What Can One Chromosome Tell us About Human Biogeographical Ancestry?

Tanjin Taher Toma*, Zachary Williams, Jeremy Dawson, Donald Adjeroh+
Lane Dept. of Computer Science and Electrical Engineering,
West Virginia University, Morgantown, WV, USA
Email: *tatoma@mix.wvu.edu, +donald.adjeroh@mail.wvu.edu

***Abstract*—We study the problem of predicting human biogeographical ancestry using genomic data. While continental level ancestry is relatively simple using genomic information, distinguishing between individuals from closely associated sub-populations (e.g., from the same continent) is still a difficult challenge. In particular, we focus on the case where the analysis is constrained to using single nucleotide polymorphisms (SNPs) from just one chromosome. We thus propose methods to construct such ancestry informative SNP panels, and assess the performance of such SNP panels from just one chromosome, for both continental-level and sub-population level ancestry prediction. We present results on the performance of the proposed methods, including a comparison with other related methods.**

Keywords—SNP, DNA, ancestry prediction, single chromosome

I. INTRODUCTION

Using DNA information to infer the ancestral origin of an individual is useful for many purposes, for instance, in detecting stratification in biomedical studies (disease or trait association) [1, 2], estimating admixture between specific ancestral populations [3, 4], determining ancestry in forensic context [5, 6], and guiding criminal investigations [7]. Such studies on ancestry identification mainly aim to identify sets of ancestry informative markers (AIMs) through analyzing the DNA sequences of different chromosomes collected from the population samples under study. Most widely used AIMs are based on single nucleotide polymorphisms (SNPs) [8] which demonstrate superior ability in predicting geographic/ethnic origin of an unknown individual compared to other markers such as short tandem repeats (STRs) [9]. While very large number of SNPs can provide nearly accurate ancestry information for multiple geographic regions, small but robust sets of SNPs are especially useful [10]. There are many published SNP panels that focused on distinguishing ancestral origins from several continental regions, e.g., Europe, America, Africa and East Asia [11]. Studies have demonstrated that many globally distributed populations, mostly continental populations can be distinguished by examining differences in allele frequencies, using the fixation index, widely known as F_{st} [12]. Although a small set (typically few hundreds) of SNPs can distinguish continental differences between individuals using the F_{st} feature [13, 14]; such panels of SNPs are less informative in detecting sub-continental differences in closely related populations [3, 15]. Apart from F_{st}-based ancestry inference, techniques based on principal component analysis (PCA) [16, 17], like EIGENSTART [16], have widespread applications. These methods represent genetic variations by principal component vectors, however, they are not highly efficient due to the requirement of genotyping very large number of SNPs (thousands to millions) to calculate the principal component vectors. For instance, Li et al [18] used 2318 SNPs for continental-level classification. Besides, many studies were able to develop small panels of SNPs to analyze ancestral origins for people from a large number of populations, e.g., 73 populations in [19] and 119 populations in [10]. However, these have typically used unsupervised learning (clustering) methods, such as STRUCTURE [20], to show which populations cluster together, without explicit prediction of the sub-populations for the individuals.

Thus, though significant progress has been made in the use of genomic data for continental-level ancestry detection, sub-continental population detection using only a few marker SNPs is still a challenge. Not much has been done on identifying sets of ancestry informative SNPs (AISNPs) that can accurately distinguish closely related sub-populations, for instance, those from the same continent. Another challenge is that of computation, and the ever limited resources available in most labs, where such ancestry classification may be needed. Thus, we add a key constraint in addressing the problem: we require that only SNPs from only one chromosome can be used in the analysis. This is important, as it will mean that the required sequencing can focus only on the specified chromosome, thus saving time and sequencing cost. Essentially, our challenge therefore is to answer the question: how much information regarding our human biological and geographical ancestry can we glean from just a single chromosome?

