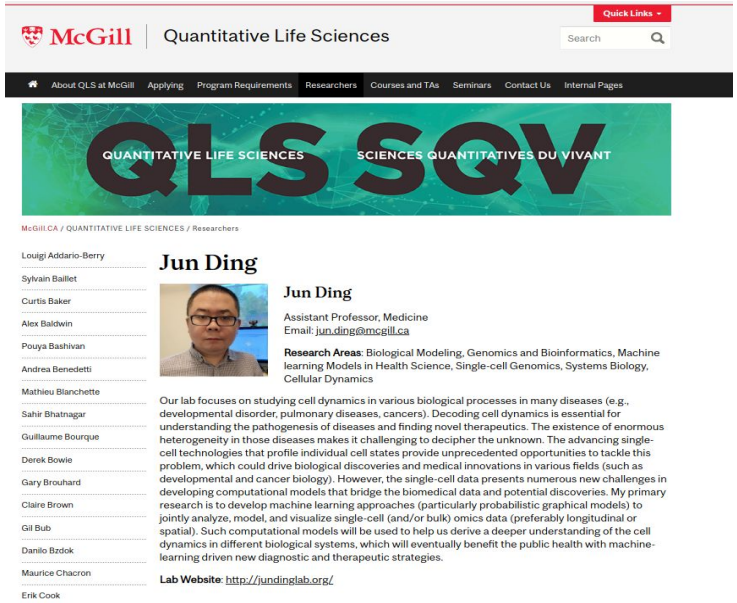
Big Biomedical Data analysis

Jun Ding, Ph.D
Assistant Professor

Department of medicine
School of Computer Science
Department of Biomedical engineering
Department of Human Genetics
McGill University
MILA-Quebec AI Institute
2024/07/08

My research



Me:
Assistant professor, Department of Medicine
QLS member
School of Computer Science
Associate member of Biomedical Engineering
McGill University
FRQS Junior 1 Researcher
MILA-Quebec AI Institute

My research:
1) Machine learning in health science
   probabilistic graphical models
   supervised/unsupervised deep neural nets
2) Single-cell Cellular dynamics
   Single-cell Transcriptomics
   Single-cell multi-omics
   Data visualizations

You?

~1 min quick introduction (In English/Chinese)

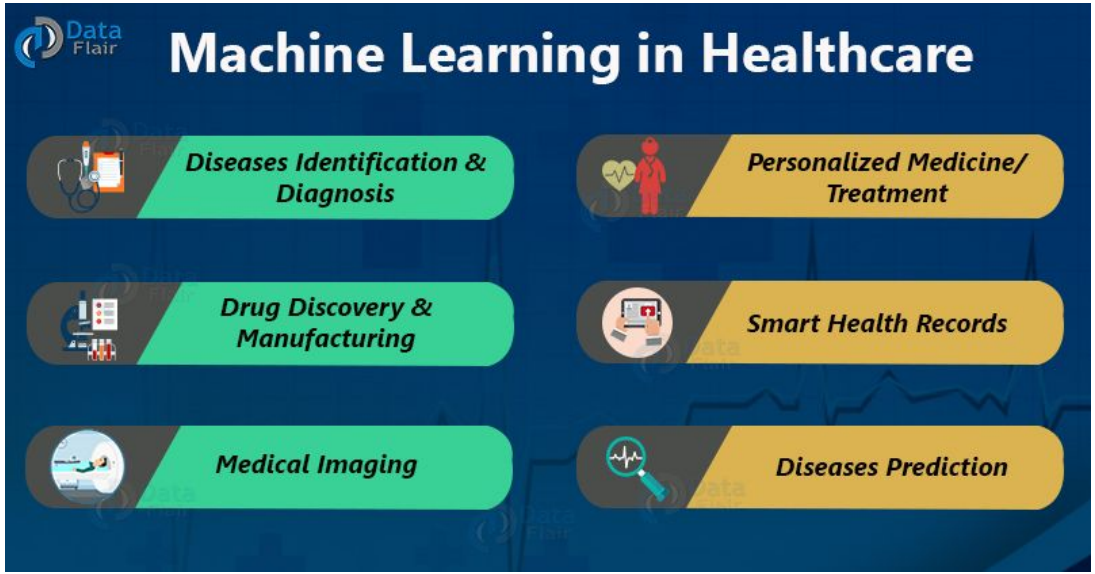
1) Who am I?
2) My major?
3) What do I want to get out from this course?

What this course is about?

Bioinformatics

Computational Biology

AI+health (machine learning in health science)



All images are from network

Discovery of cell

MICROGRAPHIA:
OR SOME
Physiological Descriptions
OF
MINUTE BODIES
MADE BY
MAGNIFYING GLASSES.
WITH
OBSERVATIONS and INQUIRIES thereupon.

By *R. HOOKE*, Fellow of the ROYAL SOCIETY.

Non possit oculo quantum contendere Linceus,
Non tamen idcirco contemnas Lippus inunctis. Horat. Ep. lib. 1.

LONDON, Printed by *Jo. Martyn*, and *Ja. Allestry*, Printers to the ROYAL SOCIETY, and are to be sold at their Shop at the *Bell* in S. *Paul's* Church-yard. MDCLXV.

Robert Hooke, 1665

Image from *Encyclopædia Britannica, Inc.*

Discovery of gene

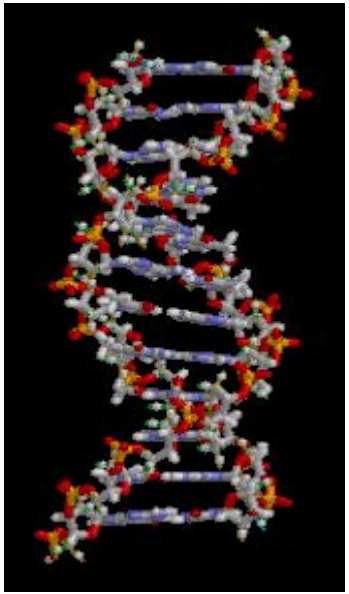
Gregor Mendel

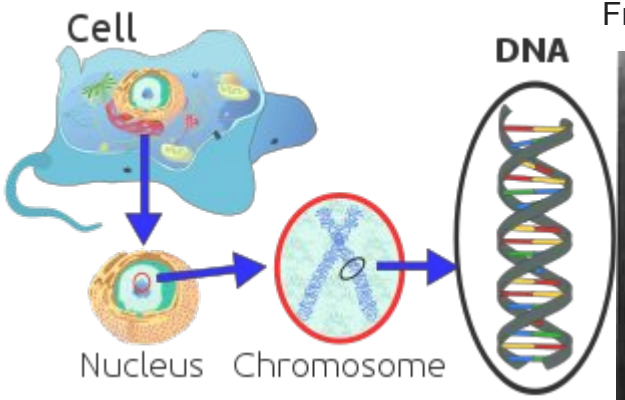
- **Gregor Mendel** was a monk in mid 1800's who discovered how genes were passed on.
- He used peas to determine the pattern of heredity

Gregor Mendel

The image was download from slideplayer.com

Discovery of DNA (and its double helix structure)



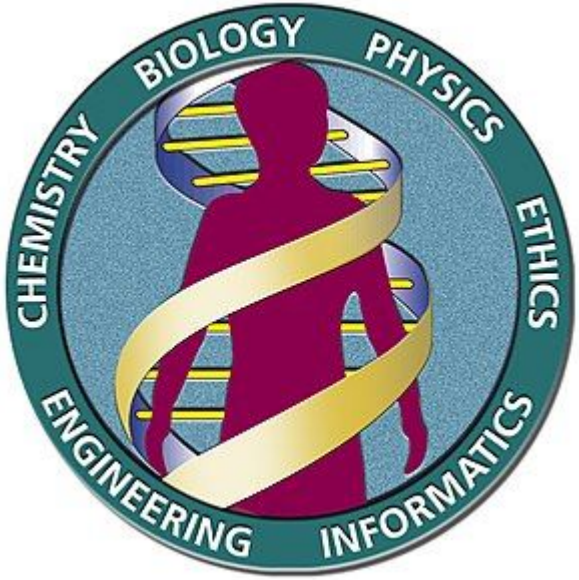


Swiss physician and biologist Friedrich Miescher

James Watson and Francis Crick

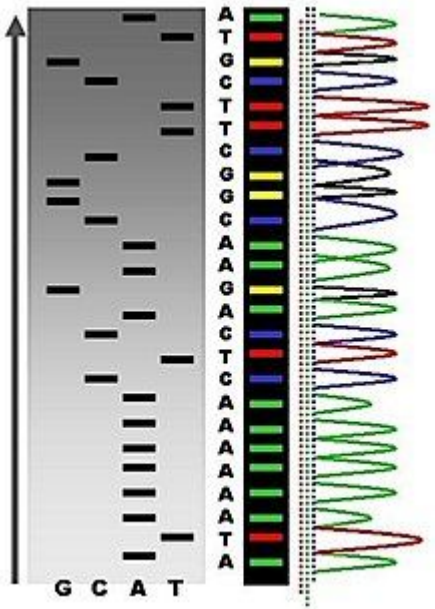
Images are from the DNA wiki page (https://en.wikipedia.org/wiki/DNA)

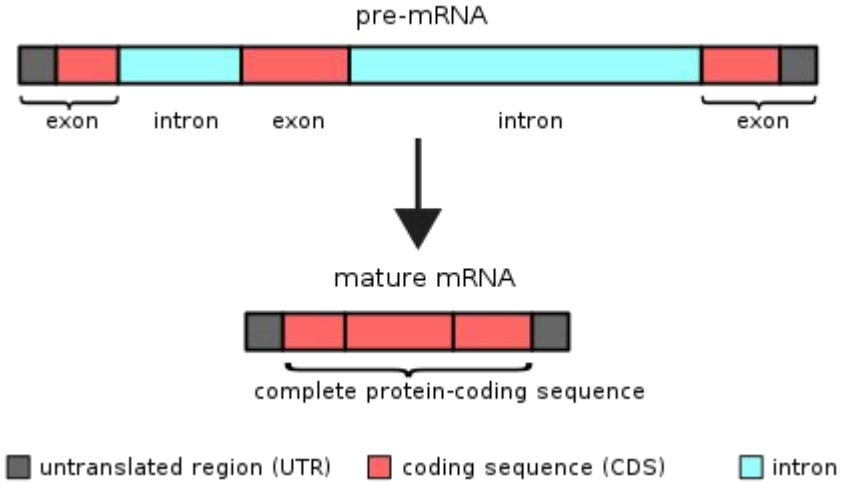
Human genome project



No.	Nation	Name	Affiliation
1		The Whitehead Institute/MIT Center for Genome Research	Massachusetts Institute of Technology
2		The Wellcome Trust Sanger Institute	Wellcome Trust
3		Washington University School of Medicine Genome Sequencing Center	Washington University in St. Louis
4		United States DOE Joint Genome Institute	United States Department of Energy
5		Baylor College of Medicine Human Genome Sequencing Center	Baylor College of Medicine
6		RIKEN Genomic Sciences Center	Riken
7		Genoscope and CNRS UMR-8030	French Alternative Energies and Atomic Energy Commission
8		GTC Sequencing Center	Genome Therapeutics Corporation, whose sequencing division is acquired by ABI
9		Department of Genome Analysis	Fritz Lipmann Institute, name changed from Institute of Molecular Biotechnology
10		Beijing Genomics Institute/Human Genome Center	Chinese Academy of Sciences
11		Multimegabase Sequencing Center	Institute for Systems Biology
12		Stanford Genome Technology Center	Stanford University
13		Stanford Human Genome Center and Department of Genetics	Stanford University School of Medicine
14		University of Washington Genome Center	University of Washington
15		Department of Molecular Biology	Keio University School of Medicine
16		University of Texas Southwestern Medical Center at Dallas	University of Texas
17		University of Oklahoma's Advanced Center for Genome Technology	Dept. of Chemistry and Biochemistry, University of Oklahoma
18		Max Planck Institute for Molecular Genetics	Max Planck Society
19		Lita Annenberg Hazen Genome Center	Cold Spring Harbor Laboratory
20		GBF/German Research Centre for Biotechnology	Reorganized and renamed to Helmholtz Center for Infection Research

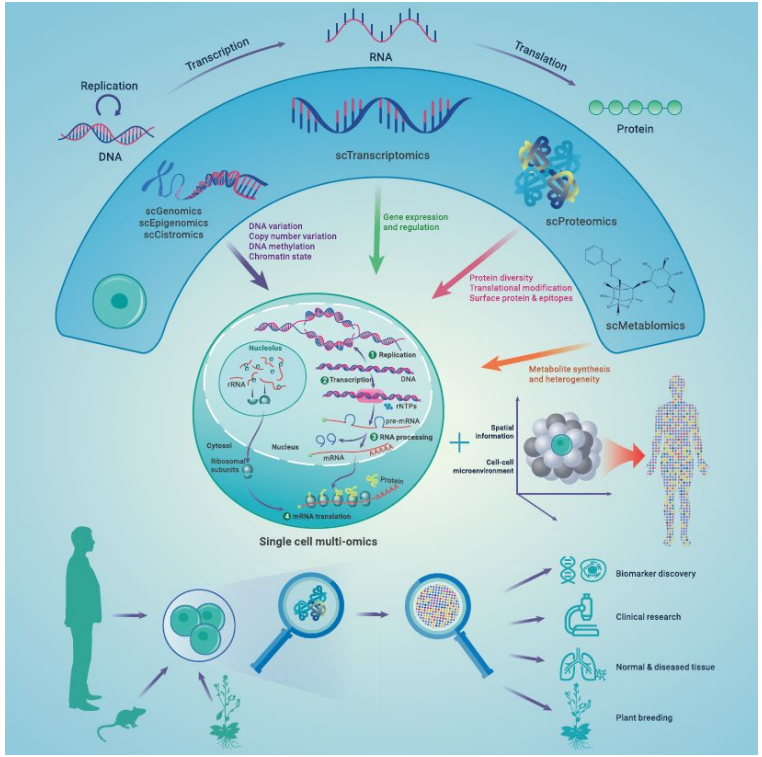
What does DNA look like?





Images are from wiki pages (https://en.wikipedia.org/wiki/DNA_sequencing)

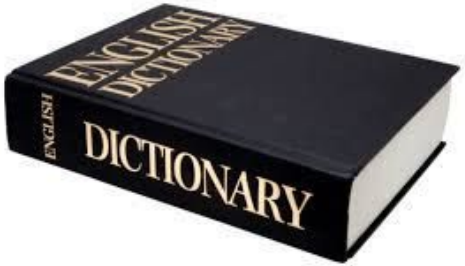
How to represent a cell (e.g., what we often call genomics)

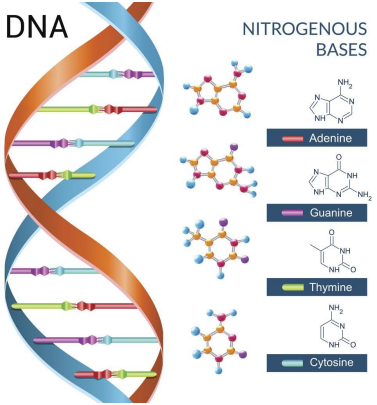


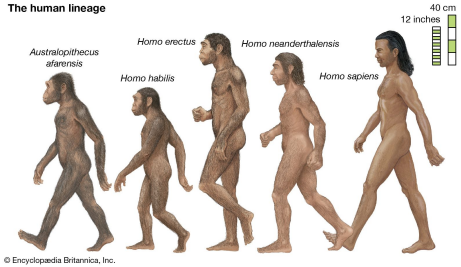
https://www.researchgate.net/publication/364396236_Single_cell_technologies_From_research_to_application/figures?lo=1

DNA data

DNA data (genome)







All images are from network

Why we need to study the DNA?

1) Better understand our human species (--> evolution)

It can answer who we are?

“Who” created us?

2) Better understand the mechanisms for many diseases

https://www.medicinenet.com/genetic_disease/article.htm

1. cystic fibrosis,
2. alpha- and beta-thalassemias,
3. sickle cell anemia (sickle cell disease),
4. Marfan syndrome,
5. fragile X syndrome,
6. Huntington's disease, and
7. hemochromatosis.

3) ** how to better “engineer” our genome**

(legal and ethical issues, might change in the near future)

RNA data

RNA is a big family

mRNAs: messenger RNA

miRNAs: microRNA

lncRNAs: long non-coding RNAs

…

mRNAs

How many?

Not sure (around 20k～30k).

What do they do?

microRNAs

How many?

~2k

What do they do?

lncRNAs

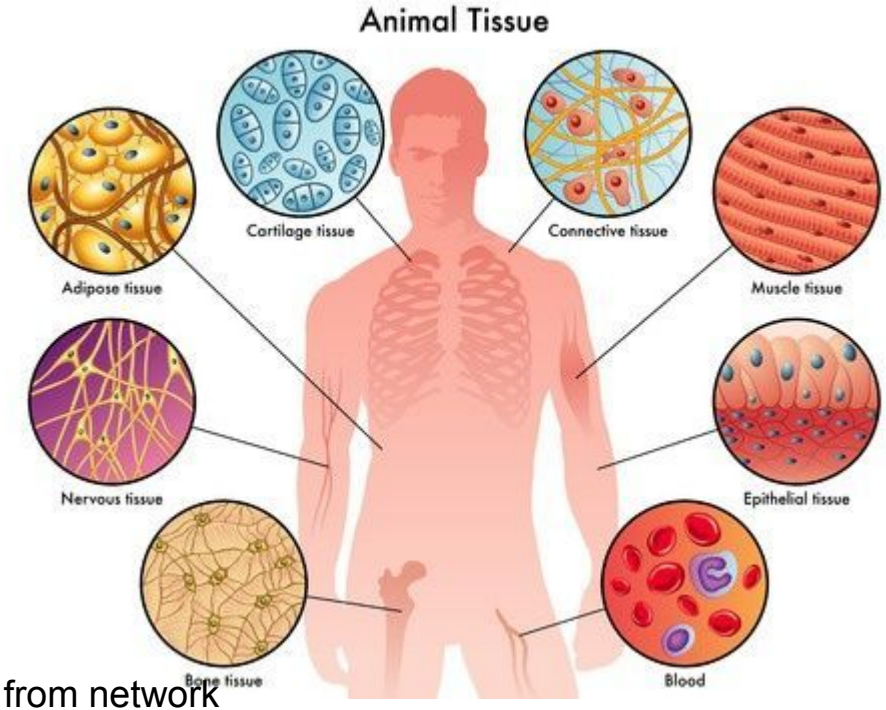
How many?

30k~60k

What do they do?

-> interact with DNA, other mRNAs, proteins.

Why we want/need to study RNAs



All images are from network

Protein data

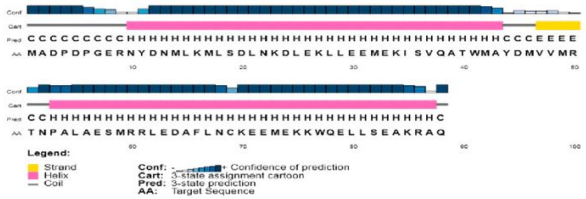
Protein sequence

1st letter	Second Letter: U		Second Letter: C		Second Letter: A		Second Letter: G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

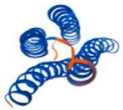
a

```
  1 ATGGCTGATCCTGACCCTGGGGAAAGAAACTATGACAACATGCTG
     M  A  D  P  D  P  G  E  R  N  Y  D  N  M  L

 46 AAAATGCTGTCAGACCTGAATAAAGACTTGGAAAAGCTGTTGGAA
     K  M  L  S  D  L  N  K  D  L  E  K  L  L  E

 91 GAGATGGAAAAAATCTCAGTGCAGGCCACGTGGATGGCCTACGAC
     E  M  E  K  I  S  V  Q  A  T  W  M  A  Y  D

136 ATGGTGGTGATGCGCACCAACCCCGCGCTGGCGGAGTCCATGCGG
     M  V  V  M  R  T  N  P  A  L  A  E  S  M  R

181 CGGCTGGAGGACGCCTTCCTCAACTGCAAGGAGGAGATGGAAAAG
     R  L  E  D  A  F  L  N  C  K  E  E  M  E  K

226 AAGTGGCAGGAGCTGCTCAGTGAGGCCAAGCGCGCGCAGTAG
     K  W  Q  E  L  L  S  E  A  K  R  A  Q  *
```

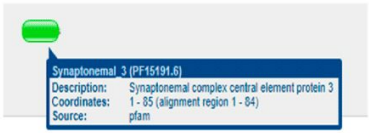
b



c



d



All images are from network

1) How many proteins

80k-400k proteins in human

2) What do they do?

Proteins are large, complex molecules that play many critical roles in the body. They do most of the work in cells and are required for the structure, function, and regulation of the body’s tissues and organs.

Other data

Other biomolecules

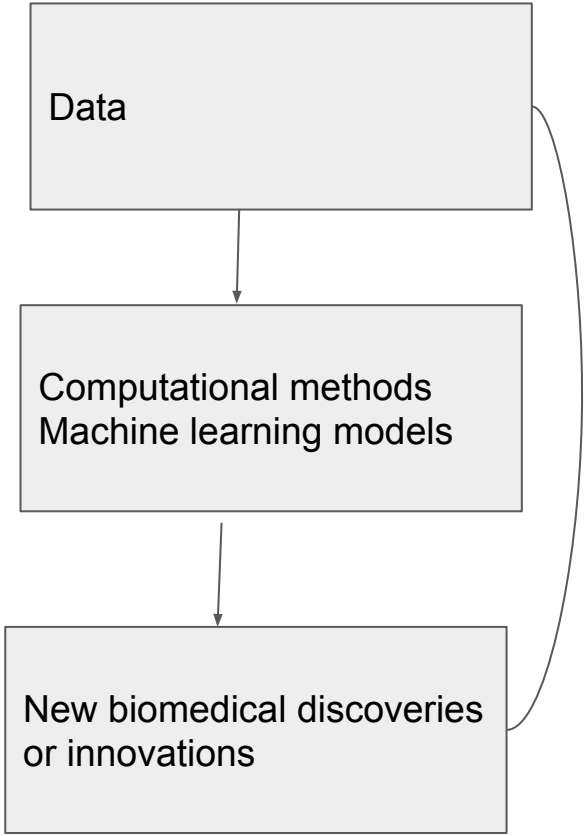
The four major types of biomolecules are carbohydrates, lipids, nucleic acids, and proteins.

Bioimaging data

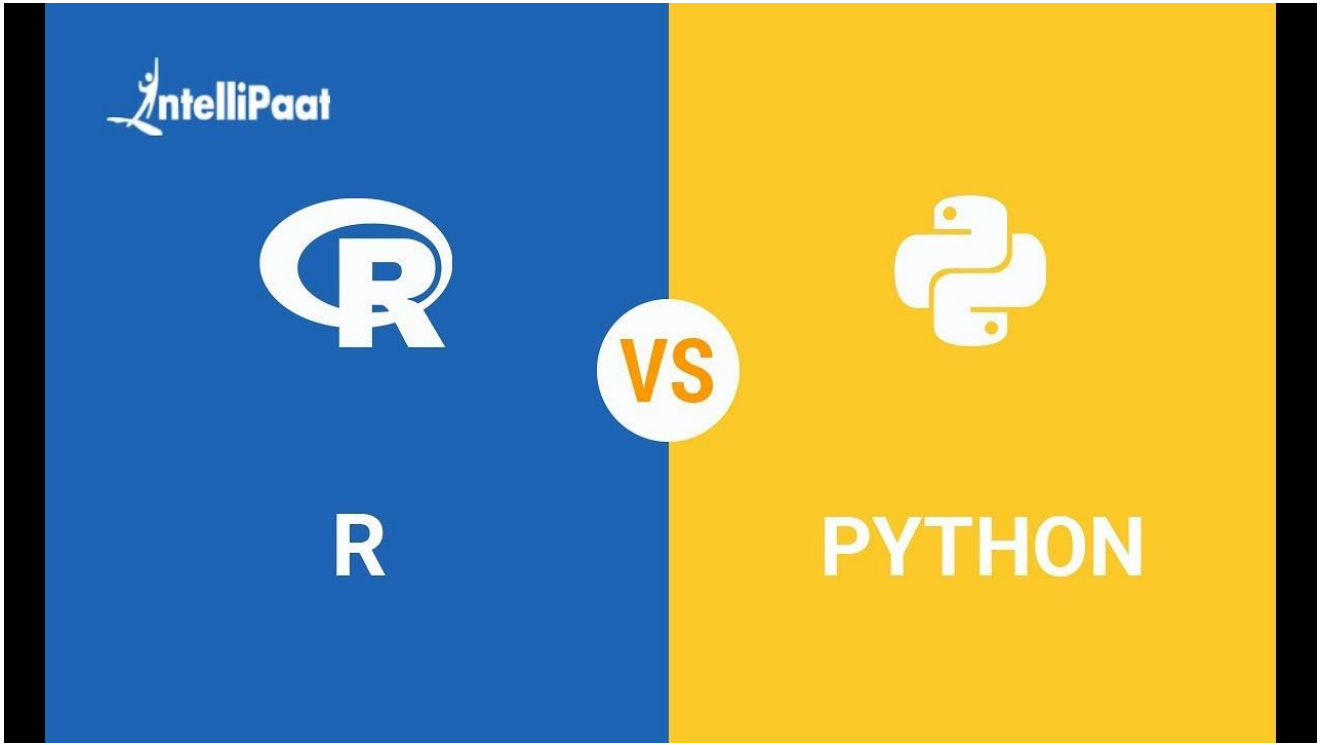
X-ray
CT-Scan
...

Clinical health record

Disease severity
Disease diagnosis
Disease progression
Disease treatment
Disease prognosis


Data
Computational methods
Machine learning models
New biomedical discoveries
or innovations

Biomedical Data analysis platform


IntelliPaat
VS
R
PYTHON

Install your python

Anaconda platform

https://www.anaconda.com/download

1) download the versions for your computer (mac or windows)

2) install anaconda

3)

Write your first bioinformatics program

Write an “encoder” function to encode your name into DNA sequence

Write a “decoder” function to decode the DNA sequence into your name

Encoding rule:

https://www.illumina.com/content/dam/illumina-marketing/documents/landing/stem/Translate%20Your%20Name%20Into%20DNA%20Code.pdf

Translate Your Name Into DNA Code

Write your name in the space below and use the table to translate it into a DNA sequence.

Our Alphabet	Amino Acid Name	Simplified Codon
A	Alanine	**GCT**
B		**GCA** (Alanine)
C	Cysteine	**TGC**
D	Aspartic acid	**GAT**
E	Glutamic acid	**GAG**
F	Phenylalanine	**TTT**
G	Glycine	**GGG**
H	Histidine	**CAT**
I	Isoleucine	**ATA**
J		**ATC** (Isoleucine)
K	Lysine	**AAG**
L	Leucine	**CTC**
M	Methionine	**ATG**
N	Asparagine	**GAC**
O		**GAT** (Asparagine)
P	Proline	**CCC**
Q	Glutamine	**GAG**
R	Arginine	**CGT**
S	Serine	**TCA**
T	Threonine	**ACT**
U		**ACG** (Threonine)
V	Valine	**GTC**
W	Tryptophan	**TGG**
X		**GTA** (Valine)
Y	Tyrosine	**TAC**
Z		**TAT** (Tyrosine)

Find one in the other team to evaluate your encoding

Who is your mate in the other team? Base-pairing of your initials

Example:

My initial: JD

=>**A**TC,**G**AT =>T,C (if no exact match, find the closest one)

Send him/her your encoded DNA sequence and your name for evaluation.

If it’s the decoding and the given name are not consistent, figure out who is wrong (winner gets 1 score, loser gets 0)

S(A)=

S(B)=

At the end of the course, we will test whether S(A)/S(B) is significantly bigger than S(B)/S(A)?

All in the winner group will have a bonus of 5 scores

DNA data analysis

1) Download a DNA sequence
2) Sequence alignment
3) Find a DNA motif
4) Identify a DNA mutation

How to download DNA sequences?

1) Ensembl

https://useast.ensembl.org/index.html

2) NCBI

3) UCSC genome browser

Sequence alignment

Example:	*Sequences*		*Alignment*
	ACCCGA		ACCCGA
	ACTA	$\Rightarrow_{\text{align}}$	AC--TA
	TCCTA		TCC-TA

images/slides partially come from course (https://math.mit.edu/classes/18.417)

Edit distance

An *edit operation* is a pair $(x, y) \in (\Sigma \cup \{-\} \neq (-, -)$. We call (x,y)

- *substitution* iff $x \neq -$ and $y \neq -$
- *deletion* iff $y = -$
- *insertion* iff $x = -$

For sequences a, b, write $a \rightarrow_{(x,y)} b$, iff a is transformed to b by operation (x, y). Furthermore, write $a \Rightarrow_S b$, iff a is transformed to b by a sequence of edit operations S.

Example

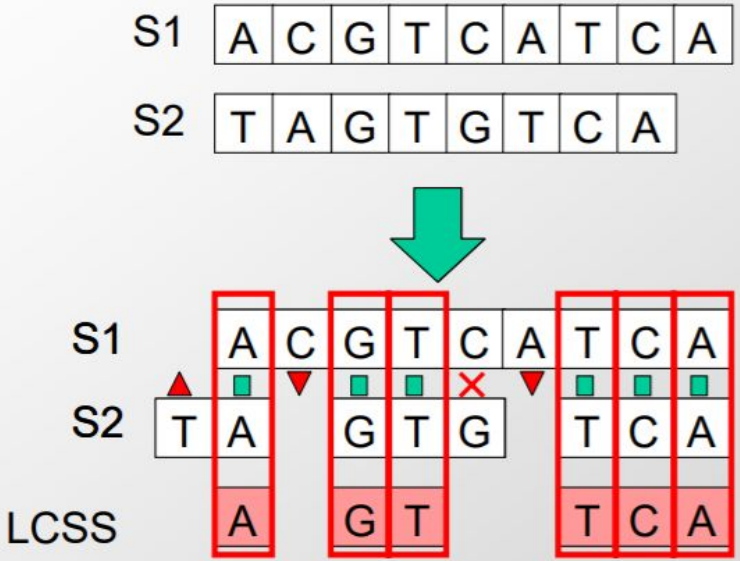
$\text{ACCCGA} \rightarrow_{(C,-)} \text{ACCGA} \rightarrow_{(G,T)} \text{ACCTA} \rightarrow_{(-,T)} \text{ATCCTA}$

$\text{ACCCGA} \Rightarrow_{(C,-),(G,T),(-,T)} \text{ATCCTA}$

images/slides partially come from course (https://math.mit.edu/classes/18.417)

Comparing two DNA sequences

- Given two possibly related strings S1 and S2
  - What is the longest common subsequence?



Edit distance:

•Number of changes needed for S1→S2

images/slides partially come from course (https://math.mit.edu/classes/18.417)
https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-096-algorithms-for-computational-biology-spring-2005/lecture-notes/lecture5_newest.pdf

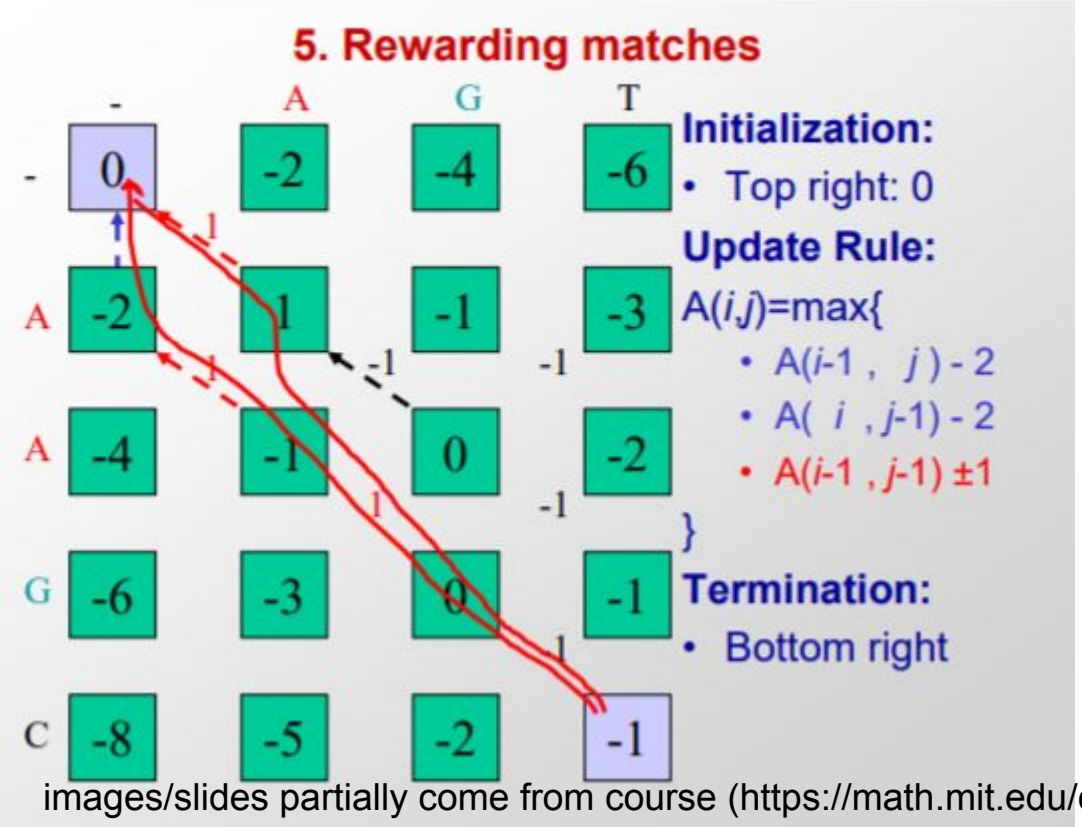
How can we compute best alignment

S1	A	C	G	T	C	A	T	C	A
S2	T	A	G	T	G	T	C	A	

- Need scoring function:
  - Score(alignment) = Total cost of editing S1 into S2
  - Cost of mutation
  - Cost of insertion / deletion
  - Reward of match
- Need algorithm for inferring best alignment
  - Enumeration?
  - How would you do it?
  - How many alignments are there?

images/slides partially come from course (https://math.mit.edu/classes/18.417)
https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-096-algorithms-for-computational-biology-spring-2005/lecture-notes/lecture5_newest.pdf

Dynamic programming



images/slides partially come from course (https://math.mit.edu/classes/18.417)
https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-096-algorithms-for-computational-biology-spring-2005/lecture-notes/lecture5_newest.pdf

Example

Align the DNA sequence encoded from your initials with someone else in your opposite team

Compare the best alignment that you got, find out whether they are consistent.

If not, figure out which answer is correct

Record the score

S(A)

S(B)

Why we want to align sequences?

Essentially, sequence alignment is used to find the distance between “sequences”

A lot of applications:

1) Assembly the genome
2) Quantify gene expression

3) Study the conservation between species

4) Understand the evolution

Python Basics

Python basics

Basic syntax

https://www.learnpython.org/

Python file inputs/outputs

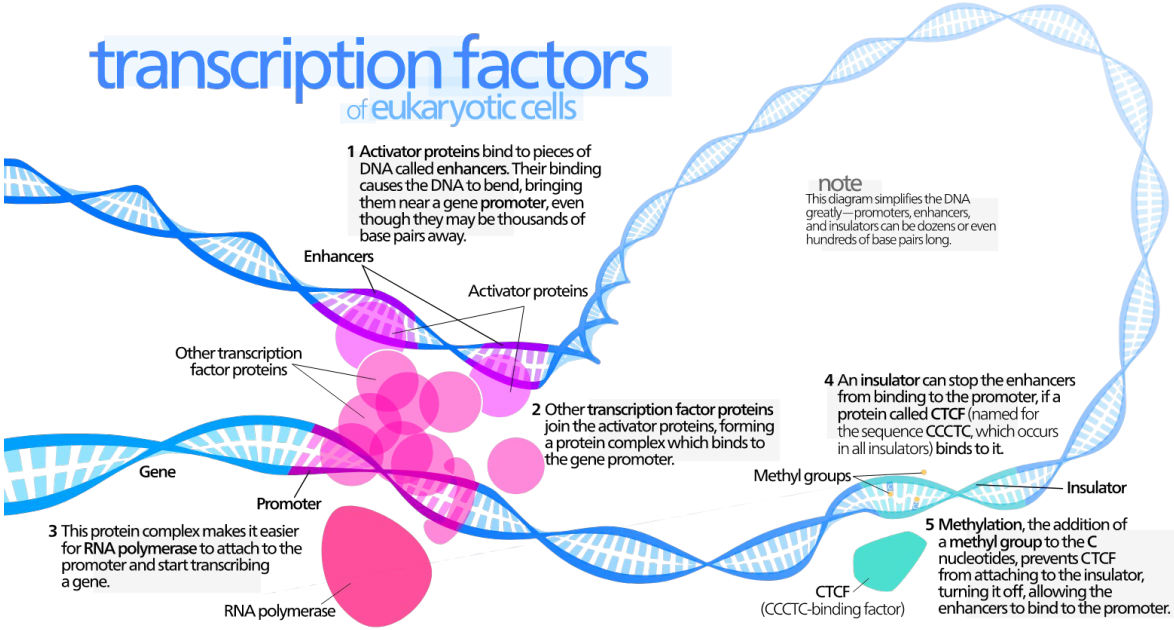
Read /Write

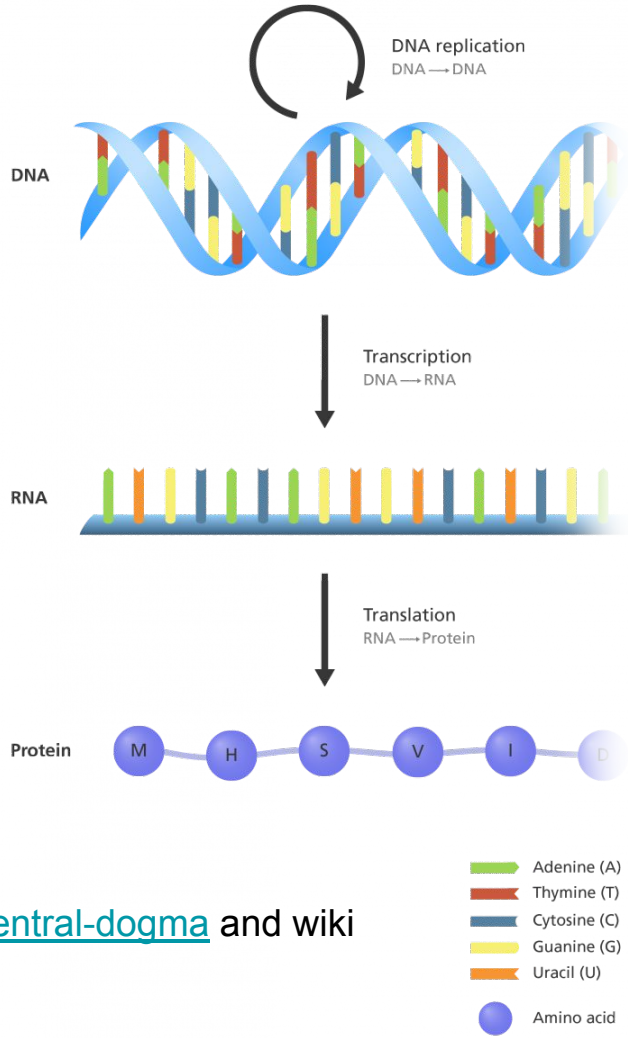
1) Write a string to a file
2) Read back the string from the file

1) Find a DNA motif

What is a DNA motif

we have to understand what is transcription factor (TF)





Images are from https://www.yourgenome.org/facts/what-is-the-central-dogma and wiki page

TF binding pattern-motif

Example: TATA BOX <-> TBP

TATABOX: TATAWAWA

TATA[A/T]A[A/T]A

=>

TATAAAAA

TATAAATA

TATATAAA

TATATATA

```
Nucleotide Code:  Base:
----------------  -----
A................Adenine
C................Cytosine
G................Guanine
T (or U).........Thymine (or Uracil)
R................A or G
Y................C or T
S................G or C
W................A or T
K................G or T
M................A or C
B................C or G or T
D................A or G or T
H................A or C or T
V................A or C or G
N................any base
. or -...........gap
```

How to identify DNA motifs?

Enrichment analysis

An example:

JUND

JUNB

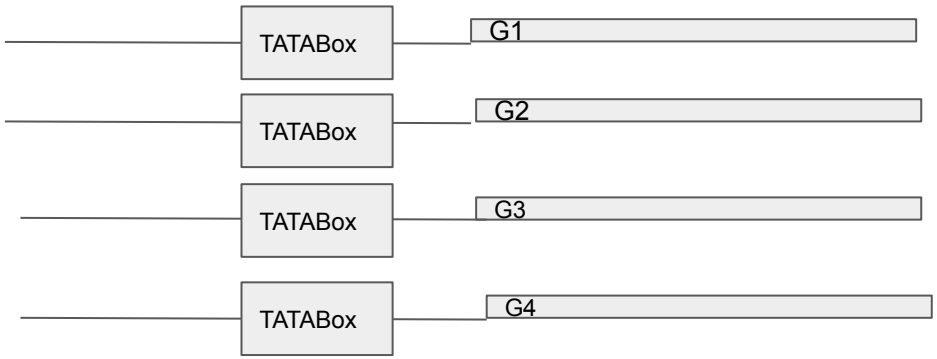
FOS

IRF1

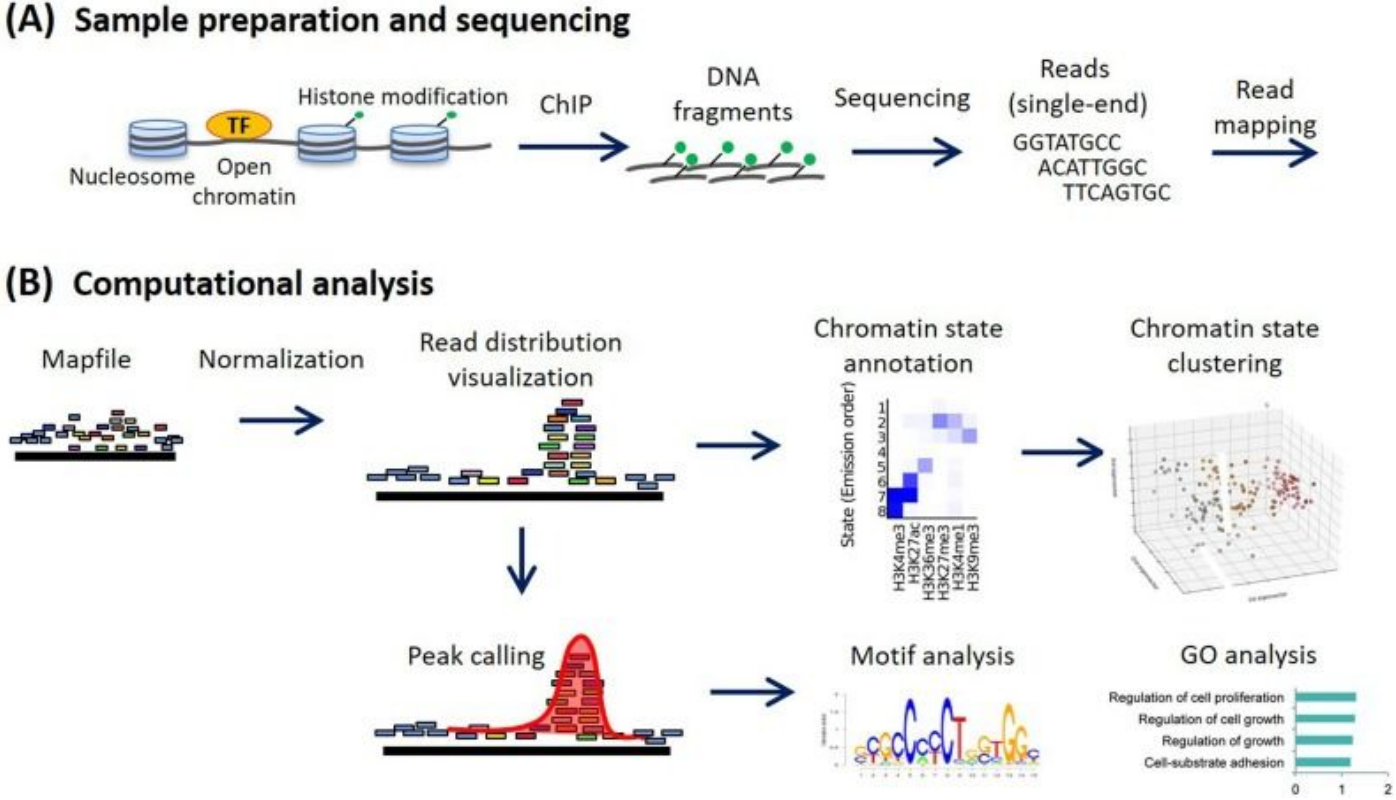
IRF2

ATF2

How to find DNA motifs?



