Sars-Cov2 (and OC43) screening comparison of HEK293 cells

April 8, 2021

Contents

Source code is provided in file $\text{sars}_{\text{cov2analysis.py}}$. The data used in this analysis might be behind or ahead of what will be published in peer-reviewed manner.

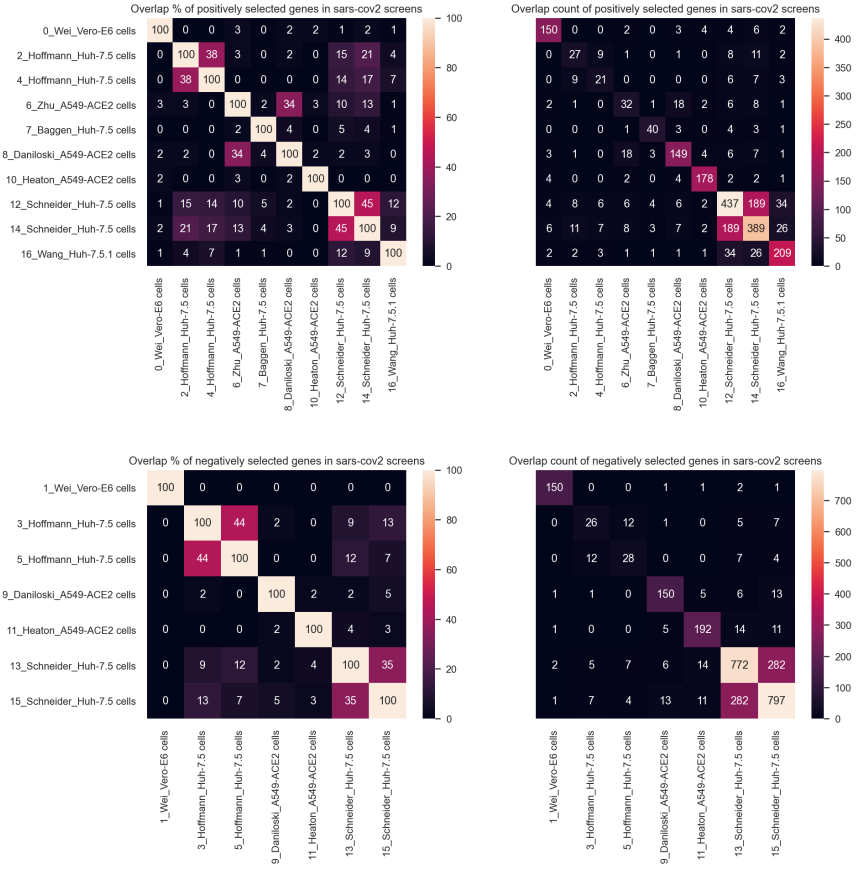
I use The COVID-19 Drug and Gene Set Library website for comparing gene sets. The website only lists "top hits" for each data set, which does not allow us to reliably report novel detected genes. However, it's a good start.

Here is an overview of the data that is provided by the website.

	author	Cell/Tissue	selection	temperature	#(reported hits)
0	Wei	Vero-E6 cells	positive	37	150
1	Wei	Vero-E6 cells	negative	37	150
2	Hoffmann	Huh-7.5 cells	positive	33	27
3	Hoffmann	Huh-7.5 cells	negative	33	26
4	Hoffmann	Huh-7.5 cells	positive	37	21
5	Hoffmann	Huh-7.5 cells	negative	37	28
6	Zhu	A549-ACE2 cells	positive	37	32
7	Baggen	Huh-7.5 cells	positive	37	40
8	Daniloski	A549-ACE2 cells	positive	37	149
9	Daniloski	A549-ACE2 cells	negative	37	150
10	Heaton	A549-ACE2 cells	positive	37	178
11	Heaton	A549-ACE2 cells	negative	37	192
12	Schneider	Huh-7.5 cells	positive	37	437
13	Schneider	Huh-7.5 cells	negative	37	772
14	Schneider	Huh-7.5 cells	positive	33	389
15	Schneider	Huh-7.5 cells	negative	33	797
16	Wang	Huh-7.5.1 cells	positive	37	209

1 Comparison based on reported gene sets

The website provides gene sets which I can easily compare.



Top 20 genes, when counting the number of occurrences per gene in published gene sets:

```
ACE2,RAB10,CTSL,SCAP,CCZ1B,WDR81,EIF4E2,CSNK2B,HS2ST1,EXT1,TAF6L,UGP2,NDST1,ACTR2,B3GAL
```

```
PRIM2     4
MRPS25    4
MIB1      4
MRPS2     3
FBL       3
          ..
TOMM70    2
ACAD9     2
CHRNB1    2
MASP1     2
```

```
WFDC12    2
Length: 331, dtype: int64
```

2 HEK293/SARS-CoV2 analysis

Since there is no standard procedure for reporting a set of top hits, it is difficult to compare reported gene sets from different groups.

In order to allow us to compare screens from different groups, we downloaded the raw data of 7 screens from 4 publications and analyzed them in the same way (MAGeCK-VISPR). I downloaded the raw data for 4 of these data sets (Hoffmann, Daniloski, Schneider, Wang). For other (preprint-)published papers, data was not provided.

For now, I also only focus on HEK293-data (and ignore VeroE6 data).

2.1 Published datasets

- Hoffmann et al. were using "An sgRNA library targeting 332 interactome genes along with 314 safe and 310 essential control sgRNAs was used to create a targeted lentiviral library." This limits comparison to those 332 genes.

- I think the screens of Schneider and Hoffmann were done in collaboration. They were sequenced on the exact same day and several authors appear on both applications. Schneider et al. used the Brunello library though. Note: Schneider only had a 37 degree control (also for the 33 degree infections). Hoffmann had dedicated control samples for both 33 and 37 degrees.

- Both Hoffmann and Schneider worked with Huh-7.5 cells

- Daniloski used the GeCKOv2 library (Set A and B combined) and worked in A549-ACE2 cells.

- Wang et al. used the GeCKOv2 library in Huh-7.5 cells, but used set A and set B separately. I'll analyze the data as if they combined them. The RRA algorithm should be able to handle it.

- Wei et al. are the only ones that used another organism (C. sabaeus).

2.2 QC of published data sets

All data sets show a mapping rate of ~70% and more and a the majority of sgRNAs were retrieved in all screens.

Comparing the PCA- and the correlation plot of Schneider and of Hoffmann with ours, we can again see how Sars-CoV-2 samples cluster with the control samples, while OC43 samples show more distinct properties.

2.3 Brief overview of top hits of data sets

- For both Schneider and Hoffmann, the OC43 infections lead to a high number of highly siginficant positive-selection hits.

For Sars-Cov2:

- Daniloski shows a couple of positive-selection hits (but no negative ones). ACE2 is the top hit
- Surprisingly, for the Schneider-analysis the non-targeting gRNAs lead to the top hit, so for now I would take the Schneider analysis with a grain of salt
- The Hoffmann dataset leads to a high number of positive- and negative-selection hits.

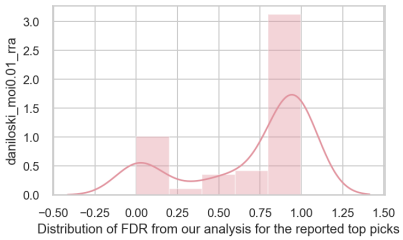
2.4 Positive selection

2.4.1 Verify reported top picks

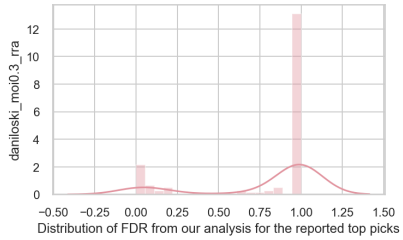
First, I wanted to control, whether the top picks reported on the maayan-lab "Covid19 Drug and Gene Set library" could be reproduced by my analysis of their raw data. I therefore plotted the FDR values from my analysis (of their raw data) for the reported genes. I expected most FDR values to be in the 0-0.2 range, which was the case for Schneider and Hoffmann, however not for Daniloski and Wang.

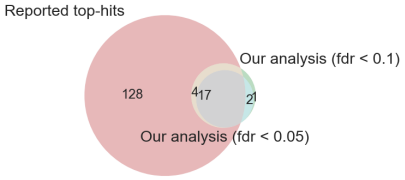
```
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Bad circle positioning")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
```

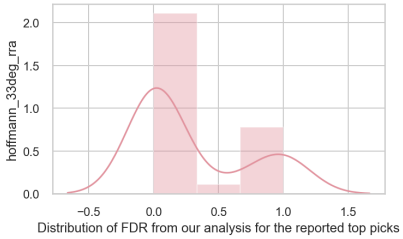
```
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Bad circle positioning")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
```

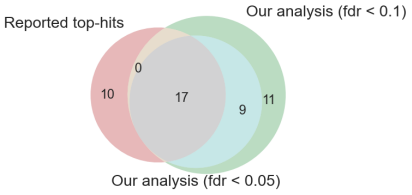

daniloski_moi0.01_rra
Distribution of FDR from our analysis for the reported top picks

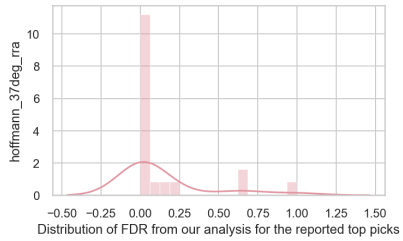

Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
124
23

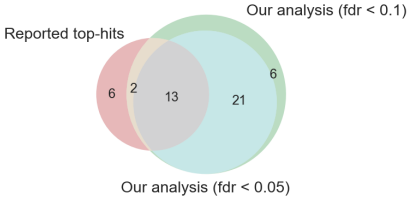

daniloski_moi0.3_rra
Distribution of FDR from our analysis for the reported top picks

