THE USE OF GENETIC POLYMORPHISMS AND DISCRIMINANT ANALYSIS IN EVALUATING GENETIC POLYMORPHISMS AS A PREDICTOR OF POPULATION GROUPS

Bruce F. Howell, J.D.

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

May 2002

APPROVED:

Olivia White, Major Professor
Gerard A. O'Donovan, Major Professor
Thomas L. Beitinger, Committee Member
Earl Zimmerman, Chair of the Department of Biology
C. Neal Tate, Dean of the Robert B. Toulouse School of Graduate Studies

Howell, Bruce F. The Use of Genetic Polymorphisms and Discriminant Analysis in Evaluating Genetic Polymorphisms as a Predictor of Population. Master of Science (Biology), May 2002, 46 pp., 8 tables, 4 matrices, references, 25 titles.

Discriminant analysis is a procedure for identifying the relationships between qualitative criterion variables and quantitative predictor variables. Data bases of genetic polymorphisms are currently available that group such polymorphisms by ethnic origin or nationality. Such information could be useful to entities that base financial determinations upon predictions of disease or to medical researchers who wish to target prevention and treatment to population groups. While the use of genetic information to make such determinations is unlawful in states and confidentiality and privacy concerns abound, methods for human "redlining" may occur. Thus, it is necessary to investigate the efficacy of the relationship of certain genetic information to ethnicity to determine if a statistical analysis can provide information concerning such relationship. The use of the statistical technique of discriminant analysis provides a tool for examining such relationship.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF MATRICES

Page

INTRODUCTION

In the United States, one hundred and fifty million Americans are provided health insurance that is based upon statistical risk factors. These Americans constitute the majority of Americans covered by health insurance and pay premiums that allow the continuation of such insurance for many others. As a result, insurance companies in this country are constantly seeking data that allow a matching of the premiums paid for the risk assumed while allowing a profit to be obtained.

Genetic information is a veritable cache of health care information. What better information to obtain for risk analysis than that of the diseases to which a person is prone based upon genetic analysis? Yet the problem is more complicated than appears at first blush. The recent revelation that the human genome contains fewer genes than originally predicted raises the question of whether or not various genetic predictors at this point in time are truly reliable. The problem is easier with genetically-based diseases such as Huntington's disease or cystic fibrosis but is more complicated with multifactoral diseases such as cancer or coronary disease.

In the mortgage industry, the concept of "redlining" undesirable property locations has been prohibited by many states. The same has been the subject of recent "genetic discrimination" laws in the insurance context. While this well-intentioned legislation is salutary, the reality is that readily available information provides an ability for any person, including those who are engaged in risk analysis, to "redline" population groups that possess genetic dispositions. The result is a potential "black market" for information which may surreptitiously be used to identify and then exclude certain population groups from insurance coverage or potentially employment. The exclusion may be patently illegal pursuant to the aforementioned laws, but data bases containing such information may be available to the unscrupulous operator

or to those who simply want to “fine tune” the risk, taking into account other risk bearing features. While no one would be so bold as to deny coverage blatantly for genetic reasons, the wealth of genetic information which is being produced and is available on a population scale provides more data which may be used in making actuarial decisions.

On the positive side, the identification of groups for the purpose of specific medicine treatments may be a good outcome for the statistical identification of population groups. This activity is already taking place in a more informal manner among the Hasidic Jewish population of New York where carriers for Tay-Sachs Disease and other diseases are routinely counseled on the genetic advisability of a proposed marriage. Identification of the potentially damaging alleles in population groups such as this group could allow the targeting a prophylactic medicine to such groups. Indeed, recently the entire genome of the Icelandic population was sold to a private company which may use such information to develop targeting drugs. (8) However, such targeting and identification poses the insurance and employability issues set forth above. Thus, intelligent legislative responses must be formulated. For example, what if a certain population group showed a lowered risk for heart disease based on genetic data? Would not a reduction in health insurance premiums be in order?

In this age of increasing bioinformatics (6), the appropriate use of population genetic information is statistically based. If the statistical basis for the conclusions of the analysis is flawed, then the fact that genetic information exists and is easily available will not matter. The assault on an individual’s privacy and the use of such information will matter for naught.

The statistical analysis regarding the use of genetic information may take many forms. However, for the purposes of this paper, only one analysis will be examined; i.e., discriminant analysis. Information available which identifies polymorphic alleles in the genetic code of

humans is the subject to which discriminant analysis shall be applied in order to determine if that statistical analysis can result in reliable predictors of population groups, which are genetic units and can be easily the subject of "genetic redlining".

Health Care Benefits in the United States are Delivered Through Actuarial Analysis

The basis for the delivery of a majority of health insurance benefits in the United States is statistical. Actuarial analysis provides the foundation for indemnity insurance, preferred provider payments and health maintenance organizations "capitation." Insurance companies use data to identify risk pools of members and classify those whose medical needs are great as "outliers".

Starting in the 1940s and 1950s, health care benefits were delivered through insurance companies and were based on actuarial underwriting which used health care history as a basis for analyzing risk. Also during this time frame, health care benefits began to be delivered through employment with many American having their benefits through their jobs.

During this period, there was a division between the insurance company which assumed the financial risk and paid the benefits, and the employer whose job it was to provide the employment and pay the premiums (usually shared with the employee). Thus, because of this division of responsibility, the information regarding the health of any particular worker was somewhat insulated from the knowledge of the employer.

The divisions between employment and insurance blurred starting before the passage of the Employee Retirement Income Security Act of 1974 ("ERISA"). Prior to the passage of ERISA, employers had begun to "self fund" health care benefits using insurance companies to administer payment of the benefits. Thus, the line between employer and insurer was not as clear. ERISA recognized the employers "self funding" of the risk of providing benefits and, thus, allowed the employer to be more in the information flow regarding the employee's physical

history. While the real onslaught of ERISA health care benefits did not start until the late 1980's, the current situation in the United States is that the vast majority of the 150,000,000 Americans who have health care benefits receive the same through ERISA plans, and the employers are intimately included in the risk analysis of insurability. Indeed, employers are considered "pools" of risk, as described above.

The ascension of ERISA has given rise to the managed care concepts that are prevalent today. Under managed care, the employer forges a relationship with the insurance company which acts as an administrator for the health care plan and seeks to provide quality health care benefits at an affordable price. Since risk must be managed and pools of high risk employees/insureds must be identified, the accumulation of health care information data is essential for managed care statistical analysis.

Predictors of Disease are Essential for Health and Life Insurance Analysis

Grouping of disease statistics by age and sex are some of the basic indicators of risk. There are, of course, many other factors in assessing risk such as life style, smoking, dangerous activities and the like. However, the basic tenet is the same – data drives the decision and is needed to provide appropriate risk assessment.

