

# Quality Control Pipeline for Massively Parallel Reporter Assays (MPRAs)

Omer Ronen

2025-09-28



# Contents

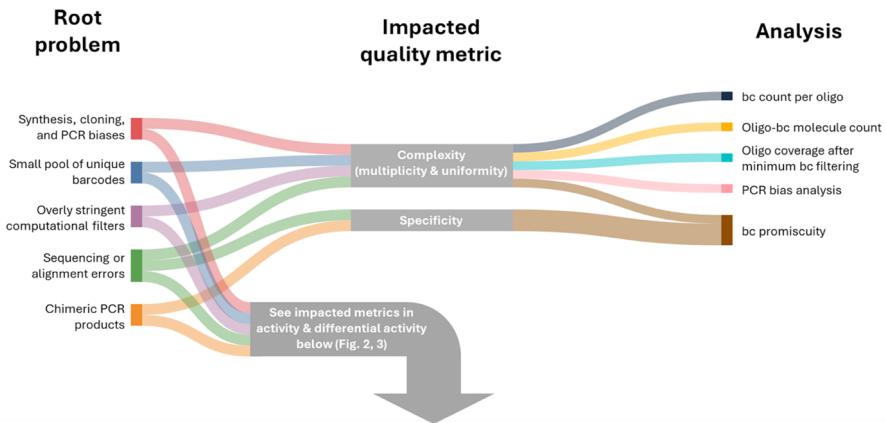
<b>Overview</b>	<b>5</b>
Usage . . . . .	5
Scripts . . . . .	6
<b>1 A guide for running the analyses notebook</b>	<b>7</b>
1.1 Associations . . . . .	7
1.2 Activity . . . . .	7
1.3 Differential Activity . . . . .	8
<b>2 Associations QC</b>	<b>9</b>
2.1 Barcodes per cCRE . . . . .	9
2.2 PCR bias - GC 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
<b>3 Activity QC</b>	<b>19</b>
3.1 Evaluating DNA and RNA complexity . . . . .	19
3.2 Evaluating reproducibility . . . . .	25



# Overview

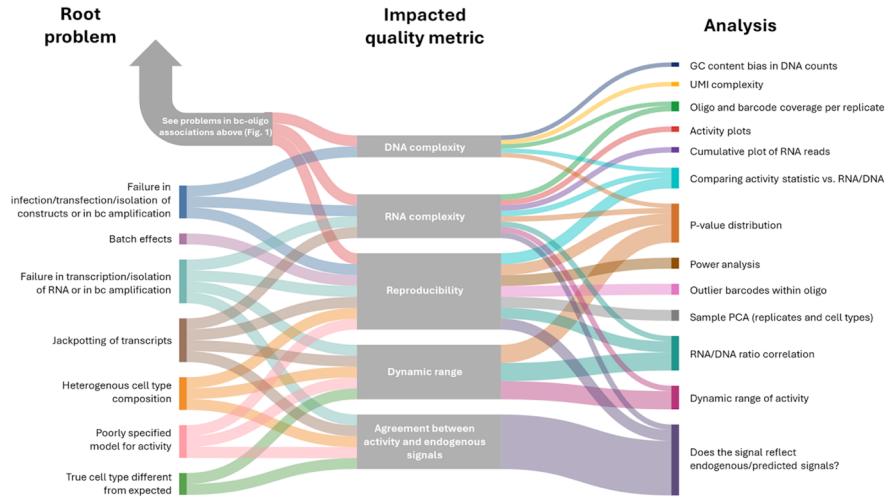
This pipeline is designed to help investigators evaluate the quality of their MPRA, quickly identify pitfalls, trace them to their source, and mitigate them. The scripts provided help ensure that the resulting MPRA data are suitable for robust statistical analysis and meaningful biological interpretation. This Bookdown accompanies our guide for best practices for MPRAAs, 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](mailto:david.gokhman@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

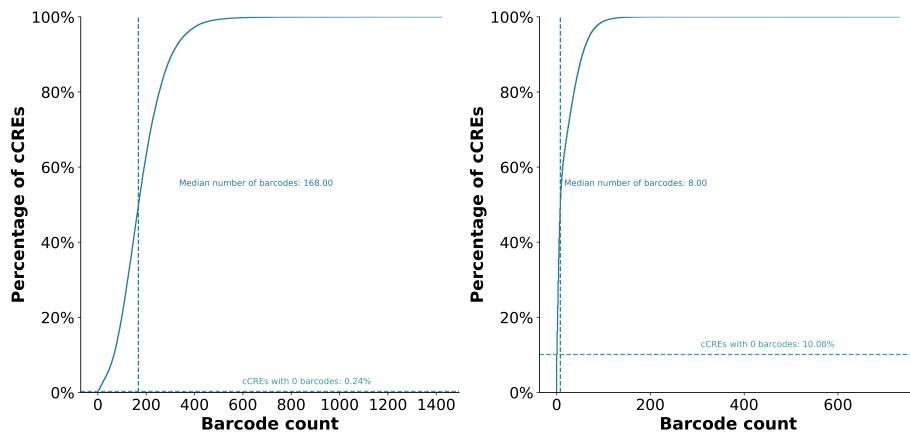
Input files

# Chapter 2

## Associations QC

### 2.1 Barcodes per cCRE

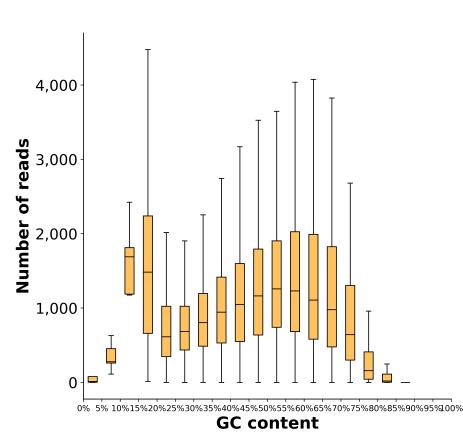
Goal: Input file: Evaluated metrics:



Legend: Interpretation:

## 2.2 PCR bias - GC content

### 2.2.1 Reads per GC bin

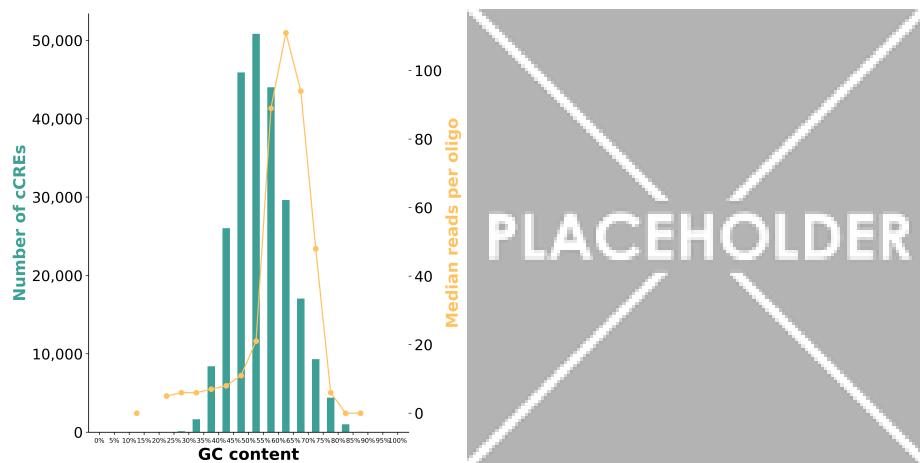


Goal: Input file: Evaluated metrics:

Legend: Interpretation:

### 2.2.2 cCRE counts

Goal: Input file: Evaluated metrics:

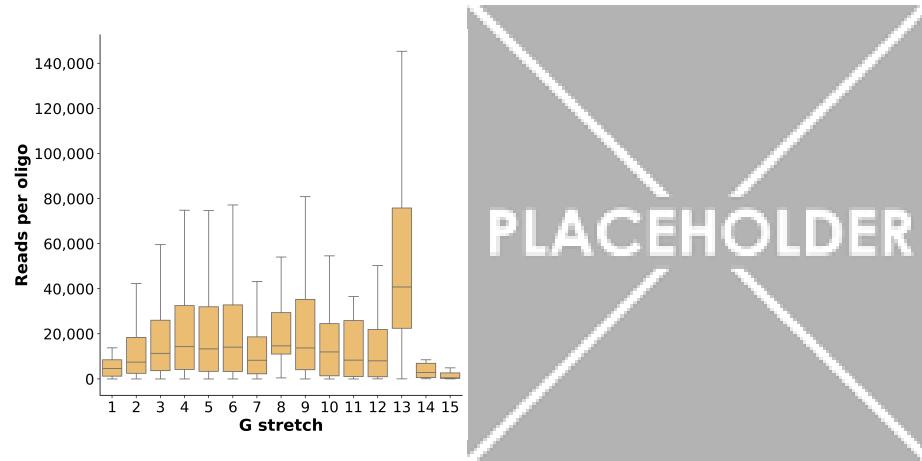


**Legend: Interpretation:**

## 2.3 PCR bias - G Stretch

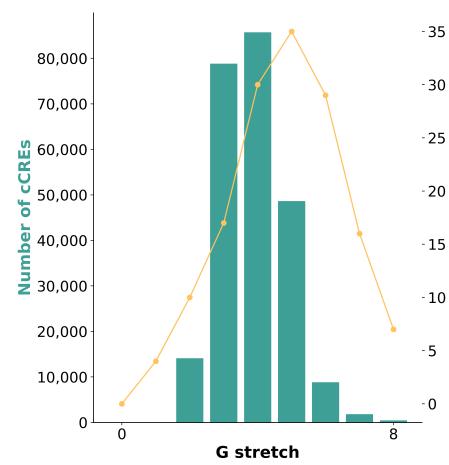
### 2.3.1 Reads per G stretch bin

**Goal: Input file: Evaluated metrics:**



Legend: Interpretation:

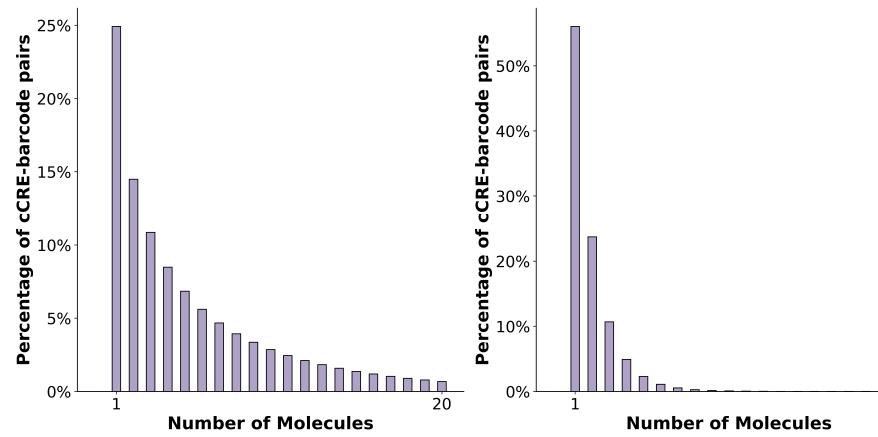
### 2.3.2 cCRE counts



Goal: Input file: Evaluated metrics:

**Legend:** Interpretation:

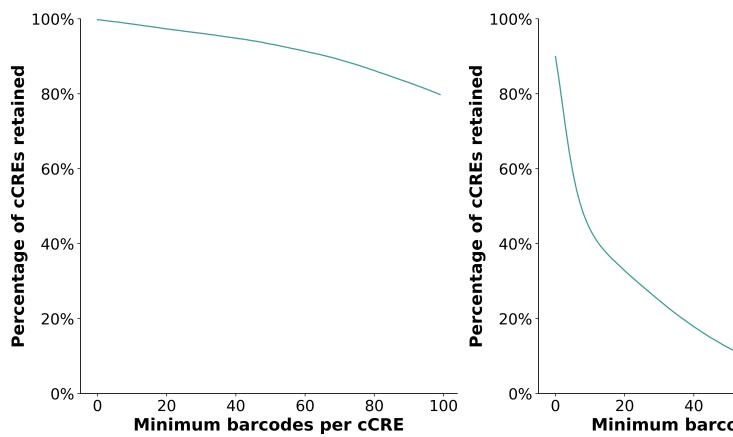
## 2.4 cCRE-barcode observations



**Goal:** Input file: Evaluated metrics:

**Legend:** Interpretation:

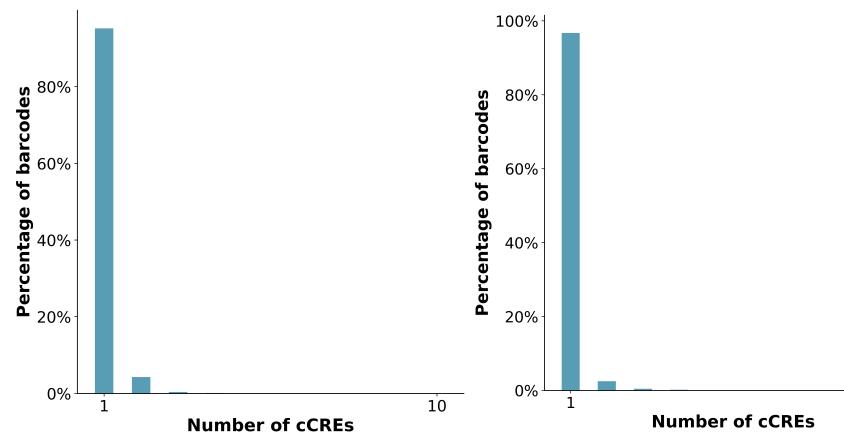
## 2.5 Retained cCREs



**Goal:** Input file: Evaluated metrics:

**Legend:** Interpretation:

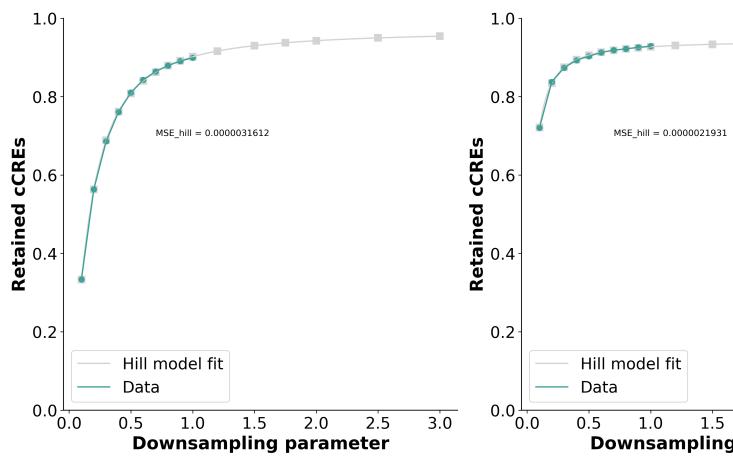
## 2.6 Barcode promiscuity



Goal: Input file: Evaluated metrics:

Legend: Interpretation:

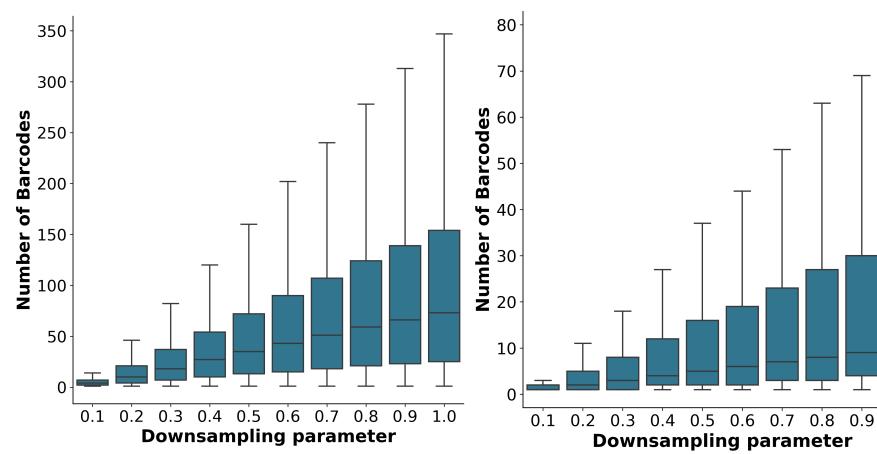
## 2.7 Downsampling - Retained cCRES



**Goal:** Input file: Evaluated metrics:

**Legend:** Interpretation:

## 2.8 Downsampling - Barcodes per cCRE



Goal: Input file: Evaluated metrics:

Legend: Interpretation:



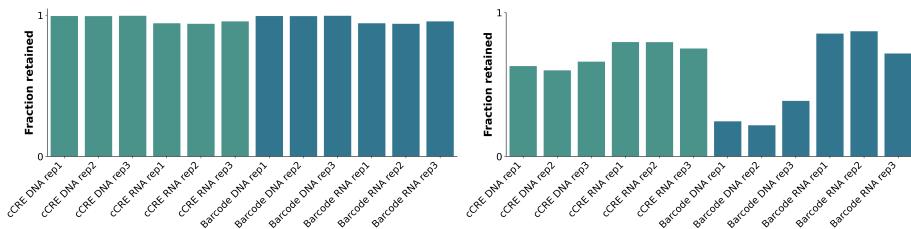
# Chapter 3

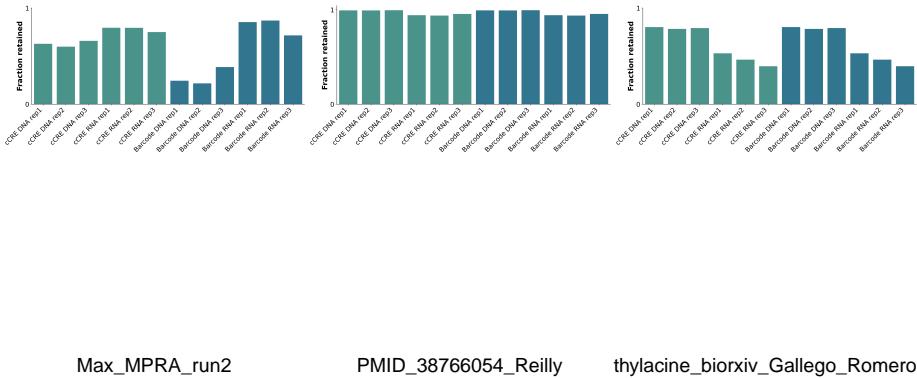
## Activity QC

### 3.1 Evaluating DNA and RNA complexity

#### 3.1.1 Retained cCREs and barcodes

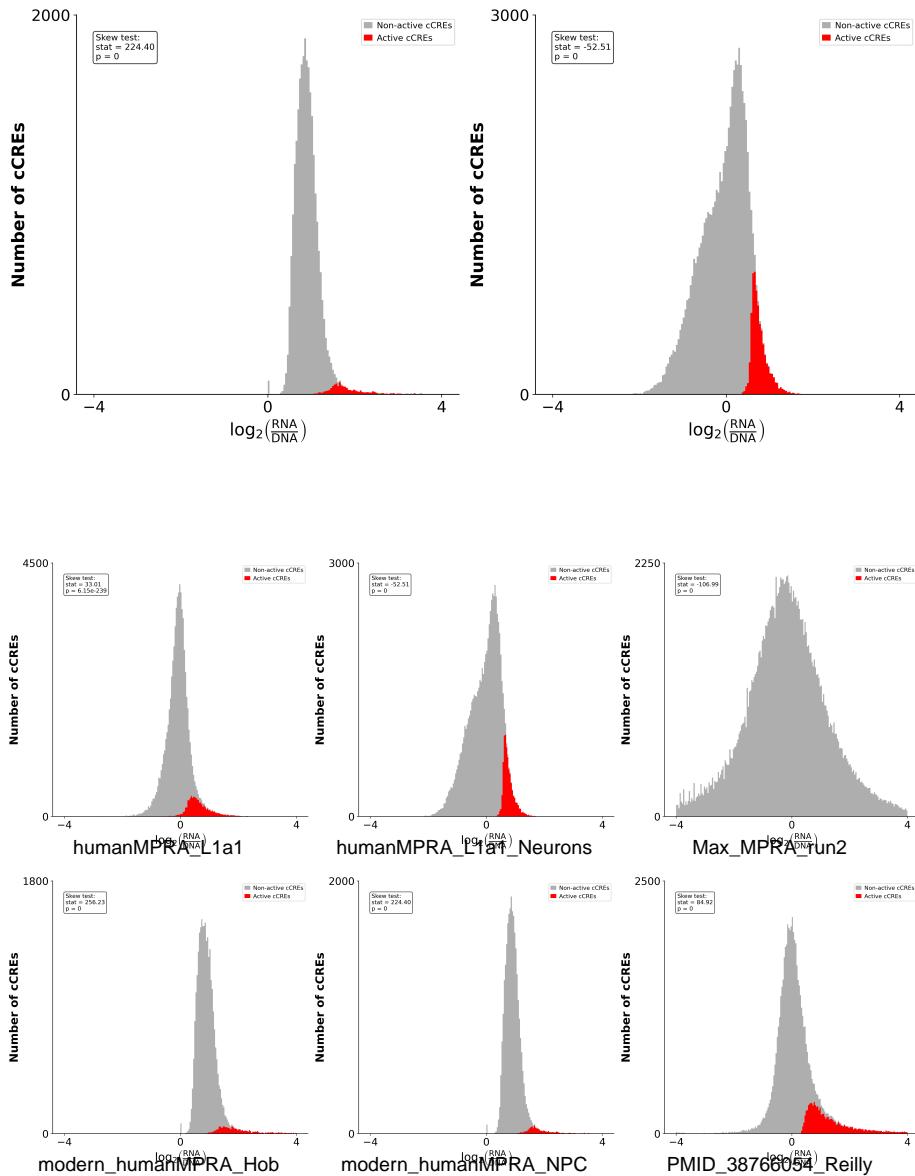
```
## Good example: PMID_38766054_Reilly  
## Bad example: Max_MPRA_run2
```





