05: Power simulations

Eneko Villanueva,Rayner Queiroz, Manasa Ramakrishna, Tom Smith

November 12, 2019

Contents

1. Introduction

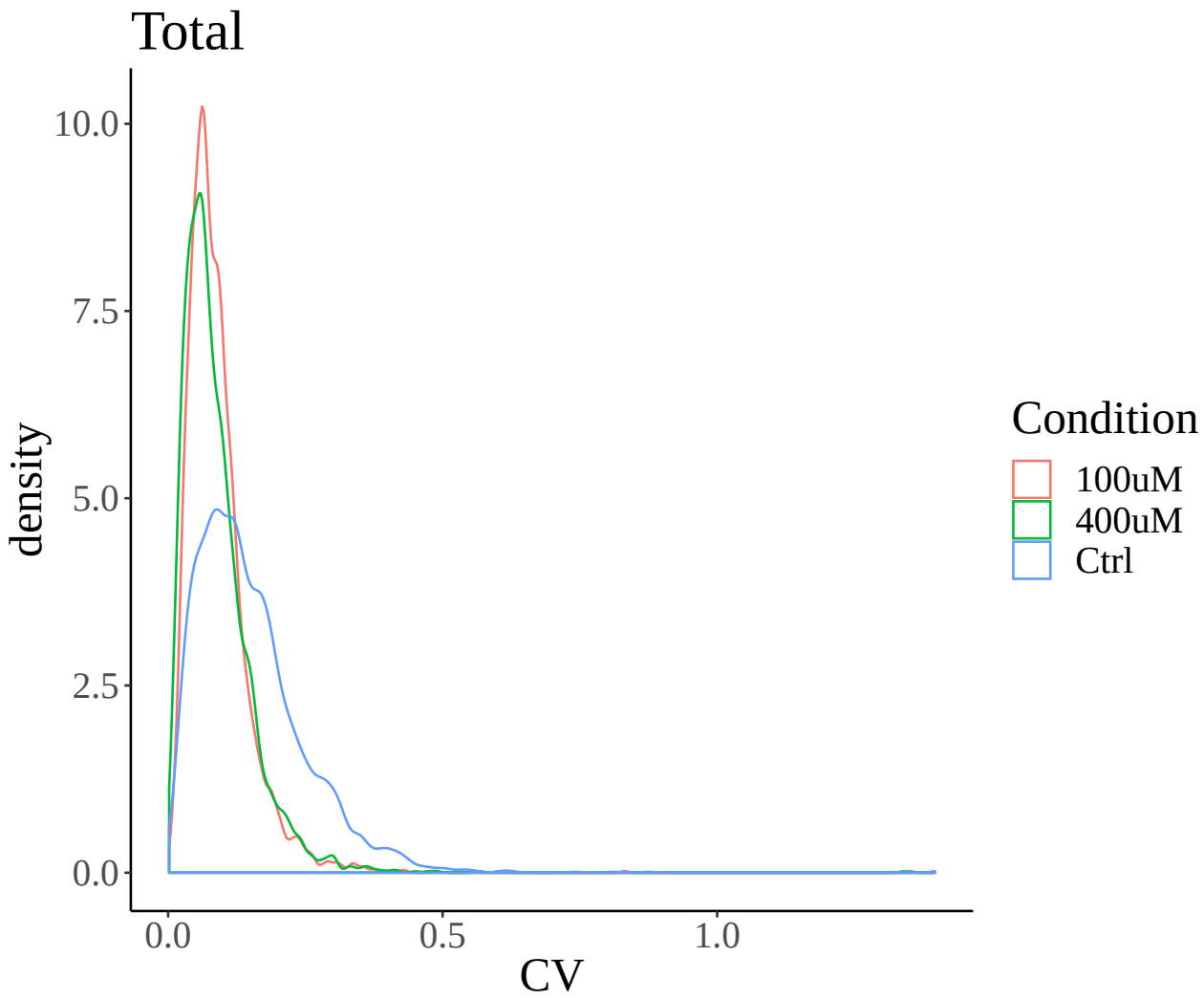
Here we simulate some datasets to estimate the number of replicates required to reach a reasonable power to detect change in RNA binding

2. Reading in normalised protein level data

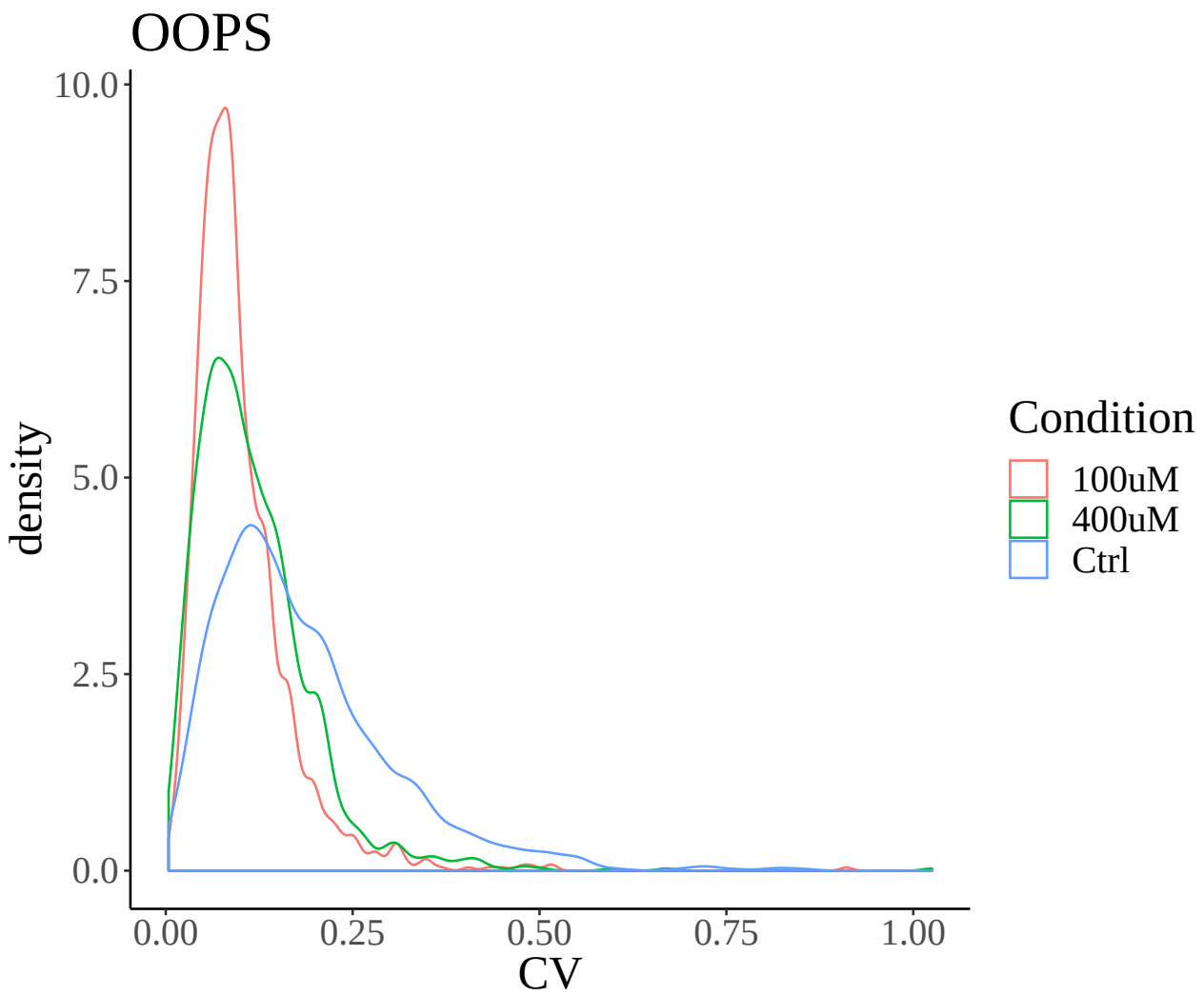
```
total_protein_quant <- readRDS("../results/total_as_res_pro_agg_norm")
oops_protein_quant <- readRDS("../results/rbp_as_res_pro_agg_norm")
```

A reminder of the % CV for the Total proteome and OOPS data

```
p_total <- plotCVs(total_protein_quant)
print(p_total$p + ggtitle("Total"))
```




```
# print(p_total + aes(CV_log))

p_oops <- plotCVs(oops_protein_quant)
print(p_oops$p + ggtitle("OOPS"))
```




```
# print(p_oops + aes(CV_log))
```

3. Simulating a dataset

Make two simulation data sets, one with 2% CV, the other with observed CV (8% & 9% for Total and OOPS, respectively). We'll just use the median CV for each data set.

```
# Median CV calculation
median_total_CV <- median(p_total$CVs$CV)
median_oops_CV <- median(p_oops$CVs$CV)

total_abundances <- 2^p_total$CVs$value
oops_abundances <- 2^p_oops$CVs$value

n_sims <- 100
max_reps <- 10

simulated_df_2_perc <- getSimulatedDF()
```

```
##   Replicate Protein_Ix Intensity  Type Condition diff
## 1         1          1  6.457367 Total   Control 0.25
```

```
## 2         2          1  6.446246 Total   Control 0.25
## 3         3          1  6.460380 Total   Control 0.25
## 4         4          1  6.479182 Total   Control 0.25
## 5         5          1  6.453655 Total   Control 0.25
## 6         6          1  6.478046 Total   Control 0.25
##
##             Total  OOPS
##   Control   21000 21000
##   Treatment 21000 21000
```

```
simulated_df_observed_cv <- getSimulatedDF(median_total_CV = median_total_CV,
    median_oops_CV = median_oops_CV)
```

```
##   Replicate Protein_Ix Intensity  Type Condition diff
## 1         1          1  7.026725 Total   Control 0.25
## 2         2          1  7.282917 Total   Control 0.25
## 3         3          1  7.177603 Total   Control 0.25
## 4         4          1  7.195054 Total   Control 0.25
## 5         5          1  7.148291 Total   Control 0.25
## 6         6          1  7.333951 Total   Control 0.25
##
##             Total  OOPS
##   Control   21000 21000
##   Treatment 21000 21000
```

```
print(dim(simulated_df_2_perc))
```

```
## [1] 84000     6
```

```
print(dim(simulated_df_observed_cv))
```

```
## [1] 84000     6
```

