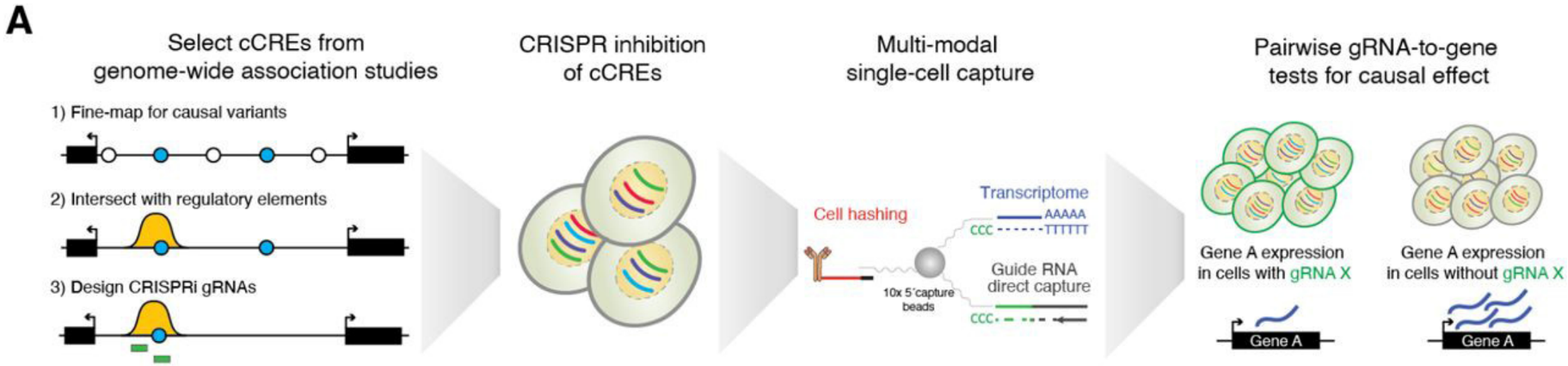
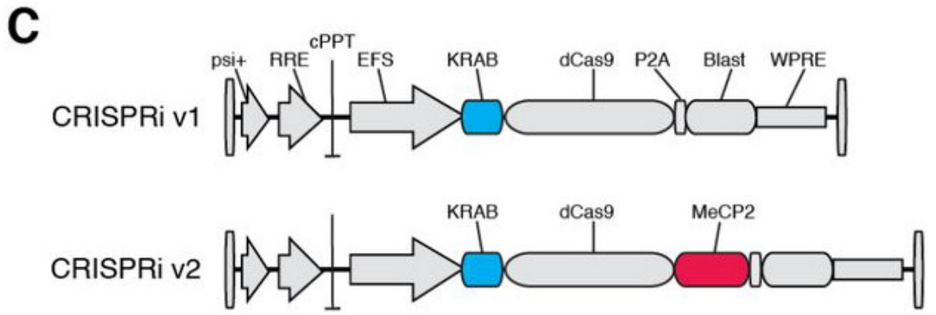
STING-seq data analysis

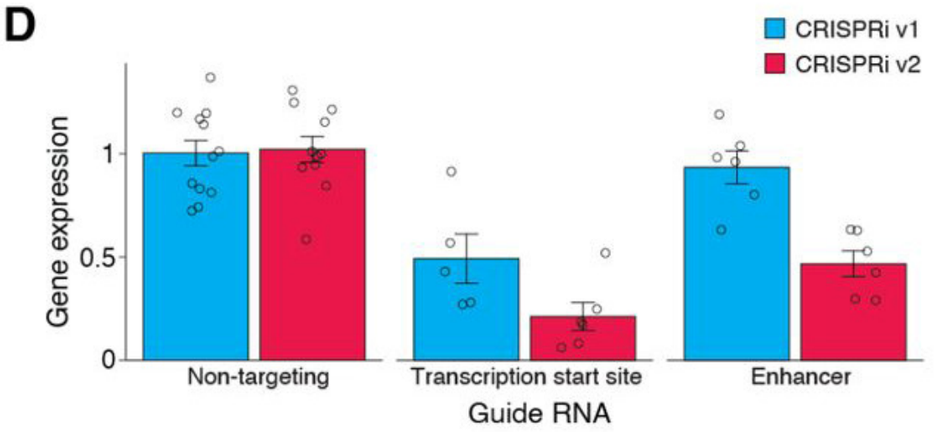
Hang Chen
8/17/2022

STING-seq concept



(c: candidate)





