Novartis Institutes for BioMedical Research
Oncology Data Science

Denoising of Single Cell Sequencing Data with Probabilistic Deep Learning

Caibin Sheng, PhD
AMLD EPFL 2022
March 28th, 2022

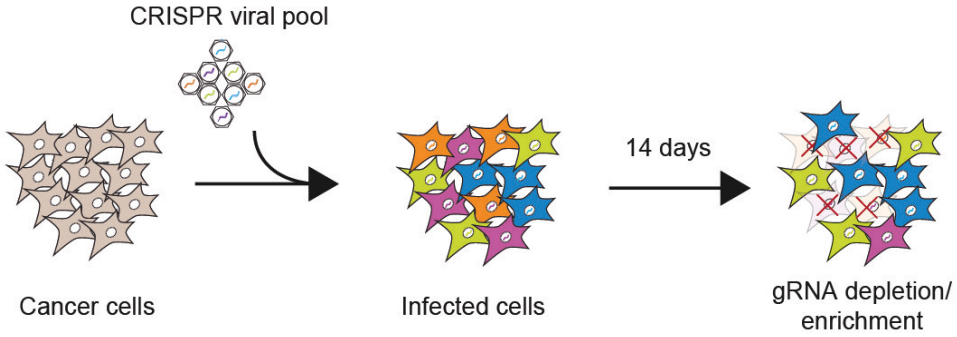
NOVARTIS | Reimagining Medicine

Overview

- Single-cell CRISPR screens in drug target identification (**WHY?**)
- A technical issue in single-cell CRISPR screens (**WHAT?**)
- How do we solve it with machine learning? (**HOW?**)
- Generalization ability?