The concern of the "patchwork quilt" of state medical information confidentiality laws that has sprung up over the years and the recent developments under HIPAA (see, infra) is that individual patient information be protected. However, protected data (i.e., that data which contains no patient identifiers) may be used to provide information regarding groups of people which, for managed care companies, provides a more detailed analysis and risk profile. The use of population data provides an overview of the risks associated with certain areas of the country and certain groups and subgroups of people. Thus, genetic data can be of use in evaluating risk

for employers and insurance companies that underwrite the risk without ever having to know the individuals who compose the group While this statistical analysis is commonplace using common physical data, the question arises as to whether or not group genetic data is of use in looking at risk pools. The major question is whether or not such genetic data will provide better risk analysis and, thus, lowered costs, or discrimination.

For exactly this concern, many states have prohibited the use of genetic information for insurance**,** health risk and employment analysis. Twenty-six states have prohibited the use of genetic technology by insurance companies and employers in the evaluation of risk factors for insured and employees, respectively. At present, the United States Congress has not passed laws that would prohibit such practices; however, through the Health Insurance Portability and Accountability Act of 1996 ("HIPAA"), the use of genetic information is prohibited through the "pre-existing" clause legislation. In addition, the HIPAA privacy regulations regulate the disclosure of genetic information of an individual.

While these efforts are good, the efficacy of the prohibition on the use of genetic information is questionable. First, even though use of genetic material in insurance risk analysis is prohibited, there may still be the use of such information illegally. Secondly, proof of discrimination based on genetic standards is very difficult at best. It is easy for an employer or insurance company to deny employment or coverage on reasons other than genetic concerns. Employers and insurers are likely to take the position that other factors weigh against a person's insurability and employability. Finally, since population data is available, such data may be a source of information for risk assessment.

Thus, it is necessary to analyze the true present efficacy of population genetic data as a predictor of disease. Many articles and papers have been written on such predictions on the

individual basis. However, the analysis on genetic data predictability using statistical methods for identifying groups is in its infancy, and the purpose of this paper is to analyze certain data using a statistical test – discriminant analysis – that is useful in providing "groupings" based on data.

GENERAL DESCRIPTION OF THE DISCRIMINANT ANALYSIS TEST

Discriminant analysis is a procedure for identifying the relationships between qualitative criterion variables and quantitative predictor variables. Discriminant analysis is a procedure for identifying boundaries between groups of objects. The boundaries are those variable characteristics which distinguish such objects in the criterion groups. The main use of discriminant analysis is to predict group membership from a set of predictors, and discriminant analysis reveals similar conclusions as regression analysis. (11)

There is a twofold benefit to the use of discriminant analysis. First, discriminant analysis can reveal which variables are related to the criterion variables. Secondly, discriminant analysis can predict values on the criterion variable when values on the predictor variables are given. (11)

Discriminant analysis is essentially an adaptation of the regression analysis techniques for situation where the criterion variable is qualitative rather than quantitative. The assumptions for the data used in discriminant analysis are (a) a random sample, (b) normal distribution, (c) homoscedasticy and (d) correlation among the data. (11)

Discriminant Analysis Procedure

The procedure for using discriminant analysis is first to classify into two or more criterion groups a number of objects that are measured on each of a number of predictor variables. For example, if groups A, B and C are to be discriminated, then objects within the groups need to be classified with each of the groups (e.g., Object A1…..Object Az, Object B1….Object Bz, Object C1….Object Cz and so forth). This can be accomplished by using an input data matrix as shown on Exhibit 1. Scores on the predictor variables (x1, x2…xz) are then run. (11)

Objects of classification means that each object possesses one of the values on the associated qualitative variable. These are essentially, then, two groups; i.e., objects belonging to groups and objects having values on a qualitative variable. (11)

Examples of the use of discriminant classification are to classify predictor variables such as credit risk versus non-risk, smoker versus non-smoker, Protestant Catholic or Jew, Democrat, Republican or Independent or, in the immediate case for , Japanese, Druze, or Dane.

Note that the groups are mutually exclusive and that the input data is not really different from multiple correlation and regression analysis. The main difference is that the objects in DA are grouped beyond correlation or regression analysis according to some meaningful criterion. Also, every object is measured on the same set of predictor variables.

A criterion variable is composed of the classification labels attached to the objects. Thus, a criterion variable can have a minimum of two values; e.g., smoker versus non-smoker, Danish versus non-Danish. The criterion variable may also have several values; e.g., Protestant, Catholic or Bhuddist, or Japanese, Danish or Druze.

The object chosen for analysis along with the criterion variable dictates the nature of the predictor variables. For example, buyers of cars being predicted might lead to seeking data on age, sex, income, geographic home location and number of children in a family. In this paper, the variables used are single nucleotide polymorphisms, single tandem repeats and other allelic sequences.

The task of discriminant analysis is to assign to the given objects a qualitative label based on information on predictor or classification variables. Predictor variables are dictated by the objects and criterion variables chosen for analysis; e.g., buyers of autos, diseases, nationality of

genotypes. The effectiveness of discriminant analysis is in the existence of predictor variables which differ in mean value from one criterion group to another. (11)

In this analysis, some assumptions are critical. First, the variance of a predictor variable must be the same in the population from which the groups are drawn. Secondly, the correlation between any two predictor variables is the same in the populations from which the criterion groups have been sampled. (11)

The Discriminant Function

In a manner similar to regression analysis, the discriminant function uses a weighted combination of predictor variable values to classify an object into one of the criterion variable groups or, alternatively, to assign the object a value on the qualitative criterion variable. The function is described as "L" which represents a derived variable defined as a weighted sum of values on individual predictor variables. Each object's score on the discrimiant function (i.e., the discriminant score) depends on such object's values on the various predictor variables. Thus, L = b1x1+b2x2+b3x3….bzxz. In this formula, "x1, x2 …xz" forth represent values on the various predictor variables, and "b1, b2 ….bz" represent the weights associated with each of the respective predictor variables. L, then, results in the object's resultant discriminant score. Note that this is the same in essence as a multiple regression equation (q.v., y`=a+bx). However, the difference is that "y`" in regression is numerical while "L" in discriminant analysis is qualitative. This difference is accomplished by utilizing a "cutoff score". (11)

The Cutoff Score

The cutoff score is a method of assigning objects to one group or another. Objects with L>X are assigned to one group and those with L<X are assigned to another group. The defining parameters in this assignment are weights and cutoff scores. Obviously, the point of the exercise

is to minimize the number of classification errors and, thus, the cutoff score with the fewest errors of classification is the best cutoff score. For example, if the frequency of a certain single tandem repeat is >0.02, such repeat could be assigned to Group A and if < 0.02, assigned to Group B. Also, if the allele count is >57, the classification would be to Group A and, if < 57, to Group B.