Morris et al., 2021

STING-seq datasets

UMI counts datasets (from Morris):

- Gene expression matrix (cDNA; gene by cell)
- gRNA expressions matrix (GDO; gRNA by cell)
- Cell surface antigens matrix (HTO; hashtag by cell)

Important parameters:

- nFeature; unique/ non-zero UMI count
- nCount; total UMI count
- Percent-mt; mitochondrial genes percentage

QC filtering

High quality and live cells:

- nFeature > 1400 UMI
- Percent-mt < 20%

Singlets:

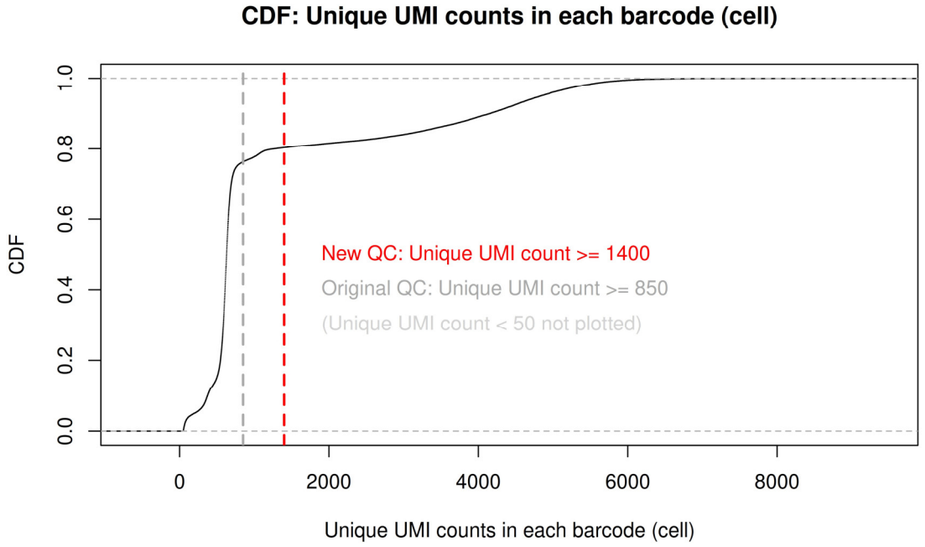
- Hashtag demuxing (Seurat)

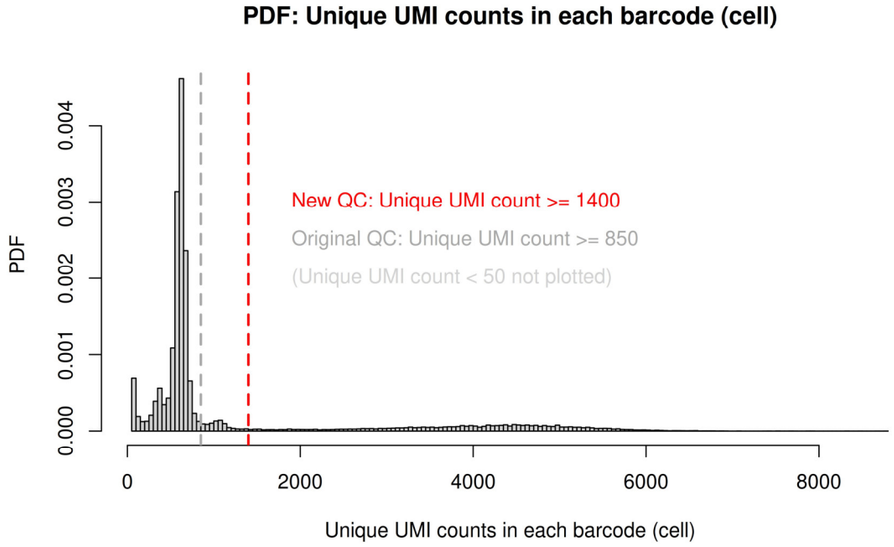
Results:

- 8,845 cells (8,588 overlapped with author’s 9,343)

