3000788 Intro to Comp Molec Biol

Week 9: Biological networks & chromatin organization

Fall 2024

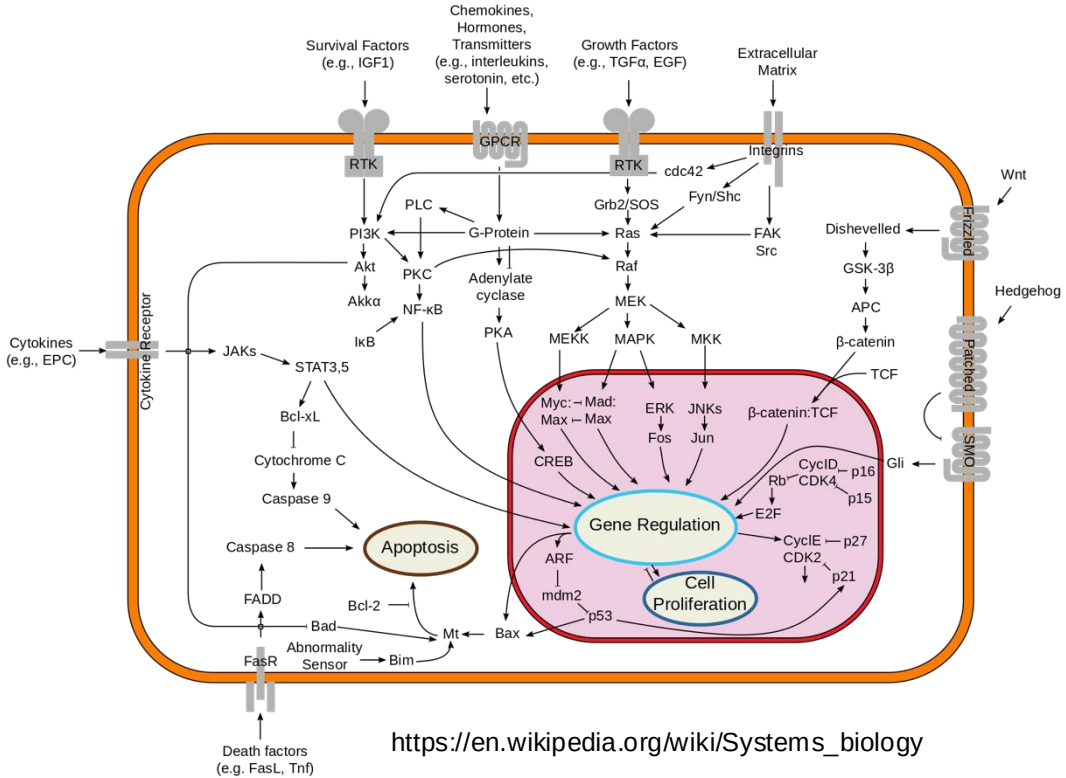
Sira Sriswasdi, PhD

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

Part 1: Systems biology

- Biological systems consists of components and interactions
- Gene expression = snapshot of gene activity
- Interaction network = mechanistic diagram
- Systems biology = understanding the dynamics of the systems
  - Multi-omics
  - Model the time-dependent dynamics

Systems biology

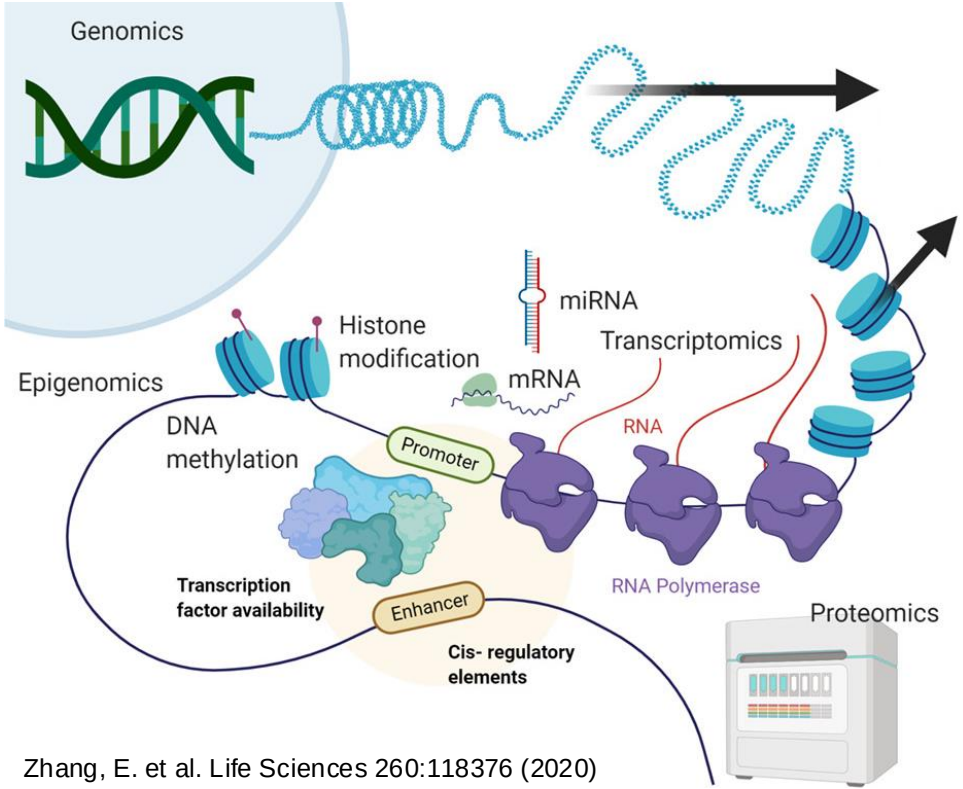


https://en.wikipedia.org/wiki/Systems_biology

- A (biological) system consists of components (genes) and rules (gene expression regulations) that control its characteristics (phenotypes)

- Systems biology = integration of data and model to fully understand the system

Multi-omics approach

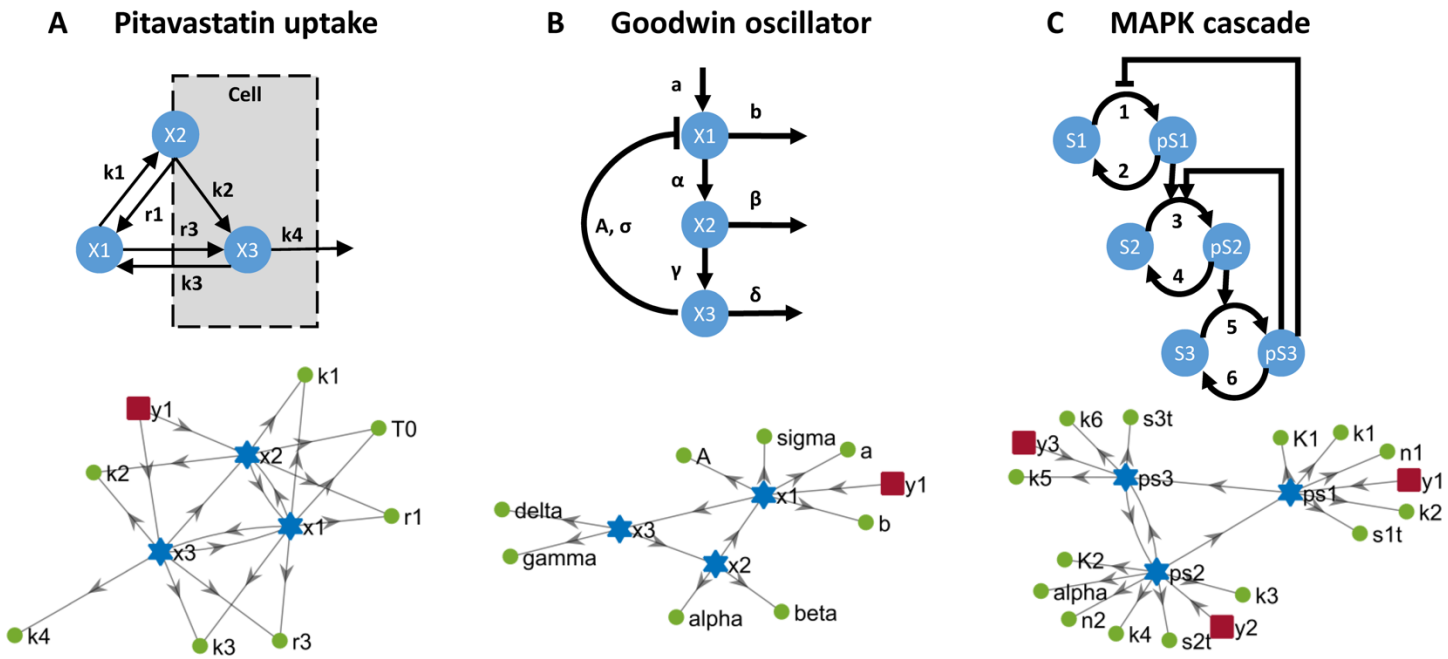


Zhang, E. et al. Life Sciences 260:118376 (2020)

- Genomics → gene state
- Epigenomics → chromatin state
- Transcriptomics → gene expression
- Proteomics → protein expression
- Other assays → protein function
- Provide mechanistic understanding

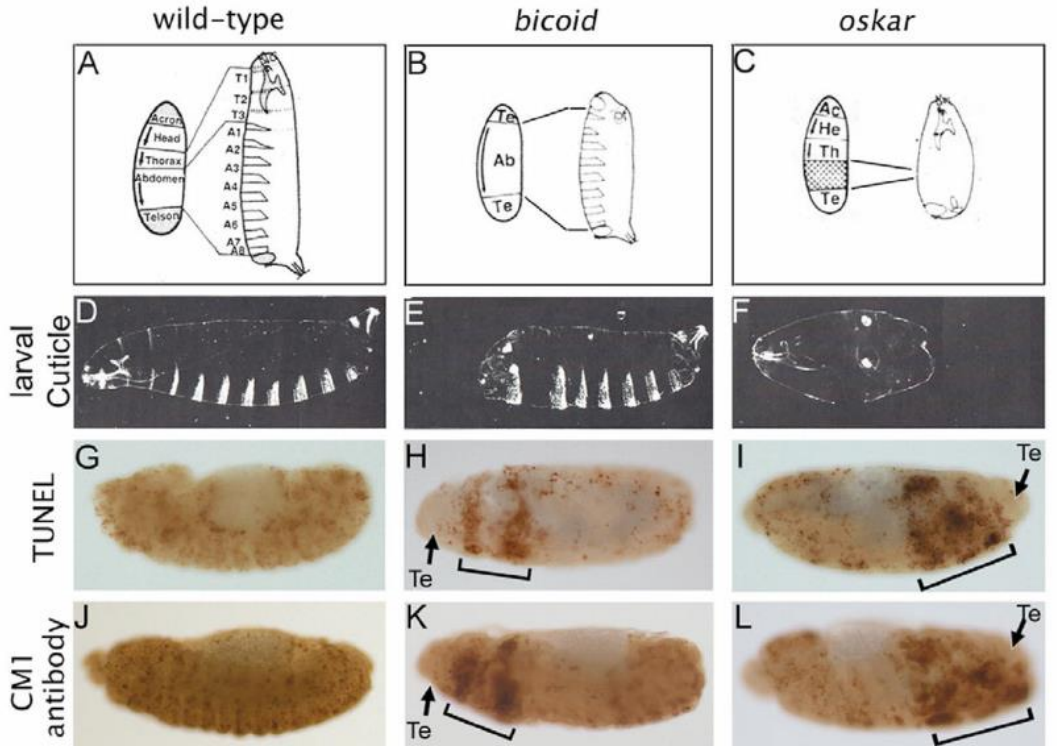
Modeling approach

Villaverde, A.F. PLoS Comp Biol 12:e1005153 (2016)



- Represent molecular interaction as network of physical-chemical reactions

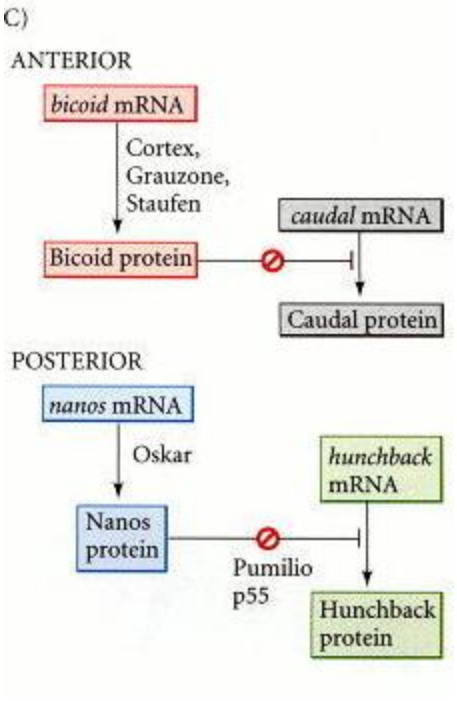
Why is systems biology interesting?

