BINF\*6970: Assignment 4

Group B: Zhikang Chen, Hoc Tran, Adedamola Ogunbona, Brittany Mac Intyre  
April 10, 2020

**Background and Objective**

Ovarian cancer rates are observed to be higher in populations where genetic propensity for tallness is also high (1). To examine the potential relationship between height and increased risk of ovarian cancer we isolated three genes to use as a proxy for: ovarian cancer (BRCA1), height (HMGA2), and general growth for human cancer and height (IGF1). We then chose two ethnicities which differ in average adult female height, 156.1 cm for East Asians (EAS) and 163.2 cm for Europeans (EUR) (**Table 1**)(2), representing ‘short’ and ‘tall’ population phenotypes, respectively. Women of European ancestry have the highest incidence rates of ovarian cancer, whereas women of East Asian populations have a lower, although gradually increasing incidence rates (3). The objective of this analysis was to attempt to classify individuals into these two ethnic groups based on their genotype information for the three representative genes: breast cancer type 1 (BRCA1), insulin-like growth factor 1 (IGF1), and high mobility group AT-hook 2 (HMGA2), to determine if variants in these genes can capture previous findings in ovarian disease states and height phenotypes related to EAS and EUR populations.

**Table 1**. Average female height by population

|  |  |  |  |
| --- | --- | --- | --- |
| Superpopulation | Country | Average Female  Height (cm) | Superpopulation  Average (cm) |
| East Asian | China | 158.0 | 156.1 |
| Japan | 158.0 |
| Vietnam | 152.2 |
| European | Finland | 165.3 | 163.2 |
| Italy | 162.5 |
| Great Britain | 161.9 |
| Spain | 163.0 |

**Dataset**

Genotype information for each gene, BRCA1, IGF1, and HMGA2, was collected from data obtained in Phase 3 of the 1000 Genomes Project (4) via the Data Slicer on the Ensembl database (5) by querying the chromosome and positional information for each gene. Populations for the East Asian (EAS) superpopulation included: Han Chinese in Beijing, China (CHB), Southern Han Chinese (CHS), Chinese Dai in Xishuangbanna, China (CDX), Japanese in Tokyo, Japan (JPT), and Kinh in Ho Chi Minh City, Vietnam (KHV). The European superpopulation (EUR) included populations from: Toscani in Italia (TSI), Finnish in Finland (FIN), British in England and Scotland (GBR), and the Iberian population in Spain (IBS). The Utah Residents (CEU) population was excluded from the EUR superpopulation dataset in this study as individuals in this population no longer lived in Europe to control for potential variation in the change in geographical location.

Filtering of the resultant variant call format (.vcf) files was performed with Bash Unix Shell command language with the bcftools module (6) on the Compute Canada Graham cluster (**Appendix A**). Six genotype .vcf files, one for each gene and superpopulation, were filtered to include only biallelic variants, and for allele frequencies equal to or greater than 0.001 for the corresponding superpopulation. Filtered single nucleotide polymorphism (SNP) counts totalled: 612 and 684 for BRCA1, 505 and 486 for IGF1, and 937 and 951 for HMGA2, for EAS and EUR, respectively.

**Analysis Methods**

Each .vcf file was read into RStudio (v. 1.2.1335) and the genotype information for each variant and each individual in the population was extracted. Genotype information was changed to numeric such that two reference alleles = 0, one reference allele and one alternative allele = 1, and two alternative alleles = 2. This resulted in matrices with SNPs as columns and individual IDs from each population as rows, which were then used to train the classifier models.

Two classification methods, elastic net and random forest, were used to classify individuals into EAS or EUR superpopulations. Individual SNP genotype (GT) data for each gene was split into 70% for the training set and 30% for the testing set. Training and test sets were indexed to ensure both methods were trained and tested on the same data. Both classifiers were trained using all SNPs as predictor variables from the labelled and filtered training set. Predictive strength of each classifier was tested on the testing data and measured with confusion matrices and corresponding AUC values.

The elastic net regularized logistic regression model was tuned on 0.1 < α < 0.9 and 0 < λ < 0.5., and the best pair of tuning variables for each GT matrix was selected by maximizing the cross-validation accuracy. The random forest method was tuned on default tuning parameters.

**Results**

Predictive strength for each classifier on the testing set was measured in confusion matrices for elastic net (**Table 2**) and random forest (**Table 3**) which show similar results for each gene between methods.

**Table 2**: Elastic net test set confusion matrices

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Predicted | | | | | |
|  |  | BRCA1 | | IGF1 | | HMGA2 | |
|  |  | EAS | EUR | EAS | EUR | EAS | EUR |
| True Class | EAS | 146 | 6 | 130 | 22 | 149 | 3 |
| EUR | 10 | 110 | 18 | 102 | 8 | 112 |

**Table 3**: Random forest test set confusion matrices

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Predicted | | | | | |
|  |  | BRCA1 | | IGF1 | | HMGA2 | |
|  |  | EAS | EUR | EAS | EUR | EAS | EUR |
| True Class | EAS | 143 | 9 | 140 | 12 | 148 | 4 |
| EUR | 24 | 96 | 41 | 79 | 15 | 105 |

A summary of prediction results is shown in **Table 4**. Sensitivity and specificity were highest for HMGA2 models over other model types for both methods. Area under the curve (AUC) was highest for HMGA2 models (0.9910 and 0.9821), followed by BRCA1 models (0.9876 and 0.9670) and then IGF1 models (0.9402 and 0.9029) for elastic net and random forest, respectively.

