Quality Control Pipeline for Massively Parallel Reporter Assays (MPRAs)

Omer Ronen

2025-09-30

Contents

Overview **5**

Usage 5

Scripts 6

1 A guide for running the analyses notebook **7**

1.1 Associations 7

1.2 Activity 7

1.3 Differential Activity 8

2 Associations QC **9**

2.1 Barcodes per cCRE 9

2.2 PCR bias - GC content 10

2.3 PCR bias - G Stretch 11

2.4 cCRE-barcode observations 13

2.5 Retained cCREs 14

2.6 Barcode promiscuity 15

2.7 Downsampling - Retained cCREs 16

2.8 Downsampling - Barcodes per cCRE 17

3 Activity QC **19**

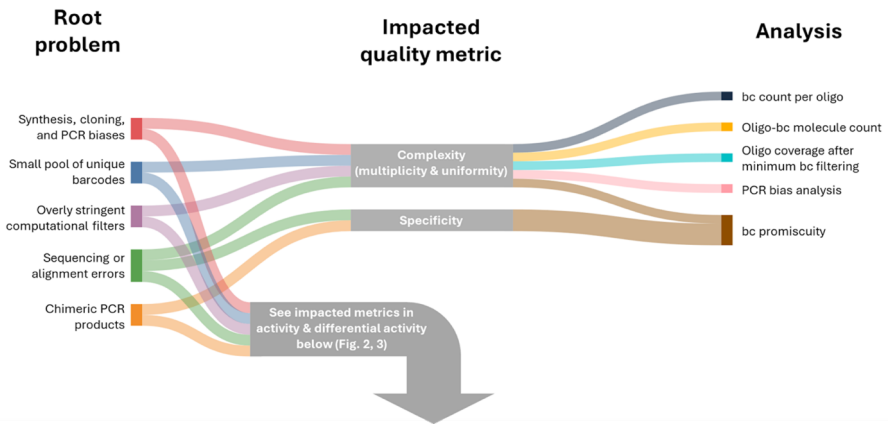
3.1 Evaluating DNA and RNA complexity 19

3.2 Evaluating reproducibility 23

Overview

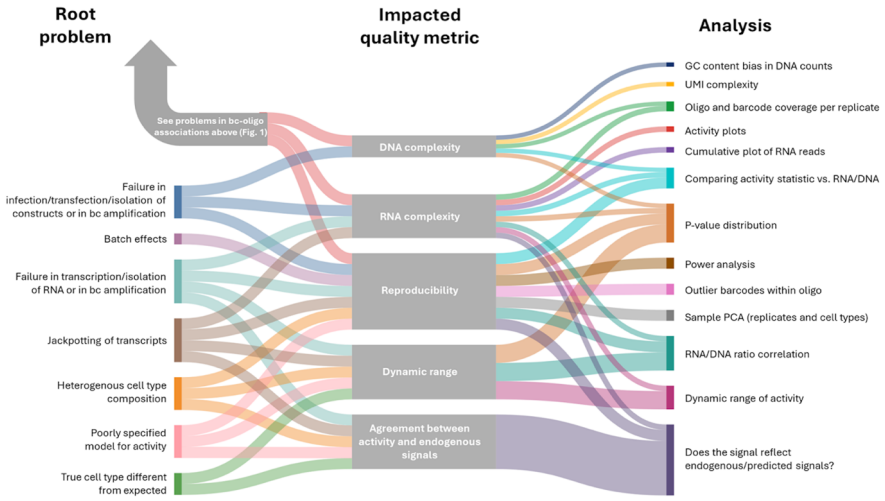
This pipeline is designed to help investigators evaluate the quality of their MPRA, quickly identify pitfalls, trace them to their source, and mitigate them. The scripts provided help ensure that the resulting MPRA data are suitable for robust statistical analysis and meaningful biological interpretation. This Bookdown accompanies our guide for best practices for MPRAs, which outlines recommendations for study design and interpretation [REF]. The manuscript covers all key experimental and analytical steps, including library design, and estimation of activity differential activity. It then describes core problems that often compromise MPRA quality, illustrating how these issues manifest in the data, and offering practical strategies for correction and optimization. Because each issue can influence multiple quality metrics, and each metric may be affected by several issues, the relationships form a many-to-many network. The figures presented below map these interdependencies and connect them to recommended diagnostic analyses.

Usage



A scheme of root problems, the impacted quality metrics and analyses for the

cCRE-barcode association step.



Root problems, impacted quality metrics and recommended analyses for the RNA and DNA quantification step.

