ALL pfMCA QC

Virginia Howick

25/05/2020

```r
setwd("/Users/vh3/Documents/PfMCA/ANALYSIS_2")
require("Matrix")
library(scater, quietly = TRUE)
require("SingleCellExperiment")
options(stringsAsFactors = FALSE)
library(plotly)
library(scran)
library(devtools)
```

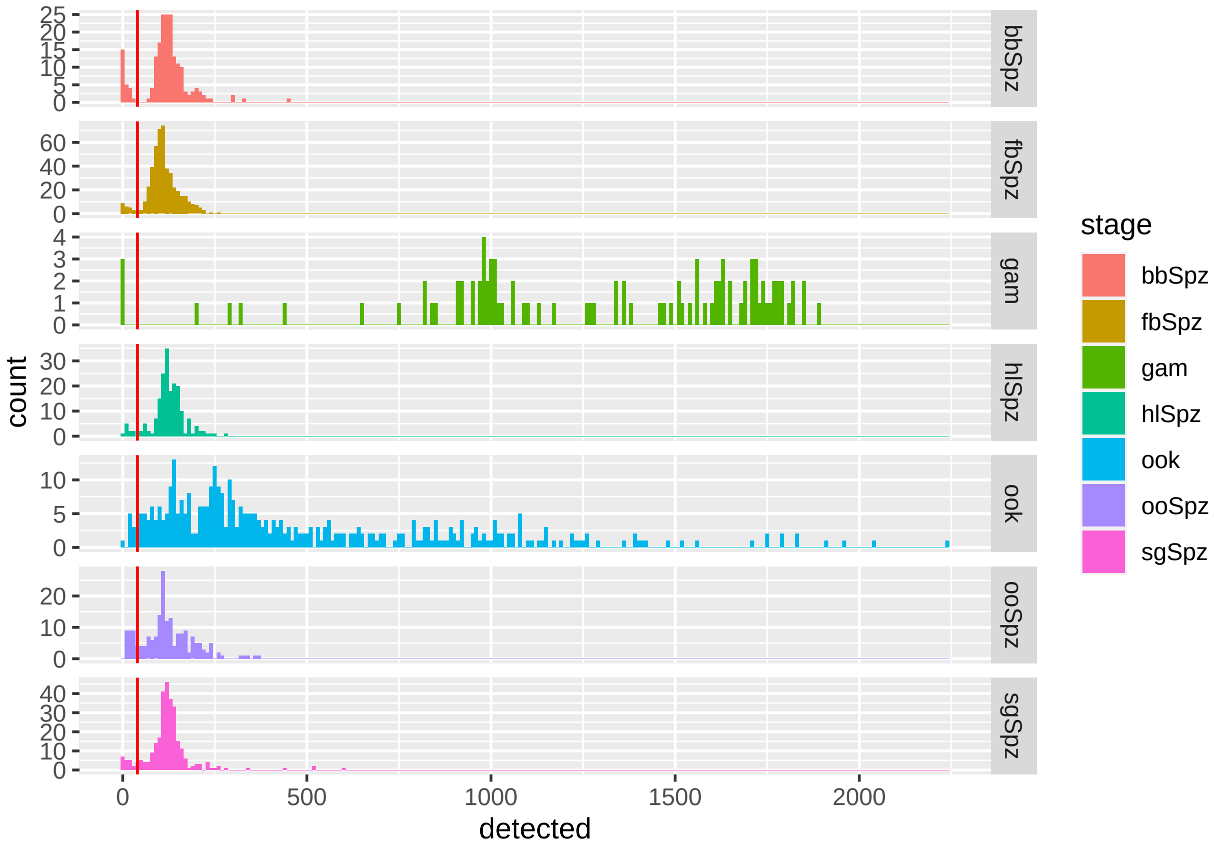
```r
molecules <- read.table("/Users/vh3/Documents/PfMCA/expression_matrices/pfMCA_counts_20200516.csv", hea
anno <- read.delim("/Users/vh3/Documents/PfMCA/expression_matrices/pfMCA_pheno.csv", header = TRUE, sep

anno <- anno[match(colnames(molecules), anno$xfilename), ]

mca <- SingleCellExperiment(assays = list(
  counts = as.matrix(molecules),
  logcounts = log2(as.matrix(molecules) + 1)
), colData = anno)
```