CHIP-Seq



Nakato, Ryuichiro, and Toyonori Sakata. "Methods for ChIP-seq analysis: A practical workflow and advanced applications." *Methods* (2020).

Map the reads to the reference genome

bowtie2

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

```
Bowtie 2 version 2.3.4.1 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r> | --interleaved <i>} [-S <sam>]

  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <i>        Files with interleaved paired-end FASTQ reads
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.
```

Peak calling method

MACS2

https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

p-value

H0: coin is fair (50% chance for head/tail)
Observation: 10 tests, 9 heads

P-value: the probability of observing 9 heads (and more) by random

p=1-pbinom(9-1,10,0.5)=0.01074219 (1%)

<cutoff (often 5% or 1%), reject the HO

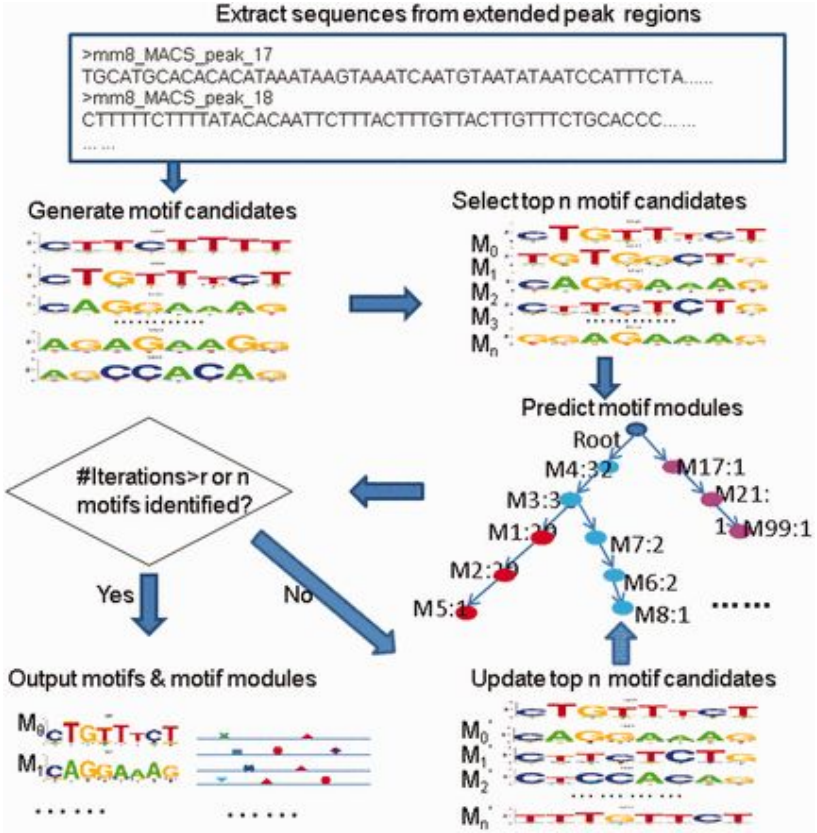
Conclusion (coin is unfair), the probability of wrong conclusion is around 1%

p=1-pbinom(7-1,10,0.5)=0.171=17%

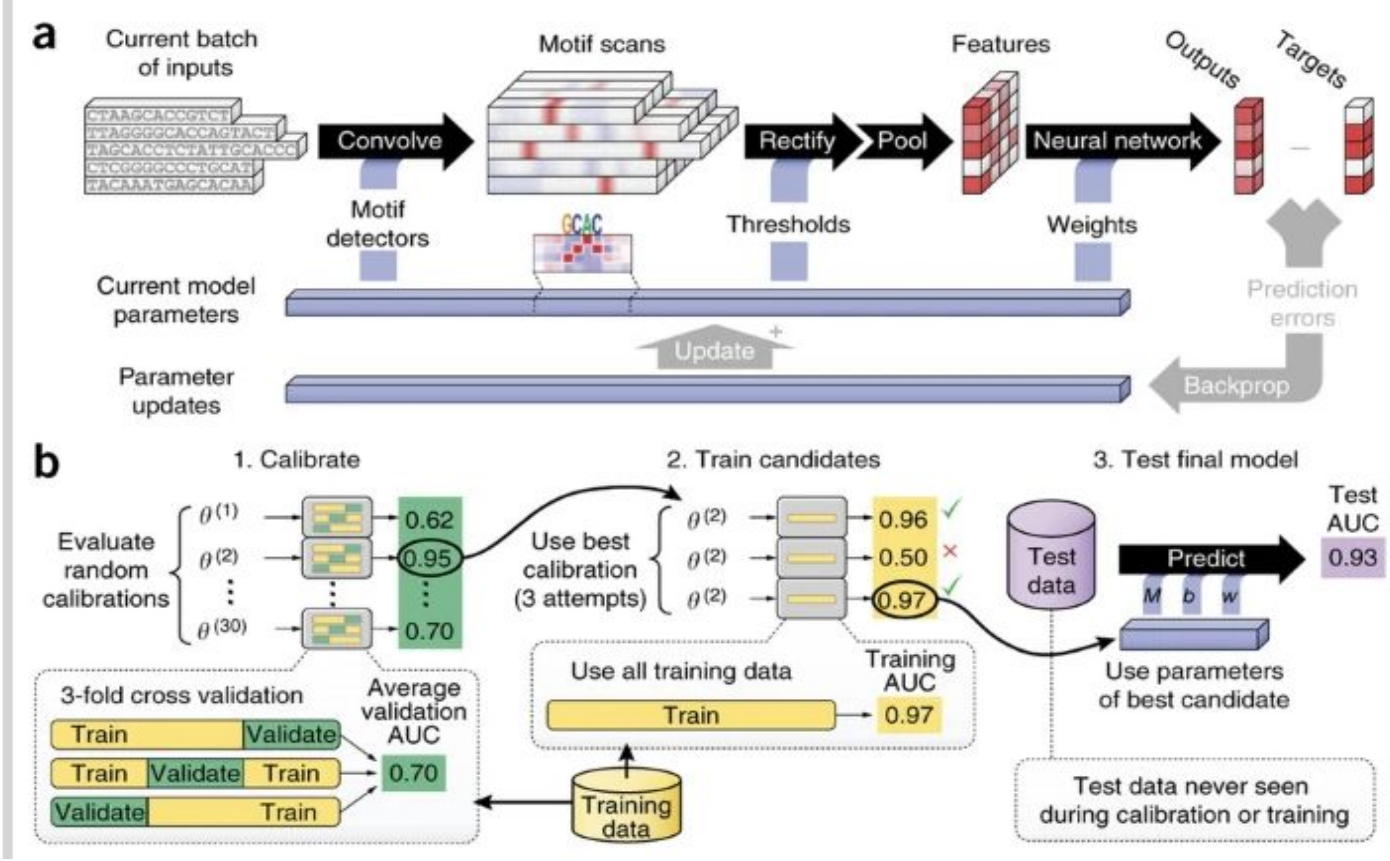
MACS2

```
jund@tiger:~$ macs2 callpeak
usage: macs2 callpeak [-h] -t TFILE [TFILE ...] [-c [CFILE [CFILE ...]]]
                      [-f {AUTO,BAM,SAM,BED,ELAND,ELANDMULTI,ELANDEXPORT,BOWTIE,
BAMPE,BEDPE}]
                      [-g GSIZE] [--keep-dup KEEPDUPLICATES]
                      [--buffer-size BUFFER_SIZE] [--outdir OUTDIR] [-n NAME]
                      [-B] [--verbose VERBOSE] [--trackline] [--SPMR]
                      [-s TSIZE] [--bw BW] [-m MFOLD MFOLD] [--fix-bimodal]
                      [--nomodel] [--shift SHIFT] [--extsize EXTSIZE]
                      [-q QVALUE | -p PVALUE] [--to-large] [--ratio RATIO]
                      [--down-sample] [--seed SEED] [--tempdir TEMPDIR]
                      [--nolambda] [--slocal SMALLLOCAL] [--llocal LARGELOCAL]
                      [--broad] [--broad-cutoff BROADCUTOFF]
                      [--cutoff-analysis] [--call-summits]
                      [--fe-cutoff FECUTOFF]
macs2 callpeak: error: argument -t/--treatment is required
```

```
$ macs2 callpeak -t
bowtie2/H1hesc_Nanog_Rep1_aln.bam \
      -c bowtie2/H1hesc_Input_Rep1_aln.bam \
      -f BAM -g 1.3e+8 \
      -n Nanog-rep1 \
      --outdir macs2
```



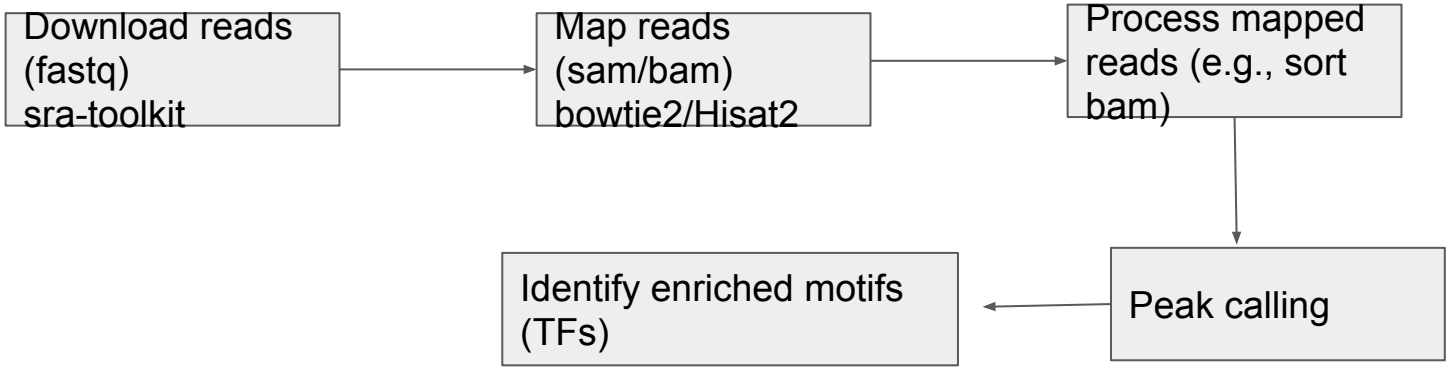
https://academic.oup.com/nar/article/42/5/e35/1055374?login=true



Deepbind

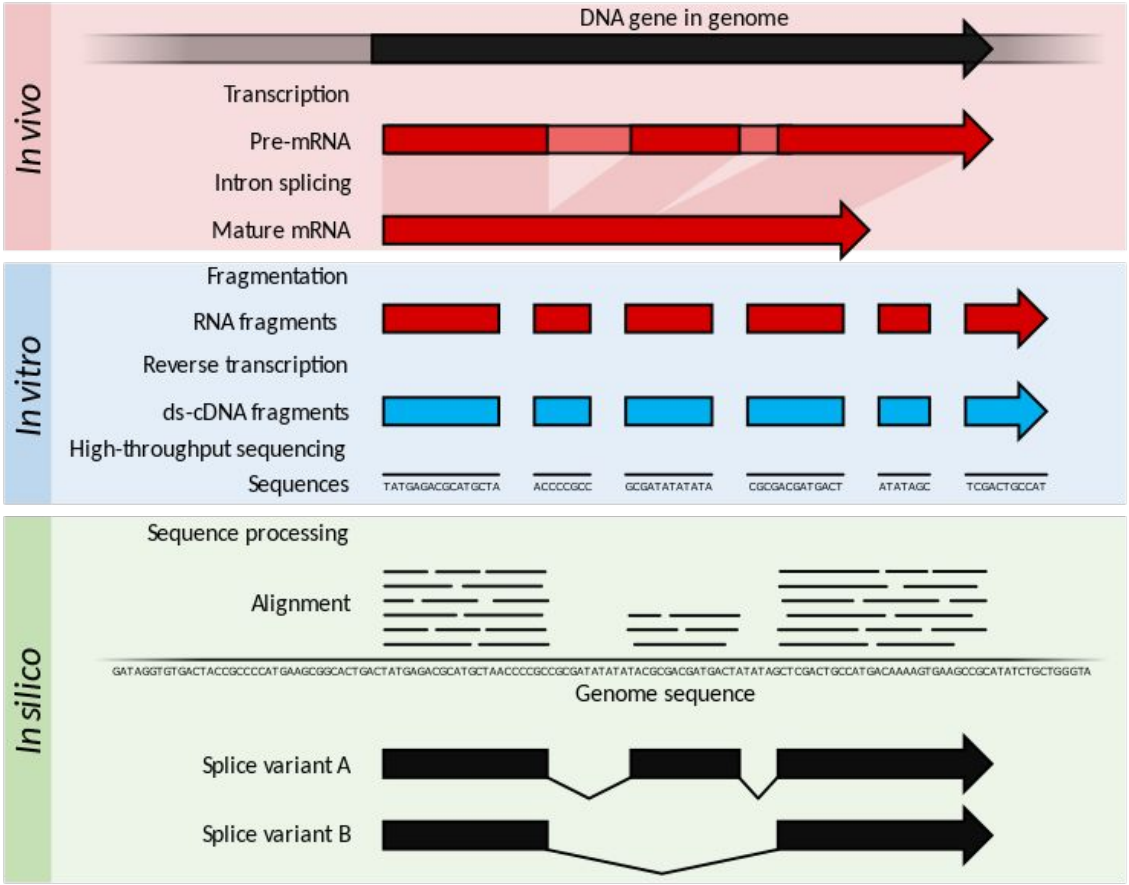
https://www.nature.com/articles/nbt.3300

Chip-seq data analysis pipeline



RNA-seq data analysis

RNA-seq


In vivo
DNA gene in genome
Transcription
Pre-mRNA
Intron splicing
Mature mRNA
In vitro
Fragmentation
RNA fragments
Reverse transcription
ds-cDNA fragments
High-throughput sequencing
Sequences
TATGAGACGCATGCTA
ACCCCGCC
GCGATATATATA
CGCGACGATGACT
ATATAGC
TCGACTGCCAT
In silico
Sequence processing
Alignment
GATAGGTGTGACTACCGCCCCATGAAGCGGCACTGACTATGAGACGCATGCTAACCCCGCCGCGATATATATACGCGACGATGACTATATAGCTCGACTGCCATGACAAAAGTGAAGCCGCATATCTGCTGGGTA
Genome sequence
Splice variant A
Splice variant B

Quality control (qc)

Depending on the quality, you might need to trim the reads

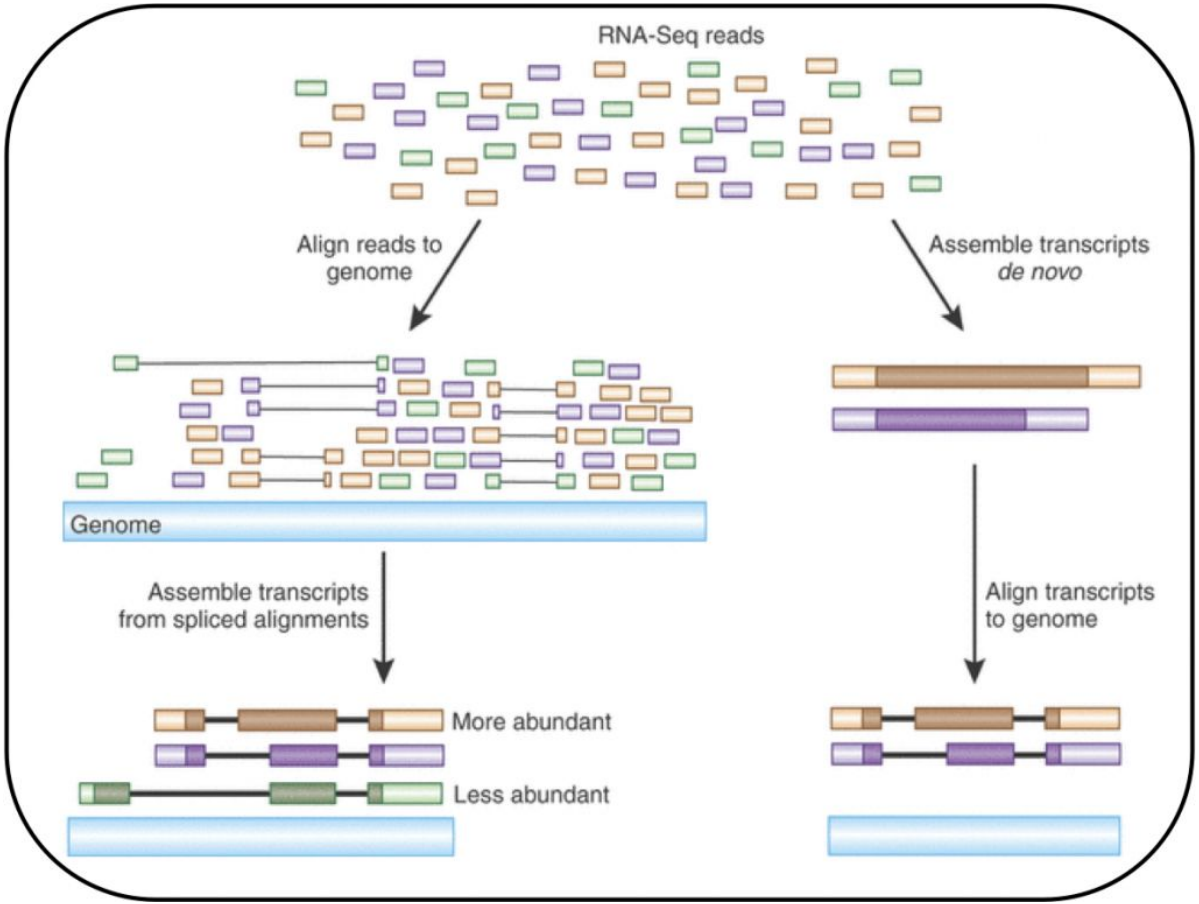
fastp

Optional

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Reads mapping

HISAT2

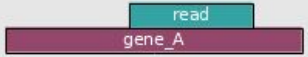
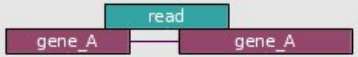
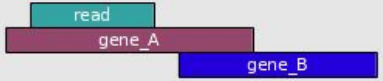
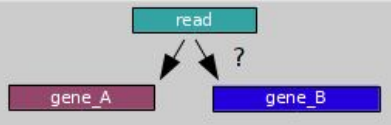


All images are from network

All images are from network

Quantify gene expression

htseq-count [options] <alignment_files> <gff_file>

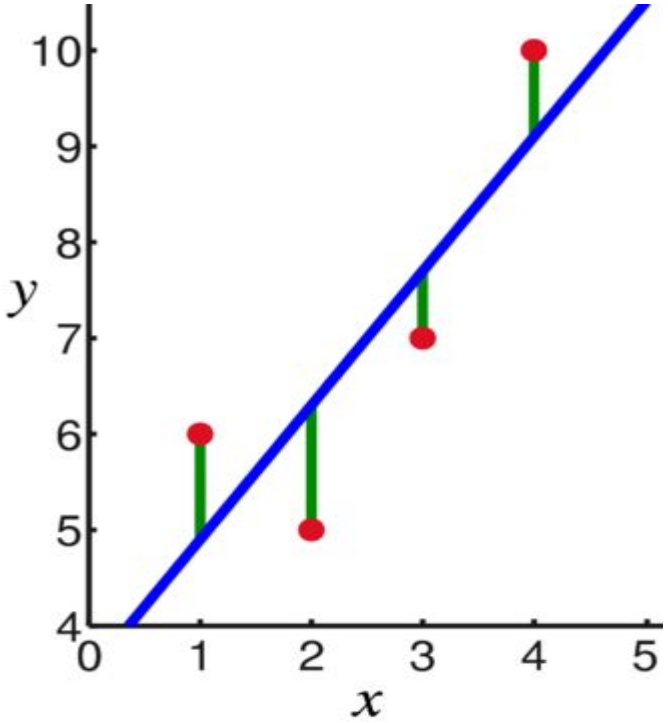
	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

All images are from network

Linear regression

Linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$



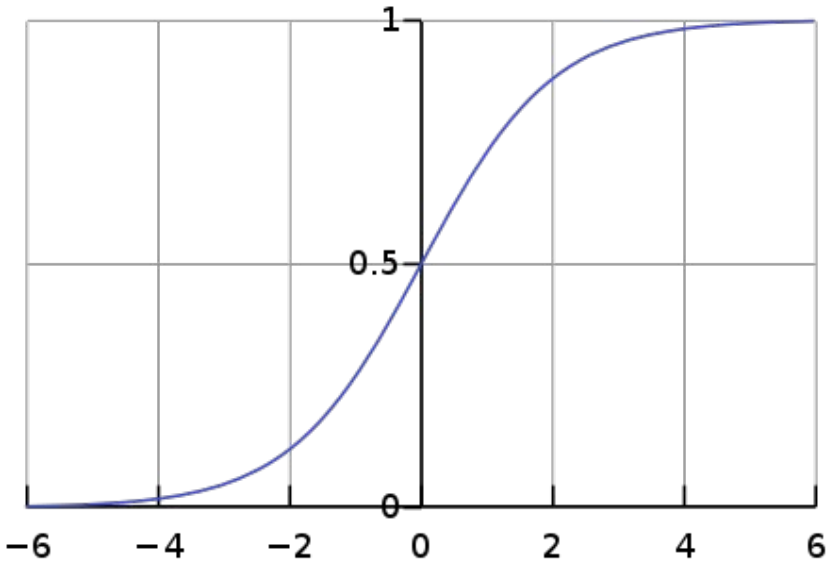
How to search the parameters?
1）Brute-force
2）Gradient descent

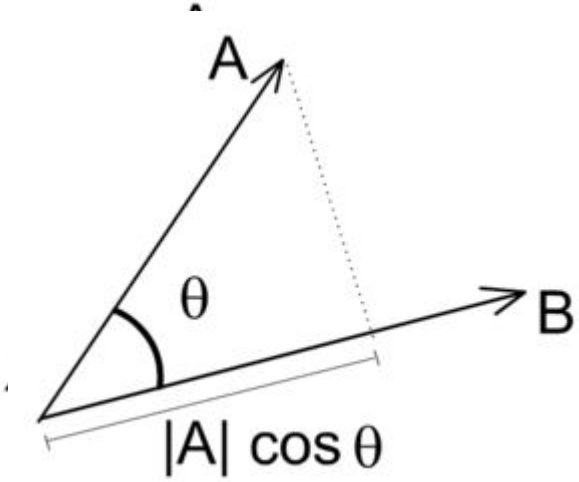
$$E = \frac{1}{n} \sum_{i=0}^{n} (y_i - \bar{y}_i)^2$$

All images are from network

Logistic regression

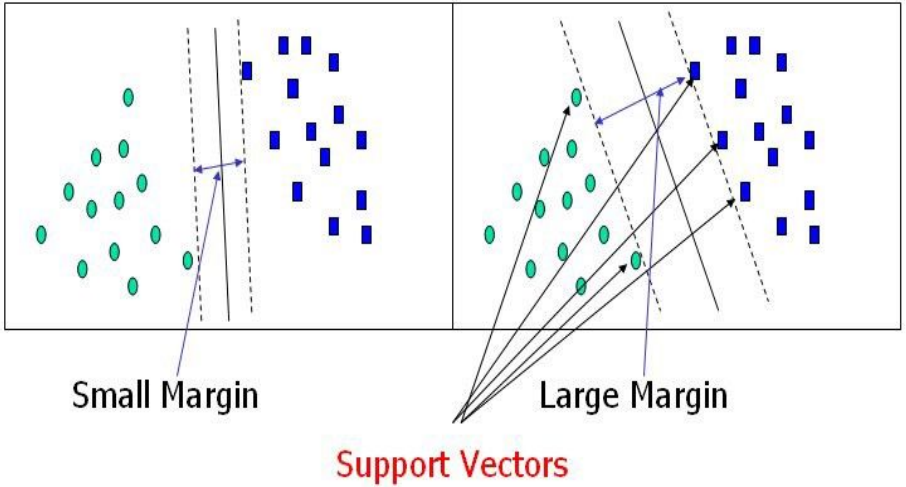
$$p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

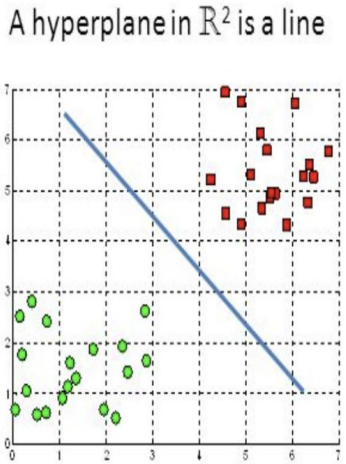


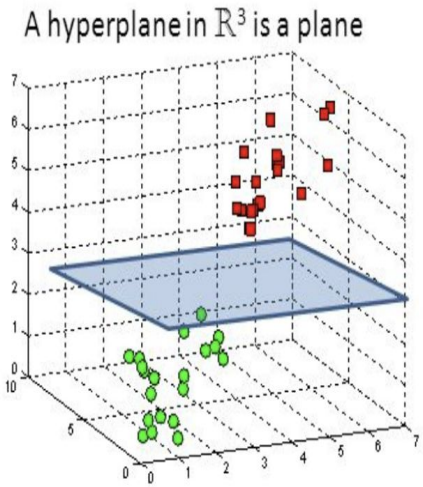


All images are from network

Support-vector machine (SVM)



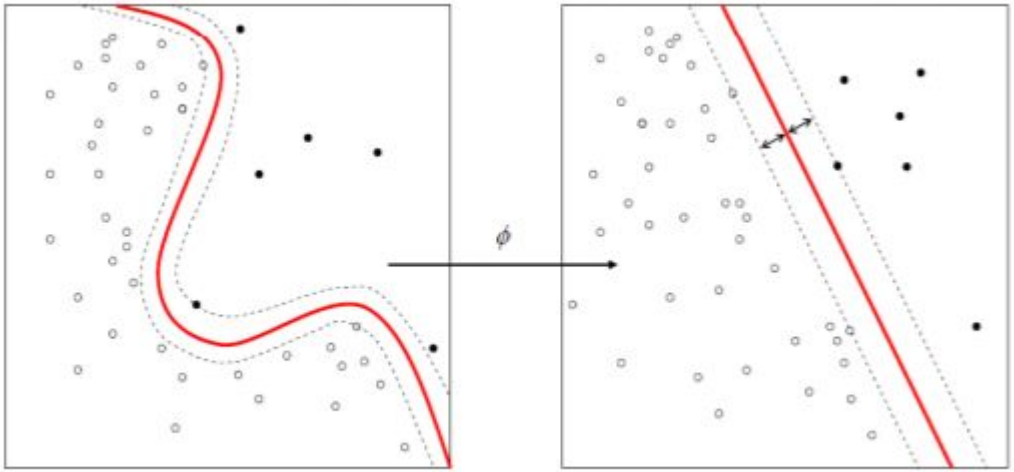




https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

All images are from network

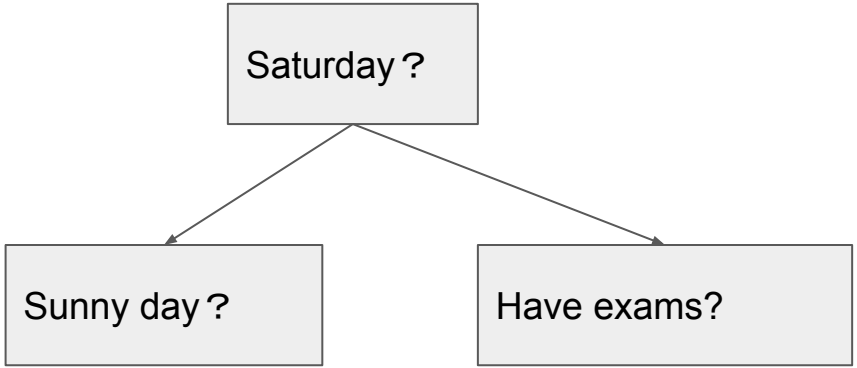
Linear/non-linear classifier



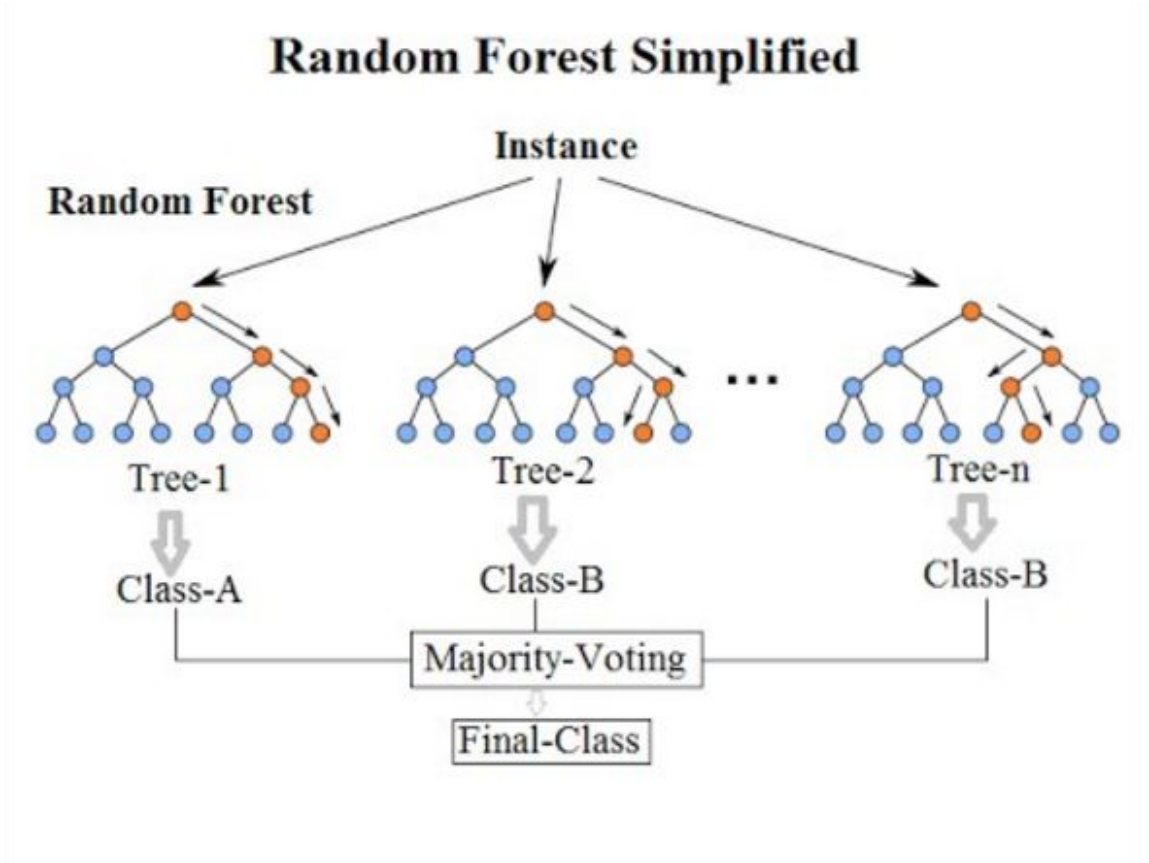
Kernel function

All images are from network

Decision tree

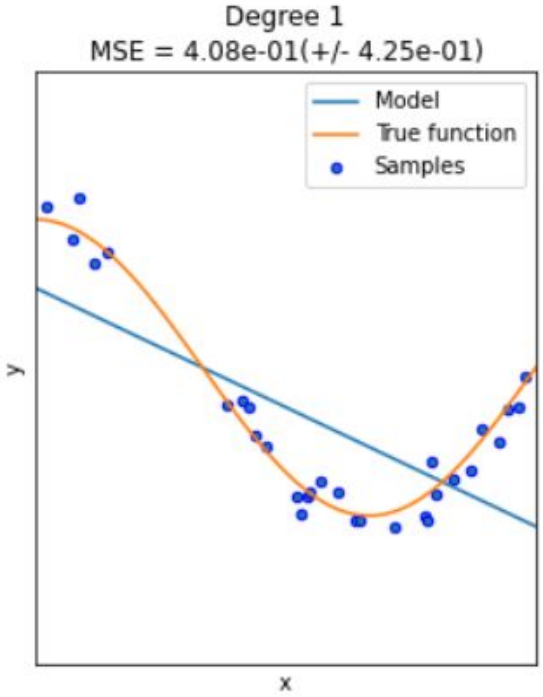


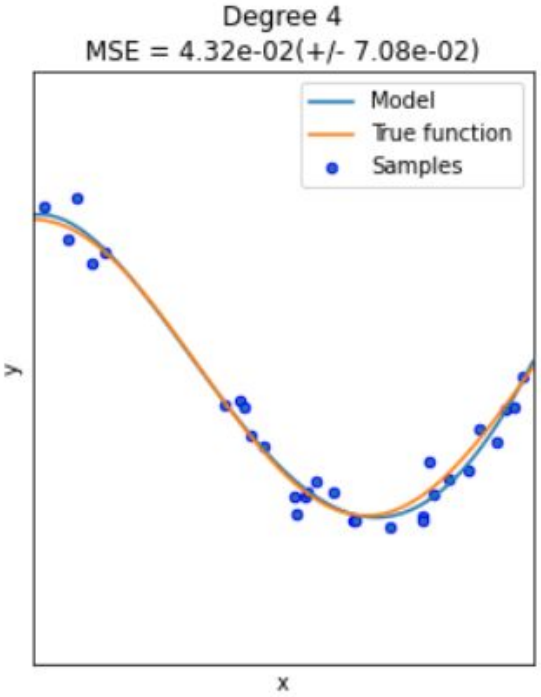
Random forest

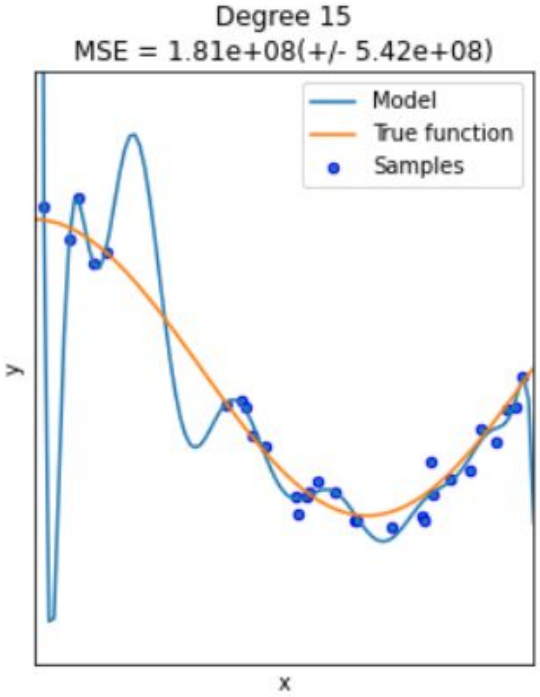


All images are from network

Pitfalls







under-fitting

good

over-fitting

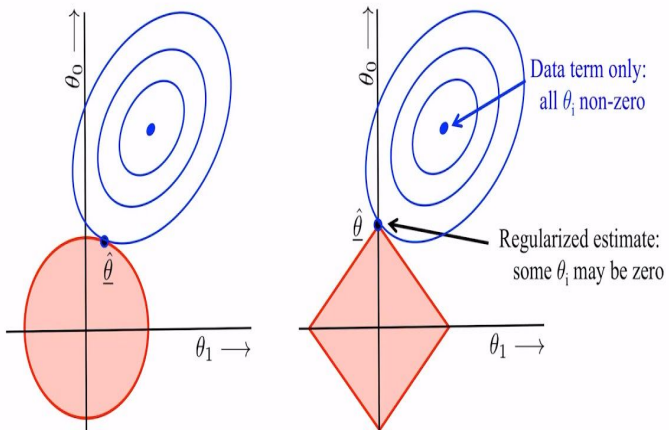
All images are from network

Regularization

L1- regularization

L2-regularization

- L1 tends to generate sparser solutions than a quadratic regularizer



All images are from network

L1 Regularization

$$\text{Cost} = \sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{ij} W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

$$\text{Cost} = \underbrace{\sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{ij} W_j)^2}_{\text{Loss function}} + \underbrace{\lambda \sum_{j=0}^{M} W_j^2}_{\text{Regularization Term}}$$

An example data

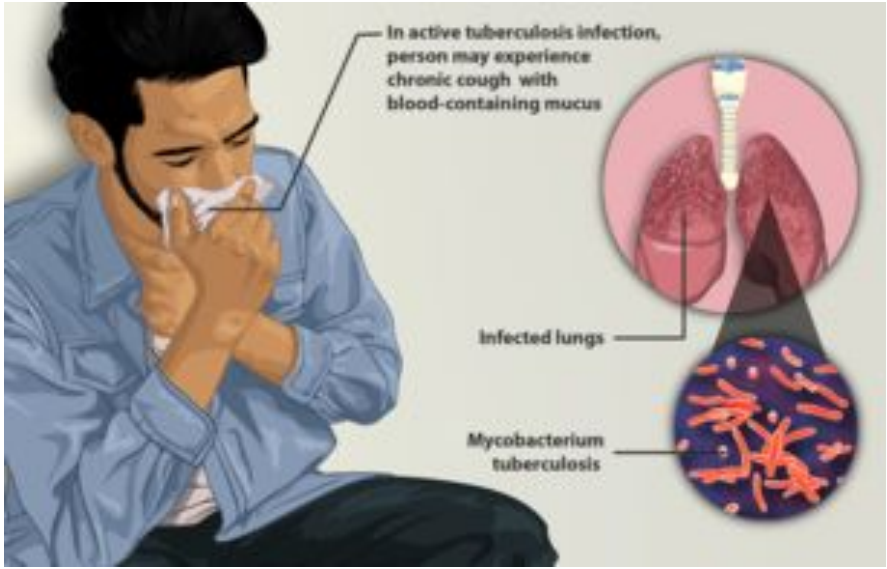
https://www.kaggle.com/uciml/pima-indians-diabetes-database

Try to download the data

wget https://filedn.com/lL2xsyY8teiHHTk3wYqUmVu/sdu_summerclass/RNA/diabetes.txt

An example study (biomarker discovery)

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162164



The image is from wiki

Homework 1:
description: you need to build a classifier (e.g., random forest or svm) for the prediction of TB on HIV patients.
steps :
1) download the dataset from the NIH GEO database
https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE162164&format=file

2) annotate the patients with its phenotype
in other words, some patients are HIV only, the rest are HIV + TB

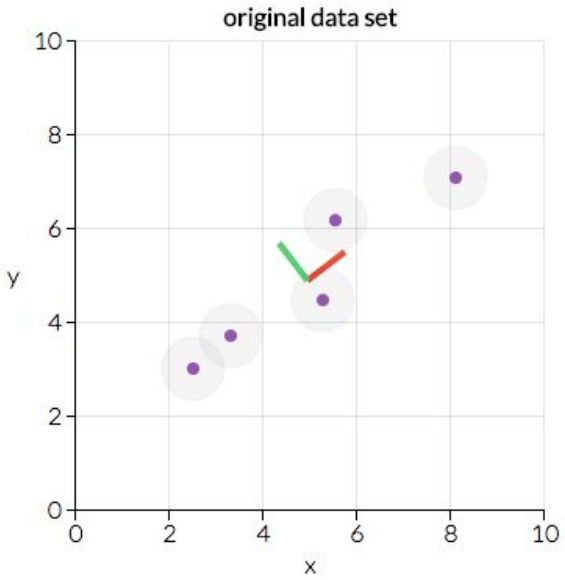
3) do some basic file reading and processing (convert float values)
4) train a model, could be tricky, the performance could be very bad
5) you need some tricks to minimize the number of features (some feature selection to reduce the feature space) , for example, if you find a gene that is not very different from HIV vs HIV+TB, then you know this feature won't be important
6) you train the model and calculate the accuracy, report it
7) write a report (jupyter notebook, detailing each of the steps and results)
you need also to tell me what is the biomarker (the most critical feature for the TB+HIV disease)

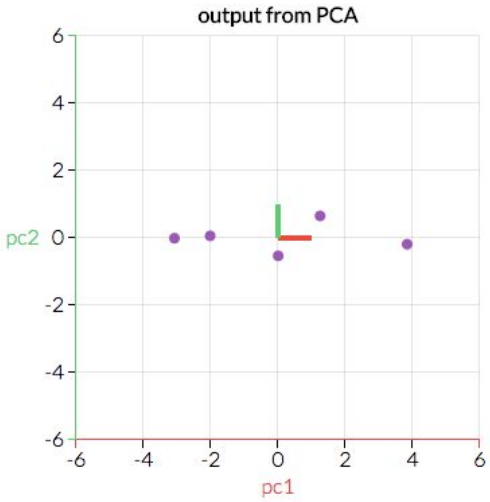
Clustering methods

Visualize the data

PCA (Essentially, it's linear transformation)

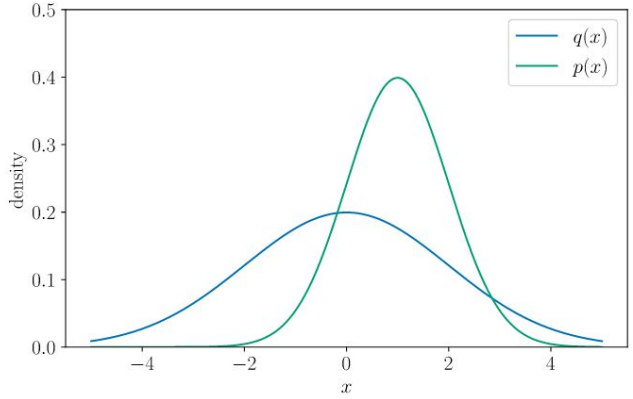
$$\sum_{i=1}^{n} ||x_i - Px_i||^2$$





Visualize the data

t-SNE (and UMAP)



