DEFINING INTERMEDIATE PHENOTYPES
BY k-MEANS CLUSTERING

by

Alfred Lin
David R. Cox
Dee Pei
Neil J. Risch
Yii-Der Ida Chen
Chin-Fu Hsiao
Lee-Ming Chuang
Chii-Min Hwu
Victor Dzau
Beatriz Rodriguez
Richard A. Olshen

Technical Report No. 212
August 2000

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California
DEFINING INTERMEDIATE PHENOTYPES
BY k-MEANS CLUSTERING

by

ALFRED LIN
DAVID R. COX
DEE PEI
NEIL J. RISCH
YII-DER IDA CHEN
CHIN-FU HSIAO
LEE-MING CHUANG
CHII-MIN HWU
VICTOR DZAU
BEATRIZ RODRIGUES
RICHARD A. OLSHEN

Technical Report No. 212
August 2000

This research was supported by grants U01 HL54527-0151 from the National Heart Lung and Blood Institute, NIH (USA); NSC-85-2331-B002-350Y, NSC-86-2314-B002-345Y, and NSC-87-2312-B002-021 Y from the National Science Council, Taiwan, ROC; and NIH grant 2 R01 CA59039

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California
Defining Intermediate Phenotypes by $k$-Means Clustering

BY ALFRED LIN  
*Department of Statistics, Stanford University, Stanford, CA*

DAVID R. COX  
*Department of Genetics, Stanford University, Stanford, CA*

DEE PEI  
*Department of Endocrinology and Metabolism, Tri-Service General Hospital, Taipei, Taiwan*

NEIL J. RISCH  
*Department of Genetics, Stanford University, Stanford, CA*

YII-DER IDA CHEN  
*Division of Endocrinology, Stanford University, Stanford, CA*

CHIN-FU HSIAO  
*National Health Research Institute, Taipei, Taiwan*

LEE-MING CHUANG  
*Department of Medicine, National Taiwan University Hospital, Taipei, Taiwan*

CHII-MIN HWU  
*Department of Medicine, Veterans General Hospital, Taipei, Taiwan*

VICTOR DZAU  
*Department of Medicine, Brigham & Women’s Hospital, Harvard Medical School, Boston, MA*

BEATRIZ RODRIGUEZ  
*Hawaii Center for Health Research, Honolulu, HI*

AND RICHARD A. OLSHEN  
*Department of Health Research and Policy, Stanford University, Stanford, CA*  
e-mail: olshen@stat.stanford.edu
**Lin et al.**

**SAPPHIRE**

Stanford, Asian and Pacific Program in Hypertension and Insulin Resistance

**Stanford**
- Yi-Der I. Chen, PhD
- Bonnie Chung
- David R. Cox, MD, PhD
- David A. Hinds, PhD
- Winnie M. Koo, MPH
- Michael Olivier, PhD
- Richard Olshen, PhD
- Robert S. Pesich
- Koustubbh M. Ranade, PhD
- Neil J. Risch, PhD

**Boston**
- Richard E. Pratt, PhD
- Victor J. Dzau, MD

**Hawaii**
- Stephen D. Arnold, MD
- Jerome P. Baraquio, BSBA
- Nikko Michele Baraquio, MA
- J. David Curb, MD, MPH
- Dominic J. Estrella, BSBM
- Joyce E. Gilbert, MPH
- John Grove, PhD
- Evelyn M. Heim, BSBA
- Kamal H. Masaki, MD
- Michael S.L. Mau, BSEE
- Craig H. Nakatsuka, MD
- Ofelia S. Rabanal
- Beatriz L. Rodriguez, MD, PhD
- G. Webster Ross, MD
- Benjamin A. Tamura, MD
- Frances K. Wakashige, BSBA