Calculate QC metrics for cells and gene, remove failed spz run

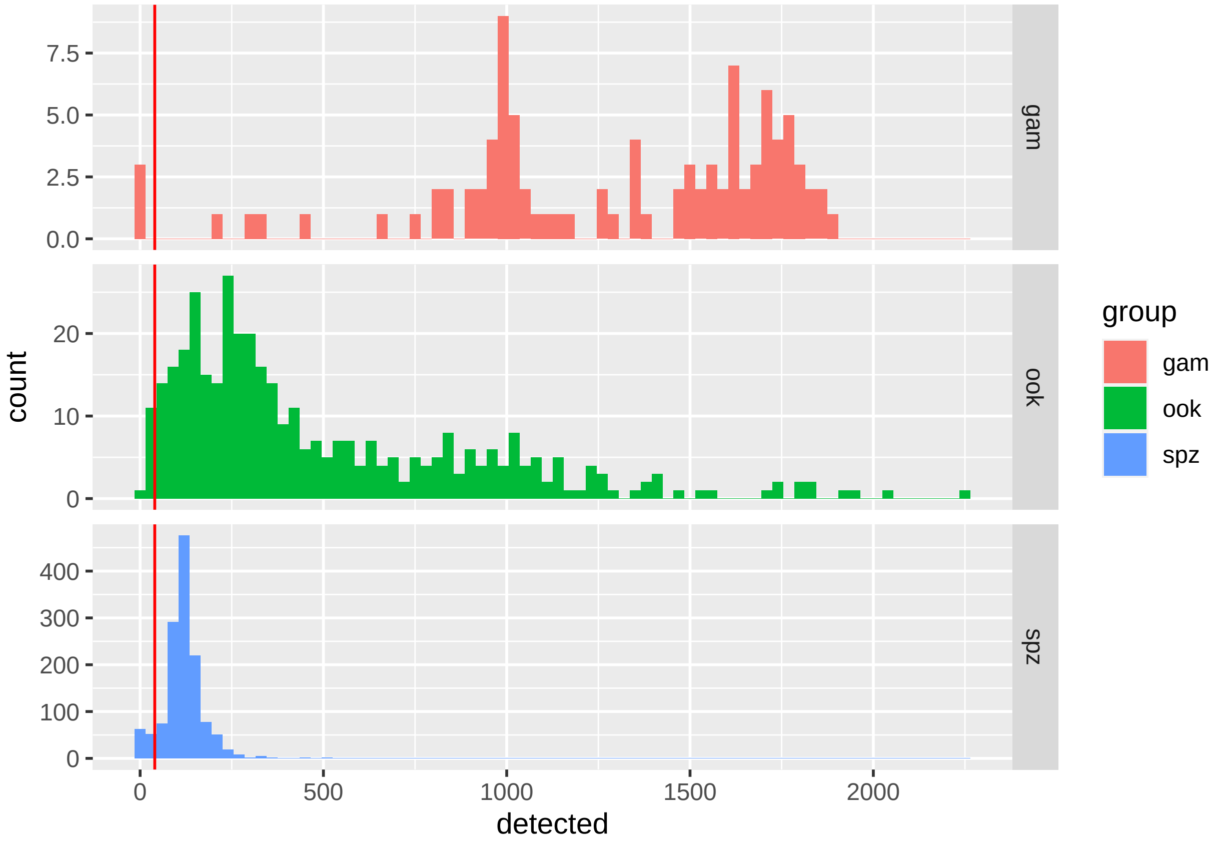
```r
CellQC <- perCellQCMetrics(mca)
FeatQC <- perFeatureQCMetrics(mca)

colData(mca) <- cbind(colData(mca), CellQC)
rowData(mca) <- cbind(rowData(mca), FeatQC)

mca <- mca[, mca$stage != "spz"]
mca <- mca[, mca$stage != "mozSpz"]
mca <- mca[, mca$stage != "ffeSpz"]
```

