Estimating genetic selection with unified approach and studying its effect on human phenome

by

Nikhil Mark Lakra

MT19218

Under the Supervision of

Dr. Vibhor Kumar

Indraprastha Institute of Information Technology Delhi

December, 2021

Estimating genetic selection with unified approach and studying its effect on human phenome

by

Nikhil Mark Lakra

MT19218

Submitted
in partial fulfilment of the requirements for the degree of
Master of Technology

to

Indraprastha Institute of Information Technology Delhi

December, 2021

Certificate

This is to certify that the thesis titled **“Estimating genetic selection with unified approach and studying its effect on human phenome”** being submitted by **Nikhil Mark Lakra** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Dr. Vibhor Kumar **December, 2021**
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Acknowledgement

With a deep sense of gratitude, I express my sincere thanks to my supervisor, Dr. Vibhor Kumar, for his valuable guidance, patience and constant encouragement during my thesis work from the beginning till the end of my work.

I express my heartiest gratitude towards my friends who helped a lot in supporting me, and lastly, I sincerely express my gratitude to my parents for constant encouragement throughout this process.

Abstract

Natural selection is a mechanism of evolution, and genetic mutations form the basis of this evolutionary mechanism. With humans migrating out of Africa and inhabiting different parts of the earth, different populations have been under different selection pressures leading to adaptations in specific genes. Several methods have been developed to identify these sites of selection, but there is a varying level of uniformity among them. In this study, a unified approach for identifying sites of selection was applied to 17 different populations belonging to the Phase III of the 1000 Genomes Project. Combining different methods which capture different signs of selection, we identified several single nucleotide polymorphism (SNP) under selection using a machine learning model. We studied SNP and the populations showing high probability scores for selection, relating the SNPs with their significant eQTLs and phenotype to gain novel insights into the adaptations provided by them. A much in-depth analysis of these SNPs could not only help in understanding the history of human evolution but also generate hypotheses for better drug selection based on ethnicity.

Table of Contents

List of Figures

List of Tables

1. Introduction

The advancement in whole-genome sequencing has enabled us to understand the evolutionary forces which have acted upon the modern human populations. Signals for positive selection in genomes can be identified by detecting regions of reduced variation, Linkage Disequilibrium (LD) patterns, shift allele frequency etc. The 1000 Genome Project is an initiative to sequence the genomes of participants of different ethnicities from around the globe. The SNP data from these various populations can be used to understand and identify the phenotypic differences between the populations.

There are pre-existing methods to detect strong selection in eukaryotic populations based on LD patterns and Cross Likelihood-Ratio (CLR) tests. Still, these methods fail when the presumptions are not met, e.g. constant population size (Koropoulis A et al., 2020). Machine Learning-based methods can be used to train and learn on these patterns and the summary statistics shown by the regions under selection to classify the SNPs.

1.1) Human Migration

Modern humans originated somewhere in the African continent between 200 to 300 thousand years ago. Two models have been proposed to explain the migration of modern humans to the Australiasian and Eurasian regions (Posth C et al. 2016). According to the first model, there was a single major migration out of the African continent. According to the second model, multiple out of Africa migrations occurred, first proposed by Lahr and Foley (Mirazo´n M et al. 1998). The first migration happened around 100 - 90 kya from the southern route through South Arabia to Southeast Asia. Then a second migration event from the Northern route, through the Levant, leading to the population of the rest of Asia and Europe around 40 and 50 kya.

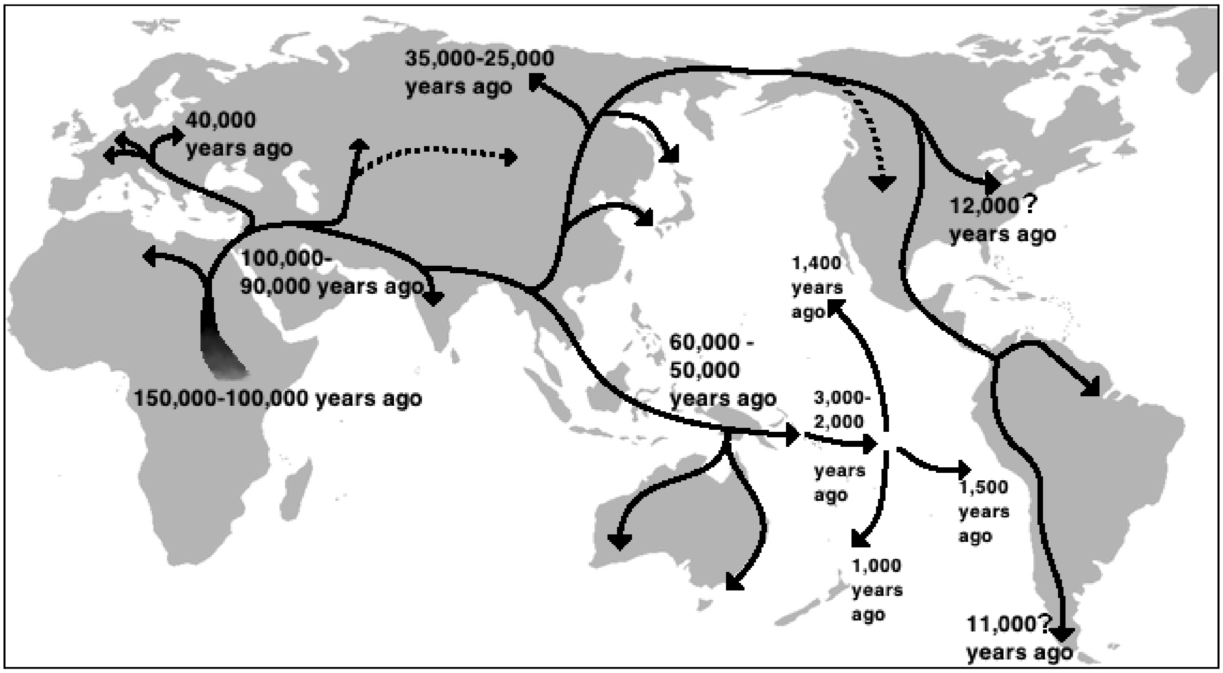


Figure 1: Out of Africa migration (source: Ephert 2011; CC-BY-SA-3.0)

1.2) Genetic drift and natural selection

Genetic drift is the change in allele frequencies in a population over the generations due to random events. Unlike natural selection, genetic drift does not depend on an allele's beneficial effects or harmful effects. Although genetic drift happens in populations of all sizes, its effects tend to be stronger in small populations (Gahan P. 2005).

The bottleneck effect is one of the extreme examples of genetic drift. The population size decreases drastically by the effects of an extreme event such as a natural disaster. The allele frequencies in the new population might not have the same frequencies as the prior population. The smaller population size also makes it more susceptible to random events causing genetic drifts (Gahan P. 2005).

Natural selection, specifically positive natural selection, is when the frequencies of the beneficial alleles increase in the population. Suppose the presence of an allele in an organism leads to an increase in its chances of survival and reproduction. It will be more likely that the organism will produce more offspring, which will inherit those traits than the organisms in that same population that do not have that beneficial trait. Through successive generations, the selected allele and nearby “hitchhiking” alleles become much more common; this event is known as a selective sweep (Schaffner S & Sabeti P, 2008).

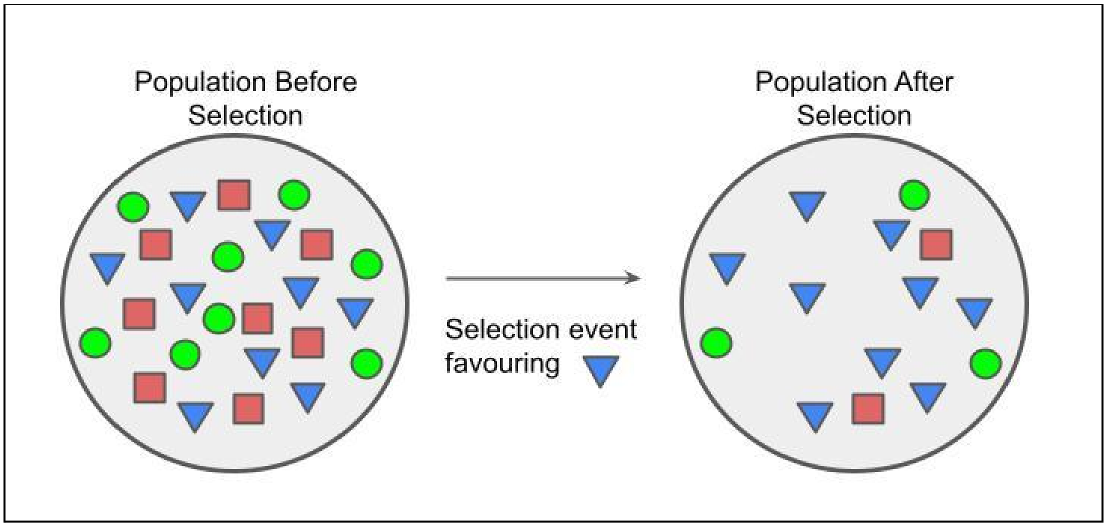

Population Before Selection
Selection event favouring ▼
Population After Selection

Figure 2: Shows a simplistic representation of different phenotypes present before and after selective an event; different shapes refer to different phenotypes that might be present in a population. Natural selection provides phenotypes with advantageous traits higher survival rates.