**Taiwai**
- Fu-Tein Chiang, MD
- Mau-Song Chang, MD
- Jaw-Wen Chen, MD
- Ying-Tsung Chen, MD
- Kai-Ming Chu, MD, PhD
- Jui-Ling Chuang, MS
- Lee-Ming Chuang, MD, PhD
- Su-Chao Ding
- Martin M-T Fuh, MD
- Low-Tong L. Ho, MD
- Chin-Fu Hsiao, PhD
- Chao Agnes Hsiung, PhD
- Chii-Min Hwu, MD
- Chi-Yuang Jeng, MD, PhD
- Chin-Fai kwok, MD
- Wen-Jane Lee, PhD
- Shih-Jong Lin, MD
- Dee Pei, MD, PhD
- Wayne H-H Sheu, MD, PhD
- Kuang-Chung Shyr, MD
- Tong-Yuan Tai, MD
- Chih-Tai Ting, MD, PhD
- Hui-Ju Tsai, MS, MPH
- Yung-Zu Tseng, MD
- Dan-Jiang Wang, MD
- Kwan-Dun Wu, MD

**Summary**

The paper is about \( k \)-means clustering as it applies to finding intermediate phenotypes in a search for genes that predispose to hypertension in an Asian and Pacific Island population. Clustering was based on variables, none marginally Guassian, that quantify levels of plasma lipids and the metabolism of insulin and glucose. Technologies for choosing “how many clusters” firmly choose two for the 597 women and somewhat less certainly two for the 535 men. In each case one cluster is clearly the “insulin resistant” cluster. In women cluster membership is significantly associated with hypertension, much less so with men. We devised permutation tests for this association that respect family structures. As well, we developed permutation tests of the familiality of variables by which the clusters were constructed and for cluster membership itself.

*Some key words:* Hypertension; Insulin resistance; \( k \)-means clustering; Permutation tests
1. INTRODUCTION

Hypertension affects 15% to 20% of the adult population and is one of the major risk factors for stroke, myocardial infarction, and renal disease. Despite intense effort, our understanding of the pathogenesis of hypertension is rather poor. Twin, adoption, and epidemiologic studies indicate that variation in blood pressure is genetically determined [Lifton:1996]. One of the goals of the Stanford Asia and Pacific Program for Hypertension and Insulin Resistance (SAPPHIRe) is to identify genes for essential hypertension. To reduce genetic heterogeneity, a possible confounding factor in any attempt to map loci for hypertension, we have focused on the relatively homogeneous Chinese and Japanese populations. The subjects upon which we actually report are groups of sibs, the proband always being hypertensive and one or more sibs hypertensive or hypotensive in senses we make precise.

