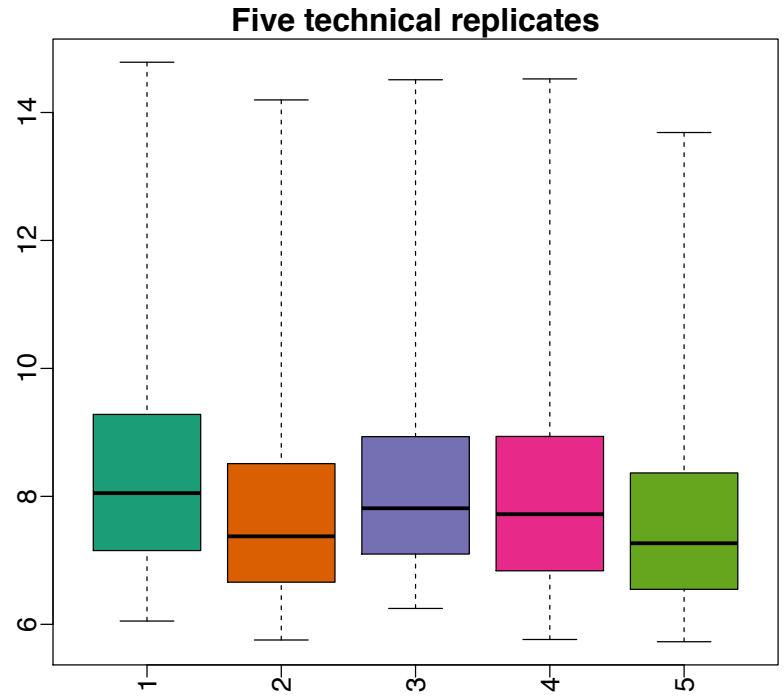
Normalization

What are we trying to fix and how do we know we fixed it?

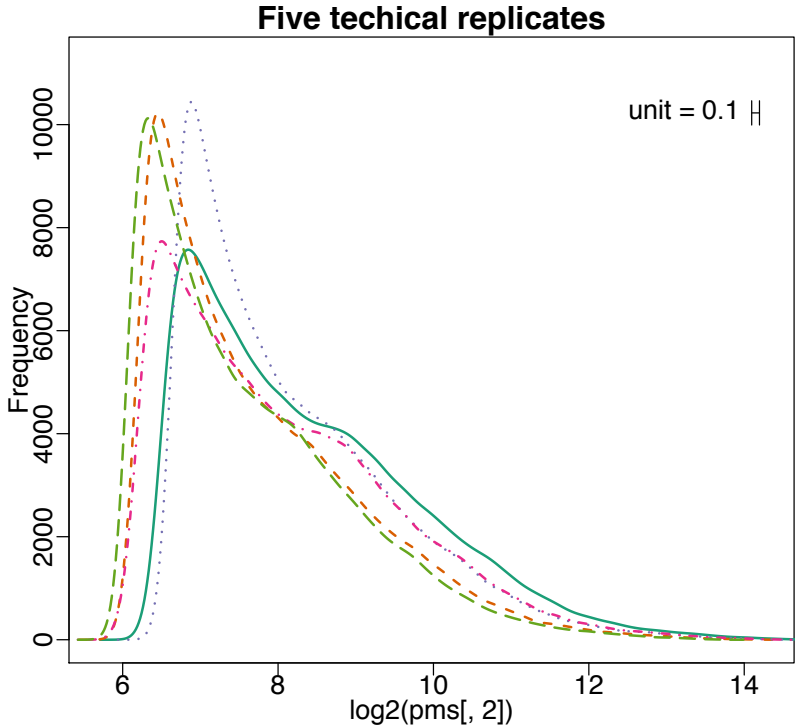
My definition of *normalization* in the context of high-throughput data

