Barnyard dataset analysis workflow

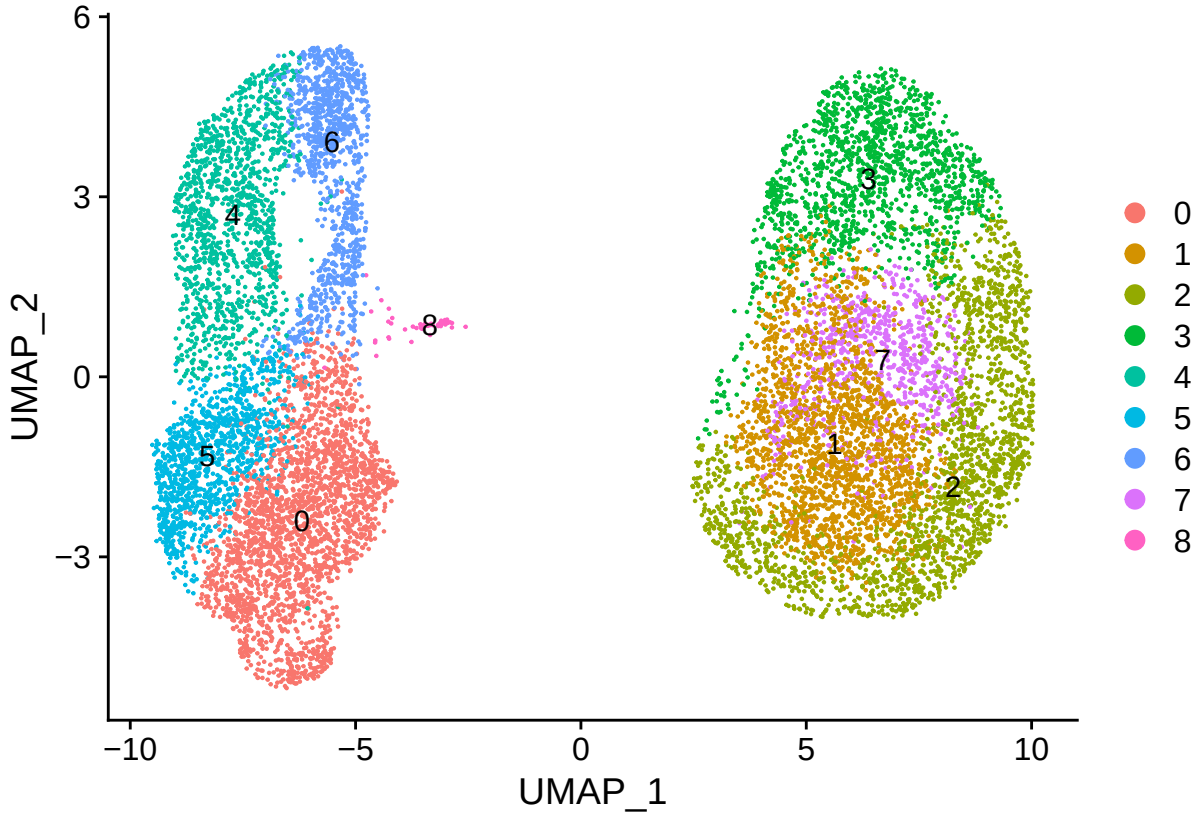
2024-04-25

Data preprocessing

Read in the scRNA-Seq data of human mouse mix dataset (obtained using scripts provided by DecontX): https://github.com/campbio/Manuscripts/tree/master/DecontX