The use of the cutoff score is subjective and depends heavily upon whether or not there is more than one predictor variable. If there is only one, the smaller the difference between the two groups on the predictor variable, the larger the overlap. If there are multiple predictor variables, weighting the various predictor variables is highly important. Such weighting derives a single predictor variable (i.e., the discriminant function). Thus, maximizing the difference to minimize the overlap is the rule. Unless there is no overlap, classification errors will occur. (11)

In the case of multiple predictor variables, in determining the weights of the predictor variables, the correlation which exists among the predictor variables are taken into account. The procedure may be generalized to any number of variables and allows the weighting of the discriminant analysis. Again, regression analysis provides an analogy; i.e., a high correlation between predictor variables and criterion variables results in reduced errors of prediction. (11)

More than Two Criterion Groups

If there are more than two criterion groups, then more than one discriminant function is needed. The rule is that one fewer discriminant function than the number of criterion groups is required unless there are fewer predictor variables than criterion groups. An example would be buyers of different automobiles; e.g., Acura, Infiniti and Volvo (criterion groups) based on income, profession, mortgages and education (predictor variables). In this case, two discriminant functions are needed. The first discriminant function discriminates Acuras from

Volvo and Infiniti buyers and the second discriminant function discriminates Infiniti buyers from Volvo buyers. The formulas for determining such discriminant functions are:

Discriminant Function One

L1 = 6y1-4y2+2y3+5y4

(with the stipulation if L > 100, assign to Volvo buyers)

Discriminant Function Two

L2 = 5y1+3y2-4y3-6y4

(with the stipulation if L < 75, assign buyer to Acura buyers)

In this exercise, the next step would be to establish the cutoff score and assign the buyers to the Acura, Infiniti or Volvo groups. Then, in sequence, the Acura buyers would be compared against all other, then the Infiniti buyers, then Volvo buyers. (11)

Stepwise Procedures

Stepwise procedures may be used with discriminant analysis as with regression analysis. Such procedures allow use of a smaller set that discriminates between or among the criterion groups in a manner that would be as well as the entire set itself. Here, one must be concerned about collinearity (i.e., the situation in which predictor variables are very highly correlated. Such problem can be avoided by not including overlapping data (e.g., sales, costs and profits). In an allelic analysis, this problem would not appear to be very significant. (11), (25).

Evaluating the Amount of Discrimination

There are several summary indices of the amount of discrimination achieved in a discriminant function evaluation. R2, the multiple correlation coefficient, is one of the indices as is Mahalanobis, D2, Wilks' Lambda and Rao's V. A meaningful evaluation of the discriminant function is in terms of the "actual errors of classification' in numbers and type. A confusion

matrix shows the tabulation of the objects' actual groups membership versus that of the predicted group membership. (Exhibit 2). What is important in the confusion matrix are the frequencies in the body of the table which reflects the associations between the predicted and actual group membership. (11)

Importance of the Predictor Variables

The relative importance of the predictor variables can be determined from the squared coefficient weights associated with each variable in the discriminant function. However, in order to do so, the discriminant function must be in the standardized "z score" form which is:

Lz = beta1z1+beta2z2…..+betzkzk

Squared beta weights reflect on the relative importance of the variables and do not reflect their absolute importance. (11)

Limitations on the Use of Discriminant Analysis

Sometimes discriminant analysis is incorrectly used where a regression analysis is more appropriate and powerful. Also, arbitrary assignment into groups by the cutoff score may overshadow the information given by the actual data. In addition, the "all or none" analysis contained in discriminant analysis can be challenged. The solution, obviously, is to do a tandem analysis of discriminant analysis and regression analysis on the data. As with correlation analysis and regression analysis, discriminant analysis is a function of the three key functions of statistical analysis – to wit, data reduction, inference and identification of association among variables. (11)

THE GATHERING OF THE DATA

The first step in the process is to gather the data for the discriminant analysis procedure; i.e., data revealing polymorphisms among populations. The gathering of this data was accomplished by visiting the ALFRED (the "Alelle FREquency Database") site on the Internet. This site is sponsored by Yale University and is found at http://alfred.med.yale.edu/alfred/index.asp. The site is designed to store and disseminate frequencies of alleles at human polymorphic sites for populations and is used for the study of population genetics and molecular anthropology. (16) The ALFRED site contains data from population groups and certain polymorphisms that occur in such groups. The format of presenting the data shows the chromosomal band position of the gene, the population name, the locus name, the locus symbol, the polymorphism name, the sample size, the sample identification, the allele name, the allele symbol, the frequency and the frequency identification. The sample and frequency identifications are used for purposes internal to ALFRED. As of 2001, more than 100,000 single nucleotide polymorphisms had been identified. (16)

The data used in this study was gathered for four population groups. Those population groups are the Druze, the Danish, the Japanese and the Europeans (mixed) and were selected on the availability of the data for common loci. For example, all of these groups reveal data for the dopamine receptor D2 (symbol DRD2) and, thus, provide a comparison of certain single nucleotide polymorphisms and single tandem repeats for that locus. Thus, a direct comparison may be made among the groups, the frequency of this data and the number of alleles for each group.

The dopamine receptor D2 was selected because of its importance in the neurotransmitter diseases of Alzheimer's Disease and Parkinson's Disease, two inflictions that are very current in

research and for which stem cell research holds some promise. Both are debilitating diseases and both have a genetic component, although there may be environmental contributors as well. In addition, the DRD2 gene has been examined as possibly having a role in the proclivity toward alcoholism. (1)

DRD2 encodes the dopamine D2 receptor which is critical in the functioning of neural circuits in the brain. The DRD2 gene spans >270 kilobases with an initial large intron of 250 kilobases. The gene is found at 11q22.3-q23.1 (2)

Dopamine receptors mediate enzymatic activities, metabolic rates and ion channels. These receptors are involved in neurological signaling. They are involved in cognitive and emotional functions and neurological disorders (17). There are five different receptors encoded by five separate genes. These genes are grouped further into two subgroups – the first comprised of the D1 and D5 receptors and the other composed of the D2, D3 and D4 receptors.

The sites within the DRD2 locus are:

5` (GAAA)n tetranucleotide STR

-141 C In/Del

Exon 8 SSCP

*Eco*RI site

Ser311Cys

TaqI D site

*Bcl*I site

Intron2 (GT)n dinucleotide STRP

Intron 1 (CT)n dinucleotide STRP

*Hinc*II site

*Taq*I A site

*Mbo*I site

*Taq*I B site

4-site haplotype (*Taq*I B, *Taq* I D, (CA) repeat, *Taq*I A)

5-site haplotype (*Taq*I B, *Taq*I D, (GT) STRP, *HincII)*

The sites chosen for this study are the 5`(GAAA)n tetranucleotide single tandem repeat polymorphism (STRP), the Intron 1 (CT) dinucleotide STRP, the –141 C In/Del polymorphism and the Intron 2 (GT)n dinucleotide STRP. A single tandem repeat polymorphism is a form of gene cluster where many identical genes lie in a tandem array. The Ins/Del polymorphism is a polymorphism that occurs due to insertion or deletion of genes. The Intron polymorphisms occur in the intron section of the DNA sequence and are the intervening sequences that are removed when the primary transcript is processed into RNA.

