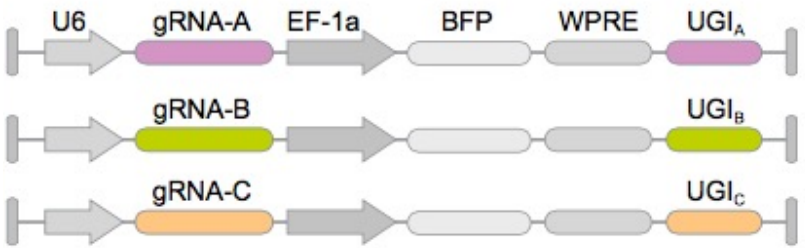
Analytical strategies for single-cell pooled CRISPR screens

Efthymia Papalexi
NYGC/NYU
Single Cell Genomics Day 2021

Current methods combine single-cell RNA-seq with pooled CRISPR screens.

Guide barcode capture

CRISP-seq

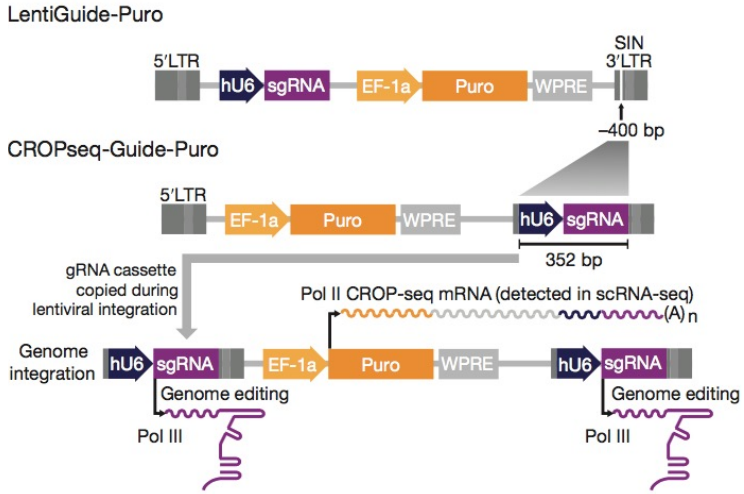


Perturb-seq

Dixit *et al.* 2016
Jaitin *et al.* 2016

gRNA polyA transcript capture

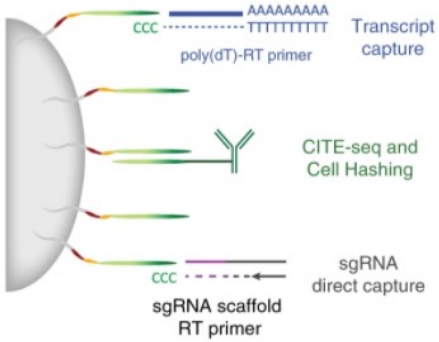
CROP-seq



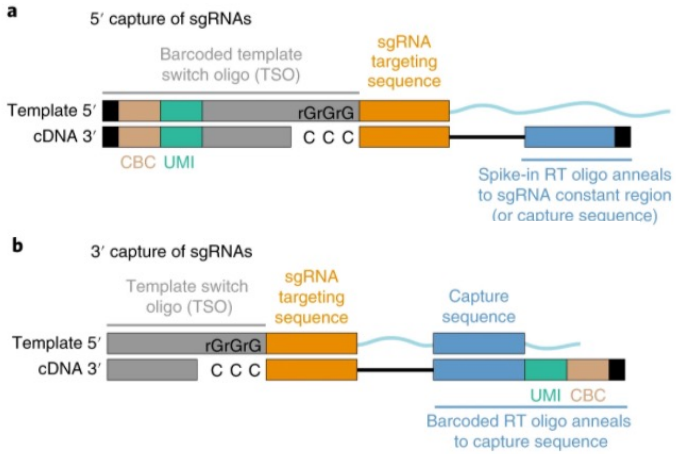
Adamson *et al.* 2016
Datlinger *et al.* 2017

Scaffold-based gRNA capture

ECCITE-seq



Direct gRNA capture Perturb-seq



Replogle *et al.* 2020
Mimitou *et al.* 2019

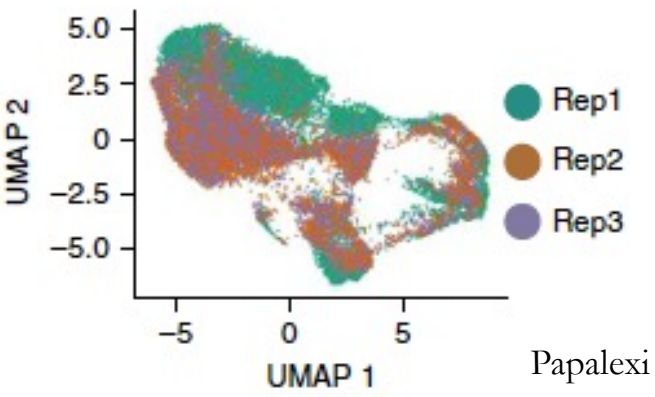
Applying single-cell pooled CRISPR screens to understand gene function.

- We can study the function of multiple genes in high throughput (10- 100s of genes at a time)
- We are no longer associating a gene to a single phenotype (surface protein expression, cell survival or drug resistance)
- We can combine gene perturbations to understand gene interactions
- Finally, we can create comprehensive maps of regulatory networks

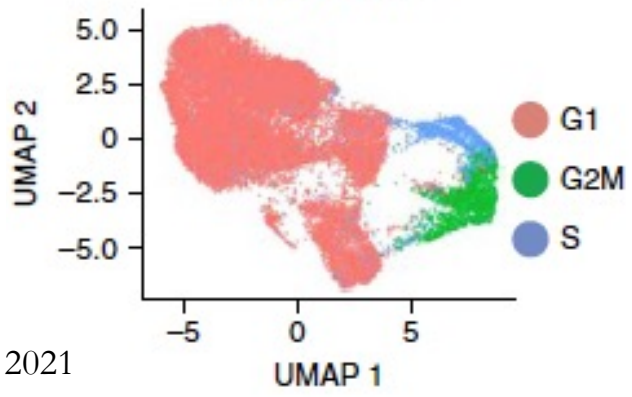
However, these types of studies present unique analytical challenges.

1. Technical variation drives mRNA-based clustering.

Biological replicate

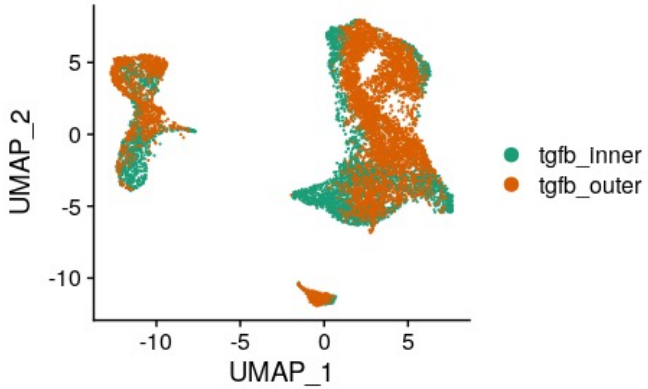


Cell cycle phase

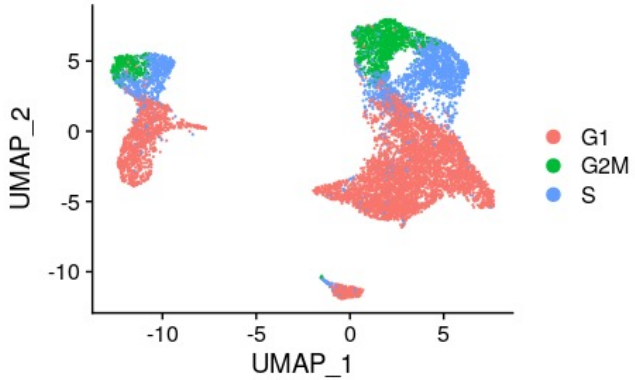


Papalexi *et al.* 2021

Cell location



Cell cycle phase



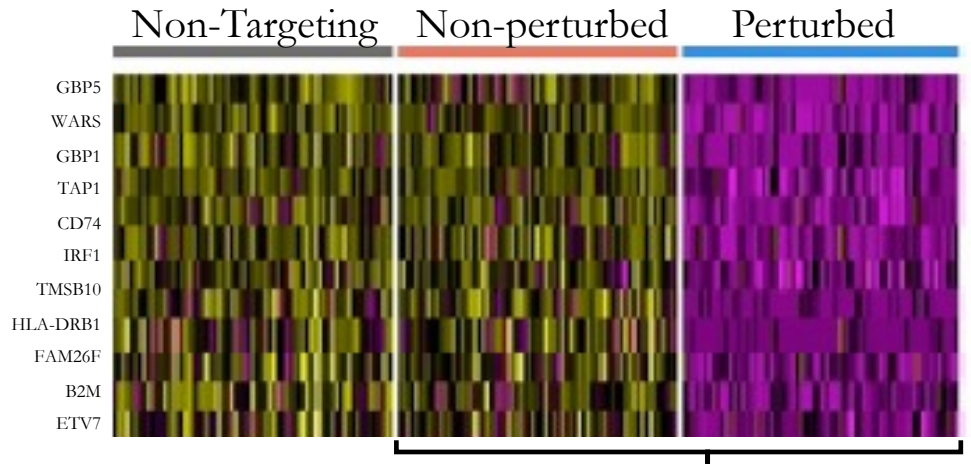
McFaline-Figueroa *et al.* 2019

- scRNA-seq clustering is often driven by technical noise (cell cycle, replicate ID, tissue origin)
- This noise masks perturbation-specific effects

