Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 552–561

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Deep Learning Approach to Biogeographical Ancestry Inference

Yue Qu[a], Dat Tran[a,*], Wanli Ma[a]

[a]*Faculty of Science and Technology, University of Canberra, ACT 2601, Australia*

Abstract

Biogeographical ancestry (BGA) inference is based on the understanding of genetic diversity distribution among population groups. BGA inference is used to detect and measure the population structure that presents the natural assignment in genetic terms, identify genetic patterns found in individuals' genotypes, and estimate an individual's BGAs. In the context of forensic, BGA inference at an individual level gives the possibilities to achieve more complete identification of missing person or suspect. Current machine learning approach to BGA inference based on Bayesian theory and principle component analysis cannot operate on the data sequence directly and require predefined features extracted from the data sequence based on prior knowledge. In this paper, we conduct a survey of the state of the art of BGA inference and propose a new approach based on deep learning to BGA inference without prior feature extraction to find hidden genetic structure and provide more accurate predictions. Our experiments conducted on the dataset for Human Genome Diversity Project (HGDP) show better results for the proposed approach.

Keywords: Biogeographical ancestry inference, machine learning.

1. Introduction

Biogeographical ancestry (BGA) inference is based on the understanding of genetic diversity distribution among population groups.

BGA indicates an individual's populations of origin with respect to the geographical regions [4, 64]. Unlike other concepts of the ancestry that are defined on the basis of linguistic, culture, or ethnic, BGAs follow a broadly continental distribution of population diversity delimited by geographical barriers, mass movements, and slightly, by the nature selection, and shown to match well to the distributions of genetic variation. BGA inference comes from the exploration of underlying genetic relationships found in human genetic data, assessing and measuring the population structure that presents the natural assignment in genetic terms, and identifying genetic proportions found in individuals' genotypes [56].

* Corresponding author. Tel.: +61-2-6201-2394.
E-mail address: dat.tran@canberra.edu.au

1877-0509
10.1016/j.procs.2019.09.210

The field of population genetics as a whole could benefit anthropology and the disease association studies in epidemiology. Assessing and measuring the ancestral genetic structure existed in modern population groups will aid anthropological studies in illuminating the patters of past human migrations [1], demographical processes [44], and the impact of natural selection [58]. In the genetic association studies and disease mapping, for instance, admixture mapping is used as a strategy to map the susceptibility in light of BGAs distributing among the group where individuals originate from admixed populations [52, 19, 62]. The strategy helps to avoid misinterpretations by reducing the bias due to populations stratification [6, 19]. When admixture mapping is used to identify disease, a well-structured population stratification supports the validation of results; otherwise, the presence of undetected population patterns will lead to spurious associations [73, 7].

In the context of forensic, the BGA inference at an individual level, the other part of BGA inference, gives the possibilities to achieve more complete identification of missing person or suspects. Yet ancestry inference can provide information about some externally visible characteristics (EVCs) such as eye, hair, and skin color [10, 34, 65]. It suggests the possibilities of unusual EVC combinations of individuals. Autosomal genotypes are human genotypes that provide comprehensive information about individuals ancestries because they cumulatively records the genome history. Current approaches to BGA inference through autosomal genotypes use genetic distances or allele frequencies as measures of variation, ancestral informative markers (AIMs) as selected features, and Bayesian or Principle Component Analysis (PCA) as inference and prediction technique. However, these current approaches have the following limitations: 1) Identifying AIMs requires domain knowledge and is labour-intensive, 2) the inference accuracy through AIMs is limited due to the absence of sharp discontinuities [33] in the neutral genetic diversity among human populations, and 3) Specifically, a set of AIMs is not always sufficient for estimating the ancestral proportion of the individuals who originate from multiple population.

Motivated by the success of deep neural networks in similar research areas [3] and the availability of human DNA sequence data, we propose to use convolutional neural networks (CNN) to BGA inference problem. The proposed approach has the following advantages: 1) The capacity of learning from data without prediction features, 2) The capacity of learning from increasingly large and high-dimensional data sets (e.g. from DNA sequencing and flow cytometry microscopy), and 3) The capacity of capturing nonlinear dependencies in genetic sequences. We conducted experiments on the dataset for Human Genome Diversity Project (HGDP) and with the basic CNN and raw input, we achieved the inference accuracy of 86% and outperformed the current techniques based on Bayesian and PCA.

The paper is organised as follows. Section 2 presents genetic approaches to BGA inference. Section 3 presents machine learning approaches to BGA inference. The proposed CNN is presented in Section 4. Our experiments are presented in detail in Section 5. Finally we conlude the paper in Section 6.

2. Genetic Approaches to BGA Inference

2.1. Human Genotypes

Haploid genotypes: As uniparentally inherited DNA, mitochondrial DNA (mtDNA) provides information about female-to-female transmitted lineage and Y chromosome is informative about male-to-male transmitted lineage [57]. As haploid DNA is undisrupted by recombination, it is widely used in (commercial) lineage tests [13, 9]. The linage-based ancestry estimation usually involves the comparison between the reference genotypes and the genotypes of unknown donors, which is justified to infer two individuals with the same haploid DNA share the same paternal or maternal ancestor [18, 35]. And the highly differentiated geographical regions signaled by haplotypes provide more explicit information of genetic factors such as the patterns of mating and recent migration events [68, 5].

