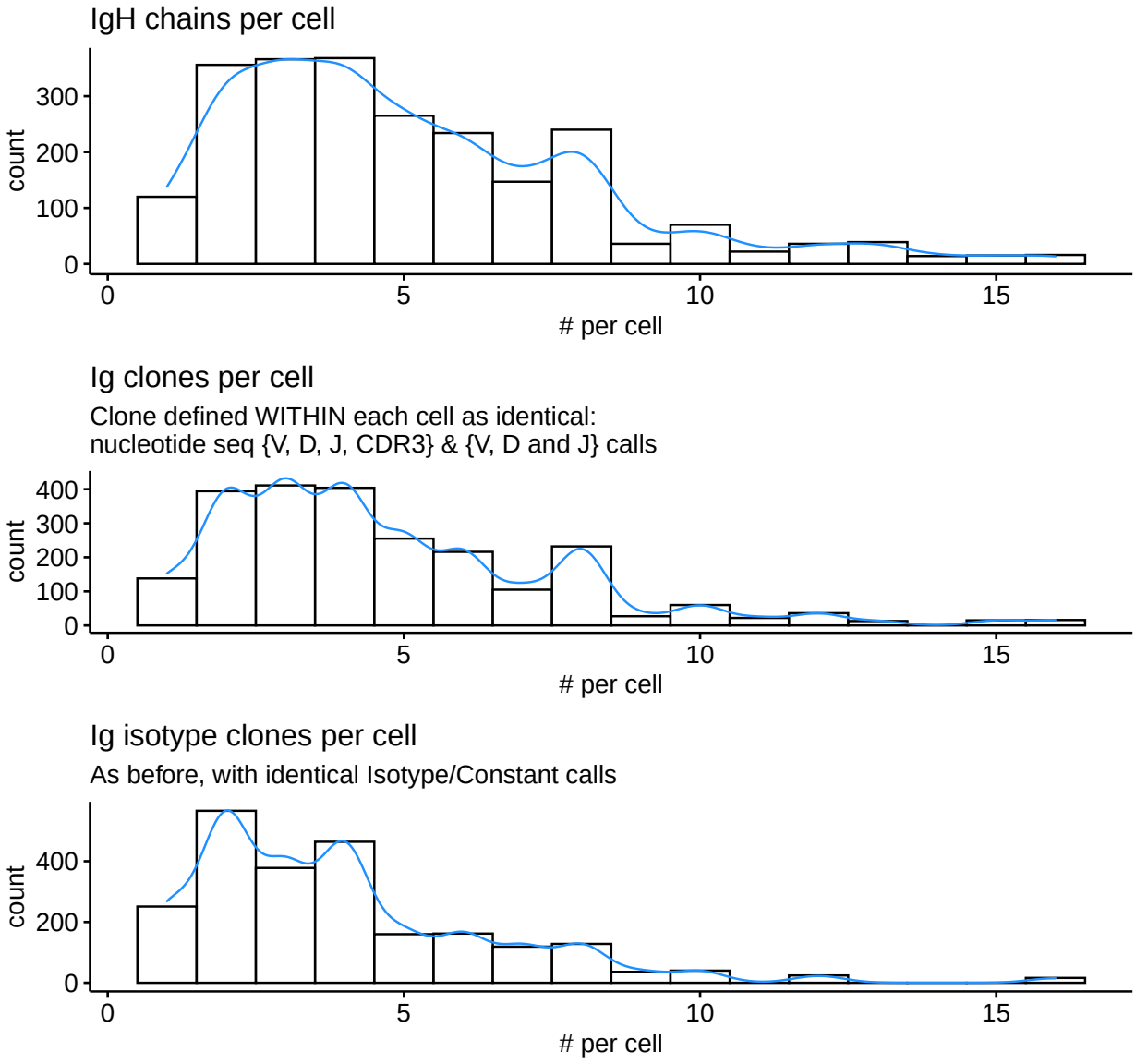
Supporting analysis: IgH chain assignment after VDJPuzzle execution

There are 960 wells (8 NTCs, 952 cells). Of which, 677 give at least one IgH chain.

There 2344 reported IgH chains, across these 677 cells. The median number of chains reported per cell is 4.

Many of these individually reported IgH chains are very similar at the nucleotide sequence level, as shown in the plots below. The first panel is a histogram of number IgH chains reported per cell. In the second panel, IgH chains are called a clone if they have identical nucleotide sequence for V, D, J segments and across CDR3, with the same allele calls for V, D and J segments (median 4). In the third panel, clones are defined more stringently, with the prior definition and the addition of the same isotype call (median 3).



