Integration of Pb and Pf data

Virginia Howick

02/07/2020

```
setwd("/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf")
library(scater, quietly = TRUE)
library(scmap)
library(Seurat)
library(scater)
library(scran)
library(cowplot)
library(gridExtra)
library(viridis)
```

```
sce <- readRDS("/Users/vh3/Documents/PfMCA/ANALYSIS_2/scmap/pborthindex_matchedstages_20200629.rds")

pfmca_orth <- readRDS("/Users/vh3/Documents/PfMCA/ANALYSIS_2/scmap/pfmca_orth_20200629.rds")

newot <- read.csv("/Users/vh3/Documents/MCA/ANALYSIS_3/ortho4.csv")

on <- newot$orth_name
newot$orth_seurat <- gsub("_", "-", on)


pbsce <- sce[rownames(sce) %in% rownames(pfmca_orth), ]
table(rownames(pfmca_orth) == rownames(pbsce))
```

```
##
## FALSE
##  4008
```

```
pbsce <- pbsce[match(rownames(pfmca_orth), rownames(pbsce)), ]
table(rownames(pfmca_orth) == rownames(pbsce))
```

```
##
## TRUE
## 4008
```

```
pbcounts <- as.data.frame(counts(pbsce))
pfcounts <- as.data.frame(counts(pfmca_orth))

allcounts <- cbind(pbcounts, pfcounts)

pbcd <- as.data.frame(colData(pbsce))
pbcd <- pbcd[c("sample_id", "Species", "ShortenedLifeStage2", "seqrunnum", "time")]
colnames(pbcd) <- c("sample_id", "species", "stage", "run", "day")
pbcd$xfilename <- pbcd$sample_id
pbcd$topcell <- pbcd$sample_id
```

```
pbcd$topcell_sls2 <- pbcd$stage
pbcd$topsim <- rep(1, length(pbcd$sample_id))


pfcd <- as.data.frame(colData(pfmca_orth))
pfcd <- pfcd[c("sample_id", "stage", "day", "run", "xfilename", "topcell", "topcell_sls2",
    "topsim")]
pfcd$species <- rep("Pfa", length(pfcd$sample_id))

pbcd <- pbcd[, match(colnames(pfcd), colnames(pbcd))]

allcd <- rbind(pbcd, pfcd)

table(rownames(allcd) == colnames(allcounts))
```

```
##
## TRUE
## 1743
```

```
mca <- SingleCellExperiment(assays = list(counts = as.matrix(allcounts)), colData = allcd)

# saveRDS(mca, '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/allpbpf_20200630.rds')
```

```
pfint <- SingleCellExperiment(assays = list(counts = as.matrix(pfcounts)), colData = pfcd)

set.seed(222)
clusters <- quickCluster(pfint)
pfint <- computeSumFactors(pfint, clusters = clusters, min.mean = 10)
pfint <- logNormCounts(pfint, log = FALSE, size_factors = sizeFactors(pfint))
pfint <- logNormCounts(pfint, log = TRUE, size_factors = sizeFactors(pfint))


# saveRDS(pfint,
# file='/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/pf_forint_20200630.rds')

pfint.seurat <- as.Seurat(pfint, counts = "counts", data = "logcounts")
```

```
## Warning: Feature names cannot have underscores ('_'), replacing with dashes
## ('-')

## Warning: Feature names cannot have underscores ('_'), replacing with dashes
## ('-')
```

```
Idents(pfint.seurat) <- "stage"
pfint.seurat <- FindVariableFeatures(pfint.seurat, selection.method = "vst", nfeatures = 500)


pbint <- SingleCellExperiment(assays = list(counts = as.matrix(pbcounts)), colData = pbcd)


pbint$species <- rep("Pbe", length(pbint$sample_id))

set.seed(222)
clusters <- quickCluster(pbint)
```

```
## Warning in (function (to_check, X, clust_centers, clust_info, dtype, nn, :
## detected tied distances to neighbors, see ?'BiocNeighbors-ties'
```

```
pbint <- computeSumFactors(pbint, clusters = clusters, min.mean = 10)
pbint <- logNormCounts(pbint, log = FALSE, size_factors = sizeFactors(pbint))
pbint <- logNormCounts(pbint, log = TRUE, size_factors = sizeFactors(pbint))

pbint.seurat <- as.Seurat(pbint, counts = "counts", data = "logcounts")
```

```
## Warning: Feature names cannot have underscores ('_'), replacing with dashes
## ('-')

## Warning: Feature names cannot have underscores ('_'), replacing with dashes
## ('-')
```

```
Idents(pbint.seurat) <- "stage"
pbint.seurat <- FindVariableFeatures(pbint.seurat, selection.method = "vst", nfeatures = 500)

# saveRDS(pbint,
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/pb_forint_2020630.rds')
```

Perform integration

We then identify anchors using the FindIntegrationAnchors function, which takes a list of Seurat objects as input, and use these anchors to integrate the two datasets together with IntegrateData.

```
p.anchors <- FindIntegrationAnchors(object.list = list(pfint.seurat, pbint.seurat),
    dims = 1:20)
```

```
## Computing 2000 integration features

## Scaling features for provided objects

## Finding all pairwise anchors

## Running CCA

## Merging objects

## Finding neighborhoods

## Finding anchors

##  Found 2152 anchors

## Filtering anchors

##  Retained 1757 anchors

## Extracting within-dataset neighbors
```

```
p.combined <- IntegrateData(anchorset = p.anchors, dims = 1:20)
```

```
## Merging dataset 2 into 1

## Extracting anchors for merged samples

## Finding integration vectors

## Finding integration vector weights

## Integrating data
```

```
## Warning: Adding a command log without an assay associated with it
```

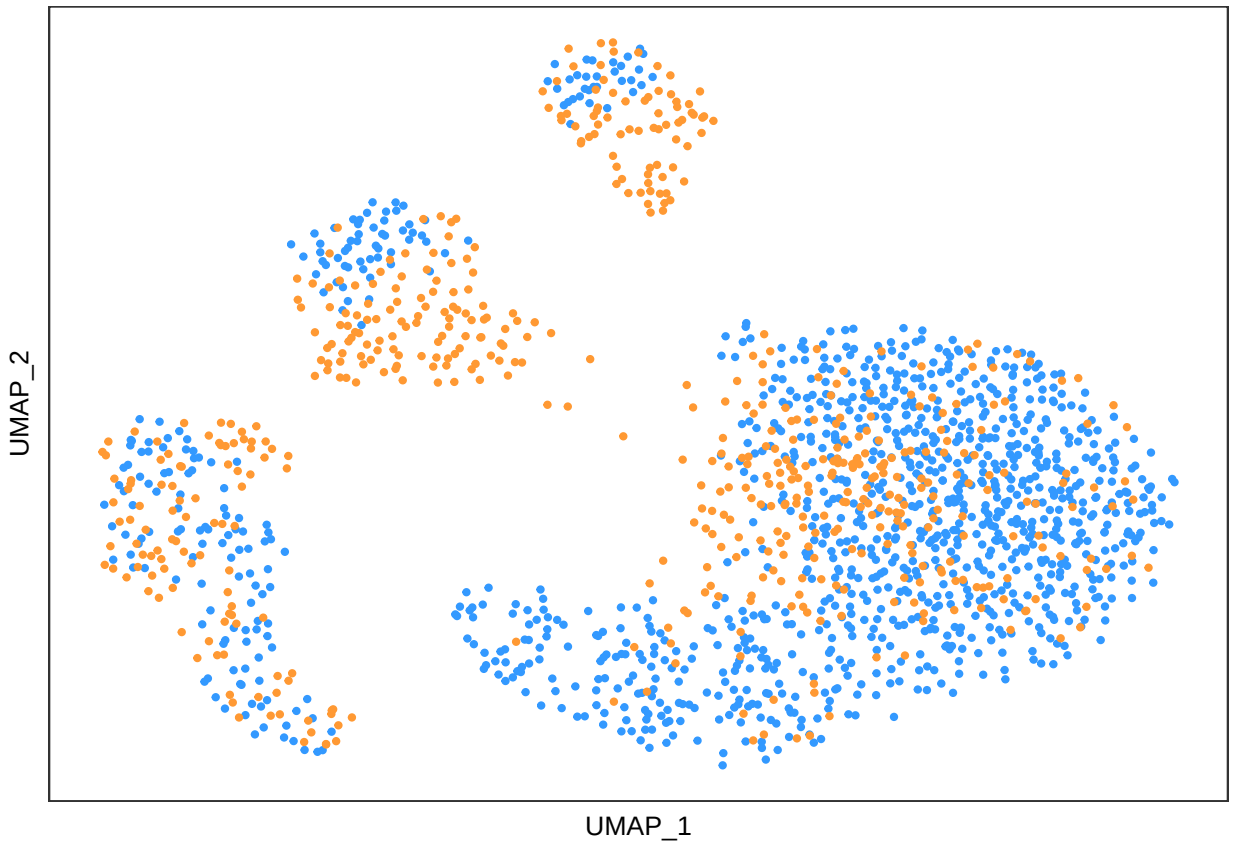
Perform an integrated analysis

Now we can run a single integrated analysis on all cells!

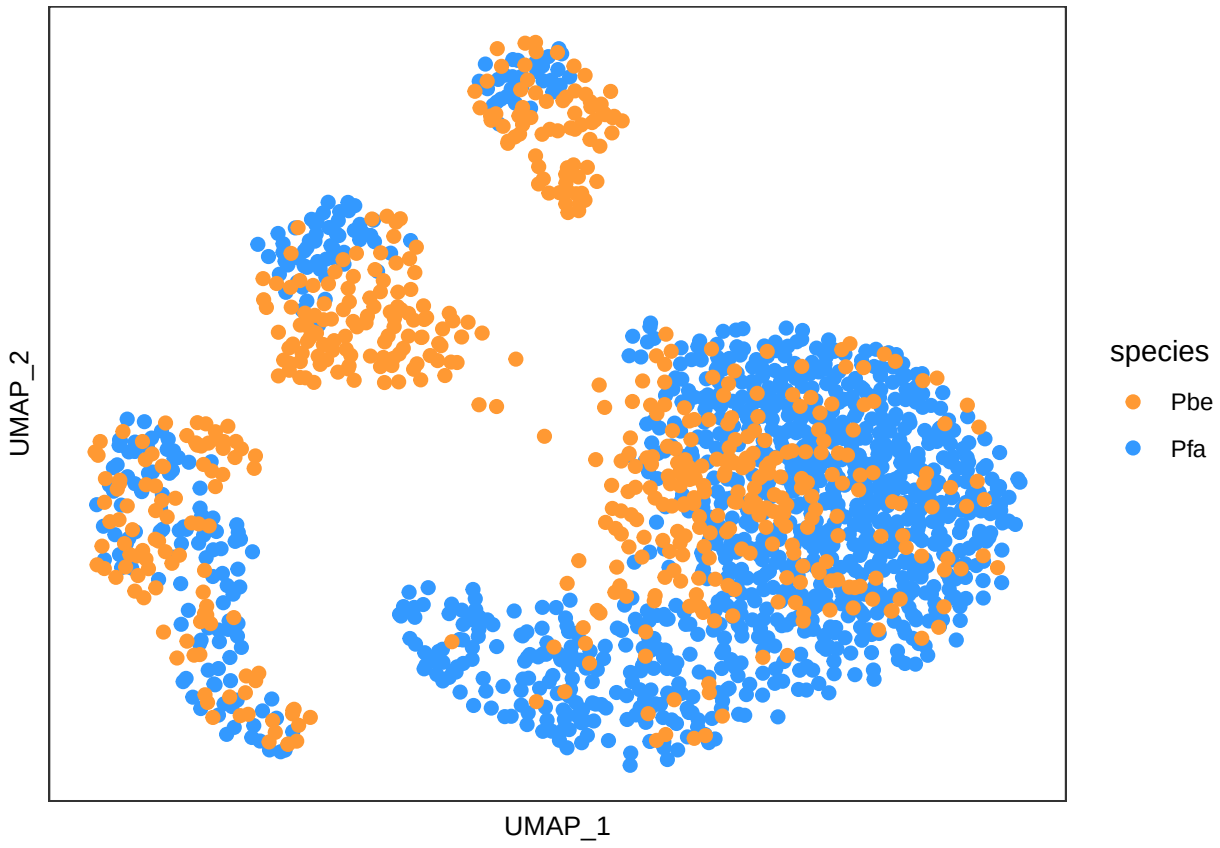
```
DefaultAssay(p.combined) <- "integrated"

# Run the standard workflow for visualization and clustering
p.combined <- ScaleData(p.combined, verbose = FALSE)
p.combined <- RunPCA(p.combined, npcs = 30, verbose = FALSE)
# t-SNE and Clustering p.combined <- FindVariableFeatures(p.combined,
# selection.method = 'vst', nfeatures = 500) hvg <- HVFInfo(object = p.combined)
# hvgfeat <- rownames(hvg)
p.combined <- RunUMAP(p.combined, reduction = "pca", dims = 1:20, umap.method = "uwot",
    n.neighbors = 5, min.dist = 2, spread = 3, seed.use = 222)
```

```
## Warning: The default method for RunUMAP has changed from calling Python UMAP via reticulate to the R
## To use Python UMAP via reticulate, set umap.method to 'umap-learn' and metric to 'correlation'
## This message will be shown once per session
## 20:38:54 UMAP embedding parameters a = 0.01251 b = 1.524
## 20:38:54 Read 1743 rows and found 20 numeric columns
## 20:38:54 Using Annoy for neighbor search, n_neighbors = 5
## 20:38:54 Building Annoy index with metric = cosine, n_trees = 50
## 0%   10   20   30   40   50   60   70   80   90   100%
## [----|----|----|----|----|----|----|----|----|----|
## **************************************************|
## 20:38:54 Writing NN index file to temp file /var/folders/jg/ylpkqzys381fzggn01b_krqw000g9y/T//RtmpB1
## 20:38:54 Searching Annoy index using 1 thread, search_k = 500
## 20:38:54 Annoy recall = 100%
## 20:38:55 Commencing smooth kNN distance calibration using 1 thread
## 20:38:56 Found 2 connected components, falling back to 'spca' initialization with init_sdev = 1
## 20:38:56 Initializing from PCA
## 20:38:56 PCA: 2 components explained 57.46% variance
## 20:38:56 Commencing optimization for 500 epochs, with 12266 positive edges
## 20:38:58 Optimization finished
```

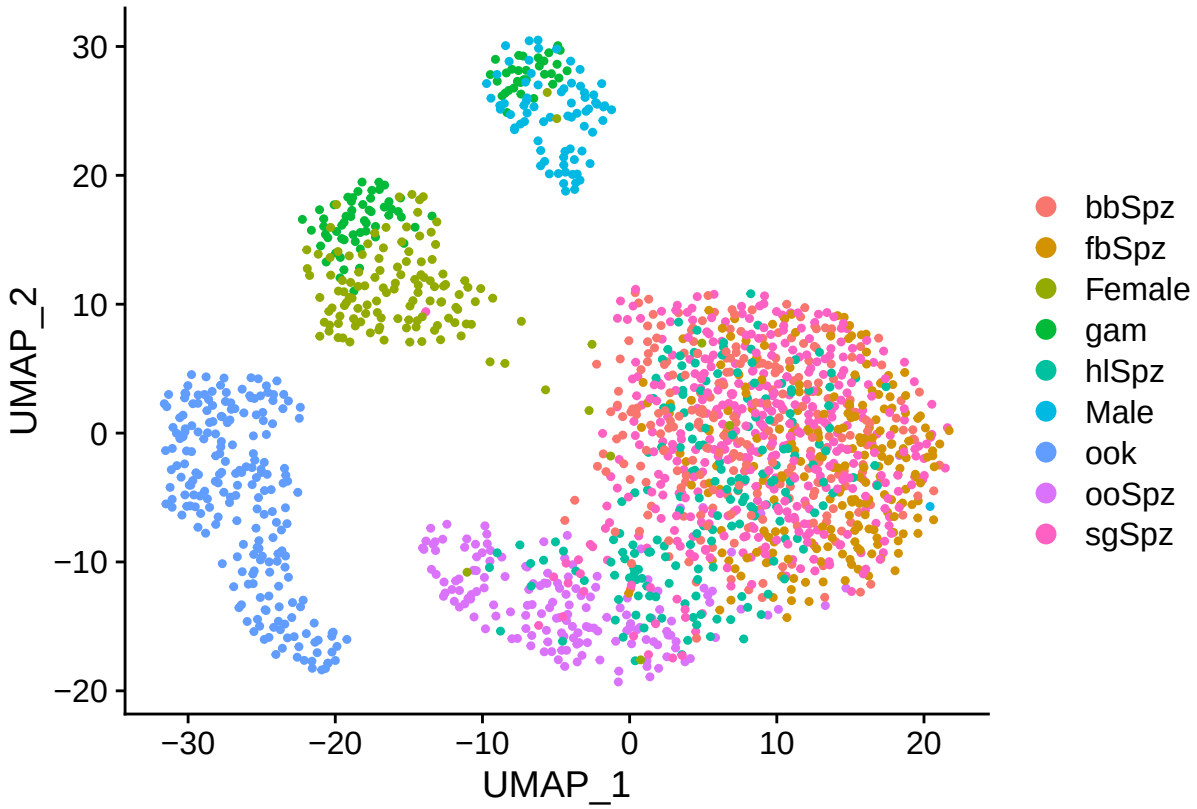
```
p <- DimPlot(p.combined, reduction = "umap", group.by = "species")
pdat <- p$data

colors = c(Pfa = "#3399FF", Pbe = "#FF9933")

ggplot(pdat, aes(UMAP_1, UMAP_2)) + geom_point(aes(colour = species), size = 0.8) +
    theme_bw() + scale_colour_manual(values = colors) + theme(axis.text = element_blank(),
    axis.ticks = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_blank(
    axis.title = element_text(size = 10), legend.position = "none")
```