NOVARTIS | Reimagining Medicine

Genetic screens in drug target identification

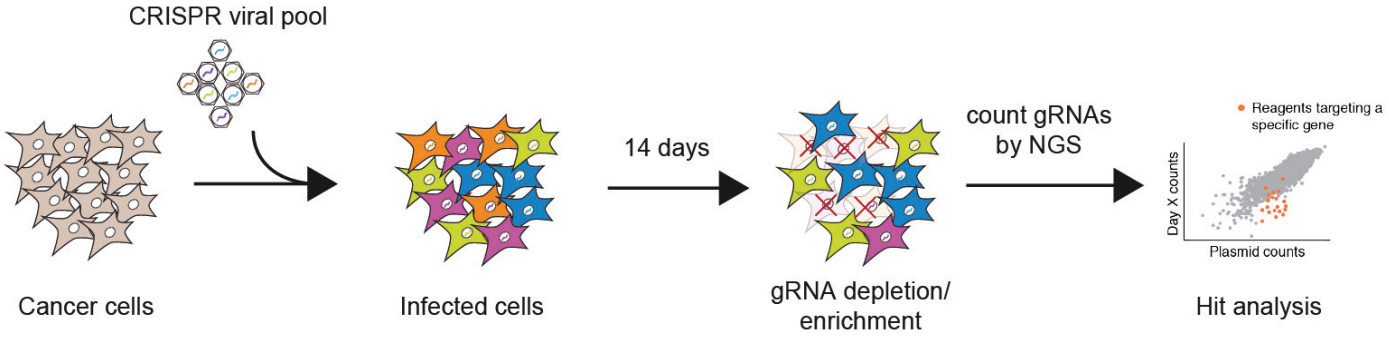


- Functional genomics identify targets regulating cancer cell proliferation

Novartis' Project DRIVE
Broad's Project Achilles
Sanger's Project Score

NOVARTIS | Reimagining Medicine

Genetic screens in drug target identification

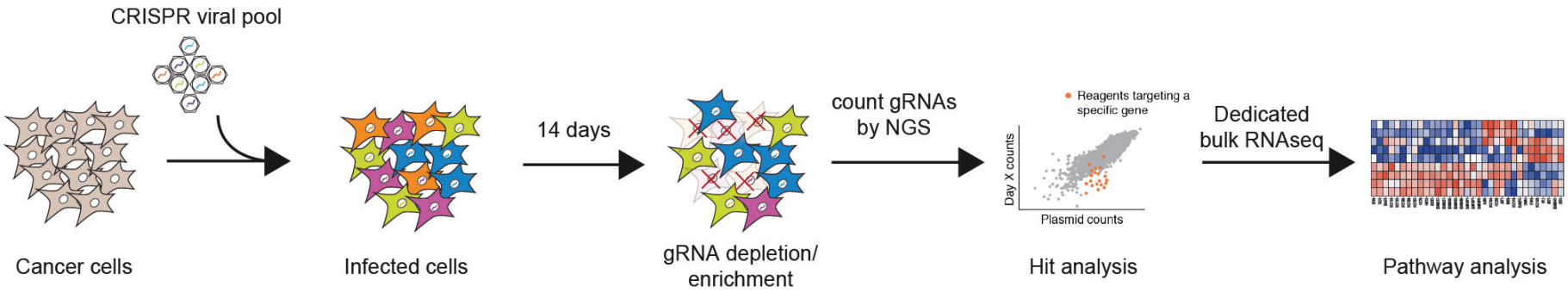


- Functional genomics identify targets regulating cancer cell proliferation

Novartis' Project DRIVE
Broad's Project Achilles
Sanger's Project Score

NOVARTIS | Reimagining Medicine

Genetic screens in drug target identification

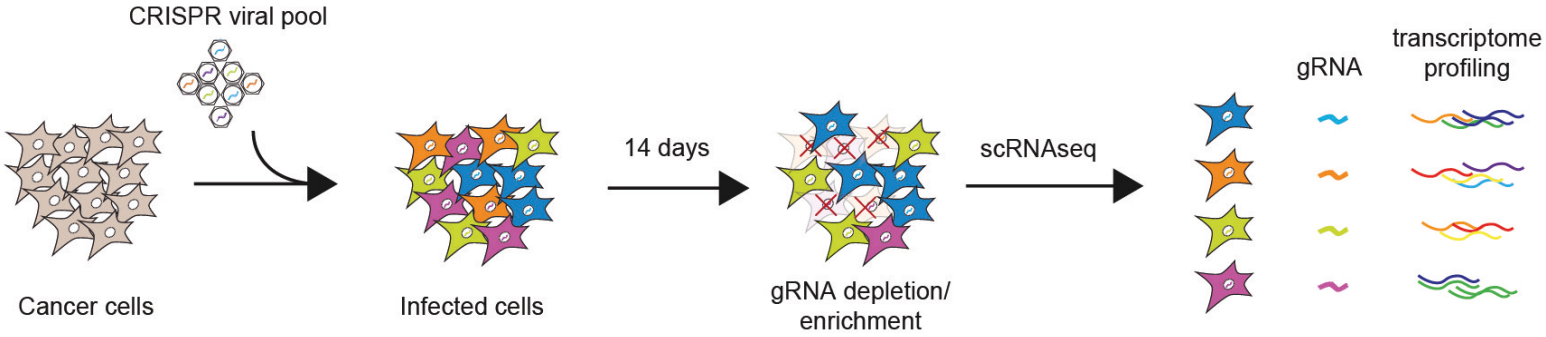


- Functional genomics identify targets regulating cancer cell proliferation
- Dedicated bulk RNA-seq validation identifies target-specific signatures
- Low efficiency

Novartis' Project DRIVE
Broad's Project Achilles
Sanger's Project Score

NOVARTIS | Reimagining Medicine

Single-cell omics facilitate drug target identification



- Single-cell CRISPR screens = functional screens + single cell RNA-seq
- Assignment of the gRNA is crucial

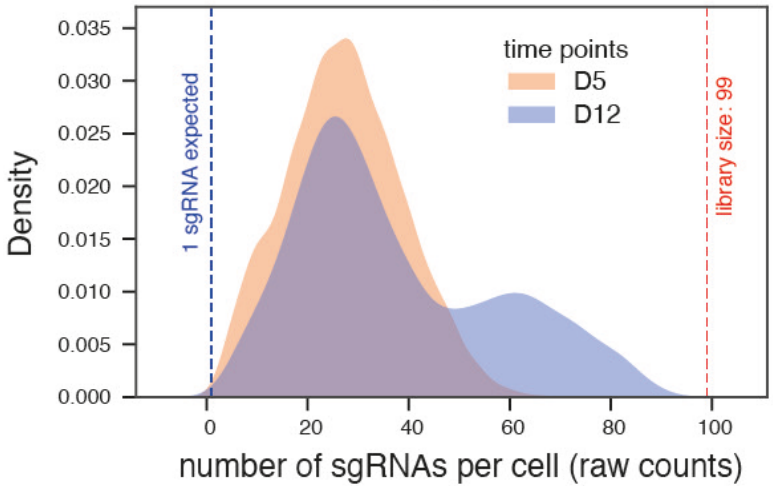
Dixit, A. et al. 2016, Cell
Adamson, B. et al. 2016, Cell

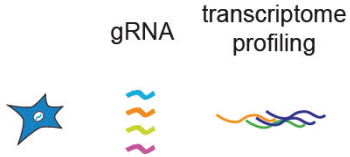
Xie, S. et al. 2017, Molecular Cell
Jaitin, D.A. et al. 2016, Cell

Datlinger, P. et al.
2017, Nature Methods

NOVARTIS | Reimagining Medicine

single-cell data is highly noisy





How can we identify the true signal?

Multiple gRNAs are detected in every cell

Sheng, C. et al. 2022, bioRxiv

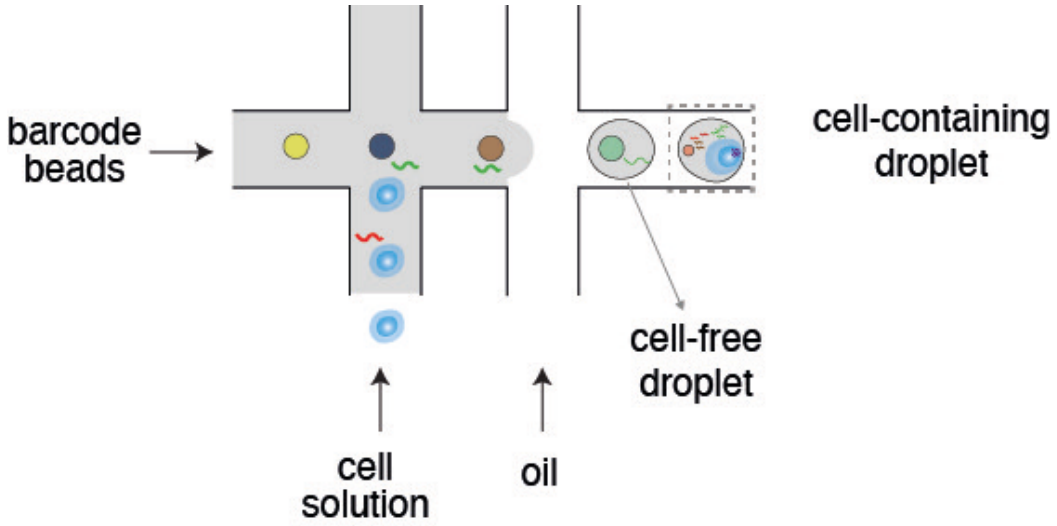
NOVARTIS | Reimagining Medicine

Overview

- Single-cell CRISPR screens in drug target identification (**WHY?**)
- A technique issue in single-cell CRISPR screens (**WHAT?**)
- How do we solve it with machine learning? (**HOW?**)
  1) Investigated the details of the technology
  2) Built a deep generative model based on the finding
  3) Optimized the model with synthetic data
  4) Validated the model in real cases
- Generalization ability?

