Perturb-seq Introduction

Eduardo Gusmao and Zhirui Hu

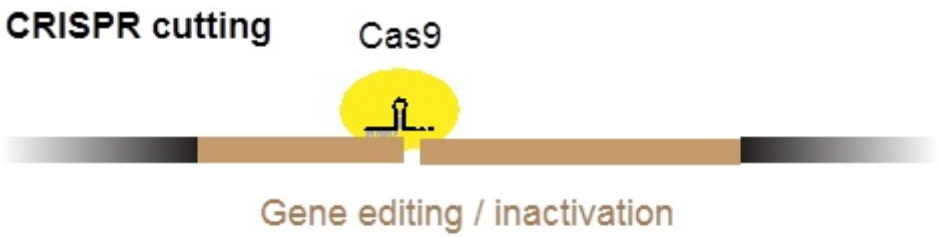
STAT316

2017-03-02

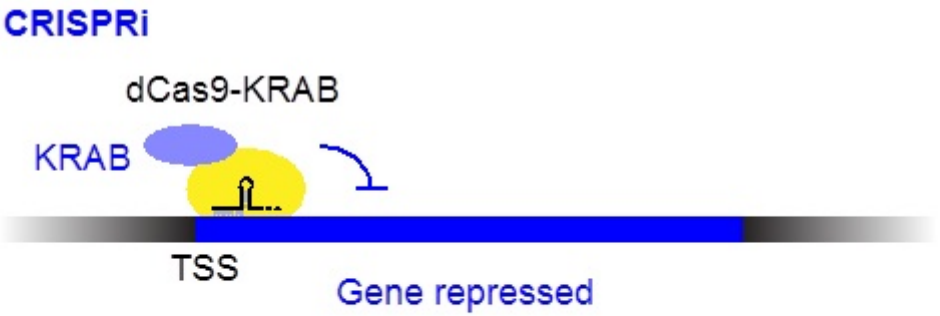
Perturb-seq Introduction

- Motivation (in Macosko EZ et al. (2015), Cell):
  - "Drop-seq could be used to provide initial insights into how genes function in the diverse cell types composing each tissue. **In addition, coupling Drop-seq with perturbations – such as small molecules, mutations, pathogens, or other stimuli – could generate an information-rich, multi-dimensional readout of the influence of perturbations on many kinds of cells**".
- Perturb-seq combines:
  - CRISPR screens (e.g. CRISPRi or CRISPR-KO).
  - Single-cell RNA-seq approaches (e.g. Drop-seq).
- Highly parallel platform for single-cell functional genomics.
  - Distinguish different perturbations that cause similar response.
  - Bulk phenotype is driven by a subpopulation.
- Robust cell barcoding strategy that encodes the identity of the CRISPR-mediated perturbation in an expressed transcript, captured during scRNA-seq.

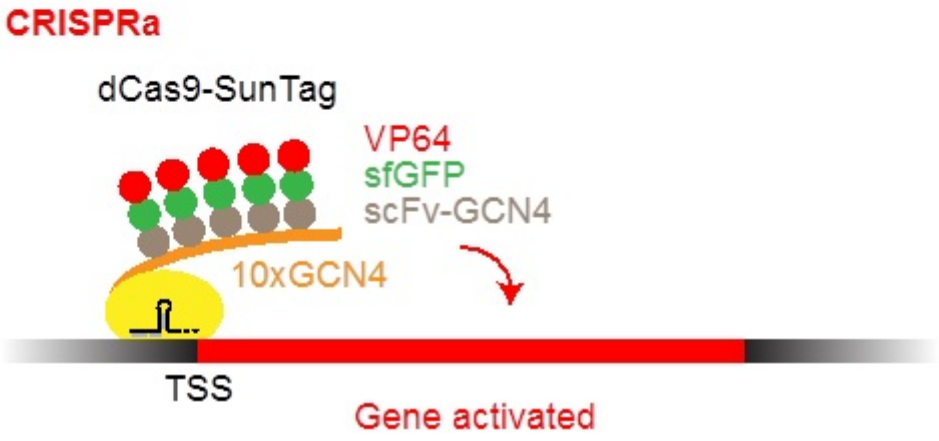
CRISPRi / CRISPRa



- CRISPR-KO
  - Functional Cas9 protein perform cuts in the DNA.
  - Cuts KO gene.

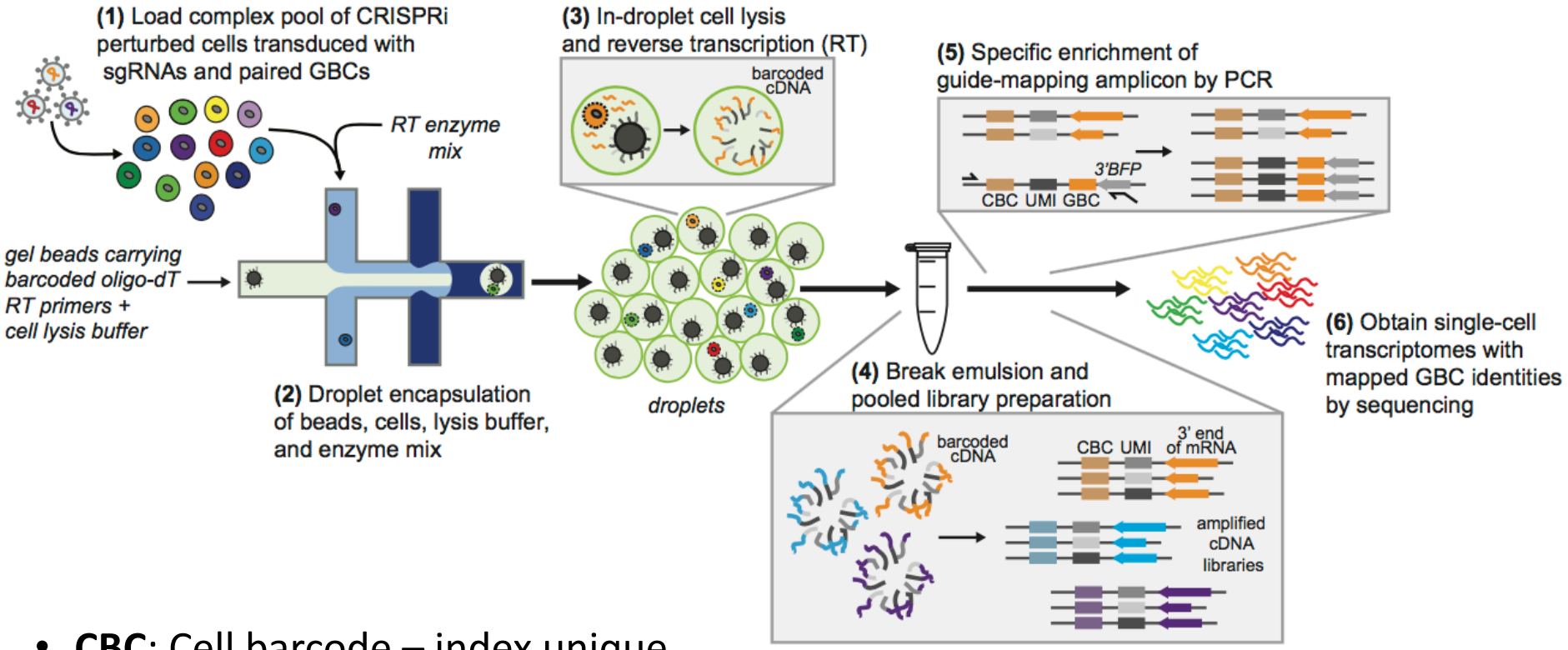


- CRISPR Interference (CRISPRi)
  - Catalytically dead Cas9 (dCas9) recruits Krüppel-Associated Box (KRAB) to gene.
  - KRAB represses gene expression.



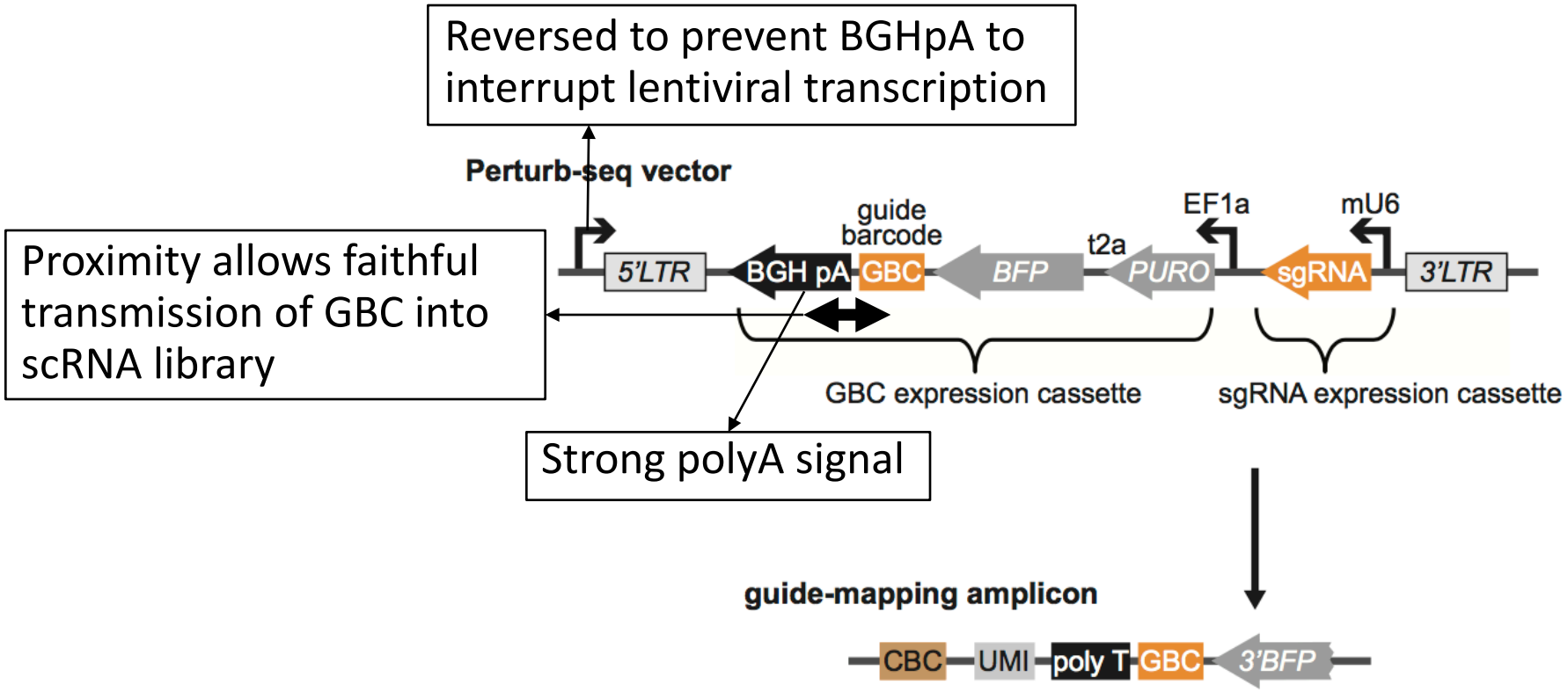
- CRISPR Activation (CRISPRa)
  - dCas9 fused to SunTag (epitope tag) containing 10 copies of GCN4.
  - GCN4 recruits VP64 transcriptional activators.

Perturb-seq Protocol



- **CBC**: Cell barcode – index unique to each bead.
- **UMI**: Unique molecular identifier – index unique to each bead oligo.
- **GBC**: Guide barcode – index unique to each sgRNA.

Perturb-seq Vector



- **CBC**: Cell barcode – index unique to each bead.
- **UMI**: Unique molecular identifier – index unique to each bead oligo.
- **GBC**: Guide barcode – index unique to each sgRNA.