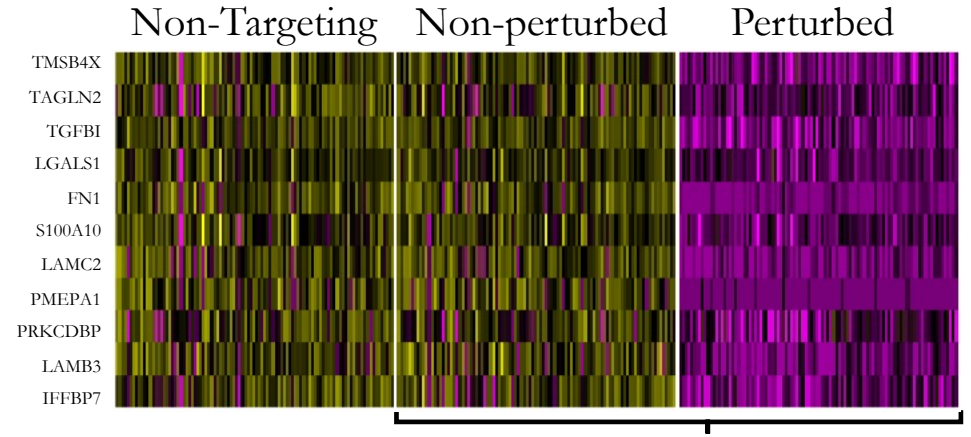
However, these types of studies present unique analytical challenges.

2. Responses to perturbation can be heterogeneous: CRISPR 'escaping' cells

- Low gRNA targeting efficiency leaves cells unperturbed.
- In frame mutations do not affect the function of the protein.
- 'Escaping' cells can be found in CRISPRko, CRISPRi and CRISPRa single-cell datasets.
- The presence of these cells hurts our ability to associate perturbations to gene signatures.



Papalexi *et al.* 2021

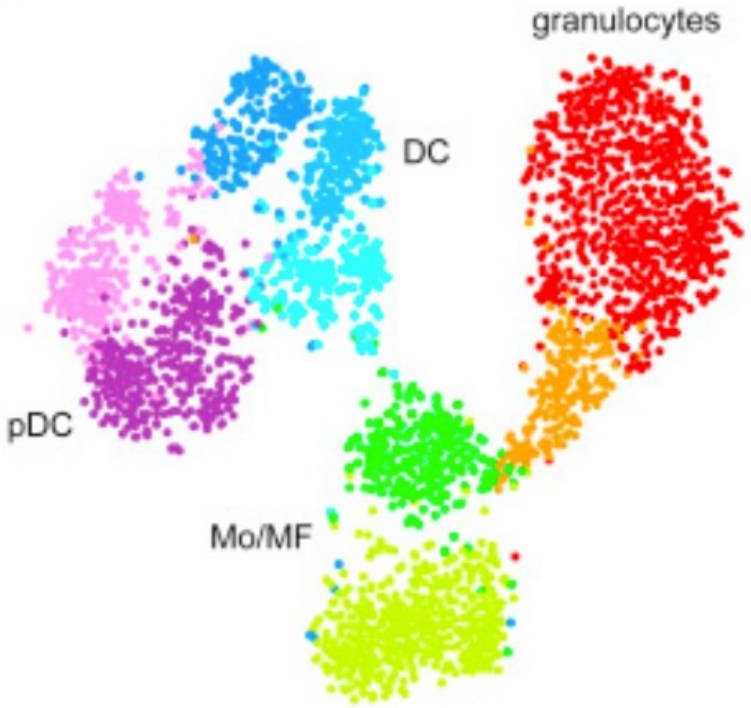


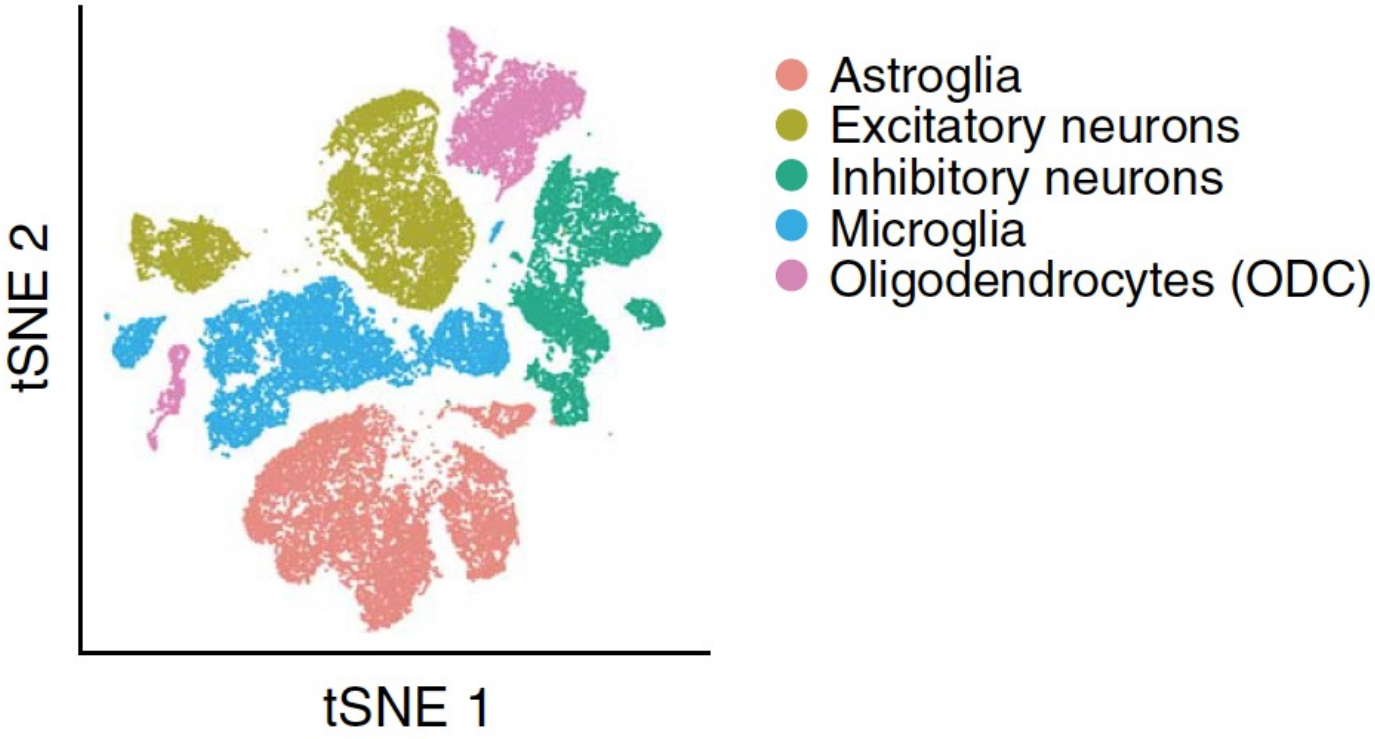
*All cells express the same gRNA

McFaline-Figueroa *et al.* 2019

However, these types of studies present unique analytical challenges.

3. Responses to perturbation can be heterogeneous: cell type-specific responses





Jaitin *et al.* 2016

Jin *et al.* 2020

There is a need for analytical strategies to overcome these challenges.