We begin by utilizing prior knowledge that the hypertensive phenotype is often associated with alterations in the levels of blood lipids, insulin, and glucose. Syndrome X or Insulin Resistance (IR) Syndrome, described by Reaven et al [Reaven:1988], is a pre-clinical syndrome in humans. Under the umbrella of this syndrome, there are a number of metabolic and hemodynamic abnormalities, including glucose intolerance, hyperinsulinemia, hypertriglyceridemia, low HDL cholesterol, and elevated blood pressure [Chen:1997], all of which are highly correlated with "insulin resistance". Whole body insulin resistance is defined as a sluggish glucose uptake by muscle in response to normal levels of insulin secreted by the pancreas. The underlying hypothesis for the observation of this clustering is based on the fundamental concept that the pathway of glucose uptake by muscle and adipose tissue in response to insulin is compromised in insulin resistant individuals [Pei:1994, Reaven:1993]. In order to compensate for this defect, beta cells in pancreatic islets hyper-secrete insulin in an attempt to maintain plasma glucose homeostasis [Reaven:1988, Laws:1992]. However, whole body hyperinsulinemia is not benign. In the presence of insulin resistance, insulin can stimulate the hepatic production of triglycerides, resulting in elevated triglyceride pool size, followed by decreased levels of HDL cholesterol, increased presence of smaller and denser LDL particles, elevation of PAI-1 levels and other atherothrombogenic factors such as fibrinogen [Chen:1997]. All of the above are significant risks for coronary heart disease [Pei:1994, Austin:1990, Juhan-Vague:1996, Modan:1995]. In addition, elevated insulin has been shown to be associated with elevated blood pressure in both cross-sectional [Haffner:1992] and prospective studies [Meehan:1994]. Although there are few experiments in humans demonstrating the direct causal effect of insulin on the elevation of blood pressure, both exercise and weight loss [Su:1995, DeFronzo:1975] have been demonstrated to result in improvement in insulin action and decrease in insulin level, associated with a decrease in blood pressure. These studies provide indirect evidence for a possible causal link between insulin and blood pressure. In animal studies, particularly in rats [Brands:1991, Krotiewsky:1979], sustained hyperinsulinemia can lead to elevated blood pressure. Available information further suggests that at least two pathways may link insulin resistance, hyperinsulinemia and elevated blood pressure. Insulin may promote renal tubular sodium resorption, thus lead to blood volume expansion [Modan:1984, Reaven:1996]. Increase in insulin concentration was also associated with significant increase in plasma catecholamine concentration, implying the presence of excess sympathetic nervous system activity in insulin resistant individuals, leading to hypertension [Ho:1990]. Insulin resistance syndrome is one of the plausible causes, not results, of high blood pressure [Reaven:1988]. Both insulin resistance and high blood pressure are markedly affected by obesity (body mass index-BMI and/or waist to hip ratio-W/H) [Manicardi:1986, Cigolini:1991]. However insulin resistance is a characteristic feature of primary hypertension independent of obesity [Pollare:1990, Shen:1988]. Non-hypertensive offspring of families with hypertension are more

We reason that the search for genes impacting blood pressure could be facilitated by grouping subjects on the basis of their plasma lipid, insulin and glucose measurements. Such a grouping is termed an intermediate phenotype. Thus, if our notion is correct, we should be able to define such intermediate phenotypes; and these phenotypes may be familial. However, lipid, glucose and insulin measurements are seen not to be Gaussian, jointly or even marginally. Random or mixed effects analysis of variance approaches to their understanding are thus suspect. Indeed, we were compelled to develop our own novel methodologies for defining the groups and for establishing their familiarity. For the former, we began with conventional algorithms for k-means clustering and adapted them to our purposes. For the later, we developed permutation tests that we believe are entirely new. We came to the problem of defining groups with no explicit preconceived notions. It may therefore seem remarkable that the groups we found - two each in men and women - can be defined in terms of the well known concept of insulin resistance.

2. SOURCES OF DATA

2.1. Participants and Setting

The SAPPHIRE Network consists of three field centers, with the Stanford field center encompassing the greater San Francisco Bay Area, the Hawaii field center covering the island of Oahu and outer islands, and the Taiwan field center consisting of a consortium of the three major hospitals located in Taipei and one branch hospital in Taichung. Seventy-eight percent of subjects recruited at the Stanford field center are of Chinese descent, while the remaining are of Japanese descent. For the Hawaii field center, only 16% are of Chinese descent while the majority, 84%, are of Japanese ancestry. For the Taiwan field center, all subjects are Chinese.

The Stanford and Hawaii field centers recruited the majority of their subjects through collaborative efforts with the Kaiser-Permanente Group. Newspaper and TV ads constitute other recruitment sources. The Taiwan field center is made up of the following major hospitals: National Taiwan University Hospital, the Veterans General Hospitals of Taipei and Taichung, and the Tri-Service General Hospital/National Defense Medical Center. The major sources of subject referrals are the Cardiology and Endocrinology Clinics in these hospitals.

2.2. Inclusion/Exclusion Criteria

The study design incorporated both concordant sib-pairs (both sibs with hypertension) and discordant sibs (one hypertensive and one hypotensive sib). Subjects were included on the following bases.

1. Current ages of subjects must be between 35 and 60 years. Subjects currently over age 60 may also be eligible provided that documentation of their hypertensive status prior to age 60 is available.

