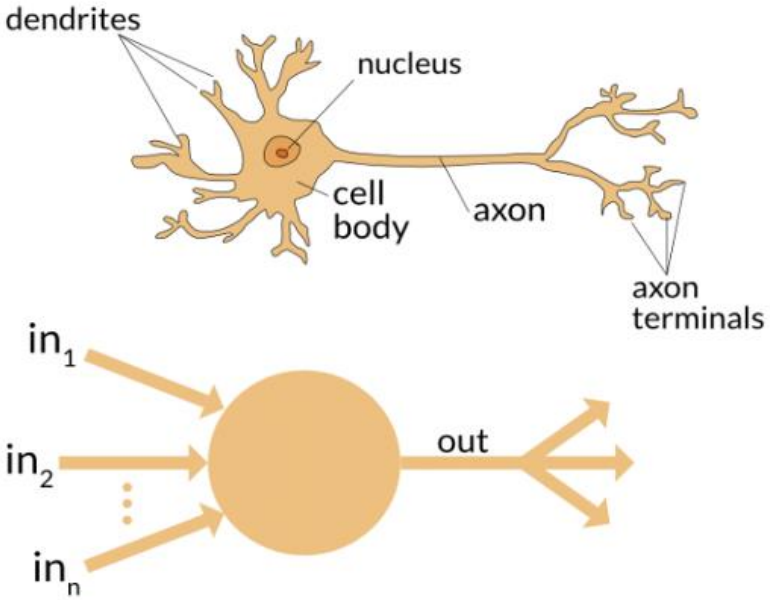
$$\text{Crossentropy} = -\frac{1}{n}\sum_{i=1}^{n} y_i \ln(\hat{y}_i) + (1 - y_i)\ln(1 - \hat{y}_i)$$