QC by stage

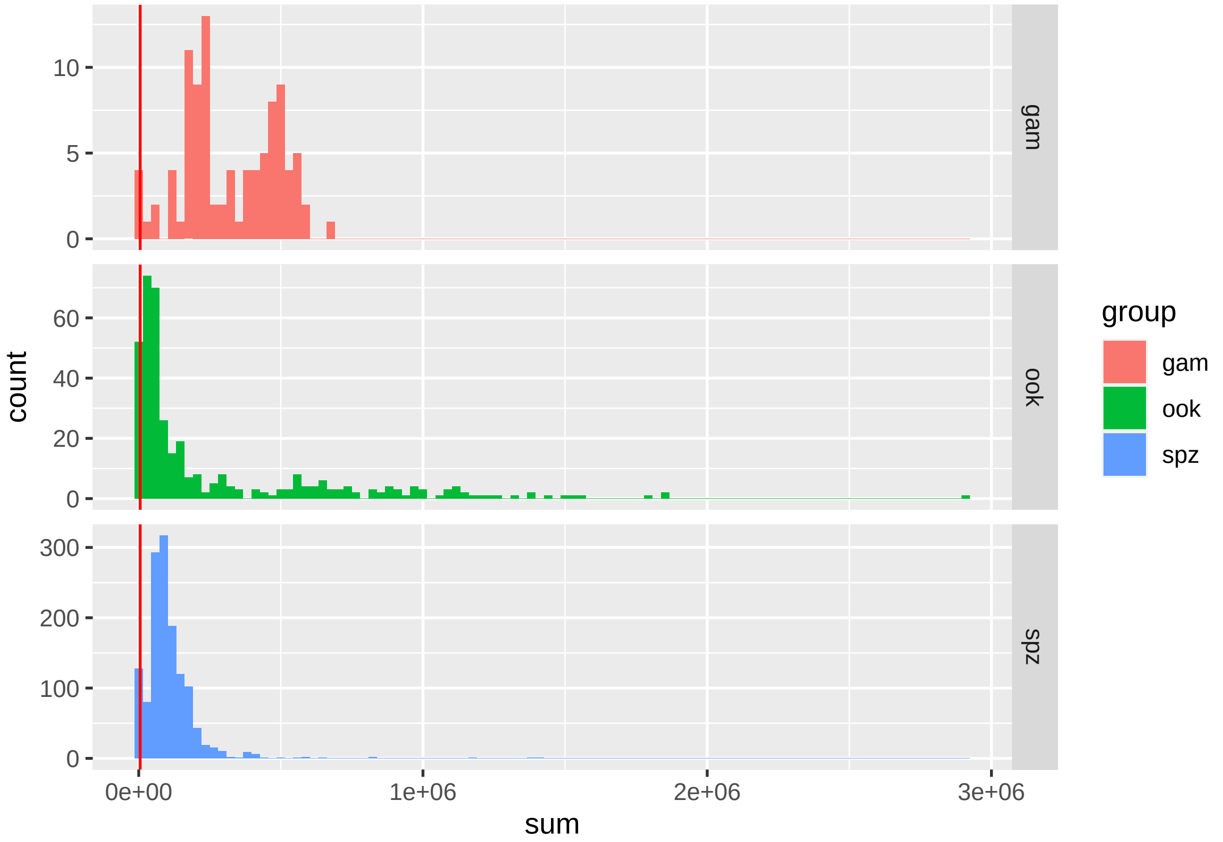
```r
tab <- as.data.frame(colData(mca))
ggplot(tab, aes(x=detected, fill = stage)) + geom_histogram(binwidth = 10) + facet_grid(stage~., scales
```



```
mca$group <- rep("spz", length(mca$sample_id))
mca[, which(mca$stage=="gam")]$group <- "gam"
mca[, which(mca$stage=="ook")]$group <- "ook"

tab <- as.data.frame(colData(mca))
ggplot(tab, aes(x=detected, fill = group)) + geom_histogram(binwidth = 30) + facet_grid(group~., scales=
```



```
tab <- as.data.frame(colData(mca))
ggplot(tab, aes(x=sum, fill = group)) + geom_histogram(bins=100) + facet_grid(group~., scales="free") +g
```



QC of single cells based on Txnal profile

```
mca <- mca[, mca$num_cells=="SC"]

mcasmall <- mca[,colData(mca)$group == "spz"]
mcamedium <- mca[, colData(mca)$group == "ook" ]
mcalarge <- mca[, colData(mca)$group == "gam" ]

#QC of gams

filter_by_total_counts <- (mcalarge$sum > 10000)
table(filter_by_total_counts)
```

```
## filter_by_total_counts
## FALSE  TRUE
##     4    91
```

```
# Filter cells with low numbers of features detected
filter_by_expr_features <- (mcalarge$detected > 500)
table(filter_by_expr_features)
```

```
## filter_by_expr_features
## FALSE  TRUE
##     7    88
```

```
mcalarge$use <- (filter_by_expr_features & filter_by_total_counts)
table(mcalarge$use)
```

```
##
```

```
## FALSE  TRUE
##     7    88
```

```
#QC of ookinetes

# Filter cells with low counts
filter_by_total_counts <- (mcamedium$sum > 5000)
table(filter_by_total_counts)
```

```
## filter_by_total_counts
## FALSE  TRUE
##    17   366
```

```
# Filter cells with low numbers of features detected
filter_by_expr_features <- (mcamedium$detected > 400)
table(filter_by_expr_features)
```

```
## filter_by_expr_features
## FALSE  TRUE
##   218   165
```

```
mcamedium$use <- (filter_by_expr_features & filter_by_total_counts)
table(mcamedium$use)
```

```
##
## FALSE  TRUE
##   218   165
```

```
##QC of Spz

# Filter cells with low counts
filter_by_total_counts <- (mcasmall$sum > 5000)
table(filter_by_total_counts)
```

```
## filter_by_total_counts
## FALSE  TRUE
##    95  1235
```

```
# Filter cells with low numbers of features detected
filter_by_expr_features <- (mcasmall$detected > 40)
table(filter_by_expr_features)
```

```
## filter_by_expr_features
## FALSE  TRUE
##   102  1228
```

```
mcasmall$use <- (filter_by_expr_features & filter_by_total_counts)
table(mcasmall$use)
```

```
##
## FALSE  TRUE
##   116  1214
```

```
mca <- cbind(mcasmall, mcamedium)
mca <- cbind(mca, mcalarge)
table(mca$use, mca$group)
```

```
##
##           gam  ook  spz
##   FALSE    7  218  116
##   TRUE    88  165 1214
```

```
#make QCed SingleCellExperiment
mca.qc.cells <- mca[ , colData(mca)$use]
meds <- tapply(colData(mca.qc.cells)$detected, colData(mca.qc.cells)$group, median)
meds
```

```
##    gam    ook    spz
## 1496.5  829.0  119.0
```

```
# Gene filtering
filter_genes <- apply(counts(mca[ , colData(mca)$use]), 1, function(x) length(x[x >= 1]) >= 2)

table(filter_genes)
```

```
## filter_genes
## FALSE  TRUE
##   730  5058
```

```
rowData(mca)$use <- filter_genes

dim(mca[rowData(mca)$use, colData(mca)$use])
```

```
## [1] 5058 1467
```

```
assay(mca, "logcounts_raw") <- log2(counts(mca) + 1)
reducedDim(mca) <- NULL


mca.qc <- mca[rowData(mca)$use, colData(mca)$use]
```

```
clusters <- quickCluster(mca.qc)
mca.qc <- computeSumFactors(mca.qc, clusters=clusters, min.mean=1)
```

