Summer 8-12-2014

Classification of Genotype and Age by Spatial Aspects of RPE Cell Morphology

Michael Boring

CLASSIFICATION OF GENOTYPE AND AGE BY SPATIAL ASPECTS OF RPE CELL MORPHOLOGY

by

MICHAEL A. BORING

Under the Direction of Yi Jiang

ABSTRACT

Age related macular degeneration (AMD) is a public health concern in an aging society. The retinal pigment epithelium (RPE) layer of the eye is a principal site of pathogenesis for AMD. Morphological characteristics of the cells in the RPE layer can be used to discriminate age and disease status of individuals. In this thesis three genotypes of mice of various ages are used to study the predictive abilities of these characteristics. The disease state is represented by two mutant genotypes and the healthy state by the wild-type. Classification analysis is applied to the RPE morphology from the different spatial regions of the RPE layer. Variable reduction is accomplished by principal component analysis (PCA) and classification analysis by the k-nearest neighbor (k-NN) algorithm. In this way the differential ability of the spatial regions to predict age and disease status by cellular variables is explored.

INDEX WORDS: Age related macular degeneration (AMD), Retinal pigment epithelium (RPE), K-nearest neighbor algorithm (k-NN), Classification, Principal component analysis (PCA)

CLASSIFICATION OF GENOTYPE AND AGE BY SPATIAL ASPECTS OF RPE CELL MORPHOLOGY

by

MICHAEL A. BORING

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2014

CLASSIFICATION OF GENOTYPE AND AGE BY SPATIAL ASPECTS OF RPE CELL MORPHOLOGY

by

MICHAEL A. BORING

Committee Chair: Yi Jiang

Committee: Xin Qi

Gensheng Qin

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2014

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor, Dr. Yi Jiang, for all that she has done to help me complete this work. At times she showed patience and understanding, at others encouragement and motivation. I am very grateful and happy to have worked with her.

I would also like to thank my other committee members Dr. Xin Qi and Dr. Gensheng Qin for taking the time to review and evaluate this thesis. Thanks to them also, along with Dr. Yichuan Zhao, for the instruction they have provided me at Georgia State University.

In addition I would like to thank the Department of Mathematics and Statistics for all their help, and to my many classmates who provided friendship and support along the way.

Finally I would like to thank my family. My parents for all the support they have given me over the years. And my siblings, Anne and Sam Boring, for the counsel they have provided for me in recent years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Age related macular degeneration is a common eye condition and the leading cause of vision loss in individuals 65 and older [1, 2]. Two forms of the condition are recognized and are labeled as “wet” and “dry”. The majority of individuals are diagnosed with the “dry” (non-neovascular) form which is characterized by the presence of drusens, yellowish spots that accumulate in and around the macula. The “wet” form (neovascular) is the more damaging form of the condition. It is characterized by the growth of new blood vessels beneath the retina that leak blood and other fluids causing permanent damage to the eye [3]. The primary causes of the AMD are aging and genetics. Age related macular degeneration currently affects more than 1.75 million individuals in the United States and rapid aging of the population will increase this number to an estimated 3 million by 2020 [4].

The principal site of pathogenesis for AMD is the retinal pigment epithelium. This pigmented cell layer sits just outside the retina and is attached to the choroid, a vascular layer that supplies blood to the retina. The RPE sheet is characterized by a regular generally hexagonal array of cells covering most of the surface [5]. Under AMD the RPE sheet shows collateral damage as the underlying photoreceptors are damaged in the disease state [5]. Previous research has attempted to quantitatively measure changes in the RPE cell morphology that occur in the disease state. This research has shown cell area and cell shape to be broad indicators of RPE cellular distortions caused by retinal degeneration [6]. Specifically in this analysis we seek to determine if particular spatial regions of the RPE sheet are more discriminatory than others in the detection of cell morphological change. Spatial regions are created by dividing the dissected RPE layer into zones and flaps. The flaps are labeled by the cardinal direction that natural unfold from the dissections. The zones are concentric circles emanating from the center of the eye outward. Figure 1, below, shows these spatial regions.

Spatial Regions

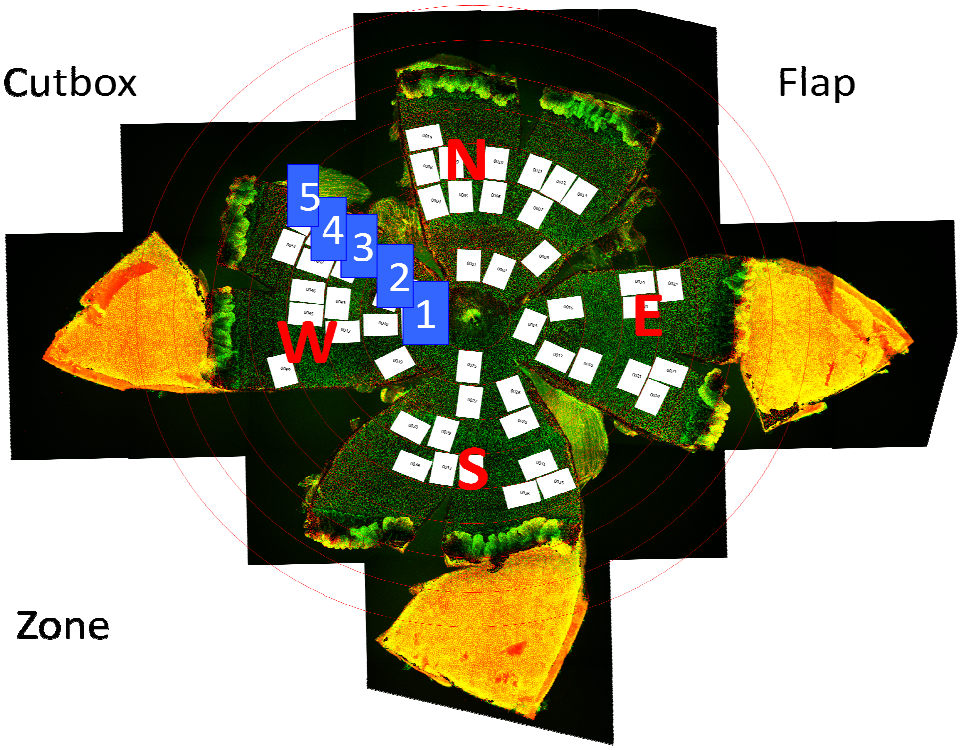

Figure 1: The spatial regions of the RPE layer. The zones (1,2,3,4,5) are concentric circles emanating from the center of the eye. The flaps are labeled as the cardinal directions (N,E,S,W). The cutboxes, segments of the layer cut for imaging and analysis, are shown as well.

