CS286: AI for Science and Engineering

Project 6: Predicting the response of cellular transcriptome to gene perturbation

PIs: Dr. Lichun Jiang (蒋立春)[1], Dr. Jie Zheng (郑杰)[2]

TA：Mr. Yimiao Feng (冯艺苗)[2]

[1] Shanghai Institute for Advance Immunochemical Studies(SIAIS),

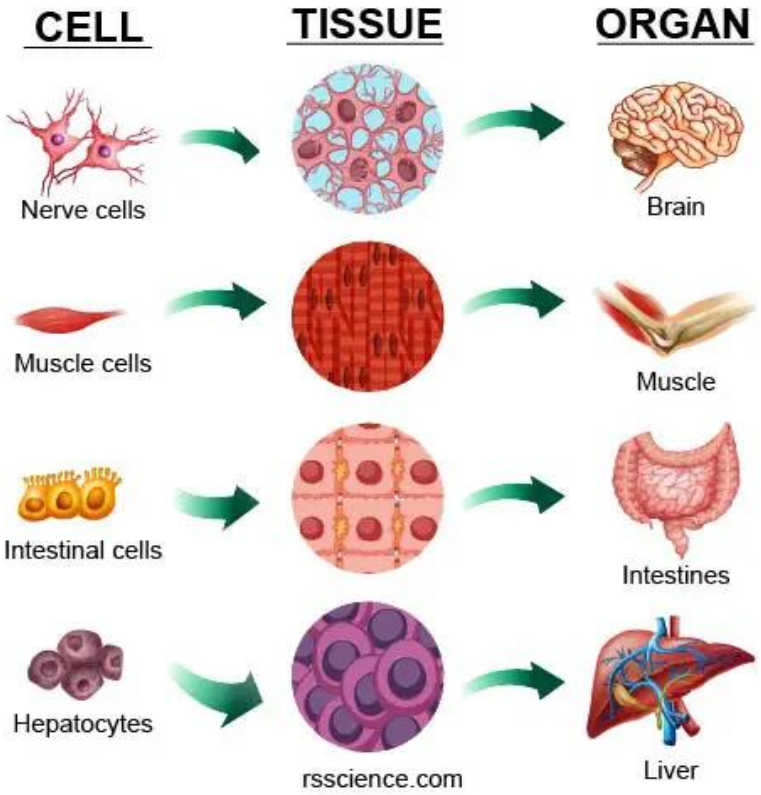
[2] School of Information Science and Technology (SIST)

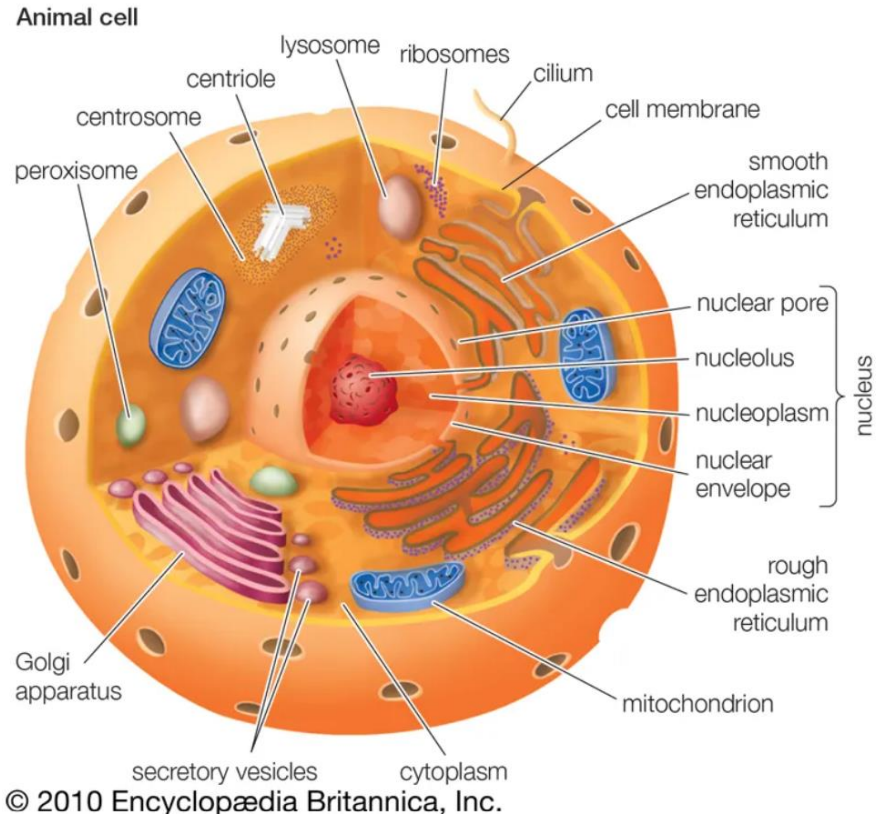
ShanghaiTech University

Fall, 2025

Biological Background of Virtual Cell

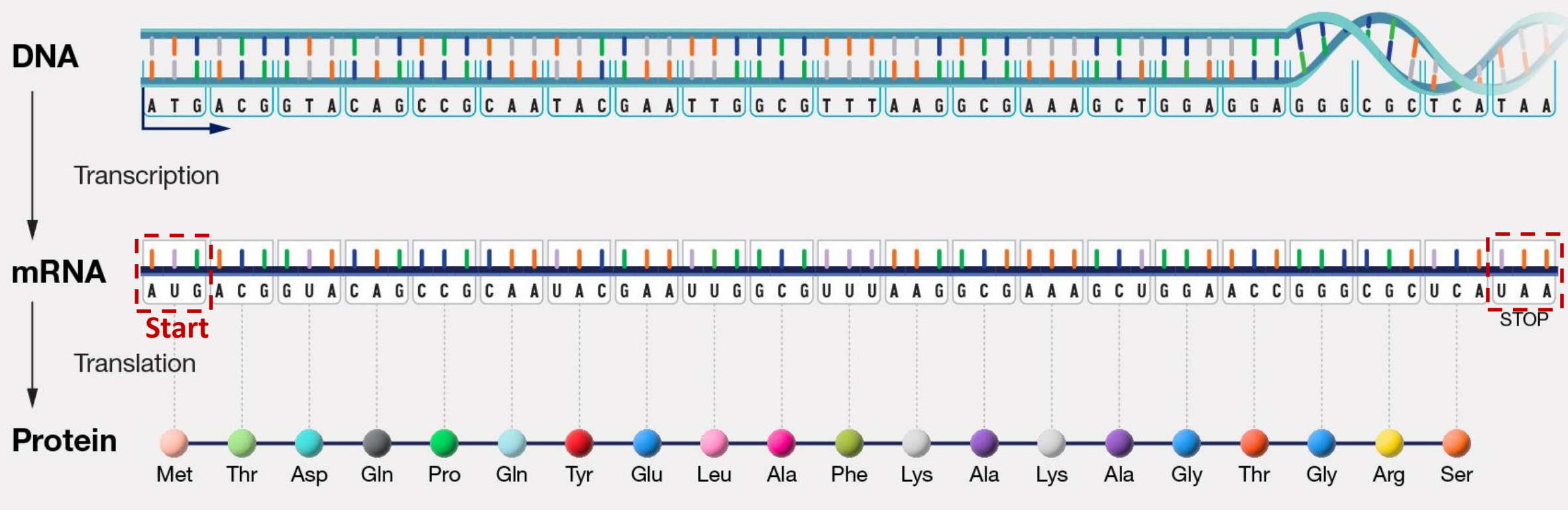
Cells are the basic units of life





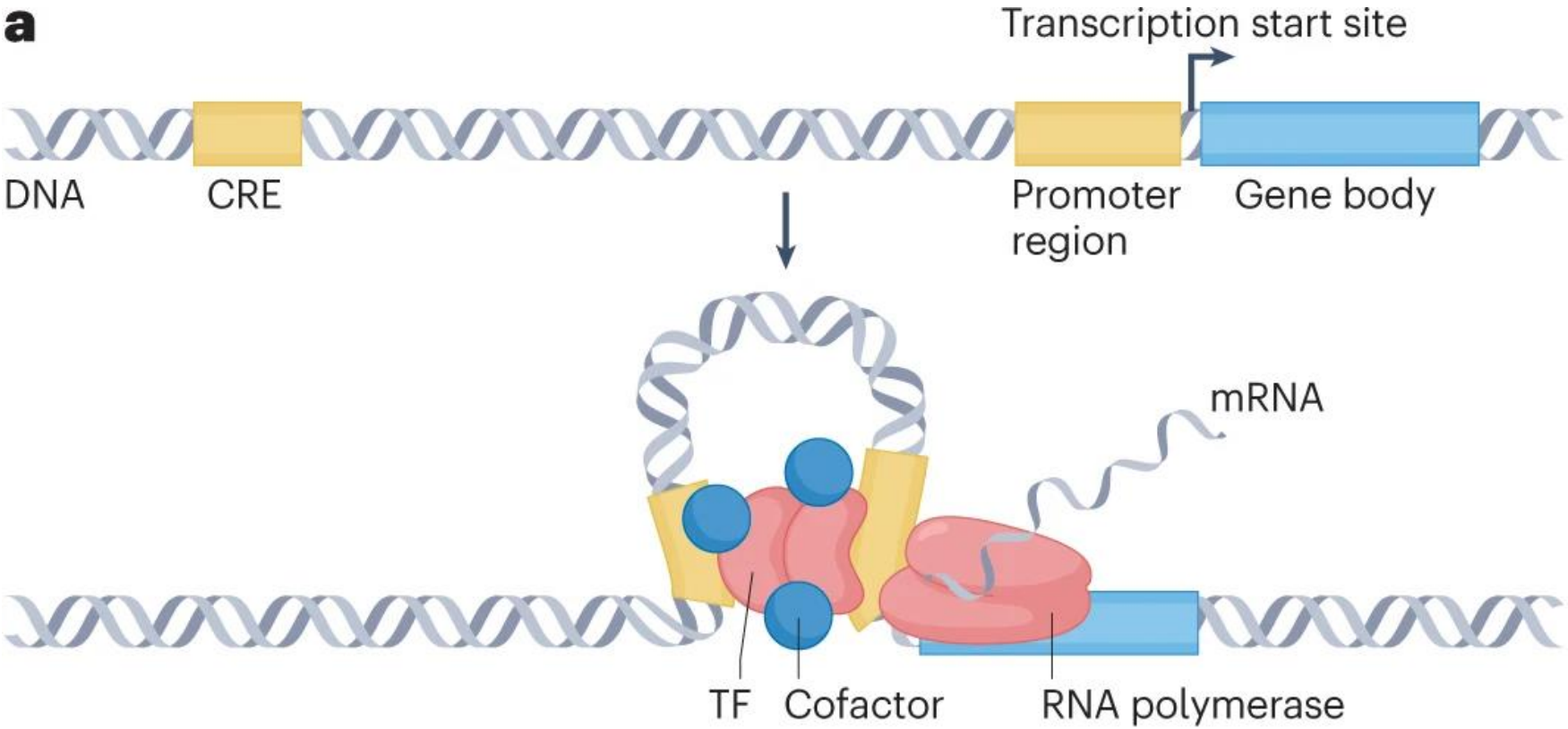
https://worksheetaidansuskm.z13.web.core.windows.net/cells-organs-tissues-and-systems.html