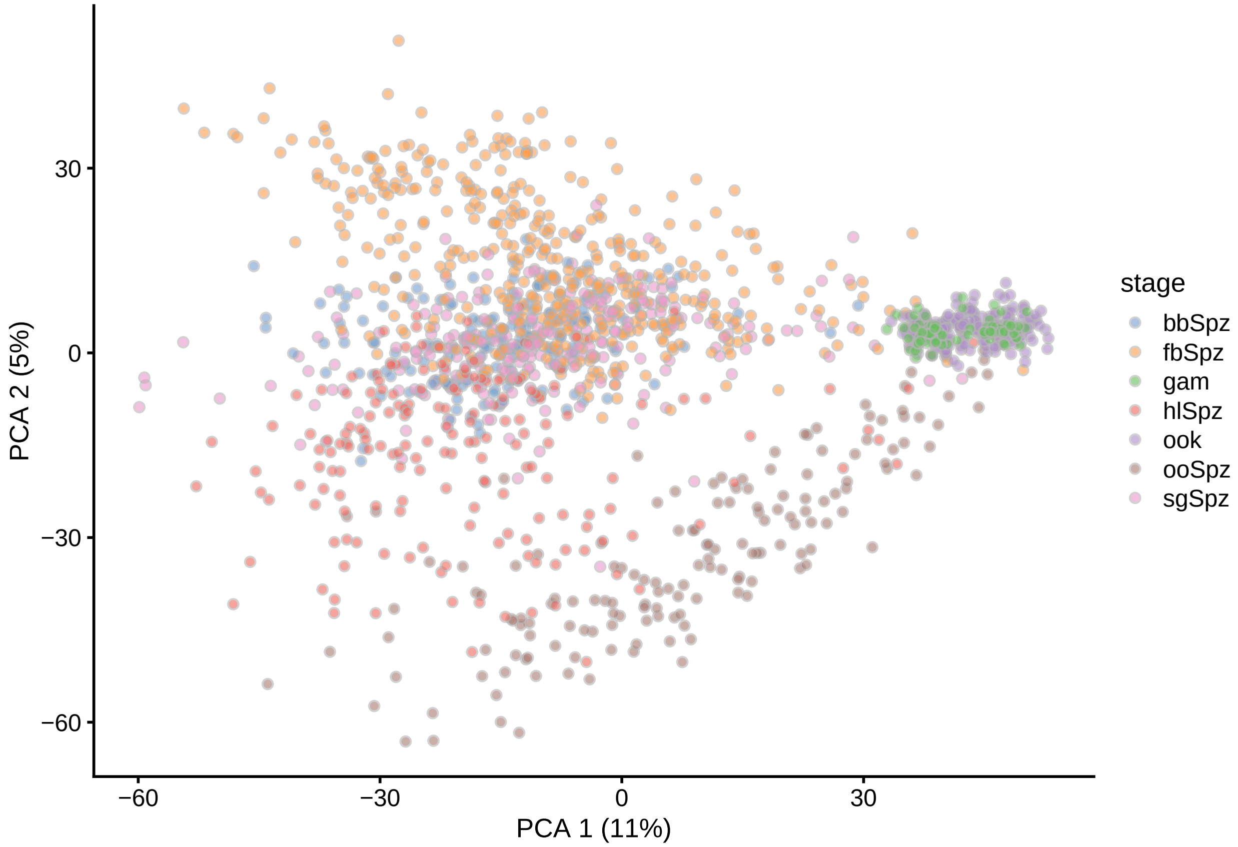
```
## Warning in FUN(...): encountered negative size factor estimates

## Warning in FUN(...): encountered negative size factor estimates

## Warning in FUN(...): encountered negative size factor estimates

## Warning in FUN(...): encountered negative size factor estimates
```

```
summary(sizeFactors(mca.qc))
```