- Accounting or correcting for sample (cell) specific technical effects

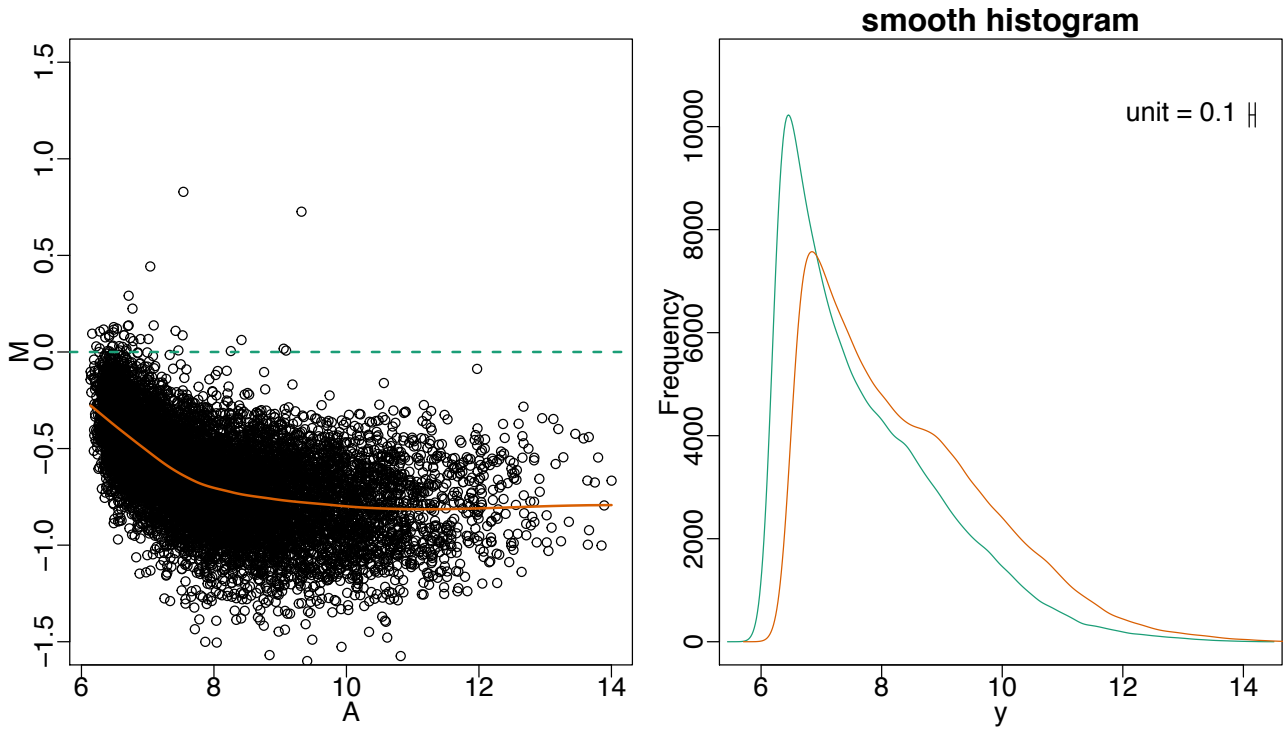
Values expected to be the same are not



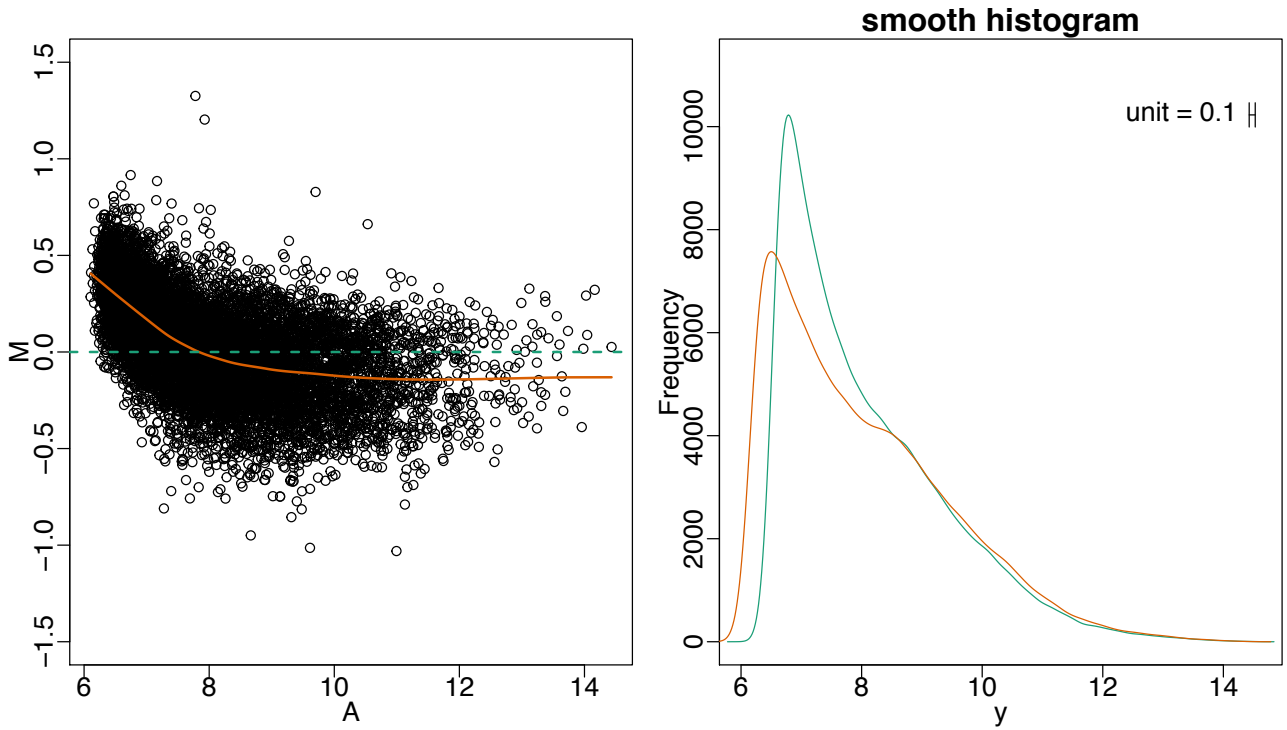
Distributions are different



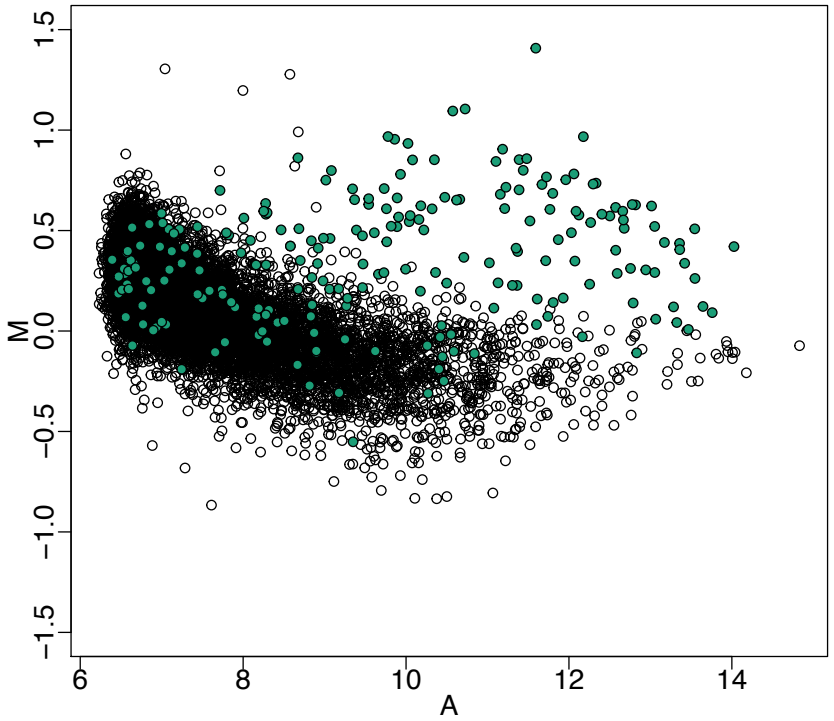
More than location and scale changes



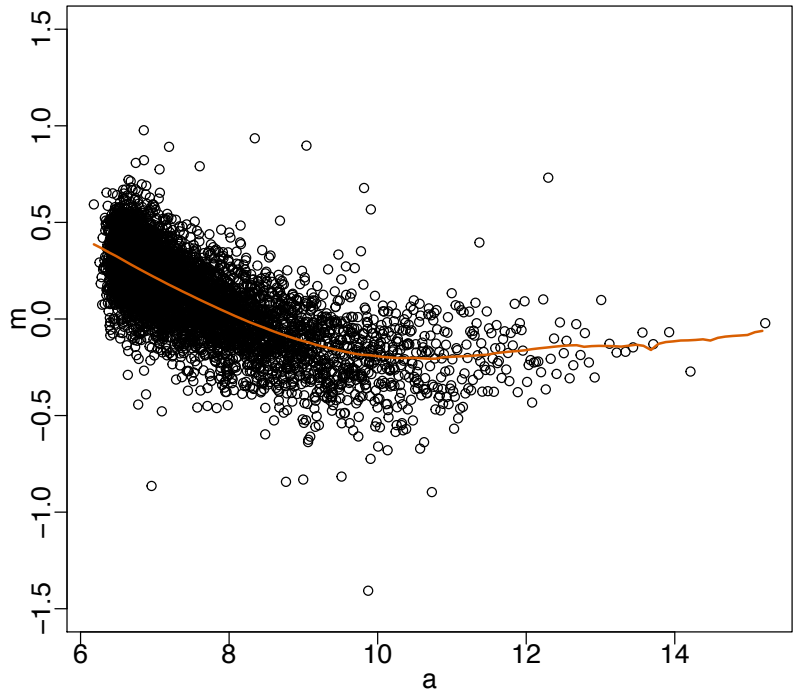
Median shifts do not solve the problem



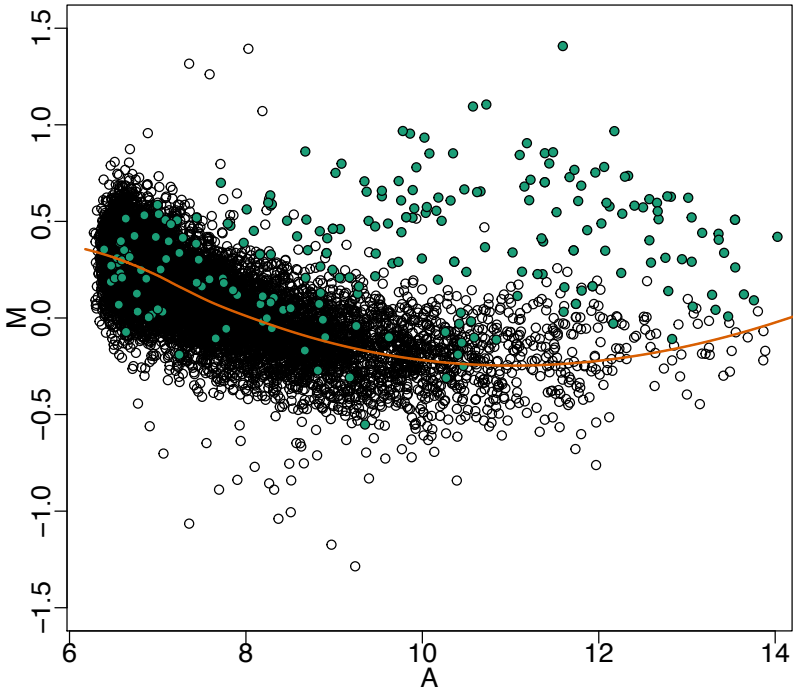
Non-linear effects



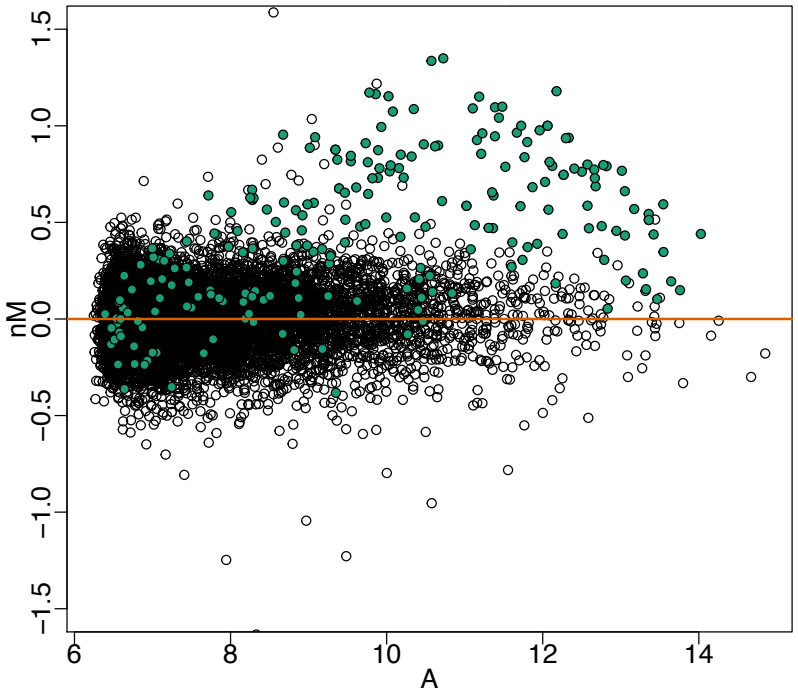
Solution 1: Local regression (loess)