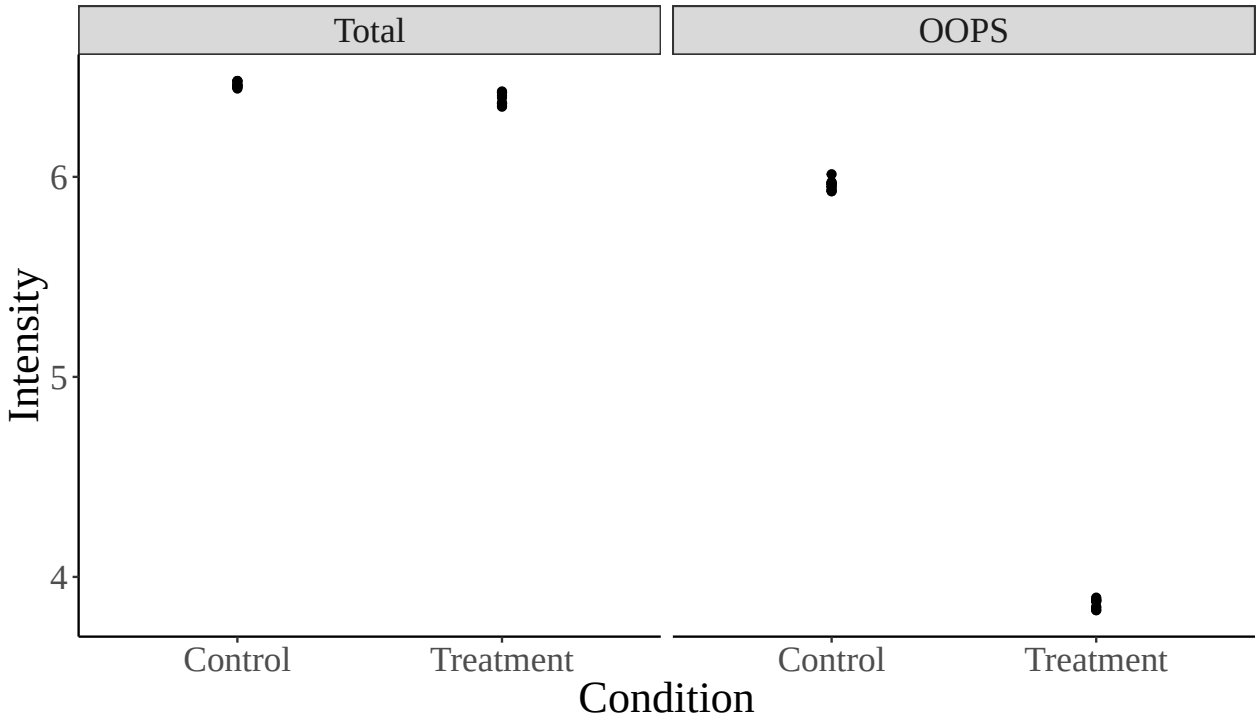
Example simulated protein abundances at 2% CV and 8/9% CV (observed)

```
print(simulated_df_2_perc %>% filter(Protein_Ix ==
    1))
```

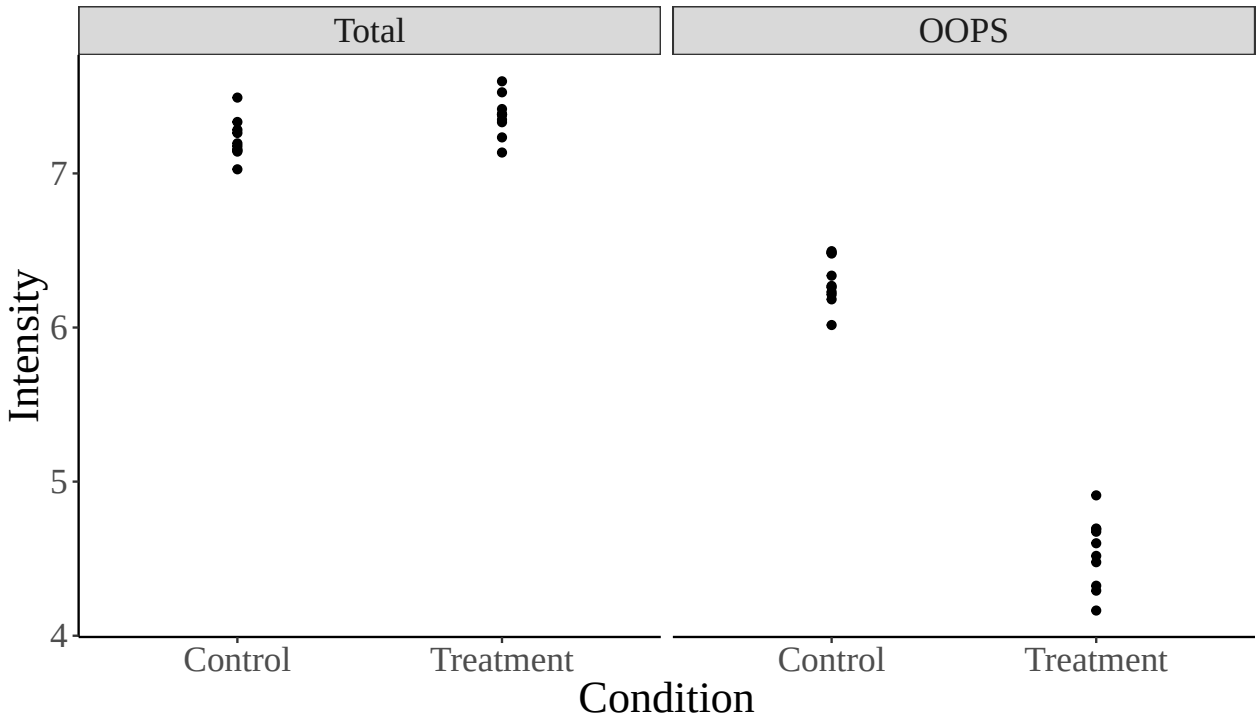
```
##    Replicate Protein_Ix Intensity  Type Condition diff
## 1          1          1  6.457367 Total   Control 0.25
## 2          2          1  6.446246 Total   Control 0.25
## 3          3          1  6.460380 Total   Control 0.25
## 4          4          1  6.479182 Total   Control 0.25
## 5          5          1  6.453655 Total   Control 0.25
## 6          6          1  6.478046 Total   Control 0.25
## 7          7          1  6.461850 Total   Control 0.25
## 8          8          1  6.453713 Total   Control 0.25
## 9          9          1  6.441395 Total   Control 0.25
## 10        10          1  6.479467 Total   Control 0.25
## 11         1          1  6.405608 Total Treatment 0.25
## 12         2          1  6.367521 Total Treatment 0.25
## 13         3          1  6.421630 Total Treatment 0.25
## 14         4          1  6.423075 Total Treatment 0.25
## 15         5          1  6.369032 Total Treatment 0.25
## 16         6          1  6.427138 Total Treatment 0.25
## 17         7          1  6.354948 Total Treatment 0.25
## 18         8          1  6.350664 Total Treatment 0.25
```

```
## 19         9          1  6.368215 Total Treatment 0.25
## 20        10          1  6.395788 Total Treatment 0.25
## 21         1          1  5.964947  OOPS   Control 0.25
## 22         2          1  5.928254  OOPS   Control 0.25
## 23         3          1  5.974137  OOPS   Control 0.25
## 24         4          1  5.970610  OOPS   Control 0.25
## 25         5          1  6.013134  OOPS   Control 0.25
## 26         6          1  5.962415  OOPS   Control 0.25
## 27         7          1  5.950149  OOPS   Control 0.25
## 28         8          1  5.936419  OOPS   Control 0.25
## 29         9          1  5.951263  OOPS   Control 0.25
## 30        10          1  5.928188  OOPS   Control 0.25
## 31         1          1  3.876568  OOPS Treatment 0.25
## 32         2          1  3.890909  OOPS Treatment 0.25
## 33         3          1  3.851315  OOPS Treatment 0.25
## 34         4          1  3.846134  OOPS Treatment 0.25
## 35         5          1  3.883921  OOPS Treatment 0.25
## 36         6          1  3.833808  OOPS Treatment 0.25
## 37         7          1  3.833791  OOPS Treatment 0.25
## 38         8          1  3.882301  OOPS Treatment 0.25
## 39         9          1  3.881435  OOPS Treatment 0.25
## 40        10          1  3.896521  OOPS Treatment 0.25
```

```
simulated_df_2_perc %>% filter(Protein_Ix == 1) %>%
    ggplot(aes(Condition, Intensity)) + geom_point() +
    my_theme + facet_wrap(~Type)
```




```
simulated_df_observed_cv %>% filter(Protein_Ix == 1) %>%
    ggplot(aes(Condition, Intensity)) + geom_point() +
    my_theme + facet_wrap(~Type)
```