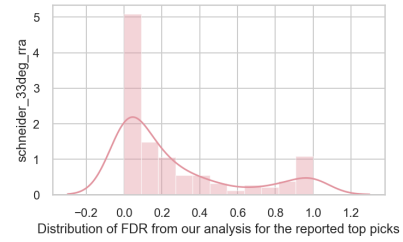

Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
128
4
17

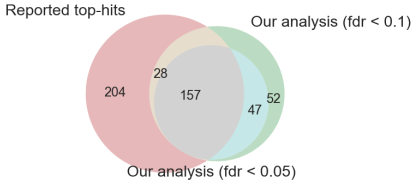

hoffmann_33deg_rra
Distribution of FDR from our analysis for the reported top picks

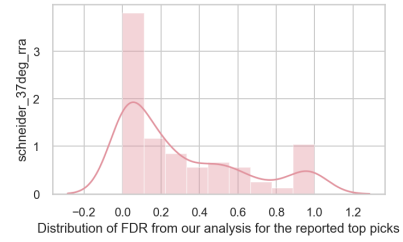

Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
10
0
17
9
11

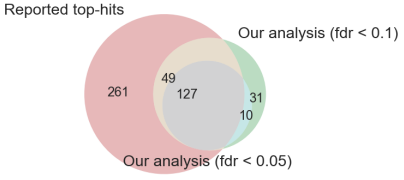

hoffmann_37deg_rra
Distribution of FDR from our analysis for the reported top picks

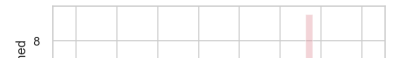

Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
6
2
13
21
6


schneider_33deg_rra
Distribution of FDR from our analysis for the reported top picks


Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
204
28
157
47
52


schneider_37deg_rra
Distribution of FDR from our analysis for the reported top picks


Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
261
49
127
31
10

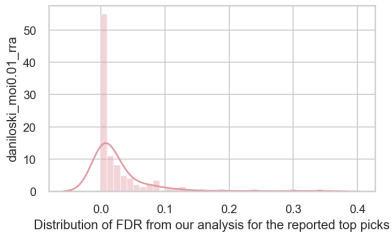


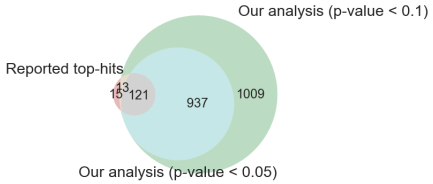

Reported top-hits

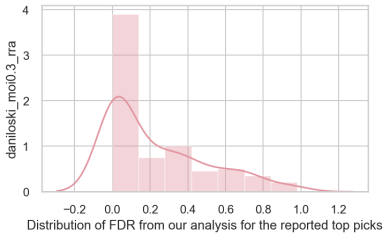
The reason, we don't see sharp peaks at ~0.0 (especially for Wang and daniloski, is that all reported publications seem to use the *p-value* instead of the *FDR* to report their picks. This means that all publications don't apply multiple-testing correction on the *p-value*, which is a common and necessary step to ensure the correct reporting of statistical significance! I'm really surprised that this kind of p-value-hacking is so common in the field. VISPR for example also sorts by the FDR.

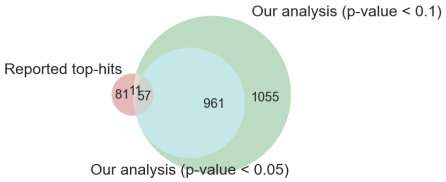
Below, I plot the same analysis, using *p-values* for selection rather than the multiple-testing-corrected *FDR*. Here the reported top picks all have expected values close to 0.0.

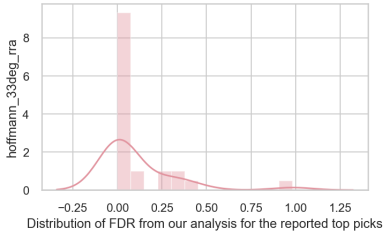
```
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Bad circle positioning")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
```

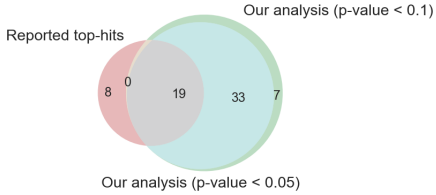

daniloski_mol0.01_rra
Distribution of FDR from our analysis for the reported top picks

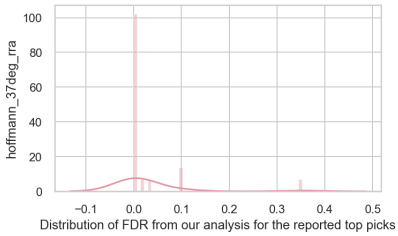

Our analysis (p-value < 0.1)
Reported top-hits
Our analysis (p-value < 0.05)
13
121
937
1009

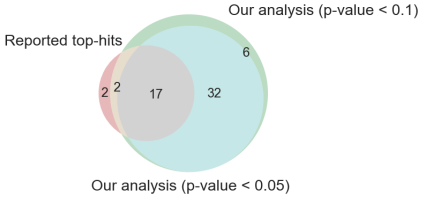

daniloski_mol0.3_rra
Distribution of FDR from our analysis for the reported top picks

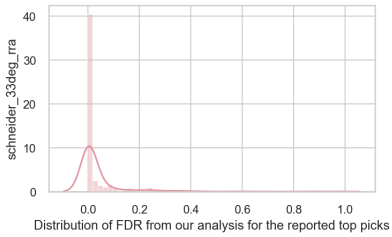

Our analysis (p-value < 0.1)
Reported top-hits
Our analysis (p-value < 0.05)
11
57
961
1055

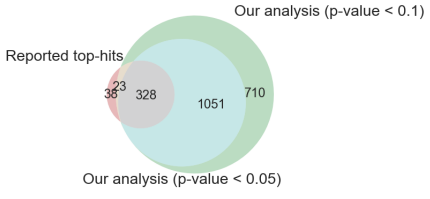

hoffmann_33deg_rra
Distribution of FDR from our analysis for the reported top picks

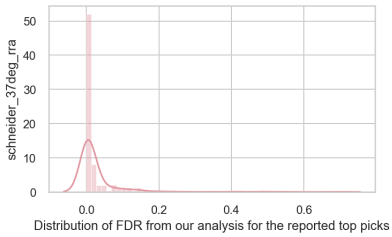

Our analysis (p-value < 0.1)
Reported top-hits
Our analysis (p-value < 0.05)
8
0
19
33
7

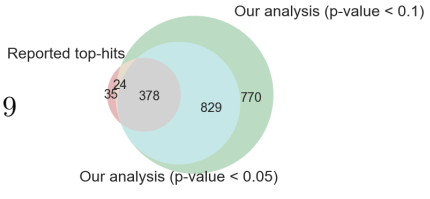

hoffmann_37deg_rra
Distribution of FDR from our analysis for the reported top picks

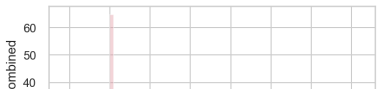

Our analysis (p-value < 0.1)
Reported top-hits
Our analysis (p-value < 0.05)
2
2
17
32
6


schneider_33deg_rra
Distribution of FDR from our analysis for the reported top picks


Our analysis (p-value < 0.1)
Reported top-hits
Our analysis (p-value < 0.05)
23
328
1051
710


schneider_37deg_rra
Distribution of FDR from our analysis for the reported top picks


Our analysis (p-value < 0.1)
Reported top-hits
Our analysis (p-value < 0.05)
24
378
829
770


ombined


Our analysis (p-value < 0.1)
Reported top-hits

2.4.2 Overlap of different studies

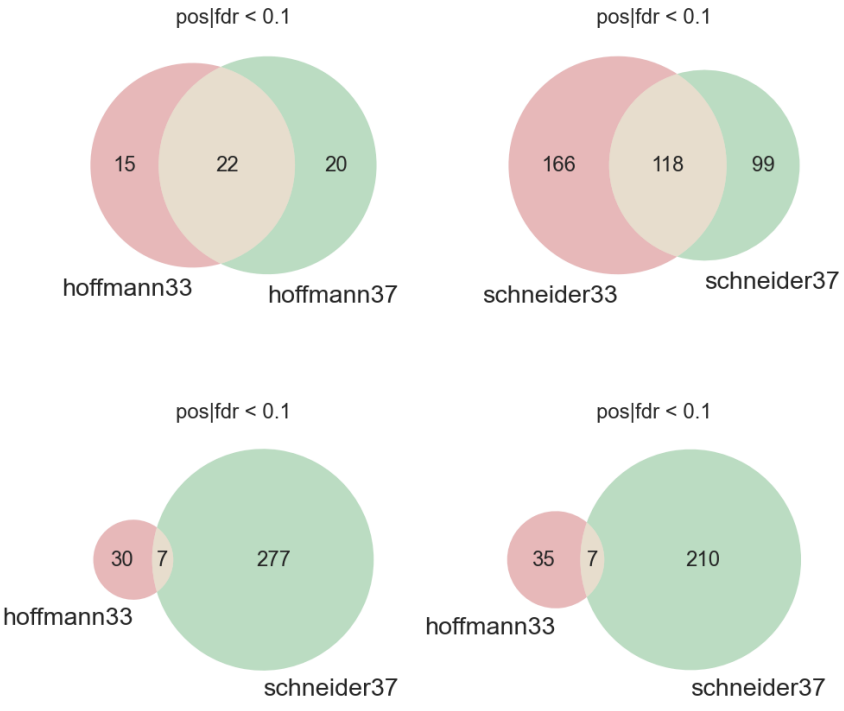
Since different groups reported their top picks with different thresholds, I use my own analysis of their data to pick selection genes and to compare them. I use different FDR and p-value thresholds for comparison. Further, since the differing experimental procedures like sequencing depth etc. can lead to a largely different variation across data sets, FDR and p-values can potentially not be compared across datasets. To overcome this limitation, I further select the top 20, 50, 100, 200 genes from each data set and compare them.