Before



After



Solution 2: Quantile Normalizatio

Example of quantile normalization

Original

2	4	4	5
5	14	4	7
4	8	6	9
3	8	5	8
3	9	3	5

Order

2	4	3	5
3	8	4	5
3	8	4	7
4	9	5	8
5	14	6	9

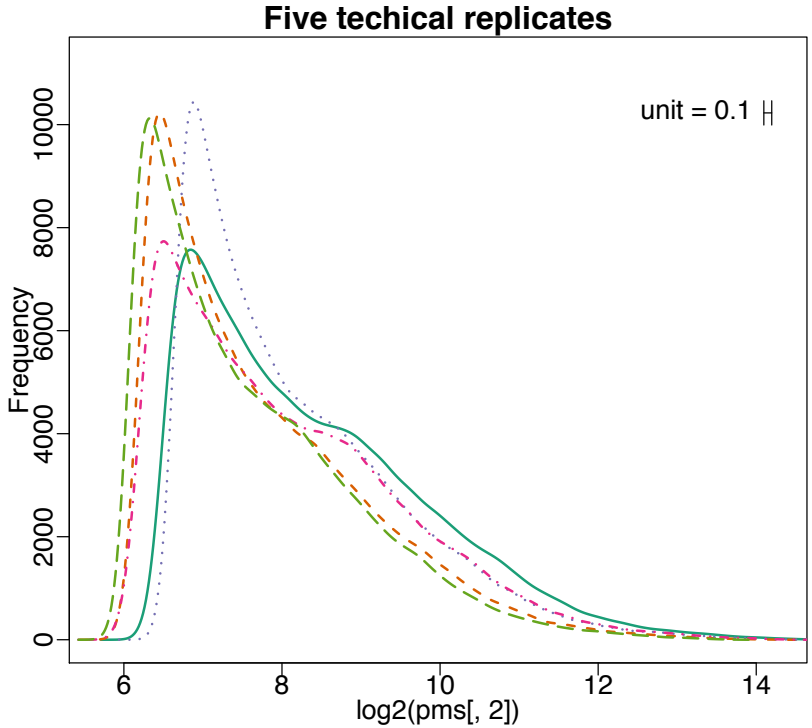
Averaged

3.5	3.5	3.5	3.5
5.0	5.0	5.0	5.0
5.5	5.5	5.5	5.5
6.5	6.5	6.5	6.5
8.5	8.5	8.5	8.5

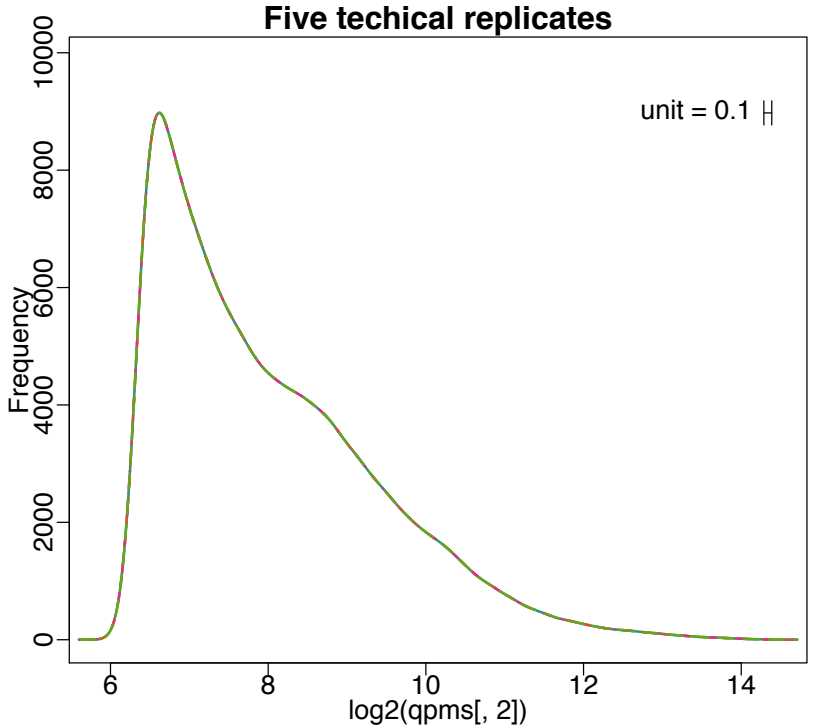
Re-order

3.5	3.5	5.0	5.0
8.5	8.5	5.5	5.5
6.5	5.0	8.5	8.5
5.0	5.5	6.5	6.5
