3000788 Intro to Comp Molec Biol

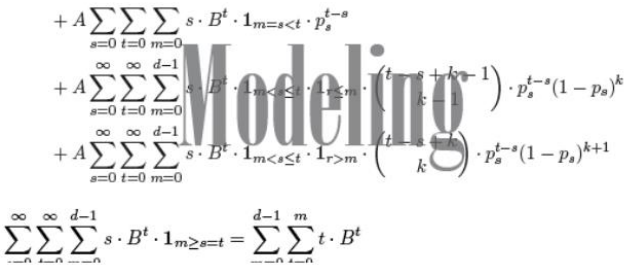
Lecture 1: Course introduction and logistics

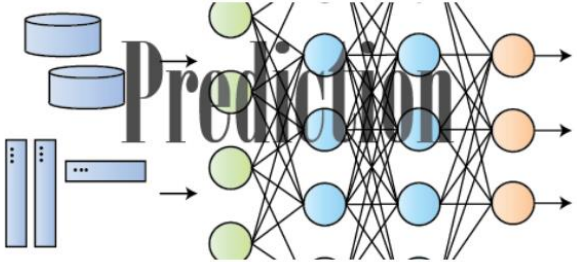
August 17, 2023

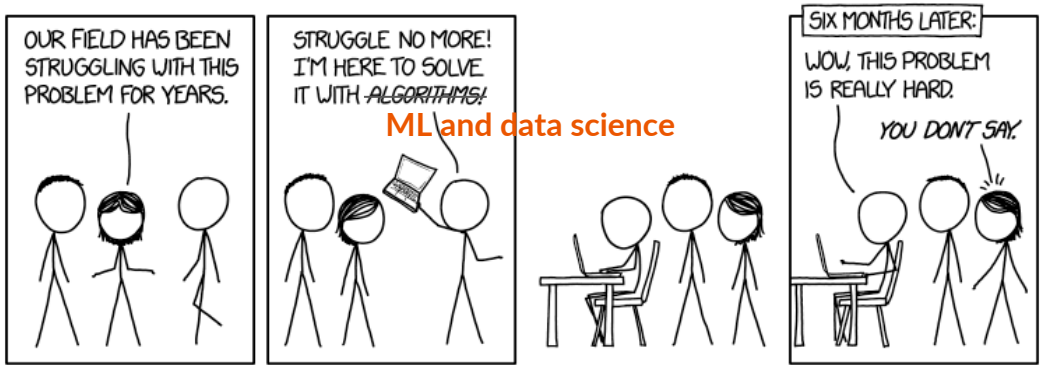
Sira Sriswasdi, PhD

- Research Affairs
- 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

- BS in Mathematics
- PhD in Computational Biology

Keywords

- Biological Networks
- Proteomics / Mass Spectrometry
- Molecular Evolution

About you

Please introduce yourself

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

About this course

- Survey broad topics in computational molecular biology
- Always ask “Why?”

- Theory and practice
- Standard & customized bioinformatics workflow

- Python programming
- Machine learning & data science

Course structure

Module	Date	Topics	Assignment
1	17 Aug – 24 Aug	Introduction & computational thinking	Problem set 1
2	28 Aug – 11 Sep	DNA sequencing & applications	Problem set 2
3	14 Sep – 28 Sep	Transcriptomics	Problem set 3-4
4	2 Oct – 2 Nov	Advanced topics - Single-cell data - Proteomics - Chromatin organization - Biological networks	Problem set 5-7
5	6 Nov – 16 Nov	Python skills for statistical analysis and visualization	Problem set 8-9
6	20 Nov – 30 Nov	Machine learning & applications	Problem set 10
		Post-course evaluation	Mock exam

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%]
  - Hands-on practice
  - Discussion

Companion courses from MIT (7.36, 6.047)

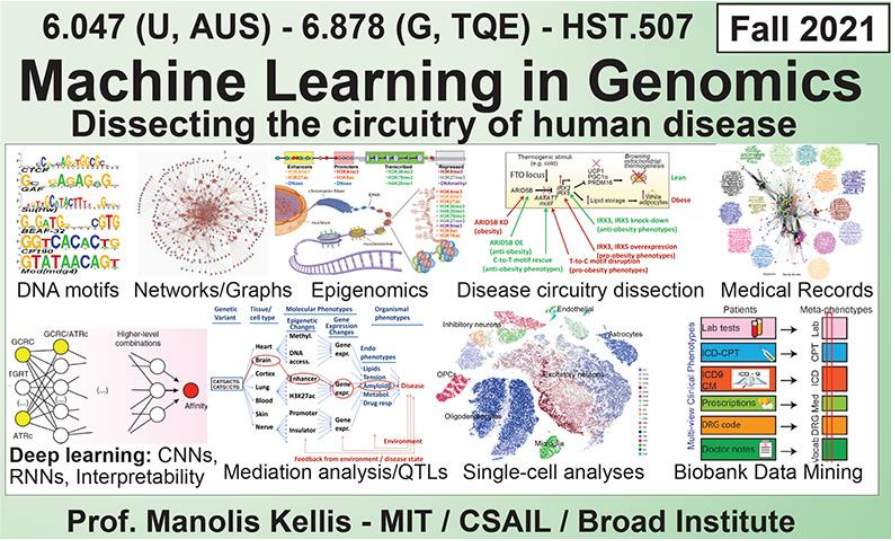
FOUNDATIONS OF COMPUTATIONAL AND SYSTEMS BIOLOGY

Lecture 1: Introduction to Computational and Systems Biology

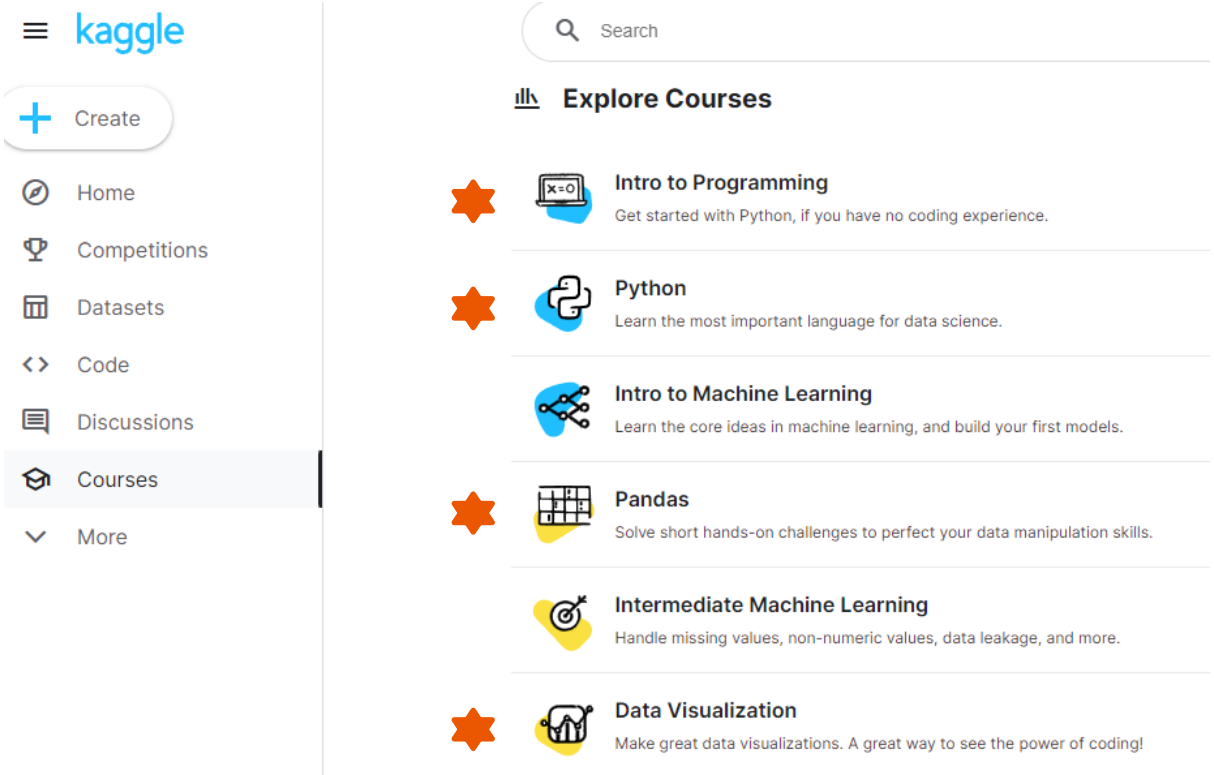
Lecture 2: Local Alignment (BLAST) and Statistics

