# Cohort Studies

Jiyu Baek, Joon Kim, Terry Kim

# Outline

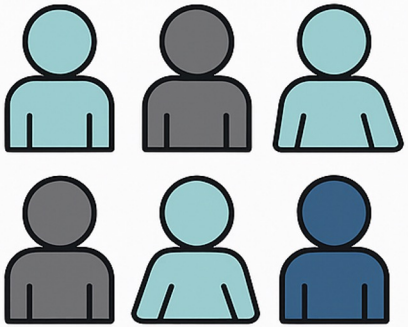
1. **Cohorts**
    a. What are cohorts?
    b. Why cohorts?

1. **Atlas Study (Liu et al.)**
    a. Background & Study Design
    b. TIME Subtype Discovery
    c. Functional and Survival Insights
    d. Clinical Implications

3. **MrVI (Boyeau et al.)**
    a. scVI Recap
    b. Challenges in scVI
    c. MrVI: An intuition
    d. Case studies

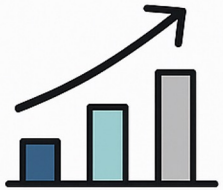
# What Are Cohorts?

A cohort is a group of individuals who share a common context



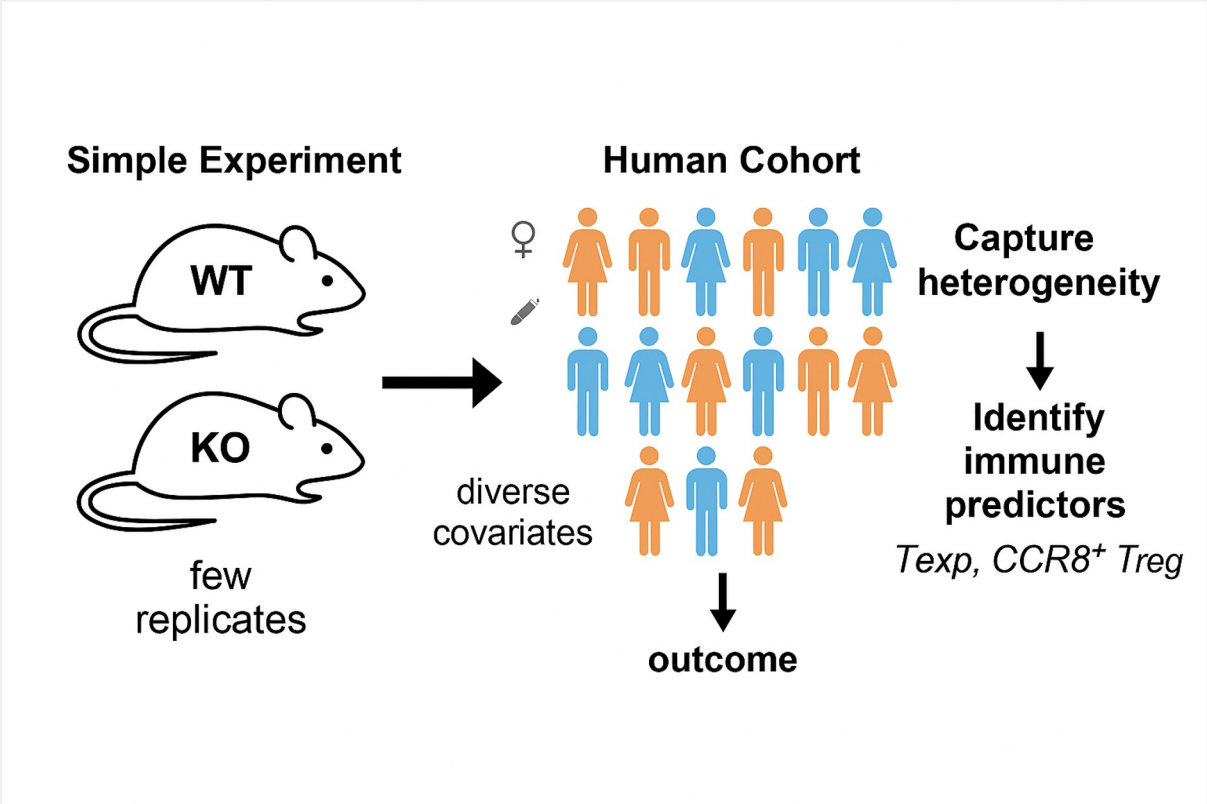
Shared condition:
NSCLC

**Different features:**
sex, PD-L1, TIME subtype



**Measure outcome:**
MPR / RFS

# Why a Cohort-Based Single-Cell Study Matters



# Outline

1. **Cohorts**
    a. What are cohorts?
    b. Why cohorts?

1. **Atlas Study (Liu et al.)**
    a. Background & Study Design
    b. TIME Subtype Discovery
    c. Functional and Survival Insights
    d. Clinical Implications

3. **MrVI (Boyeau et al.)**
    a. scVI Recap
    b. Challenges in scVI
    c. MrVI: An intuition
    d. Case studies

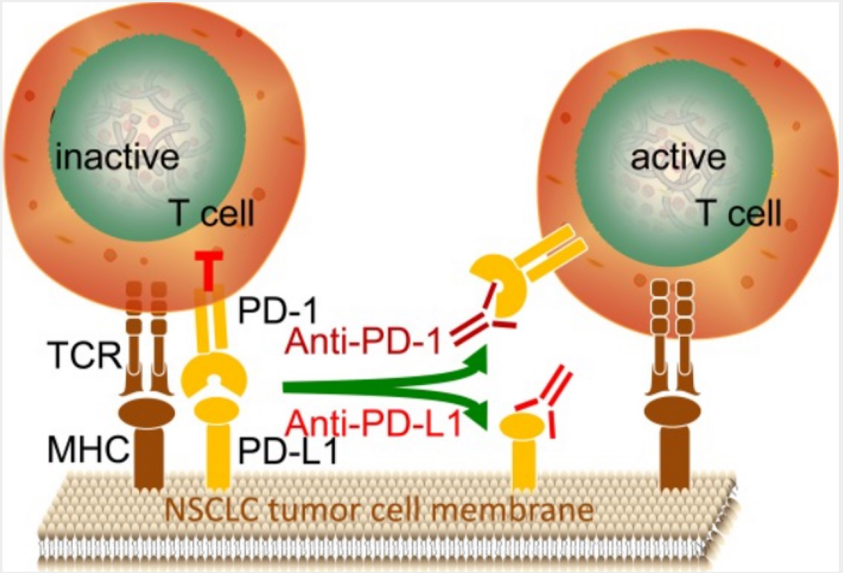
# Clinical Background and Rationale

NSCLC ≈ 85 % of lung cancer cases (LUAD vs LUSC)

PD-1/PD-L1 blockade has transformed therapy, but response rates ≈ 30–40 %

Hypothesis: **Tumor immune microenvironment (TIME) heterogeneity drives differential responses**

MPR (major pathologic response) and RFS (recurrence-free survival) as clinical results




(Zhu et al., 2020)

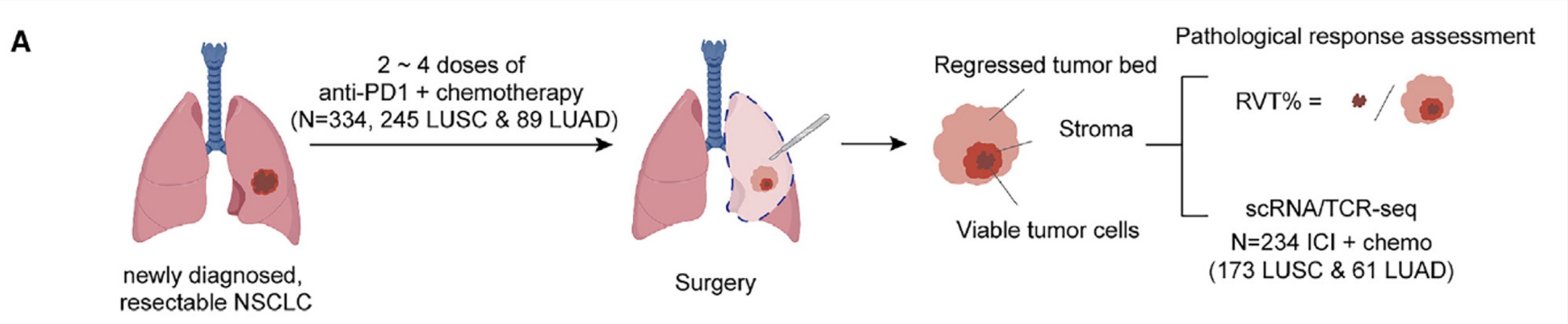
# Study Design and Cohort

234 patients with resectable NSCLC treated with neoadjuvant anti-PD-1 + chemotherapy

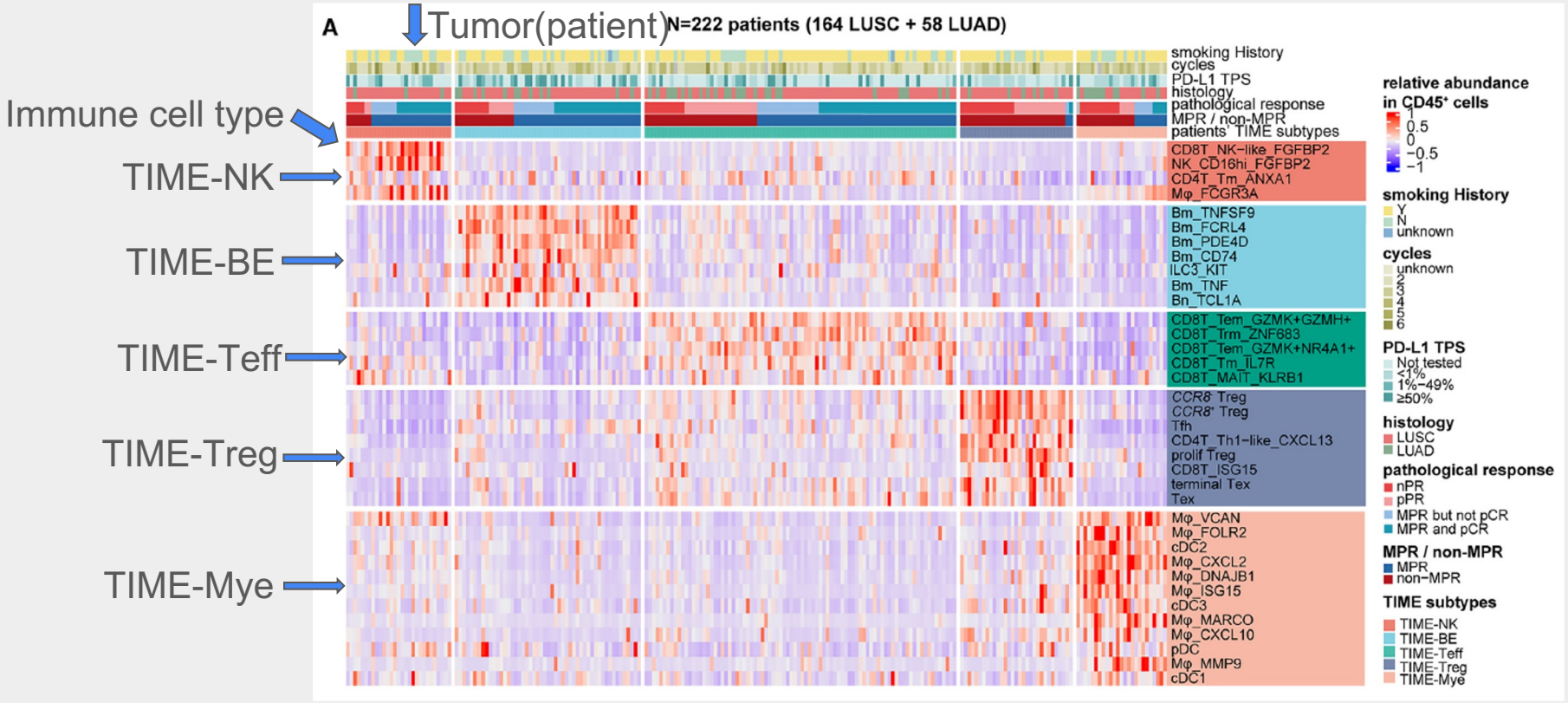
Samples collected at surgery → tumor, lymph node, and blood

**1.25 million cells** analyzed by scRNA-seq + scTCR-seq

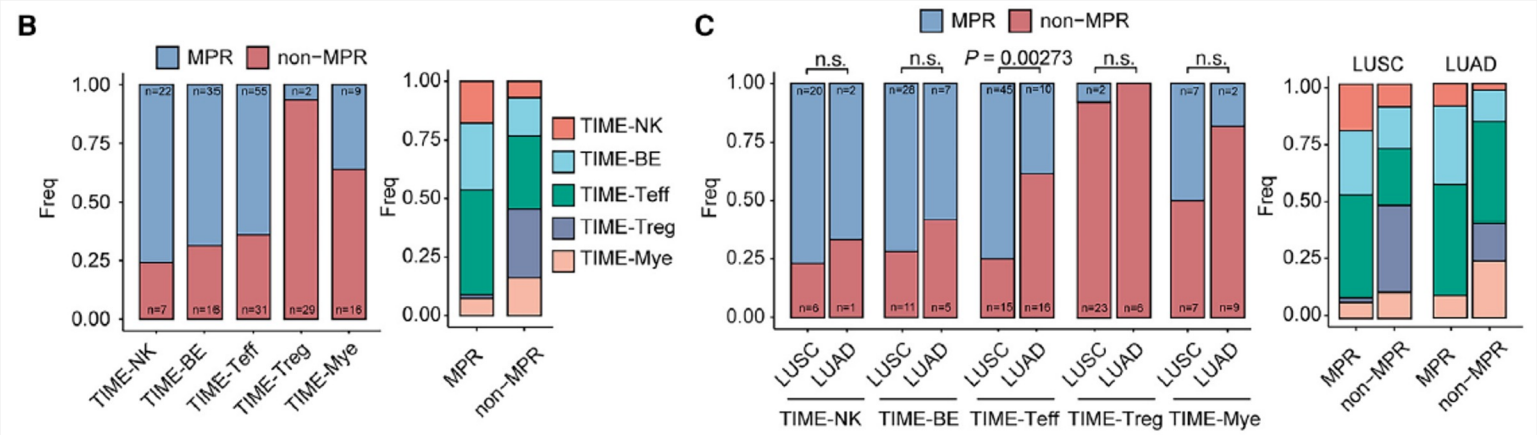
Integration and NMF-based clustering(based on immune cell frequencies) to define immune modules



