GOAL: Create a methylation aging clock to use to reliably age Madagascar bats (*E. dupreanum, P. rufus, R. madagascariensis)* from wing punches instead of dentition. This project is half wet pipeline, half dry pipeline.

Background

Epigenetic clocks allow us to measure the age of tissue by looking at changes in DNA. This works using DNA methylation. Methyl groups attach to cytosines over time. You can study changes and patterns in DNA methylation over time to determine age.

Wilkinson et al 2021 – paper that uses bat dna methylation to age bats by building epigenetic clock

Input: methylation, confirmed age

Output: model

Methods: Creation of epigenetic clocks using penalized regression

We developed epigenetic clocks for bat wing tissue by regressing chronological age on all CpGs that map to at least one of the ten bat genomes. To improve linear fit we transformed chronological age to sqrt(age + 1). Penalized regression models were created in the R package glmnet[61](https://www.nature.com/articles/s41467-021-21900-2#ref-CR61). We investigated models produced by elastic net regression (alpha = 0.5). The optimal penalty parameters in all cases were determined automatically by using a tenfold internal cross-validation (cv.glmnet) on the training set. By definition, the alpha value for the elastic net regression was set to 0.5 (midpoint between Ridge and Lasso-type regression) and was not optimized for model performance. We performed two cross-validation schemes for arriving at unbiased estimates of the accuracy of the different DNAm based age estimators. One type consisted of leaving out a single sample (LOO) from the regression, predicting an age for that sample by regressing an elastic net on the methylation profiles of all other samples and iterating over all samples. We conducted LOO analyses using all samples from all species, using all samples from each species and using all samples from several species in the same genus. The second type consisted of leaving out a single species (LOSO) from the regression, thereby predicting the age of each sample using the data for all other species.

PIPELINE – Building the clock

1. DNA extraction from wing punches
   1. In: wing punch of known-age bats (from dentition)
   2. Out: DNA
2. Methylome sequencing
   1. In: DNA
   2. Out: methylation levels (DNAm, beta values) at targeted CPG sites
   3. Using custom Illumina methylation array
      1. Biosulfite conversion is the same thing as methylation sequencing?
3. Python processing
   1. In: methylation levels raw?
   2. Out: ?
4. SeSame method to normalize beta values for each probe
   1. In: beta values (0-1)
5. Elastic net regression with known age data

PIPELINE – Using the built clock

1. Wing punch lab analysis
2. SeSAme normalization
   1. In: raw methylation levels at target CPG sites
   2. Out: normalized methylation levels

VALIDATION – WORK IN PROGRESS

1. Confirm that the Myotis samples can be processed A-Z
   1. Lab work troubleshooting – Manny/Sarah
      1. Sarah determined probe sequences based on Wilkinson et al—see /methylationbackground\_Guth
   2. There’s some python something happening here – what is it and who is going to do it in the future
   3. Dry pipeline
      1. Process raw reads (“Manny’s samples”)—modify into format to be used in a clock estimate based on Wilkinson et al
      2. Try to use a finished clock (Wilkinson et al clock) and see if I can get age estimates
         1. What is the clock, how do I use it
2. Test Manny’s pipeline on known age individuals to make sure we can get reasonable age estimates. Wilkinson is sending Manny samples—‘’the key part here is not so much the species as to get as wide and even a range of ages as possible for testing the predictive power of our approach”
   1. Wet: “for the Wilkinson samples, I’m going to be conservative and only do two Wilkinson samples for now, plus two other known-age controls from cell lines that Sarah extracted” – Manny
      1. Others will be completed in Berkeley
      2. These are already DNA, and we are going to learn how to do the methylation process
   2. Dry: predict age using wilkinson’s clock

DATASETS

**Manny’s samples:**

5 Myotis species with no known ages. “The goal for these runs was to show that the library prep, sequencing, and mapping/analysis pipeline works from A-Z. Since we didn’t want to burn through precious samples, we used as input either wing punches I could spare or cell lines that we had extracted DNA from already. The DNA from the cell lines was a control to make sure that the non-DNA carryover from wing punches didn’t inhibit or cause any issues downstream. Also, since these were the cell lines used to make the reference genomes for these species, it is the best mapping control you can ask for.

https://github.com/docmanny/methylationAging

Quick facts

5 species

* Myotis californicus
* Myotis evotis
* Myotis lucifugus
* Myotis thysanodes
* Myotis Volans

File structure

Output/methylation\_extracted

Table

Description automatically generated

DRY PIPELINE –

I looked into Wilkinson et al to see how they processed their raw methylation reads before using them in their clock to estimate ages. There is substantial documentation that they first normalized them using an R package called SeSAme. My first objective is therefore to figure out how to use SeSame and to write a pipeline to normalize some of the Myotis samples.

Based on the information above from Manny’s email, I think that I need to work with a SUB file.

CLOCK METHODS

General elastic net regression ML theory

Target variable: age

Independent variable: dna methylation levels

\*use methylation to predict age

1. Split into X and Y
   1. X = independent
   2. Y = target only
2. Split into train and test
   1. X\_train =
   2. X\_test =
   3. Y\_train =
   4. Y\_test =
3. Set up hyperparameters in elastic net
   1. Elastic net(alpha = 0.5, L1\_ratio = 0.5)
4. Fit model
   1. Enet.fit(X\_train, Y\_train)
5. Make predictions
   1. Y\_pred = Enet.pred(X\_test)
6. Evaluate
   1. MSE (y\_test, y\_pred)