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

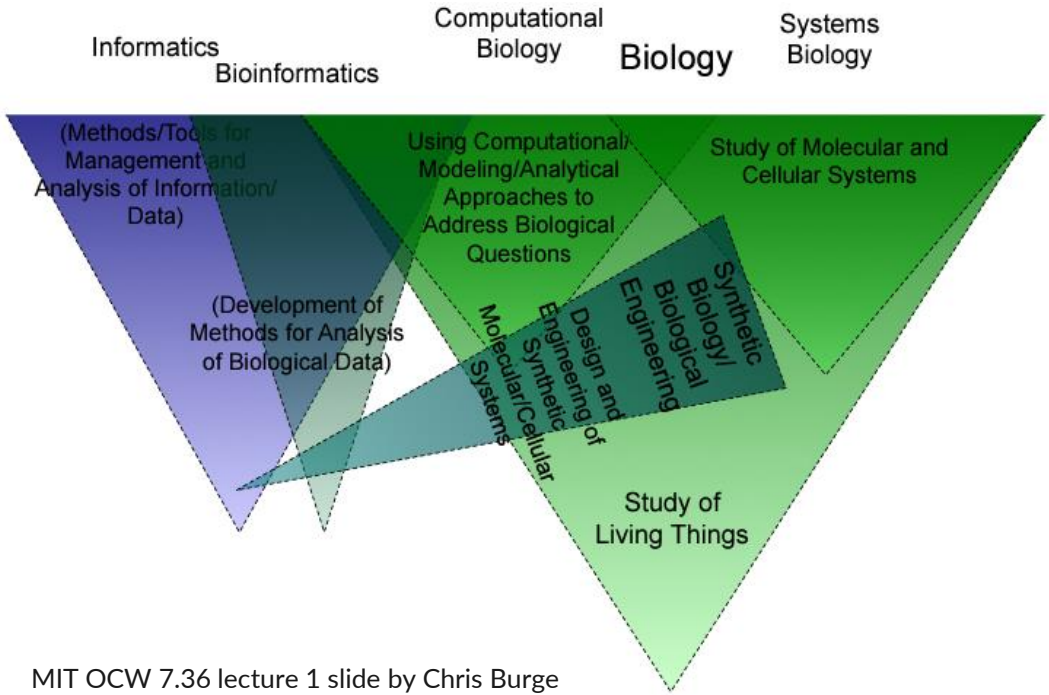
Lecture 4: Comparative Genomic Analysis of Gene Regulation



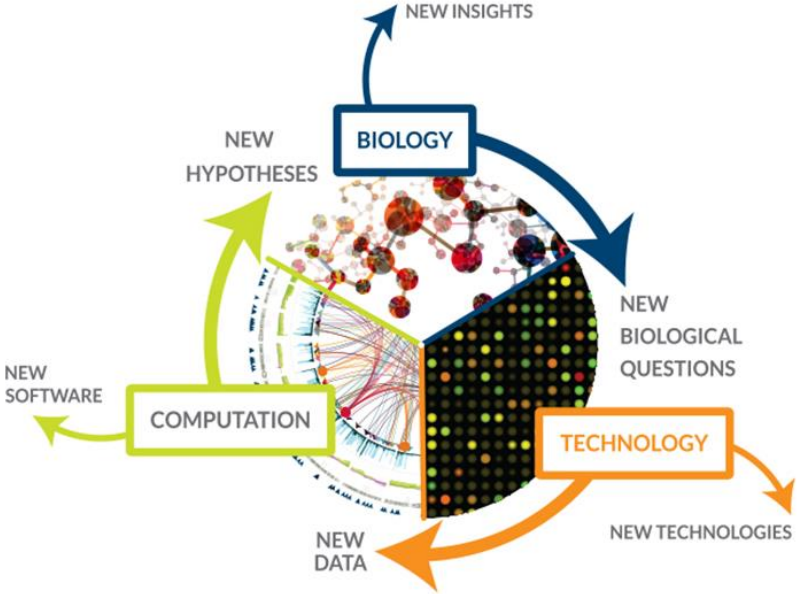
Kaggle's programming



What is computational biology?

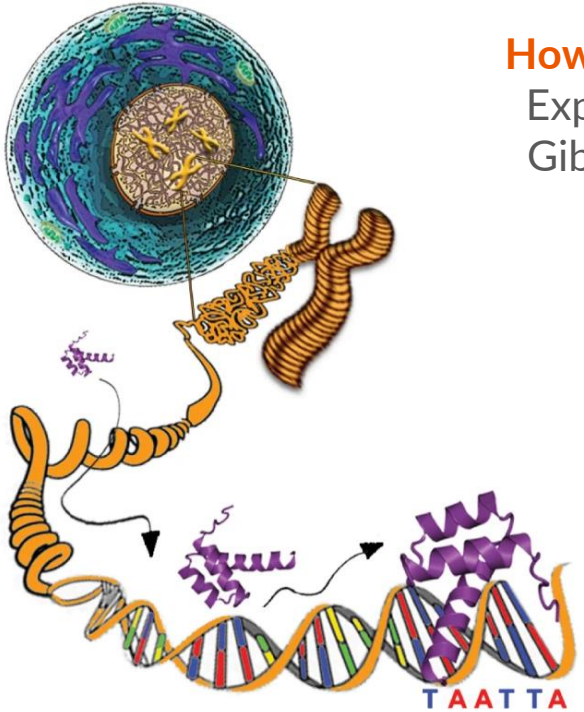


MIT OCW 7.36 lecture 1 slide by Chris Burge



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

Biology-inspired motif analysis



Castellanos *et al*. Nat Comm 2020

How to identify the motif?

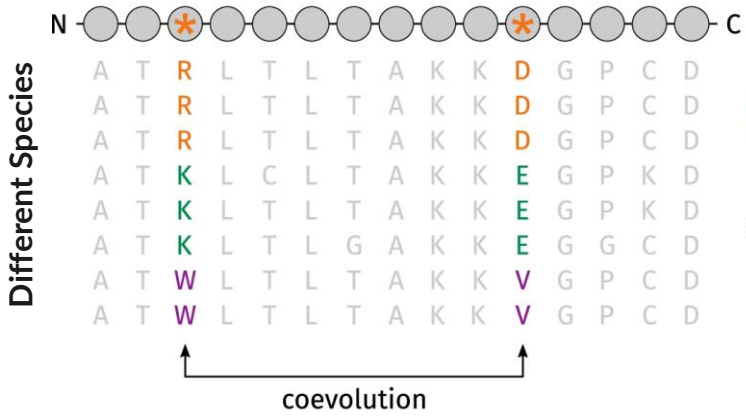
Experiment can isolate DNA sites with bound protein

Gibbs Sampling to identify similar patterns

CGGGGCTATcCAgCTGGGTCGTCACATTCCCCTT
TTTGAGGGTGCCCAATAAggGCAACTCCAAAGCGGACAAA
GGATGgAtCTGATGCCGTTTGACGACCTA
AAGGAaGCAACcCCAGGAGCGCCTTTGCTGG
AGATTATAATGTCGGTCCtTGgAACTTC
CAACTGAGATCATGCTGCATGCcAtTTTCAAC
TACATGATCTTTTGATGgcACTTGGATGAGGGAATGATGC

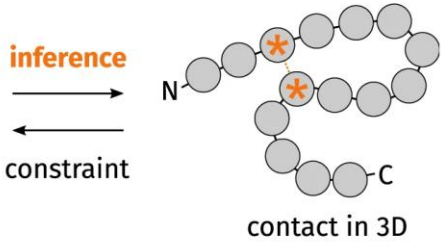
An Introduction to Bioinformatics Algorithm by Jones and Pevzner

Capturing co-evolution with mutual information



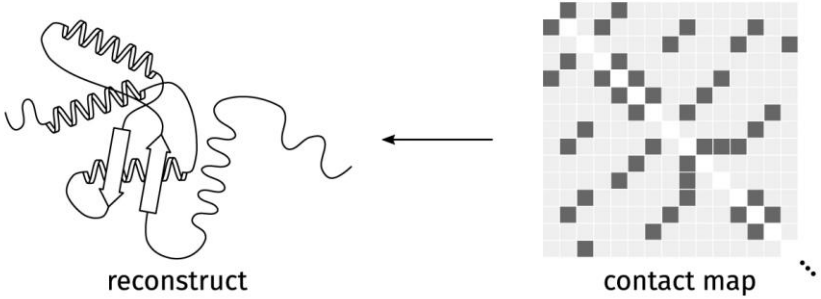
$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_{(X,Y)}(x,y) \log \left(\frac{P_{(X,Y)}(x,y)}{P_X(x)\, P_Y(y)} \right)$$

Mutual Information



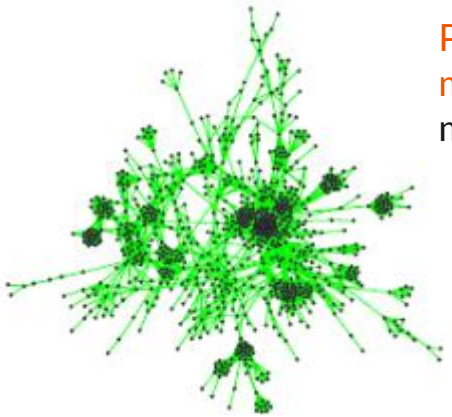
Positions in protein(s) co-evolve because they interact in some fashions – usually physical bindings