However, moving such a lineage-based inference to the conclusion that the match implies individuals' BGAs remains the dangerous of misinterpretation, because the haploid genotypes reflect the maternal or paternal fraction of an individual's total biogeographical ancestral histories. For example, although an individual's Y chromosome that contains rare Y chromosome types is able to indicate paternal genetic trace of a particular population branch, the proportion of paternally inherited component to the individual's total ancestral histories is uncertain, also the upper limit on the time of linkage is less certain [30]. Another source of the misinterpretation is the asymmetry between paternal and maternal ancestral histories. The two uniparentally inherited DNA may provide evidence for different ancestral histories which eventually signal that every individual originate from different geographic regions [17, 8].

Autosomal genotypes: Compared to haploid DNA, autosomal DNA is more widely used in BGAs estimation because of its stability and density of distribution [49]. Autosomal DNA, inherited from both parents, provides more comprehensive information about individual's ancestries because it cumulatively records the genome history [57]. This review centers on the methods using autosomal genotypes; measuring autosomal variation allows the studies of genetic admixture contributed by all of an individual's ancestors rather than just a proportion of them [23].

2.2. The Informativeness Carried by Genes

The genetic variation can be measured by either 1) Searching for genetic variations along genotypes [25, 37, 36, 40] or 2) Assessing the loci where present populations in light of allele frequencies. Whole-genome approach involves exploring and analyzing genetic patters along an individual's entire DNA sequences; although idea and has been somewhat facilitated by the sequencing methods [32, 71], it is prohibitively expensive [57].

Ancestry Informative Markers (AIMs): The informative markers, a relatively small set of single nucleotide polymorphisms (SNPs), has contributed to the BGA inference effort in a way of signaling particular population groups. The loci where exhibit extreme allele frequency differences between populations are often counted as ancestry informative markers (AIMs) [49, 23] or the ancestry-sensitive markers [27] to describe the uncertainties. The skewed allele frequency markers that are common in one population while are rare in others are also incorporated in AIMs [49]. The abundance and low mutation rates of AIMs [21] permit their practicality. The use of AIMs reduces the amount of data required for BGA inference. Compared with genome-wide SNPs or a random set of SNPs, a relatively small set of AIMs make good economic sense [16]. The alternative alleles such as Insertions and Deletions offer the similar possibilities and bring their own valuable characteristics to population structure detection and BGA inference [48]. The exploration of HGDP-CEPH has revealed Indels' ability to distinguish four major population groups: Africans, Asians, Europeans, and Naïve Americans [46]. Further, a set of 46 AIM-Indels has shown to successfully differentiate Central South Asian and Middle Eastern Asian for which SNPs do not work well [61].

The measurement of informativeness carried by markers: A common concern about the use of AIMs is to explore and select the criteria based on which one could use to identify the most optimal AIMs [27]. Ronsenberg et al [54] introduced a general measurement of the "informativeness" that multiallelic markers would provide about BGAs. This gives rise to the question of how to measure the contribution of specific markers to BGA inference. Measuring the "information amount" is tackled by an likelihood approach: the quantity can be viewed as the expected log-likelihood associated with drawing an allele randomly from a set of populations, where the correlation between admixture proportions and allele frequencies is given by G statistics [66]; the loci of high informativeness is equivalent to the loci with large values of G. The maximal correct assignment probability occurs when each allele is assigned to the population where it presents most frequently.

Identified AIM panels: There have been a handle of AIM-panels proposed. A study by Phillips et al. [49] publishes 34 AIMs from HGDP-CEPH which pronounce the allele frequency discontinuities between Africans, Europeans, and East Asian populations [49]. The group [48, 61, 59, 60] keeps their endeavor towards the identification of autosomal AIMs as well as the ancestral sensitive Indels with the measurement of "informativeness" [54] based on Wright's Fst. Keeping exploring HGDP-CEPH and 1000 Genomes datasets, Phillips et al. [48] complement the original 34-AIM-panel with 23 additional AIMs that differentiate Europeans and South Asians. Followed above studies, more ancestry sensitive indels are identified to distinguish American, Central South Asian, Middle Eastern Asian, and Oceanian populations [61, 60]. In another panel proposed by Kosoy et al. [31] there are 128 SNPs from HGDP-CEPH and 1000 Genomes, which are subsequently refined into a smaller set of 93 SNPs [43], discriminating continental subject groups including Africans, Europeans, and Amerindians. Later, a smaller panel of 55 SNPs was identified by Kidd et al. [28, 29] based on the high Fst values amongst populations. Kidd's panel is found to be capable of distinguishing African, European, East Asian, Naive American, and Oceanian groups.