In this paper, we address the problems of both continental and sub-continental ancestry identification using small SNP panels, with all SNPs in the panel coming from one single chromosome. For this study, we will focus on Chromosome 1, since this is the largest chromosome, and thus might provide the best starting point for our exercise. Thus, in this study, through analyzing the DNA information of Chromosome 1, we employed machine learning techniques and statistical approaches to identify small sets of SNPs for predicting an individual's continental and sub-continental origin. We take a three-stage approach. Initially, we apply parameter-based SNP selection, and later refined the selection by using a clustering technique (namely, DBSCAN [21]) to choose an efficient panel of SNPs. The final SNP panel is selected by applying a statistical approach based on pairwise correlation of the SNPs to identify the important AISNPs for both continental and sub-continental ancestry classification. Our continental classification is a five-class classification problem including the continents Europe, Latin America, Africa, East Asia and South Asia. Within each continent there are several closely related sub-populations and accurately distinguishing them is the challenging part. To address the sub-continental level classification problem, we focus on pairwise classification of sub-populations within each continent.

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

II. METHODS

A. Dataset and Pre-processing

For this work, we used the 1000 Genome Project, Phase 3 dataset [22] which contains information on 84.4 million variants (SNPs) from all 23 chromosomes for 2504 individuals, from 26 different sub-populations, from five continents. Table I provides a summary of different populations in the dataset. We focused on analyzing the variants from Chromosome 1 which is nearly 20.1 million SNPs. After data pre-processing steps (e.g., data cleaning), we identified continental and sub-continental ancestry informative SNPs in several stages. The DNA information for the 20.1 million variants (SNPs) from Chromosome 1 of each of the 2504 subjects resulted in a large dataset of size 61.2 GB. At the beginning, we extracted data from this large dataset and stored them in several smaller tables to be able to conduct our analysis in a MATLAB environment. 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 dip-loid, containing two nucleotides, from different combinations of the four nucleotide bases (A, C, G, T)). Next, we performed data cleaning operations on the extracted data based on the following criteria:

- The SNP loci which contain more than one reference nucleotides have been removed.
- If an alternate allele nucleotide also exists in the reference allele, corresponding SNP position is excluded from the analysis.
- SNP loci where each of the two nucleotides from all the individuals in the dataset both match with the reference allele's nucleotide are excluded from the analysis.

The above steps resulted in the removal of around 13 million SNPs in the cleaning stage. We then performed further analysis using the remaining SNPs. For the purpose of SNP selection, we removed a person's allele information from a SNP position, if the person's two nucleotides at the given position are the same as the reference allele's nucleotide. Consequently, two different sets of SNPs have been observed in the analysis. In one set, each SNP contains the same allele information among all individuals, although this allele information is different from the reference nucleotide. We call this SNP set the 'Similarity Set'. In contrast, in the other set, allele information is not the same among all individuals at the given SNP position. We call this set the 'Dissimilarity Set'. Since, for ancestry identification, we need to distinguish among populations with respect to some attribute/feature(s), SNP loci which demonstrate greater variation in DNA information among individuals will lead to better identification performance. Thus, we have chosen only the 'Dissimilarity Set' of SNPs for further analysis.

B. SNP Selection

Stage 1: Parameter-based SNP Selection:

At the beginning, we aimed to identify important markers for each of the 26 populations from the 'Dissimilarity Set' of SNPs. Consequently, we generated a structure array where each row allocates information from one SNP position containing 26 different fields corresponding to the 26 different populations.

TABLE I. 26 POPULATIONS IN THE DATASET

Population Code	Population Name	Continent	Sample Size
PUR	Puerto Rican	America	104
CLM	Colombian	America	94
PEL	Peruvian	America	85
MXL	Mexican-American	America	64
GBR	British	Europe	91
FIN	Finnish	Europe	99
IBS	Spanish	Europe	107
CEU	CEPH	Europe	99
TSI	Tuscan	Europe	107
CHS	Southern Han Chinese	East Asia	105
CDX	Dai Chinese	East Asia	93
KHV	Kinh Vietnamese	East Asia	99
CHB	Han Chinese	East Asia	103
JPT	Japanese	East Asia	104
PJL	Punjabi	South Asia	96
BEB	Bengali	South Asia	86
STU	Sri Lankan	South Asia	102
ITU	Indian	South Asia	102
GIH	Gujarati	South Asia	103
ACB	African-Caribbean	Africa	96
GWD	Gambian	Africa	113
ESN	Esan	Africa	99
MSL	Mende	Africa	85
YRI	Yoruba	Africa	108
LWK	Luhya	Africa	99
ASW	African-American SW	Africa	61