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



Bittrich *et al.* Sci Rep 2019

Using graph theory to study biological networks

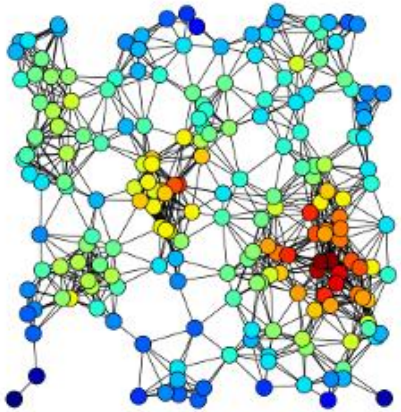


Proteins involved in many interactions might be important

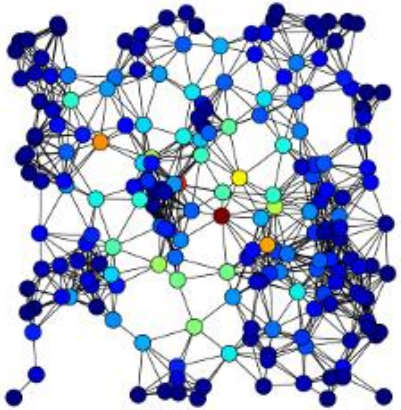
Proteins that connect other proteins might be important

Node = Protein
Edge = Protein-protein interaction

Degree

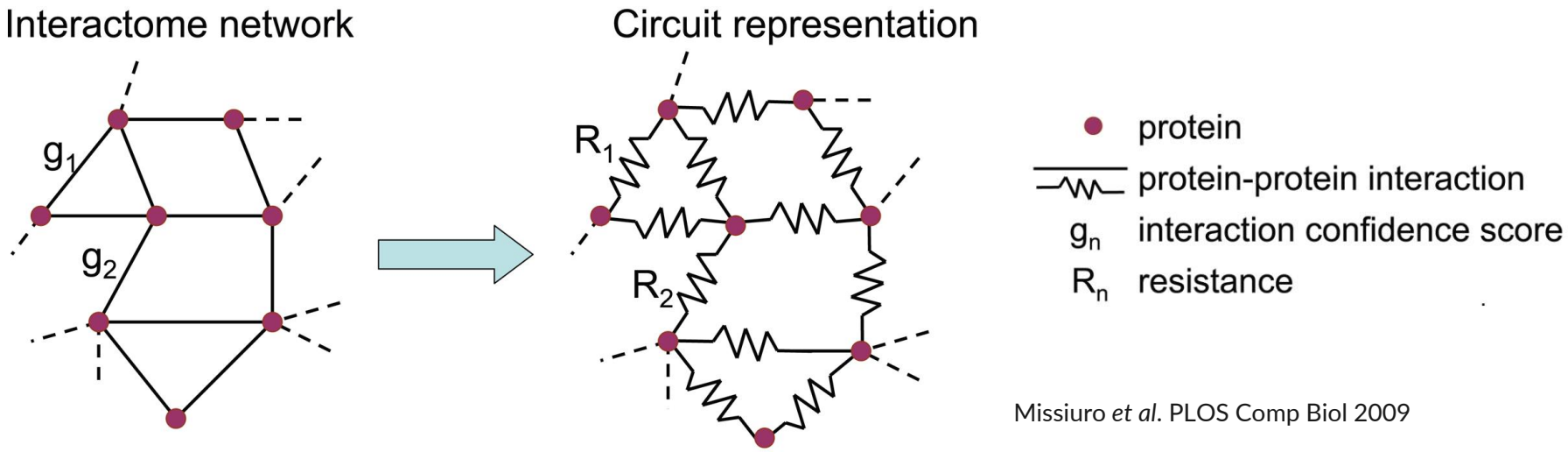


Betweenness



Images from wikipedia

Electrical circuit model for biological signal flows

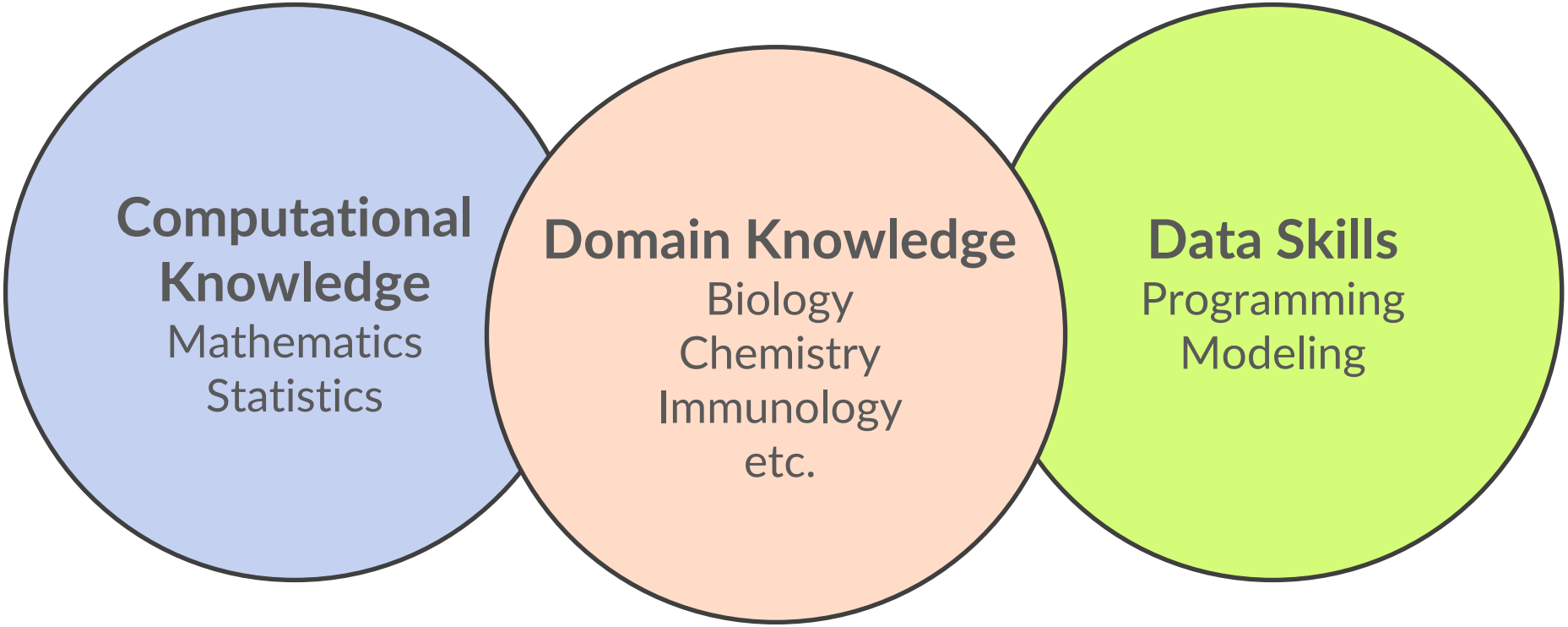


Missiuro *et al.* PLOS Comp Biol 2009

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 1: 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

- Which test to use?
  - Paired and unpaired *t*-tests
  - What about Mann-Whitney U? Wilcoxon rank-sum? Sign test?
  - Likelihood ratio test & maximum likelihood principle
  - Permutation test

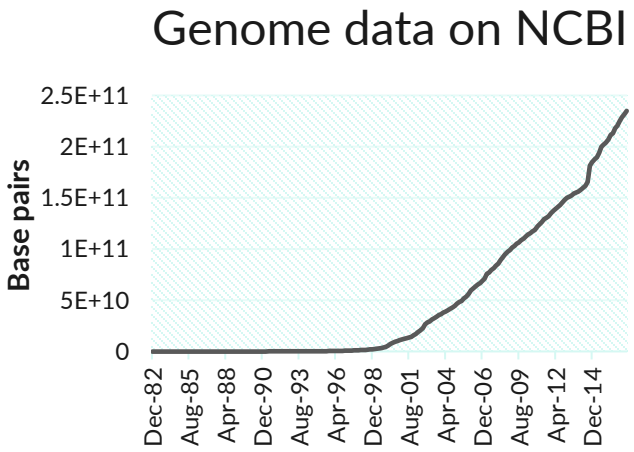
Module 2: DNA sequencing & applications

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

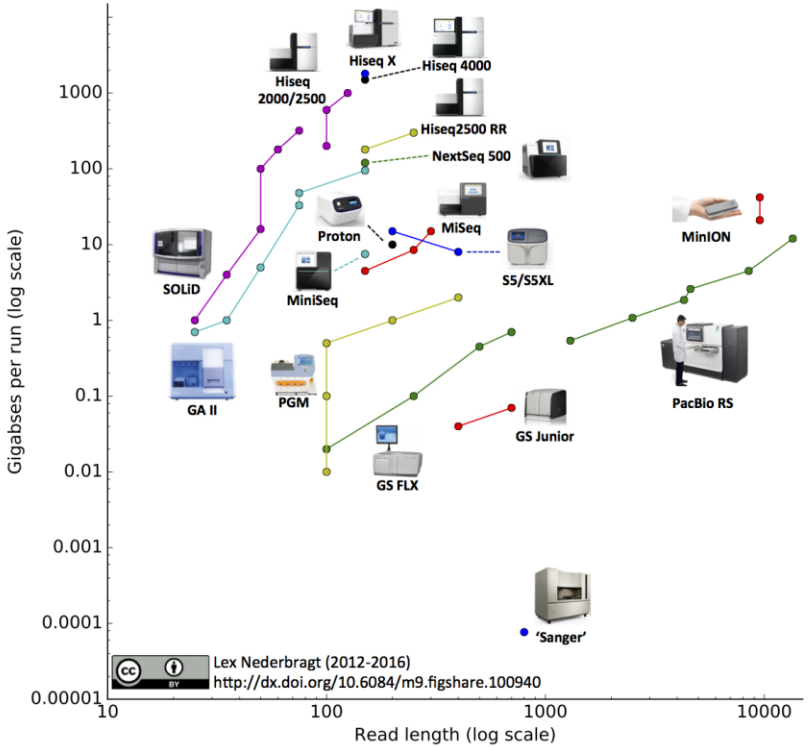
- First draft of human genome, 26 June 2000
  - BLAST sequence alignment

