Untitled Draft

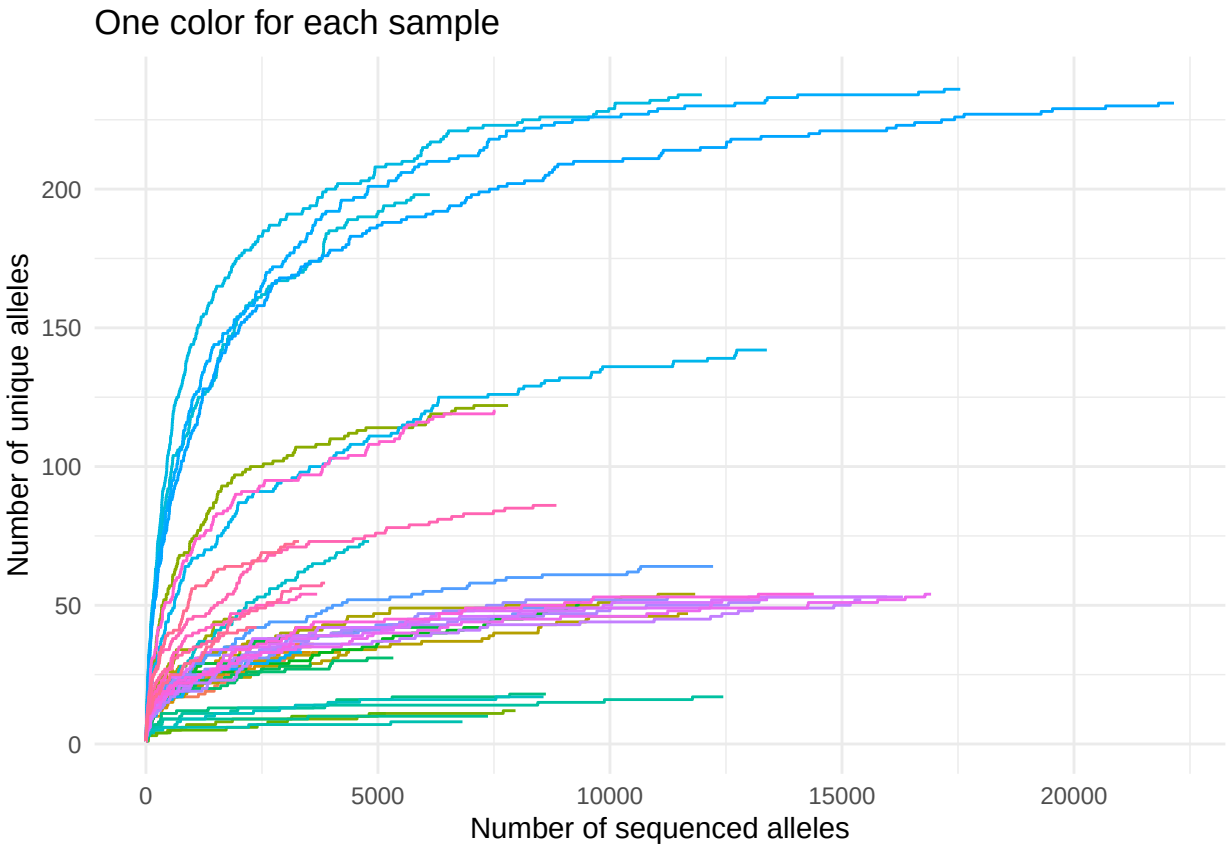
Report Author

01 October, 2020

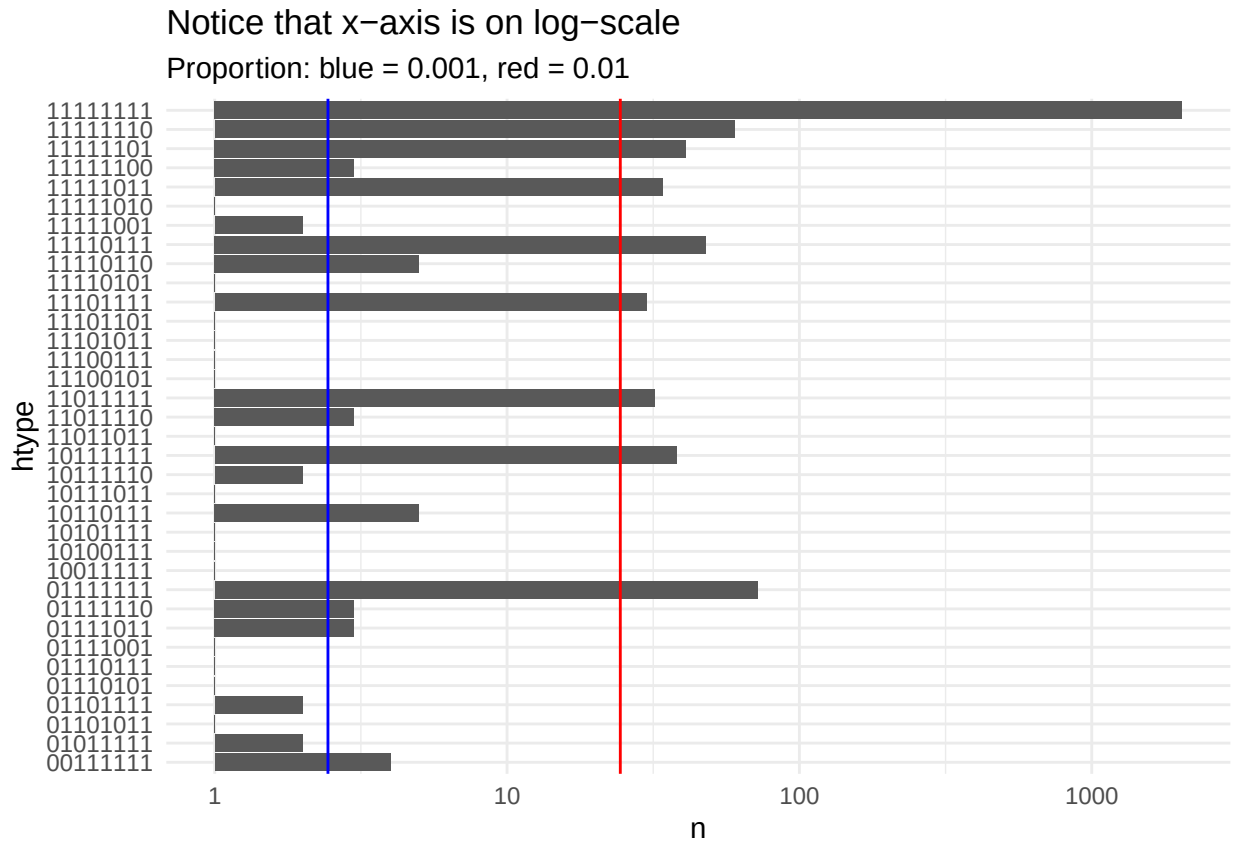
Data

```
## # A tibble: 374,923 x 6
##    sample       amplicon chr     start readseq htype
##    <chr>           <dbl> <chr>   <dbl>   <dbl> <chr>
##  1 1_M.bscflags        1 chr1  3567622       1 11111111
##  2 1_M.bscflags        1 chr1  3567622       2 11111111
##  3 1_M.bscflags        1 chr1  3567622       3 11111111
##  4 1_M.bscflags        1 chr1  3567622       4 00111111
##  5 1_M.bscflags        1 chr1  3567622       5 01111111
##  6 1_M.bscflags        1 chr1  3567622       6 11111111
##  7 1_M.bscflags        1 chr1  3567622       7 11111111
##  8 1_M.bscflags        1 chr1  3567622       8 11111111
##  9 1_M.bscflags        1 chr1  3567622       9 11111111
## 10 1_M.bscflags        1 chr1  3567622      10 11111111
## # ... with 374,913 more rows
```

Number of unique reads by number of reads



Count of alleles in 1_M.bscflags sample



Proportion 0.05

```
## # A tibble: 47 x 4
##    sample          amplicon n_dist htype_length
##    <chr>              <dbl>  <int>        <int>
##  1 1_M.bscflags           1      1            8
##  2 2_M.bscflags           1      1            8
##  3 3_M.bscflags           1      1            8
##  4 33_BS.bscflags         1      3            8
##  5 34_BS.bscflags         1      1            8
##  6 4_M.bscflags           1      1            8
##  7 5_M.bscflags           1      1            8
##  8 6_M.bscflags           1      1            8
##  9 60_M.bscflags          1      2            8
## 10 61_D.bscflags          1      3            8
## # ... with 37 more rows
```