5.5	6.5	3.5	3.5

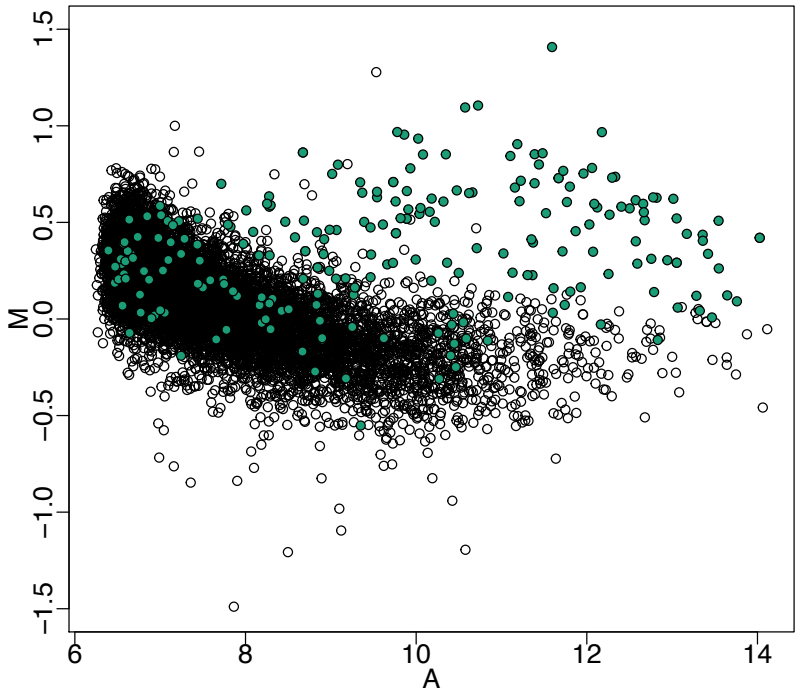
Densities are forced to be identical



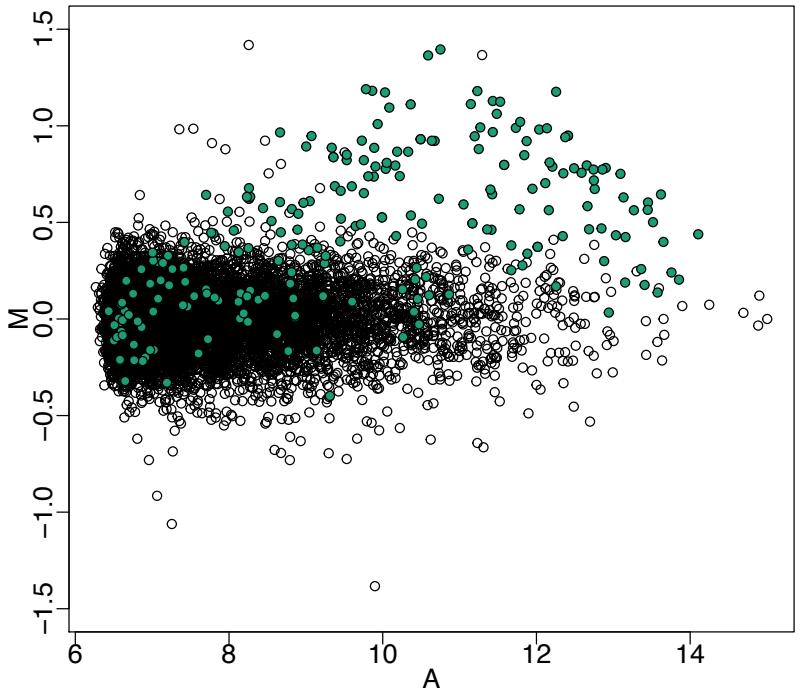
Densities are forced to be identical



But differential expression can be preserved

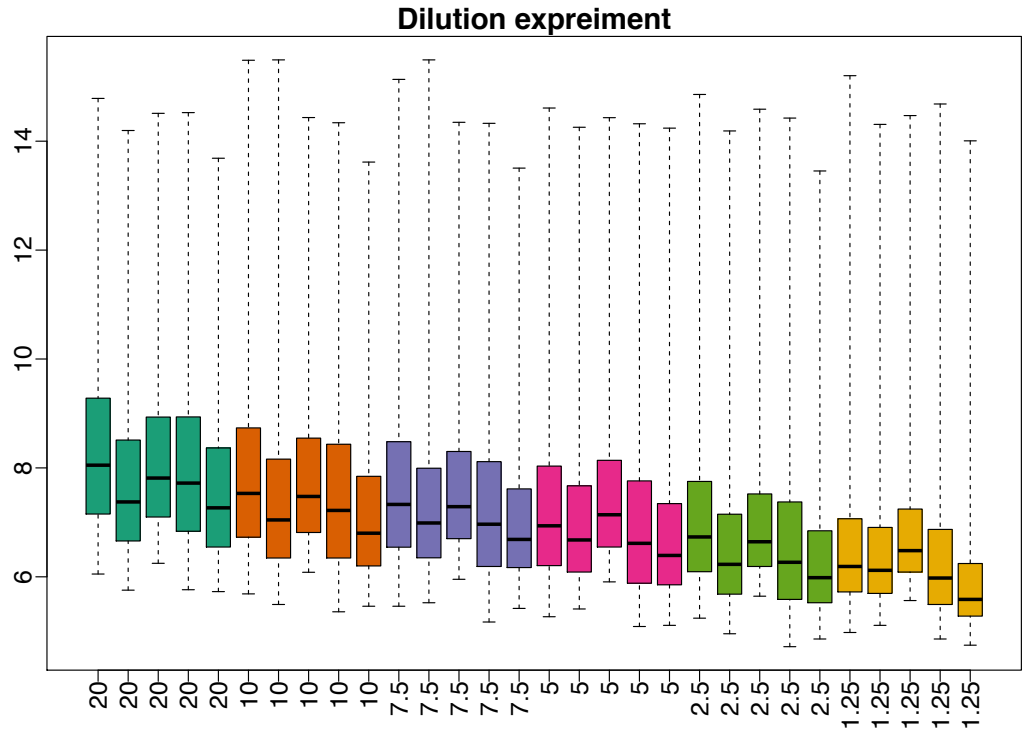


But differential expression can be preserved

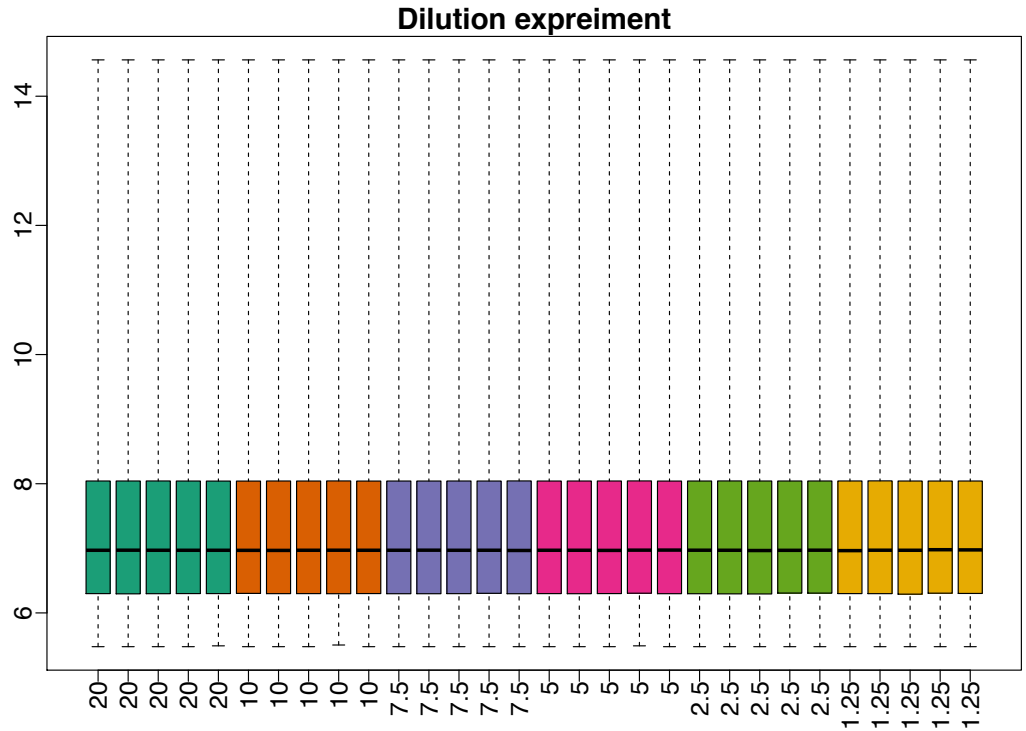


When not to normalize

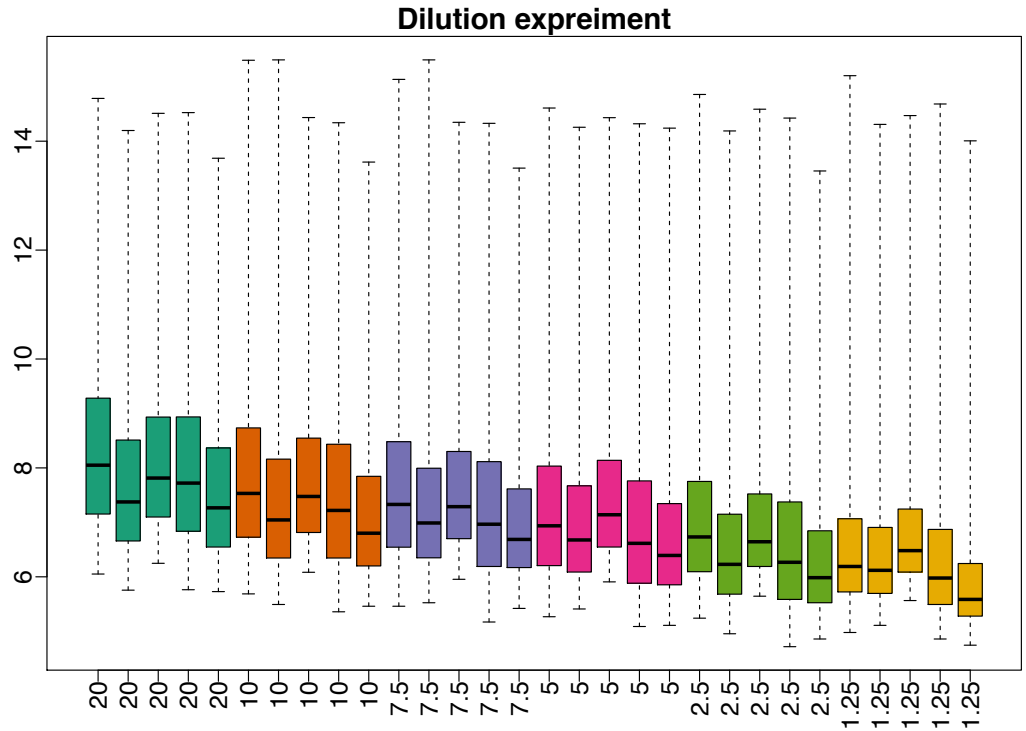
Dilution experiment