- Gene structure annotation
  - Exome sequencing
  - Oligo nucleotide microarray

Improvements in DNA sequencing



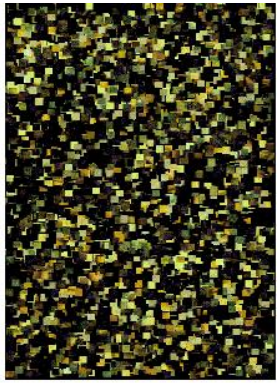
Plotted with data from
https://www.ncbi.nlm.nih.gov/genbank/statistics/



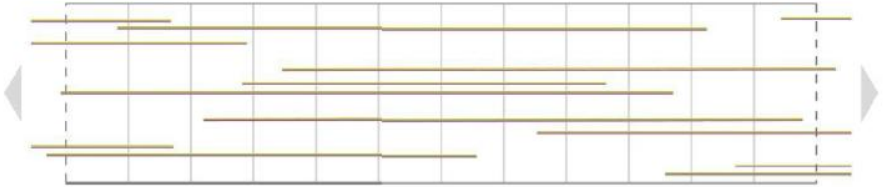
flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/

Combining short read and long read data

Short Reads



Long Reads



Power of long-read techniques

- How long is long?

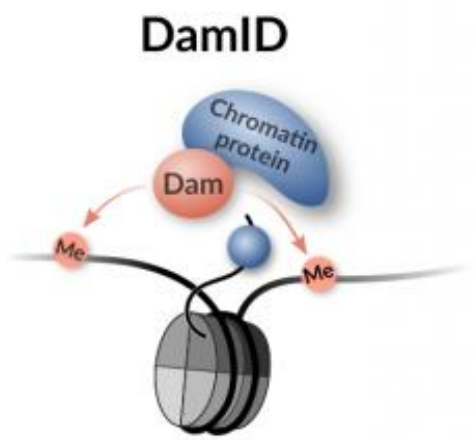
- More than A/T/C/G sequencing

- Resolving isoforms and haplotypes
- Resolving repetitive genomic regions

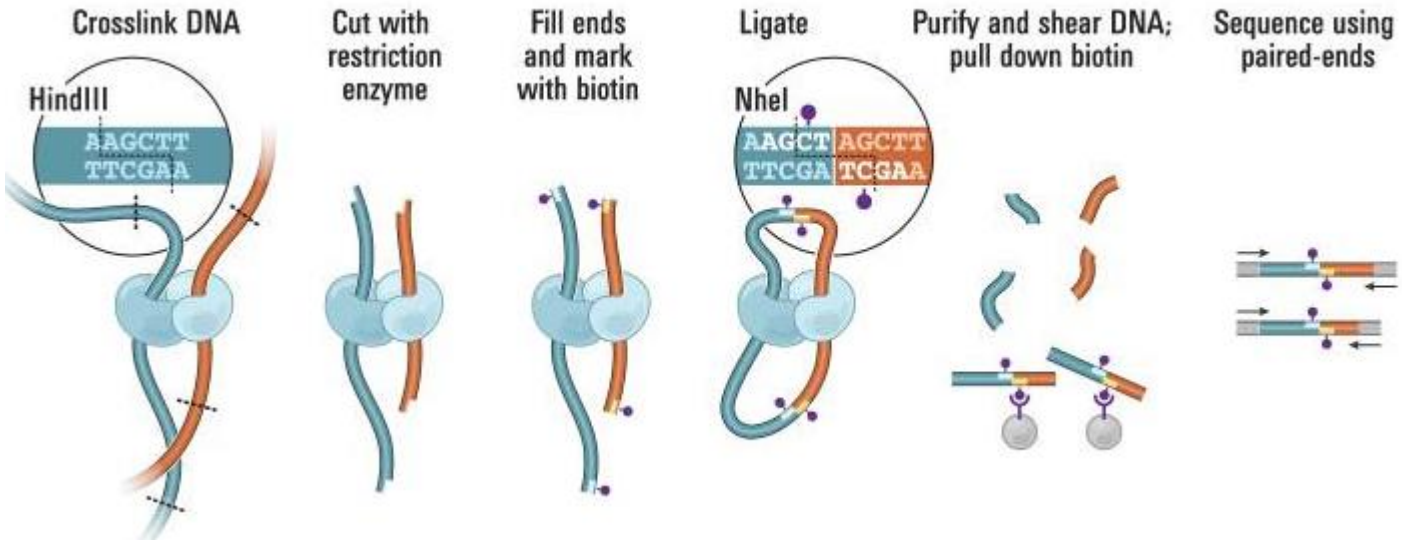
- No DNA amplification