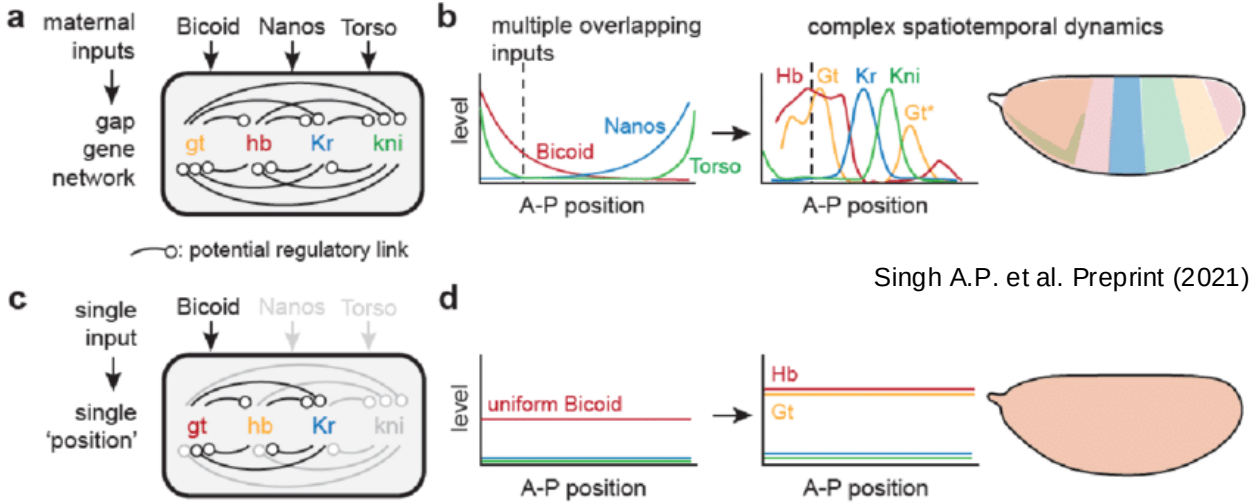


- How does gene and protein know where to activate in the body during development?
- Each cell starts in identical state!

Werz, C. Development 132:5343-52 (2006)

Why is systems biology interesting?





Singh A.P. et al. Preprint (2021)

- Gene regulatory dynamics lead to specific expression pattern over time

https://www.ncbi.nlm.nih.gov/books/NBK10039/

A simple enzymatic reaction

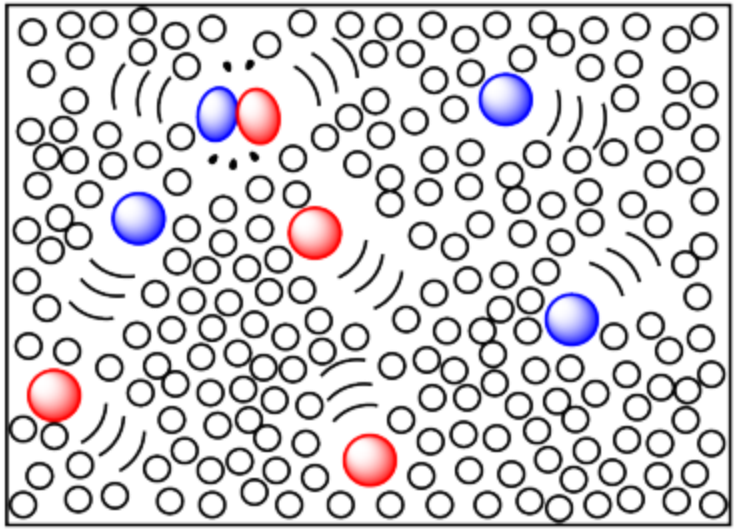
$$E + S \underset{k_{off}}{\overset{k_{on}}{\leftrightarrow}} ES \xrightarrow{k_{react}} E + P$$

- $\frac{d[S]}{dt}$ = Rate of change in S = $-k_{on}[E][S] + k_{off}[ES]$
- $\frac{d[E]}{dt}$ = Rate of change in E = $-k_{on}[E][S] + k_{off}[ES] + k_{react}[ES]$
- $\frac{d[P]}{dt}$ = Rate of change in P = $k_{react}[ES]$

A key mental image

Rate of E + S → ES =
Rate of E meeting S x
Rate of binding

Rate of E + S → ES =
[E][S] x k



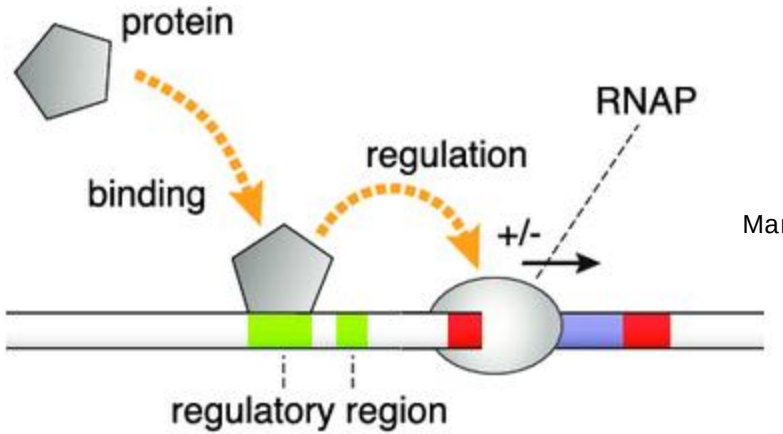
https://employees.csbsju.edu/cschaller/Reactivity/kinetics/rkphase.htm

Rate of E meeting S scales linearly with [E] and [S]

Rate of binding is a constant for E and S

- Molecules must find each other in 3D to bind and interact
- Reaction takes time (and typically energy)

From enzymatic reaction to transcription



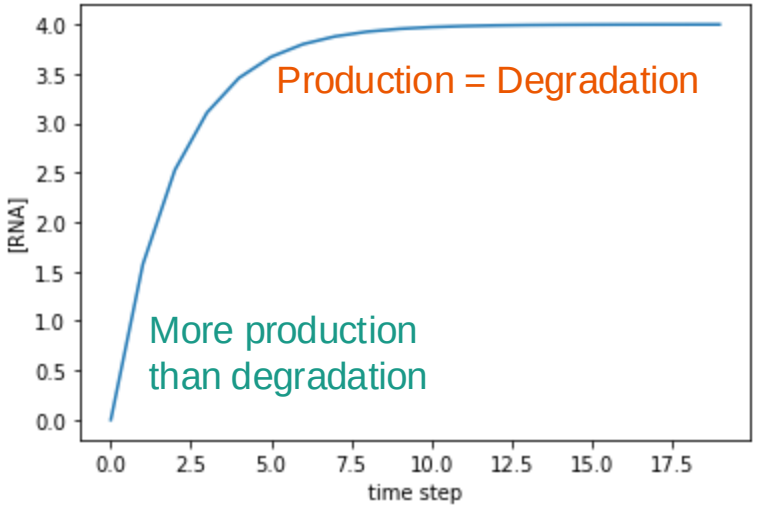
Marbach, D. et al. EvoBIO 2007

- $TF + DNA_{inactive} \leftrightarrow TF\text{-}DNA$
- $RNAP + TF\text{-}DNA \rightarrow RNAP + TF\text{-}DNA + RNA$
- $RNA \rightarrow$ degraded RNA

A differential equation for gene expression

$$DNA \xrightarrow{k_{transcription}} DNA + RNA$$

$$RNA \xrightarrow{k_{degradation}} \emptyset$$



- $\frac{d[RNA]}{dt}$ = rate of change of RNA = $k_{transcription} - k_{degradation}[RNA]$
- What would the graph of [RNA] looks like?

Simulating differential equation

- $\frac{d[RNA]}{dt} = k_{transcription} - k_{degradation}[RNA]$

- Start at an initial condition $x(0) = x_0$,
  - Calculate $x(1) = x(0) + x'(0) = x_0 + k_{transcription} - k_{degradation}x_0$
  - Calculate $x(2) = x(1) + x'(1) = x(1) + k_{transcription} - k_{degradation}x(1)$
  - $= x_0 + 2 \cdot k_{transcription} - k_{degradation}(x_0 + x(1))$
  - [RNA] after 2 timesteps = initial [RNA] + 2 units of transcription – combined degradation of [RNA] at both timesteps

A toy example

- $\frac{d[RNA]}{dt} = 4 - 0.2\,[RNA]$

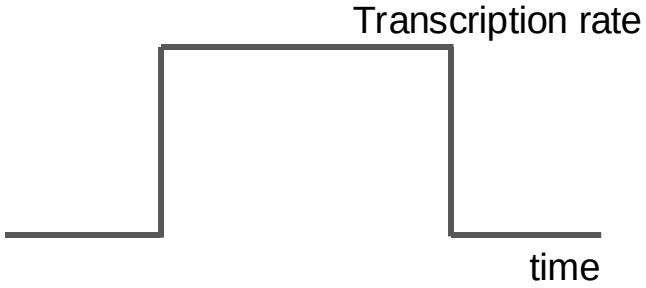
- Start at an initial condition $x(0) = 100$,
  - $x(1) = 100 + 4 - 20 = 84$
  - $x(2) = 84 + 4 - 17 = 71$
  - $x(3) = 71 + 4 - 14 = 61$
  - $x(4) = 61 + 4 - 12 = 53$
  - $x(5) = 53 + 4 - 10 = 47$

Another toy example

- $\frac{d[RNA]}{dt} = 4 - 0.2\,[RNA]$

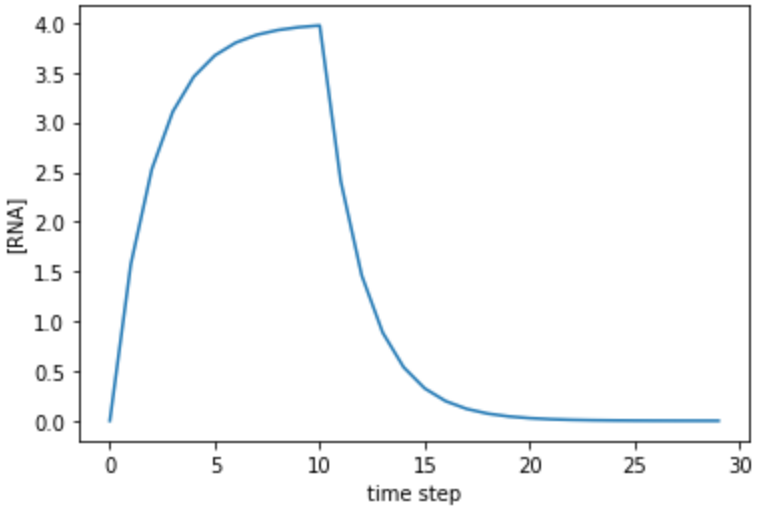
- Start at an initial condition $x(0) = 0$,
  - $x(1) = 0 + 4 - 0 = 4$
  - $x(2) = 4 + 4 - 1 = 7$
  - $x(3) = 7 + 4 - 1 = 10$
  - $x(4) = 10 + 4 - 2 = 12$
  - $x(5) = 12 + 4 - 2 = 14$