Issues of using AIMs for inference purpose: An accurate prediction of BGAs through AIMs depends on, but is not limited to, the selection of loci and population sampling. The degree of population differentiation, including how the populations are sampled, and the size of each sample strongly affects the accurate estimation of population structure and individuals' BGA inference. It is important to be clear about what reference panel is being used for the inference. Different genetic markers are subject to different selective forces [70]. Ideally, a AIM-panel should have a large spread in terms of allele frequency differences [14]. It is, however, usually difficult to obtain a sufficient spread of population data that present the full range of a population variation [47]. On the other hand, each segment of

individual's genome has its own ancestral history due to the recombination; different markers extracted from different segments may reveal the ancestral histories that trace to different subpopulations [57]. For example, several studies used STRUCTURE [51, 20] to detect the (unknown) population structure presented in a dataset but obtained different population resolution (different values of K) due to the use of various AIM-panels. Furthermore, the selected AIMs are ideally to differentiate most population and simultaneously to retain the differences between individuals within the sample population; since not every person from a given population has the AIMs identified for the population, while individuals from other populations may have [57]. The selection of AIMs, as pointed by Cavalli-Sforza [12], demands the adequacy of the coverage which is more important than the total number of data.

2.3. Methods of Analysis

2.3.1. Genetic Distance-Based Approach

Genetic distances and pure genetic-distance-based model: Genetic distances, e.g. the Wright's Fst, Tajima's D, T1 statistic, or the linkage disequilibrium (LD), are used widely to describe the genetic difference between two populations [12]. Loh [39] demonstrates a straightforward inference from linkage disequilibrium (LD). Their proposed model [39] harnesses the weighted LD as a function of genetic distance for the purpose of admixture estimation of African populations, based on the feature of admixture events retained by LD [42].

Genetic distances as the supplementary in BGA models: A pairwise matrix of genetic distances between all the possible pairs of populations or individuals may reveal the hidden genetic relatedness among the samples [67]. For example, Wright's Fst describes the probability of identity of alleles from a population compared with randomly chosen alleles [54]. In various BGA inference models, the Wright's Fst is used to measure how well the selected genetic markers (i.e. SNPs, Indels, etc.) distinguish different populations [49] and how much information carried by a marker [54]. This "preprocessing" procedure facilities the identification of informative markers and consequently significantly speeds up the BGA inference. The genetic distances can also usefully supplement the construction of genetic trees, i.e. a dichotomous tree, in evolution analysis because of its evolutionary meaning [12].

Issues of using genetic distances in BGA inference: Although genetic distances provide a clearer representation of population differences, the results depend heavily on the measurement. It is difficult to determine the level of confident that the clusters obtained are meaningful [51]. Meanwhile, a proportion of genetic information will be ignored during the calculation, suggesting that the genetic distance-based BGA inference models, even using a large amount of markers, are not suitable for detecting fine population structures[69].

2.3.2. Allele Frequency-Based Approach

The quantitative methods for analyzing genetic data based on allele frequencies of human population were pioneered by Cavalli-Sforza and Edwards in 1964 [63]. Some of the methods, such as likelihood analysis and variance analysis, are still widely used in BGA inference.

The method of likelihood analysis statistically infers the population structure and probabilistically assigns individuals to clusters. There have been several models, such as the Bayesian-MCMC model used by STRUCTURE [51], the PCA-Bayes used by Snipper [49], or Hidden Markov Models, found to be well fit genetic data for the inference purpose. The method based on variance analysis searches for the clusters that maximize the ratio of variances between branches [12]. Conducting a decades-long survey, Cavalli-Sforza et al. collected the genetic variants from various geographical population groups and produced an allele-frequency map for each allele. When these maps constitute a geographical maps of gene frequencies, the various levels of allele frequencies form the isopleths of the geographical map [12]. A geographical map conveys immediate information of the maxima, the minima, and the gradients of gene frequencies across space. The gradients of variation indicate how closely the populations are related, whereby they suggest the gene exchanges among populations [53, 41]. The map usefully supplements the genetic trees which further reveal individuals' population of origins [22, 50, 38].

3. Current Machine Learning Approach to BGA Inference

The analysis of DNA sequence variation and the efforts of studying from binary polymorphisms such as single nucleotide polymorphisms (SNPs) to multiple-allele loci such as short tandem repeats (STRs) have increased the

knowledge of diversity amongst and history of human populations. Despite this progress, the prompt introduction of machine learning techniques followed by statistical models alongside appropriate biological meanings have facilitated and accelerated the analytical procedure, bringing the possibility to analysis large volumes of data and yielding a blur to relatively clearer informative insights. Some of the methods, making use of allele frequency differentiation, apply clustering algorithms such as the Bayes and PCA to detect candidate population clusters and probabilistically assign individuals to one or more groups based on either likelihood-based models or phylogenetic trees. Another class of approach, making use of the power of machine learning algorithms, condense the associations among loci to make inference based on genetic distances such Fst and linkage disequilibrium (LD).

It has been shown that the people belonging to the same geographical group almost present very similar ancestral proportions [24]. This makes the statistical evaluation of genetic relationships among populations meaningful. An unsupervised learning algorithm learns the structure or useful properties of the dataset to detect possible population structure existed based on the measurements such as allele frequencies or genetic distances. A supervised learning algorithm experience the dataset that contains genotypes and geographical features, observe each sample and associated population labels, and ideally to assign new genetic profiles into the population groups based on its measurements.