Run linear model on simulated data with differing numbers of replicates

```
# Linear model on low CV data
simulated_lm_results_2_perc <- runLM_multiple_reps(simulated_df_2_perc)
```

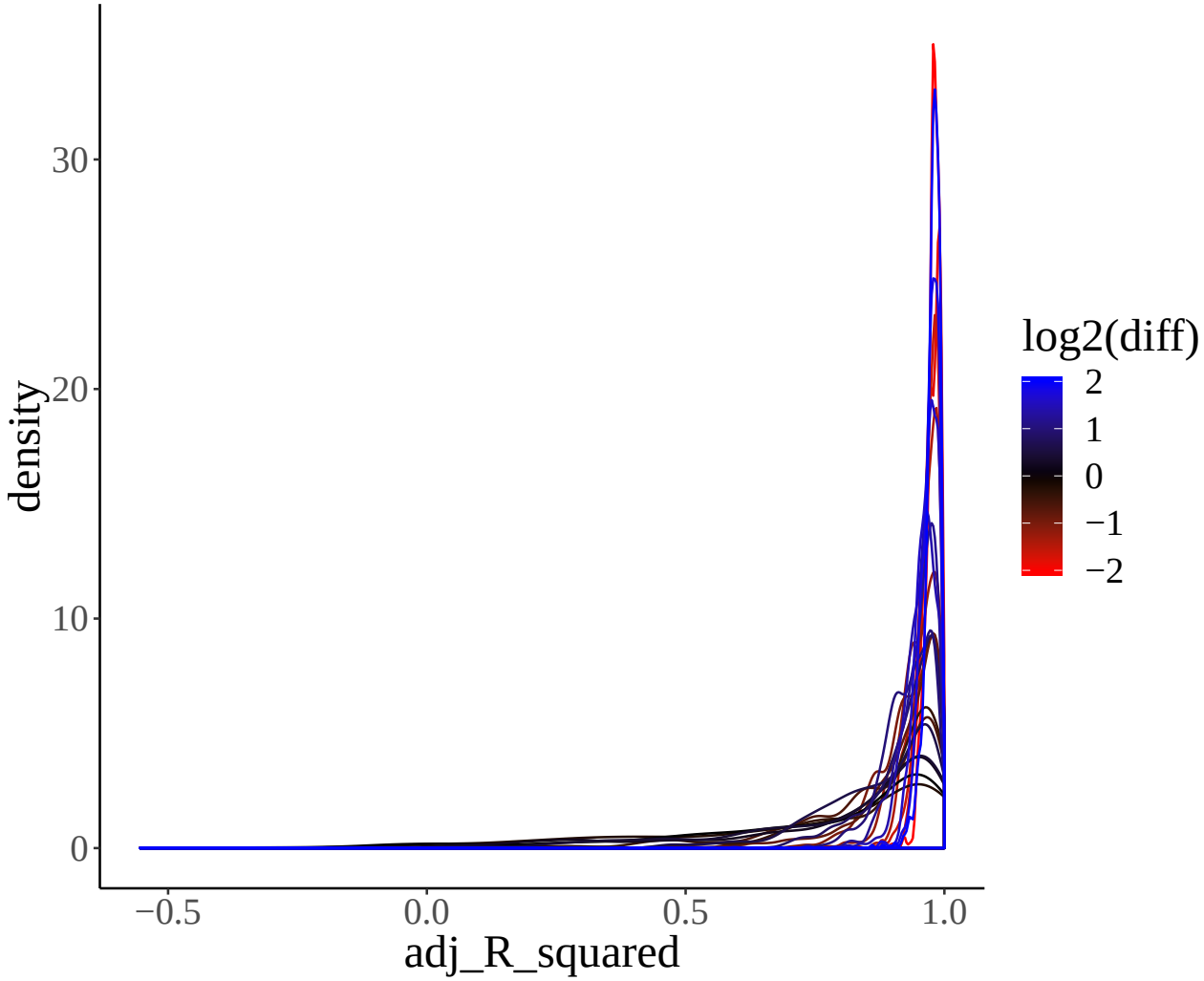
```
##   Protein_Ix Estimate Std. Error   t value      p_value adj_R_squared
## 1       2001 2.033742 0.02373922  85.67014 3.876387e-27     0.9983397
## 2       2001 2.057325 0.03536771  58.16957 5.230132e-07     0.9988550
## 3       2001 2.028396 0.02231582  90.89499 5.627983e-32     0.9982619
## 4       2001 2.017543 0.02040352  98.88211 2.304457e-41     0.9981004
## 5       2001 2.057881 0.02909659  70.72582 4.242927e-17     0.9983798
## 6       2001 2.034991 0.02017342 100.87485 2.054114e-37     0.9983696
##             BH n_reps diff
## 1 4.599104e-26      6    4
## 2 4.377740e-06      2    4
## 3 5.879982e-31      7    4
## 4 1.967219e-40      9    4
## 5 3.977744e-16      4    4
## 6 3.081170e-36      8    4
```

```
# Linear model on observed CV data
simulated_lm_results_observed_cv <- runLM_multiple_reps(simulated_df_observed_cv)
```

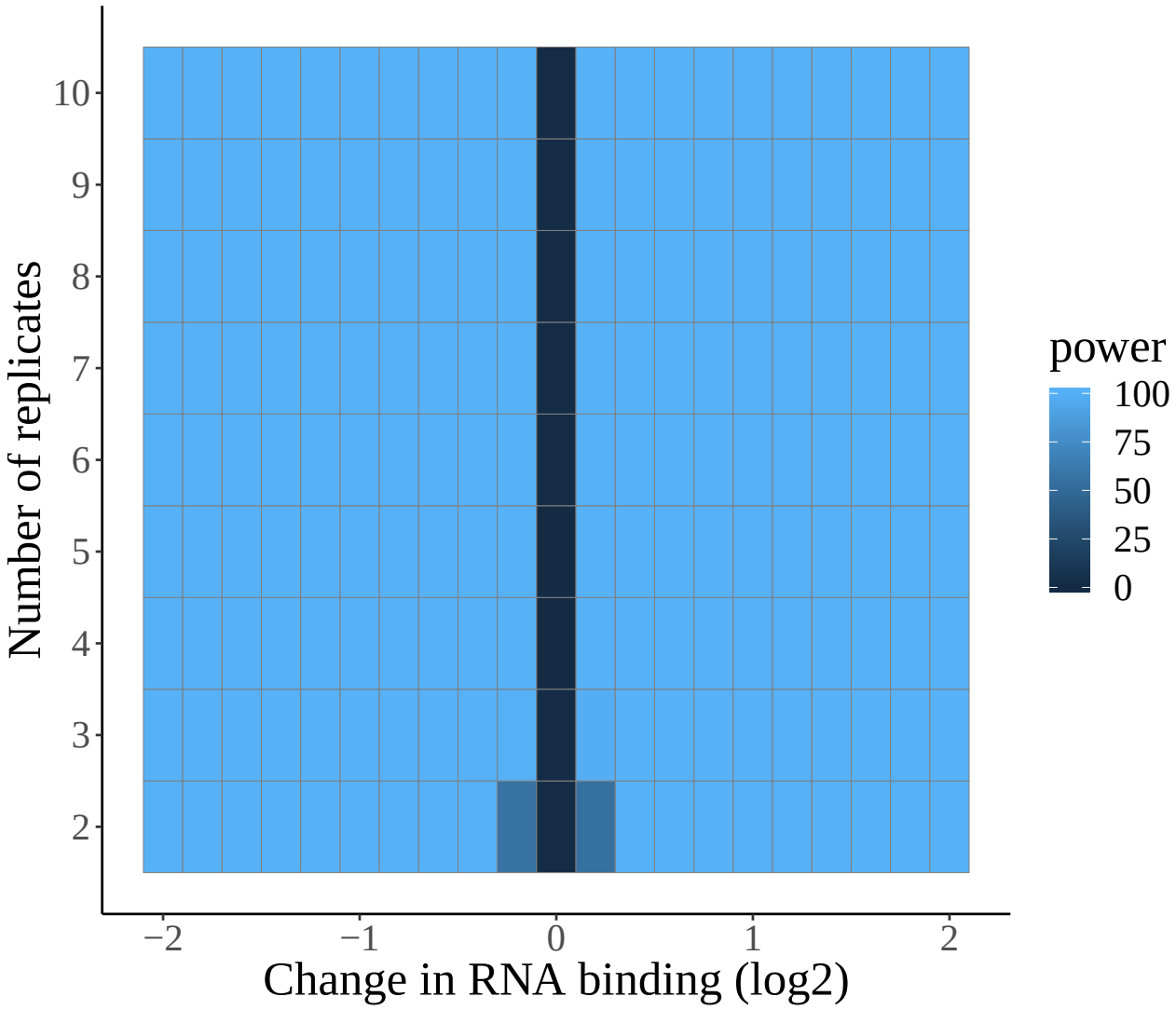
```
##   Protein_Ix Estimate Std. Error   t value      p_value adj_R_squared
## 1       2001 2.037314 0.10429140 19.534822 1.691704e-14     0.9741390
## 2       2001 2.030711 0.24620719  8.247976 1.178574e-03     0.9539845
## 3       2001 2.074395 0.11039220 18.791138 7.335403e-16     0.9659535
## 4       2001 2.056613 0.10069319 20.424545 6.382602e-20     0.9638025
## 5       2001 2.041974 0.13960993 14.626279 5.184473e-09     0.9698174
## 6       2001 2.034776 0.09918191 20.515600 2.078582e-18     0.9676138
##             BH n_reps diff
## 1 3.552579e-13      6    4
## 2 3.867022e-03      2    4
```