The 5` (GAAA) tetranucleotide STRP is an STRP that ends in the 5` sequences upstream of exon 1. The Intron 1 (CT)n dinucleotide STRP is a dinucleotide STRP located in Intron 1 and is 7608 base pairs upstream of exon 2. The –141 C In/Del polymorphism is a single nucleotide insertion/deletion polymorphism at –141 base pairs (upstream) of the start of transcription. The insertion allele corresponds to a restriction site but the deletion allele does not contain that site. The Intron 2 (GT)n dinucleotide STRP is an intron 2 dinucleotide STRP with a repeat structure varying in the dinucleotide repeat domain. This STRP is located 1311 base pairs upstream of exon 3 and 1384 base pairs downstream of the *Taq*I "D" site.

The population groups were chosen because of their commonality of polymorphisms but also because of their diversity in location and, perhaps, genetic history. The Danes, of course, are from Denmark while the Mixed Europeans contain genotypes that may also be considered to

be American. There may be some overlap between these populations. The Druze are a Middle Eastern groups of about a half a million people who live in the villages in the mountains of Syria, Lebanon, Israel and Jordan. The Japanese, of course, are inhabitants of Japan. These population groups provide diversity that will be useful as a background to run the data in order to determine if statistical significance occurs.

The data for each of the population groups is set forth on the following tables. The number in ( ) to the right of the population group name is the diploid sample size. The classification of populations into 1 and 2 as occurs in places in these tables means that the data on ALFRED was obtained from two sample groups, often as different dates.

Table 1 Intron 2 (GT)n Dinucleotide STRP

Population	Frequency by Allele Symbol					
	12	**13**	**14**	**15**	**16**	**17**
Druze 1 (2N=200)		0.285	0.150	0.385	0.180	
Druze 2 (2N=150)		0.300	0.133	0.407	0.160	
Danes 1 (2N=388)	0.003	0.186	0.098	0.451	0.263	
Danes 2 (2N=102)		0.127	0.108	0.529	0.235	
European 1 (2N=62)		0.161	0.129	0.435	0.274	
European 2 (2N=172)	0.006	0.099	0.122	0.599	0.169	0.006
Japanese (2N=100)			0.490	0.060	0.450	

Table 2 -141 C Ins/Del

Population	Frequency by Allele Symbol	
	Ins	**Del**
Druze 1 (2N=190)	0.021	0.979
Druze 2 (2N=142)	0.021	0.979
Danes 1 (2N=494)	0.119	0.881
Danes 2 (2N=180)	0.072	0.928
Europeans (2N=108)	0.056	0.944
Japanese (2N=102)	0.235	0.765

Table 3 5' (GAAA)n Tetranucleotide STRP

Population	Frequency by Allele Symbol					
	7	**8**	**9**	**10**	**11**	**12**
Druze 1 (2N=102)		0.059	0.127	0.176	0.167	0.235
Druze 2 (2N=96)		0.063	0.135	0.156	0.167	0.240
Danes 1 (2N=458)	0.002	0.124	0.061	0.212	0.214	0.188
Danes 2 (2N=64)		0.031	0.109	0.250	0.219	0.219
Europeans (2N=102)	0.010	0.108	0.039	0.275	0.196	0.176
Japanese (2N=100)		0.110	0.080	0.100	0.130	0.270
	13	**14**	**15**	**16**	**17**	**18**
Druze 1 (2N=102)	0.157	0.078	0.010			
Druze 2 (2N=96)	0.146	0.083	0.010			
Danes 1 (2N=458)	0.090	0.068	0.009	0.009	0.002	0.002
Danes 2 (2N=64)	0.078	0.063		0.031		
European (2N=102	0.147	0.029	0.020			0.010
Japanese (2N=100)	0.140	0.100	0.070			

Table 4 Intron 1 (CT)n Dinucleotide STRP

Population	Frequency by Allele Symbol					
	112	**114**	**116**	**118**	**120**	**122**
Druze 1 (2N=172)	0.006	0.029	0.826	0.064	0.006	
Druze 2 (2N=132)	0.008	0.030	0.818	0.068	0.008	
Danes 1 (2N=172)			0.762	0.058	0.029	
Danes 2 (2N=98)	0.020		0.857	0.020	0.010	
European (2N=92)	0.011	0.022	0.815	0.033	0.011	
Japanese (2N=96)	0.531	0.021	0.021			
	124	**126**	**128**	**130**	**132**	
Druze 1 (2N=172)	0.058		0.012			
Druze 2 (2N=132)	0.061		0.008			
Danes 1 (2N=172)	0.145			0.006		
Danes 2 (2N=98)	0.082				0.010	
European (2N=92)	0.098					
Japanese (2N=96)	0.385		0.031	0.010		

The Preparation of the Data

As can be seen from the preceding tables, only certain alleles in the populations had frequencies that were truly common to one another. Thus, the data selected this study had to be limited to these alleles that were common and the alleles that were not common were not used. The remaining alleles that were used are:

1. Intron 2 (GT)n dinucleotide STRP – Alleles 13, 14, 15, 16

2.-141 C Ins/Del – Alleles Ins and Del

3. 5`(GAAA)n tetranucleotide STRP – Alleles 8, 9, 10, 11, 12, 13, 14

4. Intron 1 (CT)n dinucleotide STRP – Alleles 116, 118, 120, 124

Also, note that the population samples are expressed as 2N in ALFRED. Thus, the true sample of N is equal to one-half of the population number expressed in ALFRED. It is easy to determine the number of individual genomes tested by dividing the 2N by one- half (e.g., since Druze in Intron 2 (GT)n dinucleotide STRP has a 2N of 200, 100 individuals would have been used in the study. This formulaic analysis accounts for the diploid nature of the human genome.

It is necessary to realize that the data is presented as frequencies in the tables. It is not possible to use these frequencies since, if all the data frequencies were used, each category would add up to a total of one, and analysis would be useless. However, it may be better to utilize the hard data numbers rather than the data frequencies, and, thus, it is necessary to translate the frequencies into the actual number of alleles. This transformation is done by the formula:

2(POP) X frequency = Alleles

The calculation is simple. Using Druze again from the Intron 2 (GT)n dinucleotide STRP, the population of 2N (i.e., 200) would be multiplied by the frequency for allele 13 and the product is 57 alleles (200 x 0.285 = 57).

The converted results are presented in the following tables using only the alleles to be used in this study.