NOVARTIS | Reimagining Medicine

How is the noise generated?

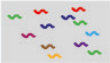


- RNAs are released from broken cells and they float around in single cell suspension
- Droplets capture not only cells but also ambient RNAs (i.e. floating RNAs)

NOVARTIS | Reimagining Medicine

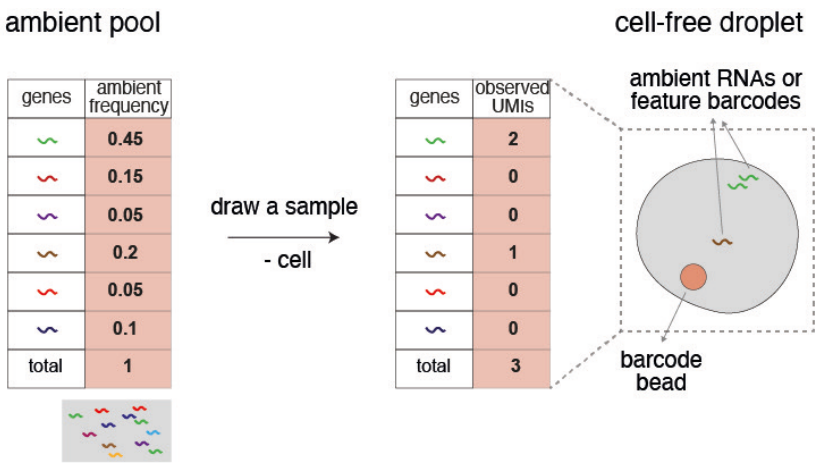
ambient pool

genes	ambient frequency
	0.45
	0.15
	0.05
	0.2
	0.05
	0.1
total	1



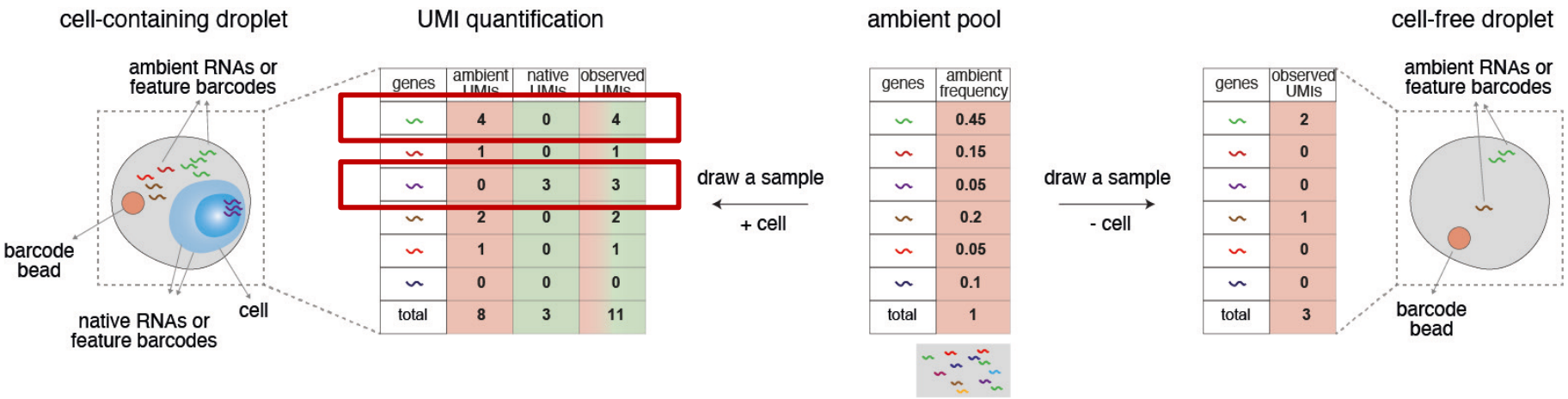
- Frequency of each RNA varies

NOVARTIS | Reimagining Medicine



- Frequency of each RNA varies

NOVARTIS | Reimagining Medicine



- Frequency of each RNA varies
- Hard to distinguish between noise and true signal by their raw counts

