The number of genes is as expected for two species. There're way more cells than we expect, which is about 1000. So what's going on?

How many UMIs per barcode?

```
tot_counts <- Matrix::colSums(res_mat)
summary(tot_counts)
```

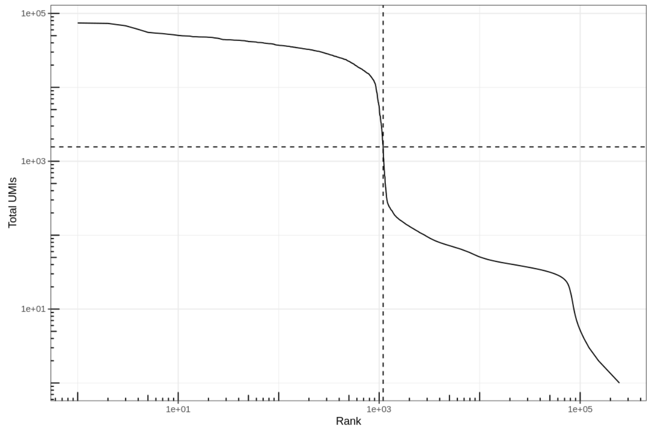
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
    0.00     1.00     1.00    76.04     8.00 74292.00 
```

The vast majority of "cells" have only a few UMI detected. Using the barcode ranking method, we are going to estimate the amount of empty droplets. There is an inflection point in the data, and we can use this point to see the transition between two components in a distribution.

```
# Compute barcode rank
bc_rank <- barcodeRanks(res_mat)
```

```
#' Knee plot for filtering empty droplets
#' 
#' Visualizes the inflection point to filter empty droplets. This function plots 
#' different datasets with a different color. Facets can be added after calling
#' this function with `facet_*` functions.
#' 
#' @param bc_rank A `DataFrame` output from `DropletUtil::barcodeRanks`.
#' @return A ggplot2 object.
knee_plot <- function(bc_rank) {
  knee_plt <- tibble(rank = bc_rank[["rank"]],
                     total = bc_rank[["total"]]) %>% 
    distinct() %>% 
    dplyr::filter(total > 0)
  annot <- tibble(inflection = metadata(bc_rank)[["inflection"]],
                  rank_cutoff = max(bc_rank$rank[bc_rank$total > metadata(bc_rank)[["inflection"]]]))
  p <- ggplot(knee_plt, aes(rank, total)) +
    geom_line() +
    geom_hline(aes(yintercept = inflection), data = annot, linetype = 2) +
    geom_vline(aes(xintercept = rank_cutoff), data = annot, linetype = 2) +
    scale_x_log10() +
    scale_y_log10() +
    annotation_logticks() +
    labs(x = "Rank", y = "Total UMIs")
  return(p)
}
```

```
options(repr.plot.width=9, repr.plot.height=6)
knee_plot(bc_rank)
```



The inflection point looks like a reasonable number of cells.

```
# Filter the matrix
res_mat <- res_mat[, tot_counts > metadata(bc_rank)$inflection]
res_mat <- res_mat[Matrix::rowSums(res_mat) > 0,]
dim(res_mat)
```

```
37755 · 1095
```

We get that around 1095 cells are real from the inflection point analysis. This is reasonable because the cell capture rate is claimed to be around 65%, but often can be lower.

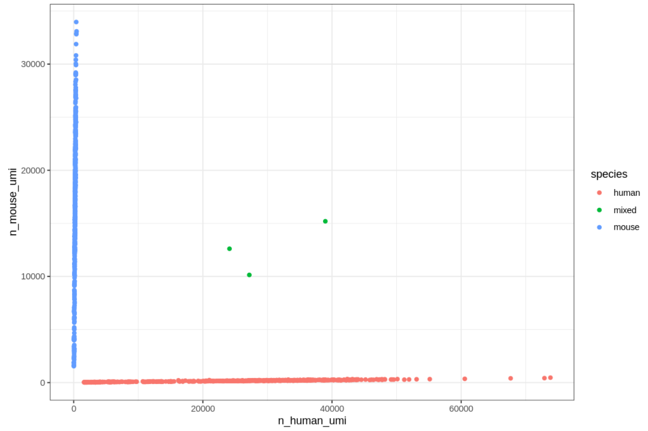
## Cell species

How many cells are from humans and how many from mice? The number of cells with mixed species indicates doublet rate.