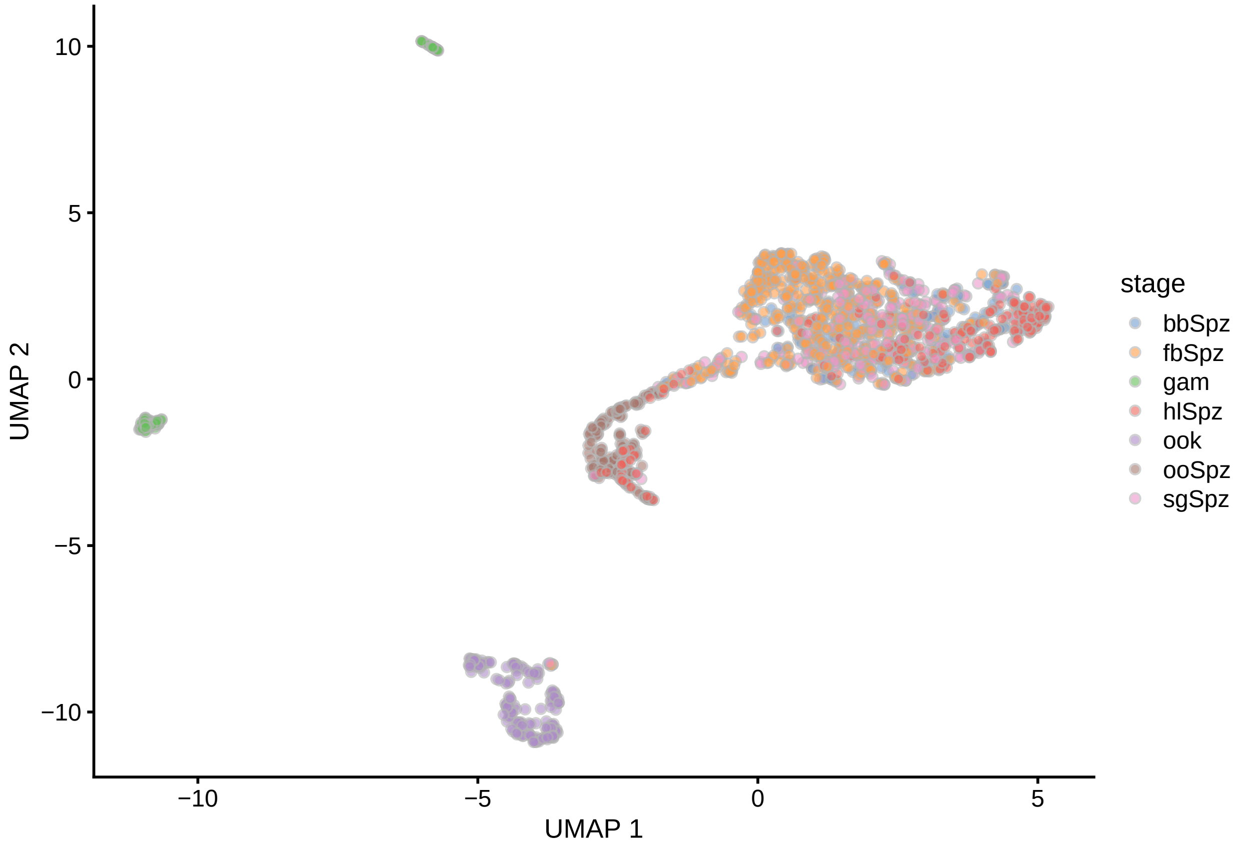
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00004  0.00363  0.00822  1.00000  0.02965 35.23496
```