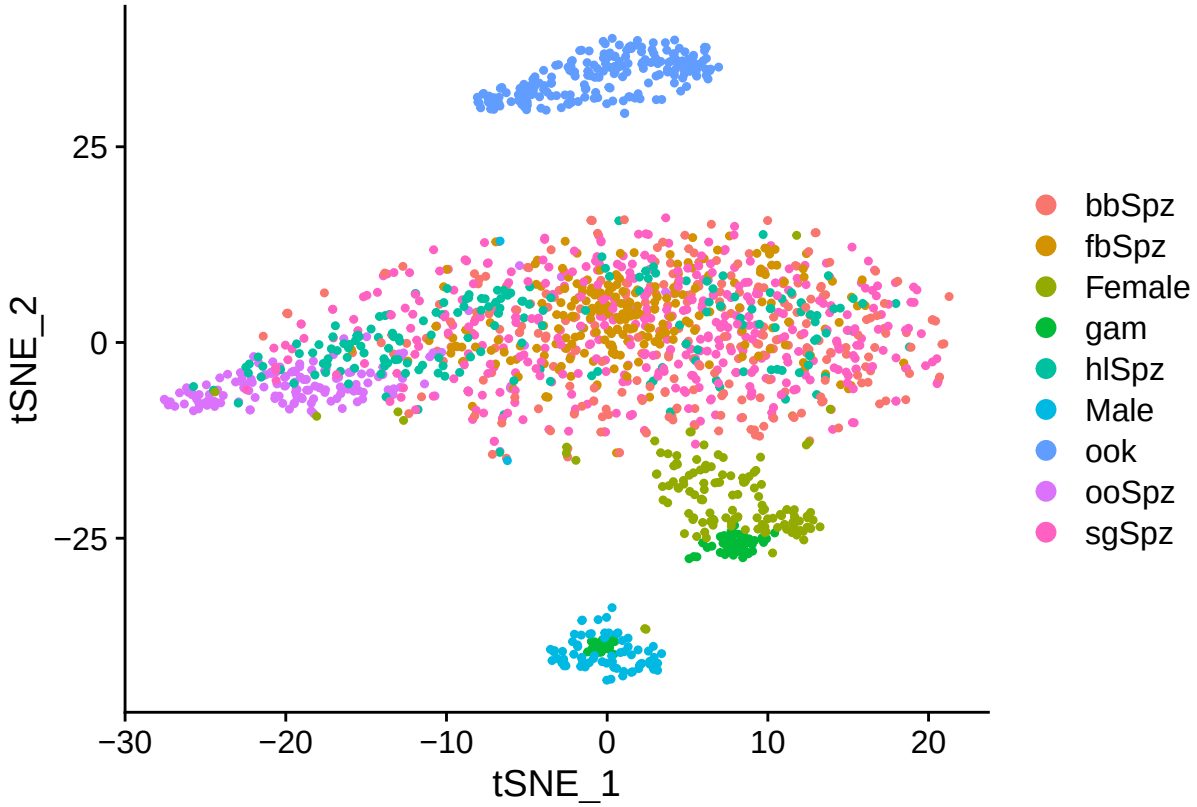

```
ggplot(pdat, aes(UMAP_1, UMAP_2)) + geom_point(aes(colour = species), size = 2) +
    theme_bw() + scale_colour_manual(values = colors) + theme(axis.text = element_blank(),
    axis.ticks = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    axis.title = element_text(size = 10))
```




```
DimPlot(p.combined, reduction = "umap", group.by = "stage")
```



```
p.combined <- RunTSNE(p.combined, reduction = "pca", dims = 1:20)
DimPlot(p.combined, reduction = "tsne", group.by = "stage")
```




```
p.combined <- FindNeighbors(p.combined, reduction = "pca", dims = 1:20)
```

```
## Computing nearest neighbor graph
## Computing SNN
```

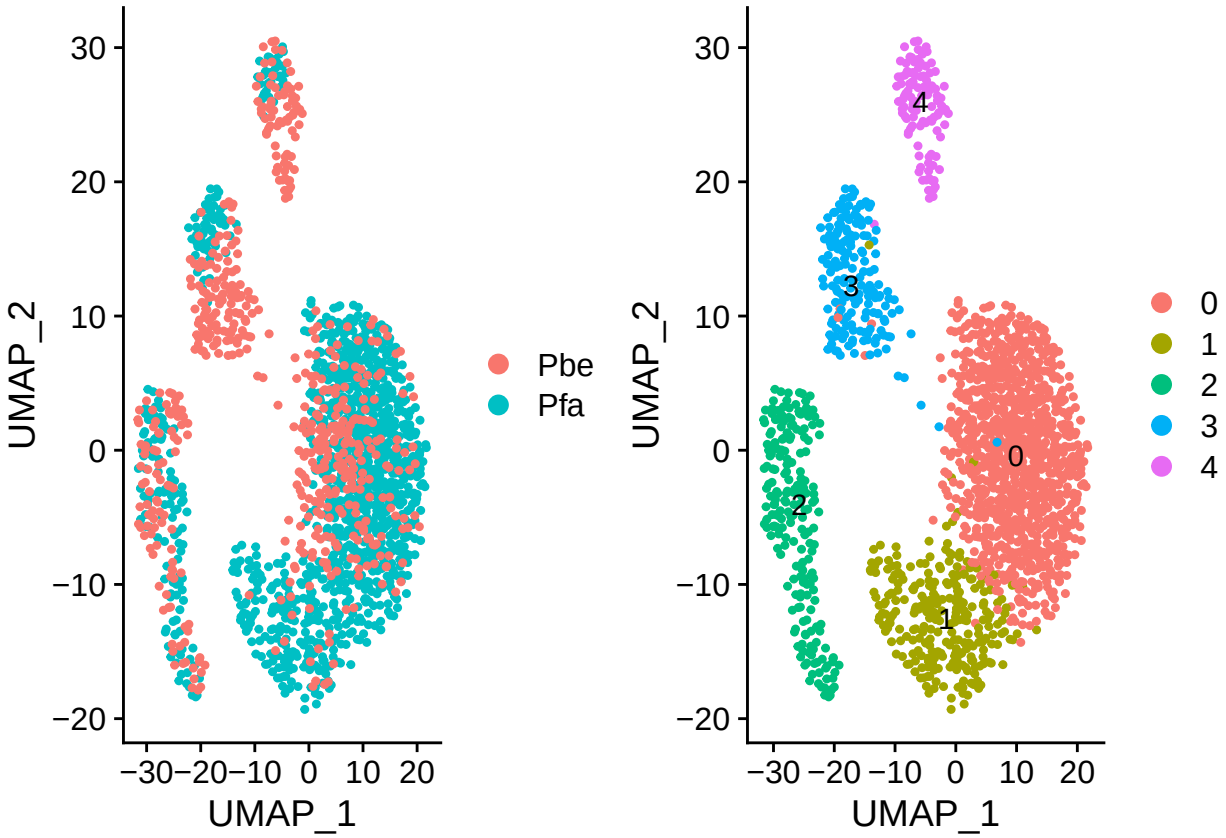
```
p.combined <- FindClusters(p.combined, resolution = 0.5)
```

```
## Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck
##
## Number of nodes: 1743
## Number of edges: 80219
##
## Running Louvain algorithm...
## Maximum modularity in 10 random starts: 0.8274
## Number of communities: 5
## Elapsed time: 0 seconds
```

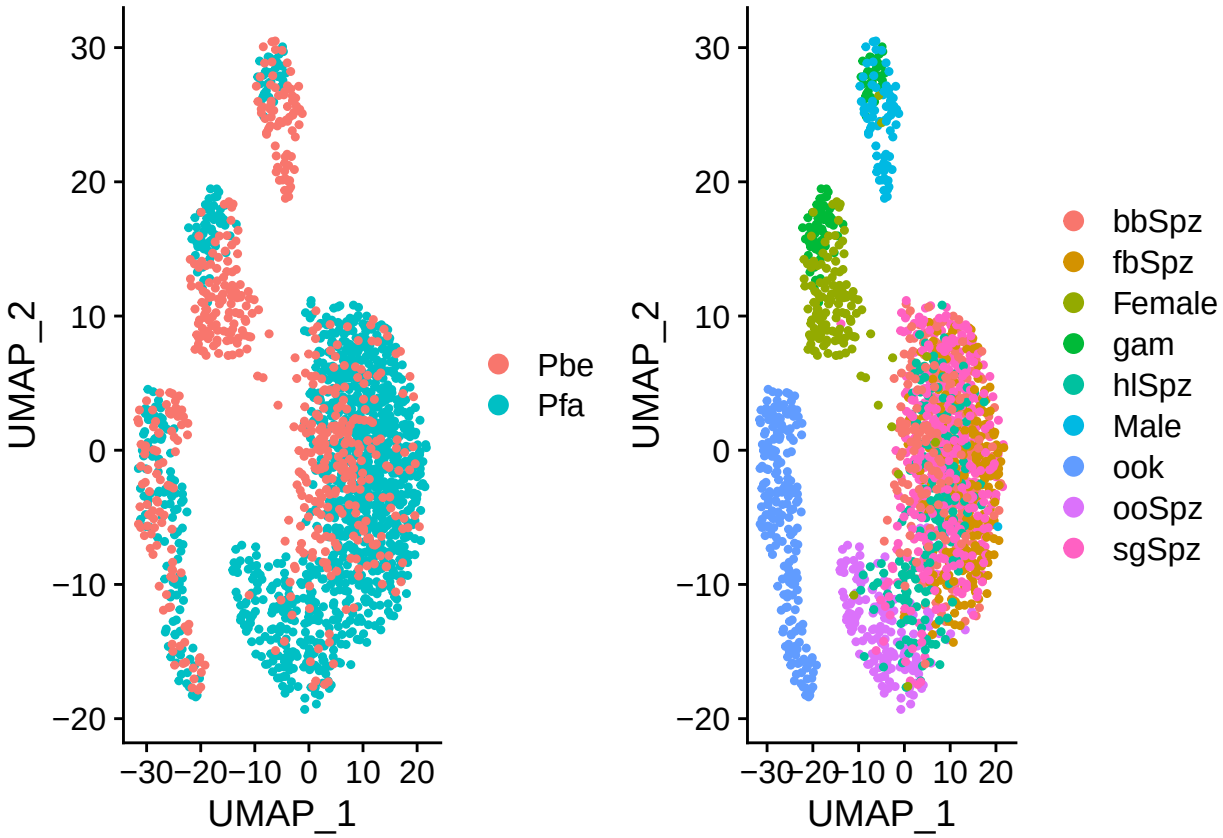
```
p1 <- DimPlot(p.combined, reduction = "umap", group.by = "species")
p2 <- DimPlot(p.combined, reduction = "umap", label = TRUE)
```

```
## Warning: Using `as.character()` on a quosure is deprecated as of rlang 0.3.0.
## Please use `as_label()` or `as_name()` instead.
## This warning is displayed once per session.
```

```
plot_grid(p1, p2)
```

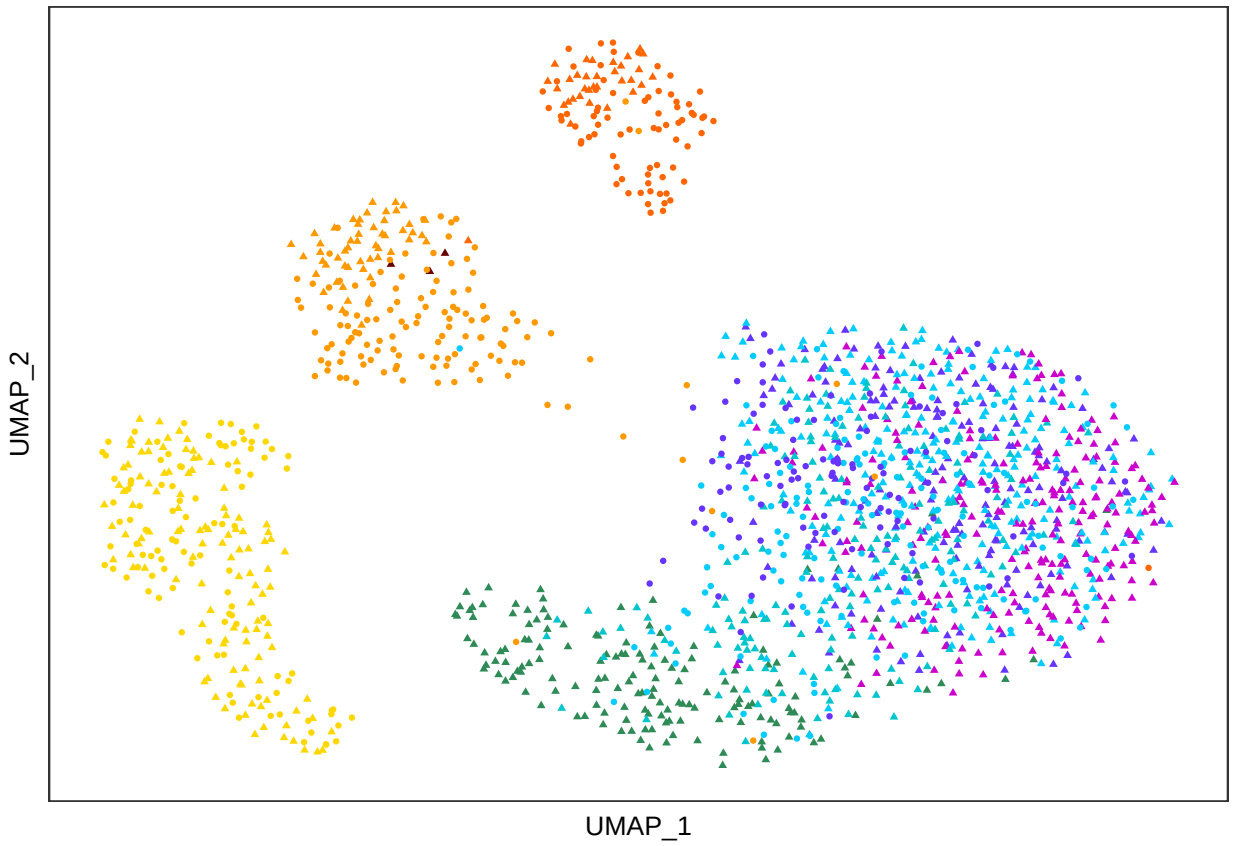



```
p1 <- DimPlot(p.combined, reduction = "umap", group.by = "species")
p2 <- DimPlot(p.combined, reduction = "umap", group.by = "stage")
plot_grid(p1, p2)
```

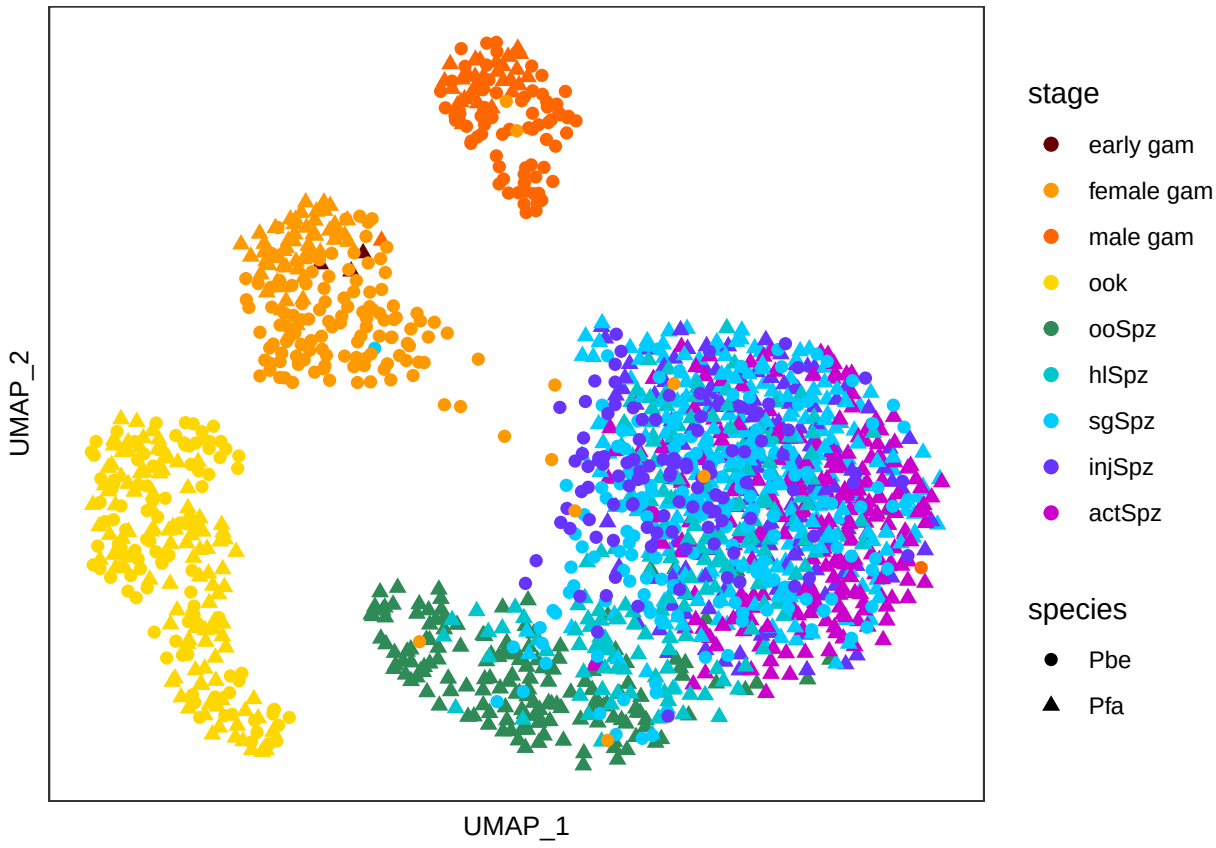



```
sex <- read.csv("/Users/vh3/Documents/PfMCA/ANALYSIS_2/20200522_PfGams_SC3_colData.csv",
    header = TRUE, row.names = 1)

md <- p.combined@meta.data
stage2 <- md["stage"]
stage2$xfilename <- rownames(stage2)

female <- rownames(subset(sex, sex == "female"))
male <- rownames(subset(sex, sex == "male"))
weird <- rownames(subset(sex, sex == "weird"))

stage2[which(stage2$xfilename %in% female), ]$stage <- "Female"
stage2[which(stage2$xfilename %in% male), ]$stage <- "Male"
stage2[which(stage2$xfilename %in% weird), ]$stage <- "early_gam"

p.combined$stage2 <- stage2$stage
p <- DimPlot(p.combined, reduction = "umap", group.by = "stage2")
pstage <- p$data
pstage$species <- pdat$species

colors = c(Male = "#FF6600", Female = "#FF9900", early_gam = "#660000", ook = "gold",
    ooSpz = "seagreen", hlSpz = "turquoise3", sgSpz = "#00CCFF", bbSpz = "#6633FF",
    fbSpz = "#CC00CC")

ggplot(pstage, aes(UMAP_1, UMAP_2)) + geom_point(aes(colour = stage2, shape = species),
    size = 0.8) + theme_bw() + scale_colour_manual(values = colors, breaks = c("early_gam",
    "Female", "Male", "ook", "ooSpz", "hlSpz", "sgSpz", "bbSpz", "fbSpz"), labels = c("early gam",
```

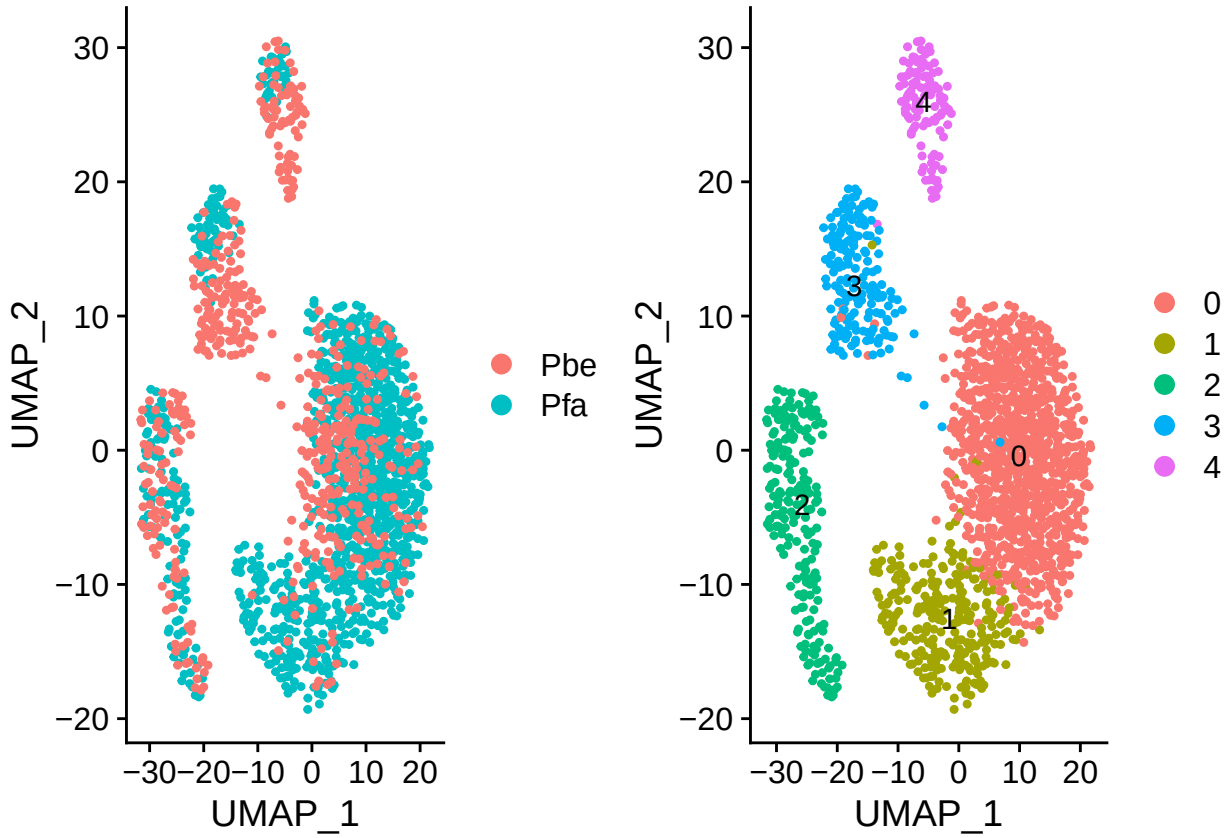
```
    "female gam", "male gam", "ook", "ooSpz", "hlSpz", "sgSpz", "injSpz", "actSpz")) +
    theme(axis.text = element_blank(), axis.ticks = element_blank(), panel.grid.major = element_blank()
        panel.grid.minor = element_blank(), axis.title = element_text(size = 10),
        legend.position = "none")
```