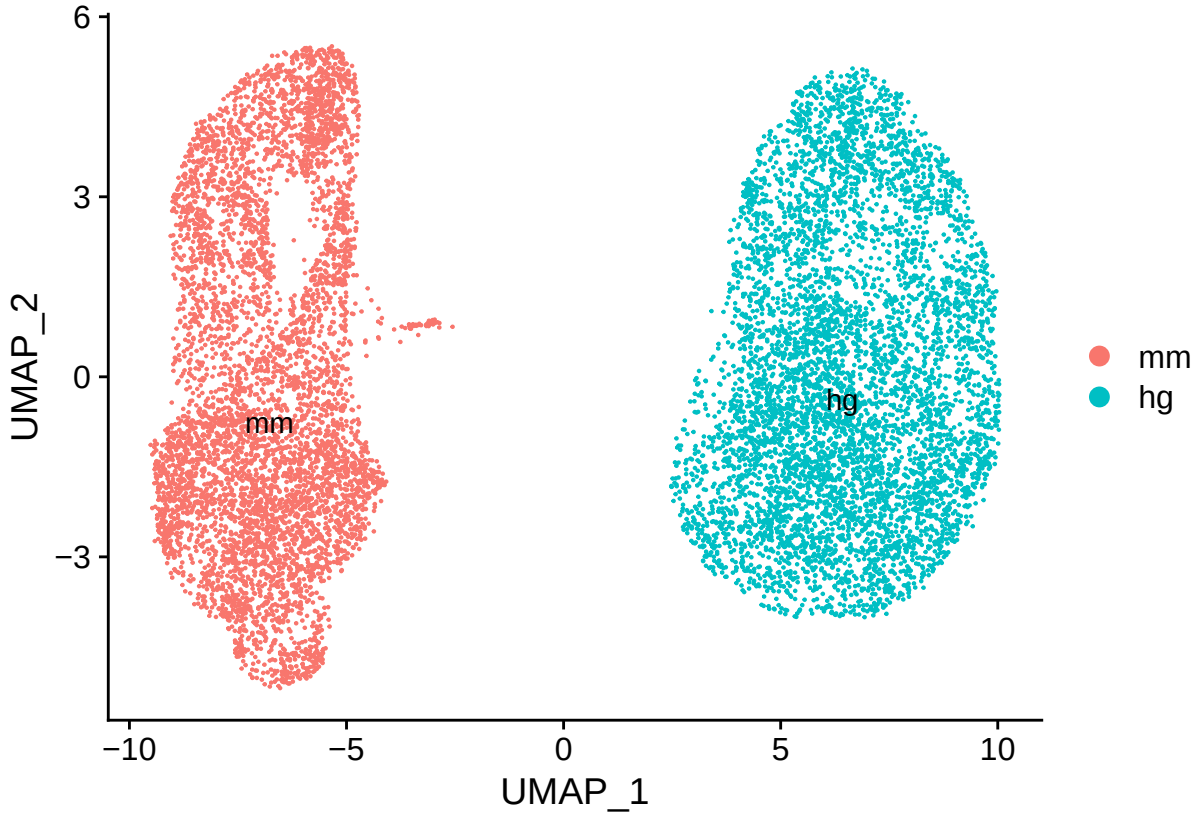
```
library(Seurat)
library(scCDC)
data <- readRDS("decontx_human_mouse_mix/Human_Mouse_Mix/specific_cells/specificCellsWithLabel.rds")
hm_mix <- data$count
```

```
hm_mix <- CreateSeuratObject(hm_mix)
hm_mix <- NormalizeData(hm_mix)
hm_mix <- FindVariableFeatures(hm_mix)
hm_mix <- ScaleData(hm_mix,rownames(hm_mix))
hm_mix <- RunPCA(hm_mix,dims=1:20,verbose = F)
hm_mix <- FindNeighbors(hm_mix,dims=1:20,verbose = F)
hm_mix <- FindClusters(hm_mix,resolution = 0.5,verbose = F)
hm_mix <- RunUMAP(hm_mix,dims = 1:20)
DimPlot(hm_mix,label = T)
```



The two big clusters of cells represents cells from human and mouse. We group small clusters within the big clusters for downstream analysis.

```
new.idents <- c('mm','hg','hg','hg','mm','mm','mm','hg','mm')
names(new.idents) <- levels(hm_mix)
hm_mix <- RenameIdents(hm_mix,new.idents)
DimPlot(hm_mix,label = T)
```



Contamination Detection and correction

Check the contamination in the data

We then apply scCDC to detect GCGs and decontaminate the data.

```
GCGs = ContaminationDetection(hm_mix)
```

```
## Caculating entropy...

## Caculating expression level...

## Calculating entropy-expression relation...

## Extracting contamination degree...

## Complete detection. 290 contaminated genes found
```

```
hm_mix_corrected = ContaminationCorrection(hm_mix,rownames(GCGs))
```

