21 Jan 2019

Random Subspace Projection For Predicting Biogeographical Ancestry

Tanjin Toma

Tayo Olufemi-Ajayi
Missouri University of Science and Technology, towd2@mst.edu

Jeremy Dawson

Donald Adjeroh

Recommended Citation

T. Toma et al., "Random Subspace Projection For Predicting Biogeographical Ancestry," *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pp. 1719 - 1725, article no. 8621222, Institute of Electrical and Electronics Engineers, Jan 2019.
The definitive version is available at https://doi.org/10.1109/BIBM.2018.8621222

Random Subspace Projection for Predicting Biogeographical Ancestry

Tanjin Toma
Lane Dept. of CSEE
West Virginia University
Morgantown, WV, USA
tatoma@mix.wvu.edu

Tayo Olufemi-Ajayi
Engineering Program,
Missouri State University
Springfield, MO, USA
TayoObafemiAjayi@MissouriState.edu

Jeremy Dawson
Lane Dept. of CSEE
West Virginia University
Morgantown, WV, USA
jeremy.dawson@mail.wvu.edu

Donald Adjeroh
Lane Dept. of CSEE
West Virginia University
Morgantown, WV, USA
don@csee.wvu.edu

ABSTRACT– Human biogeographical ancestry estimation using genomic information is an important problem with applications in population stratification, admixture mapping, forensic ancestry inference, and in healthcare. Various studies have proposed panels of ancestry informative single nucleotide polymorphisms (SNPs) for distinguishing between widely separated continental populations. There has been limited investigation on identifying SNP panels for sub-continental ancestry prediction, especially given the difficult challenge of identifying SNP markers to distinguish closely associated sub-populations, for instance, within a continent. In this study, we propose an ancestry informative SNP selection algorithm exploiting the concept of random subspace projection using supervised learning. The proposed approach identifies small panels of useful SNPs for sub-continental level ancestry classification. We show results for sub-continental level classification for all five continents in our dataset.

KEYWORDS

SNP, DNA, Ancestry Classification, SNP Selection, Random Subspace Projection, Single Chromosome

1 INTRODUCTION

Estimating human biogeographical ancestry using genomic information has long been studied in the domain of bioinformatics, genetics, and forensic science. Most genetic ancestry inference techniques focused on developing methods with the aim of distinguishing some continental populations e.g., Europe, America, Africa and East Asia [1]. These studies used DNA polymorphism, namely, single nucleotide polymorphisms (SNPs) as ancestry informative markers, since SNPs exhibit substantially different allele frequencies between populations from distant geographical regions (e.g., different continents). Some widely used techniques in the domain of continental ancestry prediction include fixation index (Fst) based methods (e.g., STRUCTURE [2]), and principal component analysis (PCA) based approaches (e.g., EIGENSTART [3]). Several studies also identified small SNP panels, (typically in the dozens to hundreds of SNPs) to estimate continental genetic ancestry[4]. See also the methods in [18, 19, 201, 21]. However, very few studies are found in the literature that published SNP panels for estimation of sub-continental level ancestry. This underscores the difficult challenges in identifying the distinctive SNP markers between closely related sub-populations within a continent [5]. Predicting an individual's sub-continental ancestry is still considered a huge challenge given the similarity of genomic attributes between the closely associated sub-continental populations (e.g., distinguishing a Han-Chinese from a Dai-Chinese, or from a Japanese using genomic information). Among the few studies that have considered sub-continental level ancestry prediction, ETHNOPRED [6] proposed an ensemble classification scheme based on disjoint decision trees that can predict individual's continental and sub-continental ancestry. Although, they addressed pairwise/binary classification problem between subpopulations from Europe, East Asia and Africa., the problem of distinguishing multiple closely related sub-populations within a given continent has not been addressed. Similar to ETHNOPRED, Graydon et al. [7] addressed ancestry classification problem in terms of binary classification between similar populations as well as distinctly different populations.