3.1. Bayesian-based BGA Inference Systems: STRUCTURE and SNIPPER

3.1.1. STRUCTURE

The most widely used Bayesian-based BGA Inference system is STRUCTURE [51]. STRUCTURE characterizes each population by a set of allele frequencies. It assumes that the allele at each locus in each sample is drawn from an identical probability distribution, and the markers are at complete linkage equilibrium and are of Hardy-Weinberg equilibrium within populations. These make an approximation for the populations that are not closely correlated with one another. The approximation promises a population to not have very similar allele frequencies with other populations and the inference of this population structure is then drawn from the allele frequency distribution among populations. The problem of inferring population structure becomes that of estimating the number of clusters K presented in a dataset. STRUCTURE applies the Bayes' rule to tackle this problem. Using the allele frequencies estimated by Gibbs sampler [15], the model updates the prior and makes inference based on the posterior which yields the likelihood of BGA membership.

STRUCTURE is designed to distinct Hardy-Weinberg populations and to assign individuals into these populations. The unsupervised learning procedure makes the model very useful when there is little geographical information. A latter study [20] further expanded STRUCTURE to allow the linkage among input markers by eliminating the linkage disequilibrium generated from the variation in ancestries among samples and the correlations in ancestry along each chromosome. This assumes that various clusters have a descended ancestral population. Being tested using HGDP-CEPH dataset, the correlated model is shown to perform better in distinguishing similar but distinct clusters [55].

Another fundamental problem is the choice of K. STRUCTURE is designed to detect the presence of population structure amongst samples. It distinguishes five continental populations in HGDP-CEPH database successfully, although STRUCTURE indicates 6 clusters with one of them corresponds to individual populations [56]. The choice of K that best fits the data is not always easy to achieve. In the original study [51], STRUCUTRE was tested using a dataset that consisted of two population groups. The model did indicate the *true* clustering result where $K = 2$, while $K = 5$ also showed relatively high posterior probability. This may reflect the subpopulations however, it is difficult to interpret. Allele frequency distribution did not indicate additional discrete populations. When K is set to a large value to detect subpopulations, the clustering results are found to be erratic. Being set $K > 6$ while with all the other factors held constant, STRUCTURE produces various clustering results for HGDP-CEPH, in particular various clusters would be identified or excluded [43].

The mechanism for genetic clustering and BGA inference introduced by STRUCTURE provides a much coherent solution to this question. And somehow the problems arising suggest the direction for further developments of Bayesian-based BGA inference. One of the cases suggested by the authors [51] where a model may produce better performance involves the use of AIMs. The use of such prior geographical information helps to determine on the value of K [27] and relieves the difficulty of admixture proportion Q inference, since there would be quite clear information about the likely values of K and Q.

3.1.2. SNIPPER

Another widely used Bayesian-based BGA inference model is SNIPPER [49] that takes the AIMs as input, identifies genetic clusters using Principle Component Analysis (PCA), and infers individual's BGAs using the Bayes' rule based on allele frequencies. The original study [49] and a serious of latter studies of the same group [48, 61, 59, 60] endeavour to identify the autosomal AIMs as well as the ancestry sensitive indels with the measurement of *informativeness* based on Wright's Fst. A set of 34 AIMs from HGDP-CEPH was firstly identified to pronounce allele frequency discontinuities between Africans, Europeans, and East Asian populations [49]. Keeping analysing HGCP-CEPH and 1000 Genomes datasets, Phillips et al. (2013) complement the original 34-AIM-panel with 23 AIMs that differentiate Europeans and South Asians. Followed above studies, more ancestry sensitive indels are identified to distinguish American, Central South Asian, Middle Eastern Asian, and Oceanian populations [61, 60].

Similar to those of STRUCTURE, SNIPPER works on the assumptions of Hardy-Weinberg equilibrium and linkage disequilibrium (LD), which have been measured and ensured by the identified AIMs. To assign individuals into population classes, Snipper adopts simple Bayes' rule: the likelihood of an individual originates from a single population is given by maximizing the posteriori where allele frequency distributions are estimated from the training set HGDP-CEPH. An individual is assigned to a single population of origin that has the highest probability. However, the geographically close-site populations, such as Central South Asians and Middle Eastern Asians, are rarely able to be assessed for clear divergence using the simple Bayes classification rule, due the difficulty of obtaining a sufficient spread of autosomal informative SNPs [48, 61]. SNIPPER adopts a PCA clustering process to increase the separation between genetically closed populations; meanwhile, introduces the use of AIM-Indels to BGA inference. A set of 46 AIM-Indels [61] successfully differentiate Central South and Middle Eastern Asians. The set of 34 AIM-SNPs [48] and the set of 46 AIM-Indels [61], when combined with Y and mitochondrial DNA variation [60], have shown to be a powerful set fo markers for BGA inference amongst Asian, African, Eurasian (Middle East Asian, South/Central Asian, and European), naive American, and Pacific regions [47].