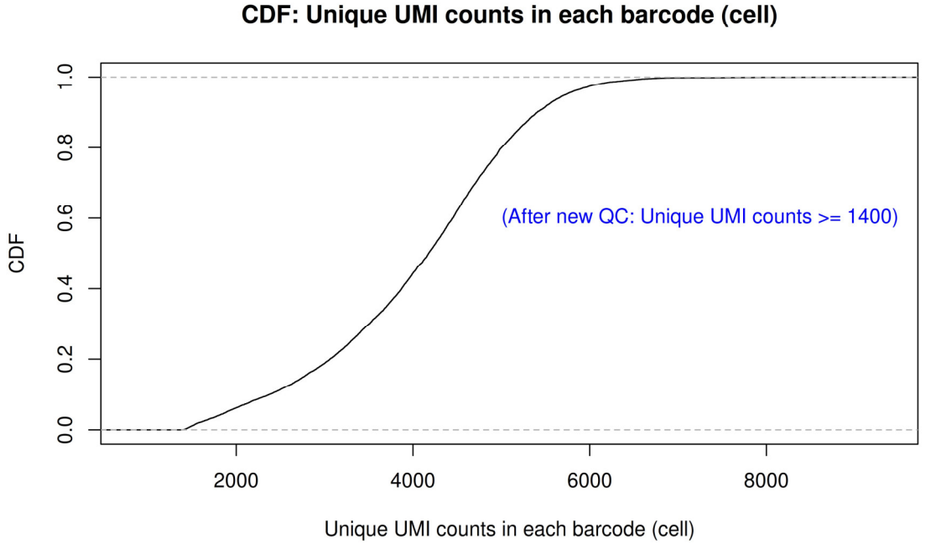
QC results (CDF/PDF)