Central dogma of molecular biology



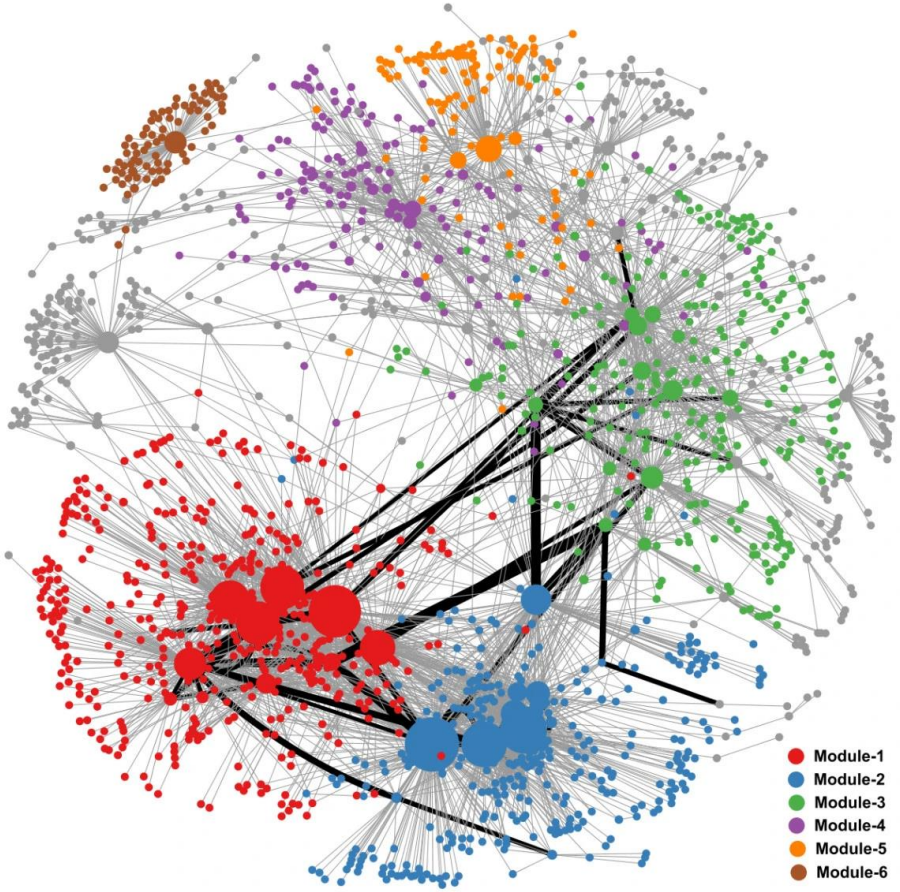
https://www.genome.gov/es/genetics-glossary/Central-Dogma

Gene regulatory network at molecular level



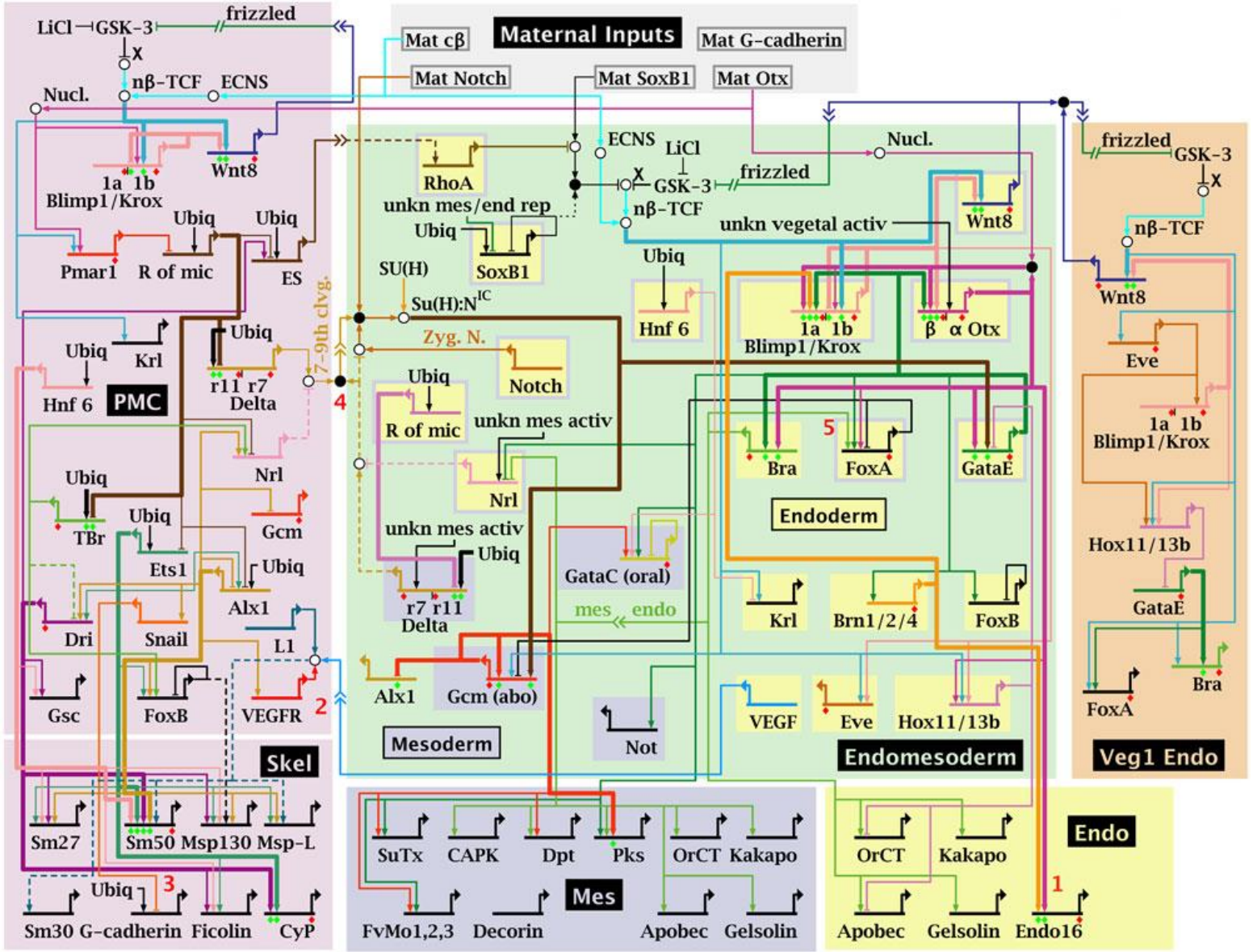
Nat Rev Genet **24**, 739–754 (2023).

Gene regulatory network is not random



Gu, Z., Zhang, C. & Wang, J. Gene regulation is governed by a core network in hepatocellular carcinoma. *BMC Syst Biol* **6**, 32 (2012)

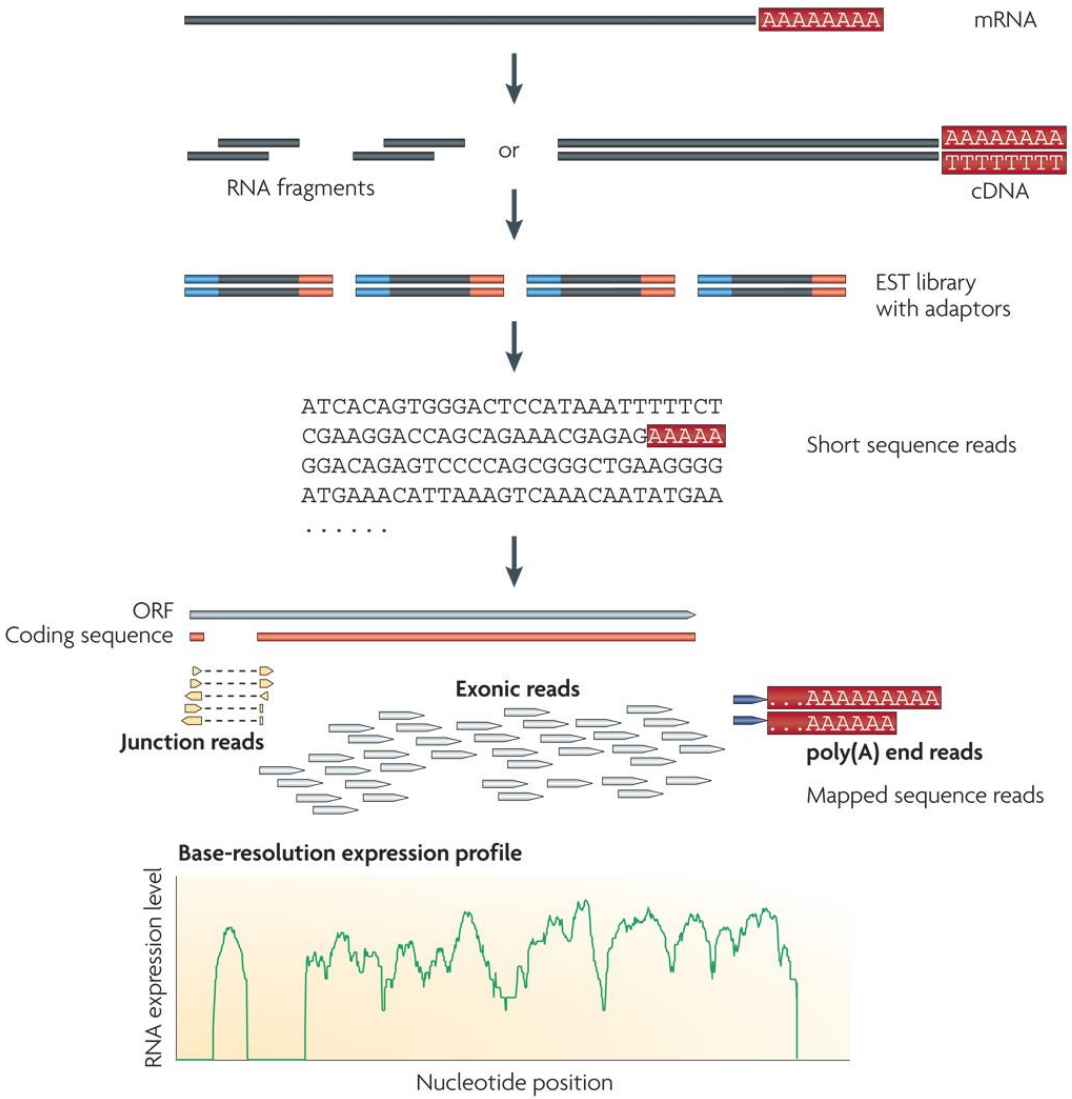
The sea urchin gene regulatory network



随机波动
细胞类型
外界环境
...