3.2. Principle Component-Based BGA Inference System EIGENSTRAT

The use of principle components (PCs) and derived methods such as principle-coordinate analysis and multidimensional analysis in genetic populations were firstly proposed by Cavalli-Sforza et al. (1994) [12]. Almost all sets of allele frequencies formed by populations contain some redundancy that can be measured by the correlation amongst genes and populations. PC-based approaches offer another simple mode of analysing population-by-allele-frequency data by representing allele frequencies as a weighted average of populations. By replacing the allele frequencies with their PC values, the resulting PC representation describes the linear compound of the original gene frequencies– of which about 40% to 50% genetic information retains with the first two PCs [12].

Patterson et al. (2006) [45] proposed a framework, EIGENSTRAT, exemplifying Principle Component Analysis (PCA)'s feasibility of detecting and qualifying the population structure presented in a biallelic dataset. EIGENSTRAT calculates on genes instead of populations. The samples are formed in a designed matrix where each row represents a sample and the columns are indexed by locus, where the number of samples is significantly smaller than the number of markers used. Performing PCA step by step, the framework outputs a set of eigenvectors and eigenvalues of which the set of eigenvectors corresponding to the largest set of eigenvalues are expected to expose the population structure.

The basic EIGENSTRAT is designed under the assumption that the markers are of linkage equilibrium. The model can be modified to accommodate markers with linkage disequilibrium by introducing a theoretical statistical parameter (the effective number of markers) that will approximate the covariance matrix to a Wishart-like matrix. The model has also been modified to allow microsatellites inputs. The ideal is similar with that of the linkage-equilibrium model. A mimic marker is defined out of each allele, presenting the number of occurrences of the allele for a sample. However, the results show that PCA does not suit this case [45].

EIGENSTRAT demonstrates the use of statistically significant axes for detecting the existence of population structure in a dataset. It is also of interest in ordination in human genetics, where it has been promoted as a sensitive and computationally efficient model-free alternative.

4. Proposed Deep Learning Approach

The current machine learning techniques including PCA and Bayesian that have applied to BGA inference cannot operate on the data sequence directly. These techniques require pre-defined features extracted from the data sequence based on prior knowledge (e.g. the presence or absence of single nucleotide polymorphisms and allele frequencies). The labour work is intensive and requires domain knowledge, and is limiting for high dimensional genetic data. Deep neural networks help to bypass this manual feature extraction. Moreover, deep neural networks capture nonlinear dependencies in the sequences and span wider sequence context at multiple genomic scales.

Convolutional neural networks (CNN) is used for regulatory genomics to leverage the variation among individuals to map the traits [2, 26] and to split the sequence into windows centered on the trait of interest [72]. The key advantage of CNN is the ability to directly train the model on larger sequences [3]. CNN allows direct training on the DNA sequences without the need of predefined features.

Alipanahi et al. [2] uses CNN to predict the sequence specificities. Their inference model innovates on training the model directly from raw DNA sequences by applying a one-dimensional convolutional layer. Zhou and Troyanskaya [72] considered CNN to jointly learn the diverse chromatin factors and to predict multiple chromatin states in parallel. From the similar insight, Kelley's CNN-based framework [26] successfully retrieves both known and novel sequence motifs of DNase I hypersensitivity.

5. Experiments

We conducted experiments to retrieve SNPs from a large volume of genotyped SNPs in which all SNPs are ascertained in a clearly documented way. We used the dataset for Human Genome Diversity Project (HGDP) that contains the various types of data: high dimensional, genome-wide SNPs, microsatelites, indels, copy number variants (CNVs), whole-genome shotgun sequences, exome sequences, haploid genotypes, high-coverage sequences, and the raw human sequences. This dataset provides a resource of 1063 lymphoblastoid cell lines (LCLs) from 1050 individuals in 52 world populations and corresponding milligram quantities of DNA which was banked at the Foundation Jean Dausset-CEPH in Paris. These LCLs were collected from various laboratories by the HGDP and CEPH in order to provide unlimited supplies of DNA and RNA for studies of sequence diversity and history of modern human populations [11].

We used a subset of 93 AIM-SNPs [43] which is selected from HGDP-CEPH Stanford dataset– a high dimension SNPs in autosomes and haploids typed acorss 1042 individuals from 52 populations across 7 population groups. The genotypes include those from 121 Africans, 160 Europeans, 207 Central-South Asians , 176 Middle-East Asians , 235 East Asians , 108 Naïve Americans, and 35 Oceanians. We first clustered all data samples into the following population groups: Africa (AFR), Europe (EUR), Asia (South Central Asia (CSA), Middle East Asia (MEA), and East Asia (EAS)), Naïve America (NAM), and Oceania (OCE) . Although the dataset provides the population tag, we clustered the samples to check whether the selected markers are sufficient to present the population stratification. Population structure was examined using STRUCTURE v2.3.4 [51].