Proportion 0.02

```
## # A tibble: 47 x 4
```

```
##    sample         amplicon n_dist htype_length
##    <chr>             <dbl>  <int>        <int>
##  1 1_M.bscflags          1      3            8
##  2 2_M.bscflags          1      3            8
##  3 3_M.bscflags          1      3            8
##  4 33_BS.bscflags        1      3            8
##  5 34_BS.bscflags        1      6            8
##  6 4_M.bscflags          1      4            8
##  7 5_M.bscflags          1      5            8
##  8 6_M.bscflags          1      5            8
##  9 60_M.bscflags         1      3            8
## 10 61_D.bscflags         1      5            8
## # ... with 37 more rows
```

Proportion 0.01

```
## # A tibble: 47 x 4
##    sample         amplicon n_dist htype_length
##    <chr>             <dbl>  <int>        <int>
##  1 1_M.bscflags          1      9            8
##  2 2_M.bscflags          1      9            8
##  3 3_M.bscflags          1      9            8
##  4 33_BS.bscflags        1      3            8
##  5 34_BS.bscflags        1      9            8
##  6 4_M.bscflags          1      9            8
##  7 5_M.bscflags          1      9            8
##  8 6_M.bscflags          1      9            8
##  9 60_M.bscflags         1      6            8
## 10 61_D.bscflags         1      9            8
## # ... with 37 more rows
```

Proportion 0.001

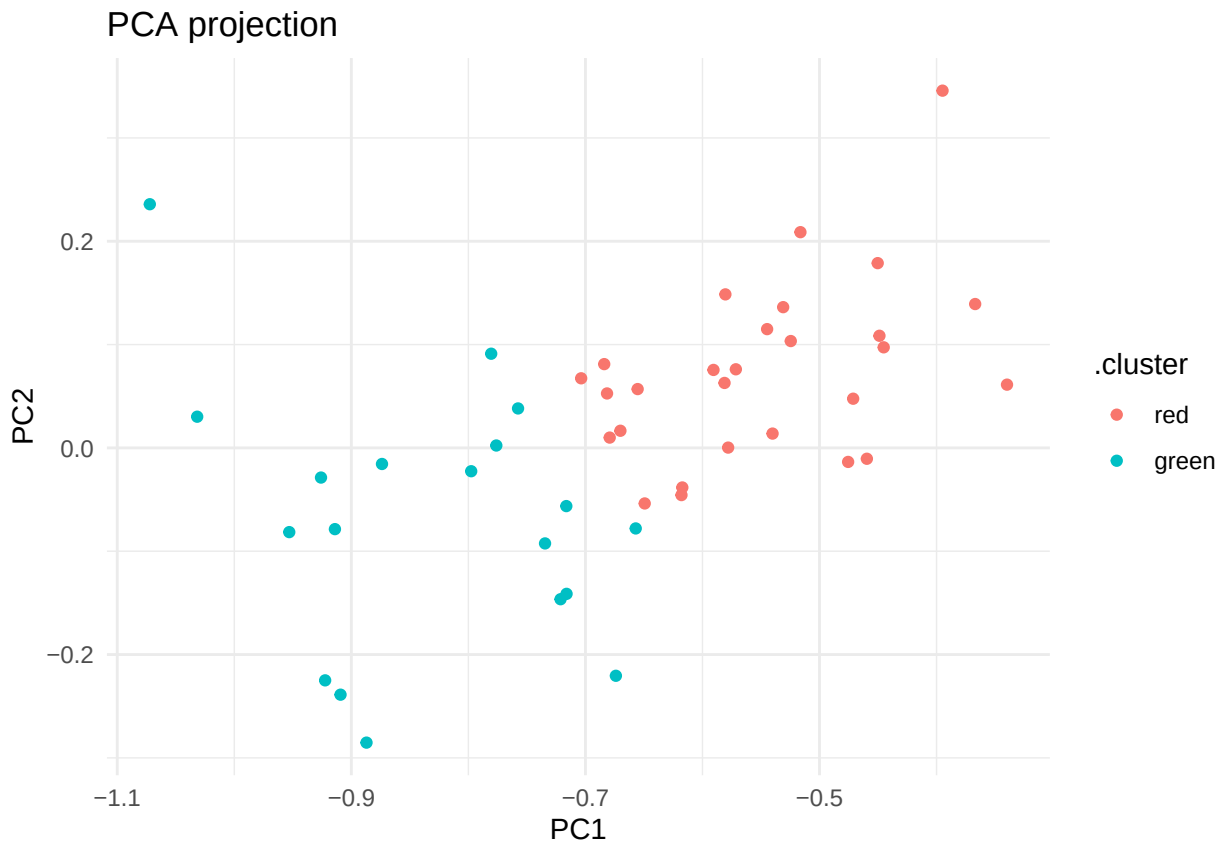
```
## # A tibble: 47 x 4
##    sample         amplicon n_dist htype_length
##    <chr>             <dbl>  <int>        <int>
##  1 1_M.bscflags          1     16            8
##  2 2_M.bscflags          1     13            8
##  3 3_M.bscflags          1     13            8
##  4 33_BS.bscflags        1      3            8
##  5 34_BS.bscflags        1     27            8
##  6 4_M.bscflags          1     15            8
##  7 5_M.bscflags          1     21            8
##  8 6_M.bscflags          1     11            8
##  9 60_M.bscflags         1     19            8
## 10 61_D.bscflags         1     28            8
## # ... with 37 more rows
```