1.1 Explanation of data:

The data used for this analysis came from collaborative research between John Nickerson's lab at the Emory Eye Center and work from Dr. Jiang and Dr. Qi of the Mathematics and Statistics Department at Georgia State University. RPE morphology data come from mice associated with three different genotypes of various ages that are placed in the three age groups. Table 1 lists the mice by age

and genotype. The age groups were created to evenly distribute the sample size among three groups thus representing young, middle, and old ages.

Table 1: Number of mice in various genotype and age groups

		Genotype			
		C57BL/6J	Rd10	RPE65	Total
Age (days)	$p \leq 60$	6	27	0	33
	$60 < p \leq 180$	8	26	3	37
	$p > 180$	12	5	16	33
	Total	26	58	19	103

The sample size in this analysis is the one hundred and three (103). The three genotypes serve as examples of various states of disease progression. Mice are used as a model organism for the study of this disease despite not having a macula. However, mice are a model for AMD because biological changes in the mouse retina from specific induced mutations are similar to what is found in humans with the disease, specifically in the RPE layer [8]. Associated with using mice as a model are advantages including cost-effectiveness, the ease of genetic manipulation, and accelerated life cycles [9].

The C57BL/6J genotype is the wild-type for this study. It is the most widely used inbred strain and it is a general purpose and background strain [10]. It is the control, the healthy model. The retinal degeneration 10 (rd10) mutant phenotype results from a missense point mutation in the Pde6b protein [10]. This phenotype shows retinal degeneration beginning as early as sixteen days after birth. This mutant strain is commonly used to study retinal diseases. It represents the diseased state in this analysis. The third genotype RPE65 is also a disease model but with more slow retinal degeneration and represents an intermediate phase of disease progression [10].

Previous analysis has shown quantitative differences in RPE sheet morphology can be used to accurately discriminate rd10 from C57BL/6J strains, despite age acting as a confounding variable. Functional principal component analysis (FPCA) is used to reduce the dimensions of the data while several classification methods are used to distinguish between genotype groups. These analyses show

that morphometric variables from the RPE layer can be used to accurately classify genotypes at nearly one hundred percent. This work implies that RPE sheet morphology can act as an early biomarker for the diagnosis of eye disease even at early stages when disease symptoms are subtle [6, 7].

1.2 Explanation of analysis:

In this thesis study, we extended previous analysis [6] to specifically include spatial information to investigate the potential differences in classification. We partition the RPE sheet into flaps (N,E,S,W) and zones (1,2,3,4,5). The predictive abilities of the zones are of particular interest, because cellular degeneration is often manifested more in the outer zones

We also consider three genotypes instead of the original two, with the inclusion of the RPE65 mutant. We expect genotype classification to be more difficult having to distinguish between three classes instead of two. Previous research is limited to discriminating between two age groups, young and old. We will have three age groups of young, middle, and old. One would expect area and shape variables to prove the most significant as they have done in previous research. Spatially we would expect zones further from the center to be better at classification [5]. All flaps would likely perform equally well with potentially the North flap (superior) being the most significant.

In addition, different methodology and statistical procedures are used in this study. Instead of FPCA, traditional principal component analysis is used for dimension reduction. This approach, while not as detailed at capturing variable information, is computationally efficient. Classification analysis is done by k-NN, a simple machine learning algorithm that is also computationally efficient. These methods are sufficient to reveal the important information from the data..

2 METHODOLOGY AND RESULTS

The cellular morphometric variables used in this study come from digital images of flatmounts from the RPE layer of the one hundred and three mice. This process involves dissection of the eyes,

exposure of the RPE layer, staining of the samples, cutbox samples taken from the layer, imaging under a confocal imaging system, and finally conversion to digital images and output to comma separated values files.

2.1 RPE Flatmount Technique, Staining and Imaging [6]:

The mice used for this study were euthanized with CO_2 in accordance with Emory University IACUC guideline and ARVO guideline for treatment of animals. The left eye from each mouse was extracted and the superior side (north flap) labeled with a fine point permanent ink pen. Four radial cuts were made from the center of the cornea followed by removal of the lens, iris, and retina. From the exposed retinal layer the RPE flatmounts were stained by anti-ZO-1 tight junction to allow visualization of cells. Imaging of the flatmounts was performed using a Nikon C1 confocal imaging system. Adobe Photoshop CS2 was used to stitch together images. Cut boxes were then taken from each image [6]. The digital conversion of these cut boxes was performed using Cell Profiler [13]. The two Cell Profiler modules applied to the images were Measure Object Size Shape and Measure Object Neighbors. Eighteen of the cellular variables generated for each cell using the Cell Profiler modules are used in this analysis. These cellular variables are further organized into three types to describe the kind of information they provide. The Neighbor type gives information about the relationship of the cell to the surrounding cells, where the Area and Shape types provide information about the area and shape of the cell, respectively.

Table 2: **Description of Variables**

Measurement Variable	Variable Description	Variable Type
Number of Neighbors	Number of Neighboring cells	Neighbor
Percent Touching	Percent of the object's boundary pixels that touch neighboring cells	Neighbor
First Closest Object Number	The index of the closest object	Neighbor
First Closest X Vector	Distance in the X direction to the closest object	Neighbor
First Closest Y Vector	Distance in the Y direction to the closest object	Neighbor
Second Closest Object Number	The index of the second closest object	Neighbor
Second Closest X Vector	Distance in the X direction to the second closest object	Neighbor
Second Closest Y Vector	Distance in the Y direction to the second closest object	Neighbor
Angle Between Neighbors	The angle formed with the object center as the vertex and the first and second closest object centers along the vectors	Neighbor
Form Factor	The area of the cell divided by the area of a circle with the same perimeter	Shape
Eccentricity	The eccentricity of the ellipse is calculated as the foci length divided by the major axis length	Shape
Solidity	The proportion of the pixels in the convex hull that are also in the region	Shape
Extent	The proportion of the pixels in the bounding box that are also in the region	Shape
Orientation	The angle between the x-axis and the major axis of the ellipse	Shape
Area	The actually number of pixels in the region	Area
Major Axis Length	The length (in pixels) of the major axis of the ellipse	Area
Minor Axis Length	The length (in pixels) of the minor axis of the ellipse	Area
Perimeter	The total number of pixels around the boundary of each region in the image	Area

2.2 Statistical Analysis:

The statistical analysis consists of three parts. First some graphical analysis is used to understand the nature of the variables in question and explore their discriminating potential. Second dimension

reduction is achieved using principal component analysis (PCA). Third k-NN classification is applied to the reduced dimensional data set for both age and genotype classes.