NOVARTIS | Reimagining Medicine

A deep generative model simulates the generation of noise

$$(1-\varepsilon) \times \beta + \varepsilon \times \alpha = \theta$$

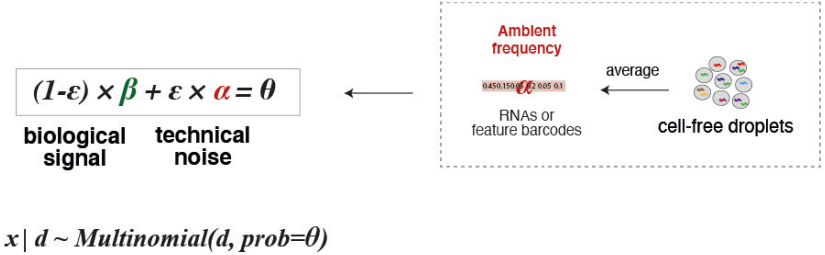
biological signal — technical noise

$$x \mid d \sim Multinomial(d, prob=\theta)$$

Sheng, C. et al. 2022, bioRxiv
https://github.com/Novartis/scAR

NOVARTIS | Reimagining Medicine

A deep generative model simulates the generation of noise



$$x \mid d \sim Multinomial(d, prob=\theta)$$

Sheng, C. et al. 2022, bioRxiv
https://github.com/Novartis/scAR

NOVARTIS | Reimagining Medicine

Autoencoder

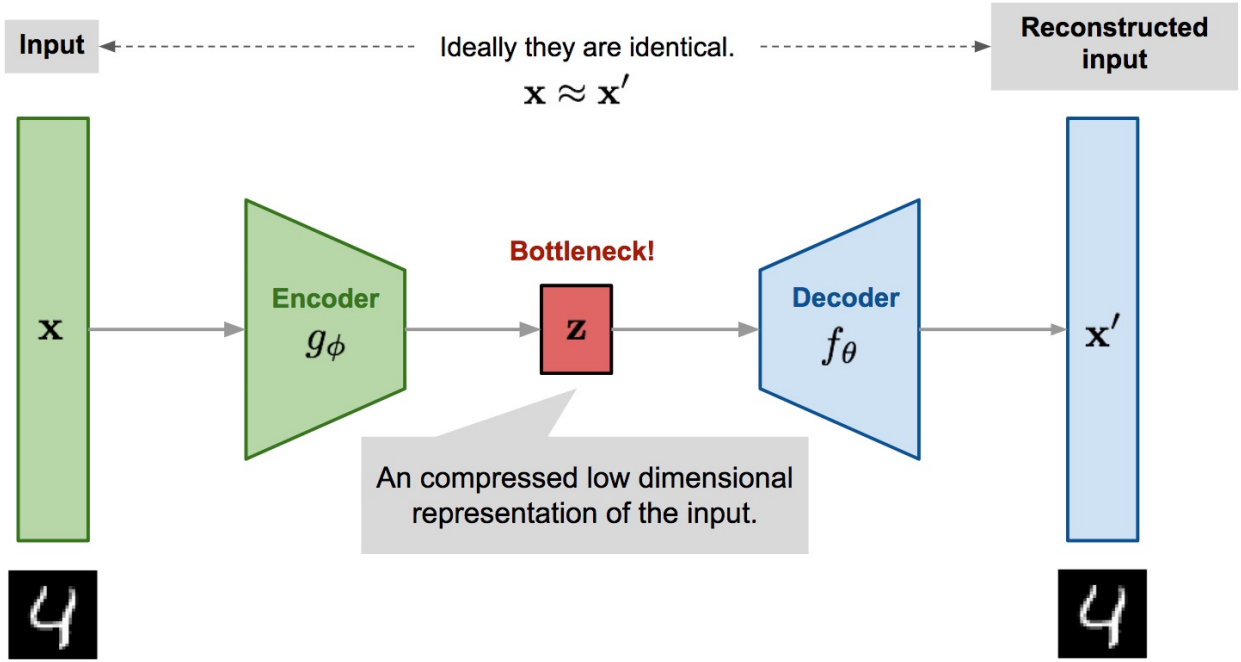
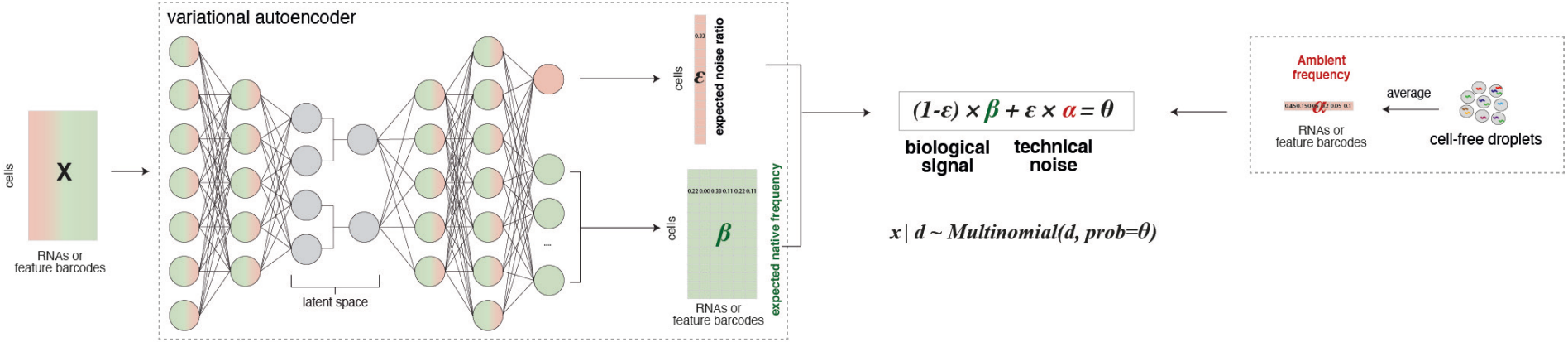