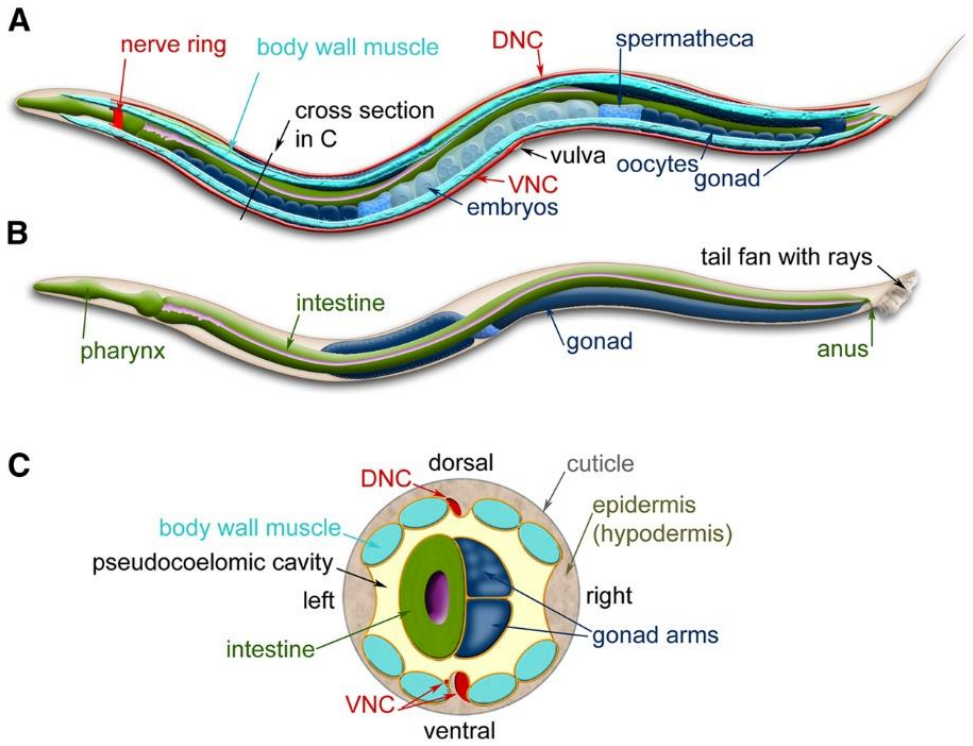
Sci. STKE **2006**,pe47-pe47(2006).

RNA Measurement on cells is relatively convenient

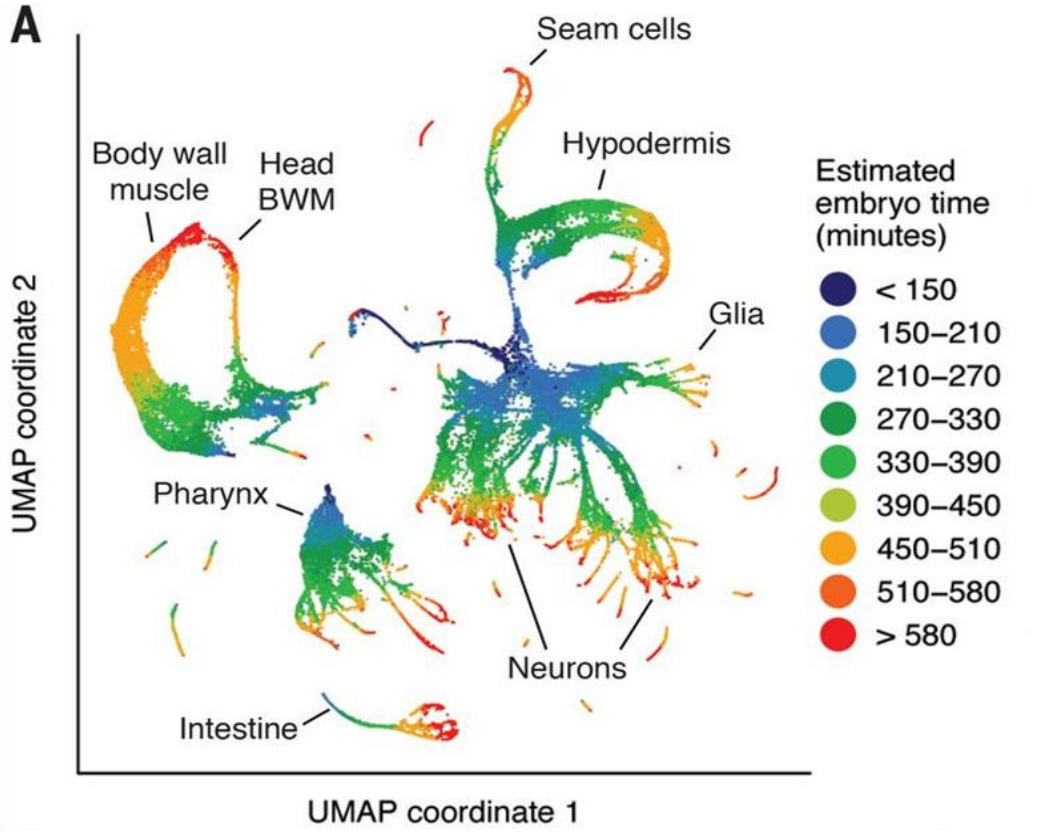


Wang et. al. Nat Rev Genet. 2009 Jan; 10(1): 57–63.

scRNA-seq zoom in single-cell RNA expression

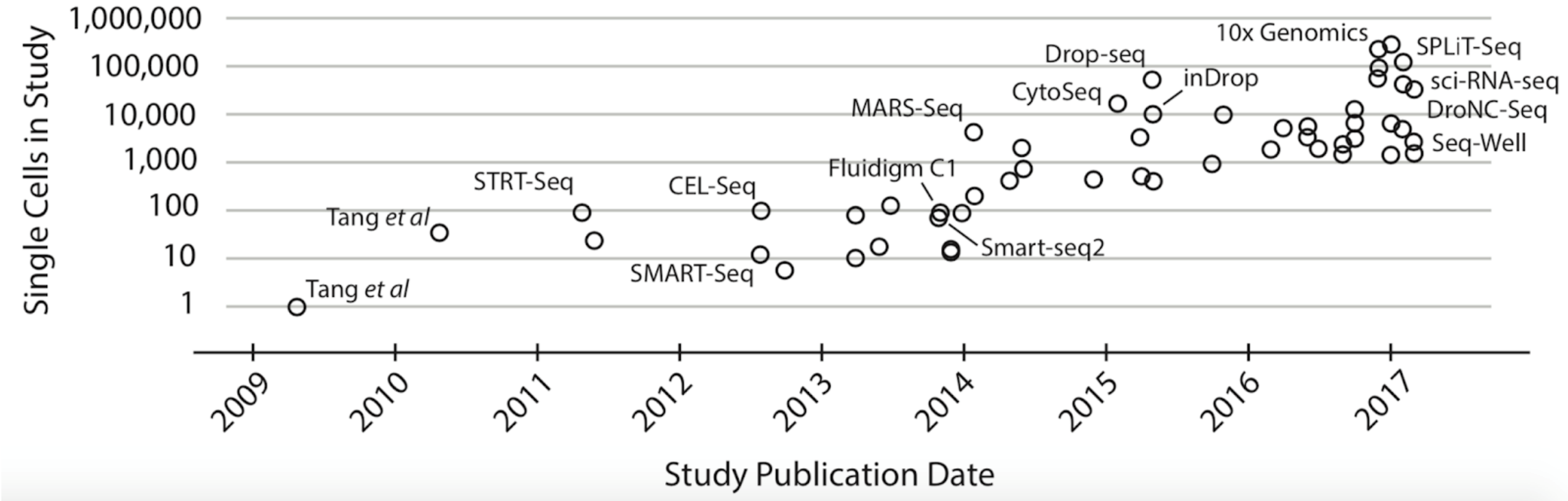


http://www.wormbook.org



Jonathan S. Packer et al. Science 2019;science.aax1971

Variety of single-cell RNA sequencing techniques



Scale limitation,Batch effects, Technique bias

Svensson V, Vento-Tormo R and Teichmann SA. (2018) Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols 13;4;599-604)