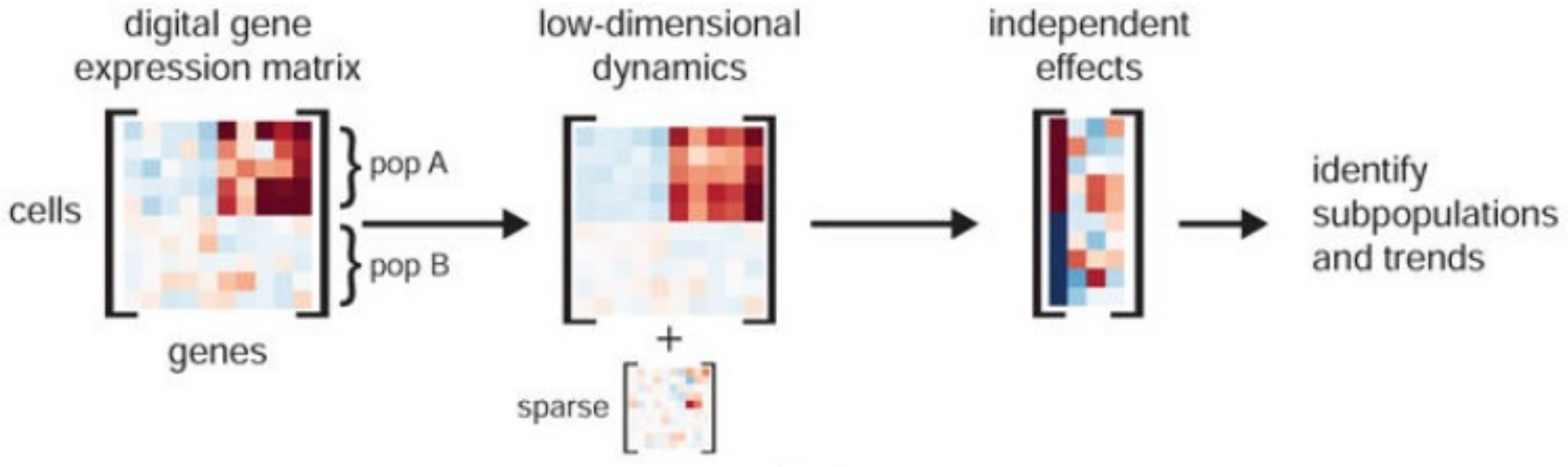
Ideally, we want to:

- Be able to identify and remove confounding sources of variation in data.
- Identify and remove CRISPR 'escaping' cells.
- Account for cell type heterogeneity and quantify the effect of each perturbation across cell types.
- Identify genes whose expression changes upon perturbation.

Current methods to address these challenges:

- Pioneering (developed in 2016): LRICA, MIMOSCA
- More recent: scMAGeCK , MUSIC, WGCNA and Mixscape.

Using low rank independent component analysis to characterize the effects of each perturbation on gene expression: LRICA

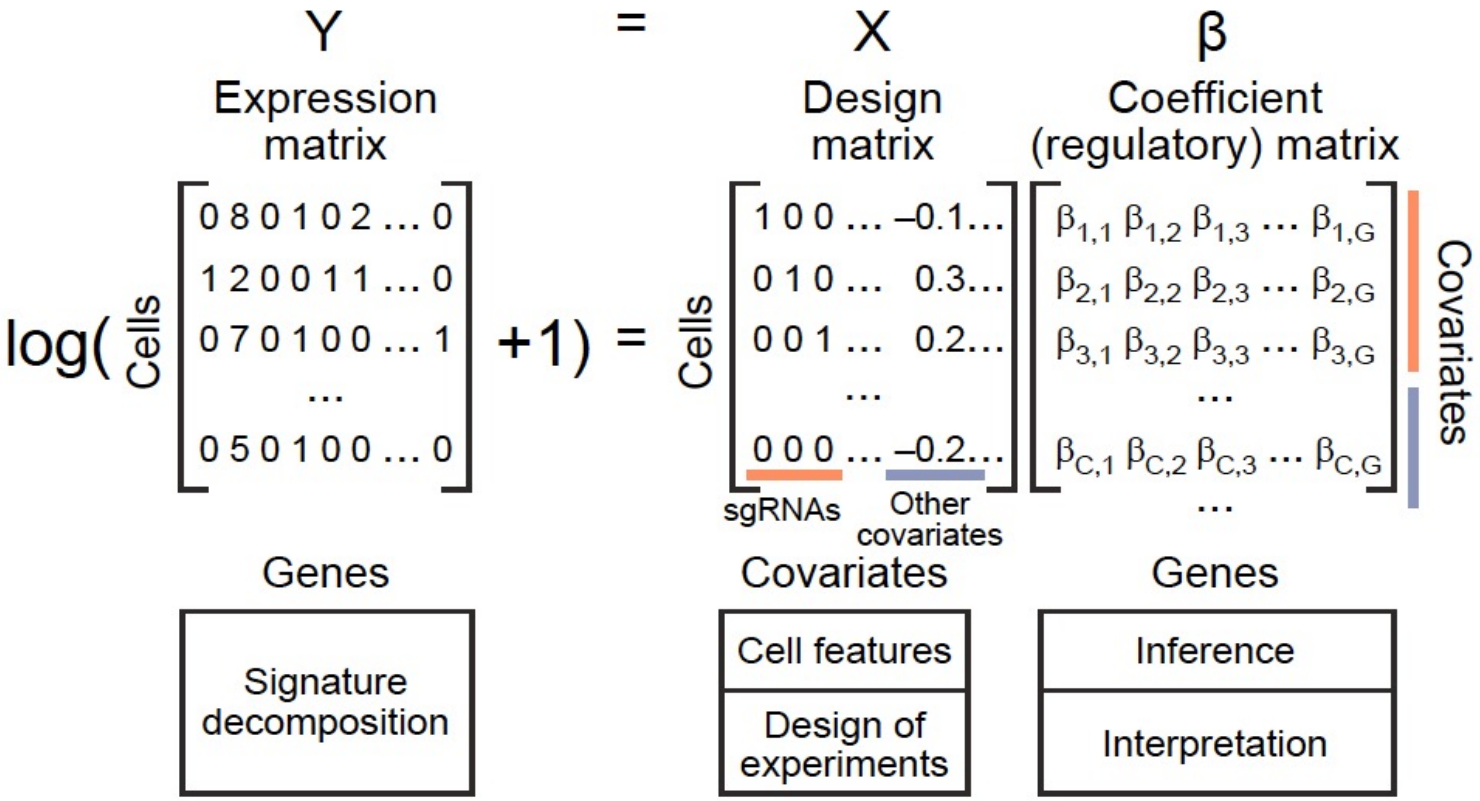


- Used robust PCA to decompose the low-dimensional dynamics of the population from the noise.
- Identify informative trends in the low-dimensional dynamics using independent component analyses (ICA).
- Associate sub-populations with specific gRNAs to each component and define genes that drive their cellular behavior.
- Train a OneClassSVM using control cells and estimate how much each perturbed cell deviates from the controls.