```
mca.qc <- logNormCounts(mca.qc, log=FALSE, size_factors=sizeFactors(mca.qc))
mca.qc <- logNormCounts(mca.qc, log=TRUE, size_factors=sizeFactors(mca.qc))

cpm(mca.qc) <- calculateCPM(mca.qc) #divide each column by its total and multiple by 1 million
assay(mca.qc, "log_cpm") <- log2(calculateCPM(mca.qc) + 1)

mca.qc <- runPCA(mca.qc,  ntop=150)
set.seed(112)
mca.qc <- runUMAP(mca.qc, ntop=150, n_neighbors = 5)
set.seed(666)
mca.qc <- runTSNE(mca.qc,  ntop = 150)
```

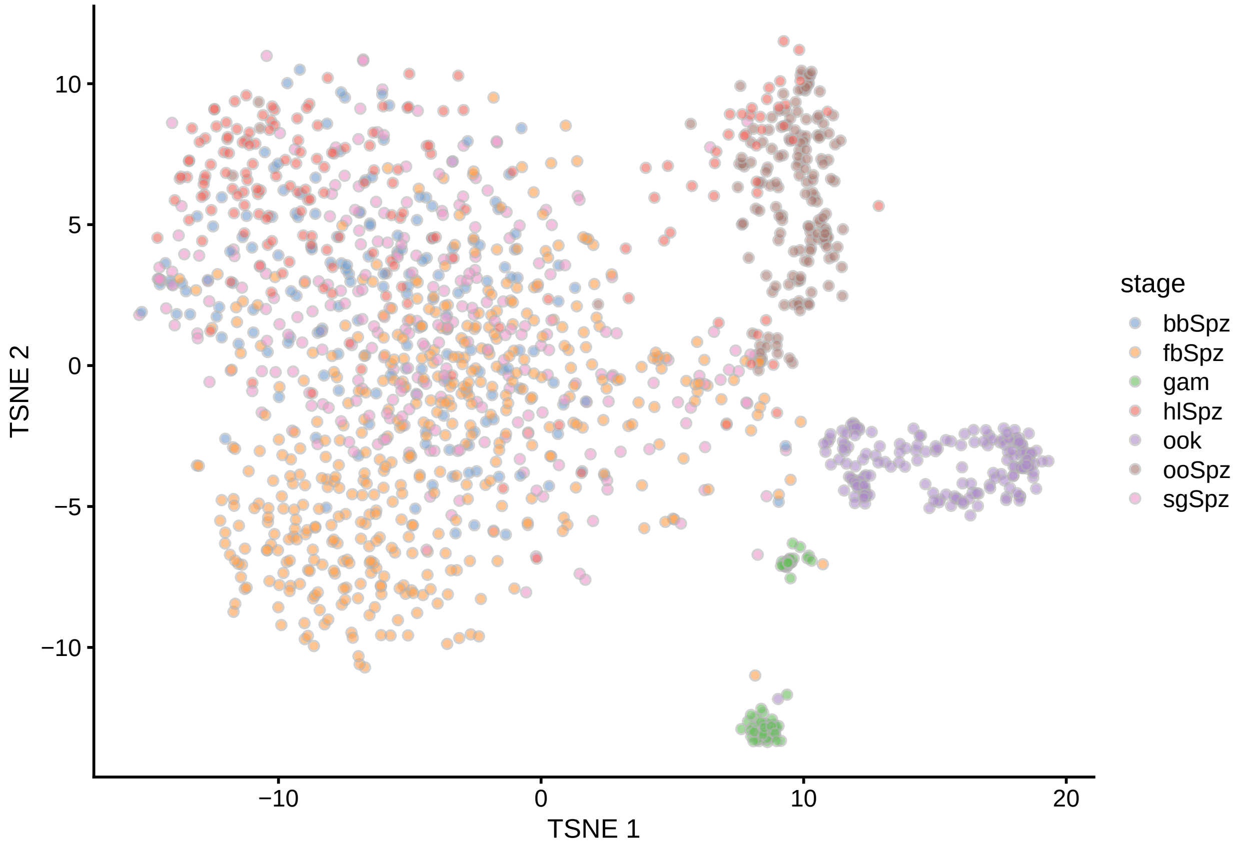
```r
plotPCA(mca.qc, colour_by = "stage")
```



```r
plotUMAP(mca.qc, colour_by = "stage")
```



```
plotTSNE(mca.qc, colour_by = "stage")
```



```
mca.qc
```

```
## class: SingleCellExperiment
## dim: 5058 1467
## metadata(0):
## assays(6): counts logcounts ... cpm log_cpm
## rownames(5058): PF3D7_0100100.1 PF3D7_0100300.1 ... mal_rna_18:rRNA
##   mal_rna_19:rRNA
## rowData names(3): mean detected use
## colnames(1467): X32706_8_5_sorted.bam X32706_8_6_sorted.bam ...
##   X31032_1_94_sorted.bam X31032_1_98_sorted.bam
## colData names(29): npgnum tag ... use sizeFactor
## reducedDimNames(3): PCA UMAP TSNE
## altExpNames(0):
```

```
assays(mca.qc)
```

```
## List of length 6
## names(6): counts logcounts logcounts_raw normcounts cpm log_cpm
```

