**BACKGROUND AND OBJECTIVE**

The objective of this project was to investigate whether samples from the 1000 genome project taken from East Asian (EAS) and European (EUR) subjects could be clustered and classified by superpopulation using the genotype data (SNPs) available for four genes: aldehyde dehydrogenase (ALDH2), cyclic adenosine monophosphate responsive element binding protein 1 (CREB1), oculocutaneous albinism type 2 (OCA2), and solute carrier family 45 member 2 (SLC45A2). *ALDH2* is involved in the detoxification of alipathic aldehydes. Inactivating mutations in *ALDH2* are some of the most prevalent race-specific human enzymopathies, potentially contributing to a number of diseases such as diabetes, osteoporosis, cancer, and cardiovascular disease (Chen et al., 2020). One such ALDH2 mutation, ALDH2\*2, has been characterized as an East Asian-specific polymorphism (Chen et al., 2020). *CREB1* is involved in transcriptional activation. Its variants are thought to involved in the aggressiveness of human colorectal cancer (Fang et al., 2016) and susceptibility to major depressive disorder (MDD) (Li et al., 2014). The *CREB1* SNPs associated with increased susceptibility to MDD have been shown to be prevalent in Europeans but largely absent in East Asian Populations (Li et al., 2014). OCA2 and SLC45A2 are both involved in pigmentation, and different variants of each have been selected in Europe and East Asia (Edwards et al., 2010; Murray et al., 2015; Quillen et al., 2018).

**DETAIL AND DESCRIPTION OF DATA**

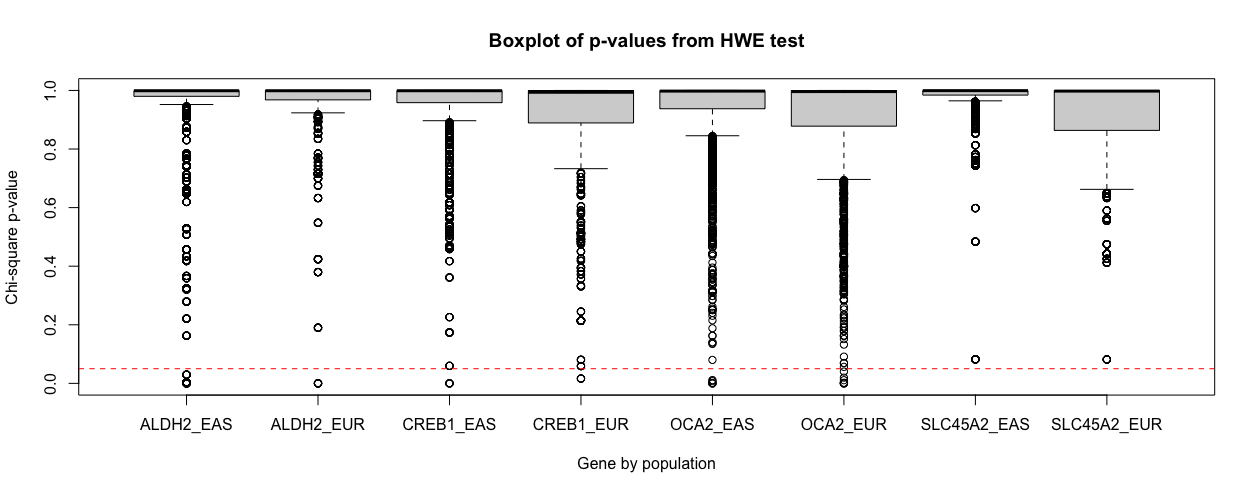


Figure 1. P-values for each gene (ALDH2, CREB1, OCA2, SLC45A2), by population (EAS, EUR), testing for Hardy-Weinberg equilibrium (HWE) using the Chi-square test are shown. The red dotted line (y = 0.05) indicates the significance threshold.