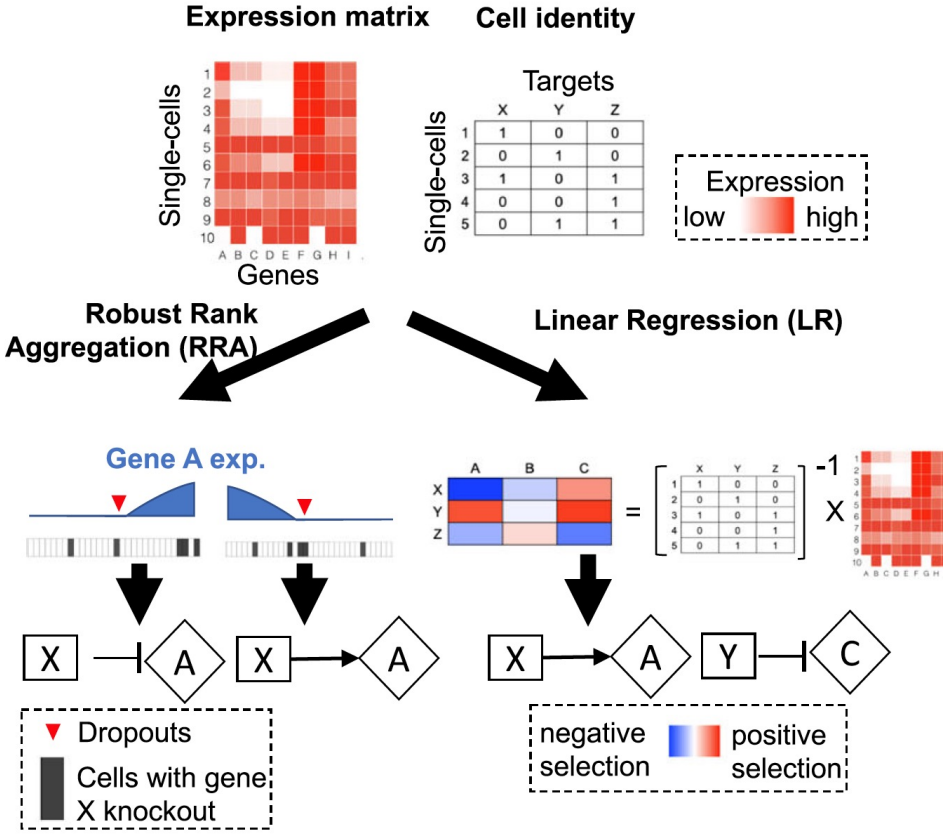
Adamson *et al.* 2016

Using linear regression to characterize the effects of each perturbation on gene expression: MIMOSCA and scMAGeCK

MIMOSCA



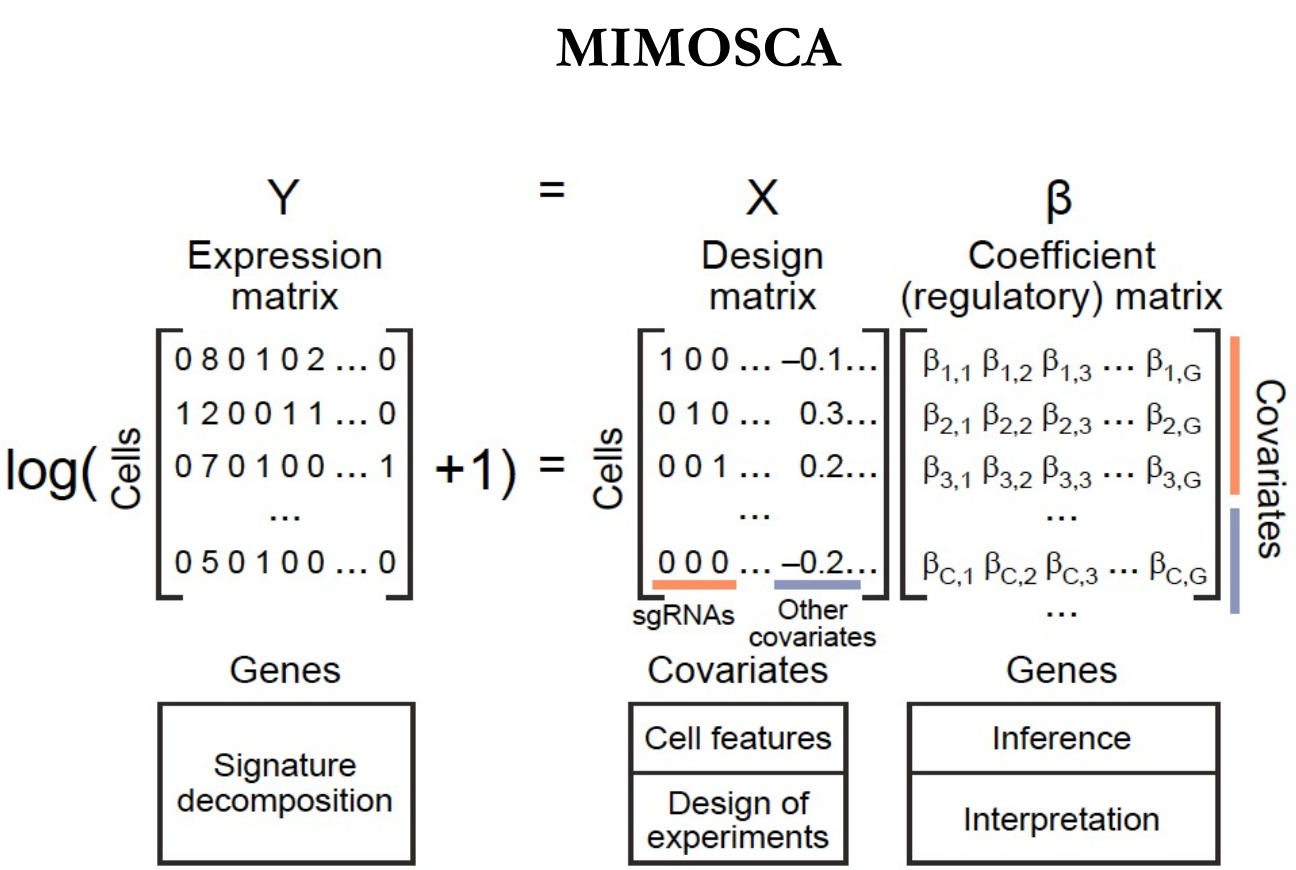
scMAGeCK



Dixit *et al.* 2016

Yang *et al.* 2020

Using linear regression to characterize the effects of each perturbation on gene expression: MIMOSCA and scMAGeCK

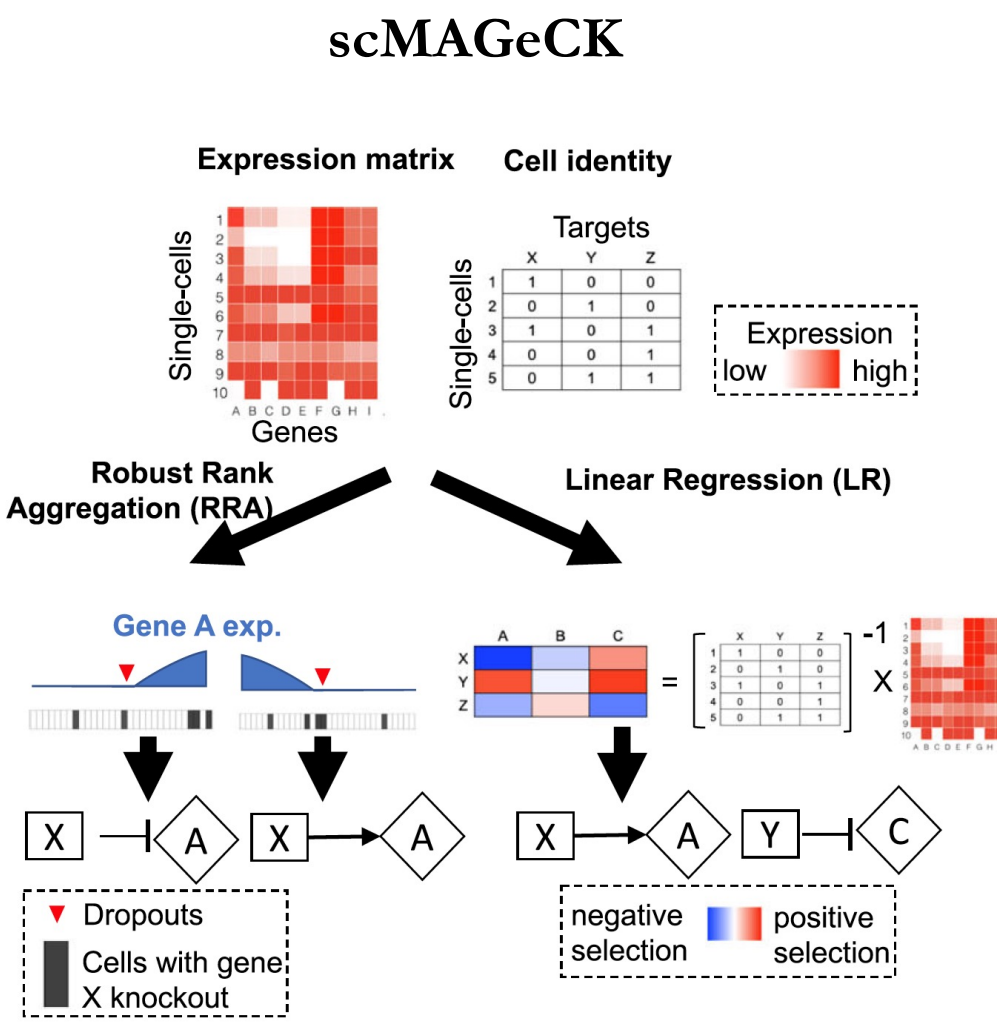


- Uses regularized linear regression to estimate the impact of perturbations on gene expression.
- Estimates the β parameters to identify which guides affect the expression of a gene.
- Can specify technical covariates (sequencing depth, cell state) in the model.
- After initial estimations can go back to identify cells that have 'escaped' perturbation and improve the mode fit.