```
#saveRDS(mca.qc, file = "pf.mca.qc_20200525.rds")
```

```
session_info()
```

```
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.0.0 (2020-04-24)
##  os       macOS Mojave 10.14.6
##  system   x86_64, darwin17.0
##  ui       X11
##  language (EN)
##  collate  en_GB.UTF-8
##  ctype    en_GB.UTF-8
##  tz       Europe/London
##  date     2020-10-13
##
## - Packages ------------------------------------------------------------------
##  package              * version  date       lib source
##  assertthat             0.2.1    2019-03-21 [1] CRAN (R 4.0.0)
##  backports              1.1.7    2020-05-13 [1] CRAN (R 4.0.0)
##  beeswarm               0.2.3    2016-04-25 [1] CRAN (R 4.0.0)
##  Biobase              * 2.48.0   2020-04-27 [1] Bioconductor
##  BiocGenerics         * 0.34.0   2020-04-27 [1] Bioconductor
##  BiocNeighbors          1.6.0    2020-04-27 [1] Bioconductor
##  BiocParallel           1.22.0   2020-04-27 [1] Bioconductor
##  BiocSingular           1.4.0    2020-04-27 [1] Bioconductor
##  bitops                 1.0-6    2013-08-17 [1] CRAN (R 4.0.0)
##  callr                  3.4.3    2020-03-28 [1] CRAN (R 4.0.0)
##  cli                    2.0.2    2020-02-28 [1] CRAN (R 4.0.0)
##  colorspace             1.4-1    2019-03-18 [1] CRAN (R 4.0.0)
##  cowplot                1.0.0    2019-07-11 [1] CRAN (R 4.0.0)
##  crayon                 1.3.4    2017-09-16 [1] CRAN (R 4.0.0)
##  data.table             1.12.8   2019-12-09 [1] CRAN (R 4.0.0)
##  DelayedArray         * 0.14.0   2020-04-27 [1] Bioconductor
##  DelayedMatrixStats     1.10.0   2020-04-27 [1] Bioconductor
##  desc                   1.2.0    2018-05-01 [1] CRAN (R 4.0.0)
##  devtools             * 2.3.0    2020-04-10 [1] CRAN (R 4.0.0)
##  digest                 0.6.25   2020-02-23 [1] CRAN (R 4.0.0)
##  dplyr                  0.8.5    2020-03-07 [1] CRAN (R 4.0.0)
##  dqrng                  0.2.1    2019-05-17 [1] CRAN (R 4.0.0)
##  edgeR                  3.30.0   2020-04-27 [1] Bioconductor
##  ellipsis               0.3.1    2020-05-15 [1] CRAN (R 4.0.0)
##  evaluate               0.14     2019-05-28 [1] CRAN (R 4.0.0)
##  fansi                  0.4.1    2020-01-08 [1] CRAN (R 4.0.0)
##  farver                 2.0.3    2020-01-16 [1] CRAN (R 4.0.0)
##  FNN                    1.1.3    2019-02-15 [1] CRAN (R 4.0.0)
##  fs                     1.4.1    2020-04-04 [1] CRAN (R 4.0.0)
##  GenomeInfoDb         * 1.24.0   2020-04-27 [1] Bioconductor
##  GenomeInfoDbData       1.2.3    2020-05-09 [1] Bioconductor
##  GenomicRanges        * 1.40.0   2020-04-27 [1] Bioconductor
##  ggbeeswarm             0.6.0    2017-08-07 [1] CRAN (R 4.0.0)
##  ggplot2              * 3.3.0    2020-03-05 [1] CRAN (R 4.0.0)
##  glue                   1.4.1    2020-05-13 [1] CRAN (R 4.0.0)
##  gridExtra              2.3      2017-09-09 [1] CRAN (R 4.0.0)
##  gtable                 0.3.0    2019-03-25 [1] CRAN (R 4.0.0)
##  htmltools              0.4.0    2019-10-04 [1] CRAN (R 4.0.0)
##  htmlwidgets            1.5.1    2019-10-08 [1] CRAN (R 4.0.0)
##  httr                   1.4.1    2019-08-05 [1] CRAN (R 4.0.0)
##  igraph                 1.2.5    2020-03-19 [1] CRAN (R 4.0.0)
##  IRanges              * 2.22.1   2020-04-28 [1] Bioconductor
##  irlba                  2.3.3    2019-02-05 [1] CRAN (R 4.0.0)
##  jsonlite               1.6.1    2020-02-02 [1] CRAN (R 4.0.0)
##  knitr                  1.28     2020-02-06 [1] CRAN (R 4.0.0)
##  labeling               0.3      2014-08-23 [1] CRAN (R 4.0.0)
##  lattice                0.20-41  2020-04-02 [2] CRAN (R 4.0.0)
##  