The prevalence of the lactase (LCT) gene in some humans, particularly those of European ancestry, shows signs of natural selection in humans. Most mammals lose their ability to process lactose when they move on from the juvenile stage into adulthood. In most individuals of European descent, the lactase activity persists through their adulthood. There are two clashing theories for lactase prevalence, the culture-historical hypothesis and the reverse-cause hypothesis (Gerbault P et al. 2011). The first proposes that lactase prevalence developed and was consequently selected after milk production and dairy consumption spread. The second hypothesis argues that only populations whose frequency of lactase prevalence was high enough adopted dairying (Gerault P et al. 2011).

1.3) Linkage Disequilibrium

Linkage disequilibrium is the non-random association of an allele to two or more loci in a given population. LD is measured by calculating the difference between the observed frequency and the expected frequency for a particular set of alleles. A particular set of alleles are said to be in linkage disequilibrium if the observed frequency is much higher or lower than the expected frequency (independent and random association).

1.4) GWAS

Genome-wide association studies are used to identify a relationship between single nucleotide polymorphisms (SNPs) to a disease or a trait. The studies are done by comparing the genomes of different individuals and looking for specific genetic markers that relate to the presence of a disease.

The 1000 Genome project was an international initiative to catalogue the most common single-nucleotide variants in the human genome from anonymised samples to be used as a resource to study rare diseases occurring in small populations (Altshuler D et al., 2012). In 2015 the study concluded phase III, having sampled over 2500 individuals from 26 populations.

Super Population	ID	Population	Total Individuals
Africa	GWD	Gambian in Western Division – Mandinka	113
Africa	ACB	African Caribbean in Barbados	96
Africa	ESN	Esan in Nigeria	99
Africa	MSL	Mende in Sierra Leone	85
Africa	LWK	Luhya in Webuye, Kenya	99
Africa	YRI	Yoruba in Ibadan, Nigeria	108
Africa	ASW	African ancestry in SW USA	61
East Asia	JPT	Japanese in Tokyo, Japan	104
East Asia	CHB	Han Chinese in Beijing, China	103
East Asia	KHV	Kinh in Ho Chi Minh City, Vietnam	99
East Asia	CDX	Chinese Dai in Xishuangbanna, China	93
East Asia	CHS	Han Chinese South, China	105
Europe	CEU	Central European ancestry from Utah	99
Europe	GBR	British from England and Scotland	91
Europe	TSI	Toscani in Italia	107
Europe	FIN	Finnish in Finland	99
Europe	IBS	Iberian populations in Spain	107

Table 1: The information for the 1000 genome population used in the study.

1.5) Previous work done

PopHumanScan is a database catalogue that compiles and annotates all candidate regions under a selection of the human genome. It contains eight summary statistics for the 22 non-admixture human populations of Phase III of the 1000 Genome Project (Murga-Moreno J, 2019).

Machine Learning has changed several existing fields, but only a few studies have used machine learning and applied it to population genetics. Recently Koropoulis et al. compared existing selection detection algorithms to various ML classifiers, which showed that in cases where the underlying assumptions of these algorithms no longer hold true, they perform worse, resulting in higher false-positive rates (Koropoulis et al., 2020). The study tested classifiers such as K-nearest neighbours (KNN), Logistic Regression (LR), Support Vector Machines (SVM), Random forests (RF) and Naive Bayesian Classifiers, and their results show that even trivial models based on these classifiers can outperform existing methods in certain situations.

Pybus et al. developed and applied hierarchical boosting, a machine learning framework. Four different boosting functions were sequentially considered within a hierarchical decision tree; the boosting function itself is a linear regression function of the scores of individual positive selection tests (Pybus et al., 2015). The framework was applied to three reference human populations from The 1000 Genome Project to generate a genome-wide classification map of selective sweeps in genomic regions (Pybus et al., 2015). Schrider, D. R., & Kern, A. D. trained a support vector machine using empirical data using functional and nonfunctional parts of the genome, which was able to identify purifying sweeps with high accuracy of ~88% on populations of phase I of 1000 genome project(Schrider D R & Kern A D et al.; 2015). Compared to previous methods, we used previously reported SNPs as our training data and used statistical scores that do not require a genetic map as features, applying it to 17 reference populations from the phase III of 1000 Genome Project to predict the selection in the SNPs reported in the GWAS Catalog. Since our summary statistics do not require a genetic map we can train our model to be SNP specific and report selection directly on the SNPs rather than reporting genomic regions like previous methods.

2. Machine Learning for Detecting Selection

There have been multiple methods for estimating selection. However, their results are not consistent with each other. Moreover, due to the lack of enough number of positive and negative control sets, it has been a non-trivial task to benchmark them properly or judge their consistency. Here we collected a set of positively selected genomic locations in humans and used it to develop and test our model for estimating selection. Our approach has been to include the knowledge base of previously proposed methods and selection identification to develop a unified approach.

2.1) Data Collection

The SNP data which is used for analysis was obtained from Phase III of The 1000 Genome Project. The 1000 Genome has SNP data from 2504 individuals, 26 different populations and five different geographic regions. Only biallelic SNPs with a minor allele frequency higher than 0.05 were considered for this study. The data for SNPs under selection was obtained from PopHumanScan (https://pophumanscan.uab.cat/), which was used to train the machine learning classifier.

Gwas catalogue was founded by National Human Genome Research Institute (NHGRI) in 2008; it is a publicly available collection of curated genome-wide SNP-trait associations. The Catalog contains over 6000 GWAS comprising more than 138000 variant-trait associations from > 4000 publications.

2.2) Statistics description

The selection tests performed in this study are based on the Long Haplotypes (iHS, RSB, XP-EHH, nS_L, XP-nS_L), allele frequency spectrum (DDAF) and population differentiation (F_{ST}, PBS). By using the different approaches to identify selection in SNPs associated with the human genome, a much more complete and comprehensive method can be achieved to identify signatures of selection occurring in the different populations. Site frequency spectrum based methods such as Tajima’s D and Fay and Wu’s H were not used in this study.

2.3) Long Haplotype Scores

Extended Haplotype Homozygosity (EHH) is used to determine long-range haplotypes; it is a reliable way to detect regions under positive selection pressure (Bomba L et al., 2015). There is an overall reduction in the haplotype diversity if a region is under positive selection pressure.

Cross population EHH (XP-EHH) and the ratio of (site-specific)EHHS between populations (Rsb) are pairwise is the log-ratio of population scores computed on two populations. Integrated haplotype homozygosity score (iHS) is the log-ratio of ancestral EHH to derive EHH of an SNP marker.

The number of segregating sites by length (n_{SL}), and XP-n_{SL} being the cross-population statistic of n_{SL}, is another method used to identify the increasing haplotype homozygosity which occurs when a region is under positive selection. The method is based on the distribution of fragment lengths between sites with the distribution of the number of segregating sites between all pairs of chromosomes and is based on the ratio of haplotype homozygosity for ancestral allele and derived alleles (Ferrer-Admetlla A et al. 2014).

2.4) High Allele Frequency

A relatively higher derived allele frequency might be a sign of positive selection in a population (Wang G and Speakman J, 2016). The Difference in Derived Allele Frequency (DDAF) between the population of interest and the other population was carried out in two ways. Global DDAF is the difference between AF of the population and AF in all of the 17 populations used in the study. Local DDAF is calculated between AF of the population and the AF of each of the super populations (Africa, East-Asia & Europe).

2.5) Population Differences

Fixation Statistic (F_{ST}) is a summary statistic used to study the population structure in two populations. There are many interpretations and descriptions(Cockerham 1969; Nei 1973; Hudson et al. 1992) of F_{ST} ever since it was first introduced by Sewell Wright(Wright, 1949). According to a study(Bhatia G, 2013), Hudson's estimator was recommended when used to study the populations in the 1000 Genome project and HapMap3 project. Hudson's estimator is not sensitive to the ratio of population sizes between two populations and also does not overestimate the F_{ST} value when compared to Winer and Cockerham estimator or Nei's estimator.