The kernel smoothing density function in R is applied to every variable for each spatial region. Some examples are shown in figures 1 (Extent) and 2 (Form Factor). Figure 2 shows separation in the density curves of C57BL/6J and rd10 suggesting that Extent can be used to easily discriminate between these two genotypes from east flap cellular data.

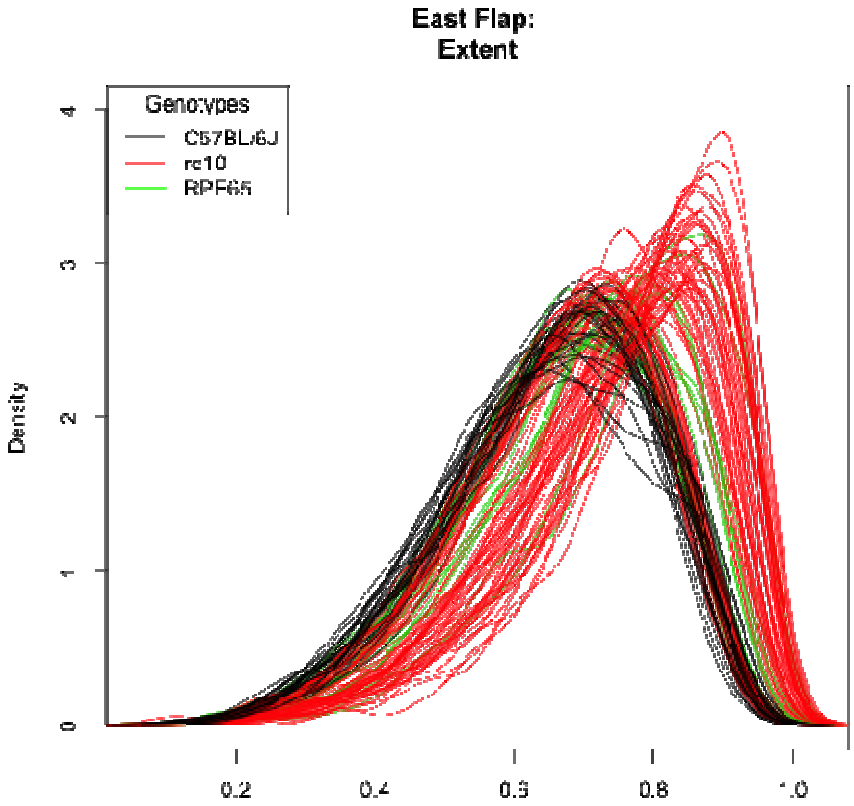

Figure 2: Kernel smoothed density graphs of the Extent variable from the East Flap. Distinction between curves suggest discrimination between C57BL/6J and rd10 is possible using this variable from the East Flap spatial region.

Figure 3 shows a more dispersed graph of densities suggesting weaker age discrimination by the Form Factor variable from North Flap data. However, there are still some regions where the two sets of

curves can be separated. These figures are examples from a more thorough graphical exploration as the first phase of the analysis.

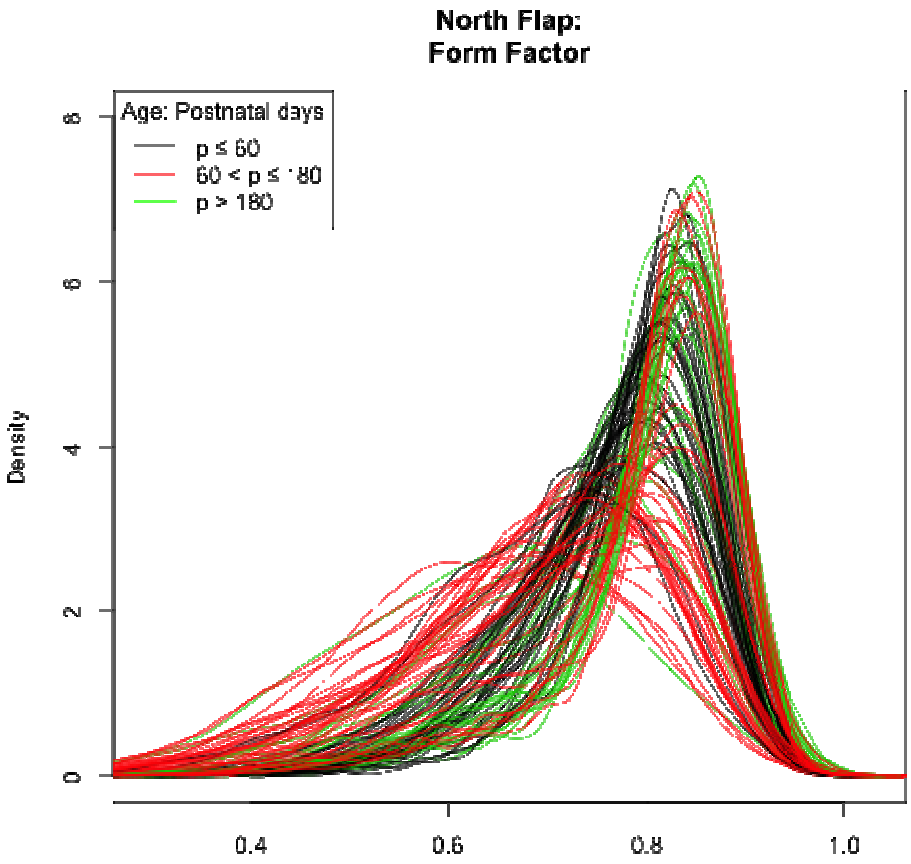

Figure 3: Kernel smoothed density graphs of the Form Factor variable.

The second part of this analysis involves extraction of quantile information and variable reduction. A vector of length sixteen is created to store the quantile data for the variable in question. The quantile data consists of the 20^{th} quantile to the 80^{th} quantile and every 4^{th} quantile increment in between. This quantile vector holds information that represents the trend in the distribution. This sixteen-dimensional vector is further reduced to two dimensions by principal component analysis (PCA). PCA is a statistical technique that reduces the dimensions of a data set while retaining the major

differences in variation among the observations [14]. Classification analysis can now be more easily achieved with a smaller dimensional data set. For all of the one hundred and three observations, the data for a particular variable from a particular spatial region is reduced to a two dimensional vector that can plotted as a point. Figures 3 and 4 show the graphical representation of this reduced data set. The first principle component score is plotted on the x-axis against the second principal component score on the y-axis. These figures further highlight classification abilities by spatial regions.