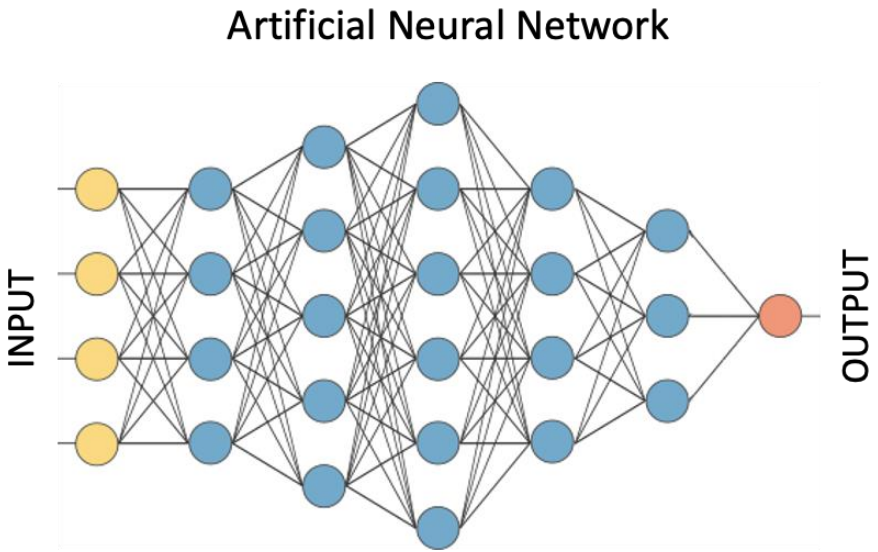
Optimization Algorithm



https://medium.com/analytics-vidhya/gradient-descent-b0dc1af33517

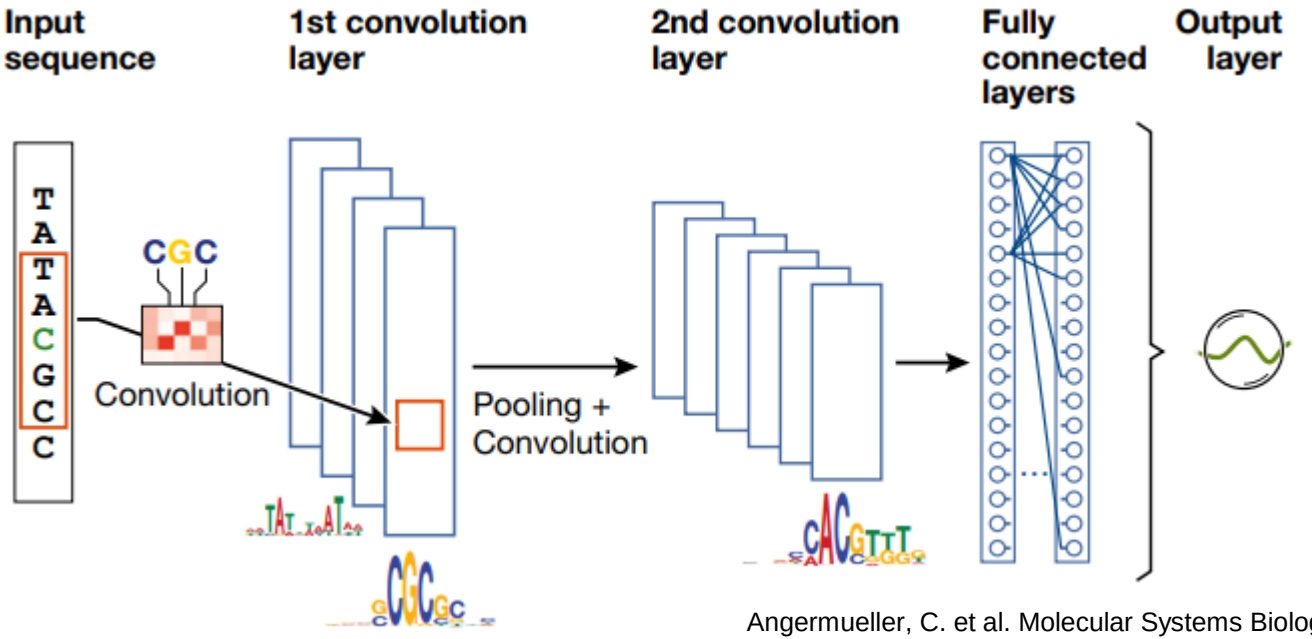
Artificial neural network





- Network of simple computation nodes: $out = f(w_1in_1 + w_2in_2 + \dots + w_nin_n)$