2. Japanese or Chinese ancestry, that is all four grandparents Japanese or all four grandparents Chinese. Hypertension is defined as follows: systolic BP (SBP) greater than or equal to 160 mm Hg or diastolic BP (DBP) greater than or equal to 95 mm Hg or taking
two medications for high blood pressure (Stage II hypertension). Alternatively, the subject could have uncontrolled hypertension, i.e., be taking one medication for high blood pressure and either SBP greater than or equal to 140 or DBP greater than or equal to 90 mm Hg.

Discordant sib-pairs where one sibling is hypertensive (as above) and the other hypotensive were also included in the study. Hypotension is defined as BP in the bottom 30% of the age and sex-adjusted BP distribution, which in our population translates into the following BP values:

1. For males under 45 years, SBP less than or equal to 115 mm Hg and DBP less than or equal to 76. For males over 45, SBP less than or equal to 122 and DBP less than or equal to 78 are used.

2. For females younger than 45, hypotension is defined as SBP less than or equal to 107 and DBP less than or equal to 70 mm Hg. For those over 45, the cut-off is SBP less than or equal to 118 and DBP less than or equal to 75.

There is no upper age cut-off for hypotensive sibs as long as both SBP and DBP readings are below the limit. However, the hypotensive sib must meet the lower age cut-off greater than or equal to 35 years. Families were excluded from the study if they met any of the following criteria:

1. One of the affected sibs is adopted (i.e. no parent in common) or if the sibs have only one parent in common.

2. Both parents have been treated for hypertension before the age of 60. If offspring reports about parents' hypertensive status are conflicting, then a single reliable report of hypertension in both parents before age 60 is reason for exclusion. This exclusion criterion, however, does not apply to discordant sib-pairs.

3. Individuals known from previous data to be diabetic were excluded. However, diabetes uncovered as a result of SAPPHIRE lab work did not lead to exclusion.

4. Severe kidney disease (except stones and remote infections), meaning creatinine greater than 1.5 mg/dl, unless documented proof that the subject met inclusion criteria prior to increase in creatinine levels.

5. A body-mass index (BMI) greater than 35.

6. In addition, the following conditions are considered as cause for exclusion: ongoing (or within the past 6 months) treatment for cancer; terminal illness (life expectancy fewer than 6 months); liver cirrhosis or any other chronic illness; pregnancy or fewer than 6 months post-partum.

For sib-pairs meeting the entry criteria, additional sibs meeting the same criteria (either hypertensive or hypotensive) were also recruited. In some families, sib recruitment is not yet complete, and/or some sibs were found upon examination not to meet criteria for hypertension or hypotension but still had blood drawn. However, most families have at least two hypertensive and/or hypotensive sibs.

Subjects without completed data were excluded from our analysis. In total, 431, families were phenotyped, of which 316 are Chinese and 115 Japanese. These included 116 families with 1 sib, 150 with 2 sibs, 67 with 3, 44 with 4, 17 with 5, 17 with 6, 14 with 7, 3 with 8, and 3 with 10 sibs, or a total of 1132 subjects.
3. METHODS

3.1. Measures and Pre-processing Data

Patients completed a survey on medical history and submitted bodily fluids for laboratory analysis. The following variables were extracted: Age, Sex, Waist-to-Hip Ratio, Body Mass Index, Triglycerides, Total Cholesterol, High-Density Lipoproteins, OGTT Glucose at -10, OGTT Glucose at 60, OGTT Glucose at 120, OGTT Insulin at -10, OGTT Insulin at 60, and OGTT Insulin at 120.

Informal review by colleagues has alerted us to possible criticism that perhaps everything we study is mediated by how obese and how old our study subjects are. To address this possibility, we first regressed each variable on a quadratic function of BMI and AGE \((y = a_0 + a_1BMI + a_2AGE + a_3(BMI \cdot AGE) + a_4BMI^2 + a_5AGE^2 + \epsilon)\) and used the resulting residuals as raw data for subsequent analyses. All regressions except that for cholesterol for men are highly significant by conventional Gaussian-based testing. However, these residuals will be shown not to be Gaussian (Figures 1 and 2); and this brings us to our next topic.

