

Use of Machine Learning Techniques for SNP Based Prediction of Ancestry

by

Dominic J. Allocco

B.A. Economics
Harvard University, 1995

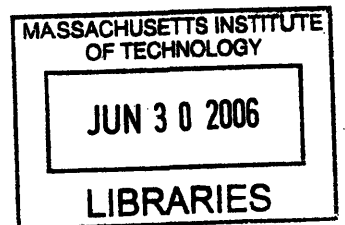
M.D. Medicine
Stanford University, 1999

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND
TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF

MASTER OF SCIENCE IN BIOMEDICAL INFORMATICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
AND
HARVARD MEDICAL SCHOOL

MAY 2006
[June 2006]

Copyright 2006 Dominic J. Allocco. All rights reserved.



ARCHIVES

This work was funded by the National Library of Medicine Fellowship in Medical Informatics. The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author: _____
Harvard-MIT Division of Health Sciences and Technology
May 6, 2006

Certified by: _____
Isaac Kohane, M.D., Ph.D.
Associate Professor of Pediatrics and Health Sciences and Technology
Thesis Supervisor

Accepted by: _____
Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Sciences and Technology

Use of Machine Learning Techniques for SNP Based Prediction of Ancestry

by

Dominic J. Allocco

Submitted to the Harvard-MIT Division of Health Sciences and Technology in May of 2006 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Biomedical Informatics

ABSTRACT

Some have argued that the genetic differences between continentally defined groups are relatively small and unlikely to have biomedical significance. In this study, the extent of variation between continentally defined groups was evaluated. Small numbers of randomly selected single nucleotide polymorphisms from the International HapMap Project were used to train classifiers for prediction of ancestral continent of origin. Predictive accuracy was then tested on independent data sets. A high degree of genetic similarity implies that groups will be difficult to distinguish, especially when only a limited amount of genetic information is used. It is shown that the genetic differences between continentally defined groups are sufficiently large that one can accurately predict ancestral continent of origin using only a minute, randomly selected fraction of the genetic variation present in the human genome. Genotype data from only 50 random single nucleotide polymorphisms can be used to predict ancestral continent of origin in the primary test data set with an average accuracy of 95%. Single nucleotide polymorphisms were also characterized as being in introns, coding exons, regulatory regions and regions coding for untranslated mRNA and classifiers constructed using only single nucleotide polymorphisms from a specific category. Predictive accuracy was similar across all of the classifiers created in this manner. Single nucleotide polymorphisms useful for prediction of ancestral continent of origin are common and distributed relatively evenly throughout the genome. These findings demonstrate the extent of variation between continentally defined groups and argue strongly against the contention that genetic differences between groups are too small to have biomedical significance.

Thesis Supervisor: Isaac S. Kohane

Title: Associate Professor of Pediatrics and Health Sciences and Technology

1. Introduction and Background

A major goal of both the Human Genome Project and the subsequent International HapMap Project was to provide a foundation for understanding genetic variation in the human genome (Collins et al. 1998; 2003). Multiple studies over the past several decades have consistently concluded that only 5-15% of human genetic variation can be explained by differences between populations (Lewontin 1972; Nei and Roychoudhury 1974; Rosenberg et al. 2002; Cavalli-Sforza and Feldman 2003). Although this proportion is relatively small, recent studies have shown that when individuals are grouped on the basis of genetic similarity, group membership corresponds closely to ancestral geographic origin (Rosenberg et al. 2002) or self-identified race or ethnicity (Tang et al. 2005).

There has been considerable discussion about the implications of these findings in the context of larger debates about population sub-structure, race and genetics (Calafell 2003; Haga and Venter 2003; Jorde and Wooding 2004; Serre and Paabo 2004; Tishkoff and Kidd 2004; Risch 2006). One major aspect of the debate concerns the extent of genetic variation between groups. This debate is becoming less theoretical as targeted therapies like BiDil (a branded combination of hydralazine and isosorbide dinitrate approved by the FDA for the treatment of congestive heart failure in African-Americans) are introduced. Some have argued that the genetic differences between continentally defined groups are relatively small and thus unlikely to have biomedical significance (Lewontin 1972; Cooper et al. 2003). Providing support for this view, others have noted that accurate classification of ancestral origin, while possible, requires large numbers of genetic markers (Calafell 2003; Jorde and Wooding 2004; Serre and Paabo 2004; Tishkoff and Kidd 2004). The opposing viewpoint holds that there are significant genetic

differences between groups and that these genetic differences may account for differences in risk of disease between populations (Burchard et al. 2003). Proponents of this viewpoint also note that an understanding of genetic variation between groups is important to avoid confounding in genotype-phenotype association studies (Risch et al. 2002; Tang et al. 2005).

It is shown here that the genetic variations that distinguish continentally defined groups are common and pervasive throughout the genome. Genetically similar groups should be difficult to distinguish using only a small number of genetic markers. In this study, I demonstrate that one can accurately predict ancestral continent of origin using only a minute, randomly selected fraction of the genetic variation present in the human genome. Genotype data from 50 random single nucleotide polymorphisms (SNPs) is sufficient to predict ancestral continent of origin in our primary test data set with an average accuracy of 95%. These findings demonstrate the extent of variation between continentally defined groups and argue strongly against the contention that differences between groups are too small to have biomedical significance.

2. Materials and Methods

The goal of this study was to investigate genetic variation between continentally defined populations. The study has several parts. First, I used all available genetic information to characterize population sub-structure. Second, I investigated how much genetic information was necessary to predict ancestral continent of origin. Third, I evaluated whether certain categories of SNPs were more useful in predicting ancestry. Finally, I evaluated the distribution throughout the genome of SNPs useful for predicting ancestral continent of origin.

Four different data sets were used for the analysis. The International HapMap Project is a large collaborative effort that has made publicly available genotype data for 270 individuals from four different populations: Yoruba in Ibadan, Nigeria (YRI); Japanese in Tokyo, Japan (JPT); Han Chinese in Beijing (CHB), China; and Utah residents with ancestry from northern and western Europe (CEU) (2003). The non-redundant data sets from public release 19 which contained data for phases I and II were used. Thirty trios were genotyped for both the Yoruba and European populations. The children were excluded from our analysis because they did not represent independent genotypes. One individual from the Japanese population who did not have phase I data was also excluded. Thus there were 60 Yoruba, 60 European-Americans, 45 Han Chinese and 44 Japanese from the HapMap included in this study. For the purpose of our analysis, the Han Chinese and Japanese subjects were grouped together. At the time of our study, genotype data that had passed quality control filters was available for almost 4 million SNPs.