Population Branching Statistic (PBS) is similar to F_{ST}, but the distance is calculated from two other populations, whereas in the case of F_{ST} it is a pairwise comparison. Both F_{ST} and PBS have been shown to detect recent natural selection. A study conducted comparing exomes of Danes, Chinese Hans and Tibetians, PBS was able to detect signs of high altitude adaptation in Tibetians; the divergence between the Chinese Hans and the people of Tibet is estimated to be only 2750 years ago (Yi X et al., 2010).

2.6) Training and test data

A dataset was made using selected SNPs reported in the PopHumanScan, as the positive data. Random SNPs were matched and picked with the same global allele frequencies and from the same chromosomes as the positive data to create the negative data. Isolation forest, which is an unsupervised classifier used to identify outliers, was used to further remove these SNPs from the negative training data if the Isolation forest classifier classified it as an outlier. Selection in SNPs is much less expected than being non-selected SNP; a rough ratio of 1:4 was maintained between the positive and negative data.

2.7) Dealing with imbalanced data

An imbalance in the training and test data was created because it is much less likely for an SNP to be under selection than not being under any selection pressure. Class imbalance in the training data can lead to unwanted biases in the model, e.g. a model that only predicts the majority regardless of the input features of the sample. To avoid this, metrics such as Area Under Curve of Receiver Operating Characteristic (AUC ROC), balanced accuracy and Matthew's Correlation Coefficient (MCC) were used to benchmark the model.

After the classifier predicts the class of the sample, it can fall into 4 different categories:
True positive (TP): Correctly predicted positive samples predicted as positive.
True negative (TN): Correctly predicted negative samples predicted as negative.
False-positive (FP): Wrongly predicted negative samples predicted as positive.
False-negative (FN): Wrongly predicted positive samples predicted as negative
.

Evaluation metrics used in the study

Matthew's correlation coefficient is a reliable statistic when dealing with binary classification and class imbalance. MCC calculates Pearson's product-moment correlation coefficient between the predicted values and the actual values (Chicco D & Jurman G, 2020).

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}}$$

The receiver operator characteristic curve is plotted between True positive rate on the y-axis vs False positive rate on the x-axis with varying classification thresholds.
True Positive Rate (TPR) = TP / (TP + FN)
False Positive Rate (FPR) = FP / (FP + TN)

Stratified K-Fold cross-validation

Cross-validation is a resampling method used to evaluate the machine learning model. K-fold cross-validation method splits the data into K subsets, 1 of the subsets is used as the validation data, and the rest k-1 subsets are used as the training data for the model. Stratified K-fold cross-validation maintains the original class ratio during the resampling of the subsets.

Probability Calibration

In cases with unbalanced data, the model, on average, predicts a higher probability for the major class than the minor class. Since the Random forest algorithm predicts a class label, the probabilities generated by it are not calibrated to handle class imbalance. Platt scaling is a simple linear regression-based method that is used to transfer the calculated scores from the uncalibrated model to the calibrated probability scores.

3. Results and Case Studies

3.1) Baseline scores of various classifiers

From 9 other classifiers, Random Forest was picked as the final classifier for the model since it performed best (MCC score of 0.876 and balanced accuracy score of 0.942).

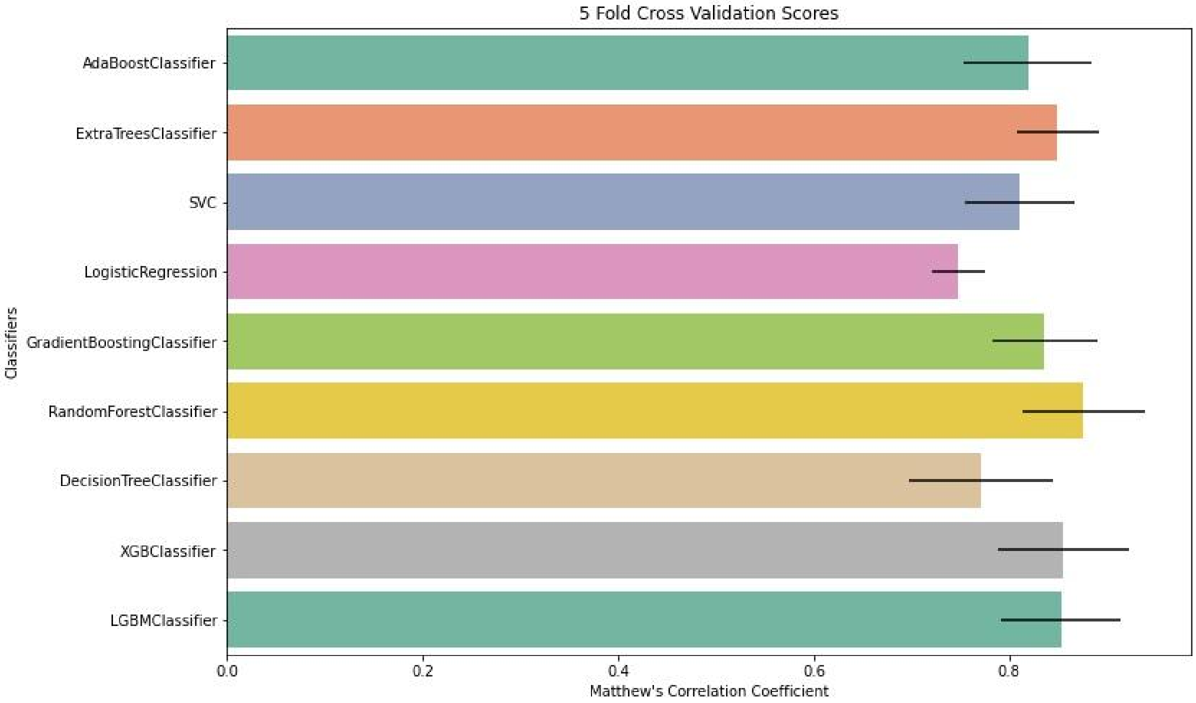


Figure 3: Plot showing the Avg. Mathew's Correlation Coefficient Score for each of the classifiers.

Classifiers	Avg. MCC Score (5FCV)	Standard Deviation
RandomForestClassifier	0.876	0.063
XGBClassifier	0.855	0.067
LGBMClassifier	0.853	0.061
ExtraTreesClassifier	0.849	0.042
GradientBoostingClassifier	0.836	0.054
AdaBoostClassifier	0.819	0.066
SVC	0.811	0.056
DecisionTreeClassifier	0.771	0.074
LogisticRegression	0.748	0.027

Table 2: Table showing the Avg. Mathew's Correlation Coefficient Score for each of the classifiers.

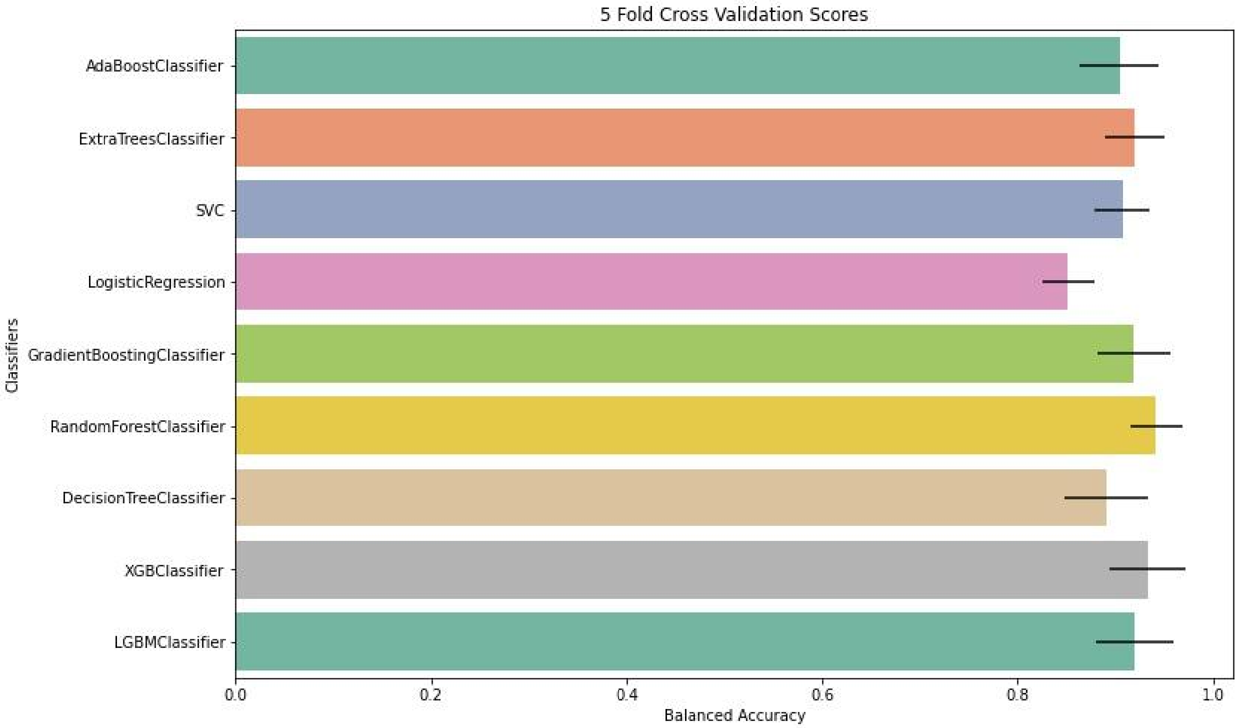


Figure 4: Plot showing the Average Balanced Accuracy Score for each of the classifiers.

