UCLA

UCLA Electronic Theses and Dissertations

Title

CanidPredict: A Platform for Prediction and Visualization of Traits using Canid methylomes

Permalink

https://escholarship.org/uc/item/2z08f55m

Author

Ramasamy Jayaseelan, Ponmathi Chamundeswari

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

CanidPredict: A Platform for Prediction and Visualization of Traits using Canid methylomes

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Bioengineering

by

Ponmathi Chamundeswari Ramasamy Jayaseelan

2022

ABSTRACT OF THESIS

CanidPredict: A Platform for

Prediction and Visualization of traits using Canid methylomes

by

Ponmathi Chamundeswari Ramasamy Jayaseelan

Master of Science in Bioengineering

University of California, Los Angeles, 2022

Professor Matteo Pellegrini, Chair

Studies of canid epigenetics can shed light on the relationship of the DNA methylome and mammalian longevity. The strong link between human and domesticated dog biology motivates the use of canids as a model for studying aging. Previous studies have examined factors such as age, sex, and weight and their associations with targeted bisulfite sequencing data using multivariate machine learning models. Creating visualization tools of these models facilitates community use of these resources. In this thesis, we describe an R-based web application platform (Shiny) that executes regression and classification models designed using buccal swab data obtained from 217 samples from previous studies. We developed CanidPredict, which allows users to upload their individual samples using either a CGMap file or a calculated methylation matrix and view predictions regarding age, sex, weight, sterilization status, and behavioural status as well as the general model performance for each trait. In addition, the goal of CanidPredict was to also create an application design which is user-friendly, can be used both on a server or locally, and can be scaled up to include further raw file formats and additional trait associations. CanidPredict is available at https://singlecell.mcdb.ucla.edu/DogAge/.

The thesis of Ponmathi Chamundeswari Ramasamy Jayaseelan is approved.

Jennifer Lynn Wilson

Xia Yang

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2022

Table of Contents

List of Figures

Acknowledgements

I would like to express my thanks to my professor, Matteo Pellegrini, for helping me and guiding me through this project and my master's degree. I would also like to thank my thesis committee members, Dr. Xia Yang and Dr. Jennifer Wilson, for their additional guidance. In addition, thank you to the members of the Pellegrini Lab for assisting me whenever I had questions during my project.

1. Introduction

Computational tools to understand mammalian longevity are vital for studies of human health and disease. In order to better understand mammalian aging, DNA methylation is often used, as it is a form of epigenetic modification of the genome which shows patterns that change throughout a lifespan. In mammals, these patterns allow for the creation of epigenetic clocks, or age estimation methods, which have been well-established and widely used to study aging and potential aging interventions [1]. Canines can be a useful model for the study of methylation changes across a lifespan. Canine epigenetics are important in understanding lifespan due to similar development, aging processes, and life stages between humans and *Canis lupus familiaris* (domesticated dogs) [2]. The variation in traits in dogs significantly impacts canine lifespan. For instance, variation in traits such as weight and reproductive capability has been studied as important factors of aging in dogs [3]. More recently, researchers have seen that dog breed is not necessarily a predictor of behavior. A genetic analysis of over 18000 canines found several regions of the genome associated with behavioral tendencies, refuting the idea that breeds dictate certain traits [4].

Thus, the relationship between canid epigenetics and traits is a key point of study. Using targeted bisulfite genomic sequencing, a method of methylation measurement which provides information about a set of loci in the genome, previous studies have been able to obtain DNA samples from dogs and wolves in a minimally invasive manner. From this high dimensional data set, researchers can then predict traits such as age, sex, and weight for individual samples to show how the consistently changing methylome is affected by these traits [5].

While previous studies have shown how multivariate machine learning models can be applied to methylation data, platforms for visualization of these models are not yet available. For experimentalists wanting to run these models in a fast and user-friendly environment, visualization platforms can greatly aid basic research. For this thesis project, a web application to generate visualizations from user-uploaded methylation data was created to further advance research in canine epigenetics.

2. Background

2.1 Measuring the impact of factors on canid epigenomes

In a previous study by Rubbi et al., they developed prediction models for aging biomarkers using DNA methylation profiles measured through buccal swabs from several canid breeds [5]. Using targeted bisulfite sequencing from buccal swabs, they measured DNA methylation profiles and also collected data regarding age, sex, weight, and sterilization status, of each dog. These factors in particular are important to consider as previous work has shown them to influence the lifespan of dogs. For example, higher body weight and larger breeds were previously shown to age more rapidly [6]. In an epigenome-wide study, body mass index (BMI) was also seen as a factor associated with methylation, showing how environmental state impacts the methylome [7]. Furthermore, in a study by Hoffman et al. looking at the mechanism behind cost-of-reproduction in dogs, they showed the impacts of reproductive capability on age at death, and how sterilization played a role in increasing lifespan [8]. Once data regarding these factors were collected, Rubbi et al. successfully associated methylation status to traits by training various prediction models and validating them through leave one out cross validation. In addition, they were able to identify single nucleotide polymorphisms (SNPs) and performed hierarchical clustering to construct genetic relationships from the samples. These studies motivate further research into the relationship between phenotype and genotype and the methylome.

2.2 CanidPredict

In this thesis project, we aimed to supplement current studies determining the ability of canid epigenomes to predict the aforementioned factors by introducing a visualization and modeling tool for researchers in this field. To do so, we developed CanidPredict, a web application platform using R Shiny (v1.7.1) which predicts the traits of user uploaded individual samples using various regression and classification-based machine learning models [9,10]. Equipped to predict age, sex, weight, sterilization status, behavioural traits, as well as genetic relationships in a phylogenetic tree, CanidPredict provides users with visualization of these models as well as a gauge of prediction error.

In addition, for increased accessibility for users, CanidPredict also provides the option to calculate methylation matrices using CGmap files, a standard file format in the field. When methylation data is uploaded, users are provided with average coverage and data quality information as well as internal filtering of the methylation sites based on coverage. The CanidPredict platform provides researchers a visualization of current traits and their impacts on the methylome, and as well as a framework that can be expanded to include future options for trait prediction (Figure 1).

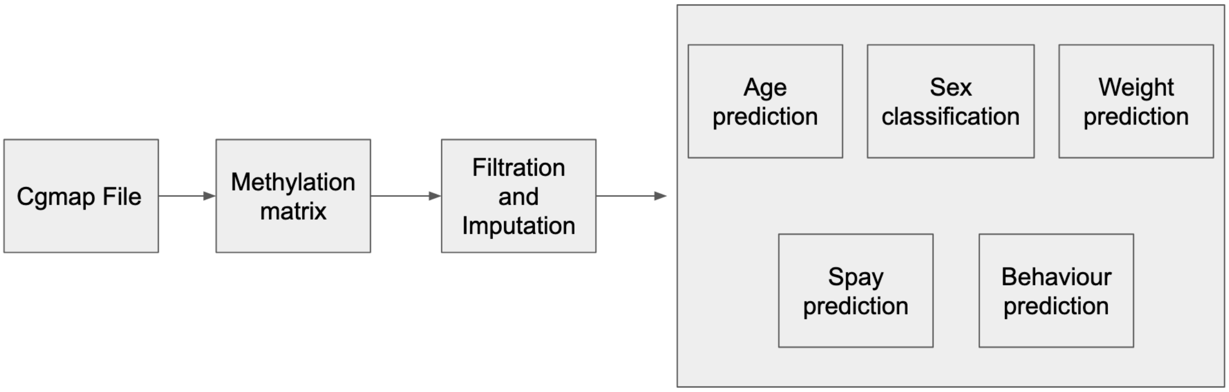


Figure 1: Application Workflow Diagram

3. Results

3.1 Data Collection and Processing

Methylation data from 217 canids (207 buccal swab samples from dogs and 10 blood samples from gray wolves) was collected in Rubbi et al. After DNA extraction from the buccal swab or blood sample, targeted bisulfite sequencing (TBS-seq) was performed using a NovaSeq6000 to obtain the methylome data. The steps undertaken by Rubbi et al. after the FASTQ files were obtained were to first use cutadapt for adapter removal. Additionally, the files were aligned to the canfam4 genome and call methylation was performed. The Align function and the CallMethylation function in BSBolt (v1.3.0) [11], a python-based software platform for bisulfite analysis which uses the BWA for read alignment, were utilized. Rubbi et al. then used the weighted average of CpG methylation near the probe center in order to determine the probe region DNA methylation values. This resulted in the aggregate matrix of the samples' methylation levels at each site, containing 217 samples and 5609 methylation sites, which we then filtered and used to fit the models for prediction and classification.