Perturb-seq Outputs

- Single-cell expression matrix in MatrixMarket format.

```
%%MatrixMarket
%
32738 65337 237812947
32709 1 59
32707 1 7
32706 1 71
32704 1 1
32703 1 56
32702 1 28
32700 1 29
…
```

Cell (CBC)

Gene

Expression (UMI)

- Mapping between CBCs and GBCs

```
32738 IARS2_pDS090
32709 SPCS3_pDS402
32707 CHERP_pDS024
32706 SLMO2_pDS433
32704 XRN1_pDS411
32703 EIF2B2_pDS463
32702 SYVN1_pDS442
32700 63(mod)_pBA580
32738 MANF_pDS027
32709 SRP68_pDS40
…
```

Cell (CBC)

Perturbation (GBC)

Perturb-seq Caveats

- Within a 50,000 single-cell library:
  - 10 minutes of machine time.
  - 1 day of library preparation.
  - 2 days of sequencing.
  - 6 months of data analysis!
- Bioinformatics & statistical analyses become the bottleneck.
- How to know if the cell really was properly Perturbed?
  - Not every cell with a knockout has a phenotype.
  - Not every cell with a phenotype is a knockout.

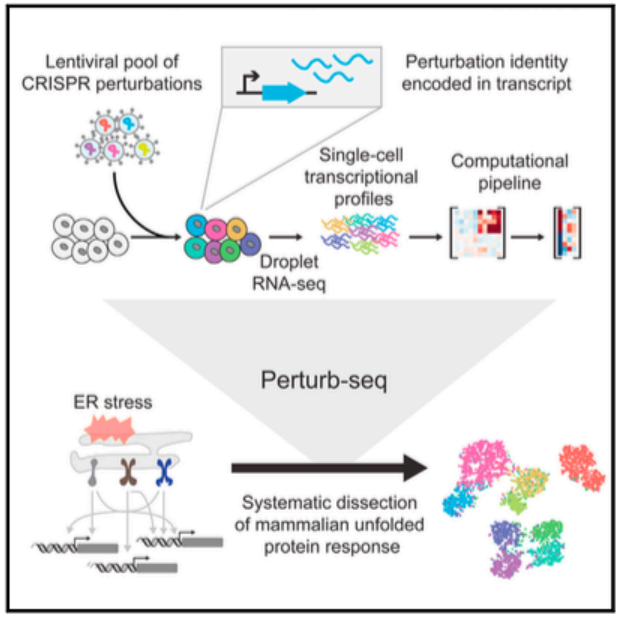
Adamson et al. Paper Discussion

Resource

Cell

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

Graphical Abstract



Authors

Britt Adamson, Thomas M. Norman, Marco Jost, ..., Oren Parnas, Aviv Regev, Jonathan S. Weissman

Correspondence

jonathan.weissman@ucsf.edu

In Brief

A strategy for barcoding CRISPR-mediated perturbations allows pooled expression profiling via single-cell RNA sequencing. Application to the mammalian unfolded protein response then enabled systematic delineation of the transcriptional arms of the response and functional clustering of genes affecting ER homeostasis.

Adamson et al., 2016, Cell *167*, 1867–1882
December 15, 2016
http://dx.doi.org/10.1016/j.cell.2016.11.048

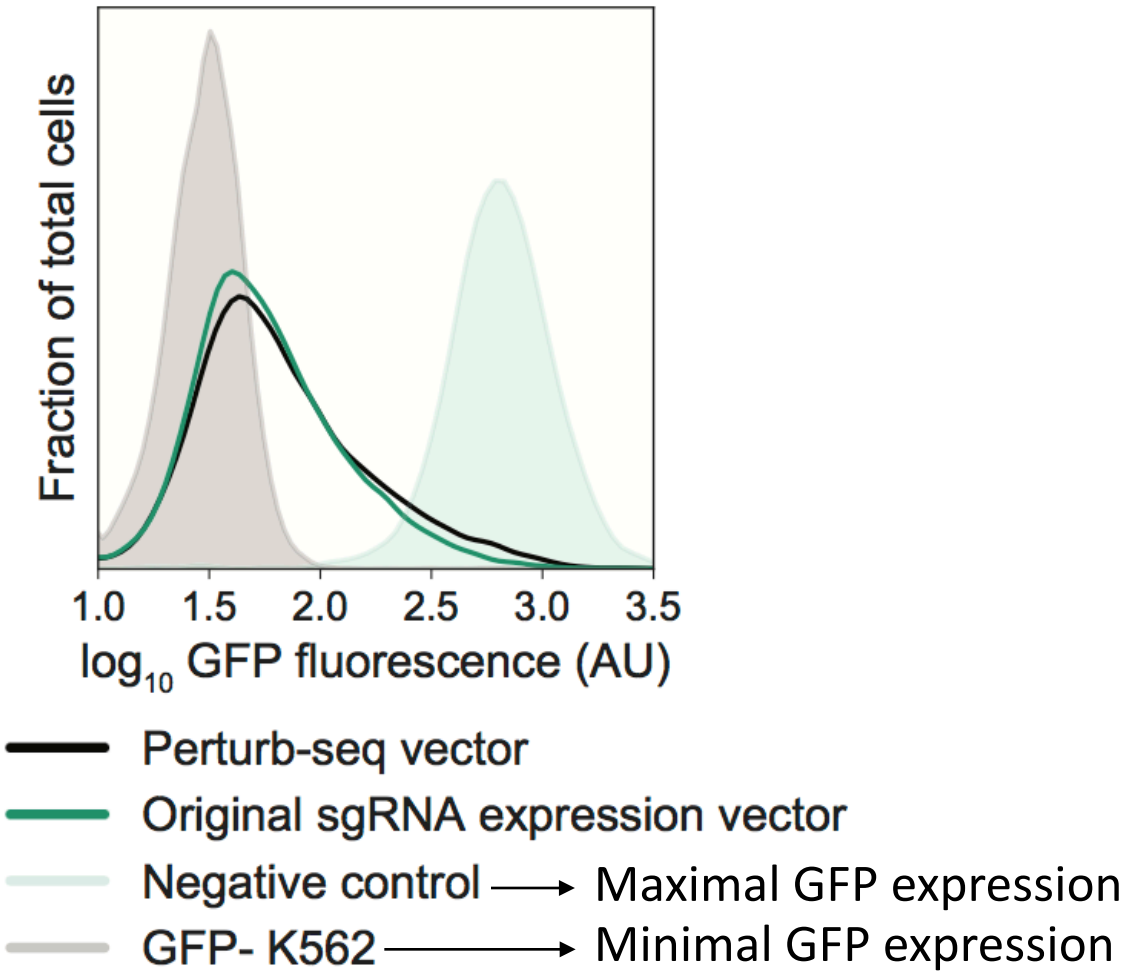
Eduardo Gusmao and Zhirui Hu

STAT316

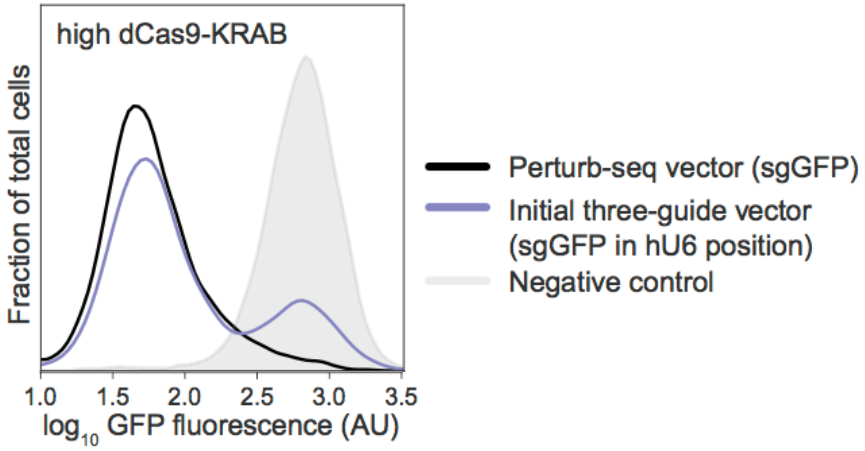
2017-03-02

Initial Perturb-seq Vector

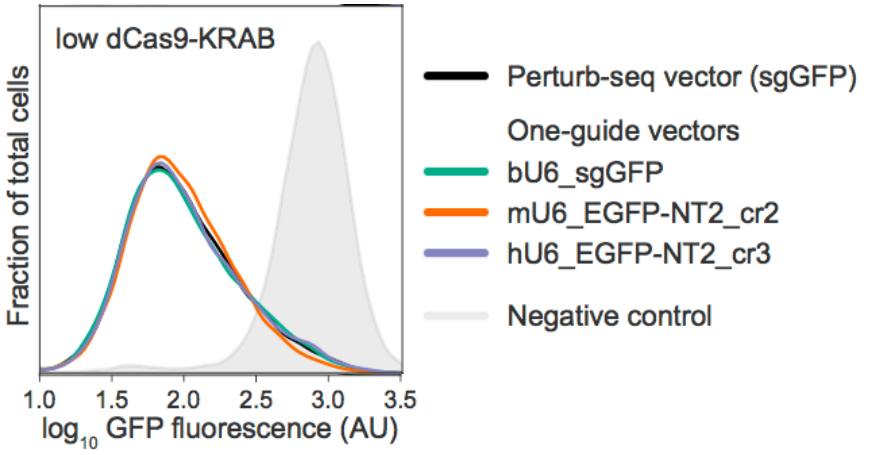
- Initial Perturb-seq vector successfully recapitulates original sgRNA vector.