Some other studies, such as, LAMP [8] and WINPOP [9] have considered the sub-continental level ancestry inference problem from the perspective of admixed populations. These studies demonstrate progress in terms of estimating sub-continental ancestry, however, there is still space for further improvement through involving more population groups and addressing multinomial classification problem in

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

the presence of closely related subpopulations in a continent. In this study, we will address both continental-level and sub-continental level ancestry estimation problems. Sub-continental level ancestry estimation has been considered as a multi-class classification problem where the instances from different classes are very close to each other based on genomic attributes. Here, we identified ancestry informative SNP markers from just one chromosome (chromosome 1), considering that such resource-constrained environment might arise in many real-world applications, such as forensic identification, or criminal investigation.

Efficient selection of ancestry informative SNPs is the key to successful ancestry prediction. The removal of redundant and noisy SNP features is essential prior to applying a learning algorithm. Here we propose a SNP marker selection technique incorporating the concept of random subspace projection and supervised learning. The proposed approach is an iterative technique that uses the supervised learning algorithm itself to evaluate the usefulness of the SNPs. A multi-layer perceptron neural network architecture with softmax activation at the output has been applied in such supervised SNP selection technique. We apply this approach of SNP selection to address the sub-continental level ancestry classification problem, involving many closely associated sub-populations.

2 RELATED WORK

In bioinformatics, many studies deal with high dimensional data involving a large number of features and a limited number of samples. This is popularly known as the 'large p, small n problem'. For example, microarray datasets measure the gene activity of thousands of genes while the number of samples is limited to several hundreds [10]. Ancestry classification also falls into this group considering the presence of millions of SNP markers and very few samples. Due to the high dimensionality of data and existence of many noisy features, traditional pattern recognition techniques often fail to adequately deal with these large p small n problems. Traditional classification algorithms, such as support vector machine (SVM) and the k-nearest-neighbor (KNN) classifiers cannot perform well in the presence of increasing number of noisy features, in spite of their ability to handle large number of features. Therefore, various techniques have been proposed to address these problems caused by high dimensional feature space including classifier aggregation and feature selection. One approach to these types of problems is the random subspace method [11], which provides improved classification accuracy by aggregating the power of multiple classifiers. It selects a random subset of features in each pass of the algorithm and constructs a decision tree classifier to predict the unknown samples. Li et al. [12] proposed another technique for high dimensional data classification, where the random subspace idea is exploited to generate the individual classifiers on the low dimensional subspaces and base classifiers are assigned different weights according to their individual performances. Apart from the classifier aggregation techniques, another type of approach in handling high dimensional data is pre-classification. Here, feature selection is aimed at removing the noisy features and then selecting the features that are most discriminative among the different classes. Random KNN [13] is one such feature selection technique. Random KNN consists of an ensemble of k-nearest neighbor base classifiers, each constructed from a random subset of the input features. The optimum subset of features is selected through ranking the features using a support measure and further applying a two-stage backward elimination procedure. In addition, there are many popular gene ranking algorithms which also followed random subspace method, such as, the RSM-GR algorithm [14], where support vector machine was used as the base classifier.

In this study, we propose a SNP selection algorithm following the notion of random subspace projection. The proposed iterative approach considers the potential interaction among the SNPs in the random subspace. By random selection of SNPs in lower- dimensional subspace for many iterations, we eventually identify the best performing SNPs for ancestry classification.

3 METHODOLOGY

3.1 Dataset and Pre-processing

The pre-processing stages for this study is based on prior work in [15], where a correlation-based SNP selection algorithm was proposed for ancestry classification. The dataset used for this work is from the 1000 Genomes Project, Phase III[16]. The dataset contains information on SNP variants from all 23 chromosomes for 2504 individuals, from 26 different sub-populations, from five continents (see Table II under the Results section). We focused on analyzing the variants from Chromosome 1 which is nearly 20.1 million SNPs. For each SNP, we extracted their position/loci number, rsID, reference allele, alternate allele (s), and allele information of all 2504 subjects (each person's allele is diploid, containing two nucleotides, from different combinations of the four nucleotide bases (A, C, G, T)). Data cleaning and preprocessing was performed following the procedure outlined in [15]. The result is a list of 6404 candidate discriminative SNPs, which provide the starting point for our analysis.