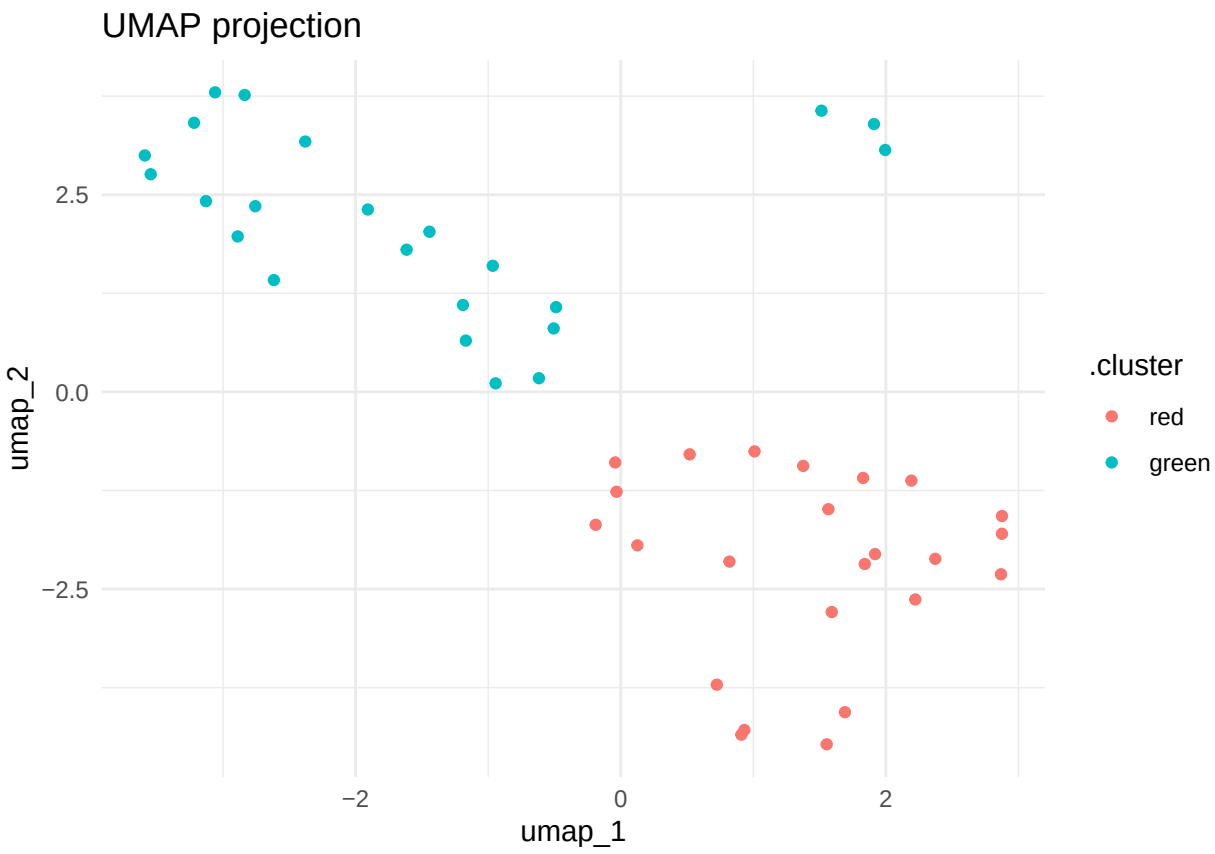
Clustering

We calculated the number of patterns that appears at least 1% of the times for each sample and amplicon. This gives us data on the following form:

```
## # A tibble: 1,583 x 4
##    sample          amplicon n_pattern htype_length
##    <chr>              <dbl>     <int>        <int>
##  1 1_M.bscflags           1         9            8
##  2 2_M.bscflags           1         9            8
##  3 3_M.bscflags           1         9            8
##  4 33_BS.bscflags         1         3            8
##  5 34_BS.bscflags         1         