Assignment 1

Gene expression data

We are going to use part of the data published by Blackmore et al. (2017), The effect of upper-respiratory infection on transcriptomic changes in the CNS. The goal of the study was to determine the effect of an upper-respiratory infection on changes in RNA transcription occuring in the cerebellum and spinal cord post infection.

The dataset is stored as a comma separated value (CSV) file. Each row holds information for a single RNA expression measurement, and the columns represent:

Column	Description
gene	The name of the gene that was measured
sample	The name of the sample the gene expression was measured in
expression	The value of the gene expression
organism	The organism/species - here all data stem from mice
age	The age of the mouse (all mice were 8 weeks here)
sex	The sex of the mouse
infection	The infection state of the mouse, i.e. infected with Influenza A or not infected.
strain	The Influenza A strain; C57BL/6 in all cases.
time	The duration of the infection (in days).
tissue	The tissue that was used for the gene expression experiment, i.e. cerebellum or spinal cord.
mouse	The mouse unique identifier.
ENTREZID	The gene ID for the ENTREZ database
product	The gene product
ensembl_gene_id	The ID of the gene from the ENSEMBL database
external_synonym	A name synonym for the gene
chromosome_name	The chromosome name of the gene
gene_biotype	The gene biotype
phenotype_description	The phenotype description of the gene
hsapiens_homolog_associated_gene_name	The human homologous gene

First, you need to load the data

```
data <- read.csv(file="Rnaseq2.csv")
```

Question 1) How many genes do you have in this dataset? Note, there might be multiple occurrences of a gene.

Question 2) How many genes have no phenotypic description associated with them? Note, "phenotype_description" should capture this information.

Question 3) What genes (gene names) of infected male mice are present on chromosome 16? You would need a combination of multiple columns to do this task.

Question 4) How many common genes are presents in all three time points?

Question 5) The log-transformation is widely used in biomedical research to deal with skewed data. Write a function that applies Log2(x + 1) on the gene expression values and append the transformed values as a column "log_exp" to the current data. Print the first 5 rows of the resulting data matrix.

Question 6) Plot the distribution of log-transformed expressions using histogram (bins=50).

Question 7) Having boxplot for samples will allow a quick overview for spotting irregularities (i.e. checking if the replicates within a sample-group show similar expression profiles). Use Boxplot to compare the distribution of log-transformed gene expression values (from Question 5) across multiple samples broken down based on sex. Note, you would need to use the following columns when plotting: "sample" and "sex".