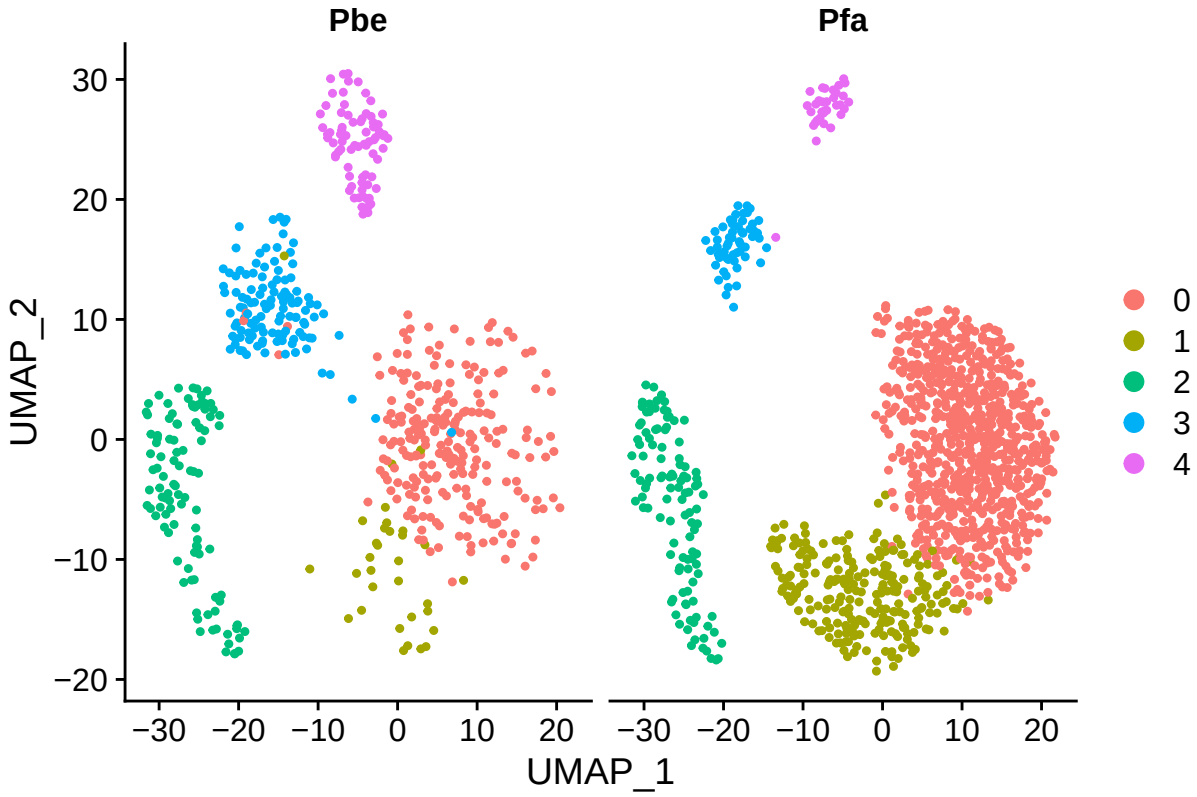

```
ggplot(pstage, aes(UMAP_1, UMAP_2)) + geom_point(aes(colour = stage2, shape = species),
    size = 2) + theme_bw() + scale_colour_manual(values = colors, breaks = c("early_gam",
    "Female", "Male", "ook", "ooSpz", "hlSpz", "sgSpz", "bbSpz", "fbSpz"), labels = c("early gam",
    "female gam", "male gam", "ook", "ooSpz", "hlSpz", "sgSpz", "injSpz", "actSpz")) +
    theme(axis.text = element_blank(), axis.ticks = element_blank(), panel.grid.major = element_blank()
        panel.grid.minor = element_blank(), axis.title = element_text(size = 10)) +
    labs(colour = "stage")
```



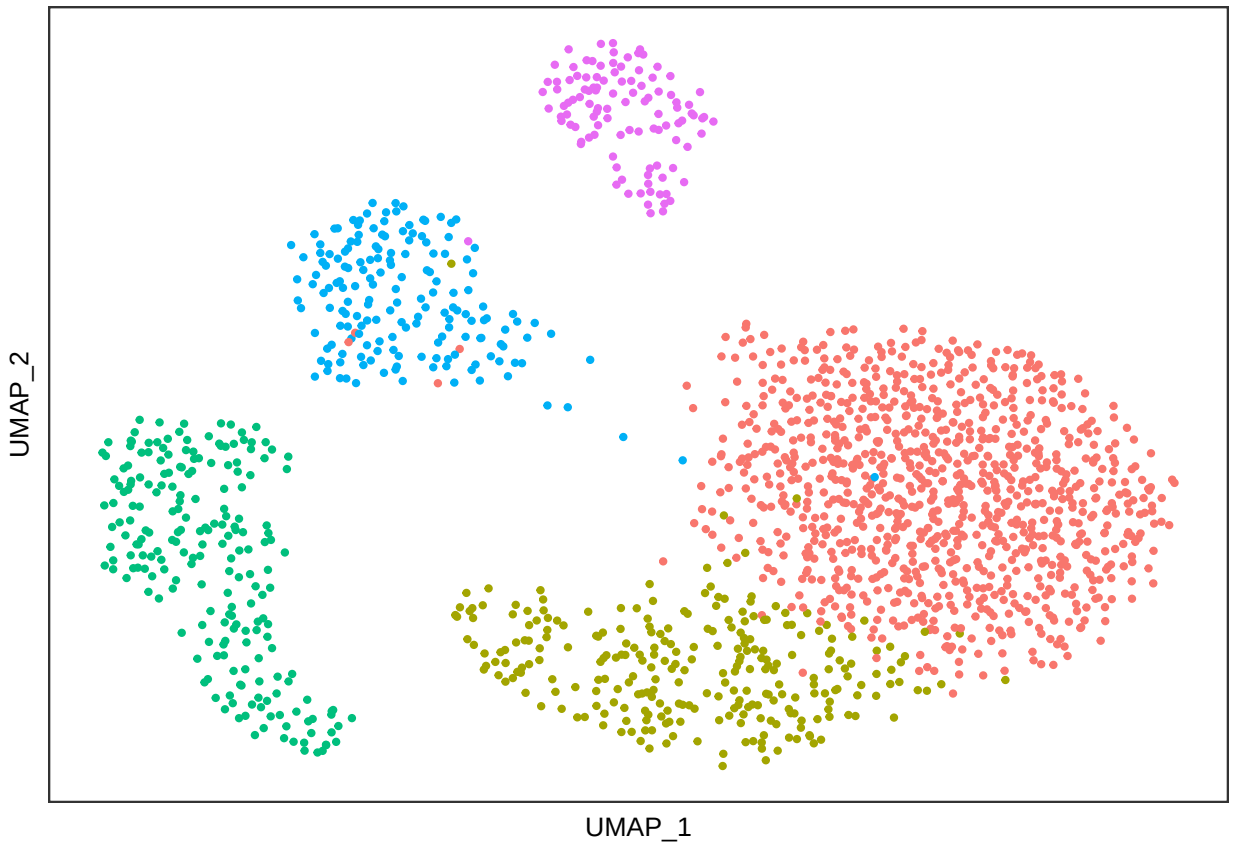

```
p1 <- DimPlot(p.combined, reduction = "umap", group.by = "species")
p2 <- DimPlot(p.combined, reduction = "umap", label = TRUE)
plot_grid(p1, p2)
```



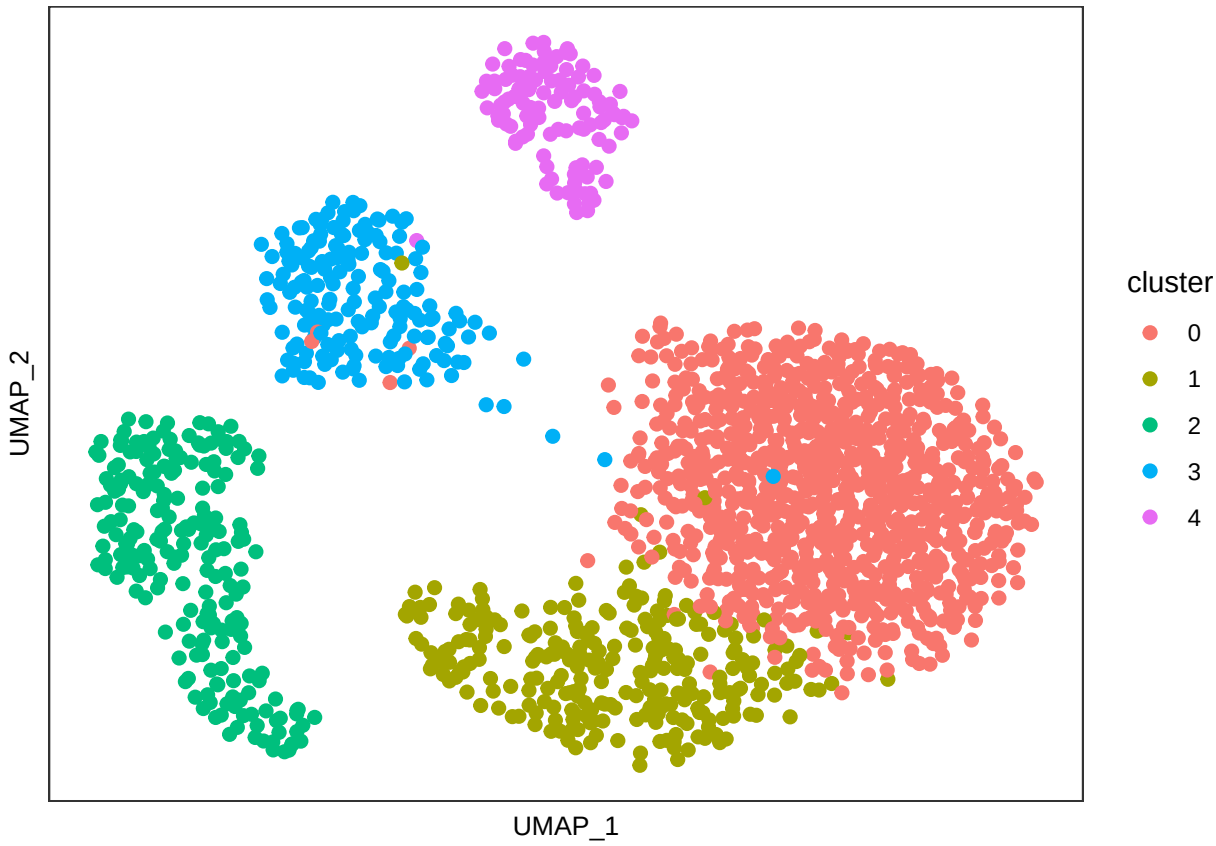

```
DimPlot(p.combined, reduction = "umap", split.by = "species")
```



```
p <- DimPlot(p.combined, reduction = "umap", label = TRUE, group.by = "seurat_clusters")

pclust <- p$data

ggplot(pclust, aes(UMAP_1, UMAP_2)) + geom_point(aes(colour = seurat_clusters), size = 0.8) +
    theme_bw() + theme(axis.text = element_blank(), axis.ticks = element_blank(),
    panel.grid.major = element_blank(), panel.grid.minor = element_blank(), axis.title = element_text(si
    legend.position = "none")
```




```
ggplot(pclust, aes(UMAP_1, UMAP_2)) + geom_point(aes(colour = seurat_clusters), size = 2) +
    theme_bw() + theme(axis.text = element_blank(), axis.ticks = element_blank(),
    panel.grid.major = element_blank(), panel.grid.minor = element_blank(), axis.title = element_text(si
    labs(colour = "cluster")
```




```
table(p.combined@meta.data$species, p.combined@meta.data$seurat_clusters)
```

```
##
##         0   1   2   3   4
##   Pbe 235  32 103 116  75
##   Pfa 747 235 112  55  33
```

Identify conserved cell type markers

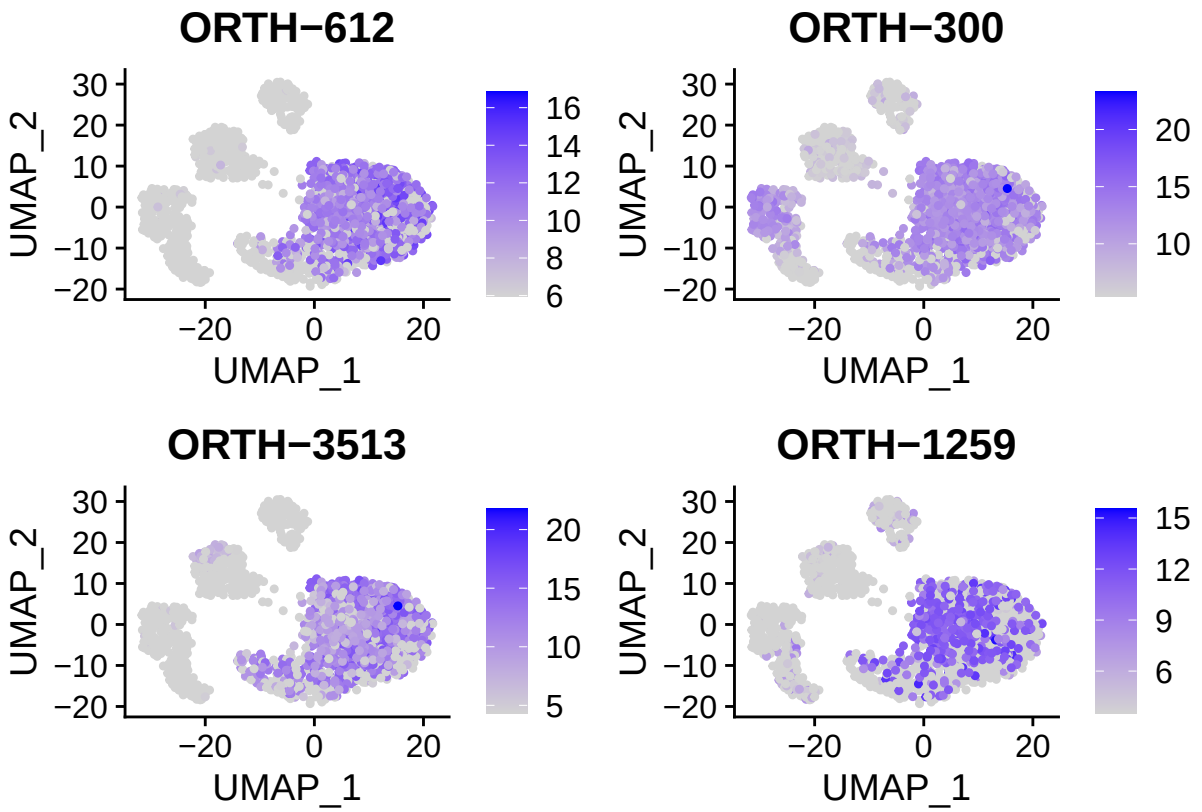
To identify canonical cell type marker genes that are conserved across conditions, we provide the FindConservedMarkers function. This function performs differential gene expression testing for each dataset/group and combines the p-values using meta-analysis methods from the MetaDE R package.

```
DefaultAssay(p.combined) <- "RNA"
Idents(p.combined) <- "seurat_clusters"

spz.markers <- FindConservedMarkers(p.combined, ident.1 = 0, grouping.var = "species",
    verbose = FALSE, logfc.threshold = 0.25, only.pos = TRUE)
head(spz.markers)
```

```
##              Pbe_p_val Pbe_avg_logFC Pbe_pct.1 Pbe_pct.2 Pbe_p_val_adj
## ORTH-612  2.001129e-69     2.3237047     0.923     0.248  8.020524e-66
## ORTH-300  3.145221e-33     0.6811077     0.979     0.758  1.260605e-29
## ORTH-3513 1.290491e-55     3.0568578     0.817     0.242  5.172286e-52
## ORTH-1259 1.928447e-79     3.0029170     0.949     0.227  7.729214e-76
## ORTH-3075 2.278279e-74     2.6825296     0.928     0.160  9.131341e-71
## ORTH-1457 2.275106e-72     2.1463712     0.949     0.236  9.118623e-69
```

```
##                Pfa_p_val Pfa_avg_logFC Pfa_pct.1 Pfa_pct.2 Pfa_p_val_adj
## ORTH-612  8.965743e-104      2.262306     0.861     0.209 3.593470e-100
## ORTH-300   3.376191e-99      4.346816     0.909     0.549  1.353177e-95
## ORTH-3513  1.387819e-88      3.015026     0.898     0.515  5.562380e-85
## ORTH-1259  4.263297e-01      1.579077     0.236     0.301  1.000000e+00
## ORTH-3075  1.198453e-26      2.474744     0.408     0.129  4.803398e-23
## ORTH-1457  1.367484e-44      1.672112     0.945     0.460  5.480878e-41
##                max_pval minimump_p_val
## ORTH-612  2.001129e-69  1.793149e-103
## ORTH-300  3.145221e-33   6.752382e-99
## ORTH-3513 1.290491e-55   2.775639e-88
## ORTH-1259 4.263297e-01   3.856893e-79
## ORTH-3075 1.198453e-26   4.556557e-74
## ORTH-1457 1.367484e-44   4.550211e-72
```

```
FeaturePlot(p.combined, features = c("ORTH-612", "ORTH-300", "ORTH-3513", "ORTH-1259"),
    min.cutoff = "q9", reduction = "umap")
```