Classifiers	Avg. MCC Score (5FCV)	Standard Deviation
RandomForestClassifier	0.942	0.027
XGBClassifier	0.933	0.039
ExtraTreesClassifier	0.920	0.031
LGBMClassifier	0.920	0.039
GradientBoostingClassifier	0.919	0.037
SVC	0.907	0.028
AdaBoostClassifier	0.904	0.041
DecisionTreeClassifier	0.891	0.042
LogisticRegression	0.852	0.027

Table 3: Table showing the Avg. Balanced Accuracy Score for each of the classifiers.

3.2) Feature Selection & Hyperparameter Tuning

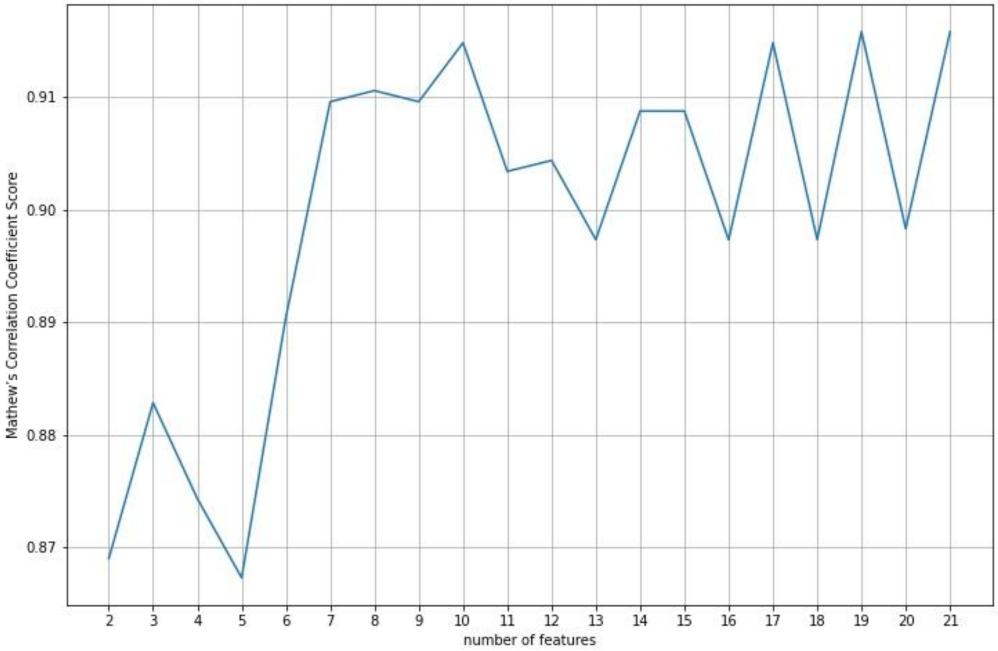


Figure 5: The graph shows Mathew's Correlation Coefficient scores for m features ($2 \le m \le 21$).

Number of features	MCC Score
2	0.84404
3	0.85786
4	0.84935
5	0.84233
6	0.86546
7	0.88459
8	0.88557
9	0.88459
10	0.88982
11	0.8784
12	0.87937
13	0.87234
14	0.87853
15	0.88376
16	0.87233
17	0.88982
18	0.86615
19	0.89081
20	0.87331
21	0.89078

Table 4: The Matthew's Correlation Coefficient scores for m features ($2 \leq m \leq 21$).

To increase the accuracy of the classifier mutual information was used to rank the features from most informative to the least informative. Matthew’s Correlation Coefficient scores were calculated using the Random Forest Classifier in sklearn using python. The top m features ($2 \leq m \leq 21$) were used successively, and their performance was evaluated on unseen test data. From the 21 features, when 19 features were used produced the highest MCC score of 0.89081.

The search space for hyperparameters of the Random Forest Classifier was tuned using the RandomizedSearchCV, and the final values were adjusted using GridSearchCV for the highest MCC score; both the methods are implemented in the sklearn python package. After feature selection and hyperparameter tuning the balanced accuracy, the score remained the same at 0.94 but the MCC score increased to 0.89.

Parameter	Value
n_estimators	90
max_features	sqrt
min_sample_leaf	1
min_sample_split	2
criterion	Gini
class_weight	balanced
random_state	42

Table 5: The final parameters used for the Random Forest Classifier.

3.3) Results

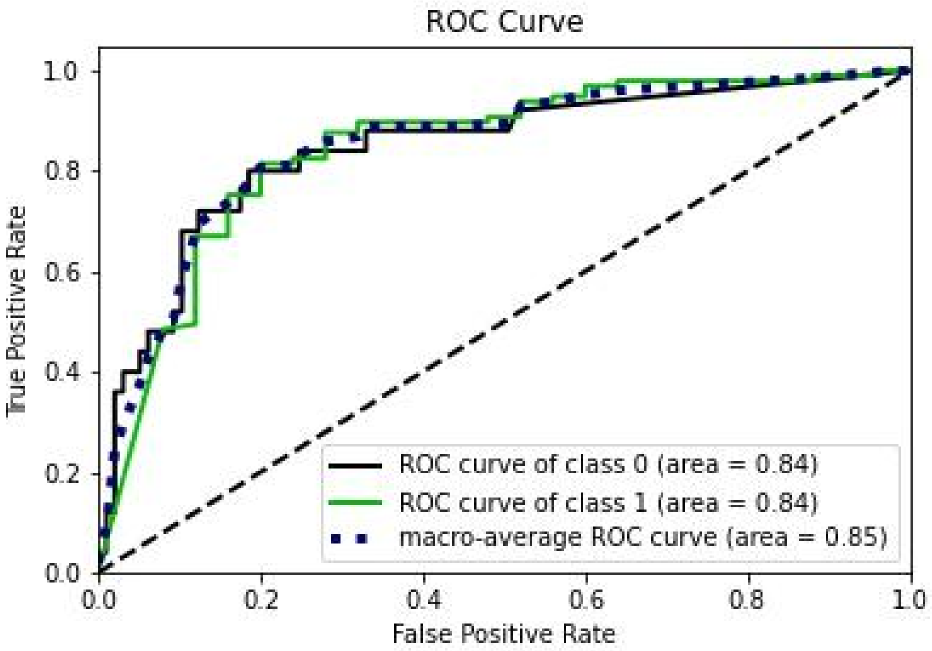
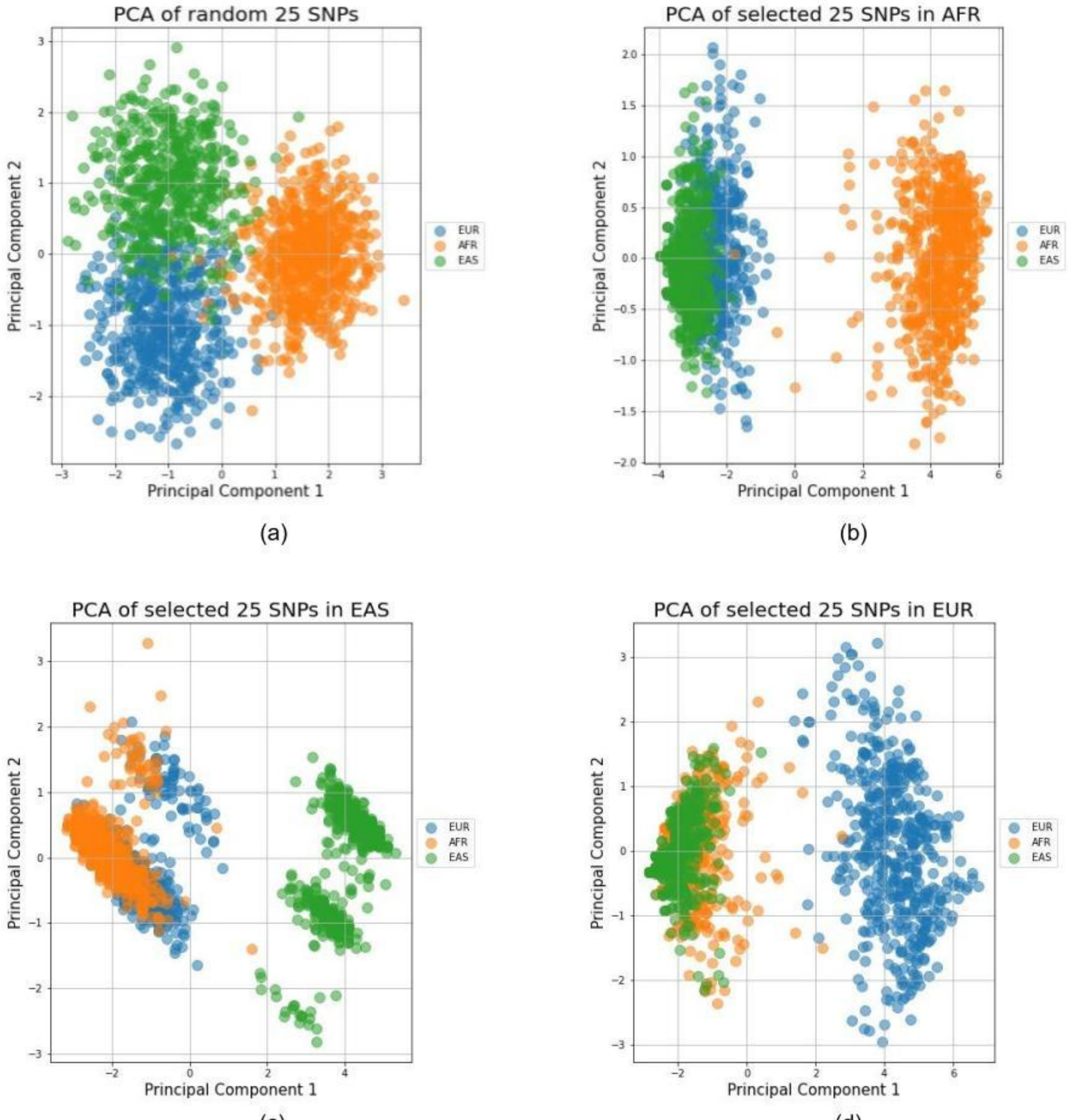