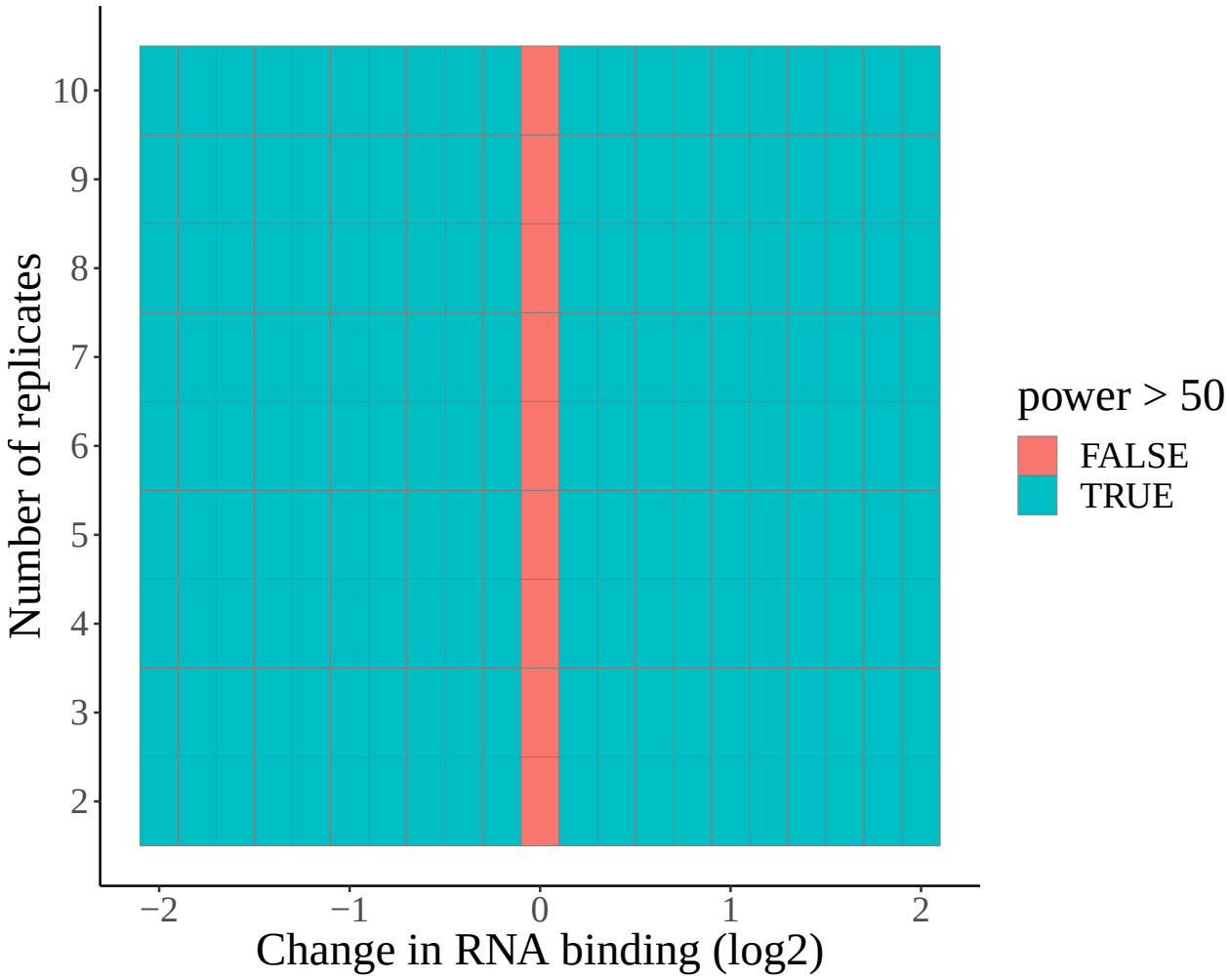
```
## 3 7.940385e-15      7    4
## 4 5.930737e-19      9    4
## 5 4.698827e-08      4    4
## 6 2.582854e-17      8    4
```

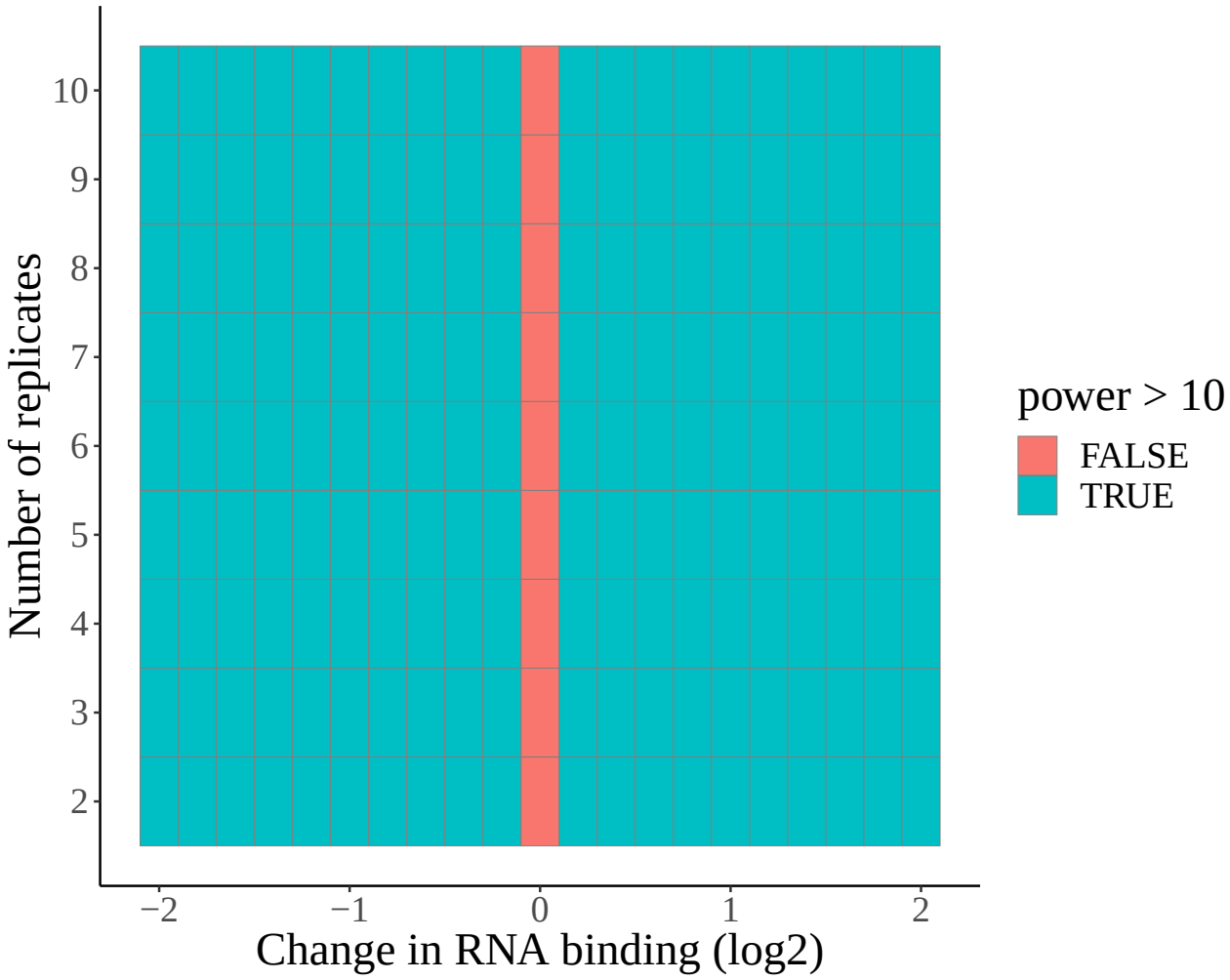
```
# Plot of simulated data using observed CV
ggplot(simulated_lm_results_observed_cv, aes(adj_R_squared,
    colour = log2(diff), group = log2(diff))) + geom_density() +
    scale_colour_gradient2(high = "blue", low = "red",
        mid = "black") + my_theme
```




```
# 2% CV simulated data
p <- plot_results(simulated_lm_results_2_perc)
```


Number of replicates
10
9
8
7
6
5
4
3
2
−2
−1
0
1
2
Change in RNA binding (log2)
power
100
75
50
25
0


Number of replicates
10
9
8
7
6
5
4
3
2
−2
−1
0
1
2
Change in RNA binding (log2)
power > 50
FALSE
TRUE