3.2. Cluster Analysis

We argue elsewhere that we seek intermediate phenotypes on the basis of measurements that quantify plasma lipids, insulin, and glucose. There are 10 such variables in all; so reducing, or summarizing, them is imperative. One conventional approach is by principal components [Seber:1984], or more generally factor analysis [Seber:1984], methods for which inferences tend to be based on Gaussian assumptions that here are untenable. Therefore, we turn to something less parametric: cluster analysis. The idea is that subgroups defined by the clusters can vary in terms of their associations with hypertension, thereby enabling us to narrow the focus of our search for predisposing genes.

The variables that underwent clustering are residuals that were standardized by subtracting their individual means and dividing by respective marginal standard deviations. This renders any conclusions unchanged by, that is to say equivariant to, scaling of data by addition and subsequent multiplication. For reasons argued elsewhere [Lenert:1999, Lin:1999, Sugar:1998] we prefer \(k\)-means [Gersho:1992] to agglomerative [Kaufman:1990] clustering. In the engineering context of lossy data compression, \(k\)-means clustering is called vector quantization. The remaining conventional approach to clustering, that of mixture models, is implicit in \(k\)-means, which does not require anything like the usual assumption of Gaussian components to the mixture.

For given \(k\), by \(k\)-means clustering one chooses \(k\) cluster centers, which are called codewords in engineering, \(\{c_1, \ldots, c_k\}\), so as to minimize the average squared distance from a randomly chosen observation to its nearest cluster center [Gersho:1992]. Average squared distance is called distortion. The feature space of variables, here 10-dimensional, is thus partitioned into what are termed Voronoi regions, that are defined by the nearest neighbor criterion. The cluster center of lowest index is used to break ties. Implementation of \(k\)-means clustering is tricky, though there is a standard algorithm due to S. Lloyd [Lloyd:1982]. The Lloyd algorithm begins with arbitrary choice of \(k\) cluster centers. The defining minimization property means that each center should be a centroid of its respective Voronoi region. The algorithm works by alternating centroids and regions. Because of the convexity of squared distance, the algorithm is a descent algorithm and converges so long as the expected squared norm of an observation is finite [Gersho:1992]. (Obviously this is not a problem with applications.) However, the minimizing cluster centers can depend on the initialization. Therefore, it is necessary to try many choices so as to ensure that a
Phenotypes

global minimum has been achieved. Furthermore, any positive definite distance matrix will do, though the added generality was irrelevant here.

To select the optimal number of clusters based on the data, we plotted the ten-fold cross-validated [Breiman:1993] distortion as a function of the number of clusters (distortion curve — Figures 3 and 4) to identify where adding another cluster had an “insignificant” effect on distortion. Although distortion is always reduced by increasing the number of clusters, at some point the reduction in distortion by adding another cluster becomes so small as to be meaningless in terms of the subject matter. To find this point, Sugar et al. [Sugar:1998] suggest a broken-line regression approach, and Lin and Olshen [Lin:1998] suggest finding the point on the graph of distortion versus number of clusters with the greatest curvature. In yet another approach [Sugar:1999] motivated by Shannon’s rate-distortion theory, distortion to the power \(-d/2\) is plotted against number of clusters, and the point with largest slope is chosen. Here \(d\) is the number of variables upon which the clustering is based. There are many other approaches to choosing the number of clusters. See especially [Gordon:1999] and [Tibshirani:2000]. In many applications, all methods yield the same answer; however, the process of choosing the “right” number of clusters remains to some extent art. One can validate choice of number of clusters graphically by “bootstrapping” [Efron:1993] the distortion curve. What we mean here is that a sample the size of our learning sample (1132) is chosen with replacement from the overall sample. The sampling is constrained so that each sample contains the same number of men, women, hypertensives, and non-hypertensives as does the original sample. For each sample graph distortion against number of clusters, and superimpose the graphs. The visual effect of doing this 100 times is powerful evidence. Cluster analyses were performed using S-Plus (Version 3.4 Release 1, MathSoft, Inc., Seattle, WA).