$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

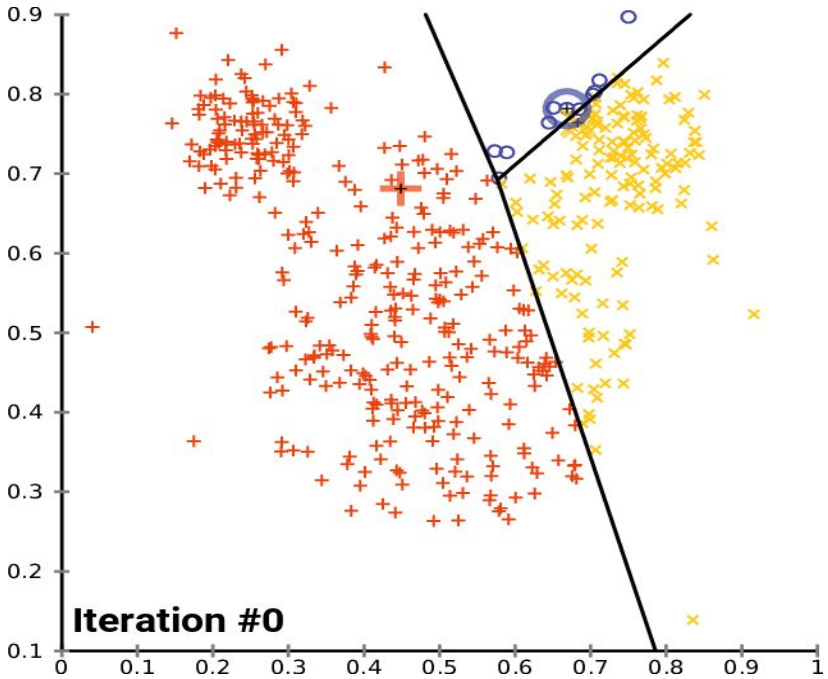
$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

Clustering methods

K-Means

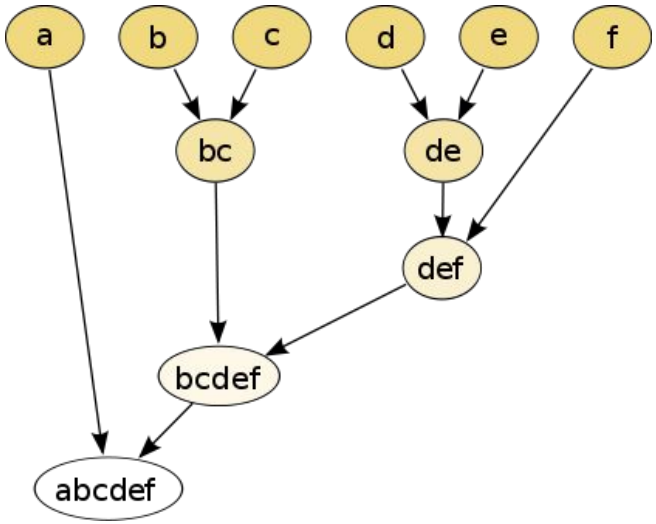
$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$



https://en.wikipedia.org/wiki/K-means_clustering#/media/File:K-means_convergence.gif `

Clustering methods

Hierarchical clustering



- The maximum distance between elements of each cluster (also called complete-linkage clustering):

$$\max\{\, d(x,y) : x \in \mathcal{A},\ y \in \mathcal{B} \,\}.$$

- The minimum distance between elements of each cluster (also called single-linkage clustering):

$$\min\{\, d(x,y) : x \in \mathcal{A},\ y \in \mathcal{B} \,\}.$$

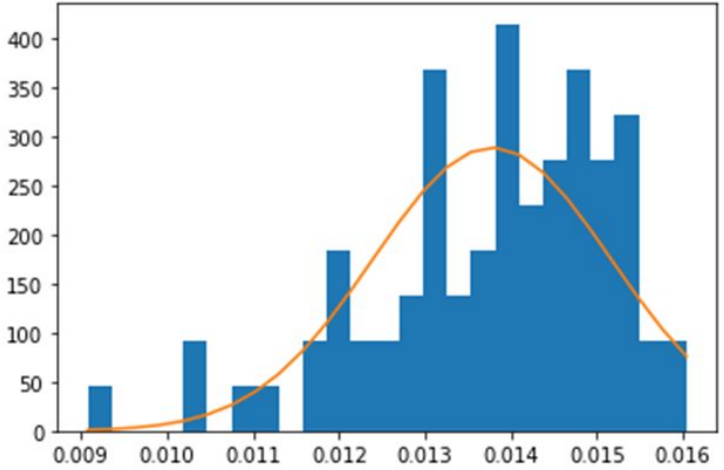
- The mean distance between elements of each cluster (also called average linkage clustering, used e.g. in UPGMA):

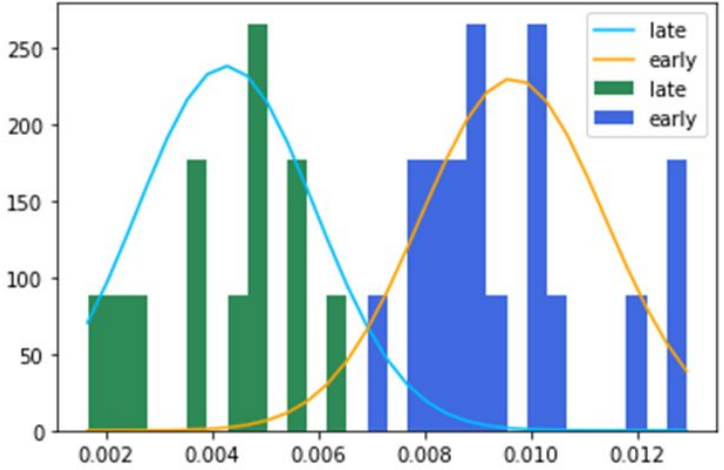
$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x,y).$$

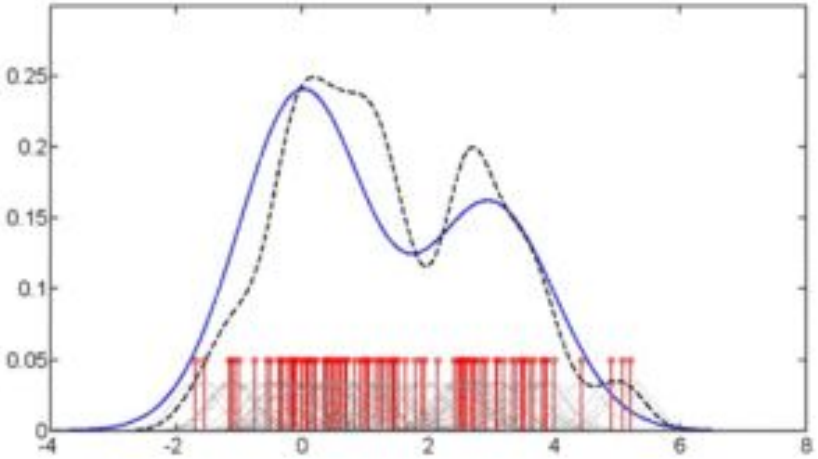
- The sum of all intra-cluster variance.

Clustering methods

Density estimator

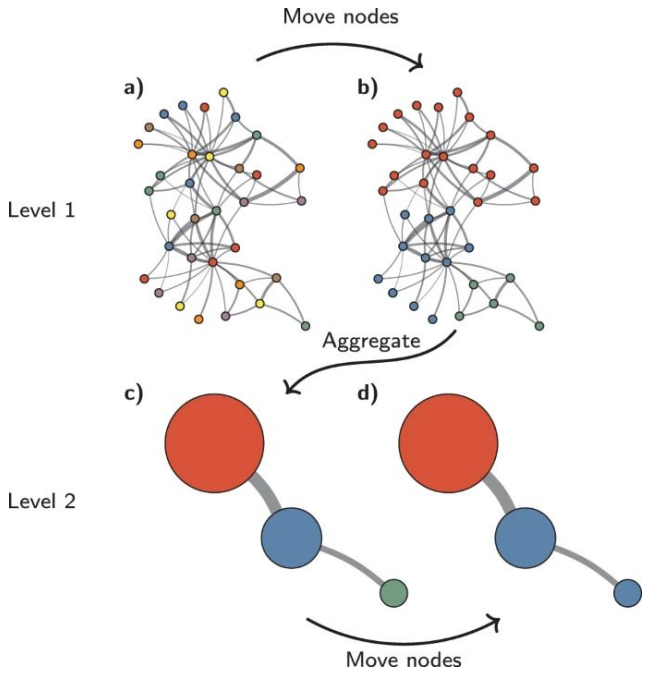


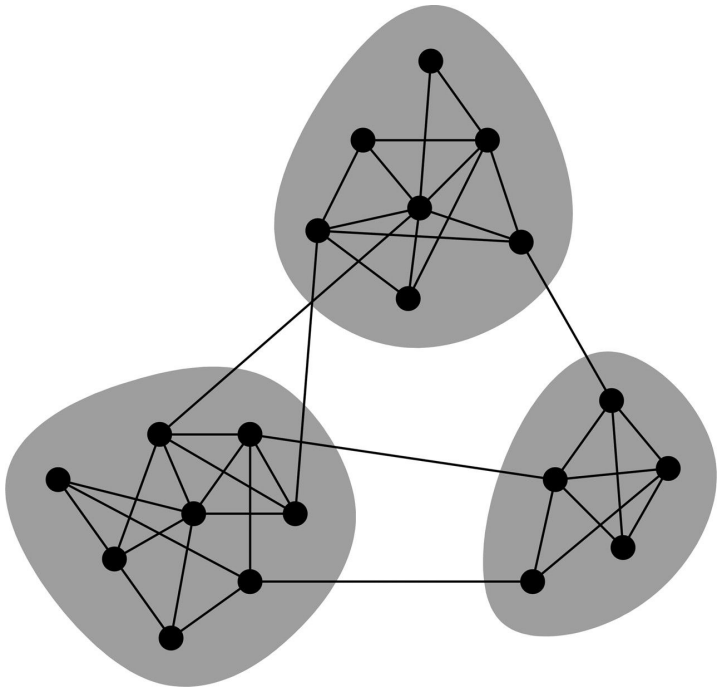




Leiden clustering (/Louvian clustering)

Modularity is a measure of the structure of networks or graphs which measures the strength of division of a network into modules (also called groups, clusters or communities).





All images are from network

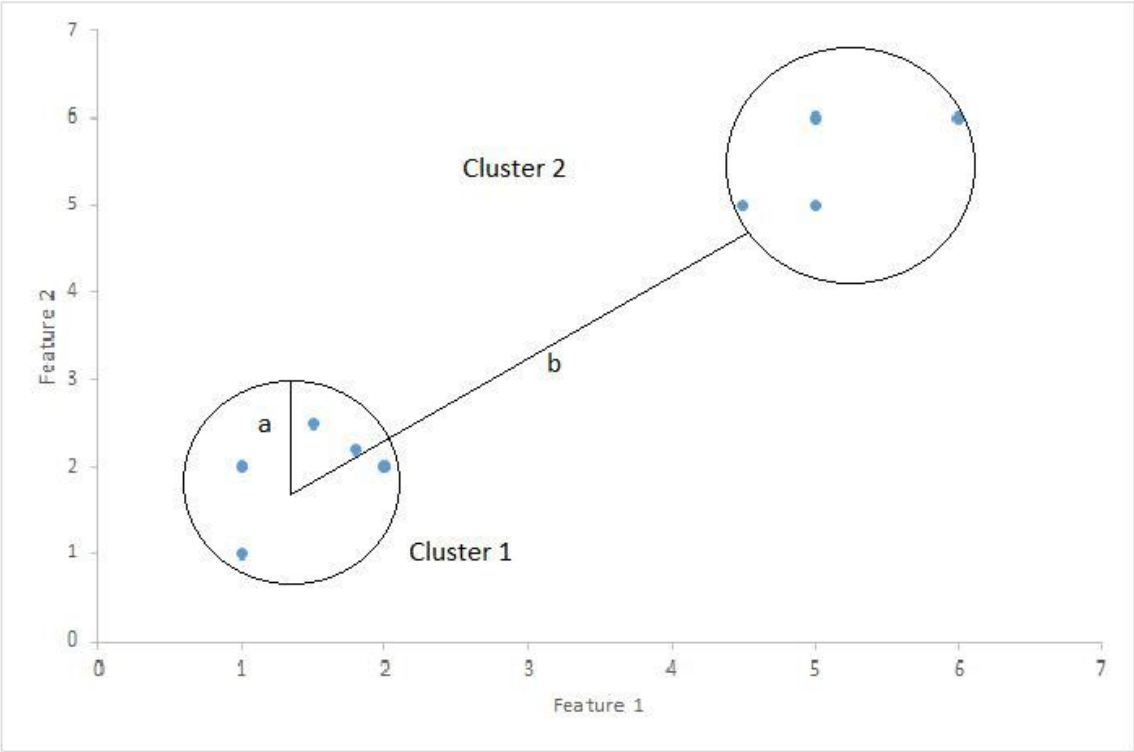
Model selection?

How to choose # of cluster (K) in the K-means?

How to choose the resolution parameter in Leiden clustering ?

Clustering evaluation metrics

Silhouette Score = (b-a)/max(a,b)

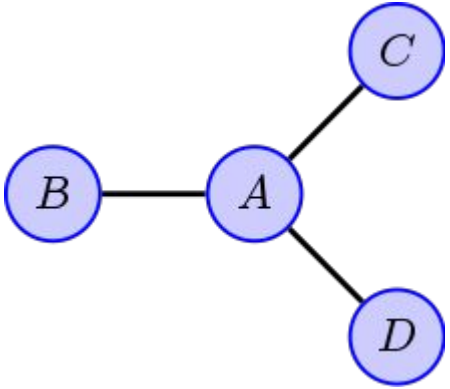


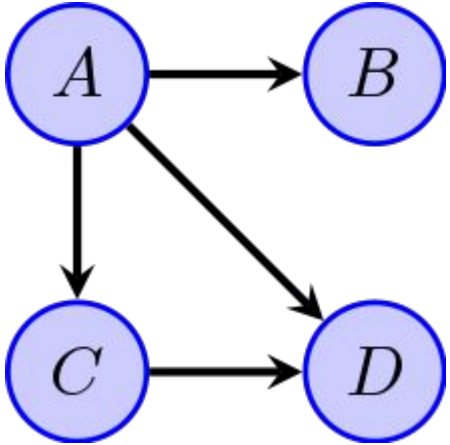
Graphical models

What is a graph

Node

Edge





Bayesian network

$$P(X_1,\ldots,X_n)=\prod_{i=1}^{n}P(X_i\mid X_1,\ldots,X_{i-1})=\prod_{i=1}^{n}P(X_i\mid Parents(X_i))$$

1) Inference

p(x|e)=p(x,e)/p(e)

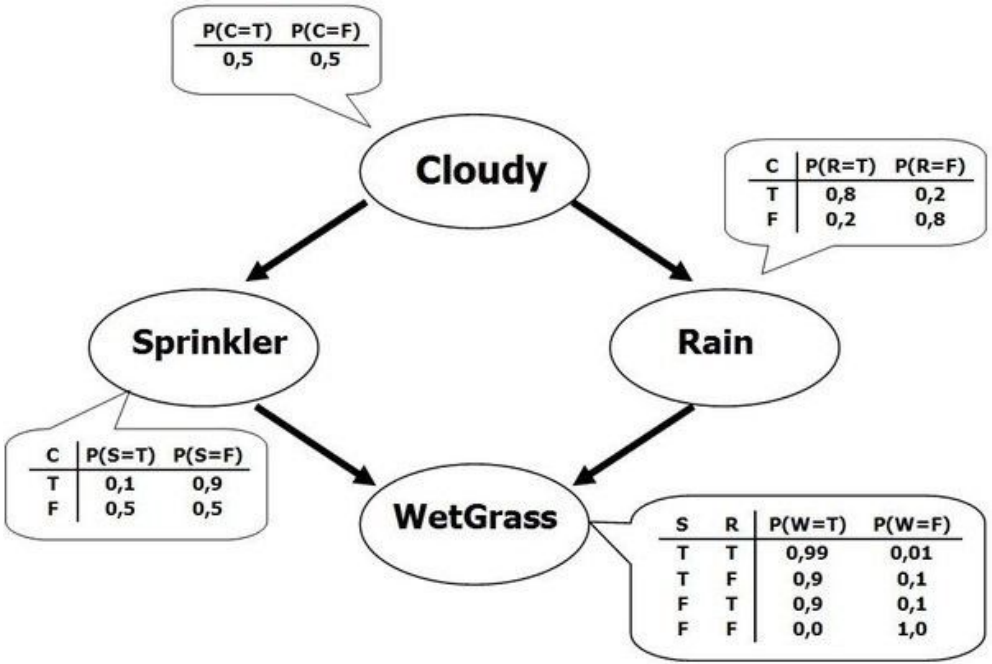
$$P(x|e)=\alpha\sum_{\forall y\in Y}P(x,e,Y)$$

p(WetGrass=True|Clouldy=True)

(2) Parameter learning

P(A|B)=N(mu,sigma)

3) Structure learning



Bayesian network

C=True=> S={0.1,0.9}

C=True=> R={0.8,0.2}

C=True => W=True

C=True=>S=True, R=True (0.1*0.8) =>0.99

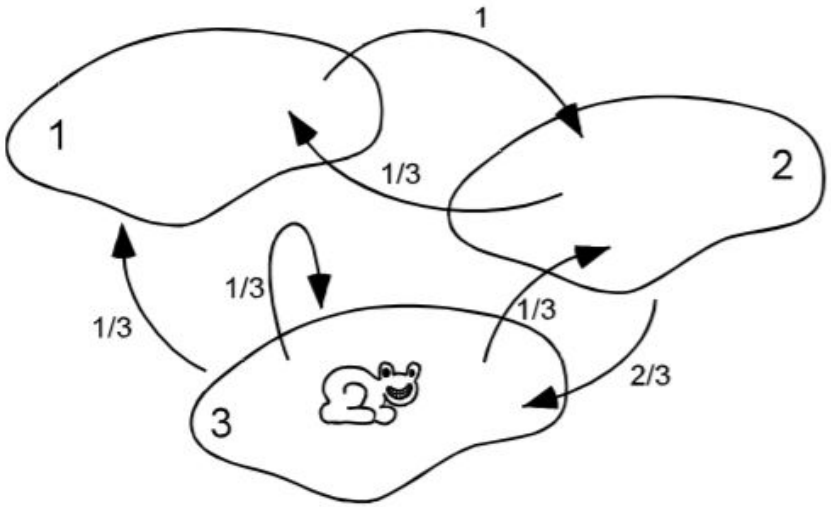
S=True, R=False (0.1*0.2) =>0.9

S=False, R=True (0.9*0.8) =>0.9

S=False, R=False (0.9,0.2) =>0

Markov chain

$$P(X_{n+m} = s | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) = P(X_{n+m} = s | X_{n-1} = i_{n-1})$$



$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1/3 & 0 & 2/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array}.$$

$$P^5 = \begin{pmatrix} 0.246914 & 0.407407 & 0.345679 \\ 0.251029 & 0.36214 & 0.386831 \\ 0.251029 & 0.366255 & 0.382716 \end{pmatrix},$$

$$P^{10} = \begin{pmatrix} 0.250013 & 0.37474 & 0.375248 \\ 0.249996 & 0.375095 & 0.374909 \\ 0.249996 & 0.375078 & 0.374926 \end{pmatrix},$$

$$P^{20} = \begin{pmatrix} 0.2500000002 & 0.3749999913 & 0.3750000085 \\ 0.2499999999 & 0.375000003 & 0.374999997 \\ 0.2499999999 & 0.3750000028 & 0.3749999973 \end{pmatrix}.$$

Diffusion

P=[p11,p12
P21,p22]

P^2=[p11*p11+p12*p21, p11*p12+p12*p22
p21*p11+p22*p21, p21*p12+p22*p22]

…

P^N

High-order Markov Chain

ACGTACTTCGAGGTTTTTAAACTACTACT

2nd transition matrix

AC->G
CT->T
GT->A
TA->C

Transition matrix in the upstream region of the following genes

JUND

JUNB

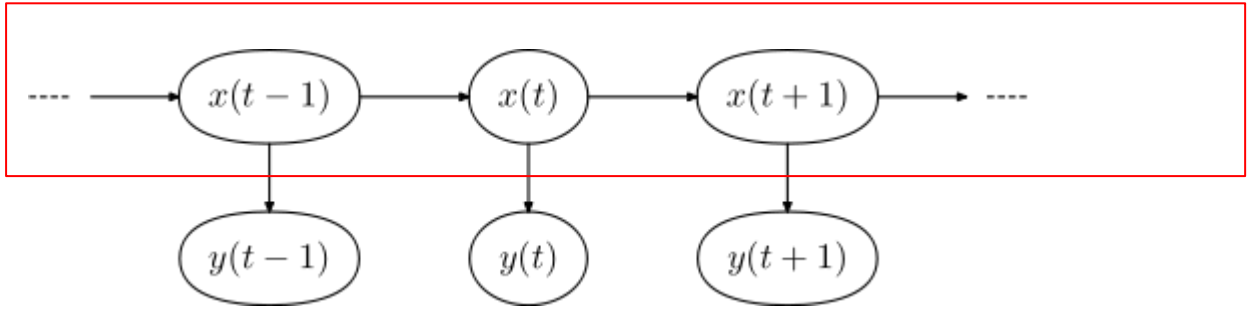
FOS

IRF1

IRF2

ATF2

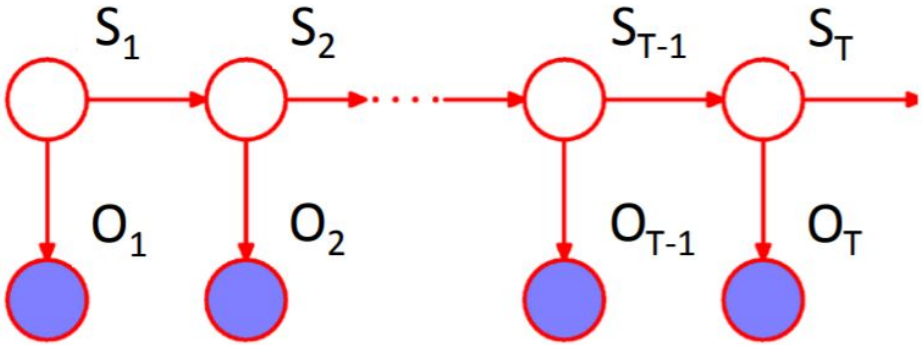
Hidden Markov Model (HMM)



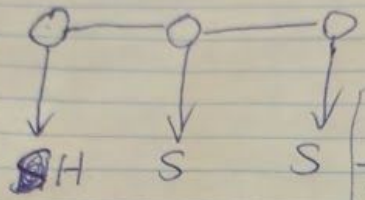
Initial probabilities
Transition probabilities
Emission probabilities P(O|S)

Inference

Given Pi, emission matrix, transition matrix => infer hidden states that fit the observation



$$p(S_1, \ldots, S_T, O_1, \ldots, O_T) = \prod_{t=1}^{T} p(O_t|S_t) \prod_{t=1}^{T} p(S_t|S_{t-1})$$



	Y
T	0.5
	0.5

EM

	Y	N
H	0.2	0.6
S	0.8	0.4

TM

	Y	N
Y	0.9	0.1
N	0.3	0.7

p-value

H0: coin is fair (50% chance for head/tail)
Observation: 10 tests, 9 heads

P-value: the probability of observing 9 heads (and more) by random

p=1-pbinom(9-1,10,0.5)=0.01074219 (1%)

<cutoff (often 5% or 1%), reject the HO

Conclusion (coin is unfair), the probability of wrong conclusion is around 1%

p=1-pbinom(7-1,10,0.5)=0.171=17%

Mann-whitney U test

H0: the probability of *X* being greater than *Y* is equal to the probability of *Y* being greater than *X*.

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j),$$

with

$$S(X, Y) = \begin{cases} 1, & \text{if } Y < X, \\ \frac{1}{2}, & \text{if } Y = X, \\ 0, & \text{if } Y > X. \end{cases}$$

How to calculate U-statistics

L: Li lei

H: Han meimei

Result:

L H H H H H H L L L L L H

U1=

U2=

L=[1,3,7,8]
H=[2,1,9,6]

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j),$$

with

$$S(X,Y) = \begin{cases} 1, & \text{if } Y < X, \\ \frac{1}{2}, & \text{if } Y = X, \\ 0, & \text{if } Y > X. \end{cases}$$

Fold change

X=[2,3,4,5]

Y=[5,6,8,12]

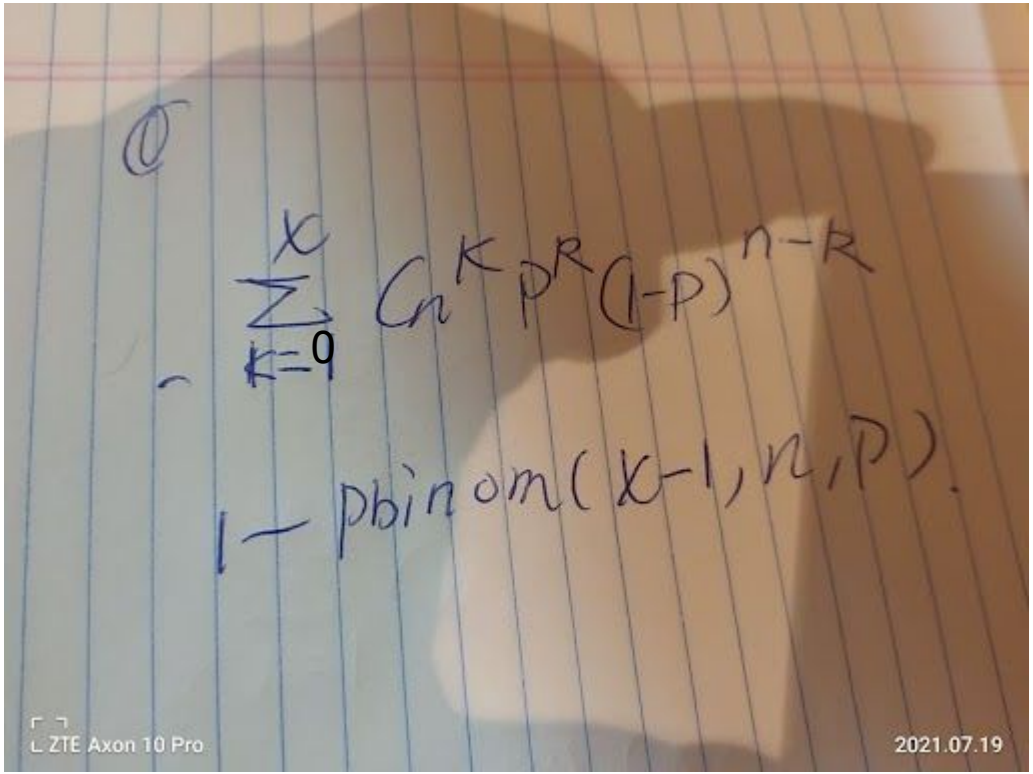
x->y

```
If mean(Y)>mean(X):
        Fold change=mean(Y)/mean(X)
Else:
        Fold change=-1* mean(X)/mean(Y)
```

Binomial test

Background frequency: 0.5
20 invites
Yes: 15 times
20: 5 times

P-value?



Single-cell genomics

Cells: the building blocks of life

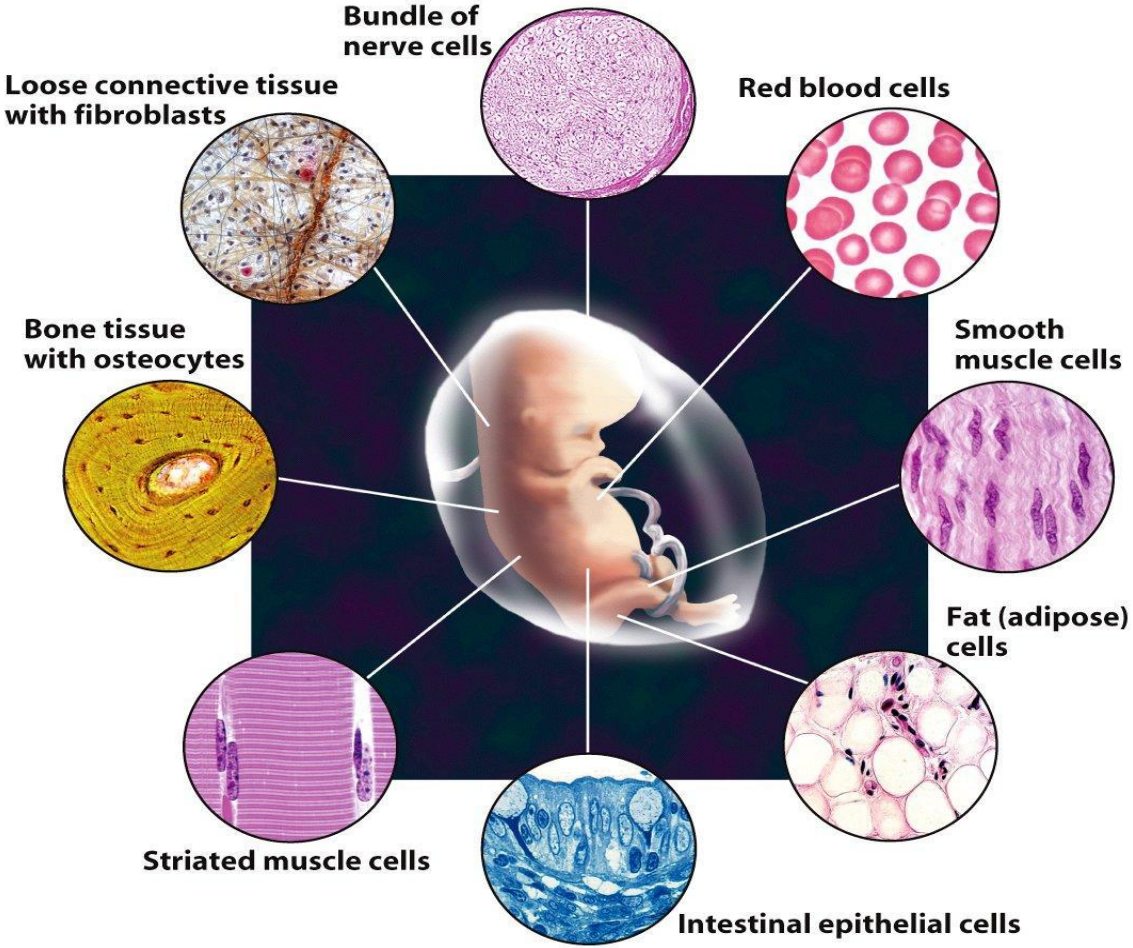
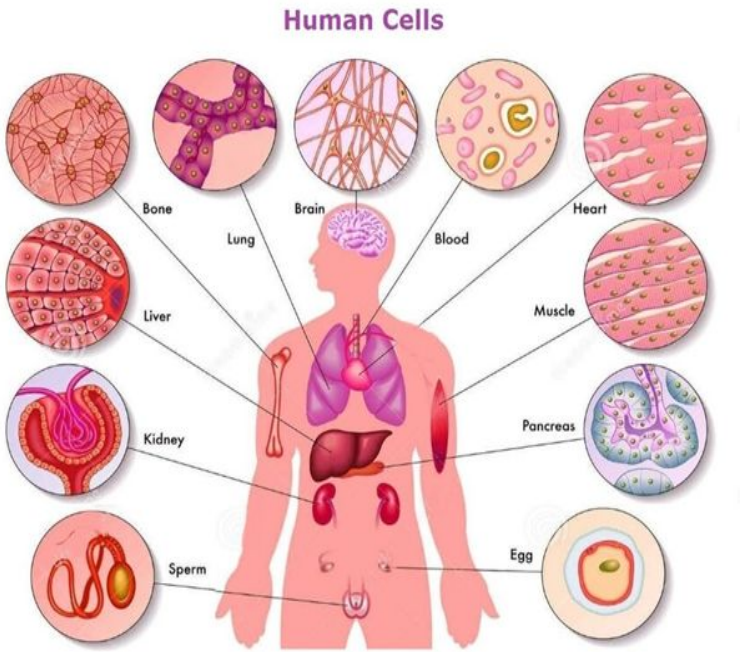
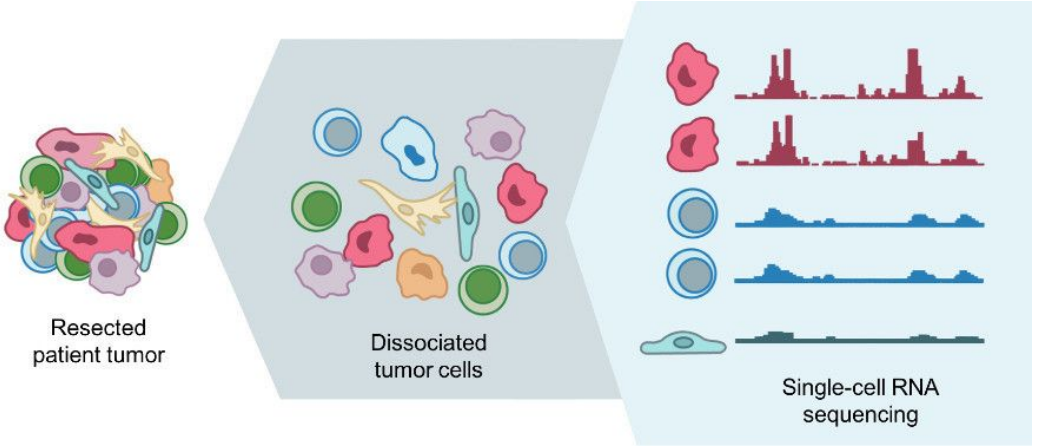


Figure 1-17 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

Why we need the single-cell?

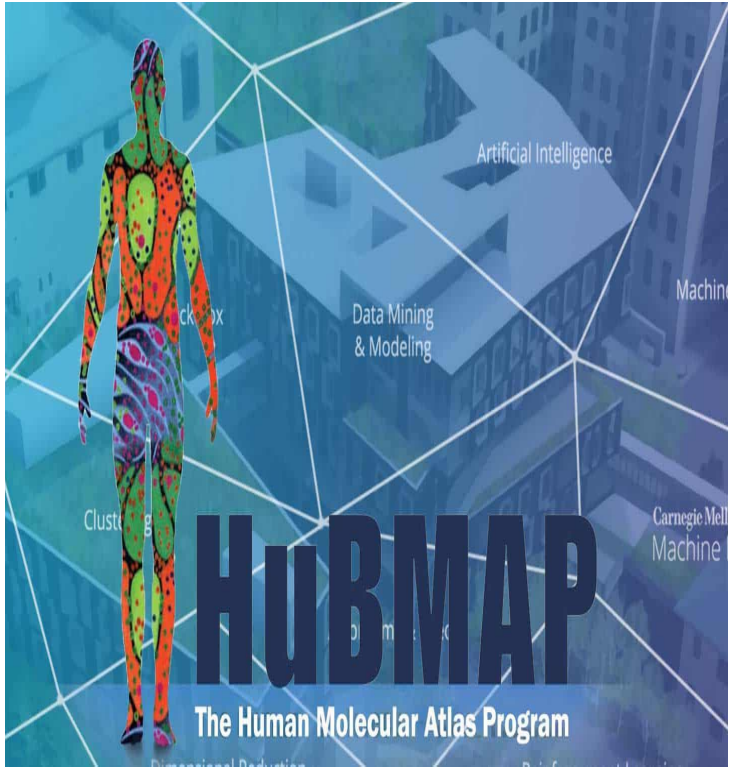


Build a google map of human body

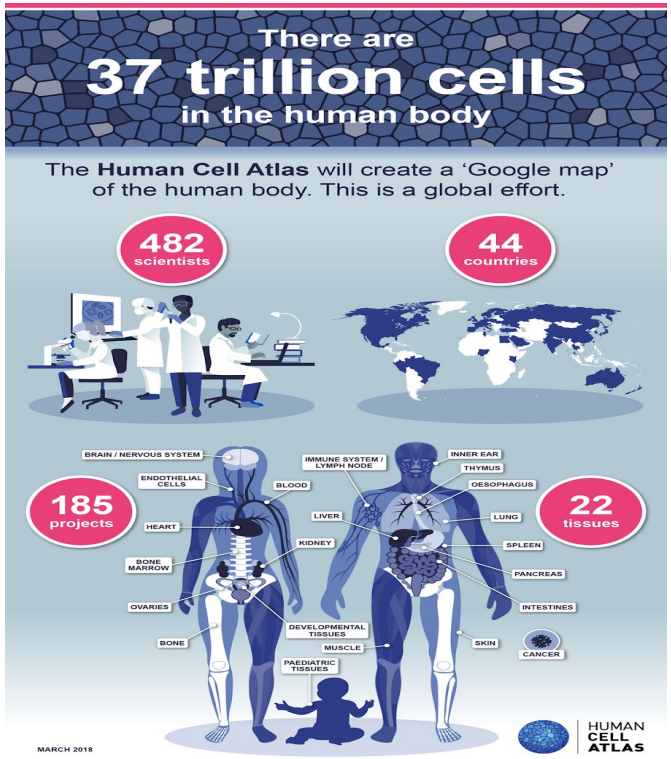


Better disease diagnosis and treatment

Cell Atlas Initiatives

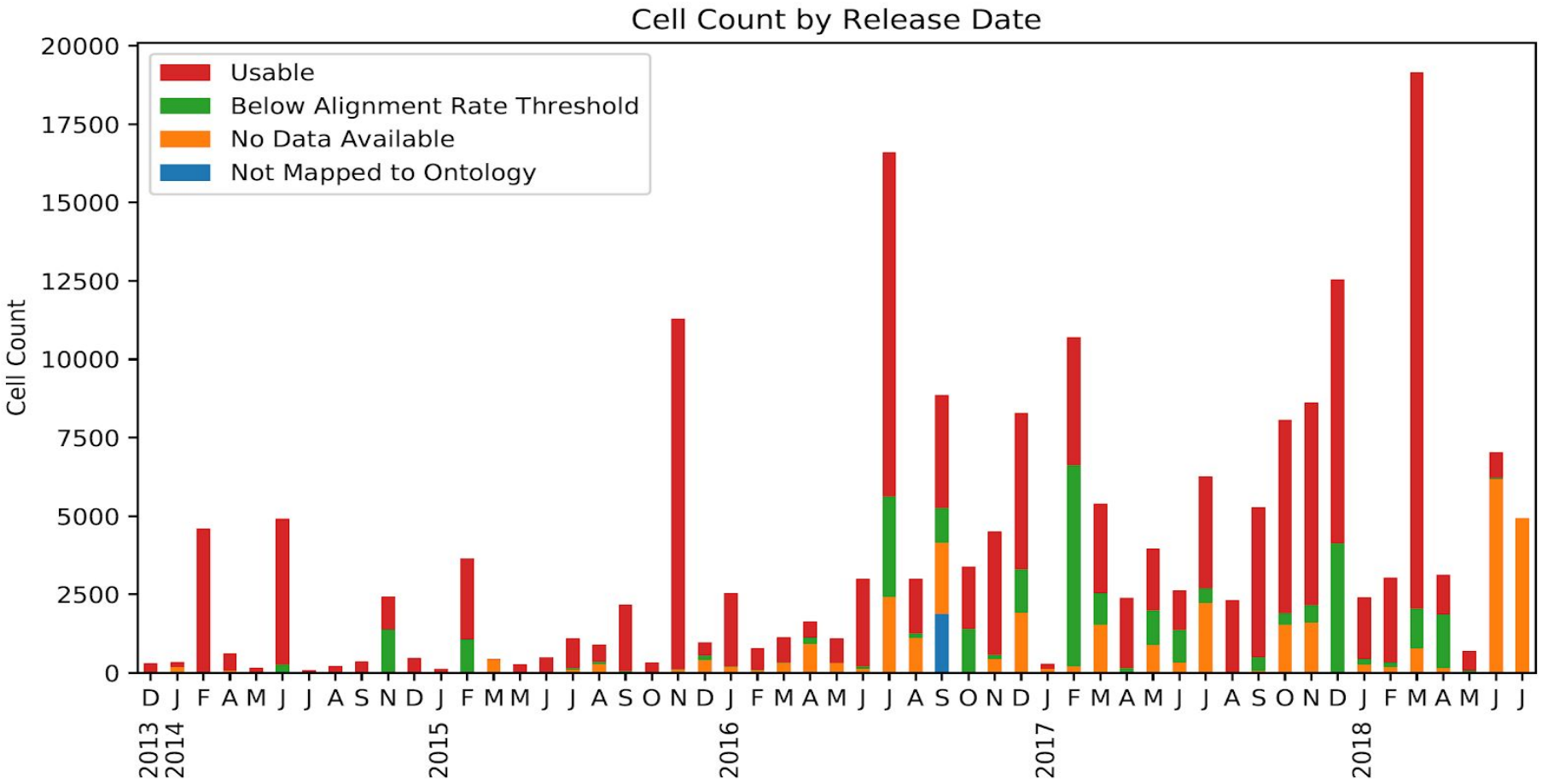


HuBMAP



Human Cell ATLAS

Single-cell data is accumulating fast



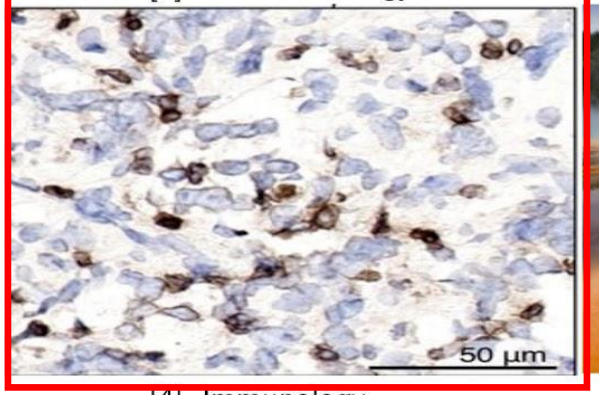
Single-cell vs. Bulk Sequencing

Single-cell

Bulk

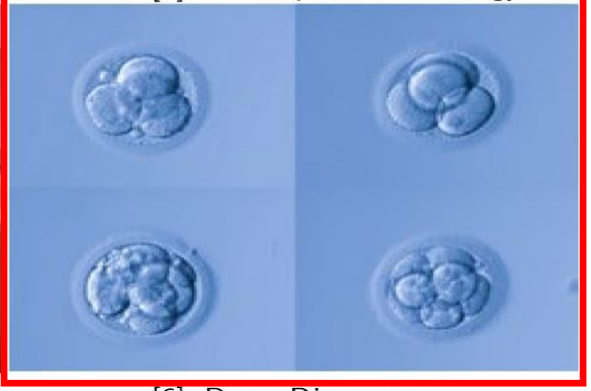
Single-cell applications in biomedical studies

[1] Cancer Biology

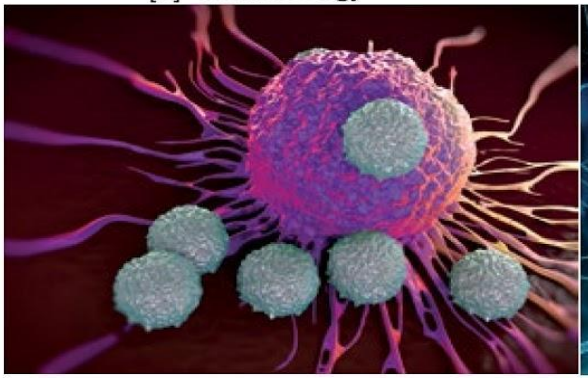


[2] Metagenomics

[3] Developmental Biology

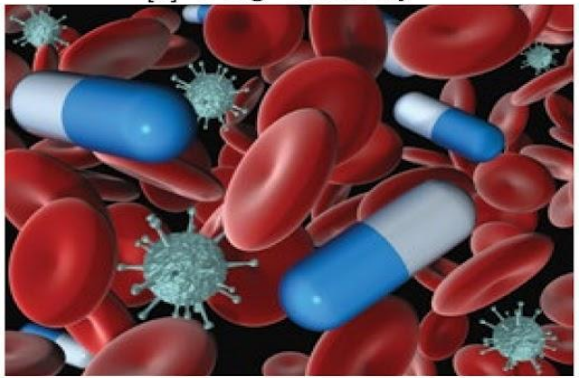


[4] Immunology



[5] Neurobiology

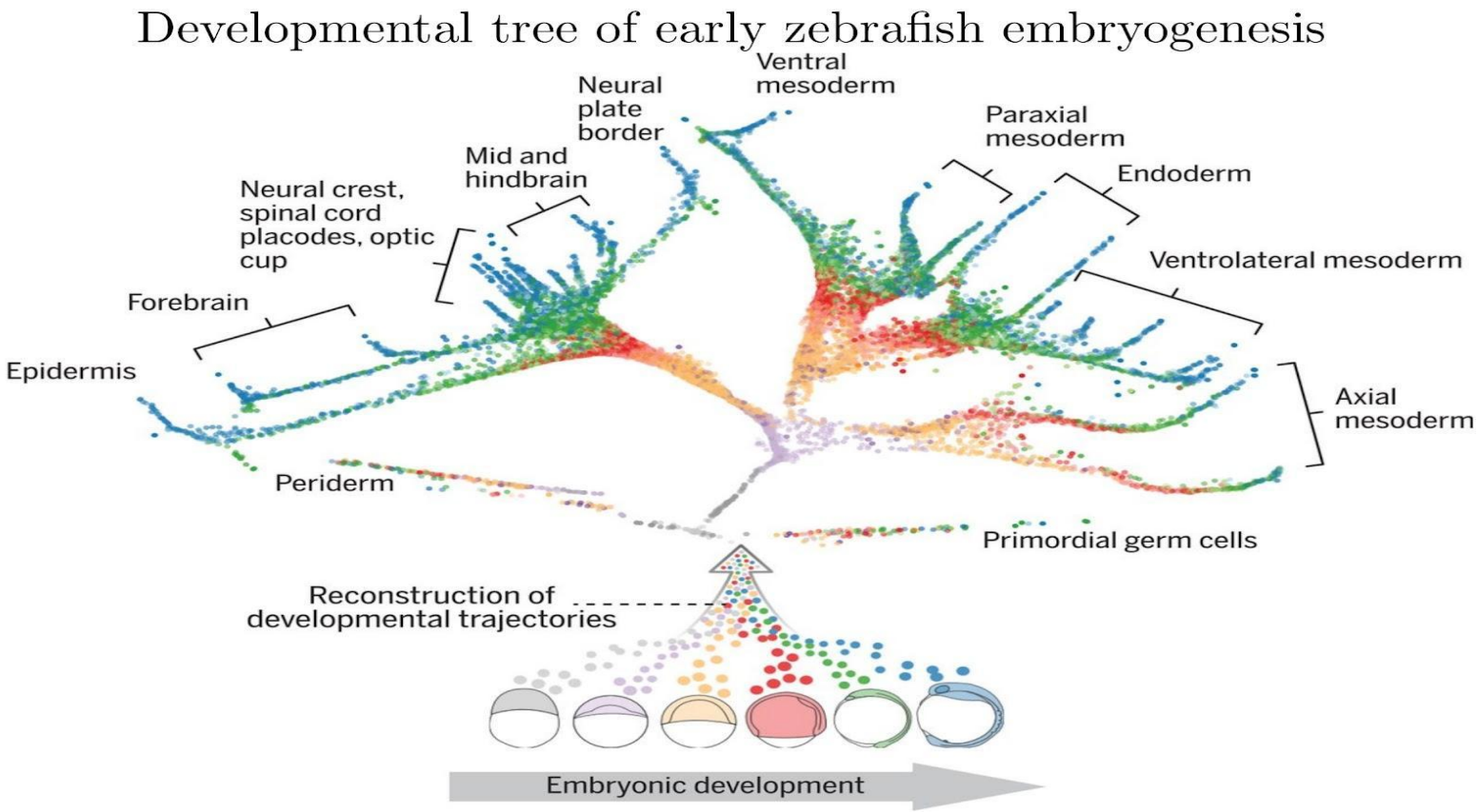
[6] Drug Discovery



[1] Engblom et al. Science. 2017 [2] Fadrosh et al. Nat Commun. 2016 [3] Treutlein et al. Nature. 2014
[4] Zheng et al. Cell. 2017 [5] Quadrato et al. Nature, 2017 [6] Heath et al. Nat Rev Drug Discov. 2016
All images retrieved from Illumina websites

Developmental Trajectory

Developmental trajectory inference methods



Farrell J. et al. Science,

Machine learning challenges?