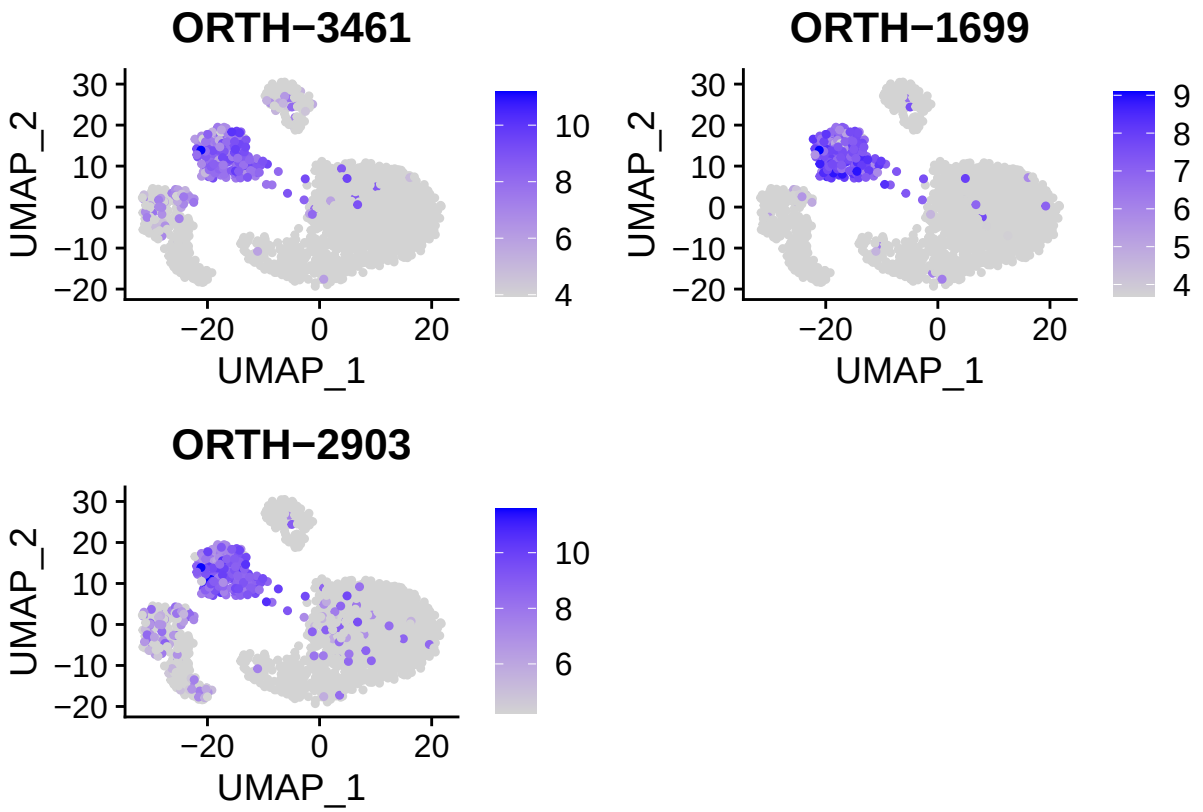



```
female.markers <- FindConservedMarkers(p.combined, ident.1 = 3, grouping.var = "species",
    verbose = FALSE, logfc.threshold = 0.25, only.pos = TRUE)
head(female.markers)
```

```
##                Pbe_p_val Pbe_avg_logFC Pbe_pct.1 Pbe_pct.2 Pbe_p_val_adj
## ORTH-3461 1.284598e-71      3.024130     0.991     0.218  5.148667e-68
## ORTH-1699 7.876107e-91      3.754212     0.991     0.099  3.156744e-87
## ORTH-2903 4.210306e-66      2.965033     1.000     0.297  1.687491e-62
## ORTH-3401 4.064024e-91      3.526841     1.000     0.099  1.628861e-87
## ORTH-3715 2.115373e-76      3.265671     1.000     0.209  8.478417e-73
## ORTH-3786 1.958535e-33      1.519718     0.690     0.137  7.849807e-30
```

```
##                Pfa_p_val Pfa_avg_logFC Pfa_pct.1 Pfa_pct.2 Pfa_p_val_adj
## ORTH-3461 1.020362e-234      6.733655     0.927     0.001 4.089613e-231
## ORTH-1699 7.030777e-227      4.524342     0.964     0.004 2.817936e-223
## ORTH-2903 7.607385e-227      5.980972     1.000     0.007 3.049040e-223
## ORTH-3401 2.572149e-226      4.848129     0.927     0.003 1.030917e-222
## ORTH-3715 1.761493e-224      6.778662     0.964     0.005 7.060062e-221
## ORTH-3786 2.661034e-222      3.740595     0.945     0.004 1.066542e-218
##                max_pval minimump_p_val
## ORTH-3461 1.284598e-71  2.040725e-234
## ORTH-1699 7.876107e-91  1.406155e-226
## ORTH-2903 4.210306e-66  1.521477e-226
## ORTH-3401 4.064024e-91  5.144298e-226
## ORTH-3715 2.115373e-76  3.522985e-224
## ORTH-3786 1.958535e-33  5.322067e-222
```

```
FeaturePlot(p.combined, features = c("ORTH-3461", "ORTH-1699", "ORTH-2903"), min.cutoff = "q9",
    reduction = "umap")
```

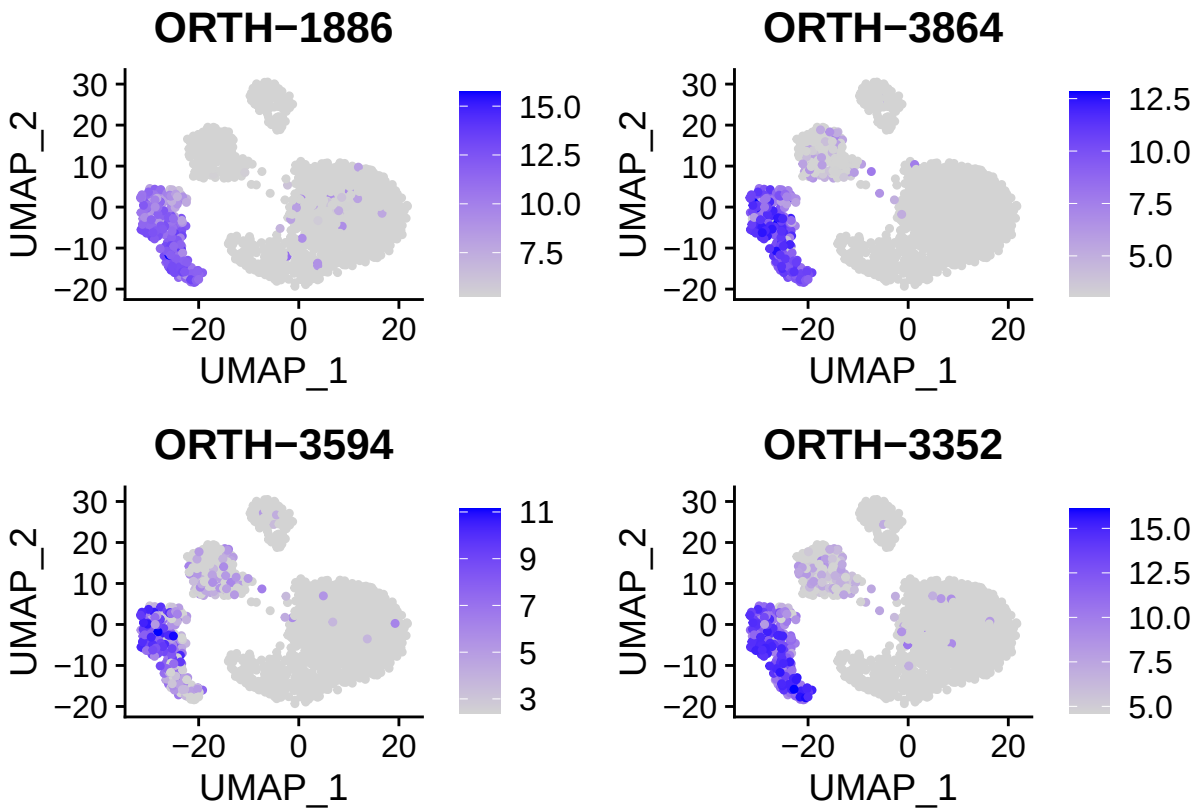



```
ook.markers <- FindConservedMarkers(p.combined, ident.1 = 2, grouping.var = "species",
    verbose = FALSE, logfc.threshold = 0.25, only.pos = TRUE)
head(ook.markers)
```

```
##                Pbe_p_val Pbe_avg_logFC Pbe_pct.1 Pbe_pct.2 Pbe_p_val_adj
## ORTH-1886 2.570544e-90      6.715708     1.000     0.100  1.030274e-86
## ORTH-3864 2.739265e-81      7.574164     1.000     0.162  1.097897e-77
## ORTH-3594 2.698085e-63      5.938248     0.922     0.162  1.081393e-59
## ORTH-3352 3.728015e-68      8.581700     0.971     0.238  1.494188e-64
## ORTH-1420 7.697858e-29      2.276606     0.786     0.262  3.085301e-25
## ORTH-3772 4.714039e-73      8.242011     1.000     0.221  1.889387e-69
```

```
##                Pfa_p_val Pfa_avg_logFC Pfa_pct.1 Pfa_pct.2 Pfa_p_val_adj
## ORTH-1886 1.782472e-249      7.614759     1.000     0.004 7.144149e-246
## ORTH-3864 5.692443e-240     10.068608     0.991     0.007 2.281531e-236
## ORTH-3594 5.833132e-236      6.880828     0.929     0.001 2.337919e-232
## ORTH-3352 9.612408e-236     11.691013     0.982     0.008 3.852653e-232
## ORTH-1420 1.521611e-229      5.848942     0.929     0.004 6.098615e-226
## ORTH-3772 1.014920e-225     11.985702     0.884     0.000 4.067801e-222
##                max_pval minimump_p_val
## ORTH-1886 2.570544e-90  3.564944e-249
## ORTH-3864 2.739265e-81  1.138489e-239
## ORTH-3594 2.698085e-63  1.166626e-235
## ORTH-3352 3.728015e-68  1.922482e-235
## ORTH-1420 7.697858e-29  3.043221e-229
## ORTH-3772 4.714039e-73  2.029841e-225
```

```
FeaturePlot(p.combined, features = c("ORTH-1886", "ORTH-3864", "ORTH-3594", "ORTH-3352"),
    min.cutoff = "q9", reduction = "umap")
```

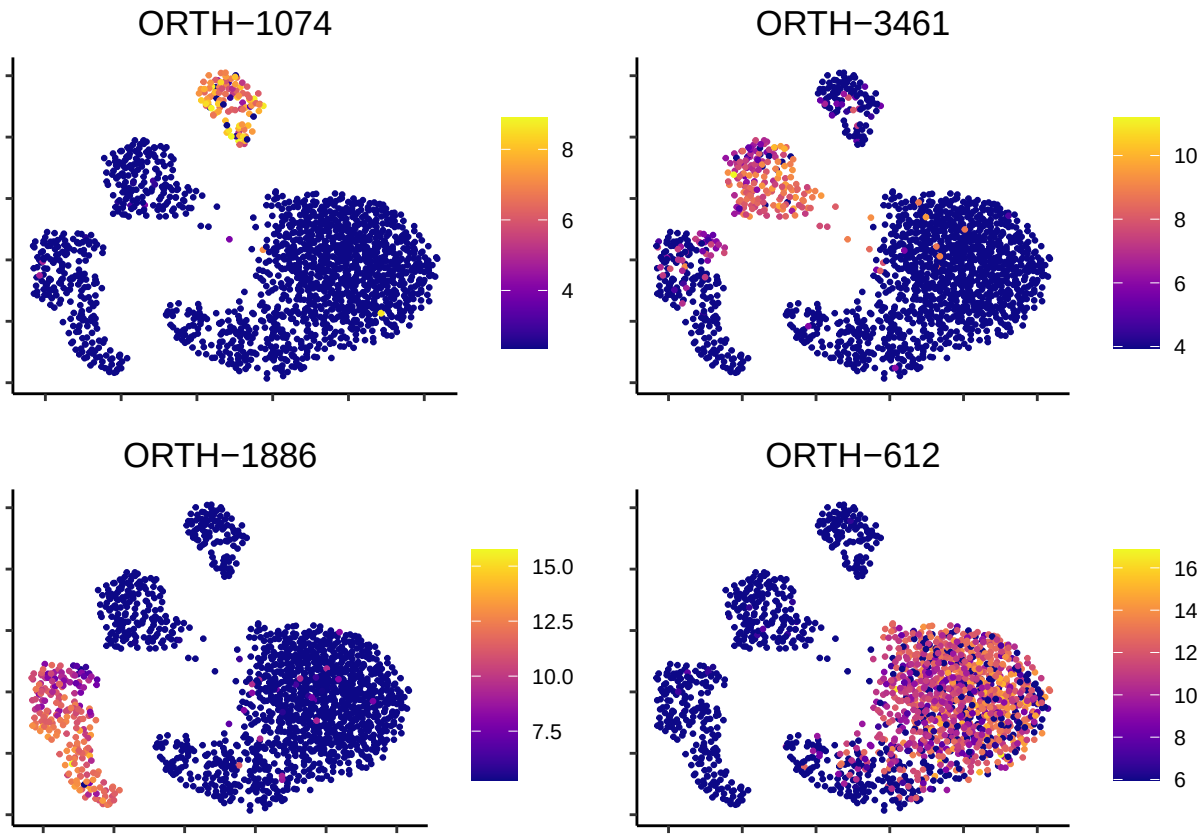



```
male.markers <- FindConservedMarkers(p.combined, ident.1 = 4, grouping.var = "species",
    verbose = FALSE, logfc.threshold = 0.25, only.pos = TRUE)
head(male.markers)
```

```
##                Pbe_p_val Pbe_avg_logFC Pbe_pct.1 Pbe_pct.2 Pbe_p_val_adj
## ORTH-1074  3.344615e-94     4.2150774     0.987     0.047  1.340522e-90
## ORTH-3921  1.399571e-76     6.7757226     0.987     0.117  5.609481e-73
## ORTH-2900  1.000801e-41     1.1562713     0.947     0.261  4.011209e-38
## ORTH-1862  3.428901e-12     0.8611099     0.600     0.224  1.374303e-08
## ORTH-291  3.464003e-115     7.9173867     1.000     0.012 1.388372e-111
## ORTH-3918  3.019281e-54     4.5638147     0.960     0.218  1.210128e-50
```

```
##               Pfa_p_val Pfa_avg_logFC Pfa_pct.1 Pfa_pct.2 Pfa_p_val_adj
## ORTH-1074 6.383493e-219      6.846696     0.848     0.000 2.558504e-215
## ORTH-3921 6.352508e-210      7.201473     1.000     0.007 2.546085e-206
## ORTH-2900 3.765386e-207      6.552619     0.970     0.006 1.509167e-203
## ORTH-1862 1.064799e-204      6.319949     1.000     0.008 4.267715e-201
## ORTH-291  8.749907e-195      1.638727     0.909     0.005 3.506963e-191
## ORTH-3918 1.180799e-193      5.649033     0.879     0.004 4.732642e-190
##                max_pval minimump_p_val
## ORTH-1074  3.344615e-94  1.276699e-218
## ORTH-3921  1.399571e-76  1.270502e-209
## ORTH-2900  1.000801e-41  7.530771e-207
## ORTH-1862  3.428901e-12  2.129598e-204
## ORTH-291  3.464003e-115  1.749981e-194
## ORTH-3918  3.019281e-54  2.361598e-193
```

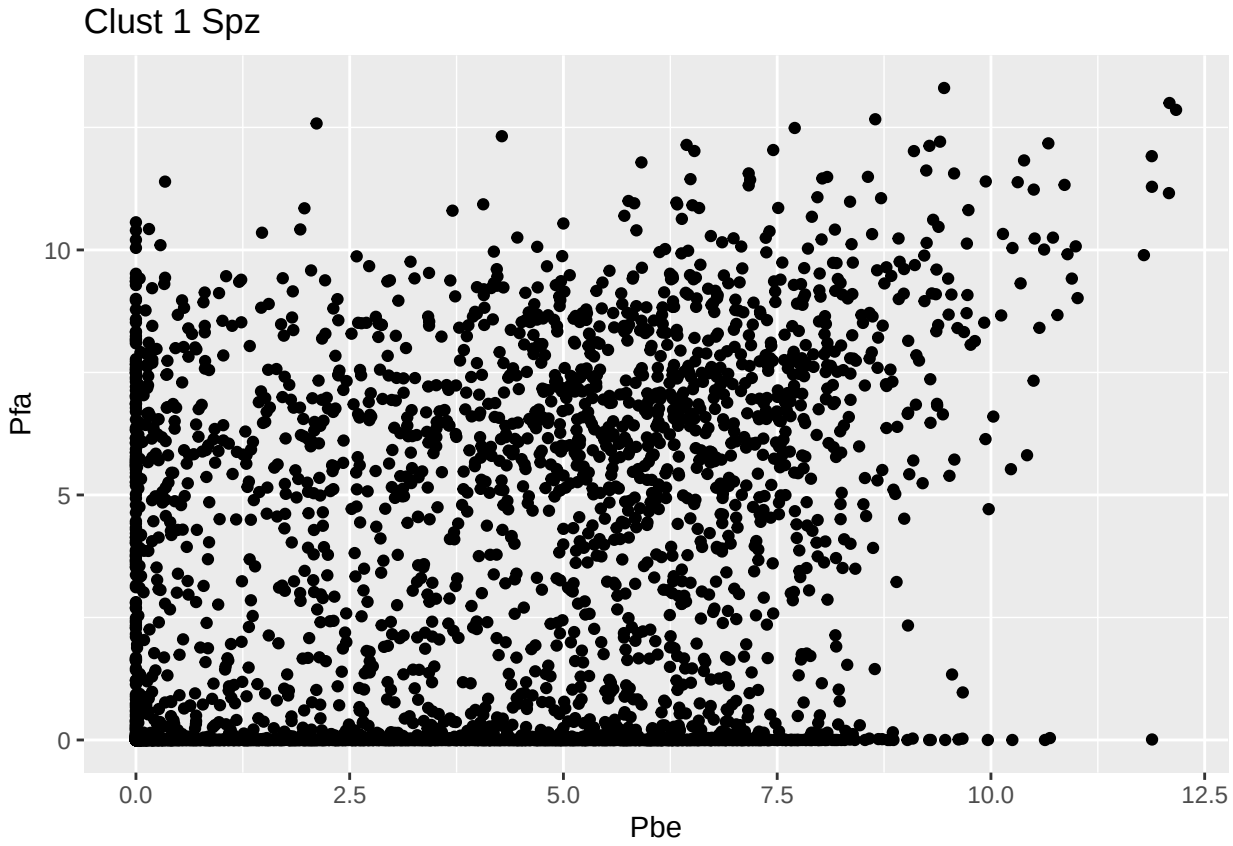
```
p <- FeaturePlot(p.combined, features = c("ORTH-1074", "ORTH-3921", "ORTH-2900",
    "ORTH-1862"), min.cutoff = "q9", reduction = "umap")

s <- FeaturePlot(p.combined, features = c("ORTH-612"), min.cutoff = "q9", reduction = "umap")
sdat <- s$data
ps <- ggplot(sdat, aes(UMAP_1, UMAP_2)) + geom_point(aes_string(colour = "ORTH.612"),
    size = 0.5) + labs(title = "ORTH-612", x = element_blank(), y = element_blank()) +
    scale_colour_viridis(option = "C") + theme_classic() + theme(axis.title = element_text(size = 8),
    legend.text = element_text(size = 8), legend.title = element_blank(), axis.text = element_text(size
    axis.text.x = element_blank(), axis.text.y = element_blank(), plot.title = element_text(hjust = 0.5)

f <- FeaturePlot(p.combined, features = c("ORTH-3461"), min.cutoff = "q9", reduction = "umap")
fdat <- f$data
pf <- ggplot(fdat, aes(UMAP_1, UMAP_2)) + geom_point(aes_string(colour = "ORTH.3461"),
    size = 0.5) + labs(title = "ORTH-3461", x = element_blank(), y = element_blank()) +
    scale_colour_viridis(option = "C") + theme_classic() + theme(axis.title = element_text(size = 8),
    legend.text = element_text(size = 8), legend.title = element_blank(), axis.text = element_text(size
    axis.text.x = element_blank(), axis.text.y = element_blank(), plot.title = element_text(hjust = 0.5)

o <- FeaturePlot(p.combined, features = c("ORTH-1886"), min.cutoff = "q9", reduction = "umap")
odat <- o$data
po <- ggplot(odat, aes(UMAP_1, UMAP_2)) + geom_point(aes_string(colour = "ORTH.1886"),
    size = 0.5) + labs(title = "ORTH-1886", x = element_blank(), y = element_blank()) +
    scale_colour_viridis(option = "C") + theme_classic() + theme(axis.title = element_text(size = 8),
    legend.text = element_text(size = 8), legend.title = element_blank(), axis.text = element_text(size
    axis.text.x = element_blank(), axis.text.y = element_blank(), plot.title = element_text(hjust = 0.5)

m <- FeaturePlot(p.combined, features = c("ORTH-1074"), min.cutoff = "q9", reduction = "umap")
mdat <- m$data
pm <- ggplot(mdat, aes(UMAP_1, UMAP_2)) + geom_point(aes_string(colour = "ORTH.1074"),
    size = 0.5) + labs(title = "ORTH-1074", x = element_blank(), y = element_blank()) +
    scale_colour_viridis(option = "C") + theme_classic() + theme(axis.title = element_text(size = 8),
    legend.text = element_text(size = 8), legend.title = element_blank(), axis.text = element_text(size
    axis.text.x = element_blank(), axis.text.y = element_blank(), plot.title = element_text(hjust = 0.5)



grid.arrange(pm, pf, po, ps, nrow = 2, ncol = 2)
```



Identify differential expressed genes across conditions

Now that we've aligned the stimulated and control cells, we can start to do comparative analyses and look at the differences induced by stimulation. One way to look broadly at these changes is to plot the average expression of both the stimulated and control cells and look for genes that are visual outliers on a scatter plot.

```
spz.cells <- subset(p.combined, idents = "1")
Idents(spz.cells) <- "species"
avg.spz.cells <- log1p(AverageExpression(spz.cells, verbose = FALSE)$RNA)
avg.spz.cells$gene <- rownames(avg.spz.cells)

ggplot(avg.spz.cells, aes(Pbe, Pfa)) + geom_point() + ggtitle("Clust 1 Spz")
```



Because we are confident in having identified common cell types across condition, we can ask what genes change in different conditions for cells of the same type. First, we create a column in the meta.data slot to hold both the cell type and stimulation information and switch the current ident to that column. Then we use FindMarkers to find the genes that are different between stimulated and control B cells.