```
gene_species <- ifelse(str_detect(rownames(res_mat), "^ENSMUSG"), "mouse", "human")
mouse_inds <- gene_species == "mouse"
human_inds <- gene_species == "human"
# mark cells as mouse or human
cell_species <- tibble(n_mouse_umi = Matrix::colSums(res_mat[mouse_inds,]),
                       n_human_umi = Matrix::colSums(res_mat[human_inds,]),
                       tot_umi = Matrix::colSums(res_mat),
                       prop_mouse = n_mouse_umi / tot_umi,
                       prop_human = n_human_umi / tot_umi)

# Classify species based on proportion of UMI, with cutoff of 90%
cell_species <- cell_species %>%
  mutate(species = case_when(
    prop_mouse > 0.9 ~ "mouse",
    prop_human > 0.9 ~ "human",
    TRUE ~ "mixed"
  ))

ggplot(cell_species, aes(n_human_umi, n_mouse_umi, color = species)) +
  geom_point()
```



We can see clearly here that not many of the cells are doublets. The doublets are represented by the green points in the graph, which are much less in quantity than the UMIs for explicitly mice and humans. There's a clear separation between the UMIs for the mice and those for the humans. The cutoff chosen was 90%, which essentially filters out the doublets, which are less than either 90% mouse or human.

```
cell_species %>%
  dplyr::count(species) %>%
  mutate(proportion = n / ncol(res_mat))
```

A tibble: 3 × 3

| species | n | proportion |
|---|---|---|
| <chr> | <int> | <dbl> |
| human | 566 | 0.516894977 |
| mixed | 3 | 0.002739726 |
| mouse | 526 | 0.480365297 |

This data tells us that only around 0.3% of cells are doublets. Note that this is only accounting for mixed cells; it is possible to have a doublet with data from the same species.

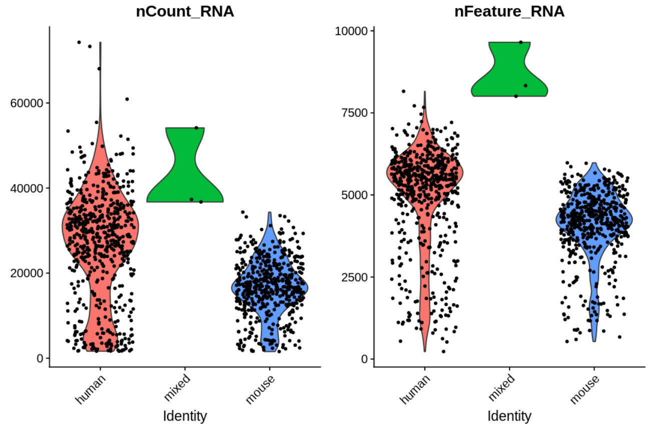
## Dimension reduction

```
seu <- CreateSeuratObject(res_mat, min.cells = 3) %>%
  SCTransform(verbose = FALSE)

# Add species to meta data
seu <- AddMetaData(seu, metadata = cell_species$species, col.name = "species")
```

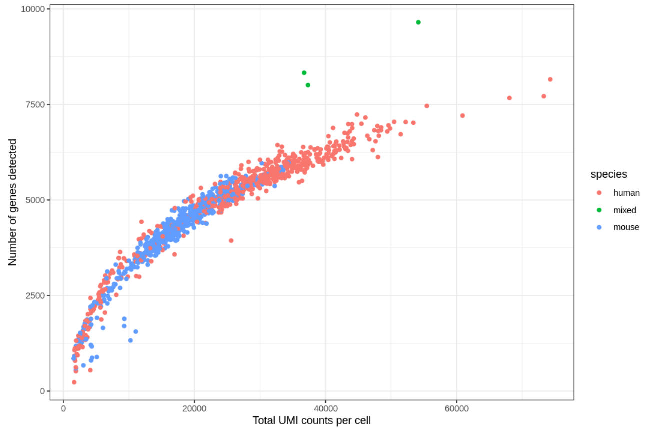
See how number of total counts and number of genes expressed are distributed.

```
VlnPlot(seu, c("nCount_RNA", "nFeature_RNA"), group.by = "species")
```



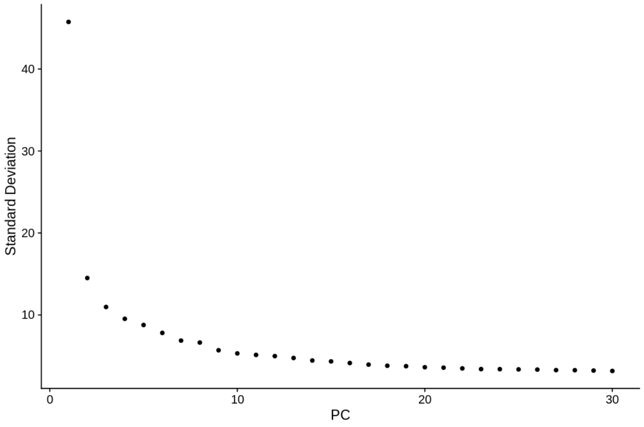
Another QC plot