Time-dependent activation

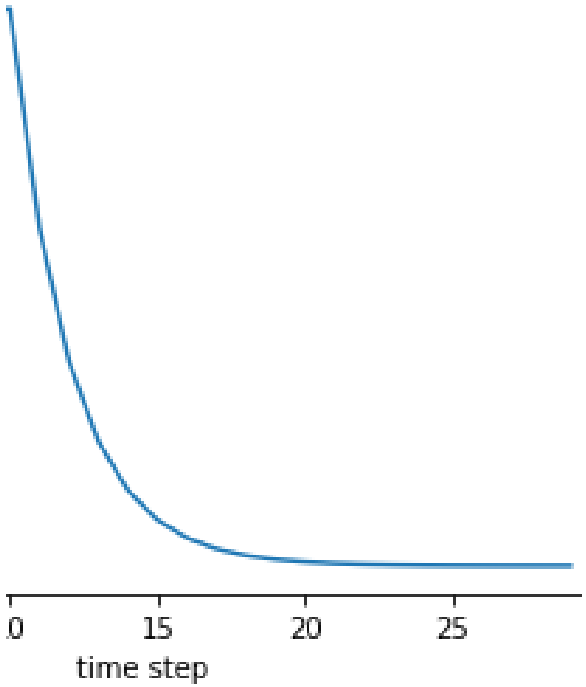



```python
def time_transcription(rna, time, k_trans, k_deg):
    if time < 10:
        return k_trans - k_deg * rna
    else:
        return - k_deg * rna
```

- First half of the dynamics is the same as before
- Second half is when RNA degrades until the [RNA] falls back to the basal level

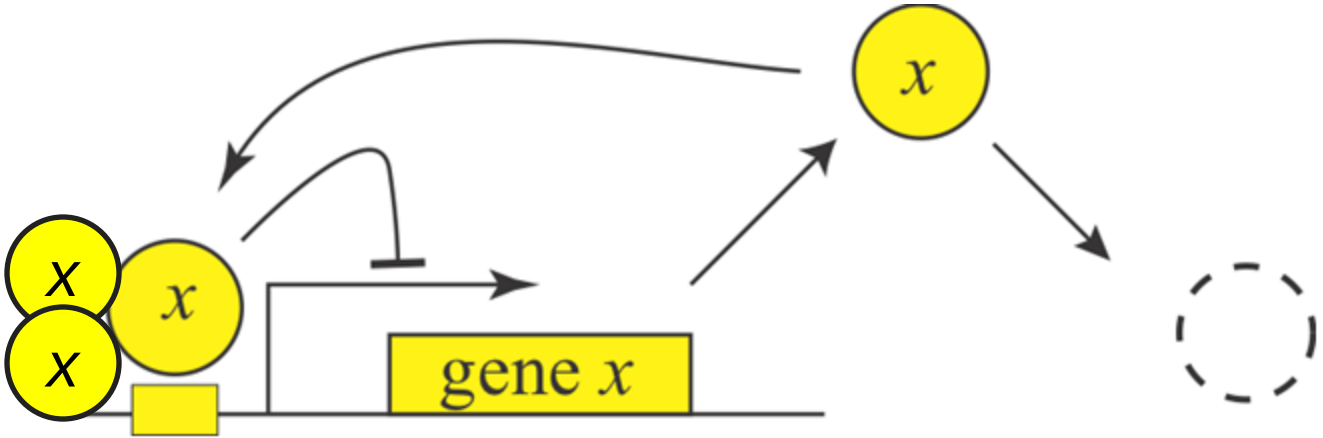


Exponential decay



- $\frac{d[RNA]}{dt} = -k_{\text{degradation}}[RNA]$
- Fast decay in the beginning because there are a lot of RNA molecules
- Slower decay at the end
- $\frac{de^{-kt}}{dt} = -ke^{-kt}$

Negative auto-regulation



http://be150.caltech.edu/2019/handouts/03_small_circuits.html

- $\frac{d[X]}{dt} = k_{transcription}[DNA_{active}] - k_{degradation}[X]$
- $[DNA_{active}]$ also depends on $[X]$ through binding dynamics
- Two variables!

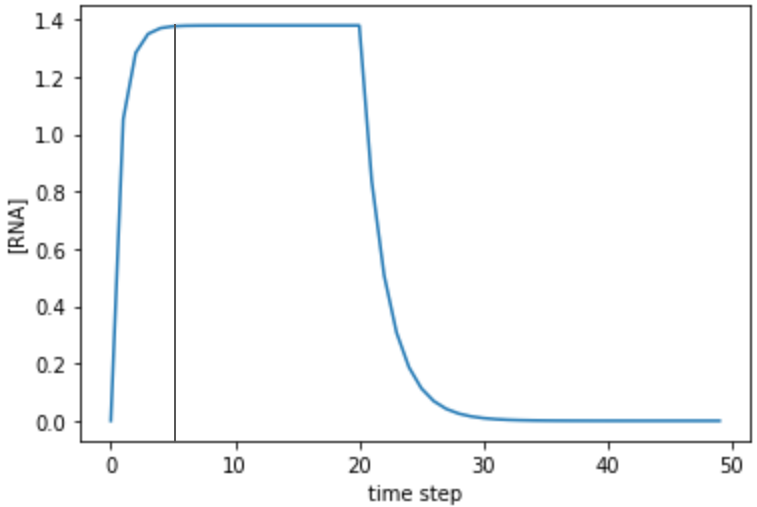
Negative auto-regulation

$$nX + DNA_{active} \underset{k_{unbind}}{\overset{k_{bind}}{\leftrightarrow}} DNA_{inactive}$$

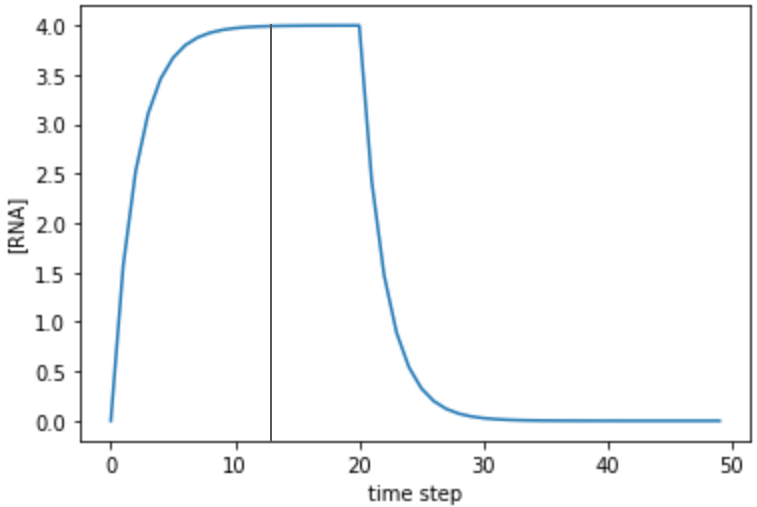
- $\frac{dDNA_{inactive}}{dt} = k_{bind}[X]^n[DNA_{active}] - k_{unbind}[DNA_{inactive}]$
- At binding equilibrium, fraction of unbound (active) DNA $= \frac{k_{bind}[X]^n}{k_{unbind}+k_{bind}[X]^n}$
- Also known as Hill function

Faster response time with negative auto-regulation

With negative auto-regulation

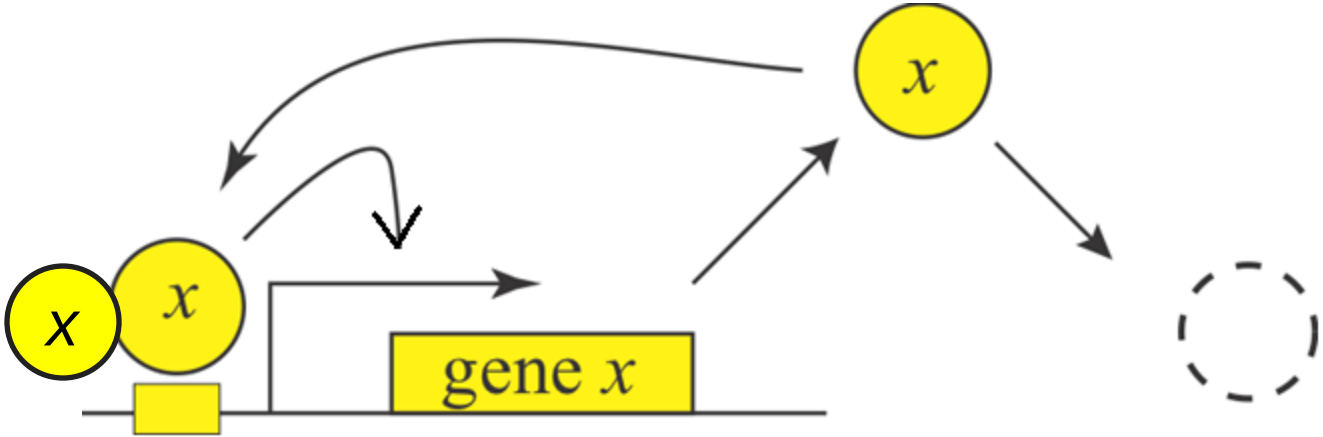


Without negative auto-regulation



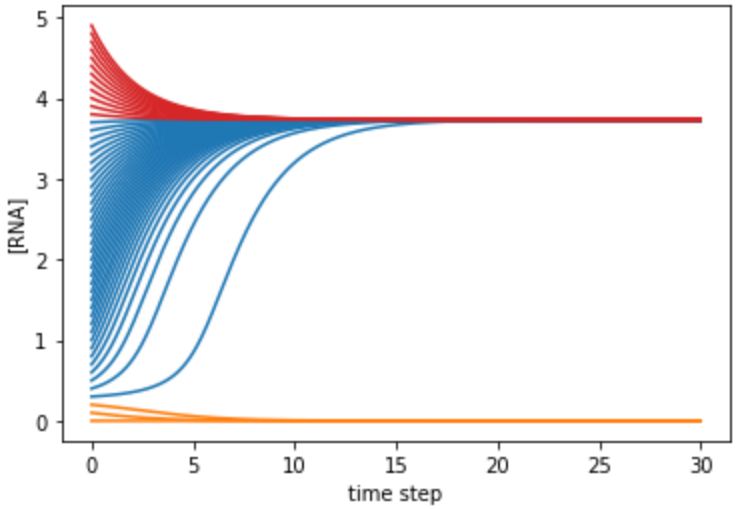
- $\frac{d[X]}{dt} = \frac{k_{transcription}}{1+(k[X])^n} - k_{degradation}[X]$

Positive auto-regulation



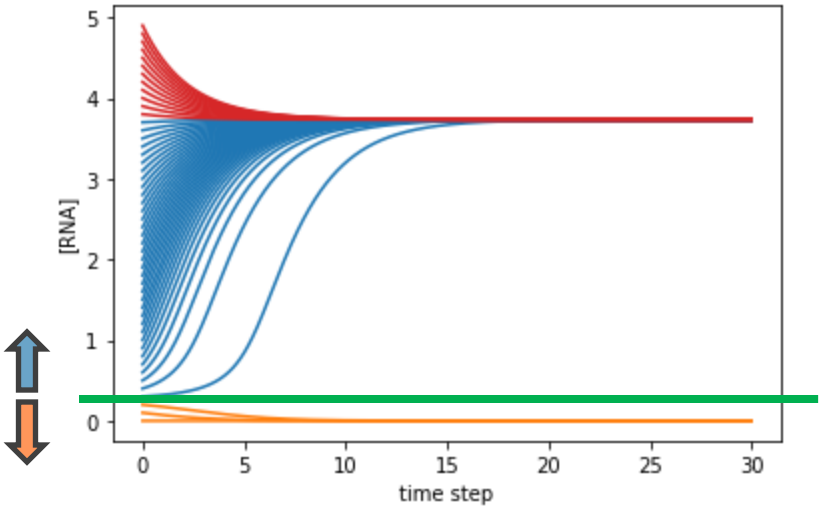
- $\frac{d[X]}{dt} = \frac{k_{transcription}(k[X])^2}{1+(k[X])^2} - k_{degradation}[X]$
- Can you guess how this system will behave?
  - How many equilibria are there?

Bistability



- For **low** [X], the degradation term wins over the transcription and [X] decays down to zero
- For **intermediate** [X], the transcription term wins over degradation and [X] increases until reaching the other equilibrium
- For **high** [X], the degradation term wins over the transcription and [X] decays down to the nearest equilibrium

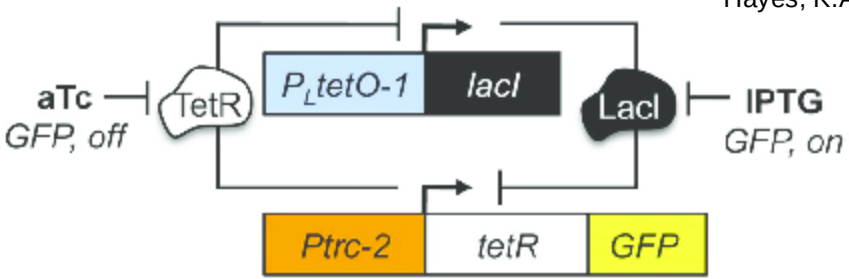
Bistability as memory



- Given a cell in low expression state
- A drug treatment raises the expression level in the cell pass the green threshold
- After stopping the treatment, the cell will “**memorize**” that it is now in the high expression state

Memory from gene toggle switch

Hayes, K.A. et al. J of Cell Biology 187:589-96 (2009)



- Two genes repressing each other
- **Two equilibria**: High LacI or high GFP

- If GFP is ON, how can we turn it off? Adding aTc or IPTG?
- Once GFP is OFF, can we stop supplying those molecules to the cell?

Two-gene linear system

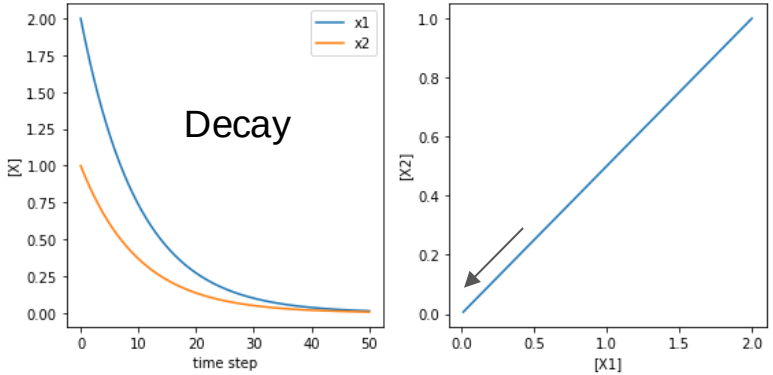
$$\frac{d[X1]}{dt} = k_{1,1}[X1] + k_{1,2}[X2]$$

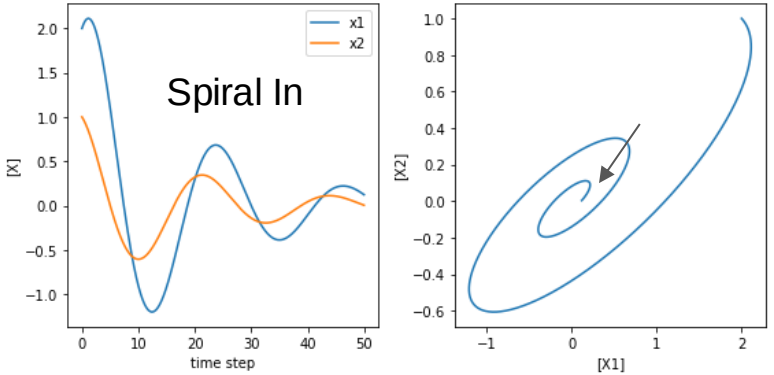
$$\frac{d[X2]}{dt} = k_{2,1}[X1] + k_{2,2}[X2]$$

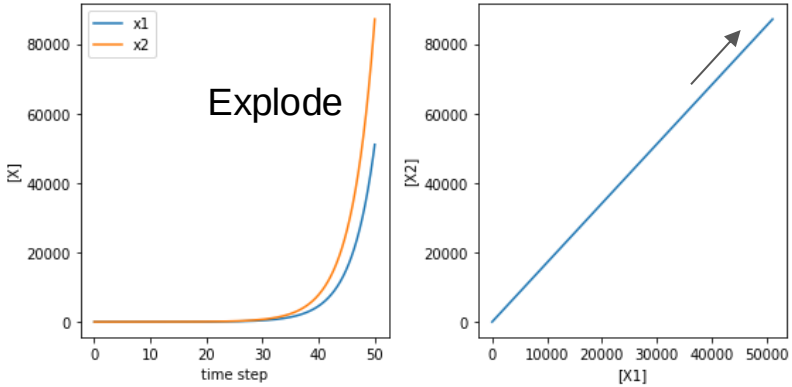
```python
def two_loci_linear(rna, time, k11, k12, k21, k22):
    dx1_dt = k11 * rna[0] + k12 * rna[1]
    dx2_dt = k21 * rna[0] + k22 * rna[1]

    return [dx1_dt, dx2_dt]
```

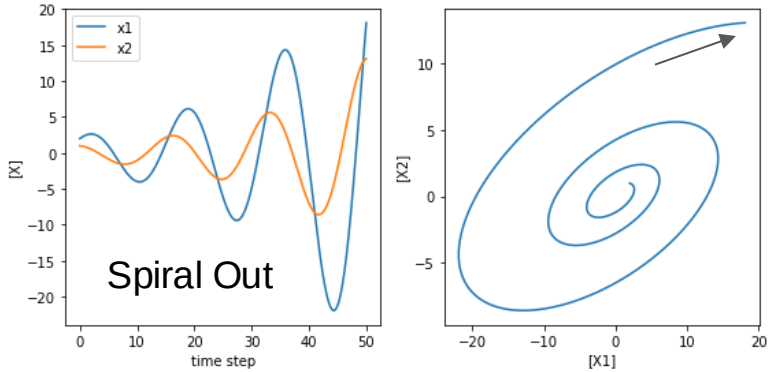
- A simple two-gene system with only linear effects
- Depending on $k_{i,j}$, interesting dynamics can be derived
  - Actually, depending on **Trace** and **Determinant** of the matrix $\{k_{i,j}\}$

Some possible dynamics

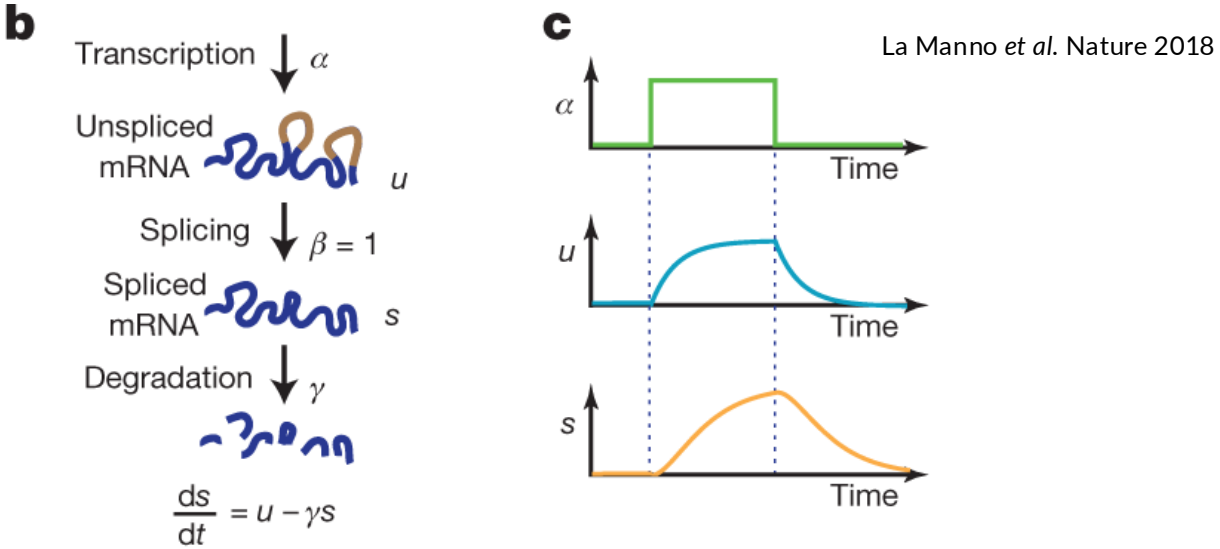








RNA velocity model



La Manno *et al.* Nature 2018

- **Two-step process**: From DNA to unspliced RNA to spliced RNA
- $\frac{d[U]}{dt} = \alpha - \beta[U], \frac{d[S]}{dt} = \beta[U] - \gamma[S]$
- Not the same dynamics as simplifying $\frac{d[S]}{dt} = \alpha - \gamma[S]$

Part 2: Useful online resources

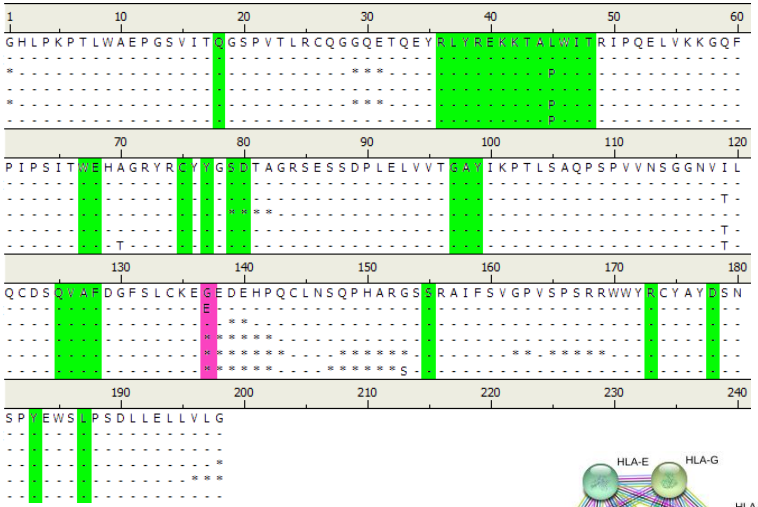
- Use public datasets to develop hypothesis
- New findings can still be discovered in published data
- Be careful of hidden confounding and design factors
- Use multiple tools to confirm

Some motivations

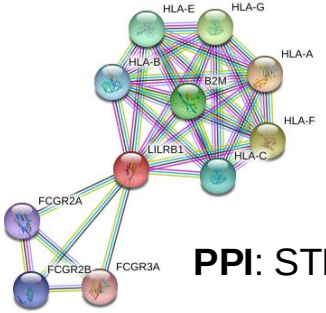
- Quick answers to quick questions
  - Impact of a newly identified mutation?
  - Where does a TF bind on genome?
- Free hypothesis generation
  - What genes are consistently up-regulated in a certain disease?
- Combine data from multiple modalities
  - Omics
  - Clinical
  - Imaging

Image from boston.lti.cs.cmu.edu

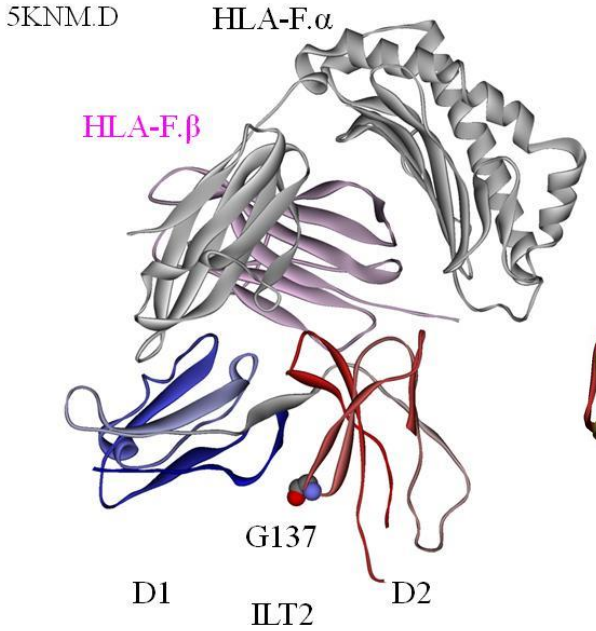
Example: G137E mutation on ILT2



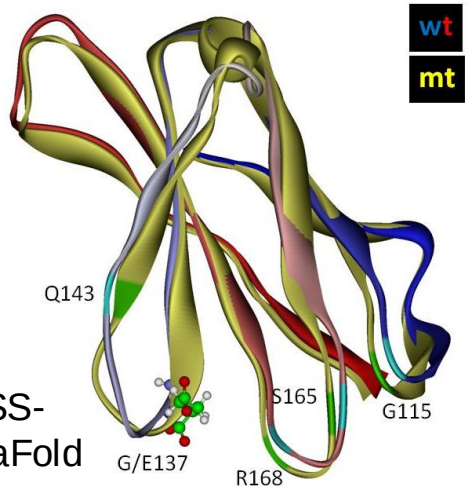
Sequence: Look up variant on ClinVar and alignment to reported sequences



PPI: STRING

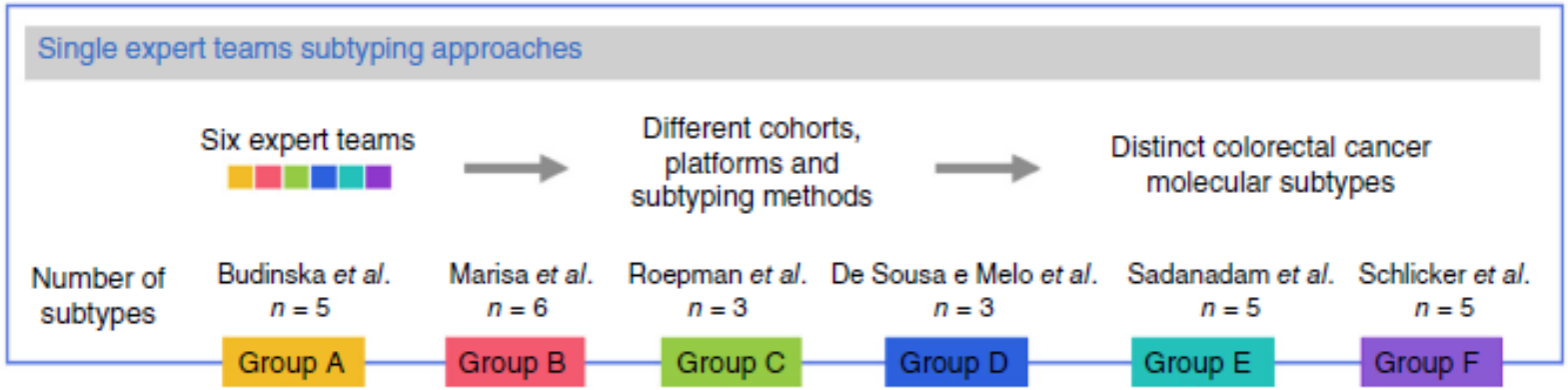


Structure: Highlight mutated residue on structure from PDB.



Modeling: SWISS-MODEL or AlphaFold

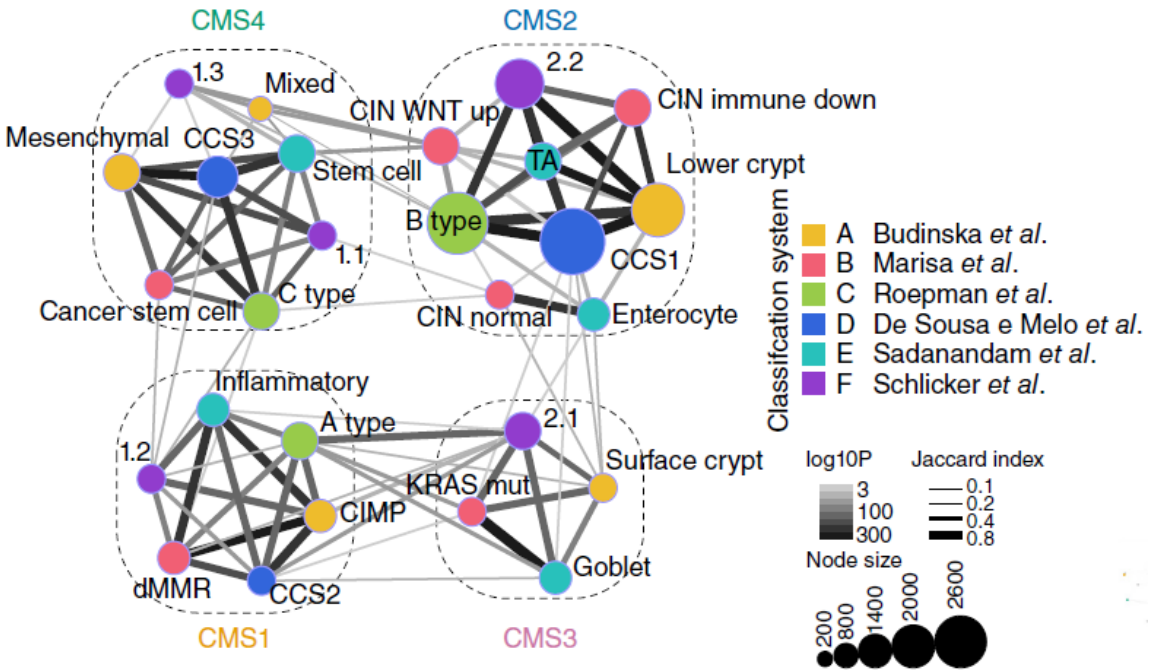
Example: Consensus cancer subtyping



Guinney *et al*. Nat Medicine 21: 1350-1356 (2015)

- Different clinicians characterize cancer patients differently
- But there should be a common molecular basis!

Example: Consensus cancer subtyping

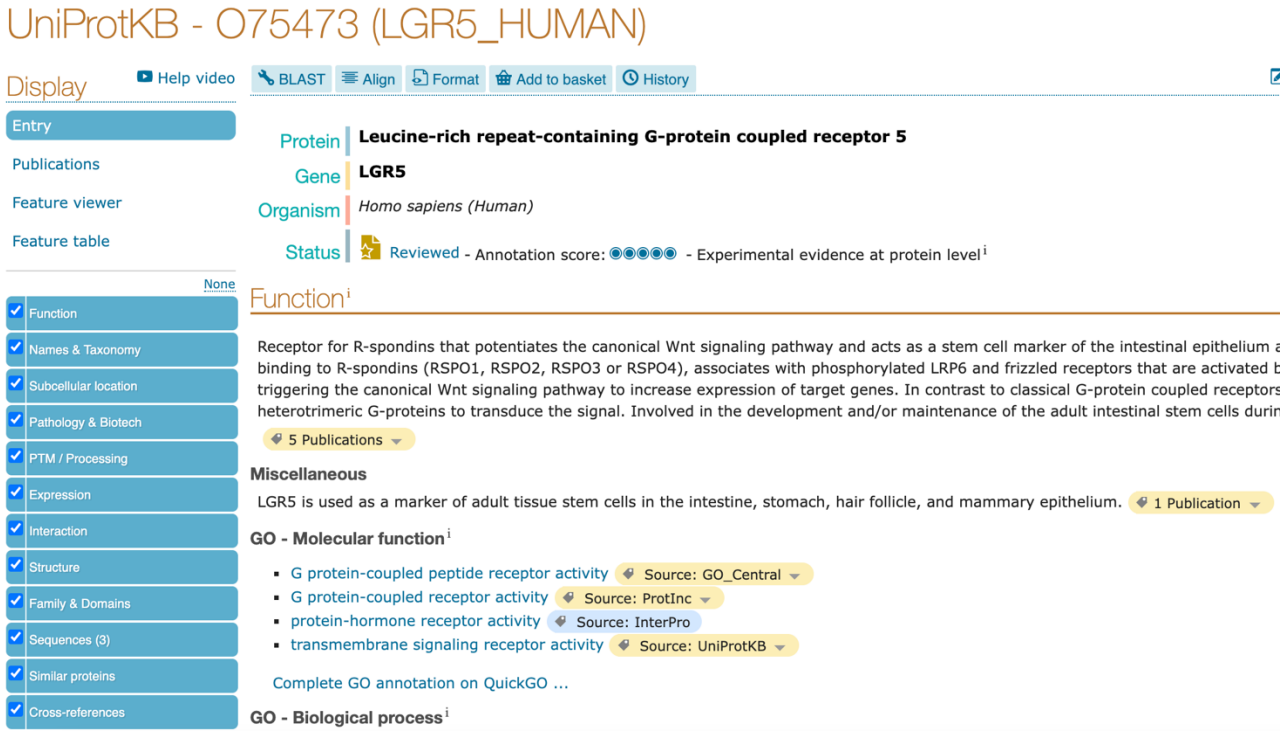


- Compare clinician's decisions on a common group of patients
- Identify 4 major groups: Consensus Molecular Subtype (CMS)

Guinney *et al*. Nat Medicine 21: 1350-1356 (2015)

Knowledgebase

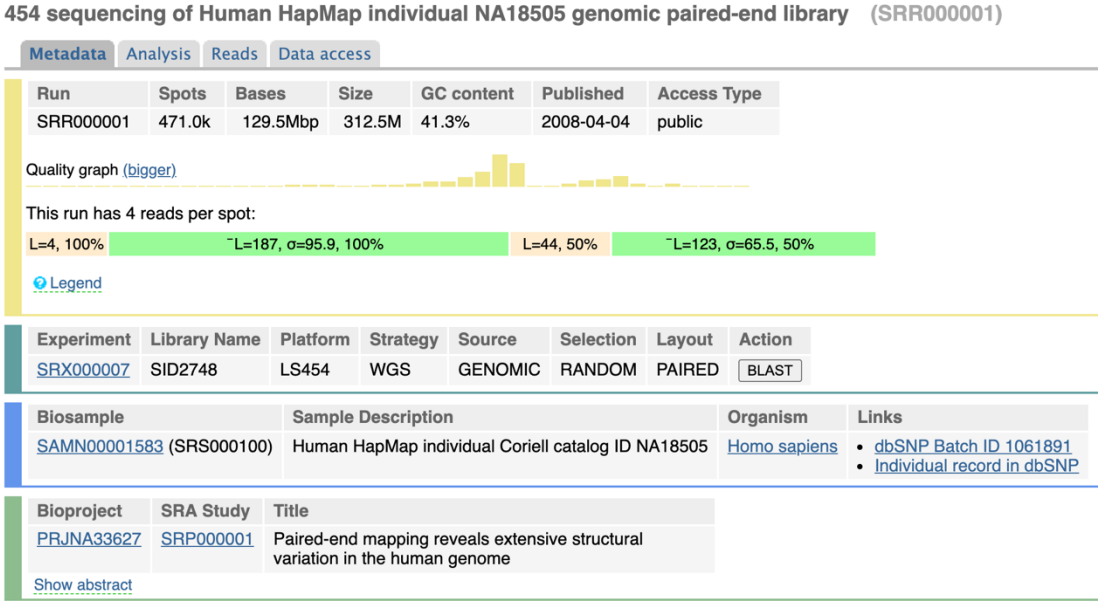
- NCBI/GenBank
- Ensembl
- Uniprot
- GeneCard
- ENCODE
- Human Protein Atlas