Figure 6: Shows the plot for ROC Curve for the Random Forest Classifier on the test data.

The final classifier showed a weighted recall score of 0.82. The Hierarchical boosting method used in the study by Pybus et al. showed a recall of 0.985 on a dataset that simulated recent selection on a single population with central European ancestry. The random forest classifier used here was trained and tested on an empirical dataset consisting of 17 different populations from 3 different super populations.

From the 1,710,625 data points (100625 SNPs and 17 populations) given to the classifier obtained from the GWAS Catalog. A total of 988 unique SNPs had a prediction probability above 0.9. Europe had 66 unique SNPs, East-Asia and Africa had similar counts at 452 and 476 unique SNPs, respectively.



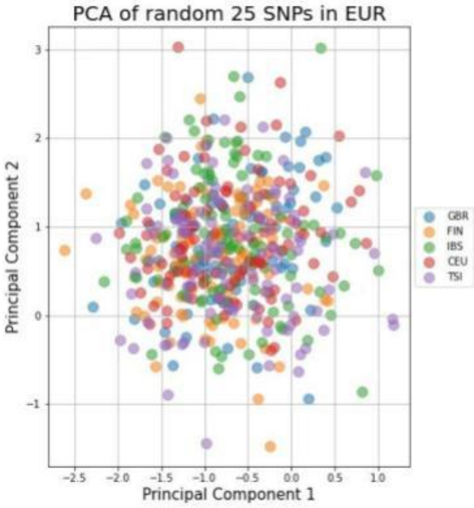
(a)

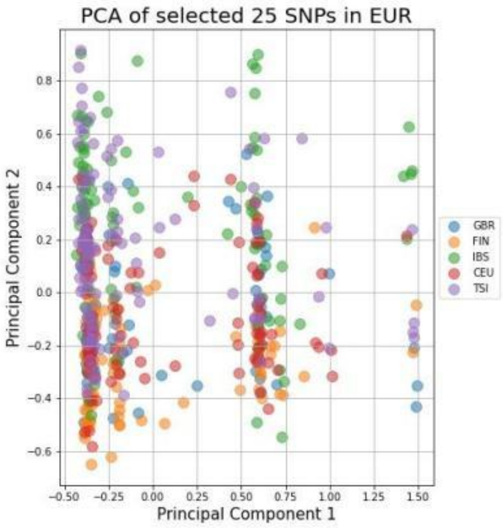
(b)

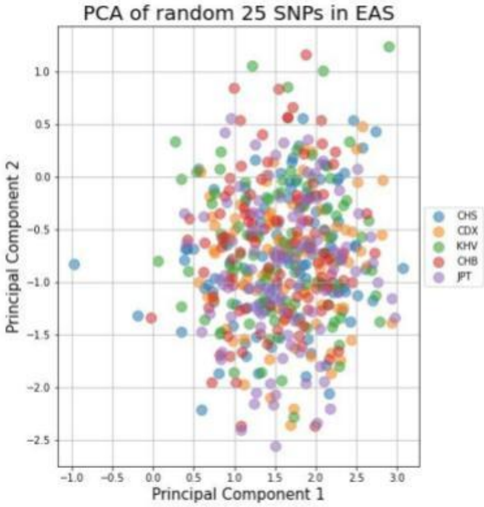
(c)

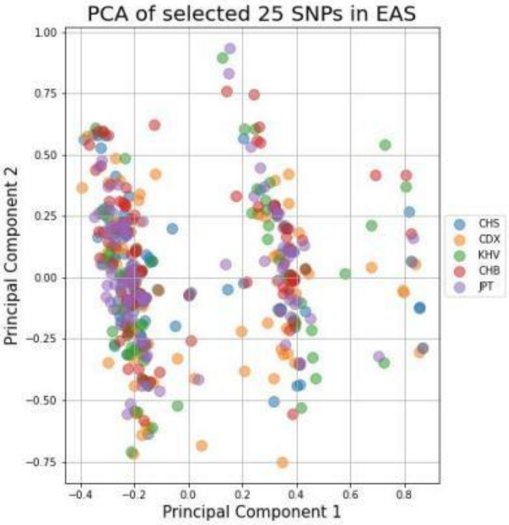
(d)

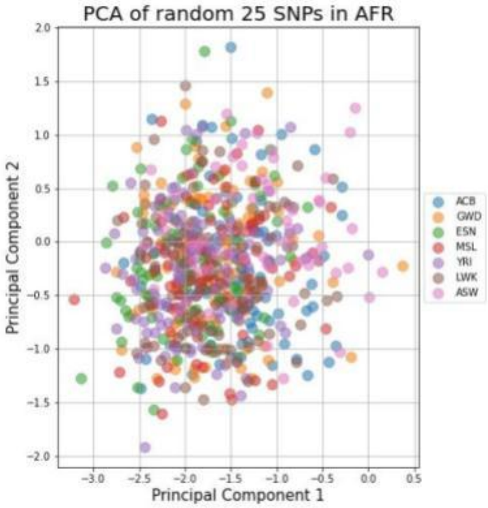
Figure 7: (a) PCA plots of random 25 SNPs. (b) PCA plot of 25 SNPs with the highest probability of selection in AFR. (c) PCA plot of 25 SNPs with the highest probability of selection in EAS. (d) PCA plot of 25 SNPs with the highest probability of selection in EUR.