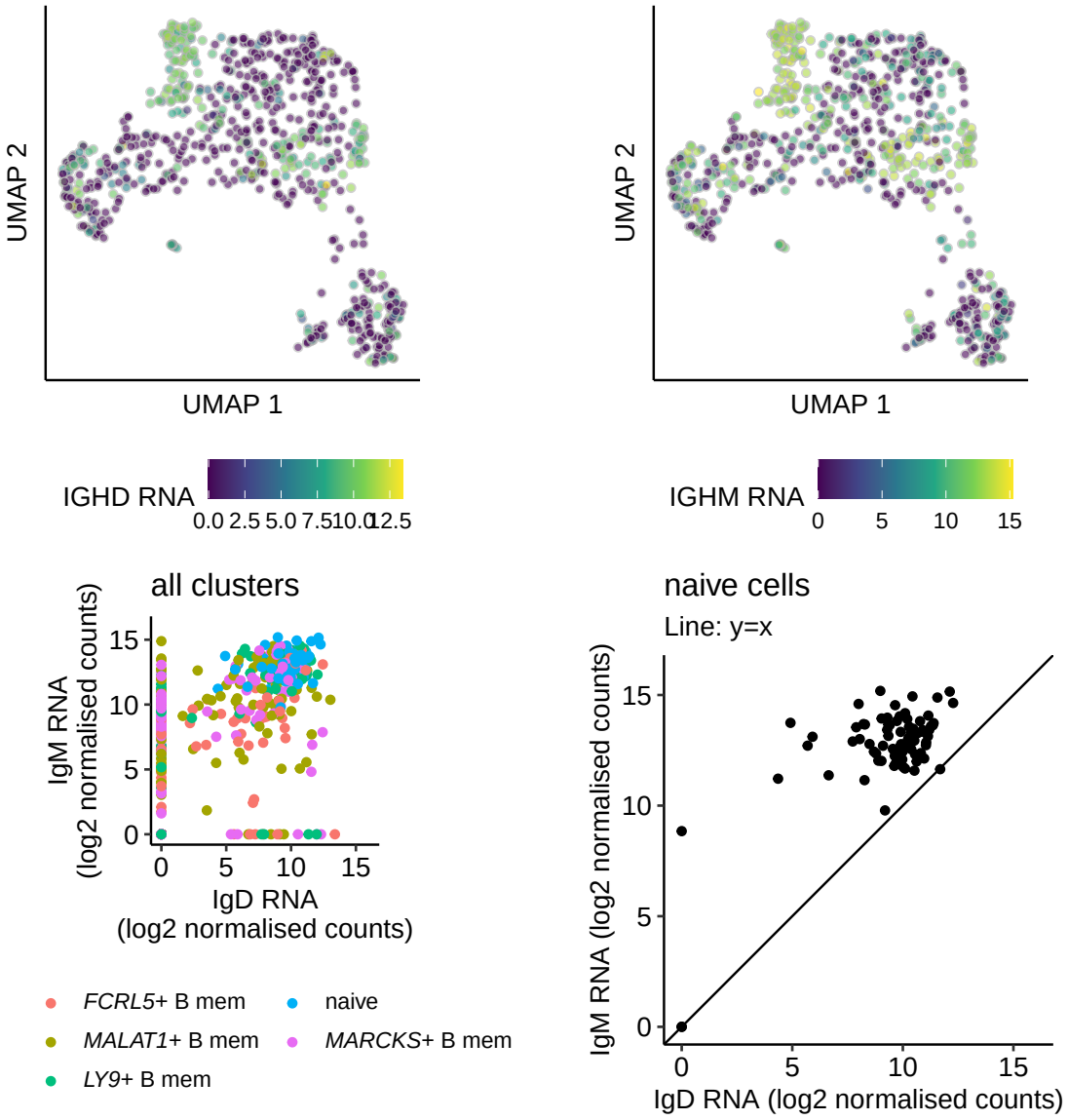
Together these data suggest that many differences between assemblies of the IgH chain are outside of V, D, J or constant region. These different assemblies may reflect small additions/substractions as Trinity makes *de novo* contigs, or may reflect sequencing errors. At the biological meaningful parts of antibody these IgH chains look identical, so it is reasonable to collapse these assemblies together, or adopt one of them for the analyses in Figures 3 and 4.

Existing approach

In Figures 3 and 4, a simple filtering strategy was applied. The most highly expressed IgH chain was used (and other IgH chains discarded): this reduced the dataset from 2344 IgH chains (across 677 cells) to 1212 IgH chains (across 677 cells). Quantitation was performed by kallisto as part of VDJPuzzle. There are 275 cells where the IgH chains expression is estimated as 0 by kallisto. Of these 61 cells have only 1 IgH reported, and there are 214 cells have >1 IgH reported. For these 214 cells, the first IgH is taken (arbitrarily).