Accession IDs

ID system	Accession
Gene Symbol	LGR5
HGNC gene	4504
Entrez (NCBI)	8549
RefSeq transcript	NM_00367
Ensembl	ENSG00000139292
Uniprot protein	O75473
OMIM	606667

Data repository



454 sequencing of Human HapMap individual NA18505 genomic paired-end library (SRR000001)

Metadata | Analysis | Reads | Data access

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR000001	471.0k	129.5Mbp	312.5M	41.3%	2008-04-04	public

Quality graph (bigger)

This run has 4 reads per spot:

L=4, 100% | ¯L=187, σ=95.9, 100% | L=44, 50% | ¯L=123, σ=65.5, 50%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX000007	SID2748	LS454	WGS	GENOMIC	RANDOM	PAIRED	BLAST

Biosample	Sample Description	Organism	Links
SAMN00001583 (SRS000100)	Human HapMap individual Coriell catalog ID NA18505	Homo sapiens	• dbSNP Batch ID 1061891 • Individual record in dbSNP

Bioproject	SRA Study	Title
PRJNA33627	SRP000001	Paired-end mapping reveals extensive structural variation in the human genome

Show abstract

ProteomeXchange ProteomeCentral

<<< Full experiment listing

PXD027487

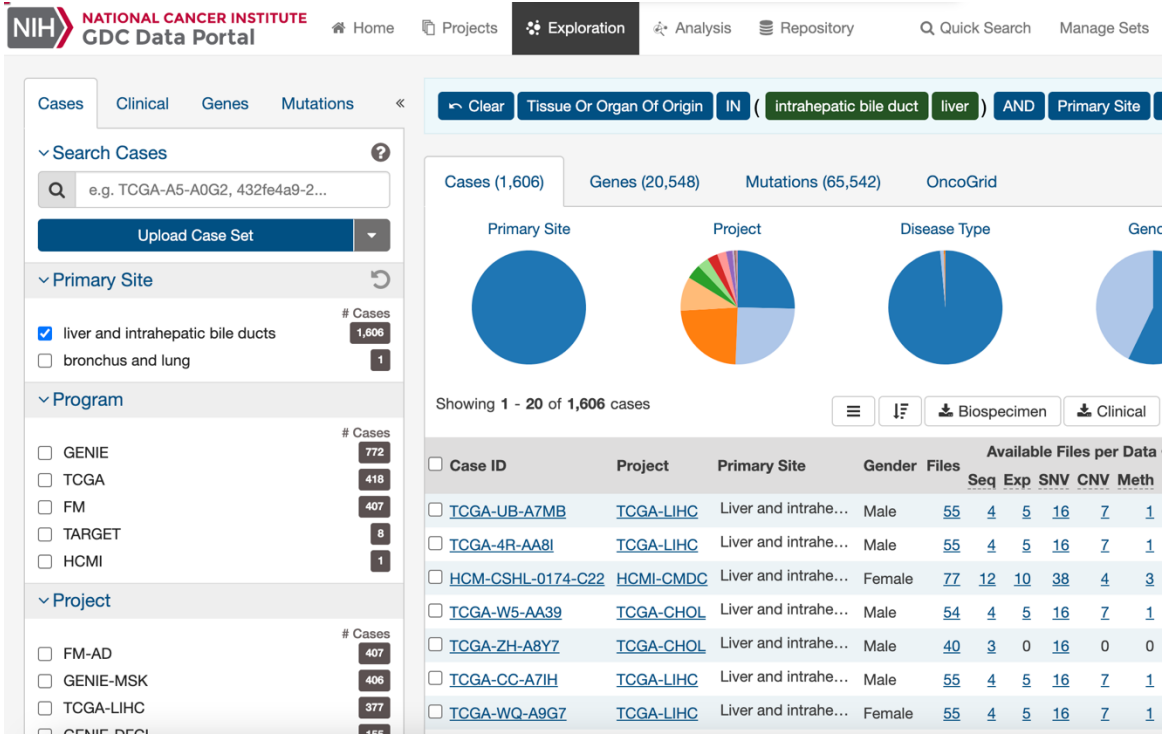
PXD027487 is an **original dataset** announced via ProteomeXchange.

Dataset Summary

Title	Interactome of GIGYF2 and EIF4E2 with and without SMG1i treatement
Description	Translation of messenger RNAs (mRNAs) with premature translation termination codons produces truncated proteins with potentially deleterious effects. This is prevented by nonsense-mediated mRNA decay (NMD) of these mRNAs. NMD is triggered by ribosomes terminating upstream of a splice site marked by an exon-junction complex (EJC), but also acts on many mRNAs lacking a splice junction after their termination codon. We developed a genome-wide CRISPR flow cytometry screen to identify regulators of mRNAs with premature termination codons in K562 cells. This screen recovered essentially all core NMD factors and suggested a role for EJC factors in degradation of PTCs without downstream splicing. Among the strongest hits were the translational repressors GIGYF2 and EIF4E2. GIGYF2 and EIF4E2 mediate translational repression but not mRNA decay of a subset of NMD targets and interact with NMD factors genetically and physically. Our results suggest a model wherein recognition of a stop codon as premature can lead to its translational repression through GIGYF2 and EIF4E2.
HostingRepository	PRIDE
AnnounceDate	2021-10-12
AnnouncementXML	Submission_2021-10-12_00:39:40.790.xml
DigitalObjectIdentifier	
ReviewLevel	Peer-reviewed dataset
DatasetOrigin	Original dataset