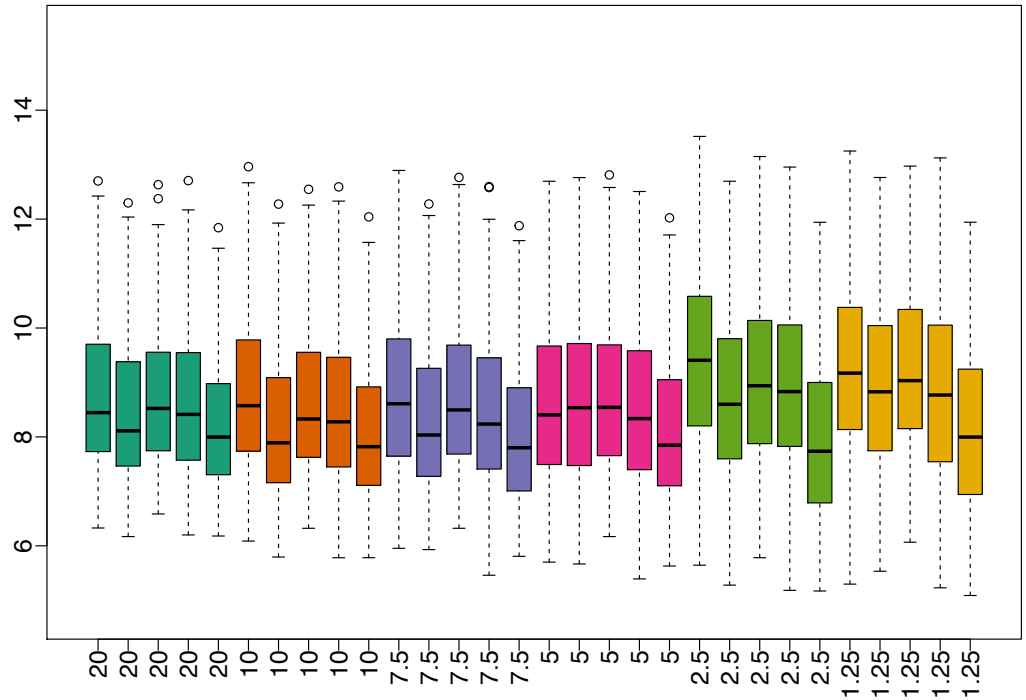
After quantile normalize



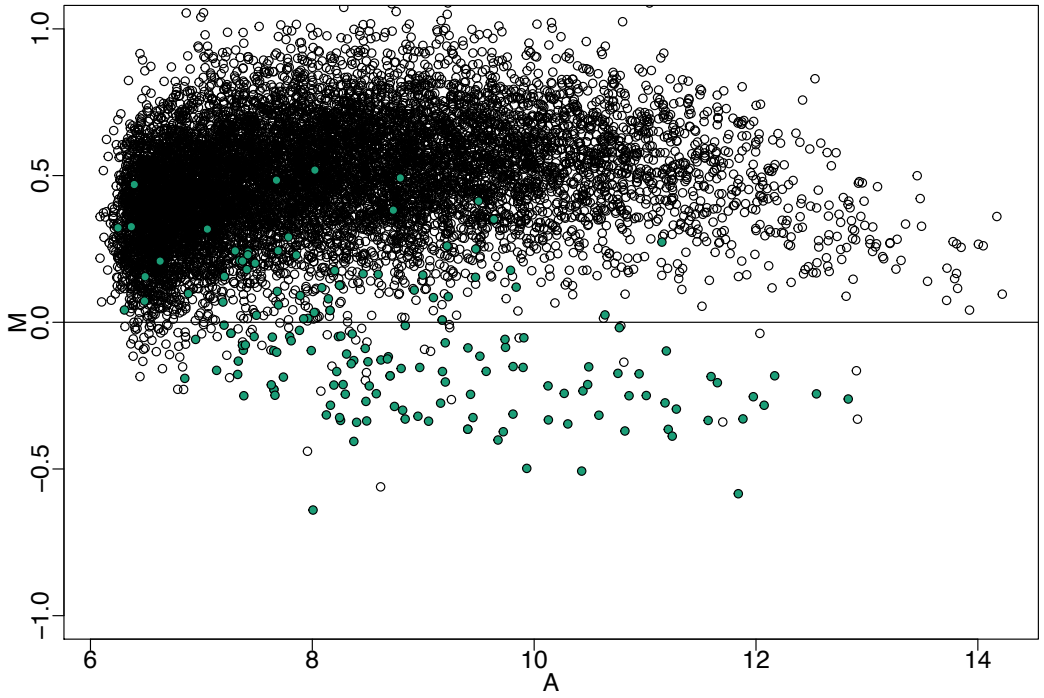
Dilution experiment



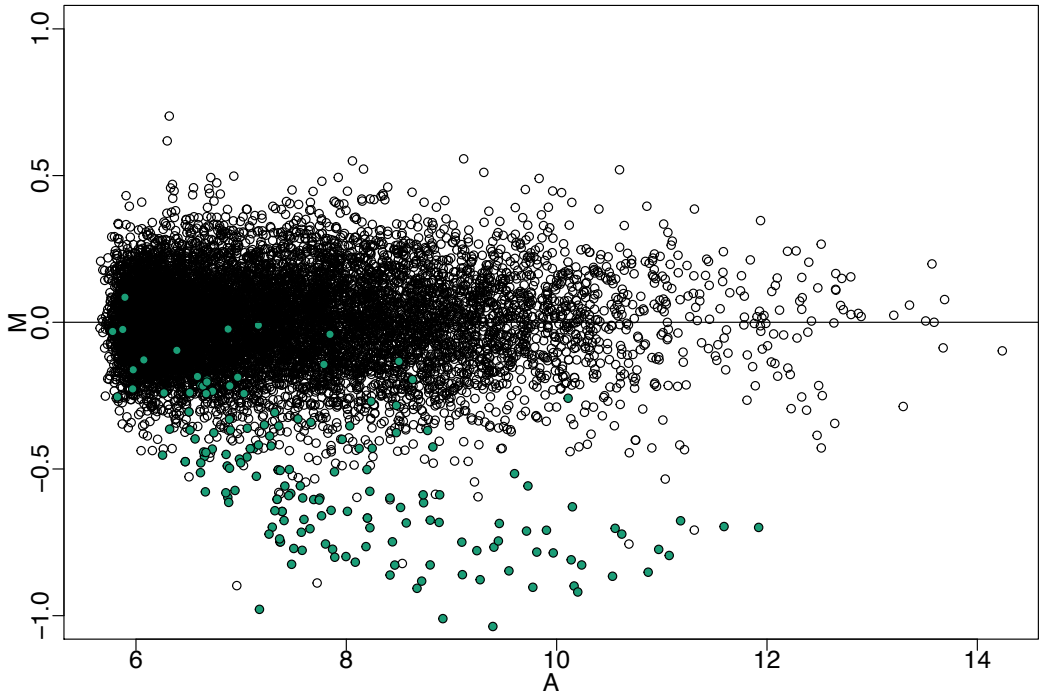
Spike-in controls



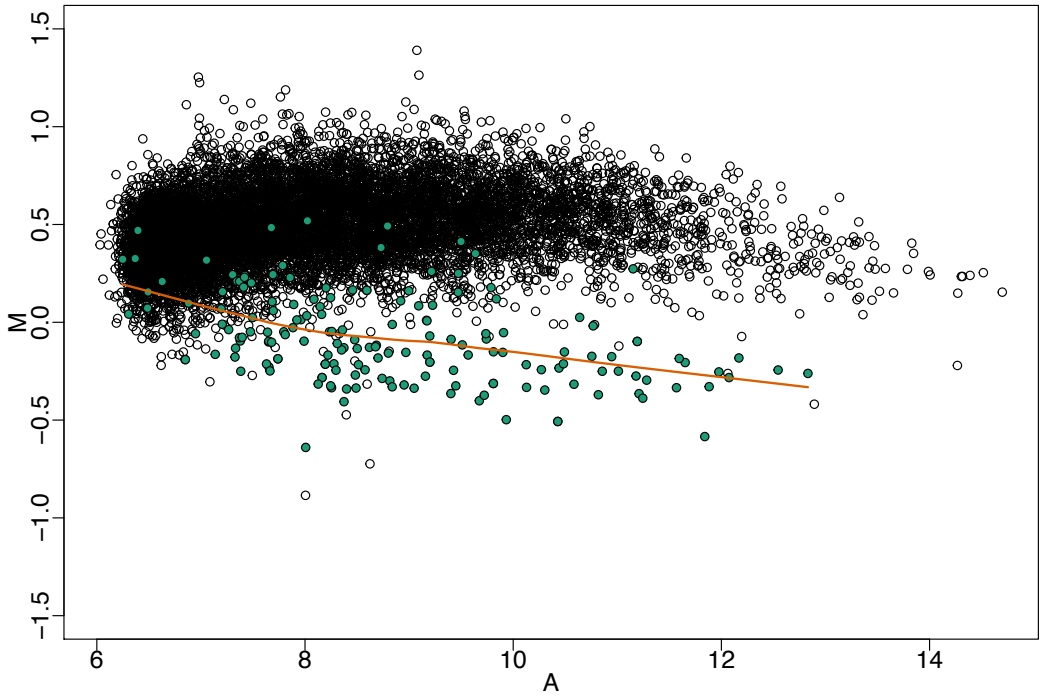
Most genes should be twice as large



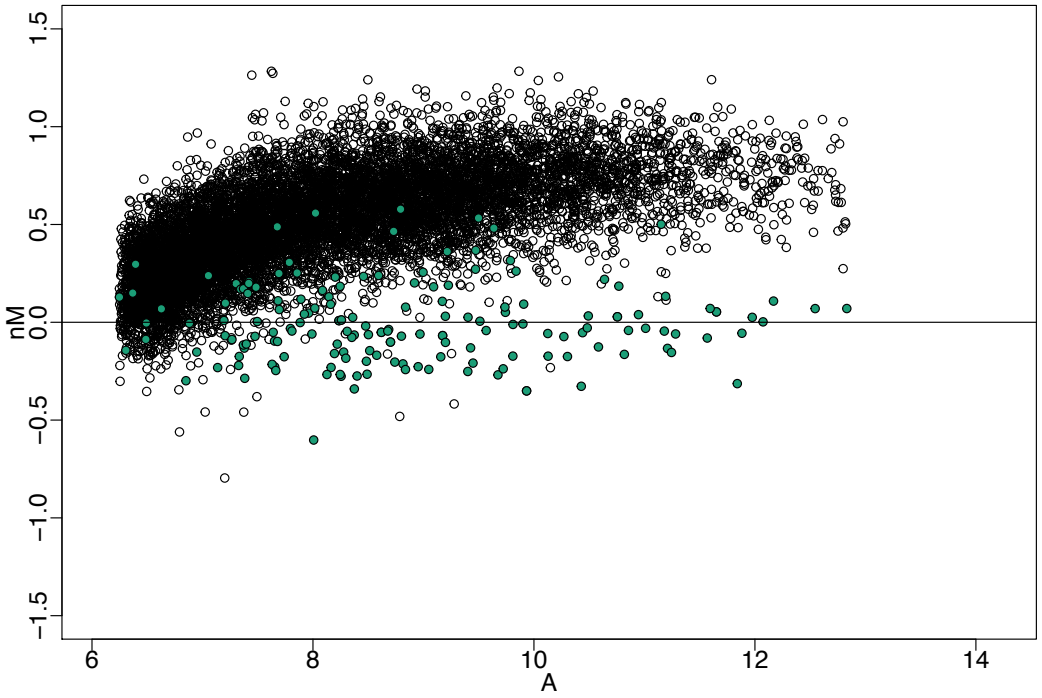
Quantile washes away signal



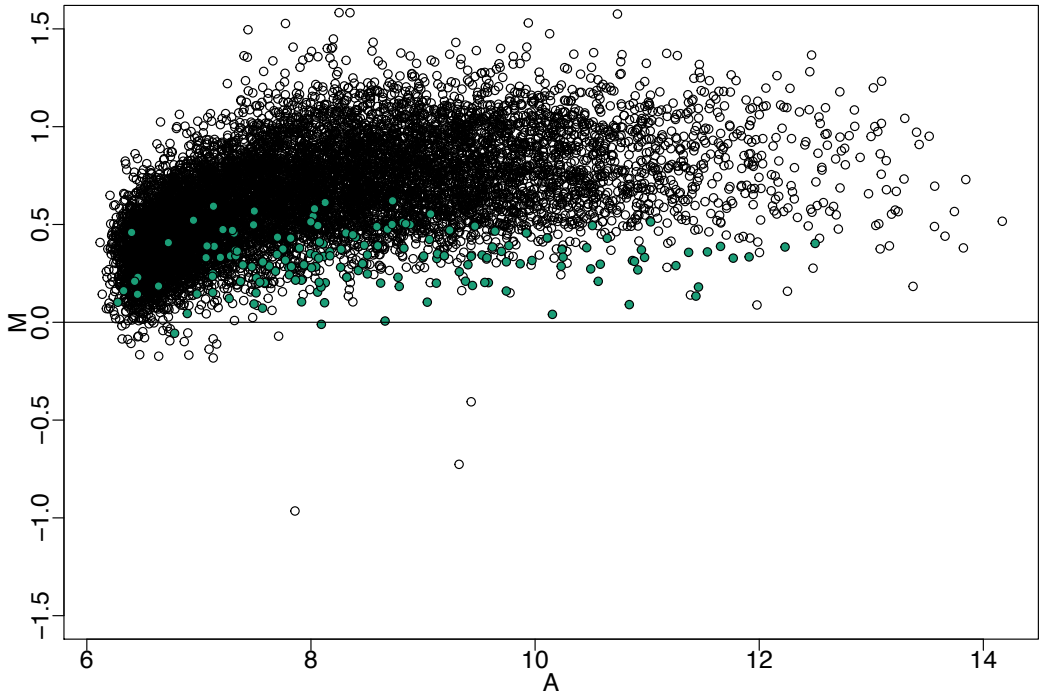
We can normalize using spike-in controls