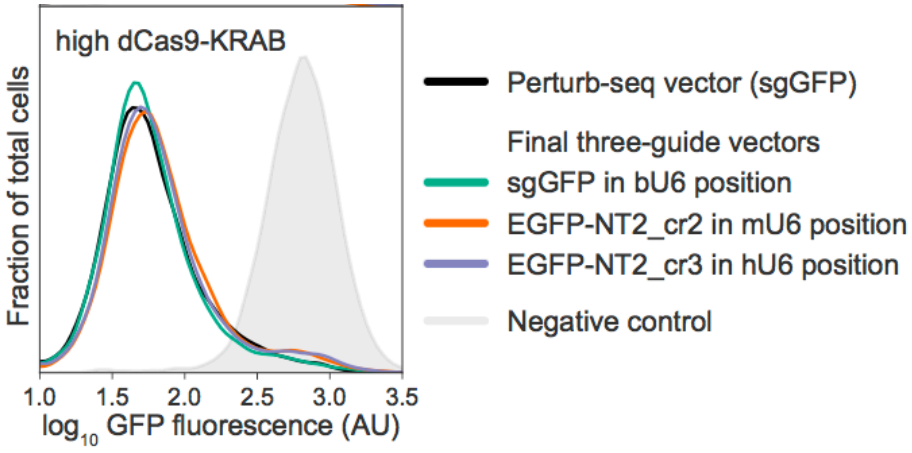


Three-Guide Perturb-seq Vector

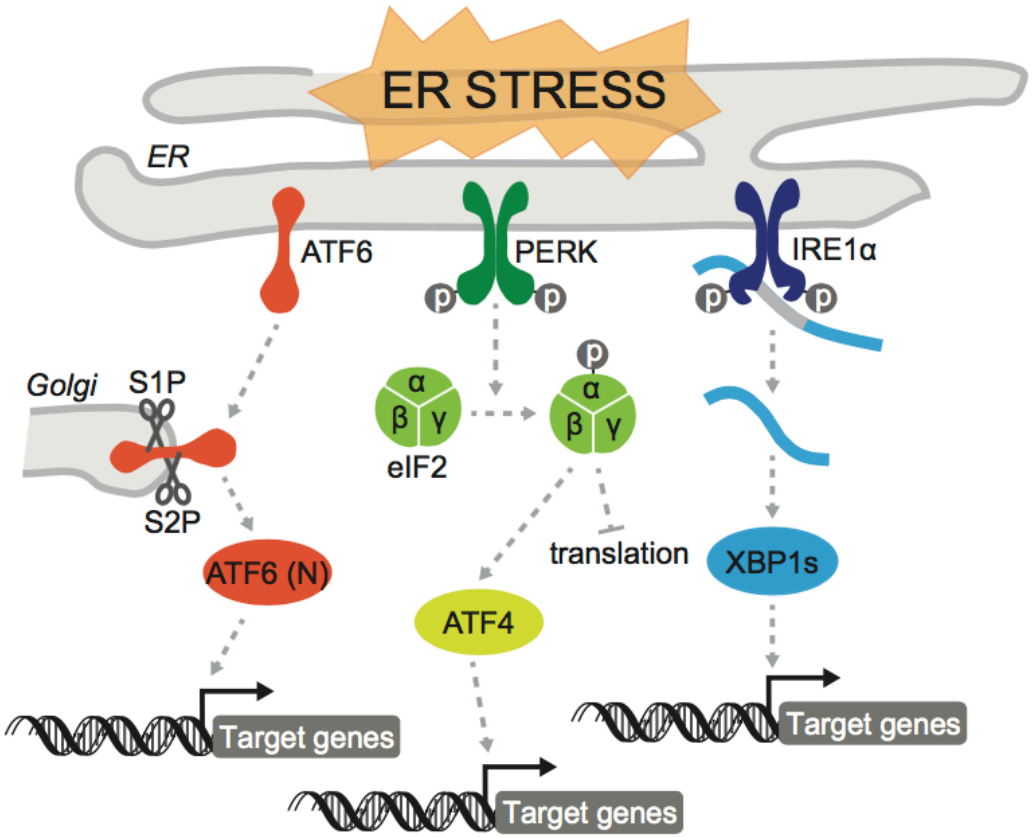


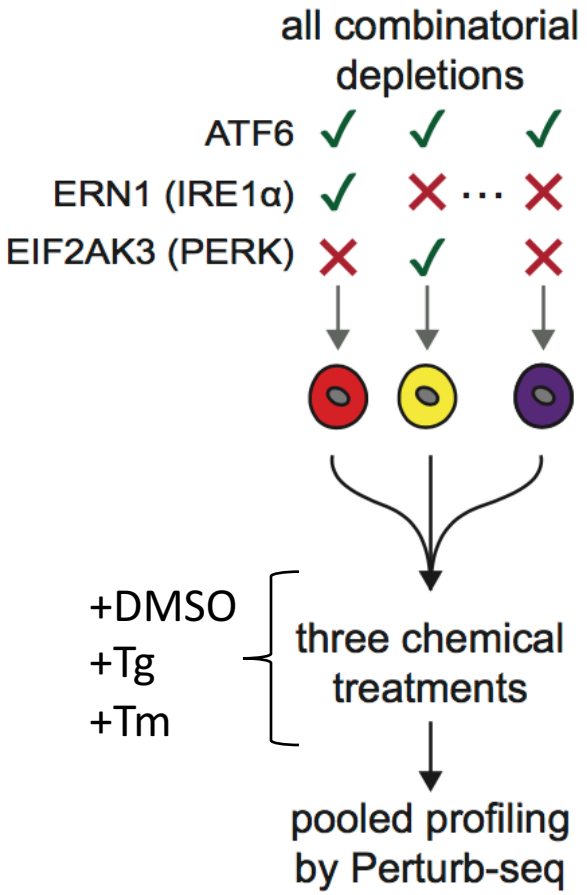
Final three-guide Perturb-seq vector





Perturb-seq to Study the UPR Epistasis





Data Analysis

- Sequencing
  - 10X Framework.
- Perturbation-identity mapping
  - Paired-end alignment of amplicon (CBC-UMI --- primer+GBC) to library of expected GBC sequences.
  - Criteria to associate CBC to GBC:
    1. Be in the upper mode of coverage distribution.
    2. Attested to by at least 50 raw reads.
    3. Attested to by at least 3 different UMIs.
- Expression normalization
  1. Scale all cells to have the same median number of total UMIs (each row of the matrix is standardized to the same sum).
  2. z-normalized with respect to control: $x_{norm} = \frac{x - \mu_{control}}{\sigma_{control}}$.
  3. Normalization with respect to control cells within same lane.
- Low cell count & inviable cell removal.

Data Analysis

- Low-Rank Independent Component Analysis (LRICA):
  1. **Sparse matrix decomposition:**
     - Remove noise and sparsity.
     - Robust PCA to solve the optimization problem:

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 \text{ subject to } X = L + S$$

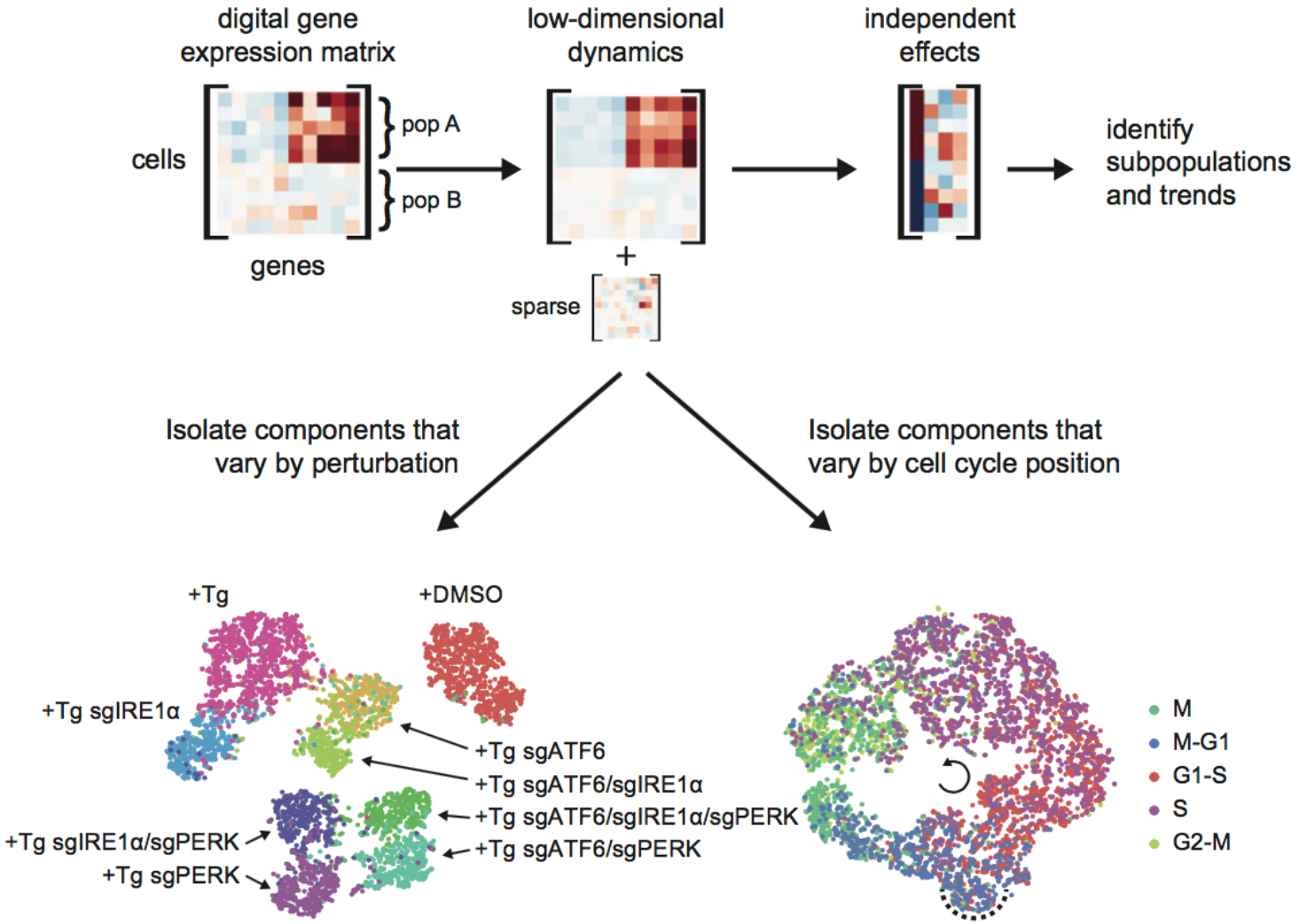
  2. **Independent Component Analysis (ICA):**
     - Isolate major trends within L.
     - Expression of a given gene (y_i) can be decomposed as a linear sum of various effects (s_1 to s_n), statistically independent from each other:

$$y_j = a_j s_1 + a_{j2} s_2 + \cdots + a_{jn} s_n$$

     - Cell's expression profile $\boldsymbol{y}$ over all genes: $\boldsymbol{y} = A\boldsymbol{s}$ $\quad Y = AS$
     - "Mixing matrix" $\boldsymbol{A}$ describes which genes correspond to which effects.
     - Process reversible: $\boldsymbol{s} = W\boldsymbol{y}$

Data Analysis

- Low-Rank Independent Component Analysis (LRICA):



Data Analysis

- Identification of differentially expressed genes:
  - **Kolmogorov-Smirnov test/metric.**
  - If $F_{perturbed}$ and $F_{control}$ are the CDFs for a given gene in the perturbed and control distribution, the test statistic is:

$$D = \sup_x \left| F_{\text{perturbed}}(x) - F_{\text{control}}(x) \right|$$

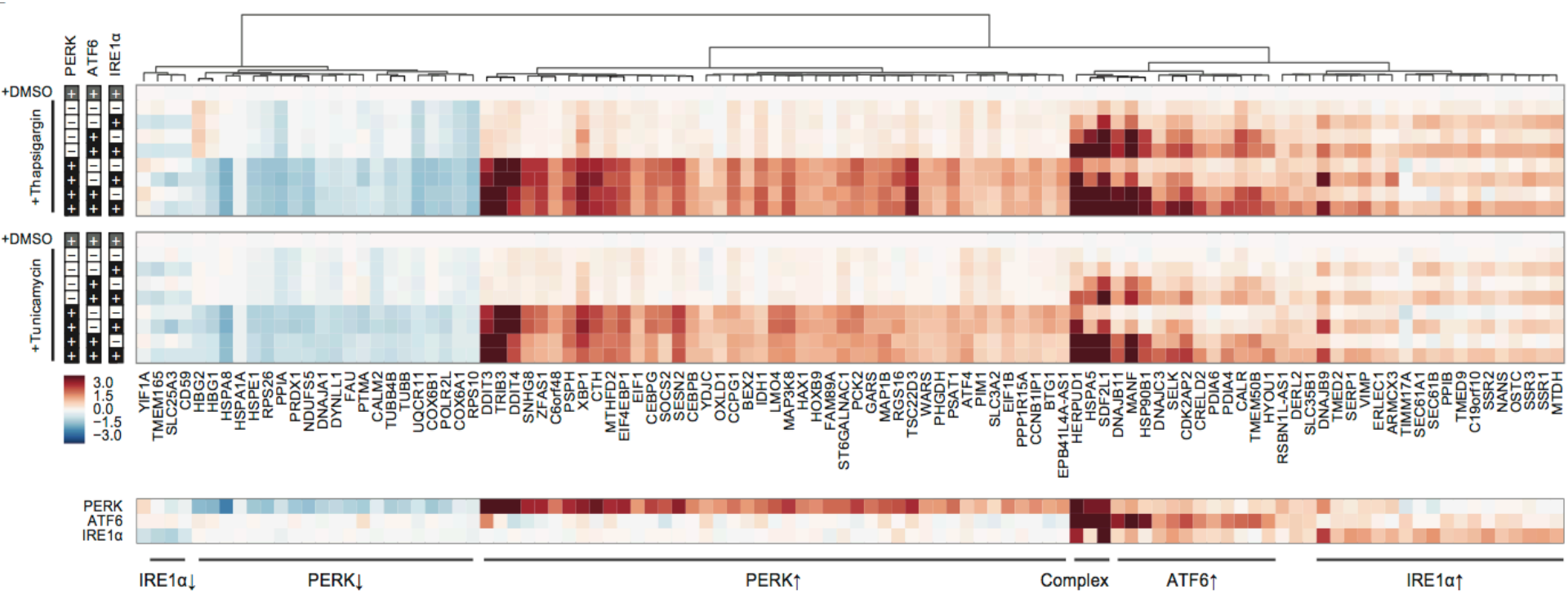
  - p-values can be calculated standardly.
  - Given the large scale of single-cell data, many genes are significant without being actually interestingly perturbed.
  - Solution: In some cases, put a direct threshold on the test statistic D itself.