Each field associated with one population group contains relevant information regarding that group, such as, number of individuals of that group existing at that SNP position (since we removed individuals from a SNP position based on the similarity of their allele with reference nucleotide) and corresponding allele information of those individuals. Next, we calculate two parameters 'α' and 'β' at each dissimilar SNP position, for each of the 26 populations, viz:

$$\alpha_i = \frac{n_p^i}{n_p} \quad \text{and} \quad \beta_i = \frac{f_p^i}{n_p^i}$$

where, p=1, 2, … ,26

n_p^i = No. of individuals of population type p existing at SNP i

n_p = Total no. of individuals of population p in training data

f_p^i = Frequency of occurrence of the allele that appears most in population p at SNP i

For any population p, a SNP position i is considered important if at that position the product α×β=1 (i.e., α=1 and β=1). Based on the values of parameters α and β, we identify the best distinguishing SNPs for each population. After we obtain the set of important SNPs for each population, we take the union of all the 26 sets. The result is a unique set of 38,532 ancestry informative SNPs. From these 38K SNPs, we further removed the SNPs which contain the same allele information across all individuals from all 26 populations in the training set, since SNPs showing no variations between different population groups are not informative in distinguishing them. At the end of this stage, we have 34,631 ancestry informative SNPs in total, all from Chromosome 1.

Stage 2: Outlier-Based SNP Selection:

To further reduce the number of SNPs, we apply a cluster-based approach on the results from Stage 1. In particular, we take a contrarian approach: we group the SNPs using a clustering technique. In doing so, we also indirectly identify those SNPs that could not be grouped comfortably into any particular cluster. These are the outlier SNPs that do not seem

to be similar to other SNPs, and thus represent good candidates for use in discriminating between ancestries. Here, we use DBSCAN [35] as the basic clustering technique. Given a set of data points in some space, DBSCAN clustering method groups together points that are closely packed together, marking the points as outliers that lie alone in low-density regions. In our problem, SNPs that contain similar ancestry information are clustered together, while some SNPs are identified as outliers with seemingly unique ancestry information. These outlier SNPs are considered good candidates for distinguishing among populations.

Here, we apply DBSCAN clustering on the 34K SNPs extracted in the previous stage of selection. The algorithm requires three inputs: data matrix *D*, radius parameter (ε) and neighborhood density threshold (*MinPts*). Data matrix *D* has 34K number of rows associated with 34K SNPs and each SNP is considered as an object with *l* dimensions, where *l* denotes number of training individuals. Each dimension belongs to the allele information of a training subject represented by a number between 1-16, since four nucleotides {A, C, G, T} generate 16 possible allele symbols {AA, AC, …, TT}. The radius parameter ε is measured as the Euclidean distance between two *l*-dimensional SNP objects and the neighborhood density threshold *MinPts* defines the minimum number of points required to form a cluster. For this problem, we have empirically chosen *MinPts*=2 and ε=0.1. Using DBSCAN clustering technique, we have obtained 2378 clusters and 6404 outliers. These 6404 outlier SNPs constitute our new set of candidate SNPs for ancestry identification.

Stage 3: Correlation-based SNP Selection:

As we obtain the set of 6404 SNPs from the clustering technique, we measure the overall 26-class ancestry prediction performance for each individual SNP marker. That is, we perform ancestry prediction using each of the 6404 SNPs, independent of the other SNPs. Of course, we do not expect to produce very good performance for a single SNP. However, the relative performance of the SNPs is a crucial piece of information for our approach. Consequently, a performance matrix *X* is generated with *m*=6404 rows, where each row of the matrix is allocated for one SNP representing a six-dimensional vector,

$$\underline{x}^{(i)} = [x_1^{(i)}\ x_2^{(i)}\ x_3^{(i)}\ x_4^{(i)}\ x_5^{(i)}\ x_6^{(i)}]$$

The first element in the vector contains the accuracy in 26-class classification using SNP *i*. The next five elements of the vector are related to the five continents, whereby each element records the percentage of test individuals correctly predicted for the given continent. Classification into 26 populations by each SNP has been conducted for 80%-20% train-test split, with *n*=2504 individuals. We have used an allele-context feature to represent each SNP during classification, where each SNP's allele-context feature belongs to three possible values: 0, 1, 2. Here, '0' means both nucleotides from an individual at a given SNP location (say *i*) are the same as the reference nucleotide; '1' means that one of the two nucleotides is different from the reference nucleotide; and '2' means that both nucleotides of the person are different from the reference nucleotide. Allele-context feature vector a and class-label vector b are denoted for both train and test sets as follows:

$$\underline{a}_{train}^{(i)} = [a_1^{(i)}\ a_2^{(i)} .....a_l^{(i)}]^T \text{ and } \underline{a}_{test}^{(i)} = [a_1^{(i)}\ a_2^{(i)} .....a_{(n-l)}^{(i)}]^T$$

$$b_{train} = [b_1\ b_2 .....b_l]^T \text{ and } b_{test} = [b_1\ b_2 .....b_{(n-l)}]^T$$