# Identification of 5 TIME Subtypes



# Pathological Response Across TIME Subtypes

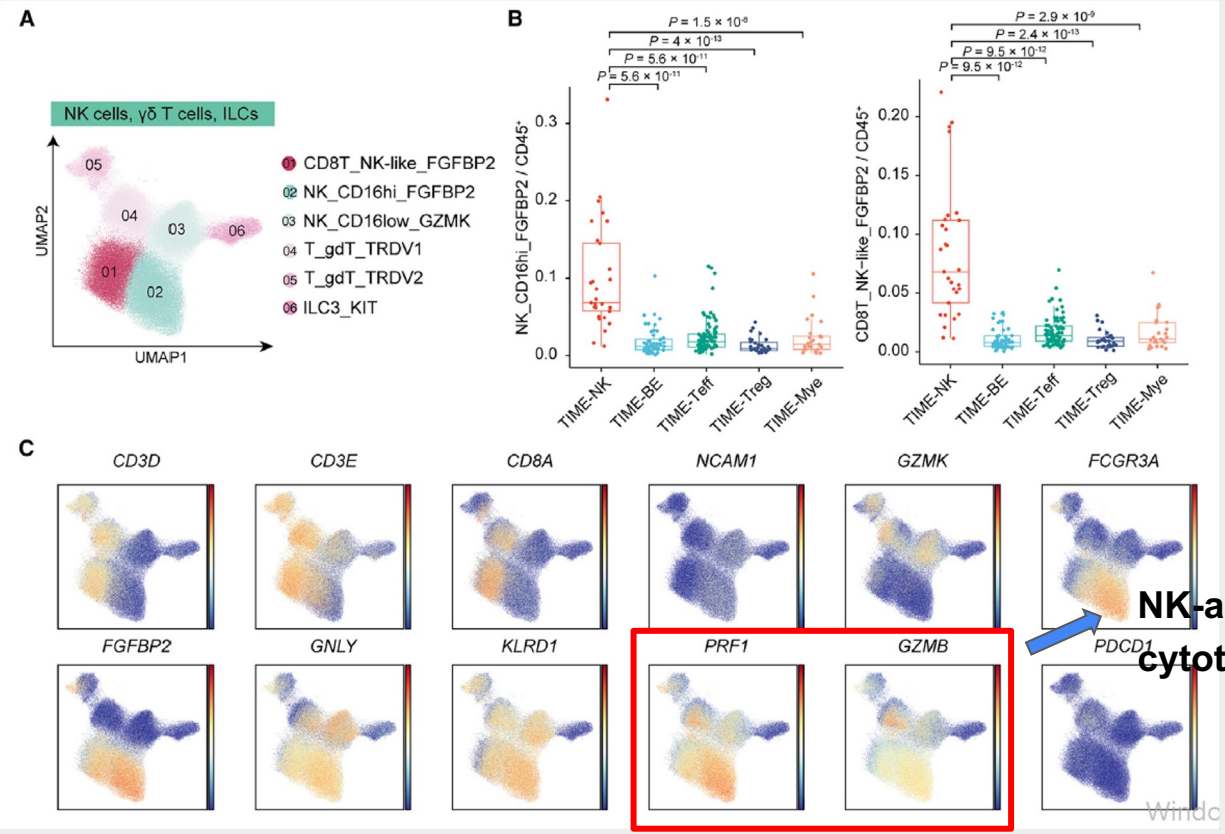



**TIME-NK and TIME-BE → high MPR / PRR**

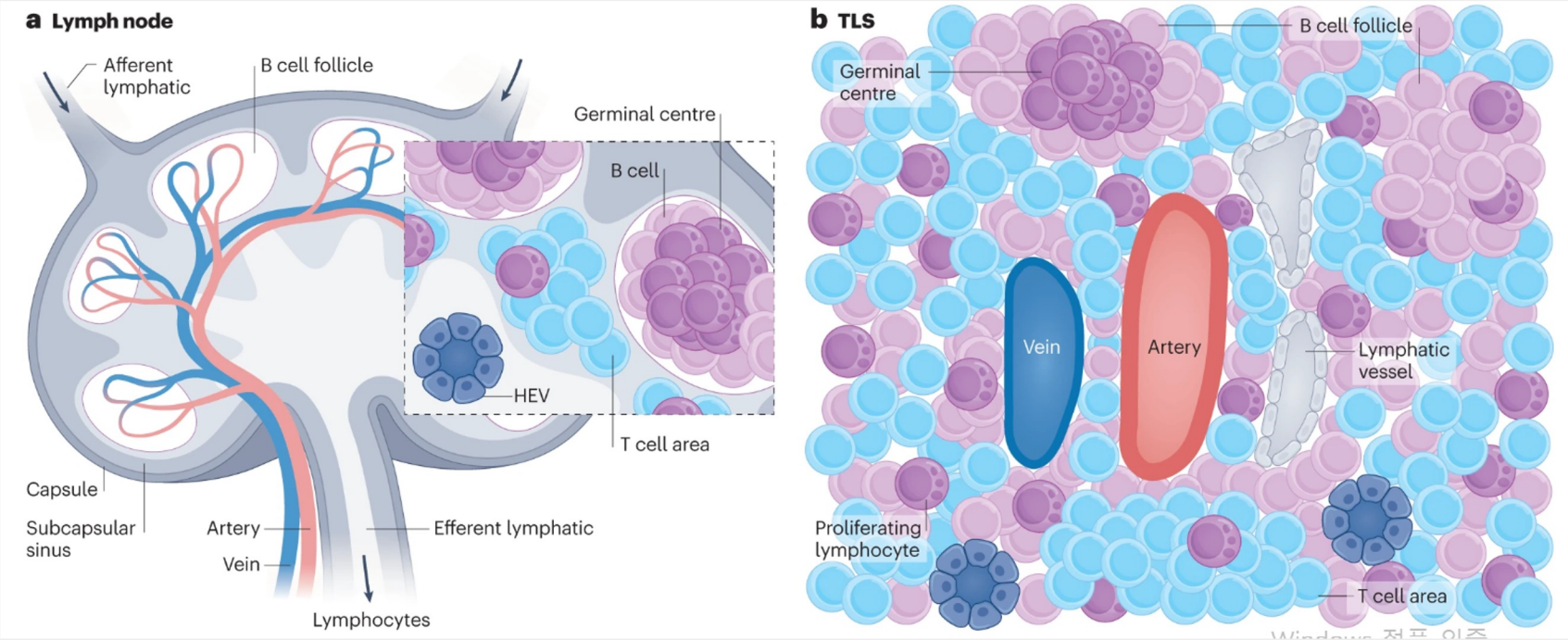
**TIME-Treg → low MPR and poor response**

Indicates immune composition predicts pathologic response

# TIME-NK:FGFBP2+ NK-like CD8 T cells

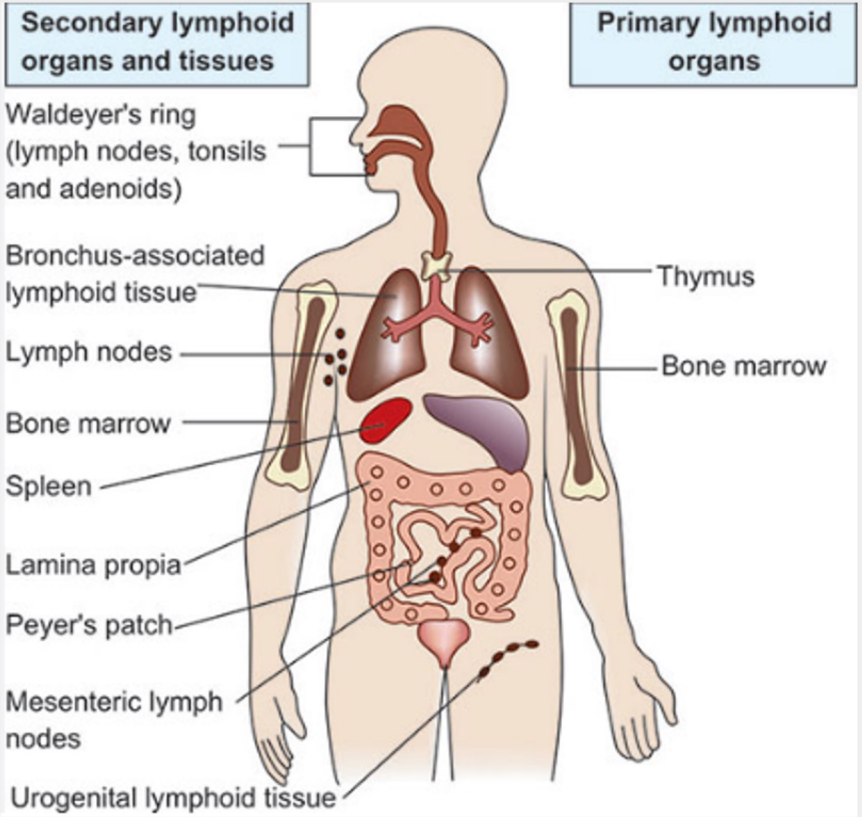


# TIME-BE: B Cells and Tertiary Lymphoid Structures (TLS)




(Yuki et al., 2023)

# TIME-BE: B Cells and Tertiary Lymphoid Structures (TLS)

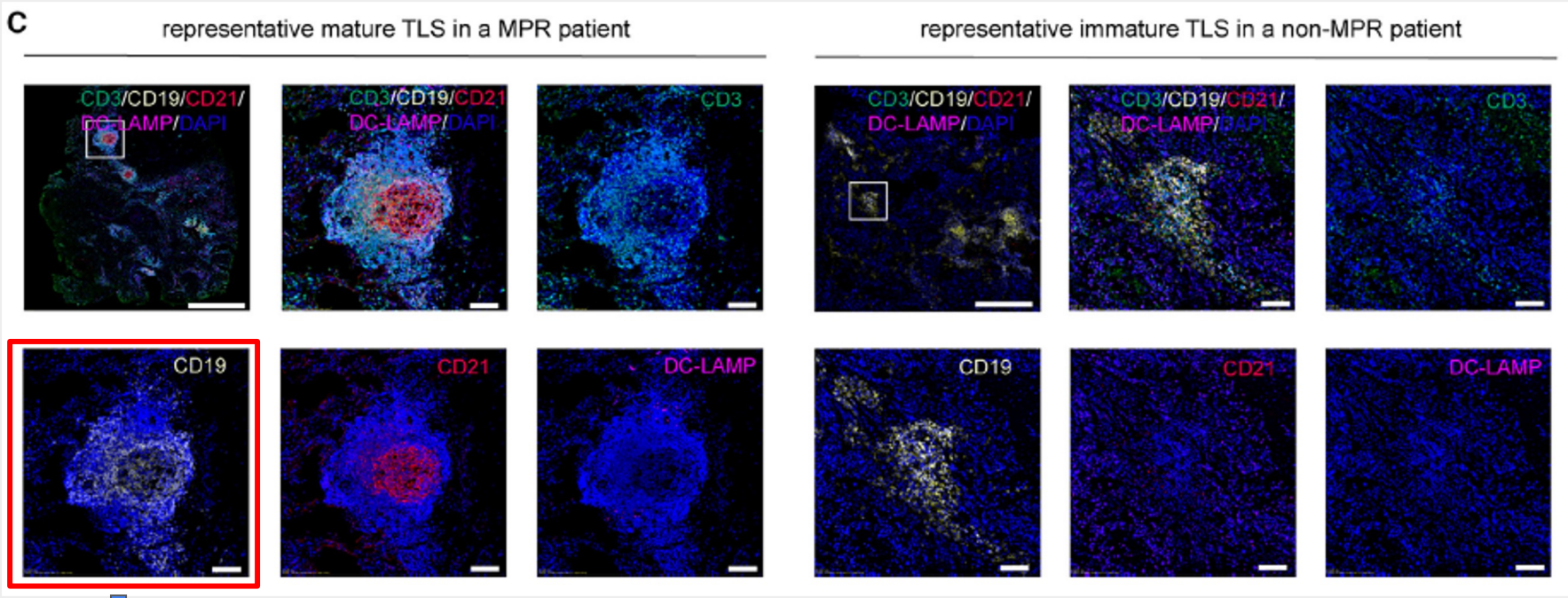



**Primary lymphoid organs** — thymus and bone marrow, where lymphocytes *develop.*

**Secondary lymphoid organs** — lymph nodes and spleen, where immune cells *activate and interact.*

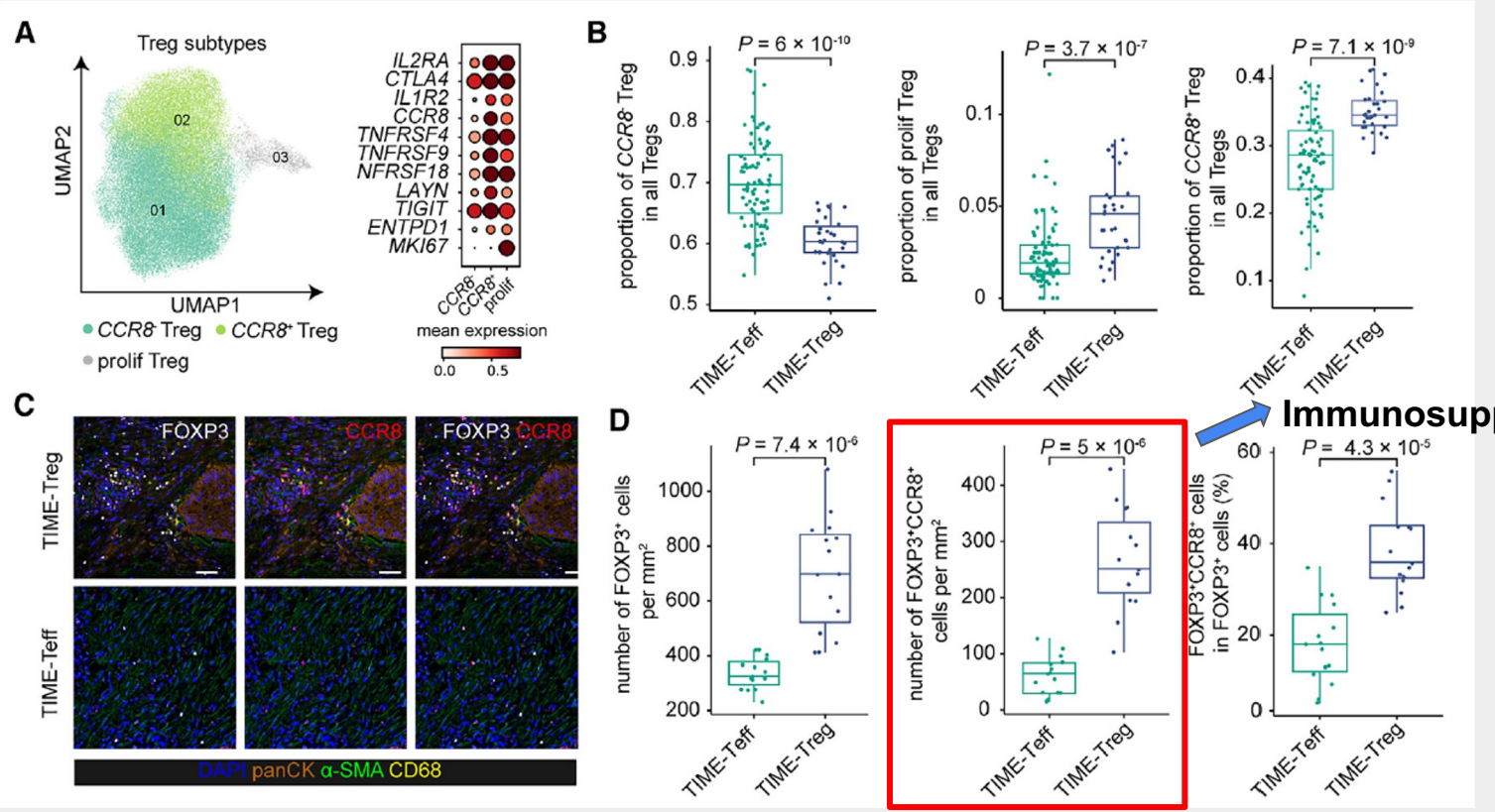
**Tertiary lymphoid structures (TLS)** — ectopic, formed in tissues like tumors under chronic inflammation.