Figure 4 shows grouping of principal component scores for the Eccentricity variable. Here we can see the wild type and rd10 genotypes principal component scores clump separately from one another, suggesting distinguishing classification is possible. Every flap shows a fairly equal ability to distinguish between the groups for this variable. This figure is an example of more extensive graphical analysis done at this stage of the thesis research.

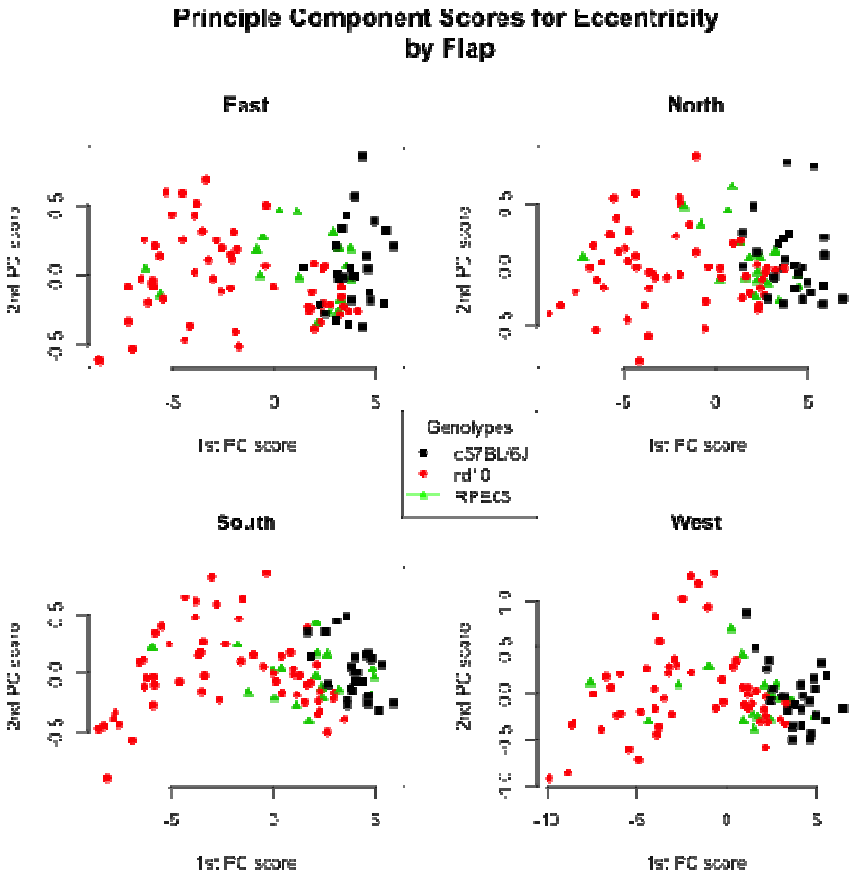

Figure 4: Plots of the first two principal component scores for the Eccentricity variable color coded by genotype. All four flaps are presented for comparison

Figure 5 (below) shows principal component scores for the Solidity variable in all five spatial zones. It can be noted that the points for each genotype separate more from one another as we move outward through the zones. In particular the points for the rd10 mice begin to spread out and move away from the other two, particularly in Zone 5. This graphical trend can be seen in many of the variables, particularly the shape variables. Figure 5 shows this trend using the Solidity variable.

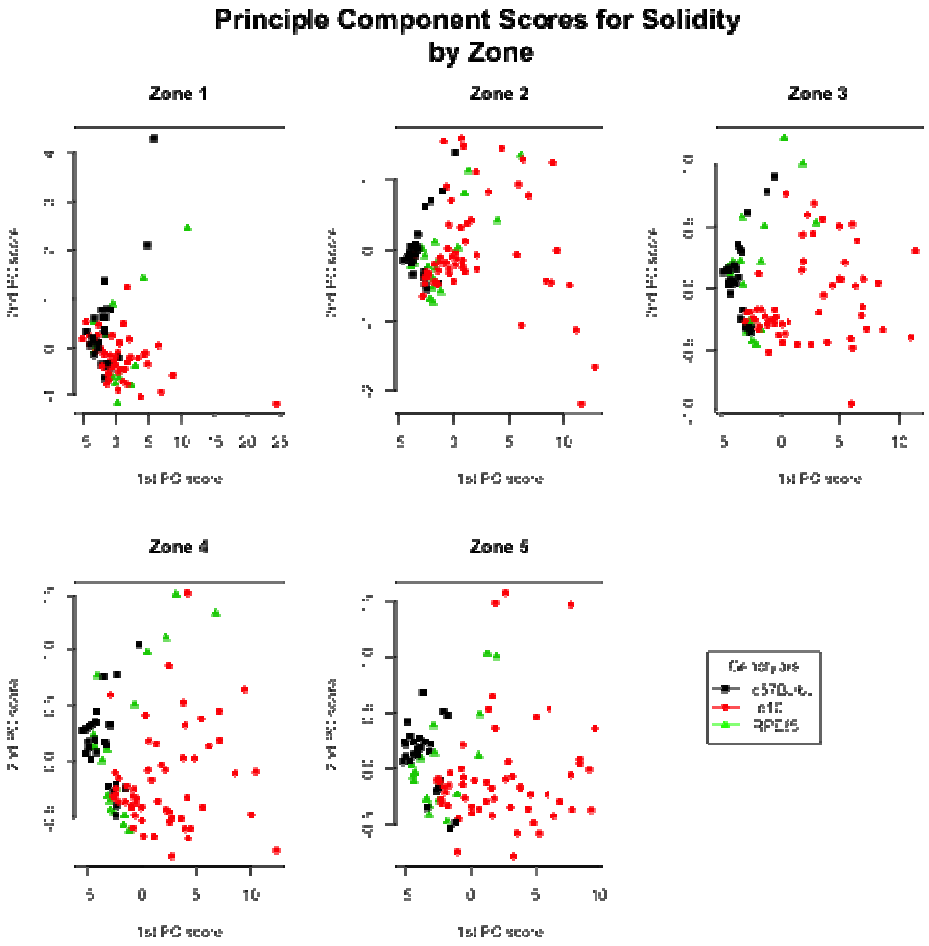

Figure 5: Plots for the first two principal component scores for the Solidity variable. Data from all five zones are plotted for comparison.

In the third part of the analysis the k-nearest neighbor (k-NN) algorithm is used for classification. K-NN is a non-parametric classification method that is among the simplest of all machine learning algorithms [12]. This algorithm works by using some distance metric, here and most commonly the standard Euclidean metric, to find the k-th nearest neighbors in the feature space to a new instance. Deciding which class this new instance belongs is done by a simple majority vote. Figure 6 provides an example.

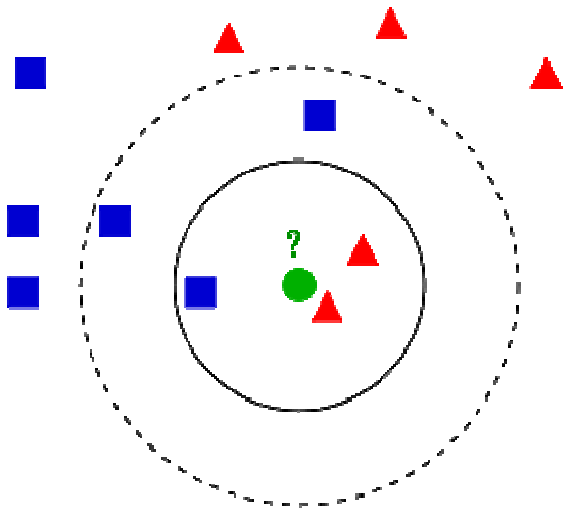

Figure 6: A graphical example of the k-NN algorithm. The green circle is data from an unknown class, classified as being from the red class when k is equal to three and the blue class when k is equal to five [12].

For each variable in each spatial region the predictive algorithm is applied in the following manner. The best parameter value for k is found using the leave one out cross validation method. For this the R package caret is used [11]. Parameter values for k are tested from one to fifteen in order to cast broad possibilities for the parameter.

Once a particular k is found, the data is split into a training group and a testing group. Once again the caret package is used to create this partition. Here there is a balanced split of the data with the attempt to preserve the overall class distribution [11]. The training set is created using eighty percent of the data, the remaining twenty percent serve as the testing set. Roughly then eighty observations are used to build the training model to predict the class output of the remaining twenty observations. The class prediction for the testing set is compared against the actual classes and a misclassification error rate is calculated. The prediction rate is taken to be the compliment of the

misclassification error rate. This procedure is looped one thousand times and the mean and standard deviation are taken to represent the overall prediction rate for the variable. This procedure, including parameter tuning, is repeated for every one of the eighteen variables in all nine spatial regions for both genotype and age group classes. The resulting three hundred and twenty four prediction rates are summarized in the tables that follow.

2.3 General Description of Results:

The mean and standard deviation of the prediction rate is the primary means of labeling a variable a good predictor. To analyze how the spatial regions differ in their overall classification abilities, a cutoff rate for what is a good predictor is established. This cutoff rate for a good predictor variable is any rate within one standard deviation of seventy percent and every rate above that level. This seventy percent level is relatively arbitrary and is chosen to more effectively demonstrate different classification abilities of the regions. Since the prediction is between three classes, a random guess would provide a prediction rate of 33%. Seventy percent is relatively good considering the nature of the data.

The next four tables describe the predictive abilities of the spatial regions. Table 3 and Table 4 show the ability of the flaps and zones to predict genotype, respectively. Table 5 and Table 6 show the age prediction of these same regions. Additional analysis is provided to explain which of the three variable types and individual variables themselves predict the most often.

Table 3

Genotype Classification Power by Flap

Flap	# Above Cutoff	Best Predictor Variable	Variable Type	Prediction Rate	Rate Standard Deviation
East	13	Form Factor	Shape	0.790684	0.057887
North	11	Extent	Shape	0.821211	0.042106
South	12	Form Factor	Shape	0.785	0.067266
West	11	First Closest X Vector	Neighbor	0.779684	0.055713

From Table 3 we can see that all flaps predict quite well and in relatively the same numbers. More than half of the eighteen variables predict above the cutoff value in every flap region. In the east and south flaps the top three variables are all of the shape type. In all flaps every one of the five shape variables are above the cutoff value. The neighbor type variables are more often not significant and the area variables are almost all above the cutoff, except for the area variable (table 2) itself, which is not over the cutoff in any of the flaps.

Table 4

Genotype Classification Power by Zone

Zone	# Above Cutoff	Best Predictor Variable	Variable Type	Prediction Rate	Rate Standard Deviation
Zone 1	8	Minor Axis Length	Area	0.885611	0.05949
Zone 2	10	Minor Axis Length	Area	0.847789	0.06069
Zone 3	9	Eccentricity	Shape	0.811	0.061358
Zone 4	9	Form Factor	Shape	0.813421	0.035791
Zone 5	7	Eccentricity	Shape	0.828947	0.144915

Table 4 presents some surprising results with Zone 5 having the least number of significant variables. Despite this the top four variables in Zone 5 were all shape variables with predictive rates above eighty percent before adding standard deviation. All four are among the top ten in the total best predictors of genotype by Zone. In general, the variables that are good predictors of genotype are higher than those from the flaps. The shape variables appear more often than those of the other types.

Table 5

Age Group Classification Power by Flap

Flap	# Above Cutoff	Best Predictor Variable	Variable Type	Prediction Rate	Rate Standard Deviation
East	2	Percent Touching	Neighbor	0.65545	0.089531
North	3	Solidity	Shape	0.66435	0.054356
South	6	Solidity	Shape	0.64195	0.101725
West	2	Form Factor	Shape	0.699895	0.078374

Age classification is much less than genotype classification. The variables that do make the cutoff are primarily of the shape type. The rates are considerable lower than in genotype classification

Table 6

Age Group Classification Power by Zone

Zone	# Above Cutoff	Best Predictor Variable	Variable Type	Prediction Rate	Rate Standard Deviation
Zone 1	0	Second Closest Object Number	Neighbor	0.592778	0.052528475
Zone 2	9	Major Axis Length	Area	0.74575	0.091639704
Zone 3	6	Eccentricity	Shape	0.73005	0.052030805
Zone 4	4	Perimeter	Area	0.7197	0.105655445
Zone 5	4	Solidity	Shape	0.684211	0.045175458

The zones have many more significant predictor variables. The prediction rates are generally higher as well. Area and shape variables once again perform the best. Zone 1 has no significant variables whereas Zone 2 has the most for any spatial region in prediction of age group. This might suggest the cells closer to the macula are more similar across ages but not across genotypes, and that ageing itself does not bring about as much significant differences in the macular RPE as disease does.

3 CONCLUSION AND DISCUSSION

The classification analysis shows that, between the two spatial regions, zones are better predictors than flaps. There are more significant variables from flaps in the classification of genotypes, meaning there are more above the established cutoff value. However, the significant variables from the zones have higher prediction rates. The variables from the zones that are significant, in particular the shape variables, recorded the highest prediction rates in the entire analysis. The flaps are relatively poor predictors of age while the zones are reasonably good, except for zone 1 which showed no significant variables.

The morphometric RPE data classifies genotype more easily than age, as indicated by the higher prediction rates and the number of variables that are found to be significant. Age was treated as a confounding variable in the previous analysis. The ability to classify genotype is the more important part of the classification, in that it more directly detects the disease state. These results support the previous finding in that the focus should be on genotype classification.

In all four classification analyses the shape variables perform the best, with the area variables second best. This supports both previous research and biological expectations. The primary differences that show up in the cellular RPE layer with the degenerative condition are in the distortion of the regular hexagonal pattern of the cells.

The zonal classification results did not show the most outer region to be the best predictor. This despite graphical evidence from plots like figure 5 showing increased separation of principal component coordinates in zone 5. In the genotype classification the zones appear to be relatively equal at discrimination, at least in the number of significant variables. It is possible the zones are equally good at classifying, but it is also possible there exists differential classification abilities that were not discovered by this analysis. Different methods, including functional principal component analysis along with various classification techniques such as linear discriminant analysis and support vector machine, may be used in addition to this analysis to find a more definitive answer. The biological plausibility that outer zones show greater variation between healthy and diseased tissue is strong enough that further analysis is needed to reach a final conclusion.

The classification rates in this analysis were lower than those in previous work. There are perhaps several reasons for this discrepancy. The additional of a third genotype, the RPE65 strain, created difficulty in discrimination. When this genotype is removed the classification rate between C57BL/6J and rd10 moves into the high nineties, in more general agreement with the previous work. It would make more sense to separately classify rd10 and RPE65 with the wild-type, as opposed to a three genotype classification.

There are numerous options for future work in this field and even with this particular set of data. Testing the wild-type genotype against each of the other two genotypes separately would be an obvious first analysis. This analysis would act more like a controlled scientific experiment, changing one variable at a time instead of two. The classification directly between C57BL/6J and RPE65 would provide information about abilities to detect even more subtle differences in RPE sheet morphology using these types of methods.

REFERENCES

[1] Klein R, Klein BE, Linton KL (1992) Prevalence of age-related maculopathy. The Beaver Dam Eye Study. *Ophthalmology* 99: 933–943.

[2] Klein R, Klein BE, Jensen SC, Meuer SM (1997) The five-year incidence and progression of age-related maculopathy: the Beaver Dam Eye Study. *Ophthalmology* 104: 7–21.

[3] O'Connell SR, Bressler N (1999) Age-related macular degeneration. In: Regillo G, Flynn Jr. H, eds. Vitreoretinal diseases: The essentials. New York: Thieme Medical Publishers. pp 213–240.

[4] Friedman, David S. , Benita J. O'Coleman, Beatriz Munoz, et al. with The Eye Diseases Research Group. Prevalence of Age-Related Macular Degeneration in the United States. 2004; *Arch Opthalmol*. 122: 564-672

[5]Chrenek MA, Dalal N, Gardner C, et al. Analysis of the RPE sheet in the rd10 retinal degeneration model. *Adv Exp Med Biol*. 2012;723:641–647.

[6]Jiang Y, Qi X, Chrenek MA, et al. Functional principal component analysis reveals discriminating categories of retinal pigment epithelial morphology in mice. *Invest Ophthalmol Vis Sci.* 2013;54:7274–7283. DOI:10.1167/iovs.13-12450

[7] Yu Jie, (2012) Classification of genotype and age of eyes using RPE cell size and shape, MS. Thesis, Georgia State University.

[8] Rakoczy Piroska Elizabeth, Dan Zhang, Terry Robertson, et al. Progressive Age-Related Changes Similar to Age-Related Macular Degeneration in a Transgenic Mouse Model. *Am J Pathol* 2002 October; 161(4): 1515–1524.

[9] Pennesi Mark E., Martha Neuringer, and Robert J. Courtney. Animal models of age related macular degeneration. *Mol Aspects Med*. 2012 August ; 33(4): 487–509. doi:10.1016/j.mam.2012.06.003.

[10] "The Jackson Laboratory." -*a Leading Genetics Research Laboratory*. N.p., n.d. Web. 07 July 2014.

[11] "The Caret Package." *The Caret Package*. N.p., n.d. Web. 07 July 2014.

[12] "K-nearest Neighbors Algorithm." *Wikipedia*. Wikimedia Foundation, 07 May 2014. Web. 07 July 2014.

[13] Lamprecht, M.R., D.M. Sabatini, and A.E. Carpenter, (2007) CellProfiler: free, versatile software for automated biological image analysis. *Biotechniques*, 2007. 42(1): p. 71-5.

[14] "Principal Component Analysis." *Wikipedia*. Wikimedia Foundation, 07 July 2014. Web. 07 July 2014.

APPENDIX

R Code:

```
## Start Code: Read in data, Create age variable, etc…


setwd("/Users/michaelboring/Desktop/Thesis Project")
filename <- read.table("filename.txt",stringsAsFactors=FALSE)
file.des <- read.csv("RPE_profile_description.csv")
d.list <- list()
for(i in 1:123){
d.list[[i]] <- read.csv(filename[i,1])
}
age <- file.des$age
file.des1 <- file.des
file.des1$agecat <- NA
file.des1$agecat[age <= 61] <- 1
file.des1$agecat[age <= 180 & age > 61] <- 2
file.des1$agecat[age > 180] <- 3
t <- which(file.des$colnumbers==28)
s <- c(10:22,24:28)
col.names <- colnames(d.list[[1]])
```

```
## Density Plots: Figures 1 and 2


# Figure 3: East Flap Extent Genotype

flap = 1

for(i in t){

data.z <- d.list[[i]]

ab <- which((data.z[,4] == levels(data.z[,4])[flap]))

d.z <- density(data.z[ab,20])

if(i == 1){

plot(d.z,col=file.des[i,2],ylim=c(0,4),main="East Flap: \nExtent",xlab="")

legend( "topleft", c("C57BL/6J","rd10","RPE65"), lty= 1,col = c('black', 'red', 'green'),title =
 "Genotypes")

}

if (i>1){

lines(d.z,col = file.des[i,2])

}

}


# Figure 4: North Flap Form Factor Age

flap = 2

for(i in t){

data.z <- d.list[[i]]

ab <- which((data.z[,4] == levels(data.z[,4])[flap]))

d.z <- density(data.z[ab,25])

if(i == 1){

plot(d.z,col=file.des1[i,5],ylim=c(0,8),main=" North Flap: \nForm Factor",xlab="")

legend( "topleft", c("p • 60","60 < p • 180","p > 180"), lty= 1,col = c('black', 'red', 'green'),title =
```

```
 "Age: Postnatal days")
}
if (i>1){
lines(d.z,col = file.des1[i,5])
}
}


## Principal Component Plots: Figures 3 and 4

# Plotting function by flap/genotype

plotPCA <- function(flap,variable){

if (flap==1){
 Q <- c(col.names[variable],"East")
}
if (flap==2){
 Q <-c(col.names[variable],"North")
}
if (flap==3){
 Q <- c(col.names[variable],"South")
}
if (flap==4){
 Q <- c(col.names[variable],"West")
}
```

```
S <- matrix(NA,nrow=123,ncol=16)
    for(i in t){
        data.z <- d.list[[i]]
        ab <- which((data.z[,4] == levels(data.z[,4])[flap]))
        qt <- quantile(data.z[ab,variable],seq(0.20,0.8,0.04))
        S[i,] <- qt
    }
 z <- file.des[,2]
 c <- which(!is.na(S[,1]))
 z <- z[c]
 S <- na.omit(S)
 class <- as.factor(z)
 pca <- princomp(S,cor=TRUE)
 pc.comp <- pca$scores
plot(pc.comp[ ,1],pc.comp[ ,2],col=class,pch = c(15,16,17)[class] ,main=Q[2],xlab="1st PC
 score",ylab="2nd PC score")


}
# Multiplot Code
par(oma = c(0,0,3,0), mfrow=c(2,2))
plotPCA(1,20)
plotPCA(2,20)
plotPCA(3,20)
plotPCA(4,20)
par(op)
mtext("Principal Component Scores for Eccentricity \n by
 Flap",side=3,line=0,font=2,cex=1.2,outer=TRUE)
```

```
op <- par(usr = c(0,1,0,1), xpd=NA)
legend( -0.425,1.7, c("c57BL/6J","rd10","RPE65"), lty= 1, pch = c(15,16,17),col = c('black', 'red',
 'green'),title = "Genotypes")



###Classification Analysis Code


library(caret)
library(class)


## Classification Functions


# Genotype by flap


ClassKnnGeno1.2 <- function(flap,variable){
S <- matrix(NA,nrow=123,ncol=16)
    for(i in t){
        data.z <- d.list[[i]]
        ab <- which((data.z[,4] == levels(data.z[,4])[flap]))
        qt <- quantile(data.z[ab,variable],seq(0.20,0.8,0.04))
        S[i,] <- qt
    }
 z <- file.des[,2]
 c <- which(!is.na(S[,1]))
 z <- z[c]
 S <- na.omit(S)
 class <- as.factor(z)
```

```
    pca <- princomp(S,scale.=TRUE)
    pc.comp <- pca$scores
    X.train <- cbind(pc.comp[,1],pc.comp[,2])
AB <- matrix(NA,nrow=15,ncol=1)
AB[,1] <- 1:15
AB <- as.data.frame(AB)
names(AB) <- "k"
fitcontrol <- trainControl(method="LOOCV")
gg <- train(X.train,class,method="knn",trControl=fitcontrol,tuneGrid=AB)
k <- gg$bestTune[1,]
v <- vector()
    for(i in 1:1000){
        set.seed(floor(runif(1,1,5000)))
        trainIndex <- createDataPartition(z, p = 0.8, list = FALSE, times = 1)
        tr <- trainIndex[,1]
        train <- X.train[tr, ]
        test <- X.train[-tr,]
        cl <- z[tr]
        model.knn <- knn(train,test,cl,k)
        v[i] <- sum(model.knn==z[-tr])/length(z[-tr])
    }
 ms <- c(mean(v),sd(v),k)
 return(ms)
}

## ClassKnnGeno1.0 is a similar function, not listed for brevity
```

```
## Classification function for Genotype by Zone


ClassKnnGeno2.0 <- function(zone,variable){
S <- matrix(NA,nrow=123,ncol=16)
    for(i in t){
        data.z <- d.list[[i]]
        ab <- which(data.z[,5] == zone)
        qt <- quantile(data.z[ab,variable],seq(0.20,0.8,0.04))
        S[i,] <- qt
    }
 z <- file.des[,2]
 c <- which(!is.na(S[,1]))
 z <- z[c]
 S <- na.omit(S)
 class <- as.factor(z)
    pca <- princomp(S,cor=TRUE)
    pc.comp <- pca$scores
    X.train <- cbind(pc.comp[,1],pc.comp[,2])
 AB <- matrix(NA,nrow=15,ncol=1)
 AB[,1] <- 1:15
 AB <- as.data.frame(AB)
 names(AB) <- "k"
 fitcontrol <- trainControl(method="LOOCV")
 gg <- train(X.train,class,method="knn",trControl=fitcontrol,tuneGrid=AB)
 k <- gg$bestTune[1,]
```

```
v <- vector()
    for(i in 1:1000){
        set.seed(floor(runif(1,1,5000)))
        trainIndex <- createDataPartition(z, p = 0.8, list = FALSE, times = 1)
        tr <- trainIndex[,1]
        train <- X.train[tr, ]
        test <- X.train[-tr,]
        cl <- z[tr]