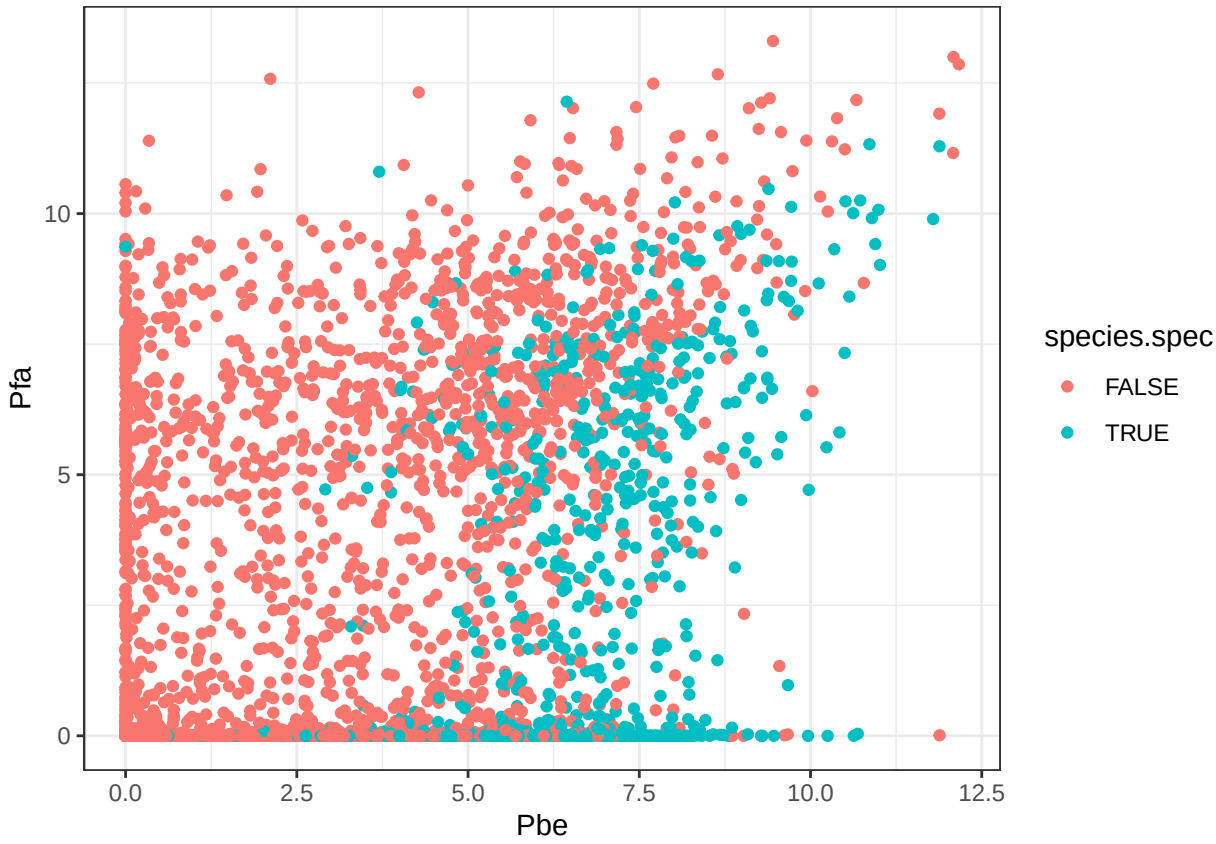
```
p.combined$clust_species <- paste(Idents(p.combined), p.combined$species, sep = "_")
p.combined$celltype <- Idents(p.combined)
Idents(p.combined) <- "clust_species"
species.spec <- FindMarkers(p.combined, ident.1 = "1_Pfa", ident.2 = "1_Pbe", verbose = FALSE)

species.spec$pbgene <- newot[match(rownames(species.spec), newot[, 12]), 2]
species.spec$pfgene <- newot[match(rownames(species.spec), newot[, 12]), 4]

head(species.spec, n = 15)
```

```
##                  p_val  avg_logFC pct.1 pct.2    p_val_adj         pbgene
## ORTH-1911 1.242975e-47  -3.796954 0.013 0.906 4.981843e-44 PBANKA_1220800
## ORTH-99   3.152783e-44  -4.708340 0.013 0.844 1.263636e-40 PBANKA_1411000
## ORTH-1752 4.076597e-44  -1.155660 0.034 0.969 1.633900e-40 PBANKA_1203800
## ORTH-3035 1.002909e-43 -10.250252 0.000 0.750 4.019658e-40 PBANKA_1002100
## ORTH-512  8.047663e-42  -9.296917 0.000 0.719 3.225503e-38 PBANKA_1454900
## ORTH-4235 1.519936e-41  -4.470804 0.004 0.750 6.091903e-38 PBANKA_0211300
## ORTH-3586 1.866733e-40  -1.939022 0.030 0.875 7.481865e-37 PBANKA_1123800
## ORTH-462  3.595004e-40  -8.665634 0.004 0.719 1.440877e-36 PBANKA_1449700
## ORTH-1188 6.256671e-40  -8.290061 0.000 0.688 2.507674e-36 PBANKA_1320200
## ORTH-3435 6.256671e-40  -9.963825 0.000 0.688 2.507674e-36 PBANKA_1107600
## ORTH-363  6.482645e-40  -6.043740 0.004 0.719 2.598244e-36 PBANKA_1439100
## ORTH-2876 1.746066e-38  -2.279791 0.013 0.750 6.998232e-35 PBANKA_0820400
```

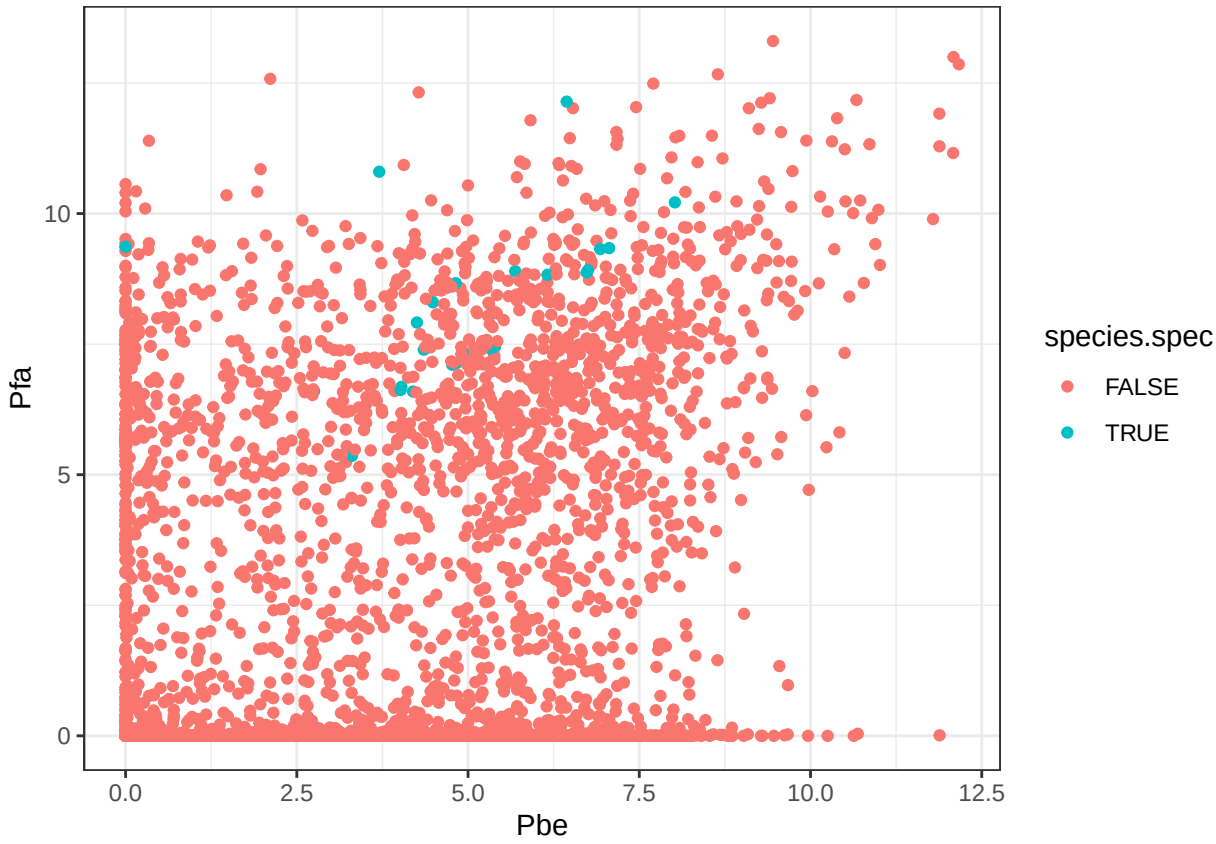
```
## ORTH-1679 4.714376e-38  -8.083441 0.000 0.656 1.889522e-34 PBANKA_0107900
## ORTH-3180 4.714376e-38  -8.527358 0.000 0.656 1.889522e-34 PBANKA_1017800
## ORTH-3270 1.774910e-37  -1.187354 0.026 0.812 7.113841e-34 PBANKA_1028500
##                  pfgene
## ORTH-1911 PF3D7_0710200
## ORTH-99   PF3D7_1312500
## ORTH-1752 PF3D7_1005600
## ORTH-3035 PF3D7_0404400
## ORTH-512  PF3D7_1241500
## ORTH-4235 PF3D7_0727200
## ORTH-3586 PF3D7_0624900
## ORTH-462  PF3D7_1235100
## ORTH-1188 PF3D7_1456500
## ORTH-3435 PF3D7_0508000
## ORTH-363  PF3D7_1224200
## ORTH-2876 PF3D7_0919500
## ORTH-1679 PF3D7_0609300
## ORTH-3180 PF3D7_1426800
## ORTH-3270 PF3D7_1414200
```

```
spz.markers$pbgene <- newot[match(rownames(spz.markers), newot[, 12]), 2]
spz.markers$pfgene <- newot[match(rownames(spz.markers), newot[, 12]), 4]

subss <- species.spec[species.spec$p_val_adj < 0.001, ]
subss2 <- subss[subss$avg_logFC > 2, ]

# write.csv(spz.markers, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/spz1_markers.csv')
# write.csv(subss, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/spz1_species_specific.csv')
# write.csv(subss2, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/spz1_pf_specific.csv')
```

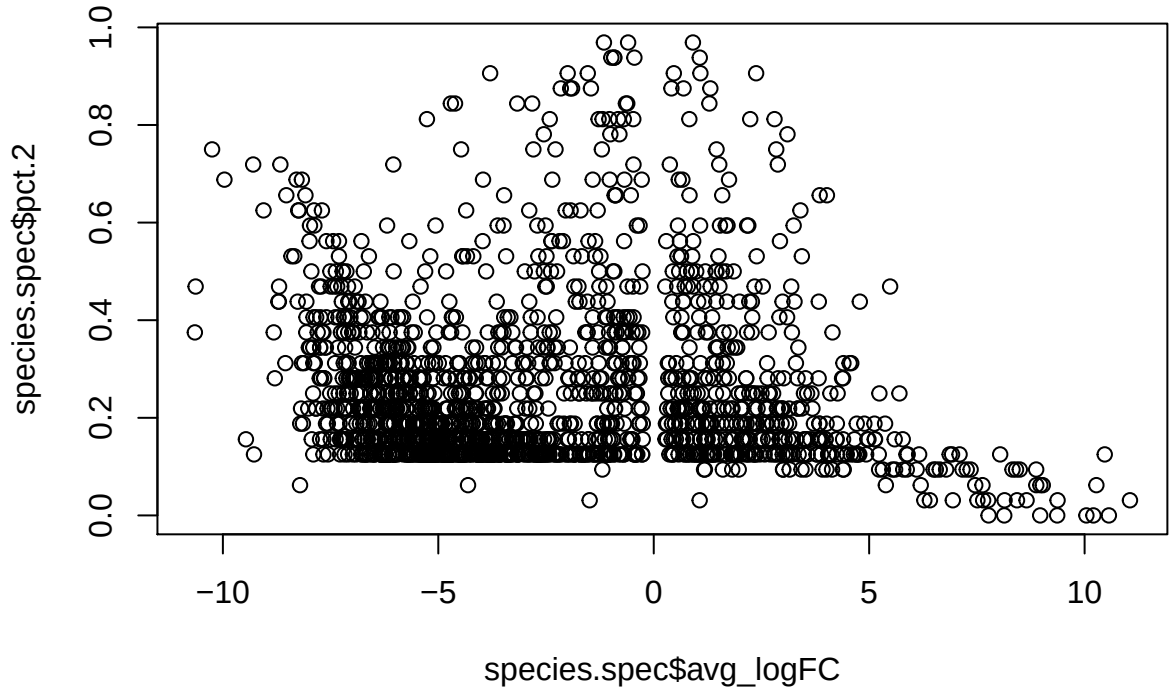
```
avg.spz.cells$species.spec <- rep("FALSE", length(avg.spz.cells$Pfa))
avg.spz.cells[which(rownames(avg.spz.cells) %in% rownames(subss)), ]$species.spec <- "TRUE"

ggplot(avg.spz.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = species.spec)) +
    theme_bw()
```



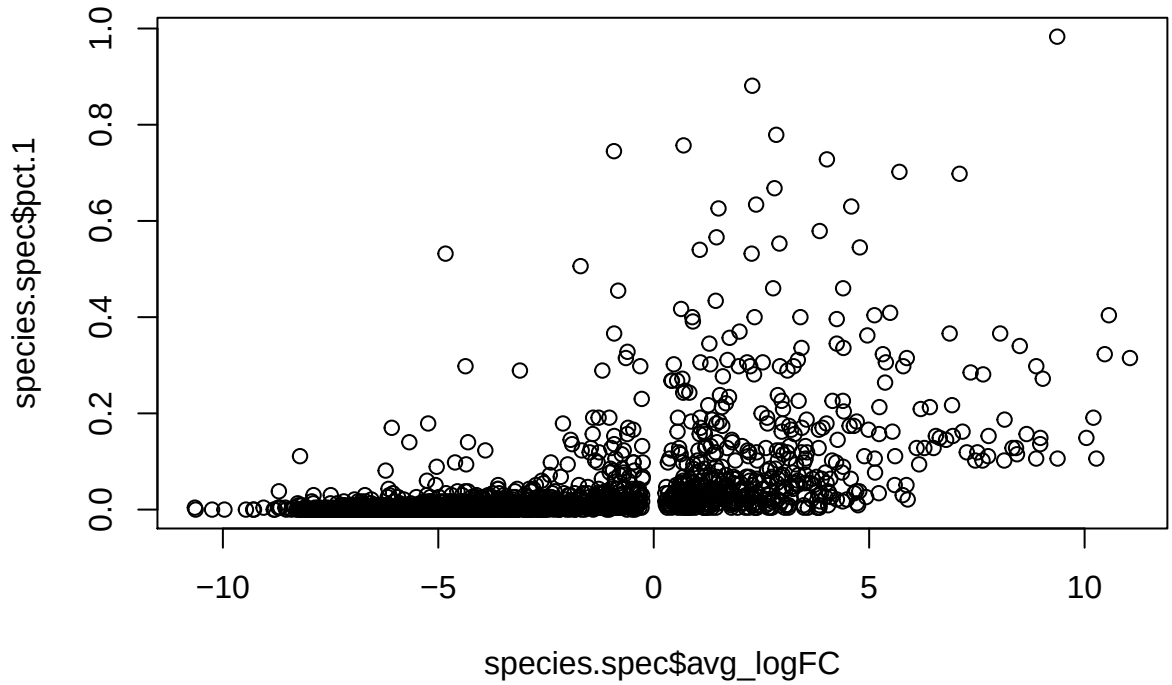

```
avg.spz.cells$species.spec <- rep("FALSE", length(avg.spz.cells$Pfa))
avg.spz.cells[which(rownames(avg.spz.cells) %in% rownames(subss2)), ]$species.spec <- "TRUE"

ggplot(avg.spz.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = species.spec)) +
    theme_bw()
```




```
plot(species.spec$avg_logFC, species.spec$pct.2)
```