# TIME-BE: B Cells and Tertiary Lymphoid Structures (TLS)

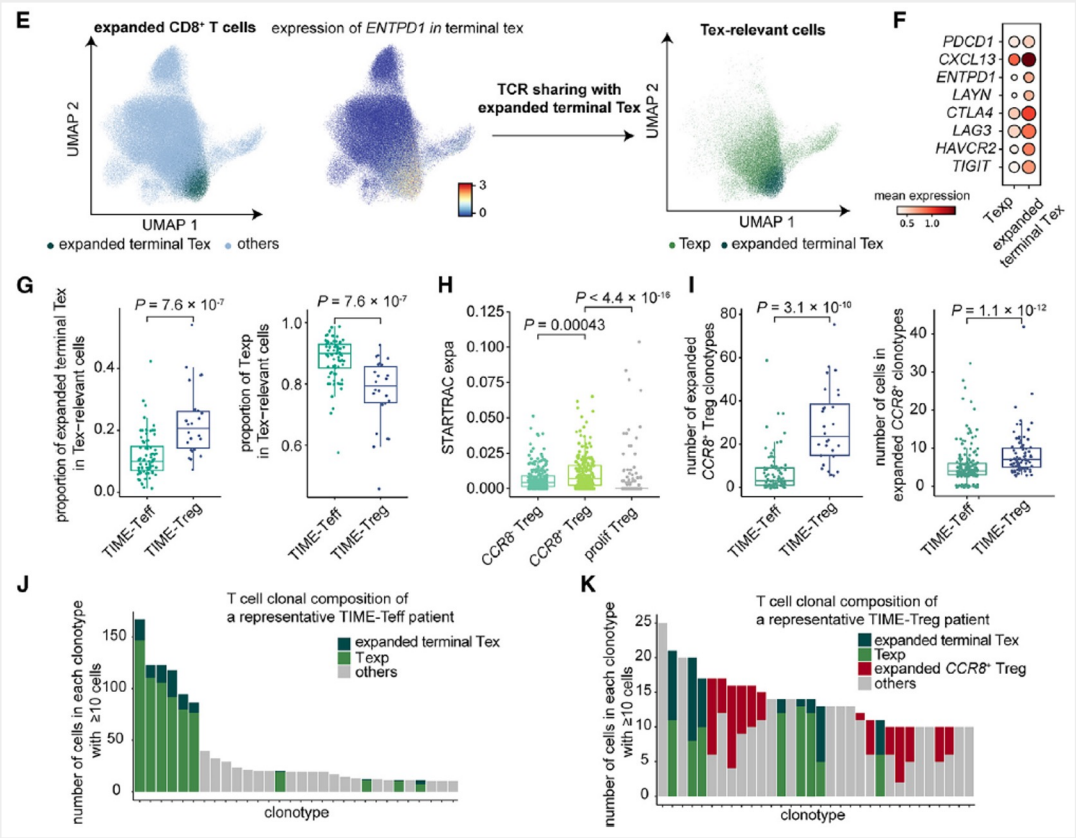



B cell aggregation!

# TIME-Teff vs TIME-Treg



# TIME-Teff vs TIME-Treg




**Texp vs Tex**
Texp cells remain functional
Tex cells are terminally exhausted.

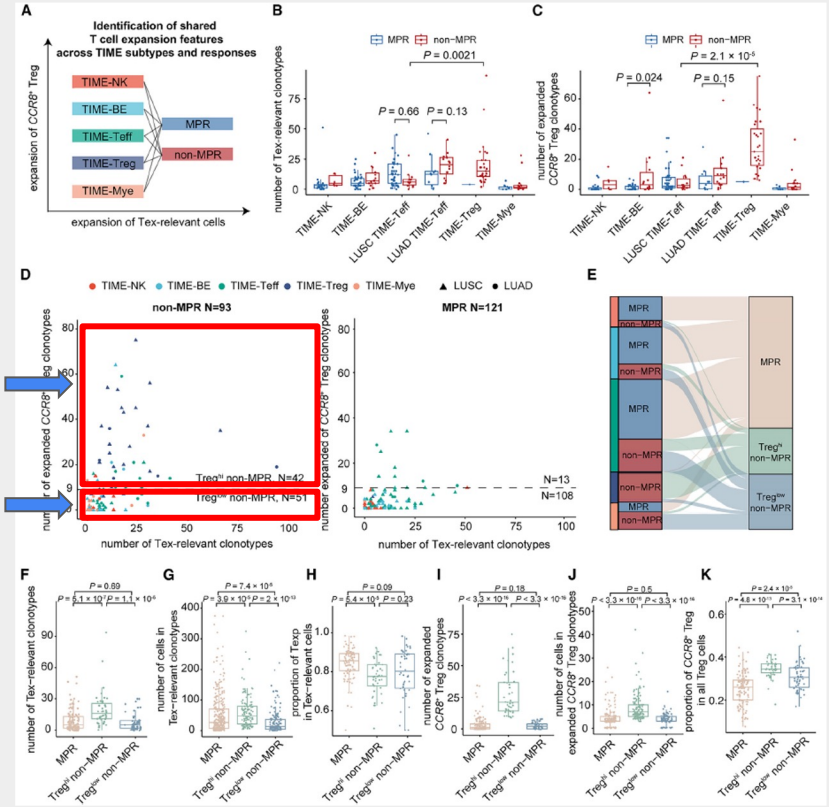
**TIME Teff vs TIME Treg**
TIME-Teff favors Texp expansion
TIME-Treg favors Tex dominance.

Effector maintenance versus exhaustion drives response outcomes.

# TCR Clonotype & Non-MPR Heterogeneity

Treg hi non-MPR

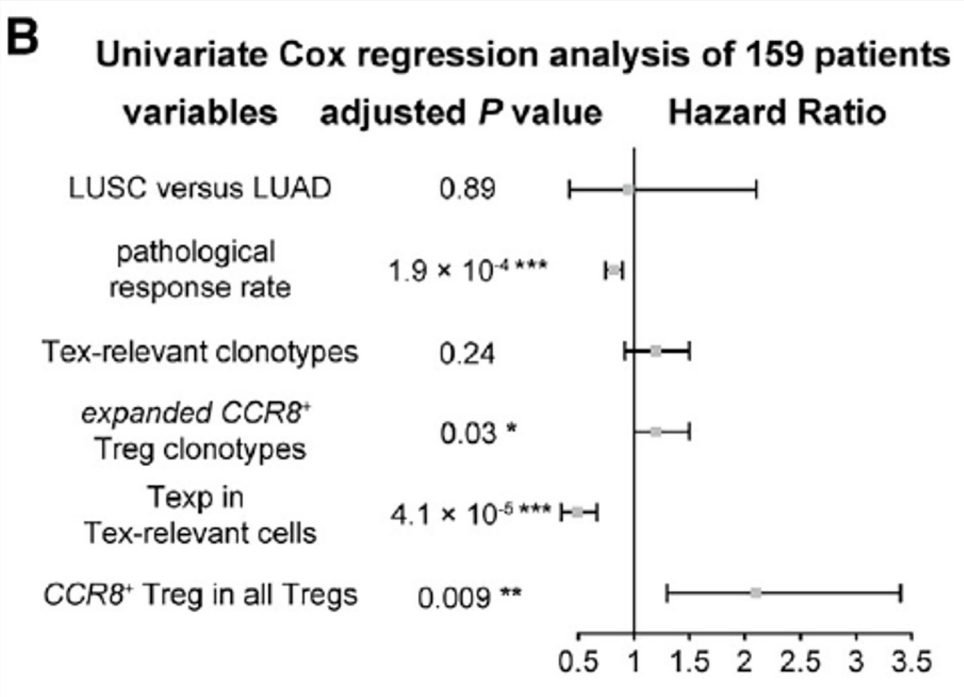
Treg low non-MPR




Non-MPR patients split into Treg hi and Treg low groups.

Treg hi tumors: expanded CCR8$^+$ Tregs, few Texp cells.

# Cox Proportional-Hazard Analysis

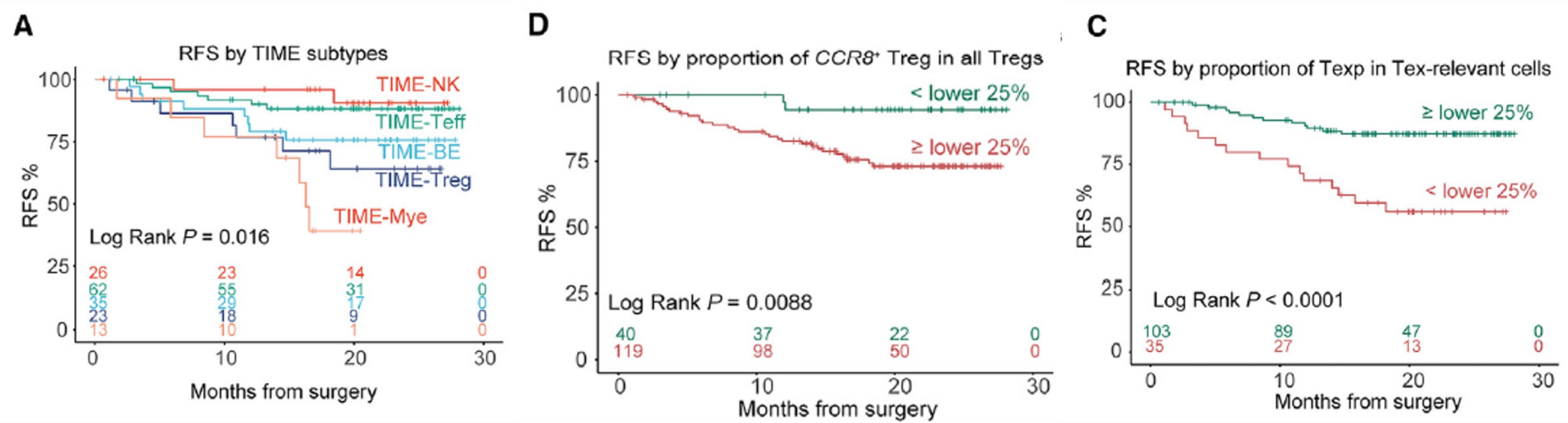



**Used to identify factors that influence recurrence risk over time**

- Models *time-to-event* data (e.g., time to recurrence)

- Estimates **Hazard Ratio (HR)** for each variable
  → **HR < 1:** lower recurrence risk
  → **HR > 1:** higher recurrence risk

# Recurrence-Free Survival (Cox-PH Analysis)




TIME-NK/Teff have best RFS

TIME-Treg/Mye relapse earlier

More Texp → lower recurrence risk; more CCR8$^{+}$ Tregs → higher risk

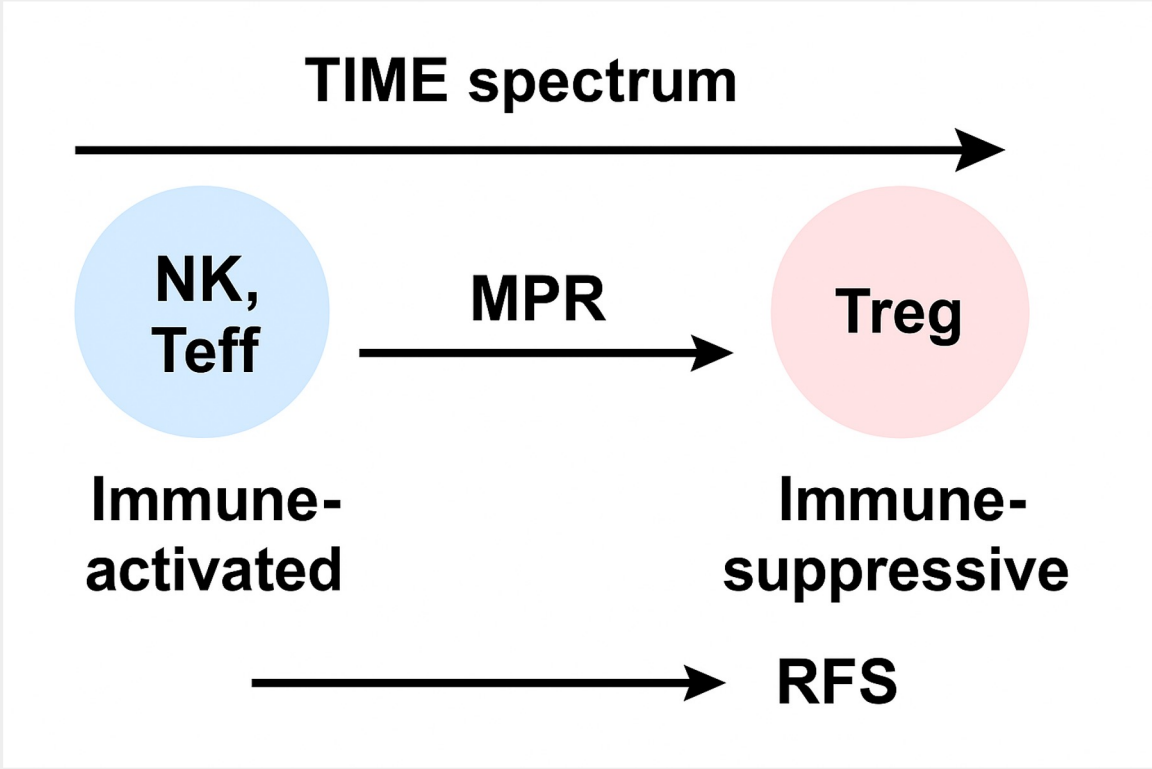
# Integrated Model and Clinical Implications