For each comparison, we plot how many genes were identified in how many data sets (multiple barplots).

The Schneider and Hoffmann data sets have the potential to skew the data:

- Hoffmann only considered a subset of all genes
- Both Schneider and Hoffmann do their work in two temperatures (33, 37 degrees), which might be quite similar.

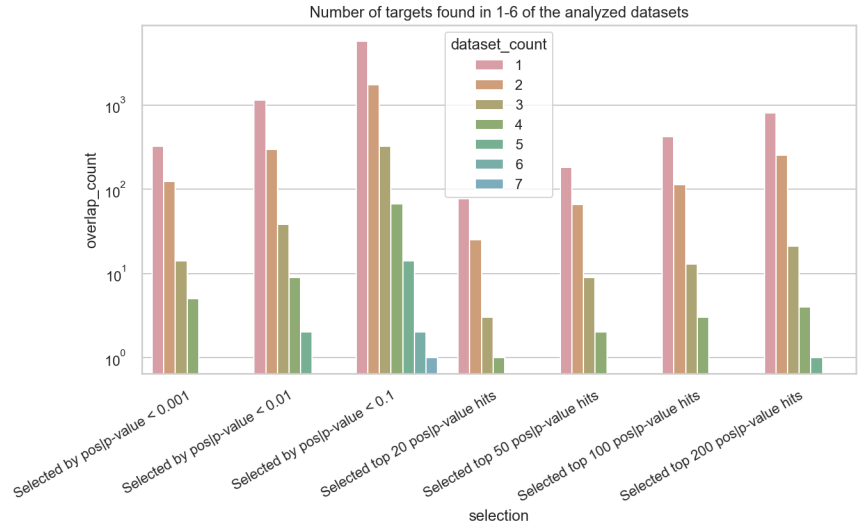
I compare the 33 and 37 degC from Schneider and Hoffmann.



Although the data sets show similarities, there are also variations and I think it's reasonable to include all datasets into the comparison. Here, we use the p-value since we gain additional statistical signifance for targets as they appear in multiple data sets.

- p-value: 0.001, 0.01, 0.1.
- Top n: 20, 50, 100, 200
- positive

```
Text(0.5, 1.0, 'Number of targets found in 1-6 of the analyzed datasets')
```



From this plot, I derive 2-3 useful sources for high-confidenece positive hits:

- When filtering by pos|p-value < 0.01, there are 12 positive selected genes identified in 4 or more distinct data sets.
- When filtering by pos|p-value < 0.1, there are ~50 positive selected genes identified in 4 or more distinct data sets.
- When filtering by pos|p-value < 0.1, there are ~13 positive selected genes identified in 5 or more distinct data sets.
- Filtering for a specific number of top hits does not seem to be a valuable approach for the identification of consensus hits.

Here are the genes which have been identified in at least 4 data sets when using a p-value < 0.01

```
SCAP      5
CSNK2B    5
CTSL      4
HS2ST1    4
ACE2      4
GPAA1     4
GNG5      4
CCZ1B     4
EIF4E2    4
RAB10     4
GIGYF2    4
Name: #(datasets with p-value < 0.01), dtype: int64
```

And here are the genes which have been identified in at least 4 data sets when using a p-value < 0.1

```
CSNK2B           7
HS2ST1           6
SCAP             6
COG3             5
SUN2             5
CTSL             5
IL17RA           5
PIGS             5
ALG6             5
SLC35B2          5
WDR91            5
RAB10            5
VPS26A           5
GPAA1            5
PYROXD1          5
ACE2             5
GNG5             5
CCZ1B            4
ARMC10           4
RNF20            4
POP1             4
CD1D             4
```

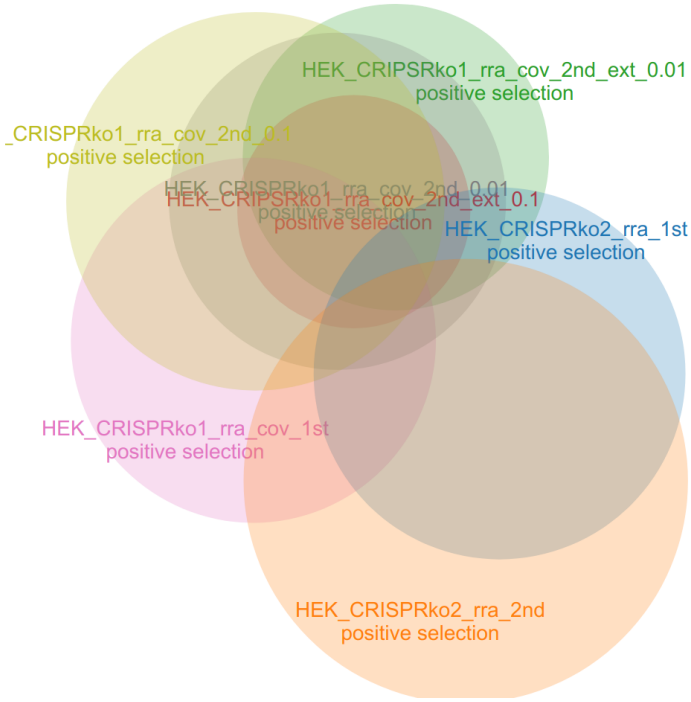
NCOA5	4
CDIPT	4
FAM175A	4
FER	4
NDC1	4
LSM6	4
GIGYF2	4
TMPRSS6	4
VPS11	4
DNMT1	4
CHCHD10	4
GNL3L	4
ZDHHC8	4
RPL17-C18orf32	4
ELOF1	4
AGRP	4
UHRF1	4
ACTR2	4
C3orf33	4
HNRNPC	4
TSTD2	4
WDR81	4
COG8	4
SLC39A9	4
ZNF671	4
CALR	4
RQCD1	4
ACSL3	4
DGCR6	4
POLR1D	4
ZNF70	4
NPC2	4
EXT2	4
NF2	4
RPL7	4
RAB14	4
VPS33A	4
FNDC3B	4
SLC30A1	4
ASCC3	4

```
CCZ1              4
GUK1              4
ALG5              4
ATP6V0D1          4
PIK3C3            4
REV1              4
ATE1              4
LDLR              4
MT2A              4
ACOT6             4
POLR3D            4
CNOT1             4
CUL2              4
EIF4E2            4
GDI2              4
CLEC16A           4
LARP7             4
INPP5J            4
VPS29             4
NPC1              4
EEF2              4
DNM2              4
Name: #(datasets with p-value < 0.1), dtype: int64
```

2.4.3 Integration of our HEK CRISPRko screen

Now that we have a good overview of published data sets, we want to put our data into this context.

1. Consensus of our data The whole HEK screening experiment has been replicated twice. Both times two infections were performed. First, I compare the different results in order to then integrate our data with published data.



A venn diagram from VISPR (FDR < 0.1) shows the following

- There is a high overlap across different conditions (e.g. 1st and 2nd infection)
- The first screen lead to very different hits compared to the second screen

Therefore I compare published hits with our hits in 3 ways:

- Union of hits from experiment 1
- Union of hits from experiment 2
- Union of hits from experiment 1 and 2 (i.e. everything shown in the Venn diagram above)

Now that we defined consensus sets for our screens and for published screens. We would like to know

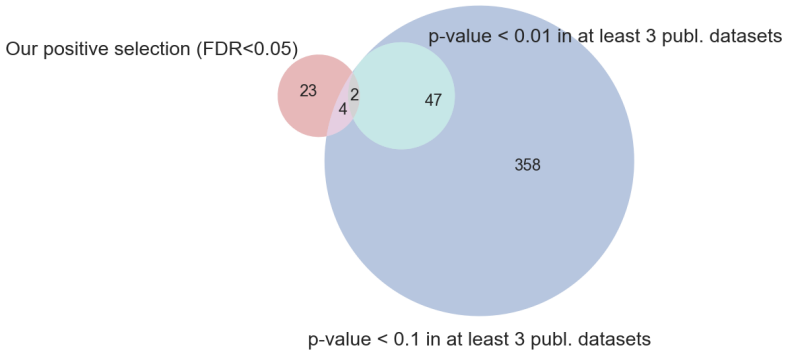
(a) What percentage of previously identified genes we are able reproduce (like this we can argue that our screening method is reliable)

(b) Which genes come up in our data but have not yet been published previously

2. Experiment 1

First, here is the overlap of our consensus and the consensus of published datasets.

```
<matplotlib_venn._common.VennDiagram at 0x7ff831db28b0>
```



There are 6 overlapping genes:

```
SLC30A1,CCZ1,DNM2,PIK3C3,SLC33A1,ALG5
```

From the 29 genes that only came up from our data, 20 have been identified in 1 or 2 published datasets.

