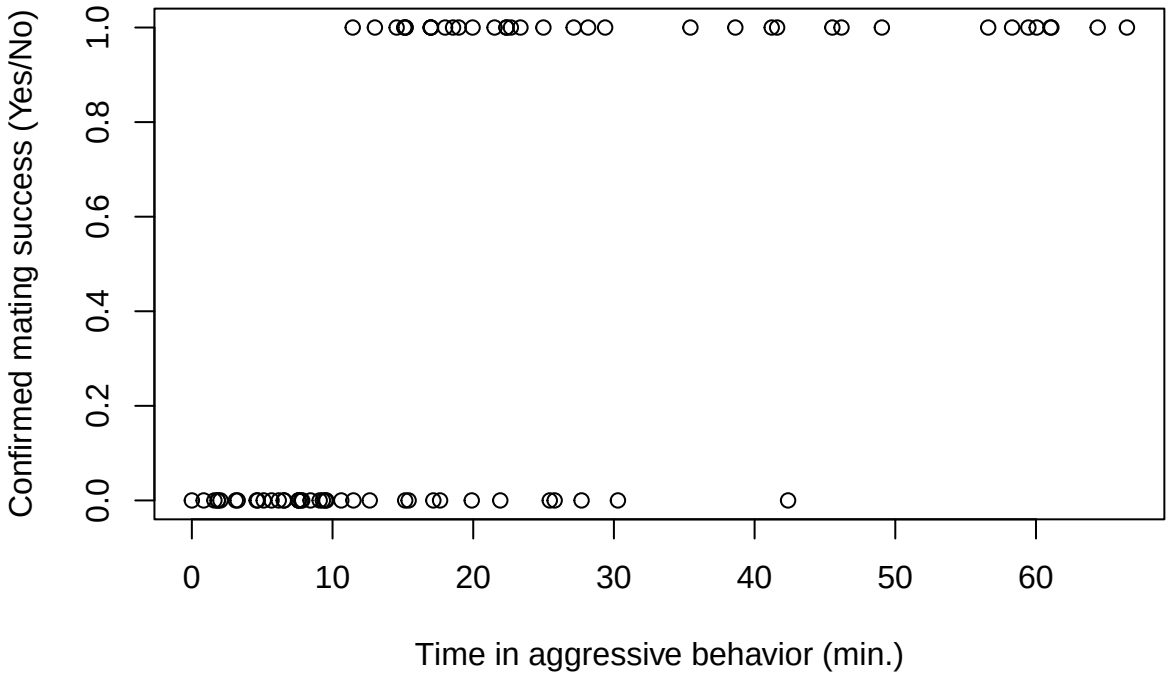
Assignment 5: brms model AND RNA-seq sample ordination // ABS (BI 610 Fall 2023)

Assignment: Your task is to use Rmarkdown to briefly address the questions below regarding two different data sets.

Please submit **both the .Rmd file *and* the resulting .html or .pdf (your choice) file.** For tips on how to write Rmarkdown reports, and an example, please see this page. You can work with other members of class, but I expect each of you to construct and run all of the code yourself. ***Submit to Canvas no later than Friday, Dec. 8th.***

Part 1: American pronghorn mating success study

Download and briefly examine the data set `pronghorn2023.tsv`. The objective of this behavioral study was to assess whether the amount of time male American pronghorn spend in aggressive interactions with other males during the breeding season has any bearing on whether they successfully mate. In the data set you will find the number of minutes (during a standard observation period) each male engaged in aggressive behavior (BATTLE), and whether each male was confirmed to have fathered at least one fawn (MATED: 0=No; 1=Yes). The researchers studied a few different pronghorn herds (HERD), and they are also interested in whether herd might generally contribute to variation in mating success. Answer the questions below in your own words (try not to use jargon) based on the code chunk used to fit a brms GLM. ***You don't need to write any code of your own, just look at the plot below and evaluate the following code.***




```
library(brms)

pronghorn <- read.delim("pronghorn2023.tsv")

brm_fit <- brm(MATED | trials(1) ~ BATTLE + (1|HERD),
               family='binomial',
               prior = prior(normal(0, 4), class = sd),
               iter = 4000, chains = 3,
               data=pronghorn)
```

