**Homework 1**

We will be working with data collected from dogs for this homework assignment. This data is split across multiple files and your job will be to assemble the data into a single matrix. Lastly, you will write a short shell script to split this matrix up into separate files for each dog breed.

Your submission will primarily be the record of commands you used, so make sure to save them somewhere. This can be most easily done by saving a portion of your history file; in your home directory this will start with a . and may be called something like .bash\_history. It may be easiest to empty this file before starting the assignment (e.g. open it with vim and type “100000 dd”).

The dogs’ saliva was used for bisulfite sequencing. This is a way of quantifying the molecules attached to DNA, often referred to as DNA methylation. The results of these sequencing experiments might make their way into future assignments or exercises.

You are expected to complete this homework using bash commands. There are multiple ways of doing this, and any series of commands that gets you to the desired product is acceptable. However, manual editing the data files (e.g. using vim) is not allowed. The bash commands listed below are likely to be helpful.

Cat

Ls

Paste

Sed

Cut

Sort

Uniq

Using \* as a wildcard character

Using > and >> to write to files (writing intermediate results to files as you assemble the final product is allowed)

Batch 1:

There are 71 files, each containing data for a single dog. Each file contains 15 fields, separated by tab characters. The corresponding meaning of each field can be found in HeaderBatch1.tsv.

Combine these files into a single spreadsheet in .tsv format. Each column should correspond to one of the 15 fields, and each row should correspond to a dog. Make sure the header is included at the top of the final file. The final matrix should look something like this:

| Sample Number | Kit Number | Kit Number | Date Of Giving | Birth Date |
| --- | --- | --- | --- | --- |
| 1 | 3.10E+13 | Adog\_1134 | 16-Aug | … |
| 2 | 3.10E+13 | Adog\_0323 | 15-Sept | … |
| … | … | … | … | … |

Batch 2:

This batch is similar to batch 1, except that the name of each dog has been removed from the text of the corresponding file. Use these to create a matrix that INCLUDES the dog's names as one of the columns. You can tell which file refers to which dog based on the file name.

Your final matrix should **not** contain the file attachment (.dog.txt). Use cut to strip away this trailing text. Names that look like “Scooby.dog.txt” will lose points, and should instead just be “Scooby”.

| Dog Name | Sample Number | Kit Number | Kit Number | Given To |
| --- | --- | --- | --- | --- |
| Scooby | 1 | 3.10E+13 | Adog\_1134 | … |
| Courage | 2 | 3.10E+13 | Adog\_0323 | … |
| … | … | … | … | … |

Batch 3:

Here, we have one file per trait, with the traits being Breed, Gender, Name and Weight. Each file has two columns separated by a comma; the first column is the Sample ID, the second column is the trait value for the corresponding dog. Again, combine these separate files into a single spreadsheet, which should look like the below.

The order of the dogs is the same in all four files. The final file should be in .tsv format, not .csv; values should be separated by tabs instead of commas.

| SampleID | Dog Name | Breed | Gender | Weight |
| --- | --- | --- | --- | --- |
| 102 | Scooby | Great Dane | Male | … |
| 104 | Courage | Yorkshire Terrier | Male | … |
| … | … | … | … | … |

Combine the batches together:

Let’s combine all 3 batches into a single spreadsheet. The columns of each spreadsheet are slightly different, so you will want to pick out only the columns shared between all 3 matrices. These will be the same traits as in batch 3: SampleID, Dog Name, Breed, Gender, Weight.

Splitting it up again:

Finally, write a shell script that splits the full matrix into separate files. Each file should be named after a dog breed, and contain all the dogs of that breed. You can put the results in the currently empty “DogsByBreed” directory. You might get a lot of files with just one animal, and some others that differ only by a few characters. For example, there is one dog that will produce the very long file name “40% pomerarian, 37% chihuahua, 12% icelandic sheepdog, 6% keeshond, 5% Mi-ki (wisdom DNA panel)”.