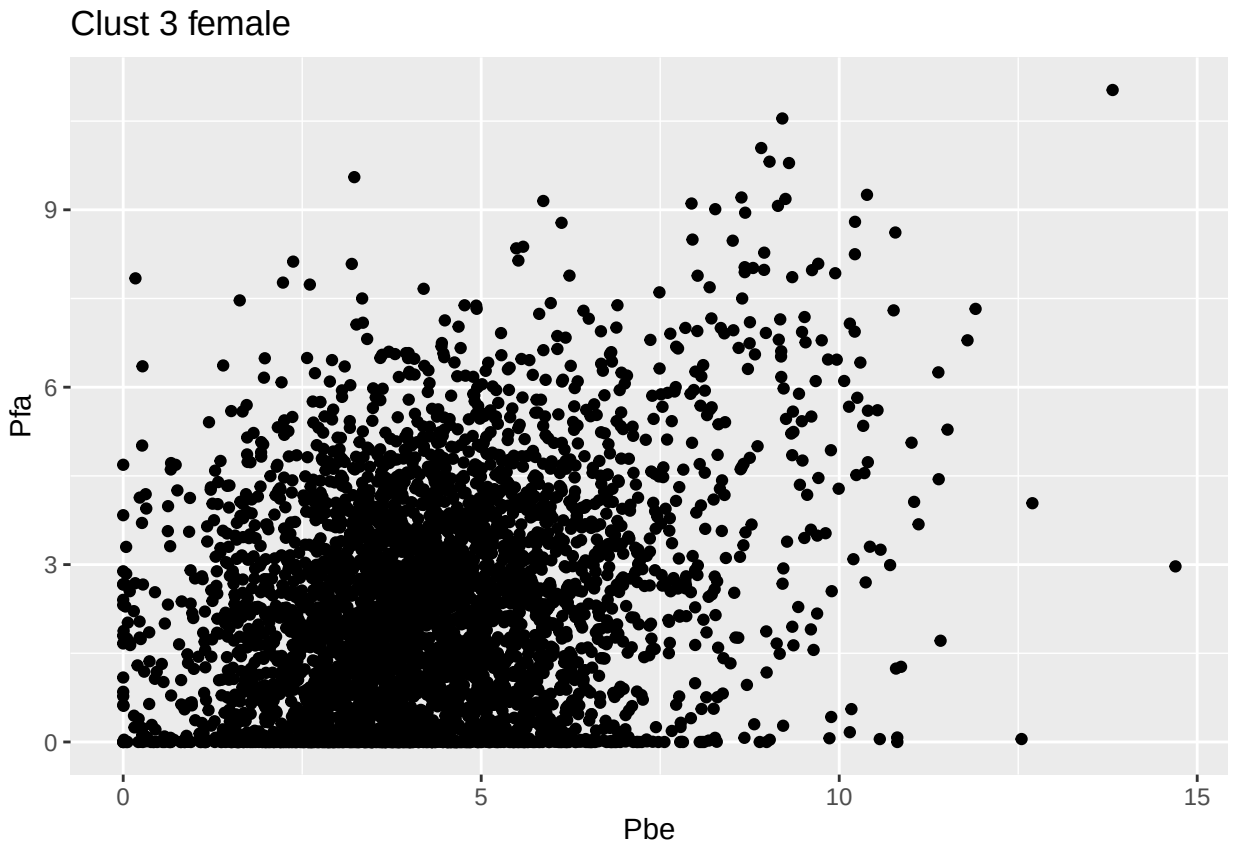

```
plot(species.spec$avg_logFC, species.spec$pct.1)
```




```
p <- ggplot(species.spec, aes(x = pct.1, y = pct.2)) + geom_point(aes(colour = avg_logFC))
```

Try the same with females

```
Idents(p.combined) <- "celltype"
f.cells <- subset(p.combined, idents = "3")
Idents(f.cells) <- "species"
avg.f.cells <- log1p(AverageExpression(f.cells, verbose = FALSE)$RNA)
avg.f.cells$gene <- rownames(avg.f.cells)

ggplot(avg.f.cells, aes(Pbe, Pfa)) + geom_point() + ggtitle("Clust 3 female")
```




```
Idents(p.combined) <- "clust_species"
species.spec <- FindMarkers(p.combined, ident.1 = "3_Pfa", ident.2 = "3_Pbe", verbose = FALSE)
head(species.spec, n = 15)
```

```
##                  p_val avg_logFC pct.1 pct.2    p_val_adj
## ORTH-1950 9.129610e-36  6.083760 0.982 0.009 3.659148e-32
## ORTH-2394 1.056291e-33  7.669609 0.964 0.034 4.233613e-30
## ORTH-1530 5.056151e-33  4.970782 0.982 0.060 2.026505e-29
## ORTH-3079 1.019999e-32  4.892329 1.000 0.112 4.088154e-29
## ORTH-4127 4.950360e-32  3.943448 0.982 0.069 1.984104e-28
## ORTH-2435 3.100052e-31  5.537888 0.982 0.129 1.242501e-27
## ORTH-2820 6.802229e-31  3.973858 0.982 0.121 2.726333e-27
## ORTH-1917 8.234486e-31  4.087523 1.000 0.138 3.300382e-27
## ORTH-389  1.254905e-29  3.542078 0.982 0.129 5.029659e-26
## ORTH-3513 2.381921e-29  5.840251 0.964 0.164 9.546738e-26
## ORTH-2941 3.527116e-29  6.324635 1.000 0.276 1.413668e-25
## ORTH-4182 5.454423e-29  3.194551 0.855 0.009 2.186133e-25
## ORTH-2881 7.944183e-29  4.687856 0.818 0.000 3.184029e-25
## ORTH-2408 1.201173e-28  5.128522 0.982 0.224 4.814303e-25
## ORTH-1279 3.720511e-28  5.749826 1.000 0.362 1.491181e-24
```

```
species.spec$pbgene <- newot[match(rownames(species.spec), newot[, 12]), 2]
species.spec$pfgene <- newot[match(rownames(species.spec), newot[, 12]), 4]

head(species.spec, n = 15)
```

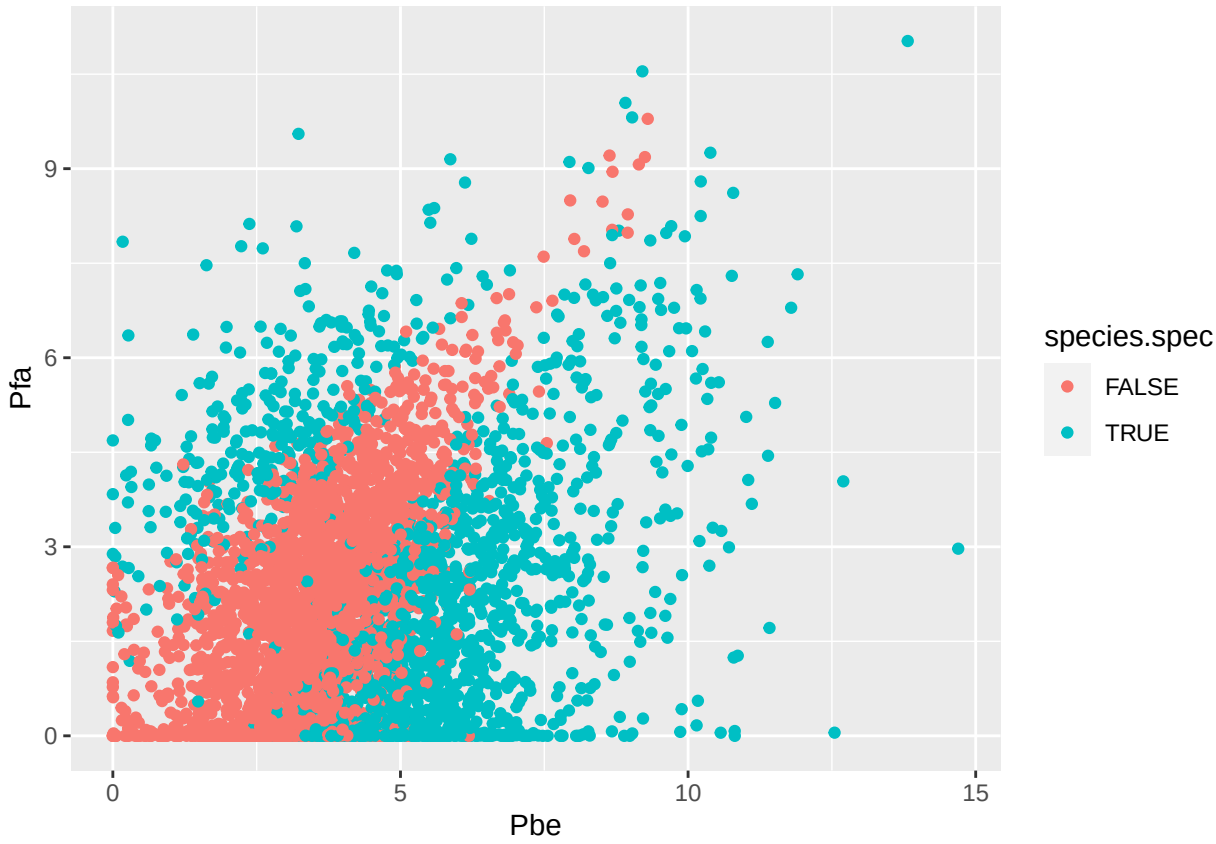
```
##                  p_val avg_logFC pct.1 pct.2    p_val_adj         pbgene
## ORTH-1950 9.129610e-36  6.083760 0.982 0.009 3.659148e-32 PBANKA_1225200
```

```
## ORTH-2394 1.056291e-33  7.669609 0.964 0.034 4.233613e-30 PBANKA_0306100
## ORTH-1530 5.056151e-33  4.970782 0.982 0.060 2.026505e-29 PBANKA_1357400
## ORTH-3079 1.019999e-32  4.892329 1.000 0.112 4.088154e-29 PBANKA_1006800
## ORTH-4127 4.950360e-32  3.943448 0.982 0.069 1.984104e-28 PBANKA_0720100
## ORTH-2435 3.100052e-31  5.537888 0.982 0.129 1.242501e-27 PBANKA_0310500
## ORTH-2820 6.802229e-31  3.973858 0.982 0.121 2.726333e-27 PBANKA_0814700
## ORTH-1917 8.234486e-31  4.087523 1.000 0.138 3.300382e-27 PBANKA_1221400
## ORTH-389  1.254905e-29  3.542078 0.982 0.129 5.029659e-26 PBANKA_1441700
## ORTH-3513 2.381921e-29  5.840251 0.964 0.164 9.546738e-26 PBANKA_1116000
## ORTH-2941 3.527116e-29  6.324635 1.000 0.276 1.413668e-25 PBANKA_0827200
## ORTH-4182 5.454423e-29  3.194551 0.855 0.009 2.186133e-25 PBANKA_0205700
## ORTH-2881 7.944183e-29  4.687856 0.818 0.000 3.184029e-25 PBANKA_0820900
## ORTH-2408 1.201173e-28  5.128522 0.982 0.224 4.814303e-25 PBANKA_0307500
## ORTH-1279 3.720511e-28  5.749826 1.000 0.362 1.491181e-24 PBANKA_1330200
##                   pfgene
## ORTH-1950 PF3D7_0805400
## ORTH-2394 PF3D7_0209000
## ORTH-1530 PF3D7_1344500
## ORTH-3079 PF3D7_0409200
## ORTH-4127 PF3D7_0418000
## ORTH-2435 PF3D7_0213600
## ORTH-2820 PF3D7_0913700
## ORTH-1917 PF3D7_0710800
## ORTH-389  PF3D7_1226900
## ORTH-3513 PF3D7_0616500
## ORTH-2941 PF3D7_0926400
## ORTH-4182 PF3D7_0107700
## ORTH-2881 PF3D7_0920000
## ORTH-2408 PF3D7_0210600
## ORTH-1279 PF3D7_1466800
```

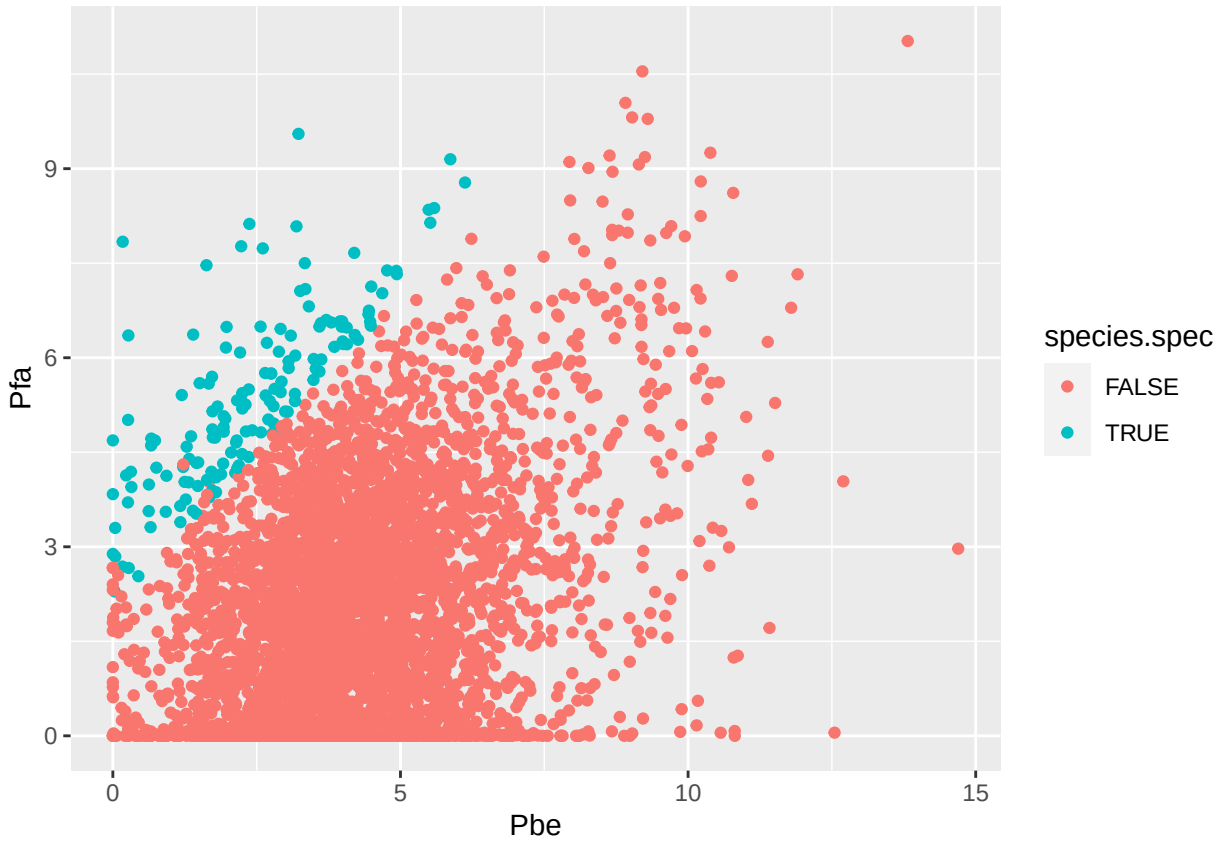
```
female.markers$pbgene <- newot[match(rownames(female.markers), newot[, 12]), 2]
female.markers$pfgene <- newot[match(rownames(female.markers), newot[, 12]), 4]

subss <- species.spec[species.spec$p_val_adj < 0.001, ]
subss2 <- subss[subss$avg_logFC > 2, ]

# write.csv(female.markers, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/female_markers.csv')
# write.csv(subss, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/female_species_specific.csv')
# write.csv(subss2, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/female_pf_specific.csv')
```

‘‘‘

```
avg.f.cells$species.spec <- rep("FALSE", length(avg.f.cells$Pfa))
avg.f.cells[which(rownames(avg.f.cells) %in% rownames(subss)), ]$species.spec <- "TRUE"
ggplot(avg.f.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = species.spec))
```




```
avg.f.cells$species.spec <- rep("FALSE", length(avg.f.cells$Pfa))
avg.f.cells[which(rownames(avg.f.cells) %in% rownames(subss2)), ]$species.spec <- "TRUE"
ggplot(avg.f.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = species.spec))
```

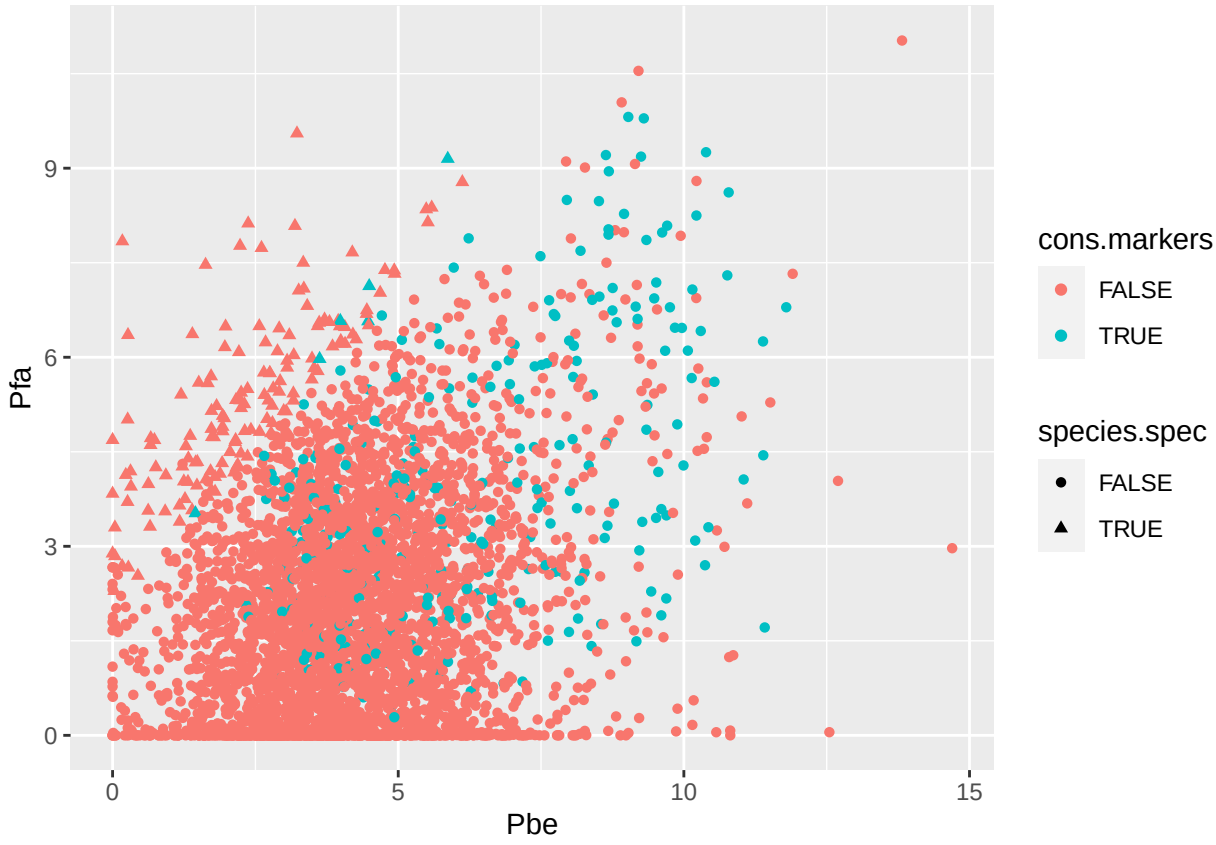



```
head(female.markers)
```

```
##                Pbe_p_val Pbe_avg_logFC Pbe_pct.1 Pbe_pct.2 Pbe_p_val_adj
## ORTH-3461 1.284598e-71      3.024130     0.991     0.218  5.148667e-68
## ORTH-1699 7.876107e-91      3.754212     0.991     0.099  3.156744e-87
## ORTH-2903 4.210306e-66      2.965033     1.000     0.297  1.687491e-62
## ORTH-3401 4.064024e-91      3.526841     1.000     0.099  1.628861e-87
## ORTH-3715 2.115373e-76      3.265671     1.000     0.209  8.478417e-73
## ORTH-3786 1.958535e-33      1.519718     0.690     0.137  7.849807e-30
##                Pfa_p_val Pfa_avg_logFC Pfa_pct.1 Pfa_pct.2 Pfa_p_val_adj
## ORTH-3461 1.020362e-234      6.733655     0.927     0.001 4.089613e-231
## ORTH-1699 7.030777e-227      4.524342     0.964     0.004 2.817936e-223
## ORTH-2903 7.607385e-227      5.980972     1.000     0.007 3.049040e-223
## ORTH-3401 2.572149e-226      4.848129     0.927     0.003 1.030917e-222
## ORTH-3715 1.761493e-224      6.778662     0.964     0.005 7.060062e-221
## ORTH-3786 2.661034e-222      3.740595     0.945     0.004 1.066542e-218
##                max_pval minimump_p_val         pbgene        pfgene
## ORTH-3461 1.284598e-71  2.040725e-234 PBANKA_1110200 PF3D7_0510700
## ORTH-1699 7.876107e-91  1.406155e-226 PBANKA_0109900 PF3D7_0611600
## ORTH-2903 4.210306e-66  1.521477e-226 PBANKA_0823200 PF3D7_0922300
## ORTH-3401 4.064024e-91  5.144298e-226 PBANKA_1104100 PF3D7_0504500
## ORTH-3715 2.115373e-76  3.522985e-224 PBANKA_1137400 PF3D7_1361300
## ORTH-3786 1.958535e-33  5.322067e-222 PBANKA_1145100 PF3D7_1369200
```