In conclusion, we can confirm and consolidate 6 positive selection genes that have already been identified in at least 3 published datasets. We further confirm 14 genes that have been previously identified in 1 or 2 published datasets. In addition, we report **9 genes that have not yet been detected** by any other group or method:

```
ERICH6B,MESDC1,WASL,SMCHD1,DISP2,RFXAP,DAZ3,CDH2,XRN1
```

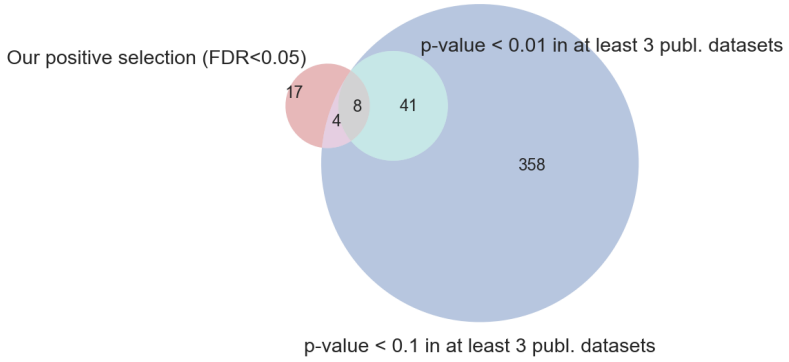
Ours 29 genes:

```
ERICH6B,SEC63,INTS6,BAX,DISP2,WDR7,CCZ1,DNM2,CDH2,ALG5,MESDC1,WASL,Non-Targeting,P
```

3. Experiment 2

   First, here is the overlap of our consensus and the consensus of published datasets

   ```
   <matplotlib_venn._common.VennDiagram at 0x7ff831c03580>
   ```



   The overlapping 12 genes are

   ```
   CCZ1B,LUC7L2,WDR91,CTSL,ACSL3,SNX27,CCZ1,WDR81,VPS35,VAC14,GDI2,ACE2
   ```

   From the 17 genes that only came up from our data, 7 have been identified in 1 or 2 published datasets.

   In conclusion, we can confirm and consolidate 12 positive selection genes that have already been identified in at least 3 published datasets. We further confirm 10 genes that have been previously identified in 1 or 2 published datasets. In addition, we report **7 genes that have not yet been detected** by any other group or method:

   ```
   SLC39A1,GNPTG,PEX10,CTBP2,SEC23IP,SLC12A9,CDH2
   ```

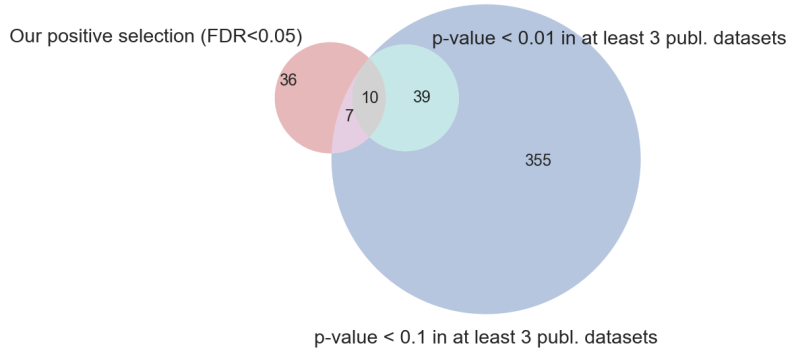
   Our 29 genes:

   ```
   SNX27,VPS35,VAC14,SLC12A9,CDH2,GDI2,CTSL,SLC39A1,GNPTAB,ARPC5L,MTMR9,UVRAG,CCZ1B,I
   ```

4. Experiment 1 and 2 combined

   First, here is the overlap of our consensus and the consensus of published datasets

```
<matplotlib_venn._common.VennDiagram at 0x7ff831bfbcd0>
```



The overlapping 17 genes are

```
CCZ1B,LUC7L2,SLC30A1,CTSL,WDR91,SNX27,CCZ1,ACSL3,DNM2,VPS35,VAC14,SLC33A1,WDR81,PI
```

From the 36 genes that only came up from our data, 21 have been identified in 1 or 2 published datasets.

In conclusion, we can confirm and consolidate 17 positive selection genes that have already been identified in at least 3 published datasets. We further confirm 11 genes that have been previously identified in 1 or 2 published datasets. In addition, we report **15 genes that have not yet been detected** by any other group or method:

```
SLC39A1,ERICH6B,MESDC1,GNPTG,WASL,PEX10,SMCHD1,DISP2,CTBP2,RFXAP,SEC23IP,DAZ3,SLC1
```

Ours 53 genes:

```
SEC63,BAX,DISP2,SNX27,DNM2,VPS35,VAC14,SLC12A9,CDH2,GDI2,CTSL,ALG5,SLC39A1,MESDC1,
```

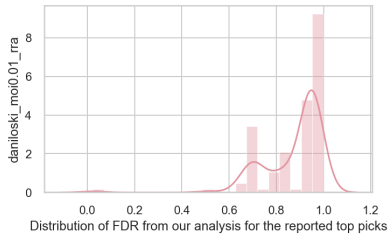
2.5 Negative selection

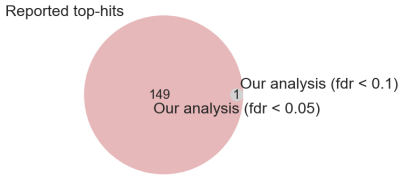
Screen data also suits for negative selection. Here, we can deduce that a gene is required/beneficial for the defense against a virus.

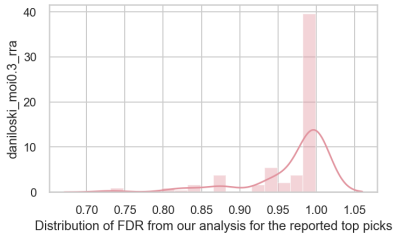
2.5.1 Verify reported top picks

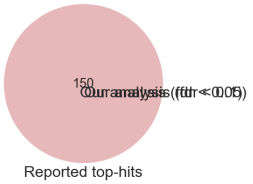
First, I wanted to control, whether the top picks reported on the maayan-lab "Covid19 Drug and Gene Set library" could be reproduced by my analysis of their raw data. I therefore plotted the FDR values from my analysis (of their raw data) for the reported genes. I expected most FDR values to be in the 0-0.2 range, which was the case for Schneider and Hoffmann, however not for Daniloski and Wang.

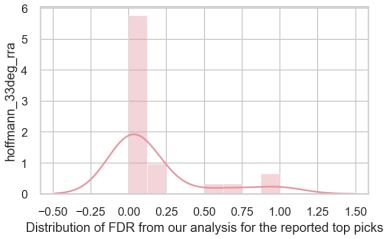
```
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Circle B has zero area")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Circle C has zero area")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
```

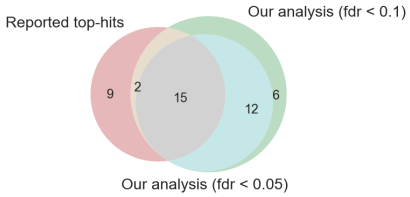

daniloski_moi0.01_rra
Distribution of FDR from our analysis for the reported top picks

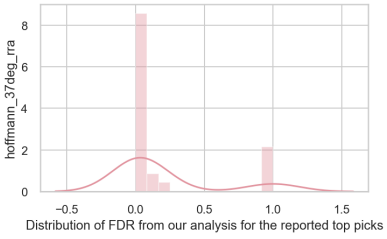

Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
149
1

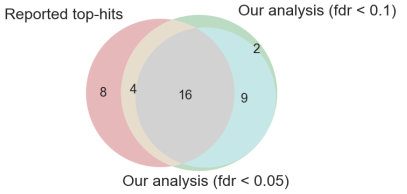

daniloski_moi0.3_rra
Distribution of FDR from our analysis for the reported top picks

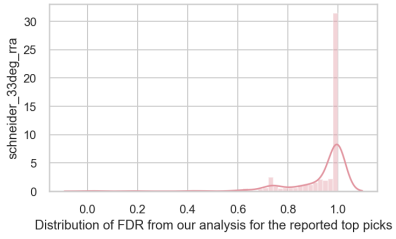

150
Reported top-hits

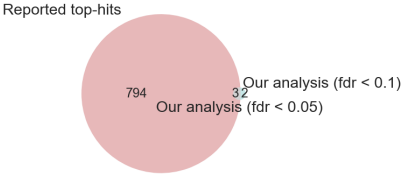

hoffmann_33deg_rra
Distribution of FDR from our analysis for the reported top picks

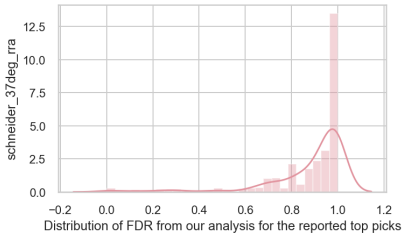

Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
9
2
15
12
6

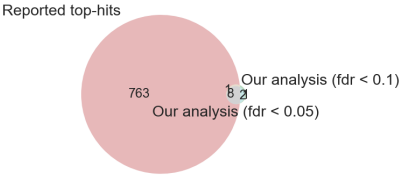

hoffmann_37deg_rra
Distribution of FDR from our analysis for the reported top picks

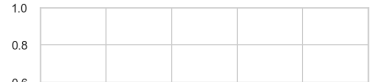

Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
8
4
16
2
9

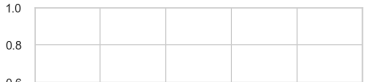

schneider_33deg_rra
Distribution of FDR from our analysis for the reported top picks


Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
794
3
2


schneider_37deg_rra
Distribution of FDR from our analysis for the reported top picks


Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
763
1
8
2

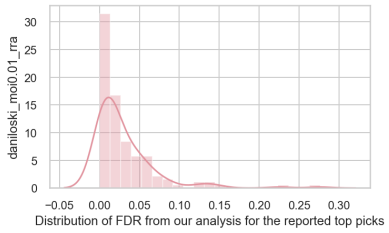


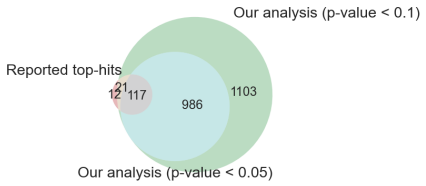


The reason, we don't see sharp peaks at ~0.0 (especially for Wang and daniloski, is that all reported publications seem to use the *p-value* instead of the *FDR* to report their picks. This means that all publications don't apply multiple-testing correction on the *p-value*, which is a common and necessary step to ensure the correct reporting of statistical significance! I'm really surprised that this kind of p-value-hacking is so common in the field. VISPR for example also sorts by the FDR.