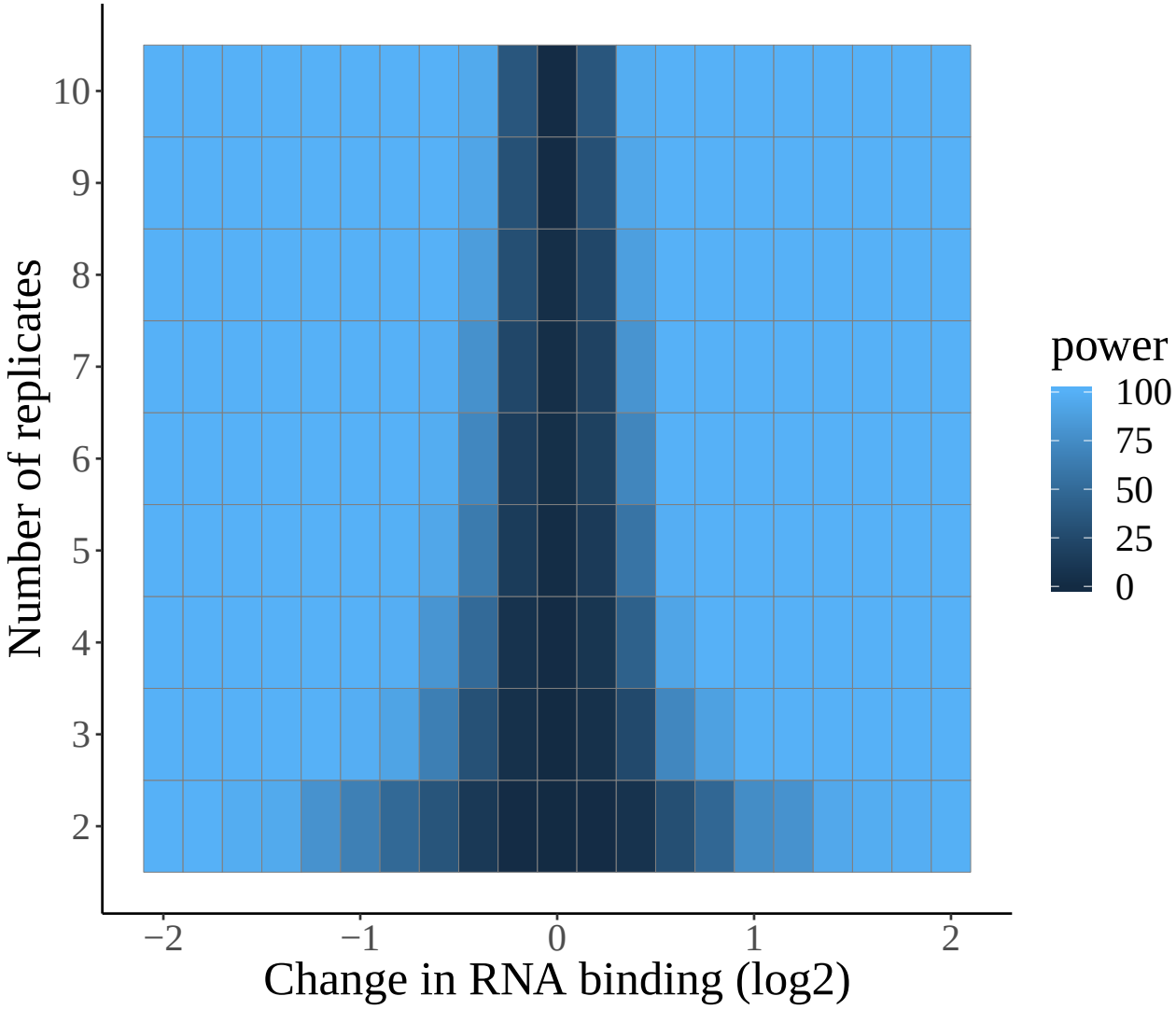


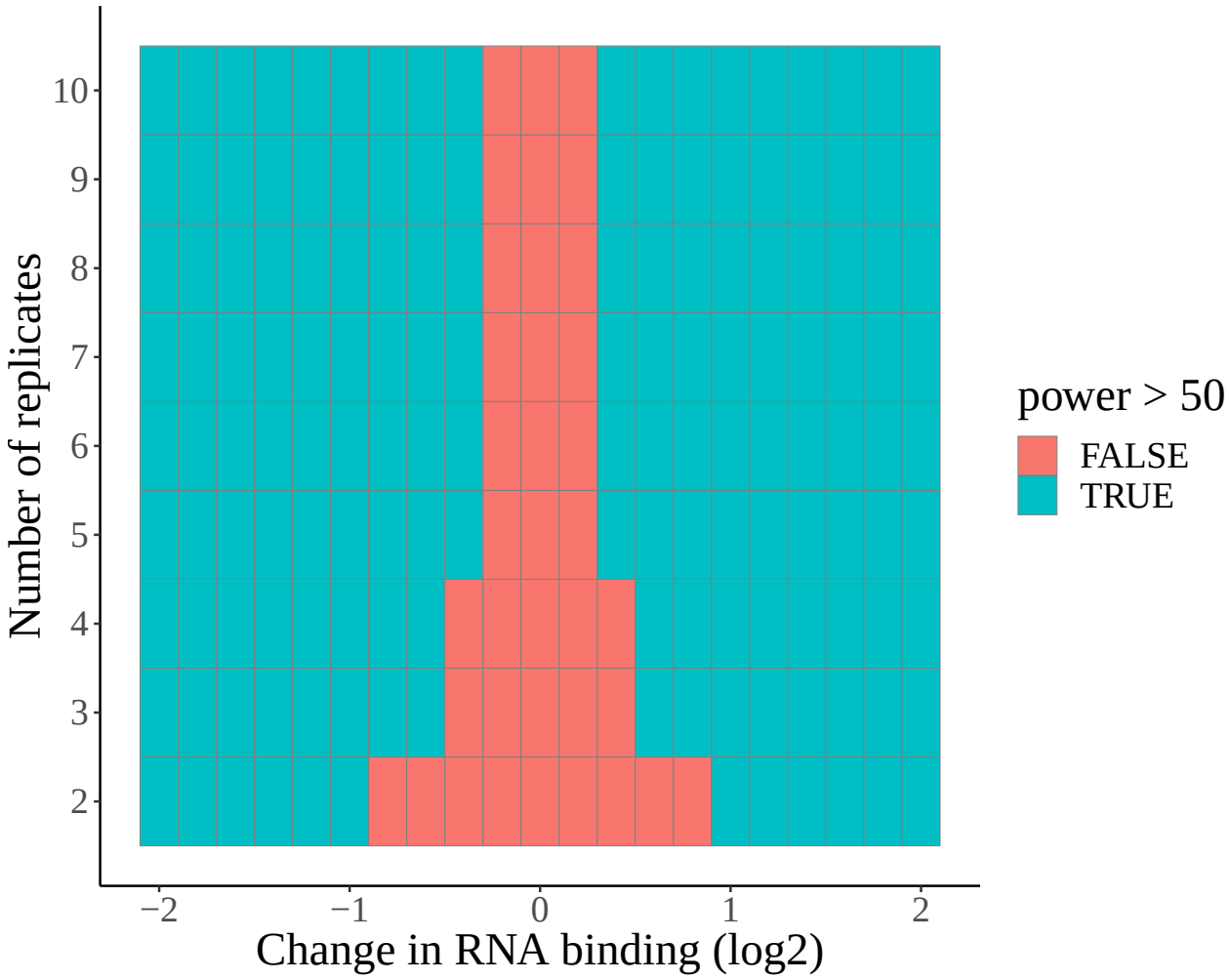

```
ggsave("../plots/power_simulations_2_perc.png", p)

# Observed CV simulated data with filtering for p
# <0.01
simulated_lm_results_observed_cv %>% group_by(n_reps,
    diff) %>% dplyr::summarise(power = sum(p_value <
    0.01))
```

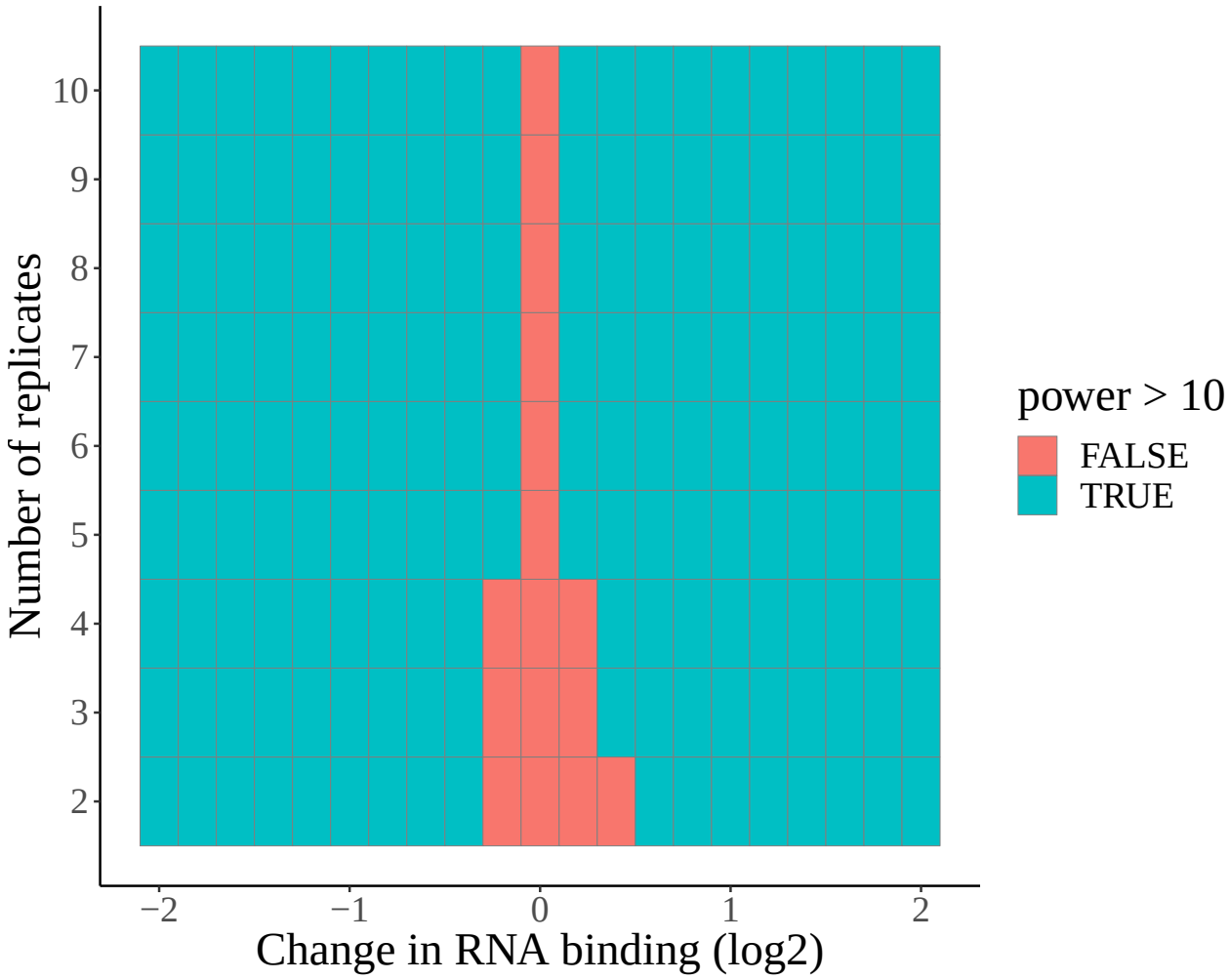
```
## # A tibble: 189 x 3
## # Groups:   n_reps [?]
##    n_reps  diff power
##     <int> <dbl> <int>
##  1      2 0.25    100
##  2      2 0.287   100
##  3      2 0.330    97
##  4      2 0.379    95
##  5      2 0.435    79
##  6      2 0.5      66
##  7      2 0.574    49
##  8      2 0.660    34
##  9      2 0.758    12
## 10      2 0.871     1
```

```
## # ... with 179 more rows
```

```
# Plot of the above
p <- plot_results(simulated_lm_results_observed_cv)
```