The second data set used consisted of 1,586,383 SNPs genotyped by investigators at Perlegen Sciences to study DNA variation in human populations (Hinds et al. 2005). Nine individuals of European ancestry in this data set were excluded because they were also genotyped in the HapMap Project. Thus, this data set included genotype data for 23 African-Americans, 15 European-Americans and 24 Han Chinese. The third data set consisted of 4,124 SNPs genotyped as part of the Innate Immunity Program for Genomic Applications (IIPGA) . This data set included data for 24 African-Americans and 23 European-Americans. The fourth data set, referred to hereafter as the CVRI dataset, was derived from DNA collected by Qing Song and Gary Gibbons at the Morehouse School of Medicine Cardiovascular Research Institute as part of a study on genetic variation. This data set contained information on 17 SNPs in 96 individuals. There was genotype data for 24 West Africans, 24 European-Americans, 24 Han Chinese and 24 African-Americans.

SNPs in the HapMap and Perlegen data sets were selected so as to be generally representative of variation throughout the genome, while SNPs in the IIPGA and CVRI data sets were selected by the original investigators for genotyping on the basis of potential involvement in the innate immune response and in cardiovascular disease, respectively. The HapMap, Perlegen and IIPGA data sets are publicly available online. The CVRI data set is described and characterized in the Appendix to this manuscript.

As an initial step in the evaluation of genetic variation between groups, I calculated the genetic distance between each individual and the three continentally defined HapMap groups. The genetic distance between two individuals at a single loci was defined to be zero if the two individuals had the same genotype, $\frac{1}{2}$ if they had one

allele in common and 1 if they had neither allele in common (i.e. $d(CG,CC)=.5$ and $d(CC,GG)=1$) (Mountain and Cavalli-Sforza 1997). The genetic distance between two individuals was calculated as the mean genetic distance over all loci genotyped in both individuals. The genetic distance between an individual and a group was defined as the mean of the pairwise genetic distances between the individual and all members of the group – except if the individual was a member of the group, the distance between the individual and him or herself was not included.

Calculating the distance between each individual and the three continentally defined HapMap groups was done to investigate population structure using a large number of genetic markers. To evaluate the extent of genetic variation between continentally defined groups, I used small, randomly selected SNP subsets to train a machine learning classifier for prediction of ancestral continent of origin. The underlying premise here was that genetically similar groups should be difficult to distinguish using only a small number of randomly chosen genetic markers.

As a classifier, I implemented the naïve Bayes algorithm (Maron 1961) in PERL. A uniform prior probability distribution over the class variable was assumed. For each test data set, I randomly selected a specified number of SNPs. The HapMap data for these SNPs was then used to train a classifier. The performance of this classifier was evaluated by determining accuracy of prediction on the other three data sets. This process was repeated 100 times for the specified number of SNPs.

I also evaluated how ability to classify ancestral origin varied as a function of SNP type. The dbSNP database (Smigielski et al. 2000) was used to identify SNPs in introns, coding exons, regulatory regions and regions coding for untranslated mRNA.

Naïve Bayes classifiers were then constructed similarly to above, but only SNPs from a single category were used.

To evaluate the distribution of SNPs where allele frequency differed significantly between groups, I computed the pairwise differences in major allele frequency among the three HapMap groups. For the purposes of our study, the most frequent allele in the Yoruba was considered to be the major allele. Each chromosome was divided into bins of a constant size. For each pairwise combination of groups, I determined the proportion of SNPs where the difference in major allele frequency exceeded a threshold value. Multiple combinations of bin size and cut-off threshold were tested.

3. Unsupervised clustering of individuals

As discussed in the Introduction, previous work by Rosenberg et al and Tang et al has shown that when a large number of microsatellite markers are used to group individuals according to genetic similarity, cluster membership corresponds very closely to ancestral continent of origin (Rosenberg et al. 2002; Tang et al. 2005). As an initial step in our investigation of genetic variation between continentally defined populations, I first attempted to replicate using SNPs the prior results seen when individuals were clustered according to genetic similarity using microsatellite markers.

In their work, both Rosenberg et al and Tang et al used the Structure program (Pritchard et al. 2000) to perform k-means clustering. K-means clustering is a powerful unsupervised clustering technique that has been used successfully in many different applications. One significant limitation of the technique is difficulty in choosing the right value of k. Choosing a value for k reflects a determination of what the “right” number of clusters is, and it is not always obvious what the best way to do this is.

In this analysis, I utilized a different methodology. I defined a pairwise measure of genetic distance between individuals. Then, for each individual, I determined the mean genetic distance to each of the three continental HapMap populations. I then plotted each individual in a three dimensional space where each of the axes represented distance to one of the three continental HapMap populations. Based both on theoretical considerations and on the prior empirical work of Rosenberg et al and Tang et al, I expected that individuals would cluster according to continent of origin. The results of these analyses for the Perlegen and IIPGA data sets are shown in Figure 1 and Figure 2.