The data used in this set of analyses were downloaded from DataSlicer, which is hosted on the ensembl GRCh37 database (http://grch37.ensembl.org/Homo\_sapiens/Tools/DataSlicer?db=core). Specifically, SNP data coming from the 1000 Genomes Project sequencing effort were used. For more detailed information on the sampling process, data production, and aims of the project, it is recommended to visit the official website (http://internationalgenome.org/). There were 504 individuals in the East Asian (EAS) super-population, and 503 individuals in the European (EUR) super-population. Four genes were studied (ALDH2, CREB1, OCA2, and SLC45A2). The number of alleles in the variant call files (vcfs) ranged from 459 for the ALDH2 gene to 5583 for OCA2. For each gene and population, allele counts were obtained. The allele counts for each gene were consolidated, and alleles with frequency < 0.001 or frequency = 1 across the entire study population were removed. In particularly, alleles with frequency = 1 across both populations are uninformative for classification. The data were checked for Hardy-Weinberg equilibrium (HWE) using a chi-squared test of genotype counts (Fig. 1). For most SNPs, the p-value is greater than 0.05, upholding the null hypothesis that the population is in HWE.

**DETAIL OF DATA ANALYSIS METHODS**

**Logistic Regression**

Logistic regression with alpha = 1 (LASSO) was used to reduce the number of variables (SNPs). The data were first scaled, such that feature coefficients (importance) could be directly compared. Populations (EAS, EUR) were then encoded into integers (0, 1, respectively). The data were divided into train, validation, and test sets. The train test was used to fit the model, validation was used to tune the lambda value, and the test set was used to determine model performance. Receiving operator characteristic (ROC) curves were created for all models, and area under the curve (AUC) was determined to visualize predictive performance. Classification thresholds were optimized for the validation and test sets; confusion matrices were generated to assess the classification performance for each.

**PCA & Unsupervised Clustering**

Principal Component Analysis (PCA) was used to reduce the dimensionality of the data for downstream analysis and visualize potential clustering of samples. Exploratory data analysis was then furthered using agglomerative hierarchical clustering and k-mean clustering. Agglomerative hierarchical clustering was performed on the original variant counts using binary distance and complete linkage which has previously performed well for two well-separated clusters of binary data (Tamasauskas et al., 2012). Agglomerative hierarchical clustering was also performed performed on the top three principal components using euclidean distance and average linkage (to reduce sensitivity to outliers) as previously employed for variant analysis (Spuesens et al., 2016). K-means, which was chosen over k-medoids because the former scales better to large datasets, was performed on the top 3 PCs. The optimal number of clusters was determined to be 2 using the gap statistic method and elbow method. Nevertheless, k-means clustering was performed for multiple values of k (2,3, and 4) each with 25 different seeding algorithms in order to look for cluster stability and previously unobserved structure in the data. To further examine cluster stability, the aforementioned was repeated for k=2 on perturbed observations by dividing the data into 5 folds and excluding one fold from each clustering analysis.

**Decision Trees**

To perform classification, we implemented decision trees and bagging. Each of the decision trees used the genotypes of the 1007 individuals for the 3672 SNPs from four different genes to generate branches. To reduce computation time both the SNPs with very rare (<0.001) and very common (>0.999) genotype frequencies were removed as they likely won’t be of a significant benefit during classification. Additionally, a portion of the data was set aside as training data, and the remainder of the observations were used during the prediction stage. Bagging was used to increase accuracy and ensure the overall model was a suitable classifier on new data. For this analysis, 300 different trees were generated, and the average predictive performance was represented by a receiver operating curve (ROC) and the resulting area under the curve (AUC).

**RESULTS**

**Logistic Regression**

The logistic regression model fitted on the training data reported a lambda.min of 0.03092 (AUC = 0.9997; # features = 26), and a lambda.1se of 0.04923 (AUC = 0.9996; # features = 11). Overall, there was very good model performance for lambda values between log(-1) and log(-5) (Fig. 2A). The fitted model was applied to the validation set using lambda.min or lamba.1se to find that both yielded AUCs of 0.9999 (Sensitivity = 0.9952, Specificity = 0.9948, Balanced accuracy = 0.9950; Fig. 2B), using an un-optimized classification threshold of 0.5. An optimal value of 0.47 was determined on the validation set, which did not change performance. The model using lambda.1se incorporated a smaller number of features. Given the overall good performance of both models, the more parsimonious model using lambda.1se was selected for further analysis. The fitted model with lambda.1se was used to predict the test set. Near-perfect classification was achieved (AUC = 1, Sensitivity = 1, Specificity = 0.9911, Balanced accuracy = 0.9955; Fig. 2C), using the optimized classification threshold of 0.47. Although the inclusion of feature interactions was considered, given the already-near-perfect performance, the analysis was concluded with single features. **X5**\_33951693\_207.0 was determined to be the most important features by far (coef = -2.3193), followed by **X12**\_111778178\_313.0, **X15**\_27926499\_6975.0, and **X2**\_207578440\_1487.0 (coef = 0.2041; -0.1226; 0.1101; respectively).