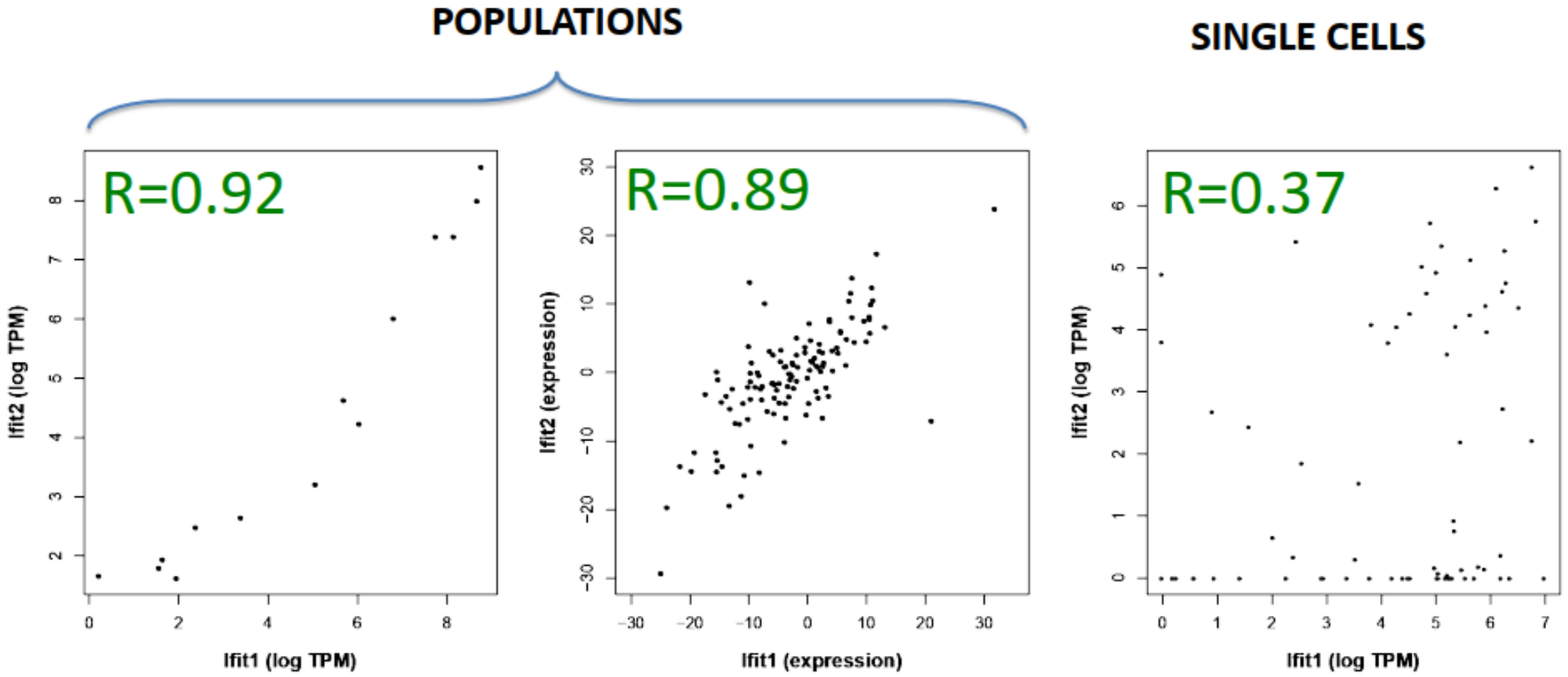
Count matrix of scRNA-seq data

Toy datasets

G/C	C1	C2	C3	…	Cn
G1	2	5	0	…	8
G2	0	19	40	…	0
G3	230	0	398	…	0
….	…	…	…	…	…
Gm	0	0	0	….	3000

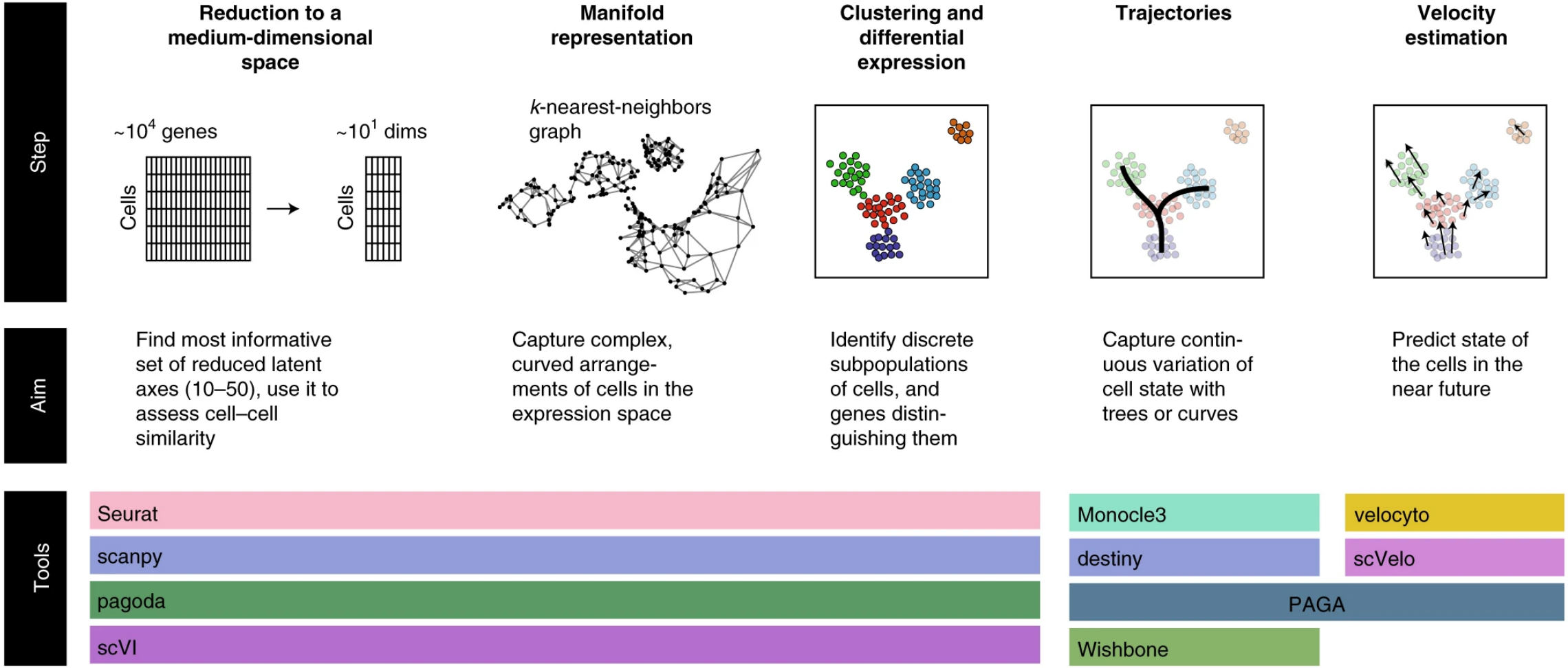
In a single piece of data:
m in the scale of 10K, **sparse matrix with a lot of zeros**
n in the scale of 10K or more, dynamic range
The n will increase with more data and studies, and technique improvement

Correlation is not well-suited for single-cell analysis——both biology and technique limit



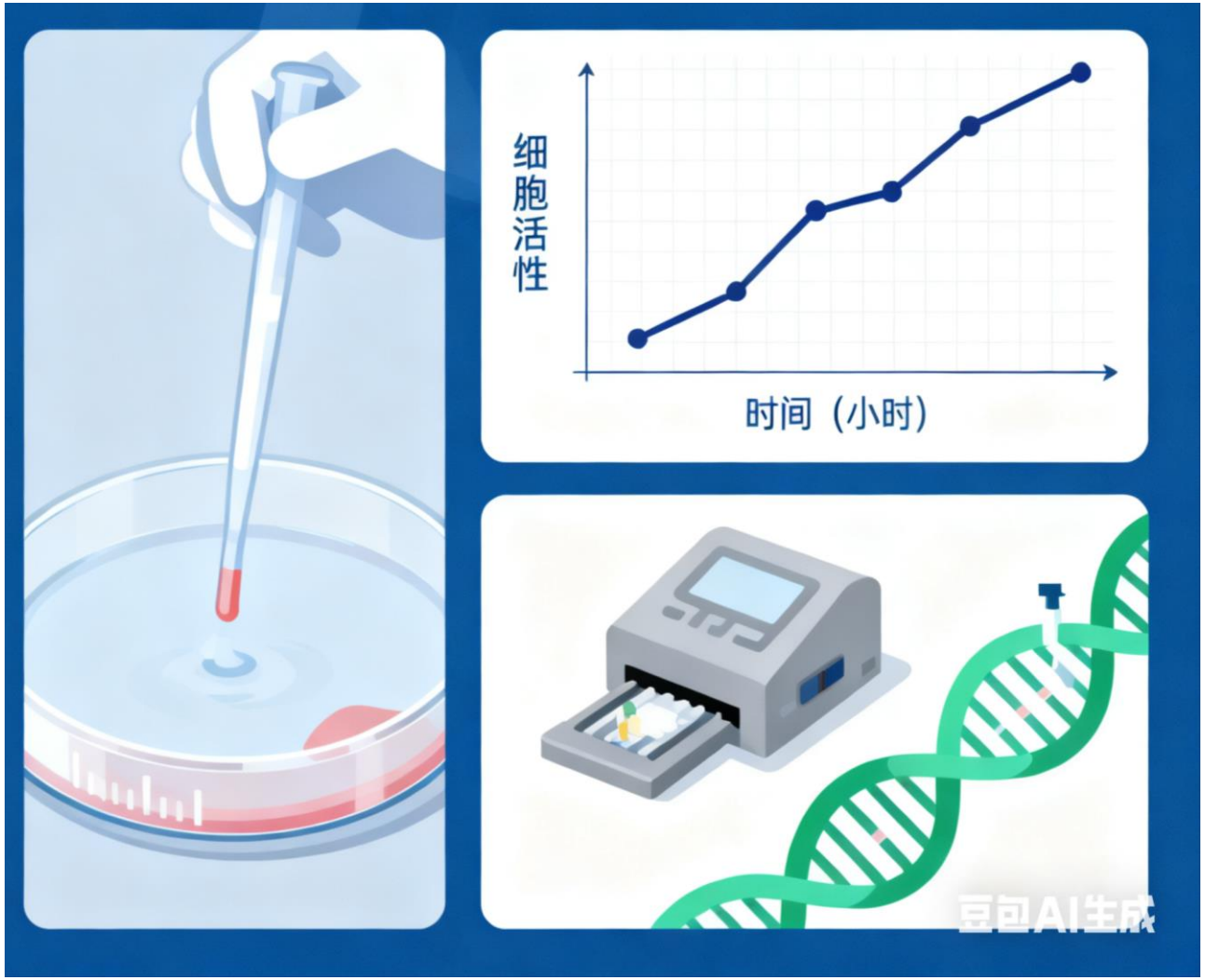
Slide courtesy of Manolis Kellis

Key analysis steps in single-cell RNA-seq analysis



Kharchenko, P.V. The triumphs and limitations of computational methods for scRNA-seq. *Nat Methods* **18**, 723–732 (2021).

