RNA-seq Data

## Agenda

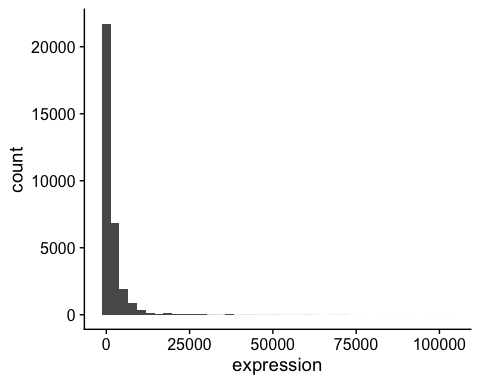
* Today’s lab
  + RNA-Seq data visualization - this lab borrows heavily from a [Carpentries incubator module on visualization](https://carpentries-incubator.github.io/bioc-intro/40-visualization.html).

## RNA-Seq data

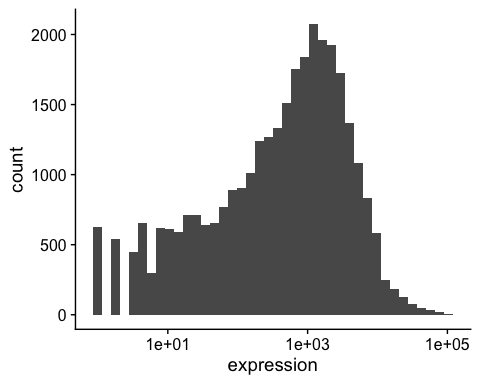
To start out with, fork the [starter code](https://classroom.github.com/a/Yz2GMIqU) and clone the repository to your local machine. Once we have done that, let’s explore the data provided:

# A tibble: 32,428 × 20  
 gene sample expression organism age sex infection strain time tissue  
 <chr> <chr> <dbl> <chr> <dbl> <chr> <chr> <chr> <dbl> <chr>   
 1 Asl GSM254… 1170 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
 2 Apod GSM254… 36194 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
 3 Cyp2d22 GSM254… 4060 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
 4 Klk6 GSM254… 287 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
 5 Fcrls GSM254… 85 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
 6 Slc2a4 GSM254… 782 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
 7 Exd2 GSM254… 1619 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
 8 Gjc2 GSM254… 288 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
 9 Plp1 GSM254… 43217 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
10 Gnb4 GSM254… 1071 Mus mus… 8 Fema… Influenz… C57BL… 8 Cereb…  
# ℹ 32,418 more rows  
# ℹ 10 more variables: mouse <dbl>, ENTREZID <dbl>, product <chr>,  
# ensembl\_gene\_id <chr>, external\_synonym <chr>, chromosome\_name <chr>,  
# gene\_biotype <chr>, phenotype\_description <chr>,  
# hsapiens\_homolog\_associated\_gene\_name <chr>, l2exp <dbl>

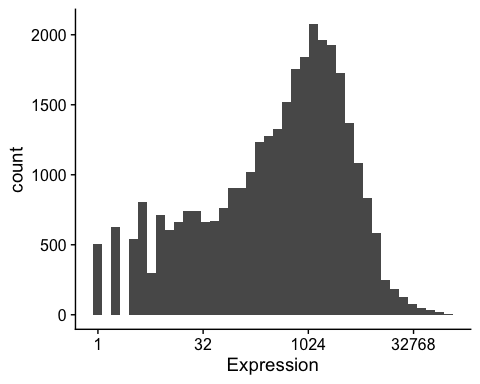
Take a look at the distribution of the expression levels:



The data should be transformed. Transforming on the scale is easy in ggplot2,

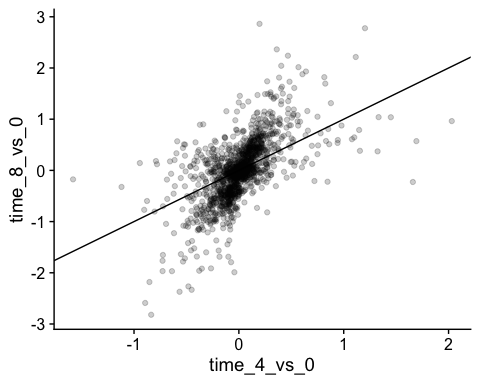


but we often use a scale when looking at expression data:

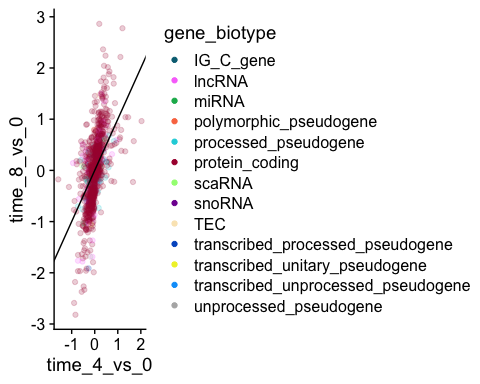


### Building plots iteratively

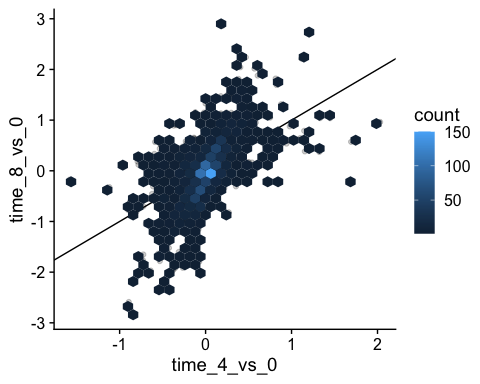
We will now draw a scatter plot with two continuous variables and the geom\_point() function. This graph will represent the fold changes of expression comparing time 8 versus time 0, and time 4 versus time 0. To this end, we first need to compute the means of the log-transformed expression values by gene and time, then the log fold changes by subtracting the mean log expressions between time 8 and time 0 and between time 4 and time 0.



We could also add color, but it begins to get a little messy with the large number of categories we have.

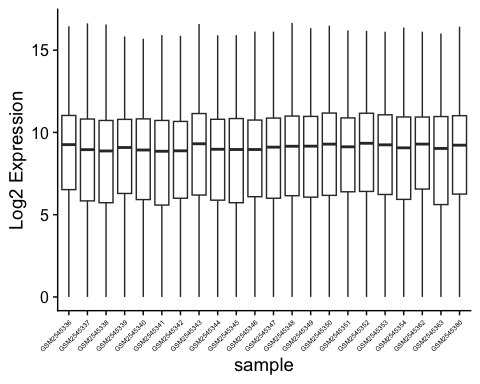


Over-plotting can be an issue with large datasets like this. One solution is using hexbin:geom\_hex().

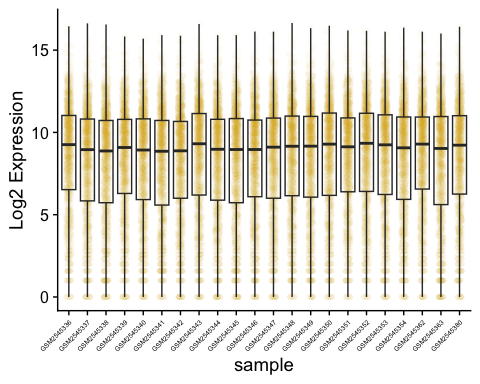


### Box plots

We can use boxplots to visualize the distribution of gene expressions within each sample:



Adding points to box plots can give us a better understanding of the underlying distributions.



We could also use geom\_violin to better see the distribution of points. Try coloring the points by time as well.

### Line plots

Let’s calculate the mean expression per duration of the infection for the 10 genes having the highest log fold changes comparing time 8 versus time 0. We can do this either by coloring each gene or using facet\_wrap.

Now we would like to split the line in each plot by the sex of the mice.

Let’s do something similar and create a plot that depicts how the average expression of each chromosome changes through the duration of infection.

The facet\_wrap geometry extracts plots into an arbitrary number of dimensions to allow them to cleanly fit on one page. On the other hand, the facet\_grid geometry allows you to explicitly specify how you want your plots to be arranged via formula notation (rows ~ columns; a . can be used as a placeholder that indicates only one row or column).

Let’s modify the previous plot to compare how the mean gene expression of males and females has changed through time: