Analysis of single cell CRISPR regulatory screens

Eugene Katsevich

Statistics Department
University of Pennsylvania

August 21, 2020

References

A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens.

M. Gasperini et al., Cell, 2019.

Conditional resampling improves sensitivity and specificity of single cell CRISPR regulatory screens.

E. Katsevich and K. Roeder, 2020. Available on bioRxiv.

Problem setup

We have a bunch of cells, indexed $i = 1, \ldots, n$.

Focusing on one enhancer and one gene, for each cell i we measure

- $X_i \in \{0, 1\}$, gRNA presence
- $Y_i \in \{0, 1, 2, \ldots\}$, gene expression (UMI count)
- $Z_i \in \mathbb{R}^d$, technical factors

Analysis goal: Determine if the gene is differentially expressed between cells with and without gRNA.

Approach 1: Negative binomial regression (Monocle2)

Assume the model

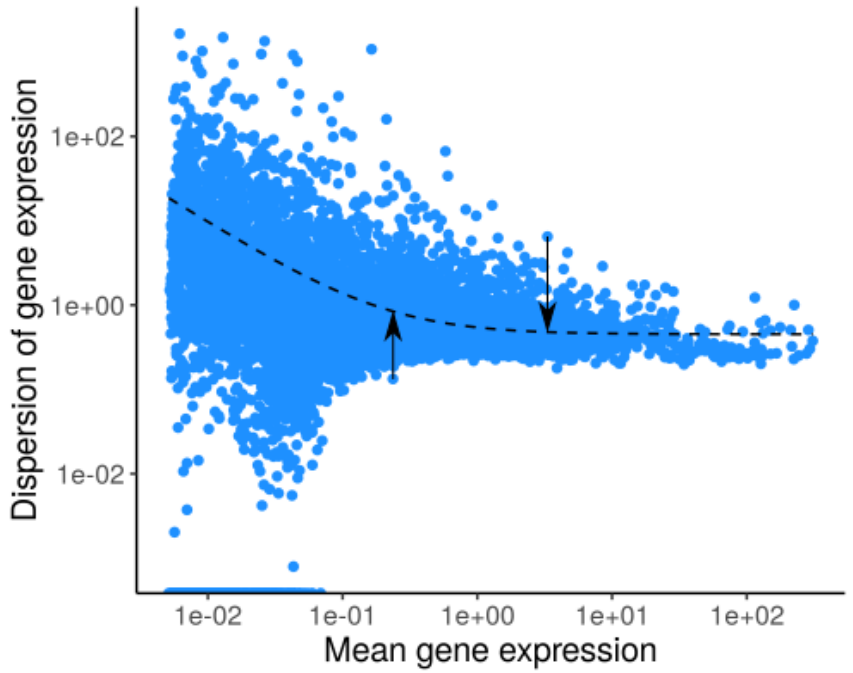
$$Y_i \overset{\text{ind}}{\sim} \text{NegBin}(s_i\mu_i, \alpha);$$

$$\log(\mu_i) = \beta_0 + X_i\beta + Z_i^T\gamma,$$

where s_i are "size factors",
α is dispersion estimate.

Null hypothesis: $\beta = 0$.

Estimating the dispersion parameter α



A sign of trouble

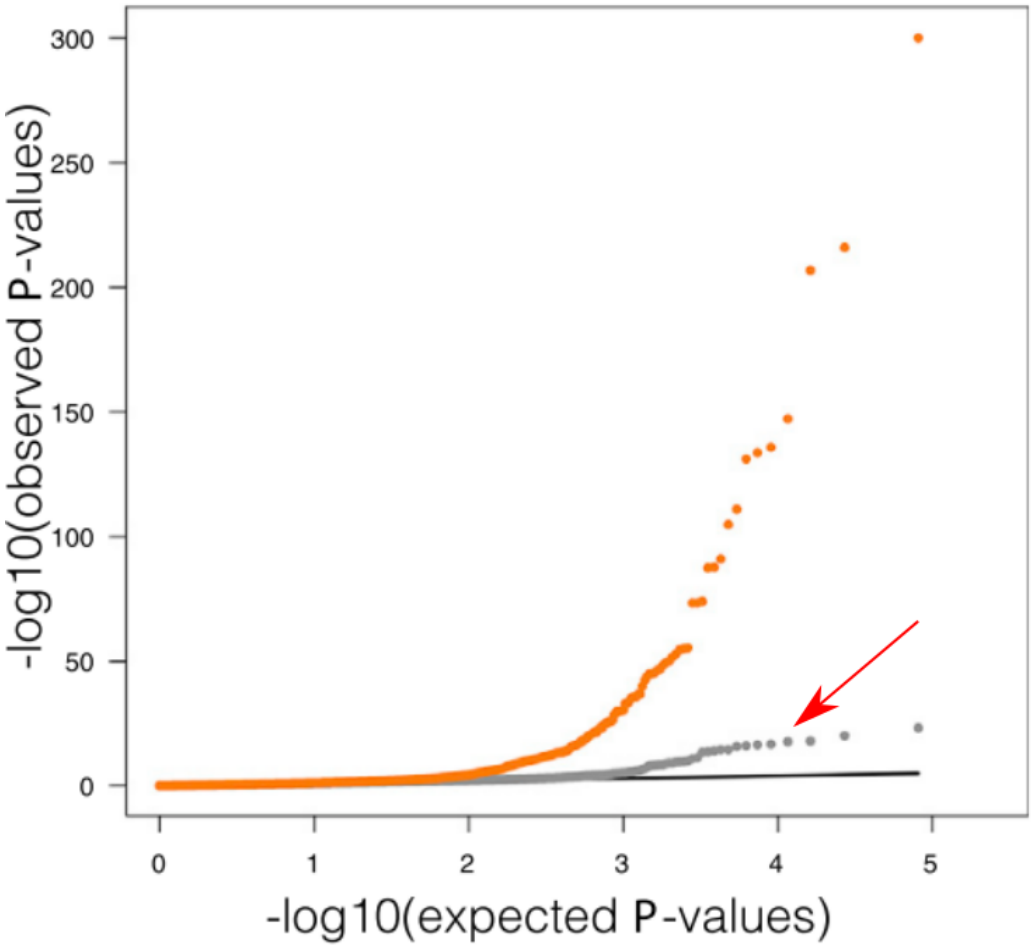


Figure 3E of Gasperini et al. (Cell, 2019).

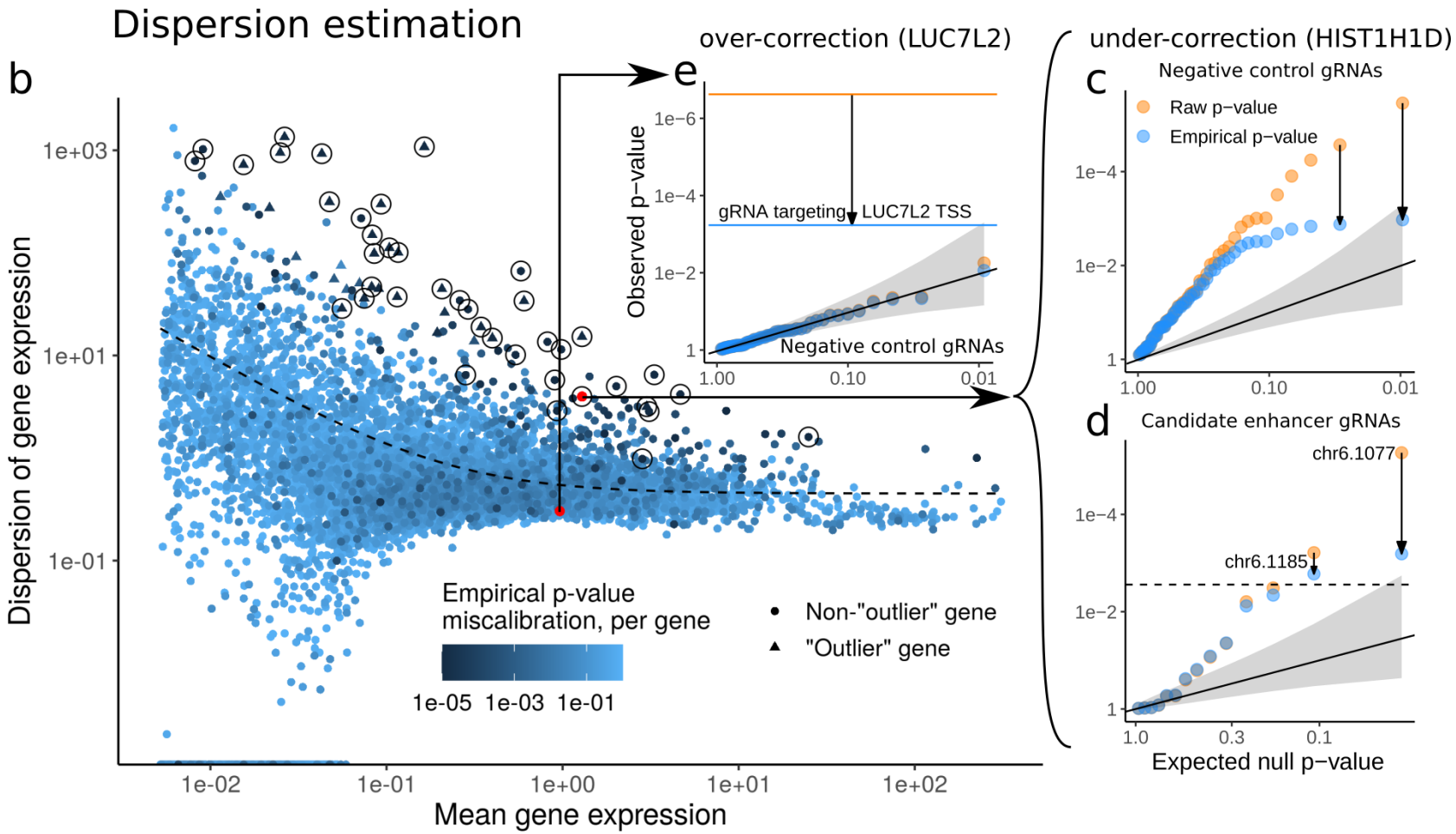
Approach 2: Build null distribution from negative controls

First, aggregate all gene / negative control gRNA p-values.

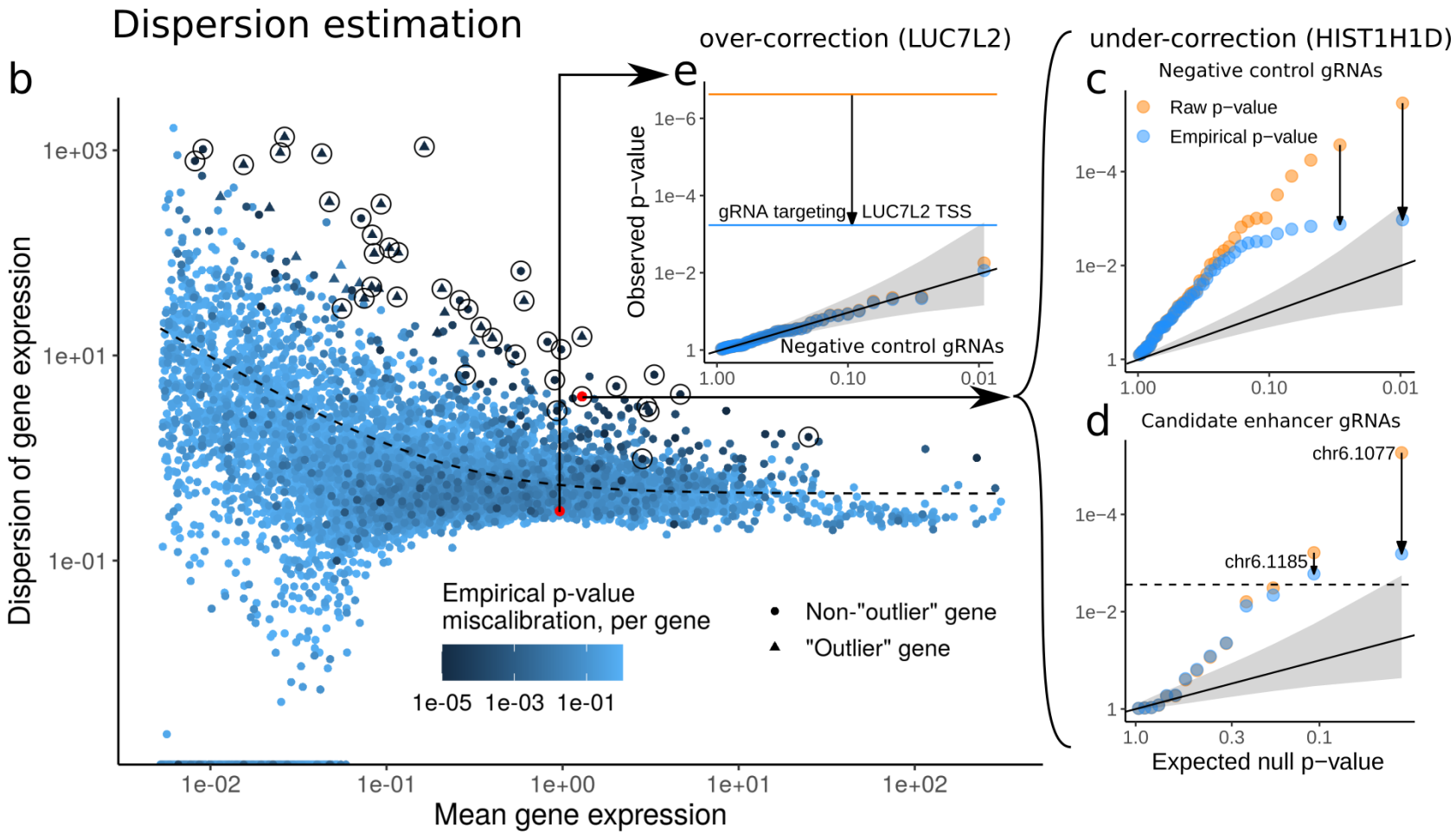
Calibrate all gene / candidate enhancer p-values against this distribution instead of uniform.

Approach taken by Gasperini et al. (2019).

A one-size-fits-all approach does not fix the problem



A one-size-fits-all approach does not fix the problem



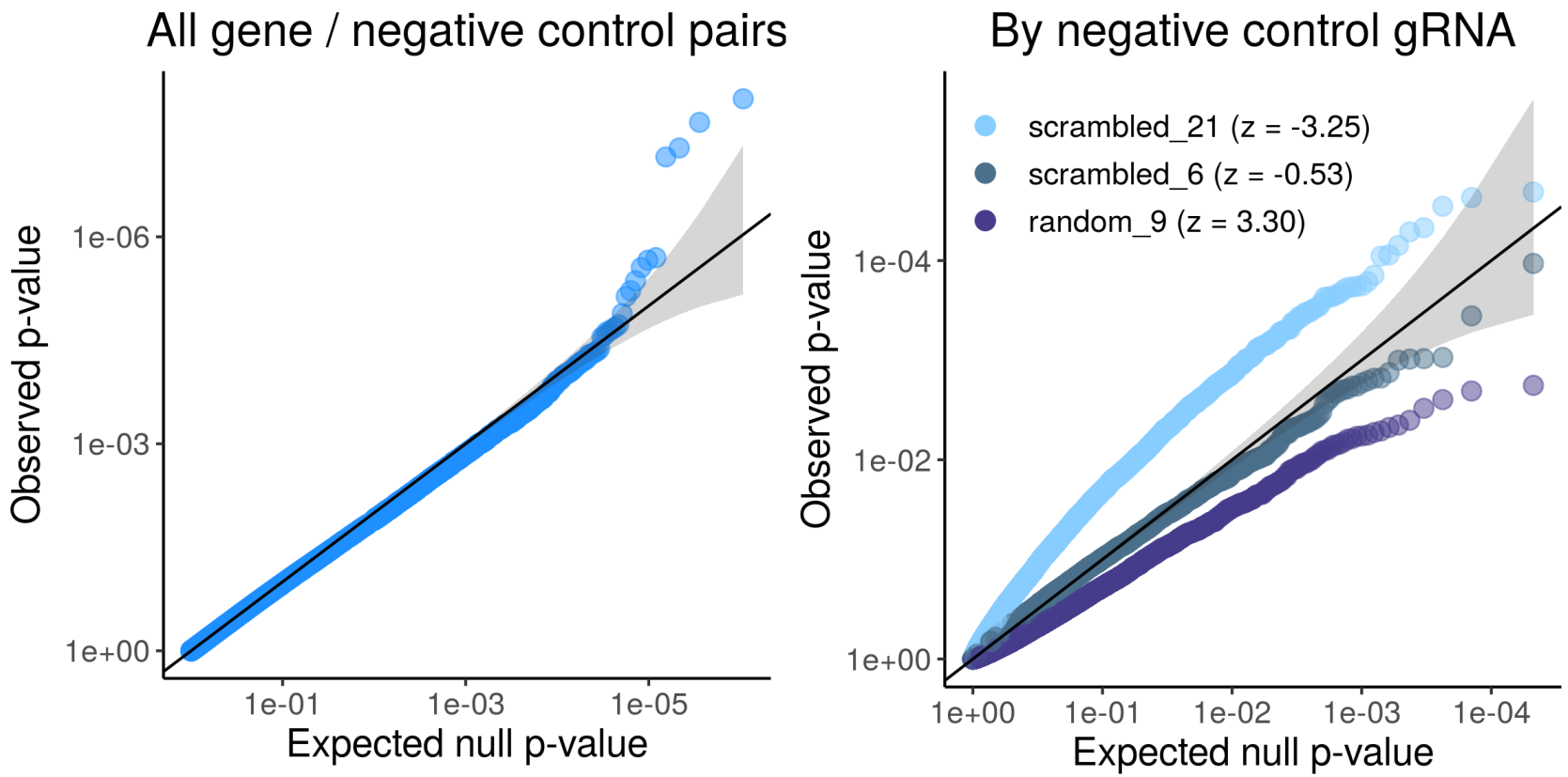
This was the starting point for our work.

Improving the negative binomial model (first try)

Remove shrinkage of dispersion estimates.

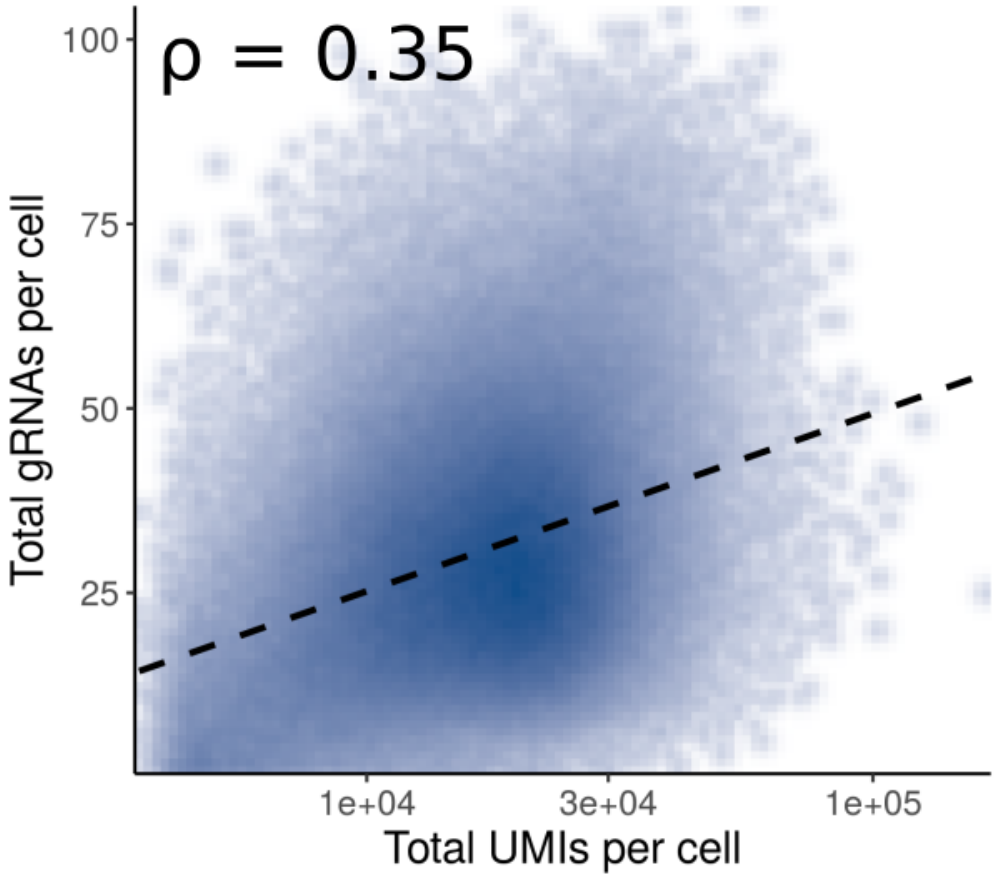
Improving the negative binomial model (first try)

Remove shrinkage of dispersion estimates.



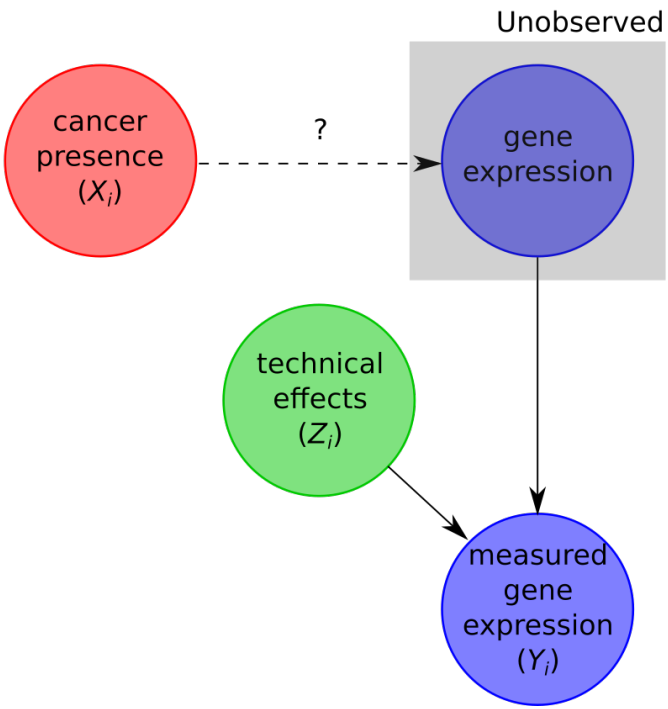
Roughly, z-values measure association of gRNA with total UMIs.

Sequencing depth impacts gRNA detection

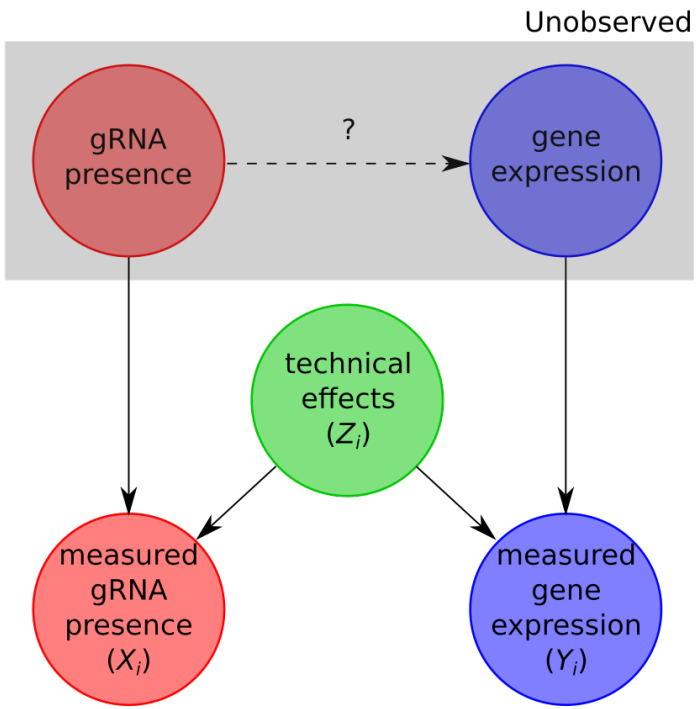


Sequencing depth acts as a confounder

Usual differential expression setup (per patient)



Single cell CRISPR screen setup (per cell)

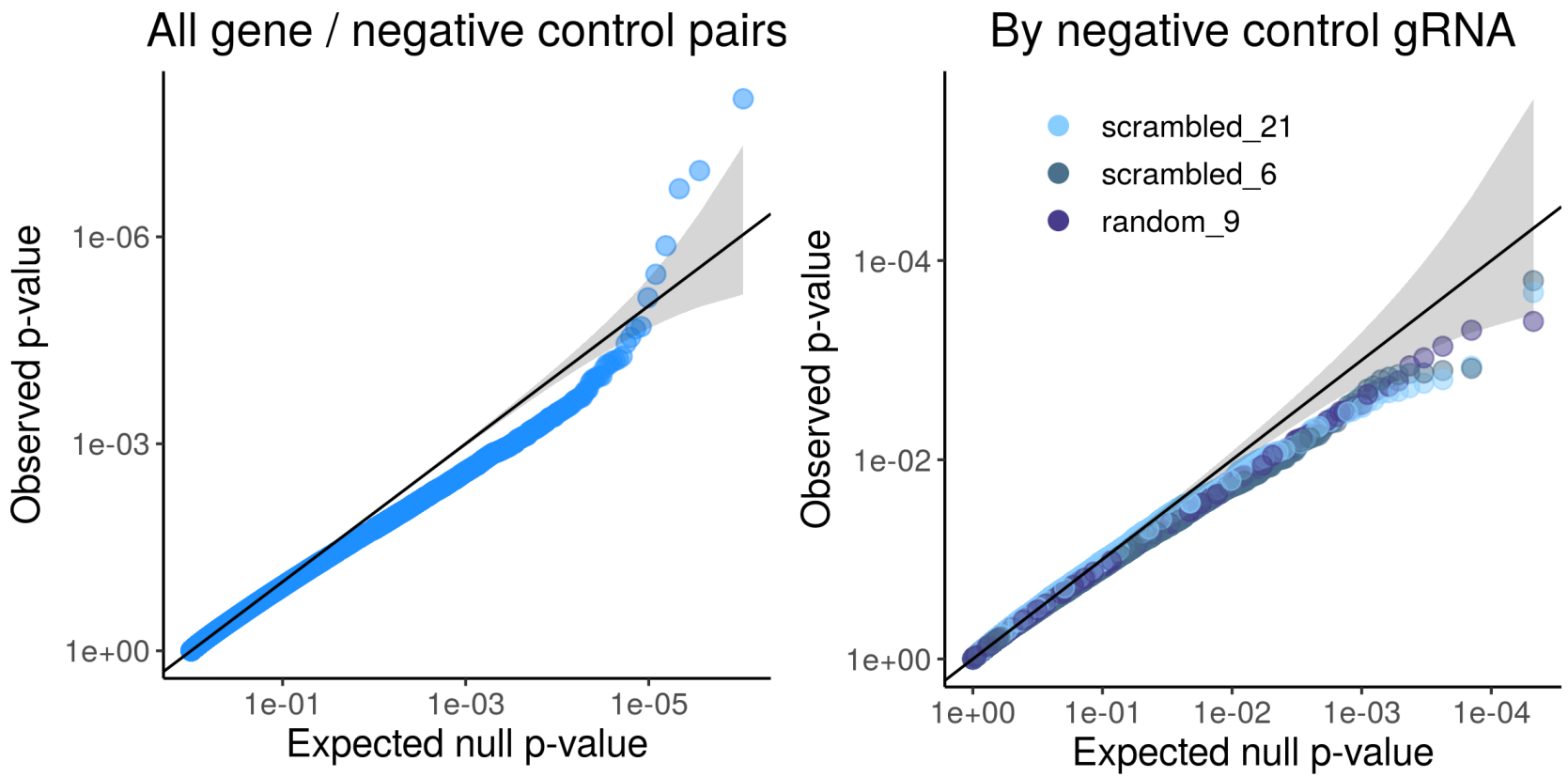


Improving the negative binomial model (second try)

Add sequencing depth as a covariate, instead of using size factors.

Improving the negative binomial model (second try)

Add sequencing depth as a covariate, instead of using size factors.



Not too bad, though it's clear some miscalibration remains.

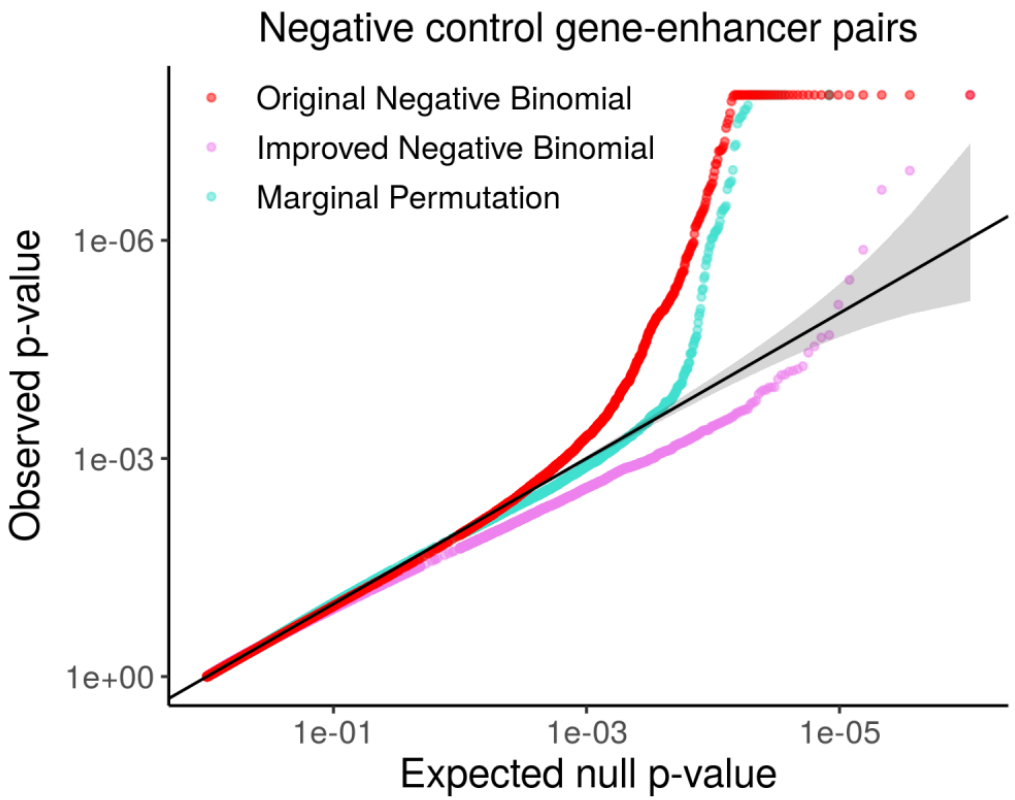
Why not just use a permutation approach?

Permute gRNA assignments among cells, e.g. as in scMAGeCK.[1]

[1] Yang et al. 2020

Why not just use a permutation approach?

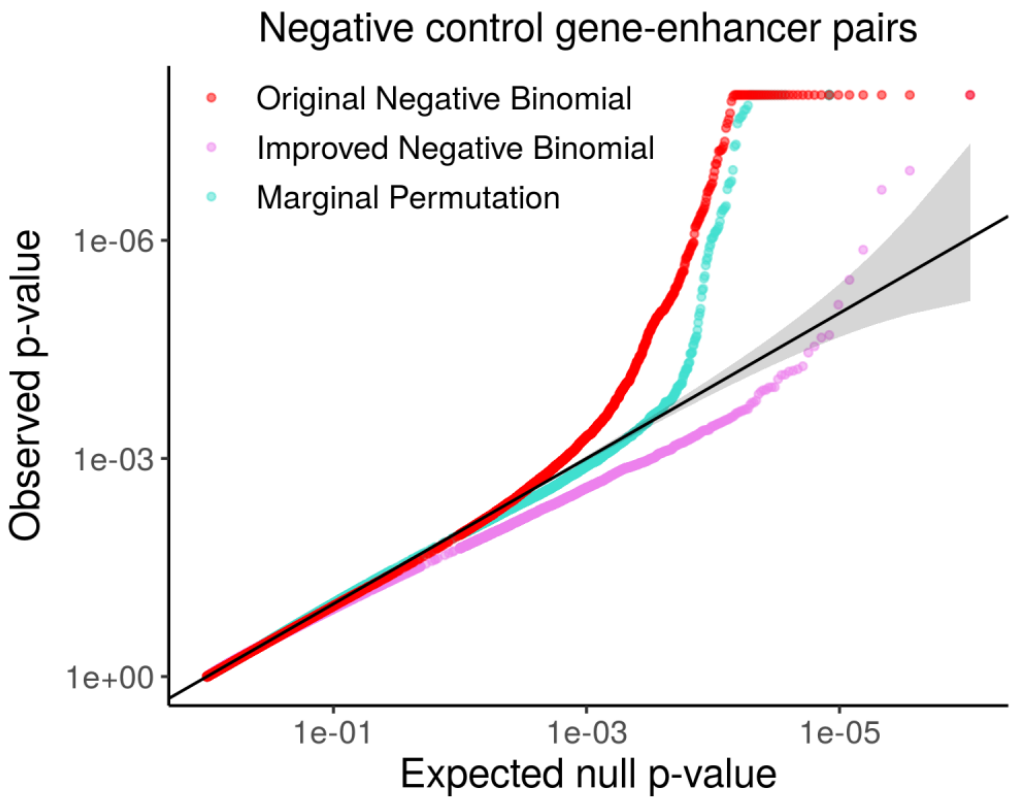
Permute gRNA assignments among cells, e.g. as in scMAGeCK.[1]



[1] Yang et al. 2020

Why not just use a permutation approach?

Permute gRNA assignments among cells, e.g. as in scMAGeCK.[1]



There is a more principled way of resampling gRNA assignments.

[1] Yang et al. 2020

The conditional randomization test (Candès et al., 2018)

Let $\pi(Z) \approx \mathbb{P}[X = 1|Z]$ be a working model for gRNA observation.

We can then calibrate any test statistic $T(X, Y, Z)$ by resampling

$$\widetilde{X}_i = \begin{cases} 1, & \text{with probability } \pi(Z_i); \\ 0, & \text{with probability } 1 - \pi(Z_i). \end{cases}$$

and recomputing $T(\widetilde{X}, Y, Z)$.

Similar to a permutation test, but takes covariates into account.

The conditional randomization test (Candès et al., 2018)

Let $\pi(Z) \approx \mathbb{P}[X = 1|Z]$ be a working model for gRNA observation.

We can then calibrate any test statistic $T(X, Y, Z)$ by resampling

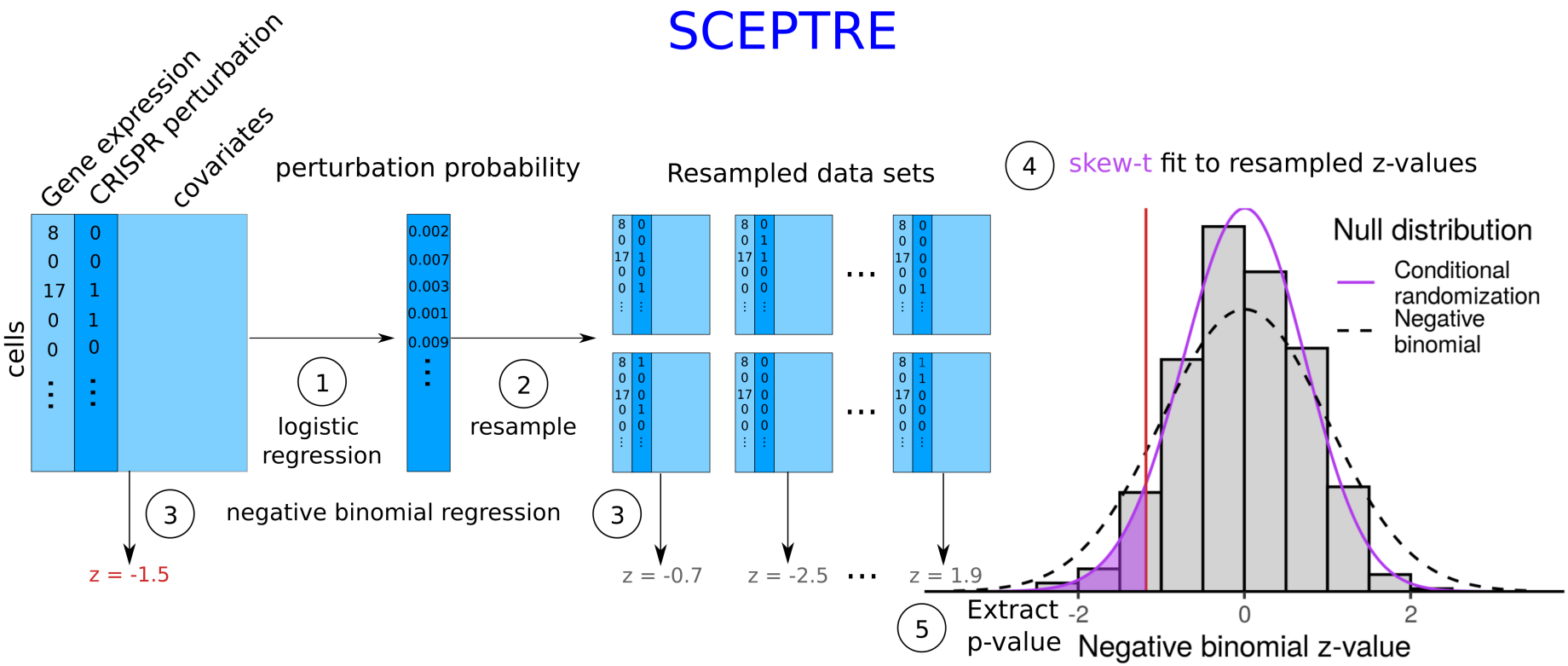
$$\widetilde{X}_i = \begin{cases} 1, & \text{with probability } \pi(Z_i); \\ 0, & \text{with probability } 1 - \pi(Z_i). \end{cases}$$

and recomputing $T(\widetilde{X}, Y, Z)$.

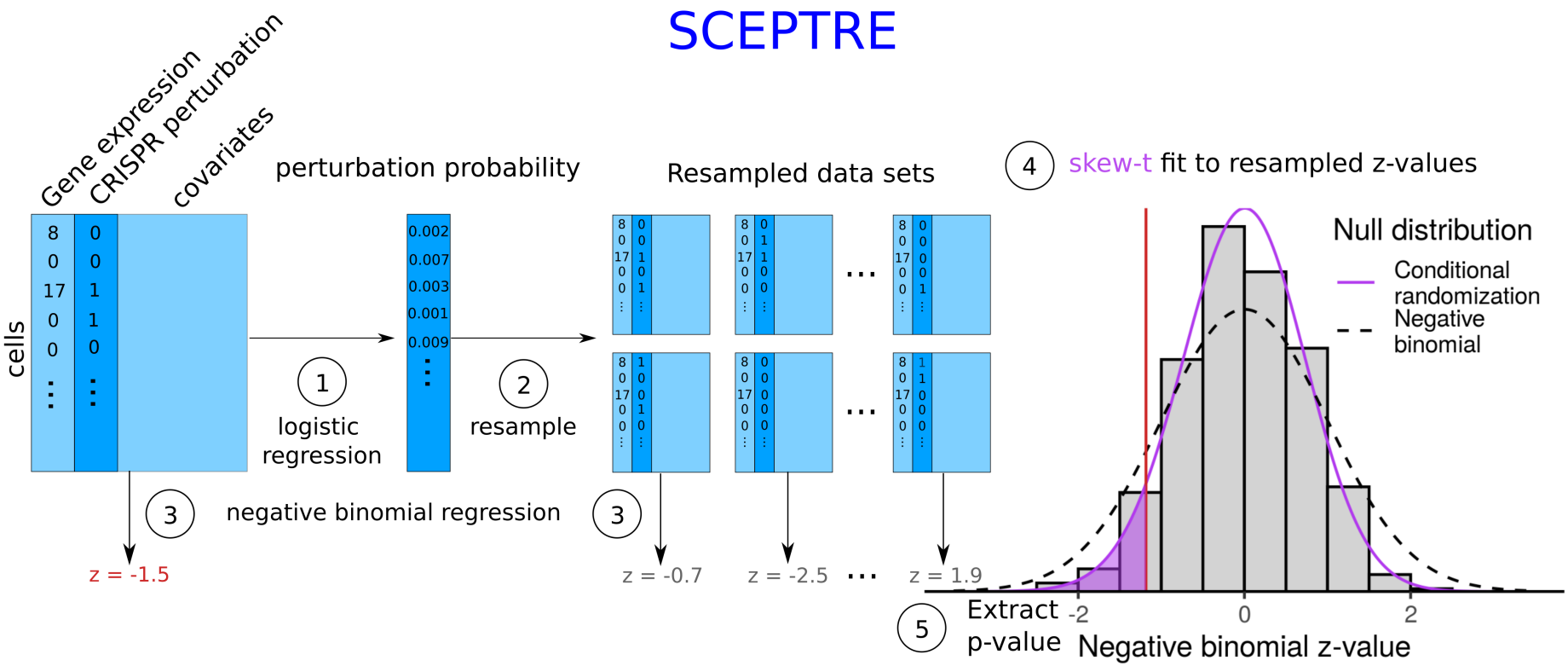
Similar to a permutation test, but takes covariates into account.

Valid calibration despite misspecifications of expression model!

New method for single cell CRISPR screen analysis

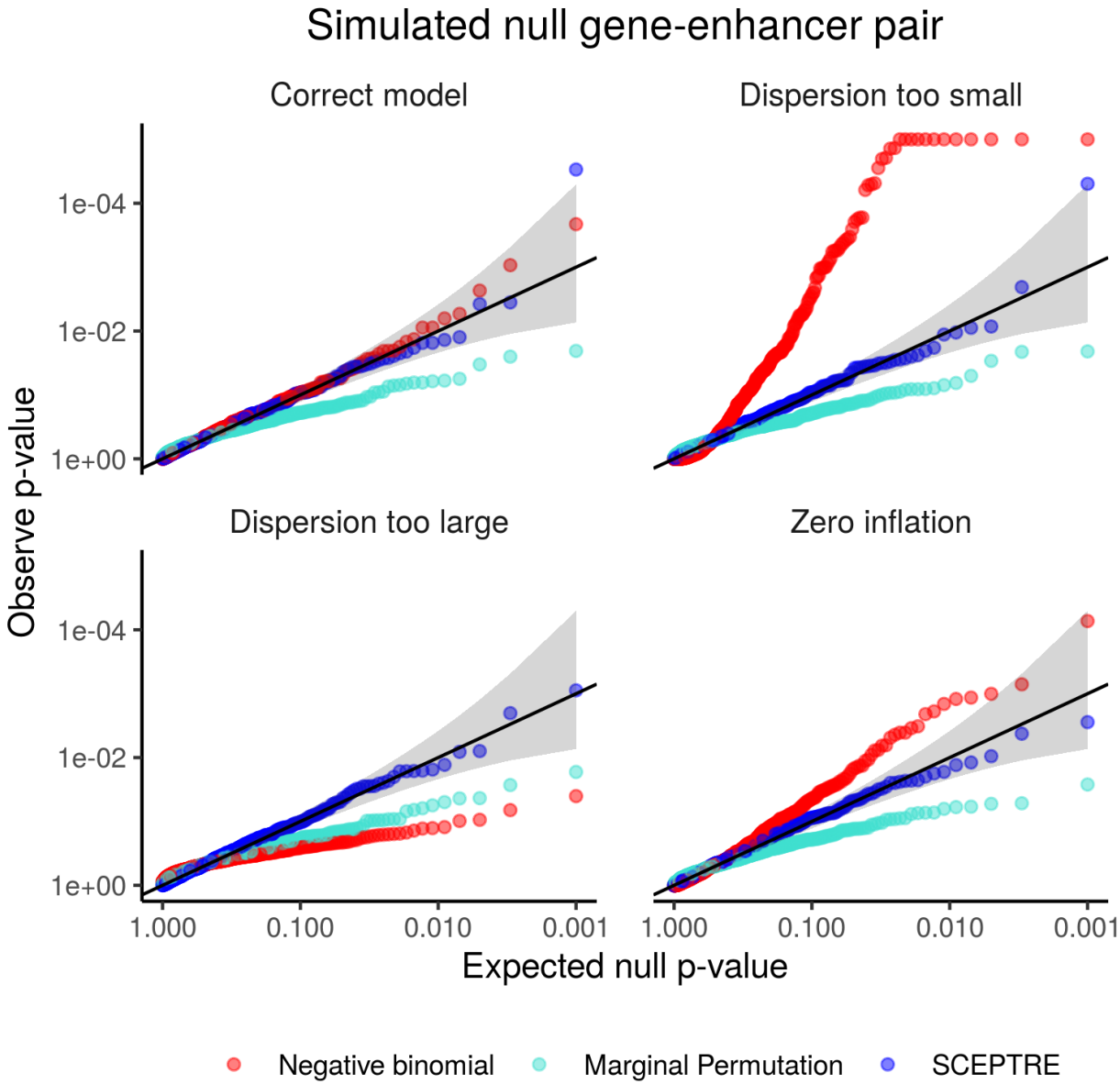


New method for single cell CRISPR screen analysis

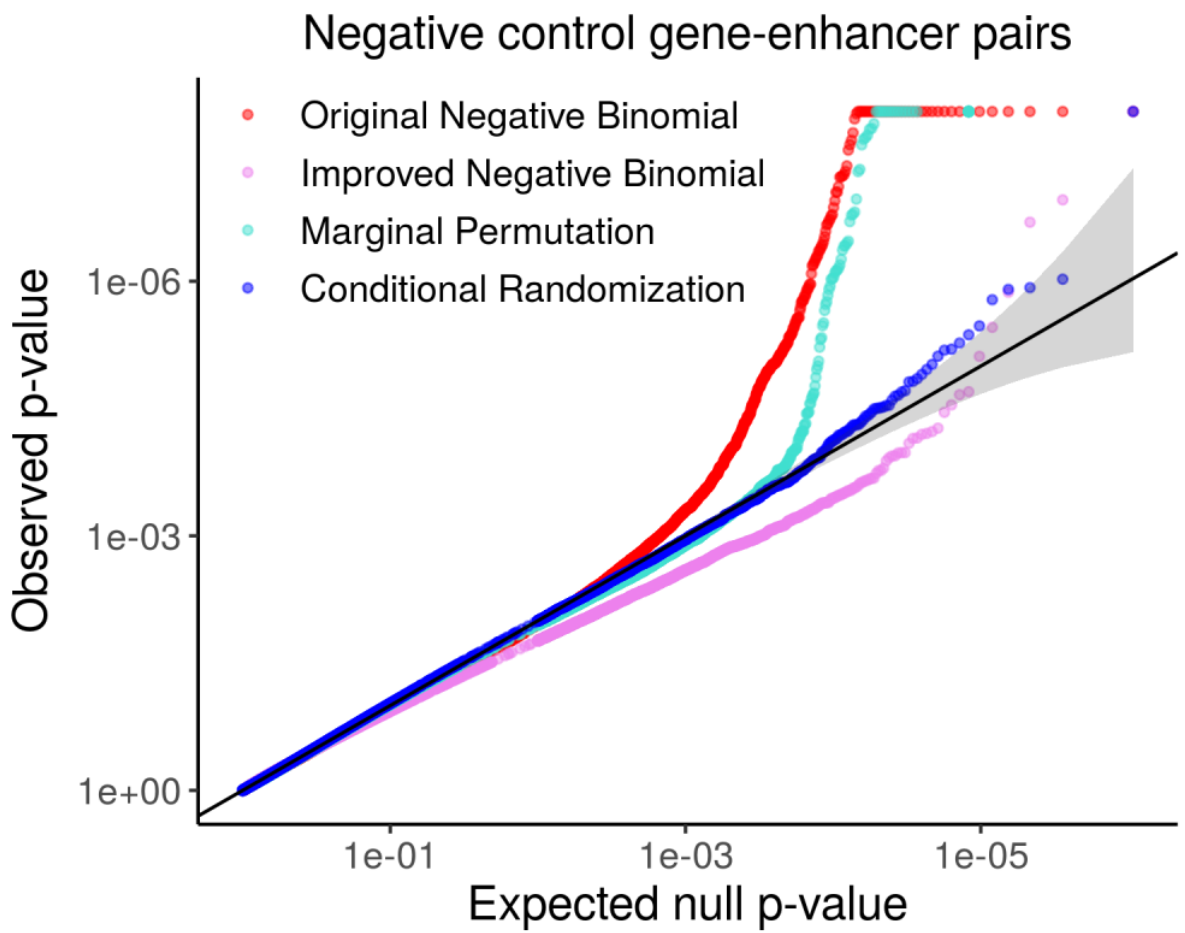


Accelerations reduce computation time for a gene-enhancer pair from 25 minutes to 19 seconds. Original approach takes 3 seconds.

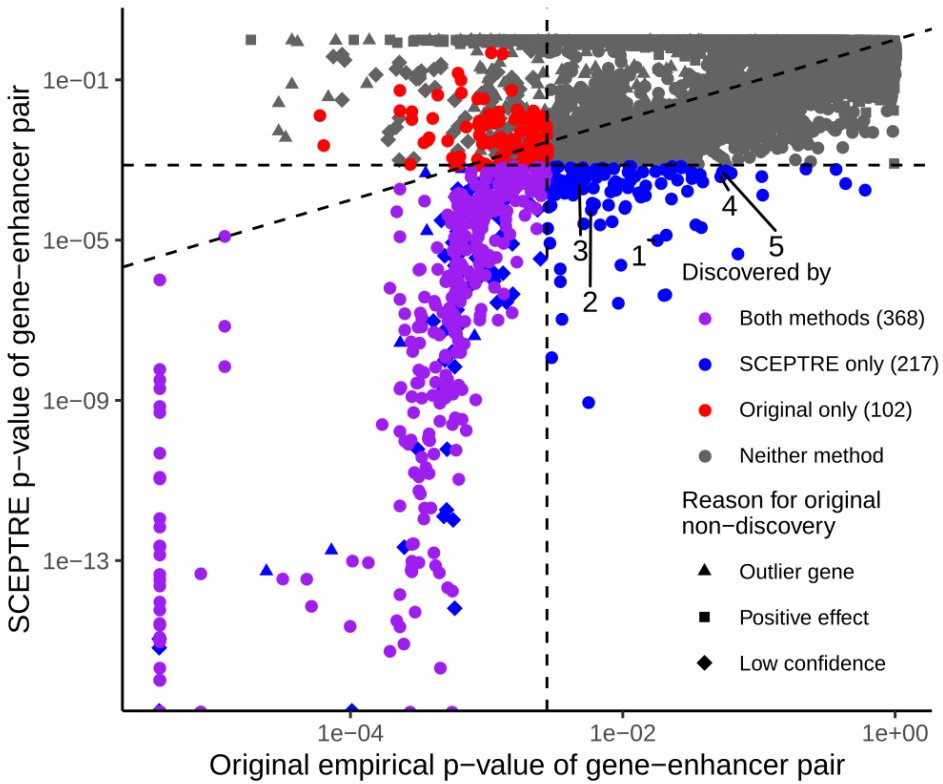
Excellent calibration on simulated data



Excellent calibration on negative control data



SCEPTRE discoveries: many different and some promising

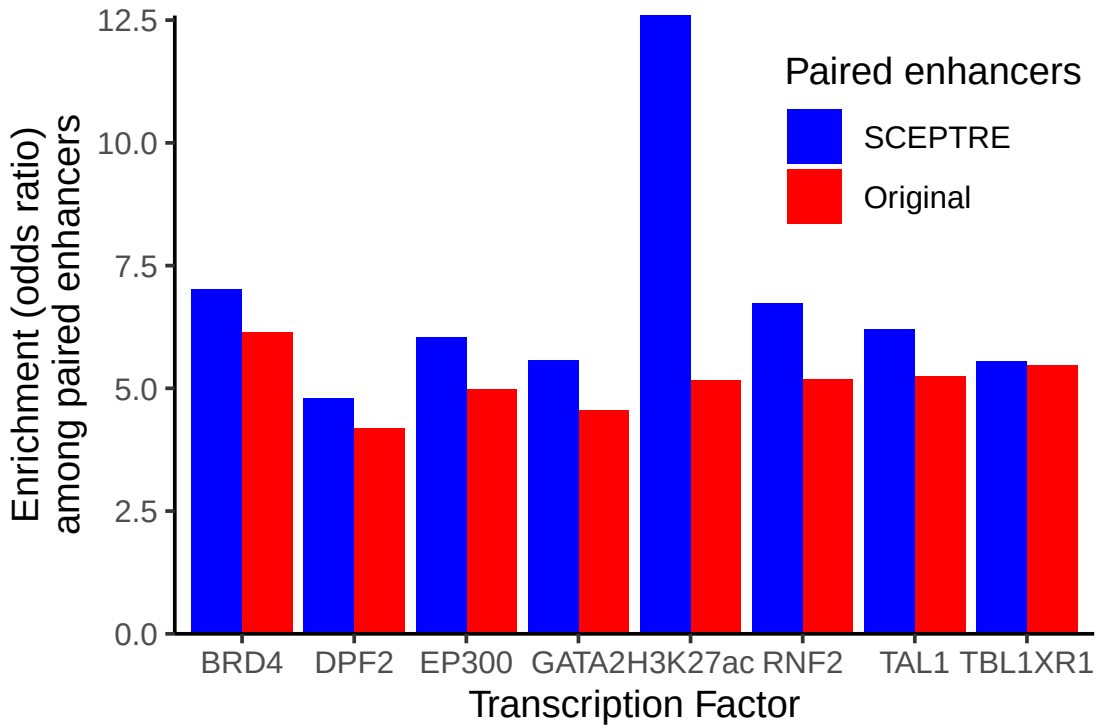


	Gene	Enhancer	SCEPTRE	Original	eQTL	eRNA
1	TOP1	chr20.1629	1.1e-05	1.8e-02	NA	6.6e-05
2	B3GNT2	chr2.2237	6.0e-05	5.8e-03	2.7e-26	NA
3	AGFG1	chr2.6820	2.6e-04	4.8e-03	5.2e-08	NA
4	EIF1	chr17.2516	4.1e-04	5.2e-02	NA	1.2e-06
5	PTPN1	chr20.2381	5.4e-04	5.3e-02	NA	2.0e-18

SCEPTRE discoveries: larger fraction within same TAD

	Gene-enhancer pairs		
	Same TAD	Total	Fraction
Original	334	470	0.71
SCEPTRE	442	585	0.76

SCEPTRE discoveries: more enriched for TF binding



Conclusions

- Single cell CRISPR screens open amazing scientific opportunities but also present new statistical challenges

Conclusions

- Single cell CRISPR screens open amazing scientific opportunities but also present new statistical challenges
- We propose a new paradigm for calibrating any test statistic without relying on validity of gene expression model

Conclusions

- Single cell CRISPR screens open amazing scientific opportunities but also present new statistical challenges
- We propose a new paradigm for calibrating any test statistic without relying on validity of gene expression model
- Improved statistical methodology yields new, biologically relevant regulatory relationships

Conclusions

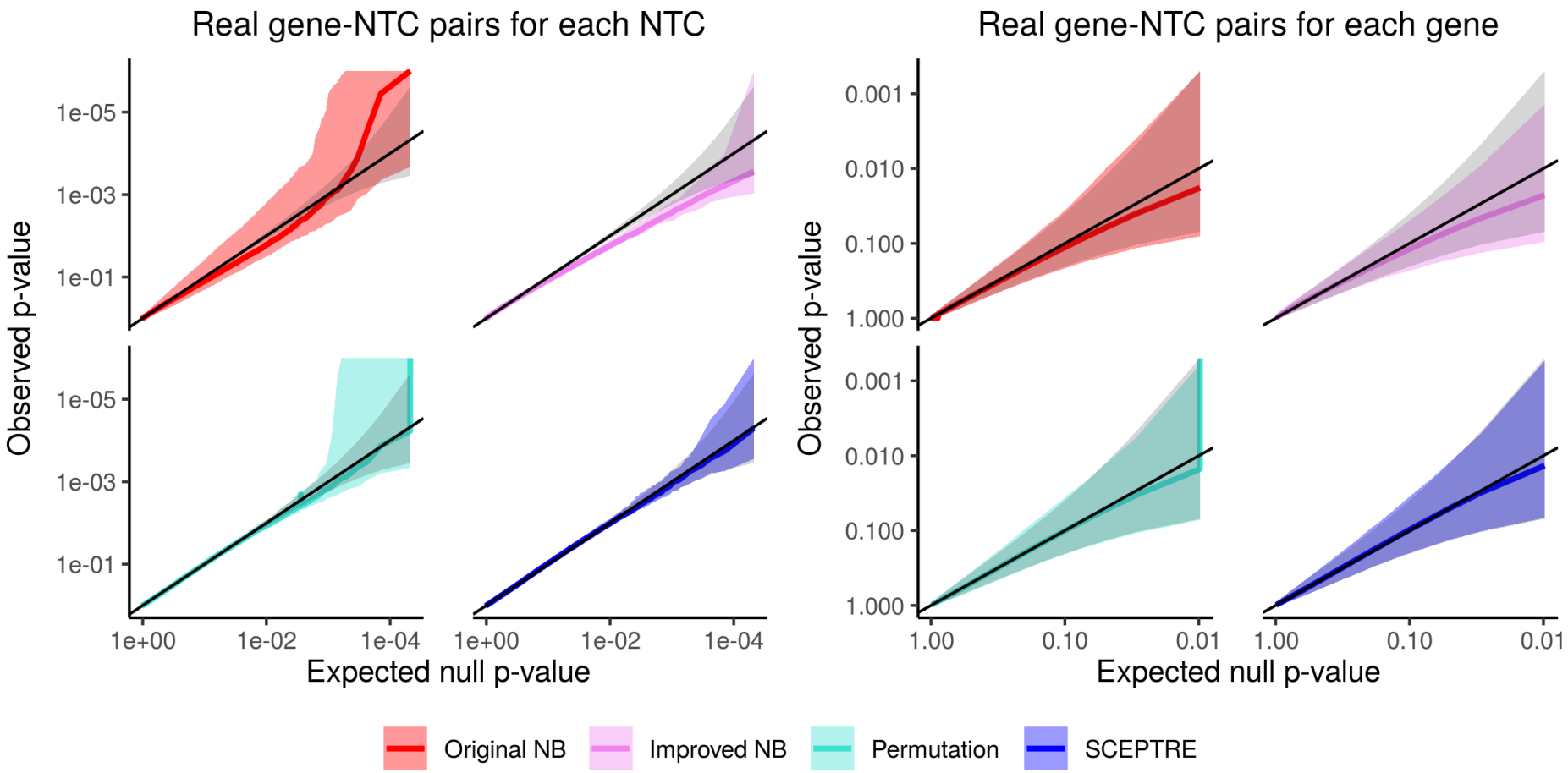
- Single cell CRISPR screens open amazing scientific opportunities but also present new statistical challenges
- We propose a new paradigm for calibrating any test statistic without relying on validity of gene expression model
- Improved statistical methodology yields new, biologically relevant regulatory relationships
- Many more statistical challenges remain, e.g. accounting for variable gRNA effectiveness and interactions among enhancers

Conclusions

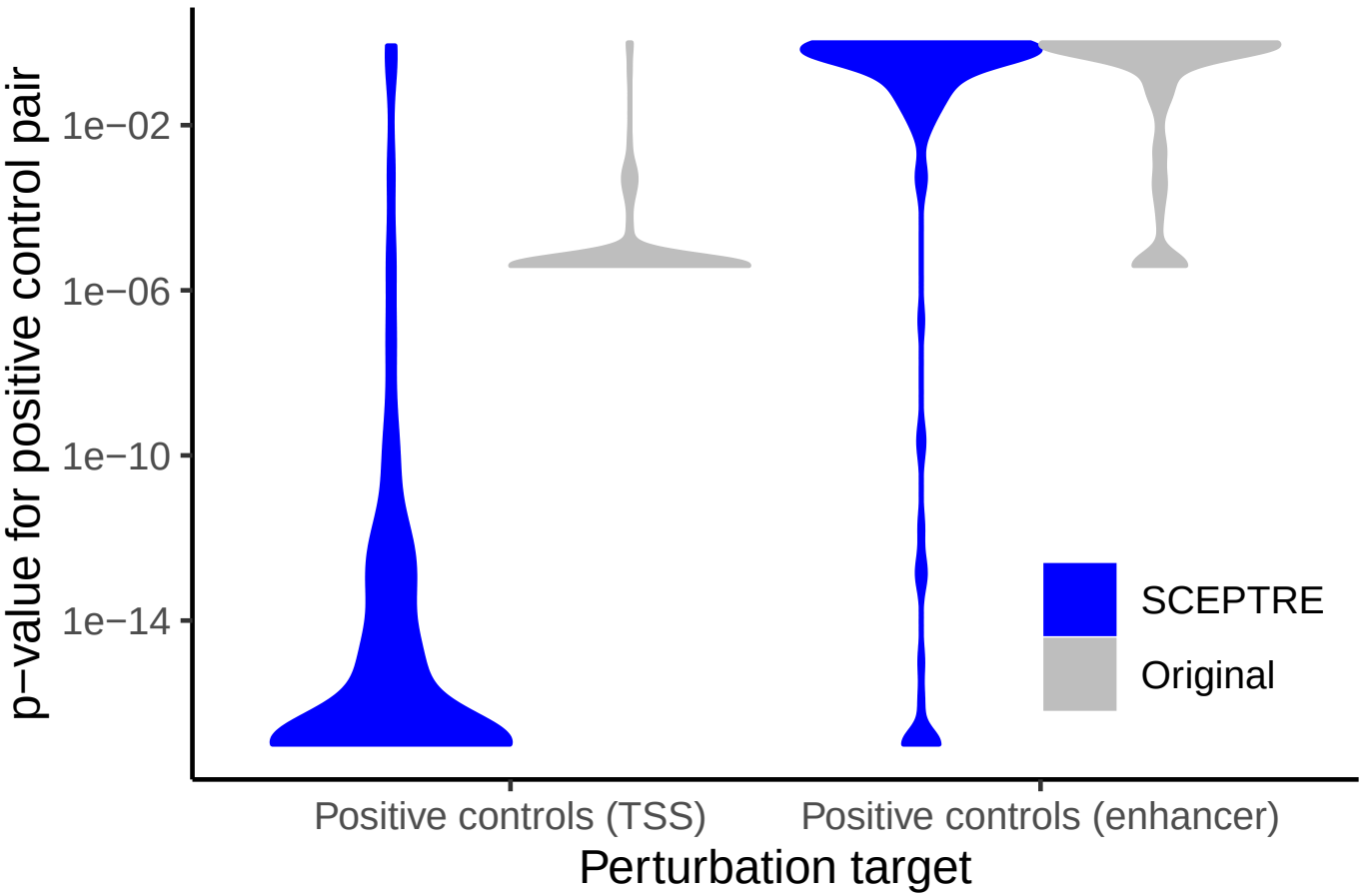
- Single cell CRISPR screens open amazing scientific opportunities but also present new statistical challenges
- We propose a new paradigm for calibrating any test statistic without relying on validity of gene expression model
- Improved statistical methodology yields new, biologically relevant regulatory relationships
- Many more statistical challenges remain, e.g. accounting for variable gRNA effectiveness and interactions among enhancers

Sincere thanks to Jay for having us, to Molly and Jacob for your help and patience, to all Shendure lab members for your feedback!

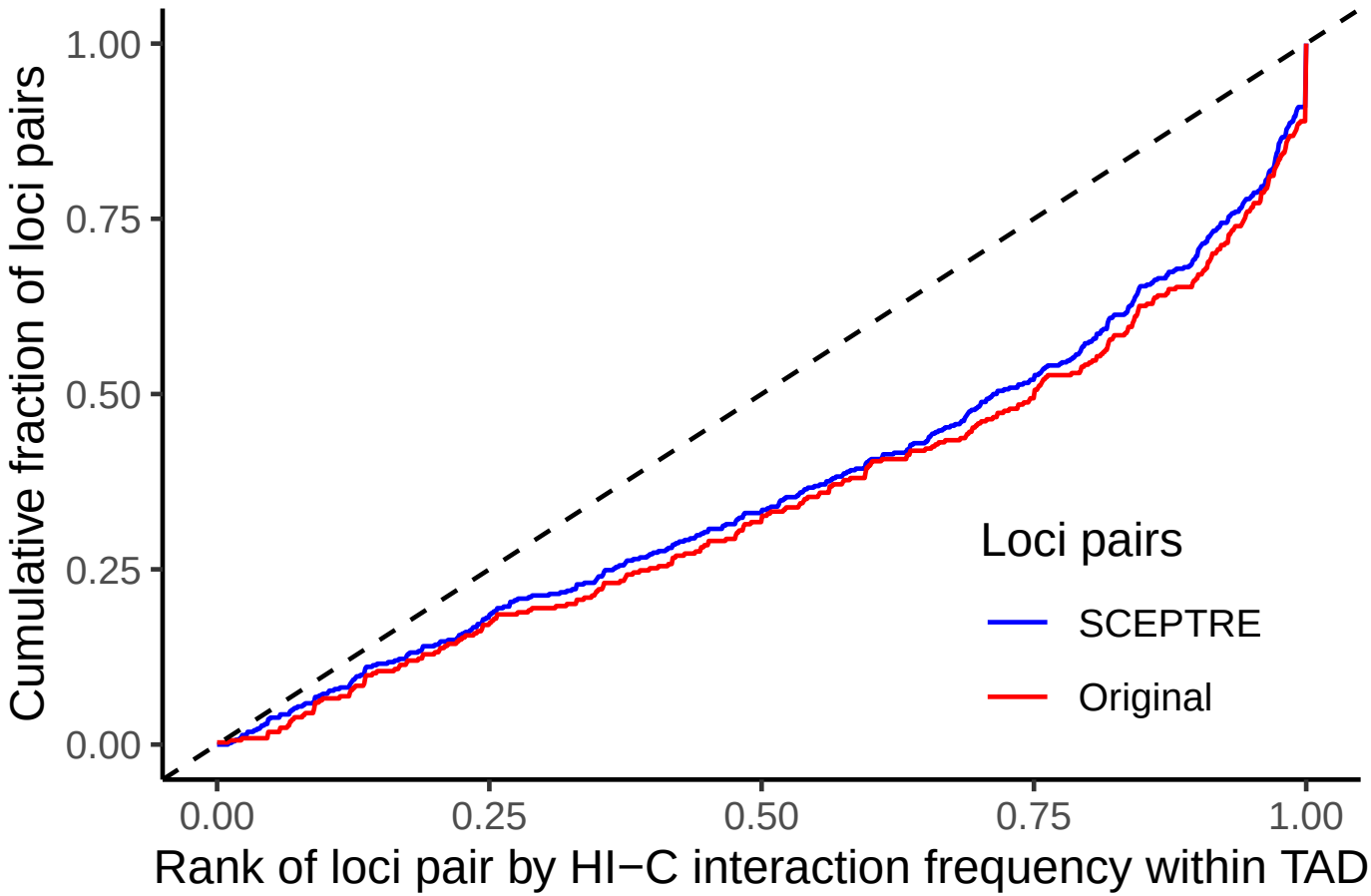
Calibration by negative control gRNA and by gene