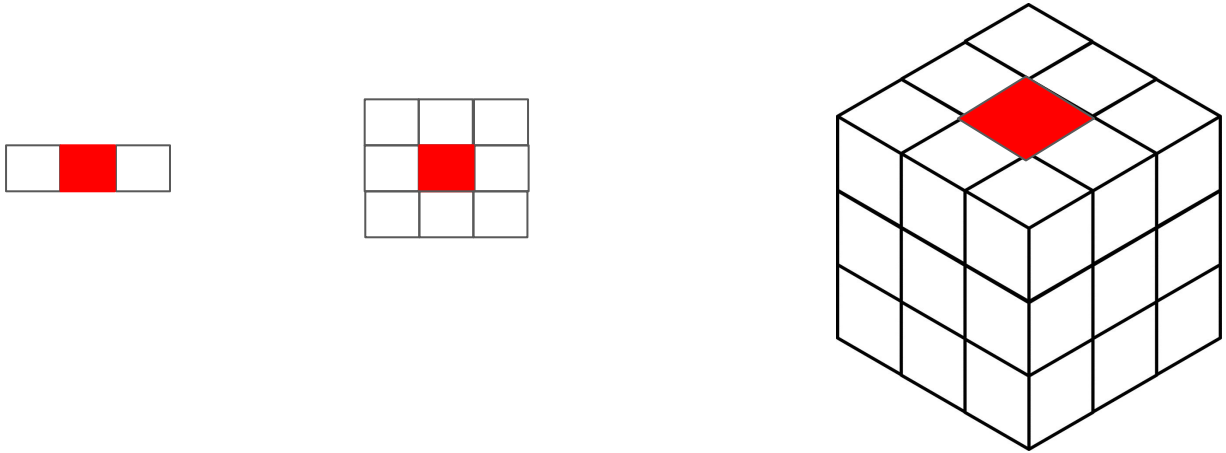
- ★ Curse of dimensionality
- ★ High noise level
- ★ Enormous heterogeneity

⇒ specific computational challenges:

- ❖ Reducing the data dimensionality
- ❖ Identifying sub-population (clustering problems)
- ❖ Reconstructing the cellular trajectories

Curse of dimensionality

Analyzing of the high dimensional data often suffers from the curse of dimensionality



The searching space increases exponentially
Neighbors of each data point also increase exponentially
Distances are on longer informative

$$\lim_{d\to\infty} E\left(\frac{\mathrm{dist}_{\max}(d) - \mathrm{dist}_{\min}(d)}{\mathrm{dist}_{\min}(d)}\right) \to 0.$$

C1: Most importantly, human eyes can't see anything beyond 3D

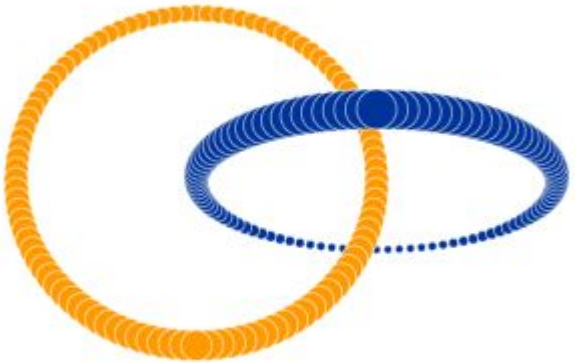
Suppose your boss gives you a single-cell dataset (10k cells by 20k genes), and told you that he wants to see what it looks like.

What is your first thought?

$$\arg\min_F |x - F(x)|$$

Dimensionality reduction

- Linear F
  PCA
- Non-linear F
  t-SNE (U-MAP)
  Auto-encoder



Non-linearly separable data

PCA

https://s3-us-west-2.amazonaws.com/articles-dimred/pca/animation.webm

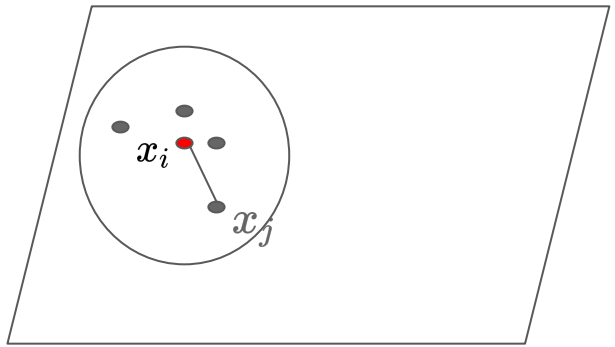
Find a linear transformation to project the data from HD to LD space that minimize the projection error.

$$\sum_{i=1}^{n} ||x_i - Px_i||^2$$

P represents the transformation matrix

t-SNE

1) Measuring the distance in higher dimensional space (Gaussian distribution)



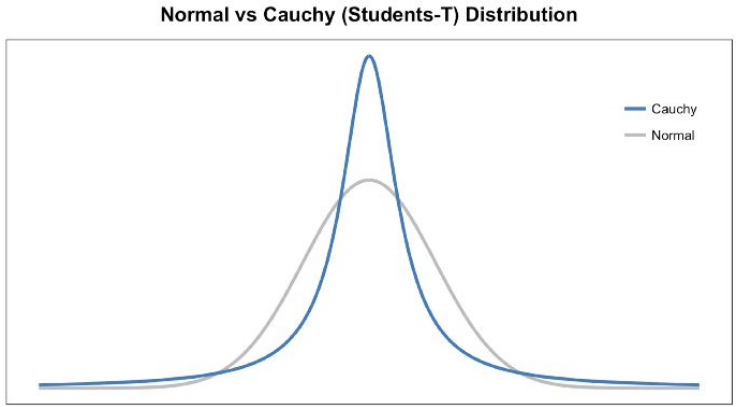
$$p_{j\,|\,i} = \frac{\exp(-||x_i - x_j||^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2 / 2\sigma^2)}$$

$$p_{i,j} = \frac{p_{i\,|\,j} + p_{j\,|\,i}}{2N}$$

$$\sum_{i,j} p_{i,j} = 1$$

t-SNE

2) Measuring the distance in lower dimensional space (long-tail student t distribution)



$$q_{ij} = \frac{\left(1+||y_i - y_j||^2\right)^{-1}}{\sum_k \sum_{l \neq k} \left(1+||y_k - y_l||^2\right)^{-1}}$$

Why not using Gaussian distribution in LD too?

The Gaussian distribution in 2D space will force all time points “together” => crowding problem

This can be mitigated by the “long-tail” student t-distribution

t-SNE

(3) The locations of the points in the LD space (y) are determined by minimizing the (non-symmetric) Kullback–Leibler divergence of the distribution P from the distribution Q.

$$C = KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Then use the gradient descent to search the y_i that minimize the KL divergence C.

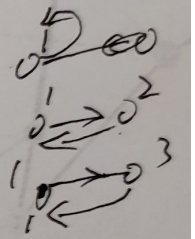
Diffusion map

1) Calculate the transition probability matrix M(i,j) (e.g., base on the distance and a chosen kernel).
2) Diffusion $M^t(i,j)$

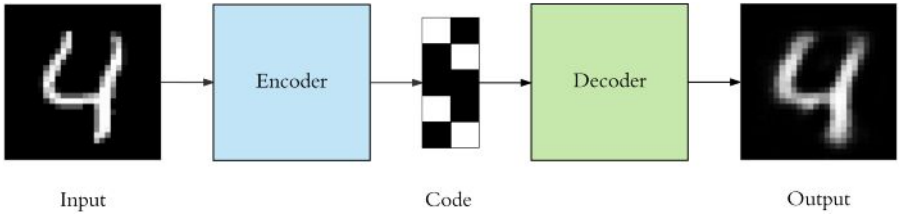
$$M^2 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

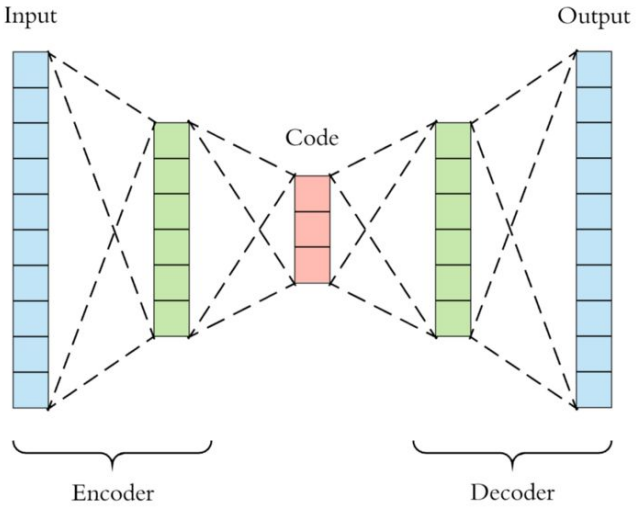
$$\Rightarrow M^2(1,1) = a_{11}a_{11} + a_{12}a_{21} + a_{13}a_{31}$$

$$M^2(1,2) = a_{11}a_{12} + a_{12}a_{22} + a_{13} + a_{32}$$



Autoencoder





$$\phi : \mathcal{X} \to \mathcal{F}$$
$$\psi : \mathcal{F} \to \mathcal{X}$$
$$\phi, \psi = \arg\min_{\phi,\psi} \|X - (\psi \circ \phi)X\|^2$$

RECAP 1

1) Dimensionality reduction techniques are commonly used in single-cell genomics
2) Popular techniques:
   Linear: PCA (Linear),
   Non-linear: Classical: t-SNE/UMAP, Neuron network: Autoencoder
3) It is usually the first step of the high-dimensional data analysis

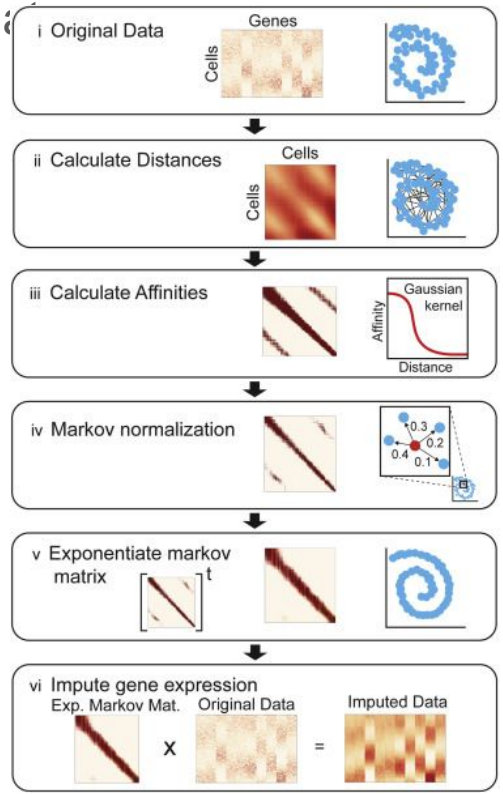
C2. Single-cell data is very noisy

The single-cell dataset (10k cells by 20k genes) data is very noisy

You want to “fix” or “clean” the data, what would you do?

Single-cell data “fixing”

MAGIC[1] is a popular method to fix the missing values (e.g., dropout) in single-cell da



$M^t(i,j)$

represents the probability that a random walk of length *t* starting at cell *i* will reach cell *j*, thus we call *t* the “diffusion time.”

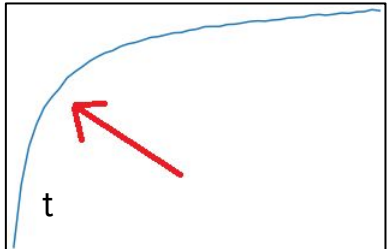
$$A(i,j) = e^{-\left(\frac{Dist(i,j)}{\sigma}\right)^2}$$

$$M(i,j) = \frac{A(i,j)}{\sum_k A(i,k)}$$

Markov transition probability from i -> j

$$D_{imputed} = M^t * D$$

R^2



[1] Dana Pe’er et al. Cell, 2018

Single-cell data denoising

Guess what is the most commonly used track?

=> remove the “bad cells”

How?

1) Remove cells with low # of expressing genes
2) Remove cells with high % of mitochondrial reads

RECAP 2

1) Single-cell imputation (e.g., MAGIC ->data fixing)
2) Single-cell data cleaning (e.g., filtering -> denoising)
3) Garbage in => Garbage out
   Always try to clean the data first before the actual modeling/analysis

C3. Single-cell data is enormously heterogeneous

The single-cell dataset (10k cells by 20k genes) data is **heterogeneous**
You want to identify all sub-populations, what would you do?

Clustering

K-Means

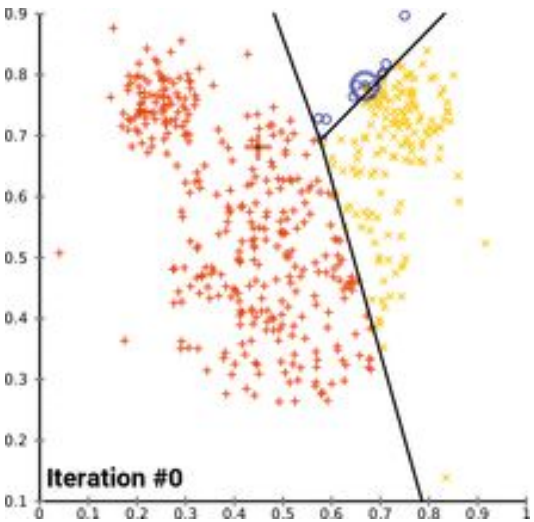
Assignment:

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \leq j \leq k\},$$

Assign the cell to the closest cluster (nearest centroid).

Update:

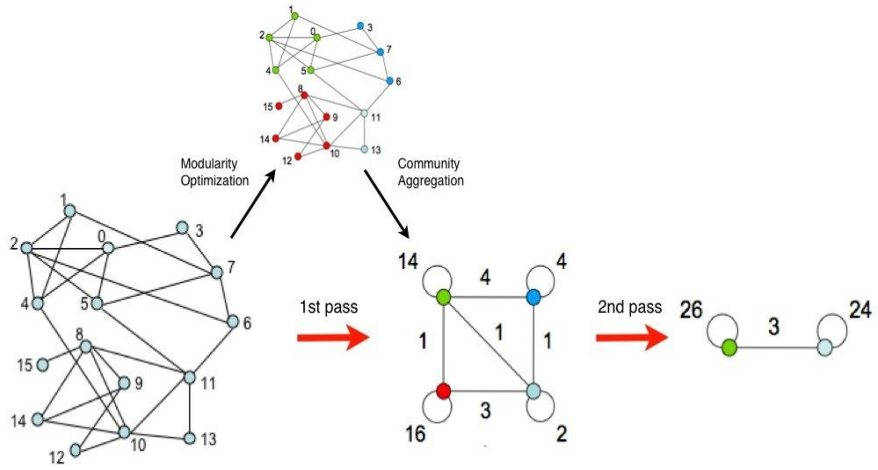
$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$



From Wiki page

Louvian/Leiden

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$



where

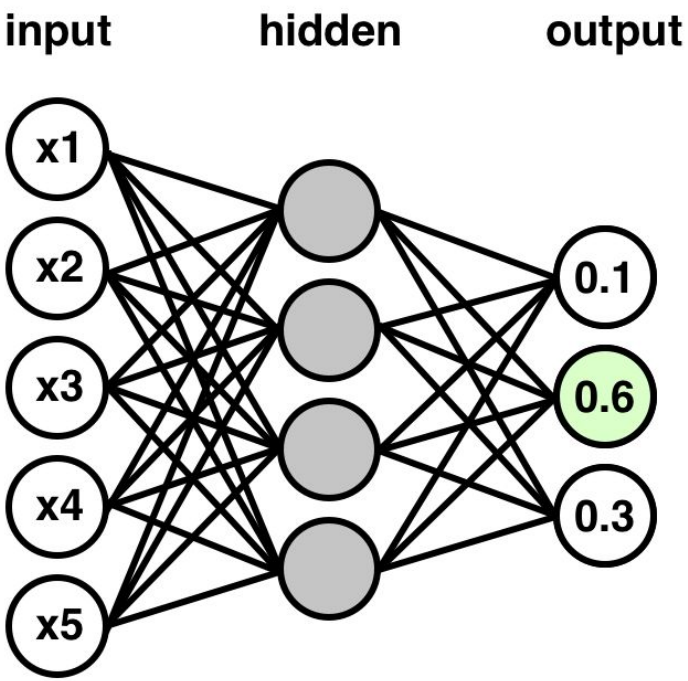
Aij is the weight of the edge between i and j.
ki is the sum of weights of the vertex attached to the vertex I, also called as degree of the node
ci is the community to which vertex i is assigned
δ(x,y) is 1 if x = y and 0 otherwise
m = (1/2)$\sum_{ij}$ Aij i.e number of links

1st step: A greedy algorithm is applied to search for the maximal Q (moving a node from community i to all its neighbors) => guarantee a local optical.

2nd step: update the weight between communities.

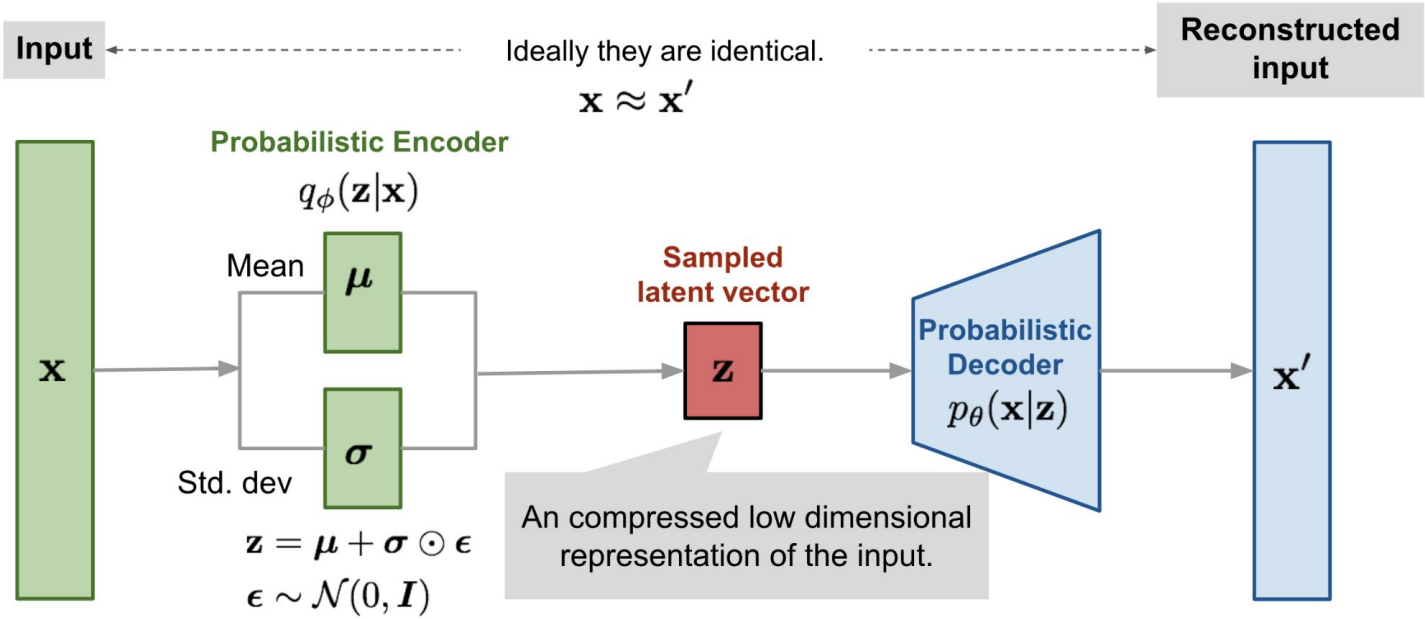
Such passes are repeatedly carried out until there is no more change in the cluster, and a maximum of modularity is achieved.

Supervised neural network for clustering



$$Loss = \sum_{i \in D} ||y_i - F(x_i)||^2$$

Variational autoencoder



How to annotate cluster?

Now, you got the clusters.
But, what are those clusters? (e.g., what cell types they are? What set of genes they are expressing)

OPEN QUESTION

A few existing solution:

(1) Use marker genes
(2) Use functional analysis (e.g., GO enrichment)
(3) Compare with expression data with known cell types

RECAP 3

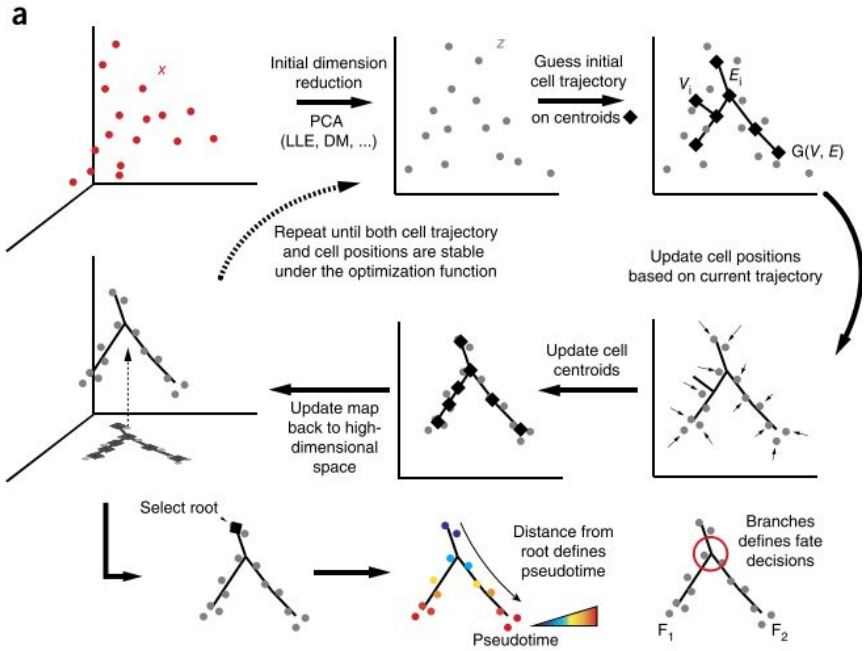
1) Clustering is the most widely used method to identify sub-populations
2) Popular methods: K-means, SOM, Louvian, Leiden, ANN (supervised)
3) No good ways to annotate clusters yet.

C4. Reconstructing trajectories from Single-cell data

How to infer the cell dynamics (the cellular state change over time) from single-cell data (often time-series)

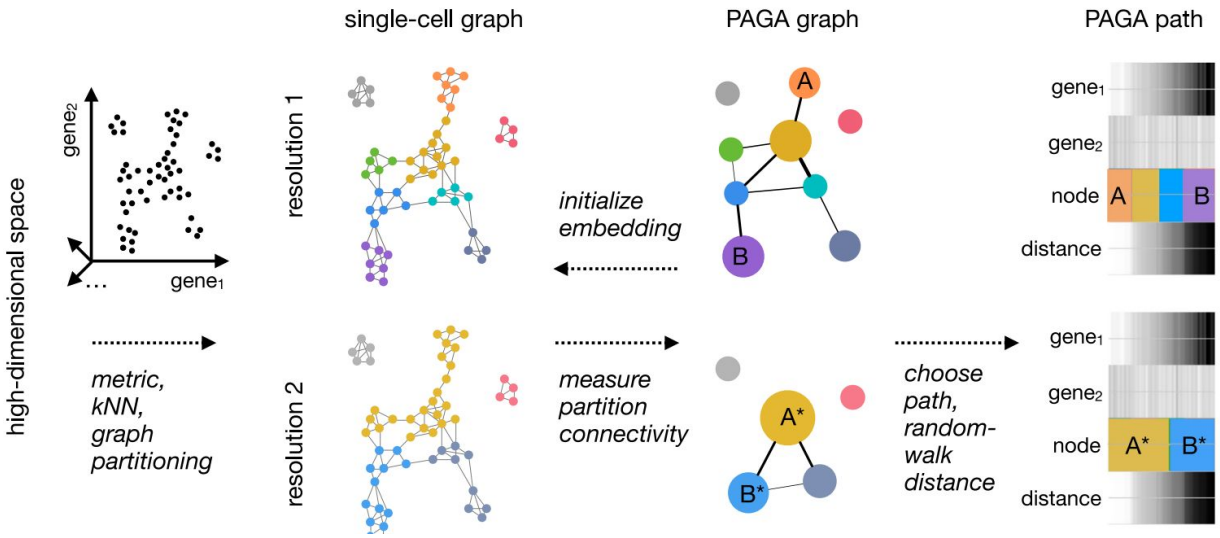
Monocle

$$\min_{\mathcal{G}\in\mathcal{G}_b}\ \min_{f_{\mathcal{G}}\in\mathcal{F}}\ \min_{Z}\ \sum_{i=1}^{N}\|\boldsymbol{x}_i - f_{\mathcal{G}}(\boldsymbol{z}_i)\|^2 + \frac{\lambda}{2}\sum_{(V_i,V_j)\in\varepsilon} b_{i,j}\,\|f_{\mathcal{G}}(\boldsymbol{z}_i) - f_{\mathcal{G}}(\boldsymbol{z}_j)\|^2$$



Qiu et al. Nature Methods, 2017

PAGA



1) Graph partitioning and abstraction
2) Pseudo-time estimation
3) Preserving Graph topology across resolutions

Wolf et al. Genome Biology, 2019

RECAP 4

1) Trajectory methods are employed to interrogate the dynamic cellular transition
2) Popular methods: Monocle, Seurat, etc.

C5. Reconstructing the regulatory networks underlying the trajectories

How to infer the transcription factors and pathways that dictate the cellular dynamics

GENIE3

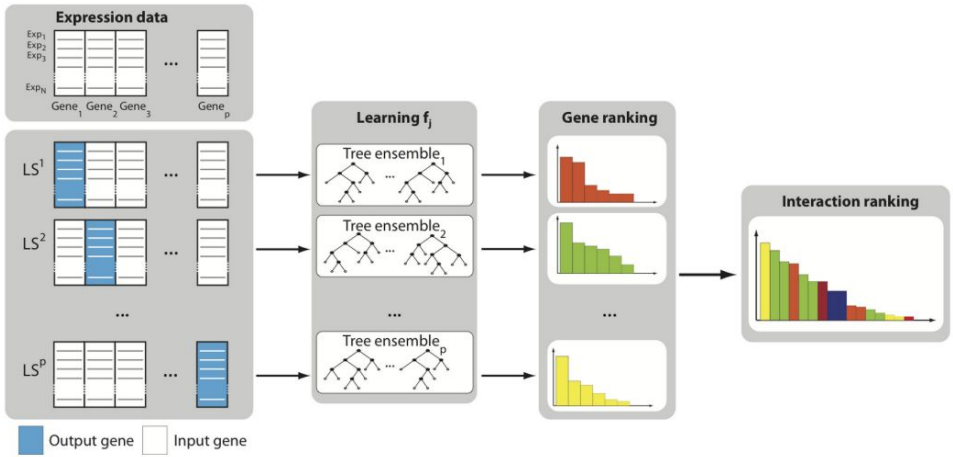
$$\mathbf{x}_k^{-j} = (x_k^1, \ldots, x_k^{j-1}, x_k^{j+1}, \ldots, x_k^p)^T,$$

we assume that we can write:

$$x_k^j = f_j(\mathbf{x}_k^{-j}) + \varepsilon_k, \forall k$$

Find f to minimize

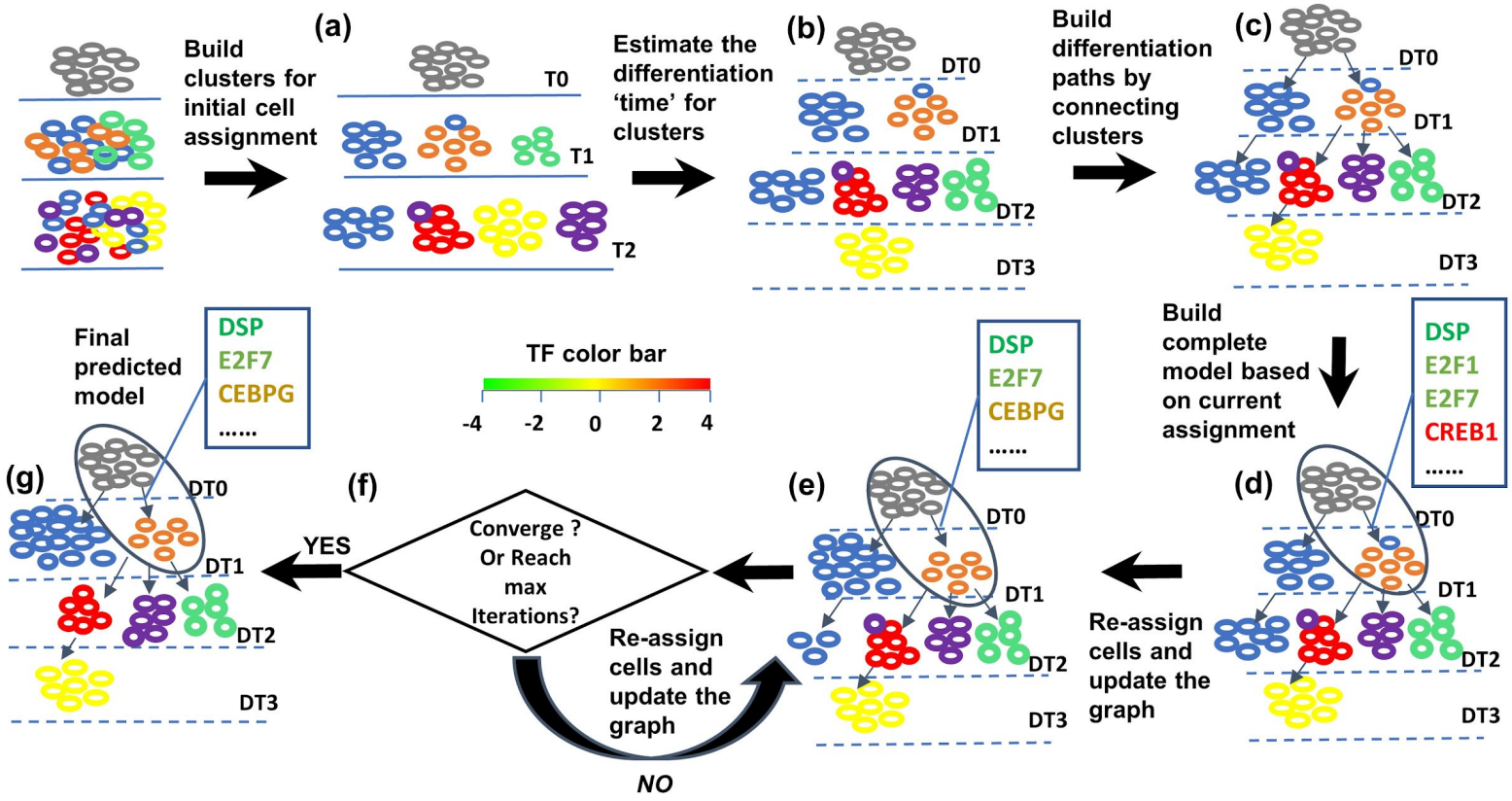
$$\sum_{k=1}^{N} (x_k^j - f_j(\mathbf{x}_k^{-j}))^2$$



SIMPLE BUT POWERFUL (Champion of the DREAM Challenge)

Huynh-Thu et al. Plos one, 2010

SCDIFF



Ding et al. Genome Research 2019

RECAP 4

1) Trajectory methods are employed to interrogate the dynamic cellular transition
2) Popular methods: Monocle, Seurat, etc.

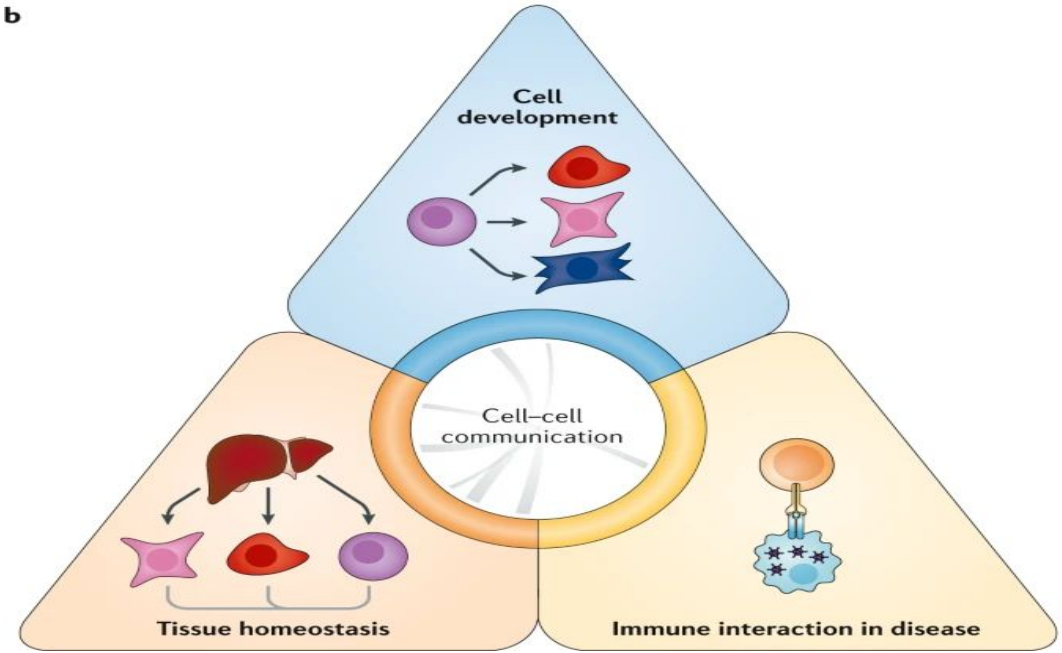
OPEN DISCUSSION

Any new ideas to reduce the data dimensionality?
Any new strategies to cluster the data points?
Any new methods to infer the gene regulatory network?

Infer cell-cell interactions from pseudotime ordering of single-cell data

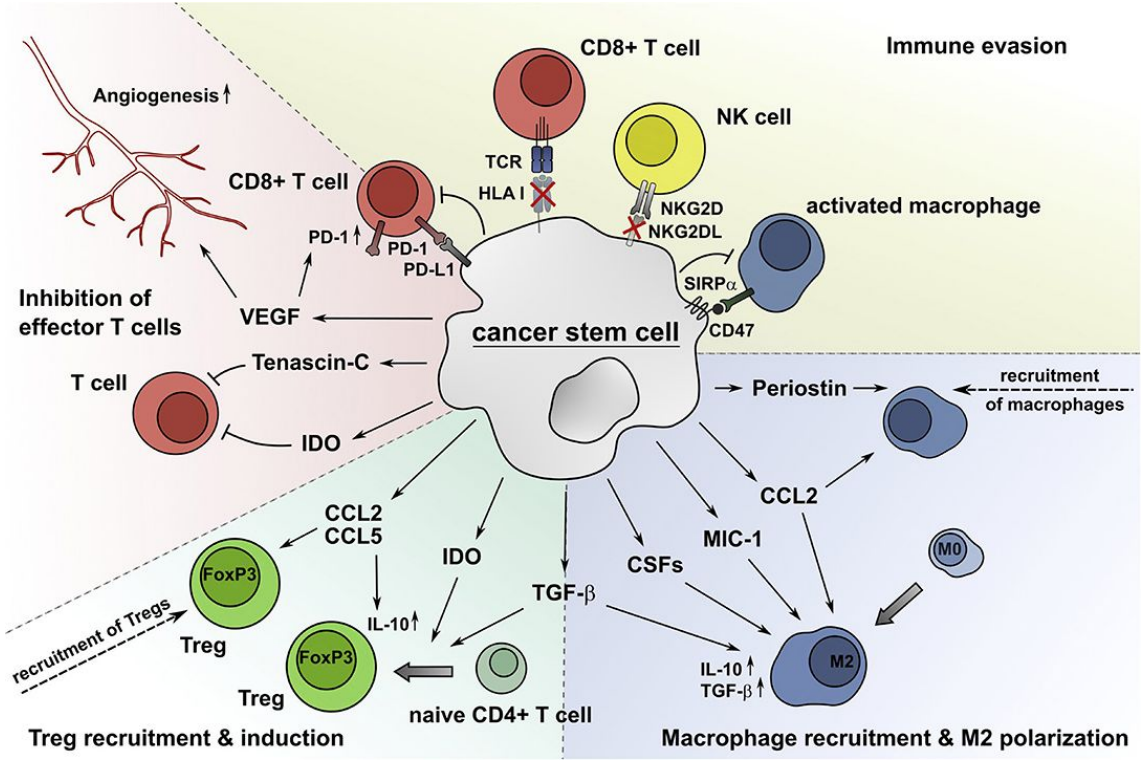
Jun Ding
Assistant professor
Department of Medicine
Department of Biomedical Engineering
McGill University

Why cell-cell interaction inference matters?



Armingol, Erick, et al. "Deciphering cell–cell interactions and communication from gene expression." *Nature Reviews Genetics* 22.2 (2021): 71-88.

Cell-cell interactions play critical roles in cancer progression



Müller, Luise, et al. "Bidirectional crosstalk between cancer stem cells and immune cell subsets." *Frontiers in immunology* (2020): 140.

Existing methods?