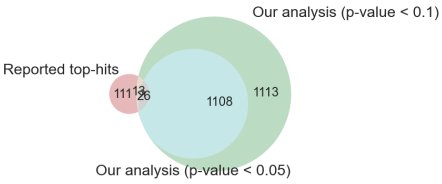
Below, I plot the same analysis, using *p-values* for selection rather than the multiple-testing-corrected *FDR*. Here the reported top picks all have expected values close to 0.0.

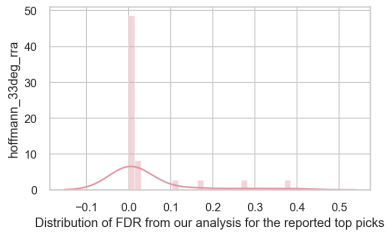
```
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Bad circle positioning")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Bad circle positioning")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
```

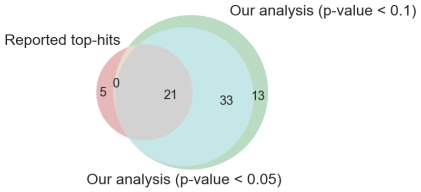

daniloski_moi0.01_rra
Distribution of FDR from our analysis for the reported top picks

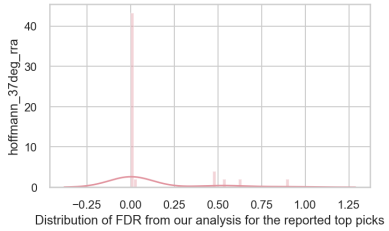

Our analysis (p-value < 0.1)
Reported top-hits
12
21
117
986
1103
Our analysis (p-value < 0.05)

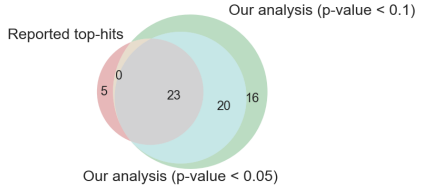

daniloski_moi0.3_rra
Distribution of FDR from our analysis for the reported top picks

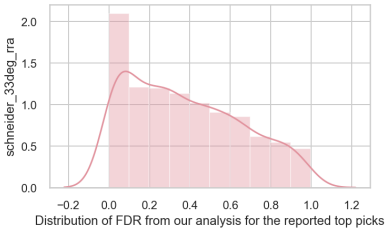

Our analysis (p-value < 0.1)
Reported top-hits
1
1
1
26
1108
1113
Our analysis (p-value < 0.05)

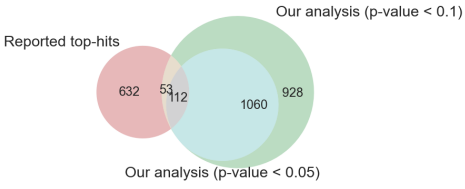

hoffmann_33deg_rra
Distribution of FDR from our analysis for the reported top picks

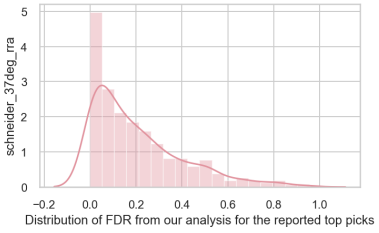

Our analysis (p-value < 0.1)
Reported top-hits
5
0
21
33
13
Our analysis (p-value < 0.05)

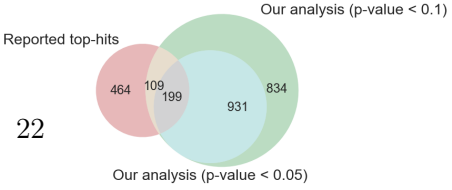

hoffmann_37deg_rra
Distribution of FDR from our analysis for the reported top picks

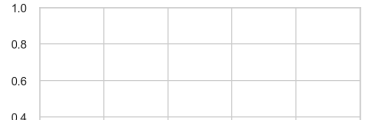

Our analysis (p-value < 0.1)
Reported top-hits
5
0
23
20
16
Our analysis (p-value < 0.05)

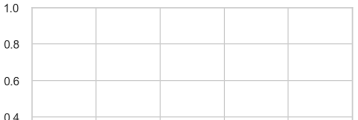

schneider_33deg_rra
Distribution of FDR from our analysis for the reported top picks


Our analysis (p-value < 0.1)
Reported top-hits
632
53
112
1060
928
Our analysis (p-value < 0.05)


schneider_37deg_rra
Distribution of FDR from our analysis for the reported top picks


Our analysis (p-value < 0.1)
Reported top-hits
464
109
199
931
834
Our analysis (p-value < 0.05)





2.5.2 Overlap of different studies

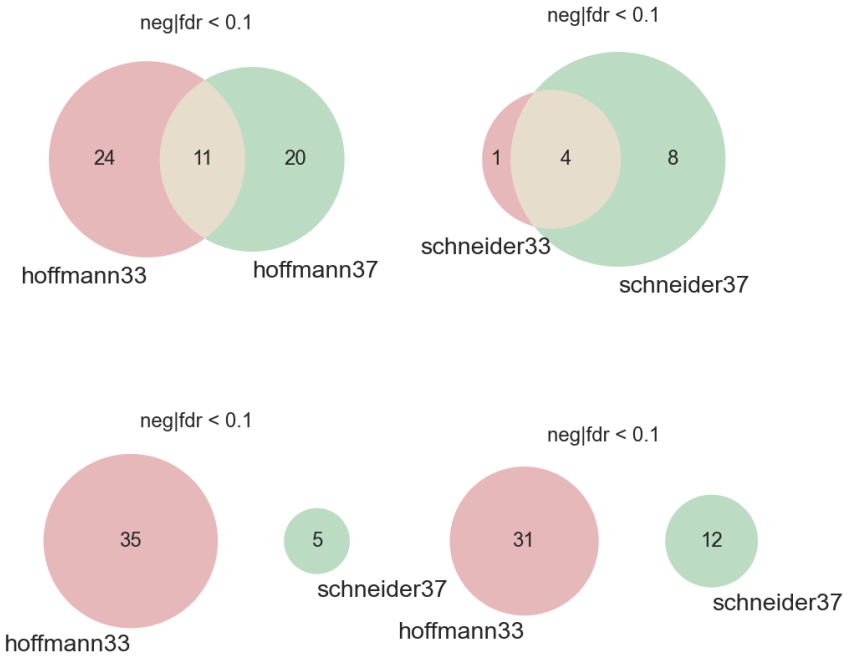
Since different groups reported their top picks with different thresholds, I use my own analysis of their data to pick selection genes and to compare them. I use different FDR and p-value thresholds for comparison. Further, since the differing experimental procedures like sequencing depth etc. can lead to a largely different variation across data sets, FDR and p-values can potentially not be compared across datasets. To overcome this limitation, I further select the top 20, 50, 100, 200 genes from each data set and compare them.

For each comparison, we plot how many genes were identified in how many data sets (multiple barplots).

The Schneider and Hoffmann data sets have the potential to skew the data:

- Hoffmann only considered a subset of all genes
- Both Schneider and Hoffmann do their work in two temperatures (33, 37 degrees), which might be quite similar.

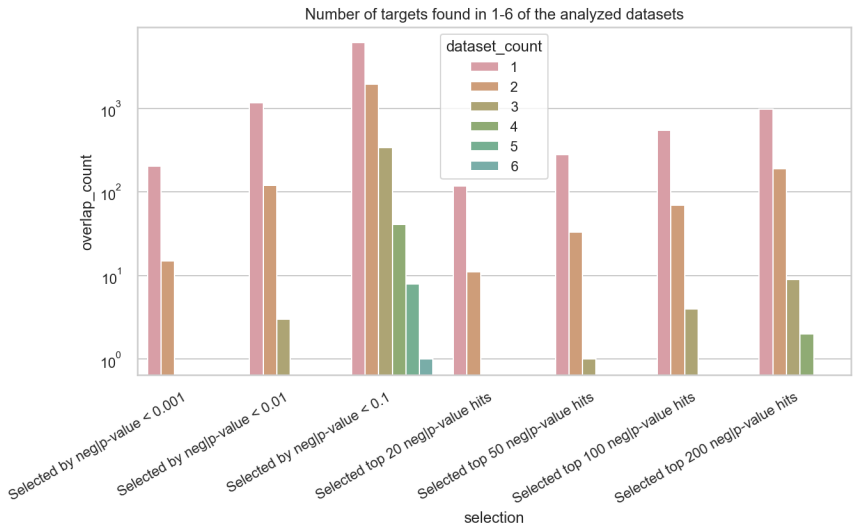
I compare the 33 and 37 degC from Schneider and Hoffmann.



Although the data sets show similarities, there are also variations and I think it's reasonable to include all datasets into the comparison. Here, we use the p-value since we gain additional statistical signifance for targets as they appear in multiple data sets.