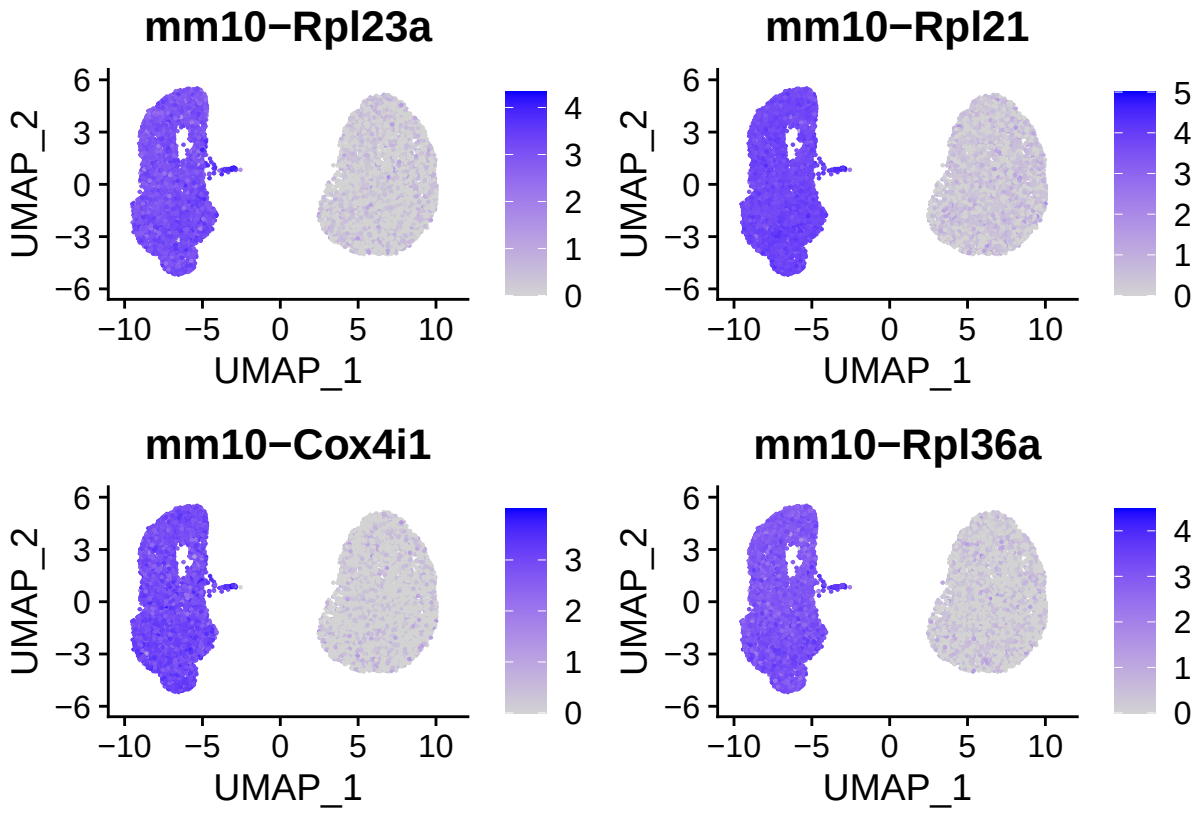
```
## Calculating correction threshold...
```

```
## Decontaminating...
```

```
DefaultAssay(hm_mix_corrected) = "Corrected"
```

290 GCGs was found in this dataset and scCDC was applied for decontamination. We select 4 GCGs as an example for visualizing the contamination pattern in the data.

```
FeaturePlot(hm_mix,rownames(GCGs)[9:12])
```



The selected four GCGs are mouse genes but were detected in both mouse and human cells, which indicates there's contamination in the data. Next, we repeat the seurat pipeline for the decontaminated data.