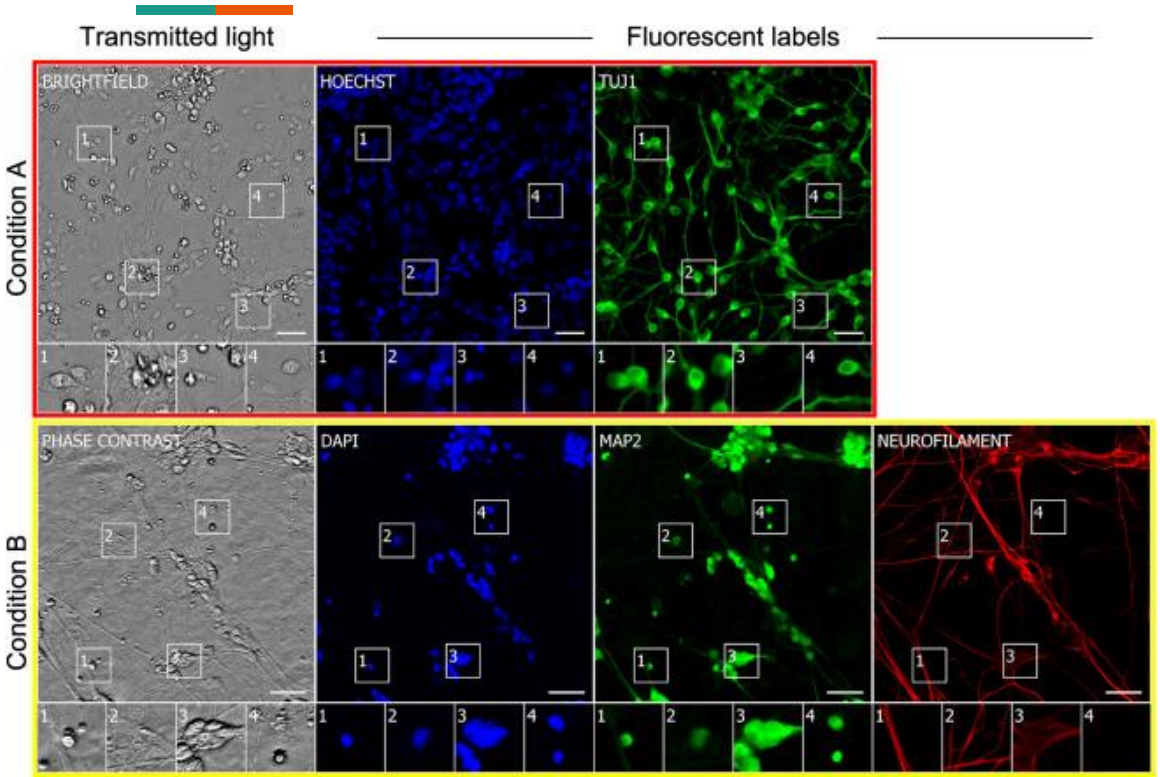
Convolution for DNA sequences



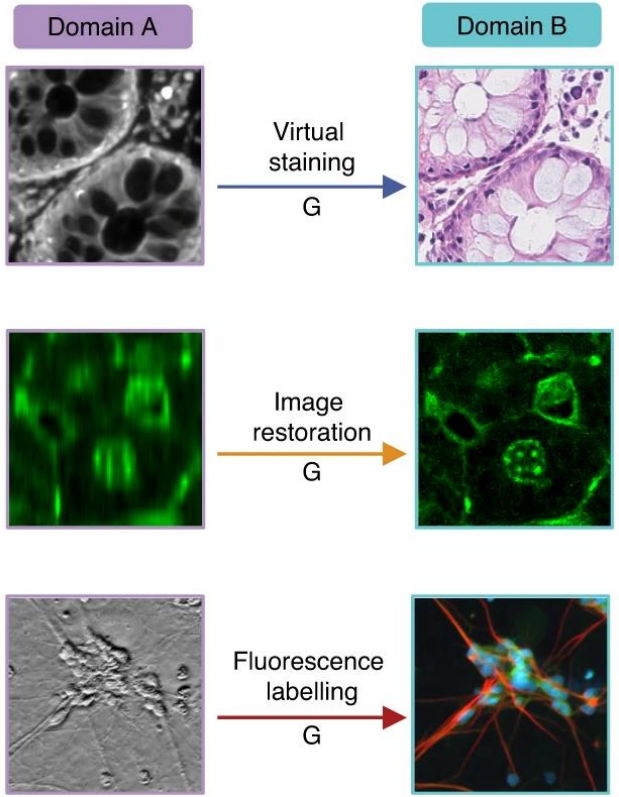
Angermueller, C. et al. Molecular Systems Biology 12:878 (2016)

- Motif = contextual pattern on DNA sequence

Virtual staining



Christiansen, E.M. et al. Cell 173:792-803.e19 (2018)



Li, X. et al. Light: Science and Applications10:44 (2021)

From bioinformatics to deep learning

Published: 24 September 2018

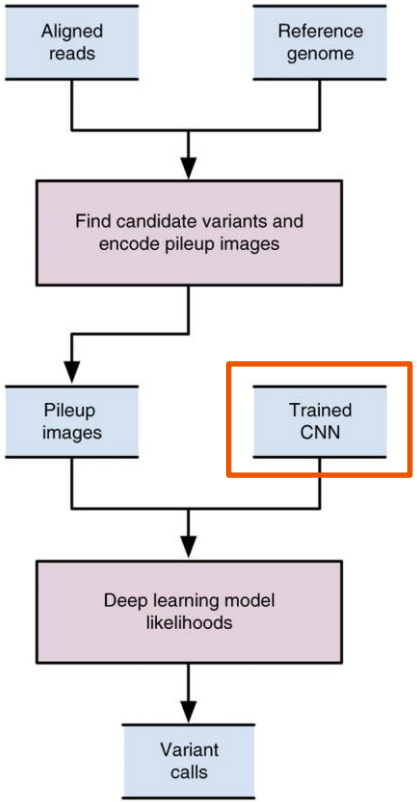
A universal SNP and small-indel variant caller using deep neural networks