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

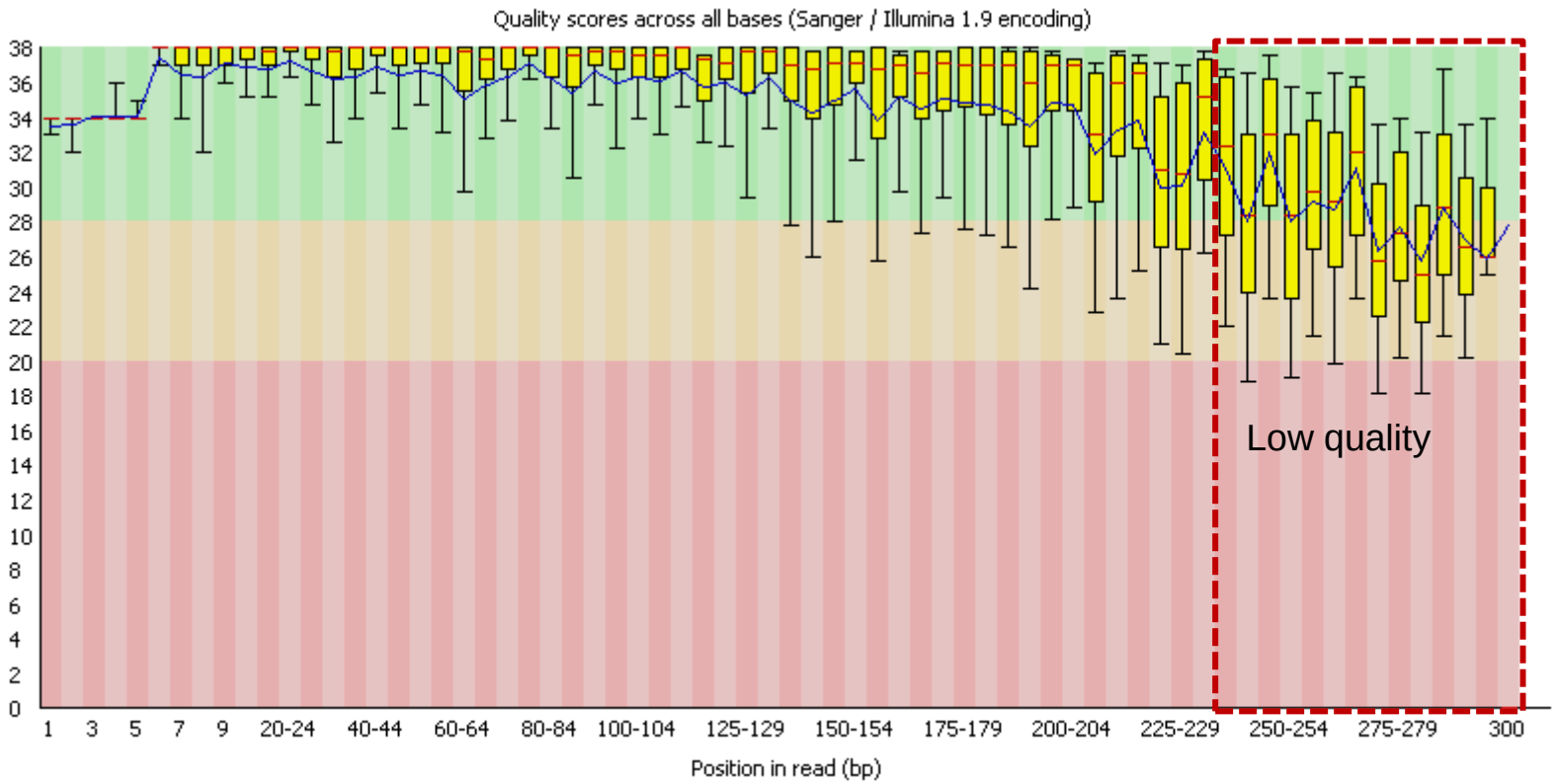


Hi-C: Chromatin folding structure



Lieberman-Aiden *et al*. Science 2009

DNA sequencing QC



Important file formats

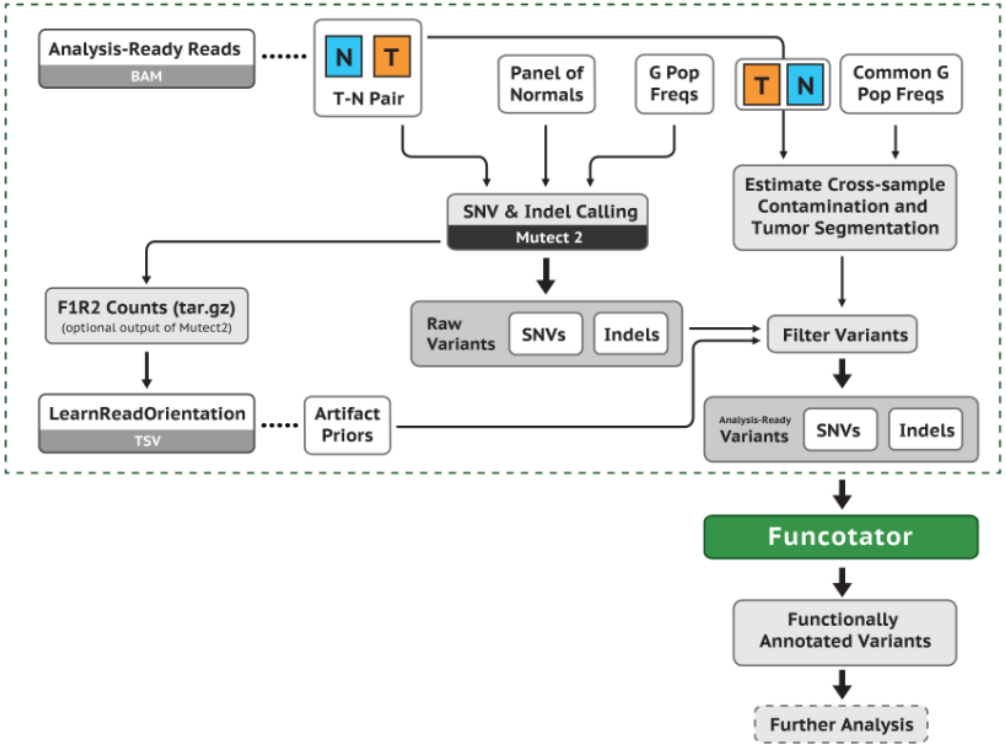
```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+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
```

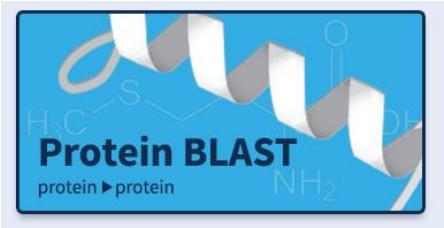
- SAM, BAM, BED, GTF, GFF, FASTA, FASTQ, VCF, MAF

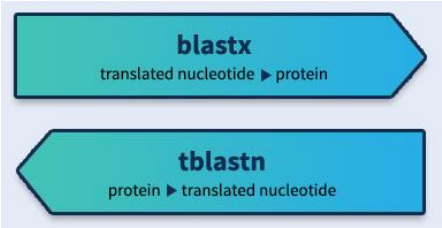
GATK variant calling pipelines

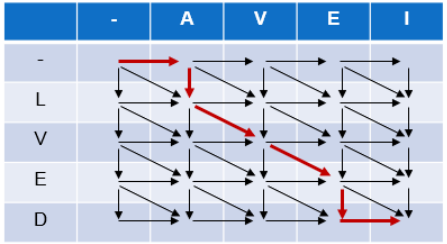


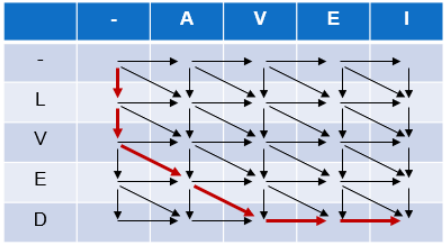
Basic Local Alignment Search Tool (BLAST)