```
ggplot(seu@meta.data, aes(nCount_RNA, nFeature_RNA, color = species)) +
  geom_point() +
  labs(x = "Total UMI counts per cell", y = "Number of genes detected")
```

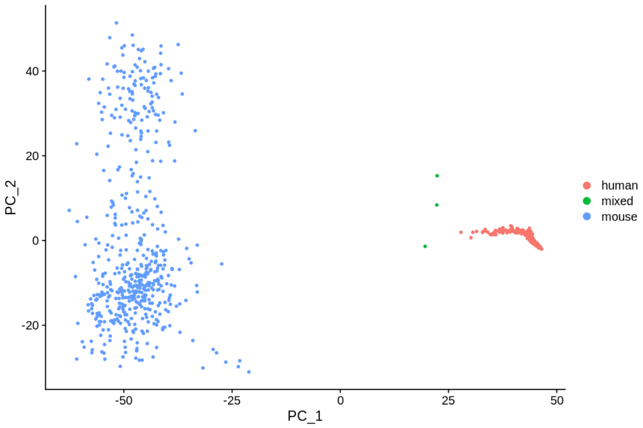


Here, the number of total counts per cell are plotted on the x axis and fraction of human/mouse transcript counts are on the y axis. There is no clean cutoff separating them, but there tends to be more cells classified as human for higher UMI counts per cell. The Internet says that the mouse genome is around 14% shorter than the human genome, which may explain why this plot shows that humans have on average, more genes detected per cell than mice do. If we look above to the plot above this, we can see more visually the distribution of total number of counts and number of genes are distributed.

```
seu <- RunPCA(seu, verbose = FALSE, npcs = 30)
ElbowPlot(seu, ndims = 30)
```

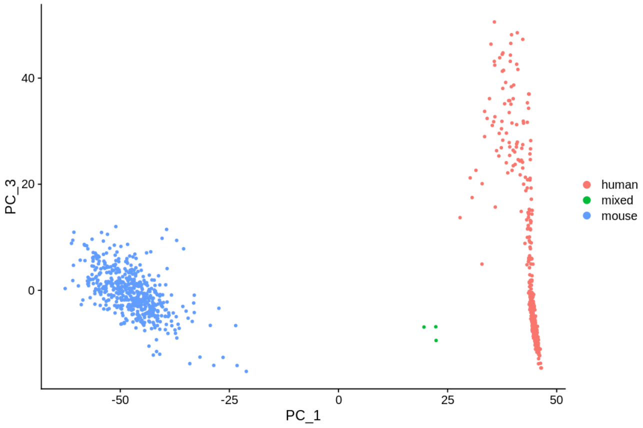


```
PCAPlot(seu, group.by = "species")
```

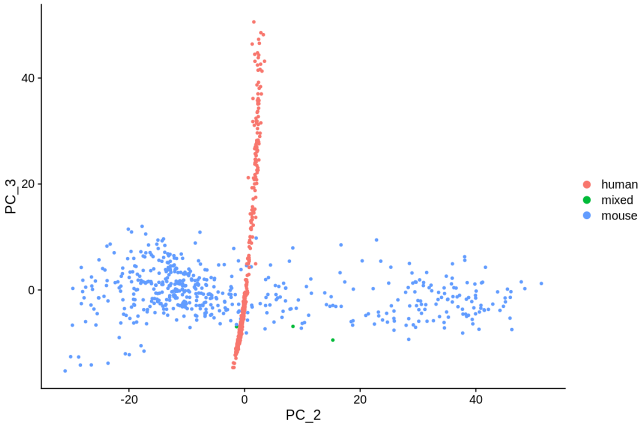


The first PC separates species, as expected. Also as expected, the doublets are in between human and mouse cells in this plot.

```
PCAPlot(seu, group.by = "species", dims=c(1, 3))
```



```
PCAPlot(seu, group.by = "species", dims=c(2, 3))
```



It is important to plot more principal components than just PC1 and PC2 since there may be other sources of variance within the data. In the two PCA plots above, we can see that PC1 vs PC3 and PC2 vs PC3 show differing representations, especially for PC2 vs. PC3, which has the human data within the mass of mouse data. There is another source of variance within the data, though the greatest source is probably shown in the PC1 vs PC2 graph, where the fact that mice and humans are different species contributes to much of variance.