Image from https://lilianweng.github.io/posts/2018-08-12-vae/

NOVARTIS | Reimagining Medicine

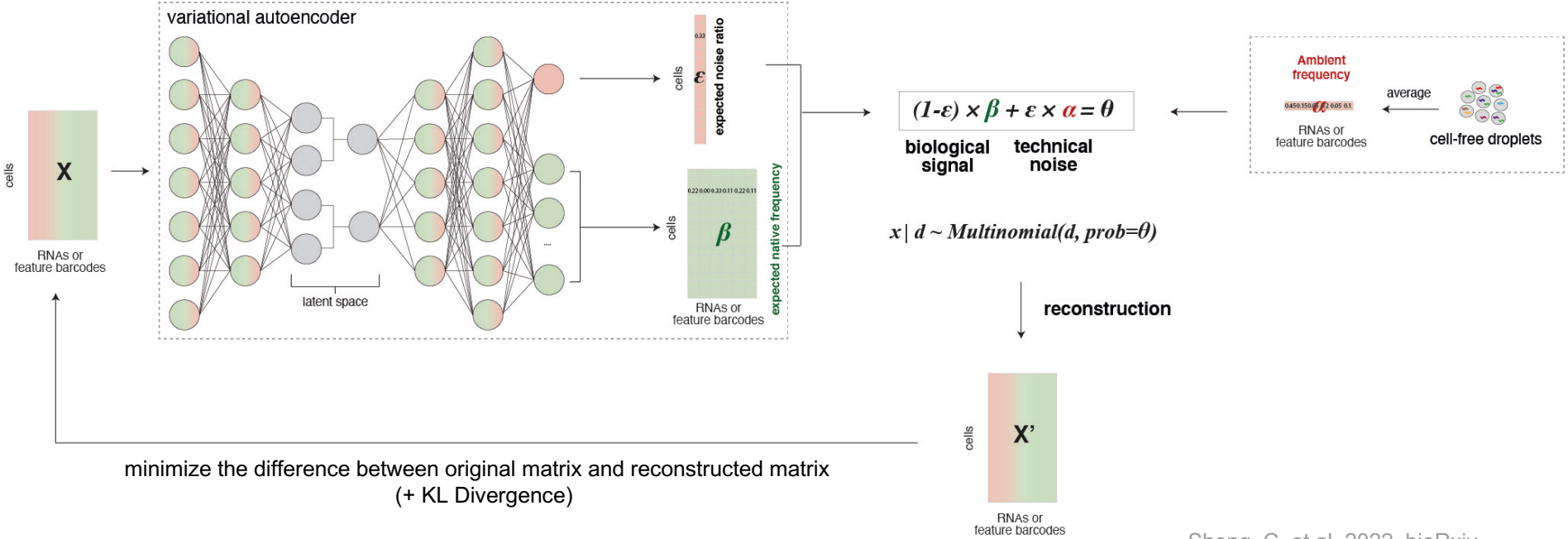
A deep generative model simulates the generation of noise



Sheng, C. et al. 2022, bioRxiv
https://github.com/Novartis/scAR

NOVARTIS | Reimagining Medicine

A deep generative model simulates the generation of noise

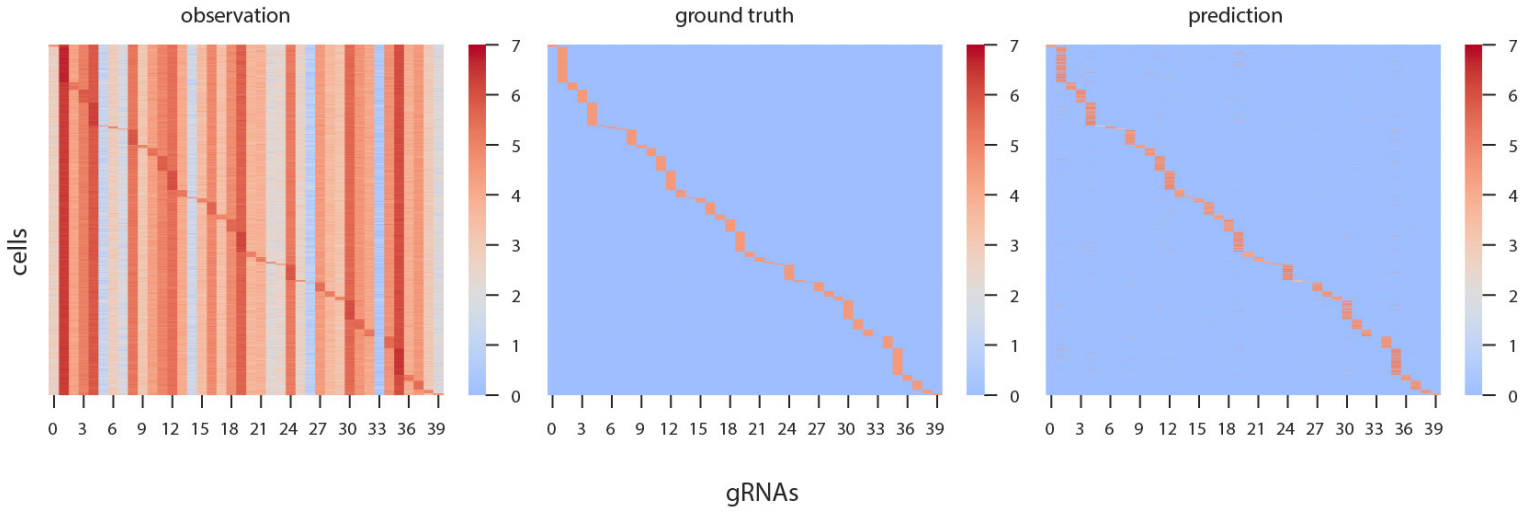


minimize the difference between original matrix and reconstructed matrix
(+ KL Divergence)