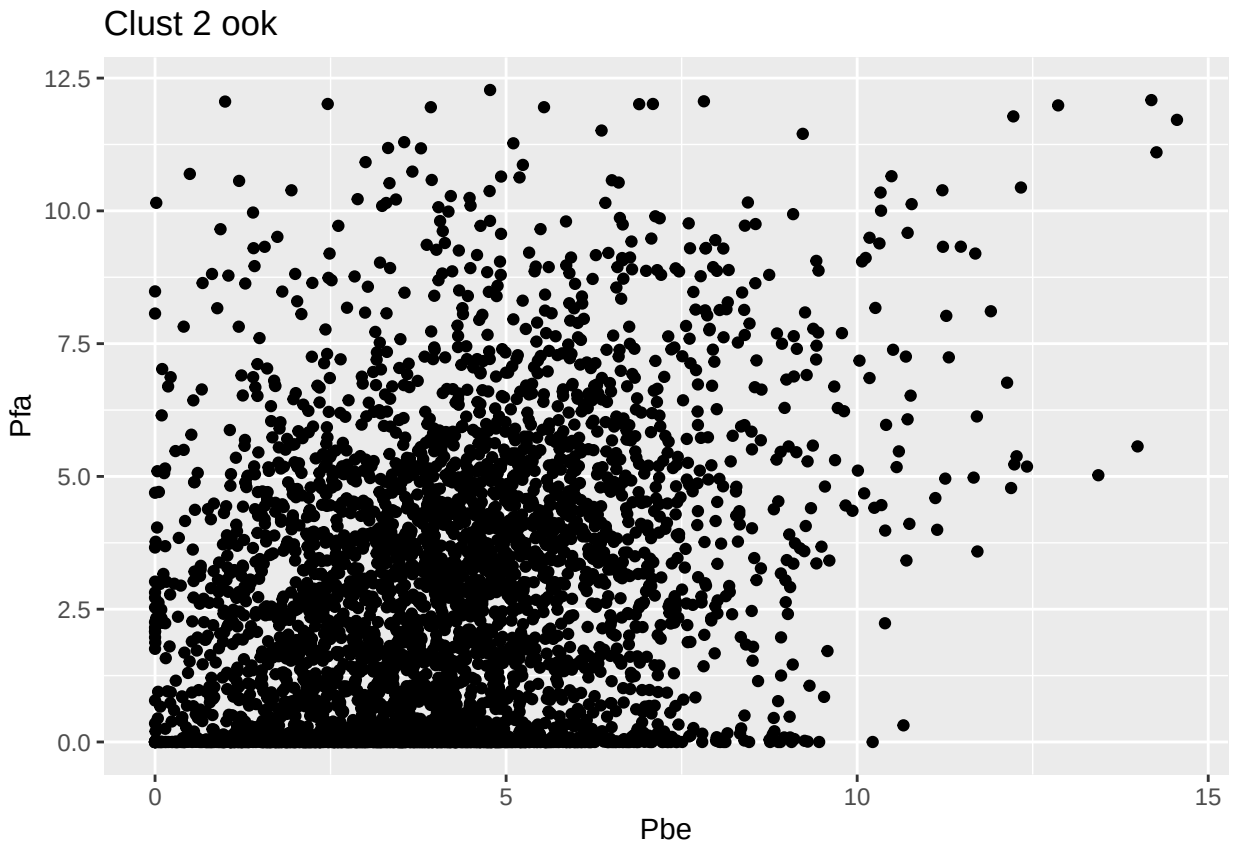
```
avg.f.cells$cons.markers <- rep("FALSE", length(avg.f.cells$Pfa))
avg.f.cells[which(rownames(avg.f.cells) %in% rownames(female.markers)), ]$cons.markers <- "TRUE"

ggplot(avg.f.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = cons.markers,
```

```
    shape = species.spec))
```



try the same with ookinetes

```
Idents(p.combined) <- "celltype"
ook.cells <- subset(p.combined, idents = "2")
Idents(ook.cells) <- "species"
avg.ook.cells <- log1p(AverageExpression(ook.cells, verbose = FALSE)$RNA)
avg.ook.cells$gene <- rownames(avg.ook.cells)

ggplot(avg.ook.cells, aes(Pbe, Pfa)) + geom_point() + ggtitle("Clust 2 ook")
```




```
Idents(p.combined) <- "clust_species"
species.spec <- FindMarkers(p.combined, ident.1 = "2_Pfa", ident.2 = "2_Pbe", verbose = FALSE)
head(species.spec, n = 15)
```

```
##                  p_val  avg_logFC pct.1 pct.2    p_val_adj
## ORTH-3338 1.785567e-42  -7.296336 0.000 1.000 7.156554e-39
## ORTH-2026 1.785567e-42  -9.458793 0.000 1.000 7.156554e-39
## ORTH-2436 1.785567e-42 -10.220866 0.000 1.000 7.156554e-39
## ORTH-2170 2.685390e-42  -8.072231 0.009 1.000 1.076304e-38
## ORTH-3819 3.995604e-42  -9.043062 0.018 1.000 1.601438e-38
## ORTH-1804 4.261695e-42  -8.262581 0.018 1.000 1.708087e-38
## ORTH-2776 8.854296e-42  -8.561868 0.036 1.000 3.548802e-38
## ORTH-462  1.038848e-41  -6.981230 0.036 1.000 4.163705e-38
## ORTH-1918 1.685066e-41  -8.655696 0.018 0.990 6.753745e-38
## ORTH-1387 1.685066e-41  -9.229355 0.018 0.990 6.753745e-38
## ORTH-1430 1.739964e-41  -9.050640 0.018 0.990 6.973774e-38
## ORTH-3520 1.756452e-41  -6.725349 0.045 1.000 7.039860e-38
## ORTH-1364 3.744489e-41  -7.665038 0.071 1.000 1.500791e-37
## ORTH-1595 4.842108e-41  -7.263648 0.009 0.981 1.940717e-37
## ORTH-3399 4.999645e-41  -7.405470 0.009 0.981 2.003858e-37
```

```
species.spec$pbgene <- newot[match(rownames(species.spec), newot[, 12]), 2]
species.spec$pfgene <- newot[match(rownames(species.spec), newot[, 12]), 4]

head(species.spec, n = 15)
```

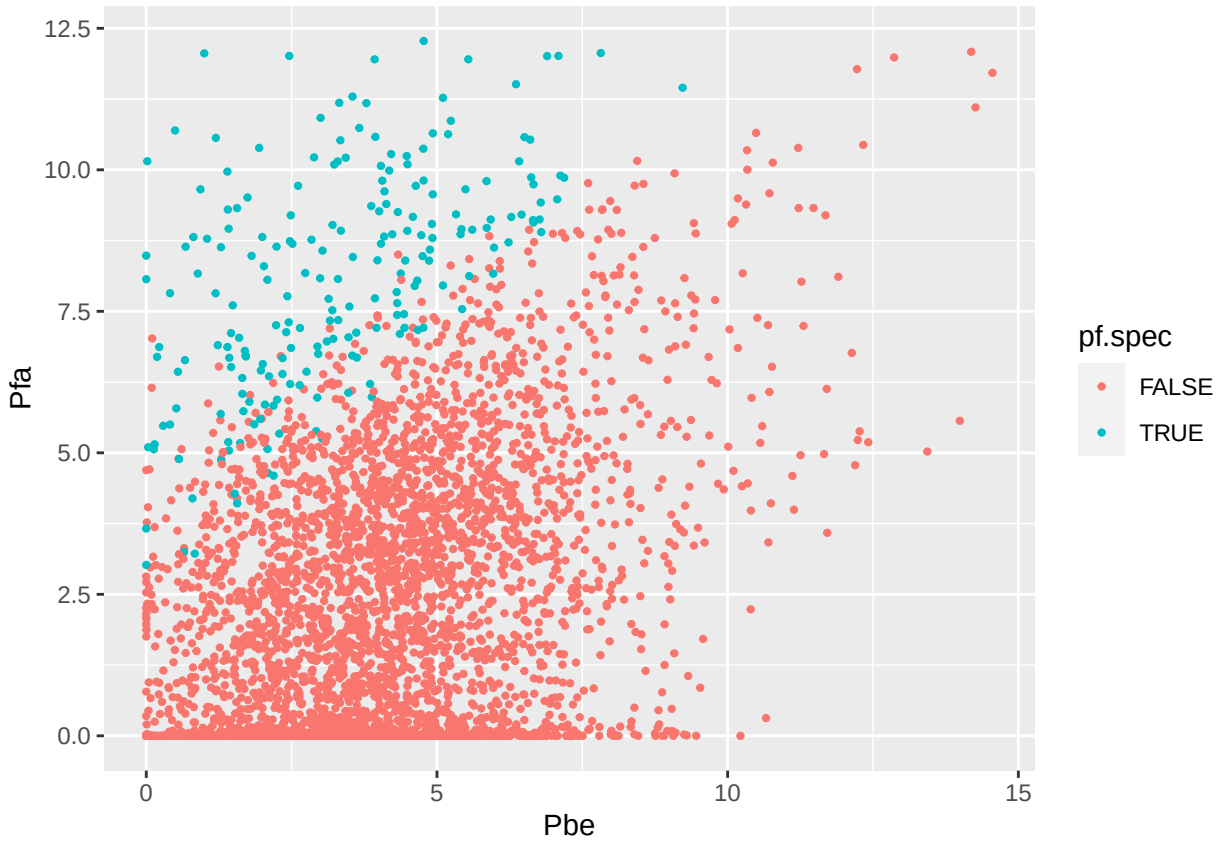
```
##                  p_val  avg_logFC pct.1 pct.2    p_val_adj         pbgene
## ORTH-3338 1.785567e-42  -7.296336 0.000 1.000 7.156554e-39 PBANKA_1036200
```

```
## ORTH-2026 1.785567e-42  -9.458793 0.000 1.000 7.156554e-39 PBANKA_1233200
## ORTH-2436 1.785567e-42 -10.220866 0.000 1.000 7.156554e-39 PBANKA_0310600
## ORTH-2170 2.685390e-42  -8.072231 0.009 1.000 1.076304e-38 PBANKA_0603400
## ORTH-3819 3.995604e-42  -9.043062 0.018 1.000 1.601438e-38 PBANKA_0403600
## ORTH-1804 4.261695e-42  -8.262581 0.018 1.000 1.708087e-38 PBANKA_1209400
## ORTH-2776 8.854296e-42  -8.561868 0.036 1.000 3.548802e-38 PBANKA_0810200
## ORTH-462  1.038848e-41  -6.981230 0.036 1.000 4.163705e-38 PBANKA_1449700
## ORTH-1918 1.685066e-41  -8.655696 0.018 0.990 6.753745e-38 PBANKA_1221500
## ORTH-1387 1.685066e-41  -9.229355 0.018 0.990 6.753745e-38 PBANKA_1342000
## ORTH-1430 1.739964e-41  -9.050640 0.018 0.990 6.973774e-38 PBANKA_1346700
## ORTH-3520 1.756452e-41  -6.725349 0.045 1.000 7.039860e-38 PBANKA_1116800
## ORTH-1364 3.744489e-41  -7.665038 0.071 1.000 1.500791e-37 PBANKA_1339600
## ORTH-1595 4.842108e-41  -7.263648 0.009 0.981 1.940717e-37 PBANKA_1364000
## ORTH-3399 4.999645e-41  -7.405470 0.009 0.981 2.003858e-37 PBANKA_1103900
##                   pfgene
## ORTH-3338 PF3D7_1406000
## ORTH-2026 PF3D7_0518400
## ORTH-2436 PF3D7_0213700
## ORTH-2170 PF3D7_1204500
## ORTH-3819 PF3D7_0305100
## ORTH-1804 PF3D7_1011000
## ORTH-2776 PF3D7_0909000
## ORTH-462  PF3D7_1235100
## ORTH-1918 PF3D7_0710900
## ORTH-1387 PF3D7_1326800
## ORTH-1430 PF3D7_1331800
## ORTH-3520 PF3D7_0617200
## ORTH-1364 PF3D7_1324400
## ORTH-1595 PF3D7_1351200
## ORTH-3399 PF3D7_0504300
```

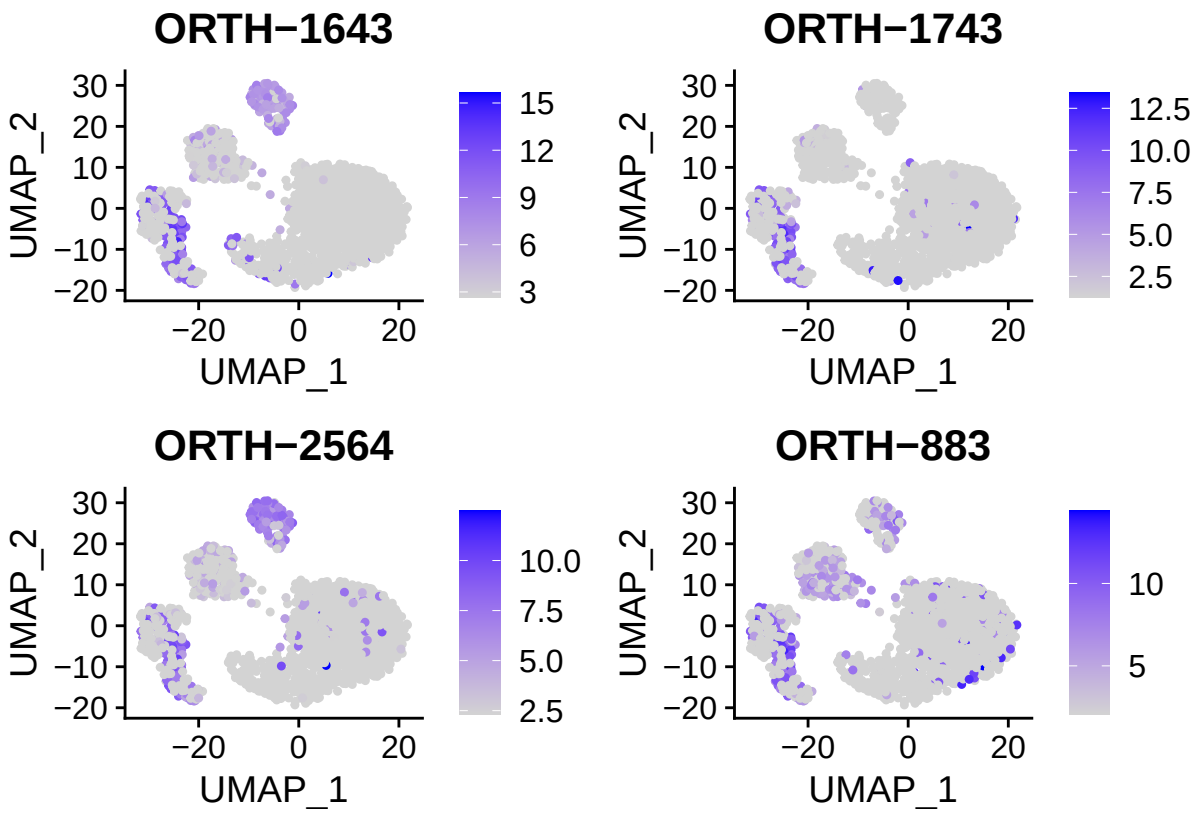
```
ook.markers$pbgene <- newot[match(rownames(ook.markers), newot[, 12]), 2]
ook.markers$pfgene <- newot[match(rownames(ook.markers), newot[, 12]), 4]

subss <- species.spec[species.spec$p_val_adj < 0.001, ]
subss2 <- subss[subss$avg_logFC > 2, ]

# write.csv(ook.markers, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/ook_markers.csv')
# write.csv(subss, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/ook_species_specific.csv')
# write.csv(subss2, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/ook_pf_specific.csv')
```

```
avg.ook.cells$species.spec <- rep("FALSE", length(avg.ook.cells$Pfa))
avg.ook.cells[which(rownames(avg.ook.cells) %in% rownames(subss)), ]$species.spec <- "TRUE"
ggplot(avg.ook.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = species.spec),
    size = 0.8)
```


```
avg.ook.cells$species.spec <- rep("FALSE", length(avg.ook.cells$Pfa))
avg.ook.cells[which(rownames(avg.ook.cells) %in% rownames(subss2)), ]$species.spec <- "TRUE"
ggplot(avg.ook.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = species.spec),
    size = 0.8) + labs(colour = "pf.spec")
```