Data Analysis

- Identification of differentially expressed genes:
  - **Random Forest Classifier**
  - Explore advantage of Perturb-seq: supervised learning can be used since cell populations are known.
  - Rationale: A gene is likely important if it's expression level can be used to accurately predict a perturbation identity.
  - Train model in 80% of cells.
  - 1000 trees implementation from Scikit-learn python package.
  - Random forest classifier assigns importance to features during training based on predictive value.
  - Take top N genes based on that importance and retrain a model with the remaining 20% cells.
  - Predict perturbation using the new model.
  - Accuracy for each gene in predicting a perturbation dictates its importance.

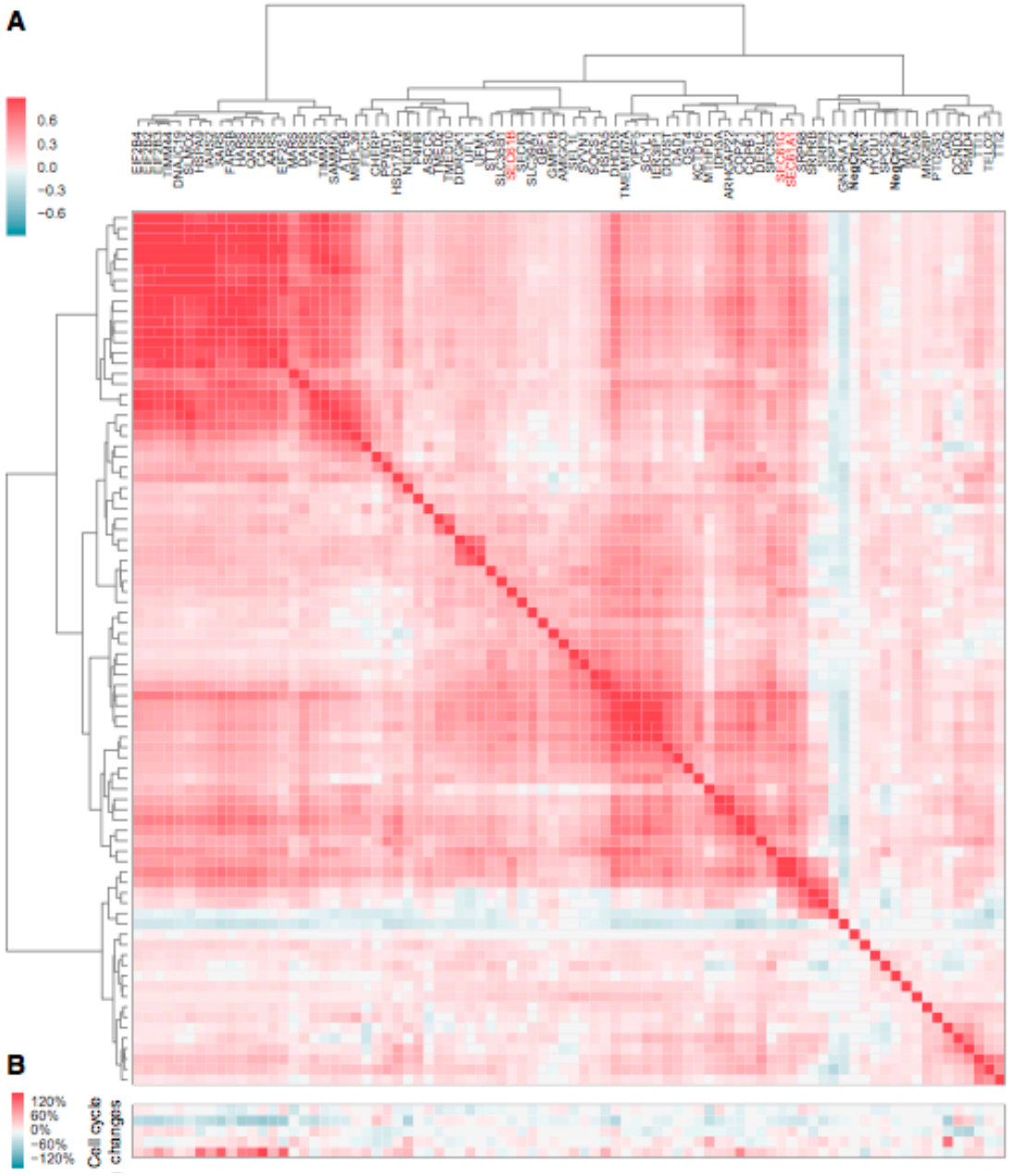
UPR Results

- Average expression of 104 DE genes (RF approach).
- Patters determine branch specificity of each gene.



Decomposition of the total response into 3 components (LRICA).

Single-Cell can Provide some Insights



aminoacyl tRNA synthetases

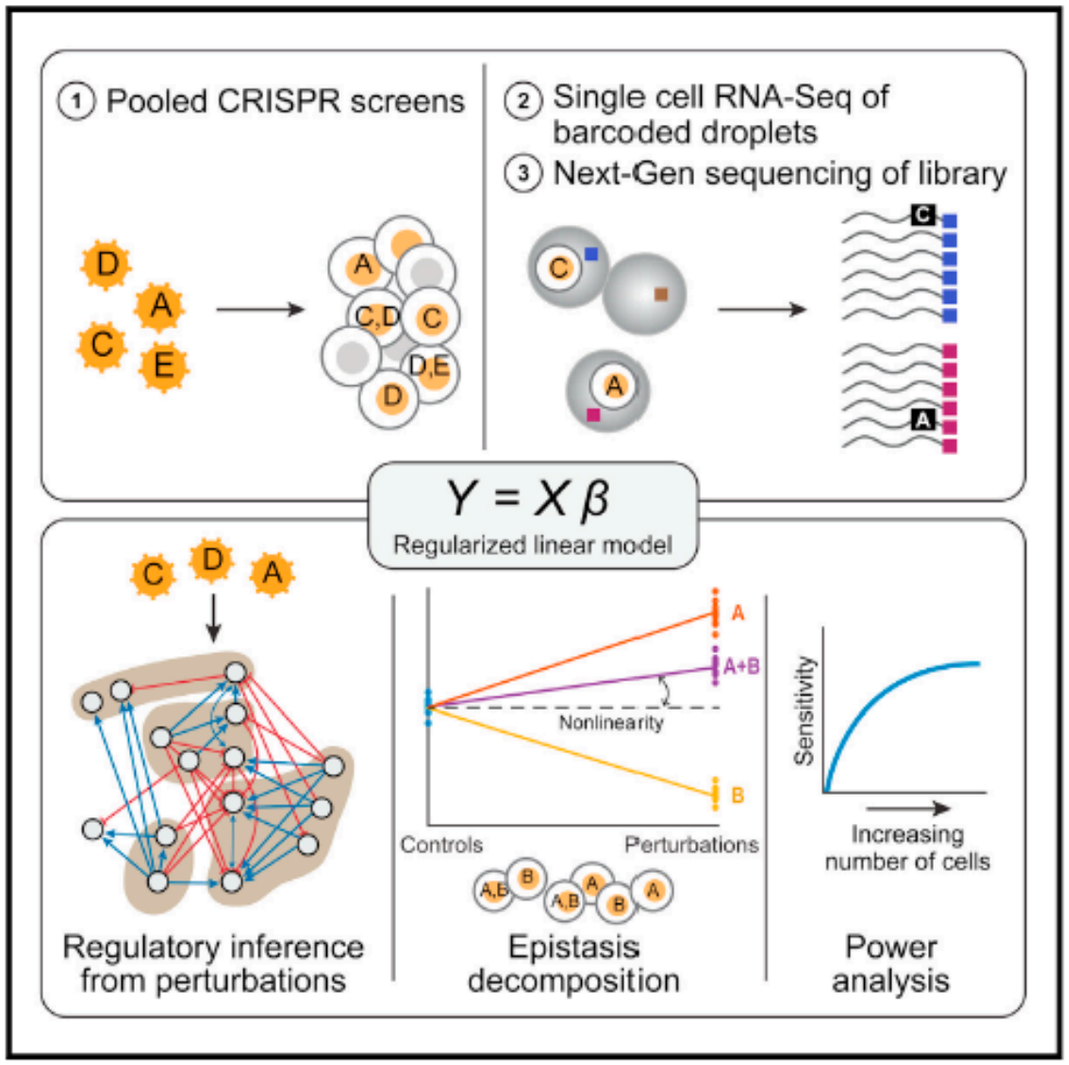
Decomposing the populations by cell-cycle position revealed that perturbation of many **aminoacyl tRNA synthetases** elicited accumulation of cells in **G2**

G2

Cell

Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens

Graphical Abstract



Authors

Atray Dixit, Oren Parnas, Biyu Li, ..., Jonathan S. Weissman, Nir Friedman, Aviv Regev

Correspondence

aregev@broadinstitute.org

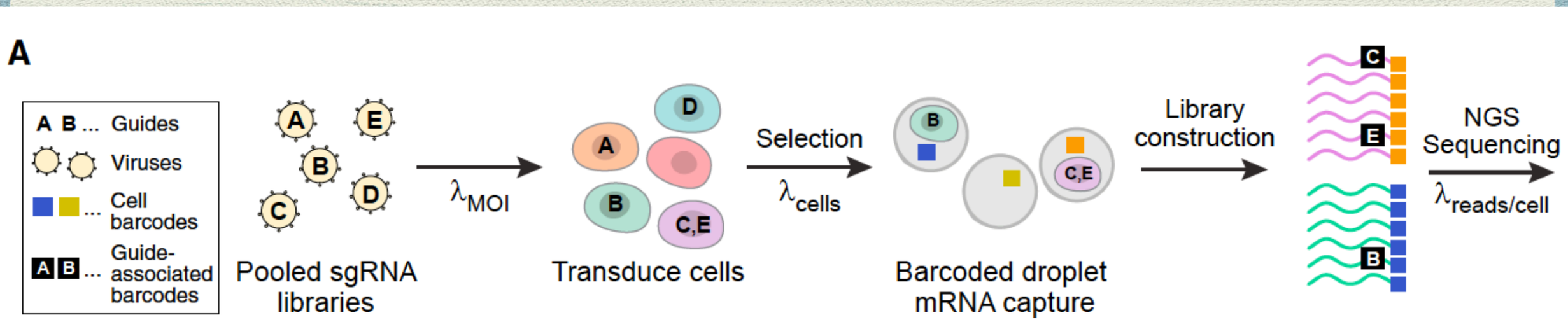
In Brief

A technology combining single-cell RNA sequencing with CRISPR-based perturbations termed Perturb-seq makes analyzing complex phenotypes at a large scale possible