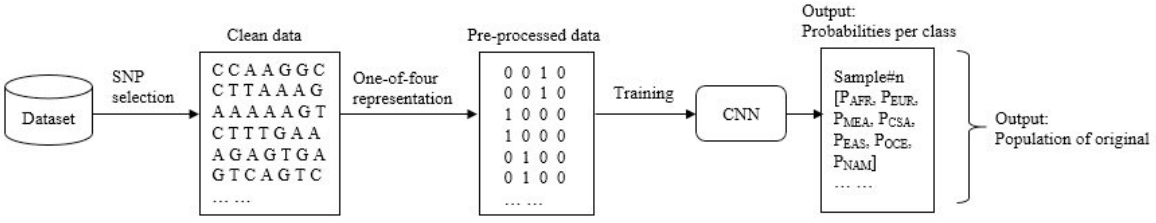

Fig. 1: Schematic overview of proposed CNN for BGA inference.

In the pre-processing stage, an array was established to present allele variation for each sample through a one-of-four representation (Figure 1). To learn DNA sequence signals, each sample is encoded as a matrix that will be then represented as an image to CNN. The four bits of each sample are considered analogously to color channels. Our proposed model allows partial genotypes. The missing locus will be encoded as -1 that represents unsigned values in the image.

We defined the structure with one convolutional layer since the size of each sample is small, only 744 (=186*4). Of 1042 samples, we randomly selected 70% samples (n=730) to train the network, 30% (n=312) is used to evaluate the performance. Considering the positive truth as the measure of prediction accuracy, the model achieves a peak performance of 86% on average (Table 1) with raw input and a 0.25 dropout threshold, after 3000 iterations. The classification accuracies of Europe (EUR), Central-South Asians (CSA), and Middle East Asia (MEA) are of 64.6%, 80.6%, and 71.7%, respectively (Table 1). These lower accuracies compared with others on the same proposed method may be from the genetic similarity [61] between the three population groups.

Table 1: Comparison [43] of classification success using 93 SNP panel

Classifier	**AFR**	**EUR**	**MEA**	**CSA**	**EAS**	**OCE**	**NAM**
STRUCTURE	99.1%	91.1%	-	7.7%	88.5%	96.8%	86.8%
PCA	87.3%	87.1%	-	18.3%	90.6%	93.5%	80.1%
CNN	100.0%	64.6%	71.7%	80.6%	97.2%	90.0%	96.9%

The proposed study provides the validation for the use of CNN in human BGA inference. We show that the proposed CNN, together with the predictive value of SNP therein, can distinguish diverse population groups with a small set of SNPs. Specifically, the proposed model successfully distinguished the three Asian sub-populations (Table 1), allowing a broader geographical scope of the 93 AIM-SNP panel.

6. Conclusion

We have presented the current genetic and machine learning approaches to BGA inference. With the larger number of available genetic data, the current machine learning techniques can make the conventional analysis more efficient and also provides an alternative of probabilistic estimation. Since this current machine learning approach requires pre-defined features extracted from the data sequence based on prior knowledge, we have proposed our deep learning approach based on Convolutional Neural Networks (CNN) that helps to bypass this manual feature extraction, captures nonlinear dependencies in the sequences and spans wider sequence context at multiple genomic scales.

References

[1] Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., Kruglyak, L., 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS biology 2, e286.

[2] Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J., 2015. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nature biotechnology 33, 831.

[3] Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O., 2016. Deep learning for computational biology. Molecular systems biology 12, 878.

[4] Bamshad, M., Wooding, S., Salisbury, B.A., Stephens, J.C., 2004. Deconstructing the relationship between genetics and race. Nature Reviews Genetics 5, 598.

[5] Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al., 2012. Fast and accurate inference of local ancestry in latino populations. Bioinformatics 28, 1359–1367.

[6] Barnholtz-Sloan, J.S., McEvoy, B., Shriver, M.D., Rebbeck, T.R., 2008. Ancestry estimation and correction for population stratification in molecular epidemiologic association studies. Cancer Epidemiology and Prevention Biomarkers 17, 471–477.

[7] Basu, A., Tang, H., Arnett, D., Gu, C.C., Mosley, T., Kardia, S., Luke, A., Tayo, B., Cooper, R., Zhu, X., et al., 2009. Admixture mapping of quantitative trait loci for bmi in african americans: evidence for loci on chromosomes 3q, 5q, and 15q. Obesity 17, 1226–1231.

[8] Bolnick, D.A., Bolnick, D.I., Smith, D.G., 2006. Asymmetric male and female genetic histories among native americans from eastern north america. Molecular biology and evolution 23, 2161–2174.