Published: 27 July 2015

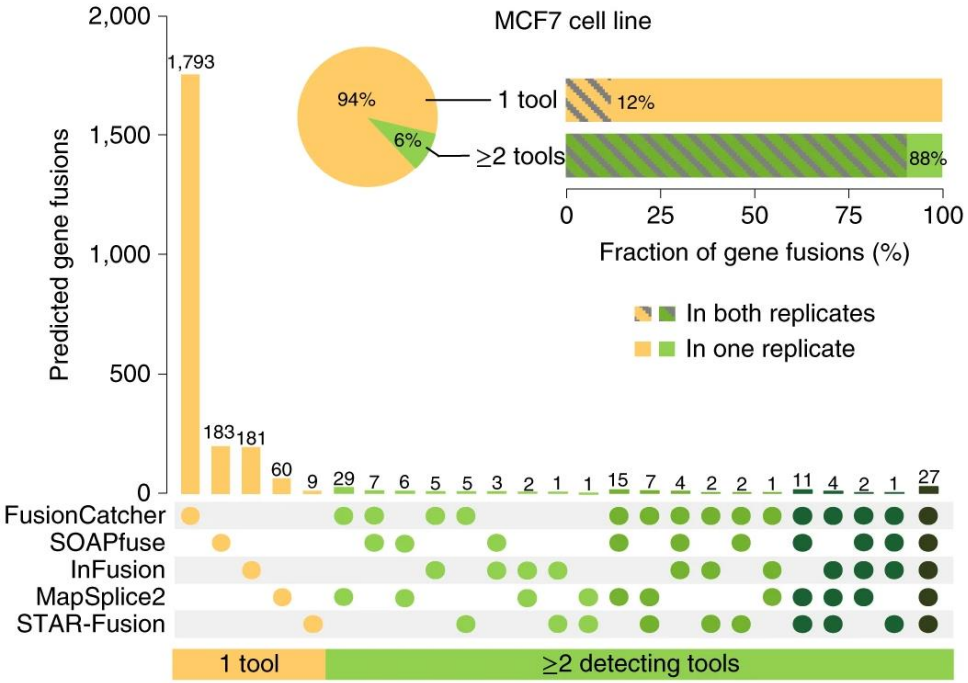
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

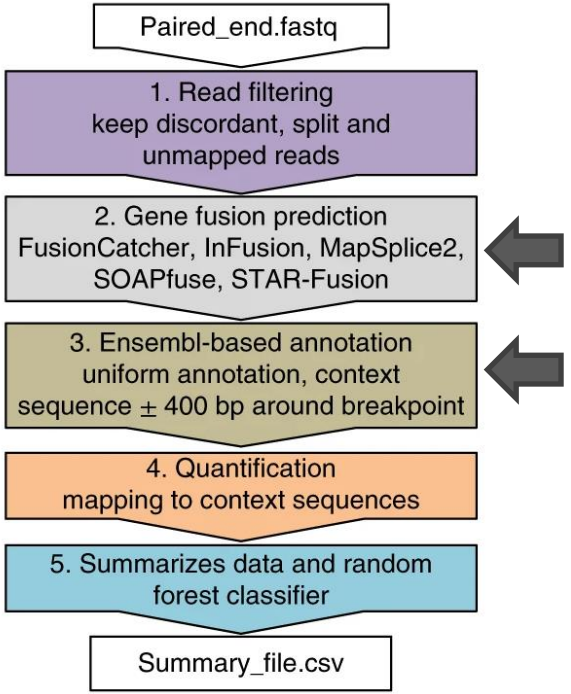
Article | Open Access | Published: 19 May 2022

Prediction of protein–protein interaction using graph neural networks



Aggregate scores from multiple tools





Weber, D. et al. Nature Biotechnology 40:1276-1284 (2022)

Any questions?