**A B C**

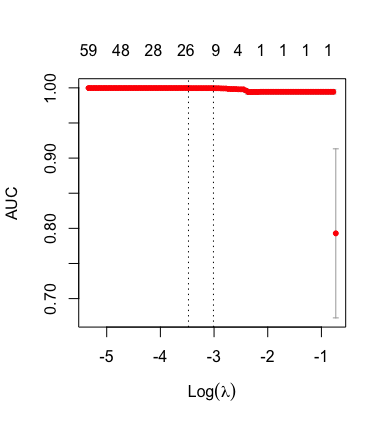
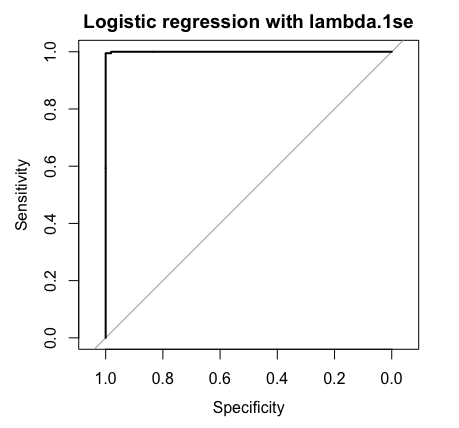
** ** 

Figure 2. **A**) Area under the curve as a function of log(lambda) for the logistic regression model fit to training data. The number of incorporated features is shown above the plot. **B**) Receiving operator characteristic curves (ROCs) for the validation set using lambda.1se, and **C**) test set using lambda.1se.

**PCA & Unsupervised Clustering**

**A B**

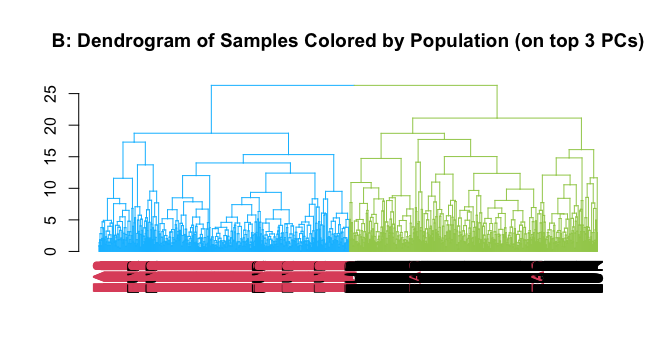
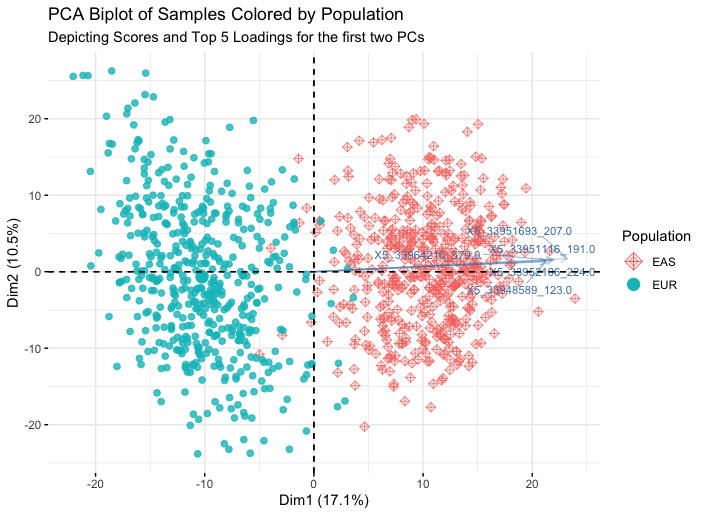


Figure 3. **A**) PCA biplot of Samples coloured by Population. **B**) Dendrograms of samples coloured by population for agglomerative hierarchical clustering performed on the top 3 principal components.