Sheng, C. et al. 2022, bioRxiv
https://github.com/Novartis/scAR

NOVARTIS | Reimagining Medicine

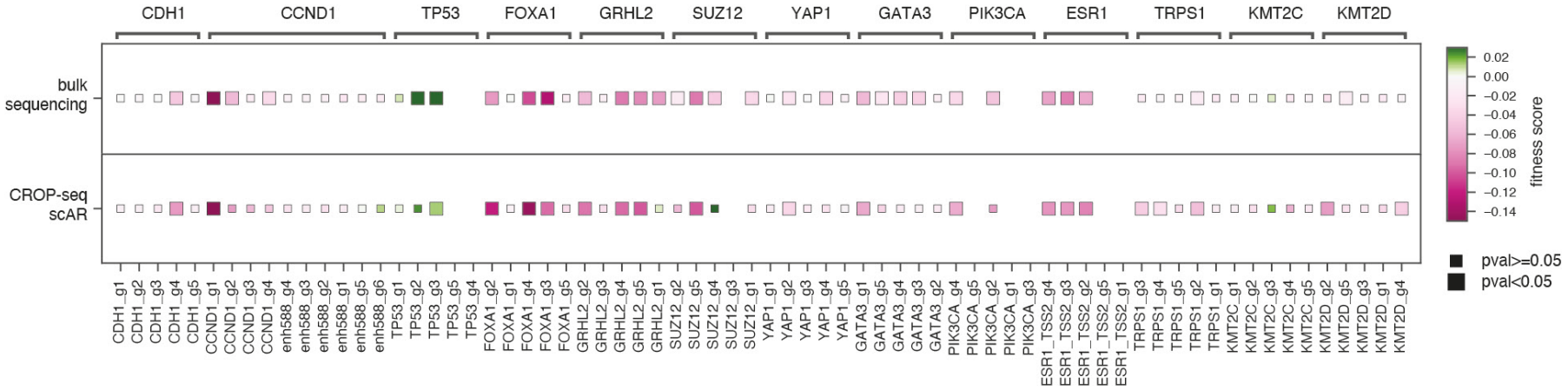
Model validation with synthetic data



Noise level: 97.5%
scAR: 89%

NOVARTIS | Reimagining Medicine

scAR enables hit analysis in single-cell CRISPR screens



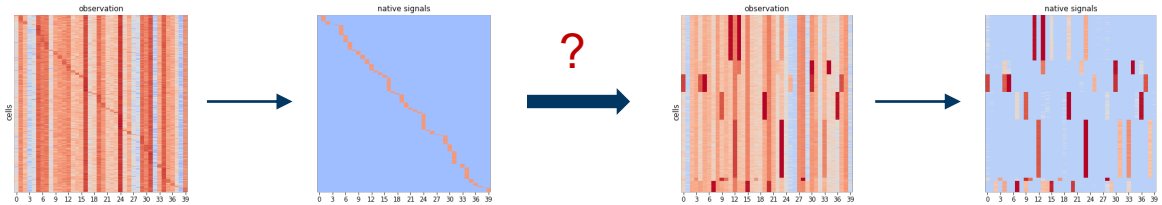
- scAR identifies most of lethal gRNAs

Sheng, C. et al. 2022, bioRxiv

NOVARTIS | Reimagining Medicine

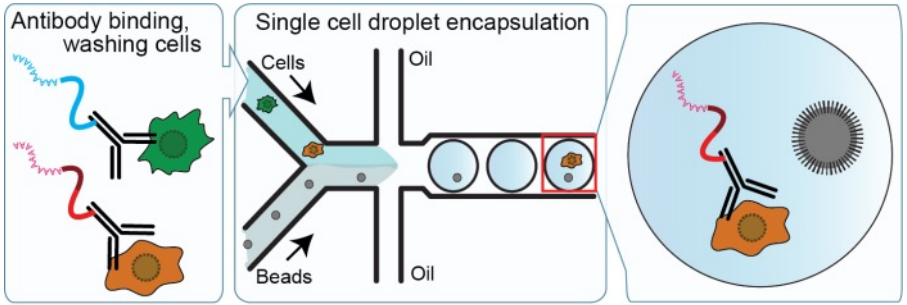
Overview

- Single-cell CRISPR screens in drug target identification (**WHY?**)
- A technique issue in single-cell CRISPR screens (**WHAT?**)
- How do we solve it with machine learning? (**HOW?**)
- Generalization ability?



NOVARTIS | Reimagining Medicine

CITE-seq



Stoeckius, M., et al. 2017, *Nature Methods*

- It allows simultaneous measurement of single cell transcriptome and proteins
- It becomes more and more popular in Immuno-oncology

NOVARTIS | Reimagining Medicine

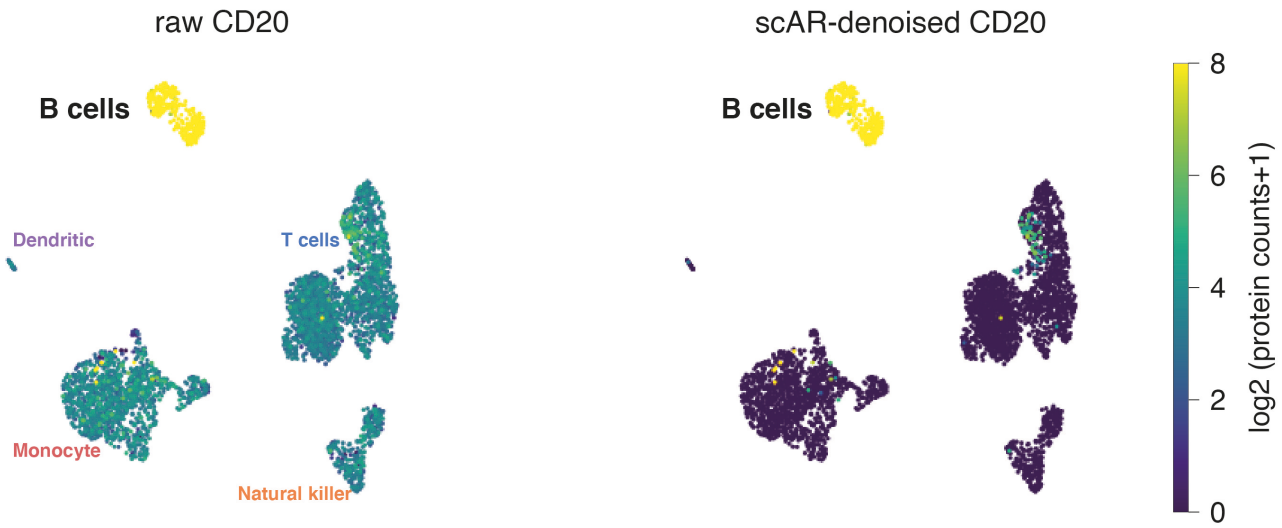
scAR removes background noise in CITE-seq