A consequence of this is that naive cells are called as IgM+ BCRs in Figure 3, but are shown to be surface IgD+ in Figure 1. It is known that naive cells co-express IgM and IgD, so this supporting analysis confirms this.

Transcriptomic data - naive cells express both IgD and IgM



VDJPuzzle - re-evaluate lower expressed IgH chains

```
## [1] "lane6963.AAGAGGCA.AAGGAGTA.cDNA190807.D9.594V.d42.L001.GRCh38.hisat2.bam"
## [2] "lane6963.AAGAGGCA.ACTGCATA.cDNA190807.E9.594V.d42.L001.GRCh38.hisat2.bam"
## [3] "lane6963.AAGAGGCA.CGTCTAAT.cDNA190807.B9.594V.d42.L001.GRCh38.hisat2.bam"
## [4] "lane6963.AAGAGGCA.CTAAGCCT.cDNA190807.C9.594V.d42.L001.GRCh38.hisat2.bam"
## [5] "lane6963.AAGAGGCA.CTCTCTAT.cDNA190807.H9.594V.d42.L001.GRCh38.hisat2.bam"
## [6] "lane6963.AAGAGGCA.GTAAGGAG.cDNA190807.F9.594V.d42.L001.GRCh38.hisat2.bam"
```

```
##    Mode    TRUE
## logical      94
```

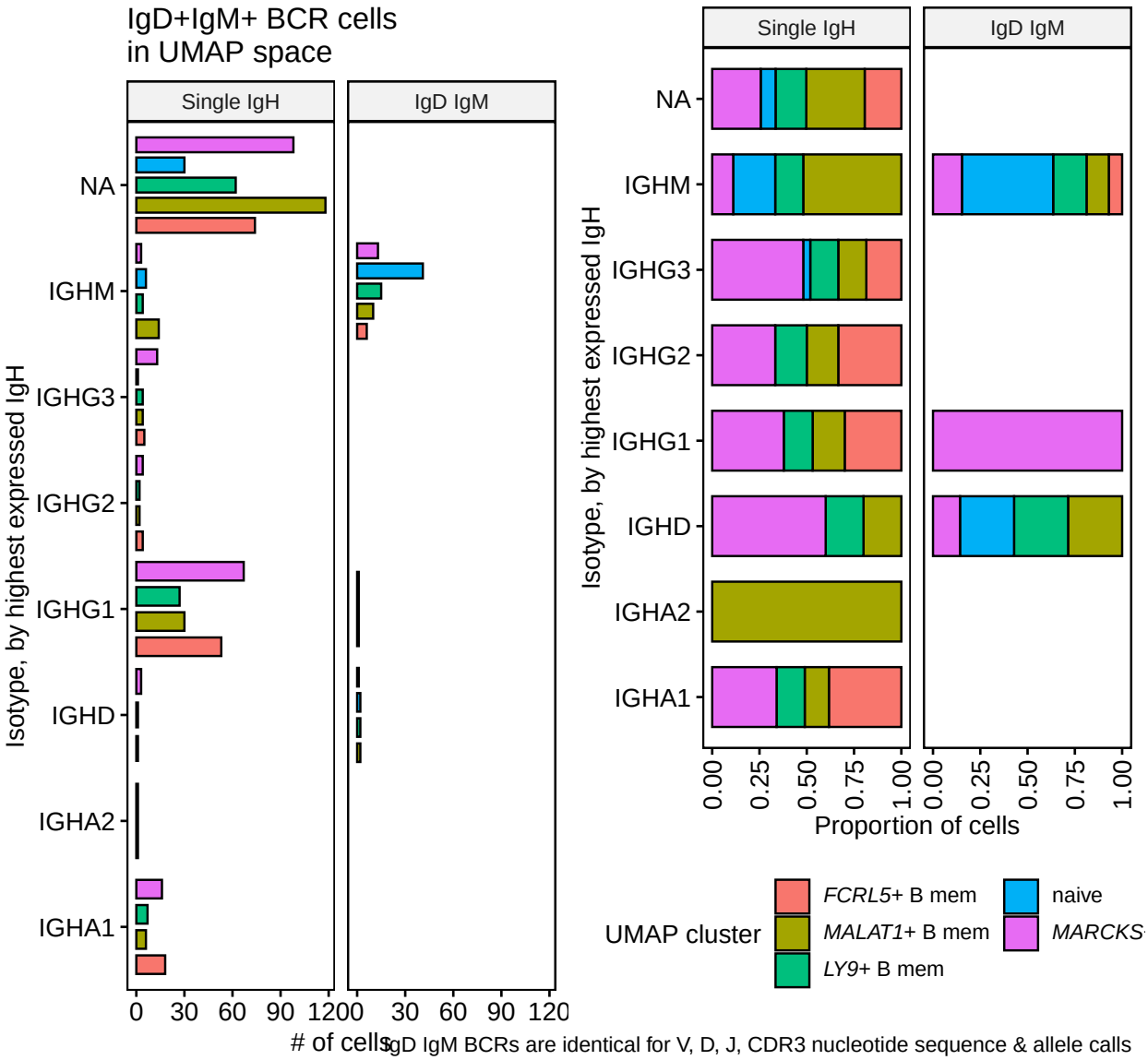
```
##    Mode   FALSE    TRUE
## logical       1      93
```