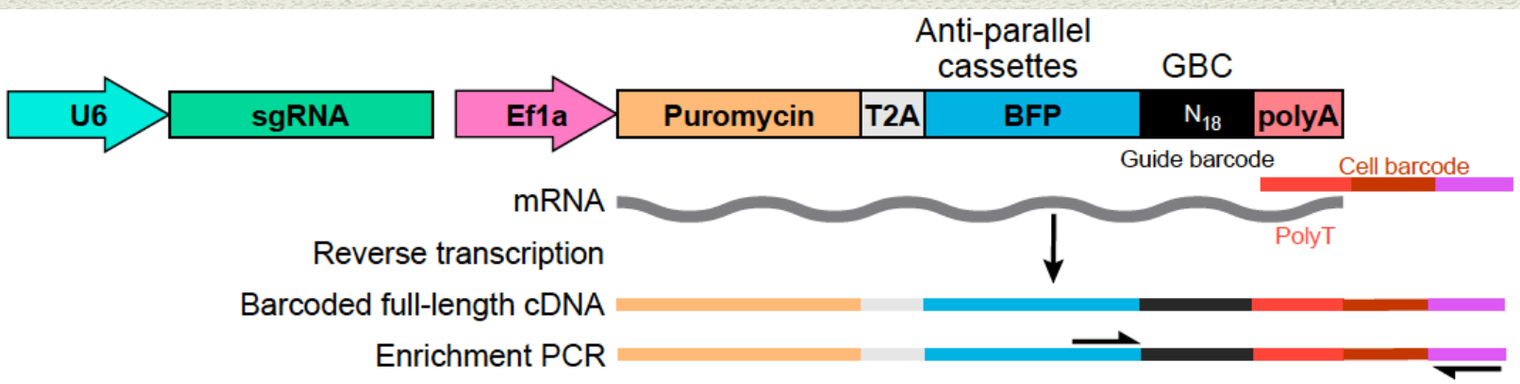
presented by Zhirui Hu

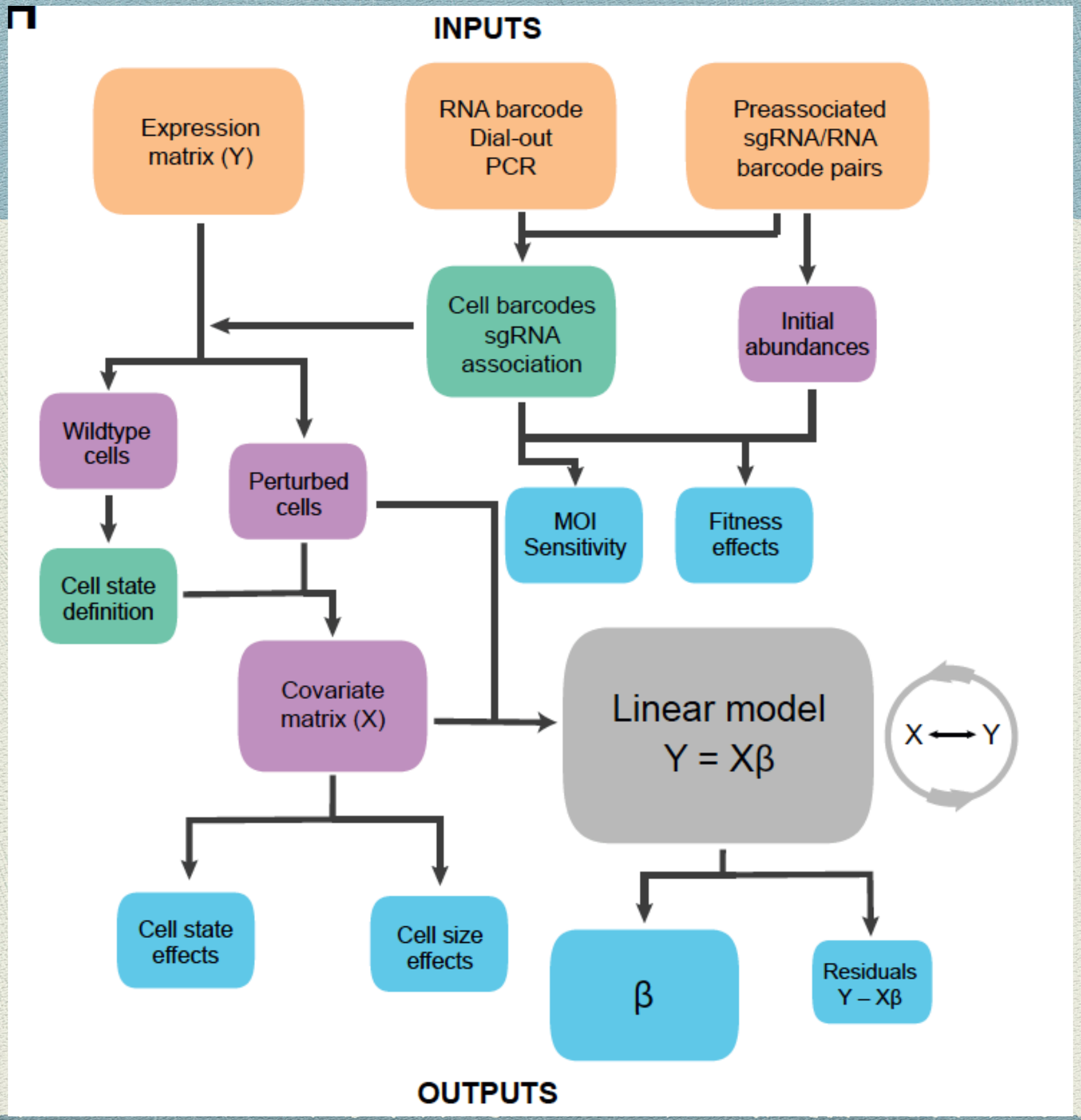
Overview

- Study single gene or epistatic effects by tuning multiplicity of infection (MOI).

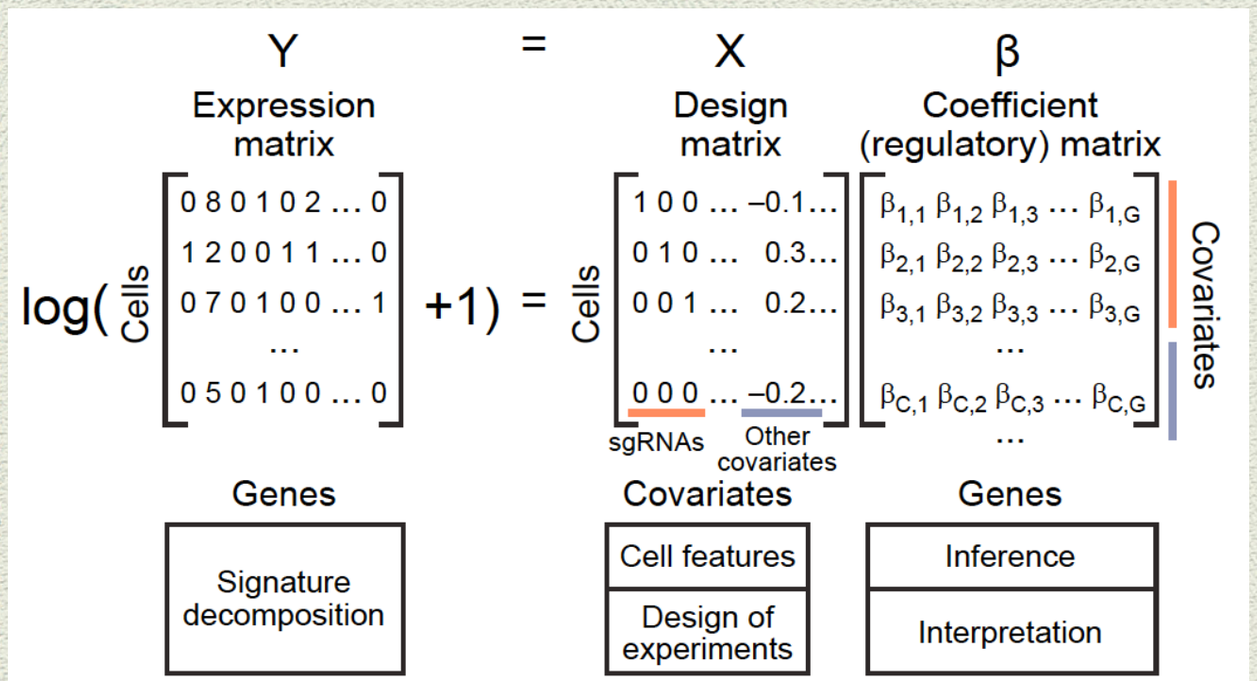


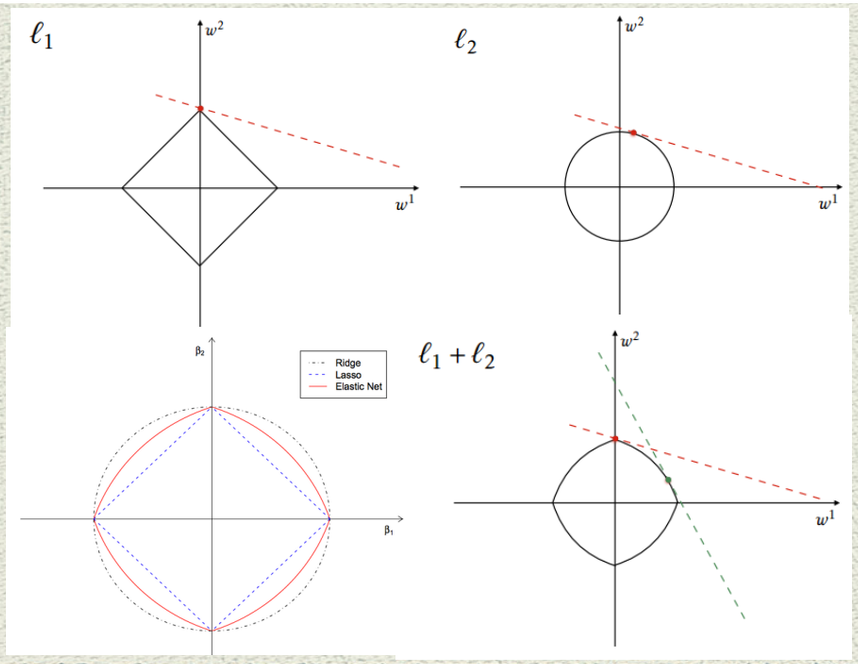
Cell type	sgRNA pool	Total cells	Time points
Mouse BMDC	Transcription factors (67 guides)	70,000	0 and 3 hr post-LPS
Human K562	Transcription factors (46 guides)	104,000	7 and 13 days
Human K562	Cell cycle regulators (36 guides)	26,000	7 days




INPUTS
Expression matrix (Y)
RNA barcode Dial-out PCR
Preassociated sgRNA/RNA barcode pairs
Cell barcodes sgRNA association
Initial abundances
Wildtype cells
Perturbed cells
MOI Sensitivity
Fitness effects
Cell state definition
Covariate matrix (X)
Linear model
Y = Xβ
X ⟷ Y
Cell state effects
Cell size effects
β
Residuals
Y – Xβ
OUTPUTS

Multi-Input-Multi-Output-Single-Cell-Analysis (MIMOSCA)





- Regularization by Elastic Net

$$\beta = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

- LASSO selects at most n variables, for p>n
- For highly correlated variables, LASSO tends to select one variable from a group

Regularized Linear regression model

- Some cells may have sgRNA but unperturbed
- Assume X has some uncertainty, use an EM-like procedure, refine covariates matrix by

$$P(X = 1|Y, \beta)$$

- Evaluate significant of coefficients by permutation test: randomized guide assignments to cells