- CD20 is a protein that exclusively expressed in B cells

Sheng, C. et al. 2022, bioRxiv

NOVARTIS | Reimagining Medicine

scAR removes background noise in CITE-seq



- CD20 is a protein that exclusively expressed in B cells

Sheng, C. et al. 2022, bioRxiv

NOVARTIS | Reimagining Medicine

Summary

- Single-cell technologies have huge potential in drug target identification, however, they suffer from substantial noise.
- We developed a ML approach (called scAR) to remove the background noise.
- We applied scAR to several single-cell technologies, and it shows high performance.

GitHub: https://github.com/Novartis/scAR/

NOVARTIS | Reimagining Medicine

Acknowledgements

ONC DS

Gang Li

Antoine de Weck

Slavica Dimitrieva

Foivos Gypas

Mathias Rechenman

Tobias Ternent

Eric Durand

Audrey Kauffmann

ONC

Rui Lopes

Giorgio Galli

Mathias Eder

Esther Ujittewaal

CBT

Guglielmo Roma

Ulrike Naumann

Annick Waldt

Rachel Cuttat

Sven Schuierer

Walter Carbone

Postdoc program

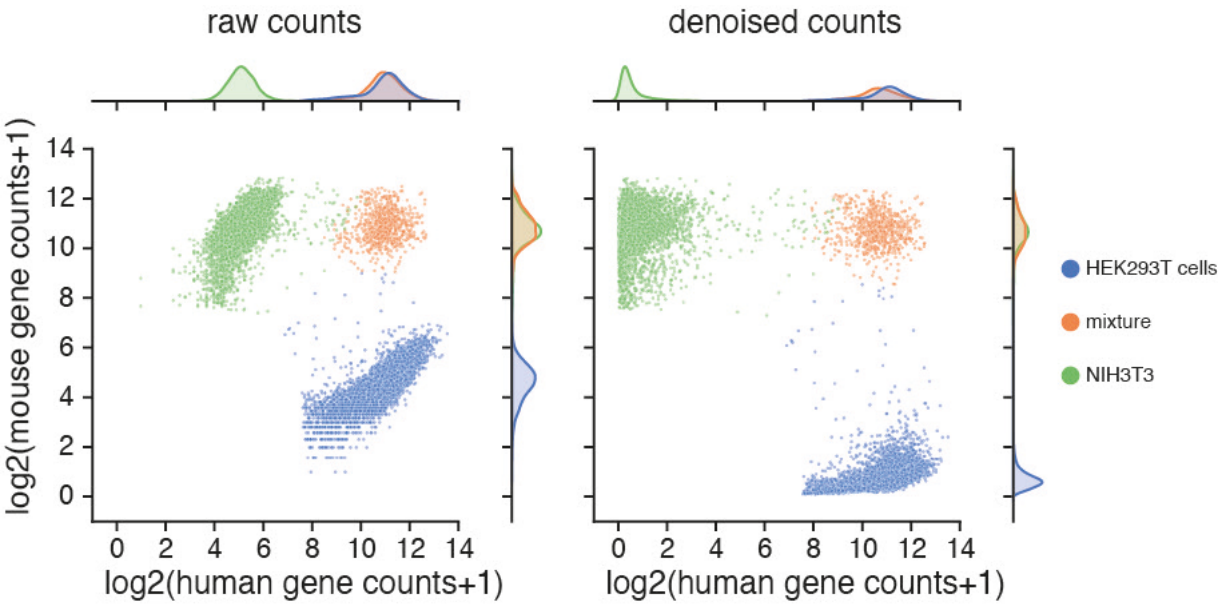
Anne Granger

OpentoWork

Caibin Sheng

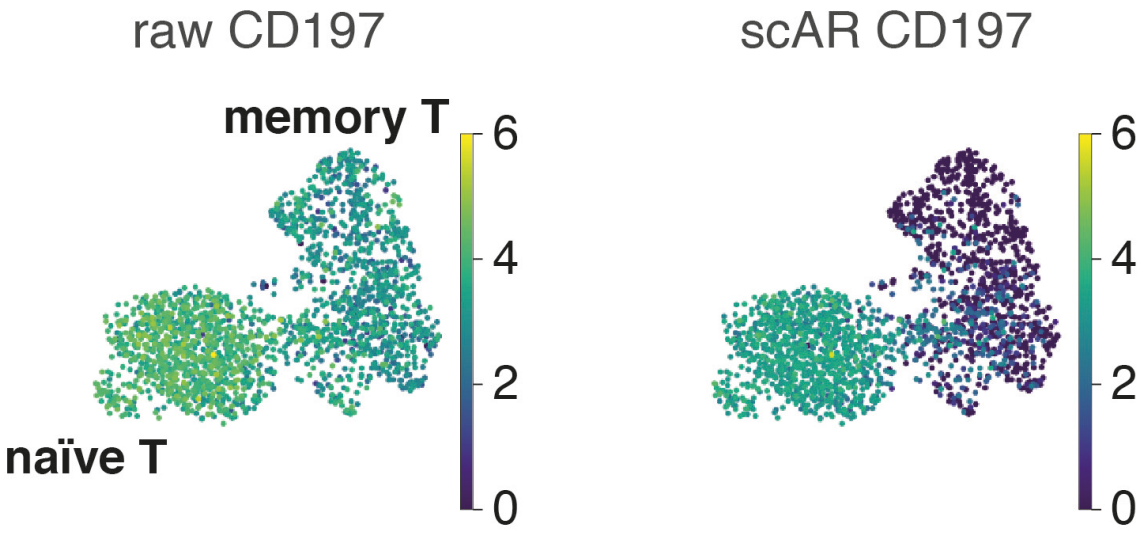
Data scientist, single cell analysis, machine learning, AI drug devel...