Here, *l*=number of training subjects
n-l=number of test subjects

Thus, for *i*=1,2,…,*m* number of SNPs, the overall performance matrix is represented as,

$$X = [\underline{x}^{(1)}\ \underline{x}^{(2)} .....\underline{x}^{(6404)}]^T$$

Once the performance matrix *X* is created, we calculate the pairwise correlation between the SNPs, using the associated performance vectors. For example, correlation of SNP *i* and SNP *k* is calculated using the Pearson's correlation coefficient as follows:

$$C = \frac{\sum_{j=1}^{5}(x_j^{(i)} - \bar{x}^{(i)})\,(x_j^{(k)} - \bar{x}^{(k)})}{\sqrt{\sum_{j=1}^{5}(x_j^{(i)} - \bar{x}^{(i)})^2}\ \sqrt{\sum_{j=1}^{5}(x_j^{(k)} - \bar{x}^{(k)})^2}}$$

Here,

$x_j^{(i)}$ =element of the vector $\underline{x}^{(i)}$ for continent *j* (j=1,2,..,5)

$\bar{x}^{(i)}$ =average of the five $x_j^{(i)}$ elements of vector $\underline{x}^{(i)}$

Now, if the correlation coefficient *C* between SNP *i* and SNP *k* is above a certain threshold *th*, that is, they are highly correlated, one of them is kept in the analysis and the other one is removed. Here, the SNP that provides better classification accuracy in the performance matrix (represented by the first element of vector $\underline{x}^{(i)}$) is considered as "non-redundant", while the other SNP is assumed redundant. The proposed correlation method for SNP selection is described below using a pseudo code.

Proposed Algorithm: Correlation-based SNP selection

```
Flag each SNP as non-Redundant
FOR i = 1 to total number of SNPs
    IF SNP(i) is non-Redundant
        FOR k = i+1 to total number of SNPs
            IF SNP(k) is non-Redundant
                Calculate correlation coefficient C between performance
                feature vectors of SNPs i and k
                IF C > threshold,
                    Flag SNP(k) as Redundant
                END IF
            END IF
        END FOR
    END IF
END FOR
Remove Redundant SNPs
```

Having described the general procedure for selecting the SNPs, the final step will be to select those that are suitable for continental-level classification, and those that are suitable for more localized discrimination between sub-populations.

a) SNP selection for continental-level classification

To find the best candidate SNPs for continental level classification, the proposed correlation based SNP selection has been exploited. First, the 6404 SNPs are ranked from highest to lowest based on their classification accuracy in the performance matrix *X* and 6404×6 performance matrix is rearranged accordingly. Following this rank of the SNPs, we create the order of the SNPs for the initial 'non-Redundant SNP set' in the algorithm and the algorithm is initialized with the best performing SNP. For a certain correlation threshold *th*, the algorithm is executed to identify the final set of non-

Redundant SNPs from the 6404 SNPs. These candidate SNPs represented by the allele-context feature are subsequently used to perform the five-continent classification for 80/20 train-test split. We carried out empirical experiments for a range of values of correlation thresholds and the threshold which provides the best classification performance with the smallest set of SNPs has been finally selected.

b) SNP selection for sub-population-level classification

When an individual's continental ancestry is known and the individual belongs to any of the two possible closely related sub-populations within that continent, the objective is to identify the accurate sub-population ancestry. In this work, we have selected candidate SNP sets for all possible pairwise classification of sub-populations within a continent exploiting the same correlation algorithm as used in the continental-level ancestry identification. Assume two sub-populations S_1 and S_2 from the same continent *j* and the goal is to identify a powerful set of candidate SNPs which will be able to distinguish individuals from these two sub-populations. Now, the 6404 SNPs are ranked from highest to lowest based on the continent *j* elements ($x_j^{(i)}$) in the performance matrix X and performance matrix is rearranged accordingly. Thus, the correlation algorithm is initialized with the best performing SNP for continent *j* and for a certain threshold the algorithm is executed to obtain the non-Redundant set of SNPs from the 6404 SNPs. Next, using the allele-context feature of these SNPs, binary classification between the two sub-populations is performed for 80/20 train-test split. Similar to continental-level classification, we tested for a range of values of correlation thresholds and chose the threshold that provides the best classification performance with a small set of SNPs.

c) Ancestry classification algorithm

Having identified the best SNP subsets, any standard classification algorithm (e.g., SVM, Random Forest, etc.) can be used for ancestry classification. In this work, we have applied softmax neural network [23] for both continental and sub-continental classification problem.

III. EXPERIMENTAL RESULTS

We performed experiments using the identified 1000 Genome dataset, with 26 sub-populations, from 5 continents. We evaluated performance of the proposed approach on both continental-level and sub-population-level ancestry prediction/ classification, as described below.

A. Continental Classification

The five-class classification into five continents -- Europe, America, East Asia, South Asia and Africa has been performed for a range of values of correlation threshold *th*=0.1 to 0.99 with an interval of 0.01. Fig. 1 depicts the overall performance in continental-level classification for *th*=0.4 to 0.99 with 0.01 interval along with the corresponding number of SNPs. The highest performance achieved is 99.19% for *th*=0.98 with 614 SNPs marked with a red square in the plot. But, since our goal is to rather use a smaller SNP panel for distinguishing continental populations, we searched for the threshold *th* that provides an optimum performance with less number of SNPs (e.g., 200 or less). From Fig. 1, we can observe the general trend in performance for the proposed approach. At *th*=0.4, the system suggests a panel of 10 SNPs, for an overall classification accuracy of about 80%. Performance generally increased with increasing correlation threshold, rising to about 94% accuracy rate, at *th*=0.82, using 93 SNPs. The best classification result is considered the one for correlation threshold *th*=0.91, resulting in a classification accuracy of 96.75% with 206 SNPs marked by the magenta square. These 206 SNPs have been considered as our final candidate SNPs for continental-level classification. The confusion matrix for five-class continental classification problem with overall performance of 96.75% is shown in Table II. Our continental classification performance has been compared with other related methods in TABLE III.

TABLE II. CONFUSION MATRIX FOR CONTINENTAL-LEVEL ANCESTRY CLASSIFICATION (OVERALL ACCURACY OF 96.75%, 206 SNPS)

Continents	Europe	America	Africa	East Asia	South Asia
Europe	**94.06%**	3.96%	0.00%	0.00%	1.98%
America	10.94%	**89.06%**	0.00%	0.00%	0.00%
Africa	0.00%	0.00%	**100.00%**	0.00%	0.00%
East Asia	0.00%	0.00%	0.00%	**100.00%**	0.00%
South Asia	1.02%	2.04%	0.00%	0.00%	**96.94%**

B. Pairwise classification between sub-populations

Table IV shows the overall pairwise classification results between sub-populations in each of the five continents in our dataset. The number of SNPs required for each classification has also been noted. From the table, it is evident that in all cases of pairwise classification of closely related populations, we can infer the ethnicity using a small panel of SNPs (less than 200) and for some instances, the accuracy is as high as 100%. For a more detailed analysis, Fig. 2(a) and Fig. 2(b), show the performance of the proposed methods with increasing correlation thresholds, using sub-populations from the continent America. The best performance has been marked with a red square in the figures. As can be observed, it is relatively easy to distinguish between individuals from certain sub-populations, even within the same continent. For instance, Fig 2(a) shows that individuals from Puerto Rico (PUR) are relatively easy to distinguish from those from Peru (PEL), achieving a 100% accuracy rate, using 56 SNPs, under our approach. However, we can also see some challenging cases, such as Columbia (CLM) and Mexico (MXL) (Fig. 2(b)), where the highest accuracy is at ~74%, using 37 SNPs. Even increasing the number of SNPs beyond 37 could not improve the result. We have shown comparative results of binary/pairwise classification of sub-populations with other studies in the literature in TABLE V. The comparative results show the proposed methods are competitive with the state-of-the-art methods, even when using information from just one chromosome.

