Pneumococcal sgRNA library efficiency exploration

Contents

1 Introduction

Here, we assess the efficiency of the CRISPRi sgRNA library designed for *Streptococcus pneumoniae* D39V in several pneumococcal strains *in silico*. All scores were computed as described in the Methods section with a separate custom R script, to be found on https://github.com/veeninglab/CRISPRi-seq, along with the data sets analyzed here. For every potential sgRNA binding site (up to eight mismatches), a genetic element was considered a hit only if its non-template strand was targeted (partially) within the gene body. Computed variables that are evaluated in this document:
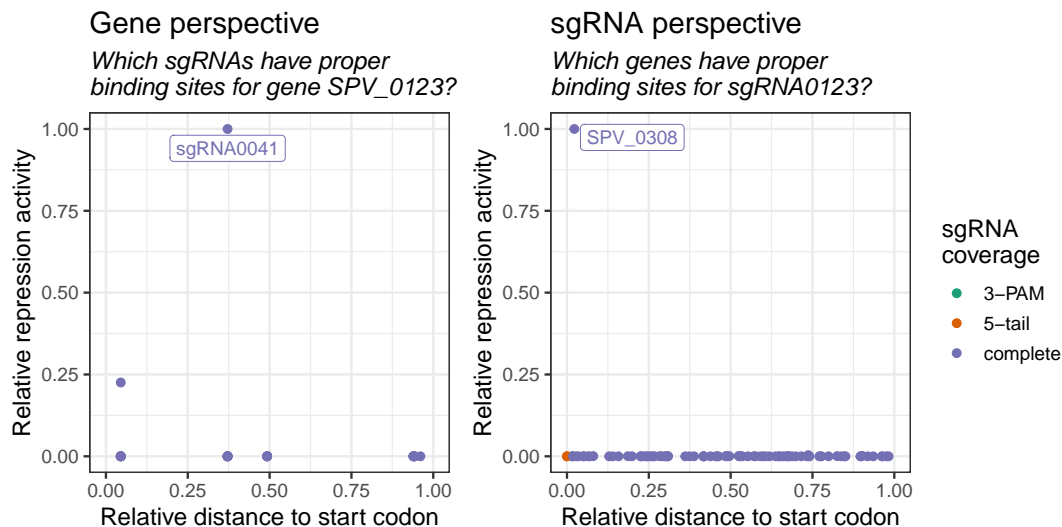
- `reprAct`: Estimated relative retained repression activity, compared to an hypothetical zero-mismatch sgRNA binding site at the same distance from the start codon, on the [0,1] interval (0 - 100% retained activity).
- `dist2SC`: Relative distance of the sgRNA binding site to the start codon of the gene. Set to the [0,1] interval with feature scaling, where 0 means binding site on the start codon or partial overlap of the 5-prime end of the gene and 1 means binding site on the far end of or partial overlap with the 3-prime end of the gene.

For all genomes, the data were retrieved from the Genbank database.

Pneumococcal strain	NCBI assembly accession
D39V	GCA_003003495.1
TIGR4	GCA_000006885.1
R6	GCA_000007045.1
Hungary19A-6 (H19A)	GCA_000019265.1
Taiwan19F-14 (T19F)	GCA_000019025.1
11A	GCA_002813955.1
G54	GCA_000019825.1

1.1 The sgRNA and the gene perspective

We assessed sgRNA library efficiency from two perspectives: per gene and per sgNRA. The plots below show the different questions they answer, with one gene and one sgRNA as example. The following sections in this document first evaluate the sgRNA perspective and then the gene perspective, per strain in the table above.

Gene perspective
Which sgRNAs have proper binding sites for gene SPV_0123?

sgRNA perspective
Which genes have proper binding sites for sgRNA0123?

2 sgRNA perspective

Question answered: How many of the sgRNAs target at least one gene, i.e., are functional?

2.1 Considering only number of mismatches

Per strain, we extracted for each sgRNA the binding site within a gene body on the non-template strand that had the minimal number of mismatches. The resulting counts are listed in the table below. Strikingly, the pipeline did not detect a zero-mismatch gene target for all 1499 sgRNAs in D39V, although the library was designed to do so for this strain specifically.