3.2 Random Subspace Projection for SNP Selection

To find an effective set of ancestry informative SNPs (AISNPs) for sub-continental level ancestry classification from the 6,404 SNPs, we propose an iterative random sampling technique. In each iteration, the algorithm randomly samples a small subset of SNPs (say M SNPs out of the 6,404 SNPs) and measures ancestry classification performance using a neural network classifier. This process of random sampling of SNPs and performance evaluation in lower-dimensional feature space continues for a large number of iterations, say N iterations. At every iteration of the algorithm, randomly selected M SNPs are used to form M-dimensional allele-context feature space for subject t in the dataset, which is denoted as $at = [a^{(1)}, a^{(2)}, .., a^{(M)}]$. Here $a^{(i)}$ representing allele-context feature of SNP i for subject t can take three possible values: 0, 1, 2, where '0' means both nucleotides from an individual at the given SNP location i, are the same as the reference nucleotide, '1' denotes that one of two nucleotides is different from the reference nucleotide, and '2' indicates that both nucleotides of that individual are different from the reference nucleotide. With the M-dimensional feature space in the form of allele-context features, multi-class ancestry classification (continental/sub-continental) is performed on the validation data. The classification accuracies for all the N iterations are stored in an $N \times 1$ vector and the corresponding SNP subsets are stored in an $N \times M$ matrix. Next a two-step ranking process is applied to identify the best discriminative SNPs. First, SNP subsets in all N iterations are ranked based on the classification performances of the individual N base classifiers. From the ranked subsets, we choose a certain number ($Q \leq N$) of top subset of SNPs and find all the unique SNPs that occur in these top Q subsets. Let m = number of unique SNPs in top Q subsets. The next step is the individual SNP ranking of each of the m SNPs. We compute the frequency of occurrence of each of the m SNPs in top Q base classifiers, which can be represented as a m-length count vector, $cQ = [c^{(1)}c^{(2)} \ldots c^{(m)}]$. Thus, each of m SNPs is assigned a rank based on their frequency of occurrence in top Q subsets. That is, a SNP is considered powerful for discriminating between populations if many of the top Q base classifiers have chosen this SNP. Finally, with these m sorted SNPs, the classifier is evaluated iteratively on the test set using top K of m SNPs by increasing K in a linear fashion, until all m SNPs are covered in the analysis.

3.2.1 One-stage Ancestry Classification: The 1000 Genomes Project, Phase III dataset used in this work contains genomes of subjects from 26 different populations, from five continents. For the one-stage 26-class ancestry classification problem, the goal is to classify the ancestry of an unknown/test individual into one of the 26 subpopulations, without initially detecting the continent of origin for the individual. To address this problem, we applied the proposed random sampling algorithm. First, we define two parameters, M and N (say, M = 50, N = 50,000). Next, we

execute all the steps of the algorithm till Step 9, when we obtain the ranking of each of the *m* unique SNPs from the top *Q* panels. Then, we initialize parameter *K* to δ = 100. Then for each *K* in the interval of 100, the top *K* SNPs are used to conduct the overall 26-class classification for 80/20 train-test split of the data. We have experimented for several discrete values of *Q*, such as, *Q* = 100, 1000, 5000, *etc*., where *Q* <= *N*. The best performance for each *Q* is recorded. Finally, the *Q* which provides the highest performance, in terms of accuracy and number of SNPs (the fewer the better computationally), is considered and the corresponding set of SNPs constitute the best candidate SNPs for one-stage 26-class ancestry classification problem.