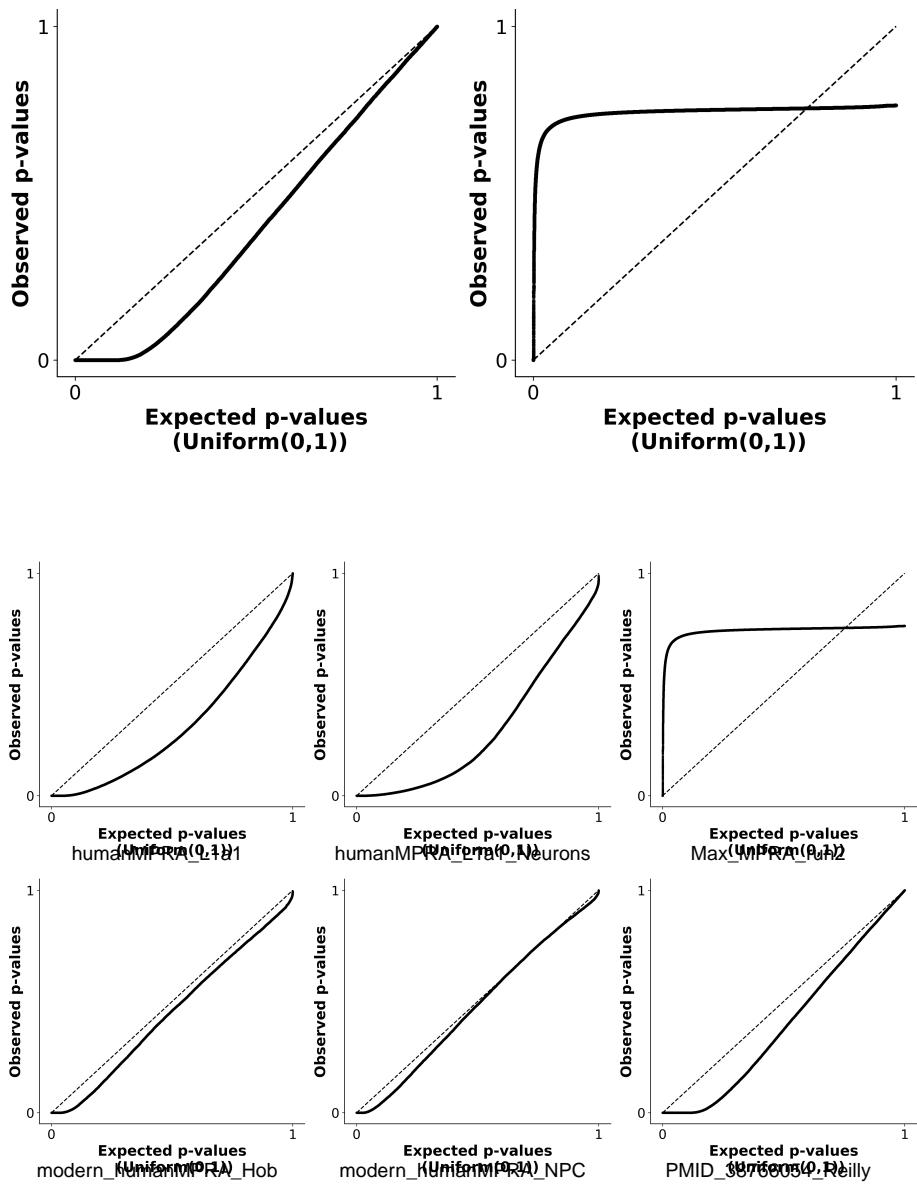
### 3.1.2 Activity distribution

```
## Good example: modern_humanMPRA_NPC
## Bad example: humanMPRA_L1a1_Neurons
```