Typical functional study——perturb and measure



图片为豆包生成

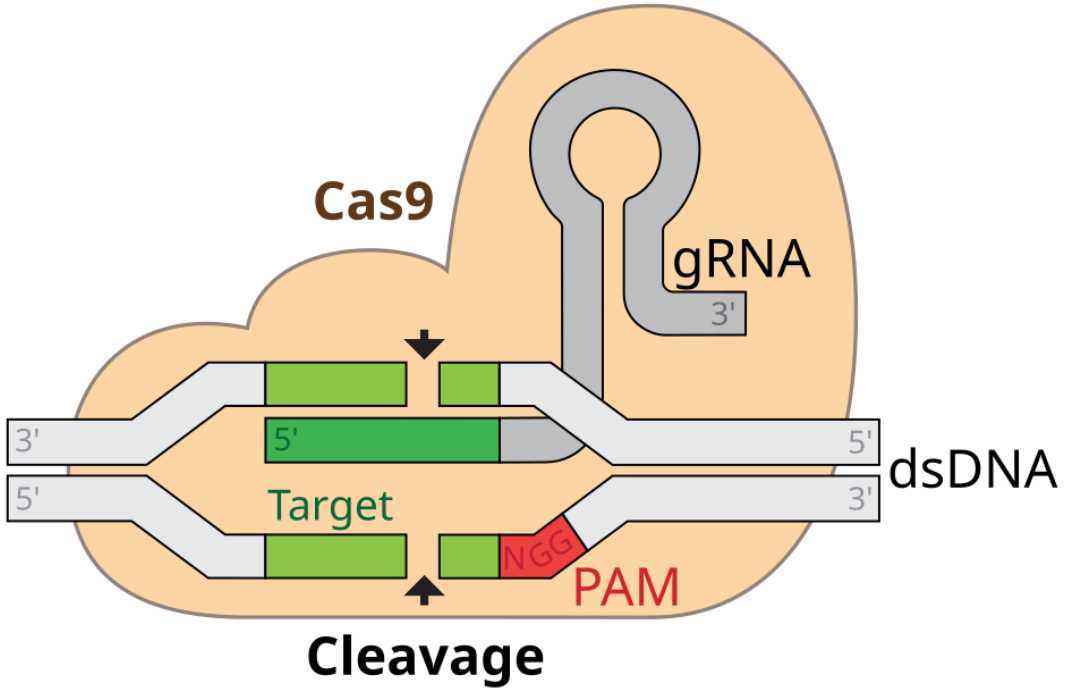
一般只有几个静态截面观察

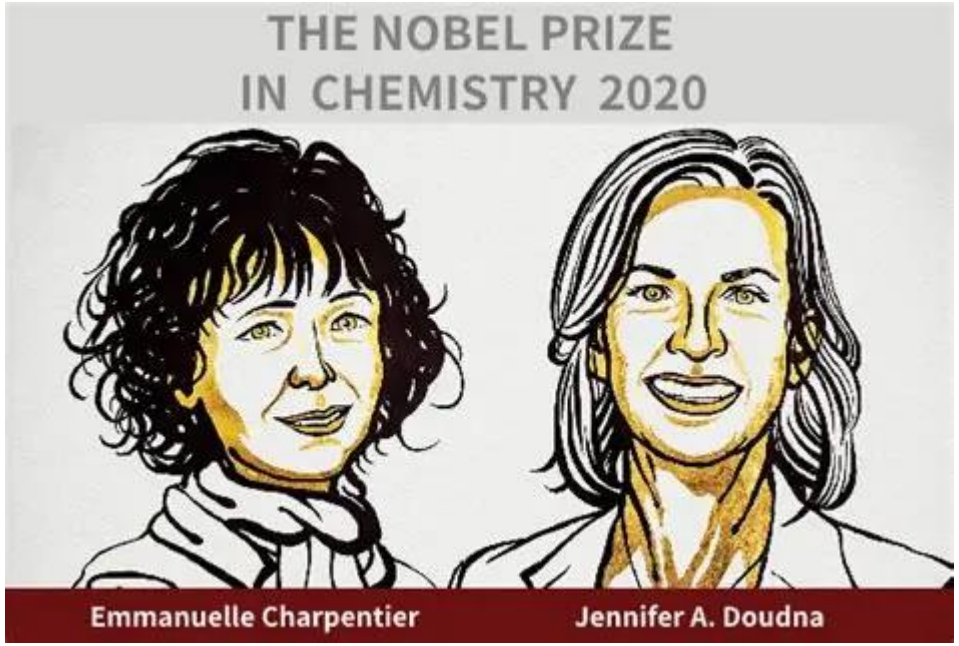
一次干扰一个因素通量低

实验周期长代价高

一般情况下同一个样品被观察就被消耗掉了

CRISPR makes perturbation much easier

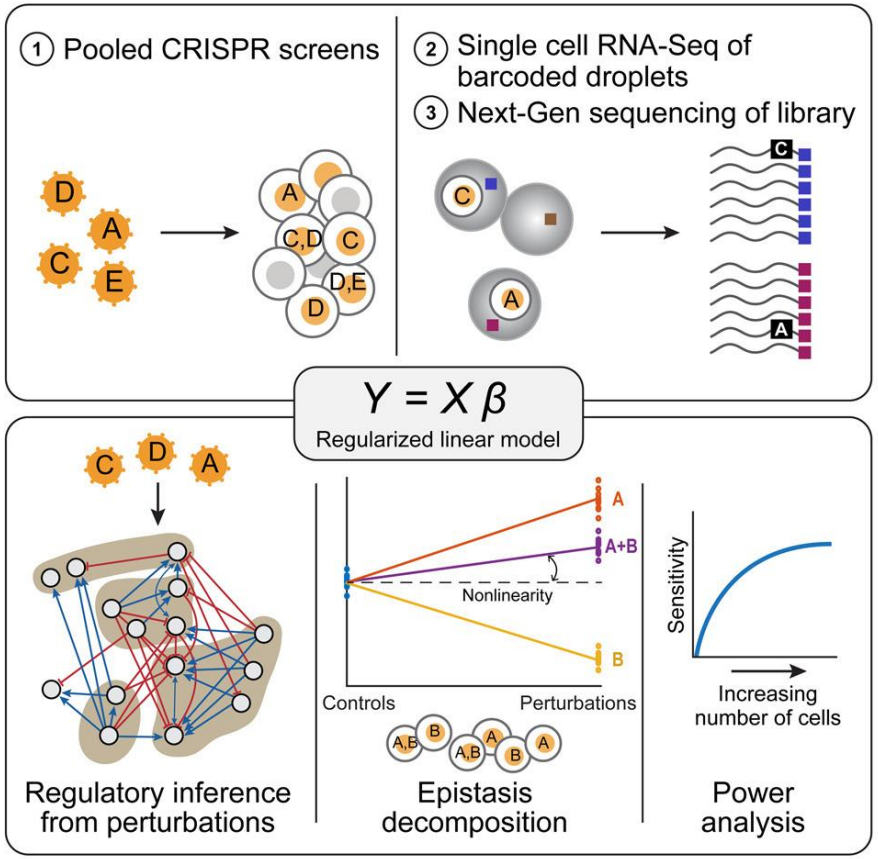




https://www.nobelprize.org/

https://www.sciencenewstoday.org/crispr-explained-gene-editing-and-the-future-of-medicine

Perturb-seq: scaled single-cell level perturbation and measurements



Data is like:

Cell 1, matrix of gene expression, A gene perturbed

Cell 2, matrix of gene expression, A gene perturbed

....

Cell N, matrix of gene expression, B gene perturbed

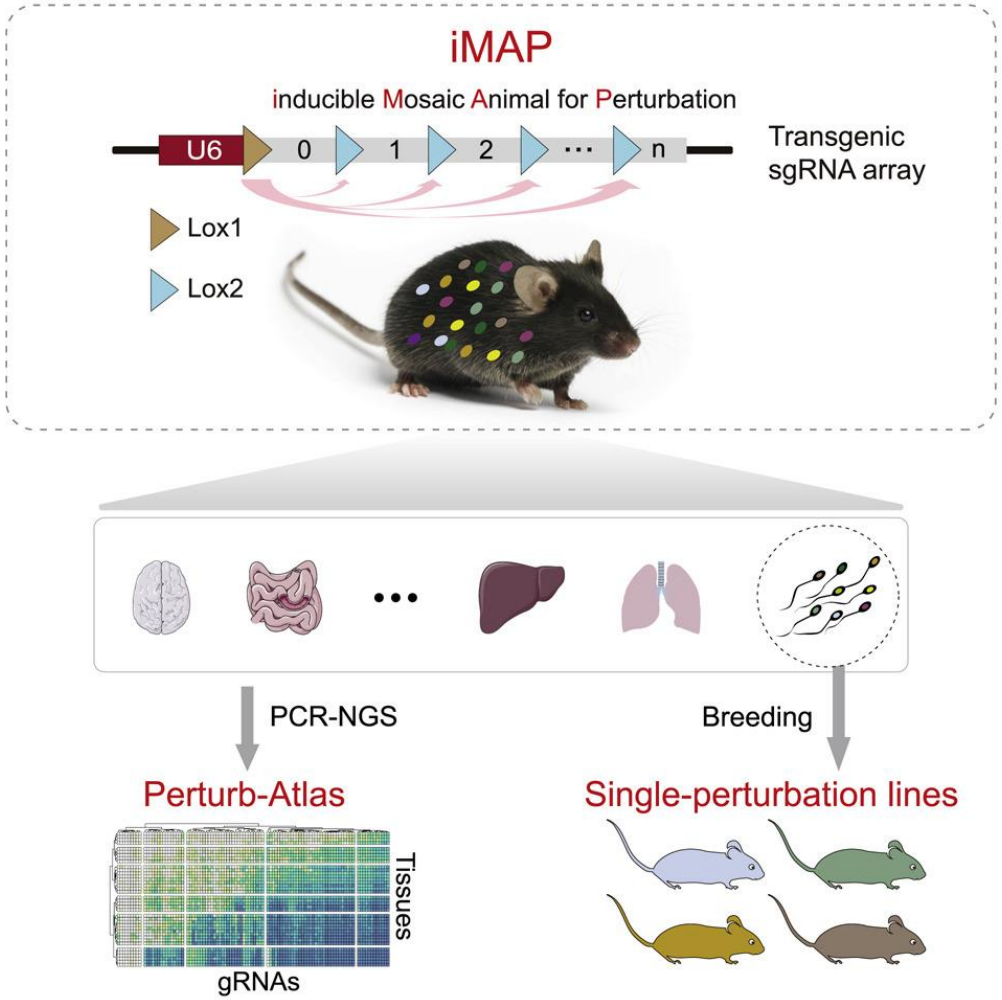
Cell N+1, matrix of gene expression, B gene perturbed

....

....

Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens
Dixit, Atray et al. Cell, Volume 167, Issue 7, 1853 - 1866.e17

In vivo perturb-seq : iMAP



我们几乎不能利用人体做实验！

Bo Liu, Zhengyu Jing, Xiaoming Zhang, *et.al.* Large-scale multiplexed mosaic CRISPR perturbation in the whole organism, Cell, Volume 185, Issue 16, 2022,Pages 3008-3024.

Virtual Cell Challenge (VCC)

From biological data to virtual cell

生物学概念□	信息科学翻译
scRNA-seq	高维、稀疏、带噪快照
CRISPR干扰	干预样本
基因调控网络	有向图 + 动力学系统
发育过程	状态机 + 轨迹
虚拟细胞	从快照+干预中反推状态转移函数

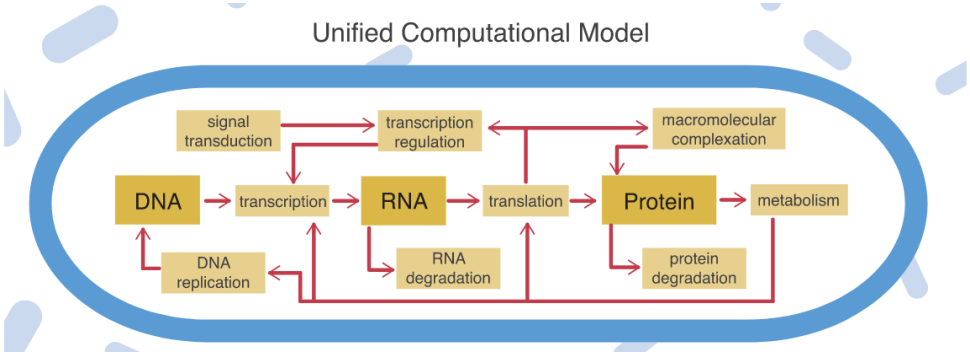
虚拟细胞 = 从稀疏干预数据中反推函数 f()，让“湿实验”在硅片上连续播放

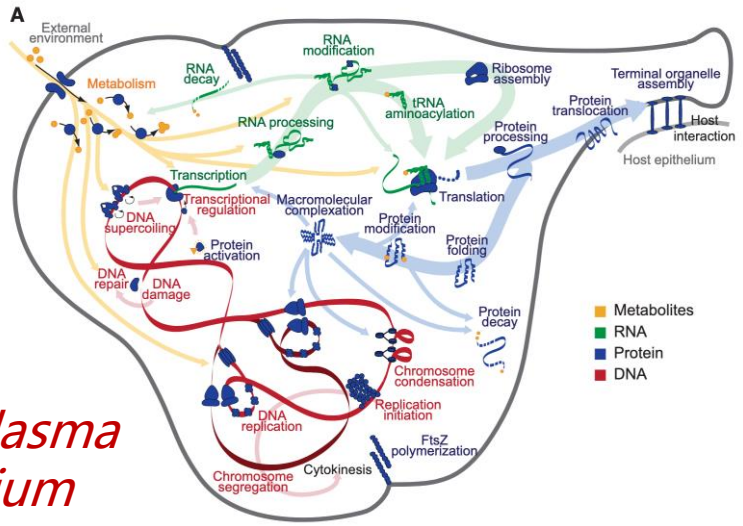
Virtual cell can make cellular study more efficient