Figure 1

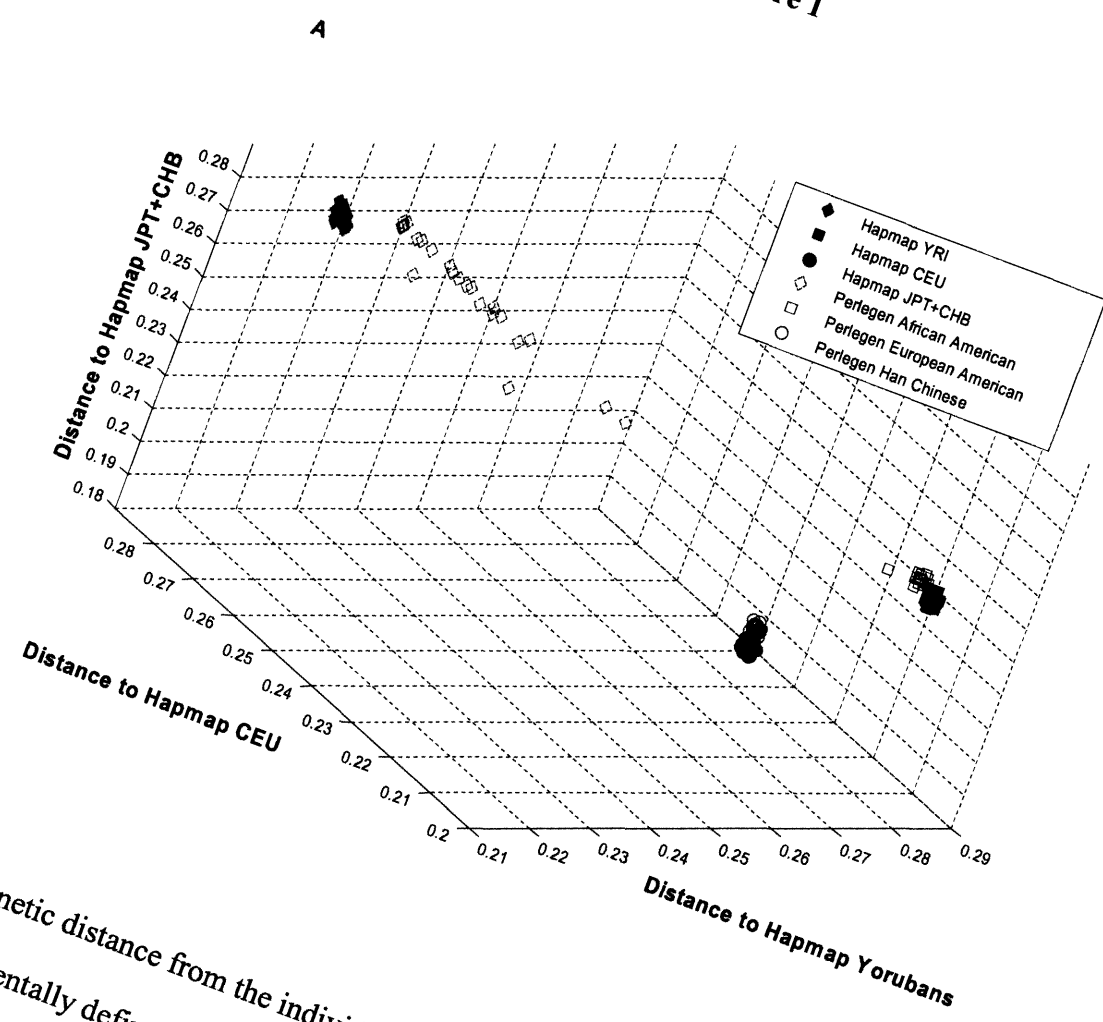


Figure 1. Genetic distance from the individuals in the HapMap and Perlegen data sets to the three continentally defined HapMap groups. 1,047,543 SNPs were used in the calculation of genetic distance. Individuals of similar geographic origin cluster together. African-Americans from the Perlegen data set lie in between the HapMap Yoruba and European-Americans, but are much closer to the Yoruba.

Figure 2

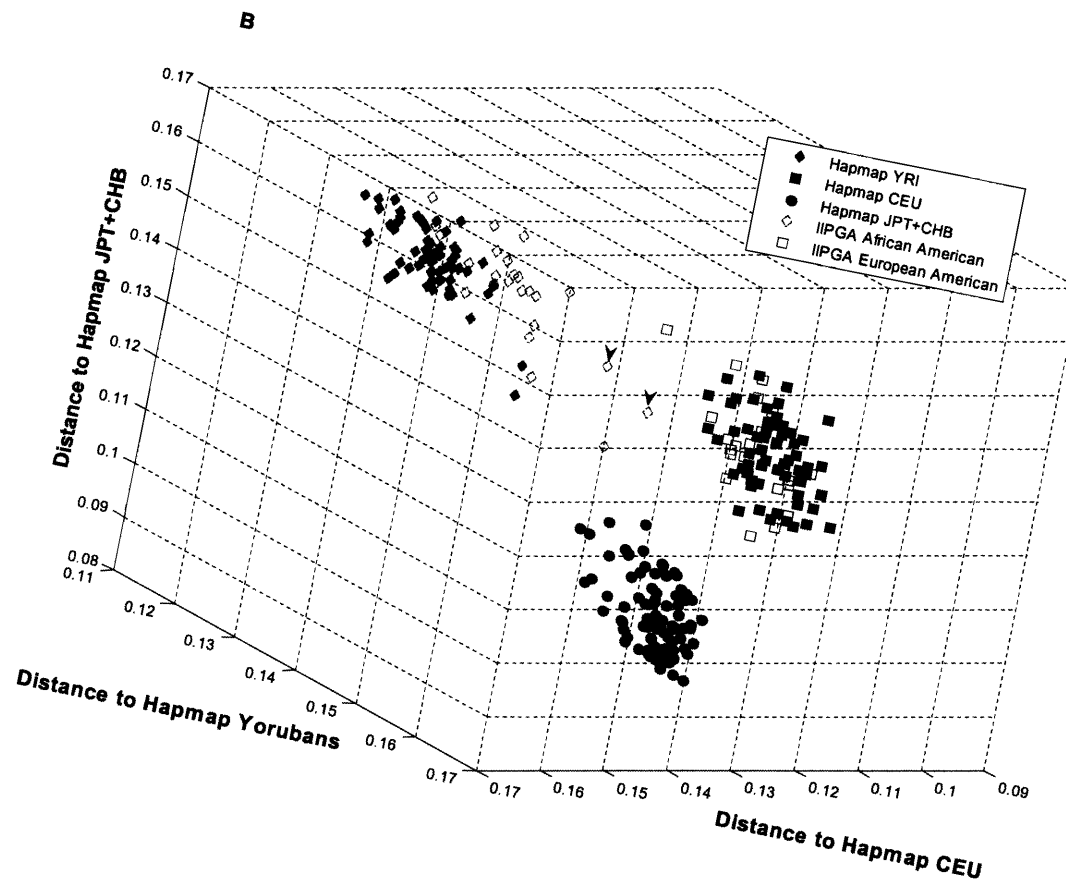


Figure 2. Genetic distance from the individuals in the HapMap and IIPGA data sets to the three continentally defined HapMap groups. 1,588 SNPs were used in the calculation of genetic distance. Individuals of similar geographic origin cluster together but less tightly than in Figure 1, reflecting the smaller number of SNPs used in the calculation of genetic distance. Individuals intermediate between clusters, like those labeled with arrowheads, may have a high degree of admixture.

Figures 1 and 2 clearly demonstrate that individuals cluster according to ancestral continent of origin, thus replicating with SNPs the previous results seen when microsatellite markers were used to do unsupervised k-means clustering. Taken together, these results indicate that there is underlying population sub-structure that is readily apparent when a large amount of genetic information is used to evaluate population sub-structure.

Examination of Figure 1 also shows that African-Americans as a group largely lie between the HapMap Yoruba and European-Americans, though much closer to the Yoruba. This suggests that African-Americans are a group of African origin that has had some admixture with people of European origin. A lesser tendency towards the Asian cluster also suggests the likelihood of admixture with people of Asian origin.

Finally, also note in Figure 2 that some individuals lie intermediate between clusters. These individuals likely have a high degree of admixture. In the next section, I describe attempts to predict ancestral continent of origin using only a small number of SNPs. I anticipated that it would be difficult to predict the continental origin of these individuals who were intermediate between clusters.

4. Predicting ancestry using a small number of randomly chosen SNPs

The analysis in the preceding section demonstrates that underlying population sub-structure can be discerned when a large amount of genetic information is used in the analysis. However, it has been argued that the differences between continentally defined populations are relatively subtle and that detecting population sub-structure requires powerful statistical techniques that integrate information across a large to number of genetic markers (Calafell 2003; Cooper et al. 2003; Jorde and Wooding 2004; Serre and Paabo 2004; Tishkoff and Kidd 2004). The implication is that these subtle genetic differences are unlikely to explain differences between continentally defined groups with regards to risk of disease or response to therapy. In this section, I empirically test the assertion that a large amount of genetic information is necessary to distinguish continentally defined groups.

For the analysis, I used genotype data from the International HapMap Project to train a naïve Bayes classifier. Naïve Bayes is a simple predictive algorithm that has been shown perform well in a wide variety of situations (Domingos and Pazzani 1997). The Perlegen, IIPGA and CVRI data sets were used as test sets. For each test data set, I chose N SNPs from the set of SNPs in common between the HapMap and the test data set and then trained a naïve Bayes classifier using the HapMap data. Predictive accuracy was then evaluated on the test set. This process was repeated 100 times for each choice of N and the mean predictive accuracy was determined. Figure 3 shows mean predictive accuracy as a function of N .

Figure 3

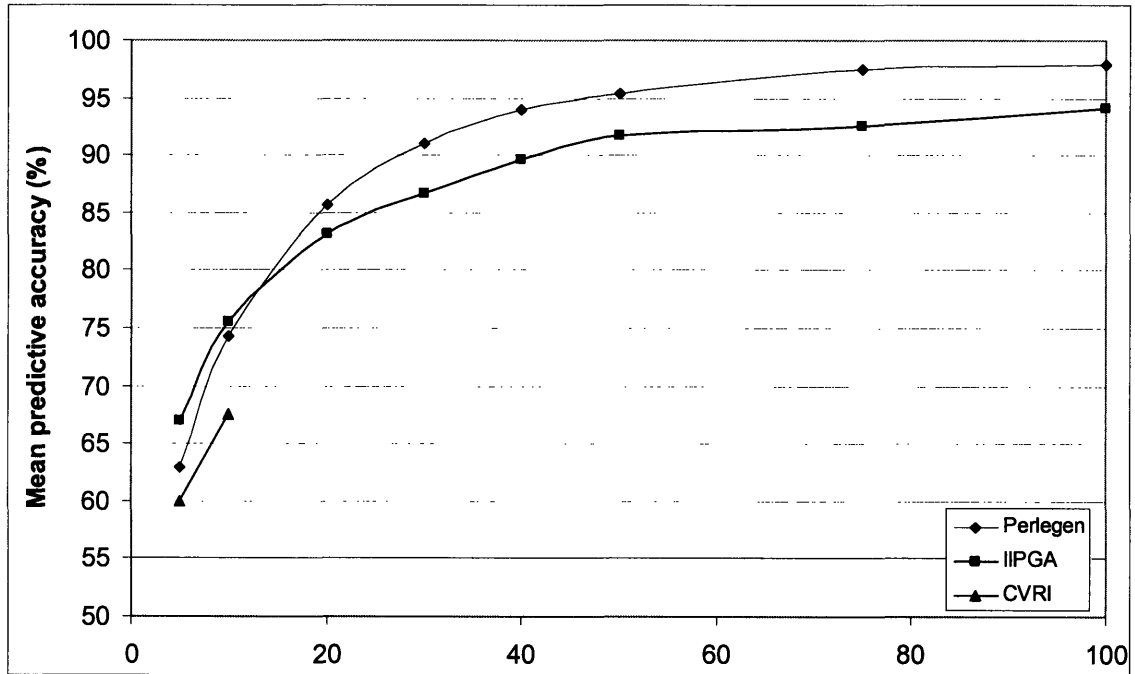


Figure 3. Mean predictive accuracy as a function of number of randomly selected SNPs used in the predictive model. Each data point represents the mean of 100 trials. Mean predictive accuracy increases rapidly as the number of SNPs used in training the classifiers grows. Predictive accuracy is quite good even when only a small number of SNPs are used.

This analysis shows that it is clearly possible to predict ancestral continent of origin using only a small number of randomly chosen SNPs. For example, when only 50 randomly selected SNPs are used, predictive accuracy is 95% in the Perlegen data set and 92% in the IIPGA data sets. Even with only 5 randomly selected SNPs, the observed predictive accuracies (ranging from 60 to 67% across the three test sets) are much higher than the 33% accuracy expected by chance.

This demonstrates it is quite possible to distinguish continentally defined groups using only a relatively small amount of genetic information. Moreover, specific markers are not necessary; randomly selected SNPs can be used to distinguish continentally defined groups. Since genetically similar groups should be difficult to distinguish when using only a limited amount of genetic information, this analysis suggests that there are significant genetic differences between continentally defined groups.

5. Predictive accuracy as a function of SNP type

The Perlegen data set is large enough to investigate genetic variation as a function of SNP type. The dbSNP database (Smigielski et al. 2000) was used to identify SNPs in introns, coding exons, regulatory regions and regions coding for untranslated mRNA. Naïve Bayes classifiers were then constructed using only SNPs from a specific category and the predictive accuracy of classifiers created using all SNPs was compared to that of classifiers created using only SNPs from a specific category.

The motivation for this analysis was to test the hypothesis that certain categories of SNPs might be more or less useful for prediction of ancestral continent of origin. There are several potential ways in which the type of SNP used for classification might be related to predictive accuracy. One might expect that in categories of SNPs where there was a higher degree of conservation, like in coding nonsynonymous SNPs or SNPs in regulatory regions, it would be more difficult to predict ancestral continent of origin than when using less conserved SNPs for prediction. However, if natural selection was a significant cause of variation between continentally defined groups, one might expect that it would actually be easier to predict ancestral continent of origin using coding nonsynonymous SNPs.

Figure 4 shows mean predictive accuracy as a function of N (the number of SNPs used to build the classifiers) with the results stratified by category of SNP. Classifiers built using all SNPs are included for comparison.

Figure 4

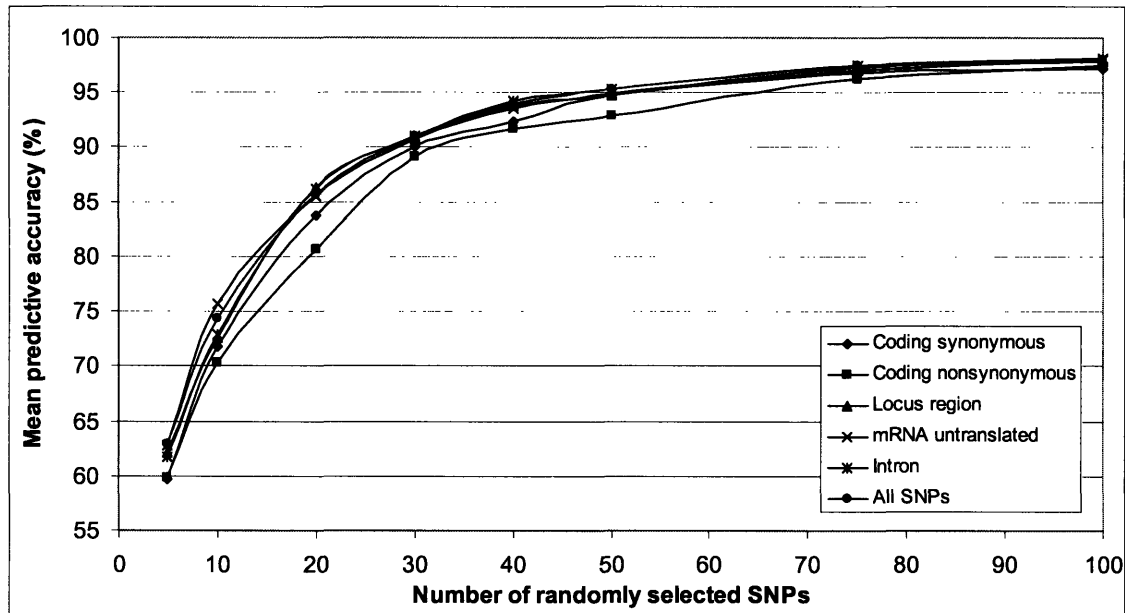


Figure 4. Mean predictive accuracy as a function of number of randomly selected SNPs used in the predictive model where results are stratified by category of SNP. Each data point represents the mean of 100 trials. Predictive accuracy is similar across all of the SNP categories examined.

It is apparent from Figure 4 that predictive accuracy is similar across all types of SNPs. This argues against natural selection being a major cause of genetic variation between continentally defined groups, as predicting ancestral continent of origin should be easier using coding nonsynonymous SNPs if that were the case. The fact that predictive accuracy is similar across all types of SNPs is at least consistent with the single-origin, or out of Africa, hypothesis. This hypothesis holds that modern day humans outside of Africa are descended from a small group that left Africa relatively recently (approximately 100,000 years ago). The observed genetic variation could result from genetic drift in relatively small founder populations.

6. Distribution of informative SNPs across the genome

I also tested whether some regions of the genome were more informative than others in predicting ancestral continent of origin. SNPs where allele frequency differs significantly between groups are most useful in predicting group membership. To investigate the distribution of these SNPs, for each SNP in the HapMap data, I calculated the major allele frequency in the three continentally defined HapMap groups. I considered the most frequent allele in the HapMap Yoruba to be the major allele. I then calculated the pairwise difference in major allele frequency between each combination of continentally defined groups. Informative SNPs were defined as those SNPs where the difference in major allele frequency was greater than a specified cut-off value. Next I divided the genome into bins and then determined the proportion of SNPs classified as informative in each bin.

The hypothesis tested was that informative SNPs would be distributed relatively evenly throughout the genome. Since the optimal cut-off threshold and bin size were not obvious, I tested multiple combinations of these parameters. Figure 5 shows the proportion of SNPs where the difference in major allele frequency exceeded 0.3. The results were binned by chromosome. Figures 6, 7 and 8 focus in on a more detailed scale. In these figures, only chromosome 1 is included. The bin size is 1,000,000 bp. The cut-off threshold is again 0.3. Figures 6, 7 and 8 show the results for the African-European, African-Asian and Asian-European comparisons respectively.

Figure 5

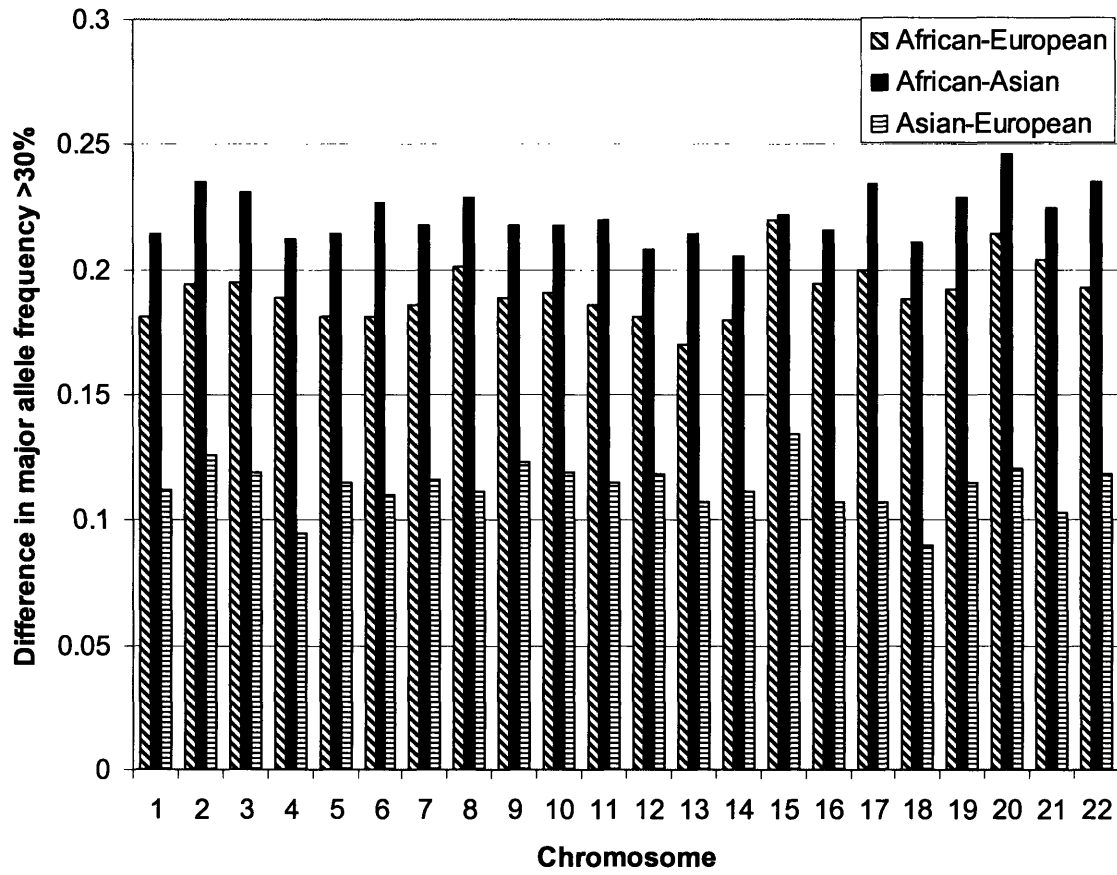


Figure 5. Distribution of informative SNPs by chromosome. In this figure informative SNPs are defined as those SNPs where the difference in major allele frequency is greater than 0.3. The proportion of informative SNPs is relatively constant across chromosome. The proportion of informative SNPs is higher for the African-European and African-Asian comparisons than for the Asian-European comparison.

Figure 6

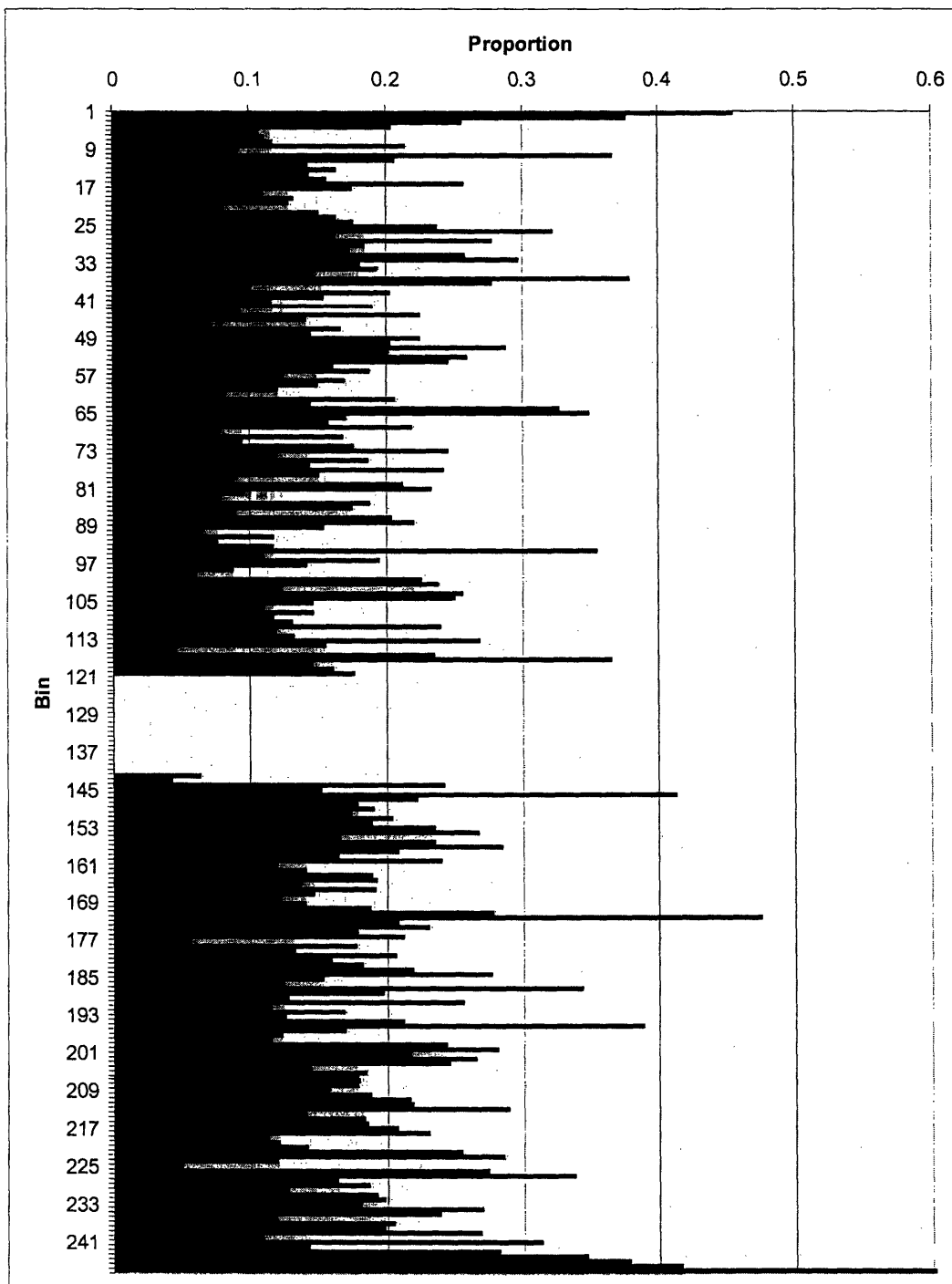


Figure 6. Proportion of SNPs on chromosome 1 where the difference in major allele frequency between the HapMap Yoruba and European-Americans is greater than 0.3. Bins represent one million base pairs.

Figure 7

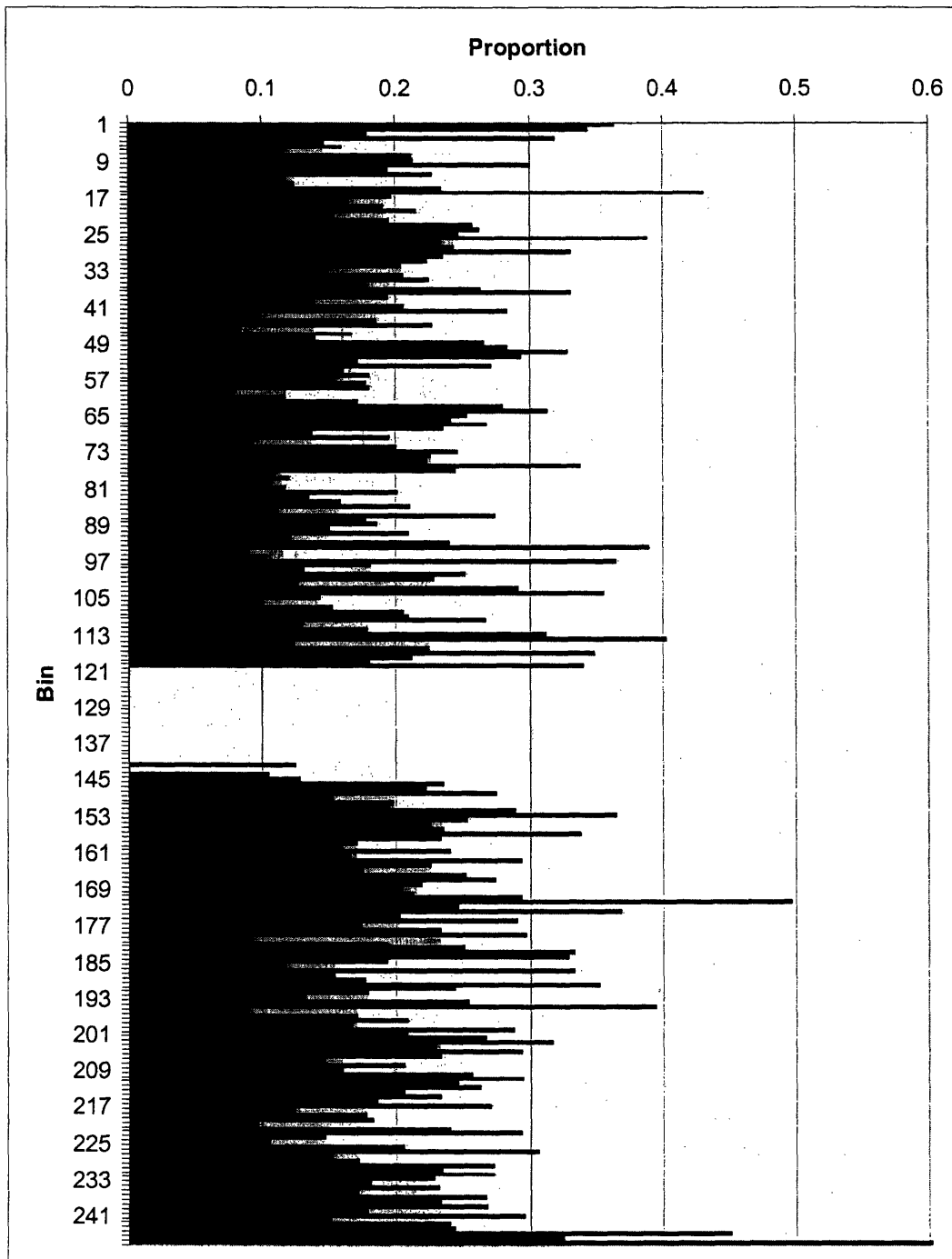


Figure 7. Proportion of SNPs on chromosome 1 where the difference in major allele frequency between the HapMap Yoruba and Asians is greater than 0.3. Bins represent one million base pairs.

Figure 8

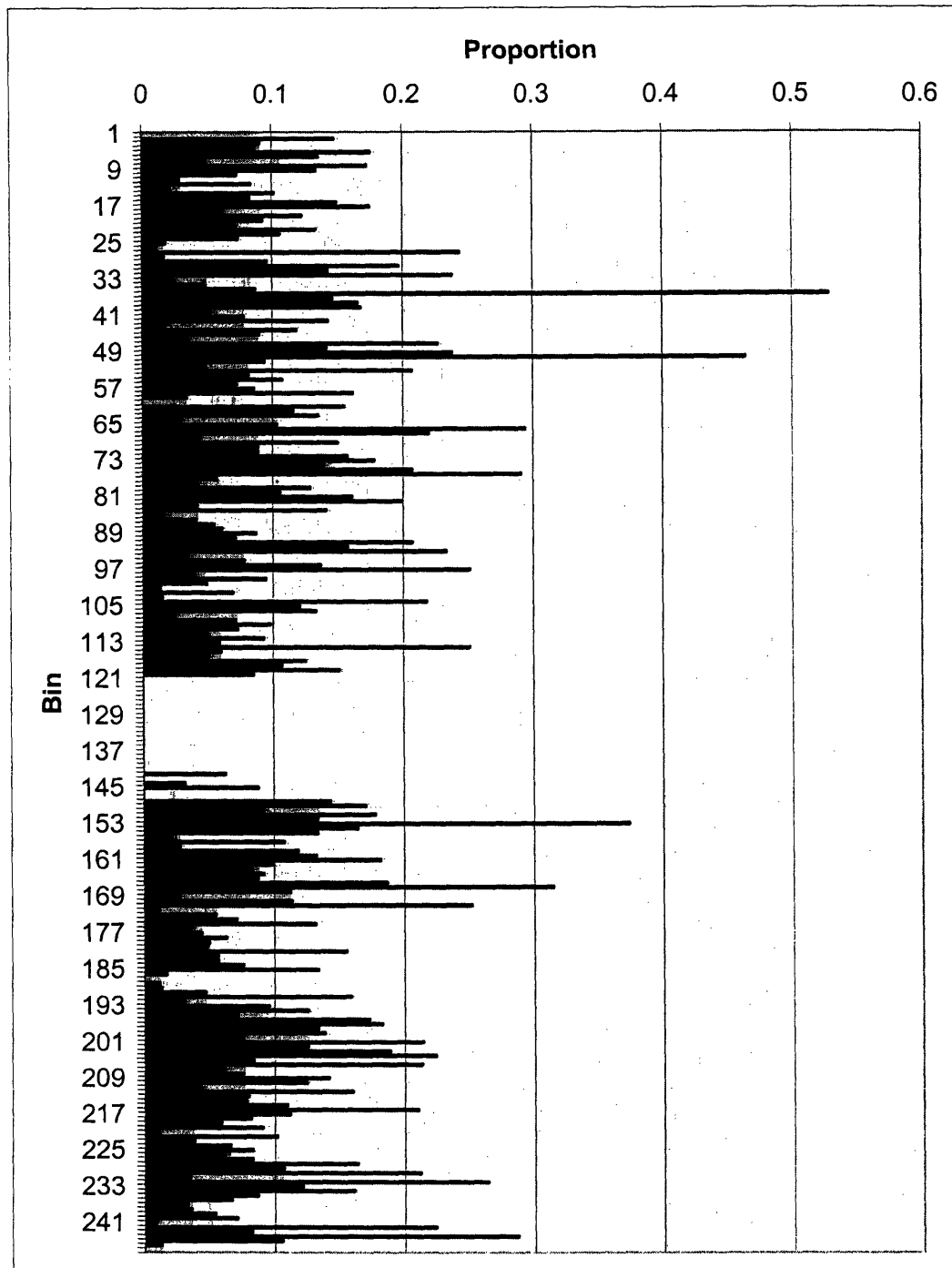


Figure 8. Proportion of SNPs on chromosome 1 where the difference in major allele frequency between the HapMap Asians and European-Americans is greater than 0.3.

Bins represent one million base pairs.

The preceding figures demonstrate that SNPs that are informative about ancestral origin are distributed relatively evenly across the genome. Informative SNPs are also quite common. When a SNP is defined as informative if the difference in major allele frequencies between the two groups under consideration is greater than 0.3, about 20% of the four million SNPs in the HapMap are informative about African ancestry compared to European or Asian ancestry. The percentage of SNPs that are informative about European versus Asian ancestry is somewhat smaller at approximately 12%, but these SNPs are still quite common.

I tested other combinations of cut-off threshold and bin size in addition to the ones described above. The general conclusion that informative SNPs were common and widely distributed throughout the genome did not change when these parameters were varied.

7. Discussion and Conclusions

In this study, I use a naïve Bayes algorithm trained using data from the HapMap to predict the ancestral geographic origin of individuals from three independent data sets. Even when a relatively small number of randomly selected SNPs are used, classification is accurate and robust. In the large Perlegen data set, predictive accuracy increases to 100% as the number of SNPs grows. This is not the case for the IIPGA test data set. Predictive accuracy as a function of the number of SNPs used, while still very good in comparison to many tests used in biology and medicine, levels off at approximately 96%. Two individuals in the IIPGA data set were classified incorrectly even if all 1588 available SNPs were used to train the model.

There are several potential explanations for these misclassifications. The two misclassifications were the individuals noted in Figure 2 to be intermediate between clusters. Thus, a likely explanation for the misclassifications is that it is inherently difficult for a classifier to assign an individual to a single continent of origin when that individual has ancestors from more than one continent.

Another potential explanation for errors in classification is that the naïve Bayes algorithm functions best when features are conditionally independent. Conditional independence is a very reasonable assumption for SNPs that are far apart or on different chromosomes. It is a more problematic assumption for SNPs that are close together. The IIPGA and CVRI data sets are composed of SNPs selected from candidate genes of interest to the original investigators. The SNPs in these two data sets are thus inherently clustered together and not evenly spread throughout the genome. The assumption of

conditional independence is therefore more likely to be violated in these data sets than in the Perlegen data. This may explain some of the misclassification errors.

A third possible explanation for the misclassification errors in our study is an imperfect gold standard. The gold standard used to define ancestral geographic origin was the original investigators' annotation. In the IIPGA and CVRI data sets, this annotation was based on self-identified race and not strictly on geographic origin. It is not expected that self-identified race should always correspond precisely to ancestral geographic origin.

Despite the limitations discussed above, this analysis demonstrates that even a minute fraction of the genetic variation in the human genome can be used to characterize ancestral geographic origin. Accurate characterization of ancestry will allow investigators conducting genetic association studies to identify stratification in study subjects and avoid false positive genotype-phenotype associations. The methodology described here is designed to predict ancestral continent of origin. It could easily be extended to make predictions about smaller units of geographic origin or individuals with a mixed background. This would require genotype data and well-characterized information about ancestral geographic origin from such individuals. Data of this kind is currently not available, but this is expected to change in the future as genotyping costs decrease. Thus it is anticipated that identification of more complex patterns of ancestry will be increasingly feasible as the amount of available data grows. This in turn will allow the development of higher resolution geographic maps and provide investigators designing genetic association studies with more powerful tools for detecting stratification.

This study demonstrates that the differences between continentally defined groups are sufficiently large that even a randomly selected, minute fraction of the genetic variation in the human genome can be used to characterize ancestral geographic origin in an accurate and reproducible manner. This argues strongly against the contention that differences between groups are too small to have biomedical significance. Understanding if and how these differences relate to risk of disease and response to therapy is one of the major challenges facing the biomedical research community.

8. Acknowledgements

This research was supported by a National Library of Medicine Training Grant for Informatics Research. I would also like to thank my research mentors Marco Ramoni and Isaac Kohane at the Children's Hospital Informatics Program and Qing Song and Gary Gibbons at the Cardiovascular Research Institute of the Morehouse School of Medicine.

9. References

- Innate Immunity PGA, NHLBI Program for Genomic Applications. 2003. The International HapMap Project. *Nature* **426**: 789-796.
- Burchard, E.G., E. Ziv, N. Coyle, S.L. Gomez, H. Tang, A.J. Karter, J.L. Mountain, E.J. Perez-Stable, D. Sheppard, and N. Risch. 2003. The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med* **348**: 1170-1175.
- Calafell, F. 2003. Classifying humans. *Nat Genet* **33**: 435-436.
- Cavalli-Sforza, L.L. and M.W. Feldman. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat Genet* **33 Suppl**: 266-275.
- Collins, F.S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters. 1998. New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**: 682-689.
- Cooper, R.S., J.S. Kaufman, and R. Ward. 2003. Race and genomics. *N Engl J Med* **348**: 1166-1170.
- Domingos, P. and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* **29**: 103-130.
- Haga, S.B. and J.C. Venter. 2003. Genetics. FDA races in wrong direction. *Science* **301**: 466.
- Hinds, D.A., L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer, and D.R. Cox. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307**: 1072-1079.
- Jorde, L.B. and S.P. Wooding. 2004. Genetic variation, classification and 'race'. *Nat Genet* **36**: S28-33.
- Lewontin, R.C. 1972. The apportionment of human diversity. In *Evolutionary Biology* (eds. T.H. Dobzhansky M.K. Hecht, and W.C. Steere), pp. 381-398. Appleton-Century-Crofts.
- Maron, M. 1961. Automatic Indexing: An Experimental Inquiry. *JACM* **8**: 404-417.
- Mountain, J.L. and L.L. Cavalli-Sforza. 1997. Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am J Hum Genet* **61**: 705-718.
- Nei, M. and A.K. Roychoudhury. 1974. Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *Am J Hum Genet* **26**: 421-443.
- Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-959.
- Risch, N. 2006. Dissecting racial and ethnic differences. *N Engl J Med* **354**: 408-411.
- Risch, N., E. Burchard, E. Ziv, and H. Tang. 2002. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol* **3**: comment2007.
- Rosenberg, N.A., J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. 2002. Genetic structure of human populations. *Science* **298**: 2381-2385.
- Serre, D. and S. Paabo. 2004. Evidence for gradients of human genetic diversity within and among continents. *Genome Res* **14**: 1679-1685.
- Smigielski, E.M., K. Sirotkin, M. Ward, and S.T. Sherry. 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* **28**: 352-355.

- Tang, H., T. Quertermous, B. Rodriguez, S.L. Kardia, X. Zhu, A. Brown, J.S. Pankow, M.A. Province, S.C. Hunt, E. Boerwinkle, N.J. Schork, and N.J. Risch. 2005. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet* **76**: 268-275.
- Tishkoff, S.A. and K.K. Kidd. 2004. Implications of biogeography of human populations for 'race' and medicine. *Nat Genet* **36**: S21-27.

Appendix: Characterization of the CVRI Data Set

RS ID #	Alleles	Major Allele	Overall	AA	Eur	HC	WA
rs7649970	C/T	C	0.818	0.771	0.813	0.938	0.750
rs7647481	A/G	G	0.807	0.771	0.792	0.938	0.729
rs7616597	A/G	A	0.865	0.708	1.000	1.000	0.750
rs4684848	A/G	G	0.620	0.354	0.813	0.938	0.375
rs4279078	C/T	C	0.859	0.958	0.938	0.542	1.000
rs3856806	C/T	C	0.813	0.896	0.771	0.667	0.917
rs2197423	A/G	G	0.818	0.771	0.813	0.938	0.750
rs1801282	C/G	G	0.917	0.958	0.813	0.938	0.958
rs17817276	A/G	A	0.813	0.896	0.604	0.813	0.938
rs17036328	C/T	T	0.818	0.771	0.813	0.938	0.750
rs17036314	C/G	G	0.672	0.688	0.708	0.583	0.708
rs12633551	A/G	G	0.911	1.000	1.000	0.646	1.000
rs12490265	A/G	G	0.776	0.896	0.604	0.813	0.792
rs12486170	C/T	T	0.802	0.896	0.604	0.813	0.896
rs11128602	C/T	C	0.630	0.604	0.708	0.542	0.667
rs11128599	A/G	G	0.625	0.708	0.708	0.563	0.521
rs10865710	C/G	C	0.699	0.667	0.729	0.688	0.750

Table Major allele frequency in the CVRI data set. AA, Eur, HC and WA stand for African-American, European-American, Han Chinese and West-African respectively.