-Cytotoxic and effector T cells

-Mature TLS

-High Texp cells

-High MPR, RFS

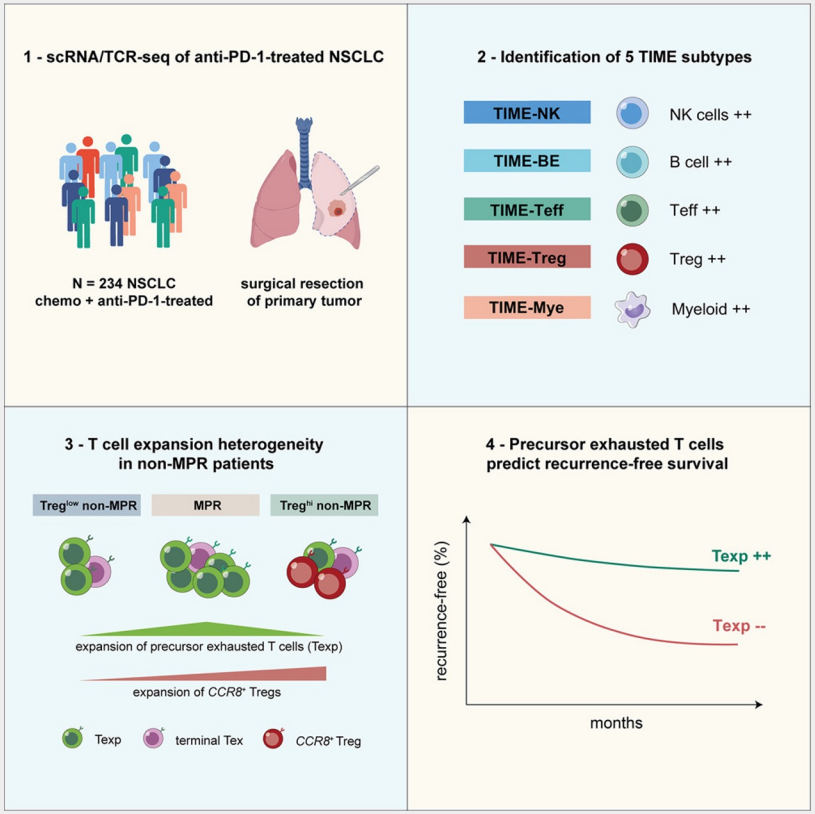



-Tregs and exhausted T cells

-Immature TLS

-Shorter RFS

# Summary and Take-Home Messages




1. The power of cohort-level single-cell analysis

1. The TIME exists on a functional spectrum

1. Distinct immune programs drive success or failure

1. Clinical implication


TAKE A
BREAK!

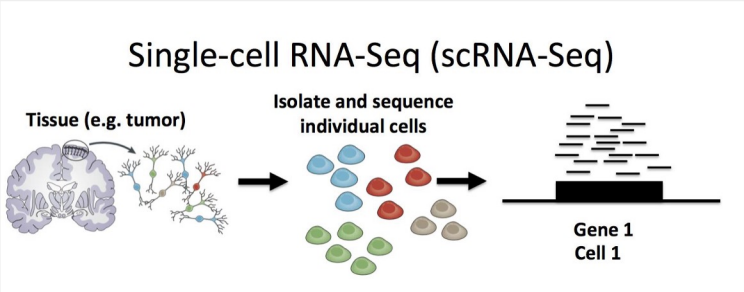
# Outline

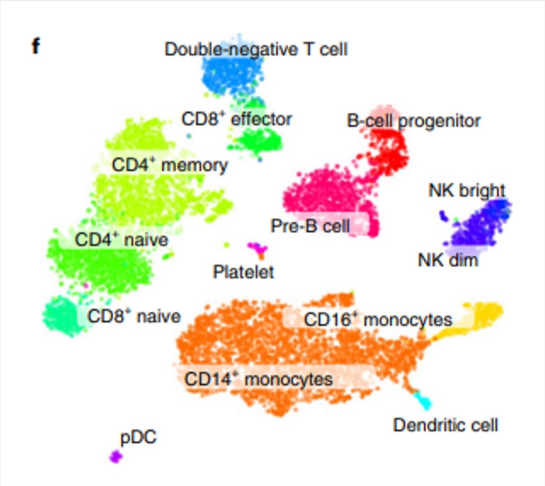
1. **Cohorts**
    a. What are cohorts?
    b. Why cohorts?

1. **Atlas Study (Liu et al.)**

3. **MrVI (Boyeau et al.)**
    a. scVI Recap
    b. Challenges in scVI
    c. MrVI: An intuition
    d. Case studies

# Recap: sc-RNA Seq allows analyzing mechanism of disease at much detailed resolution

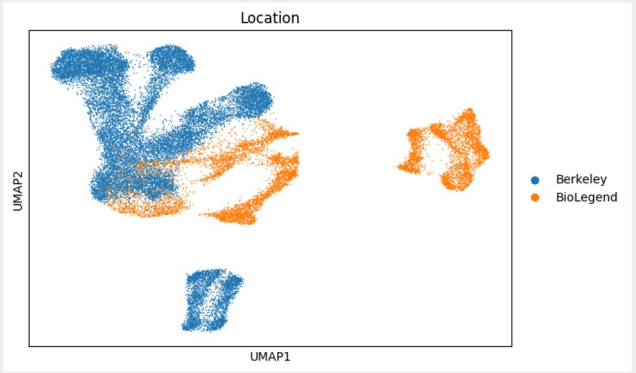


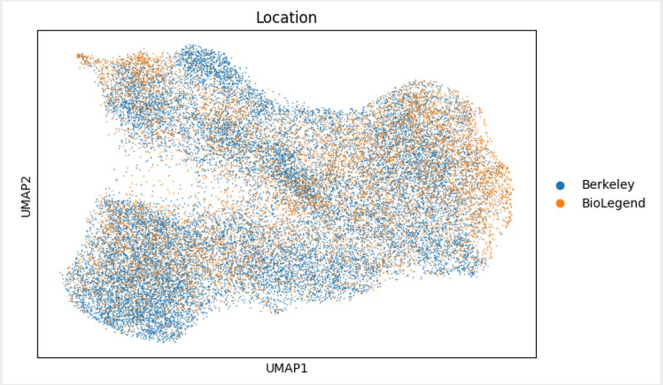



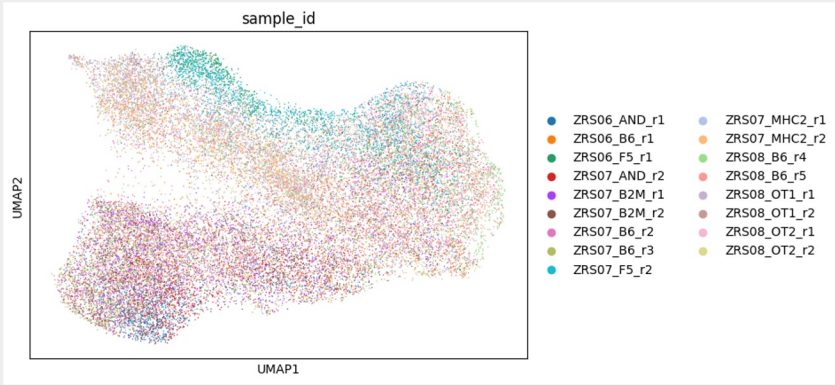

f, A t-SNE embedding of a primary peripheral blood mononuclear cell (PBMC) dataset with cell annotations. NK, natural killer, separated into CD56 bright and dim subsets. pDC, plasmacytoid dendritic cell

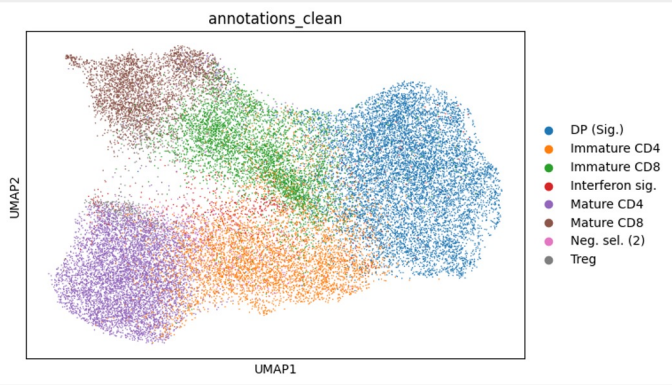
Kharchenko, Peter V. "The triumphs and limitations of computational methods for scRNA-seq." *Nature methods* 18.7 (2021): 723-732.

# scVI provided a scalable framework for the probabilistic representation and analysis of single cell




**Done? but… sample level heterogeneity?**

# Smoothies Again? Within Sample

**Single Cell:**

individual ingredients (cells) and not the smoothie (average/bulk analysis)

**PCA or even scVI:**

Latent Space, and possibly cell-level classification

However, **subtle differences inside the same cell classification** are completely ignored, in other words, “smoothied” again

**“assumes the effects they evaluate are constant”**

# Outline

1. **Cohorts**
    a. What are cohorts?
    b. Why cohorts?
2. **Atlas Study (Liu et al.)**
3. **MrVI (Boyeau et al.)**
    a. scVI Recap
    b. Challenges in scVI
    c. MrVI: An intuition
    d. Case studies

# Motivation of MrVI for Cohort Studies

Contribution:
(1) stratifying samples into groups **(HOW)**
(2) evaluating the cellular and molecular differences between groups **(WHY)**

Separate biological signal (target covariate) | technical noise (nuisance covariate)

**without requiring** a priori grouping of cells into types or states.

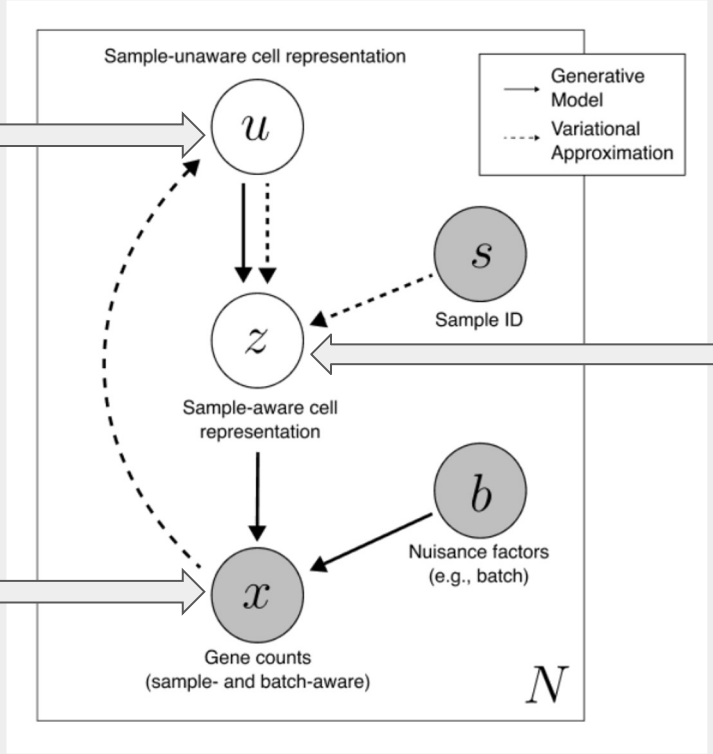
# The Diagram (you'll get sick of this…)

most pure

most noise

"Pure" state without any covariates

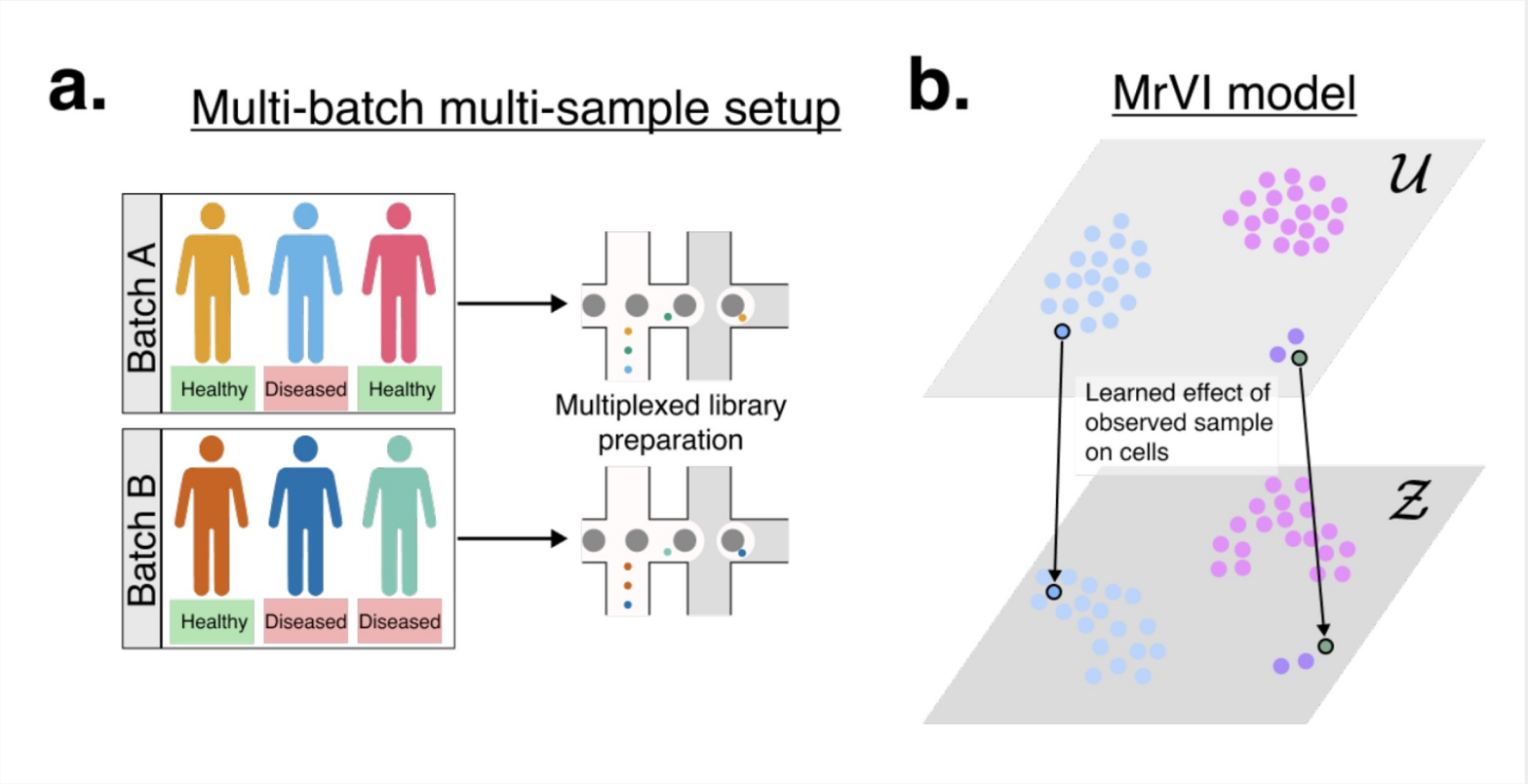
Original Data (ft. batch effects)