The quality control (QC) pipeline is organized into three chapters:

- (i) QC of the barcode association step
- (ii) QC of activity estimation
- (iii) QC of differential activity estimation

For each analysis, we provide an example of a successful and an unsuccessful dataset to illustrate how they manifest in the analysis.

We welcome questions, feedback, or suggestions. Please feel free to reach out at david.gokhman [at] weizmann.ac.il.

Scripts

All of these analyses are integrated into the quality control pipeline described in this resource, with scripts provided here: [link].

Chapter 1

A guide for running the analyses notebook

The QC pipeline has three main parts: Associations, Activity and Differential activity. For each part, there is a jupyter notebook file that enables you to run all the analyses that are presented in this book. Here we explain how to run these notebook files and what are the required inputs

1.1 Associations

First

1.1.1 Input

Input files

1.2 Activity

Second

1.2.1 Input

Input files

1.3 Differential Activity

Third

1.3.1 Input

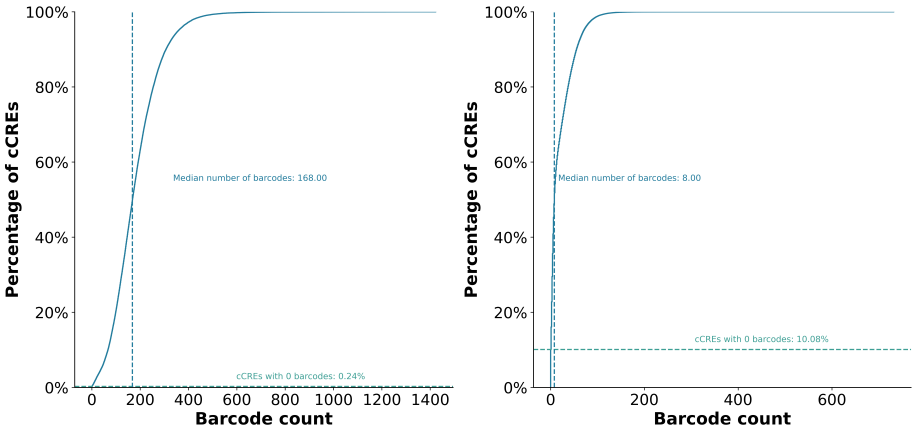
Input files

Chapter 2

Associations QC

2.1 Barcodes per cCRE

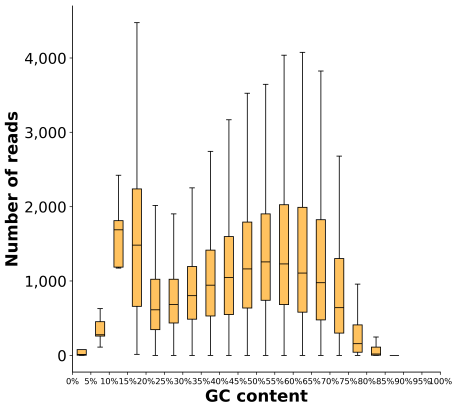
Goal: Input file: Evaluated metrics:



Legend: Interpretation:

2.2 PCR bias - GC content

2.2.1 Reads per GC bin

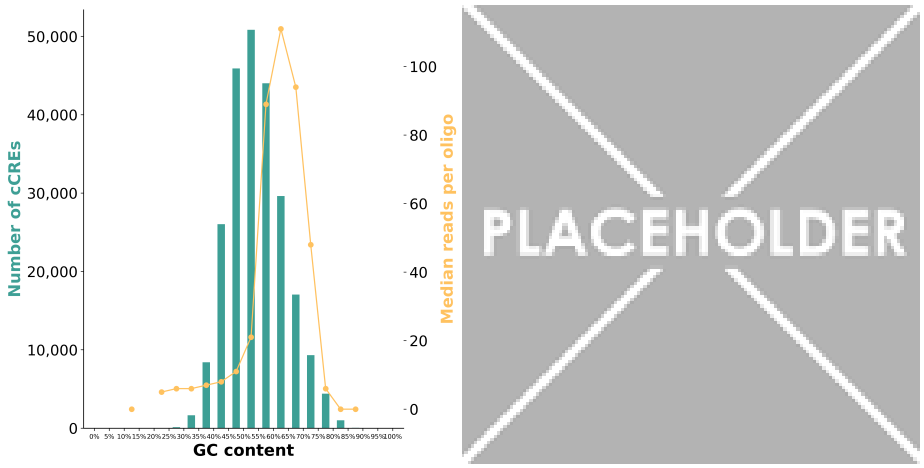


Goal: Input file: Evaluated metrics:

Legend: Interpretation:

2.2.2 cCRE counts

Goal: Input file: Evaluated metrics:

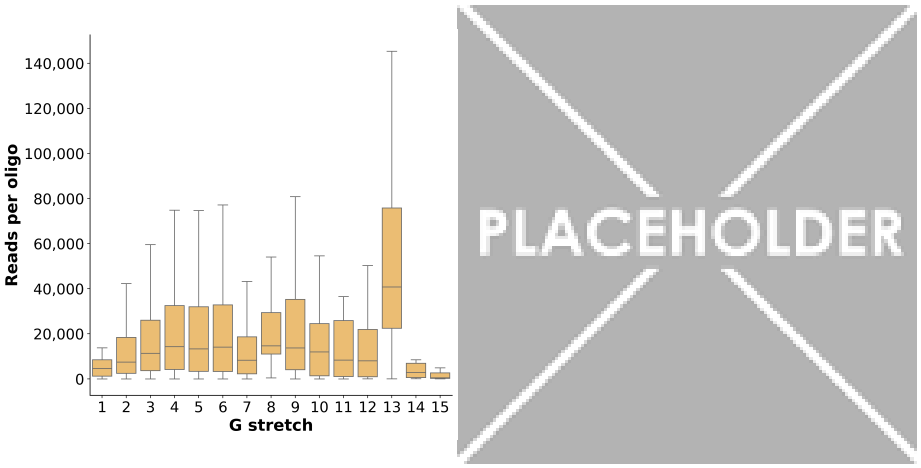


Legend: Interpretation:

2.3 PCR bias - G Stretch

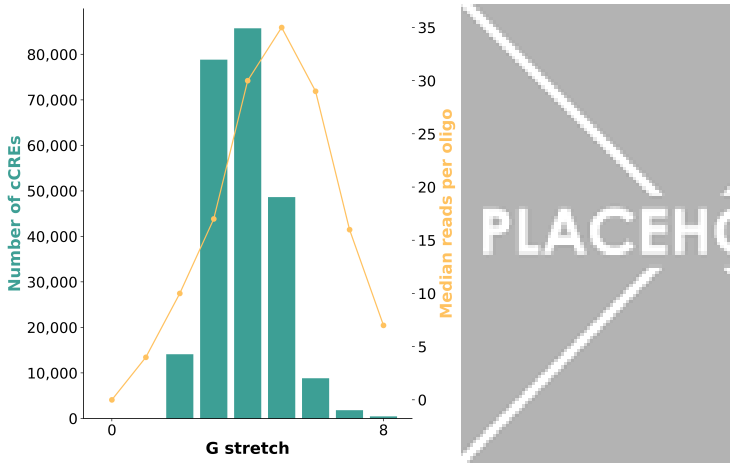
2.3.1 Reads per G stretch bin

Goal: Input file: Evaluated metrics:



Legend: Interpretation:

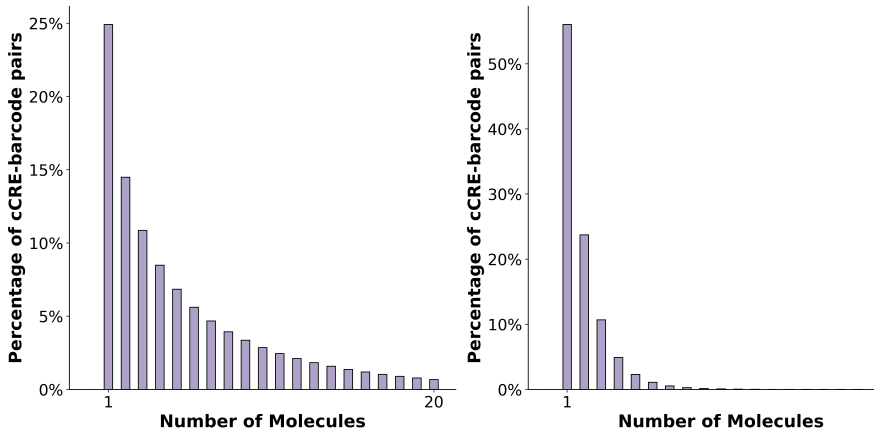
2.3.2 cCRE counts



Goal: Input file: Evaluated metrics:

Legend: Interpretation:

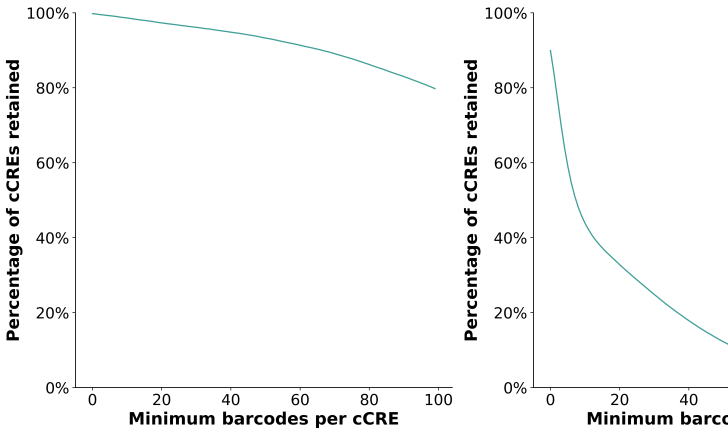
2.4 cCRE-barcode observations



Goal: Input file: Evaluated metrics:

Legend: Interpretation:

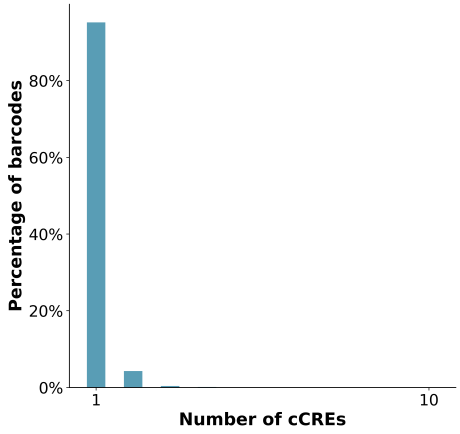
2.5 Retained cCREs

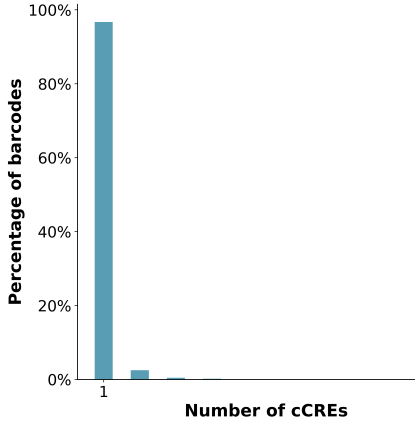


Goal: Input file: Evaluated metrics:

Legend: Interpretation:

2.6 Barcode promiscuity

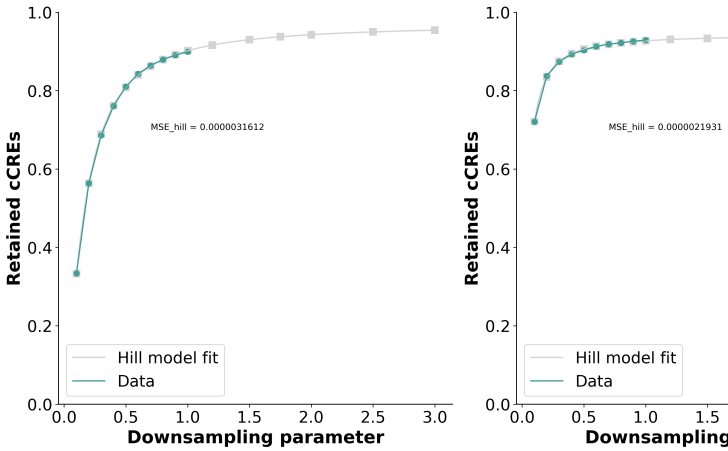




Goal: Input file: Evaluated metrics:

Legend: Interpretation:

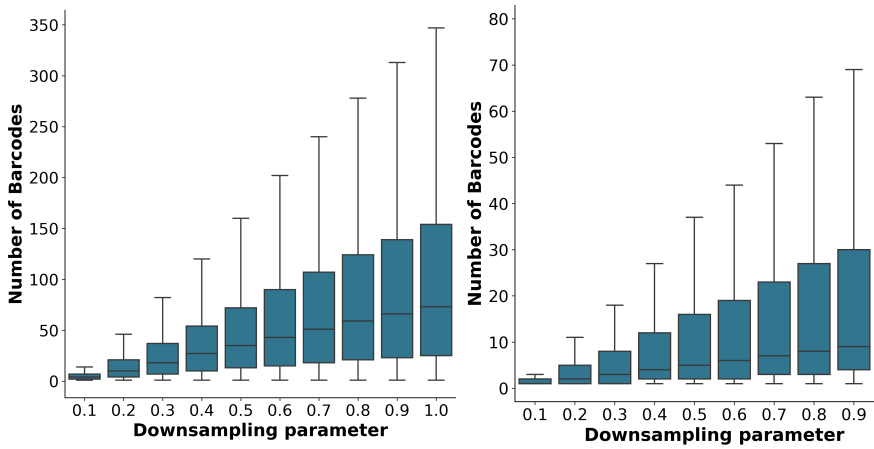
2.7 Downsampling - Retained cCREs



Goal: Input file: Evaluated metrics:

Legend: Interpretation:

2.8 Downsampling - Barcodes per cCRE



Goal: Input file: Evaluated metrics:

Legend: Interpretation:

Chapter 3

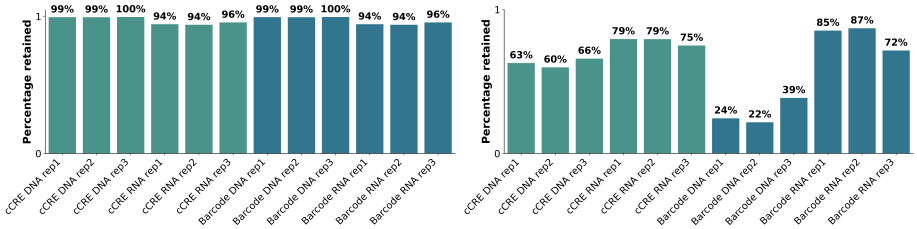
Activity QC

3.1 Evaluating DNA and RNA complexity

3.1.1 Retained cCREs and barcodes

Goal: Input file: Evaluated metrics:

```
## Good example: PMID_38766054_Reilly
## Bad example: Max_MPRA_run2
```

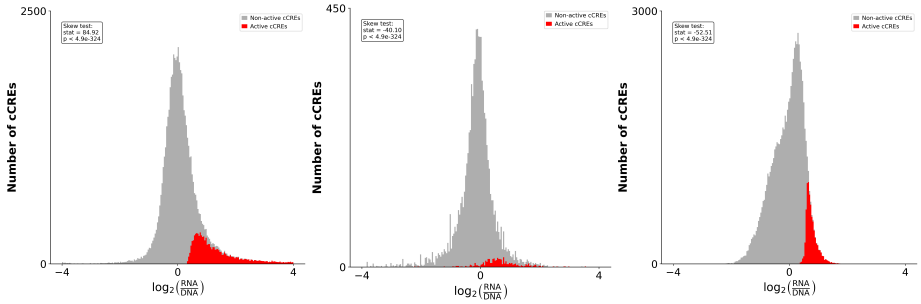


Legend: Interpretation:

3.1.2 Activity distribution

Goal: Input file: Evaluated metrics:

```
## Good example: PMID_38766054_Reilly
## Bad example: humanMPRA_L4a2
## Bad example 2: humanMPRA_L1a1_Neurons
```




```
## [1] "add arrows that indicate right tail, symmetry, or no activity detected"
```

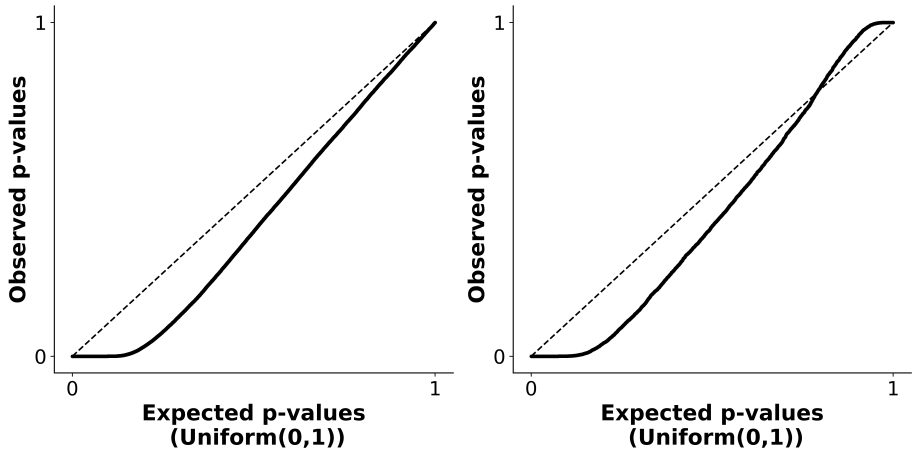
Legend: Interpretation:

3.1.3 P-value distribution

Goal: Input file: Evaluated metrics:

Problem: nothitng looks mildly bad, max looks too bad.

```
## Good example: PMID_38766054_Reilly
## Bad example: humanMPRA_L4a2
```