When agglomerative clustering was performed on the original counts, the resulting dendrogram had to be split into a greater number of branches in order to obtain clusters predominantly composed of a given population. When agglomerative clustering was performed on the principal components, the dendrogram could be cut into more simply into two branches each of which was predominantly composed of one of the two populations (98.2% EAS, and 96% EUR, respectively). For k-means clustering, k=2 largely clustered samples by population with an overall sensitivity of 0.98 and specificity of 0.99. The cluster containing predominantly EUR samples was split when k was increased to 3, and the cluster containing predominantly EAS samples was split when k was further increased to 4. When plotted on the first two PCs, k=2 appeared to result in the best separation between clusters. These findings were corroborated by the optimal number of clusters predicted by the gap statistic and elbow methods, and also by what was observed in the dendrogram for top 3 PCs. The k=2 clusters were also found to be highly stable; when the analysis was repeated on 5 sets of perturbed observations, for all sets sensitivity was between 0.984 and 0.986, specificity was between 0.991 and 0.995, and the number of misclassifications for each class was similar (7-8 for EAS and 2-4 for EUR).

**Decision Trees**

Bagging generated 300 trees, examples of which are visualized in Fig. 4A. Additionally, a singular tree was generated for comparison purposes. Interestingly, many of these trees had 3 or more SNPs when a tree generated without bagging only used 1 SNP for classification. After rounding the average bagging predictions to fit within the binary classification, bagging resulted in 4 misclassifications. To visualize prediction abilities an ROC plot was used, as seen in Fig. 4B. Bagging had a great trade-off between sensitivity and specificity with an AUC value of 0.9998. Overall, given the nature of the data, a singular decision tree would be an acceptable way to classify this data; however, the bagging trees provided a higher AUC value and therefore are considered a better classifier. Bagging is also a good practice to implement as it improves accuracy and reduces the chances of overfitting to a given set of data.

**A B**

Diagram

Description automatically generatedChart

Description automatically generated

Figure 4. **A**) Example decision tree generated during the bagging procedure, **B**) ROC plot for bagging trees which shows the trade-off between sensitivity and specificity for the final predictions.

**CONCLUSIONS AND DISCUSSION**

An advantage of logistic regression via lasso includes built-in variable selection. On the contrary, other methods like as clustering may involve dimensionality reduction via principal components analysis, which adds another layer of complexity. Logistic regression, an extension of linear regression, is also relatively easy to implement and visualize. It is also less computationally expensive than other algorithms studied (k-means, decision trees). The algorithm is robust when data is linearly separable, and presents helpful information on feature importance via model coefficients.

Overall, the predictive performance of logistic regression via lasso was very good. Accuracy, sensitivity, and specificity were above 0.990 on all train, validation, and test sets - suggesting that 1) the SNPs analyzed for the selected genes (ALDH2, CREB1, OCA2, and SLC45A2) are indeed informative for classifying individuals to super-populations, and 2) the data are linearly separable, making logistic regression is an effective method for handling this data. If performing similar types of analyses in the future, logistic regression via lasso could be recommended as a starting algorithm. Finally, the **X5**\_33951693\_207 SNP in the SLC45A2 gene was the most important feature by far, reinforcing the finding that this SNP is highly discriminatory for individuals in the EAS super-population.

Decision trees with bagging are a very simple and effective way to perform classification as they use a simple yes/no format to classify new data into given groups. This is a very intuitive way to classify, especially genotype data. Genotypes are one of three possibilities: homozygous dominant, heterozygous, or homozygous recessive. This provides the decision tree with very clear boundaries for branches and makes identification of influential SNPs more straight forward. Unfortunately, with very large datasets, which is common with genotypes, bagging is very computationally expensive. This being said, bagging provided good results and in theory would be a great application for population clustering based on SNPs. Overall, decision trees and bagging are a great tool, but logistic regression and other clustering techniques are able to identify similar patterns within the data, with far less computational effort.

**REFERENCES**

**SUMMARY OF WORK**

Task 1: Background and objective

Task 2: Detail and description of data

Task 3: Detail of data analysis methods

Task 4: Summary of results

Task 5: Conclusions and discussions

Task 6: Appendix

Task 7: Report compilation

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Natalie | Heather | Kerry |
| 1 | Editing (10%) | Editing (10%) | Writing (80%) |
| 2 | Writing (80%) | Editing(10%) | Editing(10%) |
| 3 | Writing + editing (33%) | Writing + editing (33%) | Writing + editing (33%) |
| 4 | Writing + editing (33%) | Writing + editing (33%) | Writing + editing (33%) |
| 5 | Writing + editing (33%) | Writing + editing (33%) | Writing + editing (33%) |
| 6 | Editing (10%) | Writing (80%) | Editing (10%) |
| 7 | Editing (20%) | Compilation (40%) | Compilation( 40%) |