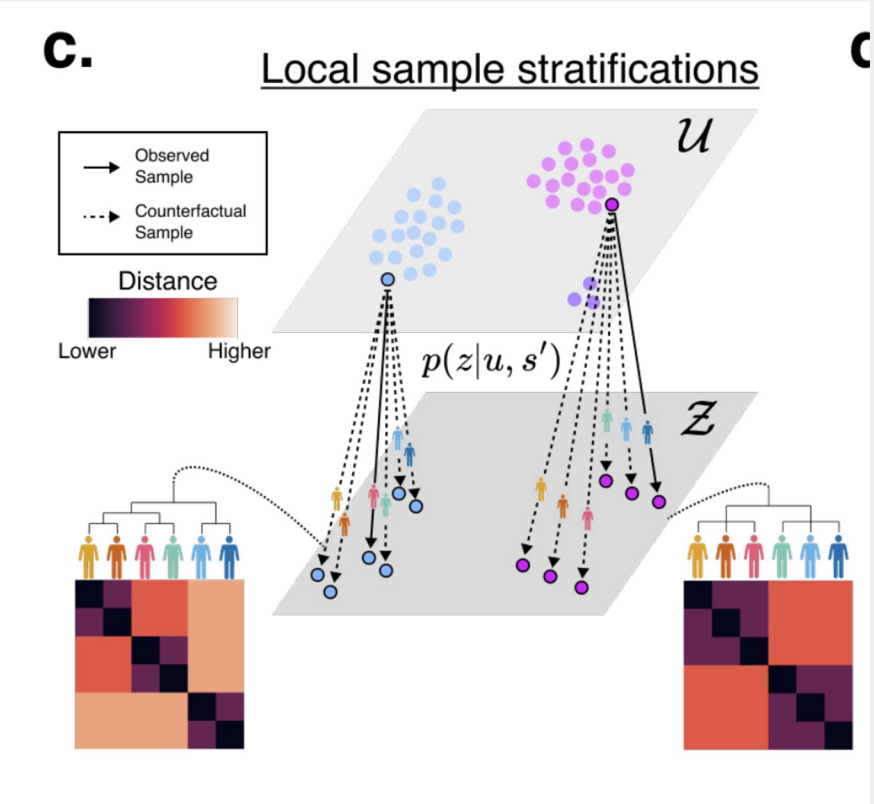

Now with covariates that we care about!

# mrVI, What Can You Do? (1/3)

(For simplicity, we assume dim(u-space) = dim(z-space))

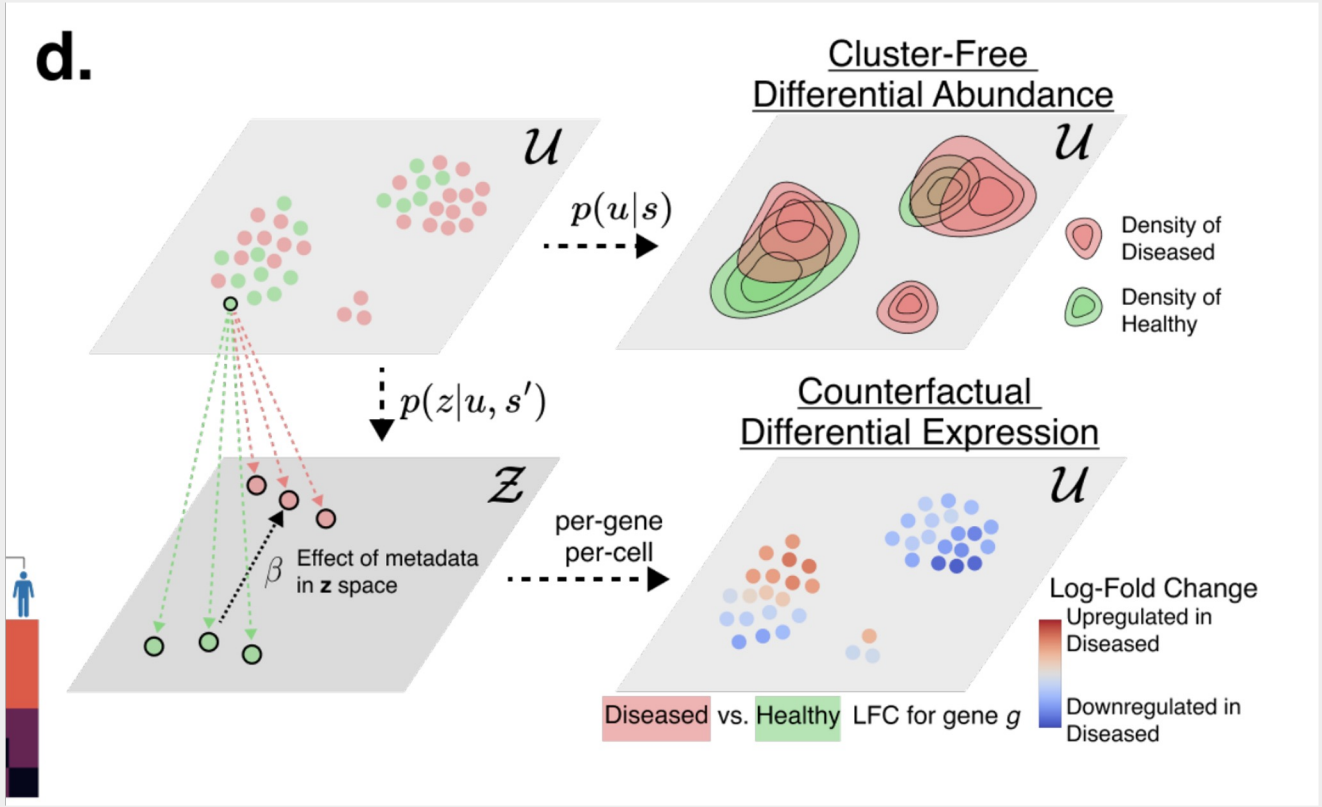


# mrVI, What Can You Do? (2/3)




"Multiverse!"

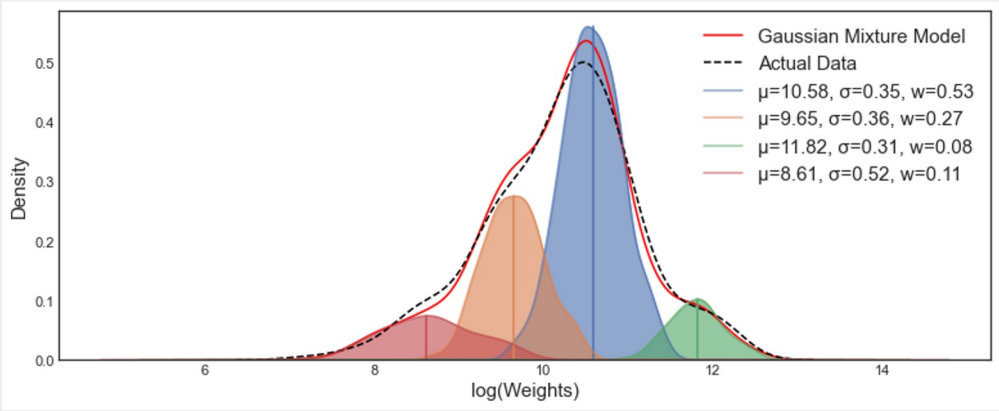
# mrVI, What Can You Do? (3/3)

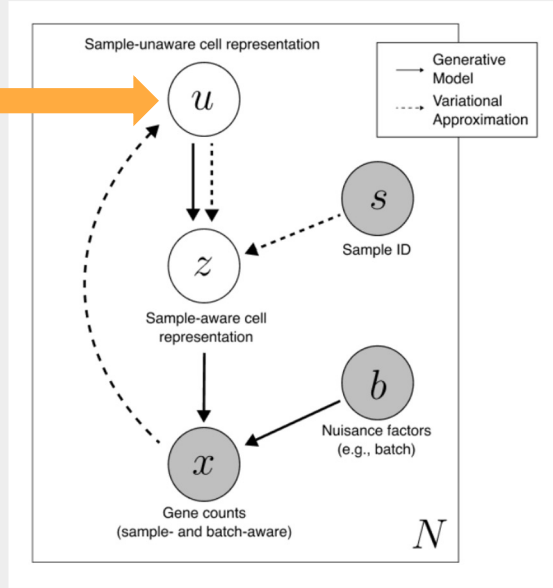



Will explain this in detail later…

# Mixture of Gaussians (very lightly)

sampling with weighted combination of Gaussians

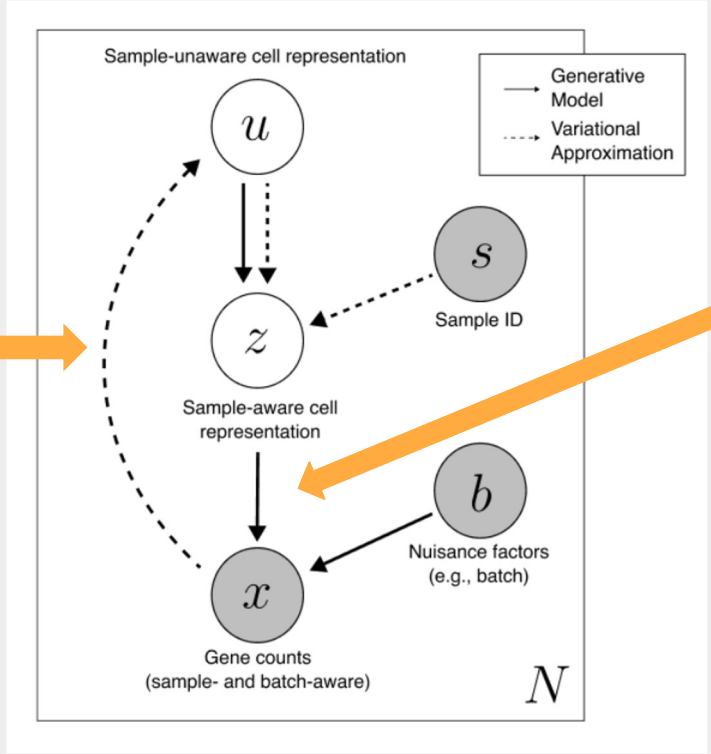



*Possibly account for distinct clusters* for different **cell types or states**

# The Diagram, but with a “VAE lens”

This is the “encoder”

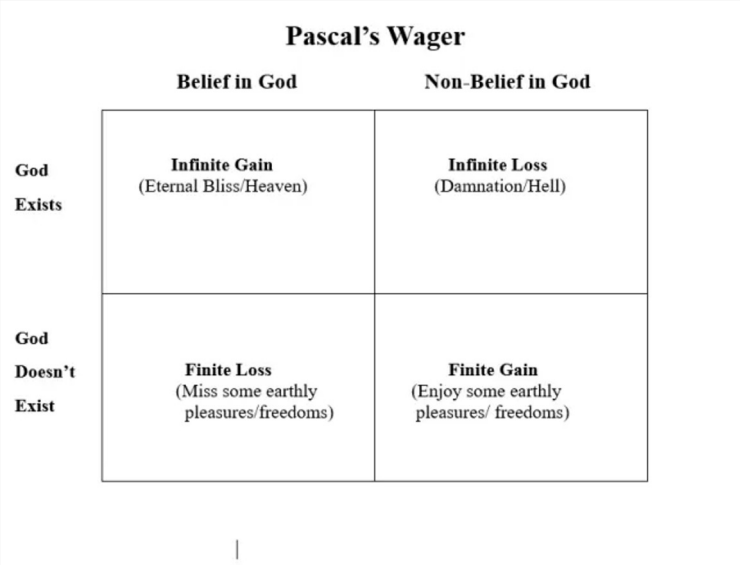



Then we try to “decode” (or “reconstruct”)

# A Piece of Philosophy

The efficacy of mrVI greatly hinges upon the existence of a *u-space* that nicely captures the essence of our dataset, in addition to the *z-space*

*If you believe in Nature…*



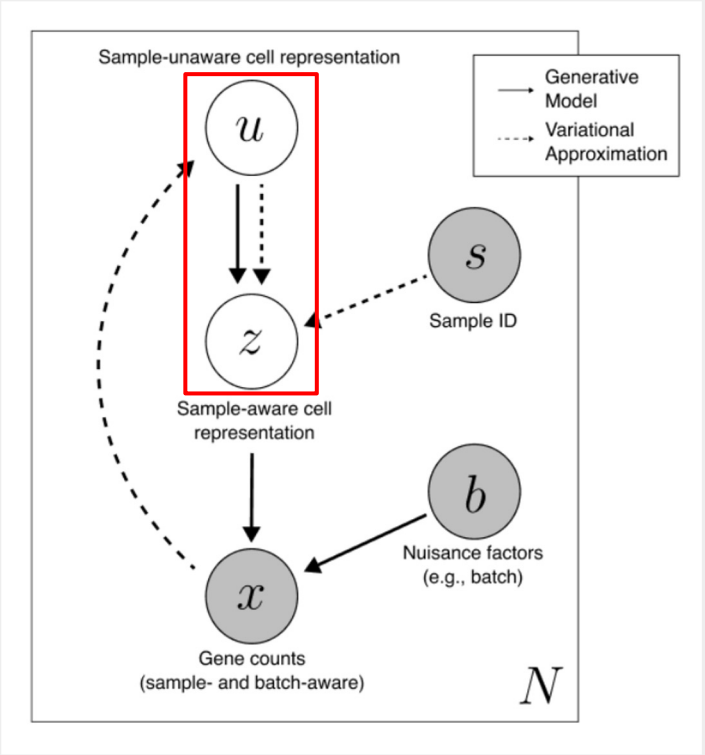

| | Belief in u-space | Non-belief in u-space |
|---|---|---|
| U-space exists | Infinite Gain<br>(Nature paper) | Infinite Loss<br>(Somebody else writes a Nature paper) |
| U-space doesn't exist | Finite Loss<br>(Experiment fails) | Finite Gain<br>(Do something else) |

# A Bad Analogy (maybe not)

U-space asks: **What** are you? (What is this cell?)

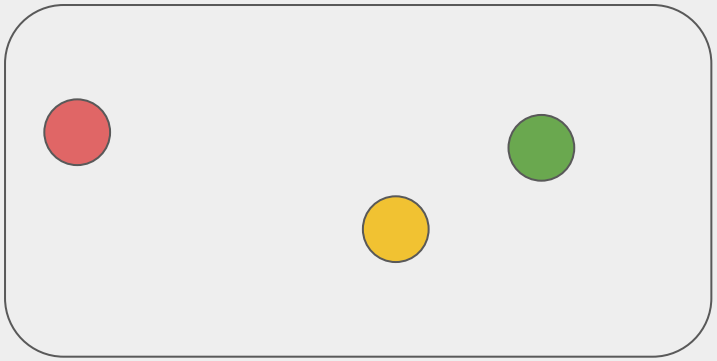
Z-space asks: What are you? And **How** are you? (How does this cell behave?)