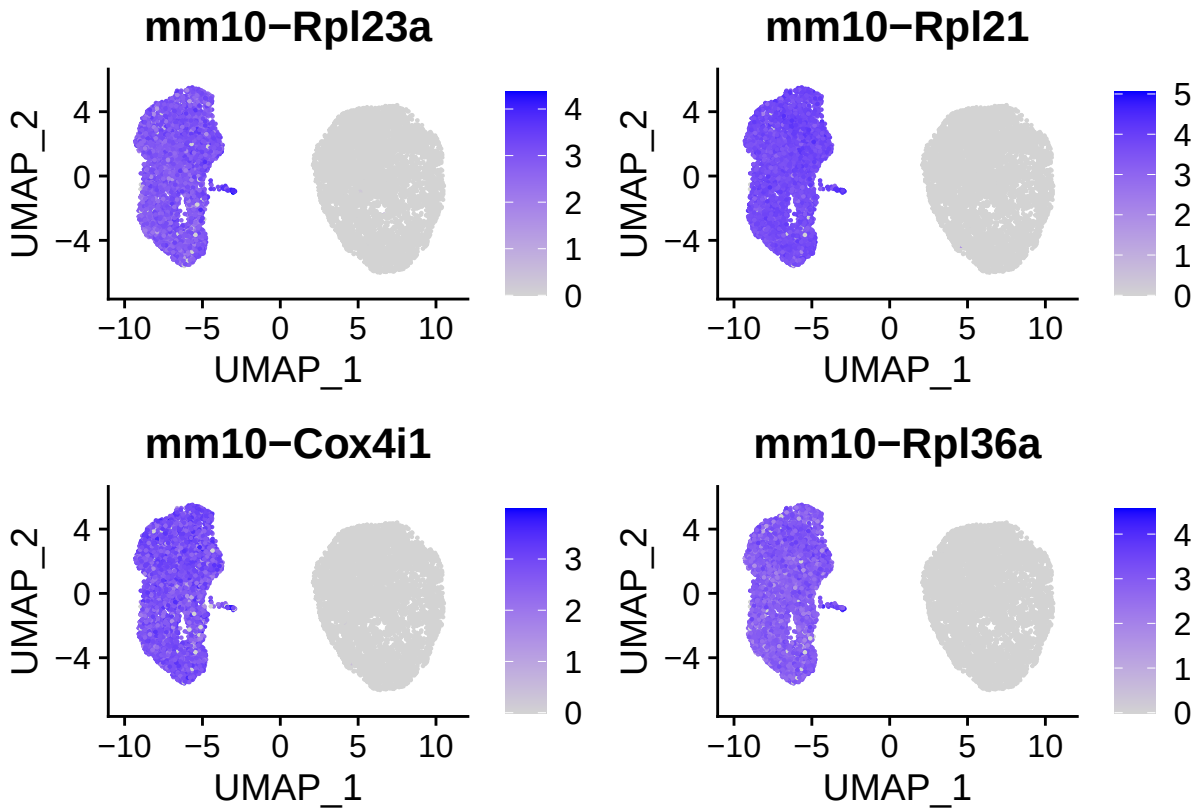
Check the decontamination result

```
hm_mix_corrected <- NormalizeData(hm_mix_corrected,
                                  normalization.method = "LogNormalize",scale.factor = 10000)
hm_mix_corrected <- FindVariableFeatures(hm_mix_corrected,
                                         selection.method = "vst", nfeatures = 2000)
all.genes <- rownames(hm_mix_corrected)
hm_mix_corrected <- ScaleData(hm_mix_corrected,
                              features = all.genes)
hm_mix_corrected <- RunPCA(hm_mix_corrected,
                           features = VariableFeatures(object = hm_mix_corrected),verbose = F)
hm_mix_corrected <- FindNeighbors(hm_mix_corrected, dims = 1:15)
```

```
hm_mix_corrected <- FindClusters(hm_mix_corrected, resolution = 0.2,verbose = F)
hm_mix_corrected <- RunUMAP(hm_mix_corrected,dims=1:15)
```

And we want to check the decontamination result, we still use the 4 GCGs as an example.

```
FeaturePlot(hm_mix_corrected,rownames(GCGs)[9:12])
```



The expression of mouse genes are only detected in mouse cells after decontamination.