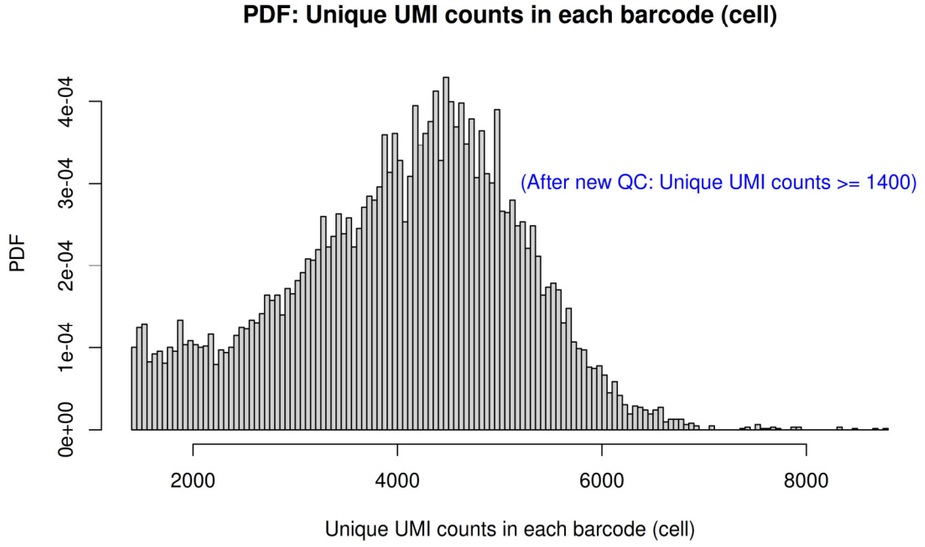
Before QC





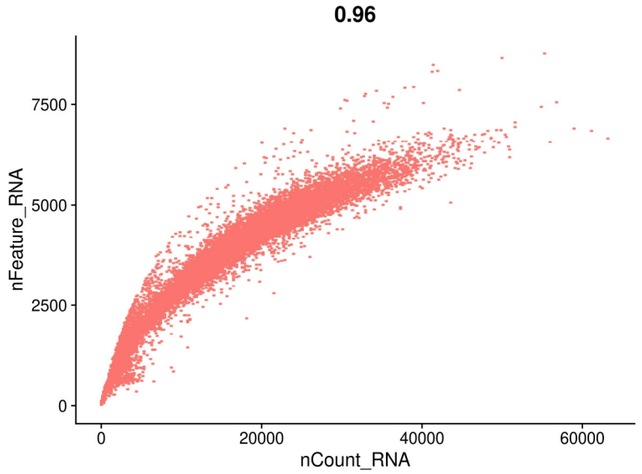
After QC

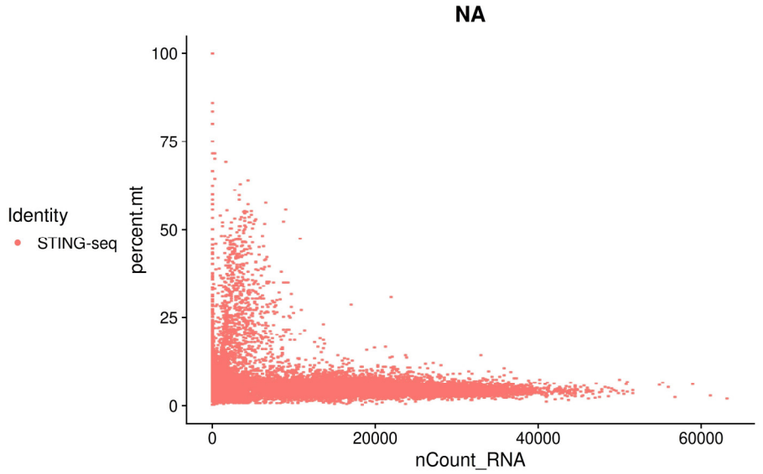


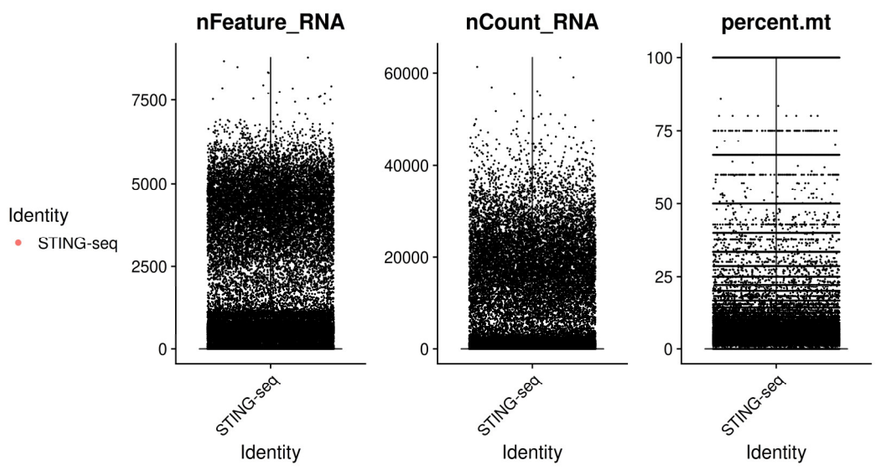


QC results (Seurat)

