Tidy gene expression and heatmap

```
library(tidyverse)
library(stringr)
```

Load it up

```
fpkm = read_csv("../datasets/stratton_fpkm.csv")
glimpse(fpkm)
```

Clean up

Let's get rid of X9 and X13, these are averages

```
fpkm = fpkm %>% select(-X9, -X13)
```

and pull out the "Blood Aged" column, but we'll keep this in case we need it later

```
blood = fpkm %>% select(`Blood Aged`)
fpkm = fpkm %>% select(-`Blood Aged`)
```

Now we can get rid of the first row

```
fpkm = fpkm %>% slice(-1)
```

Getting variable information

Now we need to extract our variable information from the column names

```
col_names = colnames(fpkm)[-1]

# get the day variable
day = str_extract(col_names, "day_\\d+")

# now the age variable
age = str_extract(col_names, "young|aged")

# we'll need a sample_ID variable for later
sample_ID = paste0("S", 1:length(age))
```

We'll create a new dataframe to hold our sample information, including a sample ID variable

```
samples = data_frame(day, age, sample_ID)
```

now we rename our columns with the new sample ID, and renaming our gene column at the same time

```
colnames(fpkm) = c("gene", sample_ID)
```

Now for the grand finale, joining the tables, first gathering the sample columns in the fpkm table and converting the expression column to numeric

```
# first let's get rid of rows with NA's
fpkm = na.omit(fpkm)
fpkm_tidy = fpkm %>% gather(sample_ID, expression, -gene) %>% mutate(expression = as.numeric(expression)
final = left_join(samples, fpkm_tidy)
final
```

Tada!

Making a heatmap

```
library(pheatmap)
set.seed(7592)
```

We'll need a matrix of expression values and a dataframe of sample info

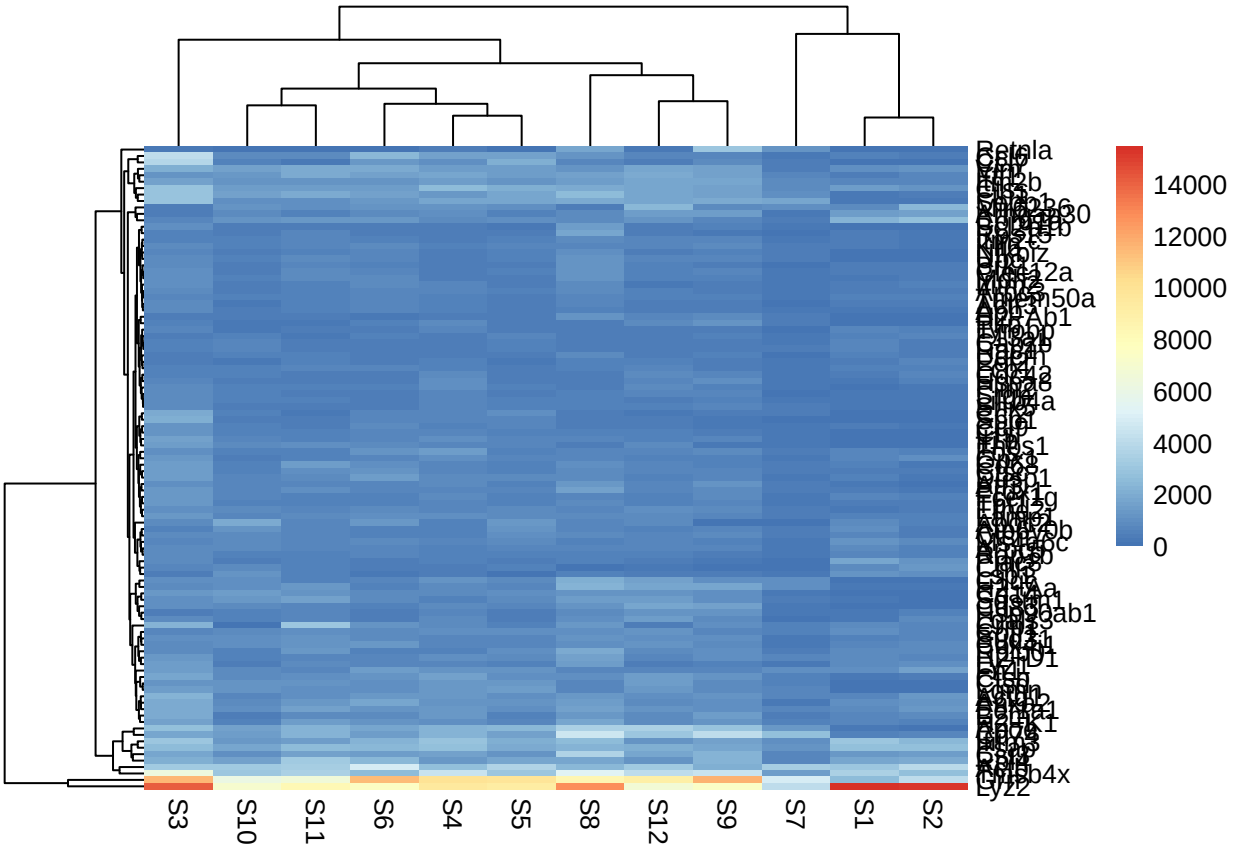
```
# filter out genes with all zeros
# get rid of duplicate genes which look to be an excel f*** up
# selct 100 top genes

fpkm_small = fpkm_tidy %>%
    group_by(gene) %>%
    filter(sum(expression) != 0) %>%
    group_by(sample_ID) %>%
    filter(!duplicated(gene)) %>%
    group_by(gene) %>% mutate(mean_exp = mean(expression)) %>%
    spread(sample_ID, expression) %>% ungroup() %>%
    top_n(100, mean_exp)


exp = as.matrix(fpkm_small[,-c(1:2)])
rownames(exp) = fpkm_small$gene
```