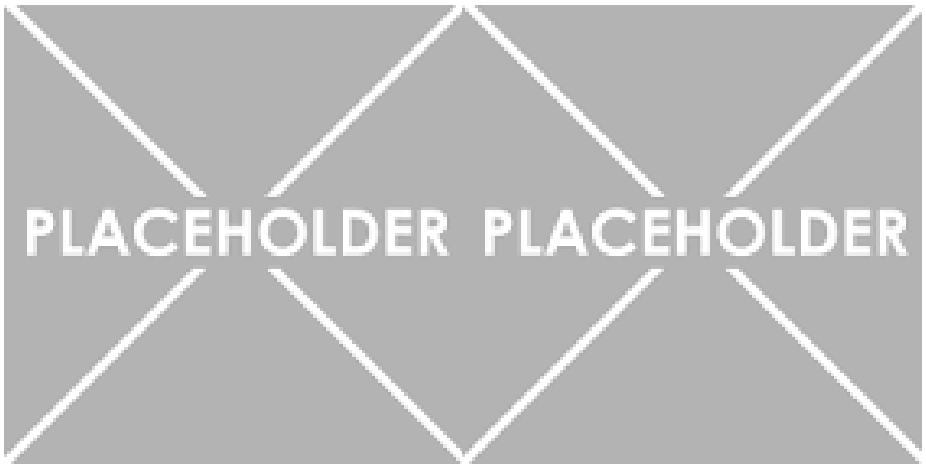


Legend: Interpretation:

3.1.4 Downsampling analysis - active cCREs

Goal: Input file: Evaluated metrics:

we should use a real downsampling - Omer is in charge of that. In the bookdown we need to mention Max's script. for Max's script - we should ask why there's a jump between the last and one-before-last downsampling. send him an email. Also ask what does "LP complexity" mean in his script.



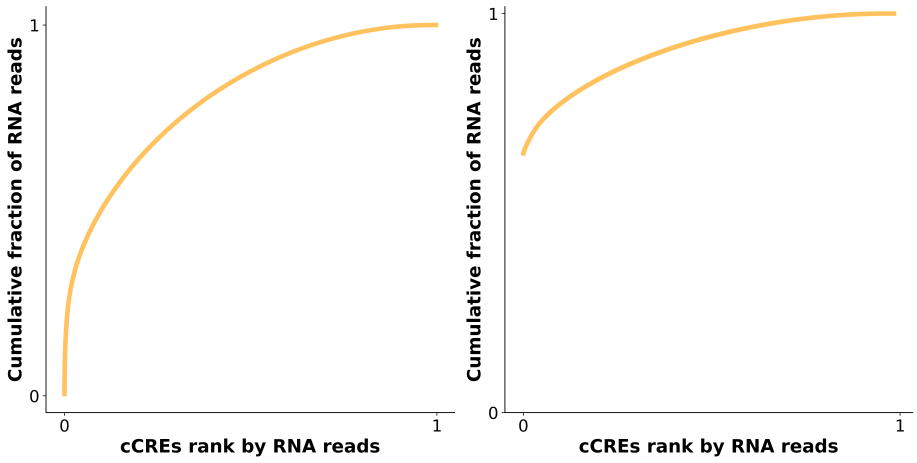
Legend: Interpretation:

3.1.5 Cumulative RNA reads

Goal: Input file: Evaluated metrics:

Add arrows in the x axis and below it “decreasing RNA reads” in illustrator.

```
## Good example: PMID_38766054_Reilly
## Bad example: d2Osteoblast_spiking_oligos
```



Legend: Interpretation:

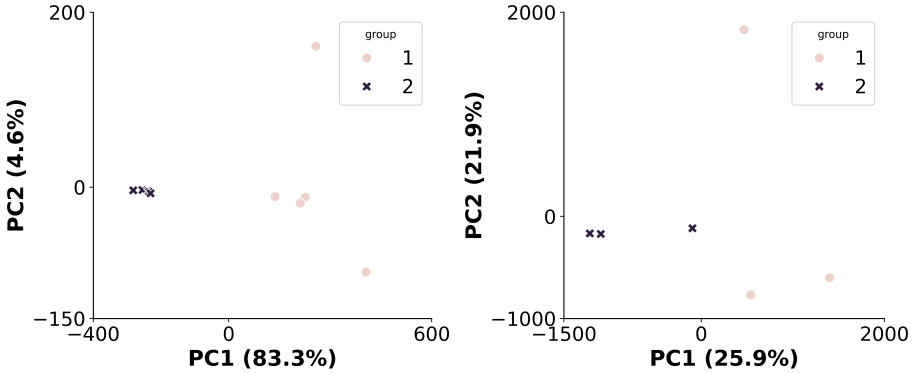
3.2 Evaluating reproducibility

3.2.1 Similarity between samples (PCA)

Goal: Input file: Evaluated metrics:

mention in the bookdown: the importance of the percentage explained by the 1st and 2nd PCs.

```
## Good example: PMID_38766054_Reilly
## Bad example: thylacine_biorxiv_Gallego_Romero
```