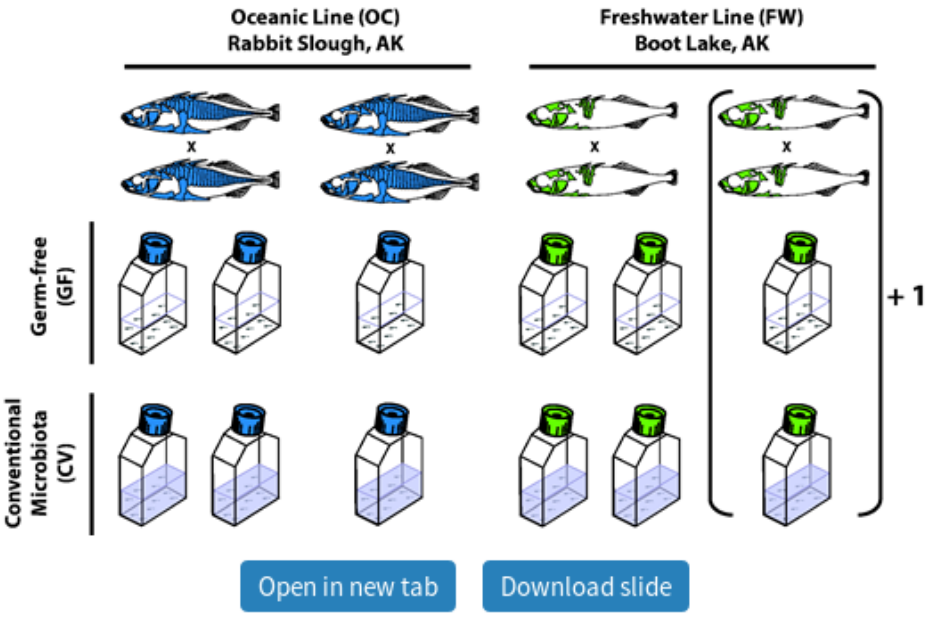
1. Name and briefly describe the type of GLM this is.
2. What is the model formula `MATED | trials(1) ~ BATTLE + (1|HERD)` set up to evaluate/estimate?

3. What is the `prior` argument telling brms to do?
4. What are the `iter` and `chains` arguments telling brms to do?

Part 2: Major determinants of stickleback gut transcriptome variation

We have gene expression data for 15,847 genes in 84 threespine stickleback fish in an experiment, summarized in this figure:

FIG. 1.—



Experimental design. An illustration of the factors and their levels included in our experiment. In total we generated 84 RNA-seq libraries (6 fish each from 14 flasks), each using mRNA isolated from the entire gut—posterior esophagus to anus—of individual fish. The two stickleback lines ("OC" and "FW") were derived from natural Alaskan populations, as indicated. Two OC families and three FW families (FW family 3 not shown, but referenced by brackets), were represented in the study. Note that for one OC family and one FW family, conventional and germ-free treatments were duplicated to examine housing ("flask") effects.

Use the data in `CVvsGF_RNAseq_CPM.tsv` and `CVvsGF_RNAseq_Metadata.tsv` to complete the following:

Your goal is to perform a PCA or nMDS (your choice) on these data to discover major sources of expression variation among fish.

Which if any of the variables in the metadata are most strongly associated with overall differences in variation (e.g., sex? population? treatment?)

In particular, to show this you should make a plot with one point for each of the 84 fish on the new axes you produce through your ordination, colored by the grouping variable. Keep in mind (especially for PCA) that normalizing the data first may improve the resolution of some patterns.

The purposes of this homework are to practice interpreting the structure of brms GLMs, running and interpreting an ordination for RNA-seq data, and communicating results in plain language.