Supporting analysis

Of 2344 IgH chains, there are 314 heavy chains that use IgM (in 123 cells), and 191 that use IgD (in 100 cells).

Of the 191 IgD+ heavy chains (191 are productive), 88 chains appear identical (present in the same cell with identity at V, D, J and CDR3 nucleotide sequences, and V,D & J allele calls), and 181 are present in the IgM heavy chains - as for IgD comparison - same sequence in the **same** cell.

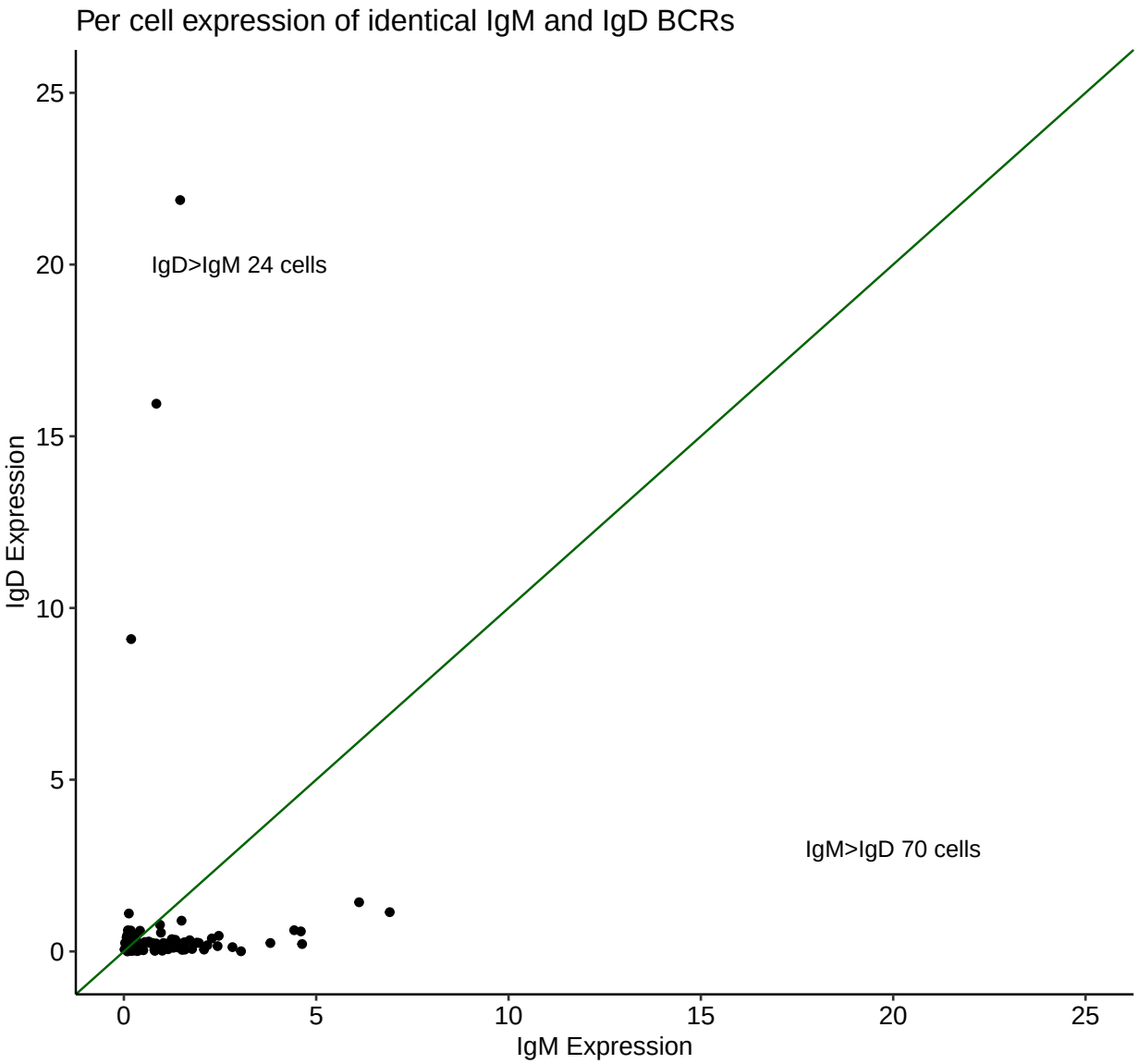
Of the 94 IgD+IgM+ cells, 93 cells have transcriptomic data that pass QC.




```
##    Mode   FALSE    TRUE
## logical      29      94
```

```
##    Mode   FALSE    TRUE
## logical       6      94
```

```
## [1] TRUE
```