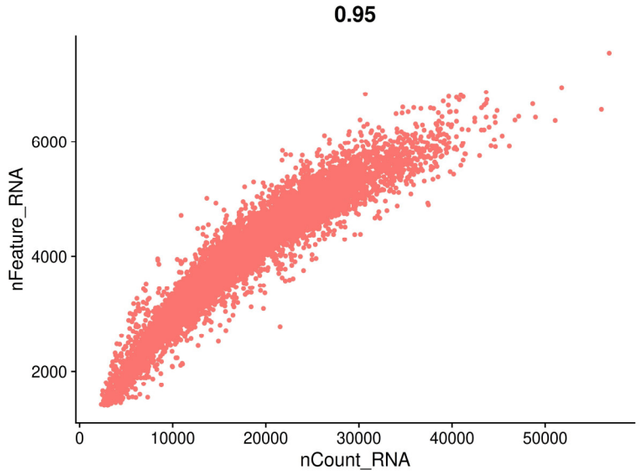
Before QC

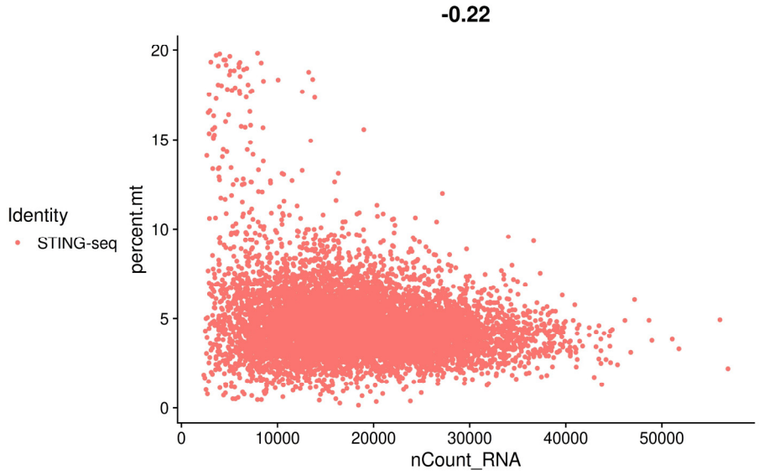


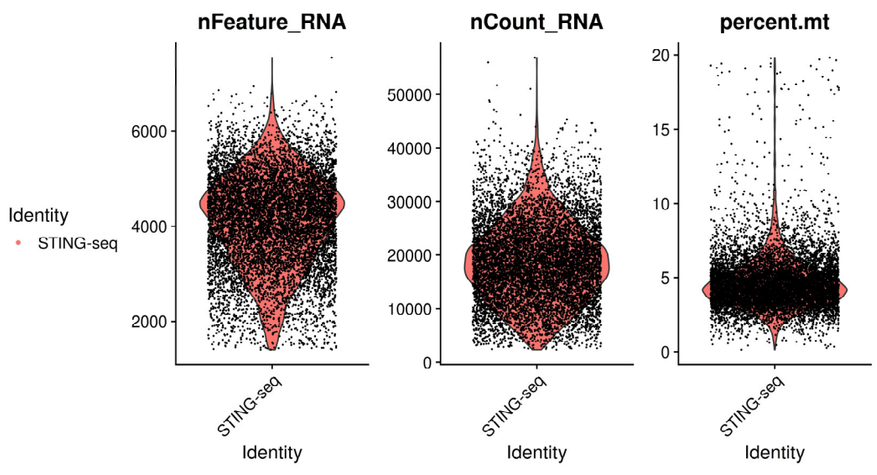




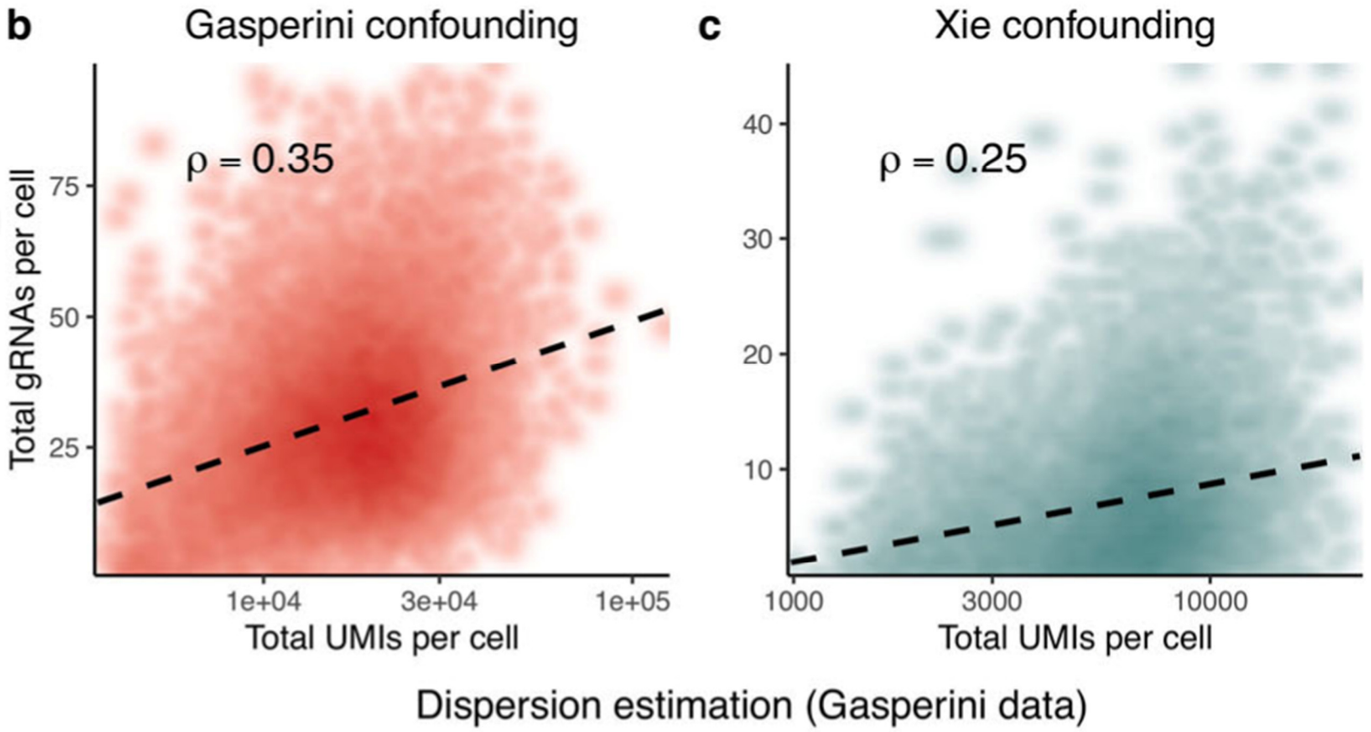
After QC







Depth: a confounder in perturb-seq



Barry et al., 2021

Gene-gRNA associations

Naive negative binomial regression:

- `gene expression ~ gRNA expression`

Negative binomial regression with total UMI as a covariate:

- `gene expression ~ gRNA expression + total UMI`

SCEPTRE:

- Negative binomial regression with covariate