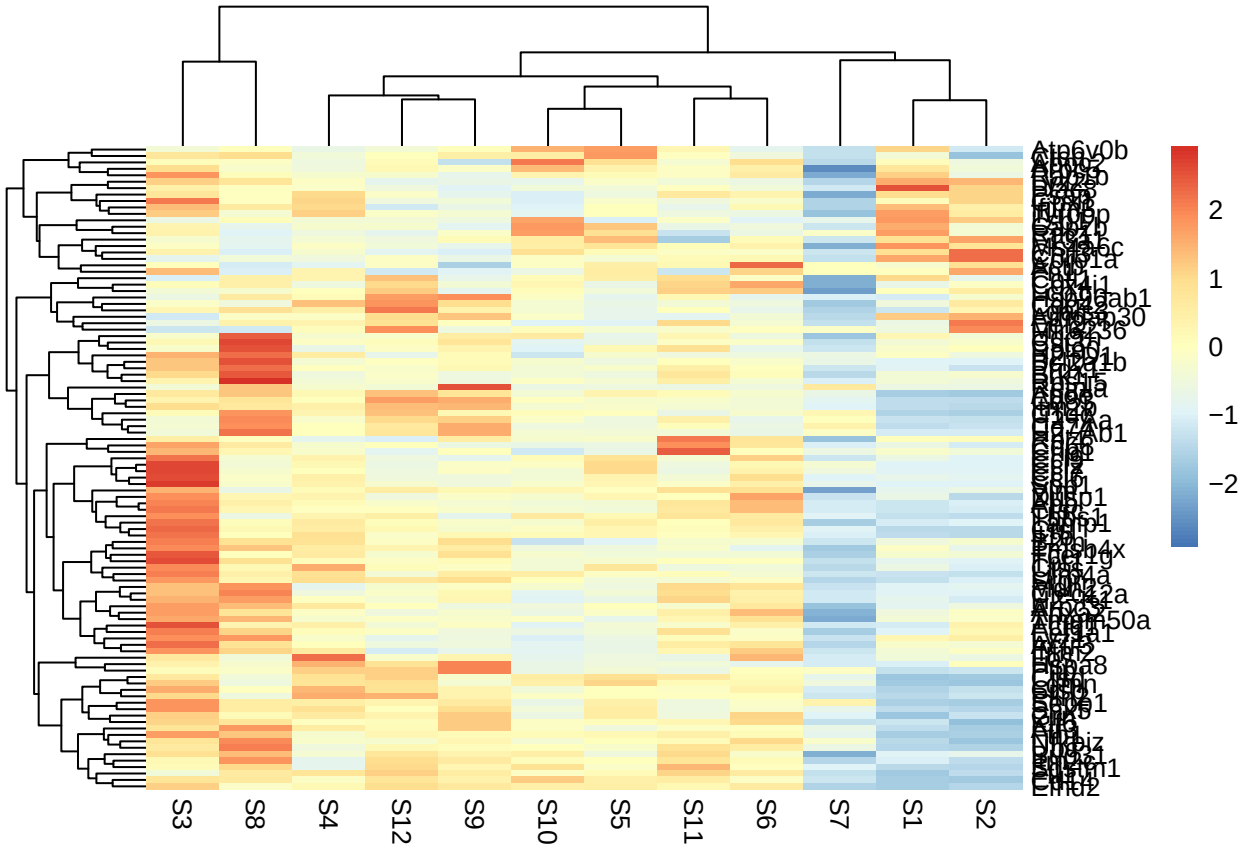
Ok, let's make a heatmap

```
pheatmap(exp)
```