Other Covariates

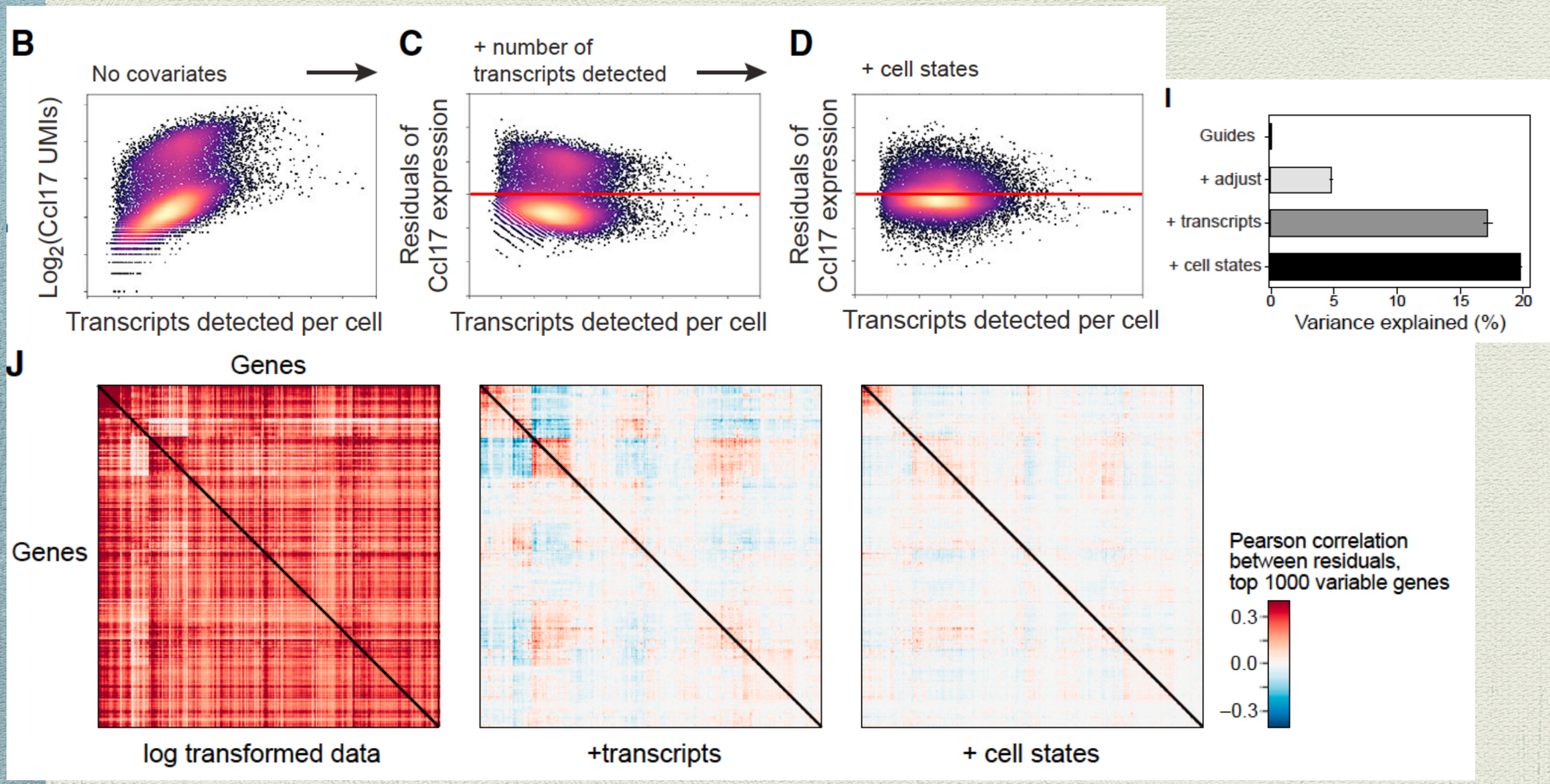
- Number of observed transcripts (cell quality) as covariates

- Cell state (e.g. cell subtype, cell cycle, etc.) by wild-type LPS-stimulated cells

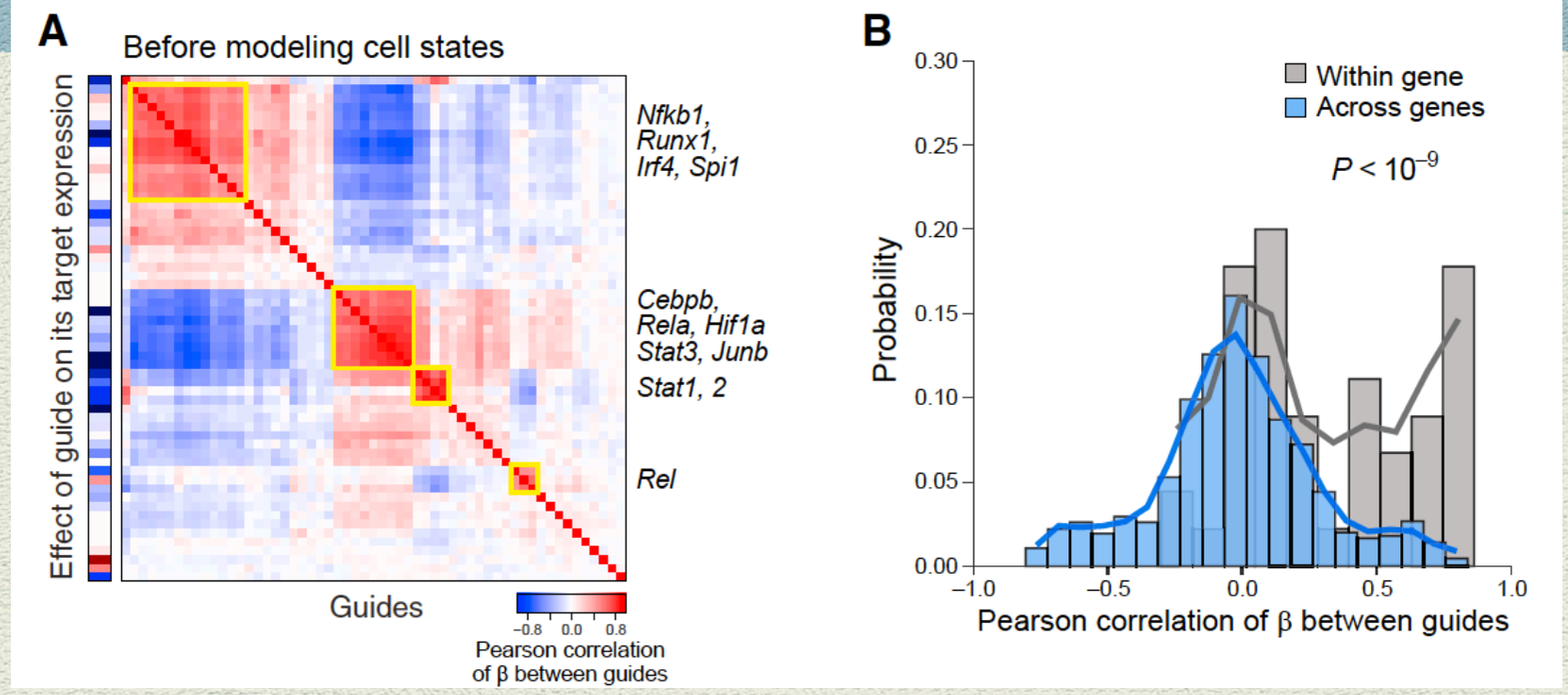
  - PCA and then clustering

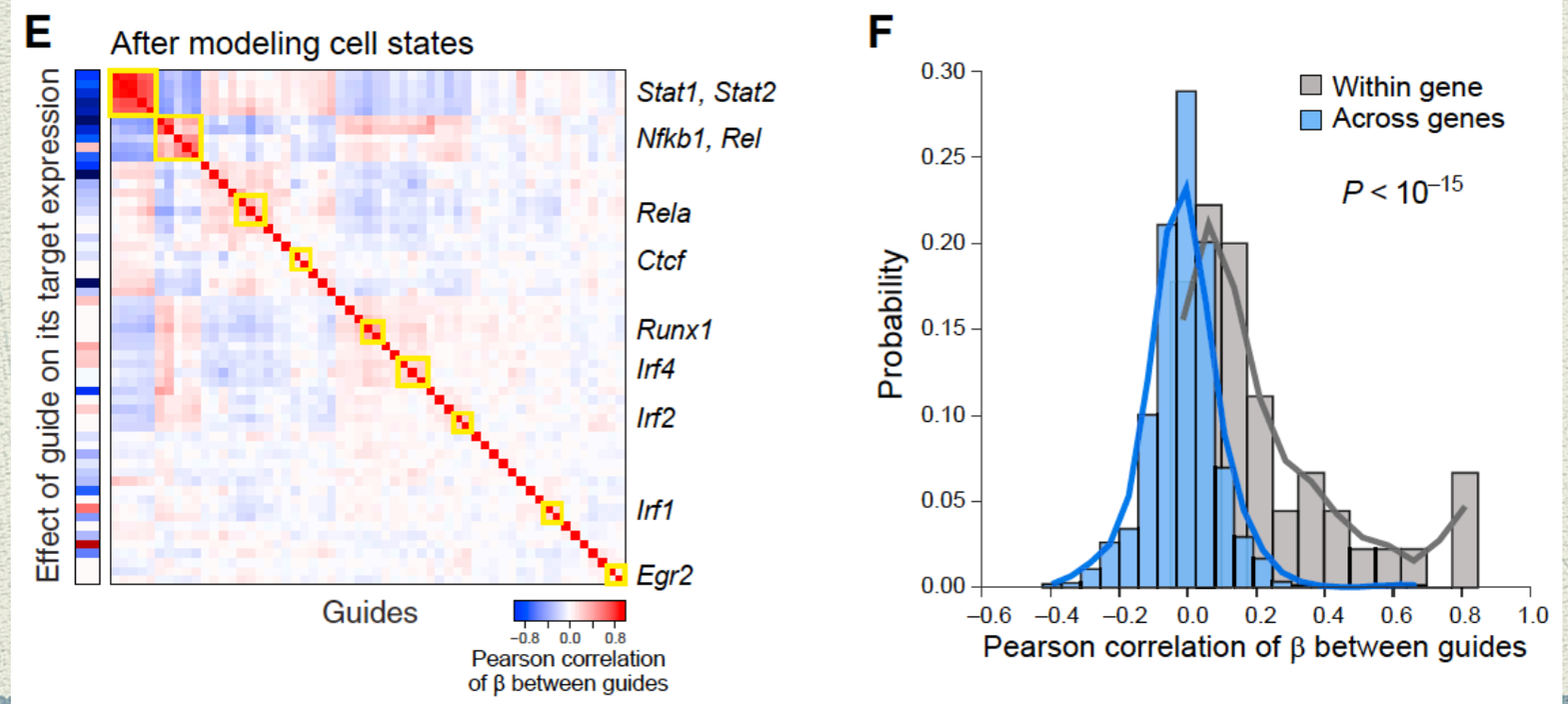
  - Project the perturb cells to these states; either continuous (cell cycle) or discrete (cell state)