- Empirical null distribution (by conditional randomization test) for *p*-value calculation

Running SCEPTRE

Input:

1. `gRNA_matrix`: gRNA by cell (QC'ed)
2. `gRNA_groups_table`: gRNA_id, gRNA_group, gRNA_type

3. `gene_matrix`: gene by cell (QC'ed)
4. `gene_gRNA_group_pairs`: gene_id, gRNA_group, pair_type
5. `covariate_matrix`: lg_gRNA_lib_size, lg_gene_lib_size, p_mito, batch

Intermediate:

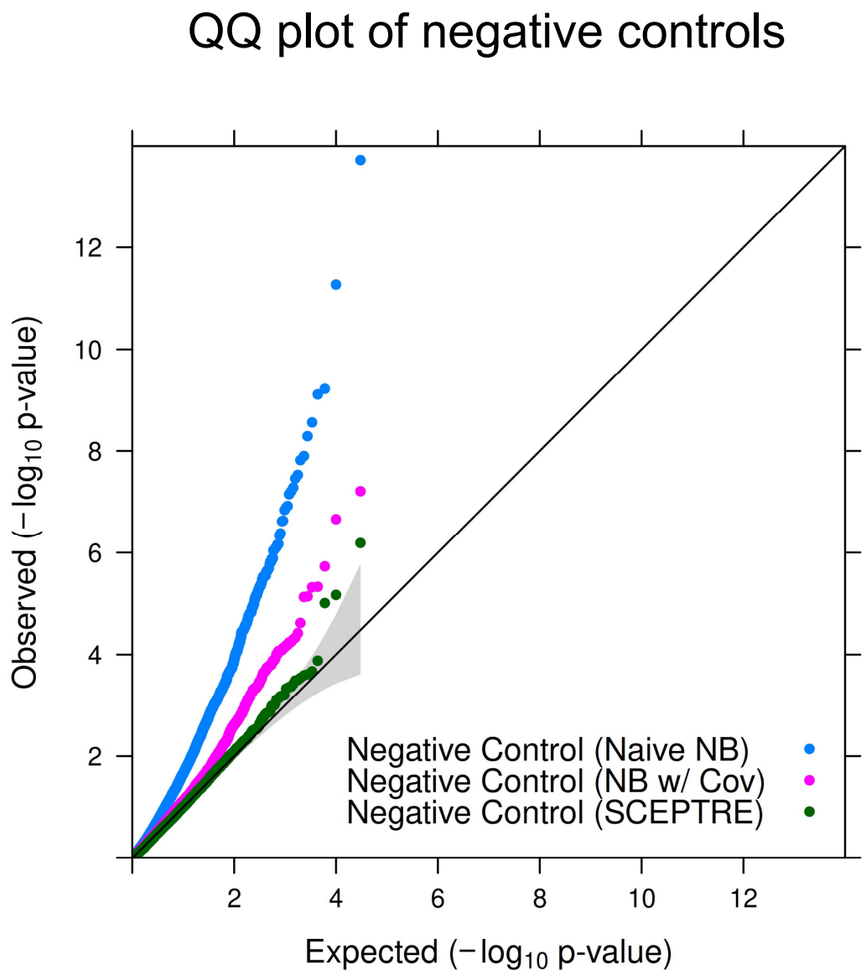
a. `perturbation_matrix`: from 1; gRNA_id by cell (binary)