Minimal number of mismatches	D39V	TIGR4	R6	H19A	T19F	11A	G54
0	1486	1143	1340	1101	1100	1060	1105
1	0	121	6	131	145	157	132
2	0	14	3	25	22	27	24
3	0	5	4	12	15	11	10
4	0	19	15	23	18	30	26
5	4	75	57	89	80	90	81
6	7	103	65	108	106	105	104
7	2	19	9	10	13	19	17
8	0	0	0	0	0	0	0
Total number of sgRNAs	*1499*	*1499*	*1499*	*1499*	*1499*	*1499*	*1499*

The reasons that 13 sgRNAs were reported to have no exact gene target in this analysis are shown in the table below. Briefly, the gene hits these sgRNAs were designed for were in all but one case not found for either of two reasons: (1) the sgRNA was designed to target the gene promoter rather than the gene body due to a lack of proper sgRNA binding sites; (2) the sgRNA targets a genetic element not annotated with a *locus_tag* key in the NCBI database and thus the target was not found (but present and as such annotated in our manually curated in-house database).

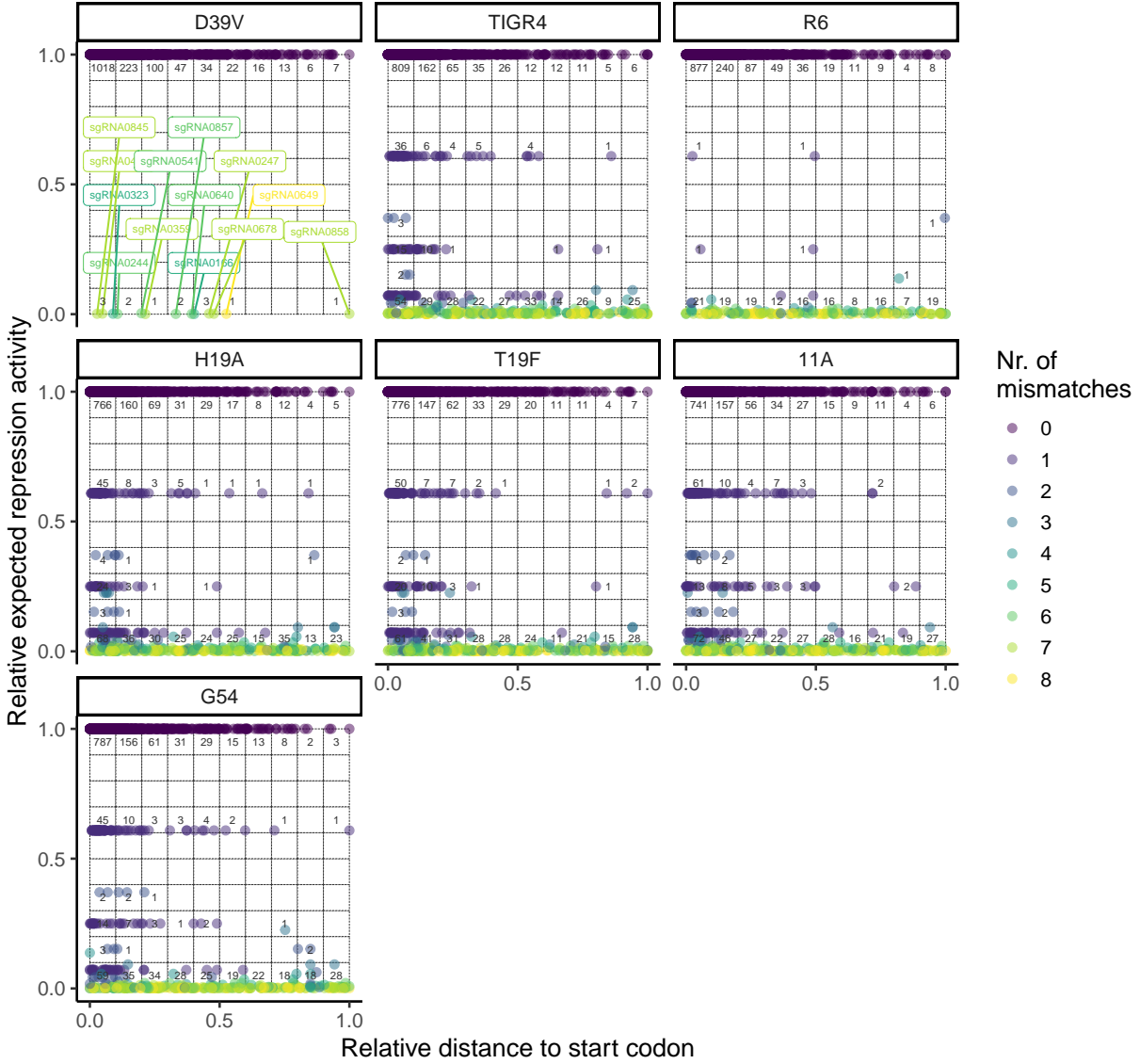| sgRNA	Targets	Reason no hit found
sgRNA0166	SPV_2167	Not annotated with locus_tag on NCBI: ncRNA srf-09 (RNAswitch-25)
sgRNA0244	SPV_0659, SPV_0660, SPV_2217	Not annotated on NCBI (binding site in SPV_2217, 676352..677447)
sgRNA0247	SPV_0664	Binds promoter region
sgRNA0323	SPV_2258	Not annotated with locus_tag on NCBI: ncRNA srf-14 (RNAswitch-26)
sgRNA0359	SPV_2270	Not annotated with locus_tag on NCBI: ncRNA srf-14 (RNAswitch-27)
sgRNA0450	SPV_2330	Not annotated with locus_tag on NCBI: ncRNA srf-14 (RNAswitch-28)
sgRNA0541	SPV_1429	Binds promoter region (partially on template strand SPV_1430)
sgRNA0640	SPV_1737	Binds promoter region
sgRNA0649	SPV_2421	Not annotated with locus_tag on NCBI: ncRNA srf-23 (RNAswitch-29)
sgRNA0678	SPV_1827, SPV_2426	Binds promoter region
sgRNA0845	SPV_0621	Design error: TCG instead of TGG
sgRNA0857	SPV_2432	Binds promoter region
sgRNA0858	SPV_2326	Binds promoter region; also exact match in promoter SPV_2325