Which leads us to believe that the **How** is encoded in (Z - U)…?!



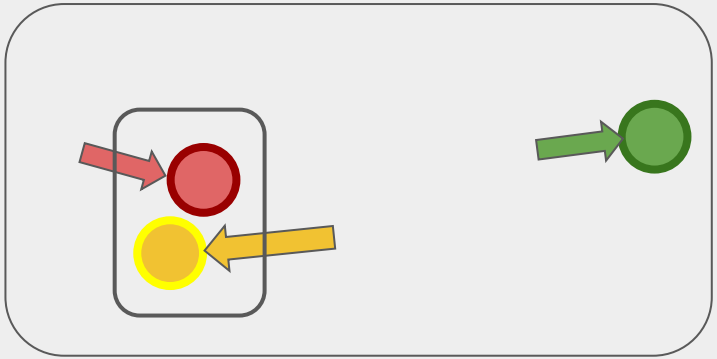
# Kind of a “Proof”

What we care about: **Arrows**!
Or in biology terms: **Sample effects!!**
Or kind of a **“(Z - U)” effect!!!**

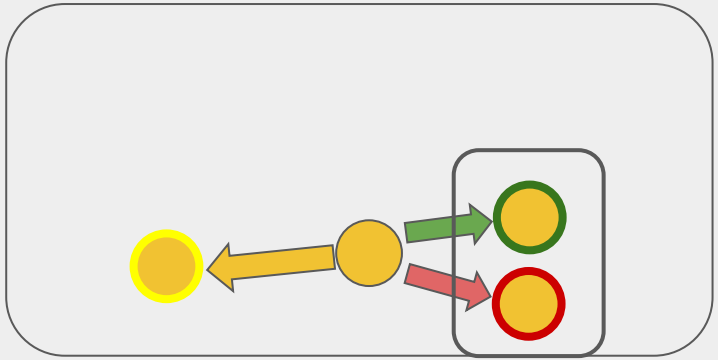


u-space

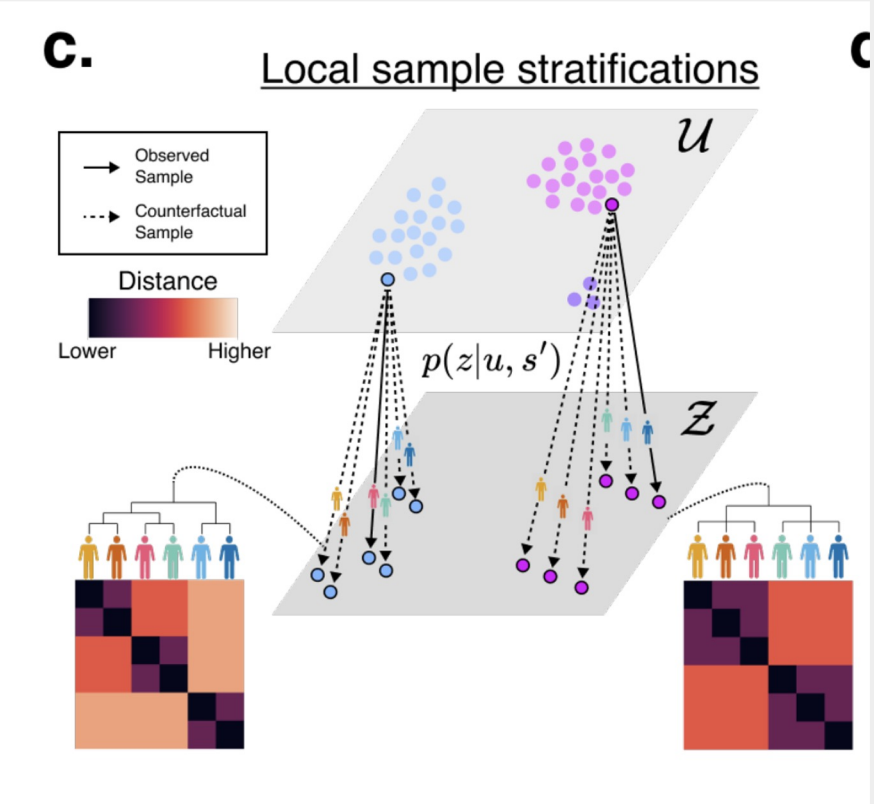
z-space (scVI)



z-space
(with counterfactuals)

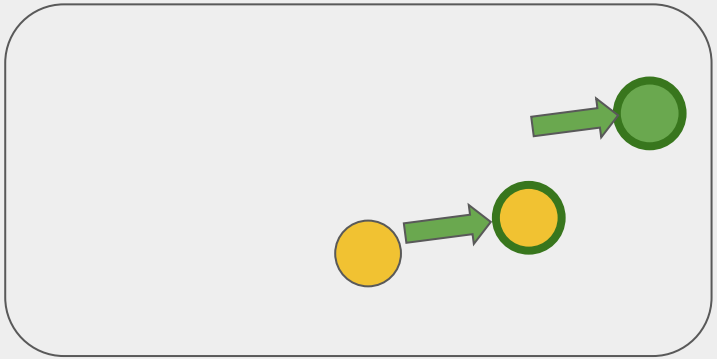


# mrVI, What Can You Do? (2/3), again




"Multiverse!"

# New Challenge, Under the Rug



What ensures that these two green arrows will be the same?

“Well… better train it good” or “No Free Lunch”

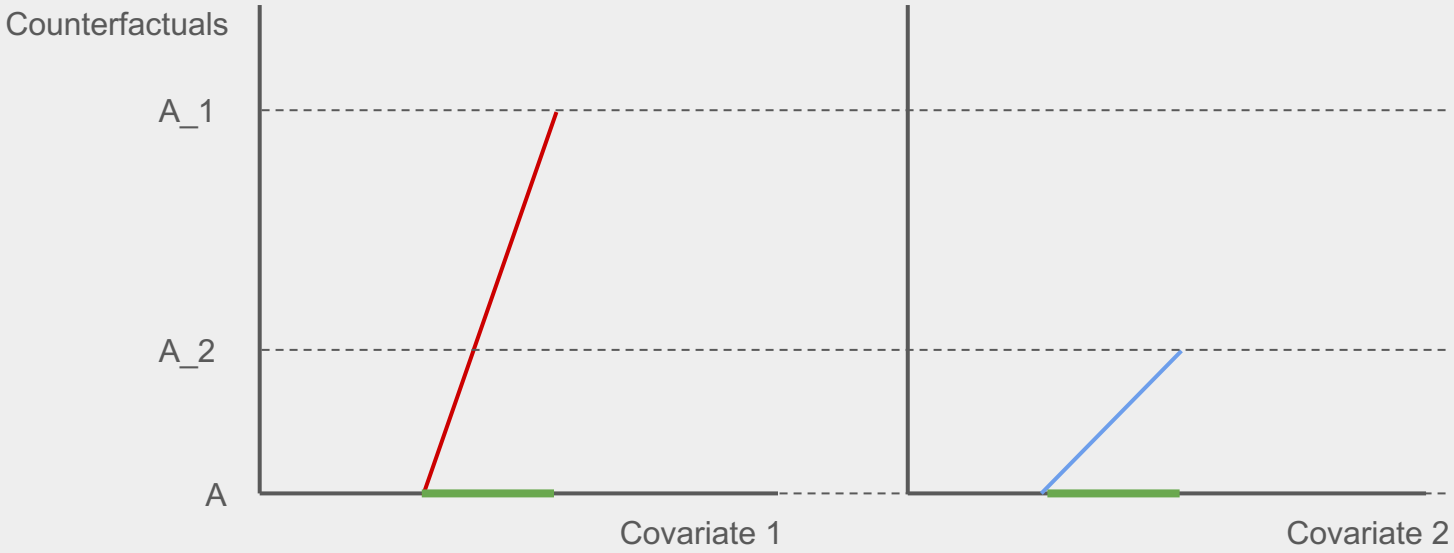
We turned one hard problem into another hard problem!

# A Bit of Math Ahead…

“Fitting a straight line can get you very far”

- Sun Tzu, probably

# How do we “measure” effects of covariates?

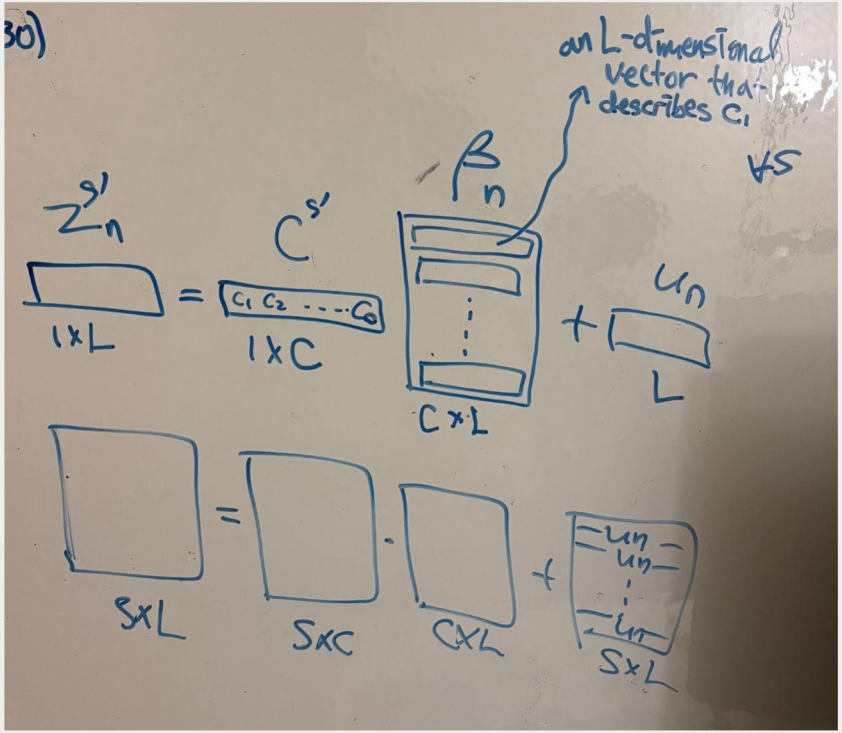



Intuition: slope is like the “sensitivity” of that covariate; little changes are significant

(high dimensions)

# Now in HD!

$$z_n^{s'} = c^{s'^T} \beta_n + u_n, \quad \forall s',$$



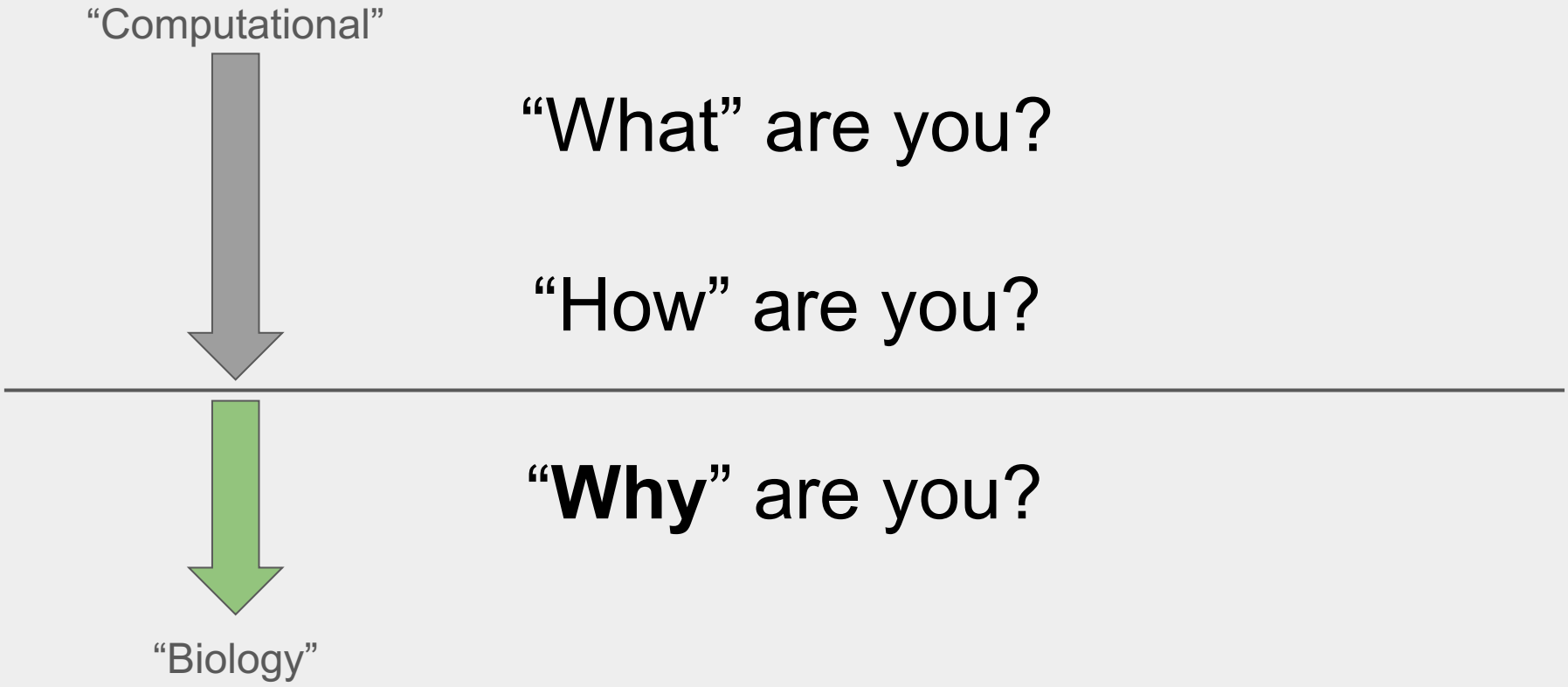

Each row of Beta_n is the “representation” of each covariate in z-space

Digression: PCA is one way?