3.2 Prediction Models Design

We designed several prediction and classification machine learning models to determine behavior traits from the aggregate methylation matrix to integrate into CanidPredict. The prediction models used either LASSO regression or Partial Least Squares Regression and the classification models used either logistic regression or support-vector machine (SVM). The models for each trait were chosen based on performance results in previous literature. Some models, such as the sterilization status classifier, required trials using multiple models due to the imbalanced class size to determine the model with the best performance.

Age prediction - In order to build the age prediction model, the R package ‘glmnet’ (v4.1.0) was used to train and validate a LASSO regression model [12]. The aggregate methylation matrix was then filtered by removing 17 of the samples with the lowest average counts as well as all columns of methylation sites with missing values, leaving a new matrix of 200 samples and 2642 methylation regions. The new matrix was then input into the function cv.glmnet with a k-fold parameter of 10, with the age in years set as the target variable. The weights, lambda value, and other parameters were all kept at their default values of ‘null’. The gamma argument, the value for balance between relaxed and regularized fitting, was specified as the default vector of values (0, 0.25, 0.5, 0.75, 1). In order to obtain individual sample predictions, Leave One Out Cross Validation (LOOCV) was used by repeating the process in a loop and setting aside a different test sample each time. The test sample was input into the function cv.glmnet.predict, with the values for the penalty parameter set to both the lambda of the cross-validated minimum mean error (lambda.min) and the greatest lambda within one standard error of cross-validated errors of lambda.min (lambda.1se). This model is then saved as an .RData file and uploaded to the server.

Sex classification - The sex classifier was built using the logistic regression model from ‘glmnet’ [12]. After removing the same samples and sites as for the age prediction, the cv.glmnet function was used with the family parameter set to ‘binomial’ to denote logistic regression with sex classes (‘1’ for female; ‘0’ for male) set as the target variable. Using the same parameters as those for the age prediction as well as the nested LOOCV method for individual sample prediction, the sex classifier was constructed and validated. This model is then saved as an .RData file and uploaded to the server.

Weight prediction - The LASSO regression model for weight prediction was built using the same functions from ‘glmnet’ and parameter specifications as the age prediction model. The model was trained on the same matrix of 200 samples and 2642 methylation sites using LOOCV, but instead the weight in pounds was set as the target variable. This model is then saved as an .RData file and uploaded to the server.

Behavior prediction - For behavior prediction, PLSR models for various behavior traits from responses to the Canine Behavioral Assessment and Research Questionnaire (C-BARQ) were constructed using the R library ‘pls’ (v2.8.0) [13]. The C-BARQ questions are a standardized tool specifically developed for behavioral evaluation and study of prevalence of behavioral issues in dogs [14] The values of the responses ranged from 0-4. Using the same matrix of 200 samples and 2642 methylation sites, each model was created using nested LOOCV, setting the target variables as each of the behavior traits and the scaling option to true. For prediction, the number of components to be used in fitting the model was set to 2 after various trials. We selected behavior traits with the highest model correlation coefficients: attachment and attention seeking behavior (R=0.424), energy level (R = 0.449), and stranger-directed fear (R=0.36). This model is then saved as an .RData file and uploaded to the server.

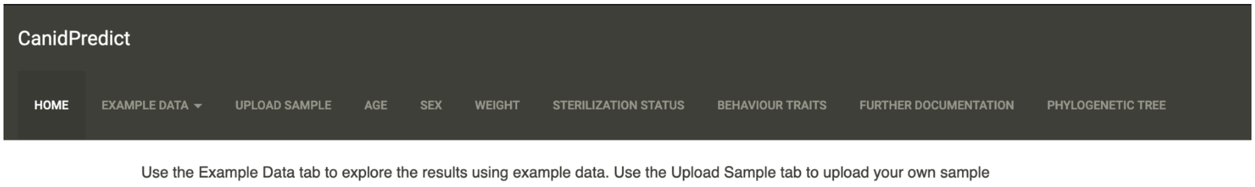
Sterilization status classification - The sterilization classifier was created using a SVM model using R library ‘e1071’ which uses the ‘libvm’ package [15,16]. Based on the steps used in building a SVM classifier using scikit-learn by Rubbi et al., additional low-read sites were removed so that the new matrix had 2420 sites remaining. As the model was performing successfully at prediction using all female samples rather than all male samples or the full sample set, the model was constructed using only the 77 female samples. This is consistent with the steps Rubbi et al. took to build their SVM classifier, as they concluded that the hormonal changes due to sterilization may impact the female epigenome to a greater extent than the male epigenome [5]. The dataset further contained imbalance classes (‘1’ for spayed; ‘0’ for intact) , with the spayed class and intact classes containing 68 samples and 9 samples respectively. In order to correct for this, oversampling was performed on the dataset using the ovum.sample function from the R library ‘ROSE’ [17]. The model was first tuned using the tune function in ‘e1071’ with a vector of values 0.001,0.01,0.1,1 for gamma, values 0.01,0.1,1,10,100 for cost and a radial kernel in order to find the optimal parameters. From the tuning, a gamma value of 0.001 and a cost of 10 were determined. Using a LOOCV nested loop, the model was built using those gamma and cost values, the oversampled data, and with the option for scaling turned off. This model is then saved as an .RData file and uploaded to the server.

3.3 Application Design

To create the app CanidPredict, we used the Shiny package, a web application creation framework within R [9]. We aimed for users to have easy access to CanidPredict, and thus the app is hosted on a readily available server (https://singlecell.mcdb.ucla.edu/DogAge), with the individual models and sample data used available on a Github repository (https://github.com/mathrj/CanidPredict).

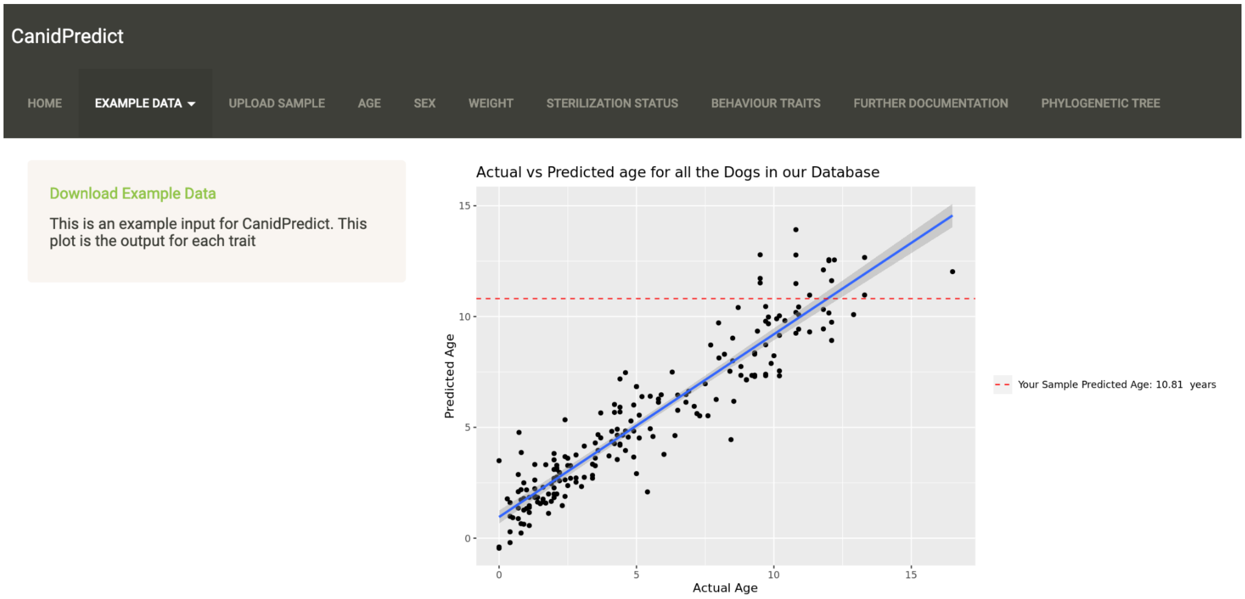
The app was designed with the user's ease and understanding as the main priority. The overall app layout is as shown below (Figure 2). With the multiple tab options such as homepage with simple instructions, an example output page, an upload page, as well as tabs for each of the models at the top of the page, the layout of the main panel results in a minimal and clean look. In addition, the example data tab provides examples of what the results in each of the modeling tabs will look like after uploading a sample file. The example file is linked on the side panel of this page for the user to download and try on their own. The example tab for the 'Age' option is shown in Figure 3.

In addition, this information can be found in the "Further Documentation" tab which links to a tutorial on the Github repository.



Use the Example Data tab to explore the results using example data. Use the Upload Sample tab to upload your own sample

Figure 2: Application Layout



Download Example Data

This is an example input for CanidPredict. This plot is the output for each trait

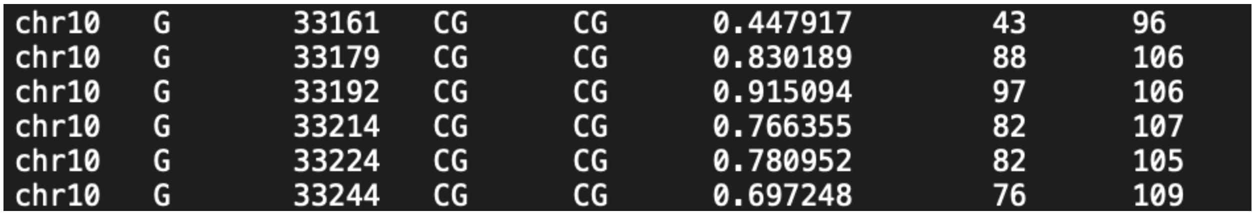
Figure 3: Age Plot using Example Data for User Reference

3.4 App Upload Page and Preprocessing

3.4.1 Data upload and Permitted File Formats

Under the “Upload Sample” tab, the user will see a file upload box in the sidebar panel. Here the user can access their local files and select the sample file of their choice to the application. The max size of uploaded files is 30 MB. In addition, the file upload box requires data files in the following formats: comma-separated values (.csv) or DNA methylome CGmap file format (.CGmap). If an incorrect file extension is used, the application will display an error "Invalid file; Please upload a .csv or .cgmap file".

CGmap files are a standard file format which provides information about DNA methylation level of covered cytosines on a reference genome [18]. Examples of parts of the CGmap file or .csv file for an expected sample are shown below in Figure 3 and 4.



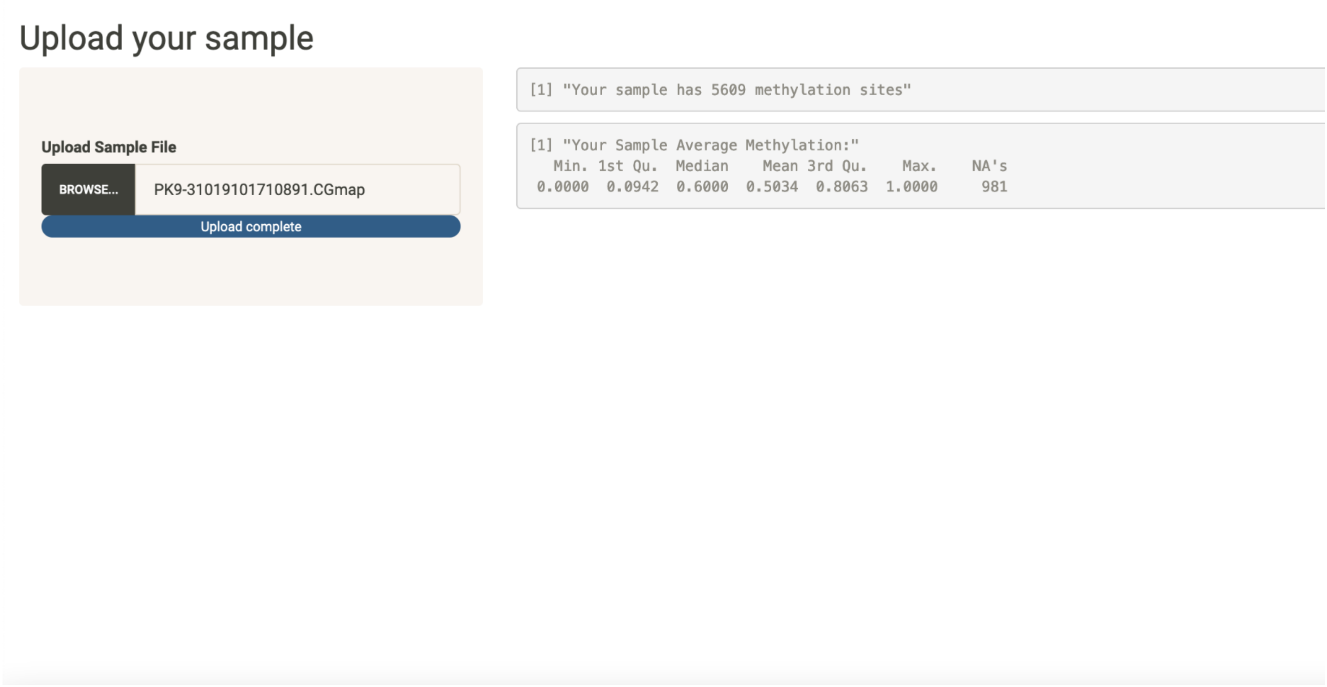
```
chr10   G       33161   CG      CG      0.447917        43      96
chr10   G       33179   CG      CG      0.830189        88      106
chr10   G       33192   CG      CG      0.915094        97      106
chr10   G       33214   CG      CG      0.766355        82      107
chr10   G       33224   CG      CG      0.780952        82      105
chr10   G       33244   CG      CG      0.697248        76      109
```

Figure 3: Example of the contents of a CGmap file of an English Springer Spaniel. The columns in the file from left to right are as follows: query template name, nucleotide on the reference genome, mapping position, dinucleotide follows, methylation level, read counts of methylated cytosines, and read counts of all cytosines [18].

	V1	V2	V3	V4	VX
1	DogWolf00110_S105.CGmap	0.074626866	0.91609589	0.160997732	…

Figure 4: Example of the proper format of a methylation matrix contained in a .csv file of an English Springer Spaniel. The first column, V1, has the filename, and the following columns pertain to each of the methylation sites.

When the sample inputs are uploaded, a summary of the sample methylation such as mean, maximum, and minimum are displayed, in addition to the number of sites with unavailable data which will either be deleted or imputed (Figure 5). User sample inputs are filtered automatically based on sites with the highest coverage and then remaining missing values are imputed using R package 'mlr' (v2.19.0) using the mean imputation method [19]. For the sterilization status classification, imputation is performed using the R package 'Hmisc' (v4.6.0) using the mean imputation method [20].



```
[1] "Your sample has 5609 methylation sites"
```

```
[1] "Your Sample Average Methylation:"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.0000  0.0942  0.6000  0.5034  0.8063  1.0000     981 
```

Figure 5: Sample Upload page after user file has been uploaded and the resulting summary information is processed.

3.5 Modeling Tabs

3.5.1 Age Prediction Model Tab

In this tab, after the user uploads a single sample, a scatter plot of the actual age (in years) from the database used in training the models versus the predicted ages obtained from the model is displayed. The layout of the tab after the application has processed is shown in Figure 6. A linear regression line is shown in blue with the 95% confidence interval shown as shading. Furthermore, when the user's sample is run through the age prediction model, the predicted age for their sample is shown as a horizontal dashed line (in red) with the value in the plot's legend.

From this, users can not only see the predicted age of their sample, but also the ages of other samples from the database of similar predicted age which provides a rough estimate of model error.

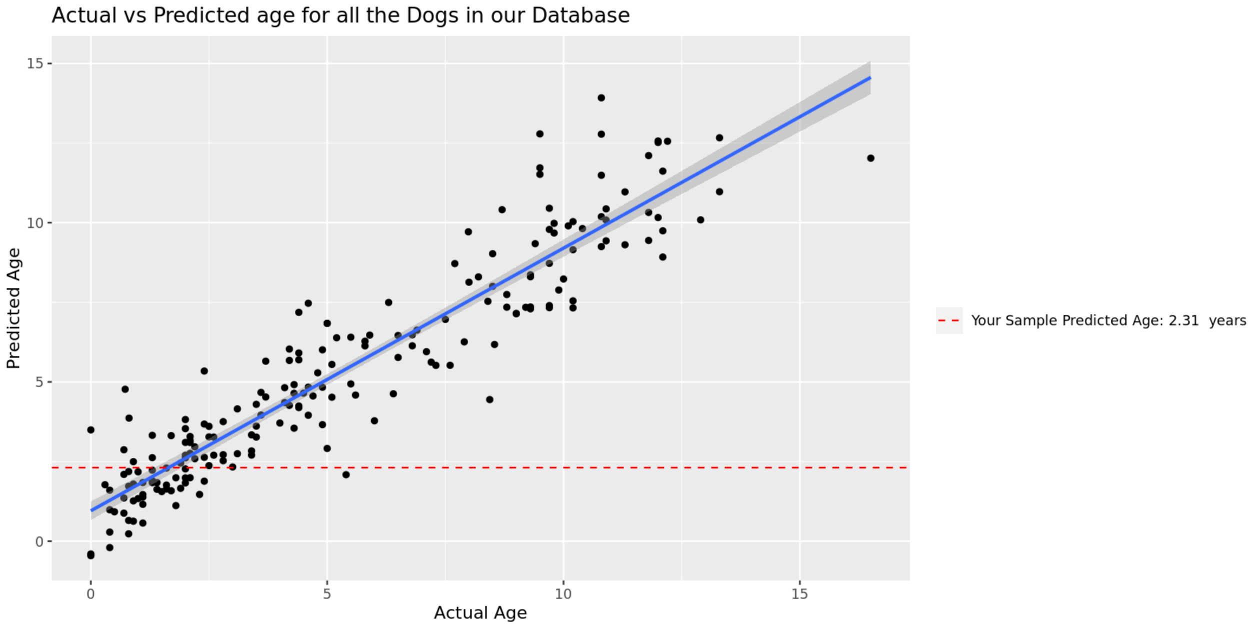


Figure 6: Sample age prediction tab (Female French Bulldog of actual age 1.9 years).

3.5.2 Sex Classification Model Tab

In this tab, after the user uploads a sample, a violin plot of the actual sex from the database used in training the models versus the predicted sex for each sample obtained from the logistic regression model is displayed. The layout of the tab after the application has processed is shown in Figure 7. Furthermore, when the user's sample is run through the sex classification model, the predicted sex for their sample is shown as a horizontal dashed line (in red) with the value and classification as male or female in the plot's legend. From this, users can see the classified sex of their sample, and also the distribution of the predicted values for both female and male samples in the database.

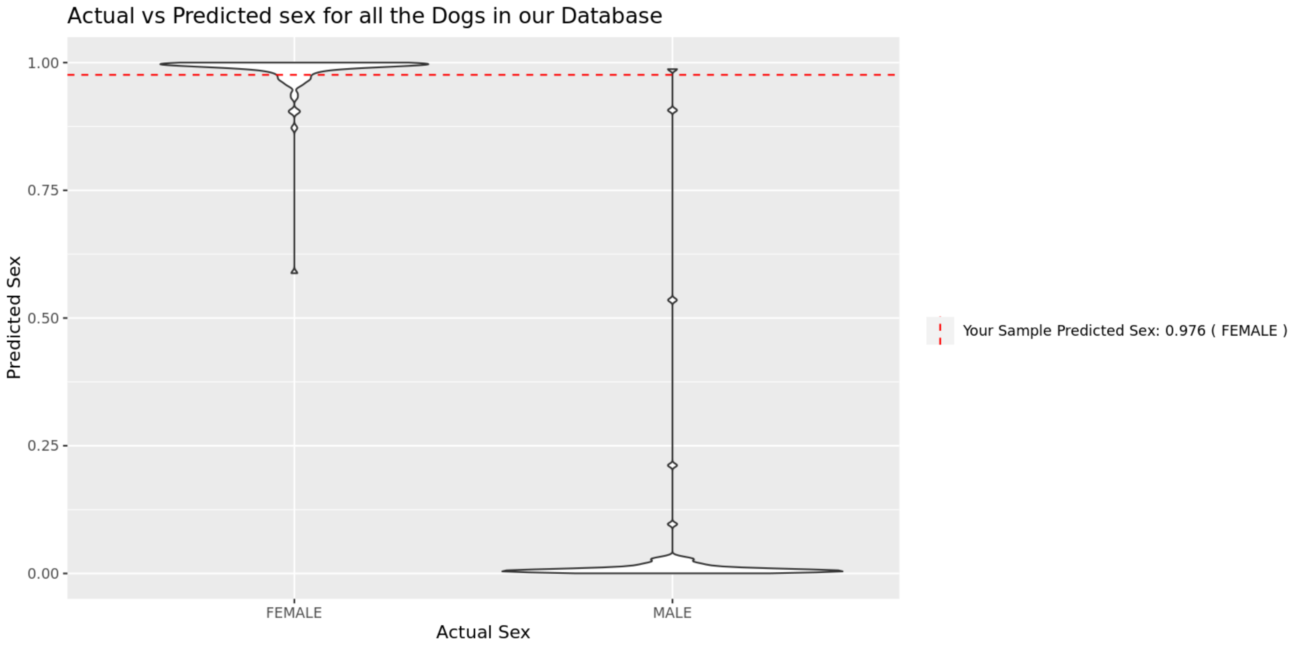


Figure 7: Sample sex prediction tab (Female French Bulldog).

3.5.3 Weight Prediction Model Tab

Similar to the age prediction model tab, the output layout of this tab is a scatter plot of the actual weight (in pounds) from the database used in training the models versus the predicted weights obtained from the model (Figure 8). Like the previous tab, the linear regression line goes through the plot (in blue) with the 95% confidence interval shown as shading and the predicted weight for the user's sample is shown as a horizontal dashed line (in red) with the value in the plot's legend.

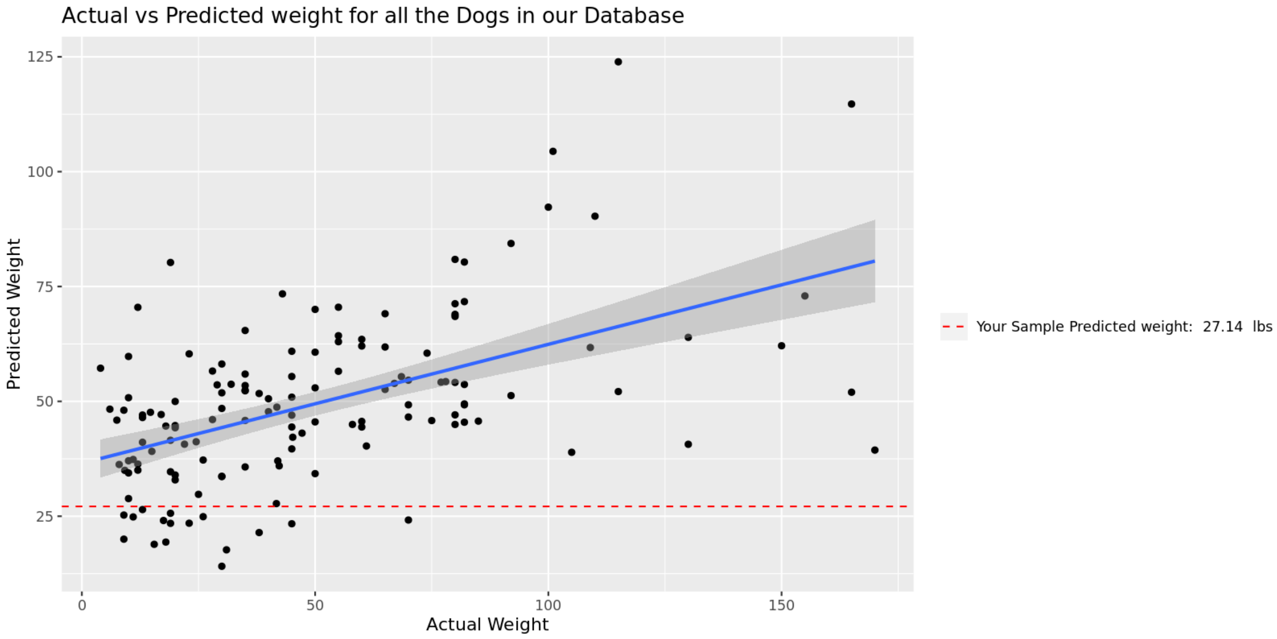


Figure 8: Sample weight prediction tab (Female French Bulldog with a weight of 26 pounds).

3.5.4 Sterilization Status Classification Model Tab

In this tab, users can classify the sterilization status of their samples. After the application processes their uploaded sample using the SVM model, a confusion matrix is displayed of the predicted and actual spayed versus intact counts as shown in Figure 9. Above the matrix, the classification output for the single sample is shown as ‘Spayed’ or ‘Not Spayed’. In addition, as the model only uses the female samples from the database, a warning is displayed to the user about the validity of the model predictions.

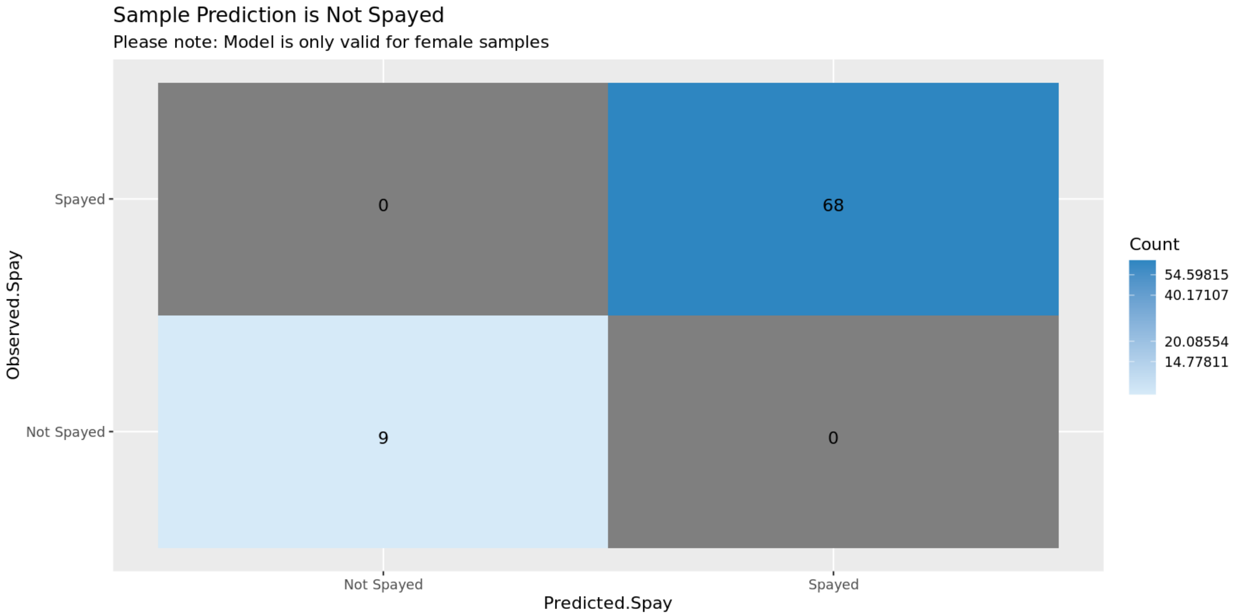


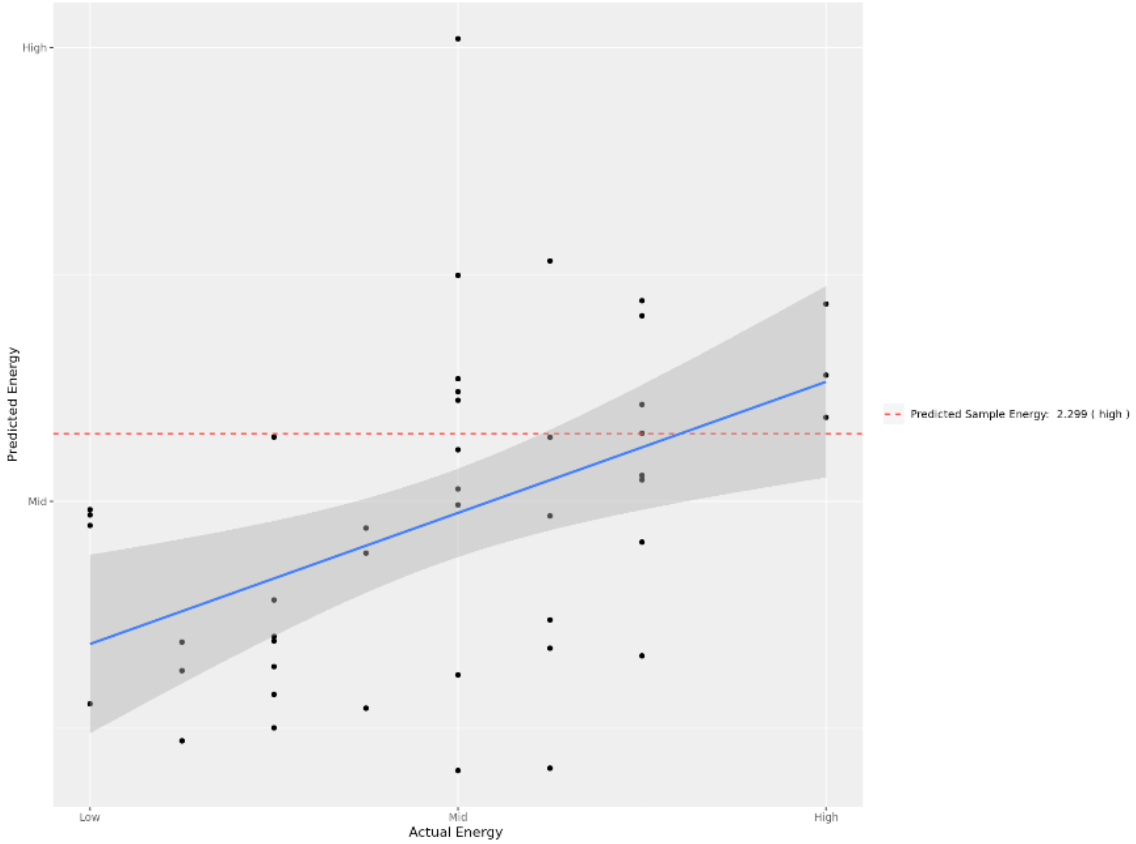
Figure 9: Sample spay prediction tab (Female French Bulldog that is not spayed).

3.5.5 Behavior Traits Prediction Model Tab

In the behavior traits model tab, three scatter plots for each behavior trait model are displayed (Figure 10). For each trait, the plot shows the actual value from the database used in training the models versus the predicted values obtained from the PLSR model. The traits included are the ones with the highest correlation seen during modeling, “Attachment/attention-seeking”, “Energy”, and “Stranger-directed Fear.” Descriptions of the traits taken from the C-BARQ website, the standard data source of dog behavioral traits, are listed as the subtitles of the scatterplots for the users’ reference. For the chart axes, we mapped quantitative questionnaire responses (i.e., 0-4) to categorical values (“Low”, “Mid”, “High”). Similar to the other tabs, the linear regression line goes through each plot (in blue) with the 95% confidence interval shown as shading and the predicted value for the user’s sample is shown as a horizontal dashed line (in red) with the value and categorization as “Low”, “Mid”, or “High” in the plot’s legend.

Actual vs Predicted Energy for all the Dogs in our Database (Corr coeff = 0.449)

Energy level: Energetic, “always on the go”, and/or playful.



Actual vs Predicted Attention for all the Dogs in our Database (Corr coeff = 0.424)

Attachment and attention-seeking: Maintaining close proximity to the owner or other members of the household, soliciting affection or attention, and displaying agitation when the owner gives attention to third parties.

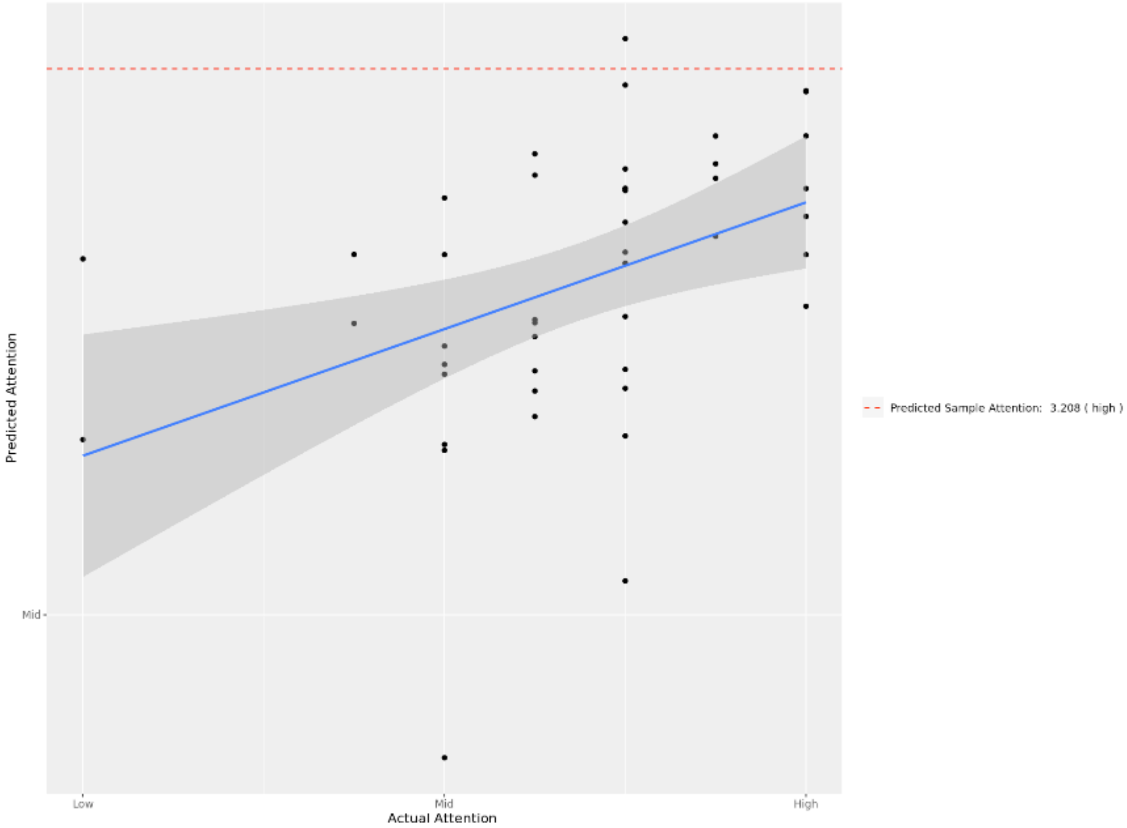


Figure 10: Sample behavior prediction tab showing outputs for “Energy level” and “Attachment and attention-seeking” behaviors (example for “Stranger-directed Fear” behavior is not shown).

3.5.6 Extended functionality

CanidPredict additionally can be extended to include processing of genotype data.

Phylogenetic tree - After the user uploads a genotype sample, it is added to a genotype matrix of all the dogs and wolves in the database and then the application will impute the data using weighted k-nearest neighbors based library in R ‘scrime’(v1.3.5) [21]. After imputation, the application will then calculate the genetic distance using the R library ‘NAM’ (v1.7.3) and apply hierarchical clustering [22]. Utilizing the ‘ape’ (v5.6.2)and ‘ggtree’ (v3.2.1) packages in R, a phylogenetic tree of the breeds in the database will be generated with the user’s sample highlighted in red. However, in its current state, this tab is conceptual and outside of the scope of this thesis project (Figure 11).

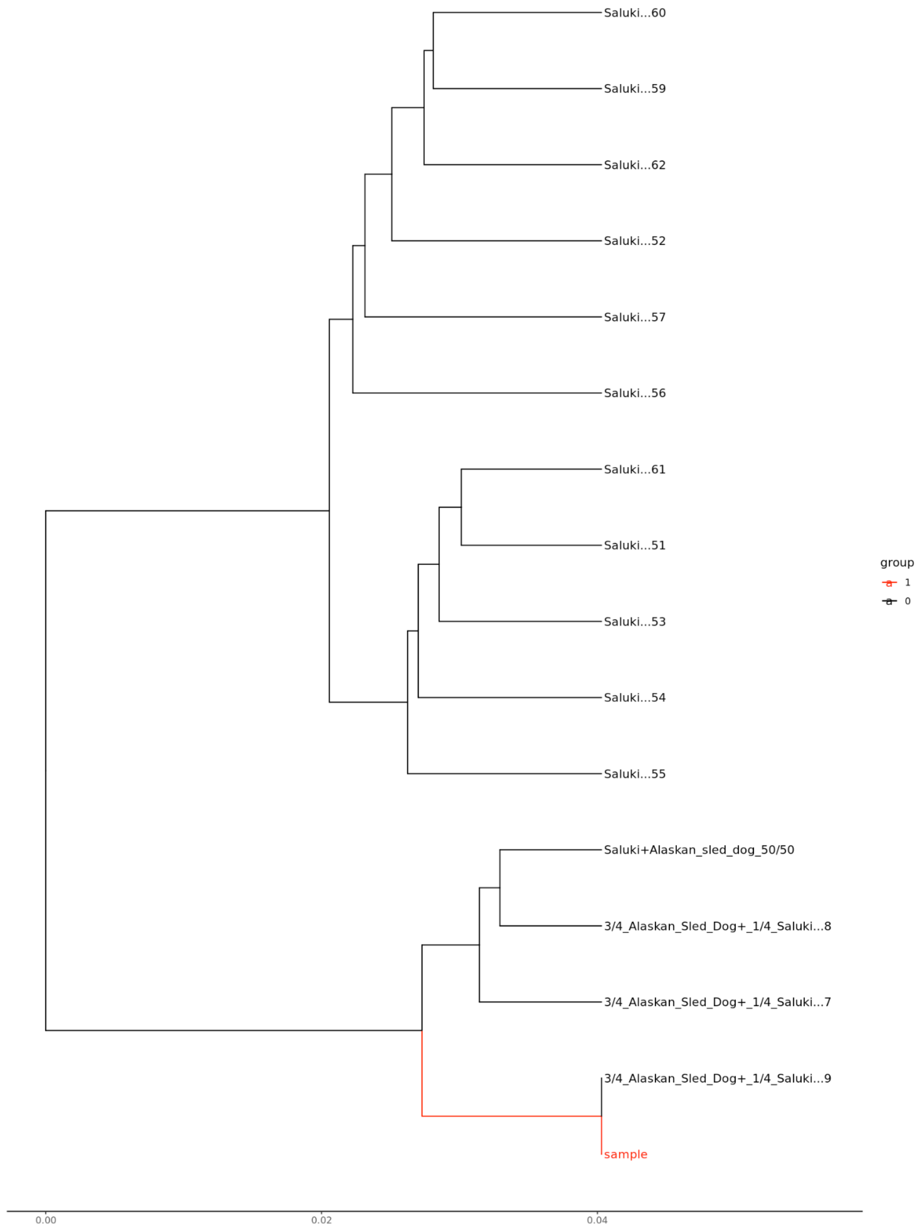


Figure 11: Example of the phylogenetic tree generated using genotype data. Users can also have the customization options of the number of tree branches or the branch width.

3.6 Application Testing and Performance Testing

In order to fine-tune the application and find regions of improvement for future iterations, the overall efficiency of CanidPredict was profiled using the R package ‘profvis’ (v0.3.7) [23]. By running the app through the profvis function, memory profiling and information about resource allocation such as memory used and time required can be obtained. This test was performed on the RStudio server (singlecell.mcdb.ucla.edu/dogagepred), which uses R version 3.6.3 running on Ubuntu. The result of the initial profvis test showed that running the app workflow from upload to analysis uses 874.5 MB memory allocation and 901.6 MB memory deallocation for a total time of 2840 ms. Generally, when there is great allocation without deallocation, it could possibly be due to a memory leak, but profvis shows little evidence of this. The aspects of the app which take the longest to run are seen in the profvis profile. In this case, in the Sterilization Status tab, the preprocessing step in which methylation sites in the sample were deleted took the longest time to complete (650 ms) when compared to the other functions in the backend.

4. Discussion

As further research uncovers how trait variation and the epigenome are associated, development of models for prediction and classification gain in importance. While previous mammalian studies in epigenetics and aging have been focused on mice or humans, using dogs as a model organism is advantageous for studying methylomes [24]. In the study conducted by Morril et al., they concluded that behavioral traits in dogs are a result of polygenic adaptation and stressed the relevance of using dog models for studies on genetic discovery [4].

R Shiny apps for interactive visualization of biological data are growing in popularity for analyzing large datasets, such as single cell RNA-Seq data, and take advantage of the plethora of statistical libraries in R [25]. Particularly, in the field of epigenetics, R Shiny apps have been applied to analyze DNA methylation and chromatin accessibility features [26]. The CanidPredict app provides users with visualization of how the model predicted other samples from the database when compared to their samples. An important future step is to consider how robust these models are to low count data, and how they respond to samples with low coverage. Further tests on sample inputs with sparse data can be conducted in order to answer these questions.

A further goal in the overall study of lifespan in dogs is to increase sample sizes, which can improve model validation and prediction performance. Using this non-invasive method of data collection utilizing buccal swabs, tests can be made available to the general public allowing for further research in the field of canid epigenetics and lifespan in the future. It should be noted that the wolf data used in creating the models was actually obtained from blood samples rather than buccal swabs, but the age prediction model still performed accurately. These results can possibly be proof of principle suggesting that these models can be applied to wolf epigenome data previously collected from blood samples and not limited to buccal swab samples [5]. Previous studies have also confirmed the validity of using both blood samples and saliva samples for obtaining DNA methylation profiles. [27]. These methods of easier data collection will allow for greater sample sizes for future studies. In addition, the CanidPredict app can also be made accessible to the public, providing users with trait information about their sample in regards to previously collected data.

4.1 Future Directions

Ongoing improvement in creating visualization tools for prediction and classification of traits is important to the field of dog epigenetics. An important aspect to CanidPredict is that in its current state, it can be scaled up to include various other prediction models and options for user customization in the future. In the future, CanidPredict can further be extended significantly to provide users with functionalities such as interfacing the platform with the BSBolt software so users can have the option to upload FASTQ files and have the genotype matrices internally constructed or to upload their own. Using the BSBolt CallVariation function and variant aggregation, similar downstream options can be applied to a FASTQ file. This will allow users to directly upload FASTQ files. Currently, using a sample in a genotype matrix form is permitted in order to create a phylogenetic tree. This option can further the user-friendliness of the app as well as reproducibility of results.

Furthermore, buccal swab data is not limited to the traits that are currently explored in CanidPredict. An interesting addition would be to use the FASTQ data from BSBolt to model bacterial associations and health in dogs, as information regarding the canine oral microbiome can also be obtained through buccal swabs. There is strong variation in bacterial populations in the canine oral microbiome, with each oral niche offering distinguished microbiota [28]. Particularly, this would be important as breed size and age of dogs are risk factors for periodontal disease, which is the leading oral disease in dogs [29]. The association of DNA methylation profiling with health status in dogs has further been researched in previous studies, specifically in canine lymphoma. Genome-wide methylation studies have shown that specific patterns in DNA hypermethylation could be traced in canines with high-grade B-cell lymphoma [30,31]. Further study into identifying specific methylation patterns in canine disease can guide future development in creating applications modeling and predicting disease risk.

References

1. Thompson MJ, vonHoldt B, Horvath S, Pellegrini M. An epigenetic aging clock for dogs and wolves. Aging (Albany NY) [Internet]. 2017 Mar 26 [cited 2022 May 18];9(3):1055–68. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5391218/.

2. Horvath S, Lu AT, Haghani A, Zoller JA, Brooke RT, Raj K, et al. Epigenetic clock and methylation studies in dogs [Internet]. bioRxiv; 2021 Mar [cited 2022 May 18] p. 2021.03.30.437604. Available from: https://www.biorxiv.org/content/10.1101/2021.03.30.437604v1.

3. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature [Internet]. 2017 Jan [cited 2022 May 18];541(7635):81–6. Available from: https://www.nature.com/articles/nature20784.

4. Morrill, K., Hekman, J., Li, X., McClure, J., Logan, B., Goodman, L., Gao, M., Dong, Y., Alonso, M., Carmichael, E., Snyder-Mackler, N., Alonso, J., Noh, H. J., Johnson, J., Koltookian, M., Lieu, C., Megquier, K., Swofford, R., Turner-Maier, J., … Karlsson, E. K. (2022). Ancestry-inclusive dog genomics challenges popular breed stereotypes. *Science*, *376*(6592). https://doi.org/10.1126/science.abk0639.

5. Rubbi L, Zhang H, Feng J, He C, Kurnia P, Ratan P, Tammana A, House S, Thompson M, Farrell C, Snir S, Stahler D, Ostrander EA, vonHoldt BM, Pellegrini M. The effects of age, sex, weight, and breed on canid methylomes. Epigenetics. 2022 May 3:1-16. doi: 10.1080/15592294.2022.2069385. Epub ahead of print. PMID: 35502722.

6. Kraus C, Pavard S, Promislow DEL. The size–life span trade-off decomposed: why large dogs die young. Am Nat. 2013;181(4):492–505.

7. Wahl S, Drong A, Lehne B, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541(7635):81–86.

8. Hoffman JM, Creevy KE, Promislow DEL. Reproductive capability is associated with lifespan and cause of death in companion dogs. PLOS ONE. 2013;8(4):e61082.

9. Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges (2021). Shiny: Web Application Framework for R. R package version 1.7.1. https://CRAN.R-project.org/package=shiny.

10. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

11. Farrell C, Thompson M, Tosevska A, Oyetunde A, Pellegrini M. BiSulfite Bolt: A bisulfite sequencing analysis platform. Gigascience. 2021 May 8;10(5):giab033. doi: 10.1093/gigascience/giab033. PMID: 33966074; PMCID: PMC8106542.

12. Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, **33**(1), 1–22. doi: 10.18637/jss.v033.i01, https://www.jstatsoft.org/v33/i01/.

13. Mevik BH, Wehrens R. The pls package: principal component and partial least squares regression in r. Journal of Statistical Software [Internet]. 2007 Jan 10 [cited 2022 May 17];18:1–23. Available from: https://doi.org/10.18637/jss.v018.i02.

14. Hsu Y, Serpell JA. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. Journal of the American Veterinary Medical Association [Internet]. 2003 Nov 1 [cited 2022 May 17];223(9):1293–300. Available from: https://avmajournals.avma.org/view/journals/javma/223/9/javma.2003.223.1293.xml.

15. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-9. https://CRAN.R-project.org/package=e1071.

16. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol [Internet]. 2011 May 6 [cited 2022 May 17];2(3):27:1-27:27. Available from: https://doi.org/10.1145/1961189.1961199.

17. Lunardon N, Menardi G, Torelli N. Rose: a package for binary imbalanced learning. The R Journal [Internet]. 2014 [cited 2022 May 17];6(1):79. Available from: https://journal.r-project.org/archive/2014/RJ-2014-008/index.html.

18. Guo W, Zhu P, Pellegrini M, Zhang MQ, Wang X, Ni Z. CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. Bioinformatics. 2018 Feb 1;34(3):381-387. doi: 10.1093/bioinformatics/btx595. PMID: 28968643; PMCID: PMC6454434.

19. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. Mlr: machine learning in r. Journal of Machine Learning Research [Internet]. 2016 [cited 2022 May 17];17(170):1–5. Available from: http://jmlr.org/papers/v17/15-066.html.

20. Frank E Harrell Jr (2021). Hmisc: Harrell Miscellaneous. R package version 4.6-0. https://CRAN.R-project.org/package=Hmisc.

21 Holger Schwender and with a contribution of Arno Fritsch (2018). scrime: Analysis of High-Dimensional Categorical Data Such as SNP Data. R package version 1.3.5. https://CRAN.R-project.org/package=scrime.

22. Xavier A, Xu S, Muir WM, Rainey KM. NAM: association studies in multiple populations. Bioinformatics 31(23):3862-3864.

23. Winston Chang, Javier Luraschi and Timothy Mastny (2020). profvis: Interactive Visualizations for Profiling R Code. R package version 0.3.7.https://CRAN.R-project.org/package=profvis.

24. Thompson MJ, Chwiałkowska K, Rubbi L, et al. A multi-tissue full lifespan epigenetic clock for mice. Aging (Albany NY). 2018 Oct 21;1(10):2832–2854.

25. Jia L, Yao W, Jiang Y, Li Y, Wang Z, Li H, et al. Development of interactive biological web applications with R/Shiny. Briefings in Bioinformatics [Internet]. 2022 Jan 17 [cited 2022 May 18];23(1):bbab415. Available from: https://academic.oup.com/bib/article/doi/10.1093/bib/bbab415/6387320.

26. Knight P, Gauthier MPL, Pardo CE, Darst RP, Kapadia K, Browder H, et al. Methylscaper: an R/Shiny app for joint visualization of DNA methylation and nucleosome occupancy in single-molecule and single-cell data. Robinson P, editor. Bioinformatics [Internet]. 2021 Dec 11 [cited 2022 May 18];37(24):4857–9. Available from: https://academic.oup.com/bioinformatics/article/37/24/4857/6298588.

27. Murata Y, Fujii A, Kanata S, Fujikawa S, Ikegame T, Nakachi Y, et al. Evaluation of the usefulness of saliva for DNA methylation analysis in cohort studies. Neuropsychopharmacol Rep [Internet]. 2019 Aug 8 [cited 2022 Jun 5];39(4):301–5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7292296/

28. Ruparell A, Inui T, Staunton R, Wallis C, Deusch O, Holcombe LJ. The canine oral microbiome: variation in bacterial populations across different niches. BMC Microbiology [Internet]. 2020 Feb 28 [cited 2022 May 18];20(1):42. Available from: https://doi.org/10.1186/s12866-020-1704-3.

29. Wallis C, Milella L, Colyer A, O'Flynn C, Harris S, Holcombe LJ. Subgingival microbiota of dogs with healthy gingiva or early periodontal disease from different geographical locations. BMC Veterinary Research [Internet]. 2021 Jan 6 [cited 2022 May 18];17(1):7. Available from: https://doi.org/10.1186/s12917-020-02660-5.

30. Hsu CH, Tomiyasu H, Lee JJ, Tung CW, Liao CH, Chuang CH, et al. Genome-wide DNA methylation analysis using MethylCap-seq in canine high-grade B-cell lymphoma. J Leukoc Biol. 2021 Jun;109(6):1089–103.

31. Ferraresso S, Aricò A, Sanavia T, Da Ros S, Milan M, Cascione L, et al. DNA methylation profiling reveals common signatures of tumorigenesis and defines epigenetic prognostic subtypes of canine Diffuse Large B-cell Lymphoma. Sci Rep. 2017 Sep 14;7(1):11591.