Therefore, we considered only sgRNA0845 to not have an exact target in D39V, due to a design error. The other 12 sgRNAs in the table above were added to the zero-mismatch sgRNAs for D39V. This yielded the following strain-wise CRISPRi library functionality summary:

	D39V	TIGR4	R6	H19A	T19F	11A	G54
Total number of sgRNAs	1499	1499	1499	1499	1499	1499	1499
sgRNAs with exact target	1498	1143	1340	1101	1100	1060	1105
Percentage of functional sgRNAs	**99.9%**	**76.3%**	**89.4%**	**73.4%**	**73.4%**	**70.7%**	**73.7%**

2.2 Considering both number and position of mismatches

Naturally, sgRNAs can still be functional with mismatches between its spacer and a binding site, albeit with lower repression activity. To refine the functionality summary of the section above, we therefore extracted for each sgRNA its "optimal" binding site within a gene body on the non-template strand. "Optimal" was here defined as the site with the highest mismatch-based expected repression activity (`reprAct`), thus taking into account both the number and position of mismatches within the sgRNA spacer. An estimated repression activity of 1 corresponds to 0 mismatches. When there were multiple sites with this maximum, the one with the smallest distance to the start codon (`dist2SC`) was chosen. These data are depicted below, split out per strain. The only sgRNAs with an expected repression of <1 in D39V were the ones for which the reasons were already covered in the section above.

Optimal* within-gene binding site per sgRNA

In dark grey: number of sgRNAs in each two-dimensional bin. *Per sgRNA, the one binding site with the highest expected repression activity (based on mismatch number and position) was selected.

Indeed, the number of sgRNAs deemed functional increases with a lower expected repression threshold. However, we hereafter conservatively consider an sgRNA to be functional only when it has at least one within-gene zero-mismatch binding site. This corresponds to the sums of the dark grey numbers in the highest row of each respective plot above and to the summary table at the end of the previous section.