scRNAseq



NOVARTIS | Reimagining Medicine

scAR improves immunophenotyping

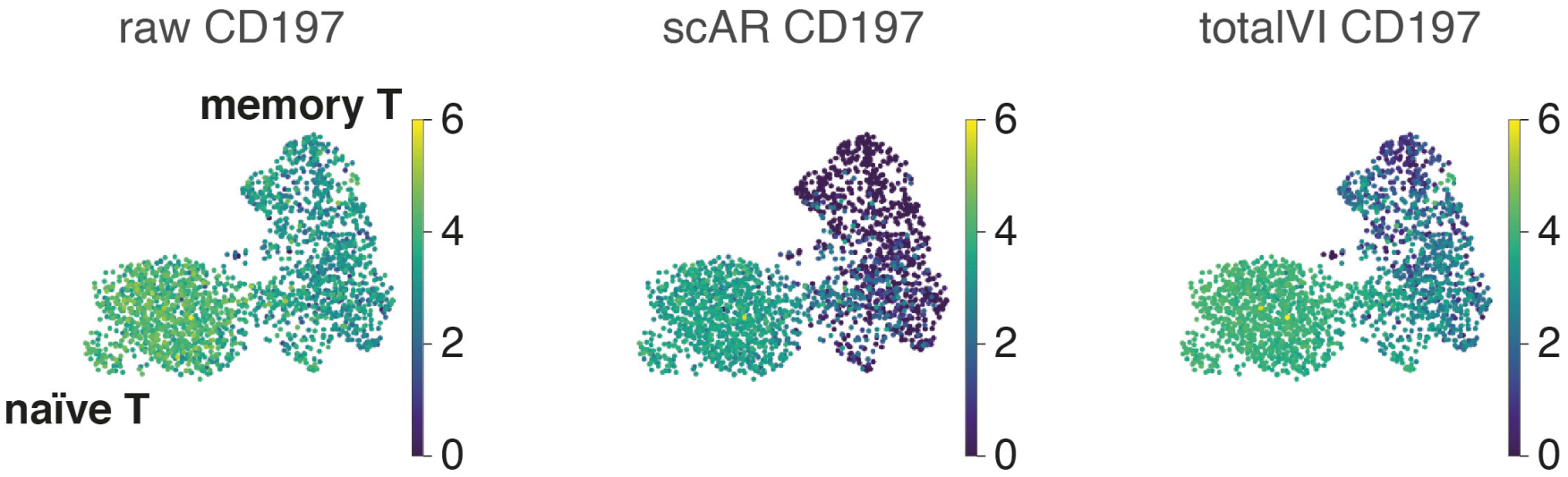


- CD197 is a protein marker to distinguish between memory T and naïve T cells

Sheng, C. et al. 2022, bioRxiv

NOVARTIS | Reimagining Medicine

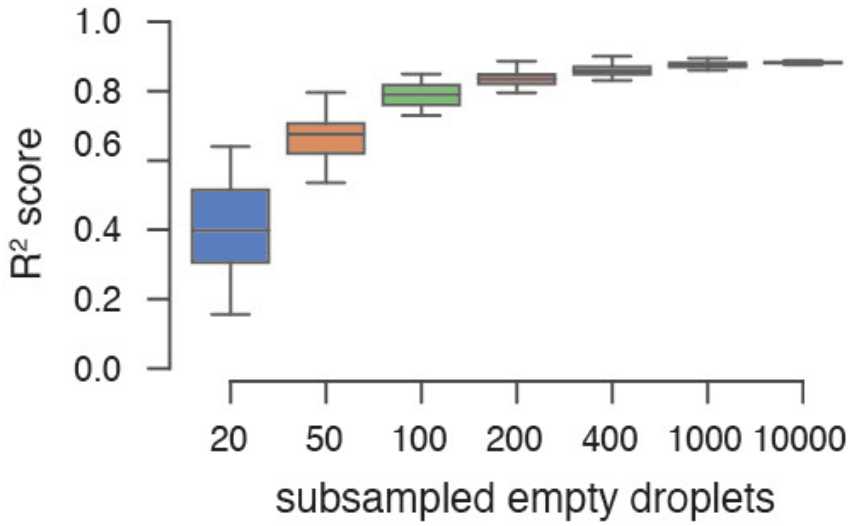
scAR outperforms the state-of-the-art approach



- CD197 is a protein marker to distinguish between memory T and naïve T cells

Gayoso, A. et al. 2021, Nature Method
Sheng, C. et al. 2022, bioRxiv

NOVARTIS | Reimagining Medicine



- Frequencies of sgRNAs are consistent in randomly sampled empty droplets
- Background signal is not random noise but deterministic

NOVARTIS | Reimagining Medicine