Including covariates



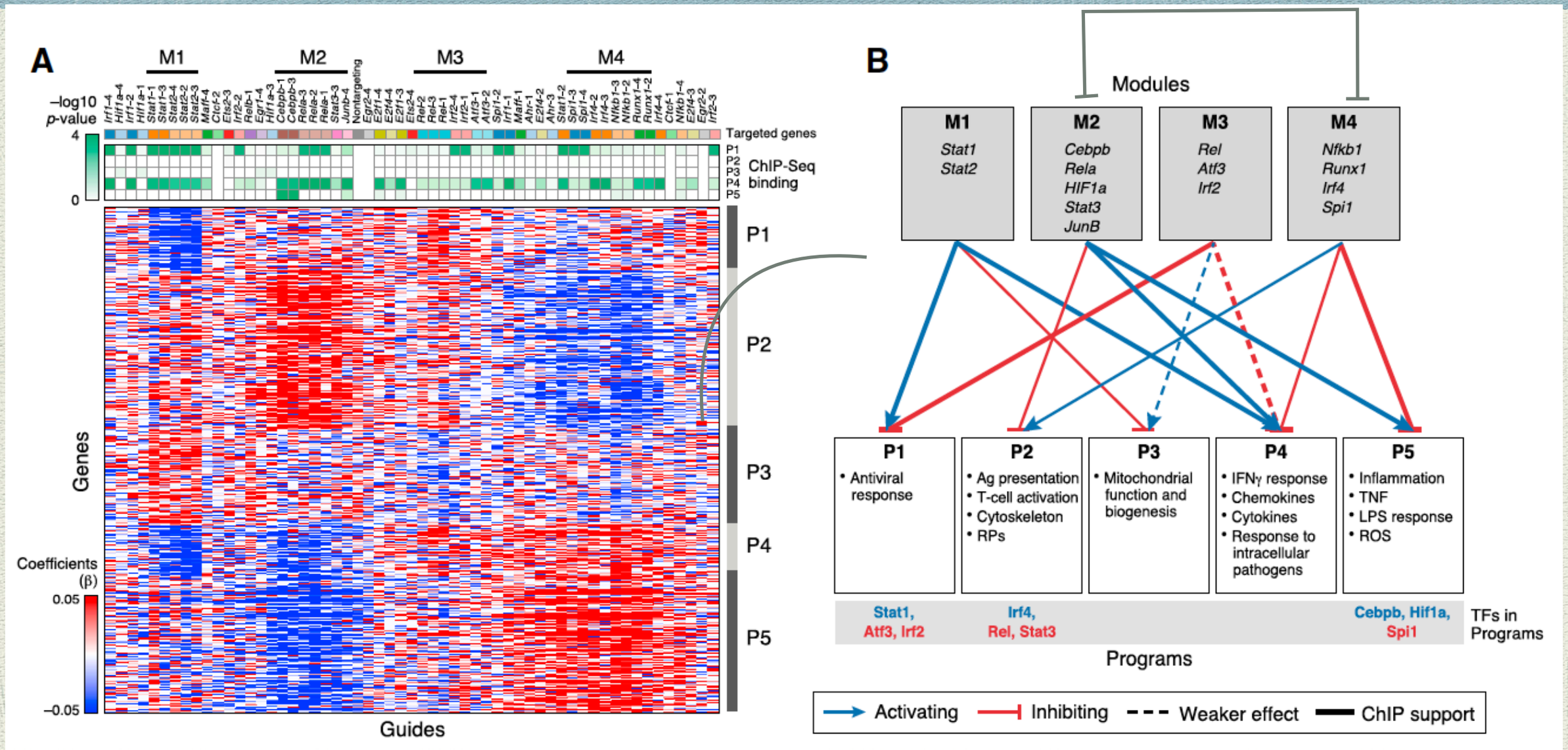
Verification of regression result





The coefficients of Guides targeting the same gene are correlated

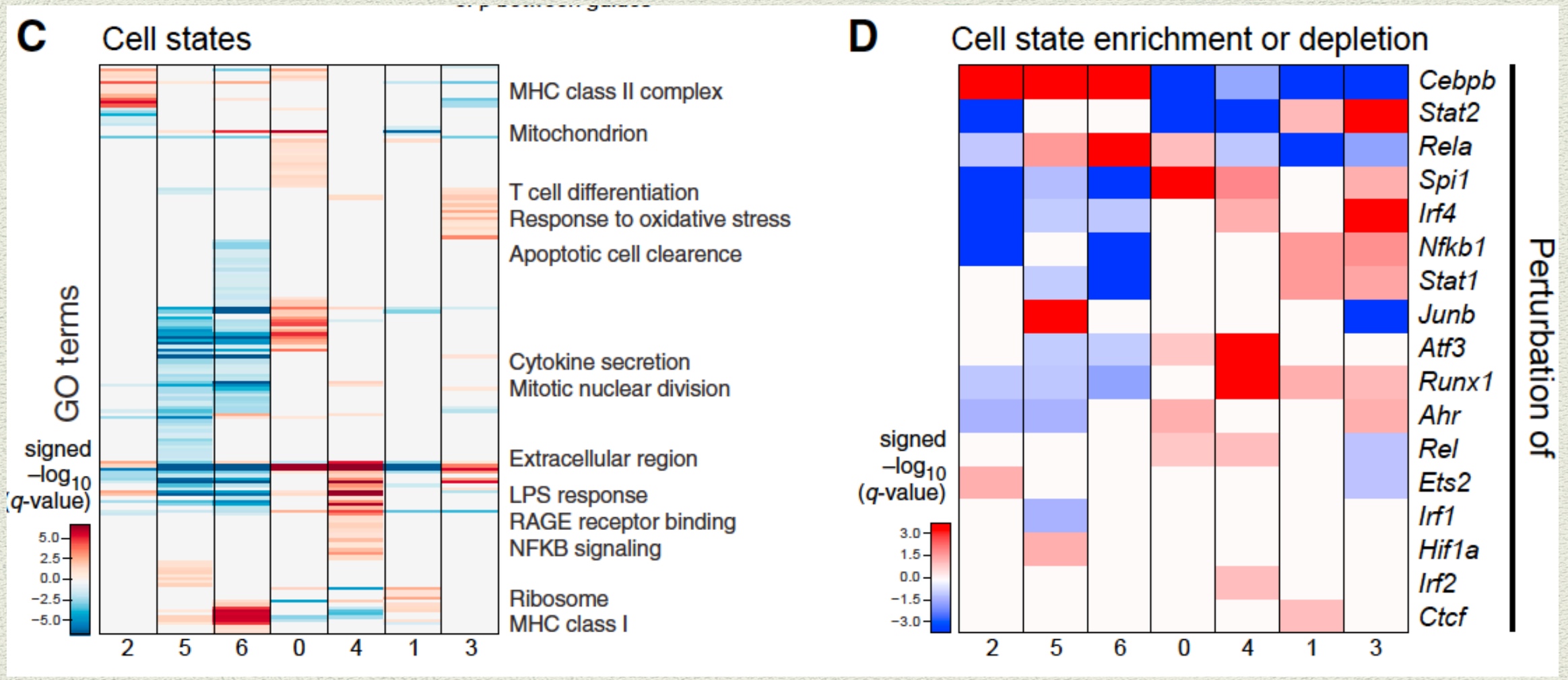
Transcriptional Program in BMPC response to LPS