- 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









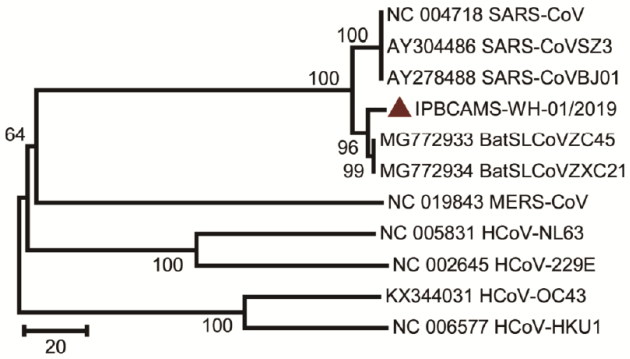
A – V E – I
– L V E D –

– – A V E I
L V E D – –

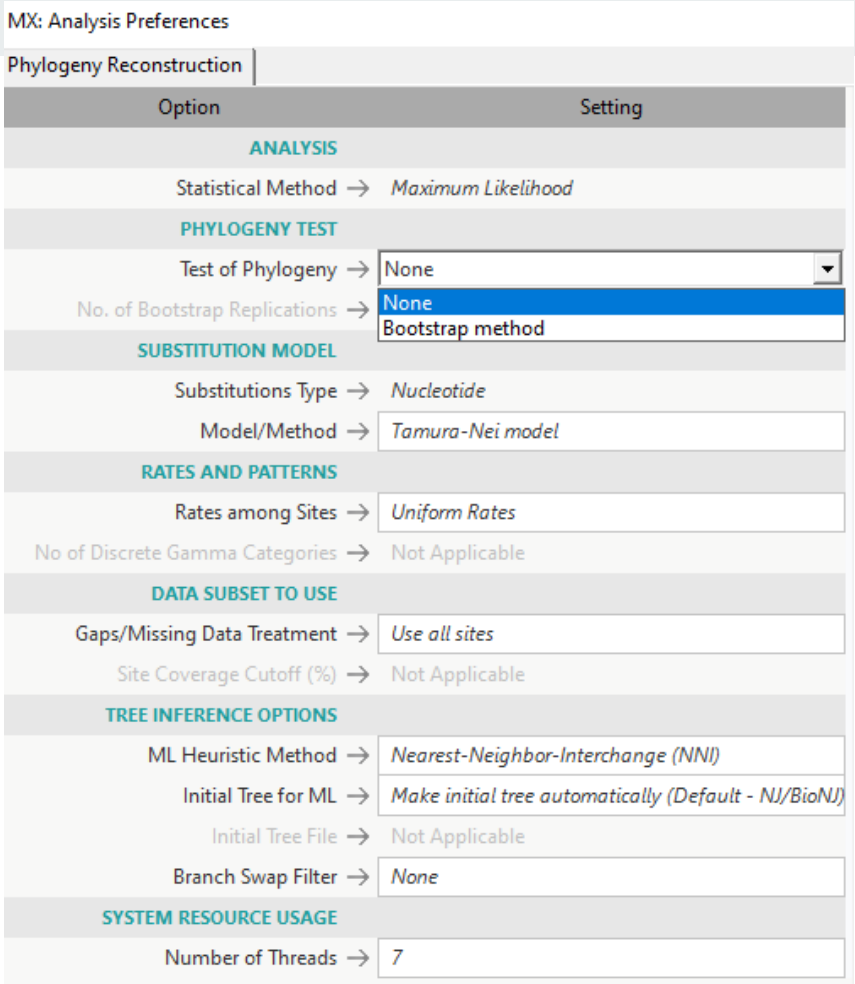
Phylogenetics

- Evolution is informative

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



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

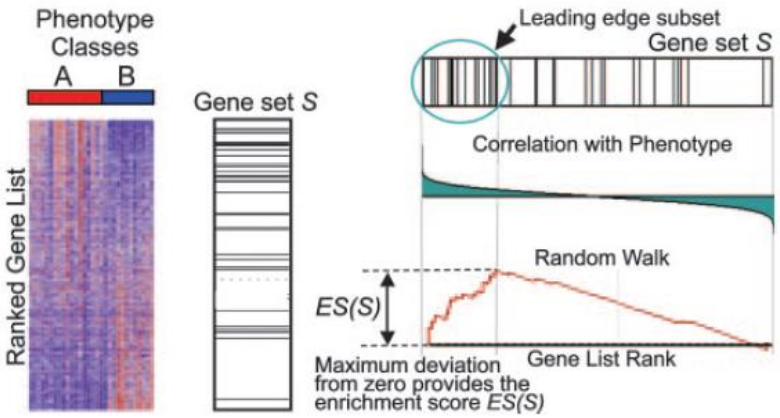


Module 3: Transcriptomics

- 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

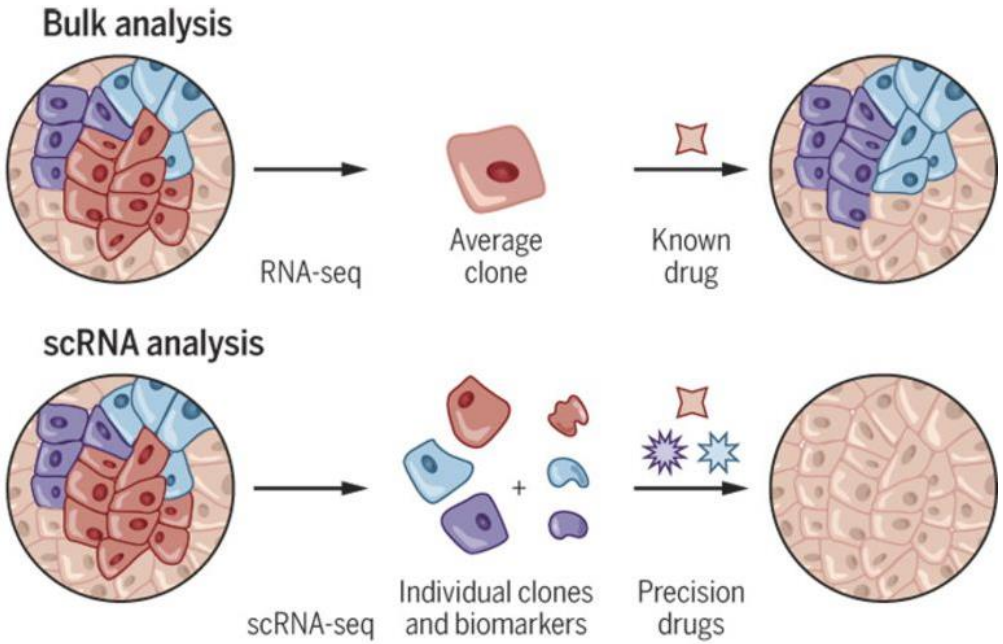
Functional enrichment analysis

- 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

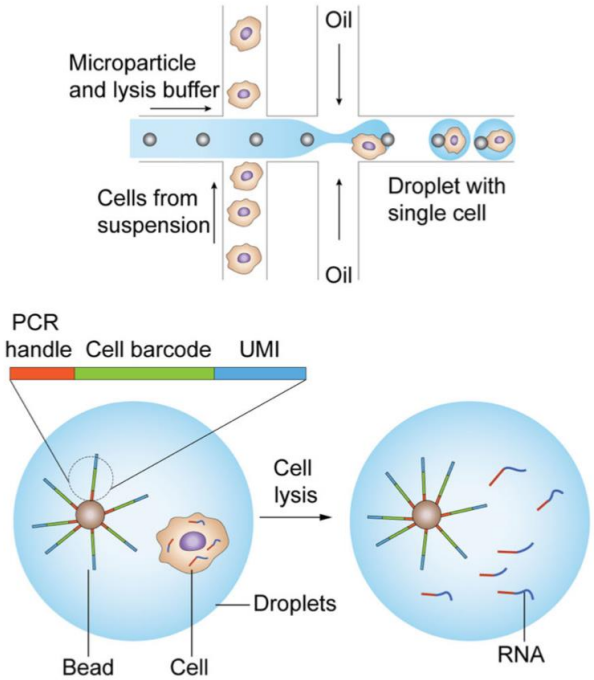


Subramanian *et al*. PNAS. 102:15545-15550 (2005)

Module 4: Advanced topics

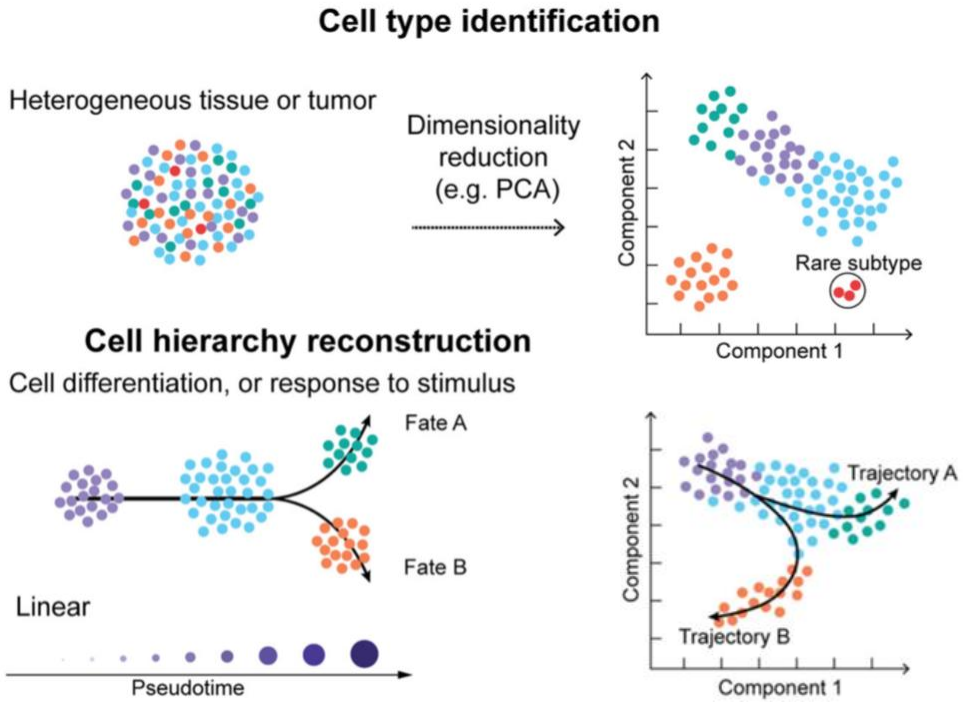


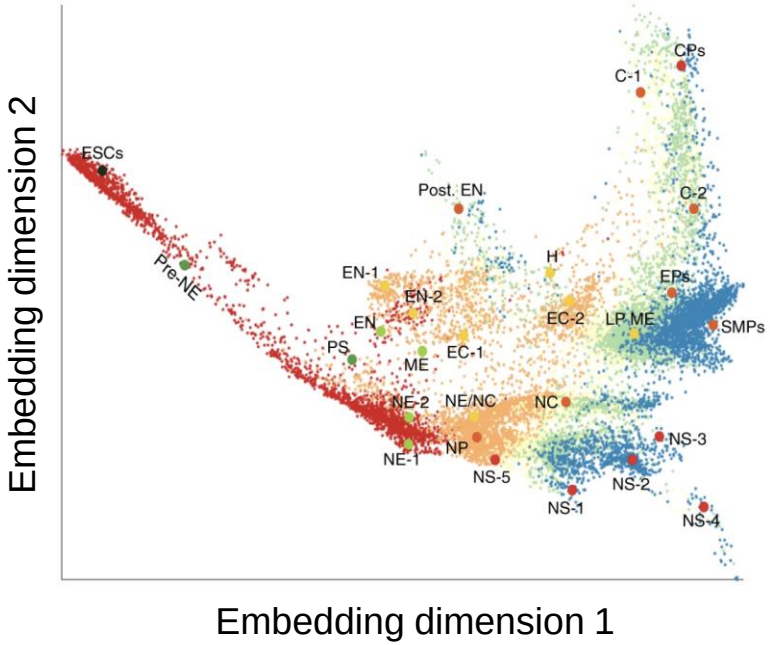
Shalek and Benson. Science Trans Med. 9:eaan4730 (2017)