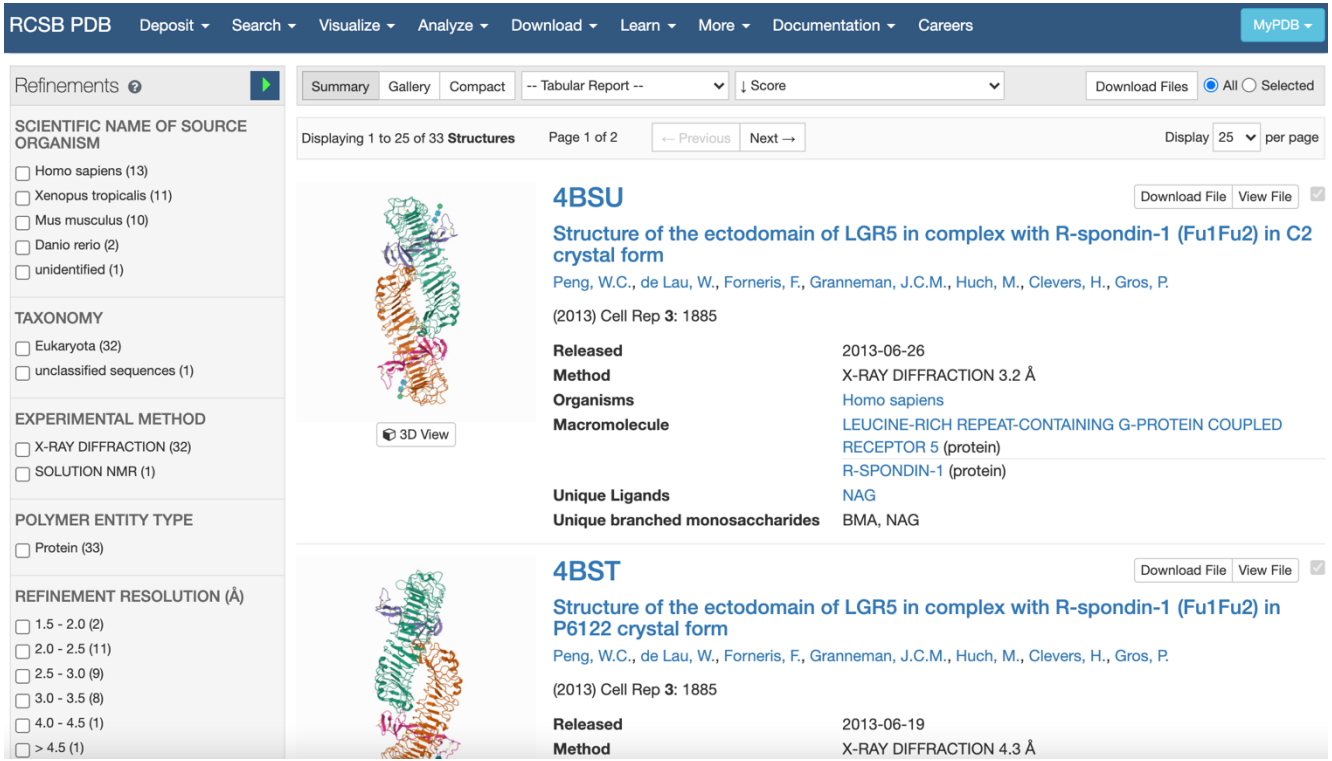
- Reanalysis with a common pipeline

Genomic Data Commons

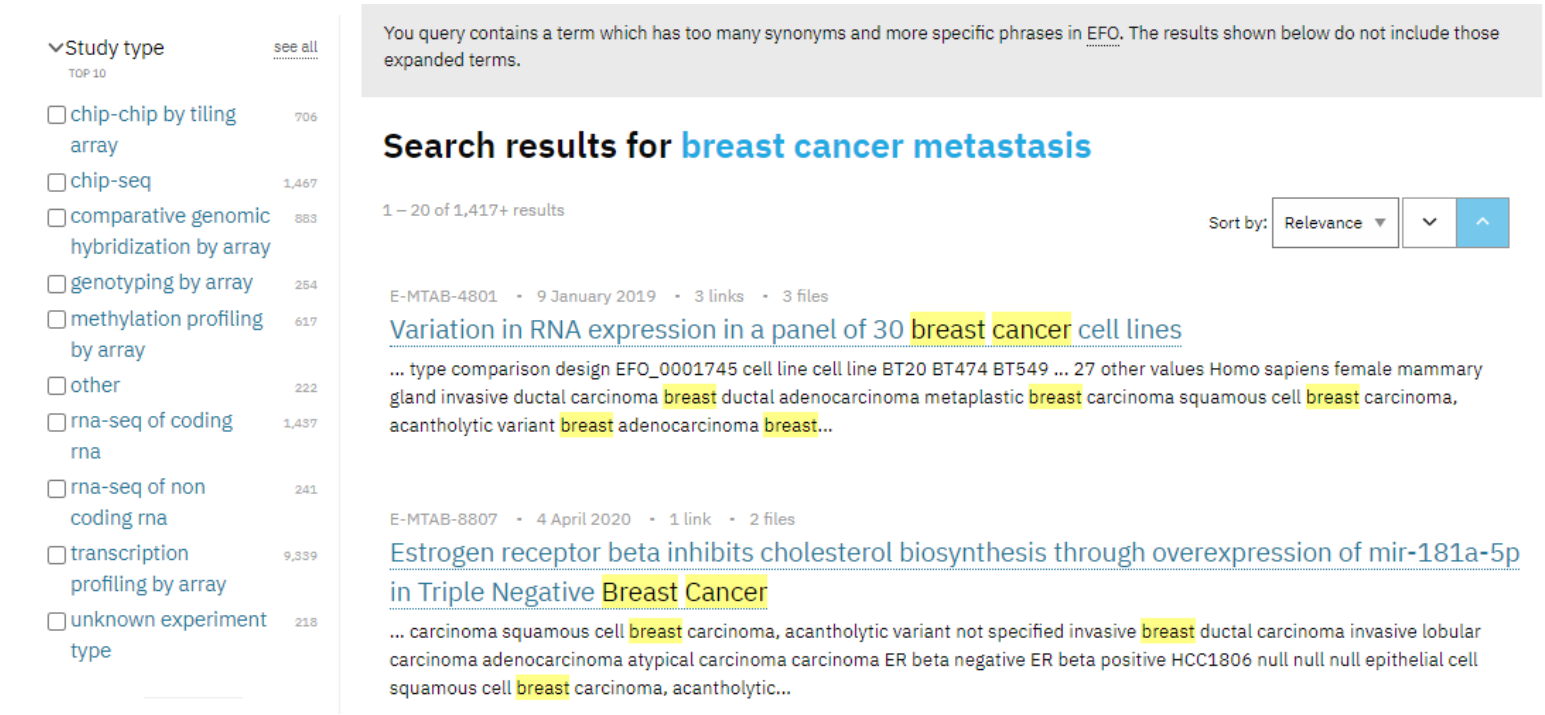


- Cancer multi-omics
  - Exome
  - RNA-seq
  - Methylation
- Clinical and demographic data
- Some with histopathological images and radiographic images

PDB: 3D structures

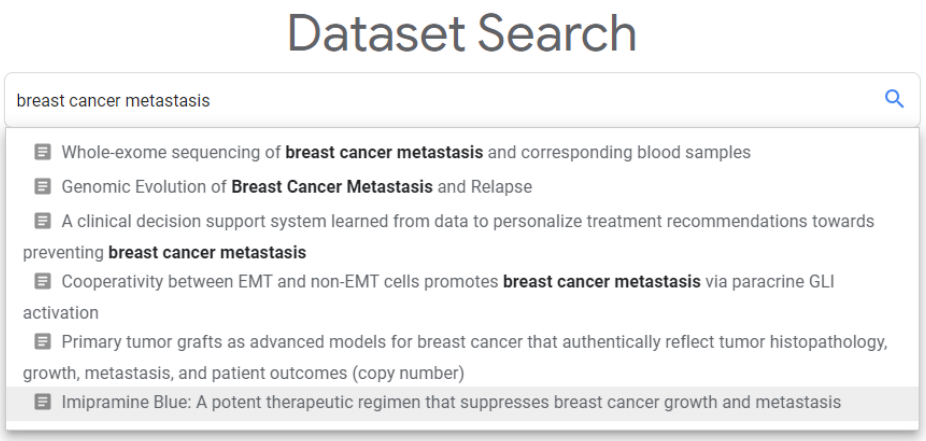


Gene Expression Omnibus / ArrayExpress



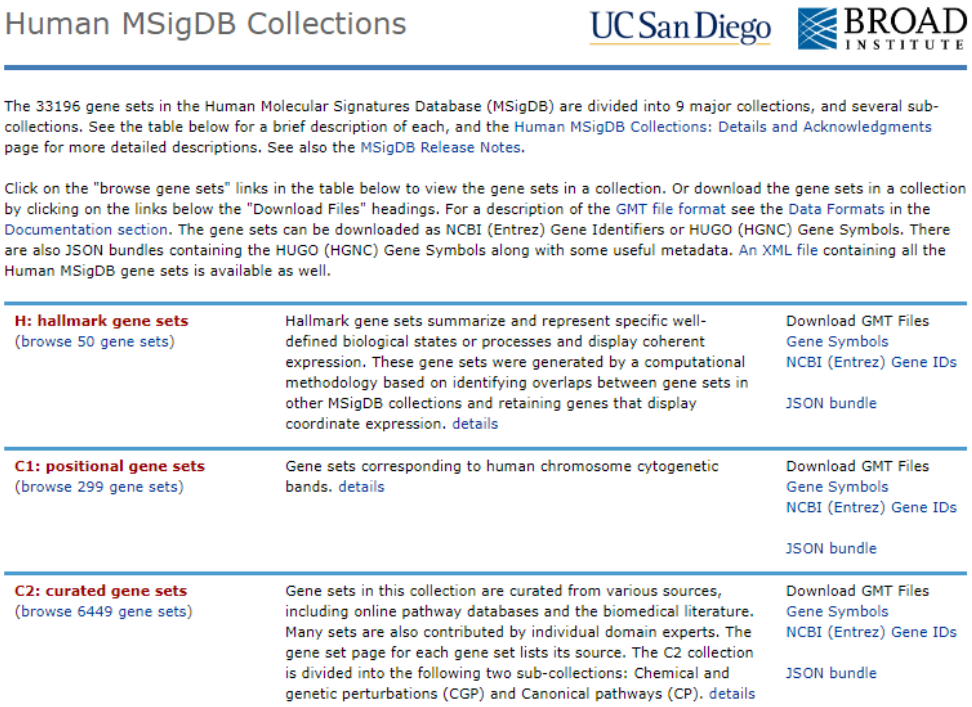
- Search for omics datasets

Google Dataset / FigShare



- Google Dataset compiles links from major public data repositories
- Figshare is a free repository that scientists often use for non-omics data

MSigDB: Curated gene sets



Human MSigDB Collections

UC San Diego BROAD INSTITUTE

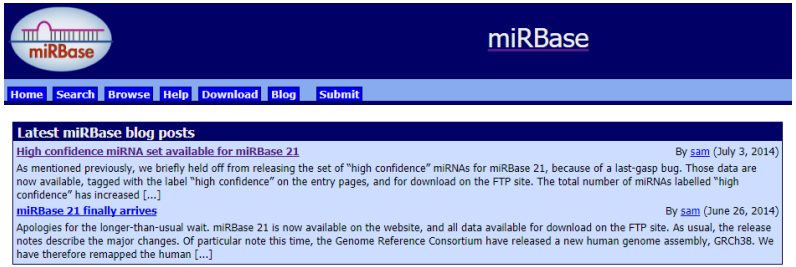
The 33196 gene sets in the Human Molecular Signatures Database (MSigDB) are divided into 9 major collections, and several sub-collections. See the table below for a brief description of each, and the Human MSigDB Collections: Details and Acknowledgments page for more detailed descriptions. See also the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

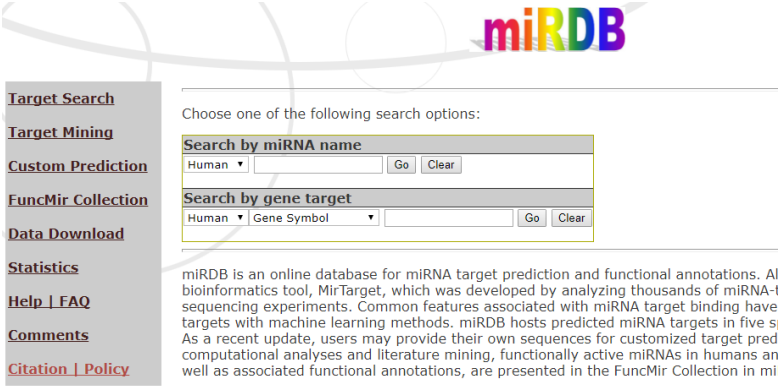
Click on the "browse gene sets" links in the table below to view the gene sets in a collection. Or download the gene sets in a collection by clicking on the links below the "Download Files" headings. For a description of the GMT file format see the Data Formats in the Documentation section. The gene 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

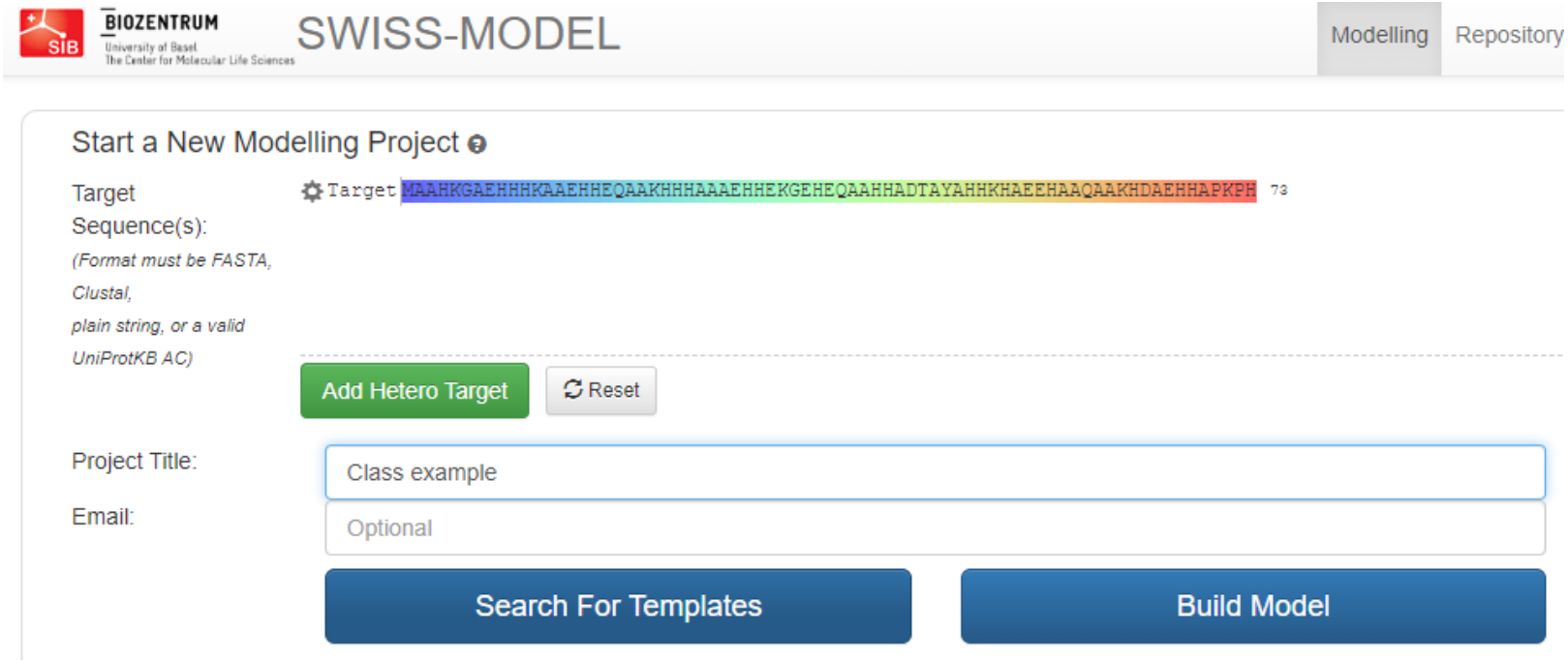
- Function-based / Disease-based / Publication-based / Genomic location-based
- For gene panel, enrichment test, etc.