We simply fit Beta_n over all counterfactual samples of z

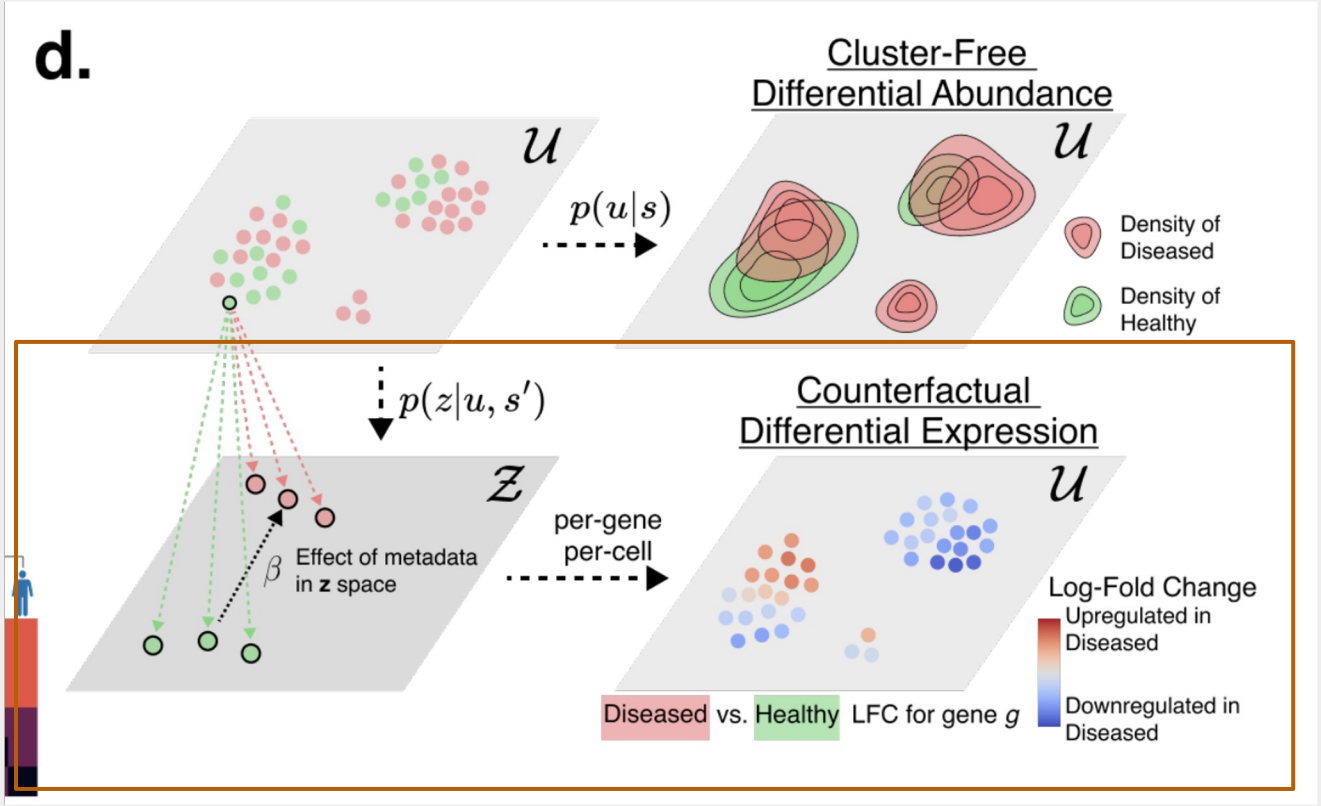
Beta_n[i] is now the “program” of the i-th covariate, or its “coefficients” (actually a vector)

Intuition: weighted sum of rows of Beta.


“Computational”
“What” are you?
“How” are you?
“Why” are you?
“Biology”


TAKE A
BREAK!

# mrVI, What Can You Do? (3/3), again



# Differential Expression!!!

(Assume a single binary covariate, a switch)

$$z_n^1 = \beta_n^1 + u_n$$

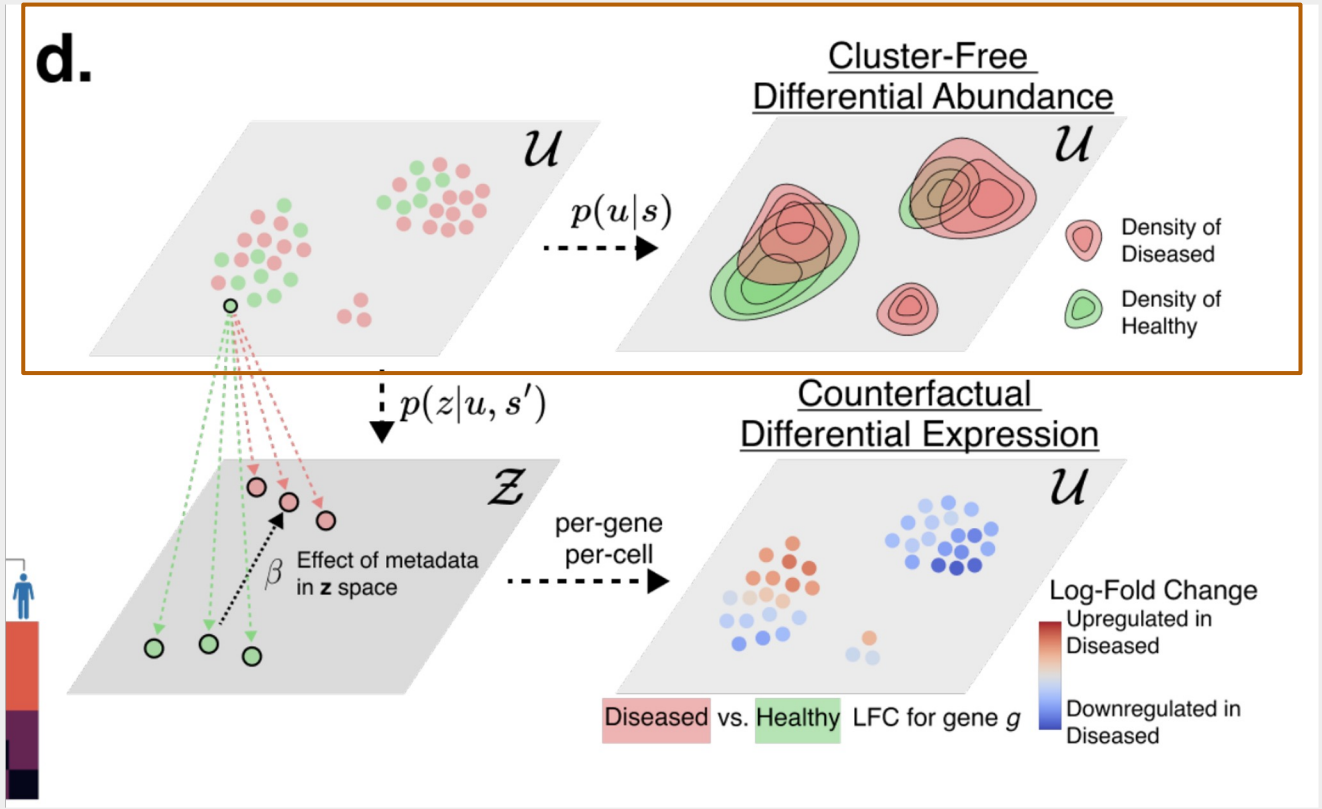
$$z_n^0 = u_n$$

We might be tempted to interpret Beta as differential expression. However, Beta cannot be interpreted as differential expression since it still lives in z(^c)-space.

Still, with the two counterfactual latent points, we can **generate** plausible x vectors (gene counts)!

We can argue that we have **eliminated unwanted batch effects** that vary across subjects (because I'm comparing with myself)

# mrVI, What Can You Do? (3/3), again, again



# Differential Abundance??? (1/3)

Remember, U-space asks: What are you? (What is this cell?)

Woah… woah… woah… Let's unpack.

“aggregated posterior distribution”
How likely is it for the cells from subject s to land onto exactly u?

u is just some cell state in u-space

$$q_s(u) := {}^{1}/_{n_s} \sum_{n:s_n=s} q(u \mid x_n)$$

s is just some subject

n are cells from that subject

# Differential Abundance??? (2/3)

$$q_s(u) := {}^1/_{n_s} \sum_{n:s_n=s} q(u \mid x_n)$$

“aggregated posterior distribution”
How likely is it for the cells from sample s to land onto exactly u?

$$q_A(u) := \frac{1}{|A|} \sum_{s \in A} q_s(u) \qquad A \subset \{1, \ldots S\}$$

This just averages that over the set of interested subjects

# Differential Abundance??? (3/3)

$$q_s(u) := {}^1/_{n_s} \sum_{n:s_n=s} q(u \mid x_n)$$

“aggregated posterior distribution”
How likely is it for the cells from sample s to land onto exactly u?

$$q_A(u) := \frac{1}{|A|} \sum_{s \in A} q_s(u)$$

$$A \subset \{1, \ldots S\}$$

This just averages that over the set of interested subjects

$$r_{AB}(u) := \log \frac{q_A(u)}{q_B(u)}.$$

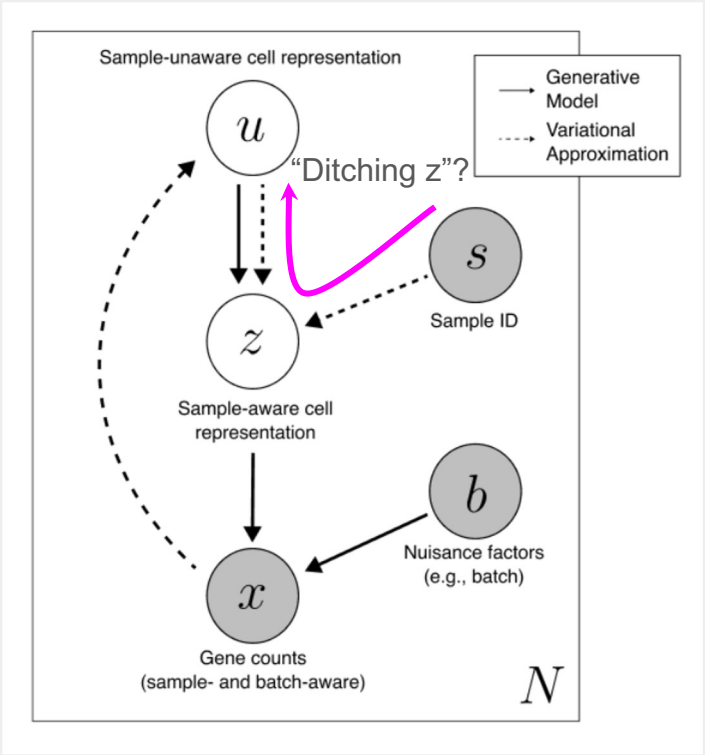
“log density ratio”
High (+) -> very likely to be in A
Low (-) -> very likely to be in B

(remember: log(1) = 0)

# Why go through all of this fuss?

1. We get a “gradient” of membership property instead of a sharp classification
2. We can argue on u-space, which does not involve neither sample nor batch effects a priori
3. We can analyze between subsets of subjects, not the entire set of them

Food for Thought: are we “ditching” z?