[9] Bolnick, D.A., Fullwiley, D., Duster, T., Cooper, R.S., Fujimura, J.H., Kahn, J., Kaufman, J.S., Marks, J., Morning, A., Nelson, A., et al., 2018. 48. the science and business of genetic ancestry testing. Beyond Bioethics: Toward a New Biopolitics , 422.
[10] Bonilla, C., Shriver, M.D., Parra, E.J., Jones, A., Fernández, J.R., 2004. Ancestral proportions and their association with skin pigmentation and bone mineral density in puerto rican women from new york city. Human genetics 115, 57–68.
[11] Cann, H.M., De Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al., 2002. A human genome diversity cell line panel. Science 296, 261–262.
[12] Cavalli-Sforza, L.L., Menozzi, P., Cavalli-Sforza, L., Piazza, A., Cavalli-Sforza, L., 1994. The history and geography of human genes. Princeton university press.
[13] Chakravarti, A., 2009. Being human: kinship: race relations. Nature 457, 380.
[14] Cheung, E.Y., Gahan, M.E., McNevin, D., 2017. Prediction of biogeographical ancestry from genotype: a comparison of classifiers. International journal of legal medicine 131, 901–912.
[15] Chib, S., 1995. Marginal likelihood from the gibbs output. Journal of the American Statistical Association .
[16] Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., Nielsen, R., 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome research 15, 1496–1502.
[17] Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglia, A., Tofanelli, S., Spedini, G., Capelli, C., 2004. Variation of female and male lineages in sub-saharan populations: the importance of sociocultural factors. Molecular Biology and Evolution 21, 1673–1682.
[18] Egeland, T., Bøvelstad, H., Storvik, G., Salas, A., 2004. Inferring the most likely geographical origin of mtdna sequence profiles. Annals of human genetics 68, 461–471.
[19] Enoch, M.A., Shen, P.H., Xu, K., Hodgkinson, C., Goldman, D., 2006. Using ancestry-informative markers to define populations and detect population stratification. Journal of Psychopharmacology 20, 19–26.
[20] Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587.
[21] Frudakis, T., Venkateswarlu, K., Thomas, M., Gaskin, Z., Ginjupalli, S., Gunturi, S., Ponnuswamy, V., Natarajan, S., Nachimuthu, P.K., 2003. A classifier for the snp-based inference of ancestry. Journal of Forensic Sciences 48, 771–782.
[22] Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y., et al., 2010. A draft sequence of the neandertal genome. science 328, 710–722.
[23] Halder, I., Shriver, M., Thomas, M., Fernandez, J.R., Frudakis, T., 2008. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Human mutation 29, 648–658.
[24] Hartl, D.L., Clark, A.G., Clark, A.G., 1997. Principles of population genetics. volume 116. Sinauer associates Sunderland.
[25] Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., Cox, D.R., 2005. Whole-genome patterns of common dna variation in three human populations. Science 307, 1072–1079.
[26] Kelley, D.R., Snoek, J., Rinn, J.L., 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research 26, 990–999.
[27] Kersbergen, P., van Duijn, K., Kloosterman, A.D., den Dunnen, J.T., Kayser, M., de Knijff, P., 2009. Developing a set of ancestry-sensitive dna markers reflecting continental origins of humans. Bmc Genetics 10, 69.
[28] Kidd, J.R., Friedlaender, F.R., Speed, W.C., Pakstis, A.J., De La Vega, F.M., Kidd, K.K., 2011. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investigative genetics 2, 1.
[29] Kidd, K.K., Speed, W.C., Pakstis, A.J., Furtado, M.R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F.R., Kidd, J.R., 2014. Progress toward an efficient panel of snps for ancestry inference. Forensic Science International: Genetics 10, 23–32.
[30] King, T.E., Parkin, E.J., Swinfield, G., Cruciani, F., Scozzari, R., Rosa, A., Lim, S.K., Xue, Y., Tyler-Smith, C., Jobling, M.A., 2007. Africans in yorkshire? the deepest-rooting clade of the y phylogeny within an english genealogy. European Journal of Human Genetics 15, 288.
[31] Kosoy, R., Nassir, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., Belmont, J.W., et al., 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in america. Human mutation 30, 69–78.
[32] Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
[33] Lao, O., van Duijn, K., Kersbergen, P., de Knijff, P., Kayser, M., 2006. Proportioning whole-genome single-nucleotide–polymorphism diversity for the identification of geographic population structure and genetic ancestry. The American Journal of Human Genetics 78, 680–690.
[34] Lee, C., Măndoiu, I.I., Nelson, C.E., 2011a. Inferring ethnicity from mitochondrial dna sequence, in: BMC proceedings, BioMed Central. p. S11.
[35] Lee, C., Măndoiu, I.I., Nelson, C.E., 2011b. Inferring ethnicity from mitochondrial dna sequence, in: BMC proceedings, BioMed Central. p. S11.
[36] Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493.
[37] Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. science 319, 1100–1104.
[38] Lipson, M., Loh, P.R., Levin, A., Reich, D., Patterson, N., Berger, B., 2013. Efficient moment-based inference of admixture parameters and sources of gene flow. Molecular biology and evolution 30, 1788–1802.
[39] Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., Berger, B., 2013. Inferring admixture histories of human populations using linkage disequilibrium. Genetics , genetics–112.
[40] Maples, B.K., Gravel, S., Kenny, E.E., Bustamante, C.D., 2013. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics 93, 278–288.
[41] McVean, G., 2009. A genealogical interpretation of principal components analysis. PLoS genetics 5, e1000686.