3 Gene perspective

Question answered: To what extent is each genome covered by the library?

4

3.1 Considering only number of mismatches

For each genetic element that was targeted by at least one sgRNA with at most eight mismatches, we extracted the binding site with the minimal number of mismatches. These gene-wise minimal mismatch binding sites are tabulated per strain below. Naturally, a large number of D39V genes has a zero-mismatch sgRNA, as the library was designed for this strain.

Minimal number of mismatches	D39V	TIGR4	R6	H19A	T19F	11A	G54
0	1554	1219	1406	1177	1162	1139	1161
1	5	128	7	133	147	176	134
2	3	17	5	37	23	37	20
3	2	7	3	13	17	26	7
4	39	54	44	62	51	52	57
5	190	247	219	308	253	274	259
6	261	409	321	470	389	397	390
7	66	170	97	167	134	184	131
8	8	23	11	18	16	20	16
Total number of genetic elements covered	*2128*	*2274*	*2113*	*2385*	*2192*	*2305*	*2175*

We calculated the percentage of the genetic elements in each genome that was covered by the library by direct targeting without mismatches. However, the true number of repressable genes will in all likelihood be much higher than that, because of polar effects: all genes in an operon are repressed if one of them is targeted.

We know for D39V that only 35 genes were not targeted (e.g. due to a lacking PAM site) and thus the effective targetable genome consisted of $2146 - 35 = 2111$ genetic elements for this sgRNA library. Assuming conserved operon structures across these strains, we used this information to extrapolate the targetable genome and covered genetic elements in all strains.

The targetable genome is simply estimated as $\frac{2111}{2146} \cdot annotated\,genes$. The estimated number of targeted genetic elements is subsequently calculated to be this number times the fraction of functional sgRNAs as computed above.

	D39V	TIGR4	R6	H19A	T19F	11A	G54
Directly targeted genetic elements	1554	1219	1406	1177	1162	1139	1161
Total annotated genetic elements	2146	2292	2116	2402	2205	2317	2186
Percentage of genetic elements directly covered	*72.4%*	*53.2%*	*66.4%*	*49%*	*52.7%*	*49.2%*	*53.1%*
Theoretically targetable genetic elements	2111	2255	2081	2363	2169	2279	2150
Percentage of functional sgRNAs	99.9%	76.3%	89.4%	73.4%	73.4%	70.7%	73.7%
Extrapolated targeted genetic elements	2110	1719	1861	1735	1592	1612	1585
Extrapolated percentage genetic elements covered	*98.3%*	*75%*	*87.9%*	*72.3%*	*72.2%*	*69.6%*	*72.5%*

Indeed, we know that the one sgRNA that is non-functional in D39V targeted one genetic element, bringing the total number of genetic elements covered by the library to $2146 - 35 - 1 = 2110$, as estimated in the table above. Although we can only compute the total number of annotated genes directly targeted by at least one sgRNA without mismatch for the other genomes, the true number of targeted genetic elements is more likely to be around the estimations in the table above, due to operon structures an polar repression effects.

3.2 Considering both number and position of mismatches

As for the sgRNA perspective analysis, genetic elements containing only an sgRNA binding site with >0 mismatches may still be repressed by the corresponding sgRNA(s), albeit less efficiently so. Again, to refine the more strict zero-mismatch results of above, we defined the "optimal" sgRNA binding site per genetic element as the one with the highest mismatch-based expected repression activity. In case of multiple binding sites with an equal relative maximum, the one with the smallest distance between PAM and start codon was opted for.