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

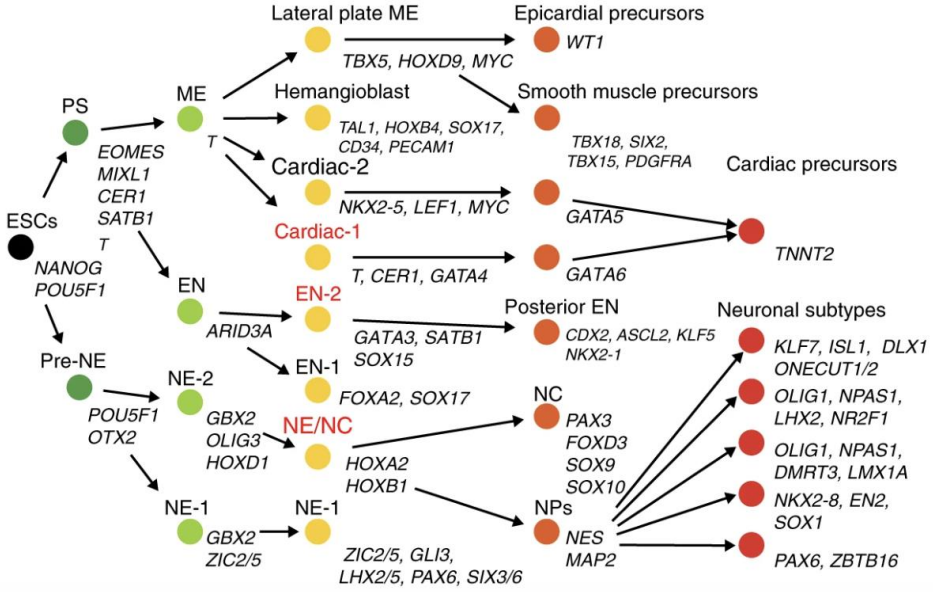
Power of single-cell data

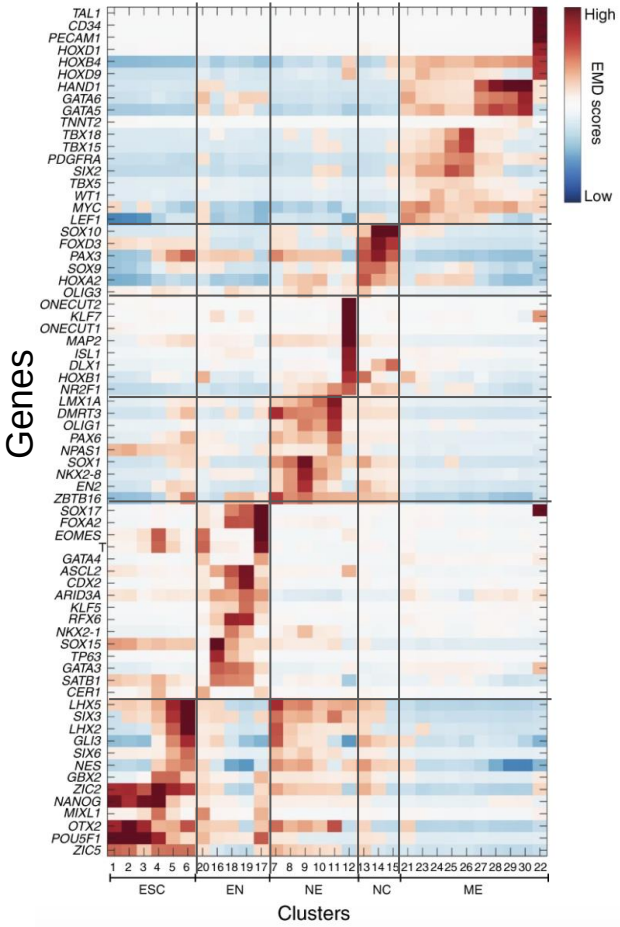




Hwang *et al.* Exp & Mol Med 50:96 (2018)

Power of single-cell data





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

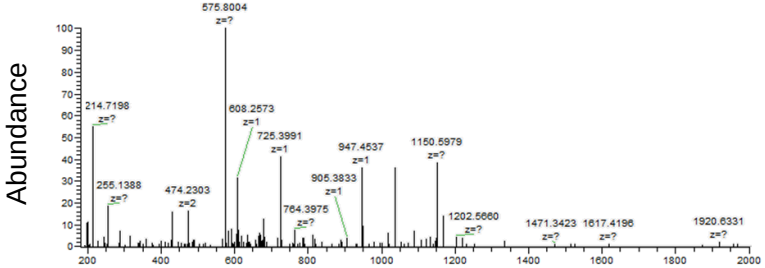
Proteomics & mass spectrometry

Peptides mixture

Mass spectrometer

Image from planetorbitrap.com

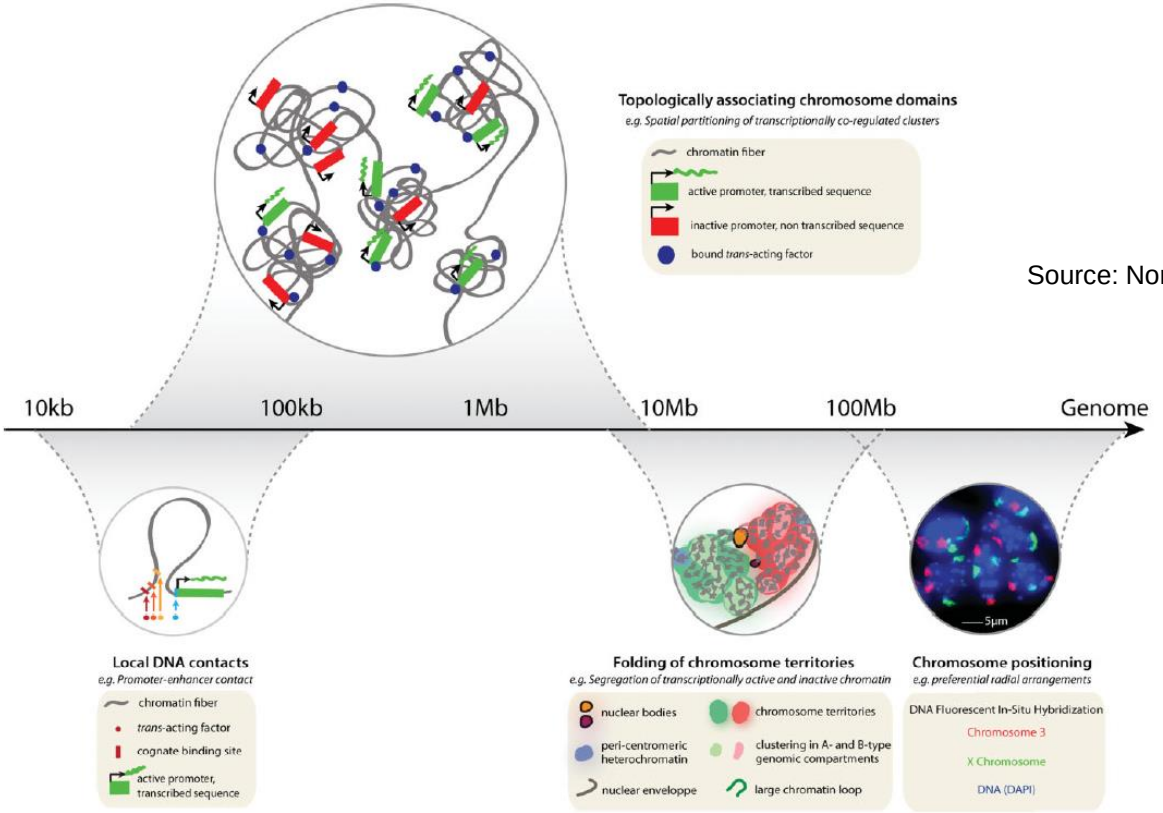
Peptide mass spectrum



SAPYSAMIADQVAQR

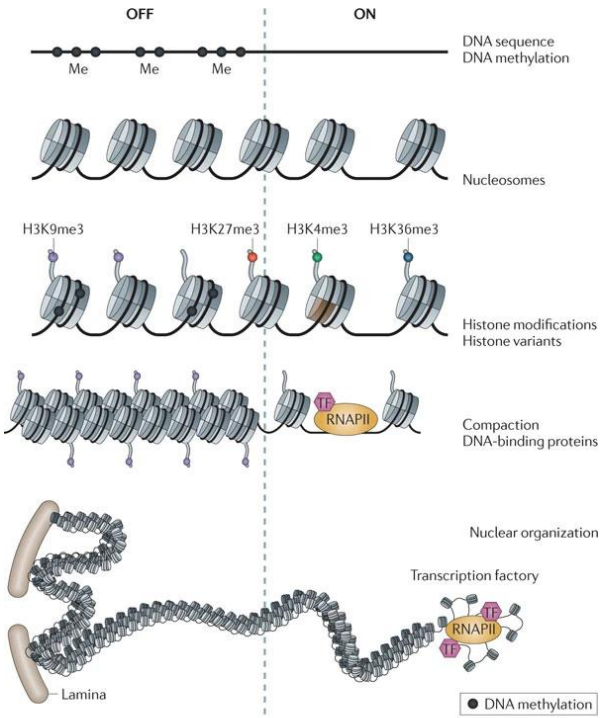
Bassani-Sternberg *et al*. Mol Cell Protomics 14(3), 658-673 (2015)