miRNA target

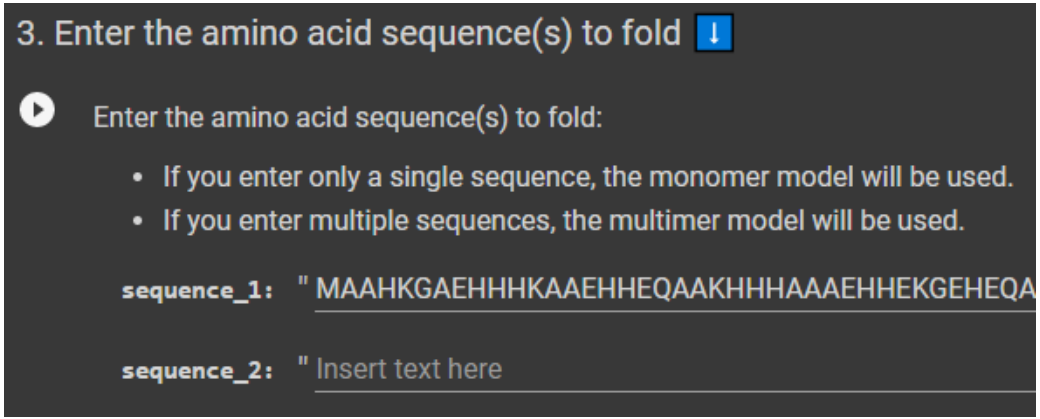




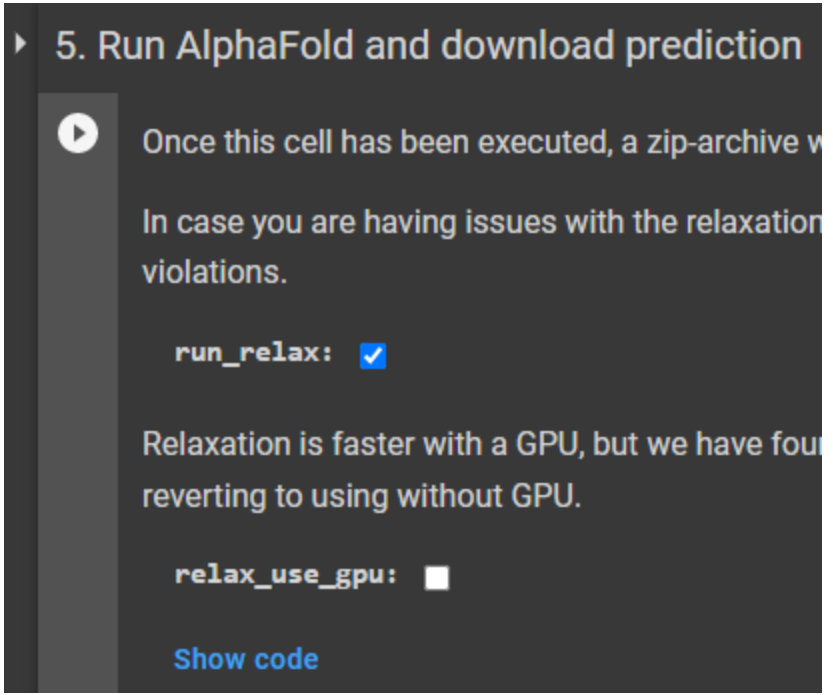
SWISS-MODEL: Homology modeling



AI-assisted protein folding (ColabFold)



https://github.com/sokrypton/ColabFold



- Accessible, simplified Python interface

GenePattern / Galaxy

Features

Powerful genomics tools in a user-friendly interface

GenePattern provides hundreds of analytical tools for the analysis of gene expression (RNA-seq and microarray), sequence variation and copy number, proteomic, flow cytometry, and network analysis. These tools are all available through a Web interface with no programming experience required.

GenePattern Notebook

The GenePattern Notebook environment extends the Jupyter Notebook system, allowing researchers to create documents that interleave formatted text, graphics and other multimedia, executable code, and GenePattern analyses, creating a single "research narrative" that puts scientific discussion and analyses in the same place.

Blog > GP updates

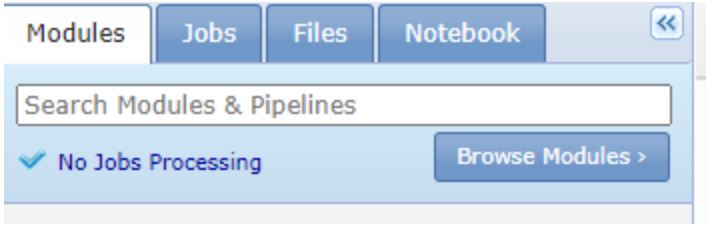
- GenePattern Coverage and Support December 23, 2021 - January 2, 2022
- End of Support for GParc
- End of support for modules using the deprecated GenePattern patch mechanism.
- End of support for Cufflinks suite of modules
- End of support for MAGeCK

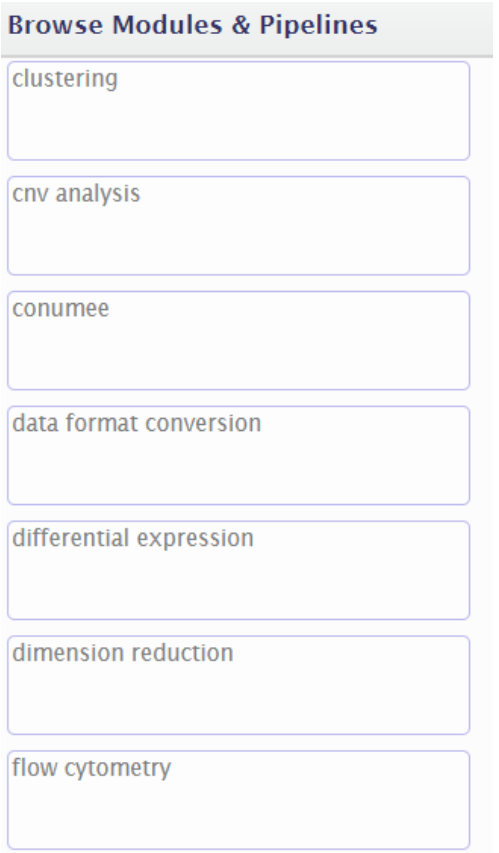
view more >

Papers > Related Publications

- Comparability and reproducibility of biomedical data
- ToppCluster: a multiple gene list feature analyzer for

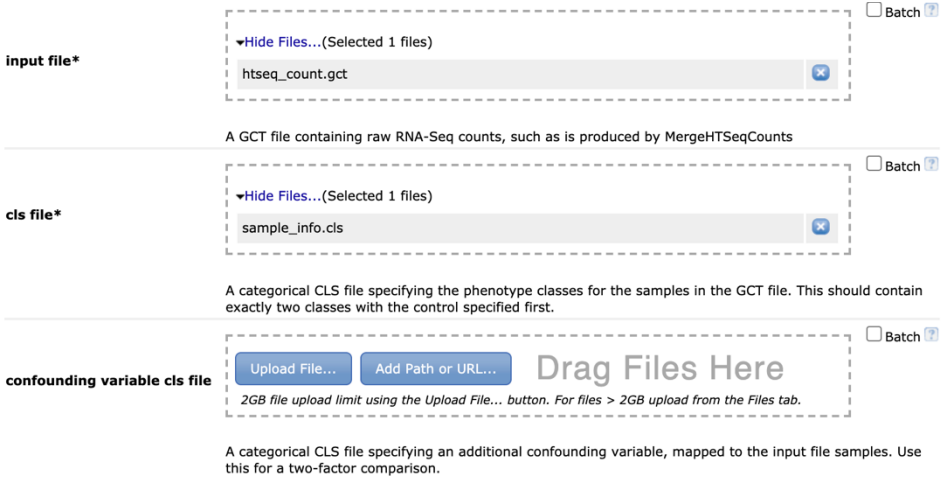
Content





- Modules = standard bioinformatics tools
- Notebook = Python environment (similar to Google Colab) with specific bioinformatics library

DESeq2 on GenePattern



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

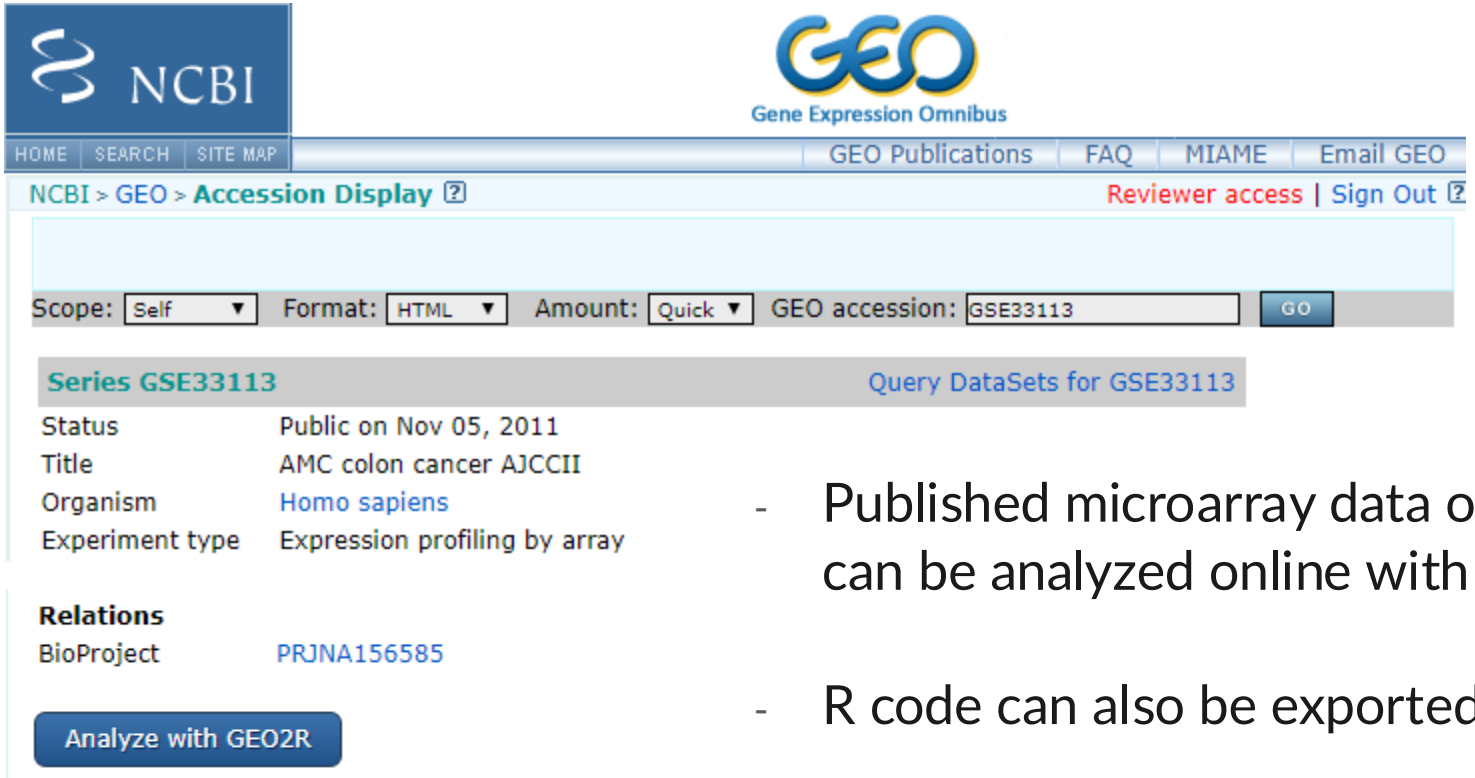
#1.2

60488	12				
Name	Description	sample1	sample2	sample3	sample4
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
```

DESeq2 output

GEO2R: Differential expression



- Published microarray data on GEO can be analyzed online with GEO2R
- R code can also be exported

Demo time!

- Let’s try some online tools:
  - ColabFold interface for AlphaFold2
  - STRING for viewing protein-protein interaction
  - GEO2R for differential expression analysis of microarray data

Any question?