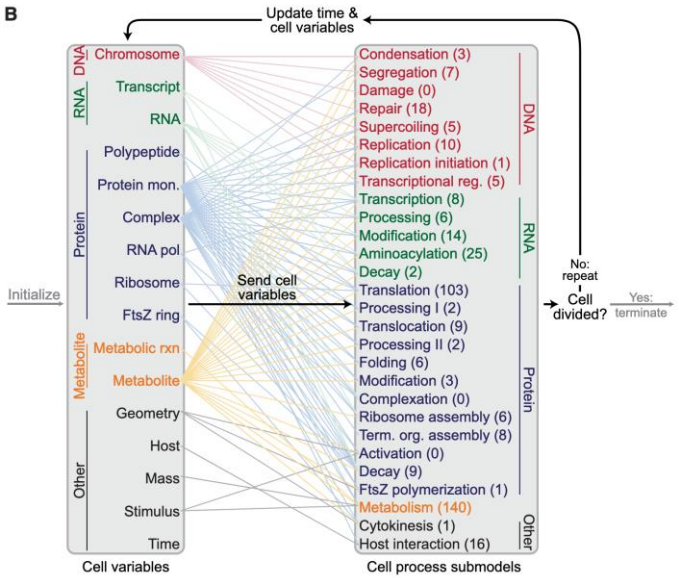
上海科技大学 ShanghaiTech University

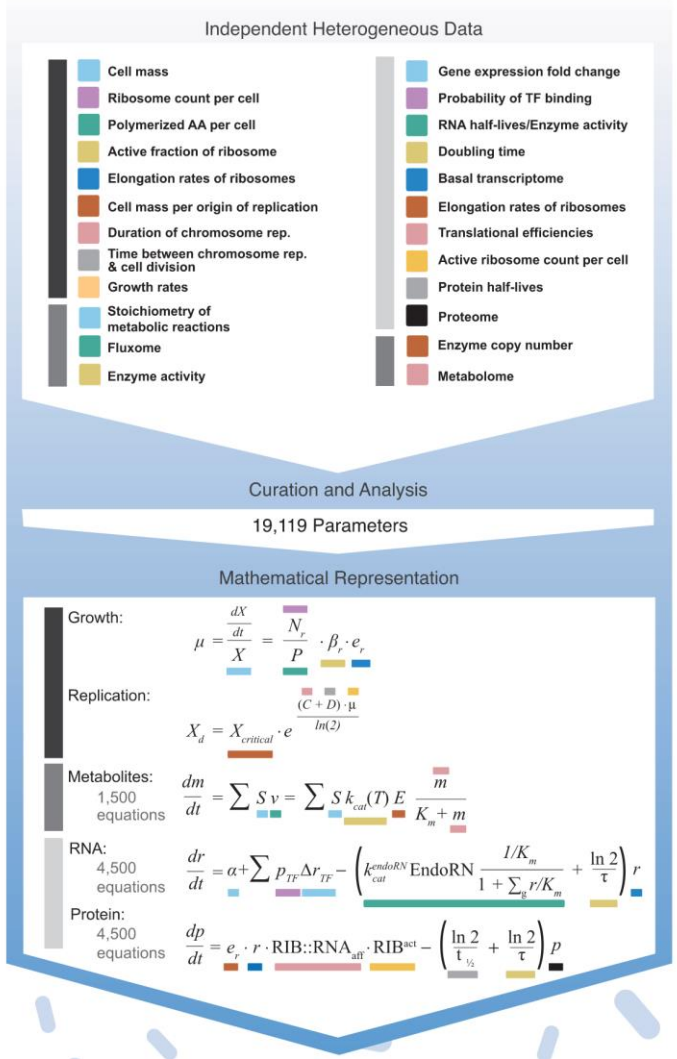
Rule-based "whole cell modeling"

E. coli









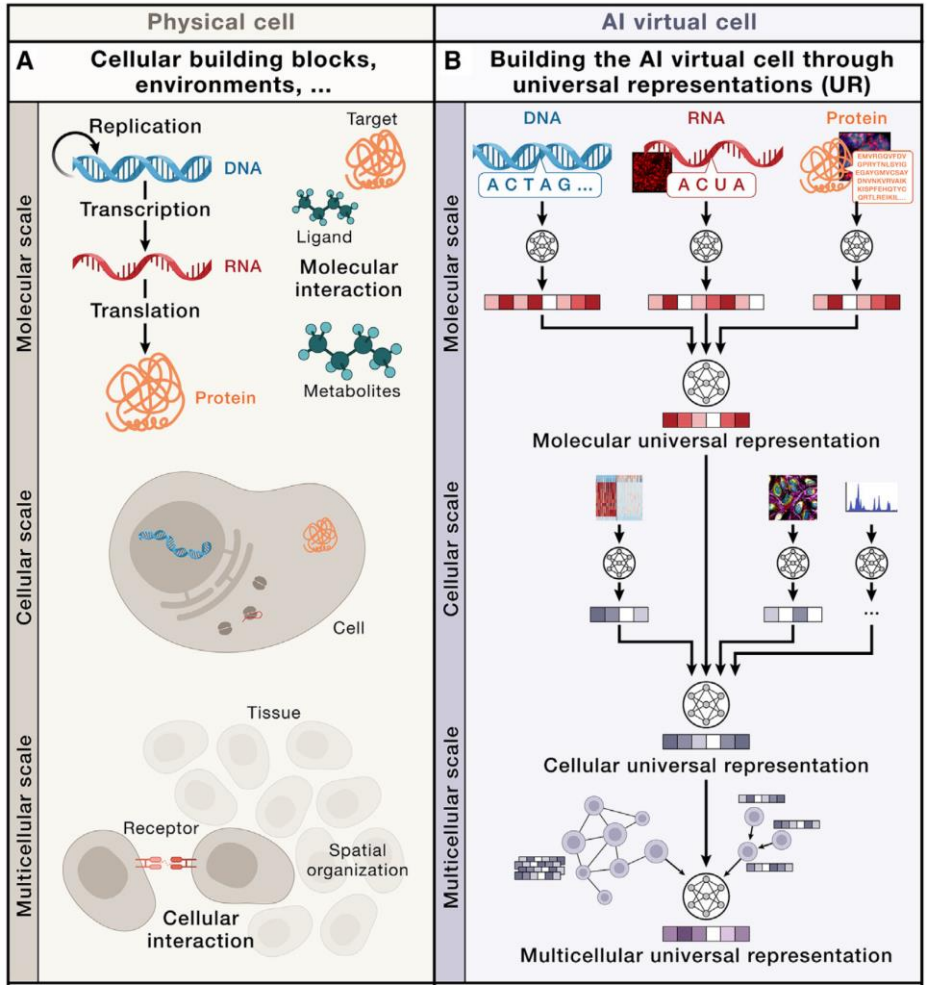
Mycoplasma genitalium

Karr, J. R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401 (2012).

Macklin, D. N. *et al.* Simultaneous cross-evaluation of heterogeneous *E. coli* datasets via mechanistic simulation. *Science* **369**, (2020).

Build the virtual cell with AI

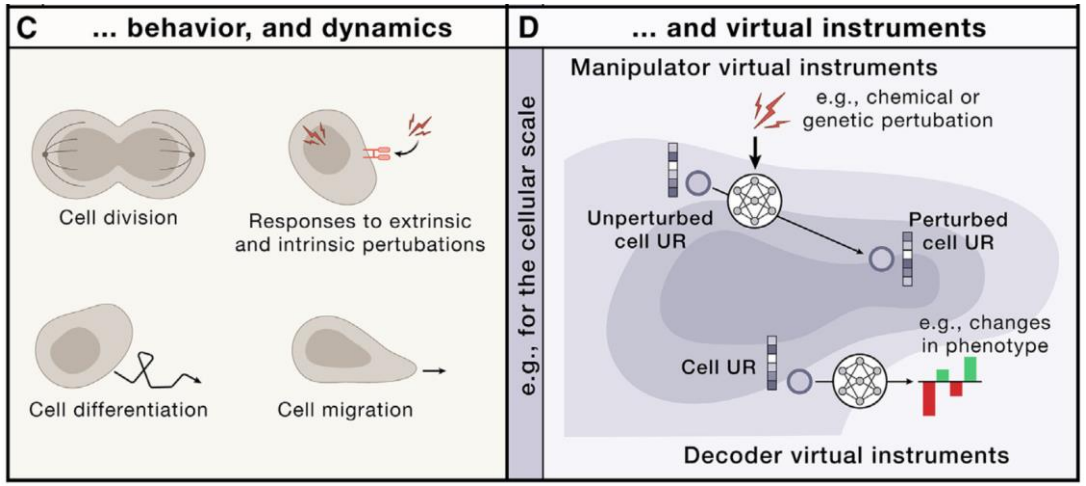
Cell
Leading Edge
CellPress
OPEN ACCESS



Perspective

How to build the virtual cell with artificial intelligence: Priorities and opportunities

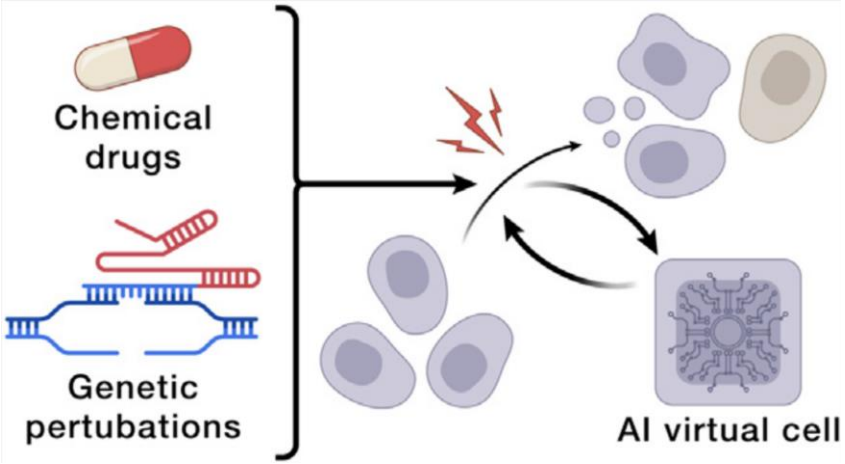
Charlotte Bunne,[1,2,3,4,50] Yusuf Roohani,[1,3,5,50] Yanay Rosen,[1,3,50] Ankit Gupta,[3,6] Xikun Zhang,[1,3,7] Marcel Roed,[1,3] Theo Alexandrov,[8,9] Mohammed AlQuraishi,[9] Patricia Brennan,[3] Daniel B. Burkhardt,[11] Andrea Califano,[10,12,13] Jonah Cool,[3] Abby F. Dernburg,[14] Kirsty Ewing,[3] Emily B. Fox,[1,15,16] Matthias Haury,[17] Amy E. Herr,[16,18] Eric Horvitz,[19] Patrick D. Hsu,[5,18,20] Viren Jain,[21] Gregory R. Johnson,[22] Thomas Kalil,[23] David R. Kelley,[24] Shana O. Kelley,[25,26] Anna Kreshuk,[27] Tim Mitchison,[28] Stephani Otte,[17] Jay Shendure,[29,30,31,32] Nicholas J. Sofroniew,[33] Fabian Theis,[34,35,36] Christina V. Theodoris,[37,38] Srigokul Upadhyayula,[14,16,39] Marc Valer,[3] Bo Wang,[40,41] Eric Xing,[42,43] Serena Yeung-Levy,[1,44] Marinka Zitnik,[45,46,47] Theofanis Karaletsos,[3,*] Aviv Regev,[2,*] Emma Lundberg,[3,6,7,48,*] Jure Leskovec,[1,3,*] and Stephen R. Quake[3,7,49,*]



Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045–7063 (2024).