Most existing methods are based on Expression thresholding

Sample or organ	Key input	Scoring	CS value	CCC score	Validation	Study focus	Ref
Cell development							
Haematopoietic cells (human)	Microarray; LRIs	Expression thresholding	Binary	No score	Functional validation	Role of CCC between differentiated haematopoietic cells and HSCs in fate decisions	15
Brain (mouse embryonic cortex)	Microarray; LRIs	Expression thresholding	Binary	No score	Functional validation	Role of microenvironment in self-renewal versus differentiation decision of neural precursor cells during neurogenesis	16
Liver and iPS cells (human)	scRNA-seq; LRIs	Expression thresholding	Binary	Normalized sum of CS	Functional validation	3D liver bud organoid from iPS cells to characterize CCC shaping hepatogenesis	24
Placenta (human)	scRNA-seq; LRIs	Expression thresholding	Binary	No score	Functional validation	CCI in the fetus–placenta interface before and after decidualization	29
Brain (mouse)	Bulk RNA-seq; LRIs	Expression thresholding	Binary	Sum of CS	Colocalization	Ligand–receptor pathways active during neural development; CCC between neural, vascular and microglial cells	22
iPS cells (mouse)	scRNA-seq; LRIs	Expression product	Continuous	Sum of CS	Functional validation	CCC at the beginning of differentiation	19
Bone marrow (mouse)	scRNA-seq; LRIs	RNA-Magnet	Continuous	No score	Colocalization	CCC and interactions between bone marrow cells	53
Tissue interactions							
Multiple lineages (human)	scRNA-seq; LRIs	Expression thresholding	Binary	Sum of CS	None	CCC between multiple cell lineages	7
Lungs (human)	popRNA-seq; LRIs	Expression thresholding	Binary	Sum of CS	Colocalization functional validation	Signals sent by mesenchymal cells in lungs that are key for self-renewal of epithelial progenitors after tissue injury	40
Heart (mouse)	scRNA-seq; LRIs	Expression thresholding	Binary	Sum of CS	Functional validation	Transcriptional profiles of non-myocyte cells in heart and their CCC	33
Lungs (mouse)	scRNA-seq; LRIs	Expression correlation (Spearman)	Continuous	No score	Expression; colocalization functional validation	CCC between and within immune and non-immune cells during development	41
Immune system and structural cells (mouse)	Low-input RNA-seq; LRIs	Differential combinations	Binary	Odds ratios	Functional validation	Role of structural cells in immune responses	58
Heart (mouse)	scRNA-seq; LRIs	Differential combinations	Binary	Sum of CS	Expression; colocalization functional validation	CCC of cardiomyocytes and non-cardiomyocytes in human heart in health and under failure	34
Placenta (human)	scRNA-seq; LRIs	CellPhoneDB	Continuous	No score	Colocalization	Key ligand–receptor pairs based on subunit architecture; CCC at maternal–fetal interface	30
Tumour microenvironment							
Melanoma (human)	scRNA-seq; LRIs	Expression thresholding	Binary	Sum of CS	None	CCI network of isolated cells	70
HNSCC (human)	scRNA-seq; LRIs	Expression thresholding	Binary	No score	Colocalization	CCC in patients with HNSCC generated by HPV or environmental carcinogens (HPV negative)	61
Five cancer types (mouse)	scRNA-seq; LRIs	Expression product	Continuous	No score	None	CCC within a tumoural microenvironment	60
Nine cancer types (human)	Microarray; LRIs	Expression correlation (Pearson)	Continuous	No score	None	Correlation between autocrine signalling pathways and mRNA levels of ligands and receptors	69
Lungs (human)	scRNA-seq; LRIs	Differential combination; expression thresholding	Binary	No score	Functional validation	Tumour–stroma CCC in lung cancer; introduced CCCExplorer	67
Ovary (human)	Microarray; LRIs; downstream target genes	Differential combination; expression thresholding	Binary	No score	Expression; functional validation	CCC between stromal and ovarian cancer cells	68
Head and neck and immune system (human)	scRNA-seq; LRIs; downstream target genes	NicheNet	Continuous	No score	None	Prediction of ligand–target links between interacting cells; tested on HNSCC data set	54

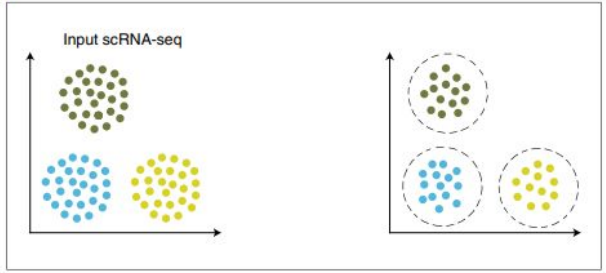
CellphoneDB

nature protocols
PROTOCOL
https://doi.org/10.1038/s41596-020-0292-x

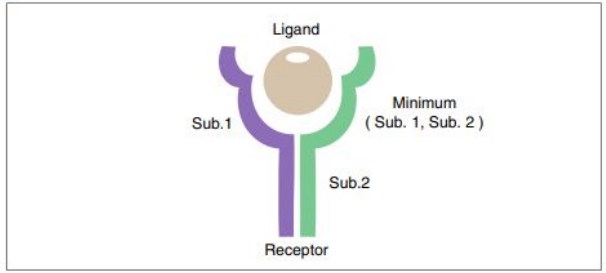
CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes

Mirjana Efremova[1], Miquel Vento-Tormo[2], Sarah A. Teichmann[1,3] and Roser Vento-Tormo[1*]

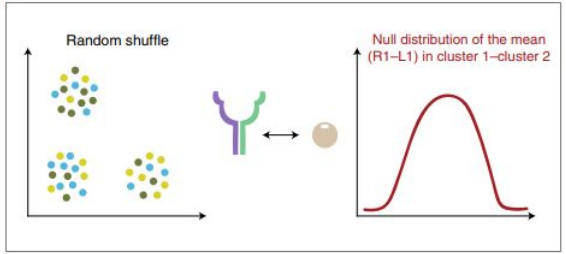
a



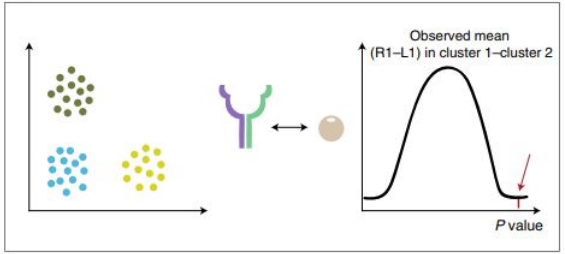
b



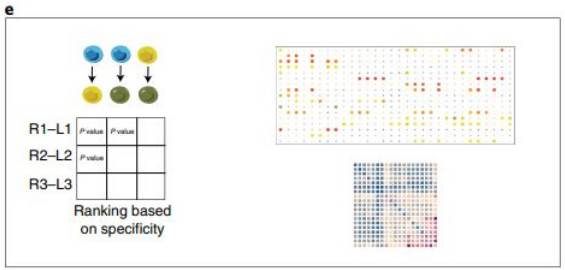
c



d



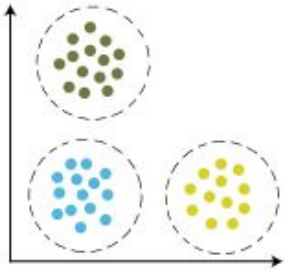
e

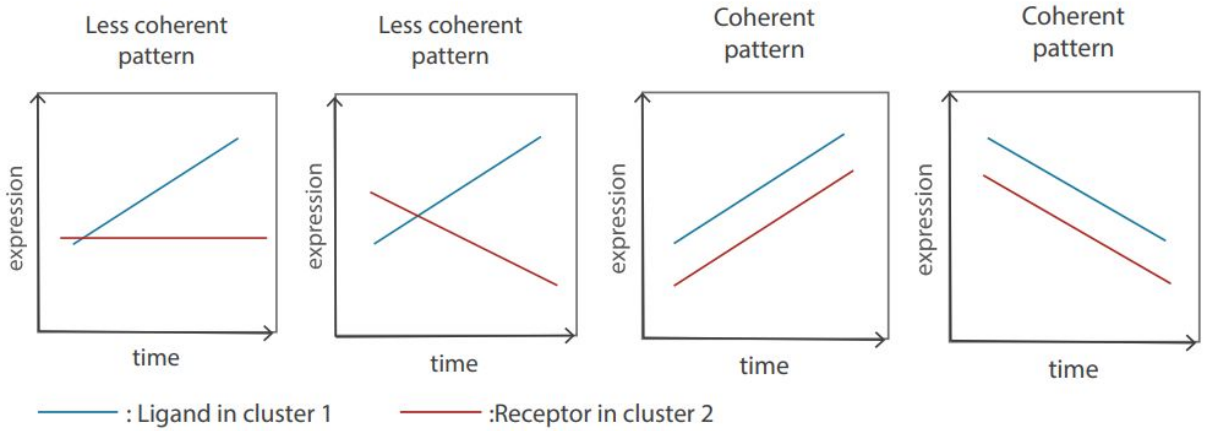


Limitations?

1) Information loss if only using the mean expression

2)Not all cells in the cluster are the same (Most biological processes are continuous)





Mean expression (ligand, receptor) based methods score the above 4 patterns the same. But, they are not!

How to address this limitation?

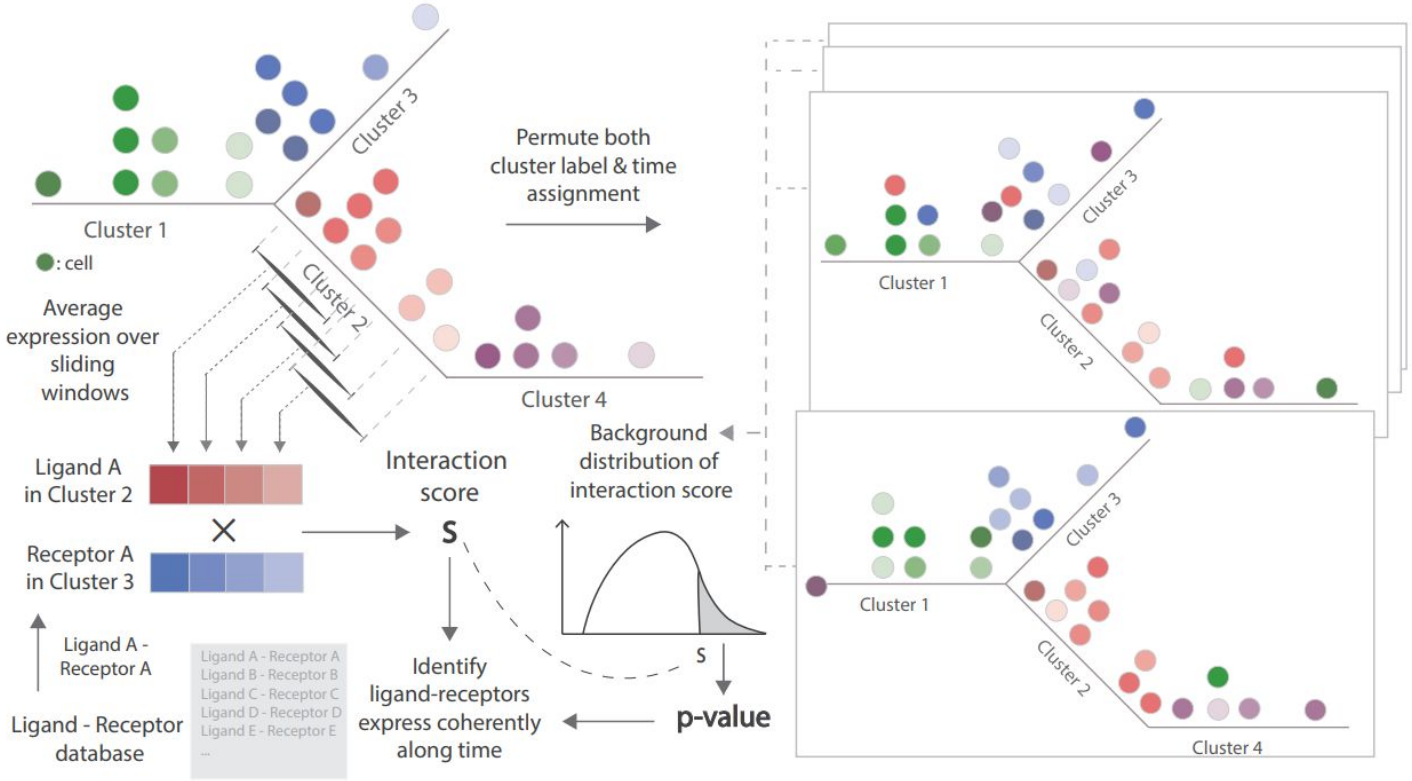
- Mean expression (Ligand, Receptor)=> two scalar value

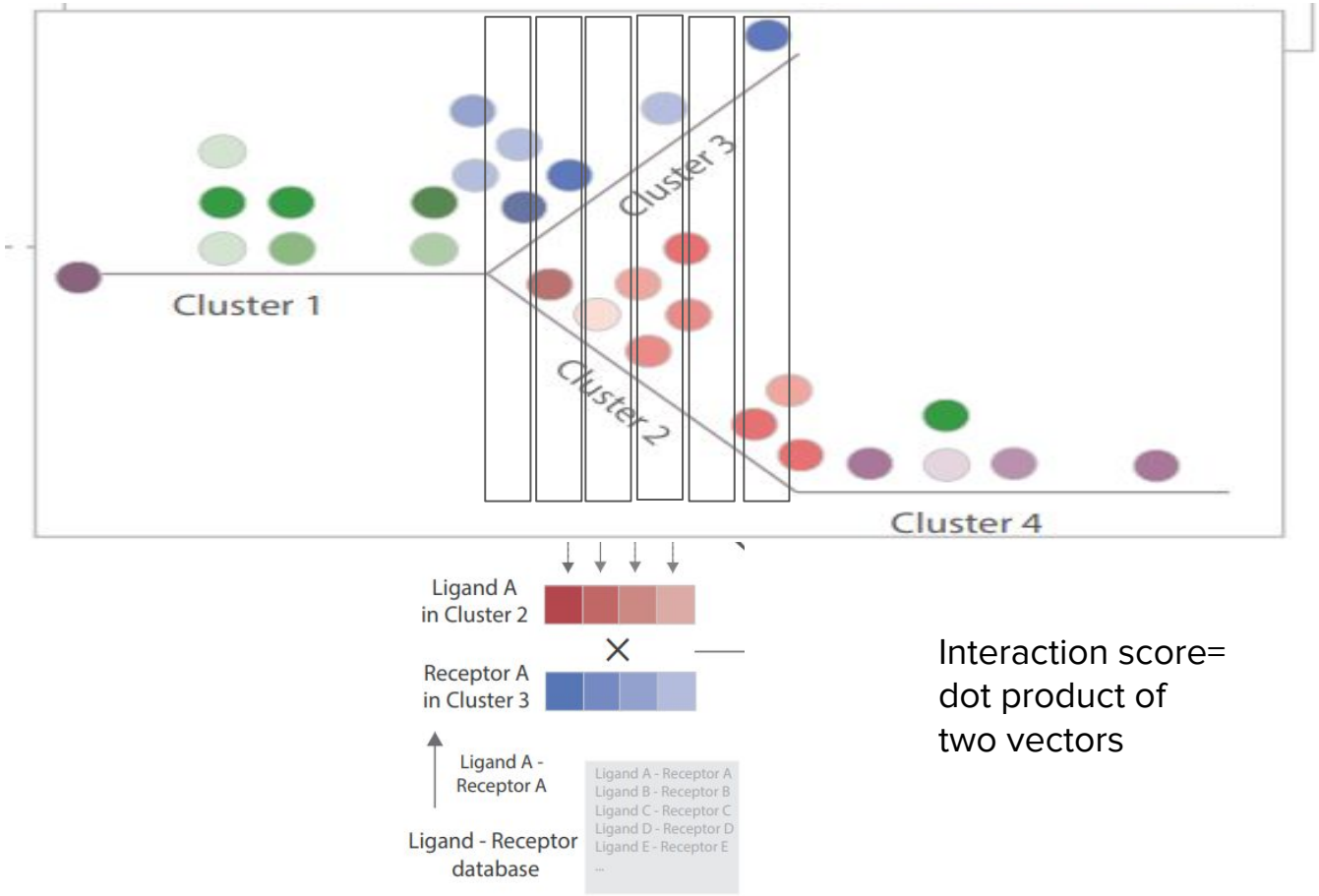
VS.

- Temporal expression (Ligand, Receptor) => two vectors

Solution: Integrate time/pseudo-time information with gene expression to infer Cell-cell interactions

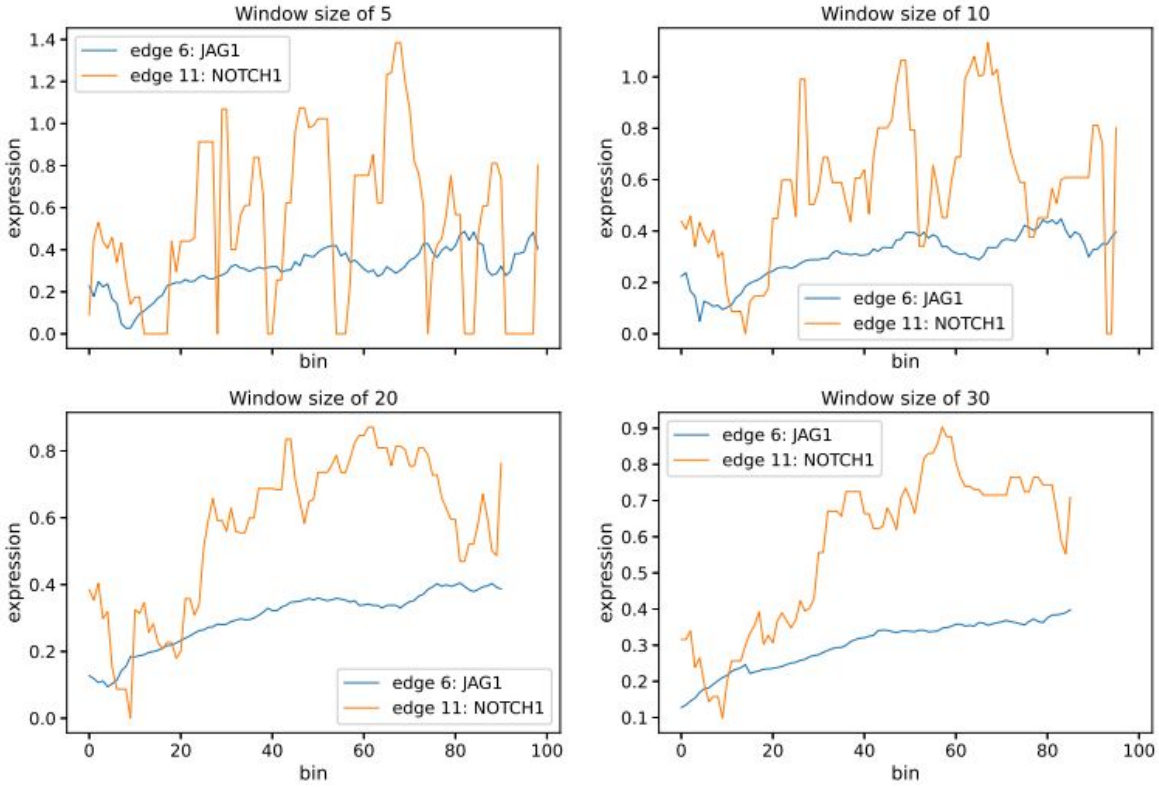
Trasig strategy



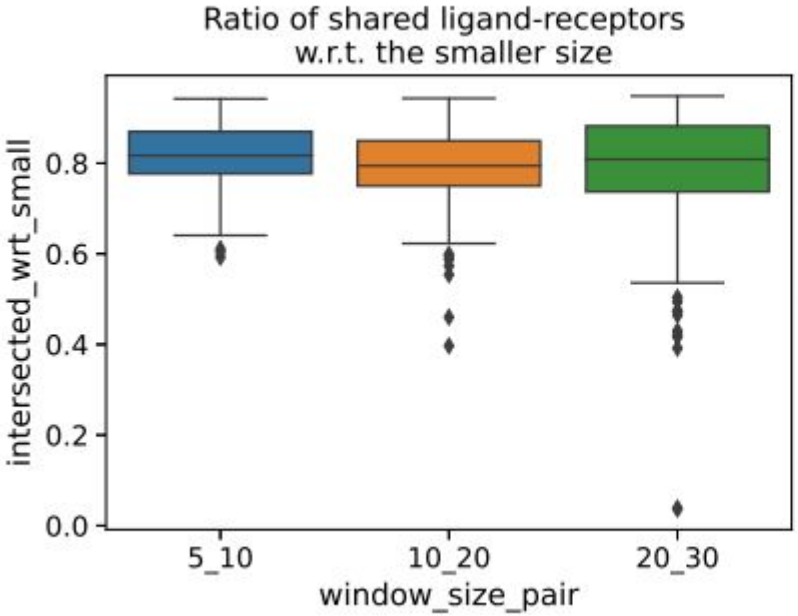


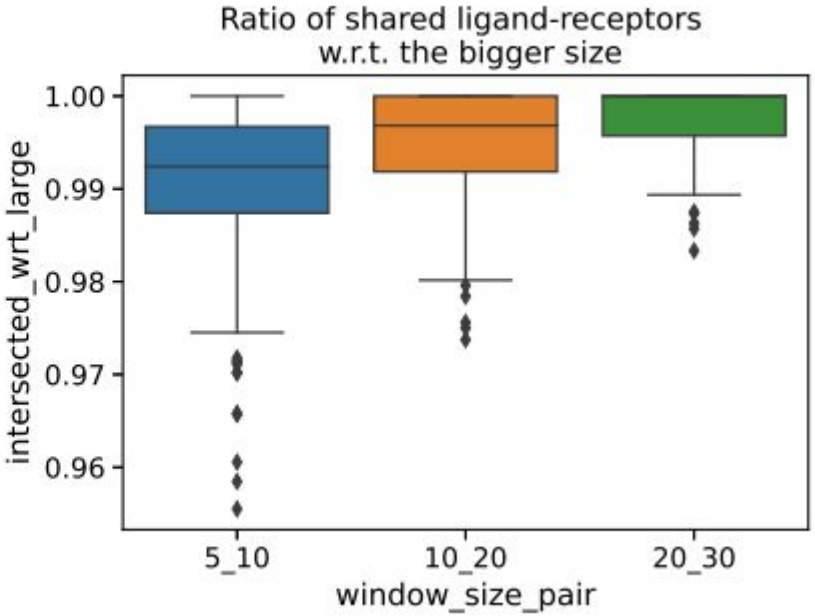
Interaction score=
dot product of
two vectors

Impact of different sizes of the sliding window

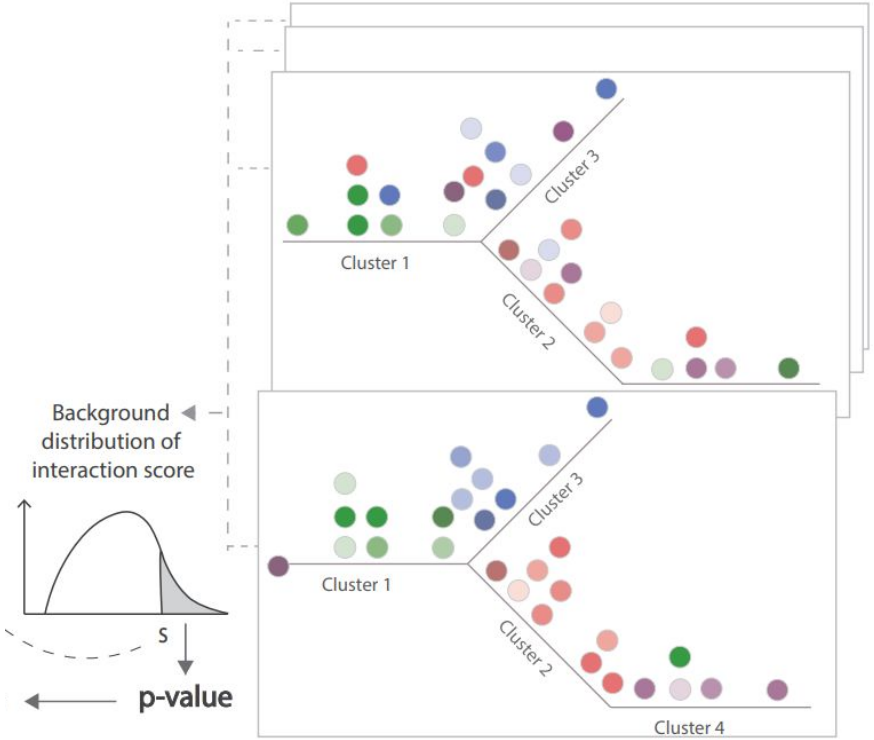


Window Size


Ratio of shared ligand-receptors
w.r.t. the smaller size
intersected_wrt_small
0.0
0.2
0.4
0.6
0.8
5_10
10_20
20_30
window_size_pair


Ratio of shared ligand-receptors
w.r.t. the bigger size
intersected_wrt_large
0.96
0.97
0.98
0.99
1.00
5_10
10_20
20_30
window_size_pair

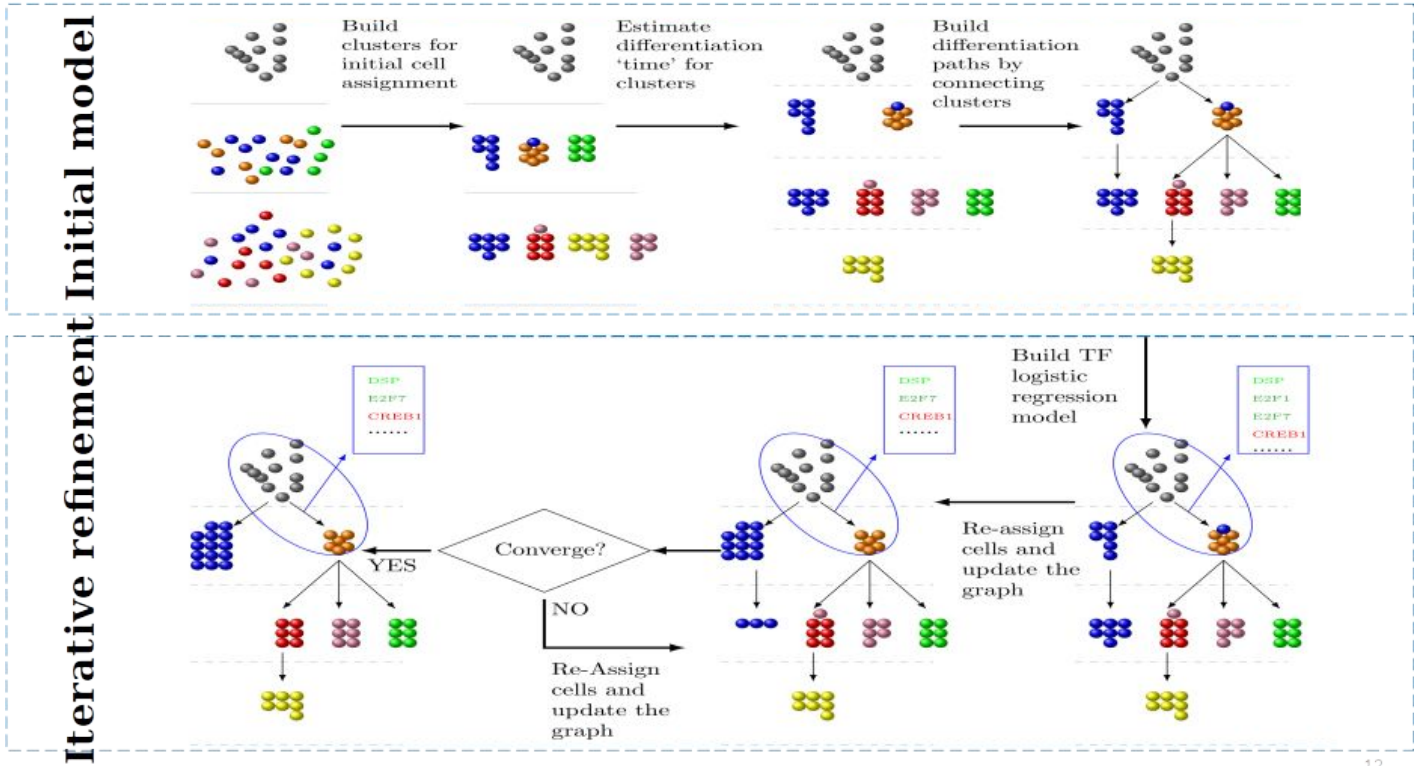
Calculating p-value for all L-R pairs



randomization => null distribution of interaction score

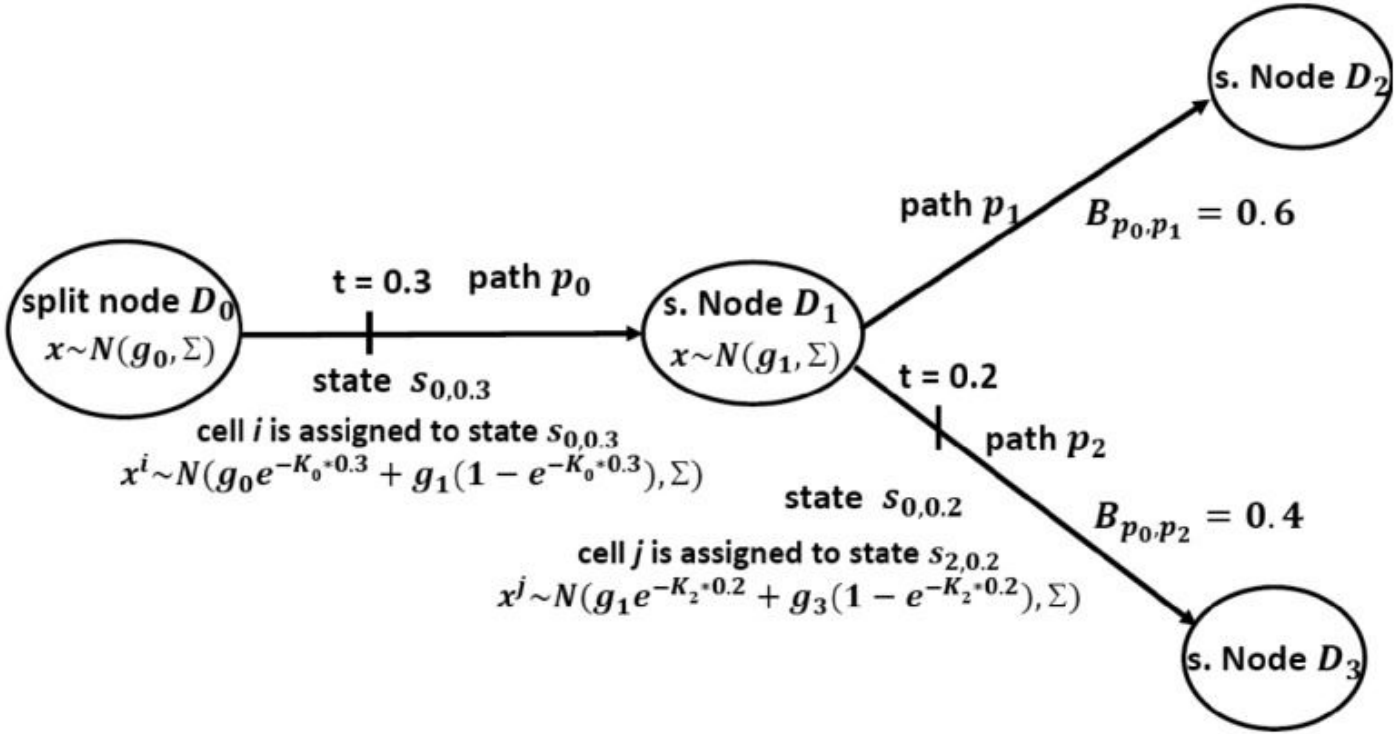
⇒ right-tail probability => p-value

How to infer the cellular trajectory? scdiff



Ding J et al. Genome Research, 2018

How to infer the cellular trajectory? cshmm



Lin, Chieh, and Ziv Bar-Joseph. "Continuous-state HMMs for modeling time-series single-cell RNA-Seq data." *Bioinformatics* 35.22 (2019): 4707-4715.

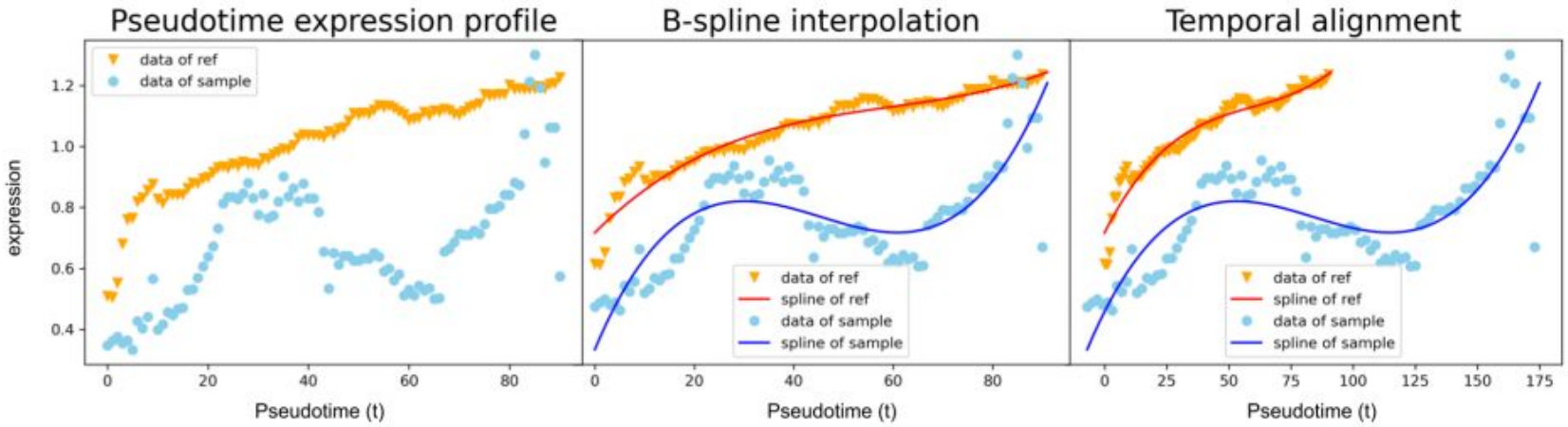
Selecting paired clusters

- Most other methods infer cell-cell interactions between all possible clusters
- Trasig: Cells can only interact if both are active at the same time

  For example, in a developmental process, cells at day E1 is unlikely to interact with cells profiled at day E16.

Temporal alignment

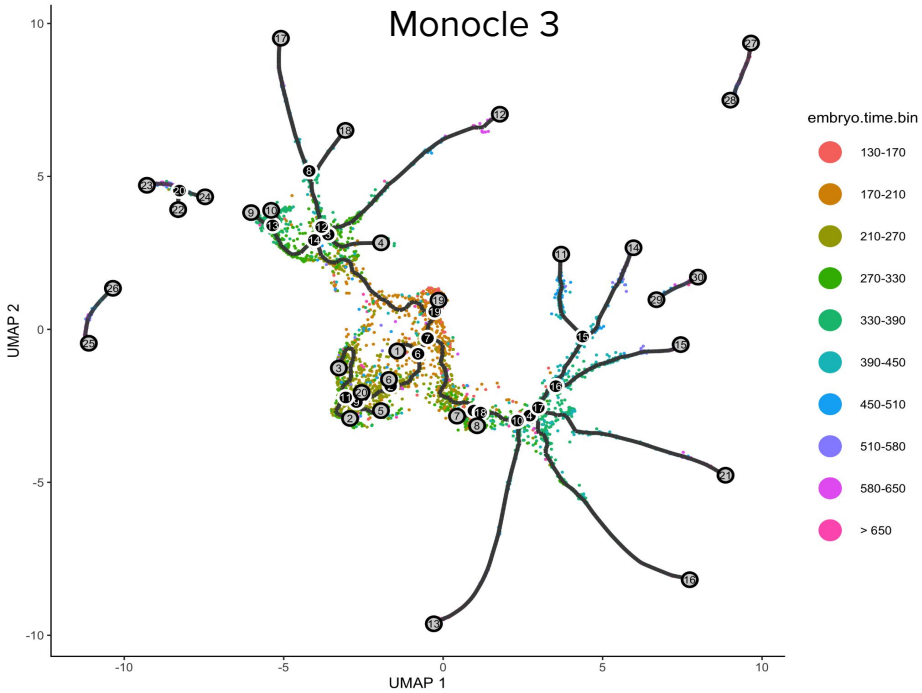
Is the time in each path of the interaction partners the same?
What if they are not? Alignment!

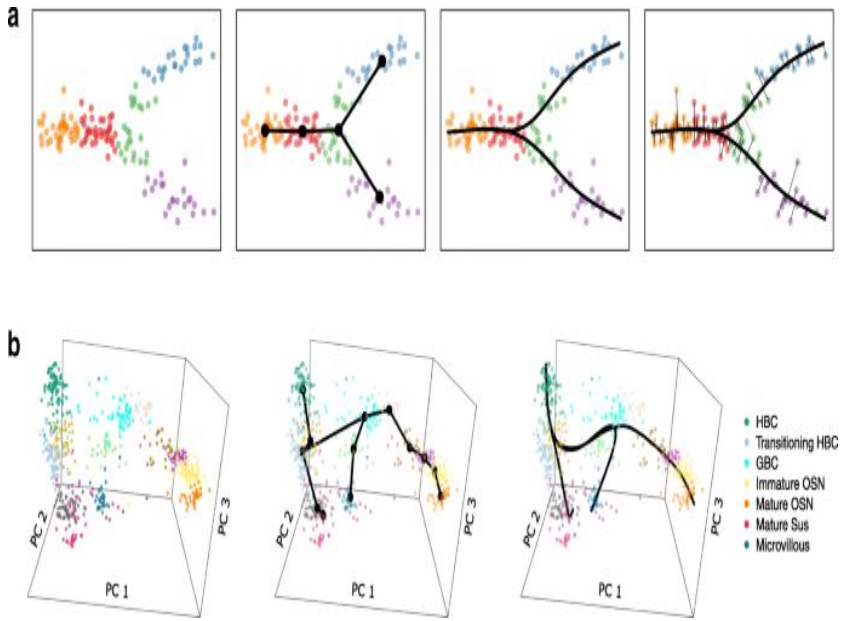


$$\tau_j(t) = \frac{(t-b_j)}{a_j}$$

Scaling and shifting

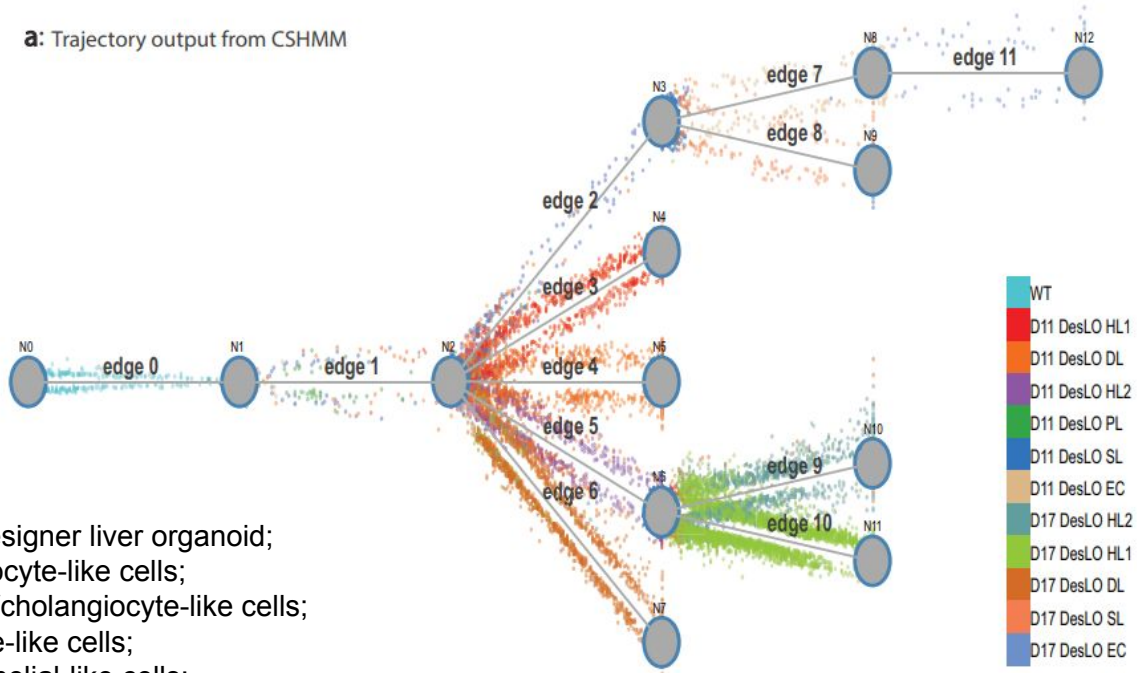
Trajectory and pseudotime inferred by other methods





Street, Kelly, et al. "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics." *BMC genomics* 19.1 (2018): 1-16.
Cao, Junyue, et al. "The single-cell transcriptional landscape of mammalian organogenesis." *Nature* 566.7745 (2019): 496-502.