TABLE III: COMPARATIVE PERFORMANCE ON CONTINENTAL-LEVEL ANCESTRY CLASSIFICATION USING SNPS

Basic Method	Data Size	Datasets Used	Classification Rate %
[24]	664	Multiple Datasets	96.1
[5]	2689	1000 Genome, HGDP, NIST	98.8
[25]	6410	Multiple Datasets	81.4
[6]	451	Own Collection	77.0 (+21.6 thresholded out)
Proposed	2504	1000 Genome Phase 3	**99.19 (614 SNPs)**
			96.75 (206 SNPs)

TABLE IV: RESULTS FOR PAIRWISE CLASSIFICATION BETWEEN SUB-POPULATIONS IN EACH CONTINENT

Continent	Sub-populations	Number of SNPs	Correlation Threshold	Accuracy (80-20)
America	PUR-PEL	56	0.76	100.00%
	PUR-MXL	44	0.72	93.33%
	PUR-CLM	89	0.83	66.67%
	CLM-PEL	96	0.84	97.06%
	CLM-MXL	37	0.69	74.07%
	PEL-MXL	96	0.84	84.00%
Europe	GBR-FIN	15	0.47	78.38%
	GBR-IBS	63	0.80	66.67%
	GBR-CEU	30	0.64	67.57%
	GBR-TSI	24	0.61	76.92%
	FIN-IBS	82	0.83	83.33%
	FIN-CEU	130	0.88	80.00%
	FIN-TSI	75	0.82	90.48%
	IBS-CEU	47	0.75	71.43%
	IBS-TSI	82	0.83	77.27%
	CEU-TSI	31	0.67	73.81%
East Asia	CHS-CDX	44	0.73	64.10%
	CHS-KHV	12	0.41	68.29%
	CHS-CHB	30	0.66	64.29%
	CHS-JPT	83	0.84	73.81%
	CDX-KHV	30	0.66	68.42%
	CDX-CHB	120	0.87	76.92%
	CDX-JPT	120	0.87	87.18%
	KHV-CHB	62	0.79	75.61%
	KHV-JPT	92	0.85	82.93%
	CHB-JPT	83	0.84	71.43%
South Asia	PJL-BEB	29	0.65	74.29%
	PJL-STU	57	0.78	62.50%
	PJL-ITU	29	0.65	70.00%
	PJL-GIH	153	0.89	100.00%
	BEB-STU	42	0.72	72.97%
	BEB-ITU	139	0.88	70.27%
	BEB-GIH	113	0.86	100.00%
	STU-ITU	29	0.65	64.29%
	STU-GIH	79	0.82	100.00%
	ITU-GIH	79	0.82	100.00%
Africa	ACB-GWD	47	0.76	76.74%
	ACB-ESN	20	0.56	79.49%
	ACB-MSL	46	0.75	71.43%
	ACB-YRI	43	0.72	80.49%
	ACB-LWK	60	0.79	79.49%
	ACB-ASW	15	0.49	81.48%
	GWD-ESN	46	0.75	77.27%
	GWD-MSL	73	0.82	72.50%
	GWD-YRI	132	0.88	100.00%
	GWD-LWK	132	0.88	100.00%
	GWD-ASW	132	0.88	96.88%
	ESN-MSL	102	0.86	69.44%
	ESL-YRI	132	0.88	100.00%
	ESN-LWK	132	0.88	100.00%
	ESN-ASW	132	0.88	96.43%
	MSL-YRI	38	0.71	100.00%
	MSL-LWK	132	0.88	100.00%
	MSL-ASW	73	0.82	91.67%
	YRI-LWK	28	0.65	78.57%
	YRI-ASW	146	0.89	90.00%
	LWK-ASW	162	0.90	85.71%

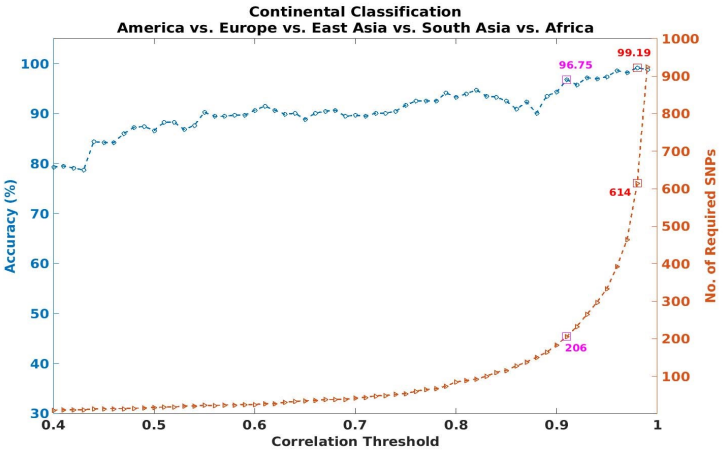

Fig 1. Continental classification results with varying thresholds

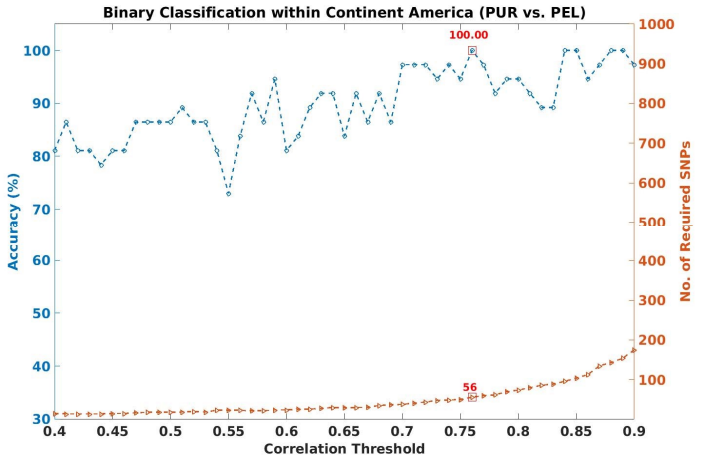

Fig 2 (a). Pairwise classification results (PUR vs. PEL) with varying thresholds

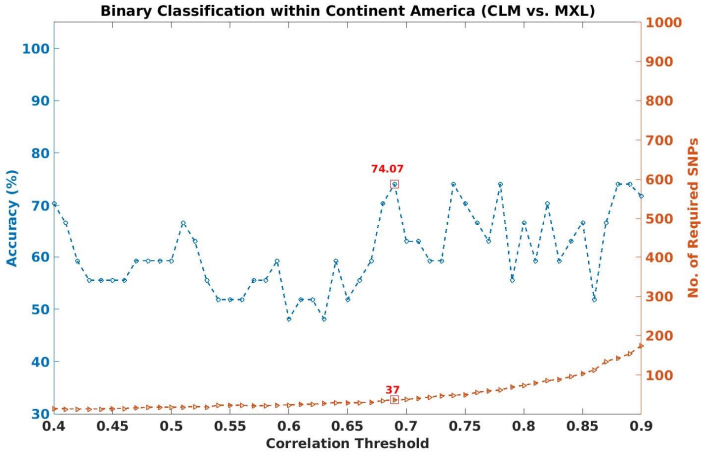

Fig 2 (b). Pairwise classification results (CLM vs. MXL) with varying thresholds

TABLE V. COMPARATIVE PERFORMANCE IN SUB-POPULATION-LEVEL ANCESTRY CLASSIFICATION

Pairwise sub-populations	Continent	Method	Data size	Datasets	Classification rate (%)	No. of attributes
CEU-TSI	EUROPE	[26]	267	HAPMAP III	86.6±2.4	180 SNPs
--	EUROPE	**PROPOSED**	503	1000 GENOME PHASE 3	76.6*	58 SNPS**
CHB-JPT	EAST ASIA	[26]	250	HAPMAP III	95.6± 3.9	877 SNPs
JPT-CHB	EAST ASIA	[27]	9104	OWN COLLECTION	74.9(77.2***)	15 STR LOCI
JPT-KOR	EAST ASIA	[27]	731	OWN COLLECTION	67.9 (63.7***)	15 STR LOCI
CHB-KOR	EAST ASIA	[27]	731	OWN COLLECTION	69.6 (62.4***)	15 STR LOCI
--	EAST ASIA	**PROPOSED**	504	1000 GENOME PHASE 3	73.3*	68 SNPS**
LWK-MKK	AFRICA	[26]	294	HAPMAP III	95.9±1.5	341 SNPs
--	AFRICA	**PROPOSED**	661	1000 GENOME PHASE 3	87.02*	87 SNPS**

*Average accuracy of all pairwise sub-population classifications within the given continent.
**Average number of SNPs required in all pairwise sub-population classifications within the given continent
*** Results obtained without normalization.