PCA of random 25 SNPs in EUR
Principal Component 2
Principal Component 1
3
2
1
0
−1
−2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0
GBR
FIN
IBS
CEU
TSI


PCA of selected 25 SNPs in EUR
Principal Component 2
Principal Component 1
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.50 −0.25 0.00 0.25 0.50 0.75 1.00 1.25 1.50
GBR
FIN
IBS
CEU
TSI


PCA of random 25 SNPs in EAS
Principal Component 2
Principal Component 1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
CHS
CDX
KHV
CHB
JPT


PCA of selected 25 SNPs in EAS
Principal Component 2
Principal Component 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−0.4 −0.2 0.0 0.2 0.4 0.6 0.8
CHS
CDX
KHV
CHB
JPT


PCA of random 25 SNPs in AFR
Principal Component 2
Principal Component 1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−3.0 −2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5
ACB
GWD
ESN
MSL
YRI
LWK
ASW

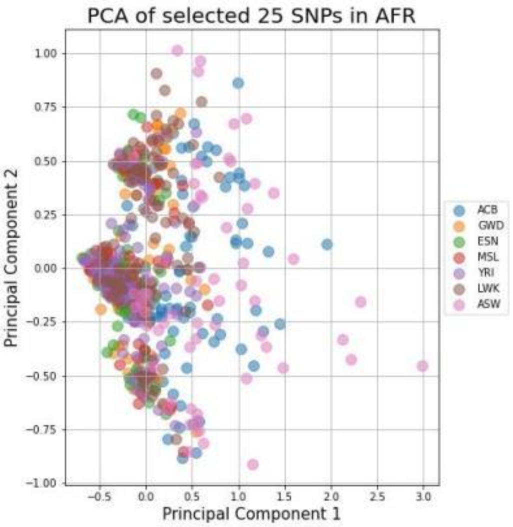

PCA of selected 25 SNPs in AFR
Principal Component 2
Principal Component 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
ACB
GWD
ESN
MSL
YRI
LWK
ASW

Figure 8: PCA plots of random 25 SNPs vs 25 SNPs with the highest probability of selection in each super population.

When we plot the highest-scoring SNPs from each of the super populations and compare them to the PCA plot of random SNPs,, thereby reducing we can observe the drift in the random SNPs. Population-specific plots with selected SNPs show a completely different pattern than when compared to the random SNPs; the populations with the selected SNPs cluster together away from the other populations on which SNPs are not selected.

3.4) Case-Studies

Warfarin maintenance dose in East-Asian population

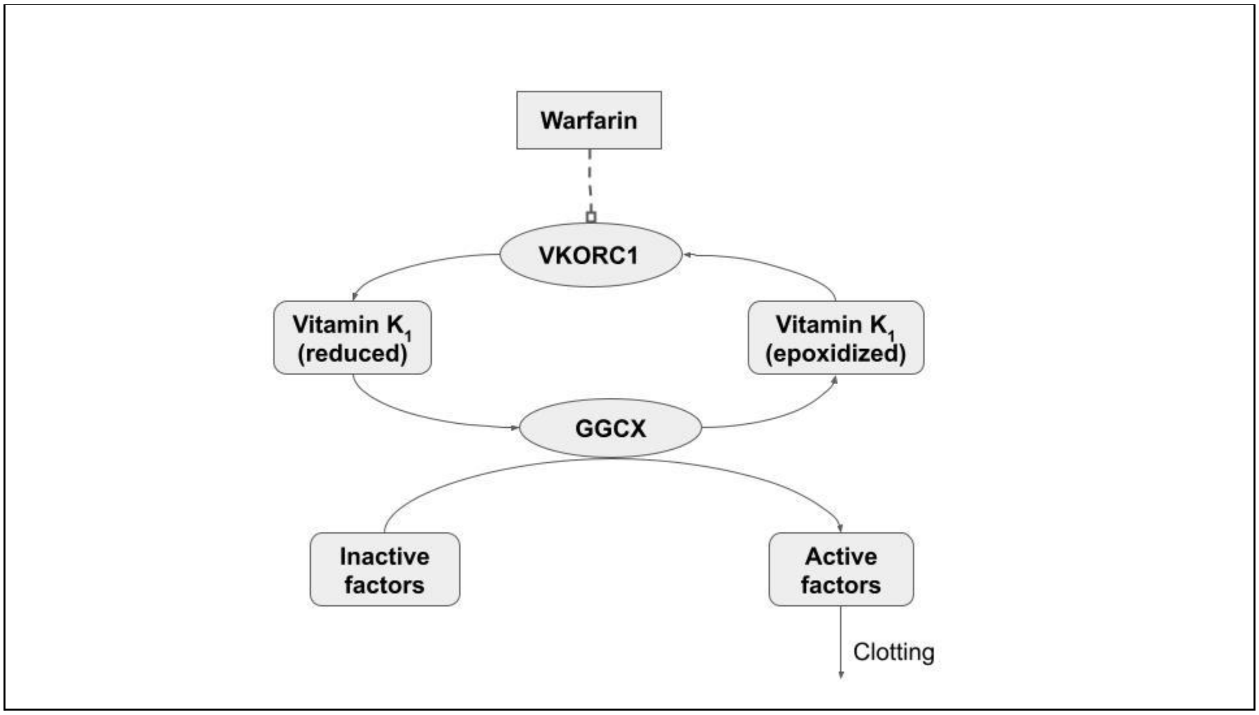


Figure 9: Shows a simplistic representation of warfarin interacting with the Vitamin K_1 clotting pathway.

Gene	SNP	P-value	Tissue
VKORC1	rs10871454	3.0e-47	Liver
VKORC1	rs9923231	7.7e-51	Liver

Table 6: Significant eQTLs associated with rs10871454 and rs9923231 reported in the GTEx database.

Warfarin is an anticoagulant that is used as a blood thinner to prevent cases of stroke in the case of atrial fibrillation or deep vein thrombosis. Warfarin functions as an antagonistic inhibitor in the vitamin K clotting pathway. Warfarin inhibits the action of Vitamin K epoxide reductase complex and thereby reduces the activation of clotting factors present in the blood. The classifier detected signals of selection in 3 sites associated with the VKORC1 gene which is present on chromosome 16. SNPs rs9923231 and rs10871454 are both C -> T mutations in the VKORC1 gene, whereas rs749671 is G -> A mutation. The presence of these mutations reduces the activity of the VKORC1 gene. Therefore required warfarin maintenance dosage in the East Asian population is much lower when compared to the Caucasian population (Lam M and Cheung B, 2012; Li S et al., 2015).

Selection of eye colour and hair colour in European populations.

The SNPs rs12916300, rs916977 & rs12913832 were found to be selected in the British, Central European and Finnish populations; these SNPs are associated with eyes and hair colours. rs12916300, rs916977 & rs12913832 are mutations in the HERC2 (Hect Domain And RCC1-Like Domain-Containing Protein 2) gene, all of which lie on chromosome 5. rs12916300 is C -> T at position 28165345 of chromosome 5; rs916977 is T -> C mutation at position 28268218 and rs12913832 is A -> G mutation at position 28120472. It has been reported to have a strong association with blue eyes in the Icelandic population(Sulem P et al., 2008). Even though the HERC2 gene does not participate in the pigmentation pathway, it is known to disrupt the expression of OCA2 (oculocutaneous albinism II), particularly in melanocytes of the iris.

The diversity in human hair and eye colour in the Baltic region is much more diverse, and as we move outwards from there, the hair tends to be black, and eyes tend to be brown (Frost P, 2006). Within this region, selection signals for blue and green eyes were observed, with signals from SNPs associated with red and blonde hair colour. The high frequency of colour selection with the high colour diversity can only be explained by the artificial selection in the population. The European population first settled in around 35 kya; this is not enough of a time period to get this amount of colour diversity due to random factors such as genetic drift, founder effect, or relaxation of natural selection(Frost P, 2006).

Selection in Toll-Like Receptor 1 (TLR1/TLR6) in the European population.

Toll-like receptors are part of our innate and adaptive immune system, and they are present on the surface membrane of macrophages, dendritic cells, and natural killer cells. TLRs are responsible for the antigen-specific adaptive immunity of the body. TLR1 and TLR6 with TLR2 have been studied to identify molecular patterns of lipoproteins and lipoteichoic acid present on the membrane of the gram-negative bacteria (Sadeghalvad M et al., 2021).

The bubonic plague pandemic occurring in Europe in the mid 14th century is one of the fatal pandemics to be recorded. The plague was caused by a gram-negative bacteria *Yersinia pestis* which spread due to fleas. It is estimated that the plague killed about 30% to 60% of the European population.