Dixit *et al.* 2016

Using linear regression to characterize the effects of each perturbation on gene expression: MIMOSCA and scMAGeCK

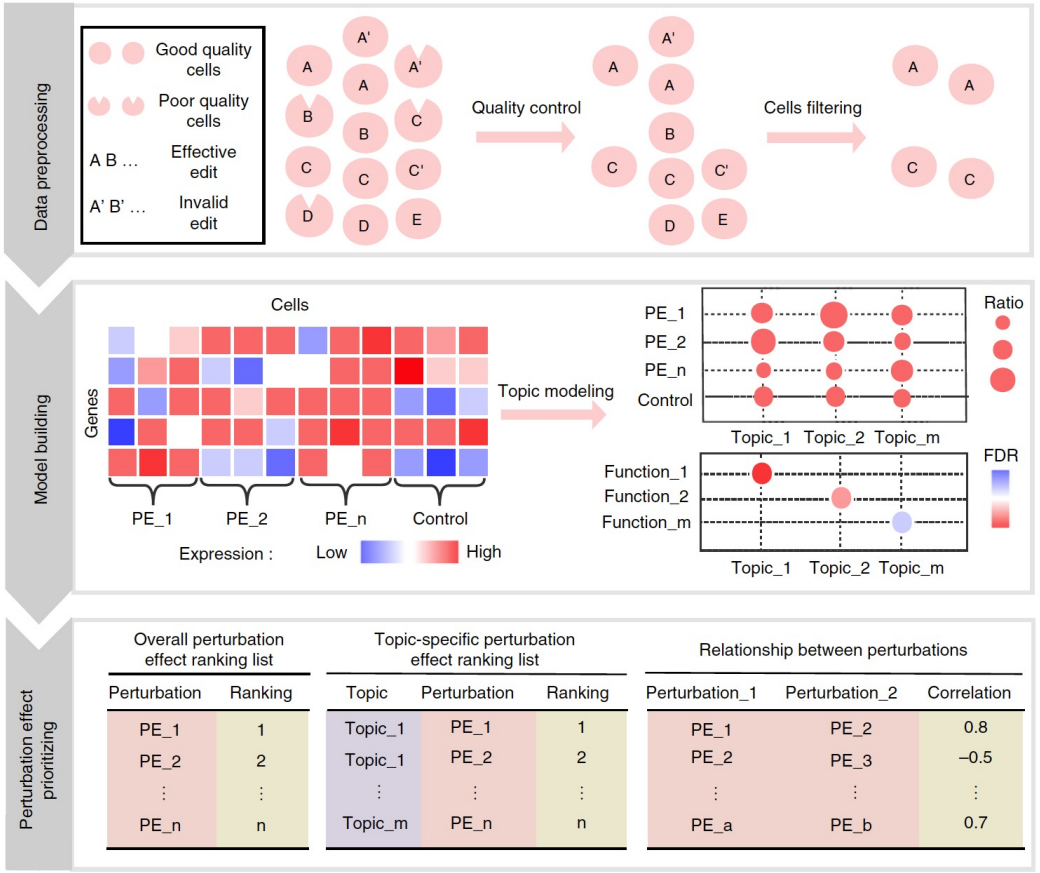
- It has in two modes: Linear regression (LR) and Robust Rank Aggregation (RRA).
- RRA: allows for detection of genes whose perturbation links to one single marker expression (virtual FACS approach).
- LR: uses linear regression to calculate selection scores for genes and correlate perturbation to shifts in gene expression.
- Does not account for technical variation and 'escaping' cells.



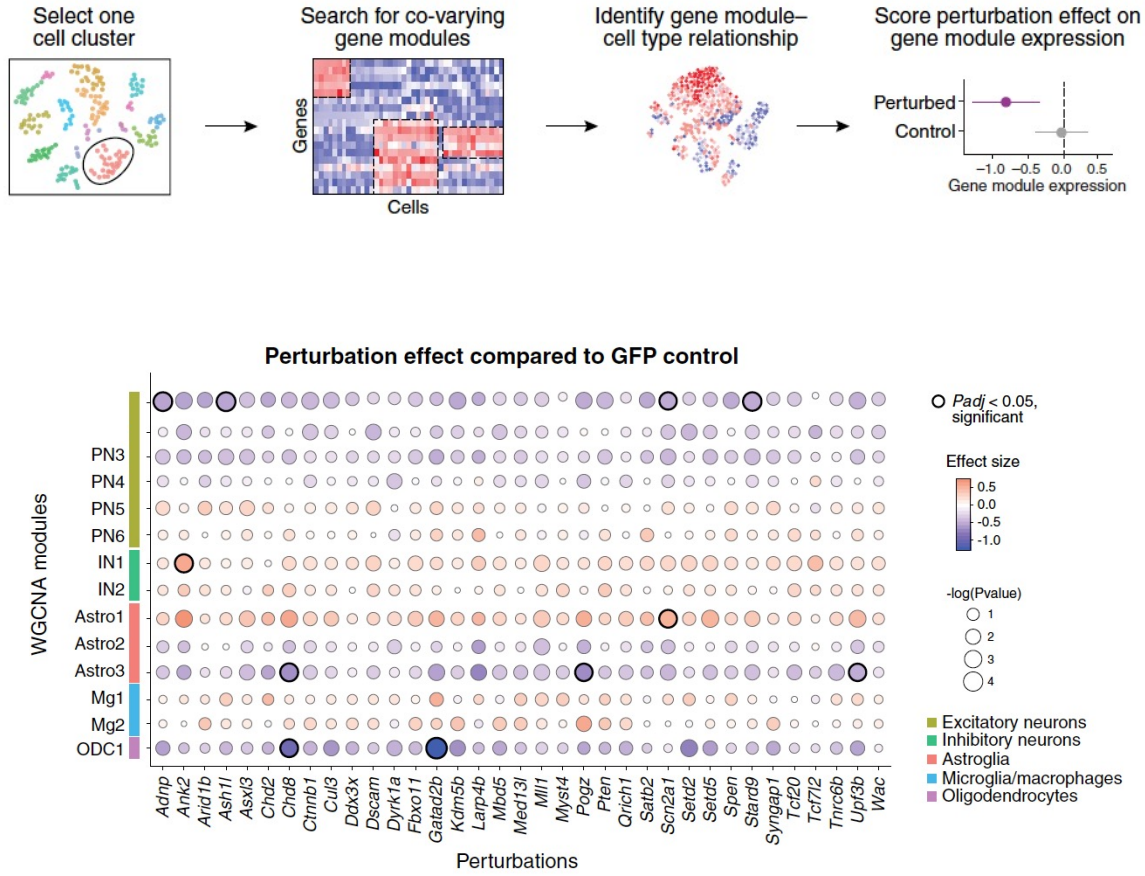
Yang *et al.* 2020

Using topic modeling to identify gene modules that correlate with each perturbation: MUSIC and WGCNA/STM