b. `combined_perturbation_matrix`: from a and 2; gRNA_group by cell (binary)

Output:

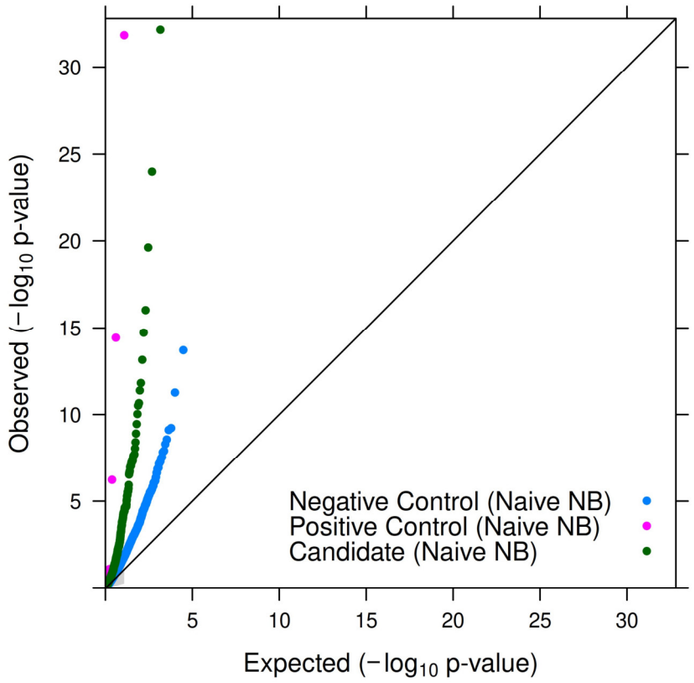
1. `result`: from 3, 4, 5, and b

Negative controls

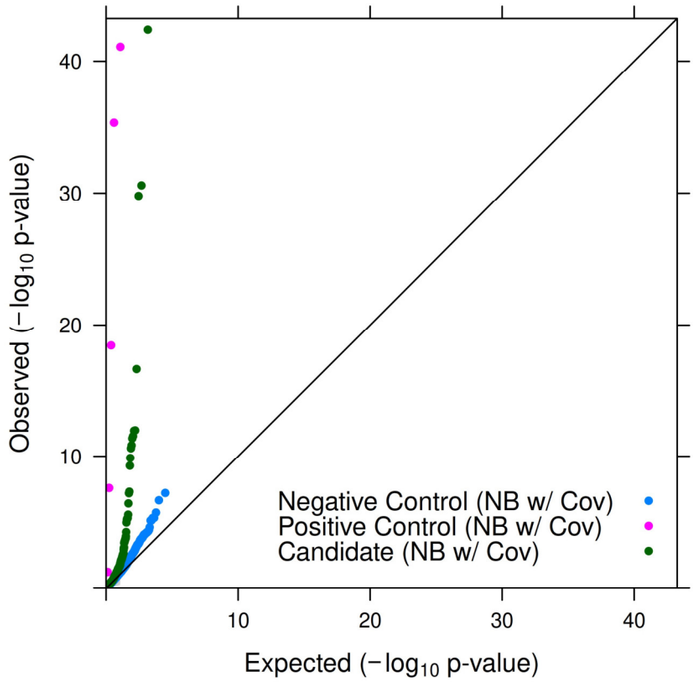


SCEPTRE results

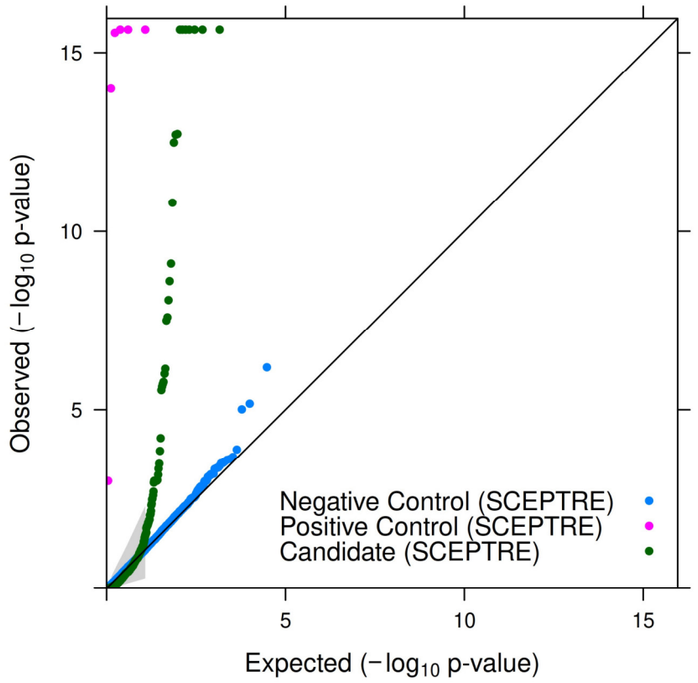
QQ plot of all gene-gRNA pairs



Naive NB



NB with Cov (depth)



SCEPTRE

SCEPTRE results

Significant gene-gRNA pairs (all overlapped with Morris' results)

Gene	cCRE SNP (gRNA target)	gRNA type	*p*-value	*Z*-value	*Log* fold change
EGFL7	rs62581550	candidate	2.44E-09	-6.13289	-0.14633
TMCO3	rs117066769	candidate	2.00E-13	-8.22509	-0.39669
ANK1	rs4737009	candidate	2.22E-16	-8.07873	-0.17905
CD52	rs4526602	candidate	2.22E-16	-11.6964	-0.76756
TFRC	rs9325434	candidate	2.79E-06	-3.97773	-0.10379
SLC7A5	rs7200918	candidate	2.22E-16	-6.44869	-0.28934
GFI1B	rs524137	candidate	1.92E-13	-9.34527	-0.37346
GFI1B	rs73660574	candidate	2.26E-06	-4.37964	-0.19432
PTPRC	rs1926231	candidate	2.22E-16	-14.3512	-1.10647
CD164	rs1546722	candidate	8.40E-09	-5.37675	-0.22988
GLRX5	rs35362007	candidate	2.22E-16	-8.45271	-0.415