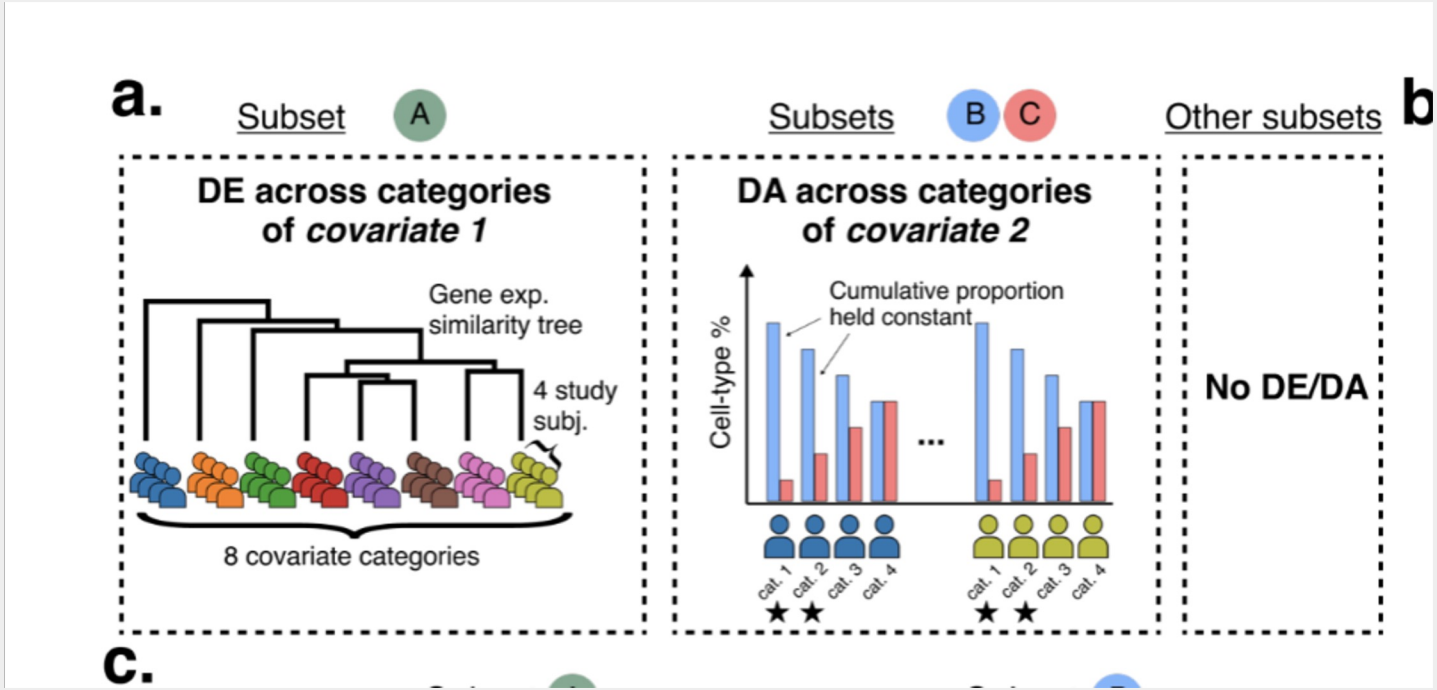
# Outline

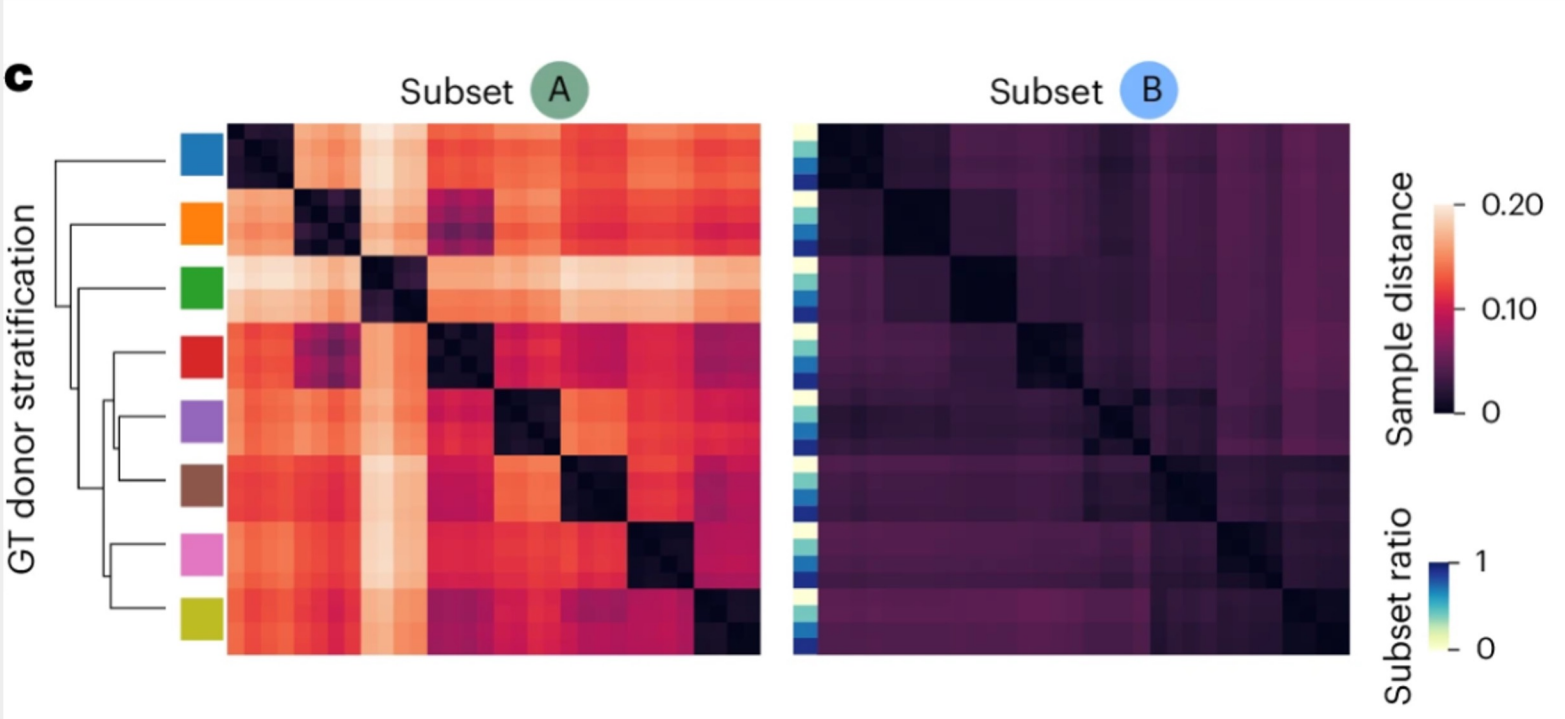
1. **Cohorts**
    a. What are cohorts?
    b. Why cohorts?

1. **Atlas Study (Liu et al.)**

3. **MrVI (Boyeau et al.)**
    a. scVI Recap
    b. Challenges in scVI
    c. MrVI: An intuition
    d. Case studies

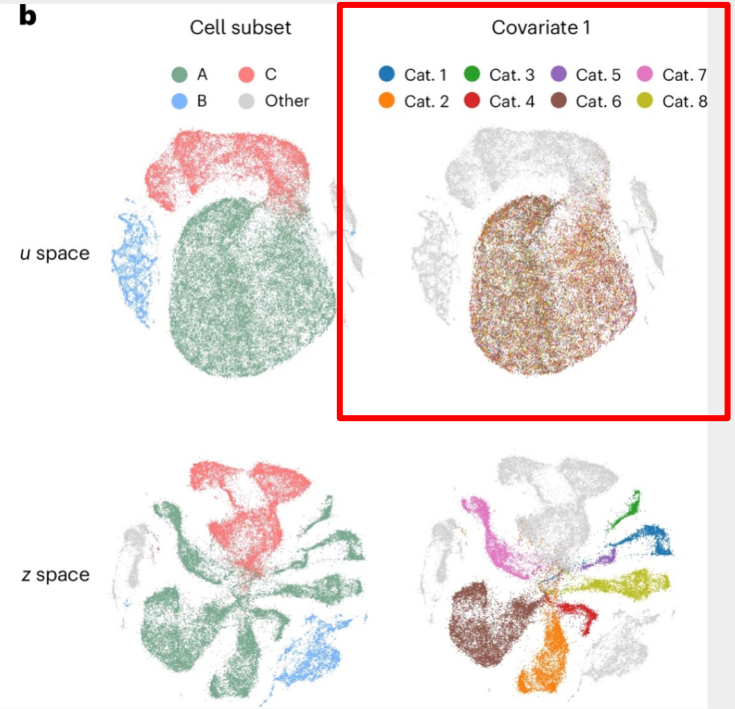
# Semi-Synthesis Dataset



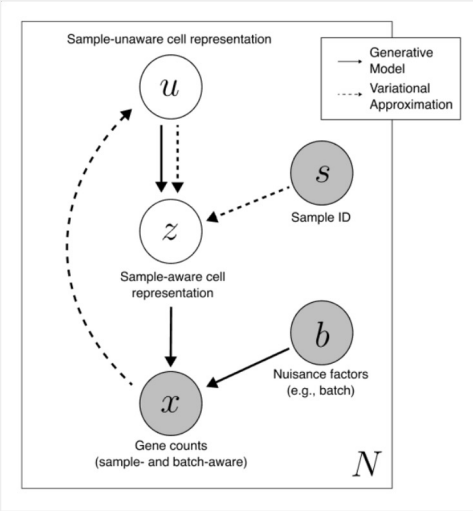

c
Subset A
Subset B
GT donor stratification
Sample distance
0.20
0.10
0
Subset ratio
1
0

Fragments into **subject-specific sub-clusters** (colored by covariate 1)

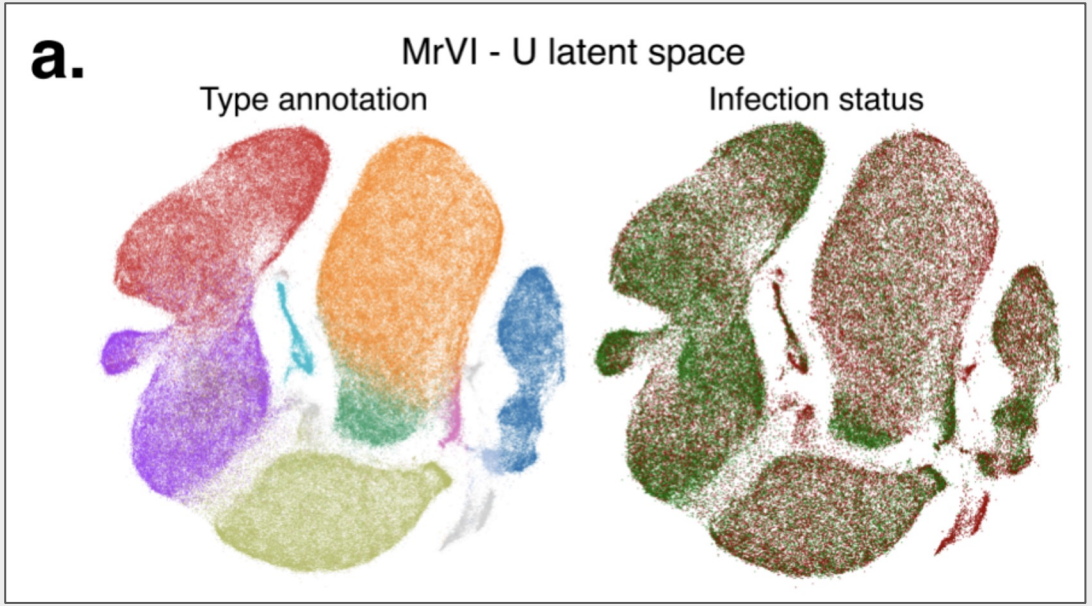
U -> Z




inter-mixed



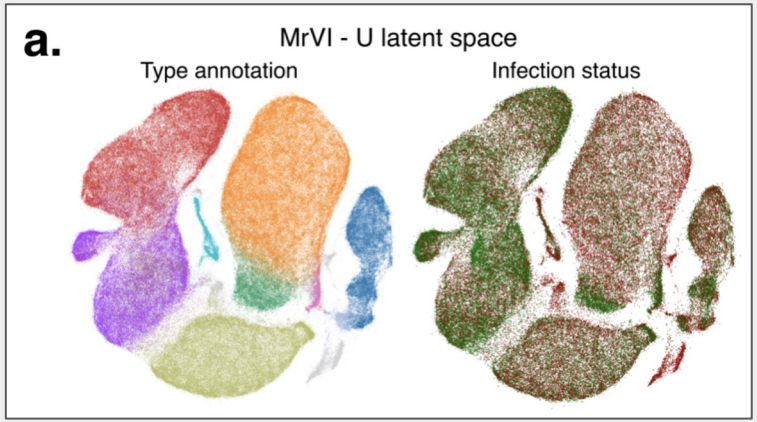
# COVID Case Study (1): **Grouping / Stratification**

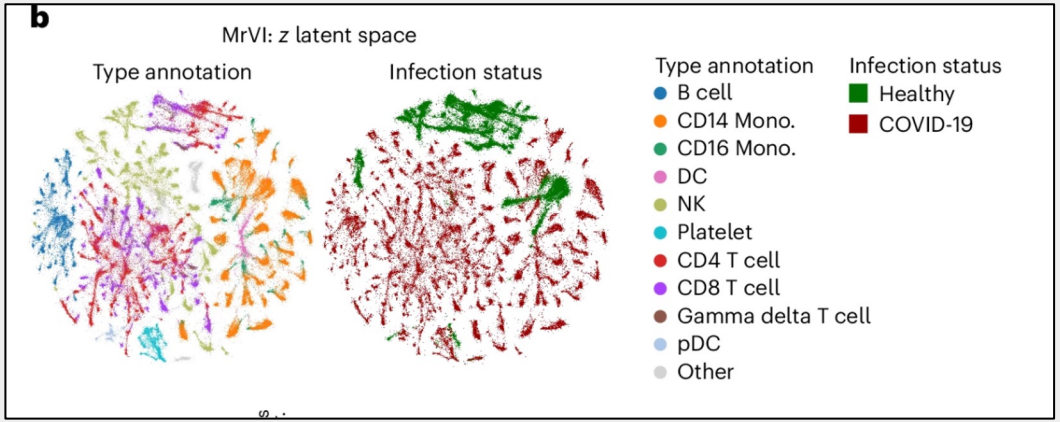



Question: Why is this the mixing of infection status important in u-space?

Question 2: What can we say about the training if they are not mixed?

# COVID Case Study (1): **Grouping / Stratification**



# Case Study (2): MrVI enables **grouping** and **characterization** of small molecules in screening assays

**sci-Plex:** high-throughput single-cell perturbation screen

Each cell tagged with a unique “barcode” indicating which drug + dose it received

## What is our ideal expectation?

| U-Space | Z-Space |
| --- | --- |
| | |

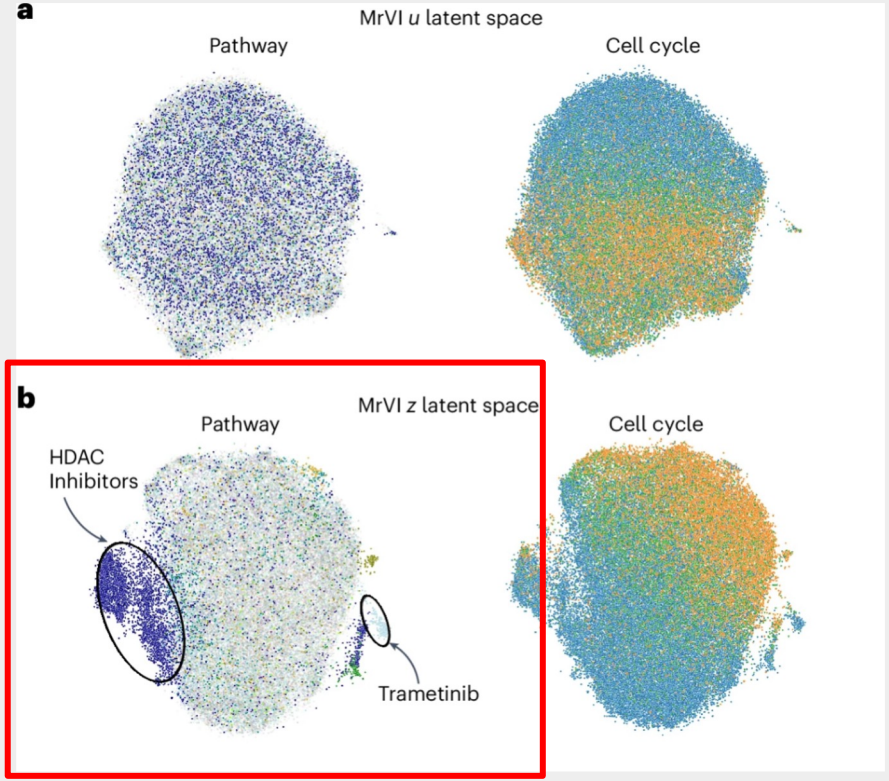
# Case Study (2): MrVI enables **grouping** and **characterization** of small molecules in screening assays

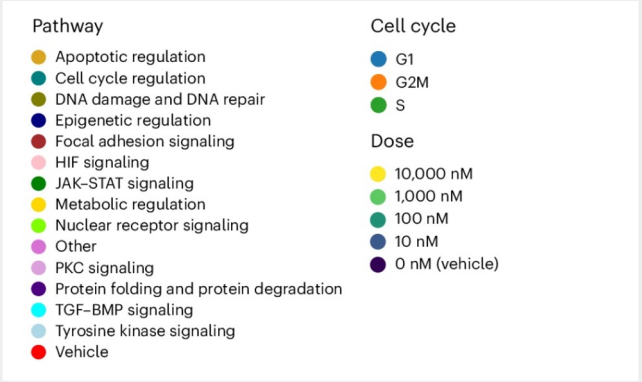
Each cell tagged with a unique “barcode” indicating which drug + dose it received

**What is our ideal expectation?**

| U-Space | Z-Space |
| --- | --- |
| **Baseline Cell Identity** | **Effect of Drugs** |
| **(What are you?)** | **(And…How are you? Behavior)** |

# Case Study (2): MrVI enables **grouping** and **characterization** of small molecules in screening assays

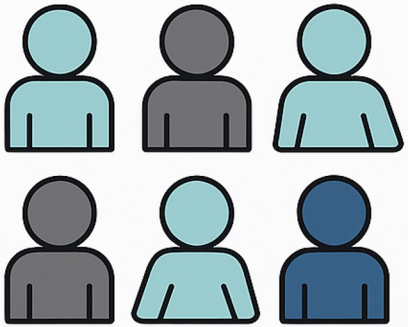



Distinct subclusters emerge,
each corresponding to specific drug classes
such as HDAC inhibitors or Trametinib (MEK inhibitor)

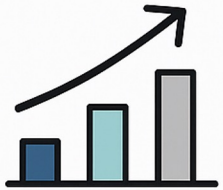
# What Are Cohorts?

A cohort is a group of individuals who share a common context



Shared condition:
NSCLC

**Different features:**
sex, PD-L1, TIME subtype



**Measure outcome:**
MPR / RFS