3.3. Testing Hypotheses

We argue that many variables we measure, despite being continuous, are not jointly or marginally Gaussian, unconditionally, or conditional upon cluster membership. These variables are driven partly by genes, and genotypes are discrete. Unconditionally or even separately by cluster membership, the distribution is a mixture of different distributions, one component per predisposing genotype. Even if the components are Gaussian, what we measure, the mixture, is not. There is only a limited range of values each variable can assume, making Gaussian assumptions even less tenable. To avoid Gaussian assumptions we employ permutation testing whenever possible.

Though the contexts in which we do permutation testing vary, the principle is always the same. We find a criterion that is “small” (respectively “large”) if a putative null hypothesis is “true,” increases (decreases) as one moves “away” from the null, and is sensitive to departures from the null hypothesis. We permute the data, always respecting family structures, and compute the criterion for each permutation. If there are \(N\) permutations in all, plus, of course, the identity permutation (the original data), and the criterion for the identity permutation has criterion that is \(k\)th largest (smallest) among the \((N+1)\) numbers, then the “p-value” of the test is \([k/(N+1)]\).

The testing has a logical order. First, there is the issue of clusters. Are they familial? Next, are the variables upon which we cluster familial? Last, is there an association between clusters and hypertension? The last two questions are addressed with criteria that are familiar: a components of variance model for one and a chi-square-like test of association for the other. However, the first question is addressed by an unusual criterion, the actual probability of observing family distributions within and across clusters. Furthermore, in all else that we do inferences for men and women are separated. Not so for testing the familiality of the clusters. For these reasons we
relegate testing familiality of the clusters to an appendix.

3.4. Testing Familiality of Variables

If clusters are familial, based as they are upon combinations of cited variables, then so too should be the individual variables. So we are interested in how familial are the variables upon which we clustered. We propose the following random effects model [Scheffe:1959] for the jth individual in the ith family:

\[ y_{ij} = \mu + a_i + e_{ij}; \]
\[ \{a_i, e_{ij}\} \text{ are independent}; \]
\[ \{a_i\} \text{ are identically distributed with } E[a_i] = 0 \text{ and } \text{Var}[a_i] = \sigma^2_a; \]
\[ \{e_{ij}\} \text{ are identically distributed with } E[e_{ij}] = 0 \text{ and } \text{Var}[e_{ij}] = \sigma^2_e. \]  

(1)

To test whether a variable is familial is equivalent to testing the following null hypothesis:

\[ H_0: \text{Variable not familial, } \sigma^2_a = 0. \]

We tested this hypothesis separately for men and women by the permutation distribution of:

\[ \hat{\sigma}^2_e = \frac{1}{\sum_{i=1}^{f} n_i} \sum_{i=1}^{f} n_i \hat{\sigma}^2_{e,i}, \]

where \( \hat{\sigma}^2_{e,i} \) = estimator of \( \sigma^2_e \) computed for the ith family,

\[ = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2; \]

\( n_i \) = number of siblings in the ith family;
\( f \) = the number of families;
\( \bar{y}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} y_{ij} \).  