Chromatin organization



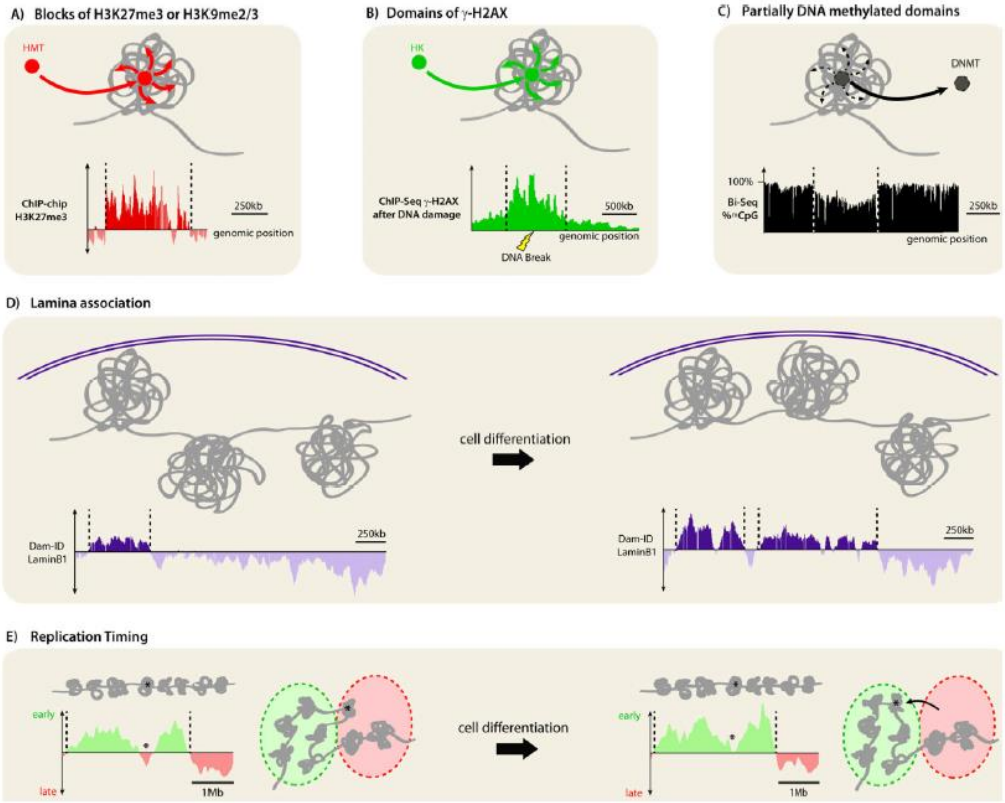
Source: Nora *et al*. Bioessay (2013)

Epigenetics and gene expression regulation



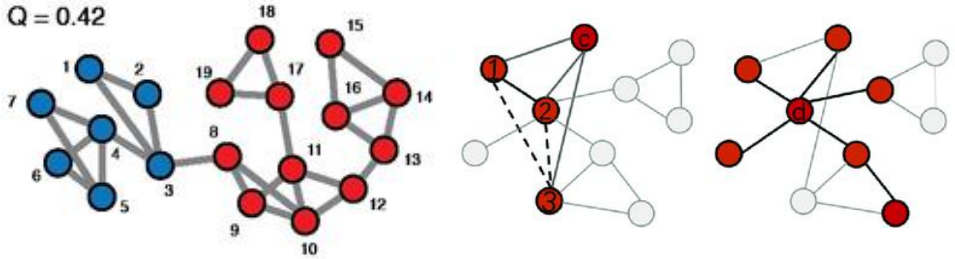
Zhou *et al*. Nat Rev Genet (2010)

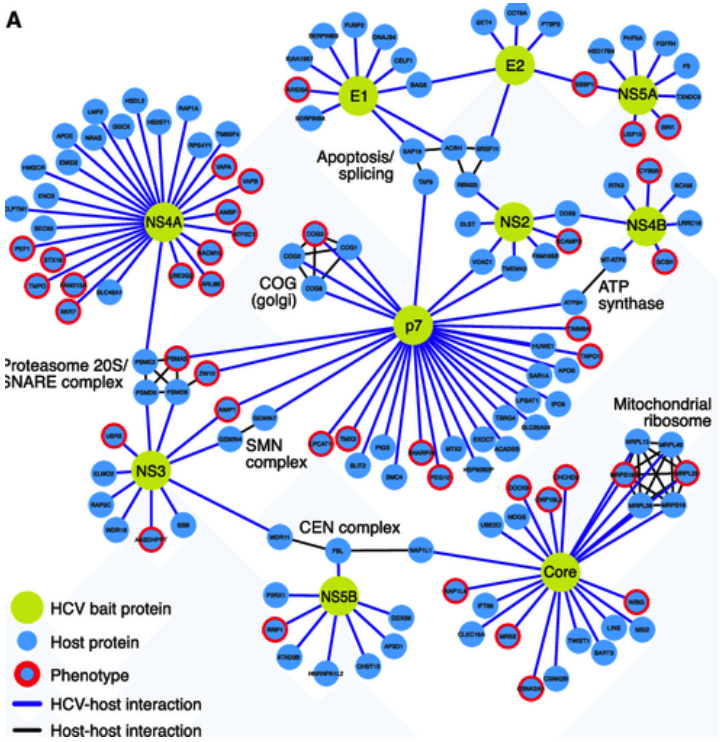
Nature Reviews | Genetics



Biological networks

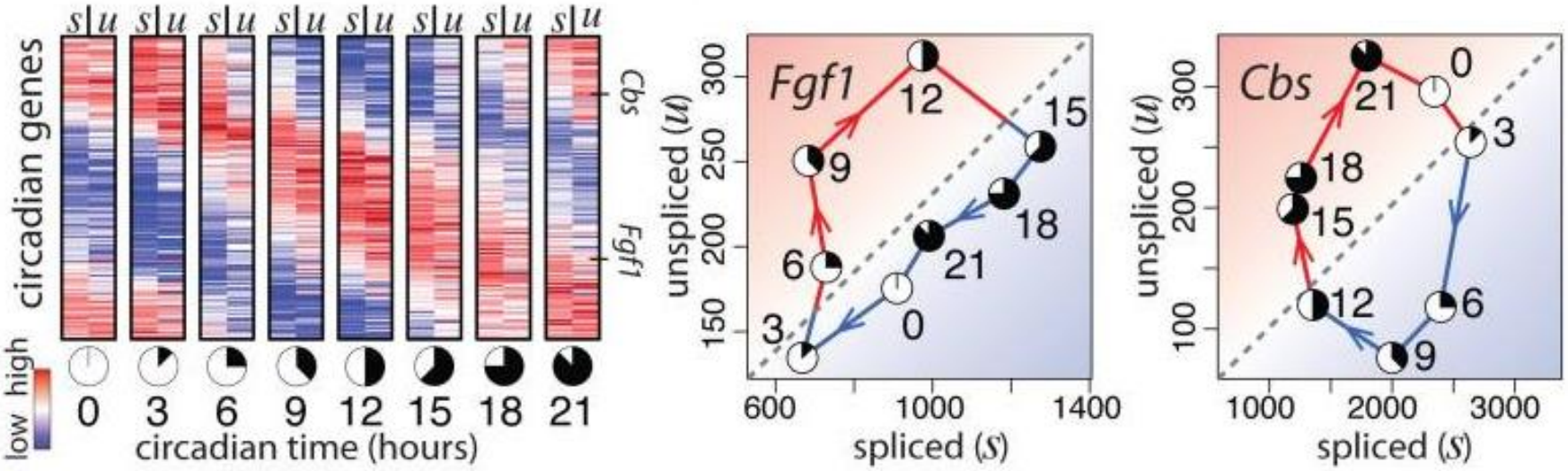
- Topological properties of networks
  - Relationship to biology
- Applications in biomedicine
- Visualization and analysis with CytoScape





Source: Ramage et al. Mol Cell (2015)

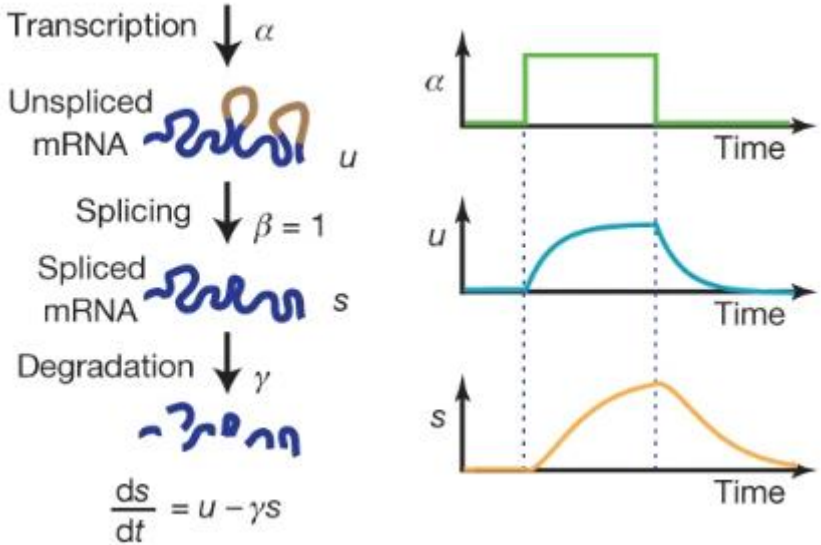
A taste of dynamics modeling in Systems Biology



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

Module 5: Python skills

- Kaggle's programming courses

- In-class practice & problem sets
  - Handling of tabular data
  - Statistical analyses
  - Visualizations

- We will spend 7 sessions on top of Kaggle's courses to get a solid foundation

Module 6: 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

Any questions?

See you next week on August 21st