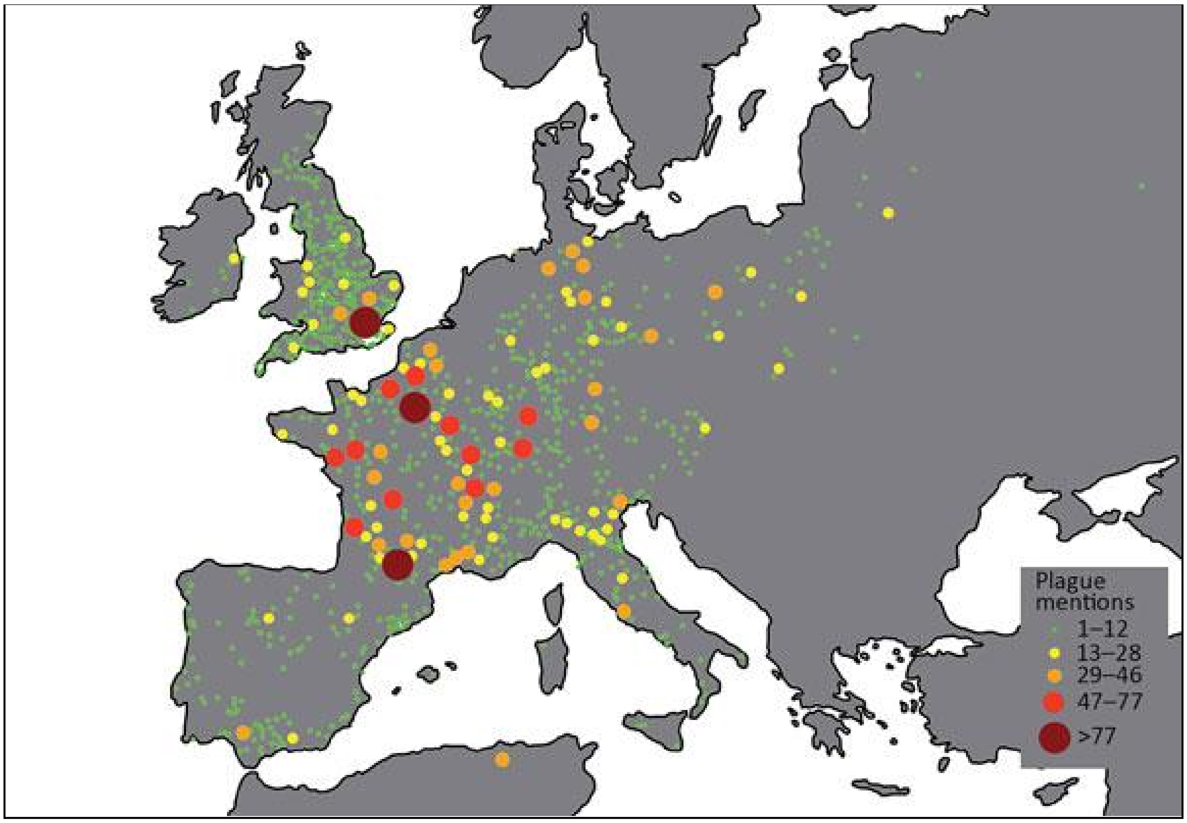


Figure 10: Plague outbreaks in Europe, 1347–1760 (source: Laayouni H 2014; PNAS License).

Gene	SNP	P-value	Tissue
TLR6	rs5743618	1.1e-18	Cultured fibroblasts

Table 7: Significant eQTLs associated with rs5743618 reported in GTEx database.

The rs5743618 SNP is a missense variant (C -> A) of the TLR1 gene which lies on chromosome 4. The missense SNP also is also in an eQTL associated with TLR6 in fibroblasts, which can produce TLRs (Bautista-Hernández L A et al., 2017). rs5743618 was reported to be under selection in three European populations, namely, Great Britain, Central Europe, and Finland. From figure 5, we can see that Central European countries and Great Britain have had higher mentions of the plague, which might cause selection in the rs574318 SNP in the TLR1 gene.

Selection in Eosinophil Count in African populations.

Eosinophils is one of the granulocytes, which are proinflammatory white-blood cells. Eosinophils play a significant role in the type-2 immune response during parasitic infections. Apart from being present in lymph nodes and the spleen, the Eosinophils are also found in the gastrointestinal tract, a common site to contract infections (McBrien C & Andrew Menzies-Gow A, 2017).

SNPs rs6472241 & rs6961605, associated with Eosinophil counts, showed selection in Gambia, Sierra Leone, and Nigeria populations. The whole European population, the Han population from Beijing and the Japanese population, which are on a higher latitude compared to previously mentioned populations, were not reported for selection in these SNPs.

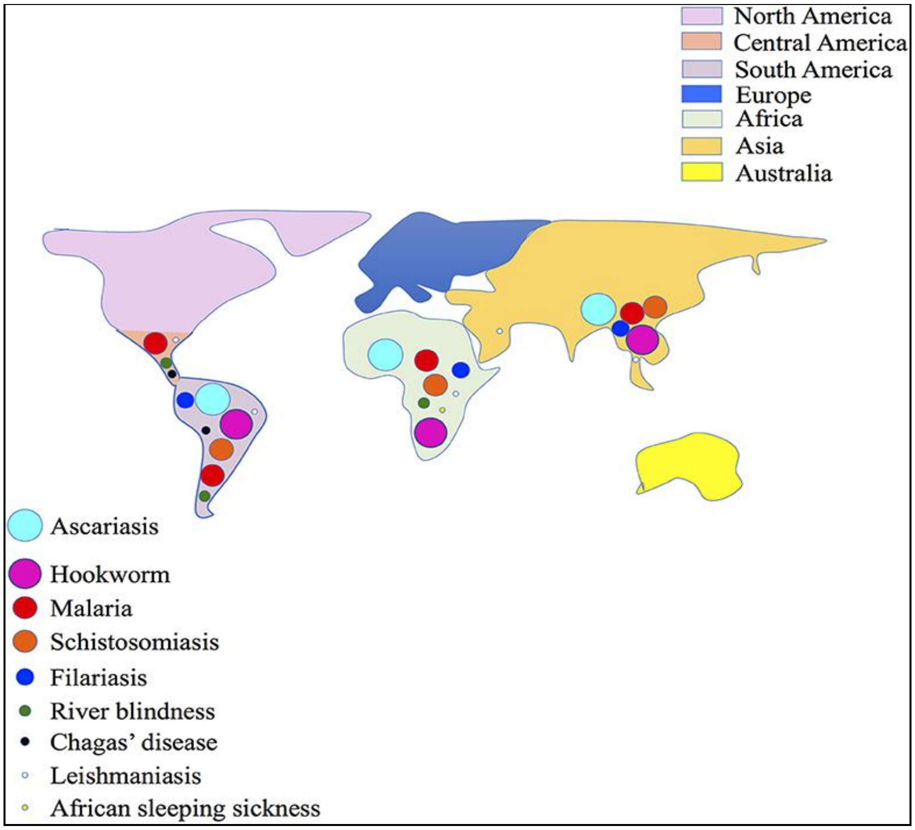


Figure 11: A representation of the worldwide distribution of parasitic infections (source: Cao B & Guiton P 2018; CC-BY).

Parasitic infections are more common as we move closer to the equator since the hot and humid climate of the tropical regions favours the growth conditions of parasitic helminths and insects.

Gene	SNP	P-value	Tissue
MTFR1	rs6472241	5.8e-14	Muscle - Skeletal
PDE7	rs6472241	1.1e-7	Esophagus - Mucosa
ZNF282	rs6961605	3.2e-62	Whole Blood

Table 8: Significant eQTLs associated with rs6472241 and rs6961605 reported in the GTEx database.

ZNF282 shows eQTL associated with rs6961605 in whole blood tissue. The zinc finger protein family is part of the gene expression pathway. The G -> C transformation in rs6472241 is linked with higher expression of the PDE7 gene in the Esophagus

Mucosa, and the PDE (cyclic nucleotide phosphodiesterase) protein family mediates and is responsible for the regulation of extracellular signalling. The presence of SNPs rs6472241 also significantly increases the expression in MFTR1 (Mitochondrial Fission Regulator 1) in the muscle cells. Studies have shown that eosinophils have about 24-36 mitochondria which is much greater than the number of mitochondria in found neutrophils (5-6), which contributes to the flexibility of the eosinophil cell in its diverse role in host defence (Porter L et al., 2018).

4. Conclusion

With this study, we aimed to use the previously known SNPs under selection in different populations and use it to identify potentially new SNPs which are under positive selection. This machine learning model made in this study was able to identify the SNPs under selection conditions in several different populations. The model successfully identified mutations, such as in the case of the VKORC1 gene, which causes slower processing of warfarin in east Asian populations or the selection in the TLR1 gene in the European population. These mutations are well studied and have been identified before with other methods. The model identified 2 mutations, rs6472241 and rs6961605, which aren't that well studied and could potentially be under selection in the western African population. Further analysis of the SNPs identified could be done to understand the mechanisms and the pathways of these SNPs, which could provide helpful insights into fields like pharmacogenetics.

To improve upon this current work, larger training data with more populations would help improve the model's performance. Incorporating other summary statistics than the ones used here as features for classification could make it possible to study rare diseases and conditions; the methods currently do not produce significant results when the MAF of an SNP is lower than 0.05, thus making it impossible to study rare diseases or conditions with this model. Further, the model assumes that the chance of selection is equal in all the populations, which could have been avoided if a separate model for each population was used, provided sufficient training data was available for each of the 17 populations. Further, the model only classifies SNPs as selected or non-selected, which might not justify the classification of certain SNPs into either category.