3.2.2 Neural Network Classifier: For classification task, we used a 2-hidden layer neural network architecture. The first hidden layer consists of 100 nodes and the second hidden layer consists of 50 nodes. In both hidden layers *ReLU* (Rectified Linear Unit) activation function is applied at the output of each hidden node. *ReLU* computes the function $f(x) = max(0, x)$, where x denotes the weighted sum of the input features to a given node. *Softmax* regression is used as activation function in the final/output layer of the network. For a K-class classification problem, number of units/nodes in the output layer of the neural network is *K*. Each of the *K* output nodes gives the probability of a certain class and probabilities from all output nodes sum to 1. Each output layer node *i* receives the weighted sum of the inputs from the previous layer with the addition of a bias term, viz:

$$z_i = \sum_j w_{i,j} x_j + b_i \quad (1)$$

where, j is the number of nodes in the previous layer. Now, to compute the *softmax* activation at each output node, exponential of the term z_i is calculated for each i,

$$t_i = e^{z_i} \quad (2)$$

Finally, activation at output node *i* is obtained by normalizing the exponential term.

$$a_i = e^{z_i} / \sum_{i=1}^{K} t_i \quad (3)$$

Thus, by normalizing the distribution, output from each node i falls in the range [0, 1]. The class associated with the highest probability value is taken as the predicted output label.

4 EXPERIMENTAL RESULTS

We have evaluated the performance of the proposed random sampling technique for both one stage and two stage ancestry classification. All the experiments were performed using the 1000 Genome Project, Phase III dataset. We report the results below.

4.1 One-stage 26-class Classification

We have demonstrated the results of one-step classification into 26 populations in Fig. 1. For our analysis, we have considered *M* = 50 and *N* = 50000. With *Q*<=*N*, the values of the parameter *Q* are chosen over a wide range, with minimum as small as 100 and the maximum equal to 50000. In Fig. 1, the classification performances are depicted for five different values of *Q* (*Q* = 100, 1000, 10000, 30000, 50000). For each *Q*, accuracy is measured on the test set using a certain *K* number of top ranked SNPs, with choice of K in the interval of 100. From the figure, it is observed that for small value of Q (e.g., Q=100), the performances over top K SNPs are relatively low, while with increasing value of Q the performances improve. The red curve in the figure demonstrates the best results in one-stage 26 class classification, which corresponds to Q=10000. With only 1900 SNPs, classification accuracy of 78.50% is achieved while Q=10000. It is also noticed that performances over top K SNPs cannot be improved further with higher values of Q (e.g., Q=30000, 50000). With Q=50000 (i.e., Q=N), represented by the green curve, the performances drop to even lower values compared to those for Q=100. In Table I, we record the results for each Q value in our experiment. For a certain value of Q, we mention two types of results: one is the number of unique SNPs available in top Q panels of N iterations and the classification accuracy achieved using all those SNPs. The other result indicates the best performance over all Q values in our experiment, in terms of number of SNPs and corresponding classification accuracy. We explain the underlying reasons behind the observed trend of the graphs in Fig. 1, for fixed M and N and with varying Q. With small Q, many SNPs have similar counts of occurrence in top Q subsets of N iterations, thus they are less likely to be properly

ranked. With higher value of Q, such as, Q=10000, 20000, we observe greater variations between the SNPs in terms of their counts of occurrence in the top panels. This results in a better ranking of the SNPs and better classification results. On the other hand, when Q is very large or close to the value of N (say, Q=50000), SNPs with very high count of the occurrence but occurring mostly in the lower panels of N iterations incorrectly achieve higher individual ranking, which deteriorates the overall classification accuracy. The above explanation strongly supports our experimental findings for one-stage 26 class classification problem, as we reach the best classification performance of 78.50% using 1900 SNPs for Q=10000. Although, it is noticed that in case of one stage 26 class classification, Q=20000 can produce a slightly higher classification accuracy 79.31%, but at the cost of much larger number of SNPs (2700 SNPs). Therefore, for the one-stage classification scheme, we considered the set of 1900 SNPs as the best candidate SNPs capable of classifying samples into one of the 26 populations with accuracy as high as 78.50%.

4.2 Choice of Parameters M and N

In our analysis, we have demonstrated all the results regarding multi-class ancestry identification (one-step or two-step 26-class classification and continental classification) for a certain value of the parameters, M and N (M=50 and N=50000). However, we have conducted empirical experiments with several other choices for M and N, and finally considered the values that yield the optimum performance in all circumstances. Our analysis showed that our choice of M=50 produced the best results (data not shown).

4.3 Comparison with Other Methods

Table II shows the results obtained from the proposed one-stage ancestry classification scheme. Here, it is noticed that although overall 26-class classification accuracy obtained is about 78.50% from the proposed one-stage approach, the individual population classification rates could vary significantly. With higher number of SNPs (e.g., 1900 SNPs for the proposed method), we achieve negligible error in continental-level identification, however, individual population classification rate drops significantly for several instances. For example, British (GBR) classification rate in one stage approach is 41.18% and African-Caribbean (ACB) classification accuracy is 42.11%.

We have also performed a limited comparison of our proposed approaches with related work. Table II, 3rd column shows corresponding results using a two-stage approach, where we first predict the continent, before classifying the sub-population within the continent. Table III presents a comparative performance of our proposed method for binary/pairwise classification of sub-populations against other related methods in the literature. The comparative results show the proposed methods are competitive with current state-of-the-art approaches.

Table I: Results for one-stage 26-class ancestry classification (M=50, N=50000)

	SNPs Coverage		Best Performances	
	No. of Unique SNPs	Accuracy (%)	No. of SNPs used	Accuracy (%)
Q=100	3378	69.98	3000	69.98
Q=500	6237	67.14	2300	73.43
Q=1000	6400	67.75	2500	72.41
Q=5000	6404	67.34	2800	74.44
Q=10000	6404	67.34	**1900**	**78.50**
Q=20000	6404	67.34	2700	79.31
Q=30000	6404	67.34	2100	77.69
Q=40000	6404	67.34	3100	75.46
Q=50000	6404	67.34	5400	68.15

5 CONCLUSIONS

In this work, we have proposed a SNP selection algorithm exploiting the approach of random subspace projection. This approach has been observed to be very effective in selecting small subsets of ancestry informative SNPs for distinguishing multiple closely associated sub-populations in the same continent. We noticed that sub-populations within continent America, East Asia and Africa are relatively easy to distinguish, whereas more difficulties arise while distinguishing between the sub-populations from South Asia and sub-population within Europe. As a future extension of this work, we may need to analyze and investigate whether other chromosomes (beyond chromosome 1) could contain better marker SNPs for ancestry estimation. Besides, it might be instructive to consider the impact of linkage disequilibrium while selecting the AISNPs. Genes which are in linkage disequilibrium may contain SNPs of similar allele information. We plan to refine our SNP selection method by ignoring the SNPs from the closely located genes which are in linkage disequilibrium. In addition, instead of using a two-layer shallow neural network as the classification scheme, we can try deep neural network architectures with more layers, such as deep belief networks.

Table II: Detailed results for proposed method

Populations	One Stage Approach	Two Stage Approach
Individual Population Classification Rates		
GBR	41.18%	52.94%
FIN	95.00%	85.00%
IBS	77.27%	72.73%
CEU	60.00%	55.00%
TSI	81.82%	86.36%
PUR	95.24%	76.19%
CLM	94.44%	83.33%
PEL	93.75%	93.75%
MXL	66.67%	66.67%
CHS	61.90%	71.43%
CDX	66.67%	66.67%
KHV	80.00%	85.00%
CHB	80.95%	85.71%
JPT	95.24%	100.00%
PJL	57.89%	73.68%
BEB	81.25%	62.50%
STU	57.14%	51.52%
ITU	57.14%	71.43%
GIH	100.00%	90.48%
ACB	42.11%	52.63%
GWD	91.67%	96.10%
ESN	100.00%	95.00%
MSL	93.75%	87.50%
YRI	90.91%	90.91%
LWK	100.00%	100.00%
ASW	50.00%	75.00%
Overall Classification Accuracy	**78.50%**	**78.70%**
Continental Classification Rates		
Europe	100%	96.04%
America	100%	90.63%
East Asia	99.01%	100%
South Asia	100%	97.96%
Africa	100%	100%
Average Continental Accuracy	**99.60%**	**97.57%**