Number of replicates
10
9
8
7
6
5
4
3
2
−2
−1
0
1
2
Change in RNA binding (log2)
power > 50
FALSE
TRUE




```
ggsave("../plots/power_simulations_observed_cvs.png",
    p)
```

4. Conclusions

So, with the observed CVs, it's hard to detect small changes in RNA binding (<2-fold). This is because we're looking at an interaction term so we require a lot more replicates to reach the same statistical power. Let's try an alternative approach where we just identify the changes in OOPS abundance and then classify the proteins according to their observed changes in total. Note, this is a less statisticall rigorous approach and the results will require a more cautious interpretation.

```
simulated_lm_results_2_perc_simple <- simulated_df_2_perc %>%
    filter(Type == "OOPS") %>% runLM_multiple_reps(reps = 2:10,
    model = "Intensity~Condition", coeff_of_interest = "ConditionTreatment")
```

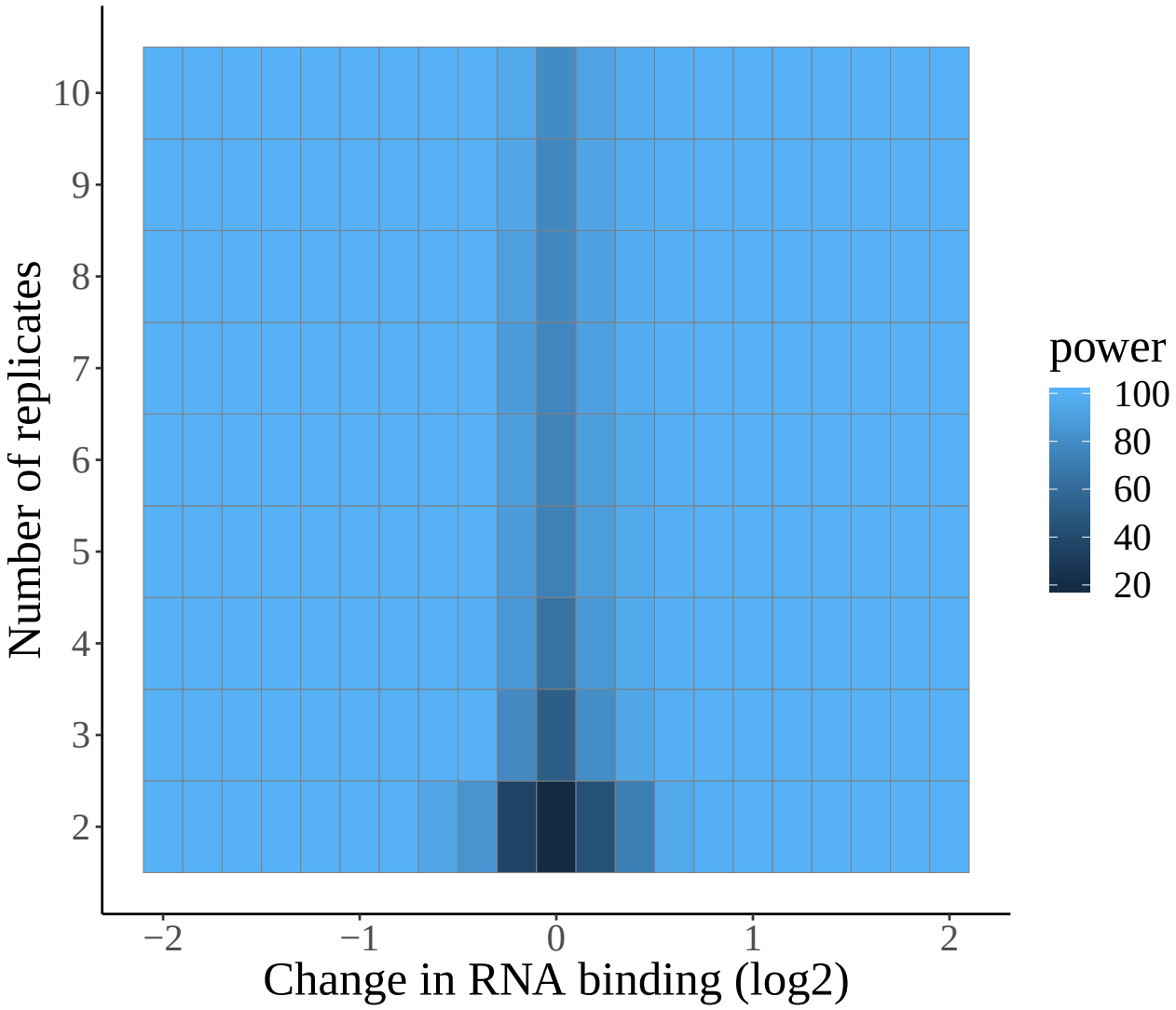
```
##   Protein_Ix Estimate  Std. Error   t value      p_value adj_R_squared
## 1       2001 1.848496 0.013575257 136.16653 1.120272e-17     0.9994071
## 2       2001 1.834030 0.009896695 185.31740 2.911713e-05     0.9999126
## 3       2001 1.841843 0.015038580 122.47454 5.887280e-20     0.9991340
## 4       2001 1.834271 0.014023990 130.79525 1.141614e-25     0.9990072
## 5       2001 1.851030 0.019507504  94.88811 9.231524e-11     0.9992231
## 6       2001 1.843882 0.013756721 134.03496 3.637057e-23     0.9991657
```

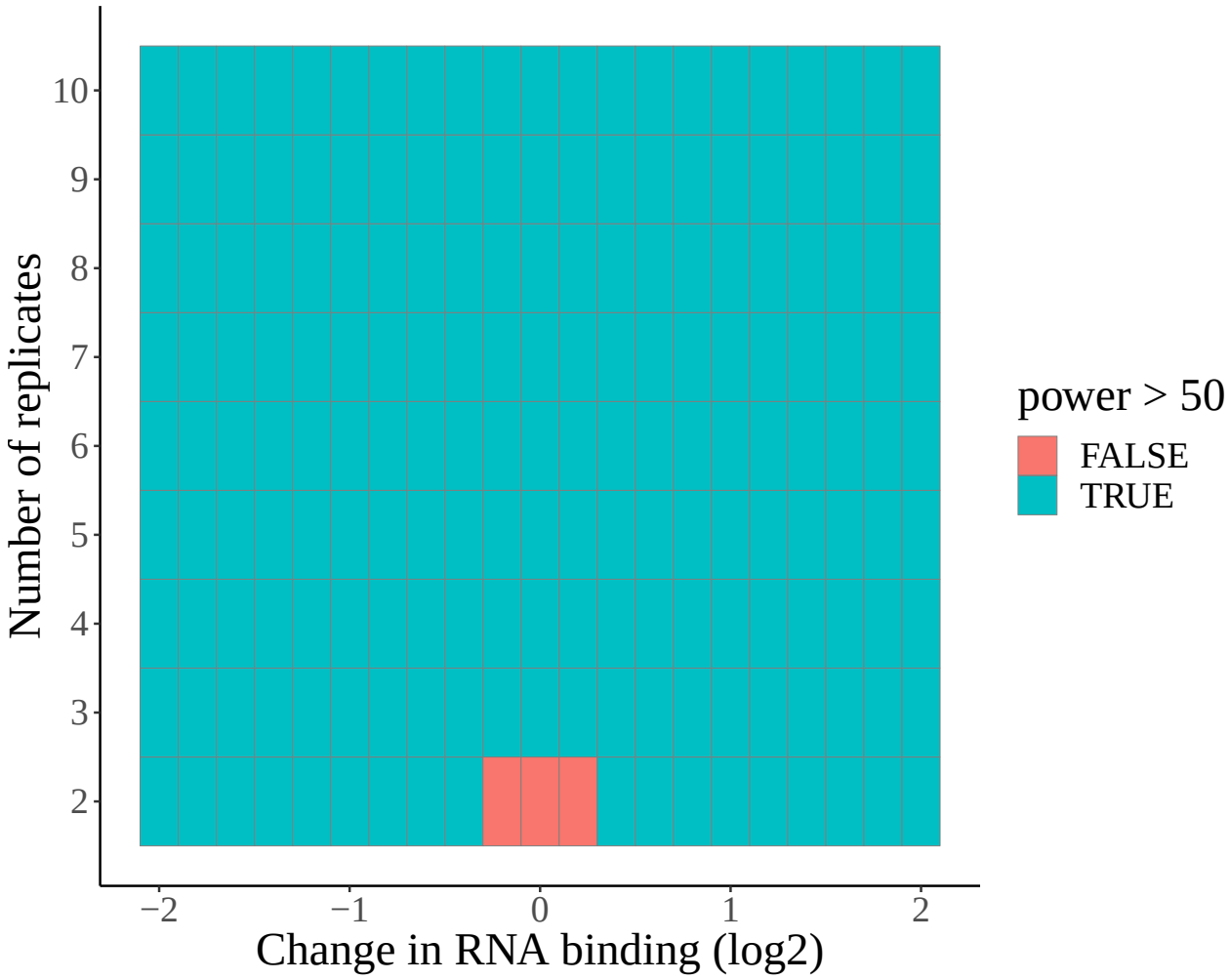
```
##             BH n_reps diff
## 1 1.618028e-16      6    4
## 2 5.164286e-04      2    4
## 3 4.154186e-19      7    4
## 4 6.393037e-25      9    4
## 5 5.296776e-10      4    4
## 6 2.871361e-22      8    4
```