lazyeval               0.2.2    2019-03-15 [1] CRAN (R 4.0.0)
##  lifecycle              0.2.0    2020-03-06 [1] CRAN (R 4.0.0)
##  limma                  3.44.1   2020-04-28 [1] Bioconductor
##  locfit                 1.5-9.4  2020-03-25 [1] CRAN (R 4.0.0)
##  magrittr               1.5      2014-11-22 [1] CRAN (R 4.0.0)
##  Matrix               * 1.2-18   2019-11-27 [2] CRAN (R 4.0.0)
##  matrixStats          * 0.56.0   2020-03-13 [1] CRAN (R 4.0.0)
##  memoise                1.1.0    2017-04-21 [1] CRAN (R 4.0.0)
##  munsell                0.5.0    2018-06-12 [1] CRAN (R 4.0.0)
##  pillar                 1.4.4    2020-05-05 [1] CRAN (R 4.0.0)
##  pkgbuild               1.0.8    2020-05-07 [1] CRAN (R 4.0.0)
##  pkgconfig              2.0.3    2019-09-22 [1] CRAN (R 4.0.0)
##  pkgload                1.0.2    2018-10-29 [1] CRAN (R 4.0.0)
##  plotly               * 4.9.2.1  2020-04-04 [1] CRAN (R 4.0.0)
##  prettyunits            1.1.1    2020-01-24 [1] CRAN (R 4.0.0)
##  processx               3.4.2    2020-02-09 [1] CRAN (R 4.0.0)
##  ps                     1.3.3    2020-05-08 [1] CRAN (R 4.0.0)
##  purrr                  0.3.4    2020-04-17 [1] CRAN (R 4.0.0)
##  R6                     2.4.1    2019-11-12 [1] CRAN (R 4.0.0)
##  Rcpp                   1.0.4.6  2020-04-09 [1] CRAN (R 4.0.0)
##  RCurl                  1.98-1.2 2020-04-18 [1] CRAN (R 4.0.0)
##  remotes                2.1.1    2020-02-15 [1] CRAN (R 4.0.0)
##  rlang                  0.4.6    2020-05-02 [1] CRAN (R 4.0.0)
##  rmarkdown              2.1      2020-01-20 [1] CRAN (R 4.0.0)
##  rprojroot              1.3-2    2018-01-03 [1] CRAN (R 4.0.0)
##  RSpectra               0.16-0   2019-12-01 [1] CRAN (R 4.0.0)
##  rsvd                   1.0.3    2020-02-17 [1] CRAN (R 4.0.0)
##  Rtsne                  0.15     2018-11-10 [1] CRAN (R 4.0.0)
##  S4Vectors            * 0.26.1   2020-05-16 [1] Bioconductor
##  scales                 1.1.1    2020-05-11 [1] CRAN (R 4.0.0)
##  scater               * 1.16.0   2020-04-27 [1] Bioconductor
##  scran                * 1.16.0   2020-04-27 [1] Bioconductor
##  sessioninfo            1.1.1    2018-11-05 [1] CRAN (R 4.0.0)
##  SingleCellExperiment * 1.10.1   2020-04-28 [1] Bioconductor
##  statmod                1.4.34   2020-02-17 [1] CRAN (R 4.0.0)
##  stringi                1.4.6    2020-02-17 [1] CRAN (R 4.0.0)
##  stringr                1.4.0    2019-02-10 [1] CRAN (R 4.0.0)
##  SummarizedExperiment * 1.18.1   2020-04-30 [1] Bioconductor
##  testthat               2.3.2    2020-03-02 [1] CRAN (R 4.0.0)
##  tibble                 3.0.1    2020-04-20 [1] CRAN (R 4.0.0)
##  tidyr                  1.1.0    2020-05-20 [1] CRAN (R 4.0.0)
##  tidyselect             1.1.0    2020-05-11 [1] CRAN (R 4.0.0)
##  usethis              * 1.6.1    2020-04-29 [1] CRAN (R 4.0.0)
##  uwot                   0.1.8    2020-03-16 [1] CRAN (R 4.0.0)
##  vctrs                  0.3.0    2020-05-11 [1] CRAN (R 4.0.0)
##  vipor                  0.4.5    2017-03-22 [1] CRAN (R 4.0.0)
##  viridis                0.5.1    2018-03-29 [1] CRAN (R 4.0.0)
##  viridisLite            0.3.0    2018-02-01 [1] CRAN (R 4.0.0)
##  withr                  2.2.0    2020-04-20 [1] CRAN (R 4.0.0)
##  xfun                   0.14     2020-05-20 [1] CRAN (R 4.0.0)
##  XVector                0.28.0   2020-04-27 [1] Bioconductor
##  yaml                   2.2.1    2020-02-01 [1] CRAN (R 4.0.0)
##  zlibbioc               1.34.0   2020-04-27 [1] Bioconductor
##
## [1] /Users/vh3/Library/R/4.0/library
## [2] /Library/Frameworks/R.framework/Versions/4.0/Resources/library
```