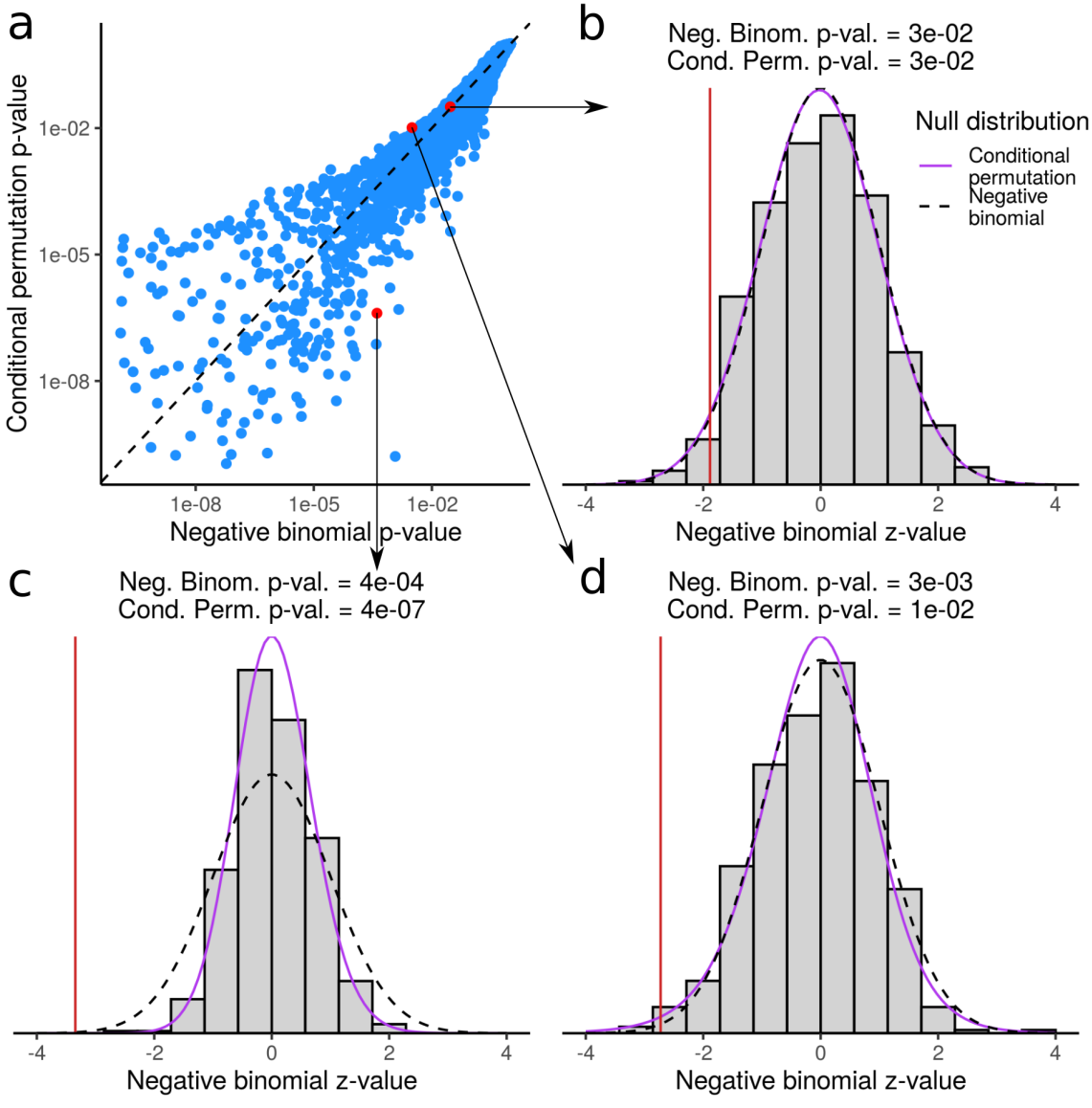
Positive controls



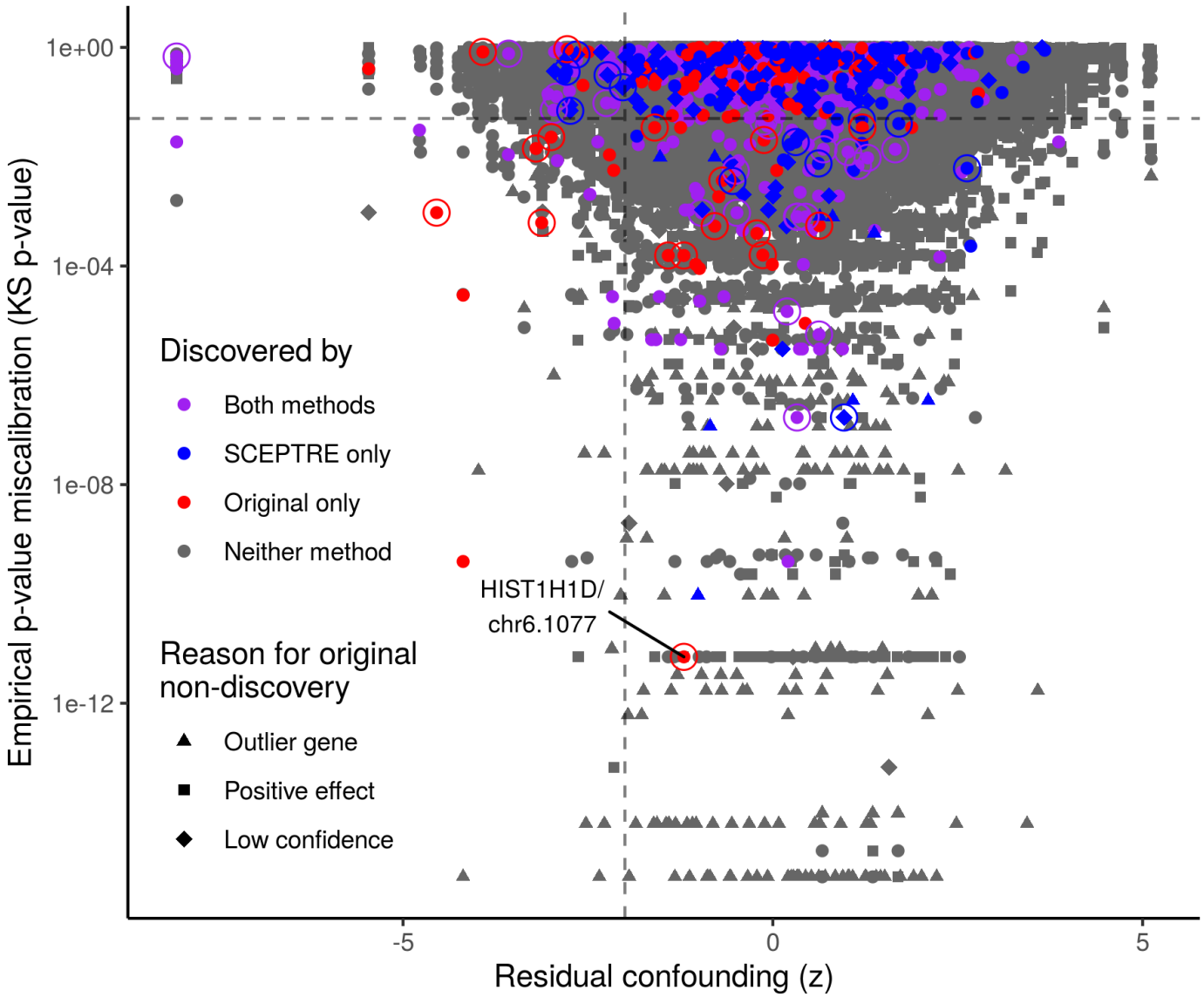
HI-C interaction frequency enrichment



Parametric vs resampling-based calibration



Potential false positives in original analysis



Details on ChIP-seq enrichment analysis