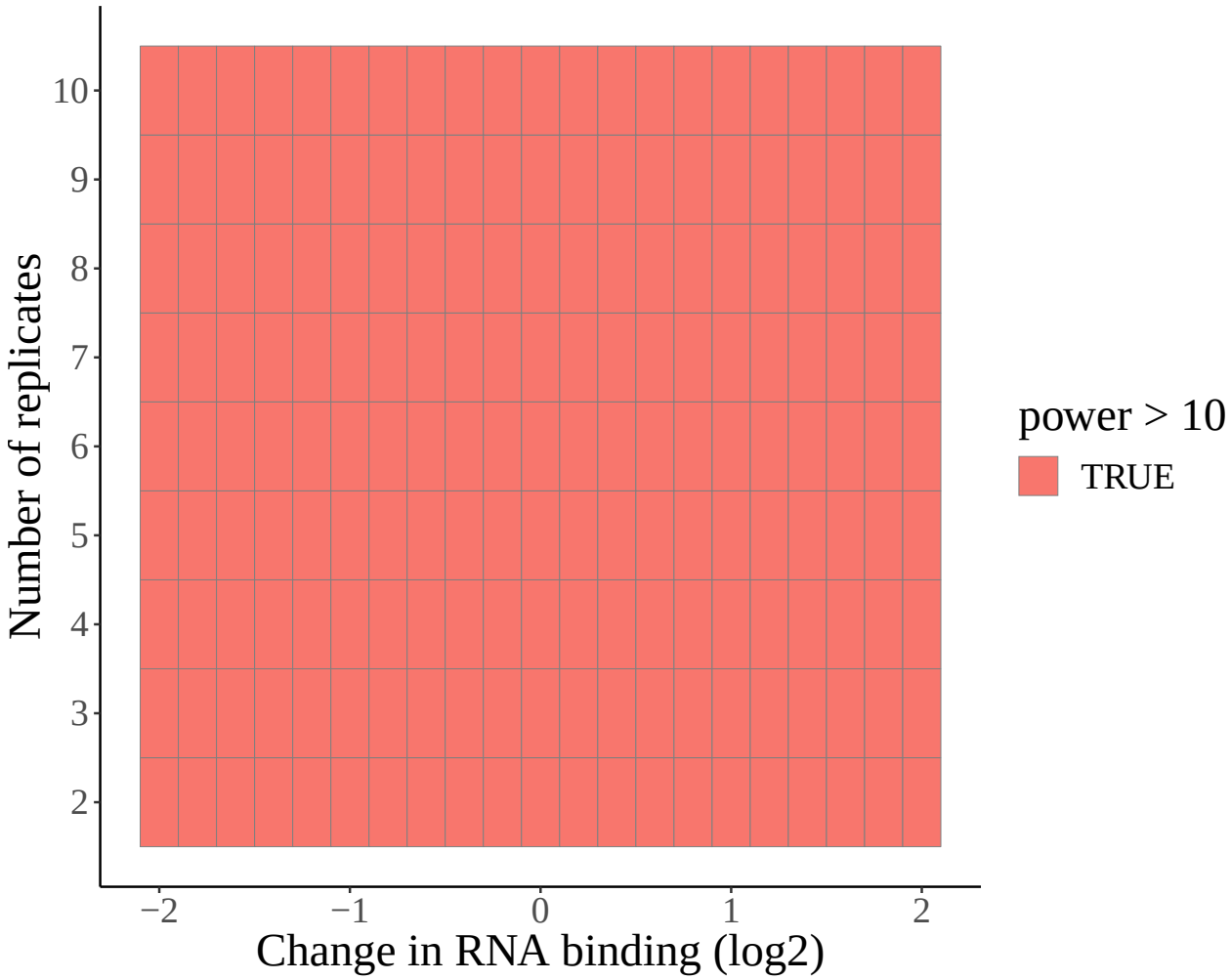
```
simulated_lm_results_observed_cv_simple <- simulated_df_observed_cv %>%
    filter(Type == "OOPS") %>% runLM_multiple_reps(reps = 2:10,
    model = "Intensity~Condition", coeff_of_interest = "ConditionTreatment")
```

```
##   Protein_Ix Estimate Std. Error  t value      p_value adj_R_squared
## 1       2001 2.094002 0.08460376 24.75070 2.648454e-10     0.9823321
## 2       2001 2.008210 0.24189352  8.30204 1.420045e-02     0.9577011
## 3       2001 2.073788 0.07384033 28.08476 2.571658e-12     0.9837653
## 4       2001 2.074202 0.06435483 32.23071 5.540956e-16     0.9838835
## 5       2001 2.069201 0.11062664 18.70437 1.507494e-06     0.9803290
## 6       2001 2.052749 0.06626730 30.97681 2.680567e-14     0.9845927
##             BH n_reps diff
## 1 2.598951e-09      6    4
## 2 2.903324e-02      2    4
## 3 3.704229e-11      7    4
## 4 9.861024e-15      9    4
## 5 8.714407e-06      4    4
## 6 4.908243e-13      8    4
```

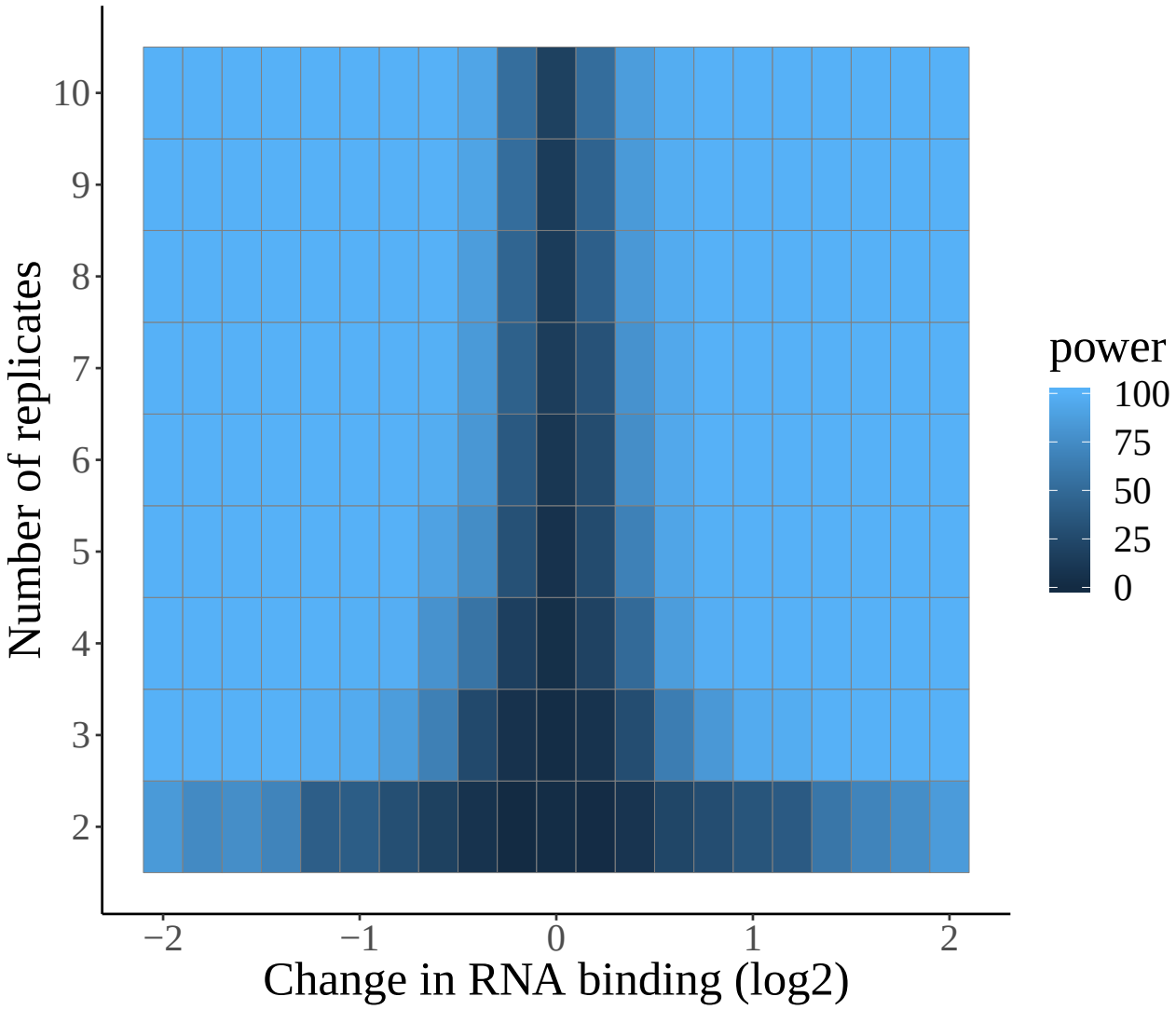
```
# Based on 2% CV
plot_results(simulated_lm_results_2_perc_simple)
```