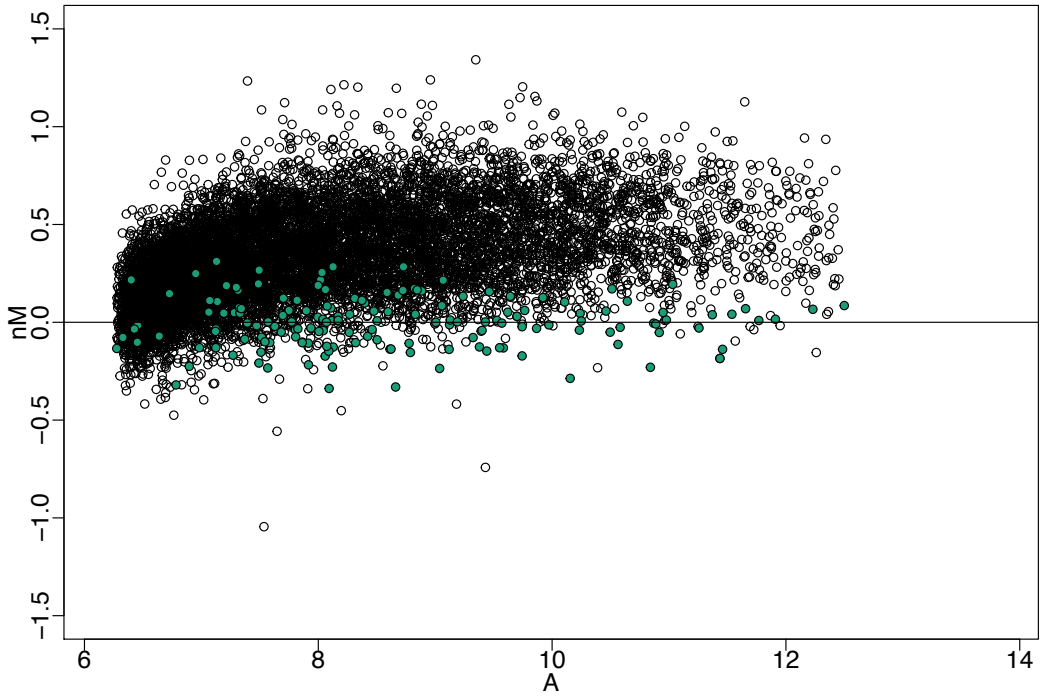


After normalization



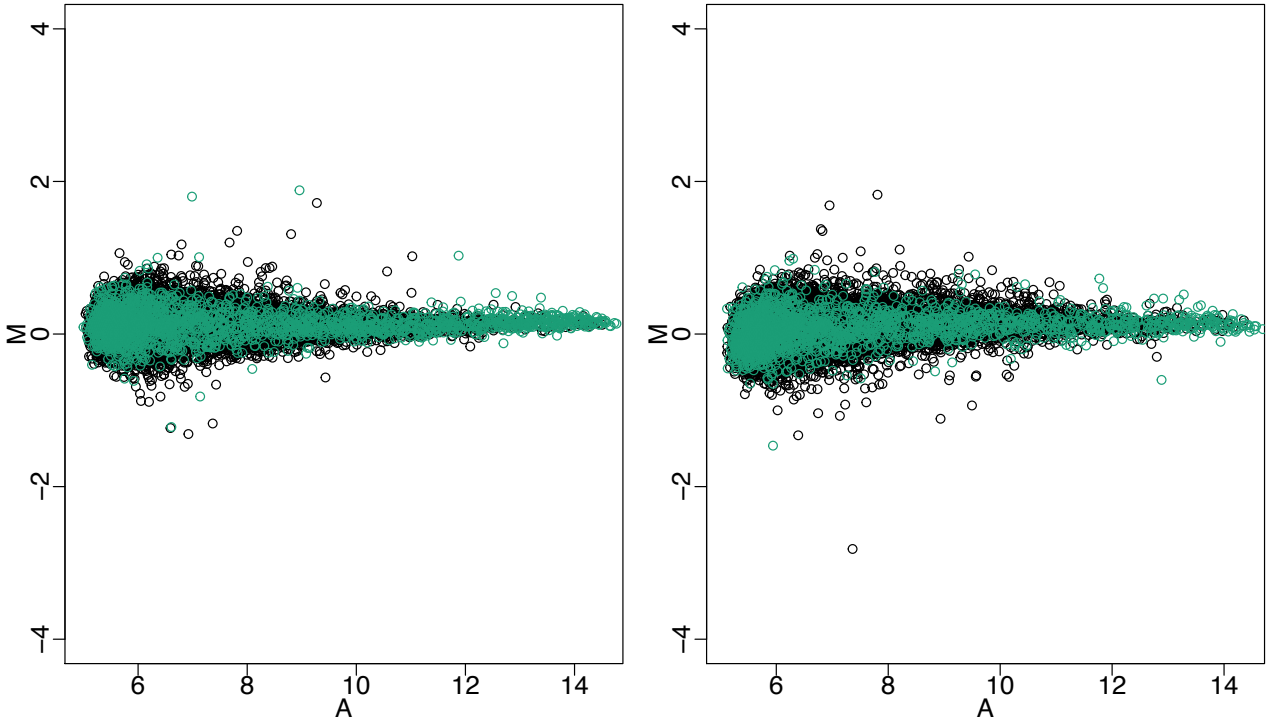
Beware: spike-in controls are not always good controls