Perturbation response prediction

- **Application scenarios:**
  - Drug target discovery
  - Gene Ontology
  - Research on Disease Mechanisms
- **Challenges:**
  - Large perturbation space (>20k genes)
  - Nonlinear and combined effects
  - Data scarcity (especially in human cells)

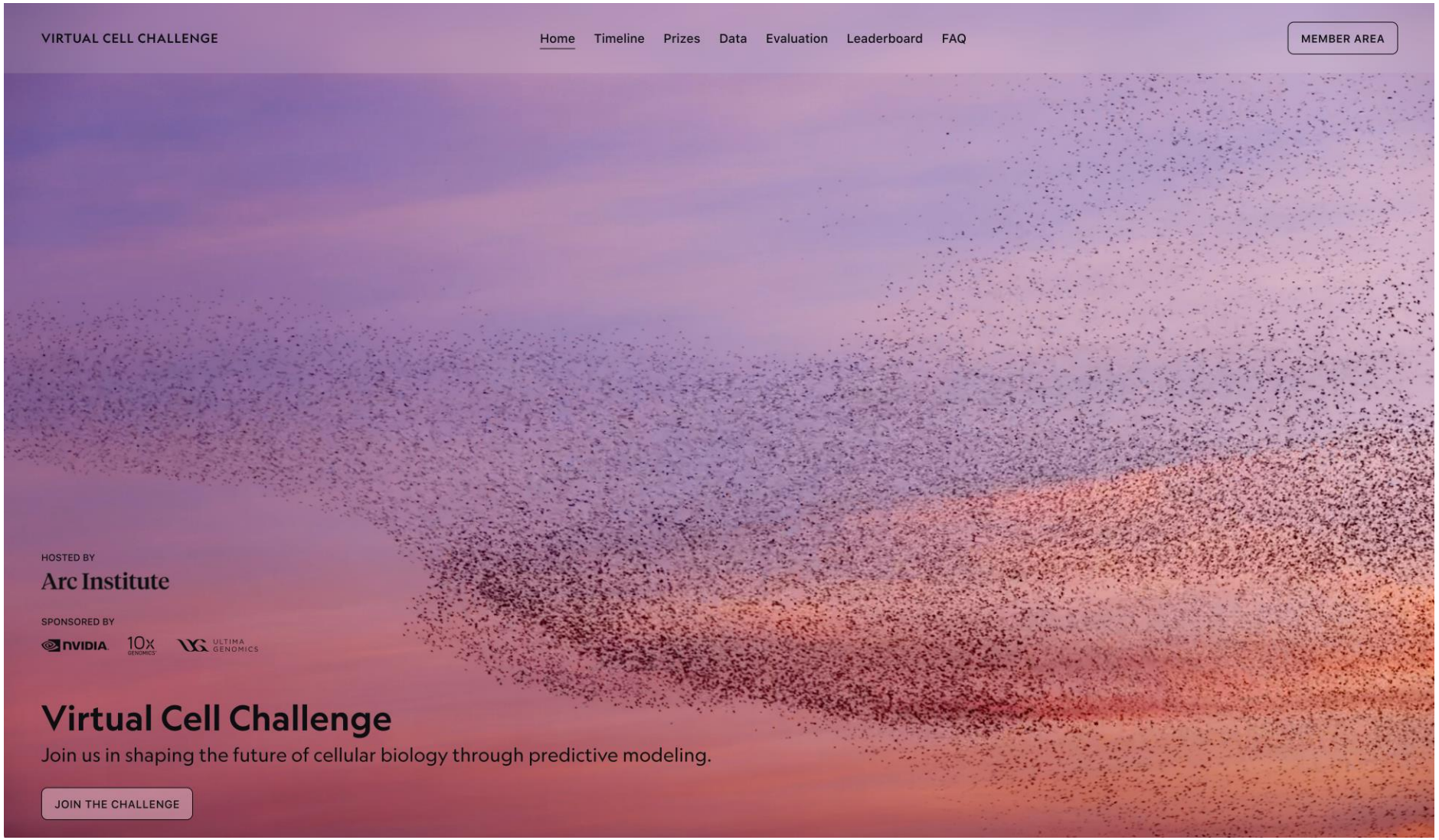


Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045–7063 (2024).

Virtual Cell Challenge (VCC)

上海科技大学
ShanghaiTech University

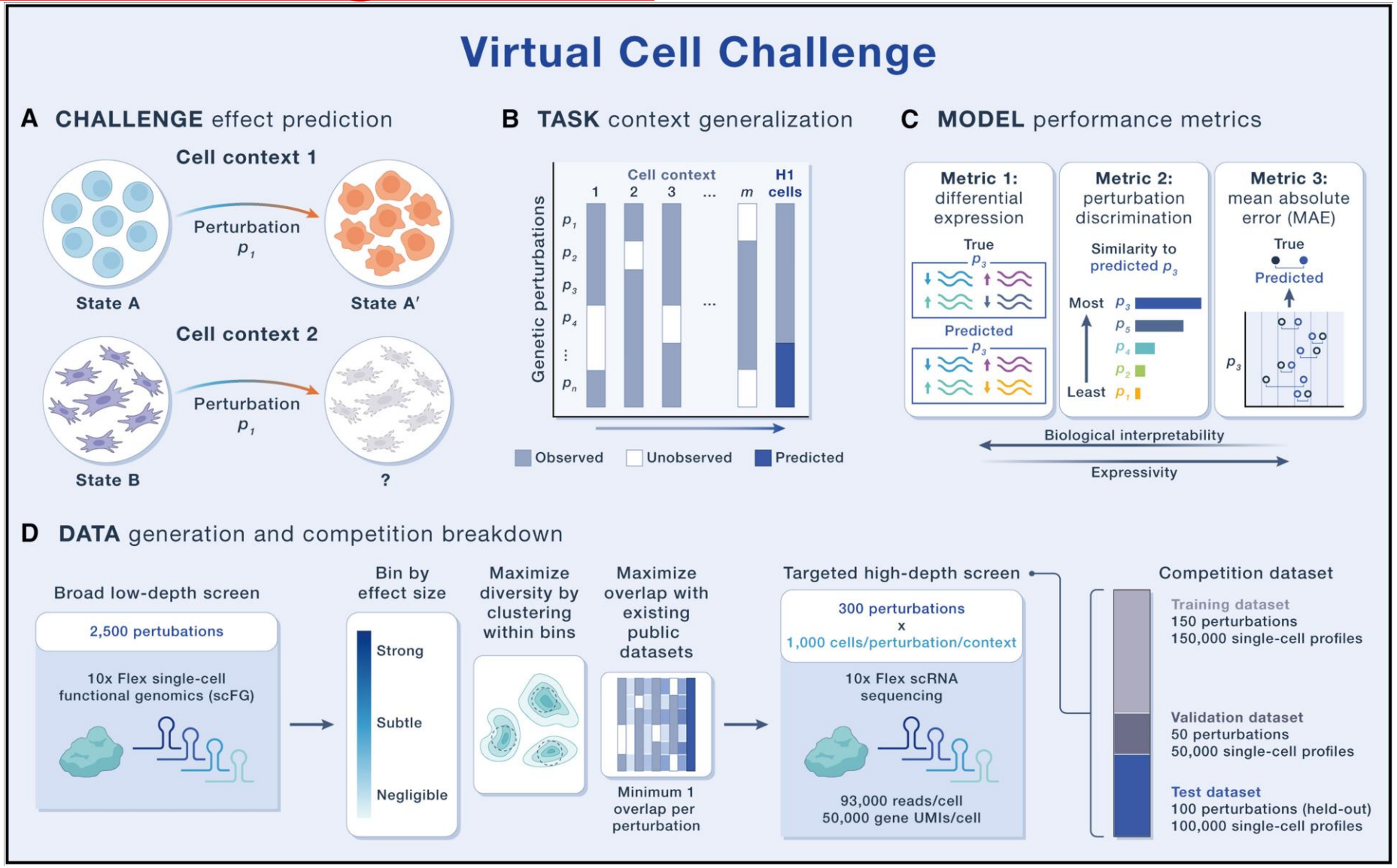
Based on single-cell RNA seq data, predict gene expression changes under unknown perturbations.



https://virtualcellchallenge.org/

Virtual Cell Challenge (VCC)

Based on single-cell RNA seq data, predict gene expression changes under unknown perturbations.



Roohani, Yusuf H., et al. "Virtual Cell Challenge: Toward a Turing test for the virtual cell." Cell 188.13 (2025): 3370-3374.

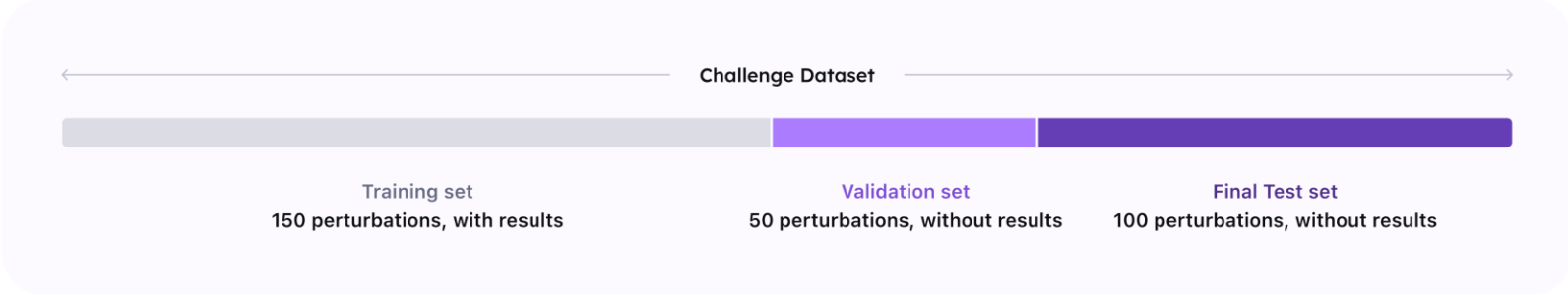
Datasets

Main dataset
H1 human embryonic stem cells (hESCs)

- **Training set**: 150 gene perturbations (~183,000 cells)
- **Validation set**: 50 gene perturbations
- **Test set**: 100 held-out perturbations, released on November 10, 2025

Additional datasets

- scBaseCount (300 M cells)
- Tahoe-100M perturbation compendium
- Public GEO / Single-Cell-Portal perturbation studies



Due to the final test set of the competition not being released until November 10th, it is necessary to upload the validation set results to the VCC website.

Arc scBaseCount ↗

A continuously updated single-cell RNA-seq database that employs a SparkAI workflow to automate discovery and standardized preprocessing of publicly available data. scBaseCount comprises over 300 million cells (and expanding), spanning 26 organisms and 72 tissues, with 150 genes specifically annotated.

TAHOE *Tahoe-100M* ↗

The world's largest single-cell dataset generated and open-sourced by Tahoe, containing 100 million cells from ~60,000 drug perturbation experiments, mapping the response of 50 cancer models to 1,100+ drug treatments.

https://virtualcellchallenge.org/datasets

Datasets

上海科技大学 ShanghaiTech University

Main dataset

H1 human embryonic stem cells (hESCs)

- **Training set**: 150 gene perturbations (~183,000 cells)
- **Validation set**: 50 gene perturbations
- **Test set**: 100 held-out perturbations, released on November 10, 2025

Additional datasets

- scBaseCount (300 M cells)
- Tahoe-100M perturbation compendium
- Public GEO / Single-Cell-Portal perturbation studies

Replogle et al., 2022 → Paper | All Datasets K562 Genome-wide (61.3 GB) • K562 (9.9 GB) • RPE1 (8.1 GB)

Genome-scale Perturb-seq targeting all expressed genes with CRISPR interference (CRISPRi) across >2.5 million human cells (K562 and RPE1). The K562 genome-wide dataset contains perturbations that overlap with most of the genes used in the Arc VCC training and validation datasets.