Trajectory on liver organoid differentiation data



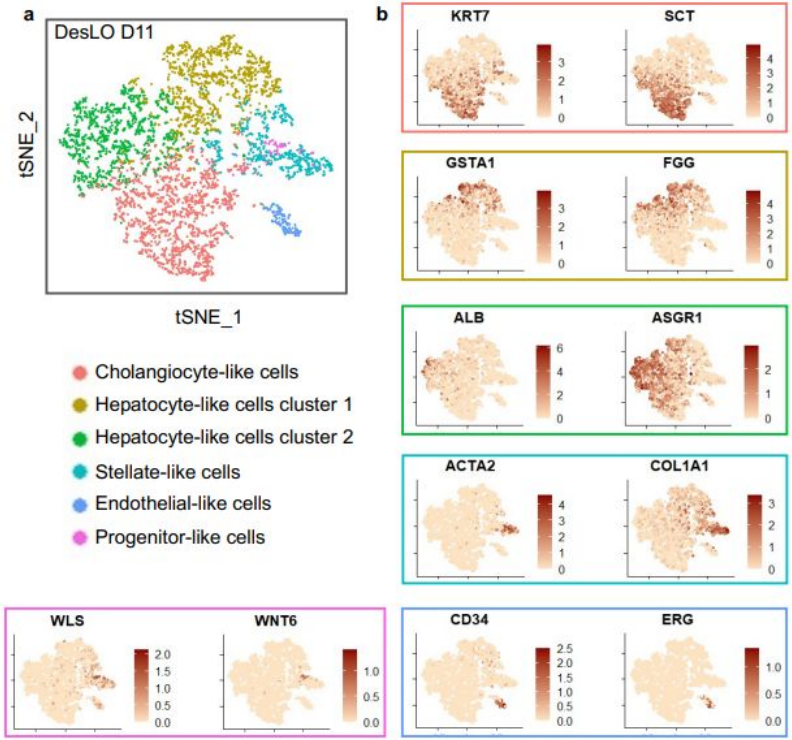
DesLO - designer liver organoid;
HL - hepatocyte-like cells;
DL - ductal/cholangiocyte-like cells;
SL - stellate-like cells;
EC - endothelial-like cells;
PL – progenitor-like,
WT: wide-type

Mo R. Ebrahimkhani, MD

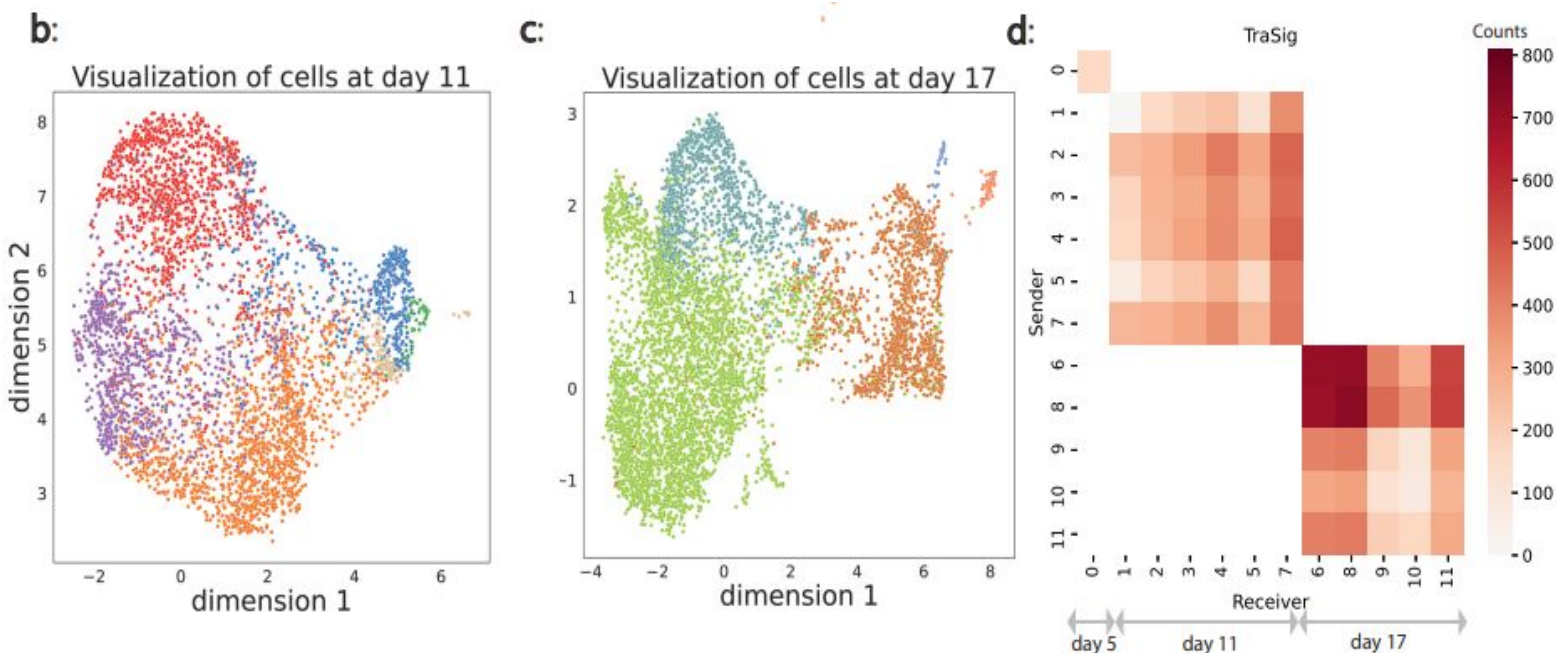
Associate professor

University of Pittsburgh

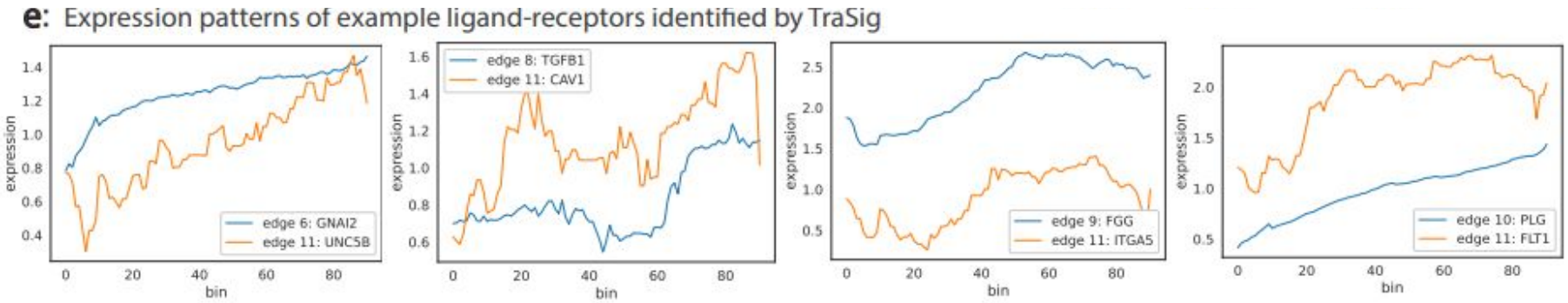
Cell type annotation



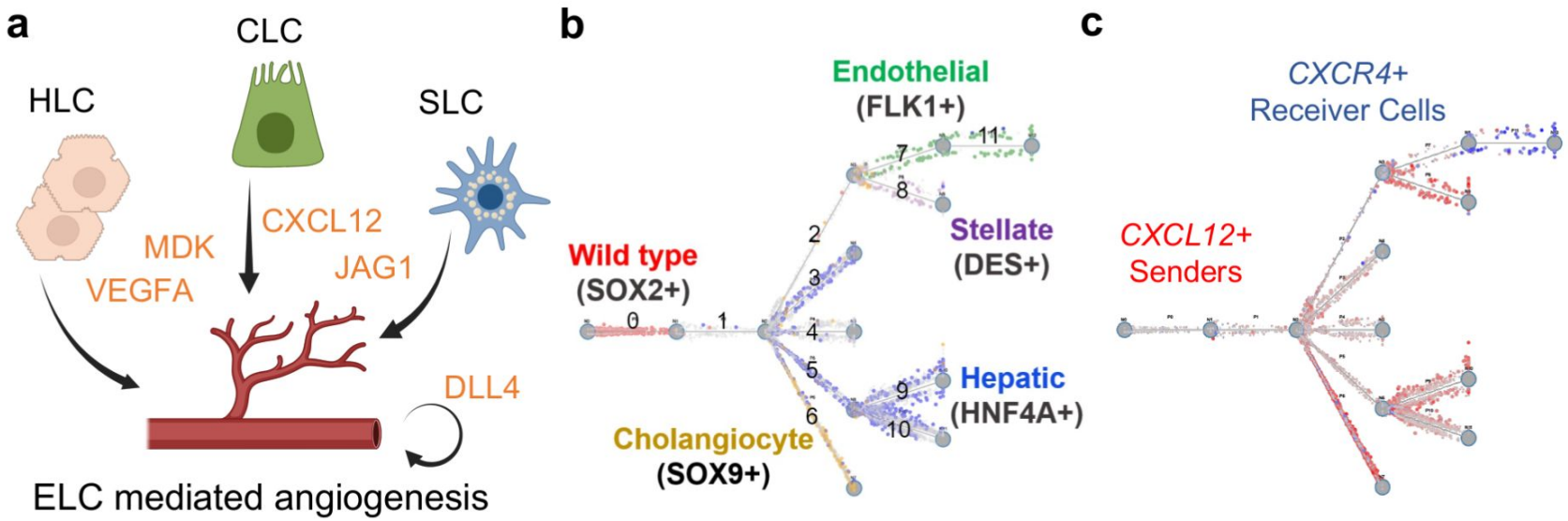
Cell-cell interactions for the liver organoid data



Temporal expression patterns for identified L-R pairs

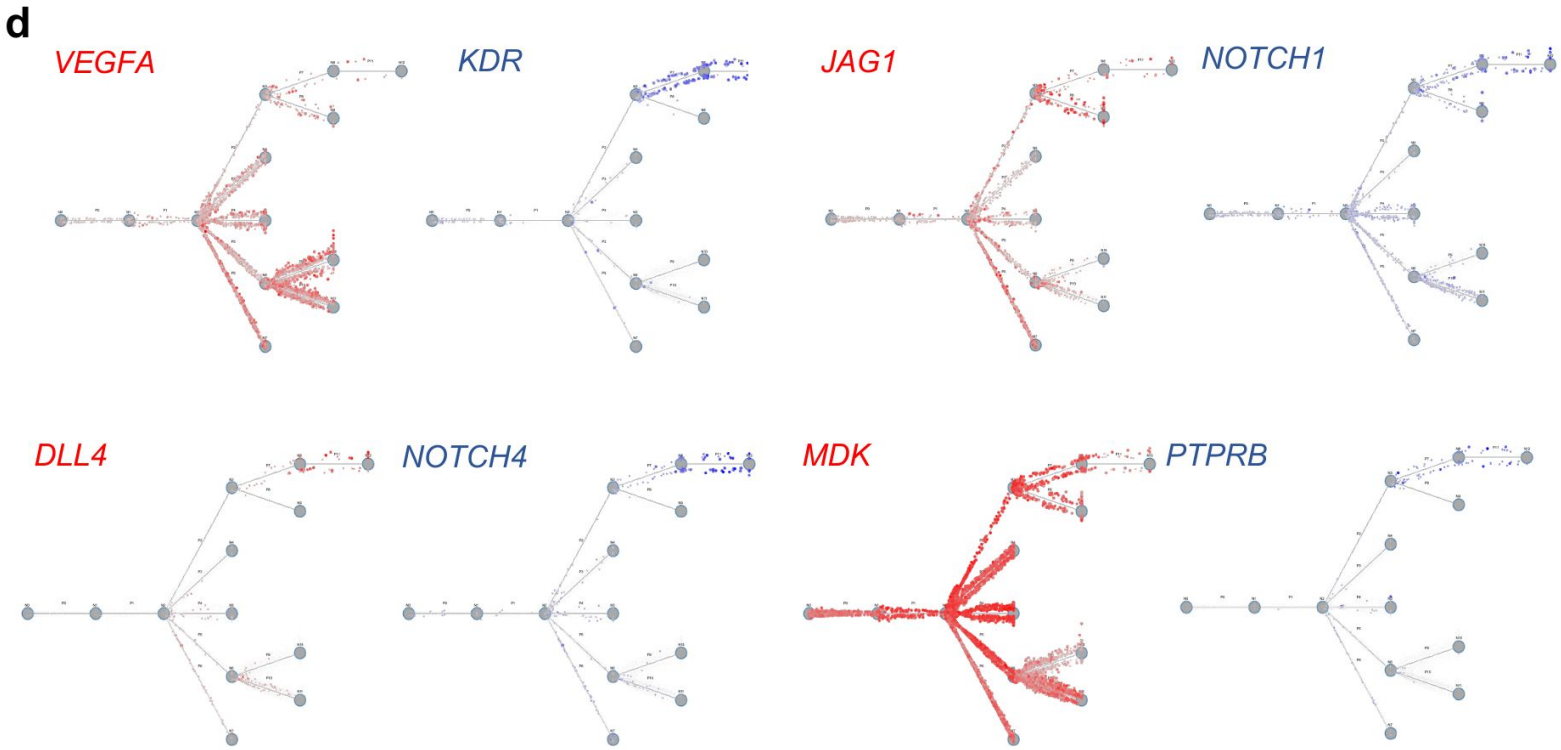


Ligand-receptor interaction predictions of interest for functional studies



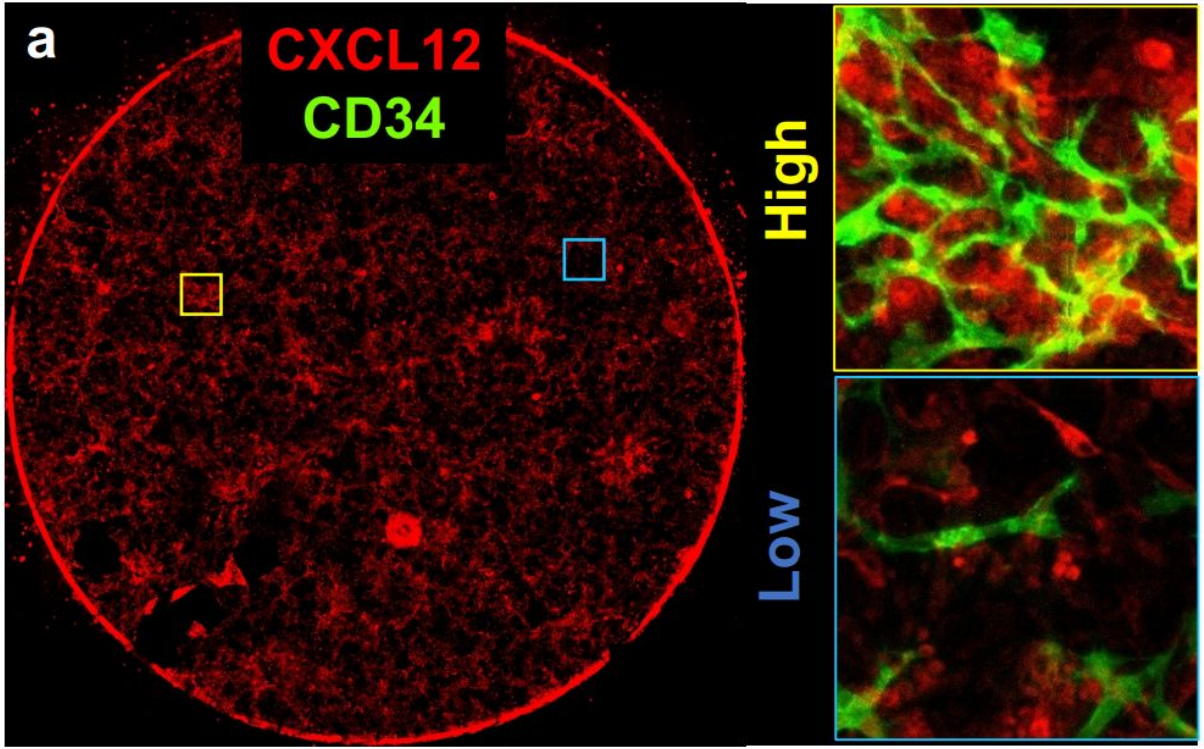
HLC, hepatocyte-like cells; CLC, cholangiocyte-like cells; SLC, stellate-like cells; ELC, endothelial-like cells

Sending and receiving cell populations



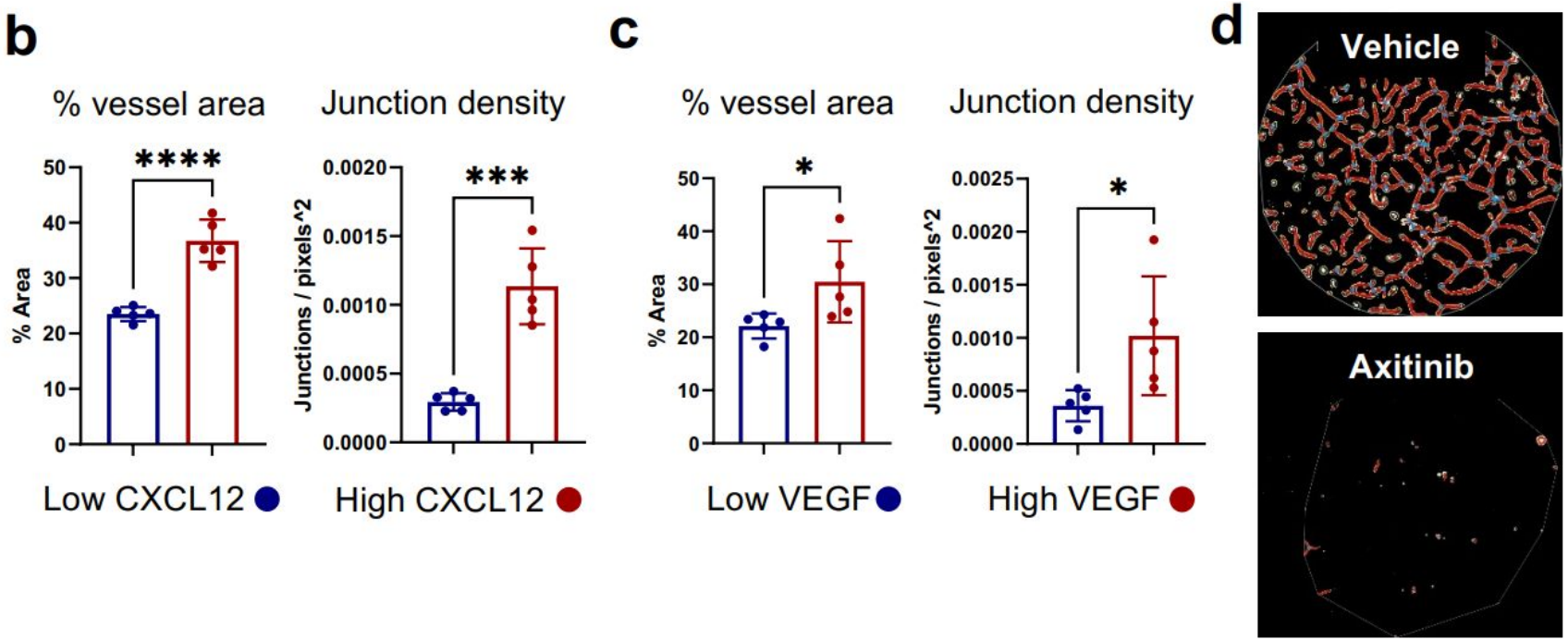
Red: sender, Blue: Receiver

Experimental validation -1



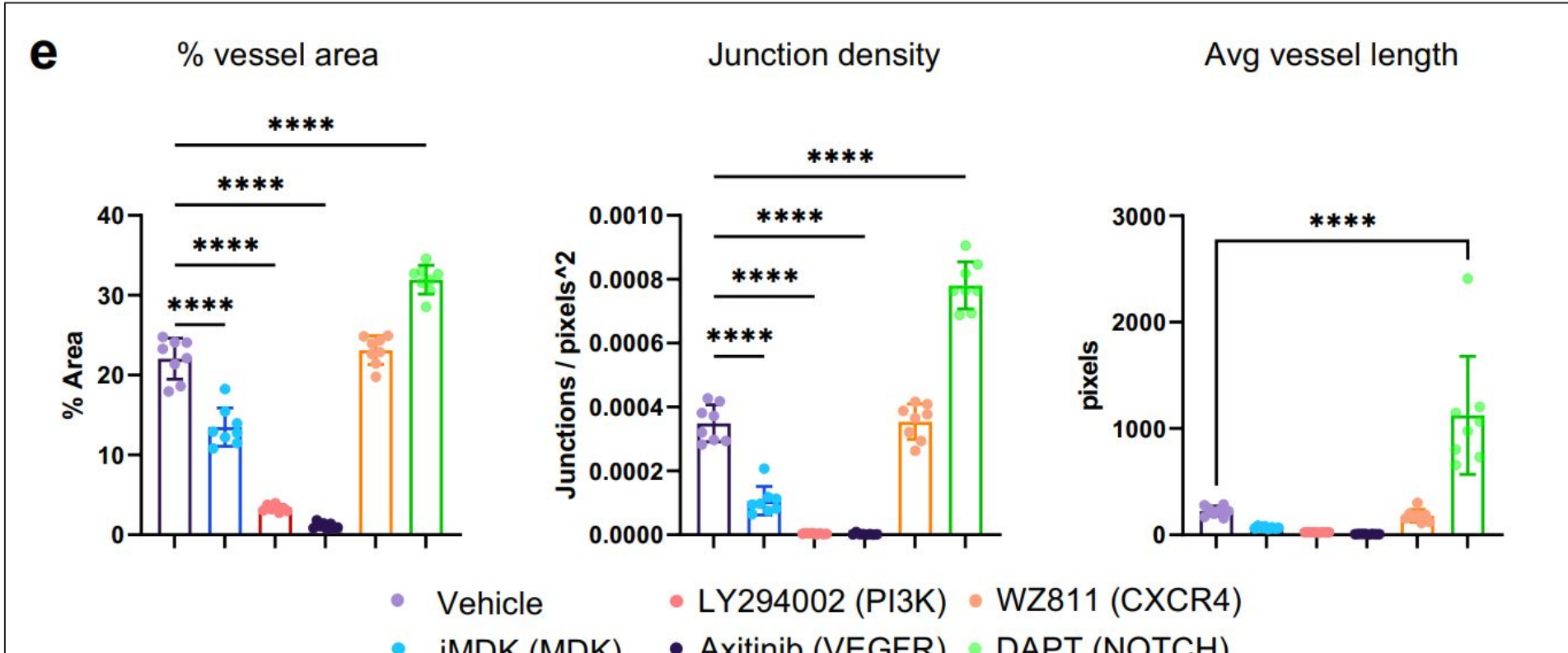
CD34:
hepatic progenitor cells

Experimental validation-2

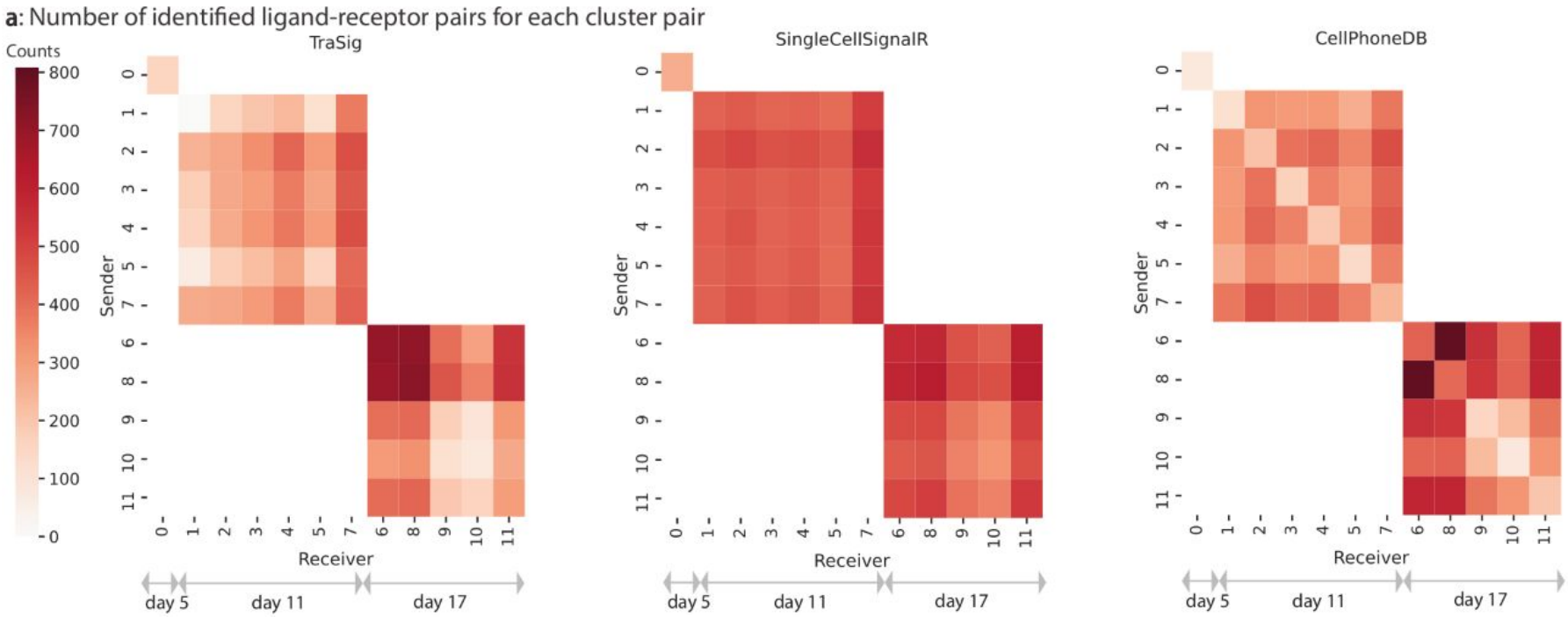


VEGF inhibitor

Experimental validation-3



Comparison with other methods-1

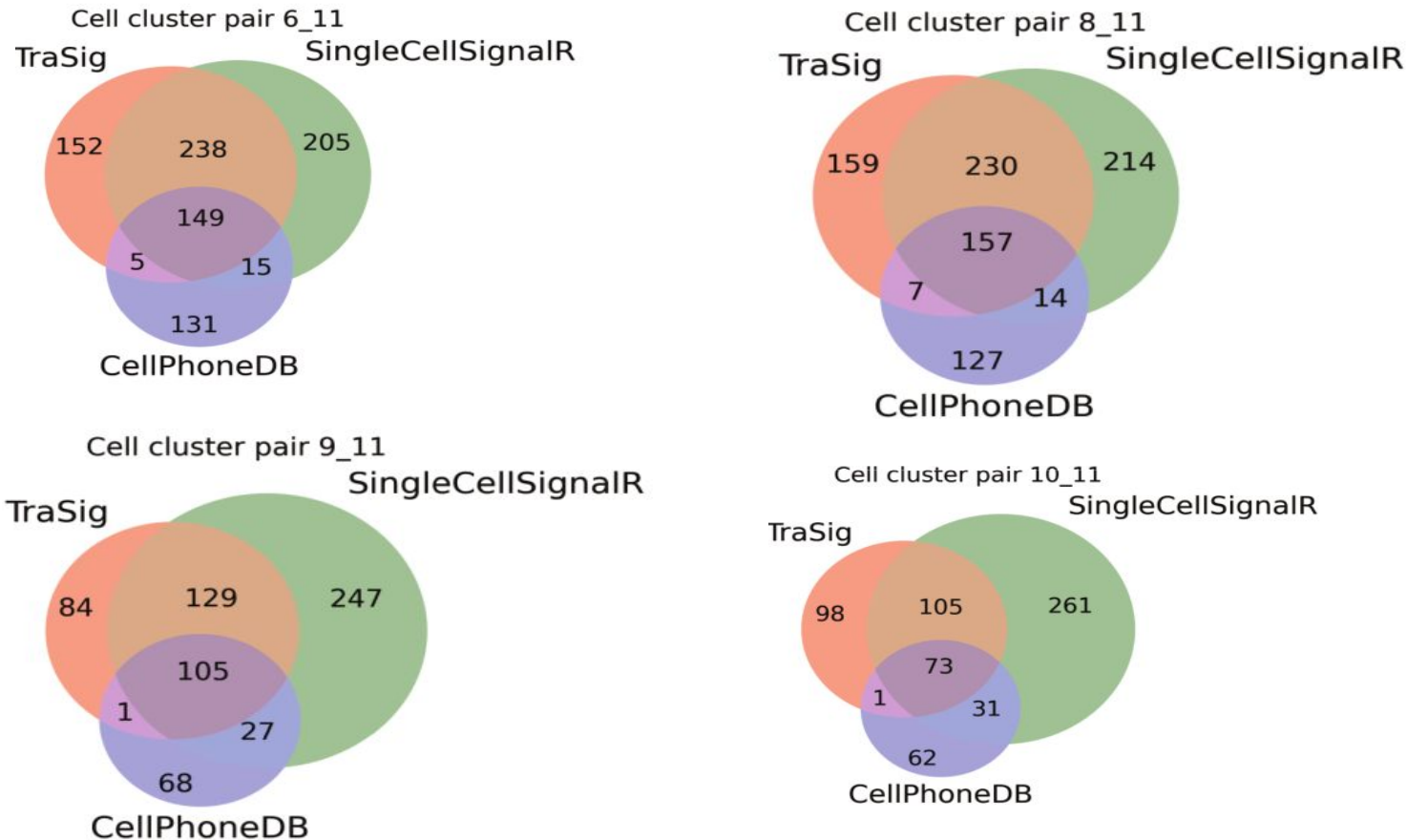


Cabello-Aguilar, Simon, et al. "SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics." *Nucleic acids research* 48.10 (2020): e55-e55.

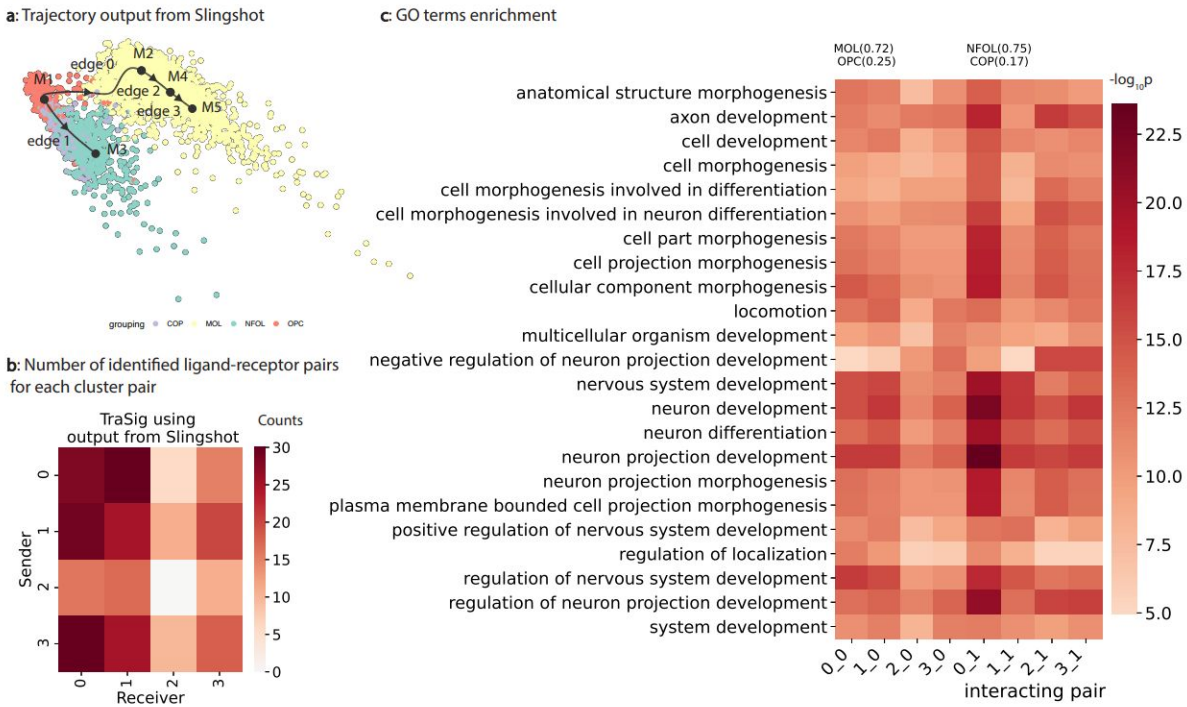
Efremova, Mirjana, et al. "CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes." *Nature protocols* 15.4 (2020): 1484-1506.

Comparison with other methods-2

Comparison with other methods-3



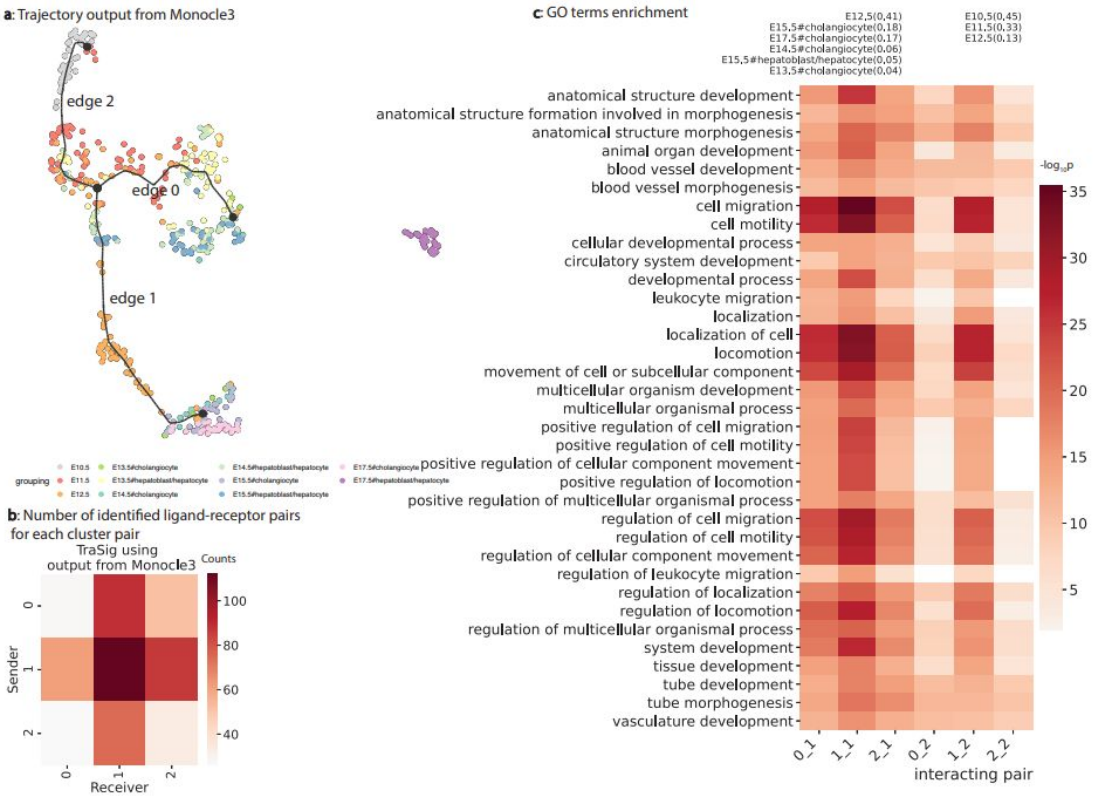
Trasig works with other pseudotime inference methods(Slingshot)



oligodendrocyte cell differentiation data [https://zenodo.org/record/1443566#.YhXJXYzMJhF]

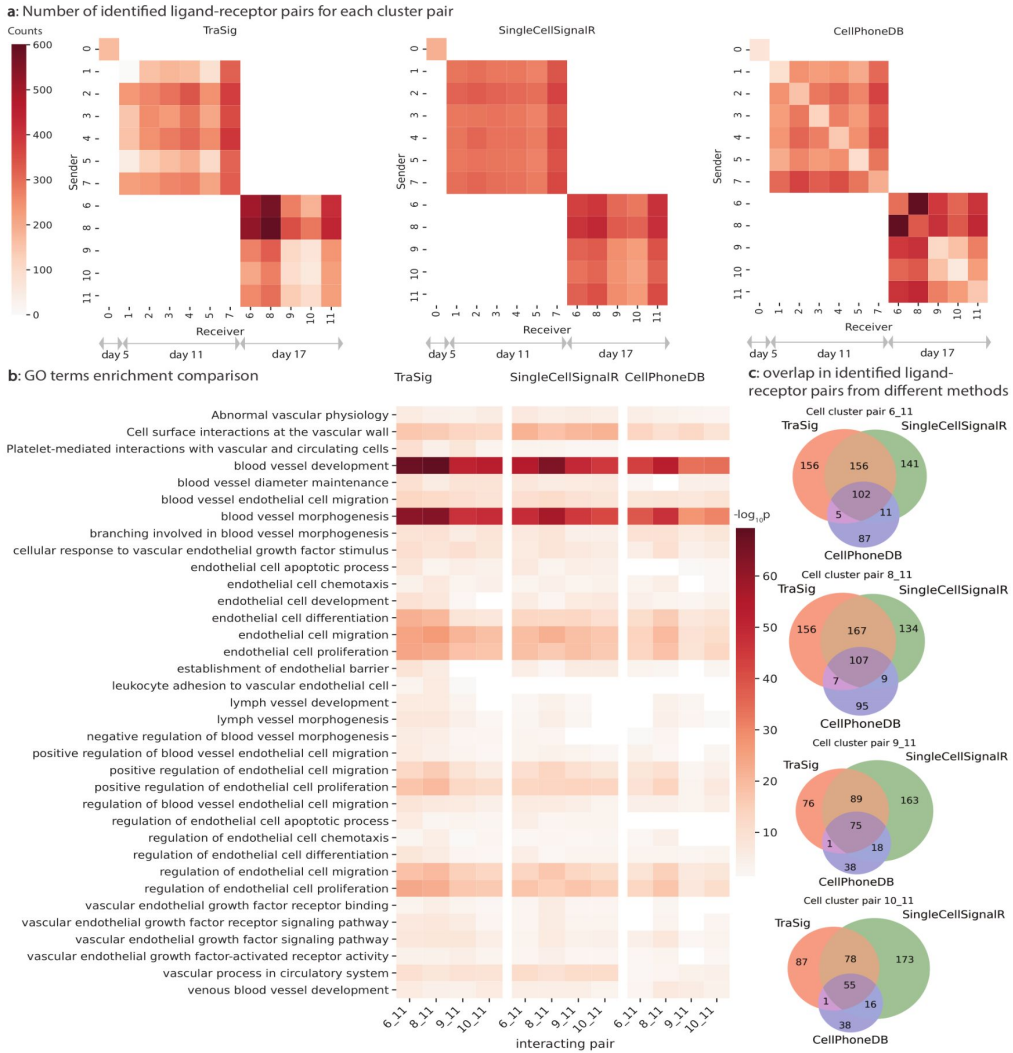
Saelens, Wouter, et al. "A comparison of single-cell trajectory inference methods." *Nature biotechnology* 37.5 (2019): 547-554.

Trasig works with Monocle3



hepatoblast differentiation data:

Yang, L. et al. A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. Hepatology 66, 1387–1401 (2017)



Comparison with another L-R database

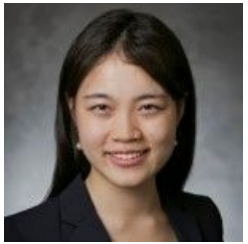
Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A. & Forrest, A. R. Predicting cell-to-cell communication networks using natmi. Nat. Commun 11, 1–11 (2020).

Thanks

Collaborators

CMU

Ziv Bar-Joseph



Dongshunyi (Dora) Li

Pitt

Mo R. Ebrahimkhani

Funding

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

THREE LAKES FOUNDATION

Fonds de recherche Santé Québec

Integration of multi-omics data for the discovery of novel regulators that modulate biological processes

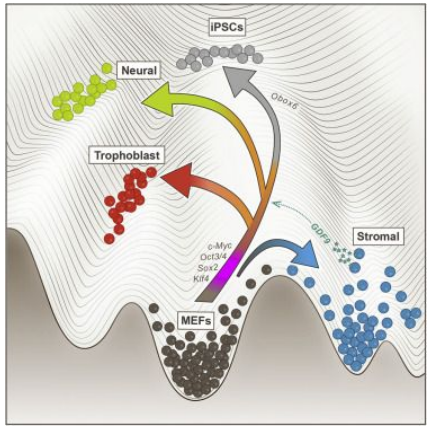
Jun Ding
Meakins-Christie Laboratories
Department of Medicine
Department of Biomedical Engineering
McGill University
02/10/2022

MEAKINS CHRISTIE

Centre universitaire de santé McGill
Institut de recherche

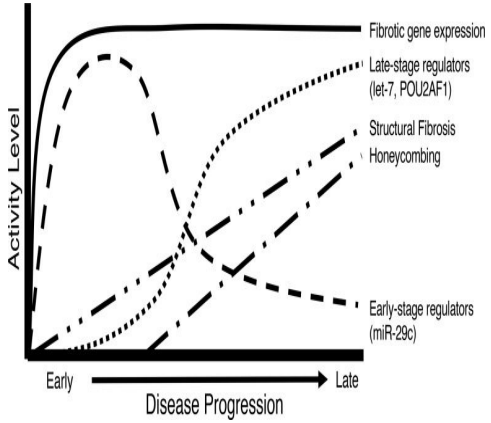
McGill University Health Centre
Research Institute

Cellular dynamics in various biological processes



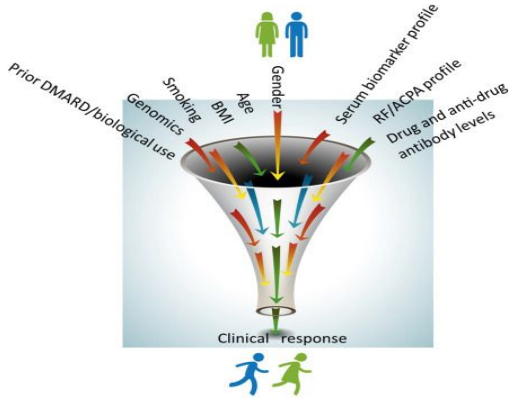
Cell differentiation & reprogramming

Schiebinger, Geoffrey, et al. "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming." *Cell* 176.4 (2019): 928-943.



Disease progression

McDonough, John E., et al. "Transcriptional regulatory model of fibrosis progression in the human lung." *JCI insight* 4.22 (2019).



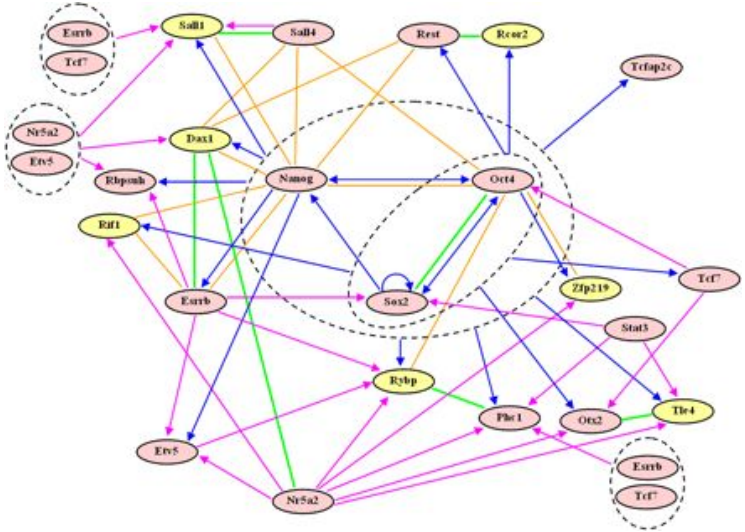
Clinical response

Wijbrandts, C. A., and P. P. Tak. "Prediction of response to targeted treatment in rheumatoid arthritis." *Mayo Clinic Proceedings*. Vol. 92. No. 7. Elsevier, 2017.

How to identify the regulators that dictate the cellular dynamics in those biological processes for “interventions”?

Regulators in Gene regulatory networks (GRN)

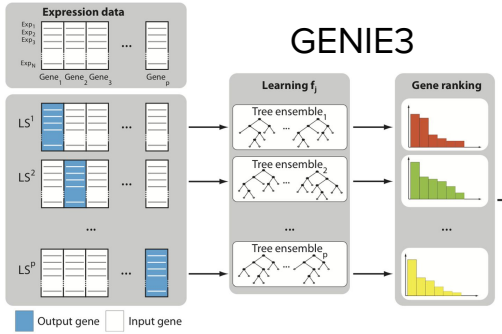
GRN inference



Zhou, Qing, et al. "A gene regulatory network in mouse embryonic stem cells." *Proceedings of the National Academy of Sciences* 104.42 (2007): 16438-16443.

Regression/Correlation/ODE based GRN inference

GENIE3

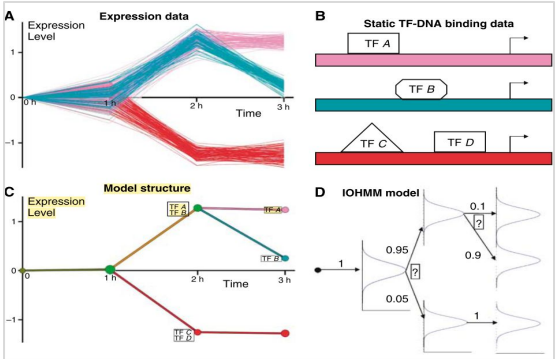


$$\sum_{k=1}^{N} (x_k^j - f_j(\mathbf{x}_k^{-j}))^2$$

Huynh-Thu, Vân Anh, et al. "Inferring regulatory networks from expression data using tree-based methods." *PloS one* 5.9 (2010): e12776.