nM
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
6
8
10
12
14
A

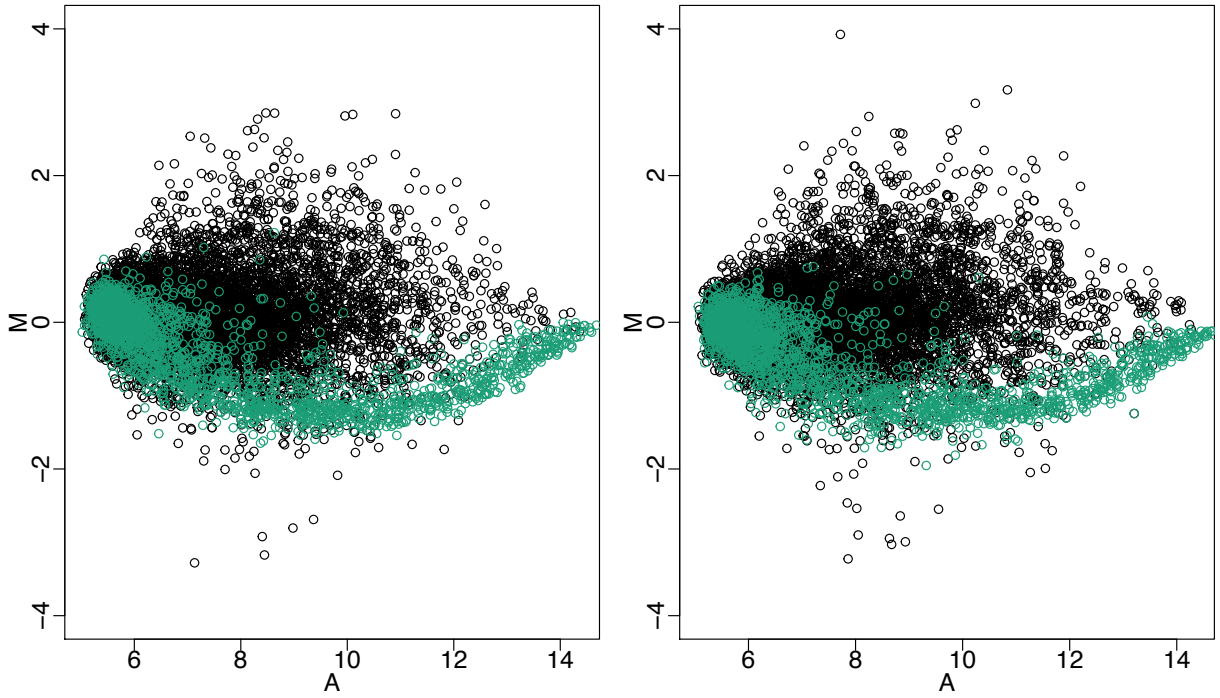
Segment 7 – …
Subset quantile

Here is an example with well-behving controls

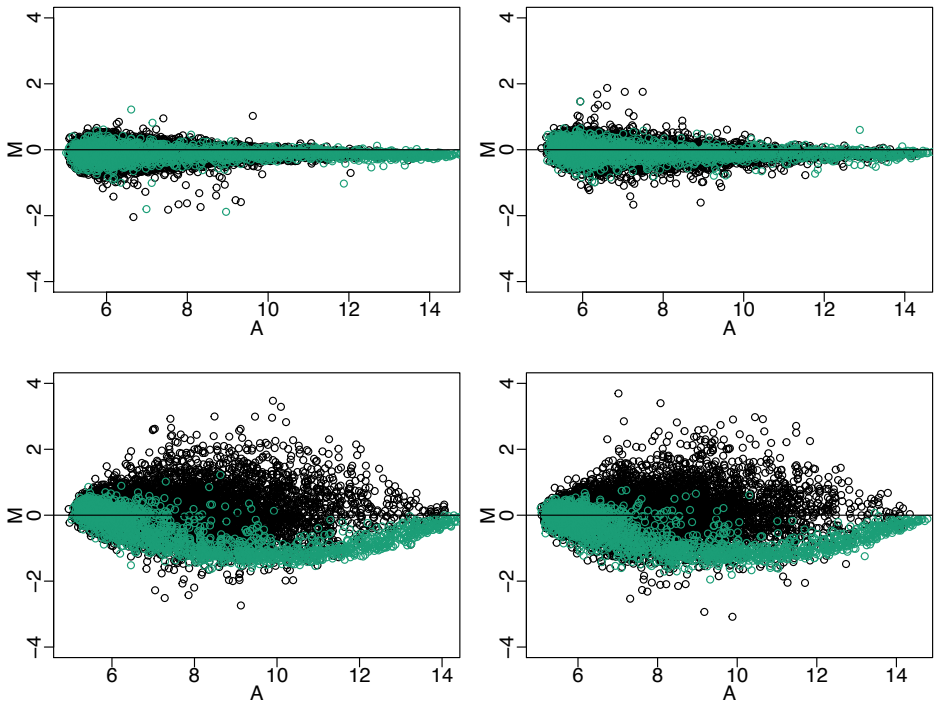


Add citation to Cell paper here

Comparing two samples known to have large global differnecs

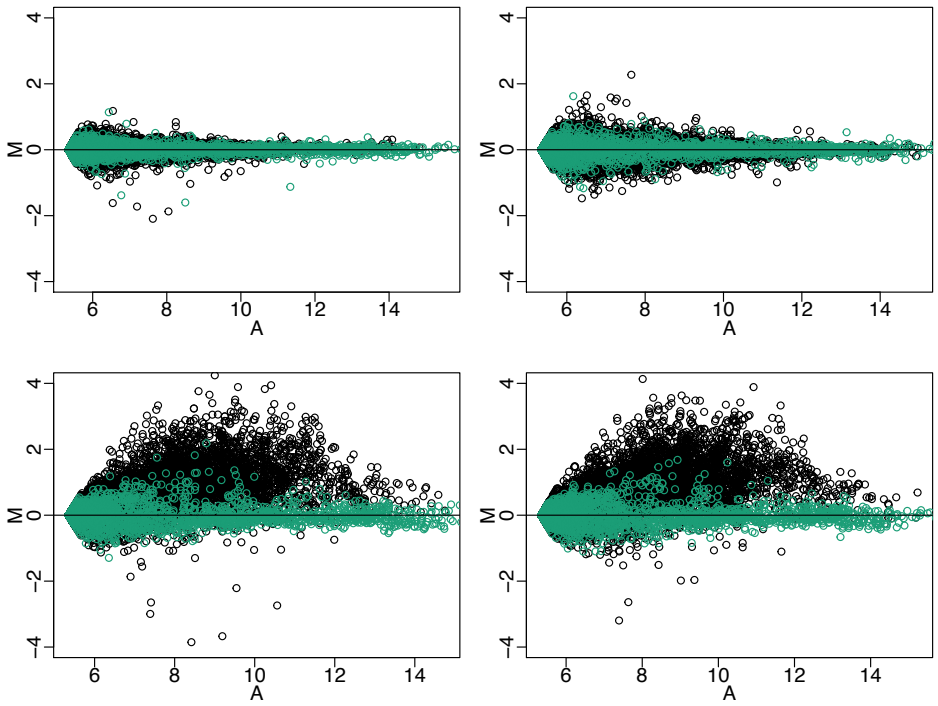


Before subset quantile normalization (SQN)



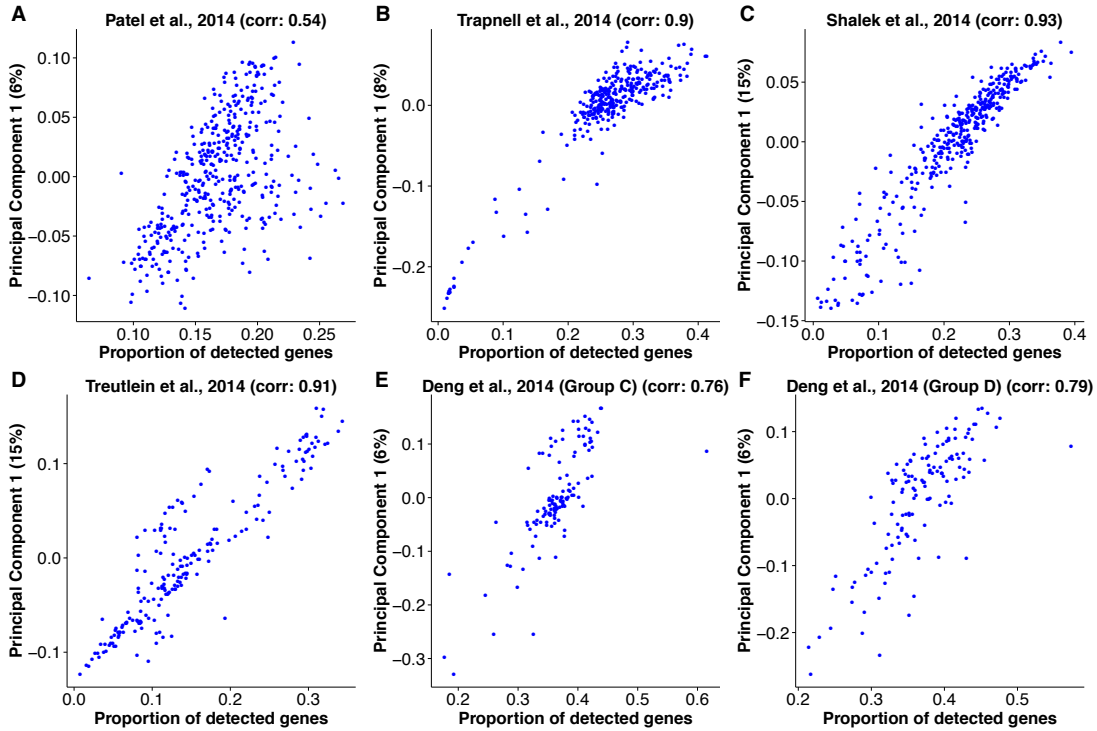
Add citation to SQN

Before subset quantile normalization (SQN)