Table 5 Intron 2 (GT)n Dinucleotide STRP

Population	**Number of Diploid Alleles**		
	14	15	16
Druze 1 (2N=200)	30	77	36
Druze 2 (2N=150)	20	61	24
Danes 1 (2N=388)	38	175	102
Danes 2 (2N=102)	11	44	28
European (2N=62)	8	27	17
European (2N=172)	21	103	29
Japanese (2N=100)	49	6	45

Table 6 -141 C Ins/Del

Population	**Number of Diploid Alleles**	
	Ins	**Del**
Druze 1 (2N=190)	4	186
Druze 2 (2N=142)	3	139
Danes 1 (2N=494)	59	435
Danes 2 (2N=180)	13	167
European (2N=108)	6	102
Japanese (2N=102)	24	78

Table 7 5' (GAAA)n Tetranucleotide STRP

Population	Number of Diploid Alleles						
	8	**9**	**10**	**11**	**12**	**13**	**14**
Druze 1 (2N=102)	6	13	18	17	23	16	8
Druze 2 (2N=96)	6	13	15	16	23	14	8
Danes 1 (2N=458)	57	28	97	98	86	41	31
Danes 2 (2N=64)	2	7	16	14	14	5	4
Europeans (2N=102)	11	4	28	20	18	15	3
Japanese (2N=100)	11	8	10	13	27	14	10

Table 8 Intron 1 (CT)n Dinucleotide STRP

Population	Number of Diploid Alleles			
	116	**118**	**120**	**124**
Druze 1 (2N=172)	142	11	1	10
Druze 2 (2N=132)	108	9	1	8
Danes 1 (2N=172)	131	10	5	25
Danes 2 (2N=98)	84	2	1	8
Europeans (2N=92)	75	3	1	9
Japanese (2N=96)	51	2	2	37

Using the basic discriminant analysis function, L, the following calculations show the trends in the population groups based upon the following Input Matrices for the alleles.

Matrix 1 Intron 2 (GT)n Dinucleotide STRP

Groups	**Objects**	**Scores**		
		x1	**x2**	**x3**
Druze	**14, 15, 16**	30	20	36
		20	61	24
Danes	**14, 15, 16**	38	175	102
		11	44	28
European	**14, 15, 16**	8	27	17
		21	103	29
Japanese	**14, 15, 16**	49	6	45

Using the formula L=b1x1+b2x2….bzxz and weighting all “b” factors by the percentage which the particular population related to all of the populations studied (to eliminate the population size bias), the results for Matrix 1 are:

1. L(Druze) = (2.39)(248) = 594.18
2. L(Danes) = (1)(398) = 398
3. L(Europeans) = (2.40)(205) = 491.6
4. L(Japanese) = (2.39)(100) = 239

Matrix 2 -141 C Ins/Del

Groups	**Objects**	**Scores**	
		x1	**x2**
Druze	**Ins/Del**	4	186
		3	139
Danes	**Ins/Del**	59	435
		13	167
Europeans	**Ins/Del**	6	102
Japanese	**Ins/Del**	24	78

The calculation of the discriminant function for Matrix 2 is as follows:

1. L(Druze) = (1.8)(232) = 598.97
2. L(Danes) = (1)(674) = 674
3. L(Europeans) = (1,8)(108) = 194.4
4. L(Japanese) = (1.8)(102) = 184.0

Matrix 3 5' (GAAA)n Tetranucleotide STRP

Groups	**Objects**	**Scores**						
		x1	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**
Druze	**8,9,10,11,12,13,14**	6	13	18	17	23	16	8
		6	13	15	16	23	14	8
Danes	**8,9,10,11,12,13,14**	57	28	97	98	86	41	31
		2	7	16	14	14	5	4
Europeans	**8,9,10,11,12,13,14**	11	4	28	20	18	15	3
Japanese	**8,9,10,11,12,13,14**	11	8	10	13	27	14	10

The discriminant functions for Matrix 3 are as follows:

1. L(Druze) = (1.85)(196) = 366.5
2. L(Danes) = (1)(500) = 500
3. L(Europeans) = (1.88)(99) = 187
4. L(Japanese) = (1.88)(93) = 175.66

Matrix 4 Intron 1 (CT)n Dinucleotide STRP

Groups	**Objects**	**Scores**			
		x1	**x2**	**x3**	**x4**
Druze	**116,118,120,124**	142	11	1	10
		108	9	1	8
Danes	**116,118,120,124**	131	10	5	25
		84	2	1	8
Europeans	**116,118,120,124**	75	3	1	9
Japanese	**116,118,120,124**	51	2	2	37

Running the formula for Matrix 4, the results are as follows:

1. L(Druze) = (1)(290) = 290
2. L((Danes) = (2.5)(266) = 661.73
3. L(Europeans) = (2.5)(87) = 220.57
4. L(Japanese) = (2.5)(92) = 230.60

ANALYSIS OF THE MANUALLY RUN ALLELE DATA

Examining the L values for the data run in the above matrices, the following results were observed.

In Matrix 1, the highest value (i.e., number of alleles weighted to reduce population size bias) of 594.18 was found in Druze with Europeans next at 491.6, Danes at 398 and Japanese at 239. Thus, if this particular allele was deleterious, the population with the greatest risk is the Druze with the Japanese having the lowest risk.

In Matrix 2, the Danes lead with 674, the Druze followed with 598.97 and the Europeans and Japanese had 194.4 and 184, respectively. Similar conclusions regarding risk are possible from these numbers.

In Matrix 3, the Danes were high with 500, the Druze next with 366.5 and the Europeans were third with 187. The Japanese came in lowest at 175.66.

In Matrix 4, the Danes led with 661.73, the Druze were next with 290, the Japanese third with 230.6 and the Europeans with 220.57.

In each example, the number for each population can be used as a cutoff score. For example, in Matrix 3 which deals with the 5` (GAAA)n tetranucleotide STRP, a cutoff score could be established for Danes of 500. Any allele count over 500 would be subjectively (and perhaps artificially) considered to be Danish. In a similar fashion, the cutoff score could be set at 366.5 for Druze, 187 for Europeans and 175.66 for Japanese for this particular polymorphism.

Thus, if a set of data was run which resulted in a discriminant function of 325.7, for example, given the arbitrary cutoff scores above, the Danish and Druze populations could be ruled out and the Europeans and Japanese considered. The idea is to establish the cutoff score that will result in the fewest errors of classification. Indeed, a simple scale of cutoff scores of

100, 200, 300, 400, 500 and so forth could be used to identify these population groups from the discriminant function produced by the alleles.

If a further discrimination was desired, one could discriminate among the particular alleles in each population. For example, in the case of Allele Matrix 2 (-141 C Ins/Del), a comparison and discrimination may be made for each of the insertions and deletions against each population group. The discriminant function would appear as follows:

(L) Druze – Ins = (1.8)(7) = 12.6

(L) Druze – Del = (1.8)(325) = 585

(L) Danes – Ins = (1)(72) = 72

(L) Danes – Del = (1)(602) = 602

(L) Europeans – Ins = (1.8)(6) = 10.8

(L) Europeans – Del = (1.8)(102) = 183.6

(L) Japanese – Ins = (1.8)(24) = 43.2

(L) Japanese – Del = (1.8)(78) = 140.4

Thus, looking at the L functions for the insertion genes, the high is 43.2 for the Japanese and the low is 10.8 for the Europeans with the Druze next at 12.6 and the Danes at 72. On the deletion side, the highest L function is 602 for the Danes with descending scores of 585 (Druze), 183.6 for the Europeans and 140.4 for the Japanese.