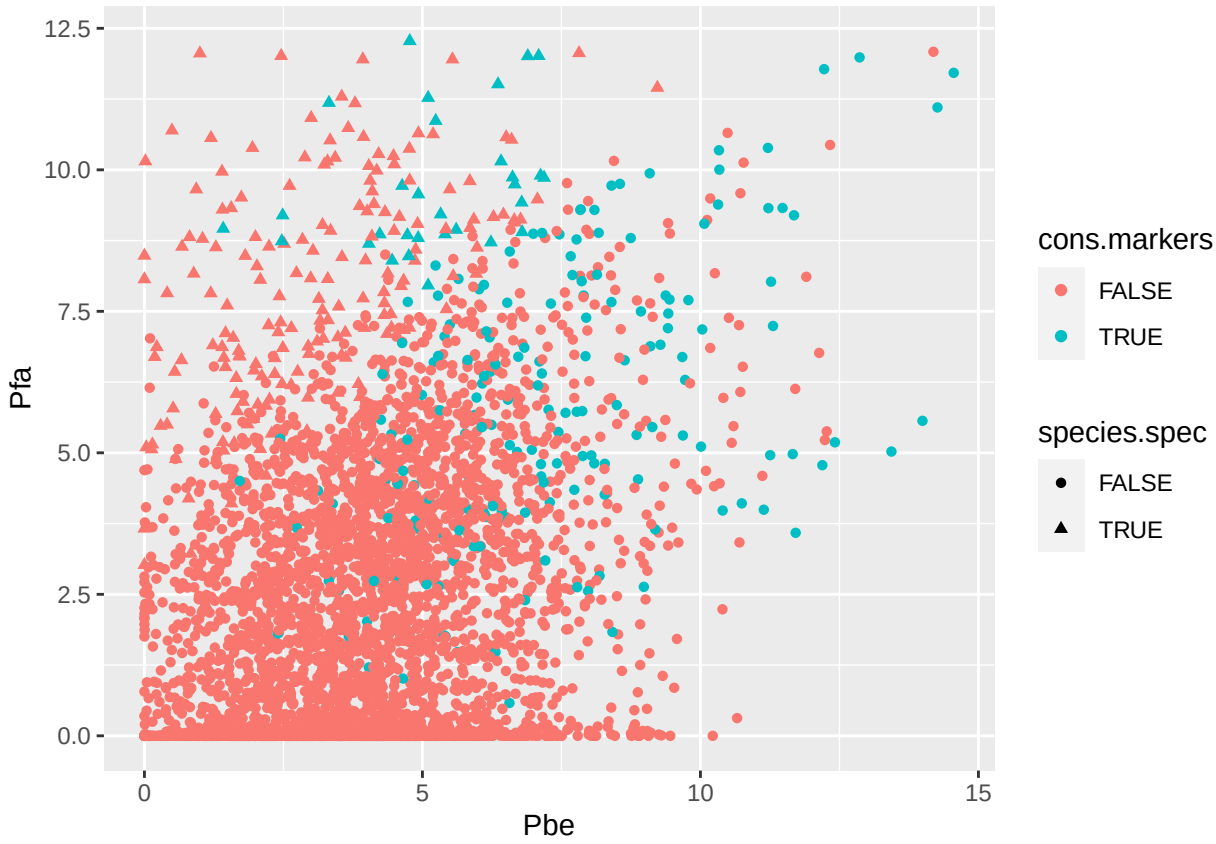

```
FeaturePlot(p.combined, features = c("ORTH-1643", "ORTH-1743", "ORTH-2564", "ORTH-883"),
    min.cutoff = "q9", reduction = "umap")
```




```
head(ook.markers)
```

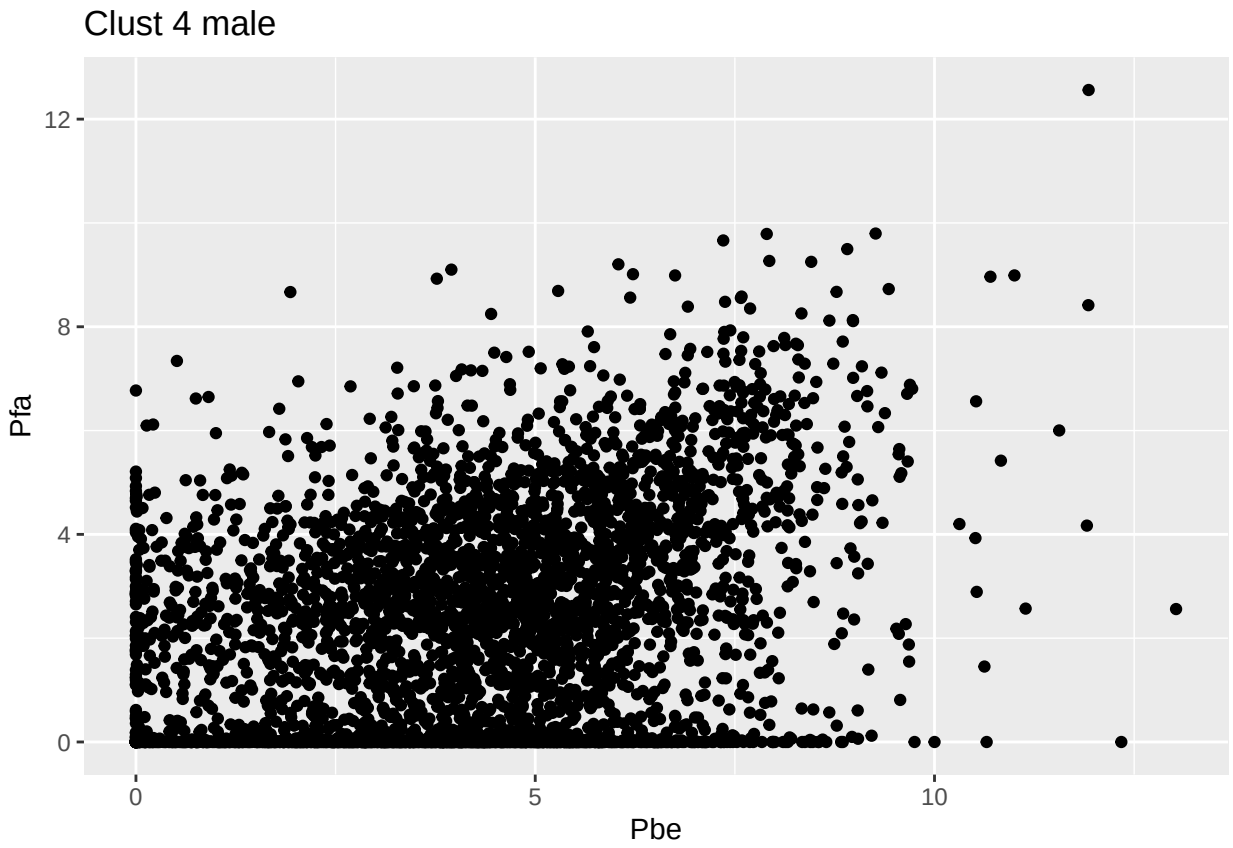
```
##              Pbe_p_val Pbe_avg_logFC Pbe_pct.1 Pbe_pct.2 Pbe_p_val_adj
## ORTH-1886 2.570544e-90      6.715708     1.000     0.100  1.030274e-86
## ORTH-3864 2.739265e-81      7.574164     1.000     0.162  1.097897e-77
## ORTH-3594 2.698085e-63      5.938248     0.922     0.162  1.081393e-59
## ORTH-3352 3.728015e-68      8.581700     0.971     0.238  1.494188e-64
## ORTH-1420 7.697858e-29      2.276606     0.786     0.262  3.085301e-25
## ORTH-3772 4.714039e-73      8.242011     1.000     0.221  1.889387e-69
##               Pfa_p_val Pfa_avg_logFC Pfa_pct.1 Pfa_pct.2 Pfa_p_val_adj
## ORTH-1886 1.782472e-249      7.614759     1.000     0.004 7.144149e-246
## ORTH-3864 5.692443e-240     10.068608     0.991     0.007 2.281531e-236
## ORTH-3594 5.833132e-236      6.880828     0.929     0.001 2.337919e-232
## ORTH-3352 9.612408e-236     11.691013     0.982     0.008 3.852653e-232
## ORTH-1420 1.521611e-229      5.848942     0.929     0.004 6.098615e-226
## ORTH-3772 1.014920e-225     11.985702     0.884     0.000 4.067801e-222
##               max_pval minimump_p_val         pbgene        pfgene
## ORTH-1886 2.570544e-90  3.564944e-249 PBANKA_1218100 PF3D7_0320400
## ORTH-3864 2.739265e-81  1.138489e-239 PBANKA_0408200 PF3D7_0310100
## ORTH-3594 2.698085e-63  1.166626e-235 PBANKA_1124700 PF3D7_0625900
## ORTH-3352 3.728015e-68  1.922482e-235 PBANKA_1037800 PF3D7_1404300
## ORTH-1420 7.697858e-29  3.043221e-229 PBANKA_1345700 PF3D7_1330700
## ORTH-3772 4.714039e-73  2.029841e-225 PBANKA_1143700 PF3D7_1367800
```

```
avg.ook.cells$cons.markers <- rep("FALSE", length(avg.ook.cells$Pfa))
avg.ook.cells[which(rownames(avg.ook.cells) %in% rownames(ook.markers)), ]$cons.markers <- "TRUE"

ggplot(avg.ook.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = cons.markers,
    shape = species.spec))
```



Males

```
Idents(p.combined) <- "celltype"
male.cells <- subset(p.combined, idents = "4")
Idents(male.cells) <- "species"
avg.male.cells <- log1p(AverageExpression(male.cells, verbose = FALSE)$RNA)
avg.male.cells$gene <- rownames(avg.male.cells)

ggplot(avg.male.cells, aes(Pbe, Pfa)) + geom_point() + ggtitle("Clust 4 male")
```




```
Idents(p.combined) <- "clust_species"
species.spec <- FindMarkers(p.combined, ident.1 = "4_Pfa", ident.2 = "4_Pbe", verbose = FALSE)
head(species.spec, n = 15)
```

```
##                  p_val  avg_logFC pct.1 pct.2    p_val_adj
## ORTH-3244 2.858468e-22   6.771585 0.939 0.000 1.145674e-18
## ORTH-466  1.381280e-21   1.249398 0.970 0.013 5.536170e-18
## ORTH-4219 2.128565e-20   4.915303 0.939 0.040 8.531287e-17
## ORTH-2929 3.759477e-19   3.271240 0.970 0.093 1.506798e-15
## ORTH-1231 1.240782e-18   3.939086 0.939 0.093 4.973056e-15
## ORTH-3266 2.484888e-18   4.625530 0.939 0.093 9.959431e-15
## ORTH-4259 3.164107e-18   3.119724 0.970 0.080 1.268174e-14
## ORTH-1895 3.389818e-18   6.734769 0.909 0.107 1.358639e-14
## ORTH-2714 3.554892e-18   2.785822 1.000 0.093 1.424801e-14
## ORTH-2823 1.191689e-17   5.150724 1.000 0.387 4.776291e-14
## ORTH-1911 2.030676e-17   3.799008 0.939 0.120 8.138951e-14
## ORTH-3171 5.749068e-17  -8.549652 0.000 1.000 2.304227e-13
## ORTH-1369 5.749068e-17 -10.000189 0.000 1.000 2.304227e-13
## ORTH-272  6.301273e-17  -8.552991 0.030 1.000 2.525550e-13
## ORTH-456  6.865298e-17  -8.570768 0.061 1.000 2.751611e-13
```

```
species.spec$pbgene <- newot[match(rownames(species.spec), newot[, 12]), 2]
species.spec$pfgene <- newot[match(rownames(species.spec), newot[, 12]), 4]

head(species.spec, n = 15)
```

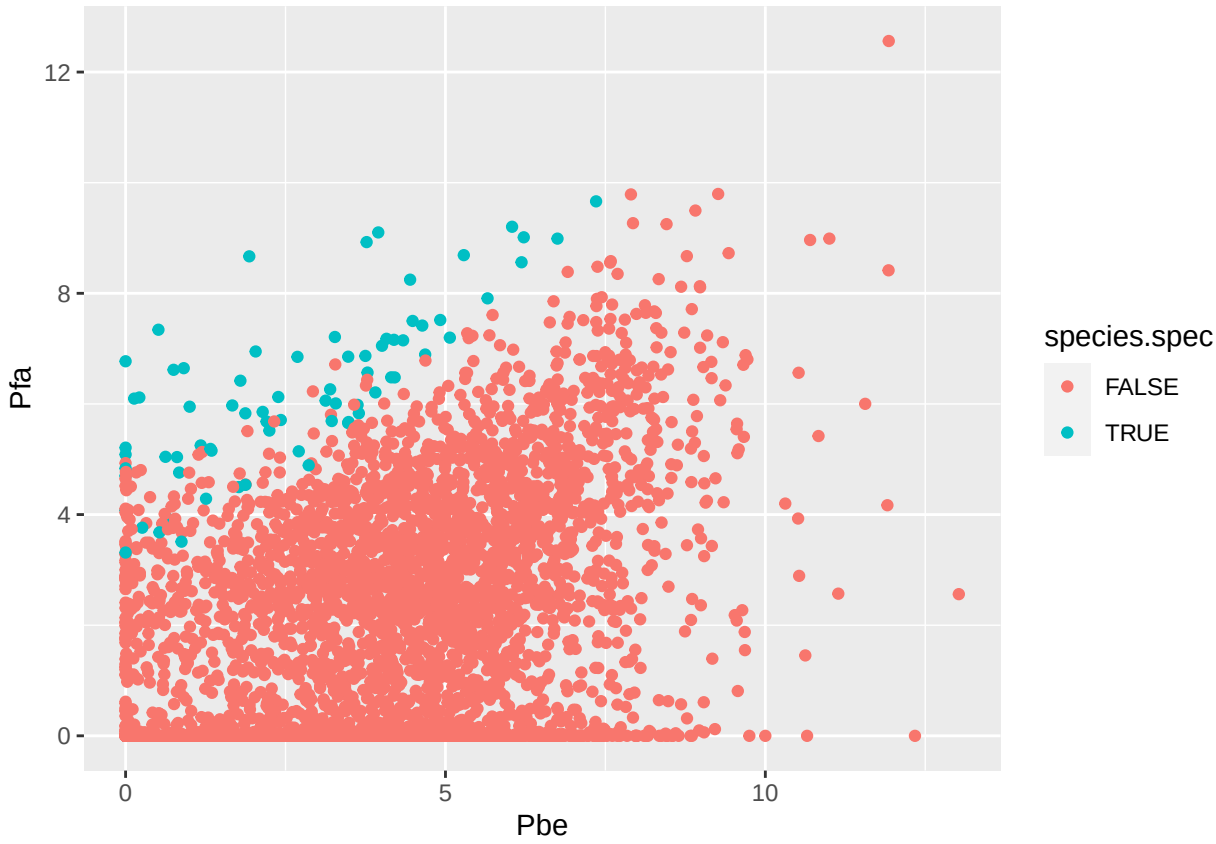
```
##                  p_val  avg_logFC pct.1 pct.2    p_val_adj         pbgene
## ORTH-3244 2.858468e-22   6.771585 0.939 0.000 1.145674e-18 PBANKA_1025200
```

```
## ORTH-466  1.381280e-21   1.249398 0.970 0.013 5.536170e-18 PBANKA_1450100
## ORTH-4219 2.128565e-20   4.915303 0.939 0.040 8.531287e-17 PBANKA_0209700
## ORTH-2929 3.759477e-19   3.271240 0.970 0.093 1.506798e-15 PBANKA_0826000
## ORTH-1231 1.240782e-18   3.939086 0.939 0.093 4.973056e-15 PBANKA_1324800
## ORTH-3266 2.484888e-18   4.625530 0.939 0.093 9.959431e-15 PBANKA_1028100
## ORTH-4259 3.164107e-18   3.119724 0.970 0.080 1.268174e-14 PBANKA_0213800
## ORTH-1895 3.389818e-18   6.734769 0.909 0.107 1.358639e-14 PBANKA_1219000
## ORTH-2714 3.554892e-18   2.785822 1.000 0.093 1.424801e-14 PBANKA_0803600
## ORTH-2823 1.191689e-17   5.150724 1.000 0.387 4.776291e-14 PBANKA_0815100
## ORTH-1911 2.030676e-17   3.799008 0.939 0.120 8.138951e-14 PBANKA_1220800
## ORTH-3171 5.749068e-17  -8.549652 0.000 1.000 2.304227e-13 PBANKA_1016700
## ORTH-1369 5.749068e-17 -10.000189 0.000 1.000 2.304227e-13 PBANKA_1340100
## ORTH-272  6.301273e-17  -8.552991 0.030 1.000 2.525550e-13 PBANKA_1429400
## ORTH-456  6.865298e-17  -8.570768 0.061 1.000 2.751611e-13 PBANKA_1449000
##                  pfgene
## ORTH-3244 PF3D7_1417500
## ORTH-466  PF3D7_1235500
## ORTH-4219 PF3D7_0103400
## ORTH-2929 PF3D7_0925200
## ORTH-1231 PF3D7_1461100
## ORTH-3266 PF3D7_1414600
## ORTH-4259 PF3D7_0729700
## ORTH-1895 PF3D7_0708500
## ORTH-2714 PF3D7_0706000
## ORTH-2823 PF3D7_0914100
## ORTH-1911 PF3D7_0710200
## ORTH-3171 PF3D7_1427900
## ORTH-1369 PF3D7_1324900
## ORTH-272  PF3D7_1213600
## ORTH-456  PF3D7_1234400
```

```
male.markers$pbgene <- newot[match(rownames(male.markers), newot[, 12]), 2]
male.markers$pfgene <- newot[match(rownames(male.markers), newot[, 12]), 4]

subss <- species.spec[species.spec$p_val_adj < 0.001, ]
subss2 <- subss[subss$avg_logFC > 2, ]

# write.csv(male.markers, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/male_markers.csv')
# write.csv(subss, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/male_species_specific.csv')
# write.csv(subss2, file =
# '/Users/vh3/Documents/PfMCA/ANALYSIS_2/pbpf/markers/male_pf_specific.csv')
```

```
avg.male.cells$species.spec <- rep("FALSE", length(avg.male.cells$Pfa))
avg.male.cells[which(rownames(avg.male.cells) %in% rownames(subss)), ]$species.spec <- "TRUE"
ggplot(avg.male.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = species.spec))
```


```
avg.male.cells$species.spec <- rep("FALSE", length(avg.male.cells$Pfa))
avg.male.cells[which(rownames(avg.male.cells) %in% rownames(subss2)), ]$species.spec <- "TRUE"
ggplot(avg.male.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = species.spec))
```




```
head(male.markers)
```

```
##               Pbe_p_val Pbe_avg_logFC Pbe_pct.1 Pbe_pct.2 Pbe_p_val_adj
## ORTH-1074  3.344615e-94     4.2150774     0.987     0.047  1.340522e-90
## ORTH-3921  1.399571e-76     6.7757226     0.987     0.117  5.609481e-73
## ORTH-2900  1.000801e-41     1.1562713     0.947     0.261  4.011209e-38
## ORTH-1862  3.428901e-12     0.8611099     0.600     0.224  1.374303e-08
## ORTH-291  3.464003e-115     7.9173867     1.000     0.012 1.388372e-111
## ORTH-3918  3.019281e-54     4.5638147     0.960     0.218  1.210128e-50
##               Pfa_p_val Pfa_avg_logFC Pfa_pct.1 Pfa_pct.2 Pfa_p_val_adj
## ORTH-1074 6.383493e-219      6.846696     0.848     0.000 2.558504e-215
## ORTH-3921 6.352508e-210      7.201473     1.000     0.007 2.546085e-206
## ORTH-2900 3.765386e-207      6.552619     0.970     0.006 1.509167e-203
## ORTH-1862 1.064799e-204      6.319949     1.000     0.008 4.267715e-201
## ORTH-291  8.749907e-195      1.638727     0.909     0.005 3.506963e-191
## ORTH-3918 1.180799e-193      5.649033     0.879     0.004 4.732642e-190
##                max_pval minimump_p_val         pbgene        pfgene
## ORTH-1074  3.344615e-94  1.276699e-218 PBANKA_1308100 PF3D7_1444200
## ORTH-3921  1.399571e-76  1.270502e-209 PBANKA_0416100 PF3D7_0905300
## ORTH-2900  1.000801e-41  7.530771e-207 PBANKA_0822900 PF3D7_0922000
## ORTH-1862  3.428901e-12  2.129598e-204 PBANKA_1215700 PF3D7_0322800
## ORTH-291  3.464003e-115  1.749981e-194 PBANKA_1431400 PF3D7_1215700
## ORTH-3918  3.019281e-54  2.361598e-193 PBANKA_0415800 PF3D7_0905600
```

```
avg.male.cells$cons.markers <- rep("FALSE", length(avg.male.cells$Pfa))
avg.male.cells[which(rownames(avg.male.cells) %in% rownames(male.markers)), ]$cons.markers <- "TRUE"

ggplot(avg.male.cells, aes(x = Pbe, y = Pfa)) + geom_point(aes(colour = cons.markers,
```

```
    shape = species.spec))
```


```
#saveRDS(p.combined, file=("pbpfcombo_SeuratInt_20200702.rds"))
```