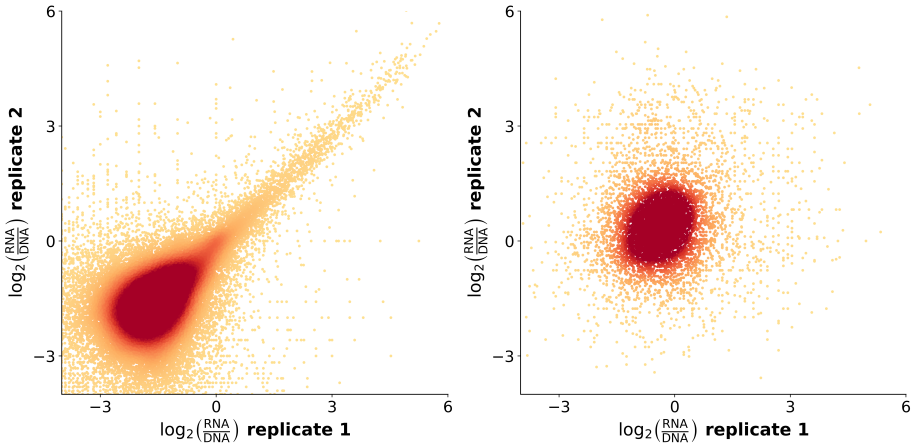


Legend: Interpretation:

3.2.2 Correlation between replicates

Goal: Input file: Evaluated metrics:

```
## Good example: thylacine_biorxiv_Gallego_Romero
## Bad example: humanMPRA_L4a2
```



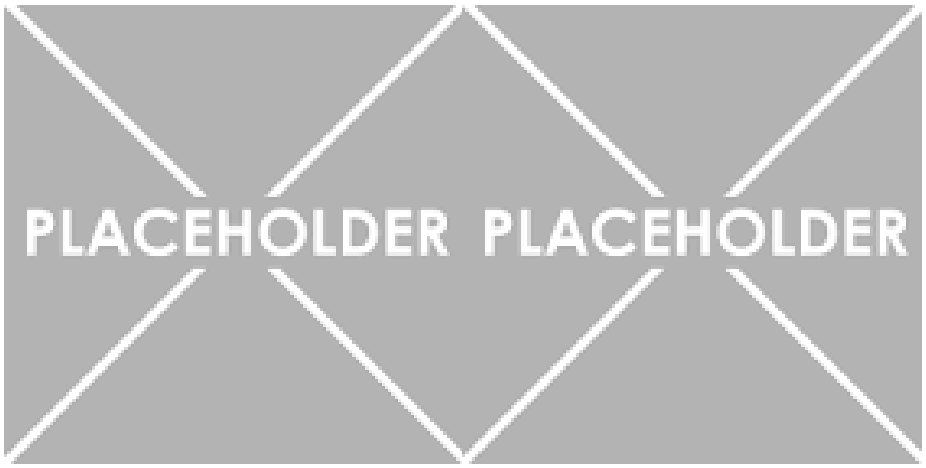

```
## Warning in rm(good_example_MPRA, bad_example_MPRA, bad_example_MPRA_2,
## analysis_name): object 'bad_example_MPRA_2' not found
```

Legend: Interpretation:

3.2.3 Variation at various activity levels

Goal: Input file: Evaluated metrics:

Omer is in charge of this part.

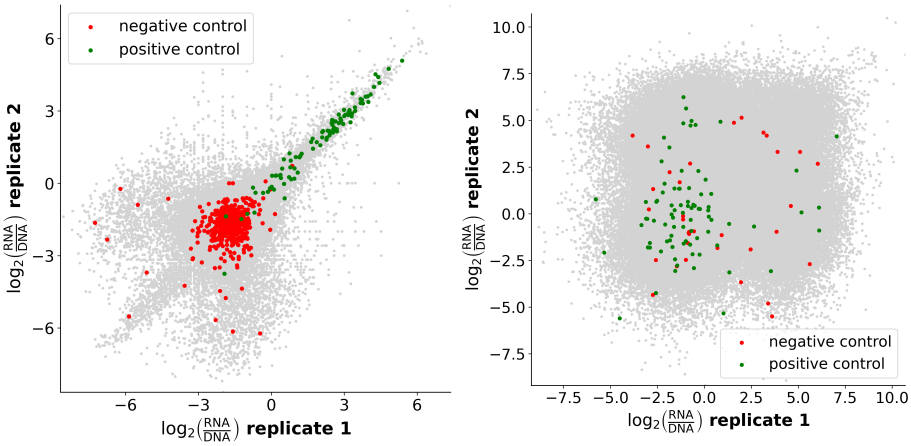


Legend: Interpretation:

3.2.4 Correlation between replicates (controls)

Goal: Input file: Evaluated metrics:

```
## Good example: PMID_38766054_Reilly
## Bad example: Max_MPRA_run2
```

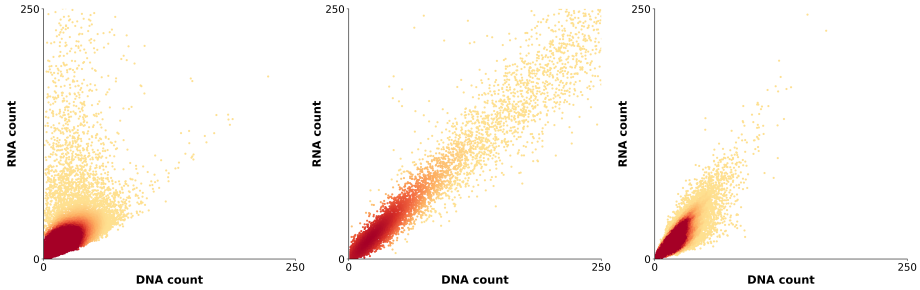


Legend: Interpretation:

3.2.5 RNA_DNA_ratio

Goal: Input file: Evaluated metrics:

```
## Good example: PMID_38766054_Reilly
## Bad example: humanMPRA_L4a2
## Bad example 2: humanMPRA_L1a1_Neurons
```

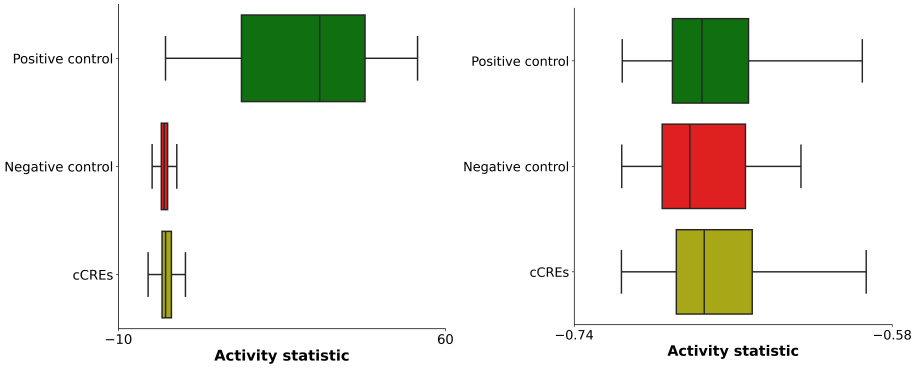


Legend: Interpretation:

3.2.6 Activity of controls - sample comparison

Goal: Input file: Evaluated metrics:

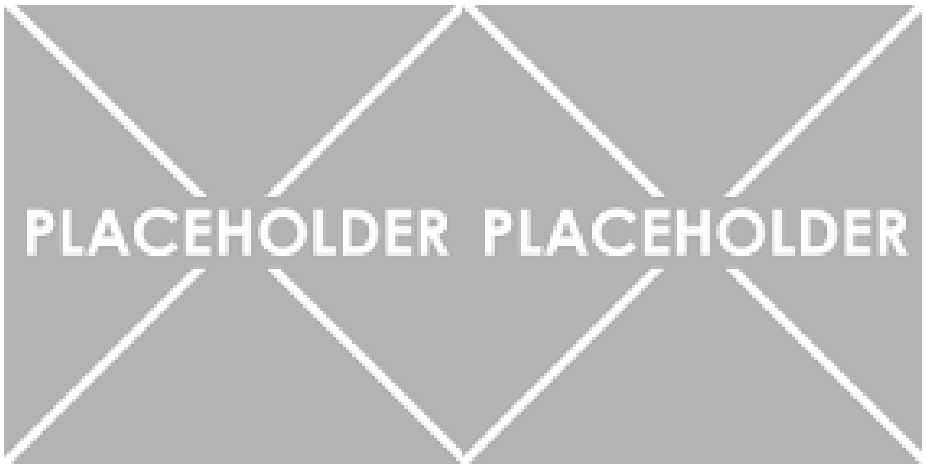
```
## Good example: PMID_38766054_Reilly
## Bad example: Max_MPRA_run2
```



Legend: Interpretation:

3.2.7 Minimizing noise [Outlier barcodes + min(DNA)] - use the mhMPRA data

Goal: Input file: Evaluated metrics:



Legend: Interpretation:

3.2.8 Outlier barcodes

Goal: Input file: Evaluated metrics:


PLACEHOLDER PLACEHOLDER