9            8
##  6 4_M.bscflags           1         9            8
##  7 5_M.bscflags           1         9            8
##  8 6_M.bscflags           1         9            8
##  9 60_M.bscflags          1         6            8
## 10 61_D.bscflags          1         9            8
## # ... with 1,573 more rows
```

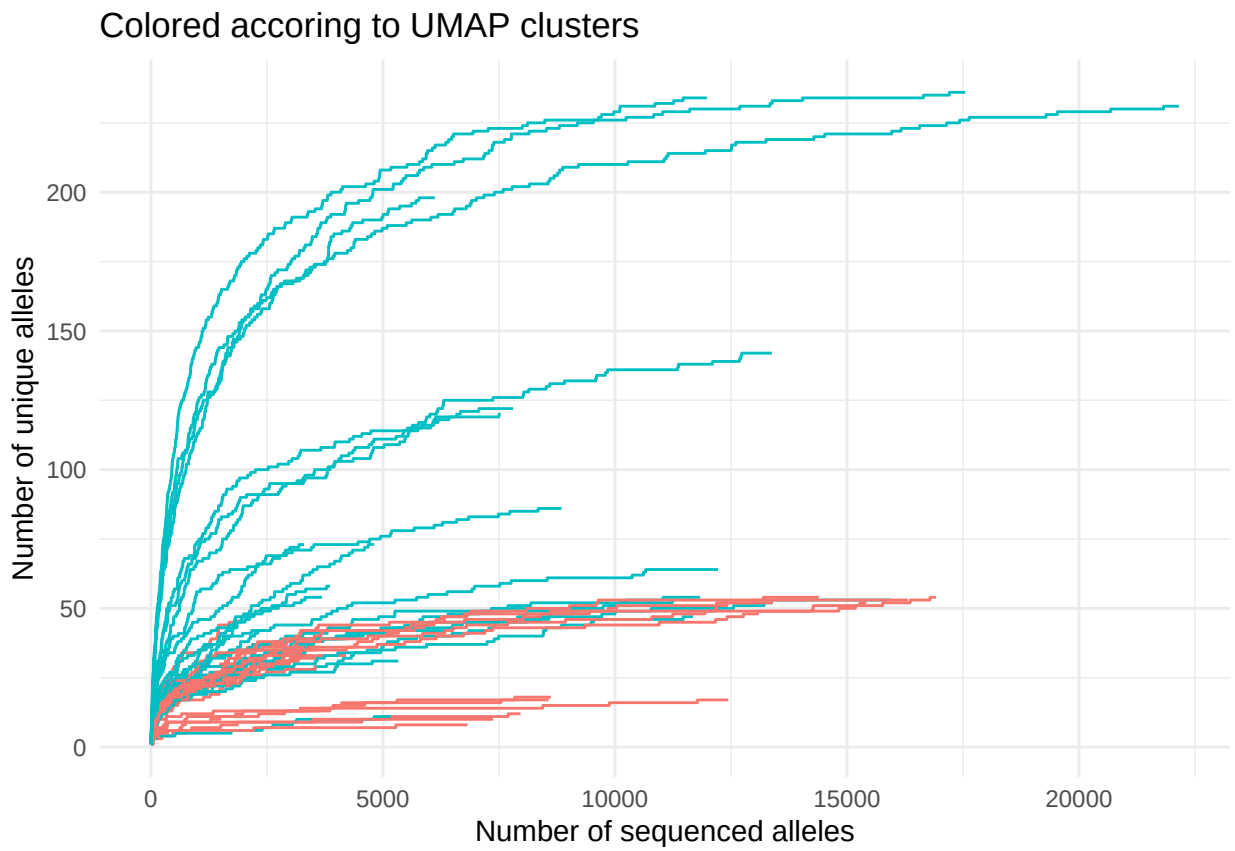
We then transform the data to have amplicons as columns and samples as rows. Do mean-imputation of missing values. Then we project the data down to a lower dimensional space and see if we can find cluster with an kmeans algorithm.



