**DETAIL AND DESCRIPTION OF DATA**

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 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 using a chi-squared test of genotype counts.

**LOGISTIC REGRESSION**

**Detail of data analysis methods**

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.

**Summary of results**

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. 1). 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), 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), 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**

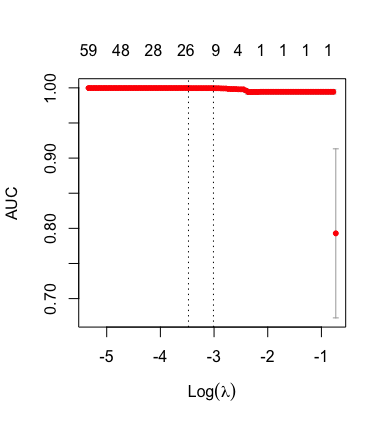
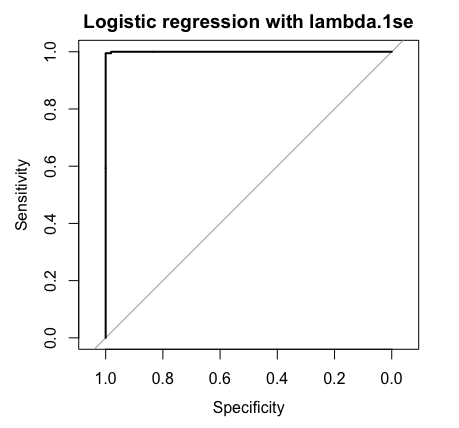
** ** 

Figure 1. **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.

**Conclusions and discussions**

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. 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.**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 (20%) | Writing (60%) | Editing (20%) |
| 7 | Editing (10%) | Main compilation (80%) | Editing (10%) |