(2)

\[ \hat{\sigma}^2_{e,i} \] is an unbiased estimator of \( \sigma^2_e \).

\[ \hat{\sigma}^2_{e,i} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \]

\[ = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} [(\mu + a_i + e_{ij}) - (\mu + a_i + \bar{e}_i)]^2 \]

\[ = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} [e_{ij} - \bar{e}_i]^2, \]  

(4)

from which it follows from elementary considerations that
\[ E[\hat{\sigma}_e^2] = \sigma_e^2. \]  

Let
\[ \sigma_e^2 = \frac{1}{\sum_{i=1}^{f} n_i} \sum_{i=1}^{f} n_i \hat{\sigma}_{e,i}^2, \]  

an unbiased estimator of \( \sigma_e^2 \). When sibs are permuted and assigned randomly to families, then if the permutation is \( \pi^{-1} \), the permuted value of \( \hat{\sigma}_{e,i}^2 \), call it \( \hat{\sigma}_{e,i}^2 \), is
\[ \frac{1}{n_i-1} \sum_{j=1}^{n_i} \left( (\mu + a_{\pi(j)/i} + \varepsilon_{\pi(j)/i}) - (\mu + \frac{1}{n_i-1} \sum_{k=1}^{n_i} a_{\pi(k)/i} + \bar{\varepsilon}_i) \right)^2, \]  

where under \( \pi^{-1} \), the \( j \)th sib in the \( i \)th family is denoted by \( \pi(j)/i \). The displayed sum is
\[ \frac{1}{n_i-1} \sum_{j=1}^{n_i} \left[ \left( \varepsilon_{\pi(j)/i} - \bar{\varepsilon}_i \right)^2 + \left( \varepsilon_{\pi(j)/i} - \bar{\varepsilon}_i \right) \left( a_{\pi(j)/i} - \frac{1}{n_i-1} \sum_{k=1}^{n_i} a_{\pi(k)/i} \right) \right]. \]  

It follows from our assumptions that the second term has expectation 0. If the \( i \)th family is mapped to itself by \( \pi^{-1} \), then
\[ E[\hat{\sigma}_{e,i}^2 | \pi] = \sigma_e^2 = E[\hat{\sigma}_{e,i}^2]. \]  

Otherwise, the third of these three preceding terms has strictly positive conditional expectation unless \( \sigma_A^2 = 0 \). By averaging \( E[\hat{\sigma}_{e,i}^2 | \pi] \) over permutations, one concludes that either \( \sigma_A^2 = 0 \), or \( E[\hat{\sigma}_e^2] > E[\hat{\sigma}_e^2] \). Hence, the null hypothesis implies that at least in expectation it is smallest among permutation values for the identity permutation.

We tested this hypothesis separately for men and women for all variables by the permutation distribution with the following algorithm:

1. \( \{y_{ij}\} \) = data. Take data and calculate the empirical estimate of \( \sigma_e^2 \). Let \( \hat{\sigma}_e^2 \) = the empirical estimate for the unpermuted data.

2. For \( t = 1 \) to \( \text{trial} \), permute individuals randomly and assign to families. Call the permuted data for the \( t \)th trial, \( \{\hat{y}_{ij}\}_{i=1}^{t} \). For this data set, calculate the empirical estimate of \( \sigma_e^2 \). Let \( \hat{\sigma}_e^2_{i=1} = \) the empirical estimate for the permuted data on the \( t \)th trial.

3. \( p \)-value for the familiality of the variable = \( \frac{\text{rank} \: \hat{\sigma}_e^2 \: \text{among} \: \{\hat{\sigma}_e^2\}}{\text{trial} + 1} \).
3.5. Testing for Association Between Cluster Membership and Hypertensive Status

We are also interested in the differences in percentage hypertensive status by cluster. But for dependencies that owe to family structure a simple chi-squared test for association would suffice. To get around the dependencies and yet have a valid test, we need to condition on the family structure of each cluster. This can be done by suitable permutation testing. For example, if Family A has 3 sibs in Cluster 1 and 5 in Cluster 2, we need to permute the data in a way that preserves the same distribution of sibs in clusters. To respect the family structures in looking for associations between cluster membership and hypertensive status, we permute hypertensive status within families.