ANKRD12	rs527890099	candidate	1.89E-06	-4.36787	-0.16145
PTPRC	rs1326279	candidate	1.58E-11	-6.41345	-0.36157
PDLIM1	rs75522380	candidate	2.22E-16	-11.8661	-0.64246
NUDT4	rs4761702	candidate	7.11E-07	-4.84094	-0.3579
CR1L	1:207821519_AACAC_A	candidate	2.22E-16	-9.845	-0.37908
PTPRC	rs1326270	candidate	3.34E-13	-6.11366	-0.2755
PIEZO1	rs10445033	candidate	7.95E-10	-7.22128	-0.35862
PTPRC	1:198630156_CCA_C	candidate	9.73E-07	-4.72977	-0.3396
TFRC	rs11712192	candidate	1.64E-06	-4.42424	-0.20131
PTPRC	rs1998843	candidate	3.08E-08	-5.40019	-0.39422
MAZ	rs34286592	candidate	2.48E-08	-5.04072	-0.33323

Some experiences and thoughts…

- Sparse matrices: better suited for scRNA-seq data, but manipulation is slightly different from dense matrices
- Loops in R: could be slow, may be replaced by C++ (`for` loops not so slow anymore comparing with `apply` nowadays)
- Scripts/packages: Unix philosophy is important, otherwise painful to debug
- Formatting: some datasets are tibbles, some columns are factors, important to check and format carefully
- Batches and other covariates: if not provided, may be inferred via PCA

Thank you!