Optimal* sgRNA binding site per gene
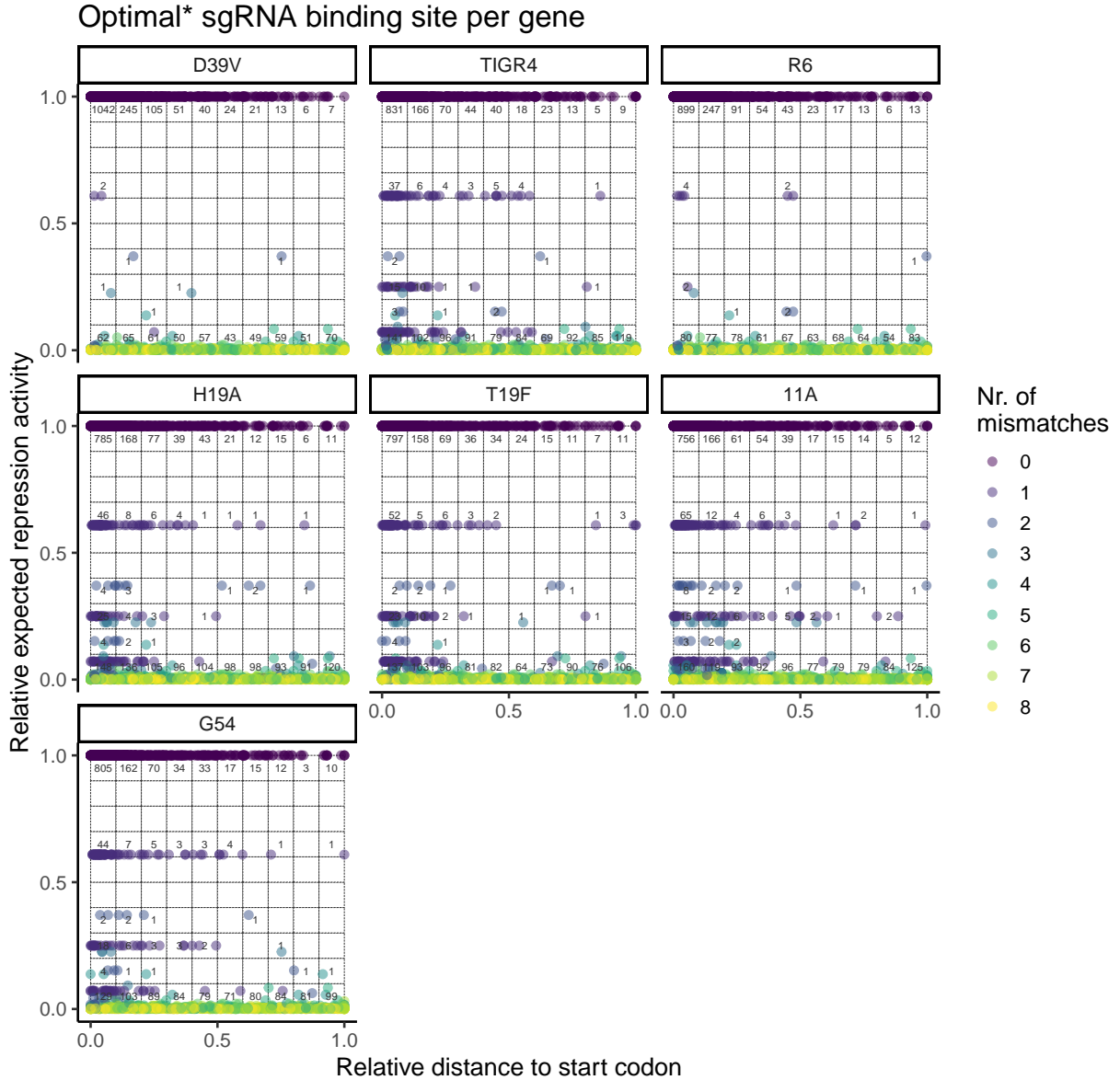
*In dark grey: number of genes in each two-dimensional bin. *Per gene, the one sgRNA binding site with the highest expected repression activity (based on mismatch number and position) was selected.*

The fact that the library was designed for D39V is reflected by the higher number of genetic elements with an optimal sgRNA binding site with both maximum mismatch-based expected repression activity and small distance to the start codon (grey numbers in the plot above). In addition, few D39V genetic elements have a moderate expected repression, which is to be expected because they were either targeted or not by design.

The relatively large number of genetic elements with good sgRNA targeting (high repression estimates) in other strains indicates conservation of genes between these strains. It should be noted that repression activity is expected to decrease drastically with increasing distance from the start codon, rendering these sgRNAs less efficient.

3.2.1 Required library size for zero-mismatch coverage

Since the CRISPRi library consists of 1499 unique sgRNAs in total and 1554 D39V genes were targeted without mismatches, it is obvious that some sgRNAs must have multiple exact gene targets. This can indeed be seen in the table below for all strains tested.