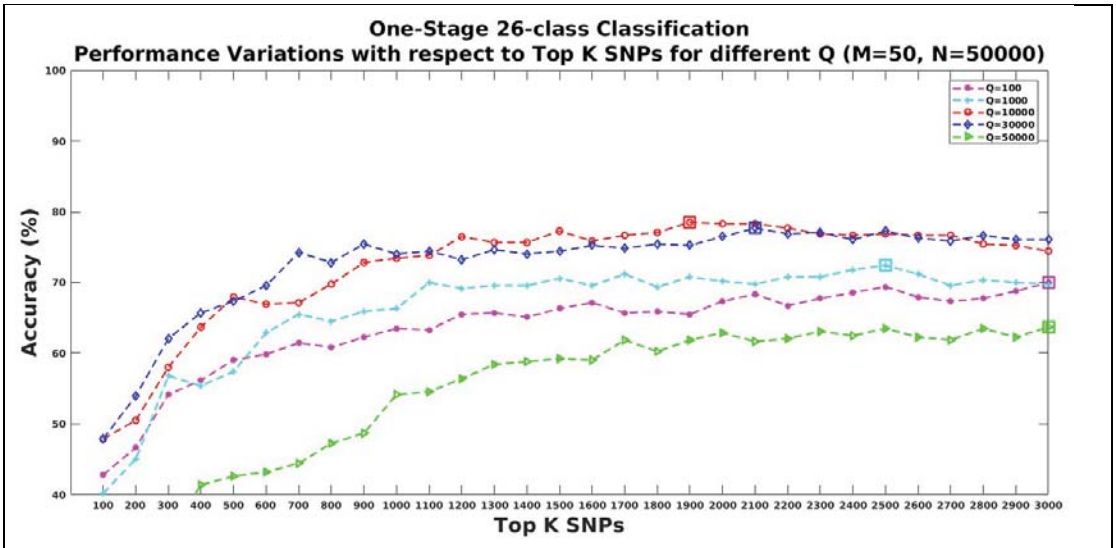

Fig. 1: One-stage 26 class classification results with varying number of top SNPs.

TABLE VI: COMPARATIVE PERFORMANCE IN SUB-POPULATION-LEVEL ANCESTRY CLASSIFICATION

pairwise sub-populations	continent	method	data size	datasets	Accuracy (%)	# attributes used
CEU-TSI	Europe	ETHNOPRED [6]	267	Hapmap iii	86.6±2.4	180 SNPs
CHB-JPT	East Asia	ETHNOPRED [6]	250	Hapmap iii	95.6± 3.9	877 SNPs
LWK-MKK	Africa	ETHNOPRED [6]	294	Hapmap iii	95.9±1.5	341 SNPs
JPT-CHB	East Asia	Bayesian [7]	9104	own collection	74.9(77.2***)	15 STR loci
JPT-KOR	East Asia	Bayesian [7]	731	own collection	67.9 (63.7)	15 STR loci
CHB-KOR	East Asia	Bayesian [7]	731	own collection	69.6 (62.4)	15 STR loci
--	Europe	SNP-Correlation[15]	503	1000 genome-III	76.6*	58 SNPs**
--	Africa	SNP-correlation[15]	661	1000 genome-III	87.0*	87 SNPs**
--	East Asia	SNP-correlation[15]	504	1000 genome-III	73.3*	68 SNPs**
--	Europe	**Proposed (one-Stage)**	503	1000 genome-III	75.3*	140 SNPs**
--	Africa	**Proposed (one-Stage)**	661	1000 genome-III	87.6*	72 SNPs**
--	East Asia	**Proposed (one-Stage)**	504	1000 genome-III	82.2*	180 SNPs**

* Average accuracy of all pairwise sub-population classifications within the given continent

**Average number of SNPs required per pairwise sub-population classification within the given continent

*** Results obtained without normalization.

REFERENCES

[1] Rami Nassir et al. “An ancestry informative marker set for determining conti- nental origin: validation and extension using human genome diversity panels”. In: *BMC Genetics* 10.1 (2009), p. 39.