BMPC: bone marrow derived dendritic cells
LPS: lipopolysaccharide

P2 and P4/5: alternative cell differentiation or maturity types

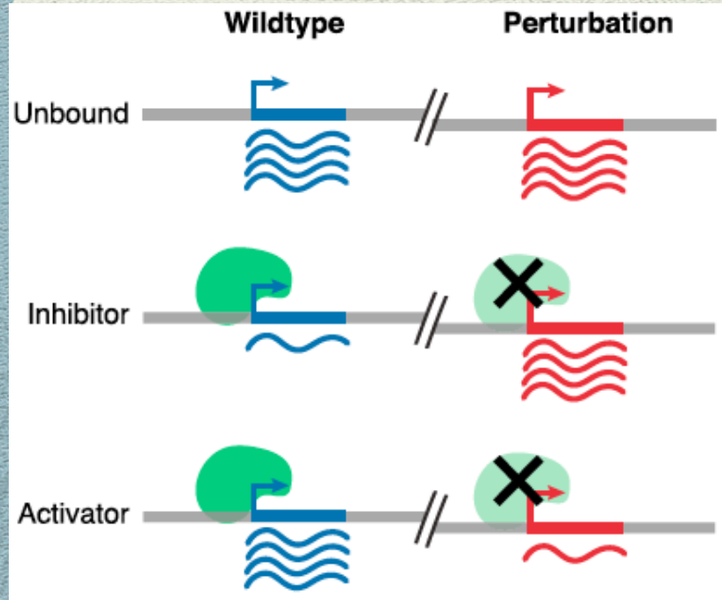
Impact on distribution of subtypes

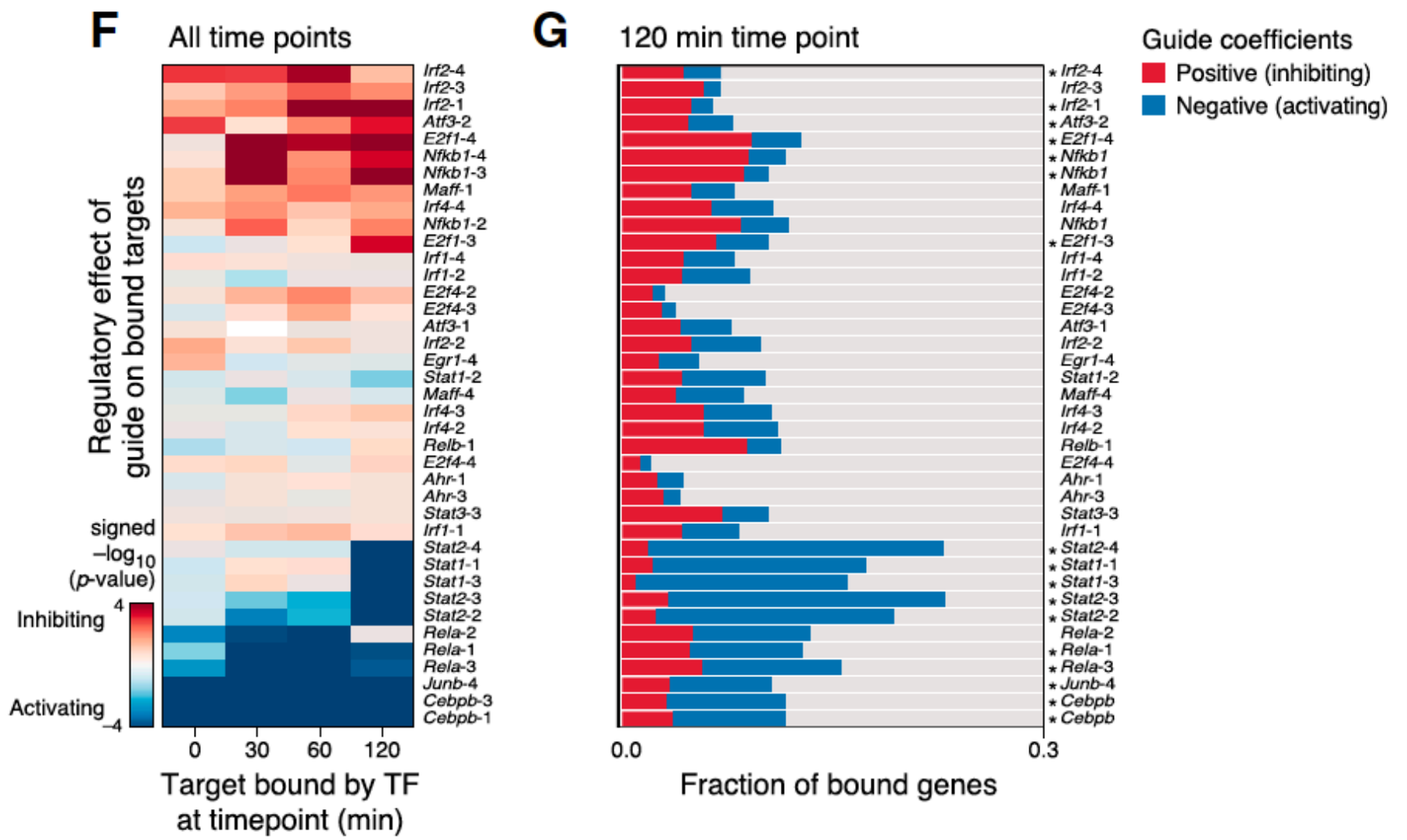


7 clusters of 1310 wild type LPS-simulated cells

TF effects on cell-state proportions

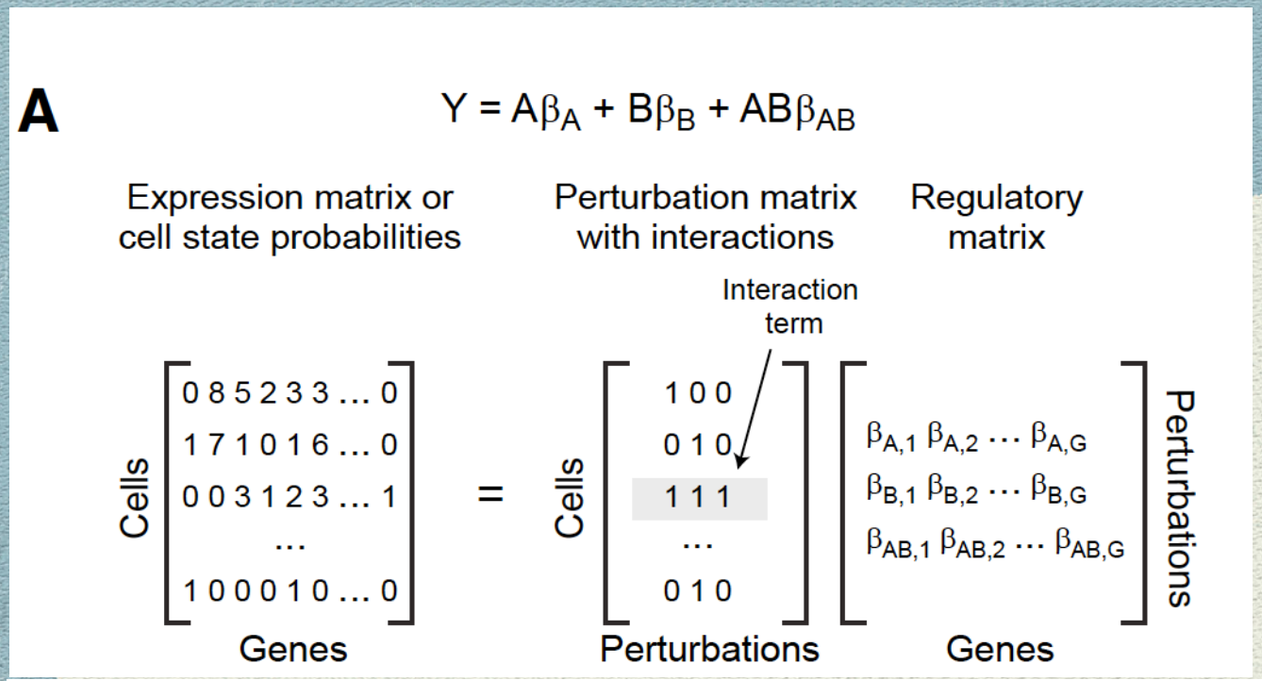
Model prediction supported by TF binding Profiles

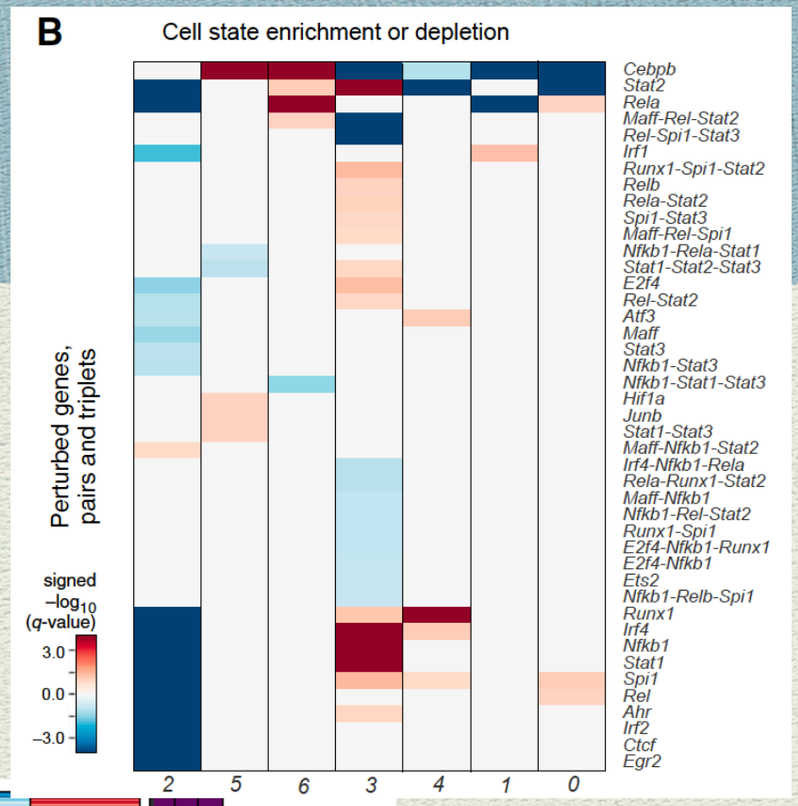


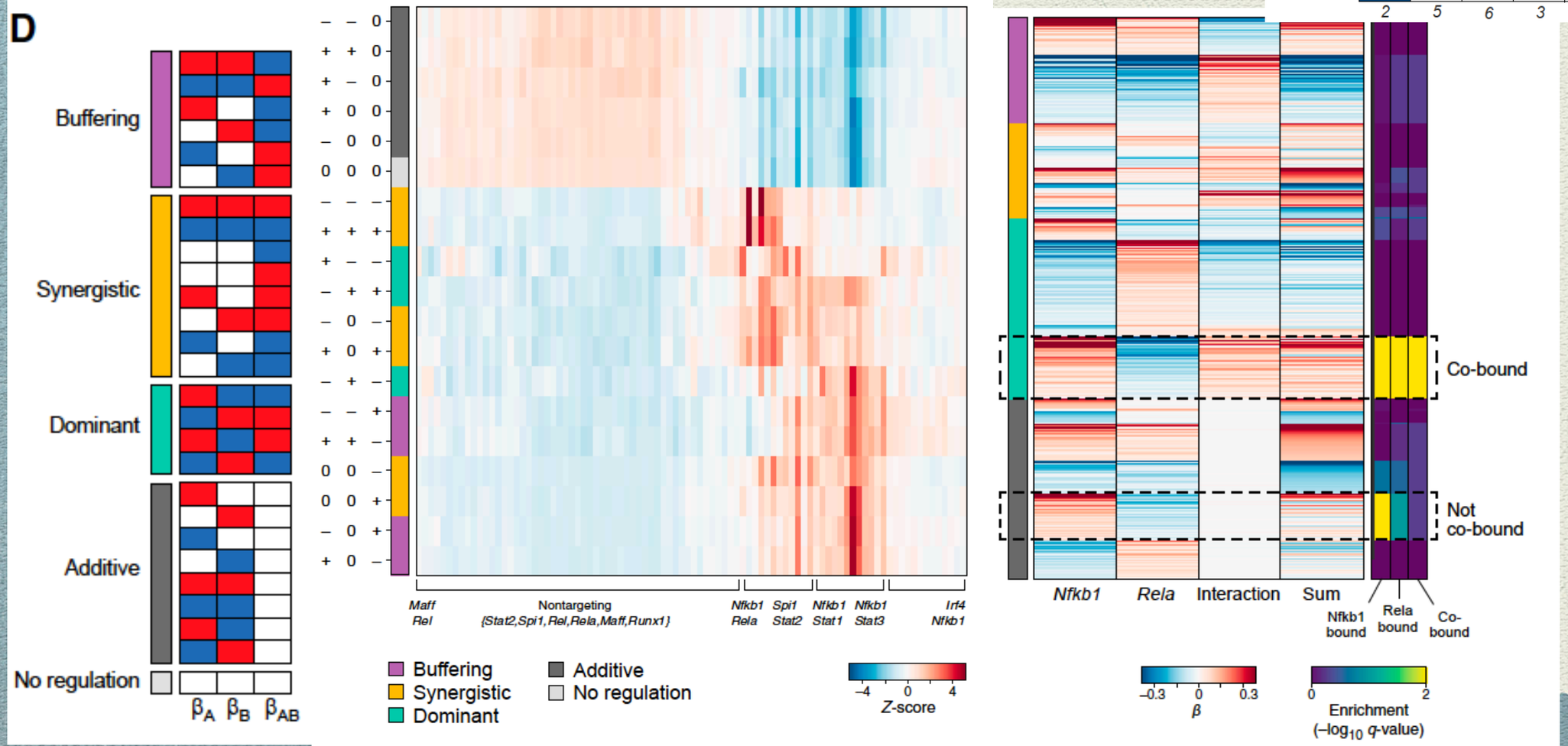


(E–G) Agreement with ChIP-seq. (E) Expected effects of TF perturbation. (F) Average regulatory effect of each guide (rows) on the genes bound by its target at four time points (columns). (G) Proportion of bound targets at 120 min post-LPS for each TF (rows) that are repressed (blue), activated (red), or unaffected (gray) by the TF's perturbation. Asterisks: significant (as in F, $p < 0.05$).

Genetic Interaction: Affect both cell states and gene expression

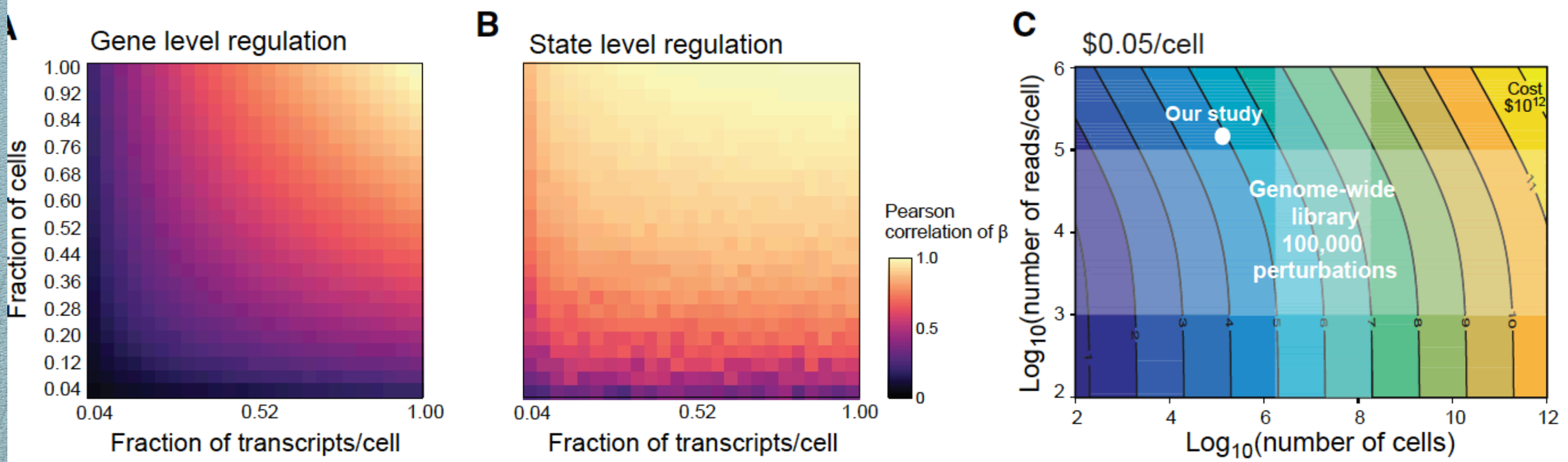






only the set with the dominant interaction is enriched for co-binding

Saturation analysis: downsampling of cells & reads



(A and B) Saturation analysis. Effect of the number of cells (y axis) and transcripts (x axis) on recovery as measured by correlation (color bar) with either the per gene (A) or cell-state signature (B) effects observed in the full data for BMDCs at 3 hr. The number of cells per perturbation (1.0) is a mean of 300 and a median of 155 and the number of transcripts per cell (1.0) is a median of 5,074.

(C) Tradespace of number of cells (x axis) and measurements per cell (y axis) required for scaling Perturb-seq.

broad survey of phenotypes can be performed with a few tens of cells per perturbation

Modularity and sparsity

Expressed barcodes mark cells derived from a common ancestor for lineage tracing

In vivo