References

1. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research. 2019 Jan;47(D1):D1005-D1012. DOI: 10.1093/nar/gky1120. PMID: 30445434; PMCID: PMC6323933.

2. Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, *23*(9), 1514. https://doi.org/10.1101/GR.154831.113

3. Bomba, L., Nicolazzi, E. L., Milanesi, M., Negrini, R., Mancini, G., Biscarini, F., Stella, A., Valentini, A., & Ajmone-Marsan, P. (2015). Relative extended haplotype homozygosity signals across breeds reveal dairy and beef specific signatures of selection. *Genetics, Selection, Evolution : GSE*, *47*(1). https://doi.org/10.1186/S12711-015-0113-9

4. Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology and Evolution*, *31*(5), 1275. https://doi.org/10.1093/MOLBEV/MSU077

5. Frost, P. (2006). European hair and eye color. A case of frequency-dependent sexual selection? *Evolution and Human Behavior*, *27*(2), 85–103. https://doi.org/10.1016/J.EVOLHUMBEHAV.2005.07.002

6. Gahan, P. B. (2005). Life: the science of biology (7th edn) W. K. Purves, D. Sadava, G. H. Orians and H. C. Heller, W. H. Freeman & Co, 1121 pp., ISBN

0-7167-9856-5 (2004). *Cell Biochemistry and Function*, *23*(3), 221–221. https://doi.org/10.1002/CBF.1179

7. Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, M., Burger, J., Swallow, D. M., & Thomas, M. G. (2011). Evolution of lactase persistence: an example of human niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1566), 863. https://doi.org/10.1098/RSTB.2010.0268

8. Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of Levels of Gene Flow from DNA Sequence Data. *Genetics*, *132*(2), 583. /pmc/articles/PMC1205159/?report=abstract

9. Koropoulis, A., Alachiotis, N., & Pavlidis, P. (2020). Detecting Positive Selection in Populations Using Genetic Data. *Methods in Molecular Biology*, *2090*, 87–123. https://doi.org/10.1007/978-1-0716-0199-0_5

10. Lam, M. P. S., & Cheung, B. M. Y. (2012). The pharmacogenetics of the response to warfarin in Chinese. *British Journal of Clinical Pharmacology*, *73*(3), 340. https://doi.org/10.1111/J.1365-2125.2011.04097.X

11. Li, S., Zou, Y., Wang, X., Huang, X., Sun, Y., Wang, Y., Dong, L., & Jiang, H. (2015). Warfarin Dosage Response Related Pharmacogenetics in Chinese Population. *PLOS ONE*, *10*(1), e0116463. https://doi.org/10.1371/JOURNAL.PONE.0116463

12. Ephert (2011), Out of Africa map, Map. https://commons.wikimedia.org/wiki/File:Human_migration_out_of_Africa.png

13. Pybus M., Luisi P., Dall'Olio G.M., Uzkudun M., Laayouni H., Bertranpetit J., Engelken J. Hierarchical boosting: a machine-learning framework to detect and

classify hard selective sweeps in human populations. Bioinformatics. 2015; 31:3946–3952.

14. Schrider, D. R., & Kern, A. D. (2015). Inferring Selective Constraint from Population Genomic Data Suggests Recent Regulatory Turnover in the Human Brain. Genome biology and evolution, 7(12), 3511–3528. https://doi.org/10.1093/gbe/evv228

15. Mirazo´n, M., Lahr, M. M., & Foley, R. A. (1998). Towards a Theory of Modern Human Origins: Geography, Demography, and Diversity in Recent Human Evolution. In *Yrbk Phys Anthropol* (Vol. 41).

16. Murga-Moreno, J., Coronado-Zamora, M., Bodelón, A., Barbadilla, A., & Casillas, S. (2019). PopHumanScan: the online catalog of human genome adaptation. *Nucleic Acids Research*, *47*(D1), D1080–D1089. https://doi.org/10.1093/NAR/GKY959

17. Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences of the United States of America*, *70*(12 Pt 1-2), 3321. https://doi.org/10.1073/PNAS.70.12.3321

18. Owen, R. P., Gong, L., Sagreiya, H., Klein, T. E., & Altman, R. B. (2010). VKORC1 Pharmacogenomics Summary. *Pharmacogenetics and Genomics*, *20*(10), 642. https://doi.org/10.1097/FPC.0B013E32833433B6

19. Posth, C., Renaud, G., Mittnik, A., Drucker, D. G., Rougier, H., Cupillard, C., Valentin, F., Thevenet, C., Furtwängler, A., Wißing, C., Francken, M., Malina, M., Bolus, M., Lari, M., Gigli, E., Capecchi, G., Crevecoeur, I., Beauval, C., Flas, D., … Krause, J. (2016). Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a late-glacial population turnover in Europe. *Current Biology*, *26*(6), 827–833. https://doi.org/10.1016/j.cub.2016.01.037

20. Schaffner, S., & Sabeti, P. (2008). *Evolutionary Adaptation and Positive Selection in Humans | Learn Science at Scitable*. Evolutionary Adaptation in the Human Lineage. Nature Education 1(1):14.

21. Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews. Genetics*, *9*(6), 477. https://doi.org/10.1038/NRG2361

22. Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P. N., Stark, M. S., Hayward, N. K. K., Martin, N. G., & Montgomery, G. W. (2008). A Single SNP in an Evolutionary Conserved Region within Intron 86 of the HERC2 Gene Determines Human Blue-Brown Eye Color. *The American Journal of Human Genetics*, *82*(2), 424–431. https://doi.org/10.1016/J.AJHG.2007.11.005

23. Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., Jakobsdottir, M., Steinberg, S., Pálsson, S., Jonasson, F., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., … Stefansson, K. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genetics 2007 39:12*, *39*(12), 1443–1452. https://doi.org/10.1038/ng.2007.13

24. Wang, G., & Speakman, J. R. (2016). Analysis of Positive Selection at Single Nucleotide Polymorphisms Associated with Body Mass Index Does Not Support the “Thrifty Gene” Hypothesis. *Cell Metabolism*, *24*(4), 531–541. https://doi.org/10.1016/J.CMET.2016.08.014

25. WRIGHT, S. (1949). THE GENETICAL STRUCTURE OF POPULATIONS. *Annals of Eugenics*, *15*(1), 323–354. https://doi.org/10.1111/J.1469-1809.1949.TB02451.X

26. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., … Wang, J. (2010). Sequencing of Fifty Human Exomes Reveals Adaptation to High Altitude. *Science (New York, N.Y.)*, *329*(5987), 75. https://doi.org/10.1126/SCIENCE.1190371

27. McBrien, C. N., & Menzies-Gow, A. (2017). The biology of eosinophils and their role in asthma. *Frontiers in Medicine*, *4*(JUN), 93. https://doi.org/10.3389/FMED.2017.00093/BIBTEX

28. Cao, B., & Guiton, P. S. (2018). Important Human Parasites of the Tropics. *Frontiers for Young Minds*, *6*. https://doi.org/10.3389/FRYM.2018.00058

29. Sadeghalvad, M., Mohammadi-Motlagh, H.-R., & Rezaei, N. (2021). Toll-Like Receptors. *Reference Module in Biomedical Sciences*. https://doi.org/10.1016/B978-0-12-818731-9.00044-6

30. Laayouni, H., Oosting, M., Luisi, P., Ioana, M., Alonso, S., Ricaño-Ponce, I., Trynka, G., Zhernakova, A., Plantinga, T. S., Cheng, S.-C., Meer, J. W. M. van der, Popp, R., Sood, A., Thelma, B. K., Wijmenga, C., Joosten, L. A. B., Bertranpetit, J., & Netea, M. G. (2014). Convergent evolution in European and Rroma populations reveals pressure exerted by plague on Toll-like receptors. *Proceedings of the National Academy of Sciences*, *111*(7), 2668–2673. https://doi.org/10.1073/PNAS.1317723111

31. Bautista-Hernández, L. A., Gómez-Olivares, J. L., Buentello-Volante, B., & Bautista-de Lucio, V. M. (2017). Fibroblasts: The Unknown Sentinels Eliciting Immune Responses Against Microorganisms. *European journal of microbiology & immunology*, *7*(3), 151–157. https://doi.org/10.1556/1886.2017.00009

32. Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., … Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature 2012 491:7422*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

33. Porter, L., Toepfner, N., Bashant, K. R., Guck, J., Ashcroft, M., Farahi, N., & Chilvers, E. R. (2018). Metabolic profiling of human eosinophils. *Frontiers in Immunology*, *9*(JUN), 1404. https://doi.org/10.3389/FIMMU.2018.01404/BIBTEX

34. Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7