Again, the idea for this comparison would be to establish a cutoff score that would limit overlap and then classify.

Therefore, it is possible to take each of the discriminant functions for each group by total for the group or by each allele and establish a cutoff score that will serve as an identifier of population based upon the weighted number of polymorphic alleles. A confusion matrix

(Exhibit 2) could then be used to run actual polymorphic data from individuals against predicted polymorphic data using the cutoff scores for different populations.

Using the SAS Data

While the manual calculations set forth above give some results from which conclusions may be drawn, the use of the SAS Program to run a discriminant analysis function and to run other tests reveals more information about the data.

The program used to input the data into SAS is as follows:

```
DATA DRD2 ALLELES;
INPUT POPULATION$ ALLELES @@;
CARDS;
[Here input the actual data using the following type of format….Z
50 D 49 E 29 J 49…..]
PROC PRINT;
ID ALLELES;
PROC UNIVARIATE PLOT NORMAL;
VAR ALLELES;
PROC CORR;
VAR ALLELES;
PROC DISCRIM;
CLASS POPULATION;
VAR ALLELES;
PROC CANDISC;
VAR ALLELES;
```

CLASS POPULATION;

PROC STEPDISC;

VAR ALLELES;

CLASS POPULATION;

RUN.

The PROC PRINT command prints out the data arranged by indicated population. The PROC UNIVARIATE PLOT NORMAL runs several tests, the most important of which is the test of whether or not the distribution of the data is normal. The PROC CORR command runs a correlation analysis. The PROC DISCRIM command runs the discriminant analysis test. The PROC CANDISC command runs a canonical discriminant analysis function which is a dimension reduction technique related to principal component analysis and canonical correlation. Finally, the STEPDISC procedure selects a subset of quantitative variables to produce a good discrimination model using forward selection, backward elimination or stepwise selection. An example of the input programs for the Intron 2, the –141 C Ins,/Del. The 5` (GAAA) and the Intron 1 data are set forth on Exhibits 3, 4, 5 and 6, respectively.

The output for each of the four data sets is set forth in Exhibits 7, 8, 9 and 10. Prior to an examination of each output, a couple of observations are in order.

First, it appears that the larger data set (Matrix 3) produces a better stepwise output. In the smaller datasets, the stepwise procedure did not complete the program.

Secondly, as mentioned above, the assumptions for discriminant analysis is that the data is a random sample, normally distributed, homoscedastic and correlated. Thus, prior to examining any further outputs, an examination of whether or not this data meets these assumptions is necessary.

It is necessary to assume here that the samples collected in the ALFRED site are random. This is truly an assumption since there is no evidence either way.

As to normality of the distribution, the Shapiro-Wilks test on the Intron 1 and 5` data shows <0.0001 (normal distribution) while the –141C and Intron 2 data show 0.0058 and 0.0339, respectively, which evidences a non-normal distribution.

Homoscedasticity means that the variances of the *y* distributions in regression analysis are all equal to one another. (20) For purposes here, this will be assumed.

Finally, the variables are assumed to be correlated since each data set shows a correlation coefficient of 1.000.

As one final preliminary note, the discriminant analysis procedure is not a real statistical test in and of itself. That is to say, while certain of the components of discriminant analysis lend themselves to the traditional tests for statistical significance, the entire analysis is not done in the traditional "null hypothesis" model. Thus, in analyzing the following data set, the emphasis will be on evaluating the data with appropriate mention of statistical significance, where appropriate.

Analysis of the 5` (GAAA)n Tetranucleotide STRP Data Set

Since the 5` (GAAA)n tetranucleotide STRP data set was the only one of the four data sets to meet all the assumptions and to allow a complete run through the stepwise procedure, this data set will be analyzed for purposes of this thesis. For ease of referral, the entire program results are set forth on the immediately following pages and is highlighted for ease of referral.

5`(GAAA)n TETRANUCLEOTIDE STRP

CREATED BY

BRUCE F. HOWELL

ALLELES	POPULATION
12	Z
59	D
11	E
11	J
26	Z
35	D
4	E
8	J
33	Z
113	D
28	E
10	J
33	Z
112	D
20	E
13	J
46	Z
100	D
18	E
27	J
30	Z
46	D
15	E
14	J
16	Z
35	D
3	E
10	J

The UNIVARIATE Procedure
Variable: ALLELES

Moments

N	28	Sum Weights	28
Mean	31.7142857	Sum Observations	888
Std Deviation	30.3227786	Variance	919.470899
Skewness	1.81657489	Kurtosis	2.68560781
Uncorrected SS	52988	Corrected SS	24825.7143
Coeff Variation	95.6123648	Std Error Mean	5.73046651

Basic Statistical Measures

Location		Variability	
Mean	31.71429	Std Deviation	30.32278
Median	23.00000	Variance	919.47090
Mode	10.00000	Range	110.00000
		Interquartile Range	23.50000

NOTE: The mode displayed is the smallest of 5 modes with a count of 2.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	5.534329	Pr > |t|	<.0001
Sign	M	14	Pr >= |M|	<.0001
Signed Rank	S	203	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.756606	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.24257	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.390083	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.409628	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	113.0
99%	113.0
95%	112.0

The UNIVARIATE Procedure
Variable: ALLELES

Quantiles (Definition 5)

Quantile	Estimate
90%	100.0
75% Q3	35.0
50% Median	23.0
25% Q1	11.5
10%	8.0
5%	4.0
1%	3.0
0% Min	3.0