Nr. of targets	D39V sgRNAs	D39V Subtotal targets	TIGR4 sgRNAs	TIGR4 Subtotal targets	R6 sgRNAs	R6 Subtotal targets	H19A sgRNAs	H19A Subtotal targets	T19F sgRNAs	T19F Subtotal targets	11A sgRNAs	11A Subtotal targets	G54 sgRNAs	G54 Subtotal targets
1	1441	1441	1096	1096	1297	1297	1054	1054	1055	1055	1015	1015	1064	1064
2	23	46	26	52	22	44	22	44	20	40	25	50	20	40
3	5	15	2	6	3	9	3	9	2	6	2	6	4	12
4	6	24	2	8	6	24	7	28	9	36	6	24	6	24
5	3	15	4	20	5	25	4	20	2	10	2	10	3	15
6	1	6	1	6	0	0	2	12	1	6	1	6	1	6
7	1	7	3	21	1	7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	2	16	0	0
9	0	0	0	0	0	0	0	0	1	9	0	0	0	0
10	0	0	1	10	0	0	1	10	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	1	12	0	0
Totals	1480	1554	1135	1219	1334	1406	1093	1177	1090	1162	1054	1139	1098	1161

Thus, zero-mismatch coverage was in all cases achieved with just a subset of the full library, even for D39V with just 1480 sgRNAs. On top of the 13 sgRNAs for which no target was identified as explained before, 6 sgRNAs were apparently not required in D39V, despite them being designed for this strain. We listed the reasons we found for this observation in the table below.

sgRNA	Targets	Reason not optimal
sgRNA0089	SPV_2129	Outperformed by sgRNA1419
sgRNA0316	SPV_2251	Outperformed by sgRNA0854
sgRNA0853	SPV_0755	Outperformed by sgRNA1340
sgRNA0867	SPV_0794	Design error: same sequence as sgRNA0850
sgRNA0952	SPV_0433	Outperformed by sgRNA1370
sgRNA1013	SPV_0690	Outperformed by sgRNA1302

So, five sgRNAs were simply outperformed by other sgRNAs. Even though they were designed to target these genetic elements specifically, other sgRNAs bind more closely to the start codon of those genetic elements, and also without mismatches. Lastly, one sgRNA was found to have an erroneous sequence, which was found to be a design error.

4 Session information

This document was generated with R Markdown.

```
## R version 3.5.2 (2018-12-20)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18362)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Dutch_Netherlands.1252  LC_CTYPE=Dutch_Netherlands.1252
## [3] LC_MONETARY=Dutch_Netherlands.1252 LC_NUMERIC=C
## [5] LC_TIME=Dutch_Netherlands.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] patchwork_1.0.0  ggrepel_0.8.1    ggplot2_3.3.0    kableExtra_1.1.0
## [5] knitr_1.24       readxl_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.2          RColorBrewer_1.1-2 cellranger_1.1.0
##  [4] pillar_1.4.2        compiler_3.5.2     tools_3.5.2
##  [7] zeallot_0.1.0       digest_0.6.23      evaluate_0.14
## [10] tibble_2.1.3        gtable_0.3.0       viridisLite_0.3.0
## [13] pkgconfig_2.0.2     rlang_0.4.0        rstudioapi_0.10
## [16] yaml_2.2.0          xfun_0.9           withr_2.1.2
## [19] stringr_1.4.0       httr_1.4.1         dplyr_0.8.3
## [22] xml2_1.2.2          vctrs_0.2.0        hms_0.5.1
## [25] tidyselect_0.2.5    webshot_0.5.1      grid_3.5.2
## [28] glue_1.3.1          R6_2.4.0           rmarkdown_1.15
## [31] purrr_0.3.2         readr_1.3.1        magrittr_1.5
## [34] backports_1.1.4     scales_1.0.0       htmltools_0.3.6
## [37] assertthat_0.2.1    rvest_0.3.4        colorspace_1.4-1
## [40] labeling_0.3        stringi_1.4.3      munsell_0.5.0
## [43] crayon_1.3.4
```