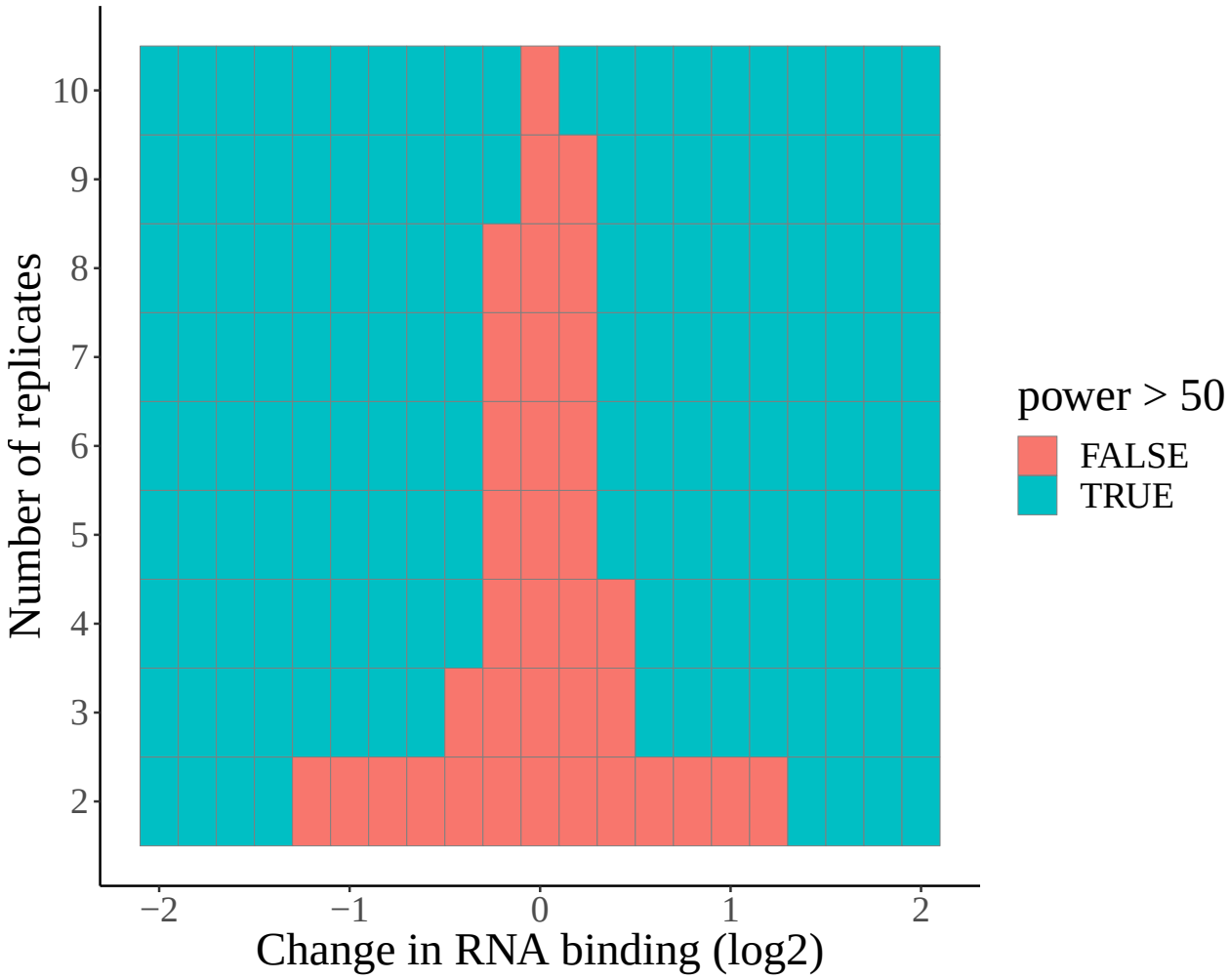

Number of replicates
10
9
8
7
6
5
4
3
2
−2
−1
0
1
2
Change in RNA binding (log2)
power
100
80
60
40
20

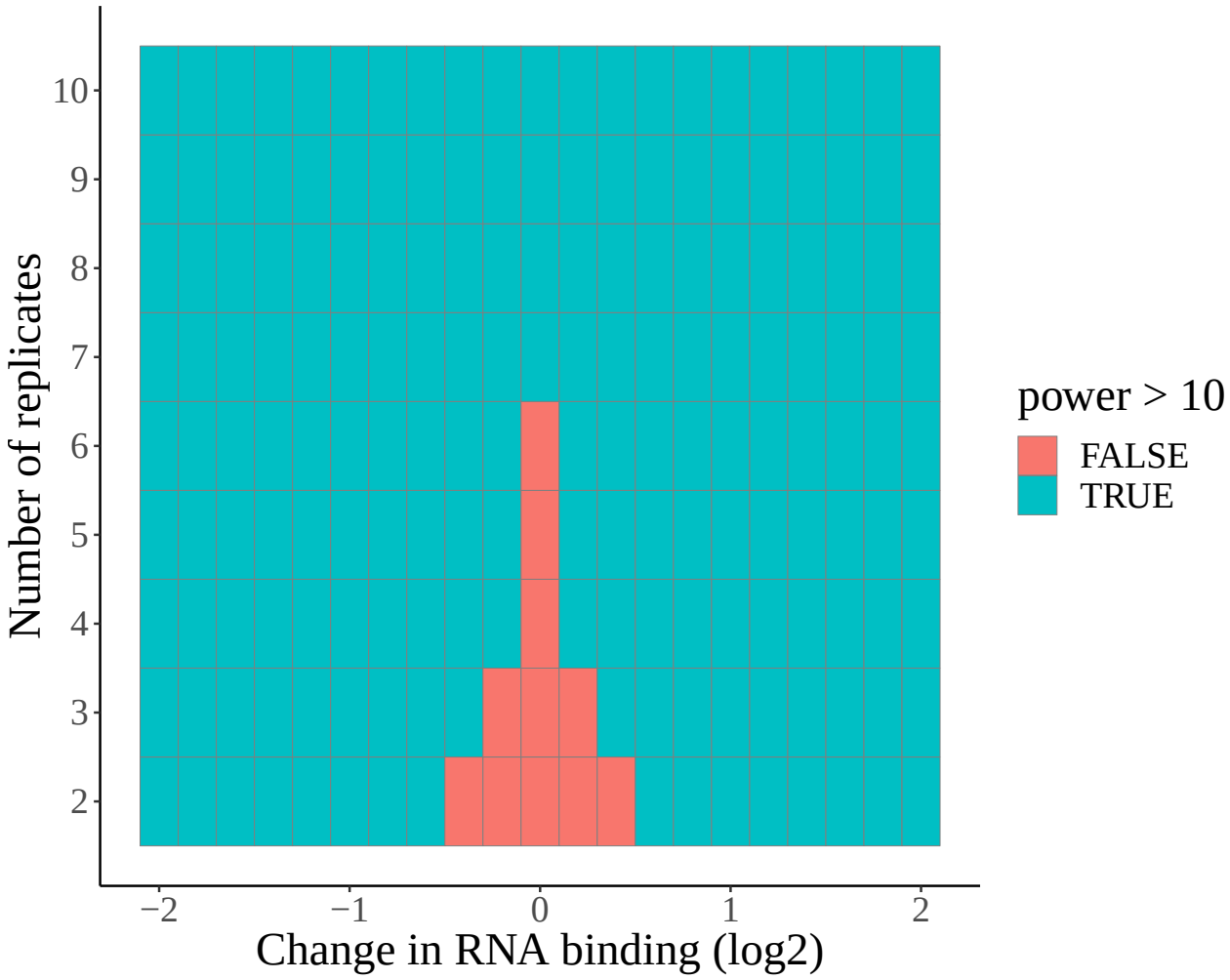

Number of replicates
10
9
8
7
6
5
4
3
2
−2
−1
0
1
2
Change in RNA binding (log2)
power > 50
FALSE
TRUE




```
# Based on observed CV of 8-9%
plot_results(simulated_lm_results_observed_cv_simple)
```


Number of replicates
10
9
8
7
6
5
4
3
2
−2
−1
0
1
2
Change in RNA binding (log2)
power
100
75
50
25
0


Number of replicates
10
9
8
7
6
5
4
3
2
−2
−1
0
1
2
Change in RNA binding (log2)
power > 50
FALSE
TRUE


Number of replicates
10
9
8
7
6
5
4
3
2
−2
−1
0
1
2
Change in RNA binding (log2)
power > 10
FALSE
TRUE