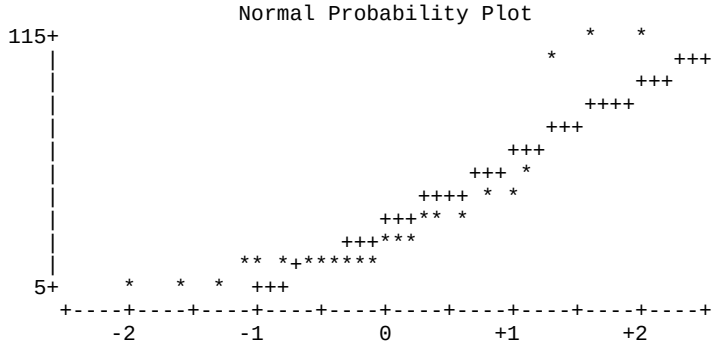
Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
3	27	46	22
4	7	59	2
8	8	100	18
10	28	112	14
10	12	113	10

```
Stem Leaf                     #             Boxplot
  11 23                       2                *
  10 0                        1                0
   9
   8
   7
   6
   5 9                        1                |
   4 66                       2                |
   3 03355                    5             +--+--+
   2 0678                     4             *-----*
   1 0011234568              10             +-----+
   0 348                      3                |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```

The UNIVARIATE Procedure
Variable: ALLELES

Normal Probability Plot



The CORR Procedure

1 Variables: ALLELES

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
ALLELES	28	31.71429	30.32278	888.00000	3.00000	113.00000

Pearson Correlation Coefficients, N = 28
Prob > |r| under H0: Rho=0

	ALLELES
ALLELES	1.00000

The DISCRIM Procedure

Observations	28	DF Total	27
Variables	1	DF Within Classes	24
Classes	4	DF Between Classes	3

Class Level Information

POPULATION	Variable Name	Frequency	Weight	Proportion	Prior Probability
D	D	7	7.0000	0.250000	0.250000
E	E	7	7.0000	0.250000	0.250000
J	J	7	7.0000	0.250000	0.250000
Z	Z	7	7.0000	0.250000	0.250000

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	5.94346

The DISCRIM Procedure

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j)$$

Generalized Squared Distance to POPULATION

From POPULATION	D	E	J	Z
D	0	8.60761	8.86713	4.94699
E	8.60761	0	0.00193	0.50366
J	8.86713	0.00193	0	0.56790
Z	4.94699	0.50366	0.56790	0

Linear Discriminant Function

$$Constant = -.5 \, \bar{X}_j' \, COV^{-1} \, \bar{X}_j \qquad Coefficient\ Vector = COV^{-1} \, \bar{X}_j$$

Linear Discriminant Function for POPULATION

Variable	D	E	J	Z
Constant	-6.69120	-0.26232	-0.23149	-1.02820
ALLELES	0.18735	0.03710	0.03485	0.07344

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.DRD2GAAAALLELES
Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function

$$D_j^2(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j)$$

Posterior Probability of Membership in Each POPULATION

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \,/\, SUM_k \exp(-.5\, D_k^2(X))$$

Number of Observations and Percent Classified into POPULATION

From POPULATION	D	E	J	Z	Total
D	4	0	0	3	7
	57.14	0.00	0.00	42.86	100.00
E	0	3	3	1	7
	0.00	42.86	42.86	14.29	100.00
J	0	1	5	1	7
	0.00	14.29	71.43	14.29	100.00
Z	0	1	1	5	7
	0.00	14.29	14.29	71.43	100.00
Total	4	5	9	10	28
	14.29	17.86	32.14	35.71	100.00
Priors	0.25	0.25	0.25	0.25	

Error Count Estimates for POPULATION

	D	E	J	Z	Total
Rate	0.4286	0.5714	0.2857	0.2857	0.3929
Priors	0.2500	0.2500	0.2500	0.2500	

The CANDISC Procedure

Observations	28	DF Total	27
Variables	1	DF Within Classes	24
Classes	4	DF Between Classes	3

Class Level Information

POPULATION	Variable Name	Frequency	Weight	Proportion
D	D	7	7.0000	0.250000
E	E	7	7.0000	0.250000
J	J	7	7.0000	0.250000
Z	Z	7	7.0000	0.250000

The CANDISC Procedure

Multivariate Statistics and Exact F Statistics

S=1 M=0.5 N=11

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.36856946	13.71	3	24	<.0001
Pillai's Trace	0.63143054	13.71	3	24	<.0001
Hotelling-Lawley Trace	1.71319282	13.71	3	24	<.0001
Roy's Greatest Root	1.71319282	13.71	3	24	<.0001

The CANDISC Procedure

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.794626	0.782871	0.070931	0.631431

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	1.7132		1.0000	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.36856946	13.71	3	24	<.0001

NOTE: The F statistic is exact.

The CANDISC Procedure

Total Canonical Structure

Variable	Can1
ALLELES	1.000000

Between Canonical Structure

Variable	Can1
ALLELES	1.000000

Pooled Within Canonical Structure

Variable	Can1
ALLELES	1.000000

The CANDISC Procedure

Total-Sample Standardized Canonical Coefficients

Variable	Can1
ALLELES	1.552973583

Pooled Within-Class Standardized Canonical Coefficients

Variable	Can1
ALLELES	1.000000000

Raw Canonical Coefficients

Variable	Can1
ALLELES	0.0512147520

Class Means on Canonical Variables

POPULATION	Can1
D	2.033957293
E	-0.899916356
J	-0.943814715
Z	-0.190226222

The STEPDISC Procedure

The Method for Selecting Variables is STEPWISE

Observations	28	Variable(s) in the Analysis	1
Class Levels	4	Variable(s) will be Included	0
		Significance Level to Enter	0.15
		Significance Level to Stay	0.15

Class Level Information

POPULATION	Variable Name	Frequency	Weight	Proportion
D	D	7	7.0000	0.250000
E	E	7	7.0000	0.250000
J	J	7	7.0000	0.250000
Z	Z	7	7.0000	0.250000

The STEPDISC Procedure
Stepwise Selection: Step 1

Statistics for Entry, DF = 3, 24

Variable	R-Square	F Value	Pr > F	Tolerance
ALLELES	0.6314	13.71	<.0001	1.0000

Variable ALLELES will be entered.

All variables have been entered.

Multivariate Statistics

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.368569	13.71	3	24	<.0001
Pillai's Trace	0.631431	13.71	3	24	<.0001
Average Squared Canonical Correlation	0.210477				

The STEPDISC Procedure
Stepwise Selection: Step 2

Statistics for Removal, DF = 3, 24

Variable	R-Square	F Value	Pr > F
ALLELES	0.6314	13.71	<.0001

No variables can be removed.

No further steps are possible.

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	ALLELES		0.6314	13.71	<.0001	0.36856946	<.0001	0.21047685	<.000

First, note that the data meets the requirements for normal distribution with a Shapiro-Wilk statistic of 0.756606 and a probability of normal distribution of <0.0001. (alpha=0.05) (Shapiro-Wilk, 0.756 <p<0.001).

Secondly, note that the data per the box plot appears to be skewed.

Thirdly, in the discrimination procedure, note that the weights given to the populations are equal. This is different than the weighting that was done manually to eliminate population size bias. This may be part of the explanation for the linear discriminant functions of the four populations being different from those functions run manually.

Using the CANDISC procedure, the observations are 28, the groups are 4 and the discriminant function is 3 since there are 4 groups. Wilks' Lambda is the proportion of the total variance in the discriminant scores not explained by differences among the groups. (25). The test results in a quantity from 0 to 1 which measures the amount of variability among the data that is not expressed by the effect of the numbers on the examined factor (e.g., the amount of variability among weight that is not accounted for by the examined diets). The Wilks' Lambda test shows that ratio of the within-groups sum of squares to the total sum of squares and, per the output, is significant. (Wilks' Lambda 0.368, p<0.001). Here, approximately 37% of the