[2] Jonathan K Pritchard et al. “Association mapping in structured populations”. In: *The American Journal of Human Genetics* 67.1 (2000), pp. 170–181.

[3] Alkes L Price et al. “Principal components analysis corrects for stratification in genome-wide association studies”. *Nature Genetics* 38.8 (2006).

[4] Judith R Kidd et al. “Analyses of a set of 128 ancestry informative single- nucleotide polymorphisms in a global set of 119 population samples”. *Investigative Genetics* 2.1 (2011), p. 1.

[5] Jacobo Pardo-Seco, Federico Martinón-Torres, and Antonio Salas. “Evaluating the accuracy of AIM panels at quantifying genome ancestry”. *BMC Genomics* 15.1 (2014), p. 543.

[6] Mohsen Hajiloo et al. “ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction”. *BMC Bioinformatics* 14.1 (2013), p. 61.

[7] Matthew Graydon, François Cholette, and Lay-Keow Ng. “Inferring ethnicity using 15 autosomal STR loci-Comparisons among populations of similar and distinctly different physical traits”. *Forensic Sci. Int’l: Genetics* 3.4 (2009), 251-4.

[8] Sriram Sankararaman et al. “Estimating local ancestry in admixed populations”. *American J. Human Genetics* 82.2 (2008), pp. 290–303.

[9] Bogdan Paşaniuc et al. “Inference of locus-specific ancestry in closely related populations”. In: *Bioinformatics* 25.12 (2009), pp. i213–i221.

[10] Paweł Teisseyre, Robert A Kłopotek, and Jan Mielniczuk. “Random Subspace Method for high-dimensional regression with the R package regRSM”. In: *Computational Statistics* 31.3 (2016), pp. 943–972.

[11] Tin Kam Ho. “The random subspace method for constructing decision forests”. *IEEE TPAMI,* 20.8 (1998), pp. 832–844.

[12] Xiaoye Li and Hongyu Zhao. “Weighted random subspace method for high dimensional data classification”. In: *Statistics and its Interface* 2.2 (2009), p. 153.

[13] Shengqiao Li, E James Harner, and Donald A Adjeroh. “Random KNN feature selection-a fast and stable alternative to Random Forests”. *BMC Bioinformatics* 12.1 (2011), p. 450.

[14] Ruichu Cai, Zhifeng Hao, and Wen Wen. “A novel gene ranking algorithm based on random subspace method”. In: *IEEE IJCNN 2007*, pp. 219

[15] Tanjin Taher Toma et al. “What can one chromosome tell us about human biogeographical ancestry?” In: *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*. IEEE. 2017, pp. 188–193.

[16] 1000 Genomes Project Consortium et al. “A global reference for human genetic variation”. *Nature* 526.7571 (2015), p. 68.

[17] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: KDD. vol. 96. 34. 1996, pp. 226–231.

[18] Fondevila, M, et al. "Revision of the SNPforID 34-plex forensic ancestry test: assay enhancements, standard reference sample genotypes and extended population studies." *Forensic Sci. Int’l: Genetics* 7.1 (2013):63-74.

[19] Gettings, Katherine Butler, et al. "A 50-SNP assay for biogeographic ancestry and phenotype prediction in the US population." *FSI: Genetics* 8.1 (2014): 101-108.

[20] Lao, Oscar, et al. "Evaluating self-declared ancestry of US Americans with autosomal, Y-chromosomal and mitochondrial DNA." *Human Mutation* 31.12 (2010).

[21] Nievergelt, Caroline M., et al. "Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel.” *Investigative Genetics* 4.1 (2013): 13.