### 3.1.3 P-value distribution

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

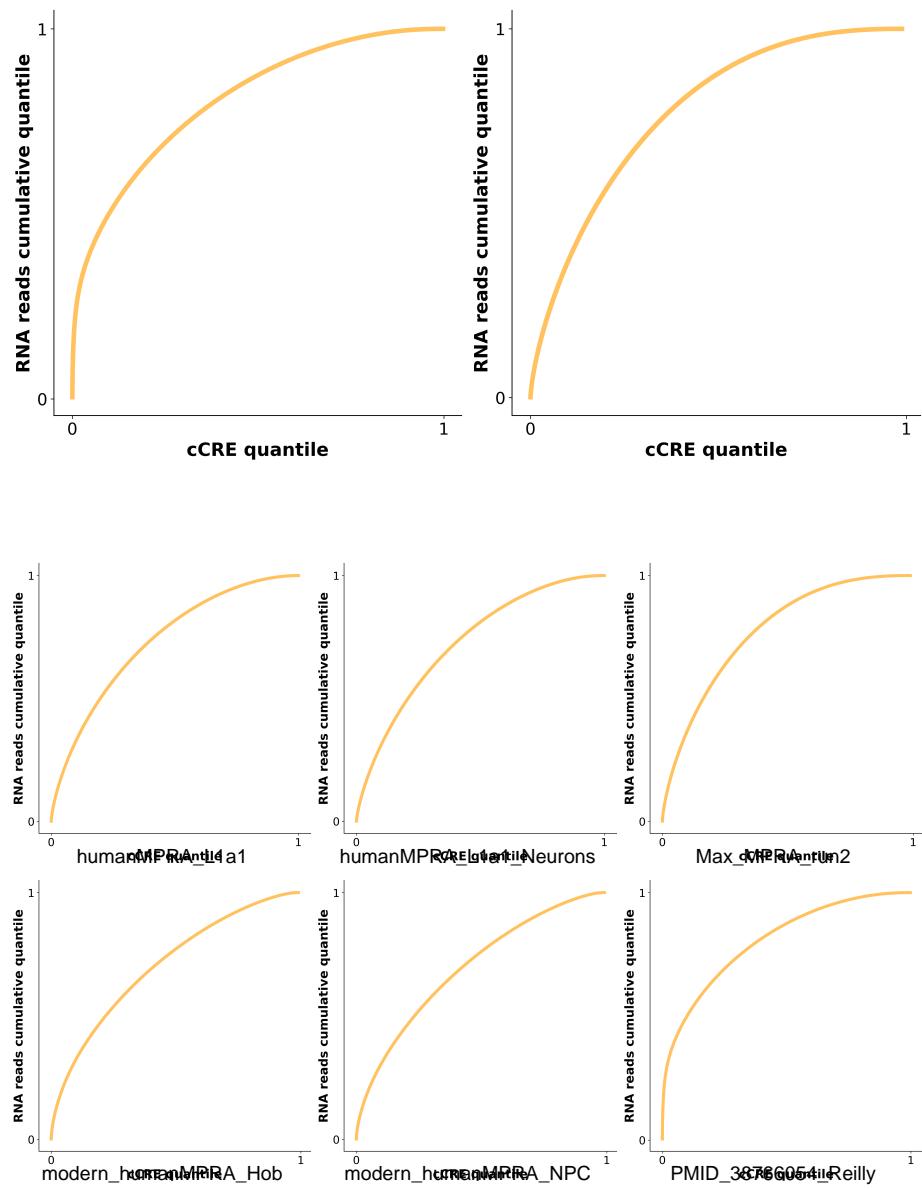


### 3.1.4 Downsampling analysis - active cCREs



### 3.1.5 Cumulative RNA reads

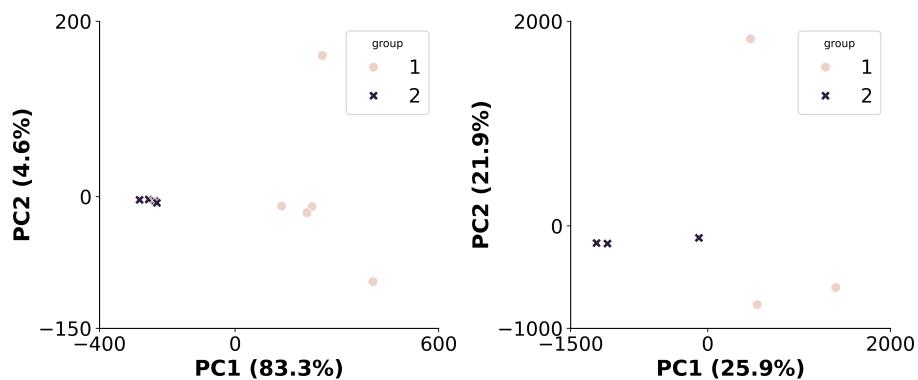
```
## Good example: PMID_38766054_Reilly  
## Bad example: Max_MPRA_run2
```

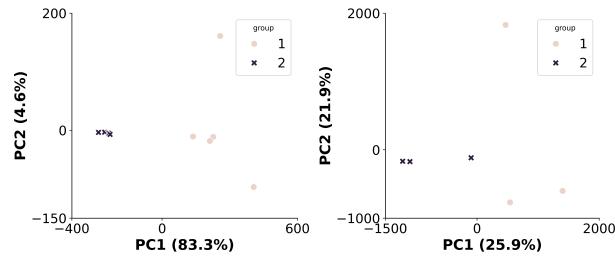


## 3.2 Evaluating reproducibility

### 3.2.1 Similarity between samples (PCA)

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

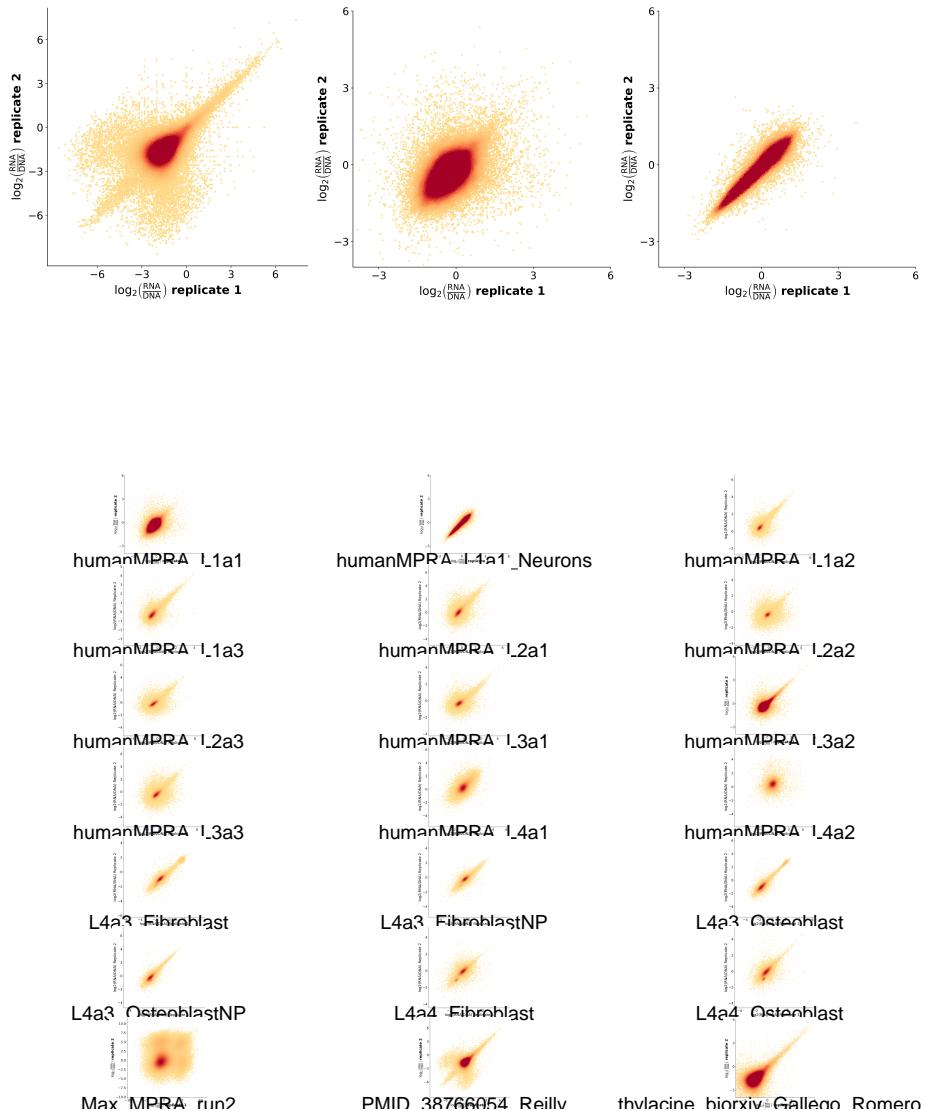




PMID\_38766054\_Reilly      thylacine\_biorxiv\_Gallego\_Romero

### 3.2.2 Correlation between replicates

```
## Good example: PMID_38766054_Reilly  
## Bad example: humanMPRA_L1a1  
## Bad example 2: humanMPRA_L1a1_Neurons
```