Nadig et al., 2025 → Paper | All Datasets HepG2 (5.2 GB) • Jurkat (8.7 GB)

Single-cell CRISPR screens of DepMap Common Essential Genes in Jurkat and HepG2 cells.

Jiang et al., 2025 → Paper | Dataset

Perturb-seq experiments in six different cancer cell lines from different tissues of origin: A549 (lung), MCF7 (breast), HT29 (colon), HAP1 (bone marrow), BxPC3 (pancreas), and K562 (bone marrow).

Srivatsan et al., 2020 → Paper | Dataset

Introduces "sci-Plex," which uses "nuclear hashing" to quantify global transcriptional responses to thousands of independent perturbations at single-cell resolution and applies it to screen three cancer cell lines exposed to 188 compounds.

McFaline-Figuero et al., 2024 → Paper | Dataset

Introduces sci-Plex-Gene-by-Environment, a platform for combined single-cell genetic and chemical screening at scale and applies it to screen combinations of chemical and genetic perturbations in glioblastoma cell lines.

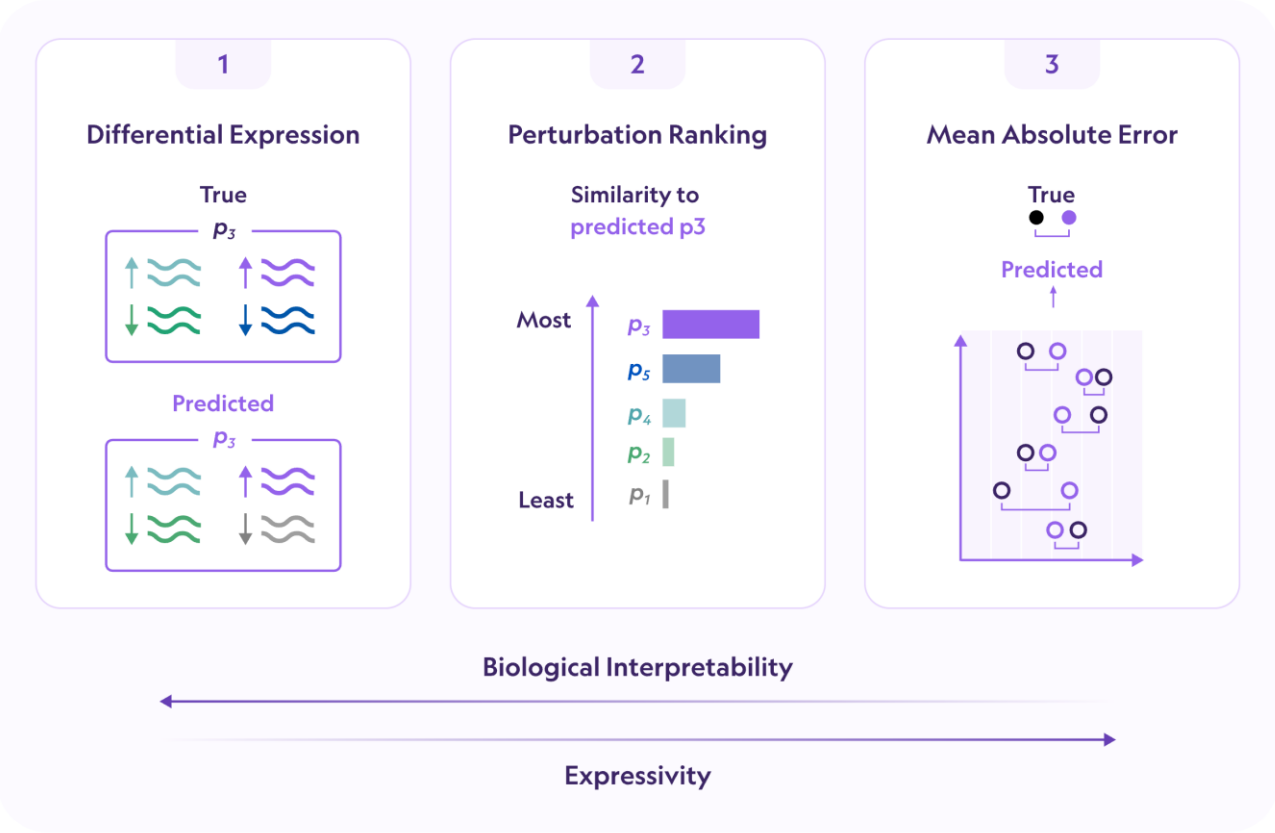
Parse-10 Million Human PBMCs in a Single Experiment → Dataset

This dataset contains 90 cytokine perturbation responses in peripheral blood mononuclear cells (PBMCs) from 12 donors ranging across 18 cell types.

https://virtualcellchallenge.org/datasets

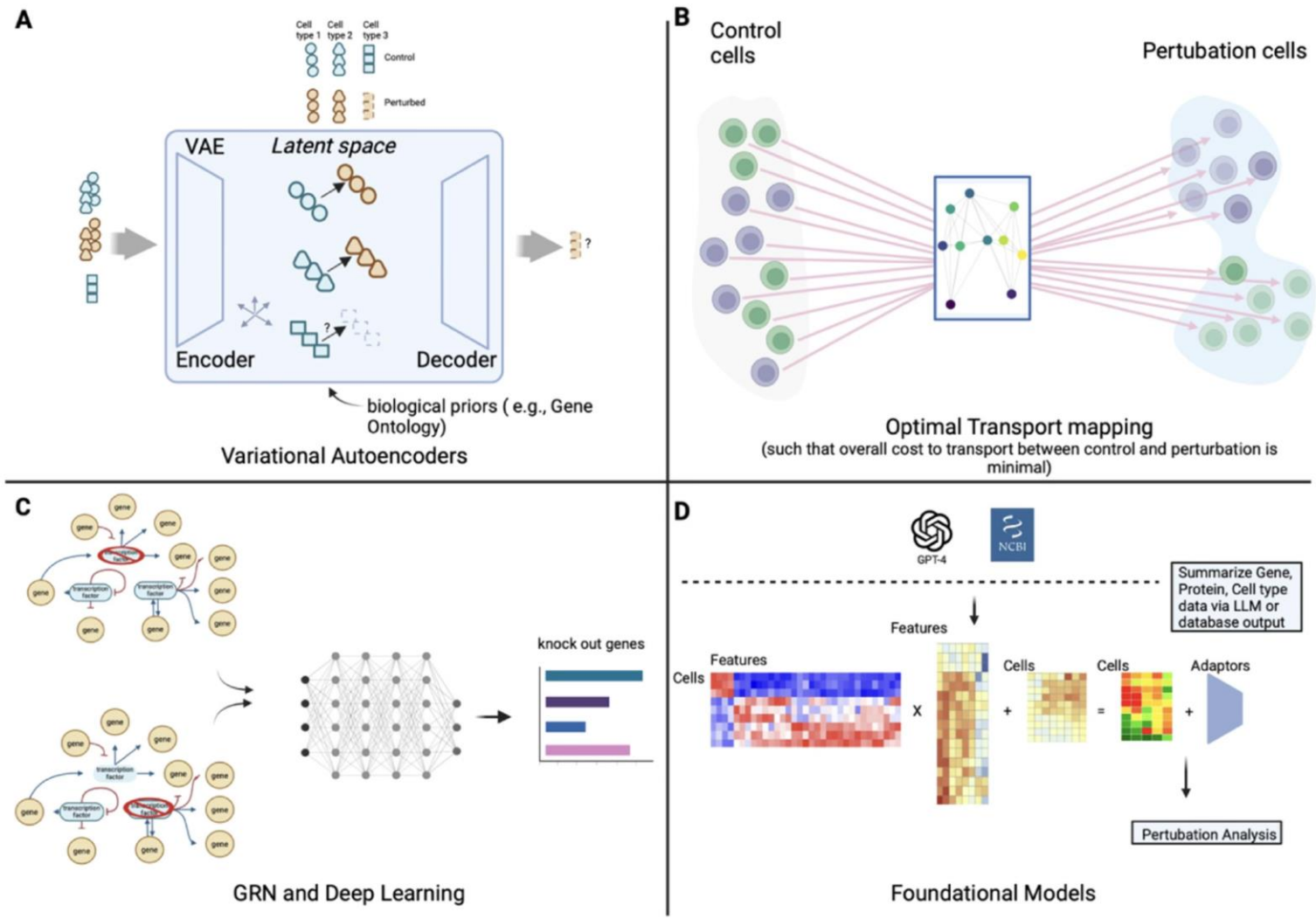
Evaluation metrics

- Differential-Expression Score (DES) evaluates how accurately a model predicts differential gene expression
- Perturbation-Discrimination Score (PDS) measures a model's ability to distinguish between perturbations by ranking predictions according to their similarity to the true perturbational effect, regardless of their effect size
- Mean Absolute Error (MAE) evaluates the overall accuracy of the predicted expression profile



Existing methods

- **Based on VAE (cVAE)**
  - CPA
  - PRNet
- **Based on optimal transport theory**
  - Cell-OT
- **Based on gene regulatory networks and deep learning**
  - GEARS
- **Based on foundation models**
  - scGPT
  - GenePert



Gavriilidis, G. I., Vasileiou, V., Orfanou, A., Ishaque, N. & Psomopoulos, F. A mini-review on perturbation modelling across single-cell omic modalities. *Computational and Structural Biotechnology Journal* **23**, 1886–1896 (2024).

Suggested modeling paths

- Fine tune pre trained single-cell foundation models (such as scGPT, Geneformer, and scBERT) using appropriate downstream tasks
- Based on the conditional variational autoencoder (cVAE), the perturbed genes can treated as conditional latent variables, and learn a distributional mapping from control to perturbed state
- Design a new prediction model for perturbation response based on basic frameworks such as Transformer or Diffusion
- Integrating gene regulatory networks or PPI modeling based on GNN

References

- C. Bunne *et al.*, How to build the virtual cell with artificial intelligence: Priorities and opportunities, *Cell*, vol. 187, no. 25, pp. 7045–7063, Dec. 2024, doi: 10.1016/j.cell.2024.11.015.
- M. Lotfollahi *et al.*, Predicting cellular responses to complex perturbations in high-throughput screens, *Mol. Syst. Biol.*, vol. 19, no. 6, p. e11517, June 2023, doi: 10.15252/msb.202211517.
- Y. H. Roohani et al., Virtual Cell Challenge: Toward a Turing test for the virtual cell, Cell, vol. 188, no. 13, pp. 3370–3374, June 2025, doi: 10.1016/j.cell.2025.06.008.
- Virtual Cell Challenge home page: https://virtualcellchallenge.org/
- scBaseCount: https://github.com/ArcInstitute/arc-virtual-cell-atlas/tree/main/scBaseCount
- Tahoe-100M: https://github.com/ArcInstitute/arc-virtual-cell-atlas/tree/main/tahoe-100M
- PerturbArena: https://luyitian.github.io/PerturbArena/index.html
- Systema: https://brbiclab.epfl.ch/projects/systema/