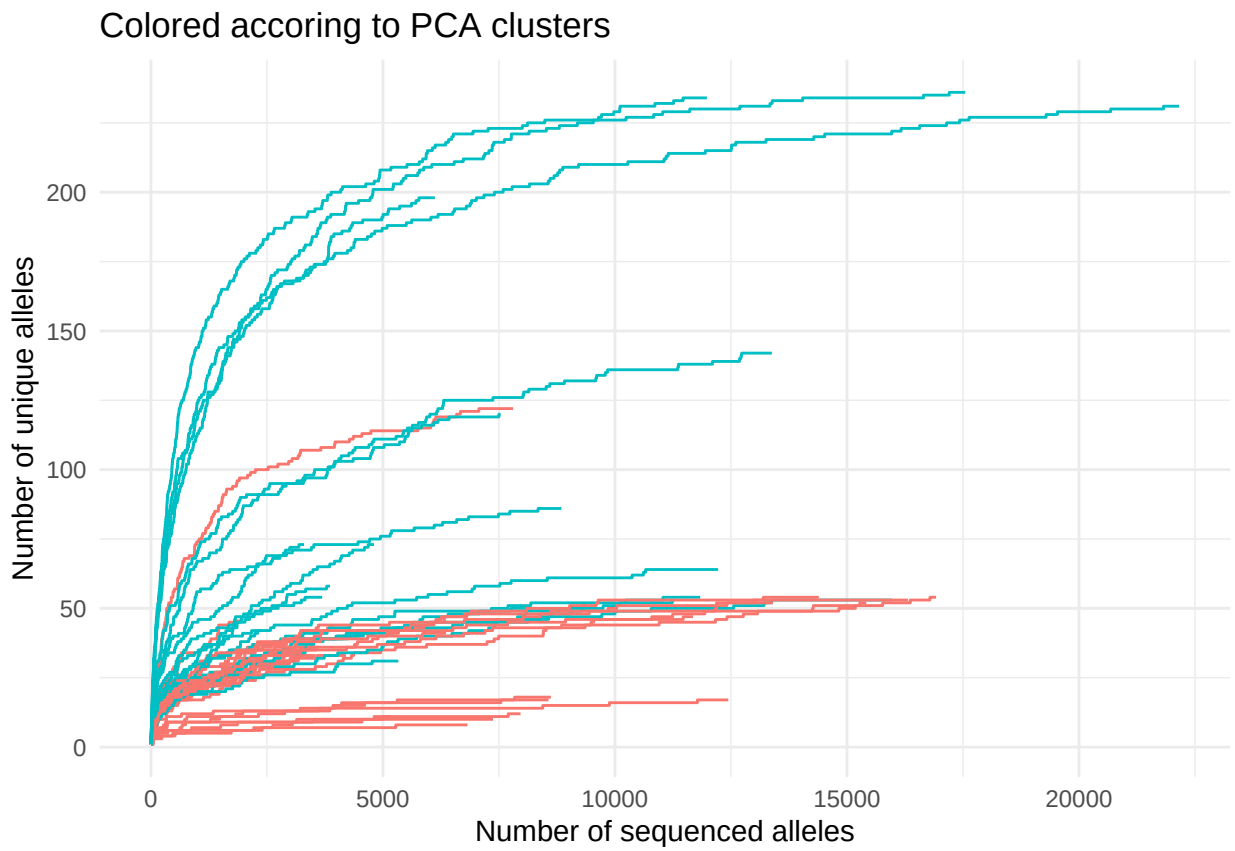


Clusters applied to first chart

We take the clusters we found in the UMAP project and color the first chart accordingly.



We take the clusters we found in the PCA project and color the first chart accordingly.



Clusters

```
## Joining, by = "sample"
## Joining, by = "sample"
```

sample	pca_cluster	umap_cluster	Type	description
1_M.bscflags	red	red	cell line	clonal passage18 subclone1
2_M.bscflags	red	red	cell line	clonal passage18 subclone3
3_M.bscflags	red	red	xenograph	clonal xenograph 4 months
33_BS.bscflags	red	red	xenograph	polyclonal xenograph 4 months
34_BS.bscflags	red	red	cell line	clonal passage24 subclone5
4_M.bscflags	red	red	cell line	clonal passage 70 subclone1
5_M.bscflags	red	red	cell line	clonal passage 70 subclone 3
6_M.bscflags	red	red	colon crypt	NA
62_D.bscflags	red	red	colon crypt	NA
64_D.bscflags	red	red	si crypt	NA
67_D.bscflags	red	red	colon crypt	NA
7_M.bscflags	red	red	colon crypt	NA
71_D.bscflags	red	red	colon crypt	NA
72_D.bscflags	red	red	colon crypt	NA
73_D.bscflags	red	red	colon crypt	NA
74_D.bscflags	red	red	colon crypt	NA
75_D.bscflags	red	red	bulk DNA	tumor J side A

sample	pca_cluster	umap_cluster	Type	description
8_M.bscflags	red	red	bulk DNA	normal colon M
86_D.bscflags	red	red	NA	NA
87_D.bscflags	red	red	NA	NA
88_D.bscflags	red	red	NA	NA
89_D.bscflags	red	red	cell line	clonal passage 105 subclone3
9_M.bscflags	red	red	NA	NA
90_D.bscflags	red	red	SURF	Tumor J
60_M.bscflags	red	green	colon crypt	NA
61_D.bscflags	green	green	colon crypt	NA
63_D.bscflags	red	green	si crypt	NA
65_D.bscflags	red	green	si crypt	NA
68_D.bscflags	green	green	colon crypt	NA
69_D.bscflags	red	green	cell line	clonal passage 70 subclone5
70_D.bscflags	green	green	colon crypt	NA
76_D.bscflags	green	green	bulk DNA	tumor J side B
77_D.bscflags	green	green	bulk DNA	normal colon J
78_D.bscflags	green	green	bulk DNA	tumor M side A
79_D.bscflags	green	green	cell line	clonal passage 105 subclone1
80_D.bscflags	green	green	bulk DNA	normal colon D
81_D.bscflags	green	green	xenograph	polyclonal
82_D.bscflags	green	green	xenograph	clonal 4 months
83_D.bscflags	green	green	xenograph	clonal 4 months
84_D.bscflags	green	green	xenograph	clonal 4 months
85_D.bscflags	green	green	NA	NA
91_D.bscflags	green	green	NA	NA
92_D.bscflags	green	green	NA	NA
93_D.bscflags	green	green	SURF	Tumor J
94_D.bscflags	green	green	SURF	Tumor J
95_D.bscflags	green	green	SURF	Tumor J
96_D.bscflags	green	green	cell line	clonal passage 105 subclone5

Reproducibility

Reproducibility receipt

```
## [1] "2020-10-01 12:59:50 PDT"

## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Mojave 10.14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