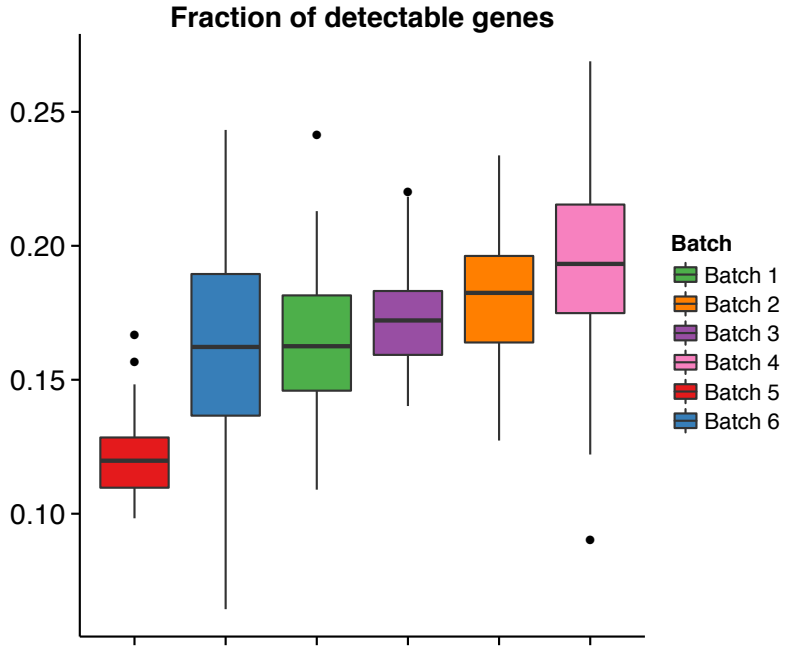


scRNAseq

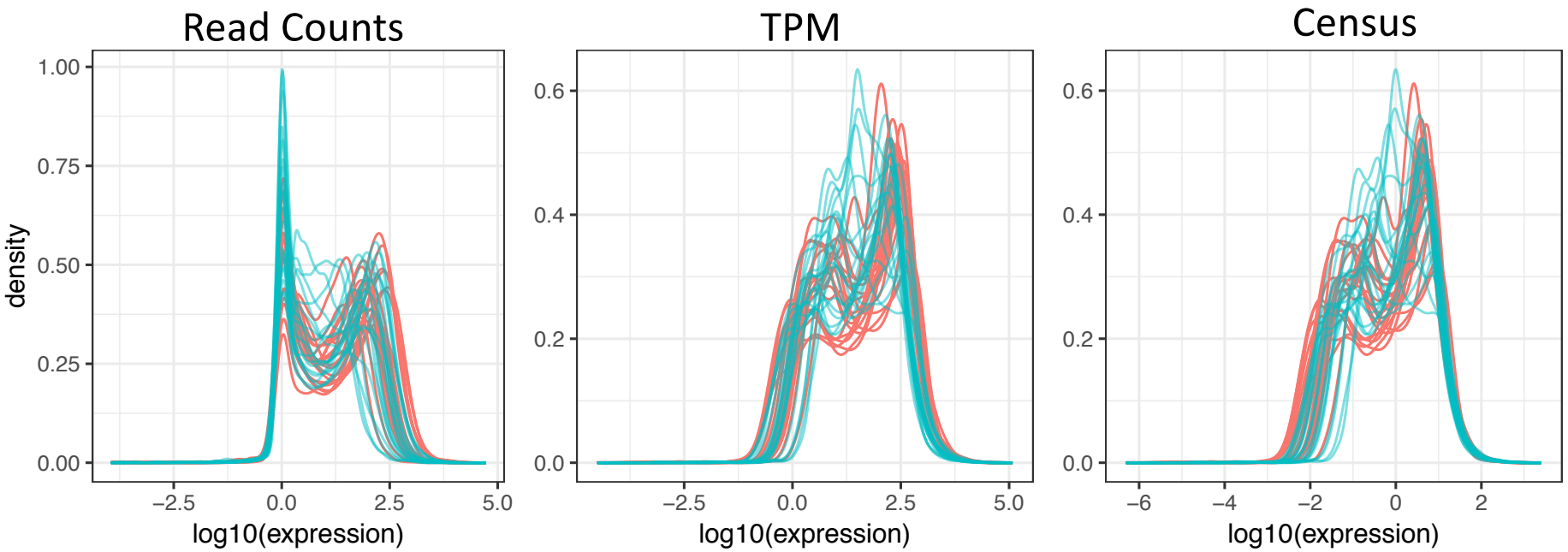
There are many zeros and they varies across sample

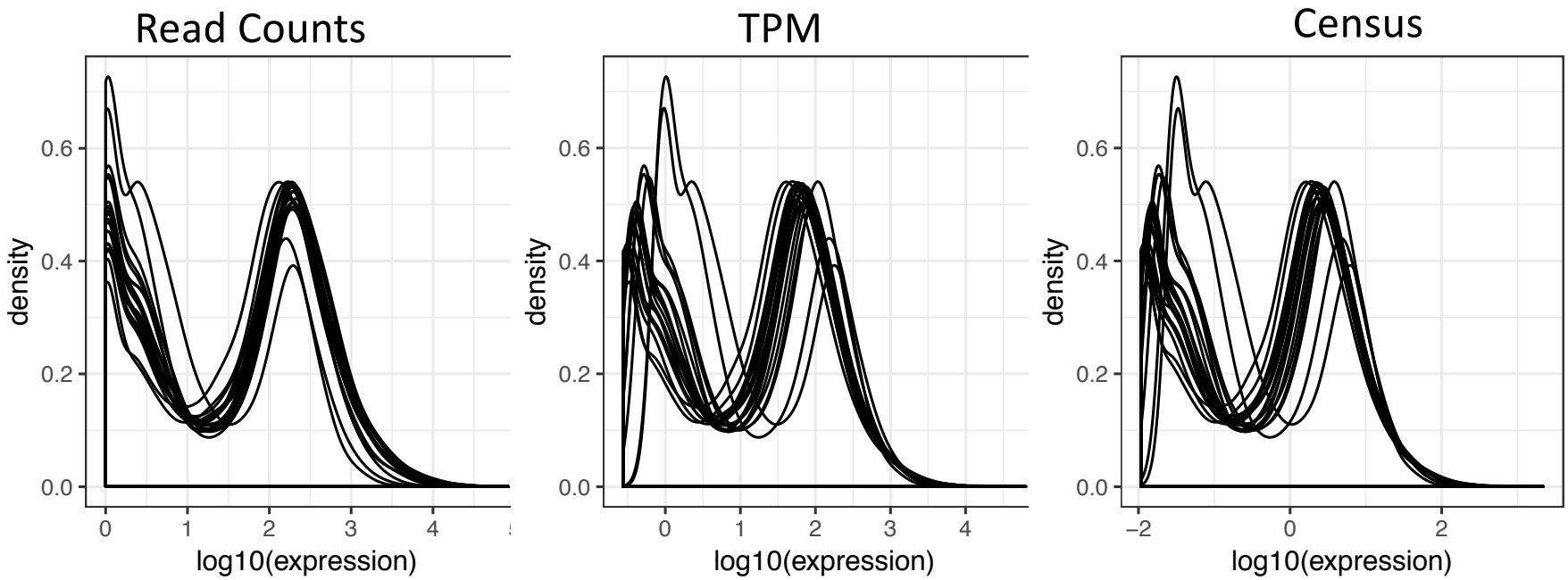


The proportion of zeros changes

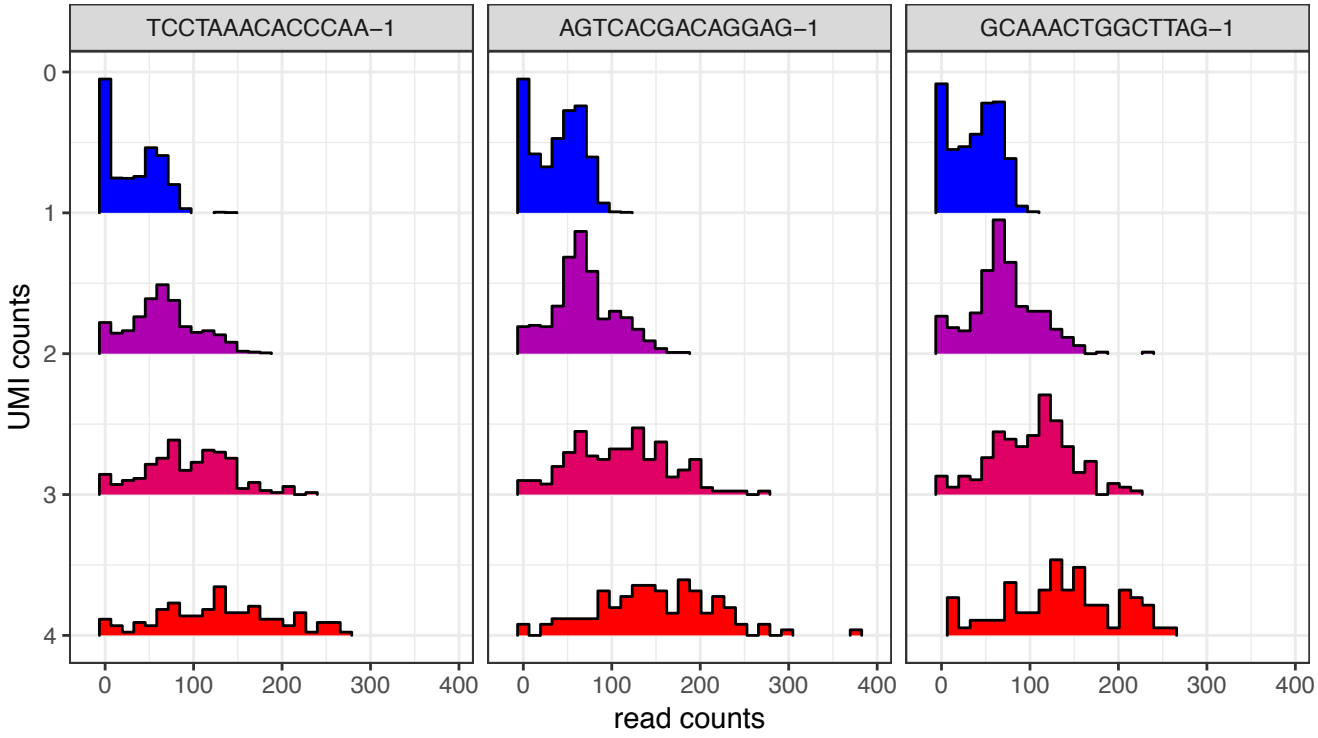


Patel et al dataset




Read Counts
TPM
Census
density
log10(expression)
0.0
0.2
0.4
0.6
0
1
2
3
4
−2

This seems to be mostly technical variability



Major challenges

- Are differences in total observed UMIs biological from technical?
- Are differences in empirical distribution biological or technical?
- Can modular approaches deal with low count data?
- Can we find useful controls?