        model.knn <- knn(train,test,cl,k)
        v[i] <- sum(model.knn==z[-tr])/length(z[-tr])
    }
 ms <- c(mean(v),sd(v),k)
 return(ms)
}
```

```
## Age Classification by Flap

ClassKnnAge1.0 <- function(flap,variable){
S <- matrix(NA,nrow=123,ncol=16)
    for(i in t){
        data.z <- d.list[[i]]
        ab <- which((data.z[,4] == levels(data.z[,4])[flap]))
        qt <- quantile(data.z[ab,variable],seq(0.20,0.8,0.04))
        S[i,] <- qt
```

```
    }
z <- file.des1[,5]
c <- which(!is.na(S[,1]))
z <- z[c]
S <- na.omit(S)
class <- as.factor(z)
    pca <- princomp(S,cor=TRUE)
    pc.comp <- pca$scores
    X.train <- cbind(pc.comp[,1],pc.comp[,2])
AB <- matrix(NA,nrow=15,ncol=1)
AB[,1] <- 1:15
AB <- as.data.frame(AB)
names(AB) <- "k"
fitcontrol <- trainControl(method="LOOCV")
gg <- train(X.train,class,method="knn",trControl=fitcontrol,tuneGrid=AB)
k <- gg$bestTune[1,]
v <- vector()
    for(i in 1:1000){
        set.seed(floor(runif(1,1,5000)))
        trainIndex <- createDataPartition(z, p = 0.8, list = FALSE, times = 1)
        tr <- trainIndex[,1]
        train <- X.train[tr, ]
        test <- X.train[-tr,]
        cl <- z[tr]
        model.knn <- knn(train,test,cl,k)
        v[i] <- sum(model.knn==z[-tr])/length(z[-tr])
```

```
    }
 ms <- c(mean(v),sd(v),k)
 return(ms)
}
## ClassKnnAge1.2 is similar function, not listed for brevity


## Classification Function for Age by Zone
ClassKnnAge2.0 <- function(zone,variable){
S <- matrix(NA,nrow=123,ncol=16)
    for(i in t){
        data.z <- d.list[[i]]
        ab <- which(data.z[,5] == zone)
        qt <- quantile(data.z[ab,variable],seq(0.20,0.8,0.04))
        S[i,] <- qt
    }
 z <- file.des1[,5]
 c <- which(!is.na(S[,1]))
 z <- z[c]
 S <- na.omit(S)
 class <- as.factor(z)
    pca <- princomp(S,cor=TRUE)
    pc.comp <- pca$scores
    X.train <- cbind(pc.comp[,1],pc.comp[,2])
 AB <- matrix(NA,nrow=15,ncol=1)
 AB[,1] <- 1:15
 AB <- as.data.frame(AB)
```

```
 names(AB) <- "k"
 fitcontrol <- trainControl(method="LOOCV")
 gg <- train(X.train,class,method="knn",trControl=fitcontrol,tuneGrid=AB)
 k <- gg$bestTune[1,]
v <- vector()
    for(i in 1:1000){
        set.seed(floor(runif(1,1,5000)))
        trainIndex <- createDataPartition(z, p = 0.8, list = FALSE, times = 1)
        tr <- trainIndex[,1]
        train <- X.train[tr, ]
        test <- X.train[-tr,]
        cl <- z[tr]
        model.knn <- knn(train,test,cl,k)
        v[i] <- sum(model.knn==z[-tr])/length(z[-tr])
    }
 ms <- c(mean(v),sd(v),k)
 return(ms)
}
## ClassKnnAge2.2 is similar function, not listed for brevity


### Create Data frame for Rates, export to CSV file
# Genotype by Flap
D<- matrix(NA,nrow=4,ncol=5)
for(i in 1:4){
D[i,1] <- 10
D[i,2] <-     i
```

```
D[i,3:5 ]<- ClassKnnGeno1.2(i,10)
}
for(j in s[-c(1,9:18)]){
D1 <- matrix(NA,nrow=4,ncol=5)
for(i in 1:4){
D1[i,1] <- j
D1[i,2] <-   i
D1[i,3:5 ]<- ClassKnnGeno1.2(i,j)
}
D <- rbind(D,D1)
}

for(j in s[-c(1:8)]){
D1 <- matrix(NA,nrow=4,ncol=5)
for(i in 1:4){
D1[i,1] <- j
D1[i,2] <-   i
D1[i,3:5 ]<- ClassKnnGeno1.0(i,j)
}
D <- rbind(D,D1)
}

D.df <- as.data.frame(D)
D.df
nam <- c("var","flap","predRate","stdev","k")
names(D.df) <- nam
```

```
write.table(D.df,file="RatesFlap.csv",sep=",",row.names=FALSE,quote=FALSE)



# Genotype by Zone
D<- matrix(NA,nrow=5,ncol=5)
for(i in 1:5){
D[i,1] <- 10
D[i,2] <-     i
D[i,3:5 ]<- ClassKnnGeno2.2(i,10)
}
for(j in s[-c(1,9:18)]){
D1 <- matrix(NA,nrow=5,ncol=5)
for(i in 1:5){
D1[i,1] <- j
D1[i,2] <-   i
D1[i,3:5 ]<- ClassKnnGeno2.2(i,j)
}
D <- rbind(D,D1)
}

for(j in s[-c(1:8)]){
D1 <- matrix(NA,nrow=5,ncol=5)
for(i in 1:5){
D1[i,1] <- j
D1[i,2] <-   i
```

```
D1[i,3:5 ]<- ClassKnnGeno2.0(i,j)
}
D <- rbind(D,D1)
}

D.df <- as.data.frame(D)
D.df
nam <- c("var","zone","predRate","stdev","k")
names(D.df) <- nam
write.table(D.df,file="RatesZone.csv",sep=",",row.names=FALSE,quote=FALSE)

# Age by Flap
D<- matrix(NA,nrow=4,ncol=5)
for(i in 1:4){
D[i,1] <- 10
D[i,2] <-     i
D[i,3:5 ]<- ClassKnnAge1.2(i,10)
}
for(j in s[-c(1,9:18)]){
D1 <- matrix(NA,nrow=4,ncol=5)
for(i in 1:4){
D1[i,1] <- j
D1[i,2] <-   i
D1[i,3:5 ]<- ClassKnnAge1.2(i,j)
}
D <- rbind(D,D1)
```

```
}

for(j in s[-c(1:8)]){
D1 <- matrix(NA,nrow=4,ncol=5)
for(i in 1:4){
D1[i,1] <- j
D1[i,2] <-   i
D1[i,3:5 ]<- ClassKnnAge1.0(i,j)
}
D <- rbind(D,D1)
}

D.df <- as.data.frame(D)
D.df
nam <- c("var","flap","predRate","stdev","k")
names(D.df) <- nam

write.table(D.df,file="RatesFlapAge.csv",sep=",",row.names=FALSE,quote=FALSE)
```

```
# Age by Zone
D<- matrix(NA,nrow=5,ncol=5)
for(i in 1:5){
D[i,1] <- 10
D[i,2] <-     i
D[i,3:5 ]<- ClassKnnAge2.2(i,10)
}
for(j in s[-c(1,9:18)]){
D1 <- matrix(NA,nrow=5,ncol=5)
for(i in 1:5){
D1[i,1] <- j
D1[i,2] <-   i
D1[i,3:5 ]<- ClassKnnAge2.2(i,j)
}
D <- rbind(D,D1)
}

for(j in s[-c(1:8)]){
D1 <- matrix(NA,nrow=5,ncol=5)
for(i in 1:5){
D1[i,1] <- j
D1[i,2] <-   i
D1[i,3:5 ]<- ClassKnnAge2.0(i,j)
}
D <- rbind(D,D1)
```

```
}

D.df <- as.data.frame(D)
D.df
nam <- c("var","zone","predRate","stdev","k")
names(D.df) <- nam

write.table(D.df,file="RatesZoneAge.csv",sep=",",row.names=FALSE,quote=FALSE)
```