Per cell expression of identical IgM and IgD BCRs
IgD>IgM 24 cells
IgM>IgD 70 cells
IgD Expression
IgM Expression
0
5
10
15
20
25

SessionInfo

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: CentOS Linux 7 (Core)
##
## Matrix products: default
## BLAS:   /bi/apps/R/3.6.1/lib64/R/lib/libRblas.so
## LAPACK: /bi/apps/R/3.6.1/lib64/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C         LC_TIME=C
##  [4] LC_COLLATE=C         LC_MONETARY=C        LC_MESSAGES=C
##  [7] LC_PAPER=C           LC_NAME=C            LC_ADDRESS=C
## [10] LC_TELEPHONE=C       LC_MEASUREMENT=C     LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] scater_1.14.5               SingleCellExperiment_1.8.0
##  [3] SummarizedExperiment_1.16.0 DelayedArray_0.12.0
##  [5] BiocParallel_1.20.0         matrixStats_0.55.0
##  [7] Biobase_2.46.0              GenomicRanges_1.38.0
##  [9] GenomeInfoDb_1.22.0         IRanges_2.20.1
## [11] S4Vectors_0.24.1            BiocGenerics_0.32.0
## [13] cowplot_1.0.0               ggpubr_0.2.4
## [15] magrittr_1.5                forcats_0.4.0
## [17] stringr_1.4.0               dplyr_1.0.2
## [19] purrr_0.3.3                 readr_1.3.1
## [21] tidyr_1.0.0                 tibble_3.0.4
## [23] ggplot2_3.3.2               tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] ggtext_0.1.1             bitops_1.0-6             fs_1.3.1
##  [4] lubridate_1.7.4          httr_1.4.1               tools_3.6.1
##  [7] backports_1.1.5          R6_2.4.1                 irlba_2.3.3
## [10] vipor_0.4.5              DBI_1.1.0                colorspace_1.4-1
## [13] withr_2.1.2              gridExtra_2.3            tidyselect_1.1.0
## [16] compiler_3.6.1           cli_2.0.0                rvest_0.3.5
## [19] BiocNeighbors_1.4.1      xml2_1.2.2               labeling_0.3
## [22] scales_1.1.0             digest_0.6.23            rmarkdown_2.0
## [25] XVector_0.26.0           pkgconfig_2.0.3          htmltools_0.4.0
## [28] dbplyr_1.4.2             rlang_0.4.10             readxl_1.3.1
## [31] rstudioapi_0.10          DelayedMatrixStats_1.8.0 farver_2.0.1
## [34] generics_0.0.2           jsonlite_1.6             RCurl_1.95-4.12
## [37] BiocSingular_1.2.0       GenomeInfoDbData_1.2.2   Matrix_1.2-17
## [40] Rcpp_1.0.3               ggbeeswarm_0.6.0         munsell_0.5.0
## [43] fansi_0.4.0              viridis_0.5.1            lifecycle_0.2.0
## [46] stringi_1.4.3            yaml_2.2.0               zlibbioc_1.32.0
## [49] grid_3.6.1               crayon_1.3.4             lattice_0.20-38
## [52] haven_2.2.0              gridtext_0.1.4           hms_0.5.2
## [55] knitr_1.26               pillar_1.4.7             markdown_1.1
```

```
## [58] ggsignif_0.6.0          reprex_0.3.0            glue_1.4.2
## [61] evaluate_0.14           modelr_0.1.5            vctrs_0.3.6
## [64] cellranger_1.1.0        gtable_0.3.0            assertthat_0.2.1
## [67] xfun_0.11               rsvd_1.0.2              broom_0.7.3
## [70] viridisLite_0.3.0       beeswarm_0.2.3          ellipsis_0.3.0
```