Hmm, that's not right! Let's try scaling

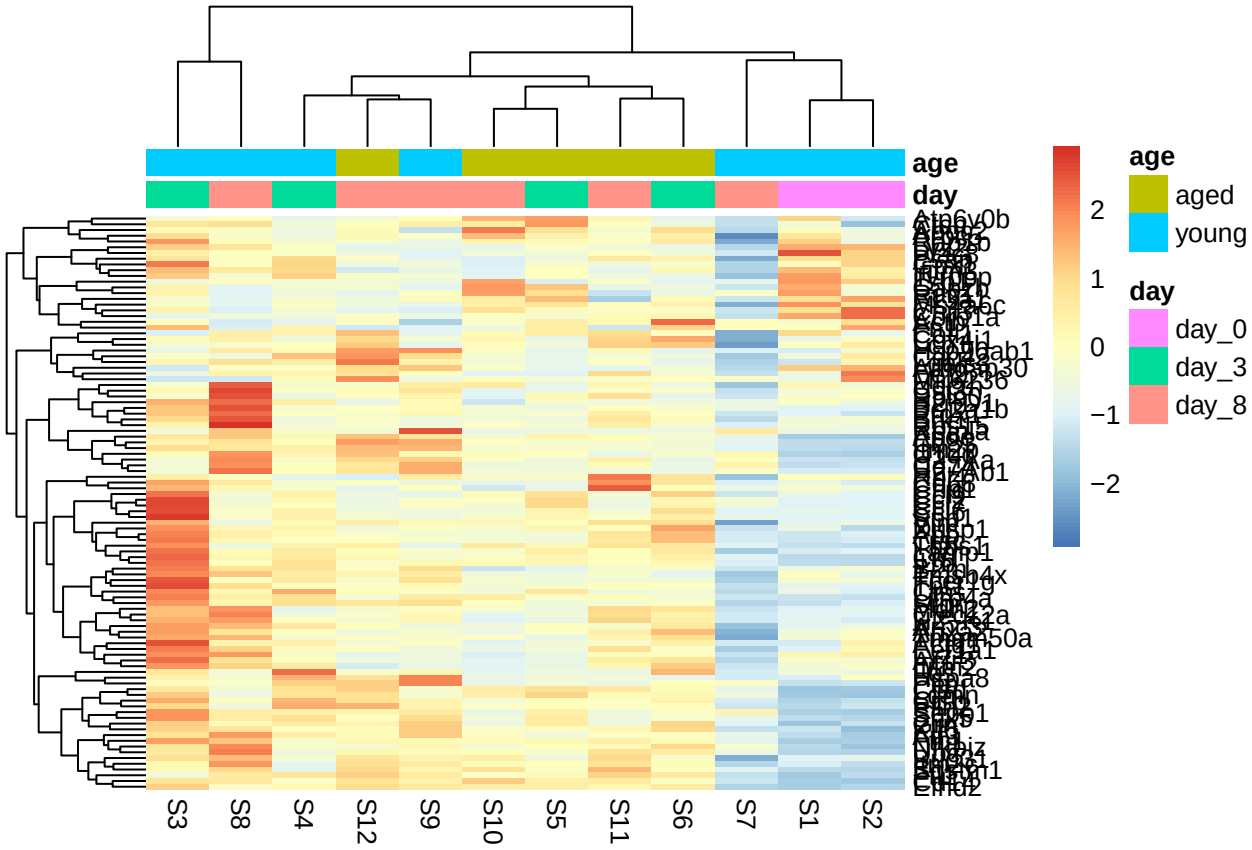
```
pheatmap(exp, scale = "row")
```



Add an annotation to the columns

```
col_anno = as.data.frame(samples[, -3])
rownames(col_anno) = samples$sample_ID

pheatmap(exp, annotation_col = col_anno, scale = "row")
```



`pheatmap` has loads of options. Read the help page and play around