MUSIC



WGCNA/STM

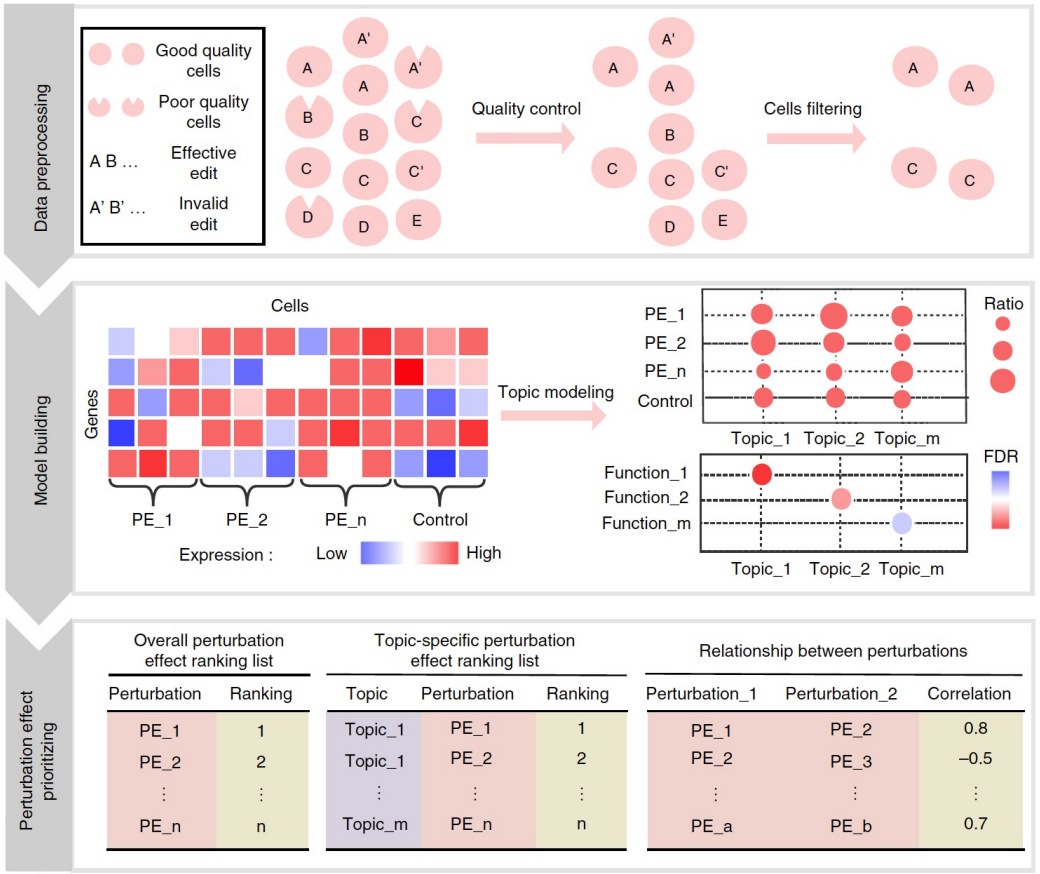


Duan *et al.* 2019

Jin *et al.* 2020

Using topic modeling to identify gene modules that correlate with each perturbation: MUSIC and WGCNA/STM

MUSIC



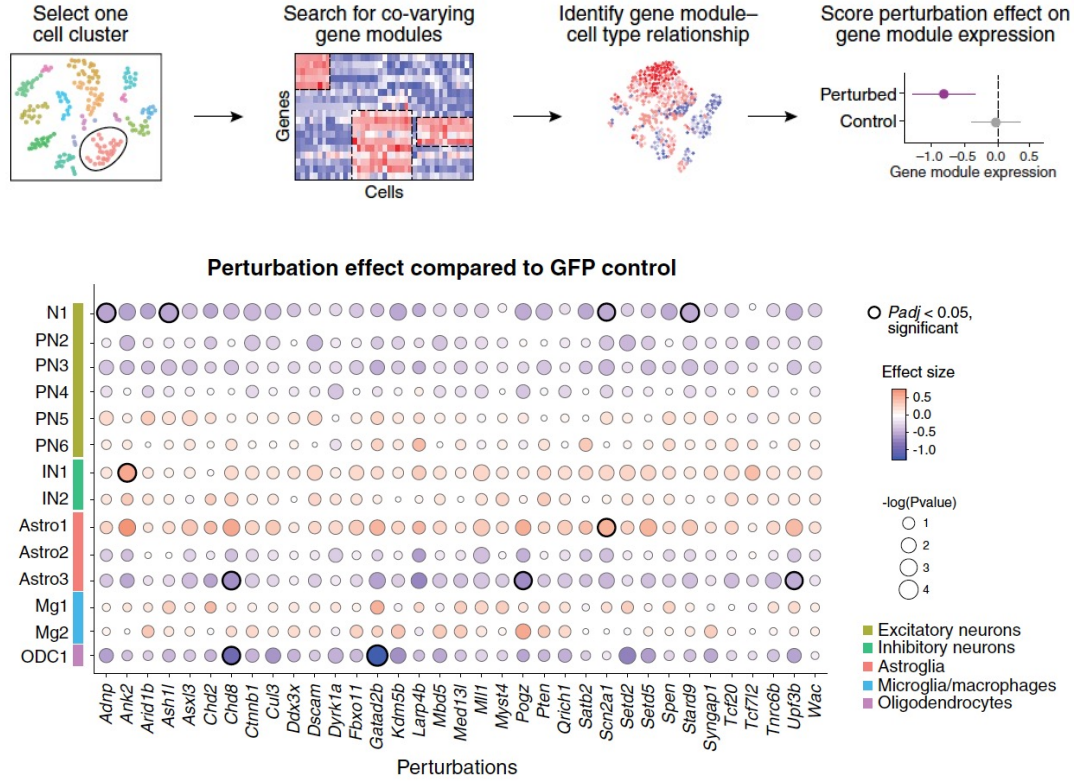
- Uses SAVER to impute missing signal and improve scRNA-seq data quality.
- Filters 'escaping' cells by looking at cells sharing the same perturbation and asking how similar they are to control cells.
- Finds topics representing specific biological functions associated with a group of highly differentially expressed genes.
- Prioritizes the perturbation effect by calculating a topic probability difference between control and targeted cells.

Duan *et al.* 2019

Using topic modeling to identify gene modules that correlate with each perturbation: MUSIC and WGCNA/STM

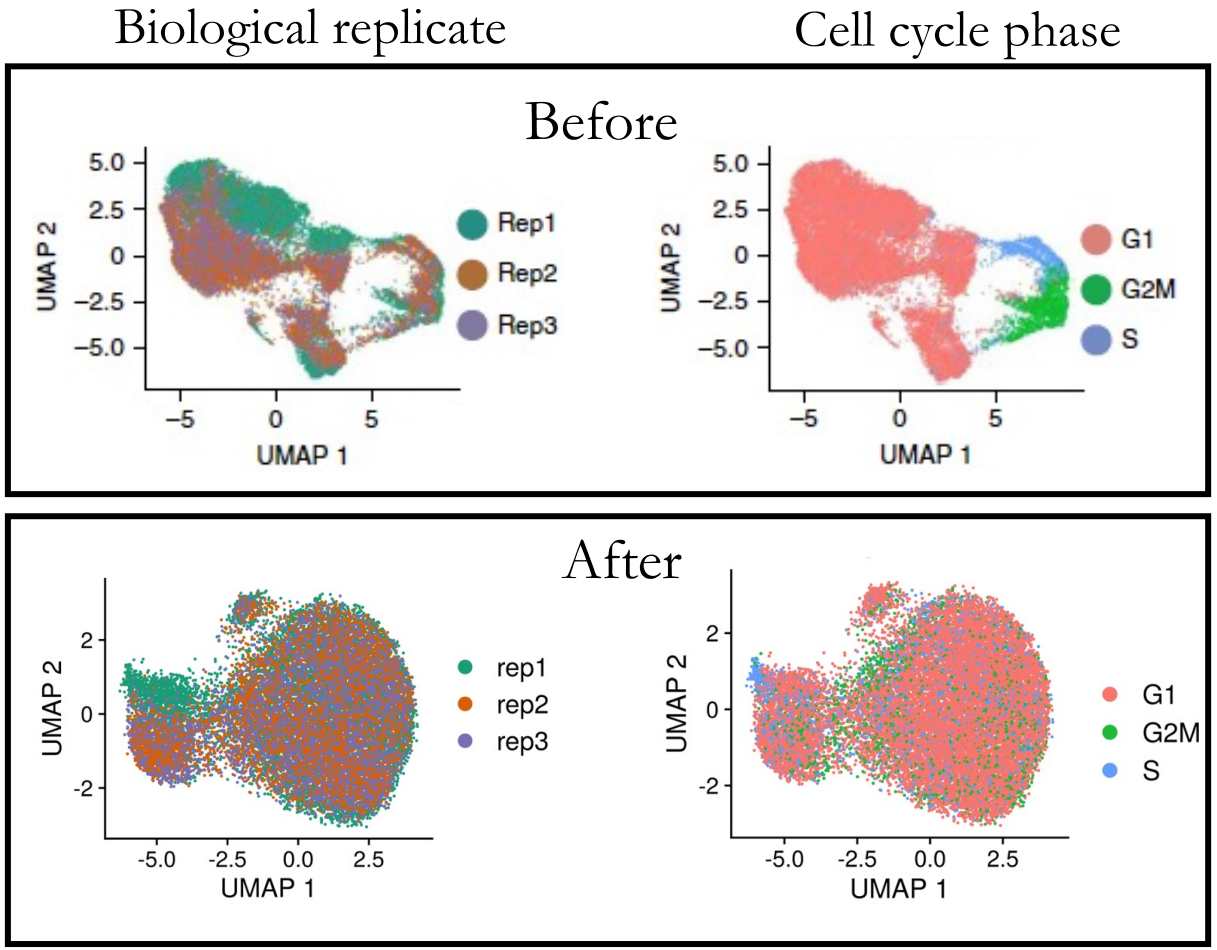
- Uses weighted gene correlation network analysis and structural topic modeling to identify modules/topics with correlated expression.
- Focuses on cell type-specific gene modules that highly correlated with one or more topics from STM.
- Calculates the effect size of each perturbation on each module using linear regression.
- Corrects for batch and sequencing depth using these as covariates in the the linear model.