- p-value: 0.001, 0.01, 0.1.
- Top n: 20, 50, 100, 200
- negative

```
Text(0.5, 1.0, 'Number of targets found in 1-6 of the analyzed datasets')
```



Here are the genes which have been identified in at least 3 data sets when using a p-value < 0.01

```
NCAPG     3
PSIP1     3
NDUFB9    3
Name: #(datasets with p-value < 0.01), dtype: int64
```

And here are the genes which have been identified in at least 4 data sets when using a p-value < 0.1

```
EXOSC8      6
VPS28       5
```

KLHL21	5
PRIM2	5
SBNO1	5
OIP5	5
STARD3	5
RNF41	5
BRI3	5
ORM1	4
VPS37B	4
LMBR1L	4
NR1H2	4
RHBDD2	4
TUBG1	4
COX8A	4
TMEM182	4
ZNF772	4
MEX3D	4
SP7	4
ZNF417	4
BCL2L1	4
ERLEC1	4
PTPN18	4
OR4X2	4
WDR45B	4
CYB561D2	4
ELF1	4
NRD1	4
RBFOX3	4
RBBP5	4
CXorf56	4
ATF7IP2	4
PRMT6	4
ARHGEF15	4
WFDC11	4
H2AFZ	4
PIK3CB	4
MRPS2	4
NCAPG	4
STAT6	4
SPINK13	4

```
KCNG4       4
HPX         4
FAM57A      4
MSRA        4
ZNF28       4
RPSA        4
ATXN7       4
MRPS25      4
Name: #(datasets with p-value < 0.1), dtype: int64
```

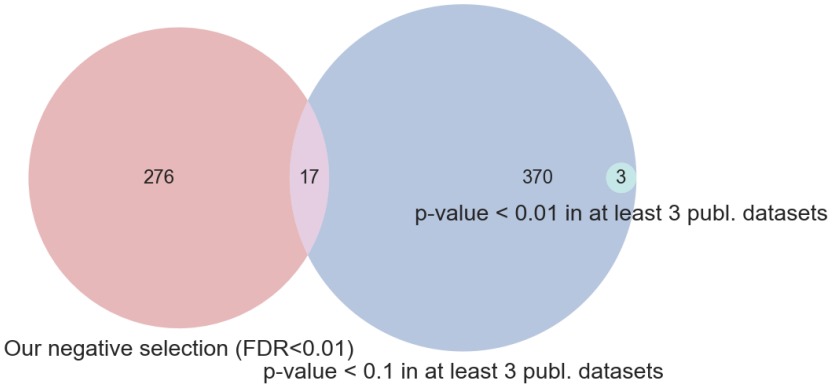
2.5.3 Integration of our HEK CRISPRko screen

Now that we have a good overview of published data sets, we want to put our data into this context.

I take the union of all reported negative selection genes from all our screens using an FDR of 0.01. Here, I don't look at experiment 1 and 2 separately.

First, here is the overlap of our consensus and the consensus of published datasets

```
<matplotlib_venn._common.VennDiagram at 0x7ff83280feb0>
```



The overlapping 17 genes are

```
NDUFB10,MRPL41,WDR33,POLR2I,MINOS1,TRAPPC11,CDCA8,ZNHIT2,VPS28,DUT,STARD3,NELFCD,HSPE1
```

From the 276 genes that only came up from our data, 160 have been identified in 1 or 2 published datasets.

In conclusion, we can confirm and consolidate 17 negative selection genes that have already been identified in at least 3 published datasets. We further confirm 160 genes that have been previously identified in 1 or 2 published datasets. In addition, we report **116 genes that have not yet been detected** by any other group or method:

```
GOSR2,CCNH,UQCR10,FAM25G,U2AF1,NMD3,TAF1C,NOP10,UQCRC1,REXO2,MRPS18A,RNH1,FAM231A,URI1
```

Our 293 negative-selection genes:

```
COX6B1,UQCR10,IARS2,FAM25G,ATP6V1B2,GNB2L1,NMD3,TAF1C,NOP10,UQCRC1,REXO2,RNH1,GADD45GI
```

3 HEK293/OC43 analysis

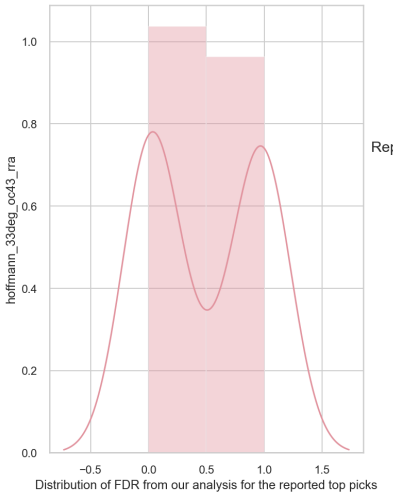
As with the HEK analysis, we can perform an OC43 analysis both for positive- and negative-selection genes.

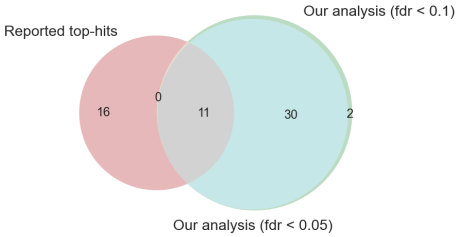
3.1 Positive selection

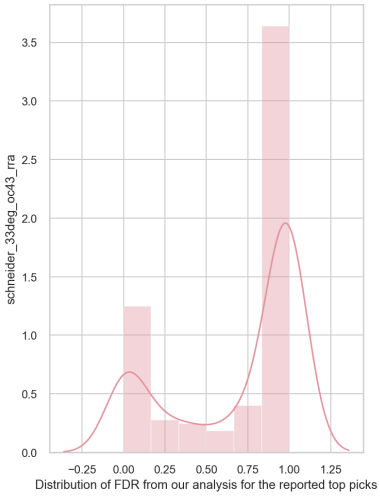
3.1.1 Verify reported top picks

First, I wanted to control, whether the top picks reported on the maayan-lab "Covid19 Drug and Gene Set library" could be reproduced by my analysis of their raw data. I therefore plotted the FDR values from my analysis (of their raw data) for the reported genes. Hoffmann and Schneider each report a data set in 33°C condition. Wang et al. did an OC43 screen in 37 °C.

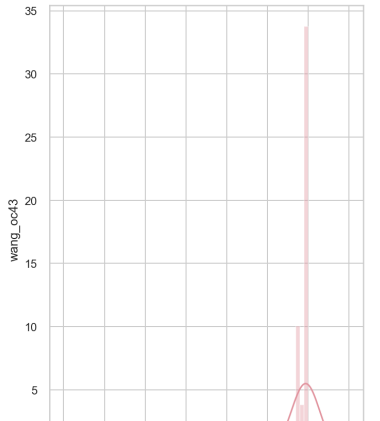
```
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Bad circle positioning")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Bad circle positioning")
```

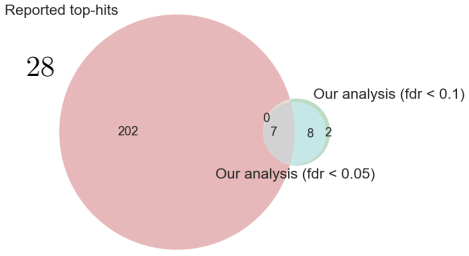

hoffmann_33deg_oc43_rra
Distribution of FDR from our analysis for the reported top picks


Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
16
0
11
30
2


schneider_33deg_oc43_rra
Distribution of FDR from our analysis for the reported top picks

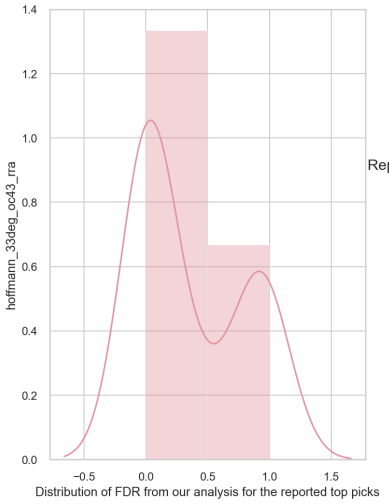

Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
316
10
63
46
12

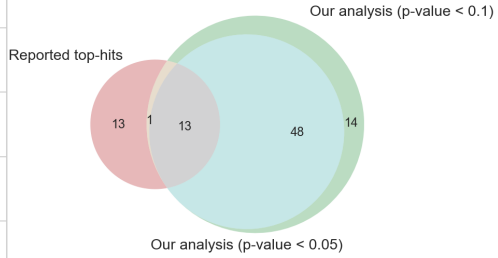

wang_oc43

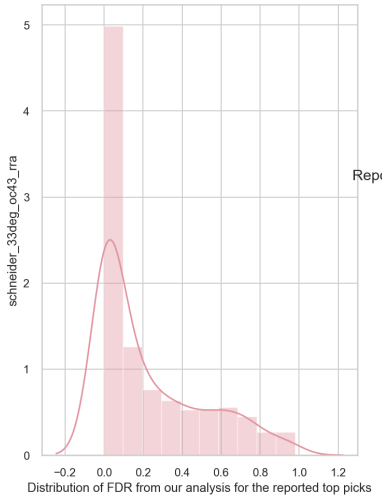

Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
202
0
7
8
2

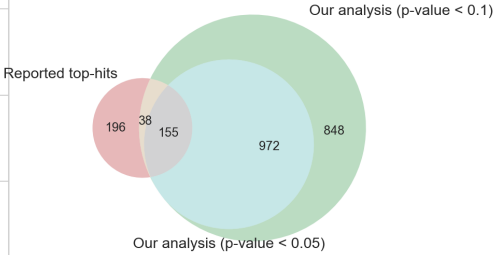
Using the p-value improves the situation only slightly