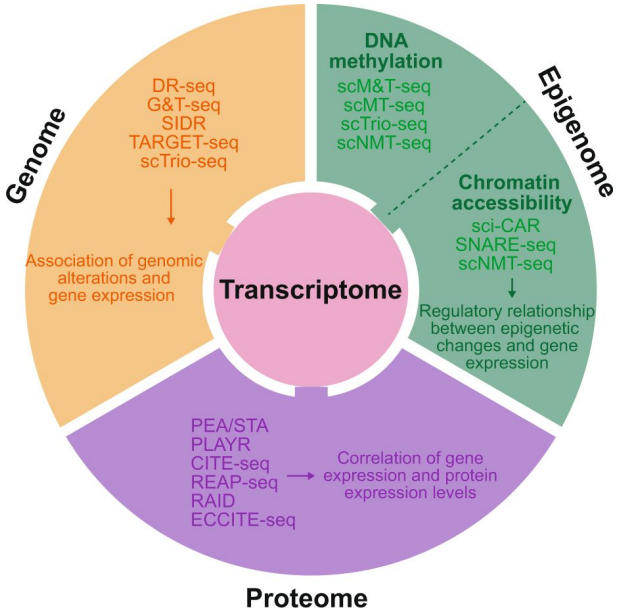
Probabilistic Graph based GRN inference

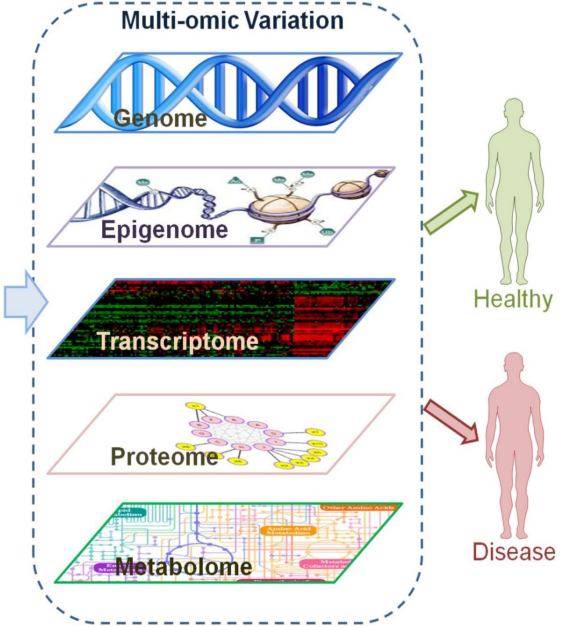
DREM

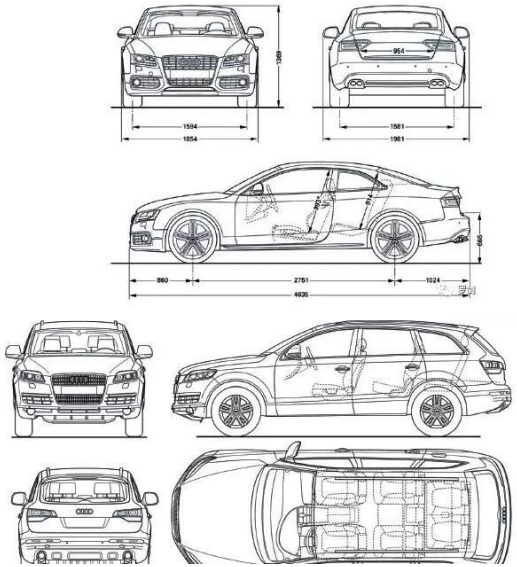


Schulz, Marcel H., et al. "DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data." *BMC systems biology* 6.1 (2012): 1-9.

RNA-seq vs. Multi-omics







Multi-omics:
Complementary views
from different perspectives

Lee, Jeongwoo, Do Young Hyeon, and Daehee Hwang. "Single-cell multiomics: technologies and data analysis methods." *Experimental & Molecular Medicine* 52.9 (2020): 1428-1442.

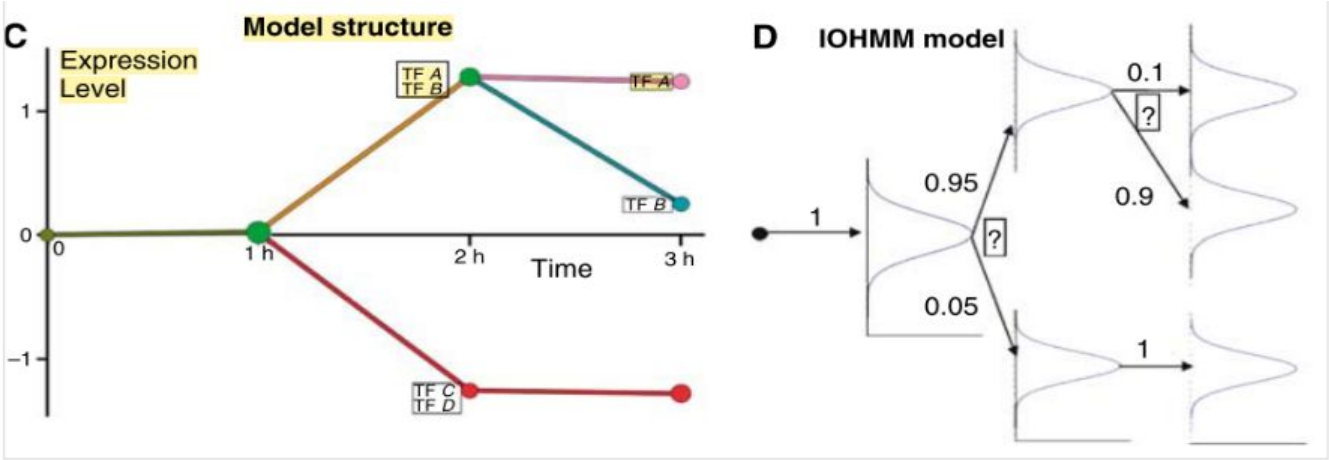
Sun, Yan V., and Yi-Juan Hu. "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases." *Advances in genetics* 93 (2016): 147-190.

Interactive Dynamic Regulatory Events Miner (IDREM)

Model Overview

IOHMM model M =(H,E,Θ,Ψ)

H,E denote the nodes and edges -> model structure.
Θ,Ψ represent the parameters for calculating the the emission and transition probabilities
-> model parameters under current structure



Θ denote the parameters for a gaussian model, which maps the **gene expression** at each node => emission probability.

Ψ denote the parameters for a regression model, which maps the input (**TF-DNA binding**) => transition probability (e.g. 0.95, 0.05 shown in D).

Ernst, Jason, et al. "Reconstructing dynamic regulatory maps." *Molecular systems biology* 3.1 (2007): 74.

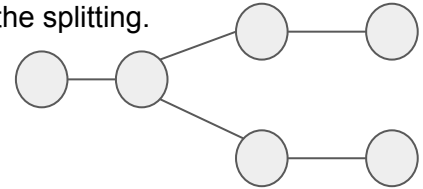
Model learning

(1) Randomly split all genes into a train set (75%) and a test set (25%).

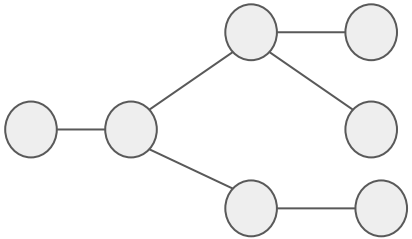
(2) Start searching the structure from a single chain.

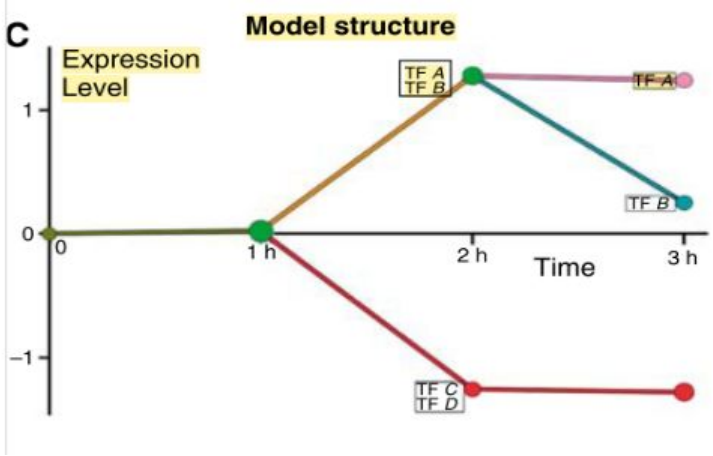
(3) Under current structure (H,E), Use Baum-welch algorithm to find the model parameters (using train set) which present the maximal likelihood on test set r(G_test|M). M is the current model M =(H,E,Θ,Ψ).

(4) Random Split the path under certain constraints (e.g. no more 3 edges coming out from node). Then calculate the score for the new model r(G_test|M_new). M_new is the model after the splitting.



(5) We keep doing the above process until the score converges.
Then, we got the final Structure. Finally, we used all genes to estimate the model parameters => Final Model M.





Score calculation

$$r(G|M) = \sum_{g \in G} log \sum_{q \in Q} \prod_{t=1}^{n-1} f_{q(t)(o_g(t)} \prod_{t=1}^{n-1} P(H_t = q(t)|H_{t-1} = q(t-1), I(g,t))$$

The first product denotes the emission probability and the second product represents the transition probability. The inner sum is
over all paths and the outer sum is over all genes in G. $I(g, t)$ is the dynamic input prior learned by integrating all omics data.

$$P(H_t = q(t)|H_{t-1} = q(t-1), \boxed{I(g,t))}$$

This probability can be calculated using a regression model

Where the omics integration happens

Integration of TF-DNA interaction data

$$RegValue(TF_x, time_z) = |expression(x, z+1) - expression(x, z)|$$

Then, we normalize all RegValue to [0,1] using the logistic function

$$f_w(x) = \frac{1}{1 + e^{-xw}}$$

To determine the sign of the regulation RegDirection (x,y,z) (Note: if the original TF-DNA file already has this information, just use it directly)

If TF up , target up (activation 1)
If TF up, target down (repression -1)
If TF down, target up (repression -1)
If TF down, target down (activation 1)

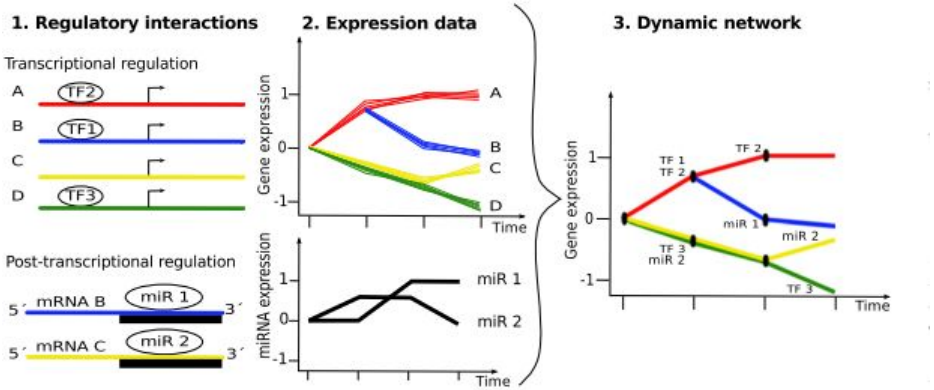
The final TF-DNA interaction value :

$Interaction(TF_x, gene_y, time_z) = RegValue(x,z) * RegDirection(x,y,z) * TFDNA(x,y,z)$

Where, TFDNA (x,y,z) is the binary to represent whether TF x is binding to gene y at time point z.

Integration of miRNA data

miRNA information was treated as a special type of “TF”, which can only repress the target expression. On the other hand, TF can either activate or repress target expression.



Integration of proteomics and PPI

It's not accurate to use gene expression level to represent the level of corresponding TF. Besides, TF regulates the gene expression via a "impact" on RNA polymerase (Pre-initialization complex-PIC). The impact was by a series of Protein-protein interactions.

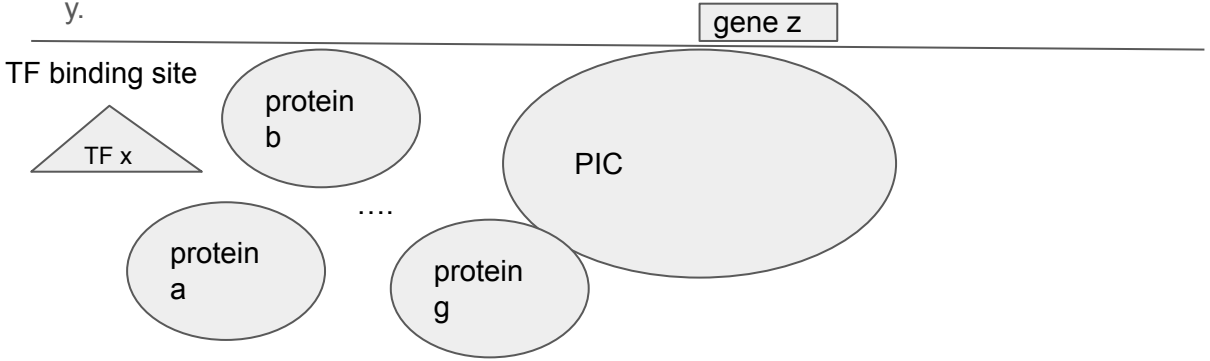
$$TF_x = \frac{1}{|Y|} \sum_{\{y \in Y\}} ProteinLevel_x * ProteinLevel_y * PPI(x, y)$$

Y : interacting proteins of x; Protein level all normalized to [0,1]

E.g.,

Case A: TF x is highly expressing, but none of its known interacting proteins are expressing.

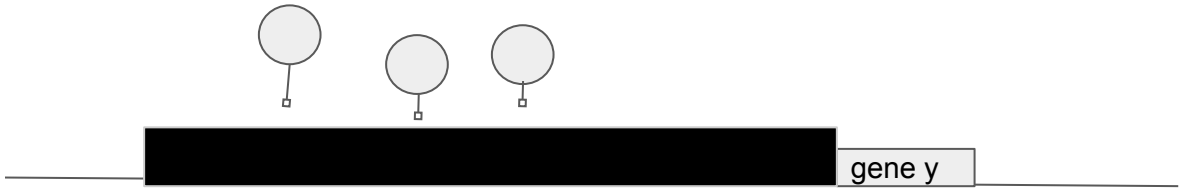
Case B: TF y is expressing and so do its interacting proteins. TF x and y are both known to regulate gene z. In this specific case, TF x is more likely active compared with y.



Integration of methylation data

In the main framework, the TF-DNA data is static, which means it's not changing during the process. This is definitely not the case in reality. Here, we can use the methylation data to get the dynamic TF-DNA binding information. The methylation in the promoter region will silence the downstream gene expression [pubmed 24555846]. Detail steps:

(1) Mapping the methylation reads and calling the methylation peaks.
(2) Compare the peaks with genomic location of TSSs of all genes.
(3) If there are peaks found in the promoter region (within upstream 10k of gene TSS), the promote of gene get methylated and we will modify all TF-DNA binding related to this gene.



The transition model will be impacted by the dynamic TF-DNA binding. As the transition model and emission model are tangling with each other, the emission model will be also impacted.

IDREM software interface

DREM - Dynamic Regulatory Events Miner

1. Data Input:

TF-gene Interaction Source: User Provided

TF-gene Interactions File: Browse...

Expression Data File: Browse...

Saved Model File: Browse...

View TF-gene Data | View Expression Data

Spot IDs in the data file | Repeat Data...

Log normalize data | Normalize data | No normalization/add 0

2. Gene Annotation Input:

Gene Annotation Source: User provided

Cross Reference Source: User provided

Gene Annotation File: Browse...

Cross Reference File: Browse...

Download the latest: Annotations | Cross References | Ontology

3. Options:

Options...

Main Panel

4. Execute:

Execute

Options

Gene Annotations | GO Analysis | DECOD Options | Expression Scaling Options | microRNA Option

Methylation Option | Proteomics Option | Filtering Options | Search Options | Model Selection Options

Methylation data File: Browse...

GTF file: Browse...

Methylation Panel

Options

Gene Annotations | GO Analysis | DECOD Options | Expression Scaling Options | microRNA Option

Methylation Option | Proteomics Option | Filtering Options | Search Options | Model Selection Options

miRNA-Gene Interaction Source: User Provided

microRNA-Gene Interaction File: Browse...

microRNA Expression Data File: Browse...

miRNA Repeat Data...

Log normalize data | Normalize data | No normalization/add 0

Filter miRNA with no expression from regulator data:

miRNA Panel

Options

Gene Annotations | GO Analysis | DECOD Options | Expression Scaling Options | microRNA Option

Methylation Option | Proteomics Option | Filtering Options | Search Options | Model Selection Options

Only Use Proteomcis Data for TFs | Use Proteomics data for all Proteins | Do Not Use Proteomics data

Proteomics Data File: Browse...

Repeat Proteomics Data File... | Log normalize data | Normalize data | No normalization/add 0

Protein-Protein Interaction File: Browse...

Proteomics PPI Panel

IDREM application in lung development

➢ **Gene expression**

The gene expression is in FPKM format with 15 time points.

e16.5	e18.5	p0.5	p1.5	p2.5	p4	p5	p7	p10	p13.5	p15	p19	p23	p28

➢ **miRNA expression**

The miRNA expression data is from NanoString technologies- ncount expression.
Based on the manual from NanoString technologies, it needs to be normalized.
http://www.nanostring.com/media/pdf/MAN_nCounter_Gene_Expression_Data_Analysis_Guidelines.pdf

Here, we used the housekeeping genes to do the normalization.
In the miRNA expression dataset, they offered the expression for a few housekeeping genes: Actb,B2m,Gapdh,Rpl19

The normalization steps:

A. First calculate the geometric mean of the expression of these housekeeping genes for each lane (sample)

$$g_{sample} = (\prod_{i \in H} g_i)^{(1/|H|)}$$

➢ Proteomics data

There are 15 time points for the proteomics data

e16.5	e18.5	p0.5	p1.5	p2.5	p4	p5	p7	p10	p13.5	p15	p19	p23	p28

Summary Table

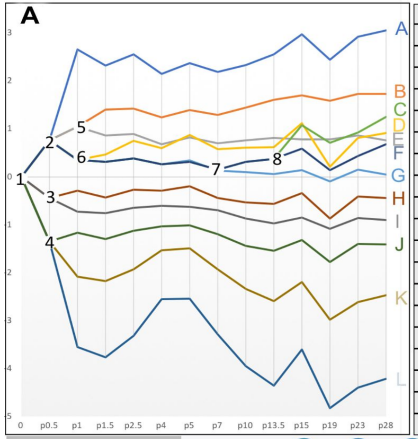
Time Point	e16.5	e18.5	p0.5	p1	p1.5	p2.5	p4	p5	p7	p10	p13.5	p15	p19	p23	p28
gene expression	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
miRNA expression	O	X	O	O	O	O	O	O	O	O	O	O	O	O	O
proteomics	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

➢ Methylation data

we have the methylation for the following 6 time point:
p0.5, p2.5,p7,p10,p19,p28

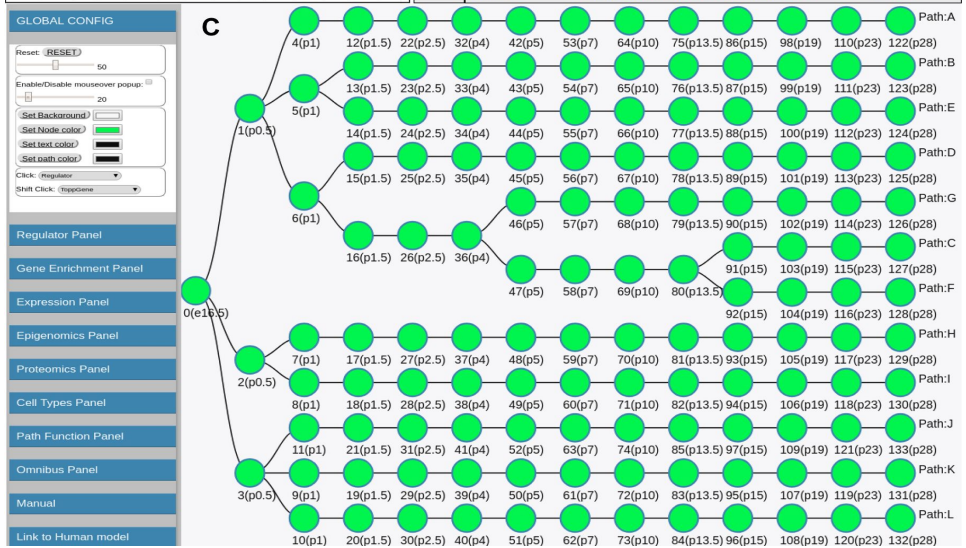
By combined all the above datasets, we decided to use the following 14 time points.

e16.5	p0.5	p1.5	p2.5	p4	p5	p7	p10	p13.5	p15	p19	p23	p28



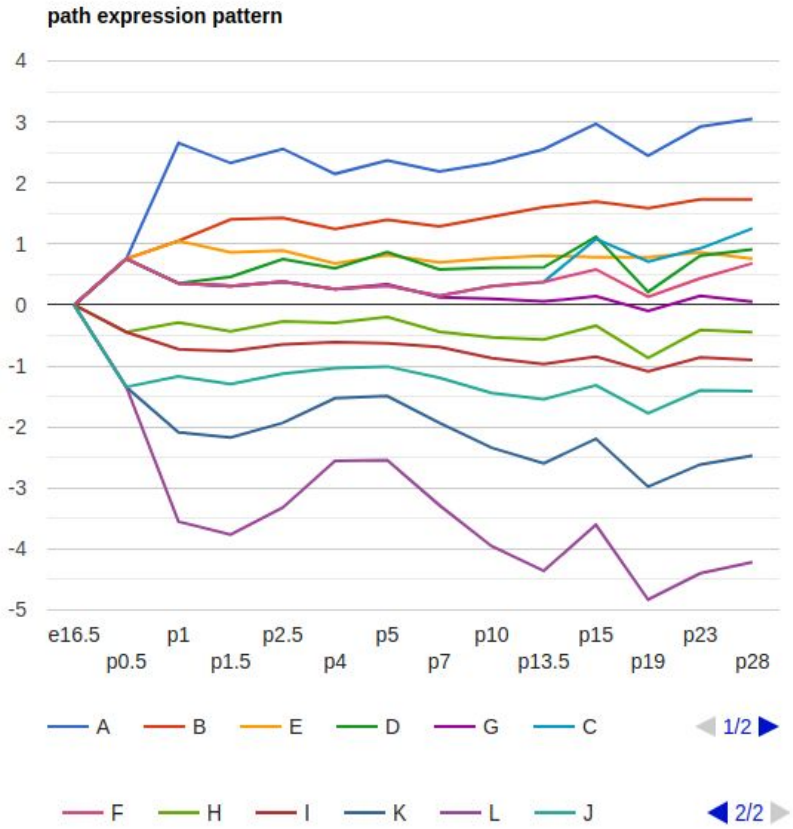
B

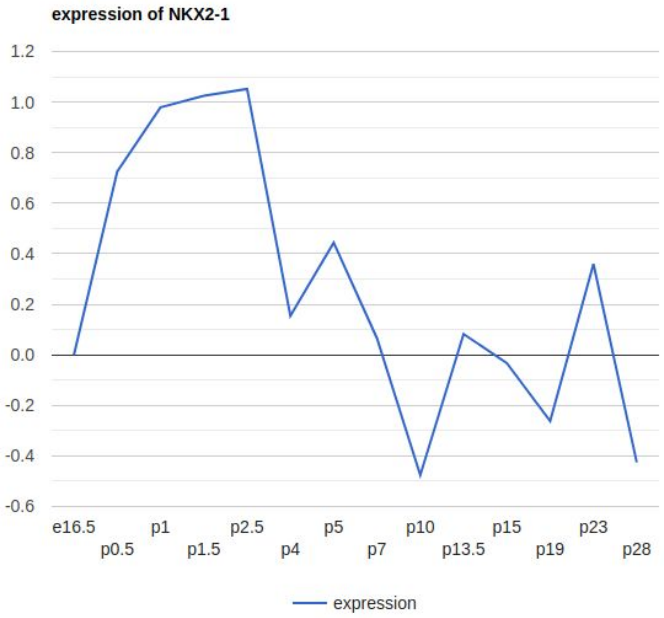
Path	Top regulators
1 – 2	SPIB,ETS1,FLI1,CUZD1,SFPI1,ETV4,ELF2,PIK3R3,ERG,PSMD12
1 – 3	miR-1192 ,miR-495,miR-590-3p,miR-186,miR-30e
1 – 4	RB1,E2F5,E2F2,miR-590-3p,TBP,E2F4,NKX3-1,TFDP1
2 – *A*	miR-451,miR-758,miR-187,IRF5,IRF3,IRF4,RELA,IRF2,SRF,IRF6
2 – 5	STAT5A,E2F1,CEBPA,CEBPB,STAT6,VDR,GATA1,PTGDR,SRPR
2 – 6	ELK4,miR-210,miR-551b,miR-433,miR-190,ZBTB16
3 – *H*	NR1I1,SMAD3,SFPI1,ETS1,SPIB1,PIK3R3,ERG,PSMD12,FSCN1,ATM1
3 – *I*	miR-192,miR-495,miR-141,miR-590-3p,miR-421
4 – *J*	miR-377,EGR2,EGR3,miR-302,miR-338-3p,HIC1,miR-302d
4 – *K*	TRAF4,ALX1,miR-132,ALX4,miR-296-3p,miR-451,POU2F1,miR-485
4 – *L*	OTX1,OTX2,POU5F1,POU2AF1,HMGA2,HOXA4,CDX1/2,NKX6-2,GATA6
5 – *B*	IRF8,IRF6,IRF2,RF5,IRF3,IRF4,miR-126-3a,IRF9,miR-183,PAX6
5 – *E*	RFX1,NKX2-1,SMAD3,SP1,SP4,RFXANK,RFX5,RFXAP,SMAD1,TRAF4
6 – *D*	IL6,TBP,RB1,miR-182,E2F4,ESRRA,GLI3,GLI1,GTF2A2,GTF2A1
6 – 7	let-7b,let-7a,let-7g,let-7i,miR-133a,miR-133b
7 – 8	IRF9,IRF7,IRF5,IRF3,IRF4,STAT1,IRF1,STAT2,IRF2,miR-539
7 – *G*	HINFP,ARNT,miR-184,XBP1,PITX2,RFX1,MYCN,MXI1,HAND1
8 – *C*	STAT2,miR-708,miR-449c,STAT4,miR-494,SRY,STAT6,IRF1,miR-539
8 – *F*	HES1,TCFAP2A,ZBTB7A,MAZ,ARNT2,miR-324,WT1,NFYB,TEAD2,ZIC2



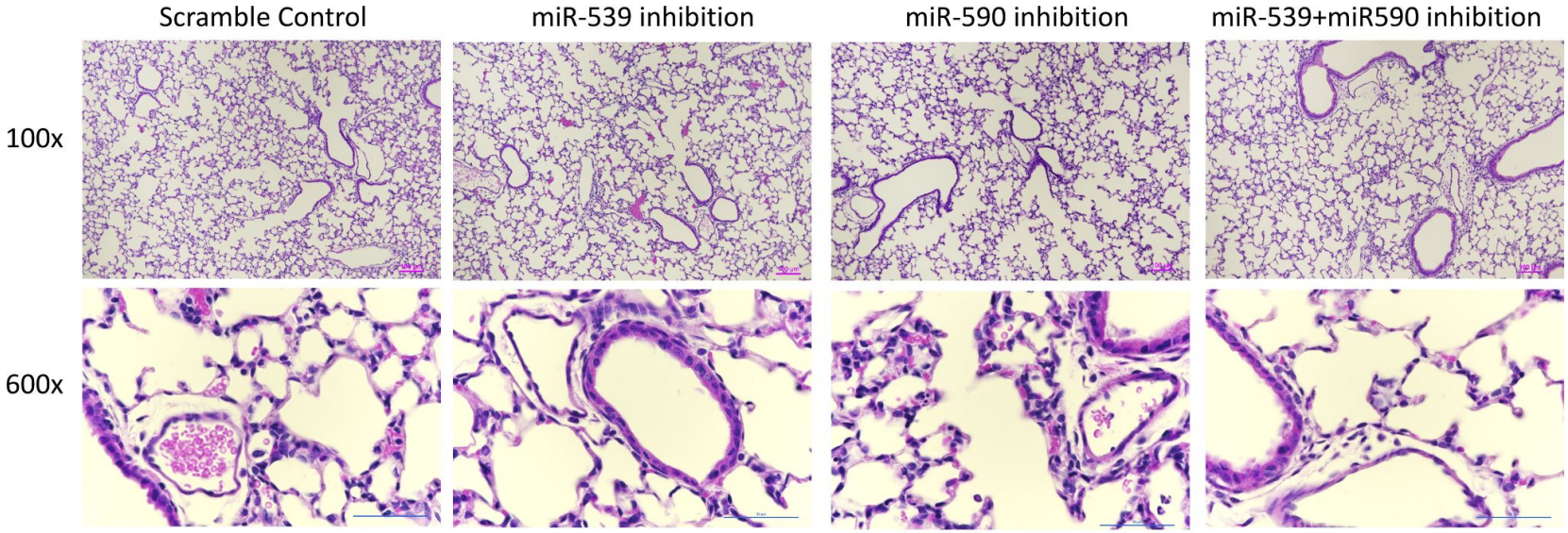
https://www.cs.cmu.edu/~jund/idrem_lung/

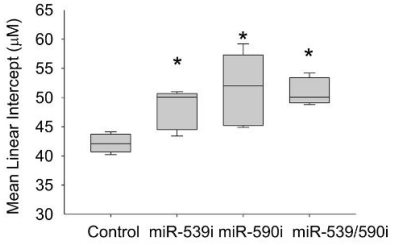
Ding, Jun, et al. "Integrating multiomics longitudinal data to reconstruct networks underlying lung development." *American Journal of Physiology-Lung Cellular and Molecular Physiology* 317.5 (2019): L556-L568.

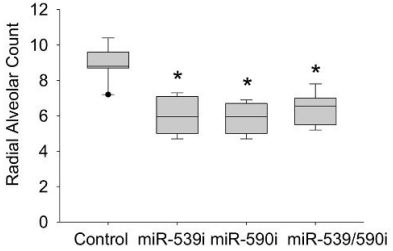

path expression pattern
4
3
2
1
0
-1
-2
-3
-4
-5
e16.5
p0.5
p1
p1.5
p2.5
p4
p5
p7
p10
p13.5
p15
p19
p23
p28
A
B
E
D
G
C
1/2
F
H
I
K
L
J
2/2

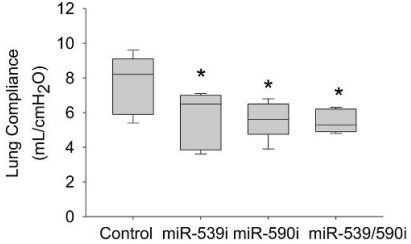

expression of NKX2-1
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
e16.5
p0.5
p1
p1.5
p2.5
p4
p5
p7
p10
p13.5
p15
p19
p23
p28
expression

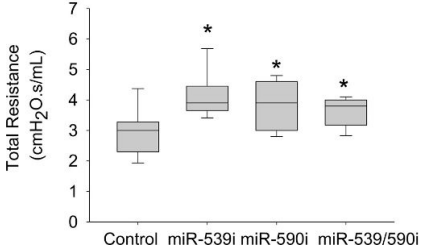
Experimental validation for the novel regulators (miR-539 and miR-590) from IDREM

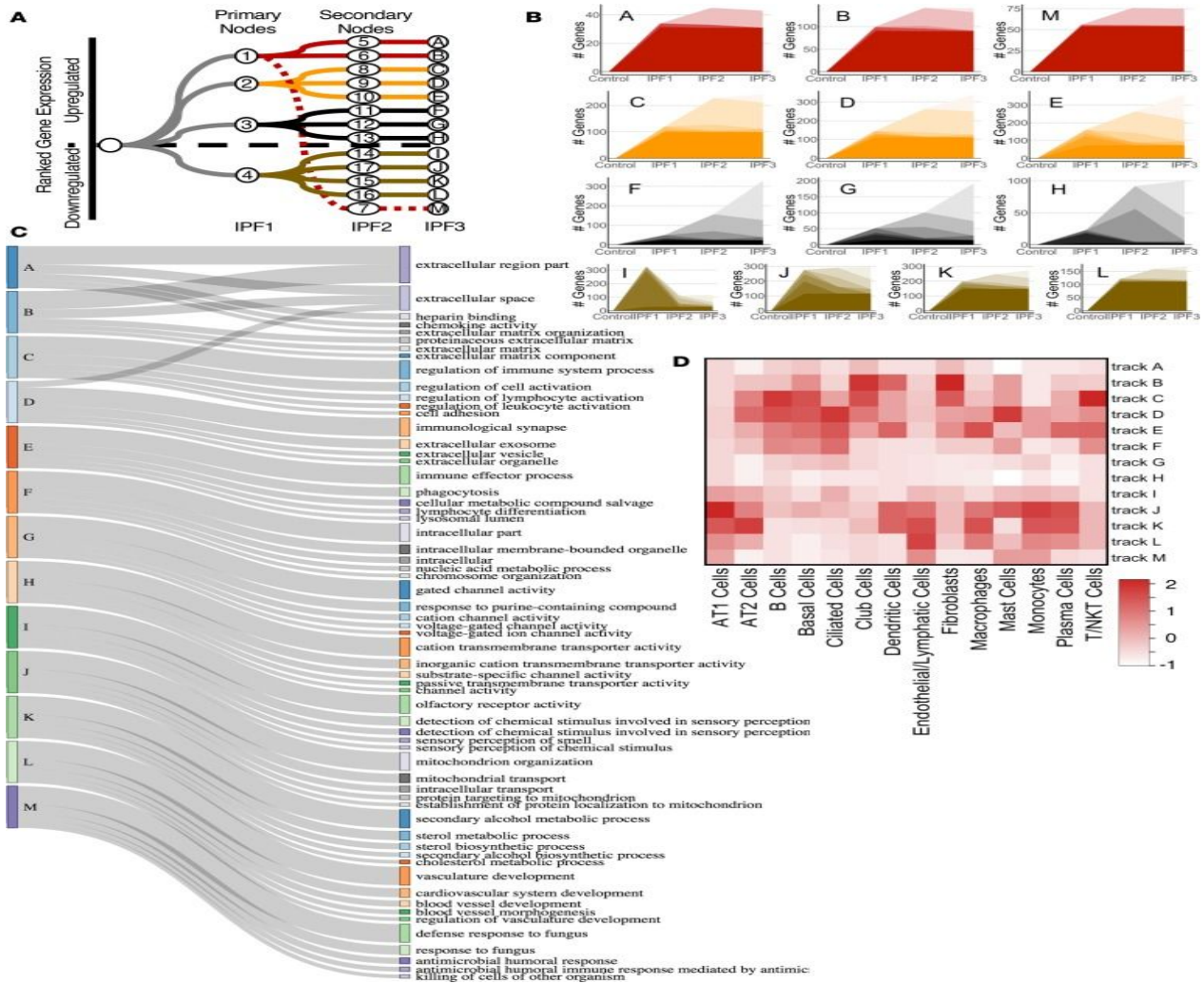




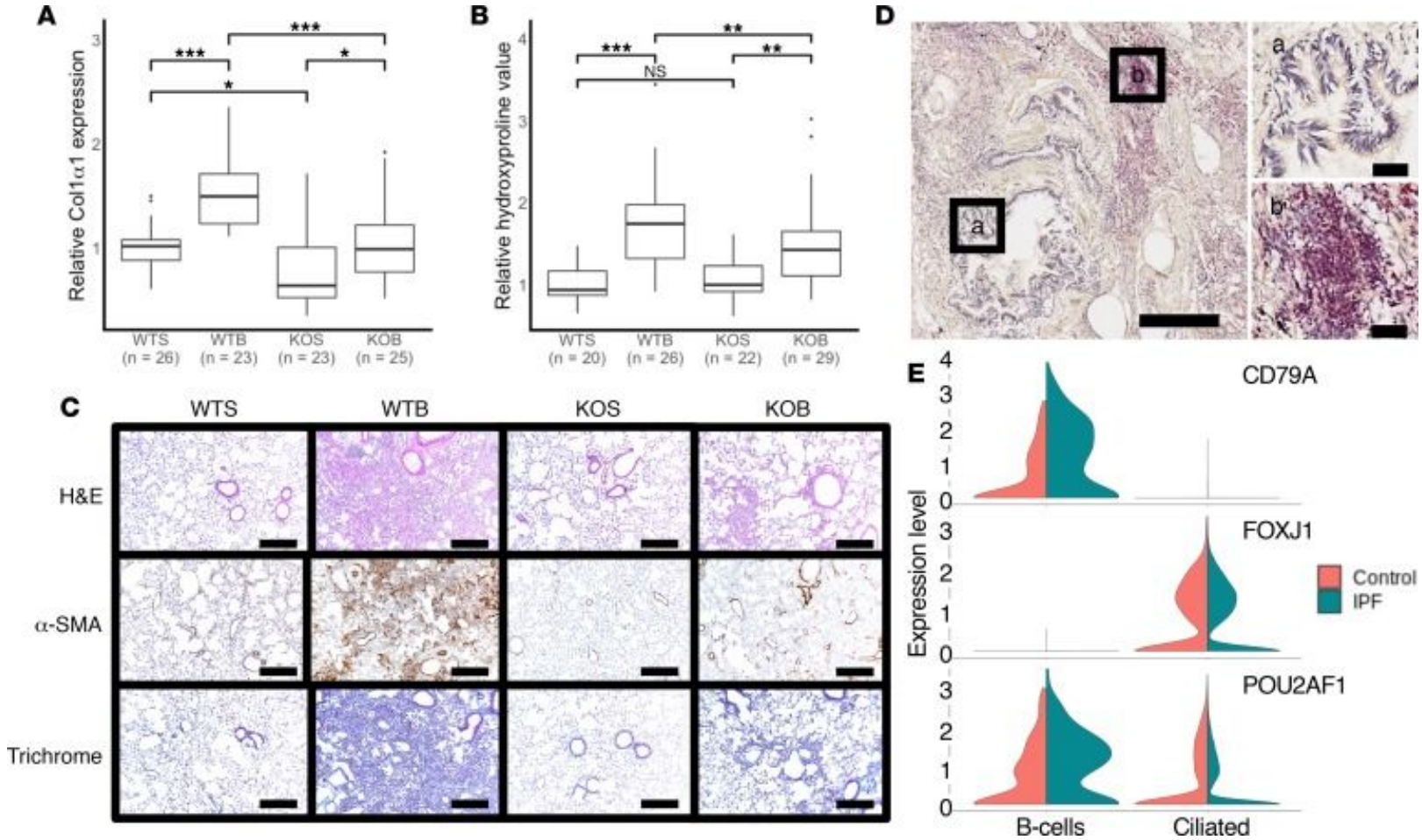








McDonough, John E., et al. "Transcriptional regulatory model of fibrosis progression in the human lung." *JCI insight* 4.22 (2019).


A
Relative Col1α1 expression
***
***
*
*
WTS (n = 26)
WTB (n = 23)
KOS (n = 23)
KOB (n = 25)
B
Relative hydroxyproline value
***
**
NS
**
WTS (n = 20)
WTB (n = 26)
KOS (n = 22)
KOB (n = 29)
C
WTS
WTB
KOS
KOB
H&E
α-SMA
Trichrome
D
a
b
E
CD79A
FOXJ1
POU2AF1
Expression level
Control
IPF
B-cells
Ciliated

Summary -I

1) Graphical models are very flexible for data integration (particularly the Input-output hidden markov)

2) Integration of multi-omics data could lead to the discovery of novel regulators for various biological processes

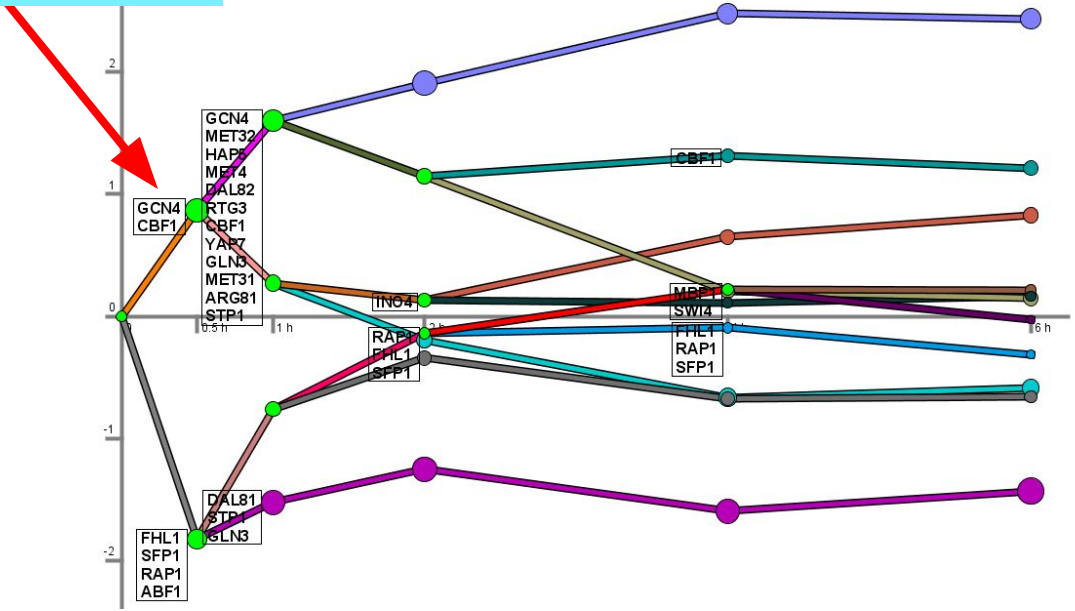
3) Interactively visualized model could promote novel biological discoveries

Multi-omic model that identifies novel drugs against COVID-19

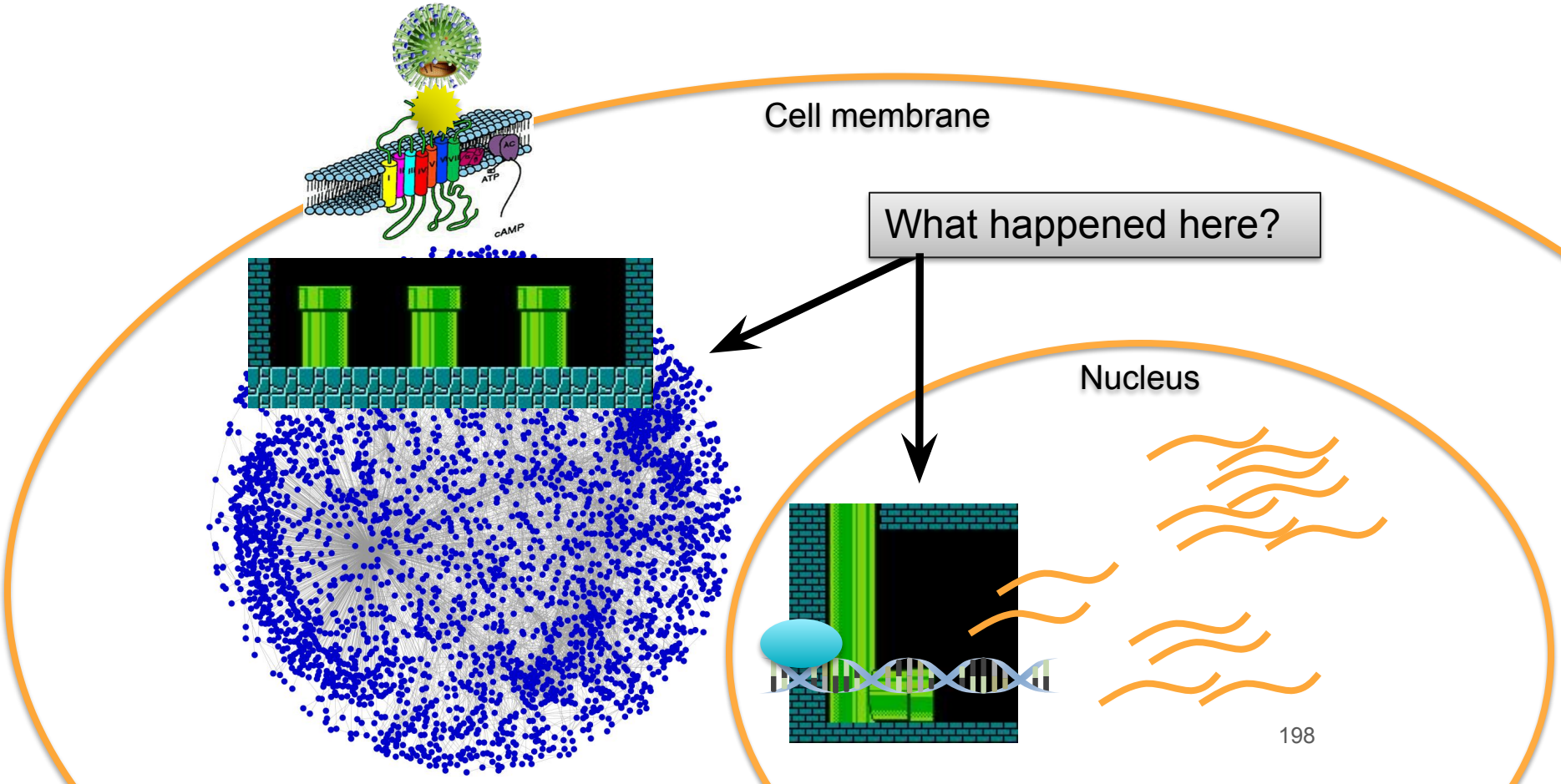
SARS-Cov2 modified SDREM analysis

DREM is useful, but several questions remain …

Who controls the master regulators?