[42] Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., Reich, D., 2011. The history of african gene flow into southern europeans, levantines, and jews. PLoS genetics 7, e1001373.
[43] Nassir, R., Kosoy, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., Belmont, J.W., et al., 2009. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC genetics 10, 39.
[44] Nielsen, R., Hubisz, M.J., Torgerson, D., Andres, A.M., Albrechtsen, A., Gutenkunst, R., Adams, M., Cargill, M., Boyko, A., Indap, A., et al., 2009. Darwinian and demographic forces affecting human protein coding genes. Genome research , gr–088336.
[45] Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLoS genetics 2, e190.
[46] Pereira, R., Phillips, C., Pinto, N., Santos, C., dos Santos, S.E.B., Amorim, A., Carracedo, Á., Gusmão, L., 2012. Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. PloS one 7, e29684.
[47] Phillips, C., 2015. Forensic genetic analysis of bio-geographical ancestry. Forensic Science International: Genetics 18, 49–65.
[48] Phillips, C., Aradas, A.F., Kriegel, A., Fondevila, M., Bulbul, O., Santos, C., Rech, F.S., Carceles, M.P., Carracedo, Á., Schneider, P., et al., 2013. Eurasiaplex: a forensic snp assay for differentiating european and south asian ancestries. Forensic Science International: Genetics 7, 359–366.
[49] Phillips, C., Salas, A., Sanchez, J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M.C., Ballard, D., Lareu, M., et al., 2007. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker snps. Forensic Science International: Genetics 1, 273–280.
[50] Pickrell, J.K., Pritchard, J.K., 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics 8, e1002967.
[51] Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.
[52] Reich, D., Patterson, N., De Jager, P.L., McDonald, G.J., Waliszewska, A., Tandon, A., Lincoln, R.R., DeLoa, C., Fruhan, S.A., Cabre, P., et al., 2005. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. Nature genetics 37, 1113.
[53] Reich, D., Price, A.L., Patterson, N., 2008. Principal component analysis of genetic data. Nature genetics 40, 491.
[54] Rosenberg, N.A., Li, L.M., Ward, R., Pritchard, J.K., 2003. Informativeness of genetic markers for inference of ancestry. The American Journal of Human Genetics 73, 1402–1422.
[55] Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., Feldman, M.W., 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS genetics 1, e70.
[56] Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W., 2002. Genetic structure of human populations. science 298, 2381–2385.
[57] Royal, C.D., Novembre, J., Fullerton, S.M., Goldstein, D.B., Long, J.C., Bamshad, M.J., Clark, A.G., 2010. Inferring genetic ancestry: opportunities, challenges, and implications. The American Journal of Human Genetics 86, 661–673.
[58] Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T., Altshuler, D., Lander, E., 2006. Positive natural selection in the human lineage. science 312, 1614–1620.
[59] Santos, C., Fondevila, M., Ballard, D., Banemann, R., Bento, A.M., Børsting, C., Branicki, W., Brisighelli, F., Burrington, M., Capal, T., et al., 2015a. Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (aim) panels: results of a collaborative ednap exercise. Forensic Science International: Genetics 19, 56–67.
[60] Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R.A., Burchard, E.G., Schanfield, M.S., Souto, L., Uacyisrael, J., Via, M., et al., 2016. Pacifiplex: an ancestry-informative snp panel centred on australia and the pacific region. Forensic Science International: Genetics 20, 71–80.
[61] Santos, C., Phillips, C., Oldoni, F., Amigo, J., Fondevila, M., Pereira, R., Carracedo, Á., Lareu, M.V., 2015b. Completion of a worldwide reference panel of samples for an ancestry informative indel assay. Forensic Science International: Genetics 17, 75–80.
[62] Seldin, M.F., 2007. Admixture mapping as a tool in gene discovery. Current opinion in genetics & development 17, 177–181.
[63] Sforza, C.L., Edwards, A.W.F., 1964. Analysis of human evolution. Genet. Today 3, 923–933.
[64] Shriver, M.D., Kittles, R.A., 2004. Genetic ancestry and the search for personalized genetic histories. Nature Reviews Genetics 5, 611.
[65] Shriver, M.D., Parra, E.J., Dios, S., Bonilla, C., Norton, H., Jovel, C., Pfaff, C., Jones, C., Massac, A., Cameron, N., et al., 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. Human genetics 112, 387–399.
[66] Sokal, R., Rohlf, F., 1995. Biometry. Freedman New York .
[67] Stevens, E.L., Heckenberg, G., Roberson, E.D., Baugher, J.D., Downey, T.J., Pevsner, J., 2011. Inference of relationships in population data using identity-by-descent and identity-by-state. PLoS genetics 7, e1002287.
[68] Tang, H., Coram, M., Wang, P., Zhu, X., Risch, N., 2006. Reconstructing genetic ancestry blocks in admixed individuals. The American Journal of Human Genetics 79, 1–12.
[69] Turakulov, R., Easteal, S., 2003. Number of snps loci needed to detect population structure. Human heredity 55, 37–45.
[70] Underhill, P.A., Kivisild, T., 2007. Use of y chromosome and mitochondrial dna population structure in tracing human migrations. Annu. Rev. Genet. 41, 539–564.
[71] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al., 2001. The sequence of the human genome. science 291, 1304–1351.
[72] Zhou, J., Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods 12, 931.
[73] Zhu, X., Luke, A., Cooper, R.S., Quertermous, T., Hanis, C., Mosley, T., Gu, C.C., Tang, H., Rao, D.C., Risch, N., et al., 2005. Admixture mapping for hypertension loci with genome-scan markers. Nature genetics 37, 177.