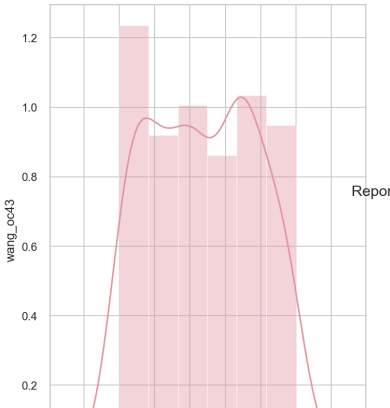
```
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
```

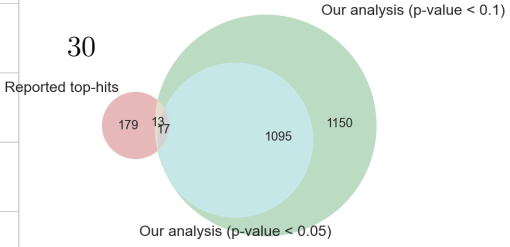

hoffmann_33deg_oc43_rra
Distribution of FDR from our analysis for the reported top picks


Our analysis (p-value < 0.1)
Reported top-hits
13
1
13
48
14
Our analysis (p-value < 0.05)


schneider_33deg_oc43_rra
Distribution of FDR from our analysis for the reported top picks


Our analysis (p-value < 0.1)
Reported top-hits
196
38
155
972
848
Our analysis (p-value < 0.05)

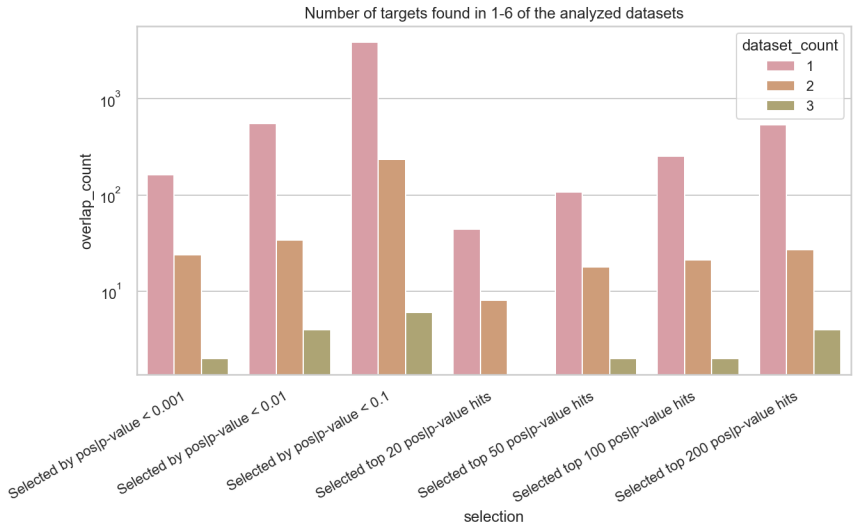

wang_oc43


Our analysis (p-value < 0.1)
Reported top-hits
179
13
17
1095
1150
Our analysis (p-value < 0.05)

3.1.2 Overlap of different studies

- p-value: 0.001, 0.01, 0.1.
- Top n: 20, 50, 100, 200
- positive

```
Text(0.5, 1.0, 'Number of targets found in 1-6 of the analyzed datasets')
```



Here are the genes which have been identified in at least 2 data sets when using a p-value < 0.01

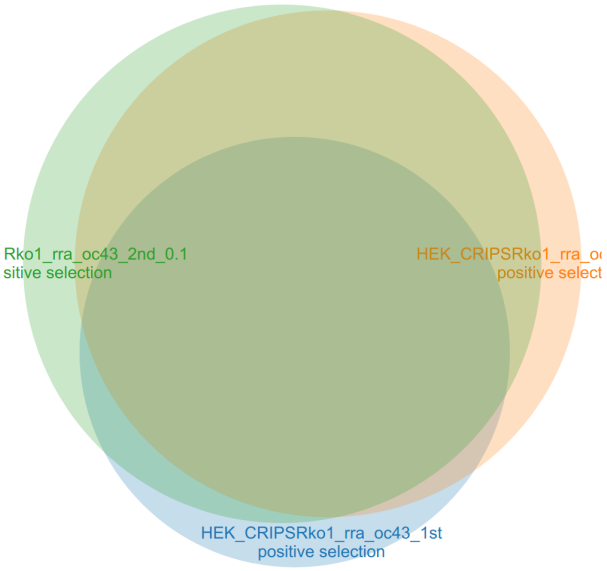
```
RAB7A      3
FAM98A     3
VPS11      3
RAB2A      3
GPAA1      2
XYLT2      2
GNPTAB     2
SPNS1      2
AHCYL1     2
MRPS25     2
CDX2       2
TSEN54     2
ZFP36L2    2
```

```
UGDH       2
WDR91      2
VPS39      2
FAM20B     2
VPS33A     2
VPS16      2
EXTL3      2
EMC1       2
WDR18      2
EXT1       2
CCZ1B      2
EXT2       2
B4GALT7    2
WDR81      2
ARL8B      2
MFSD12     2
B3GAT3     2
B3GALT6    2
ECSIT      2
NDST1      2
SCAP       2
DPH5       2
ATP6V1A    2
SLC35B2    2
PIGO       2
Name: #(datasets with p-value < 0.01), dtype: int64
```

And here are the genes which have been identified in at least 3 data sets when using a p-value < 0.1

```
GPAA1     3
FAM98A    3
RAB2A     3
VPS39     3
RAB7A     3
VPS11     3
Name: #(datasets with p-value < 0.1), dtype: int64
```

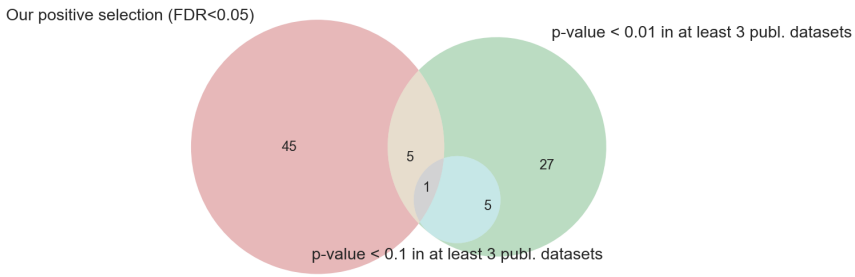
3.1.3 Integration of our HEK CRISPRko screen



The first and second infection with 0.1 and 0.01 MOI all lead to very similar results. To not miss any hits, I take the union of the three sets with FDR 0.05.

First, here is the overlap of our consensus and the consensus of published datasets.

```
<matplotlib_venn._common.VennDiagram at 0x7ff832a00bb0>
```



There are 6 overlapping genes:

```
CCZ1B,EXT2,WDR91,SLC35B2,WDR81,RAB7A
```

From the 45 genes that only came up from our data, 18 have been identified in 1 or 2 published datasets.

In conclusion, we can confirm and consolidate 6 positive selection genes that have already been identified in at least 3 published datasets. We further confirm 18 genes that have been previously identified in 1 or 2 published datasets. In addition, we report **27 genes that have not yet been detected** by any other group or method:

```
SLC35A1,DOCK5,ARID1A,C1GALT1,NXPE3,INPP5K,KRTAP5-7,SLC35D1,WRB,SEC61B,GABPB1,MDH1,CASD1
```

Ours 45 genes:

```
ADNP,NXPE3,IFITM2,GET4,INPP5K,KRTAP5-7,SLC35D1,SIGIRR,TMED2,SLC35B2,DOK2,CCZ1,ANO9,SMAR
```

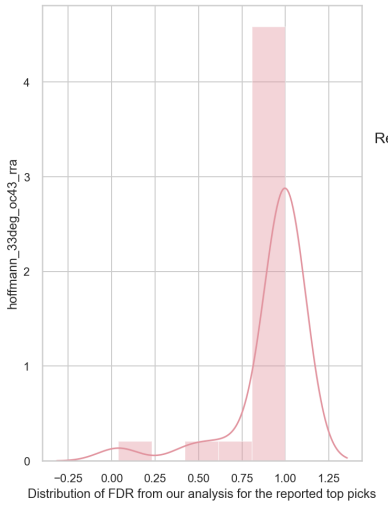
3.2 Negative selection

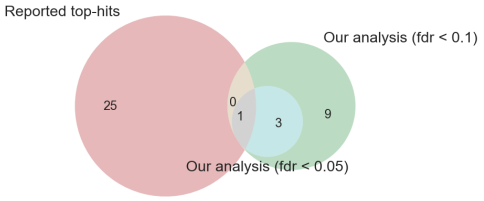
The negative selection surprisingly leads to a very low overlap with published data sets.

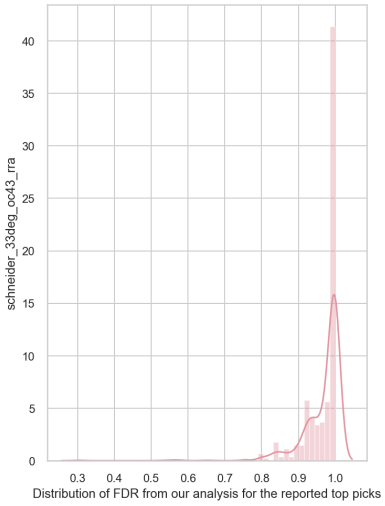
3.2.1 Verify reported top picks

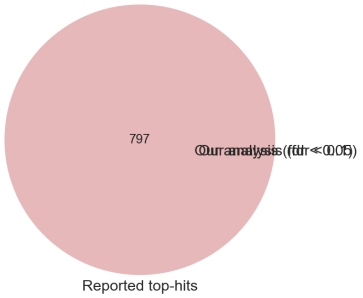
First, I wanted to control, whether the top picks reported on the maayan-lab "Covid19 Drug and Gene Set library" could be reproduced by my analysis of their raw data. I therefore plotted the FDR values from my analysis (of their raw data) for the reported genes. Hoffmann and Schneider each report a data set in 33°C condition. Wang et al. did an OC43 screen in 37 °C.

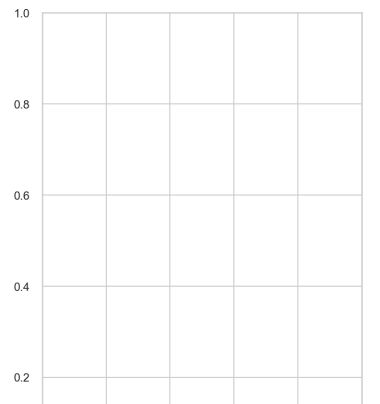
```
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Bad circle positioning")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Circle B has zero area")
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn("Circle C has zero area")
```

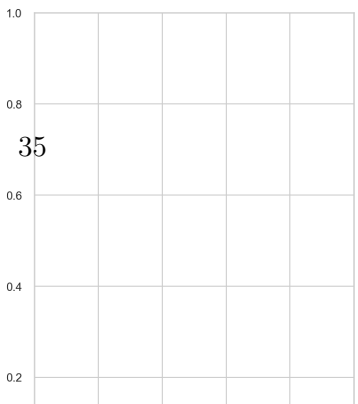

hoffmann_33deg_oc43_rra
Distribution of FDR from our analysis for the reported top picks


Reported top-hits
Our analysis (fdr < 0.1)
Our analysis (fdr < 0.05)
25
0
1
3
9


schneider_33deg_oc43_rra
Distribution of FDR from our analysis for the reported top picks

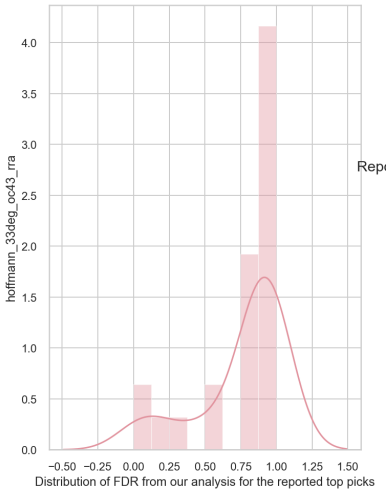

797
Reported top-hits

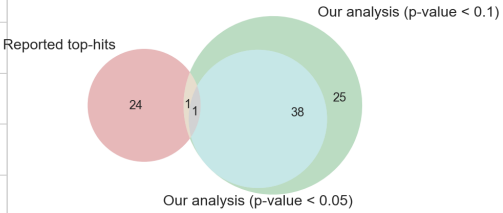


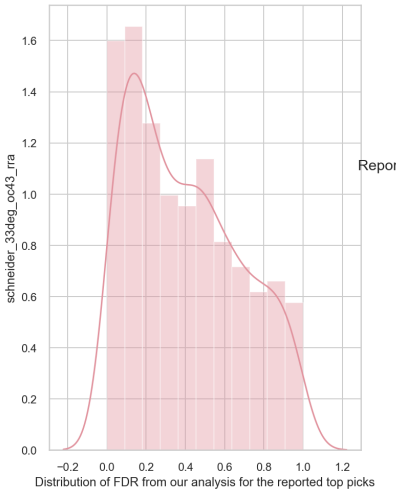


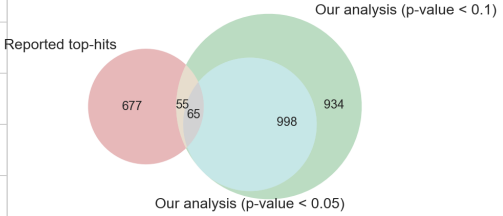
Using the p-value improves the situation only slightly

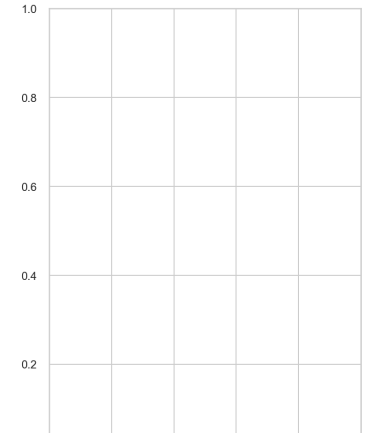
```
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
/nix/store/az3a2pmmdgy24sxkwx3lhs3zs1nsagfh-python3-3.8.5-env/lib/python3.8/site-packag
  warnings.warn(msg, FutureWarning)
```

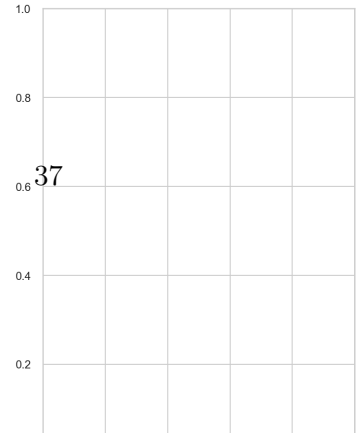

hoffmann_33deg_oc43_rra
Distribution of FDR from our analysis for the reported top picks


Reported top-hits
Our analysis (p-value < 0.1)
Our analysis (p-value < 0.05)
24
1
1
38
25


schneider_33deg_oc43_rra
Distribution of FDR from our analysis for the reported top picks


Reported top-hits
Our analysis (p-value < 0.1)
Our analysis (p-value < 0.05)
677
55
65
998
934

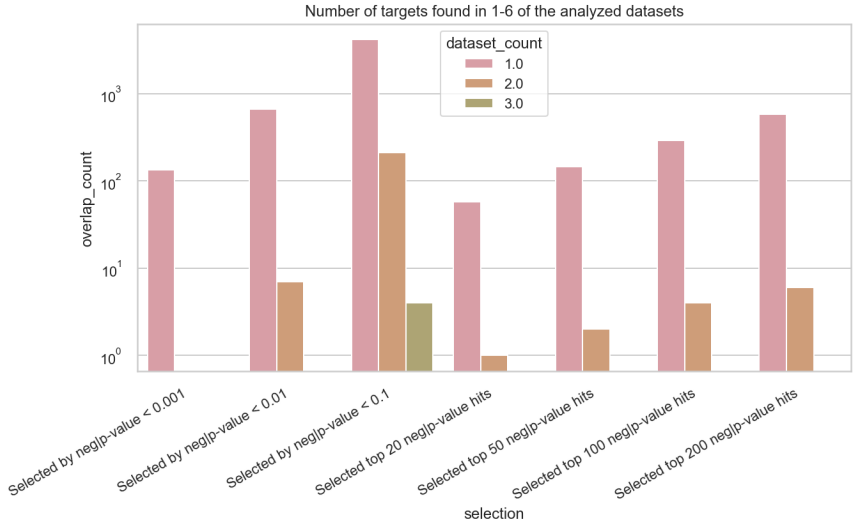




3.2.2 Overlap of different studies

- p-value: 0.001, 0.01, 0.1.
- Top n: 20, 50, 100, 200
- negative

```
Text(0.5, 1.0, 'Number of targets found in 1-6 of the analyzed datasets')
```



Here are the genes which have been identified in at least 2 data sets when using a p-value < 0.01

```
TUBA1B    2
GTF2F2    2
NUP54     2
SART3     2
TBK1      2
BCL2L1    2
GIGYF2    2
Name: #(datasets with p-value < 0.01), dtype: int64
```

And here are the genes which have been identified in at least 2 data sets when using a p-value < 0.1

```
INTS4     3
```

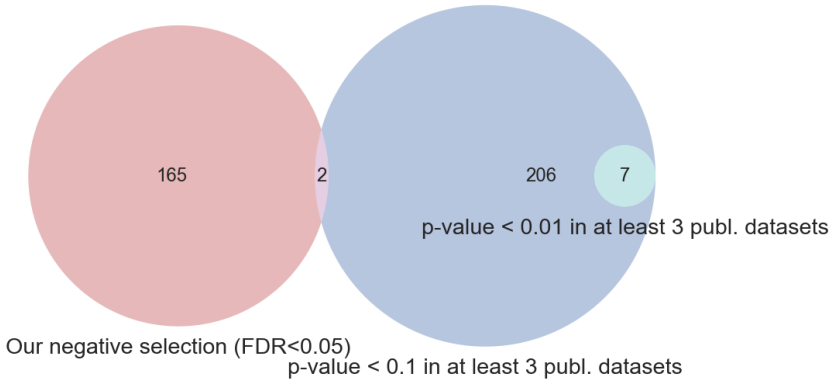
```
GTF2F2    3
PRIM2     3
YIF1A     3
RP2       2
          ..
OR4S2     2
NACA      2
SNX27     2
KIF11     2
STK11     2
Name: #(datasets with p-value < 0.1), Length: 215, dtype: int64
```

3.2.3 Integration of our HEK CRISPRko screen

The first and second infection with 0.1 and 0.01 MOI all lead to very similar results. To not miss any hits, I take the union of the three sets with FDR 0.05.

First, here is the overlap of our consensus and the consensus of published datasets.

```
<matplotlib_venn._common.VennDiagram at 0x7ff831950820>
```



There are only 2 overlapping genes:

```
HYPK,VPS28
```