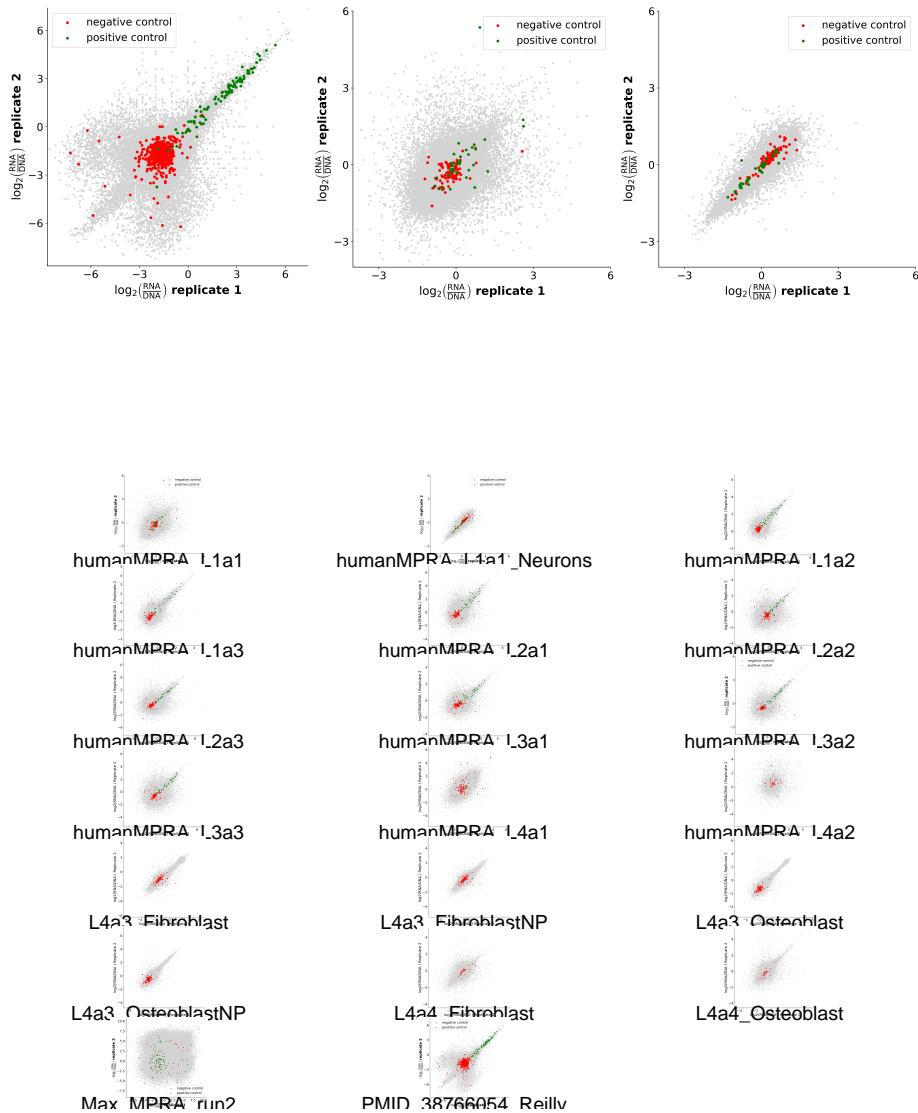


### 3.2.3 Variation at various activity levels

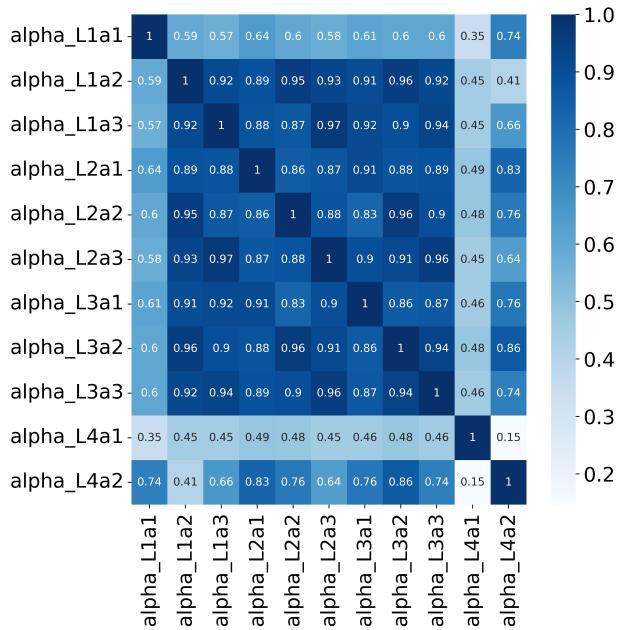


### 3.2.4 Correlation between replicates (controls)

```
## Good example: PMID_38766054_Reilly  
## Bad example: humanMPRA_L1a1  
## Bad example 2: humanMPRA_L1a1_Neurons
```