IV. CONCLUSIONS

In this work, we have developed an ancestry identification system to predict continental origin of an unknown individual and also to distinguish between closely related sub-populations within a continent. Here, only SNPs from just one chromosome (namely, Chromosome 1) have been analyzed to identify different panels of ancestry informative SNPs. Both machine learning and statistical approaches have been employed for selecting candidate SNPs. Our results demonstrate that one single chromosome (in particular, Chromosome 1), if carefully analyzed, could hold enough information for accurate prediction of human biogeographical ancestry. This has a significant implication in terms of the computational resources required for analysis of ancestry, and in the applications of such analysis, such as in studies of genetic diseases, forensics, and biometrics. An interesting further work is to investigate the performance of other chromosomes, especially the smallest chromosomes, to see if we can construct equally high-performing panels of ancestry informative SNPs using even less information.

REFERENCES

[1] Enoch MA, Shen PH, Xu K, Hodgkinson C, Goldman D: "Using ancestry informative markers to define populations and detect population stratification," *J Psychopharmacol*, 20 (2006):199-126.

[2] Araújo, Gilderlanio S., et al. "Integrating, summarizing and visualizing GWAS-hits and human diversity with DANCE (Disease-ANCEstry networks)." *Bioinformatics* 32.8 (2016): 1247-1249.

[3] Tian, Chao, et al. "A genomewide single-nucleotide–polymorphism panel for Mexican American admixture mapping." *The American Journal of Human Genetics* 80.6 (2007): 1014-1023.

[4] Sanderson, Jean, et al. "Reconstructing past admixture processes from local genomic ancestry using wavelet transformation." *Genetics* 200.2 (2015): 469-481.

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

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

[7] Krimsky, S and Simoncelli, T, *Genetic Justice: DNA Data Banks, Criminal Investigations, and Civil Liberties*, Columbia University Press, (2012).

[8] Amirisetty S., Hershey G.K., Baye, T.M. "AncestrySNPminer: A bioinformatics tool to retrieve and develop ancestry informative SNP panels", Genomics 100 (2012) 57–63

[9] N.M. Silva, L. Pereira, E.S. Poloni, M. Currat, "Human neutral genetic variation and forensic STR data", PLoS ONE 7 (2012) e49666

[10] Kidd, Judith R., 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): 1.

[11] Nassir, Rami, et al. "An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels." *BMC Genetics* 10.1 (2009): 39.

[12] Wright S: *Evolution and the Genetics of Populations, vol 2: The Theory of Gene Frequencies*, Chicago and London: University of Chicago Press; 1969.

[13] Kosoy, Roman, et al. "Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America." *Human Mutation* 30.1 (2009): 69-78.

[14] Halder, Indrani, et al. "A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications." *Human Mutation* 29.5 (2008): 648-658.

[15] Campbell, Catarina D., et al. "Demonstrating stratification in a European American population." *Nature Genetics* 37.8 (2005): 868.

[16] Price, Alkes L., et al. "Principal components analysis corrects for stratification in genome-wide association studies." *Nature Genetics* 38.8 (2006): 904.

[17] Patterson, Nick, Alkes L. Price, and David Reich. "Population structure and eigenanalysis." *PLoS genetics* 2.12 (2006): e190.

[18] Li Y, Byun J, Cai G, et al, "FastProp: A rapid pricipal componet dericed method to infer intercontenental ancestry using genetic data", *BMC Bioinformatics*, 17:122 (2016).

[19] Kidd, Kenneth K., et al. "Progress toward an efficient panel of SNPs for ancestry inference." *Forensic Science International: Genetics* 10 (2014): 23-32.

[20] Pritchard, Jonathan K., et al. "Association mapping in structured populations." *The American Journal of Human Genetics* 67.1 (2000): 170-181.

[21] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *KDD*. 96.34 (1996).

[22] 1000 Genomes Project Consortium. "A global reference for human genetic variation." *Nature* 526.7571 (2015): 68.

[23] Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

[25] 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.

[26] Hajiloo, Mohsen, 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): 61.

[27] Graydon, Matthew, François Cholette, and Lay-Keow Ng. "Inferring ethnicity using 15 autosomal STR loci—Comparisons among populations of similar and distinctly different physical traits." *Forensic Science International: Genetics* 3.4 (2009): 251-254.