Response to infection

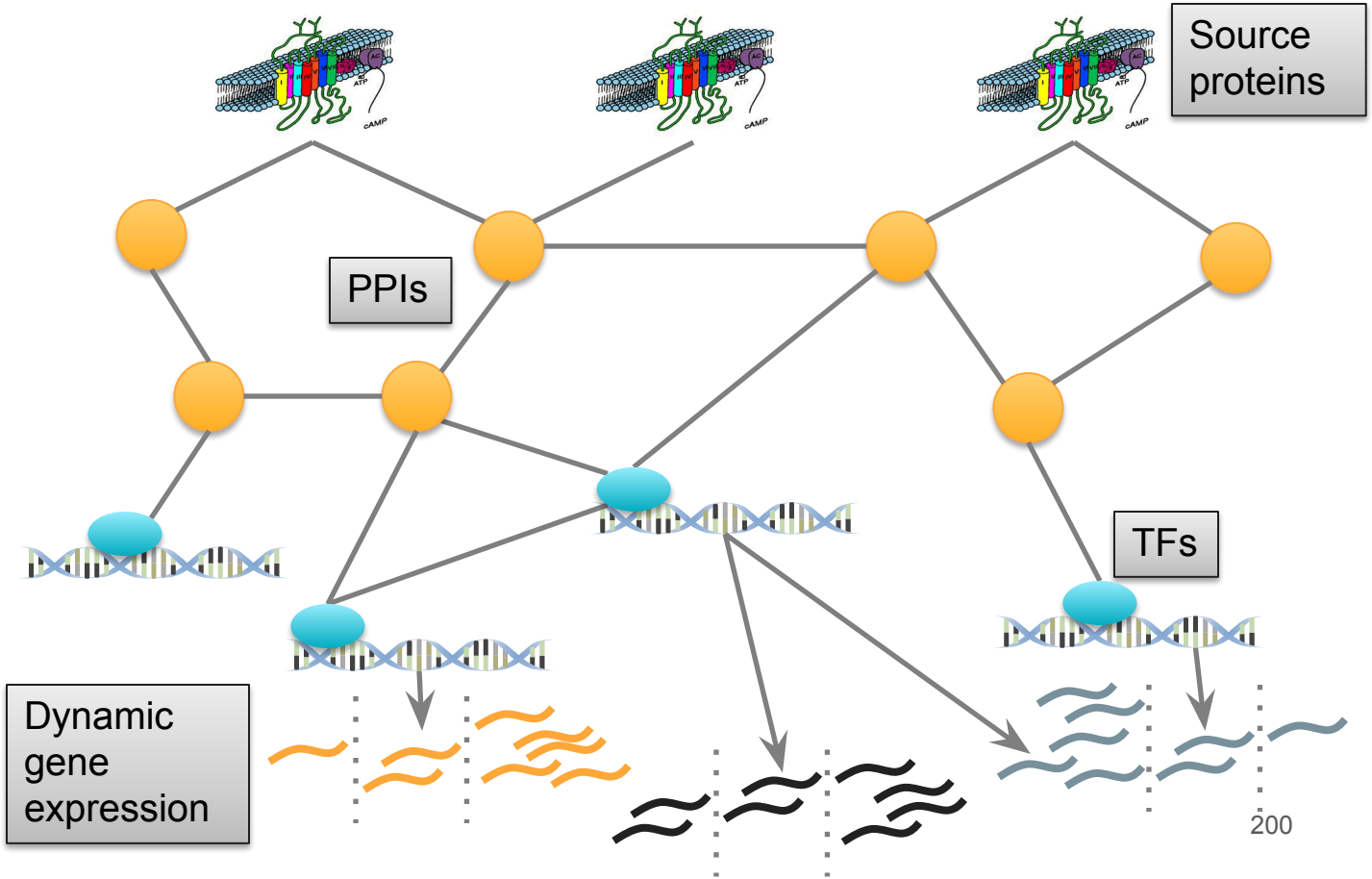


SDREM: Extending DREM to model signaling networks

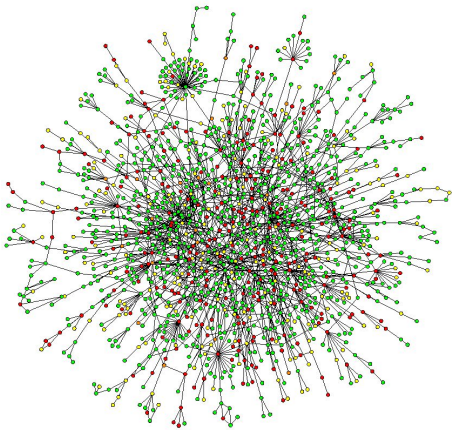
Inputs:

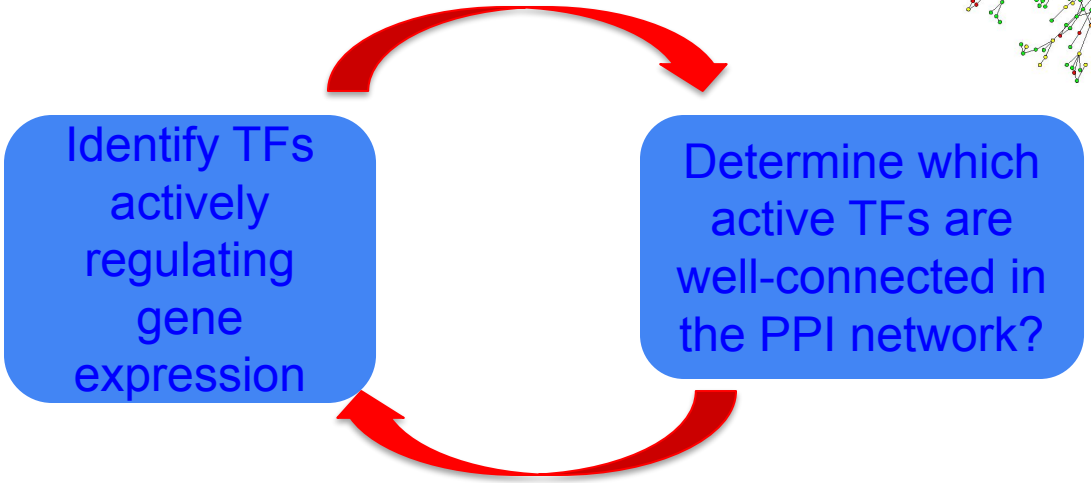
- Condition specific inputs:
  - Time series expression data following treatment
  - (A few) receptors interacting with invader or activated by condition of interest
  - Phosphorylation data
  - Protein level data
- General interaction data (not necessarily from the same condition):
  - Protein-DNA interactions
  - Motif information
  - Protein interaction networks

Inferring signaling pathways



Iterative method for reconstructing dynamic signaling and regulatory networks





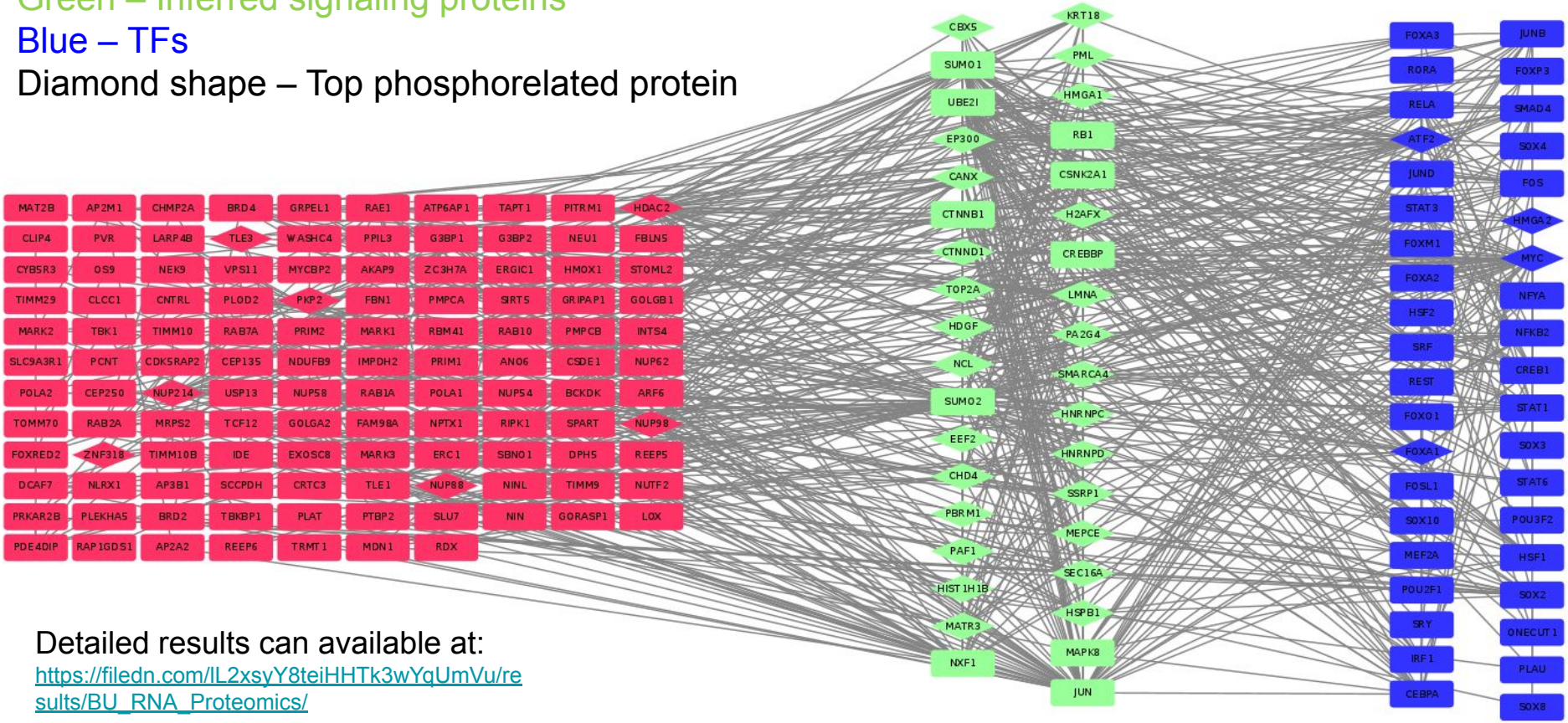
mSDREM model

Red – Source proteins (interacting with virus protein directly)

Green – Inferred signaling proteins

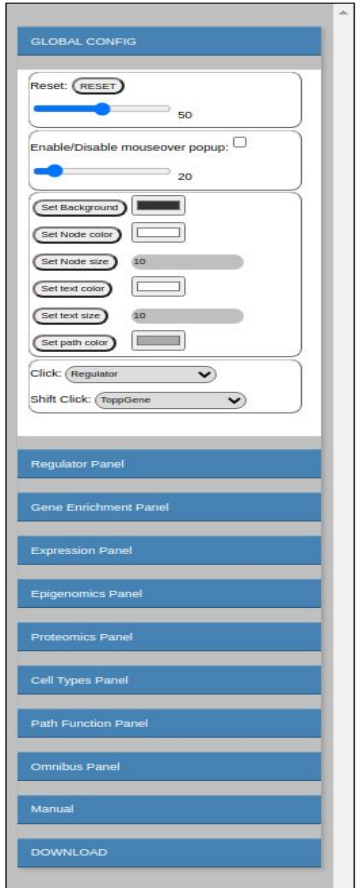
Blue – TFs

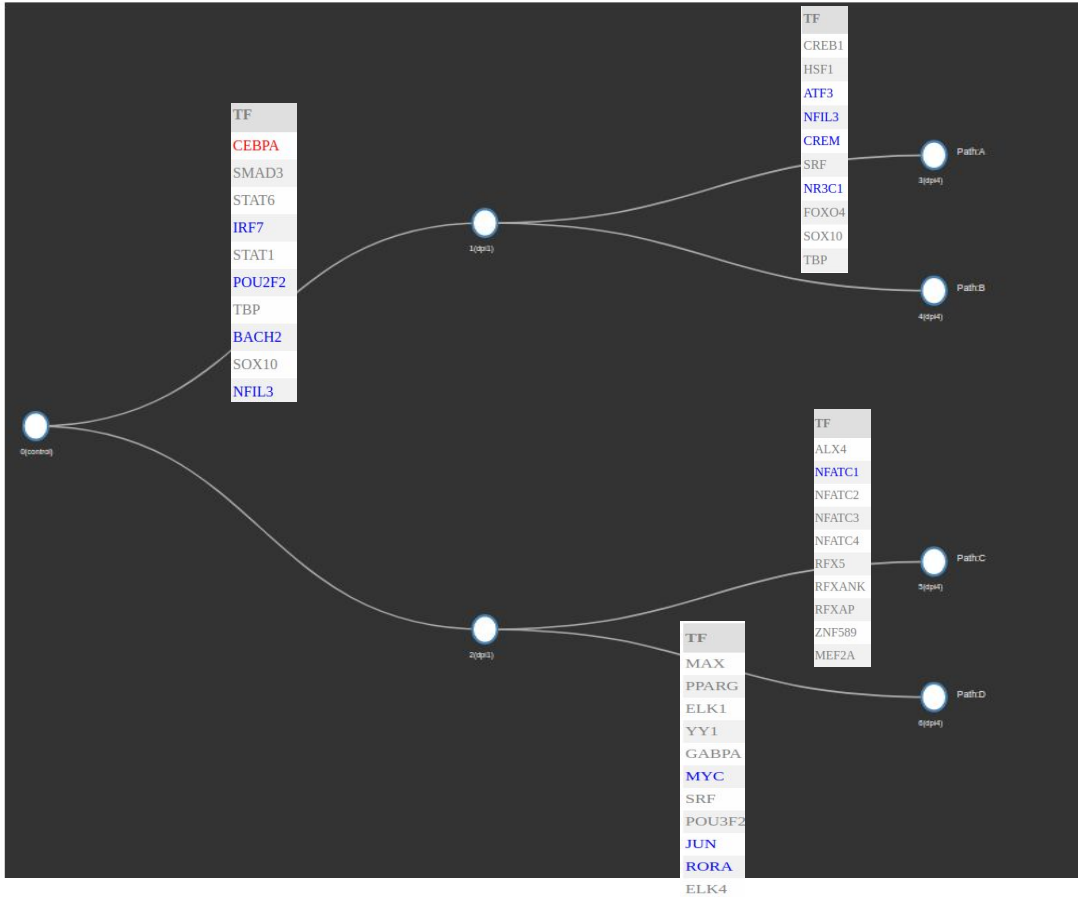
Diamond shape – Top phosphorelated protein



Detailed results can available at:
https://filedn.com/lL2xsyY8teiHHTk3wYqUmVu/results/BU_RNA_Proteomics/

The IDREM model of the RNA-seq + Proteomics data





Interactive viewer:
https://filedn.com/lL2xsyY8teiHHTk3wYqUmVu/results/BU_RNA_Proteomics/cpm.csv.log.merged.csv_viz/idrem_result.html

Using the tool you can explore gene expression levels, top TFs and the paths they regulate and protein levels of all genes.

Please refer to the manual (bottom of the panel) for a description of the iDREM model.

Top 50 proteins from msdrem single knock-out

1	Gene	Source	Top Phosphorylated	Node score
2	ATF2	N	Y	0.5338
3	NUP98	Y	Y	0.2926
4	HMGA2	N	Y	0.166
5	HDAC2	Y	Y	0.1596
6	SMARCA4	N	Y	0.1076
7	MYC	N	Y	0.0832
8	H2AFX	N	Y	0.0736
9	PKP2	Y	Y	0.069
10	EP300	N	Y	0.0586
11	CREBBP	N	N	0.0414
12	HDGF	N	Y	0.04
13	MATR3	N	Y	0.039
14	EEF2	N	Y	0.038
15	ZNF318	Y	Y	0.038
16	CHD4	N	Y	0.0342
17	JUN	N	N	0.033
18	HMGA1	N	Y	0.027
19	KRT18	N	Y	0.027
20	FOS	N	N	0.025
21	RELA	N	N	0.025
22	NCL	N	Y	0.0246
23	SUMO2	N	N	0.024
24	HSF1	N	N	0.023
25	UBE2I	N	N	0.022
26	CREB1	N	N	0.021
27	HSPB1	N	Y	0.021
28	SSRP1	N	Y	0.021
29	STAT3	N	N	0.021
30	RAB7A	Y	N	0.0204
31	CTNNB1	N	N	0.02
32	PML	N	Y	0.02
33	RB1	N	N	0.02
34	CSNK2A1	N	N	0.019
35	JUNB	N	N	0.019
36	NFKB2	N	N	0.019
37	FOXA1	N	Y	0.018
38	G3BP1	Y	N	0.018
39	MAPK8	N	N	0.018
40	PBRM1	N	Y	0.018
41	PA2G4	N	Y	0.017
42	CANX	N	Y	0.016
43	HIST1H1B	N	Y	0.016
44	SOX2	N	N	0.016
45	TBK1	Y	N	0.016
46	BRD4	Y	N	0.015
47	CEBPA	N	N	0.015
48	MYCBP2	Y	N	0.015
49	HNRNPC	N	Y	0.0146
50	TLE1	Y	N	0.0144

Please find a complete list of inferred proteins using the link below:
https://filedn.com/lL2xsyY8teiHHTk3wYqUmVu/results/BU_RNA_Proteomics/singleKnockDown_ProteinInfo.tsv

Top Phosphorylated proteins: 676 in total. They are the largest log fold change of phosphorylation (vs uninfected). Please refer to page 2 for the detailed step of getting top phosphorylated proteins.

Top protein pairs from msdrem double knock-out

1	Gene A	Gene B	Top Phosphorylated	Top Phosphorylated	Source A	Source B	Epsilon_score
2	HDAC2	NUP98	Y	Y	Y	Y	-0.054084280710472
3	NUP98	ZNF318	Y	Y	Y	Y	-0.031639974251359
4	NUP98	TCF12	Y	N	Y	Y	-0.023886891746087
5	NUP98	TLE1	Y	N	Y	Y	-0.018165796550107
6	EP300	HDAC2	Y	Y	N	Y	-0.017383578741585
7	EP300	SMARCA4	Y	Y	N	N	-0.0169261382919
8	ATF2	NUP98	Y	Y	N	Y	-0.016096475119695
9	EP300	MYC	Y	Y	N	N	-0.014742251576723
10	NUP98	PKP2	Y	Y	Y	Y	-0.014339963027297
11	NUP98	SMARCA4	Y	Y	Y	N	-0.014299673399903
12	CHD4	NUP98	Y	Y	N	Y	-0.012833236313526
13	CREBBP	EP300	N	Y	N	N	-0.011113506435719
14	EP300	ZNF318	Y	Y	N	Y	-0.010289275657974
15	BRD4	NUP98	N	Y	Y	Y	-0.010281521404165
16	NUP62	NUP98	N	Y	Y	Y	-0.009947199504145
17	CREBBP	HDAC2	N	Y	N	Y	-0.009336699939213
18	CREBBP	SMARCA4	N	Y	N	N	-0.009051626505638
19	ATF2	SMARCA4	Y	Y	N	N	-0.008642882480171
20	SMARCA4	ZNF318	Y	Y	N	Y	-0.008426600752971
21	ATF2	EP300	Y	Y	N	N	-0.008298515826746
22	CREBBP	MYC	N	Y	N	N	-0.007991135702319
23	EP300	TCF12	Y	N	N	Y	-0.007712495917196
24	ATF2	MYC	Y	Y	N	N	-0.007504625314953
25	MYC	ZNF318	Y	Y	N	Y	-0.007410521288836
26	NUP98	REEP5	Y	N	Y	Y	-0.007228736760322
27	NUP98	PLAT	Y	N	Y	Y	-0.007209177250831
28	NCL	NUP98	Y	Y	N	Y	-0.007177888353749
29	ATF2	HDAC2	Y	Y	N	Y	-0.006952181242069
30	FOXA1	NUP98	Y	Y	N	Y	-0.006639693196061
31	HDAC2	TCF12	Y	N	Y	Y	-0.006610557717119
32	NUP98	TLE3	Y	Y	Y	Y	-0.00657219011527
33	NFKB2	NUP98	N	Y	N	Y	-0.006376278210566
34	SMARCA4	TCF12	Y	N	N	Y	-0.006361744851497
35	CEBPA	NUP98	N	Y	N	Y	-0.006292984770442
36	MATR3	NUP98	Y	Y	N	Y	-0.006063401228953
37	NUP98	TBK1	Y	N	Y	Y	-0.005990594453656
38	EP300	TLE1	Y	N	N	Y	-0.005907491793951
39	HDAC2	MYC	Y	Y	Y	N	-0.005573785831696
40	CREBBP	ZNF318	N	Y	N	Y	-0.005560134845143
41	ATF2	ZNF318	Y	Y	N	Y	-0.005312975410987
42	NUP98	RELA	Y	N	Y	N	-0.005038024067538
43	HDAC2	TLE1	Y	N	Y	Y	-0.00502727805059
44	NUP98	REST	Y	N	Y	N	-0.004865073094444
45	SMARCA4	TLE1	Y	N	N	Y	-0.004838057789369
46	MYC	TCF12	Y	N	N	Y	-0.004700443394698
47	HMGA2	NUP98	Y	Y	N	Y	-0.004698165698413
48	EP300	PKP2	Y	Y	N	Y	-0.004526911503787
49	EP300	SUMO2	Y	N	N	N	-0.004422742995151
50	ATF2	TCF12	Y	N	N	Y	-0.00439821308928

Please find a complete list of inferred proteins using the link below: https://filedn.com/lL2xsyY8teiHHTk3wYqUmVu/results/BU_RNA_Proteomics/DoubleKnockDown_ProteinInfo_1k.tsv

Top Phosphorylated proteins: 676 in total. They are the largest log fold change of phosphorylation (vs uninfected). Please refer to page 2 for the detailed step of getting top phosphorylated proteins.

Protein top 100 gene list + TF

	0	1	2	3	4	5	6
0.01	0.525544	0.00018	0.194486	1	1	1	1
0.05	0.525544	9.58E-07	6.36E-10	0.050682	1	1	1
0.1	0.525544	3.69E-06	1.80E-09	0.019216	0.211704	0.028446	1
0.15	0.525544	0.000492	5.32E-11	0.000136	0.00536	0.100379	1
0.2	0.525544	0.000942	1.47E-08	1.51E-06	0.00223	0.215479	1
0.25	0.525544	0.00822	1.84E-08	2.80E-07	0.019603	0.073937	0.067354
0.3	0.525544	0.013851	3.68E-07	4.17E-08	0.007846	0.184453	0.162935

gene_symbol	hypertension	COPD	diabetes	smoke	cancer	sex	age	sum	
jun	-1	0	-1	-1	-1	-1	0	-5	GO
fos	-1	0	1	-1	-1	0	-1	-3	pos
ube2i	0	0	-1	-1	-1	0	0	-3	reg
stat6	0	0	-1	-1	-1	0	0	-3	reg
nek9	0	0	-1	-1	0	0	0	-2	reg
csde1	0	-1	0	-1	0	0	0	-2	reg
myc	-1	1	0	0	0	-1	-1	-2	reg
timm9	0	0	-1	0	0	-1	0	-2	reg
jund	-1	0	0	0	-1	0	0	-2	reg
lmna	-1	0	0	0	0	-1	0	-2	neg
foxa3	-1	0	0	0	0	0	-1	-2	pos
exosc8	0	-1	1	-1	0	-1	0	-2	pos
sox4	0	-1	-1	-1	1	0	0	-2	neg
junb	-1	1	0	0	-1	-1	0	-2	neg
rab7a	-1	-1	0	-1	0	0	1	-2	syn
reep5	0	-1	0	-1	-1	0	1	-2	neg
irf1	-1	0	1	0	-1	0	-1	-2	reg
srf	-1	1	0	-1	-1	0	0	-2	neg
mef2a	0	0	0	-1	-1	0	0	-2	neg
smarca4	0	-1	-1	-1	1	0	0	-2	neg
brd2	-1	1	-1	0	-1	-1	1	-2	pos
cbx5	1	-1	0	-1	0	0	-1	-2	reg
foxa2	-1	-1	1	-1	0	0	0	-2	pos
sox10	0	0	1	1	0	0	0	2	vira
tbk1	0	0	1	0	0	0	1	2	neg
ide	0	1	0	1	0	0	0	2	neg
rab1a	0	1	0	-1	1	0	1	2	neg
g3bp1	0	0	-1	1	1	0	1	2	neg
nup88	0	1	-1	1	1	-1	1	2	
tle3	0	1	0	-1	0	1	1	2	
sumo1	1	1	0	-1	1	0	0	2	
ctnnb1	0	0	1	0	0	0	1	2	
golga2	0	1	0	0	1	0	0	2	
hmga1	0	0	0	1	1	0	0	2	
top2a	1	0	-1	1	1	0	0	2	
cep250	1	-1	0	1	1	0	0	2	
foxm1	1	0	0	0	1	0	0	2	
pml	-1	1	0	1	0	0	1	2	
sec16a	0	1	0	0	1	0	1	3	
nup214	1	0	0	1	0	0	1	3	
h2afx	0	1	0	1	1	0	0	3	
sox2	1	0	0	1	1	0	0	3	
plau	0	0	0	1	1	1	0	3	
chd4	1	0	0	1	1	0	1	4	
csnk2a1	0	1	1	1	0	0	1	4	
ero1b	1	1	0	1	0	0	1	4	

Intersection of top genes with underlying condition genes

Enriched GO categories for intersection genes

Protein top 100 gene list + TF

	0	1	2	3	4	5	6
0.01	0.525544	0.00018	0.194486	1	1	1	1
0.05	0.525544	9.58E-07	6.36E-10	0.050682	1	1	1
0.1	0.525544	3.69E-06	1.80E-09	0.019216	0.211704	0.028446	1
0.15	0.525544	0.000492	5.32E-11	0.000136	0.00536	0.100379	1
0.2	0.525544	0.000942	1.47E-08	1.51E-06	0.00223	0.215479	1
0.25	0.525544	0.00822	1.84E-08	2.80E-07	0.019603	0.073937	0.067354
0.3	0.525544	0.013851	3.68E-07	4.17E-08	0.007846	0.184453	0.162935

gene_symbol	hypertension	COPD	diabetes	smoke	cancer	sex	age	sum
jun	-1	0	-1	-1	-1	-1	0	-5
fos	-1	0	1	-1	-1	0	-1	-3
ube2i	0	0	-1	-1	-1	0	0	-3
stat6	0	0	-1	-1	-1	0	0	-3
nek9	0	0	-1	-1	0	0	0	-2
csde1	0	-1	0	-1	0	0	0	-2
myc	-1	1	0	0	0	-1	-1	-2
timm9	0	0	-1	0	0	-1	0	-2
jund	-1	0	0	0	-1	0	0	-2
lmna	-1	0	0	0	0	-1	0	-2
foxa3	-1	0	0	0	0	0	-1	-2
exosc8	0	-1	1	-1	0	-1	0	-2
sox4	0	-1	-1	-1	1	0	0	-2
junb	-1	1	0	0	-1	-1	0	-2
rab7a	-1	-1	0	-1	0	0	1	-2
reep5	0	-1	0	-1	-1	0	1	-2
irf1	-1	0	1	0	-1	0	-1	-2
srf	-1	1	0	-1	-1	0	0	-2
mef2a	0	0	0	-1	-1	0	0	-2
smarca4	0	-1	-1	-1	1	0	0	-2
brd2	-1	1	-1	0	-1	-1	1	-2
cbx5	1	-1	0	-1	0	0	-1	-2
foxa2	-1	-1	1	-1	0	0	0	-2
sox10	0	0	1	1	0	0	0	2
tbk1	0	0	1	0	0	0	1	2
ide	0	1	0	1	0	0	0	2
rab1a	0	1	0	-1	1	0	1	2
g3bp1	0	0	-1	1	1	0	1	2
nup88	0	1	-1	1	1	-1	1	2
tle3	0	1	0	-1	0	1	1	2
sumo1	1	1	0	-1	1	0	0	2
ctnnb1	0	0	1	0	0	0	1	2
golga2	0	1	0	0	1	0	0	2
hmga1	0	0	0	1	1	0	0	2
top2a	1	0	-1	1	1	0	0	2
cep250	1	-1	0	1	1	0	0	2
foxm1	1	0	0	0	1	0	0	2
pml	-1	1	0	1	0	0	1	2
sec16a	0	1	0	0	1	0	1	3
nup214	1	0	0	1	0	0	1	3
h2afx	0	1	0	1	1	0	0	3
sox2	1	0	0	1	1	0	0	3
plau	0	0	0	1	1	1	0	3
chd4	1	0	0	1	1	0	1	4
csnk2a1	0	1	1	1	0	0	1	4
ero1b	1	1	0	1	0	0	1	4

	Homo sapiens (REF)	upload_1 (Hierarchy) NEW!					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	▲ FDR
positive regulation of nitrogen compound metabolic process	3274	30	7.22	4.15	+	5.94E-14	9.48E-10
regulation of nitrogen compound metabolic process	5895	37	13.01	2.85	+	3.28E-13	1.05E-09
regulation of transcription by RNA polymerase II	2193	25	4.84	5.17	+	2.92E-13	1.17E-09
regulation of nucleobase-containing compound metabolic process	4041	32	8.91	3.59	+	2.23E-13	1.19E-09
regulation of macromolecule metabolic process	6474	38	14.28	2.66	+	7.80E-13	1.38E-09
regulation of primary metabolic process	6088	37	13.43	2.75	+	9.62E-13	1.40E-09
negative regulation of RNA biosynthetic process	1274	20	2.81	7.12	+	7.09E-13	1.41E-09
positive regulation of macromolecule metabolic process	3621	30	7.99	3.76	+	8.87E-13	1.42E-09
positive regulation of transcription by RNA polymerase II	1259	20	2.78	7.20	+	5.71E-13	1.52E-09
negative regulation of nucleic acid-templated transcription	1272	20	2.81	7.13	+	6.89E-13	1.57E-09
symbiotic process	884	18	1.95	9.23	+	2.06E-13	1.65E-09
negative regulation of metabolic process	3098	28	6.83	4.10	+	1.24E-12	1.66E-09
regulation of RNA metabolic process	3781	30	8.34	3.60	+	2.80E-12	3.19E-09
negative regulation of RNA metabolic process	1373	20	3.03	6.60	+	2.76E-12	3.39E-09
negative regulation of transcription, DNA-templated	1219	19	2.69	7.07	+	3.75E-12	3.99E-09
positive regulation of transcription, DNA-templated	1600	21	3.53	5.95	+	4.46E-12	4.45E-09
regulation of gene expression	4873	33	10.75	3.07	+	5.67E-12	5.33E-09
positive regulation of metabolic process	3920	30	8.65	3.47	+	7.26E-12	6.10E-09
viral process	792	16	1.75	9.16	+	7.21E-12	6.40E-09
negative regulation of nucleobase-containing compound metabolic process	1477	20	3.26	6.14	+	1.03E-11	8.24E-09
negative regulation of cellular metabolic process	2607	25	5.75	4.35	+	1.38E-11	8.45E-09
negative regulation of macromolecule biosynthetic process	1500	20	3.31	6.04	+	1.36E-11	8.70E-09

Top Ranked Proteins+TFs with RNA Screen Hits Evidence

- Top ranked proteins+TFs from mSDREM analysis (179 genes)
  - 45 genes from mSDREM and condition-specific analysis

Gene	Source?	Function	Effect	Phosphorylated?
BCKDK	Y	catalyzes the phosphorylation and inactivation of the branched-chain alpha-ketoacid dehydrogenase complex	increased SARS-CoV replication	Y (24hr)
RAB7A[a]	Y	key regulator in endo-lysosomal trafficking	reduced MHV-CoV replication	N
CSNK2A1	N	serine/threonine-protein kinase that phosphorylates acidic proteins such as casein	decreased SARS-CoV replication	N
POU3F2	N	Transcription factor that plays a key role in neuronal differentiation	decreased IBV-CoV replication	N
SMAD4[*]	N	involved in transmitting chemical signals from the cell surface to the nucleus	conferred resistance to virus-induced cell death (SARS-CoV-2)	N
SMARCA4	N	encodes for BRG1, a subunit of several different protein groupings called SWI/SNF protein complexes. SWI/SNF complexes regulate gene activity thru chromatin remodeling	conferred resistance to virus-induced cell death (SARS-CoV-2)	Y (6hr & 24hr)
UBE2L	N	essential for nuclear architecture and chromosome segregation	decrease IBV-CoV replication	N

[a]not listed on weighted condition specific analysis (P-value 9.60E-02)

[*]no known drug targets but some associated compounds

P-value 1.96E-02*

P-value 4.72E-03*

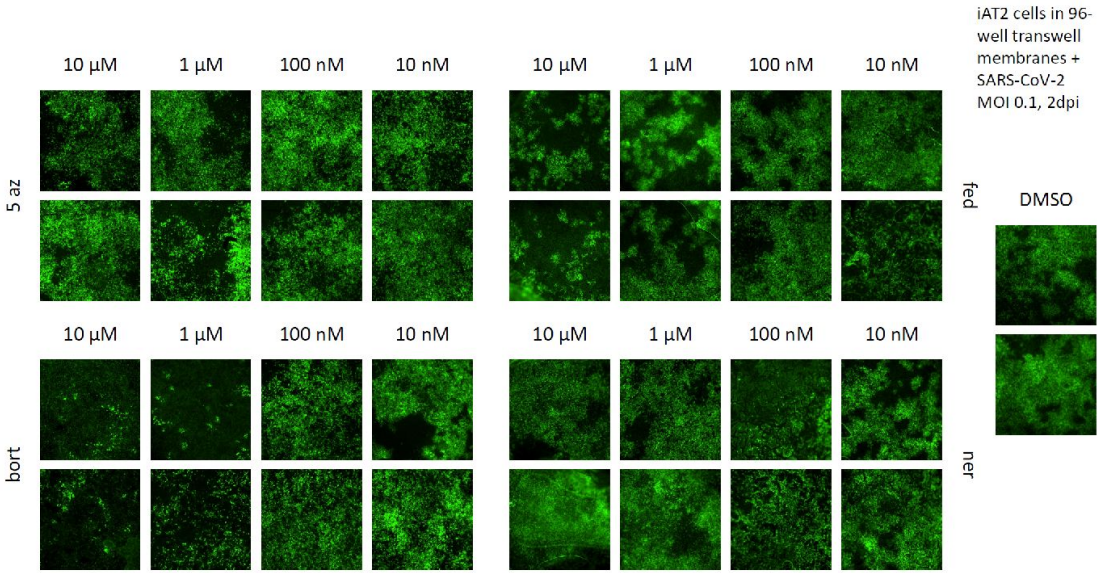
SDREM Predictions

Gene	minRank	Source	Approved drug name(s)
RELA	11	N	bortezomib; velcade
NFKB2	26	N	bortezomib; velcade
DNMT1	31	Y	azacitidine; decitabine
BRD4	32	Y	fedratinib; alprazolam
ERBB2	45	Y	ibrutinib; lapatinib; neratinib; afatinib; acalabrutinib; dacomitinib; trastuzumab emtansine; trastuzumab deruxtecan; tucatinib; pertuzumab; trastuzumab

Drugs:

1. Bortezomib (NFKb inhibitor) concentrations: 10uM, 1uM, 0.1uM, 0.01uM
    1. Stock = 50mM
    2. IC_{50} (A549s) = 0.0025 μM
2. 5-Azacytidine (DNMT1 inhibitor) concentrations: 10uM, 1uM, 0.1uM, 0.01uM
    1. Stock = 100mM
3. Fedratinib (BRD4 inhibitor) concentrations: 10uM, 1uM, 0.1uM, 0.01uM
    1. Stock = 50mM
    2. IC_{50} (Caco-2) = 2.1-6.5uM, (HEK293) 1.2uM
4. Neratinib (ERBB2 inhibitor) concentrations: 10uM, 1uM, 0.1uM, 0.01uM
    1. Stock = 10mM
    2. IC_{50} (MDA and other cancer cell lines) = <0.005uM or 1-10uM

Apical: 30 minute pre-treatment only, treat apically with virus for 1 hr, then only basolaterally for remainder of experiment



Summary -II

1) iDREM framework could be extended to study infectious disease (signaling networks + regulatory networks)

2) Integration of multi-omics data could lead to the discovery of novel drug for COVID

Thanks

Collaborators

Ziv Bar-Joseph (CMU)

Naftali Kaminski (Yale)

Funding :

CIHR IRSC
Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

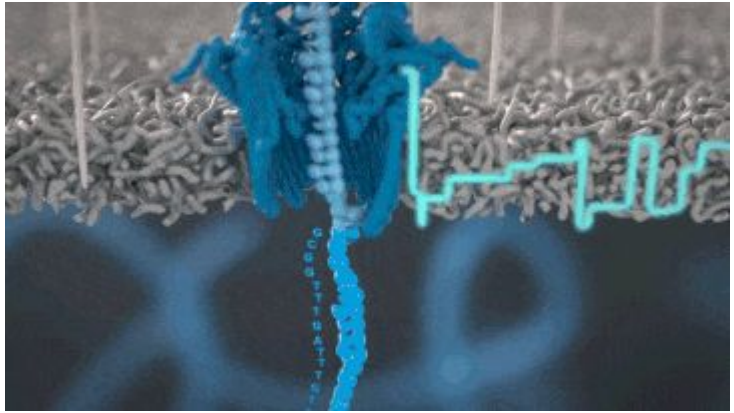
Fonds de recherche Santé Québec

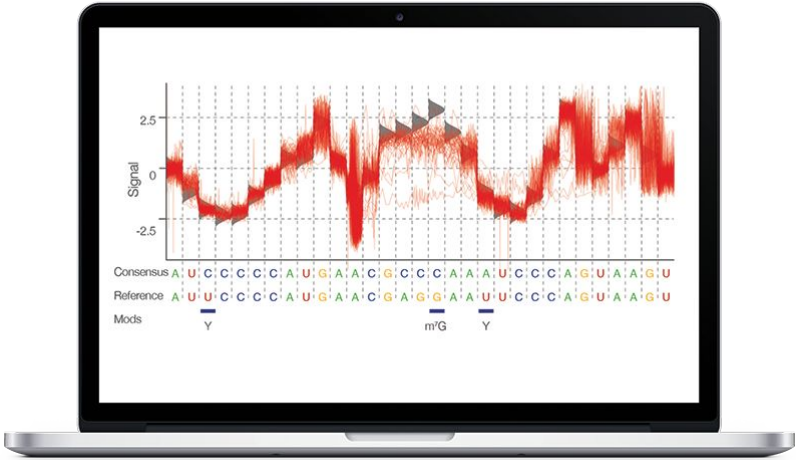
THREE LAKES FOUNDATION

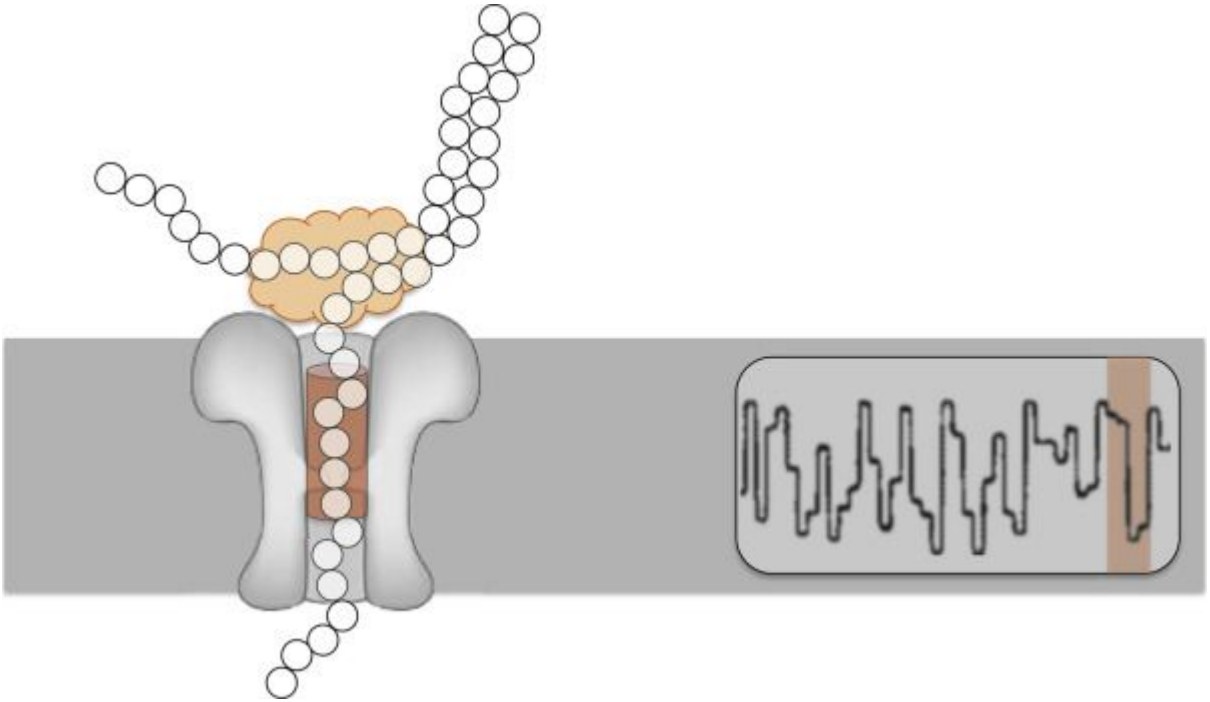
Nanopore sequencing

Third-generation sequencing

https://nanoporetech.com/applications/dna-nanopore-sequencing




2.5
0
-2.5
Signal
Consensus A U C C C C C A U G A A C G C C C A A A U C C C A G U A A G U
Reference A U U C C C C A U G A A C G A G G A A U U C C C A G U A A G U
Mods
Y
m7G
Y



https://www.sciencedirect.com/topics/neuroscience/nanopore-sequencing