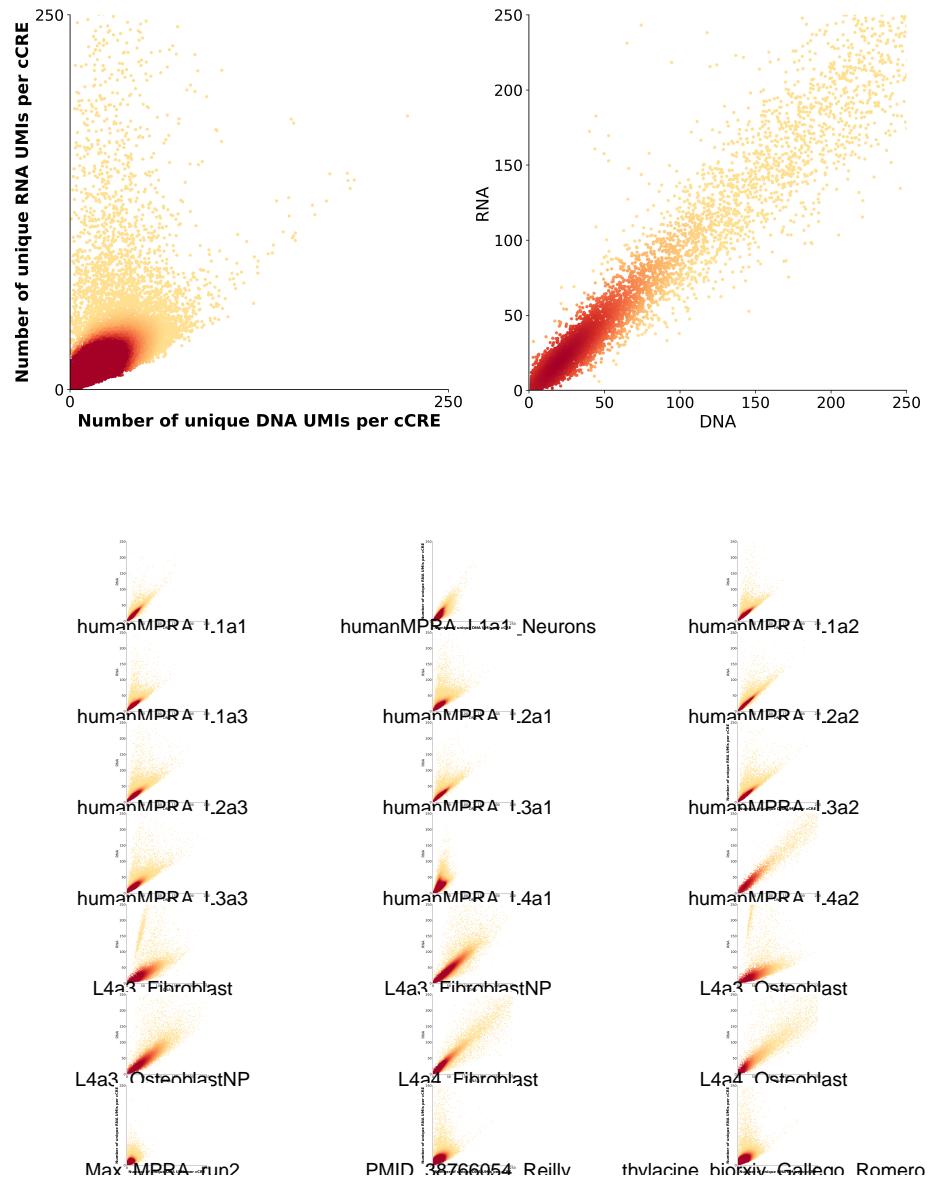
### 3.2.5 Activity of controls - sample comparison



```
## [1] "humanMPRA_L3a2"
```

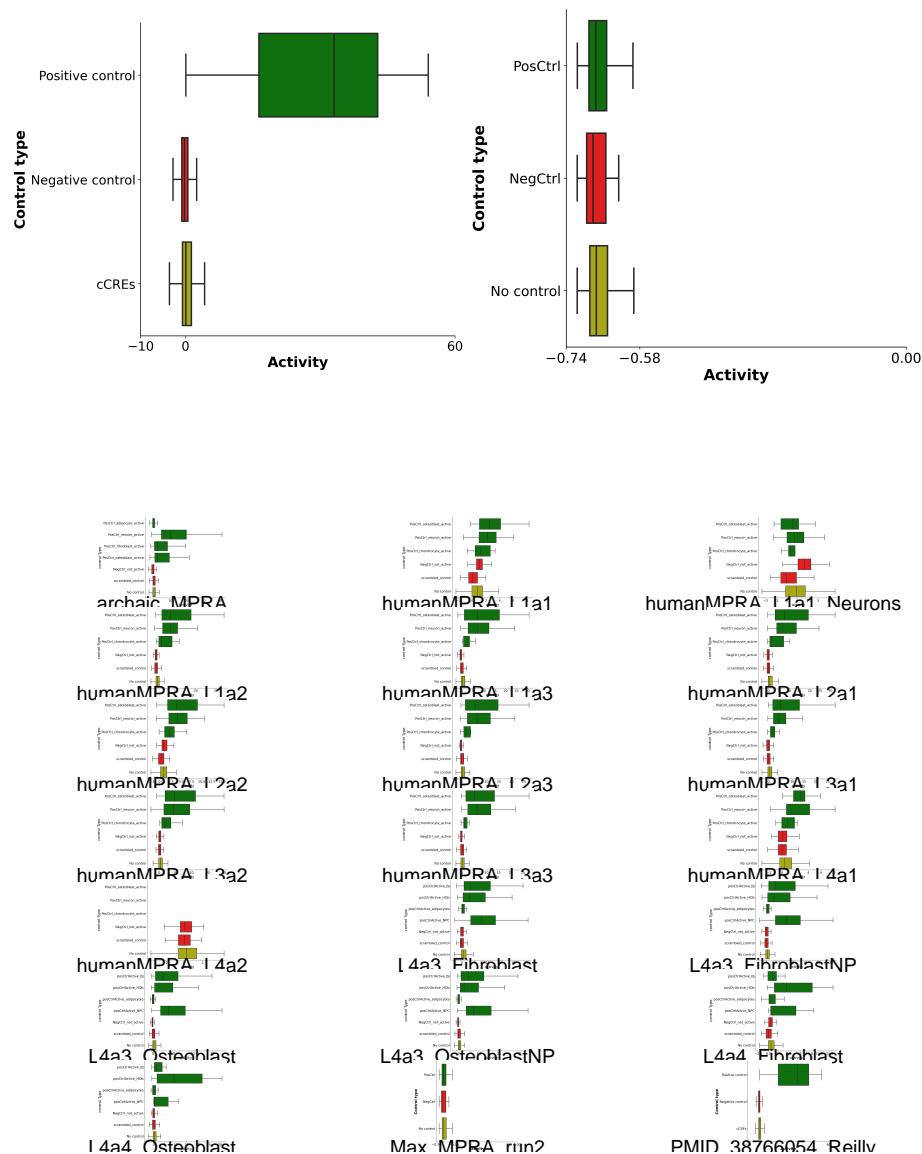
### 3.2.6 RNA\_DNA\_ratio

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



### 3.2.7 Activity of controls - sample comparison

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



### 3.2.8 Outlier barcodes + min(DNA)



### 3.2.9 Outlier barcodes

