**Journal of Statistics: Advances in Theory and Applications**
Volume 3, Number 2, 2010, Pages 125-134

# EVALUATION OF STATISTICAL TESTS FOR ETHNO-SNP SELECTION

**GAOLIN ZHENG[1], CHUNG-HAO CHEN[1] and TOM MILLEDGE[2]**

[1]Department of Mathematics and Computer Science
North Carolina Central University
Durham, NC 27707
U. S. A.
e-mail: chchen@nccu.edu

[2]Scalable Computing Support Center
Duke University
Durham, NC 27708
U. S. A.

## Abstract

**Motivation:** SNPs have shown a lot of promises in disease association, personalized medicine, and population classification studies. The completion of International Hapmap Project has facilitated the SNP-based ethno-classification. Due to the large amount of SNPs in the human genome, it is desirable to find a small set of informative SNPs for the classification task. Previous studies tried to find ethnically related SNPs from all the chromosomes and mitochondria and genotype data are usually treated as numeric data. Here, we focus on two small ethnically related genomic pieces in order to reduce noise. We apply a categorical statistical testing method to find marker SNPs. We evaluate its performance with two non-categorical statistical methods.

2010 Mathematics Subject Classification: 62.

Keywords and phrases: SNP selection, ethno-classification, chi-squared test, Kruskal-Wallis test, support vector machine.

Received May 14, 2010

**Results:** We ranked SNPs based on three statistical testing methods and used the top SNPs for ethno-classification via support vector machine. The best results were obtained with a chi-squared test of independence, where using only the top two mitochondrial SNPs resulted in a classification accuracy of 98.9%. The top 10 mitochondrial SNPs identified from all the three statistical tests were able to completely classify the three populations.

## 1. Introduction

Genetic variations such as single nucleotide polymorphisms (SNPs) can be useful for disease association studies (Burwinkel et al. [2]), personalized medicine (Burmester et al. [1]), and population studies (Park et al. [3], Zhou and Wang [7]). In order to reduce analysis complexity, finding relevant SNPs (often referred as feature selection) is one of the important tasks due to the large amount of SNPs in the human genome. Feature selection methods are usually employed to find informative SNPs. Feature selection methods can be divided into three major categories: filter methods, wrapper methods, and embedded methods. A detailed review of these selection methods is given by Saeys et al. [4]. Among the three feature selection approaches, the filtering approach is simple, efficient, and is not dependent on machine learning tools used in the classification.

When SNPs are used for population studies, they are usually treated as numeric data (Park et al. [3], Zhou and Wang [7]). Zhou and Wang used modified $T$-test and $F$-statistics to rank SNPs (Zhou and Wang [7]). Approximately 95% of samples were correctly classified into three ethnic groups using the top 100 SNPs from either modified $T$-test or $F$-statistics (Zhou and Wang [7]). Park et al. combined $t$-statistics and nearest shrunken centroid to rank the informativeness of each SNP (Park et al. [3]). They were able to completely classify the three populations using the top 82 SNPs. In this paper, we propose to use chi-squared test of independence to find marker SNPs, where genotype data are treated as nominal data. We evaluate this method with two non-categorical methods: Kruskal-Wallis test and analysis of variances (ANOVA).

Although previous studies explored SNPs from all the chromosomes and mitochondria to find informative SNPs for population classification (Park et al. [3], Zhou and Wang [7]), we focus exclusively on mitochondrial and *Y*-chromosome SNPs. *Y*-chromosome DNA has paternal lineage and is passed down from father to son. In 1999, Sutovsky et al. [5] reported that paternal sperm mitochondria (containing mitochondrial DNA) are marked with ubiquitin to select them for later destruction inside the embryo (Sutovsky et al. [5]). Therefore, mitochondria are usually inherited exclusively from the mother. Hence, these two genomic pieces are more relevant for ethnic classification task, because they are not shuffled by recombination and are mostly passed down intact.

This paper is organized as follows. Section 2 details the proposed methods. Section 3 shows the experimental results. Section 4 concludes this paper.

## 2. Methods

Figure 1 illustrates the flow chart of our proposed method. We first extract SNP data from International Hapmap Project (www.hapmap.org) download page. Discriminative SNPs are identified using one of the three statistical filters as discussed below. Top SNPs from these testing serve as SNP marker panel for subsequent classification task. Finally, we use a five-fold cross validation to evaluate the performance of our statistical testing methods in term of their prediction accuracy.

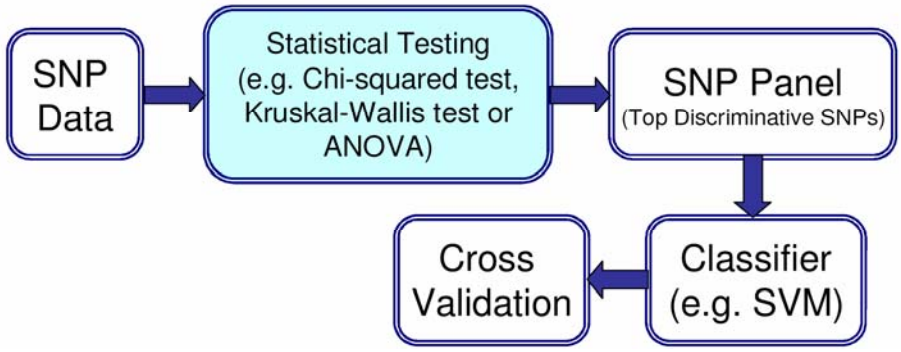


**Figure 1.** Main steps of ethno-classification process.

### 2.1. Chi-squared filter: Test of independence between ethnicity and SNP

Both ethnicity and genotype are nominal variables. A chi-squared test of independence can be used to tell, if the two nominal variables are associated or not. We first obtain a contingency table between the genotype and ethnicity for each SNP, with genotypes in the rows and ethnicities in the columns. Given an SNP contingency table, the statistic for chi-squared test of independence is given by

$$\chi_s^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}, \tag{1}$$

where $r$ is the number of genotypes for the SNP, $c$ is the number of ethnicities, $O_{i,j}$ is the observed frequency for $i$-th genotype and $j$-th ethnicity, and $E_{i\text{-}j}$ is the expected frequency, which is given by

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N}, \tag{2}$$

where $N$ is the total number of samples. The null hypothesis is that the SNP is unrelated with ethnicity. A $P$-value can be obtained by comparing the $\chi_s^2$ value against a chi-square distribution table of $(r-1)\times(c-1)$ degrees of freedom. A small $P$-value indicates an association between the SNP and ethnicity. For example, the chi-squared test of independence on SNP rs7876537 gives a $P$-value of 2.345126*e*-49. This indicates SNP rs7876537 is highly associated with ethnicity.

The method mentioned above can serve as a filter to select discriminative SNPs for the classification task. We call this chi-squared filter and apply this filter on SNPs from different genomic pieces, and we rank them based on the $P$-values in an ascending order. The top ranked genes will be used for classification.

### 2.2. Kruskal-Wallis filter: Non-parametric test for equal medians

The Kruskal-Wallis test is a non-parametric test for testing equality of population median among three or more groups. In order to perform Kruskal-Wallis test, genotype data has to be discretized first so they can be ranked (ordinal measurement requirement by Kruskal-Wallis). For a given SNP, the Kruskal-Wallis test statistic is given by

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1), \tag{3}$$

where $n_i$ is the number of observations in ethnic group $i, N = \sum_{i=1}^{k} n_i$ (the total number of observations in all $k$ groups), and $R_i$ is the sum of all the ranks of the $n_i$ observations in ethnic group $i$. For larger samples and/or for $k > 5, H$ is considered to be approximated by $\chi^2$ with $k$-1 degrees of freedom (Zar [6]). The null hypothesis is that all populations have identical distribution functions, and the alternative hypothesis is that at least two of the samples differ only with respect to location (median). For example, Kruskal-Wallis test of SNP rs7876537 gives a $P$-value of 7.5198$e$-43. Analogous to chi-squared filter, we call this the Kruskal-Wallis filter.

### 2.3. ANOVA filter: Parametric test for equal means

Discretization is also required to perform ANOVA. ANOVA makes more assumptions about data such as normality and homogeneity of variances. For a SNP, the $F$-statistic is given by

$$F = \frac{\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i} - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2}{N}}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i}} \times \frac{N-k}{k-1}, \tag{4}$$

where $X_{ij}$ is the digitized genotype value for the $j$-th sample in the $i$-th ethnic group, $k$ is the total number of ethnic groups, $n_i$ is the number of samples in the $i$-th ethnic group, and $N$ is the total number of samples in the test. The null hypothesis is that the mean genotype values are the same among all the ethnic groups. A $P$-value is obtained by comparing the $F$-statistic against an $F$ distribution table of ($k$–1) and ($N$–$k$) degrees of freedom. The smaller the $P$-value, the higher the discrimination power an SNP has. For example, an $F$-test on discretized SNP rs7876537 gives us an $F$-statistic of 343.71 with a $P$-value of less than 2.2$e$-16.

### 2.4. Classification methods

Many classification methods can be used for population classification such as neural networks, support vector machines (SVM), $k$-nearest neighbors, Bayesian classifiers, logistic regression, random forest, ensemble methods etc. We conducted some preliminary trials on neural networks, Bayesian classifier, random forest, support vector machine, and ensemble techniques such as bagging and boosting, and we did not see any drastic difference among these classifiers. In this study, we choose the well-studied SVM classifier. SVM comes with a rich set of kernel functions. Our preliminary study shows that the radial basis, linear, and sigmoid kernels perform similarly and better than the polynomial kernel. To simplify the comparative effort, we adopt SVM with linear kernel in this study to compare the performance of the three statistical filters.

## 3. Experimental Results

SNP data sets were obtained from the International Hapmap Project website (www.hapmap.org). The present study attempts to find SNPs, which distinguish among three major ethnic groups, Yoruba in Ibadan, Nigeria (YRI), a combination of Japanese in Tokyo (JPT) and Han Chinese (CHB) in Beijing (CHB + JPT), and Utah residents with ancestry from northern and western Europe (CEU). There are 90 individuals from each of the three ethnic groups. We extracted SNP data from $Y$-chromosome and mitochondria. We ran our experiments on SNPs from these two genomic pieces separately to study individual prediction performance.

### 3.1. Classification results using *Y*-chromosome SNPs

The three statistical testing methods are applied to each SNP on the *Y*-chromosome and top ranked SNPs are used for classification by SVM with a linear kernel. The performance of the three statistical filters assessed by using a five-fold cross validation accuracy is shown in Figure 2.

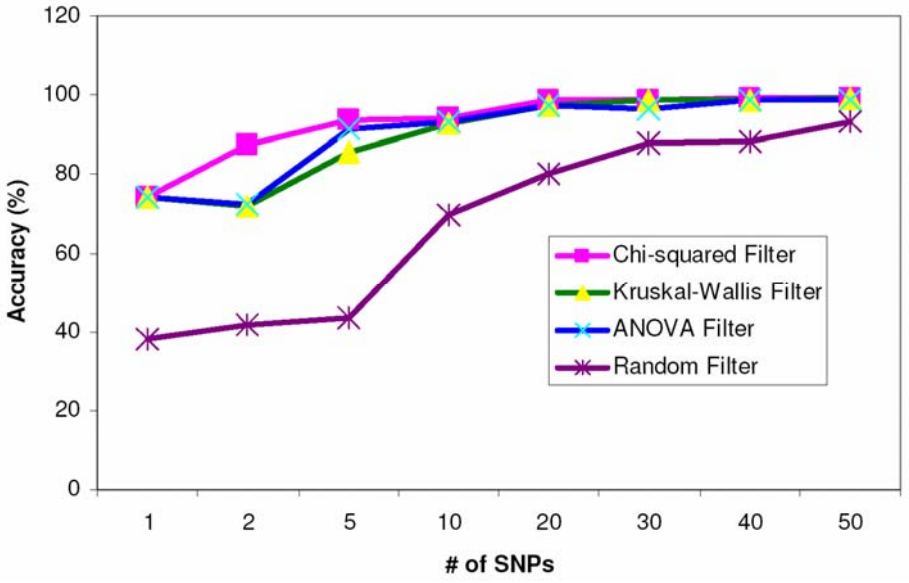


**Figure 2.** Five-fold cross validation accuracy using *Y*-chromosome SNPs filtered by three statistical filters and a random filter.

Chi-squared filter is able to find the top two *Y*-chromosome SNPs that reasonably classify the three ethnic groups. When using 5 or more top ranked SNPs, the performance of the three statistical filters are comparable and the prediction accuracy is approaching 100%. The random filter is significantly inferior to all the three statistical filters.

### 3.2. Classification results using mitochondrial SNPs

Similar experiments are conducted on mitochondrial SNPs. Figure 3 shows their performance assessed with a five-fold cross validation accuracy. Overall, chi-squared filter performs better than the other two statistical filters (Figure 3).

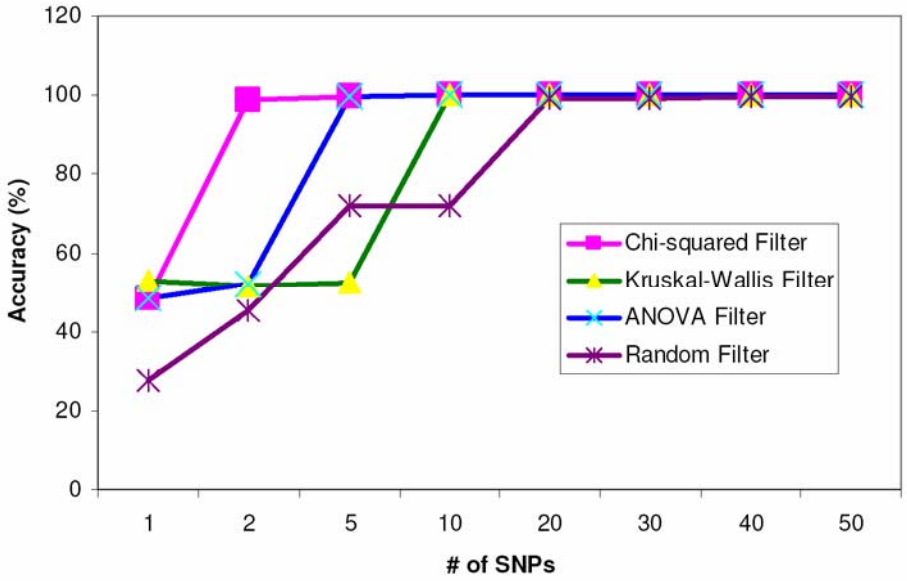


**Figure 3.** Five-fold cross validation accuracy using mitochondrial SNPs filtered by three statistical filters and a random filter.

As Figure 3 illustrates, using only one top ranked SNP does not produce satisfactory prediction results. The prediction accuracy quickly picks up, when using top two SNPs selected by the chi-squared filter. However, it is not good enough to use only 2 SNPs selected from other filters. When using ten or more top ranked SNPs, the classification performance is adequate for all the statistical filters. The top two SNPs identified by chi-squared test of independence are rs2248727 and rs28358887. Tables 1 and 2 show the contingency tables for the two SNPs. These two SNPs are complimentary to each other. SNP rs2248727 is able to tell Europeans apart from Africans and Asians, and SNP rs28358887 is able to tell Asians apart from Africans and Europeans, and together they are able to accurately classify the three populations. Comparing to SNPs from *Y*-chromosome, mitochondrial SNPs are able to reach more accurate prediction with smaller number of SNPs.

**Table 1.** Contingency table for rs2248727

| | African | Asian | European |
|---|---|---|---|
| NN | 8 | 5 | 1 |
| TT | 0 | 1 | 89 |
| CC | 82 | 84 | 0 |

**Table 2.** Contingency table for rs28358887

| | African | Asian | European |
|---|---|---|---|
| AA | 0 | 89 | 2 |
| GG | 90 | 0 | 88 |
| NN | 0 | 1 | 0 |

## 4. Conclusion

We evaluated three statistical filters to find SNPs for ethno-classification. The best performing filter was the chi-squared test of independence, which found informative SNPs without converting the genotype data. Using only the top two mitochondrial SNPs identified with a chi-squared filter, we obtained a classification accuracy of 98.9%. When ten or more top SNPs were used, the three statistical filters performed similarly. We could completely classify the three populations with only ten top mitochondrial SNPs, while earlier works on the same data set require significantly more SNPs to obtain similar results. One of the possible problems with previous studies was that rank-based SNP selection is susceptible to redundancy issues. An effort should be made to remove redundancy, if one would examine all the genomic pieces. The redundancy issue is less severe in our experiments due to our choice of the two small ethnically related genomic pieces. Although both genomic pieces are able to provide informative SNPs for satisfactory prediction with only two top SNPs, mitochondria is an excellent genomic source for ethnic classification.

## Acknowledgement

This work was supported by the National Institutes of Health [5T36GM008789-08].

## References

[1] J. Burmester and M. Sedova et al., DMET microarray technology for pharmacogenomics-based personalized medicine, Methods Mol. Biol. 632 (2010), 99-124.

[2] B. Burwinkel and K. Shanmugam et al., Transcription factor 7-like 2 (TCF7L2) variant is associated with familiar breast cancer risk: A case-control study, BMC Cancer 6 (2006), 268.

[3] J. Park and S. Hwang et al., SNP@Ethnos: A database of ethnically variant single-nucleotide polymorphisms 10.1093/nar/gkl962, Nucl. Acids Res. 35(suppl-1) (2007), 711-715.

[4] Y. Saeys and I. Inza et al., A review of feature selection techniques in bioinformatics, Bioinformatics 23(19) (2007), 2507-2517.

[5] P. Sutovsky and P. R. Moreno et al., Ubiquitin tag for sperm mitochondria, Nature 402(6760) (1999), 371-372.

[6] J. H. Zar, Biostatistical Analysis, Prentice Hall, (1999).

[7] N. Zhou and L. Wang, Effective selection of informative SNPs and classification on the Hapmap genotype data, BMC Bioinformatics 8 (2007), 484.

■