WGCNA/STM



Jin *et al.* 2020

Calculating the perturbation signature of cells to remove confounding sources of variation in the dataset: Mixscape

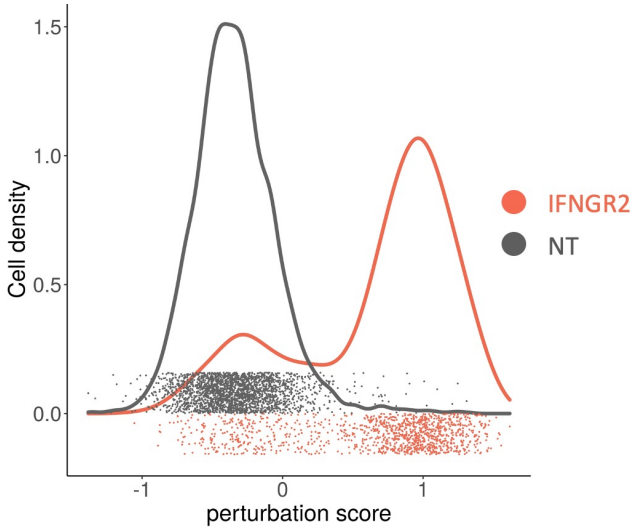


- Mixscape removes confounding sources of variation by calculating the perturbation signature of every cell.
- For every cell: finds 20 nearest non-targeting cell neighbors.
- Averages the expression of neighbors and subtracts averaged expression from cell.
- Uses perturbation signature as input to clustering and downstream analyses.

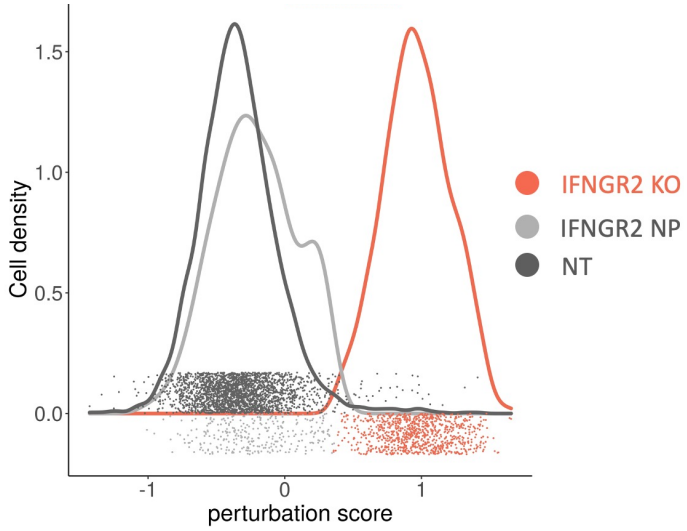
Papalexi *et al.* 2021

Using Gaussian mixture models to remove 'escaping' cells:

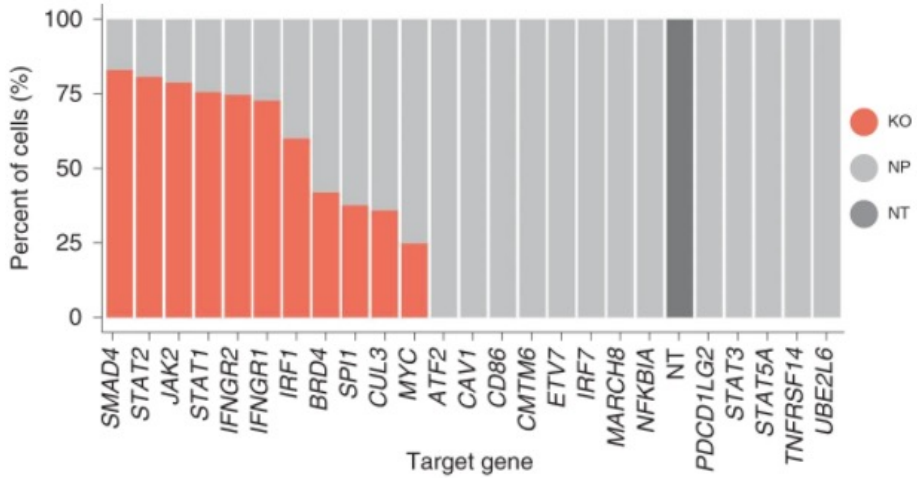
Original



Mixscape



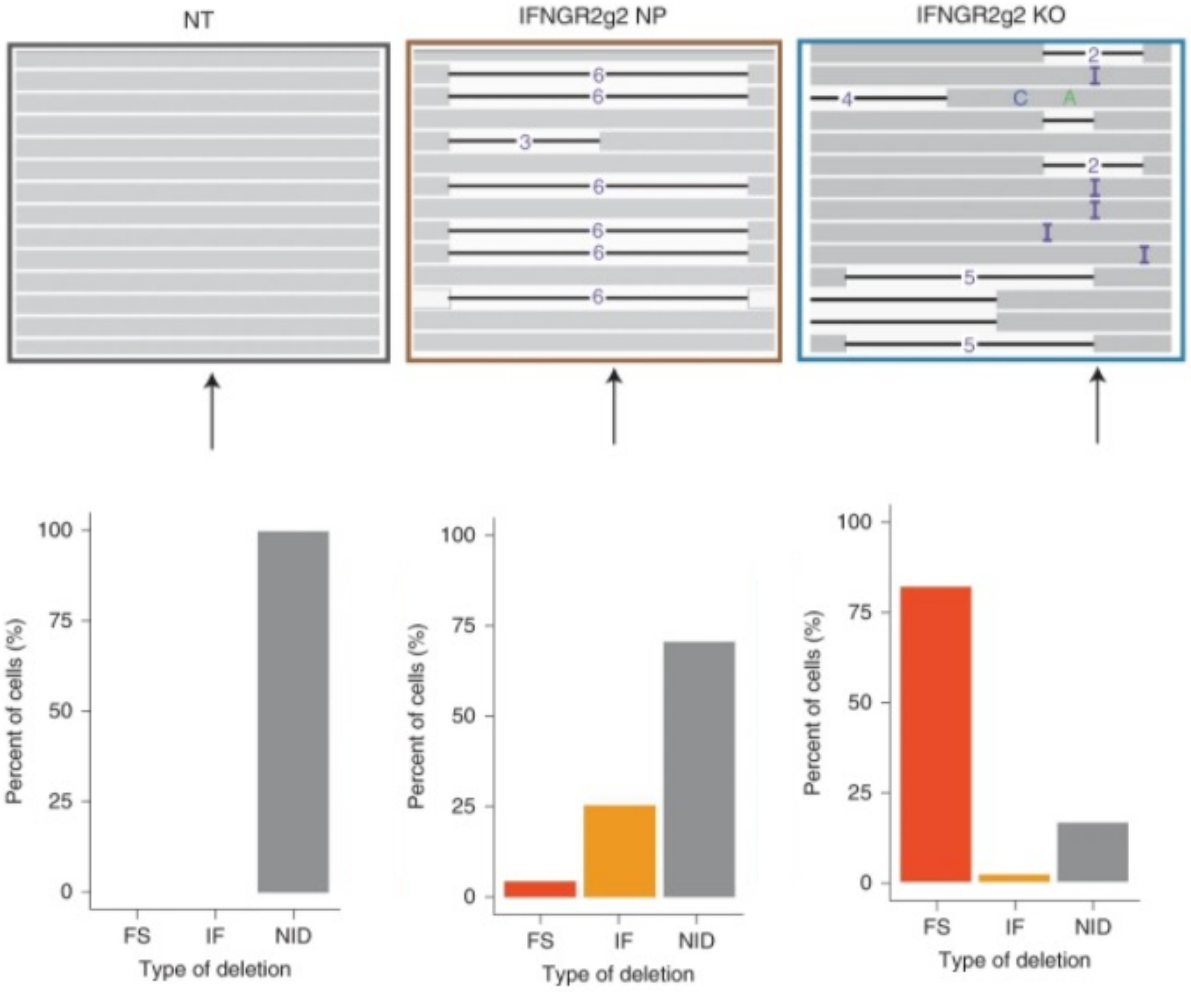
Quantifying % of KO cells



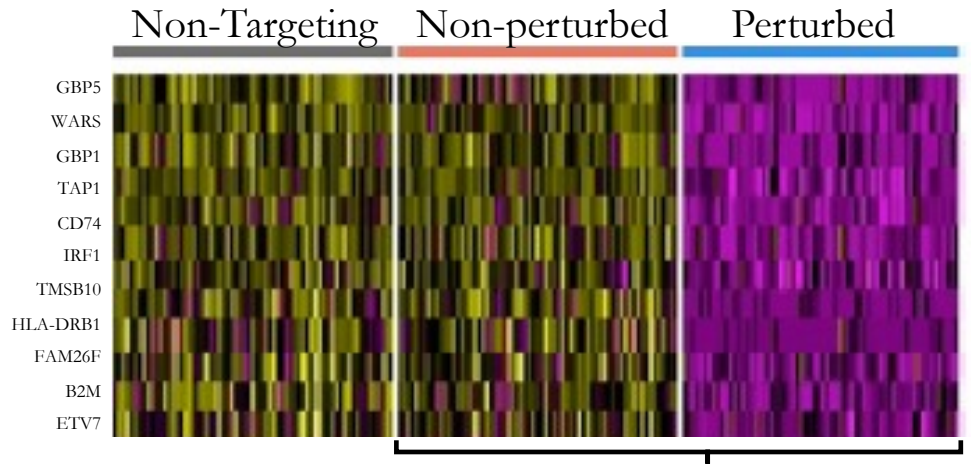
- Calculates the perturbation score of every cell to quantify how perturbed a cell is.
- Assumes that 'escaping' cells will be transcriptomically identical to non-targeting cells.
- Models each perturbation as a mixture of cellular responses, to identify and remove 'escaping' cells.

Papalexi *et al.* 2021

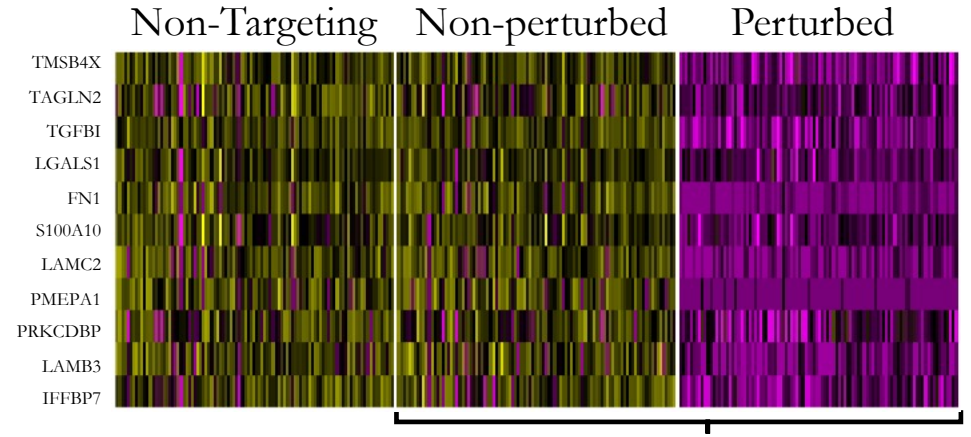
'Escaping' cells have in-frame mutations that do not result in phenotypic loss of function.



No indels In-frame Frameshift



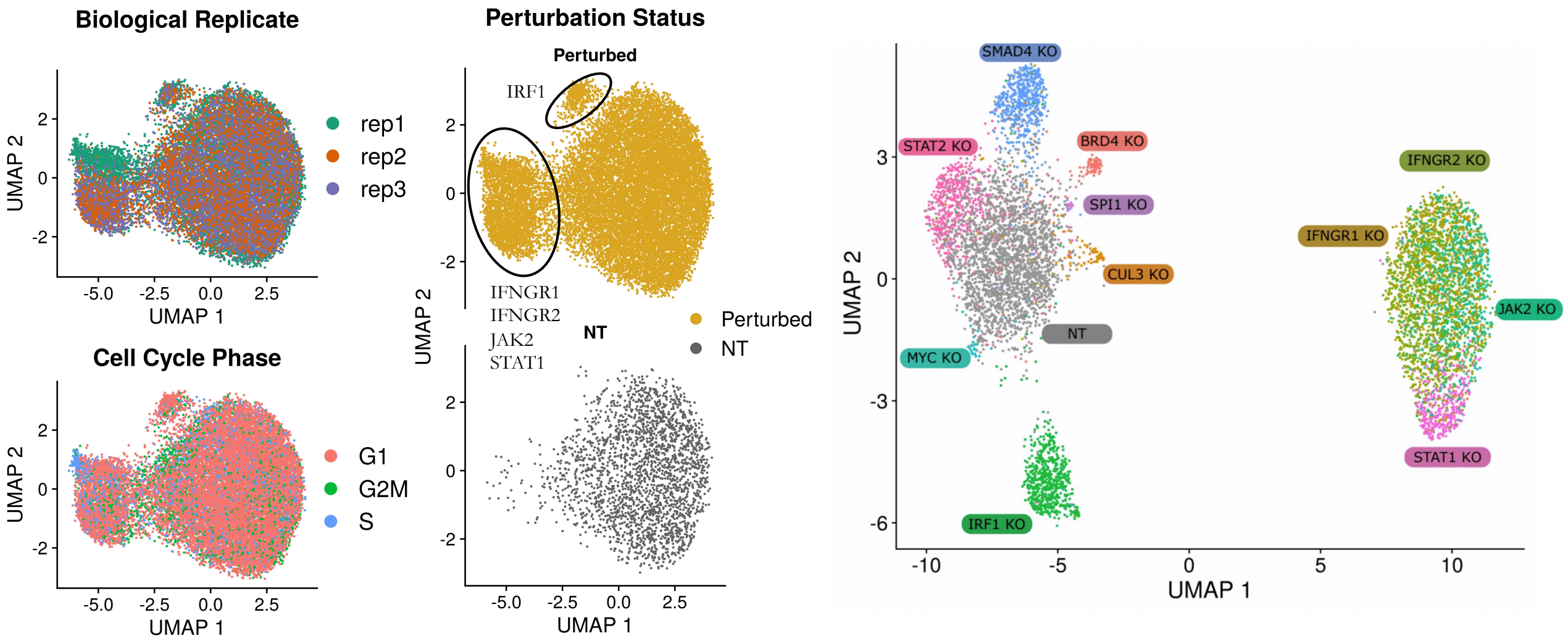
Papalexi *et al.* 2021



*All cells express the same gRNA

McFaline-Figueroa *et al.* 2019

Removing non-perturbed cells and using linear discriminant analysis helps to visualize separation across perturbations.



Papalexi *et al.* 2021

Remaining challenges:

- Can we quantify the effect of each perturbation and prioritize perturbations based on their strength?
- Can we extend current methods to characterize perturbation-specific changes to chromatin accessibility states and protein expression levels?
- How do we remove noise and 'escaping' cells in datasets with 2 or more perturbations per cell?
- Can we characterize more subtle shifts in gene expression ?

Available resources:

- **LRICA**: A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response

https://www.sciencedirect.com/science/article/pii/S0092867416316609

- **MIMOSCA**: a computational tool to estimate the impact of each perturbation of gene expression.

https://github.com/asncd/MIMOSCA

- **scMAGeCK**: links genotypes with multiple phenotypes in single-cell CRISPR screens.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1928-4

- **MUSIC**: Model-based understanding of single-cell CRISPR screening.

https://github.com/bm2-lab/MUSIC

https://www.nature.com/articles/s41467-019-10216-x

- **WGCNA/STM**: In-vivo Perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes.

https://science-sciencemag-org.proxy.library.nyu.edu/content/370/6520/eaaz6063

- **Mixscape**: Characterizing the molecular regulation of inhibitory immune checkpoints with mutli-modal single-cell screens.

https://www.nature.com/articles/s41588-021-00778-2

https://satijalab.org/seurat/articles/mixscape_vignette.html