**Table 4**: ROC curves for elastic net and random forest classifiers on each gene

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Elastic Net | | | Random Forest | | |
|  | BRCA1 | IGF1 | HMGA2 | BRCA1 | IGF1 | HMGA2 |
| Best tuning parameters | α = 0.25  λ = 0.01070 | α = 0.33  λ = 0.00022 | α = 0.15  λ = 0.08985 | mtry = 26 | mtry = 24 | mtry = 33 |
| Sensitivity | 0.9166667 | 0.8500000 | 0.9333333 | 0.8000000 | 0.6583333 | 0.8750000 |
| Specificity | 0.9605263 | 0.8552632 | 0.9802632 | 0.9407895 | 0.9210526 | 0.9736842 |
| AUC | 0.9876 | 0.9402 | 0.9910 | 0.9670 | 0.9029 | 0.9821 |

The receiver operating characteristic (ROC) curves for each gene show similar trends between the two classifiers (**Figure 1**). Models for the HMGA2 gene (blue) are the closest to perfect classification (upper right corner of the graph) and show higher sensitivity and reflect a greater AUC, 0.9910 and 0.9821 for elastic net and random forest, respectively (**Table 4**). BRCA1 models (red) have comparable, but slightly lower AUC values of 0.9876 for elastic net and 0.9029 for random forest. Models for IGF1 (green) were the least accurate for each method, with an AUC of 0.9402 for elastic net and 0.9029 for random forest.

**a)**

**b)**

A close up of a map

Description automatically generated A close up of a map

Description automatically generated

**Figure 1**. ROC curves for Elastic Net **(a)** and Random Forest **(b)** for gene models BRCA1 (red), IGF1 (green), and HMGA2 (blue).

**Conclusion**

All gene models for both elastic net and random forest classifiers performed well with an AUC > 0.90, each with a higher specificity than sensitivity. The classifying models with the best predictive ability were those for the HMGA2 gene which had the greatest AUC scores with an average of 0.9865 between both methods, only slightly higher than the average BRCA1 model score of 0.9773, indicating both HMGA2 and BRCA1 represent genotype variations that are unique to both EAS and EUR superpopulations. Thus, genetic variation similarities in the HMGA2 and BRCA1 genes found within these populations are sufficient to distinguish them, as seen in the classification models used here. The genotype information of the entire population (of both males and females) reflects the differences in observed disease and phenotypic differences, in ovarian cancer rates and average height, respectively, for adult females. Although the IGF1 models showed lower predictive ability (average AUC = 0.9215) than the other gene models, it still represents a good reference for predicting the correct ethnicity for each individual. Overall, all three genes examined here, BRCA1, IGF1, and HMGA2 represented a good reference for classifying individuals into their associated superpopulation and shows concordance with observed traits between EAS and EUR populations.

**References**

1. Dixon-Suen SC, Nagle CM, Thrift AP, Pharoah PDP, Ewing A, Pearce CL, *et al*. (2018) Adult height is associated with increased risk of ovarian cancer: a Mendelian randomisation study. *Br J Cancer*, 118:1123–1129. doi: 10.1038/s41416-018-0011-3.
2. Average Height By Country Population. (2020). Retrieved 2020-04-09, from http://worldpopulationreview.com/countries/average-height-by-country/
3. Coburn SB, Bray F, Sherman ME, *et al*. (2017) International patterns and trends in ovarian cancer incidence, overall and by histologic subtype. *Int J Cancer*. 140(11):2451–2460. doi:10.1002/ijc.v140.11.
4. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. *Nature* 526:68–74.
5. Cunningham F, Achutlan P, Akanni W, Allen J, Amode MR, Armean I, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, *et al.* (2018)Ensembl 2019. *Nucleic Acids Research*, 47:D745-751. doi: 10.1093/nar/gky1113.
6. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27 (21):2987–2993.

**Appendix A**

#Unix shell command line code

#VCF files used for this analysis can be accessed on Compute Canada at: /scratch/bcmac/stats.

module load bcftools

#Filtering shown for East Asian (EAS) population only, process repeated for each gene for the European population (EUR)

#Filter for biallelic SNPs only, with maximum alleles at 2 per SNP

bcftools view --max-alleles 2 BRCA1EAS > filt\_biallelic\_BRCA1EAS.vcf

bcftools view --max-alleles 2 EAS\_IGF1 > filt\_biallelic\_EAS\_IGF1.vcf

bcftools view --max-alleles 2 EAS\_HMGA2 > filt\_biallelic\_EAS\_HMGA2.vcf

#Filter for allele frequency >= 0.001 for the corresponding population

bcftools view -i 'EAS\_AF>0.0009' filt\_biallelic\_BRCA1EAS.vcf > filt\_ba\_freq\_BRCA1EAS.vcf

bcftools view -i 'EAS\_AF>0.0009' filt\_biallelic\_EAS\_IGF1.vcf > filt\_ba\_freq\_EAS\_IGF1.vcf

bcftools view -i 'EAS\_AF>0.0009' filt\_biallelic\_EAS\_HMGA2.vcf > filt\_ba\_freq\_EAS\_HMGA2.vcf

**Appendix B**

See attached R code.