```
## other attached packages:
##  [1] drc_3.0-1           MASS_7.3-51.6       embed_0.1.1
##  [4] recipes_0.1.13.9001 cpp11_0.2.1         readxl_1.3.1
##  [7] forcats_0.5.0       stringr_1.4.0       dplyr_1.0.2
## [10] purrr_0.3.4         readr_1.3.1         tidyr_1.1.2
## [13] tibble_3.0.3        ggplot2_3.3.2       tidyverse_1.3.0
## [16] rmarkdown_2.3       drake_7.12.4        dotenv_1.0.2
## [19] conflicted_1.0.4
##
## loaded via a namespace (and not attached):
##   [1] backports_1.1.9        plyr_1.8.6             igraph_1.2.5
##   [4] splines_4.0.2          storr_1.2.1            crosstalk_1.1.0.1
##   [7] tfruns_1.4             TH.data_1.0-10         rstantools_2.1.1
##  [10] inline_0.3.16          digest_0.6.25          htmltools_0.5.0
##  [13] rsconnect_0.8.16       fansi_0.4.1            magrittr_1.5
##  [16] memoise_1.1.0          base64url_1.4          openxlsx_4.1.5
##  [19] modelr_0.1.8           gower_0.2.2            matrixStats_0.56.0
##  [22] RcppParallel_5.0.2     sandwich_2.5-1         xts_0.12.1
##  [25] prettyunits_1.1.1      colorspace_1.4-1       blob_1.2.1
##  [28] rvest_0.3.6            haven_2.3.1            xfun_0.17
##  [31] callr_3.4.4            crayon_1.3.4.9000      jsonlite_1.7.1
##  [34] lme4_1.1-23            zeallot_0.1.0          survival_3.2-3
##  [37] zoo_1.8-8              glue_1.4.2             gtable_0.3.0
##  [40] ipred_0.9-9            V8_3.2.0               car_3.0-9
##  [43] pkgbuild_1.1.0         rstan_2.21.2           abind_1.4-5
##  [46] scales_1.1.1           mvtnorm_1.1-1          decor_1.0.0
##  [49] DBI_1.1.0              miniUI_0.1.1.1         Rcpp_1.0.5
##  [52] plotrix_3.7-8          xtable_1.8-4           progress_1.2.2
##  [55] reticulate_1.16        foreign_0.8-80         txtq_0.2.3
##  [58] StanHeaders_2.21.0-6   stats4_4.0.2           lava_1.6.7
##  [61] prodlim_2019.11.13     DT_0.15                htmlwidgets_1.5.1.9002
##  [64] httr_1.4.2             threejs_0.3.3          ellipsis_0.3.1
##  [67] farver_2.0.3           loo_2.3.1              pkgconfig_2.0.3
##  [70] uwot_0.1.8             nnet_7.3-14            dbplyr_1.4.4
##  [73] utf8_1.1.4             labeling_0.3           tidyselect_1.1.0
##  [76] rlang_0.4.7            reshape2_1.4.4         later_1.1.0.1
##  [79] munsell_0.5.0          cellranger_1.1.0       tools_4.0.2
##  [82] cli_2.0.2              generics_0.0.2         broom_0.7.0
##  [85] ggridges_0.5.2         evaluate_0.14          fastmap_1.0.1
##  [88] yaml_2.2.1             processx_3.4.4         knitr_1.29
##  [91] fs_1.5.0               zip_2.1.0              nlme_3.1-148
##  [94] whisker_0.4            mime_0.9               rstanarm_2.21.1
##  [97] xml2_1.3.2             brio_1.0.0             compiler_4.0.2
## [100] bayesplot_1.7.2        shinythemes_1.1.2      rstudioapi_0.11
## [103] curl_4.3               filelock_1.0.2         reprex_0.3.0
## [106] statmod_1.4.34         stringi_1.5.3          highr_0.8
## [109] ps_1.3.4               desc_1.2.0             lattice_0.20-41
## [112] Matrix_1.2-18          tensorflow_2.2.0       keras_2.3.0.0
## [115] nloptr_1.2.2.2         markdown_1.1           shinyjs_2.0.0
## [118] vctrs_0.3.4            pillar_1.4.6           lifecycle_0.2.0
## [121] data.table_1.13.0      httpuv_1.5.4           R6_2.4.1
## [124] promises_1.1.1         rio_0.5.16             gridExtra_2.3
## [127] codetools_0.2-16       boot_1.3-25            colourpicker_1.0
## [130] gtools_3.8.2           assertthat_0.2.1       rprojroot_1.3-2
```

```
## [133] withr_2.2.0           shinystan_2.5.0       multcomp_1.4-13
## [136] parallel_4.0.2        hms_0.5.3             grid_4.0.2
## [139] rpart_4.1-15          timeDate_3043.102     class_7.3-17
## [142] minqa_1.2.4           carData_3.0-4         shiny_1.5.0
## [145] lubridate_1.7.9       base64enc_0.1-3       dygraphs_1.1.1.6
```