variance is not explained by group differences. This is statistically significant and the null hypothesis of Druze=Danes=Europeans=Japanese is rejected.

The CANDISC program produces a canonical correlation among the data. Here, the correlation is 0.794 which, being close to 1.000, is somewhat correlated. The canonical correlation measures the association between the discriminant scores and the groups. Here, the association would appear to be high.

The Eigenvalue (25) is a ratio of the between-groups sum of squares to the within-groups sum of squares. It is a function of roots of matrices. This value measures the spread of the group centroids in the dimension of multivariate space. (20) Here, is 1.713 which is statistically significant (Eigenvalue 1.713, $p < 0.0001$).

Finally, evaluating the STEPDISC procedure, the R2 has a value of 0.631 with a probability of <0.0001. (Stepwise elimination, $R2=0.631$, $p<0.0001$). This shows that the data is capable of being subjected to the stepwise procedure and in this procedure is statistically significant.

SUMMARY OF THE SAS AND MANUAL ANALYSIS

The SAS output is somewhat less intuitive than the computation of the discriminant functions done manually. However, the SAS output is important to show different tests, such as canonical correlation and Wilks Lambda, that are necessary to test the data for the discriminant function.

Here, the linear discriminant functions for the different populations shown by the SAS output and the manual output are as follows:

Population	**SAS**	**Manual**	**SASRank**	**MRank**
Druze	0.073	366.5	3	2
Danes	0.187	500	2	1
Europeans	0.037	187	4	3
Japanese	0.348	175.6	1	4

In comparing the results produced by SAS to those produced manually, it is easy to see that they are quite different. This can only be explained by the fact that, in the SAS program, the populations were weighted equally. Also, since there is no correlation among the ranks of the populations in the above table, different results occur significantly when a cutoff score is to be used.

CONCLUSION

The discriminant analysis statistical test will produce a method of discriminating between populations based solely on the knowledge of genetic polymorphisms. Of course, data are required to be collected in order for such actual data to be compared against predetermined and established cutoff scores for various populations. Such information may be used for discrimination in beneficial or non-beneficial manners. The use to which such information is put will be determined by public policy. Such policy needs to swiftly be determined since the information regarding genetic polymorphisms and population is growing and more available each day.

SOURCES

1. Blum, K.; Noble, E.P.; Sheridan, P.J; Montgomery, A.; Ritchie, T; Jagadeeswaran, P.; Nogami, H.; Briggs, A.H.; Cohn, J.B. Allelic association of human dopamine D(2) receptor gene in alcoholism. J.A.M.A. 263: 2055-2060 (1990).

2. Eubanks, JH, Djabali, M., Selleri, L., Grandy, DK, Civelli, O., McElligott, DL., Evans, GA. "Structure and linkage of the D2 dopamine receptor and neural cell adhesion molecule genes on human chromosome 11q23". Genomics 14:1010-8. (1992) Online citation.

3. Freeman, S. and Herron, J. (Second Ed.). 2001. *Evolutionary Analysis*. Prentiss Hall, New Jersey. 704 pp.

4. Genetic Discrimination ACOEM Calls for Guidelines to Limit Use of Genetic Testing in Workplace. BNA Employment Discrimination Report, Vol. 16, No. 9: 271-272. (2001)

5. Gifford, D. Blazing Pathways Through Genetic Mountains. Science 293: 2049-2051. (2001).

6. Glover, K., Domeika, H., Christiansen, J., Miles, A. and Watts, T. 2001. *Collisions at the Intersection: Law and Bioinformatics*. BNA Health Law Reporter, Vol. 11, No. 6: 233-238.

7. Graur, D. and Li, W. 2000. *Fundamentals of Molecular Evolution* (Second Ed.). Sinauer Associates, Inc., Massachusetts. 481 pp.

8. Gulcher, J.R.; Steffansson, K. The Icelandic Healthcare Database and informed consent. N. Engl. J. Med. 2000; 342: 1827-30 (2000).

9. Hawley, R. and Mori, C. 1999. *The Human Genome; A User's Guide*. Academic Press, London, England. 415 pp.

10. Holtzman, N.; Marteau, T. Will Genetics Revolutionize medicine? N. Engl. J. Med. 343:141-4 (2000).

11. Kachigan, S. 1991. *Multivariat Statistical Analysis*. Radius Press, New York. 303 pp.

12. Kwok, P. Genetic Association by Whole-Genome Analysis? Science 294:1669-1670. (2001).

13. Lange, K. 1997. *Mathematical and Statistical Methods for Genetic Analysis*. Springer Press, New York. 265 pp.

14. Lewin, B. 2000. *Genes VII*. Oxford University Press, Oxford, England. 990 pp.

15. Mettler, L., Gregg, T., and Shaffer, H. 1988. *Population Genetics and Evolution* (Second Ed.). Prentiss Hall, New Jersey. 325 pp.

16. Osier, M., Cheung, K., Kidd, J., Pakstis, A., Miller, P. and Kidd, K. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms – an update. Nucleic Acids Research, 29:317-319 (2001).

17. Sokoloff, P., Giros, B., Martres, MP., Bouthenet, ML., Schwartz, JC. "Molecular cloning and characterization of a novel dopamine receptor (D3) as a target for neuroleptics". Nature 347:146-51. (1990) Online citation.

18. Strokstad, E. Data Hoarding Blocks Progress in Genetics. Science 295: 599. (2002)

19. Sveinbjornsdottir, S; Hicks, A. A.; Jonsson, T.; Petursson, H.; Guomundsson, G.; Frigge, M.; Kong, A; Gulcher, J.; Stefansson, K. Familial Aggregation of Parkinson's Disease in Iceland. N. Engl. J. Med. 343: 1765-70 (2000).

20. SPSS Base 10.0 Applications Guide. 1999. SPSS Inc., Chicago, Illinois. 427 pp.

21. Szathmary, E.; Jordan, F.; Csaba, P. Can Genes Explain Biological Complexity? Science 232: 1315-1316 (2001)

22. Temple, L.; McLeod, R.; Gallinger, S.; Wright, J. Defining Disease in the Genomics Era. Science 293: 807-808 (2001).

23. Winkelmann, B.R.; Hager, J.; Kraus, W.E.; Merlini, P.; Keavney, B.; Grant, P.J.; Muhlestein, J.B.; Granger, C.B. Genetics of Coronary Heart Disease: Current Knowledge and Research Principles. Duke Clinical Research Institute. Am. Heart J. 2000; 140:S1-S2. 549 pp.

24. Wood, A. J. J. Racial Differences in the Response to Drugs – Pointers to Genetic Differences. N. Engl. J. Med., Vol. 344, No. 18: 1393-1395 (2001).

25. Zar, J. 1999. *Biostatistical Analysis* (Fourth Ed.). Prentiss Hall, New Jersey. 663 pp. plus appendices.