To be precise, we tested for the difference of hypertensive status between clusters separately for men and women with the following algorithm:

1. For our original data, calculate the usual $2 \times 2$ chi-squared statistic $\hat{X}^2$ for the unpermuted data.

2. For $t = 1$ to trial, for each family, permute the assignment of the values of hypertensive status within families keeping the number of sibs for each family in each cluster the same as the original data. For each iteration, calculate the chi-squared statistic. Let $\hat{X}^2|_{t=i}$ = the empirical estimate for the permuted data on the $i$th trial.

3. p-value for the familiarity of the variable = 1 - [rank $\hat{X}^2$ among $\{\hat{X}^2\}]/[$trial + 1$].

We noticed after our analysis that the results for men differ from the results for women. Many studies have shown an association between menopausal status and hypertensive status. Thus, we looked at the association between cluster membership and hypertensive status separately for women who have not reached menopause and women who have reached menopause to explore if data for women who have reached menopause are like those for men.
4. RESULTS

Figures 1 and 2 show, for women and men respectively, scatter plots of the various variables of lipid measurements that entered our analysis. The plots suggest that these variables are not Gaussian, marginally or jointly and thus Gaussian-based random effects analysis maybe suspect. $k$-means clustering was used to separate the data because this algorithm does not assume normality of the data.

Figures 3 and 4 show the distortion plot for women and men respectively. The curvature test suggests that in both cases, there are two clusters. Results of bootstrapping the distortion curve are depicted in Figures 5 and 6 respectively, for women and men. In women, there is an obvious kink in the bootstrap distortion curves at two clusters. In men, the results are less obvious; however, the best answer is still two clusters.

Tables 1 and 2 show the mean and standard deviations of the variables upon which clustering was based, along with age and hypertensive status, for women and men respectively. In both tables, we have arranged the naming of the clusters so that Cluster 2 contains subjects that are older and have a higher incidence of hypertension. For women, Cluster 1 contains 64.3% and Cluster 2 contains 35.7% of the population. For men, Cluster 1 contains 59.4% and Cluster 2 contains 40.6% of the population. For both men and women, the larger cluster contains individuals who are hypotensive and hypertensive, the other appearing clearly to be those who seem to be insulin resistant, or in a few cases frankly diabetic. Levels of hypertension are higher in the insulin resistant cluster, for both men and women.

Permutation testing was employed to decide if the clusters are familial, and also if the variables by which clusters were defined are familial. Results show that both clusters are significantly familial (p-value less than .0001). Ten variables that define the clusters were studied as to how familial they are, separately for men and women. Results are shown in Table 3. For women, all variables were significant at most at the 5% level. For men, all variables were significant at least at the 5% level, except for total cholesterol, OGTT glucose at one hour, OGTT insulin at
Fig. 3. Ten-fold cross-validated distortion versus numbers of clusters for females.

Fig. 4. Ten-fold cross-validated distortion versus numbers of clusters for males.
Fig. 5. One hundred bootstrapped, ten-fold cross-validated distortion curves versus numbers of clusters for females. Notice the “kink” at two clusters.

Fig. 6. One hundred bootstrapped, ten-fold cross-validated distortion curves versus numbers of clusters for males. The “kink” at two clusters is noticeable but less obvious than for females.
Table 1. Means and standard deviations of variables that summarize physical status, glucose, and insulin, overall and by cluster membership, for females.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All Females</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std.</td>
<td>Mean</td>
</tr>
<tr>
<td>No. Patients in Cohort</td>
<td>597</td>
<td></td>
<td>384</td>
</tr>
<tr>
<td>Physical Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>51.58</td>
<td>8.49</td>
<td>51.51</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>24.87</td>
<td>3.60</td>
<td>24.72</td>
</tr>
<tr>
<td>Hypertensive Status</td>
<td>0.70</td>
<td>0.21</td>
<td>0.64</td>
</tr>
<tr>
<td>Waist-Hip Ratio*</td>
<td>0.85</td>
<td>0.08</td>
<td>0.84</td>
</tr>
<tr>
<td>Cholesterol*</td>
<td>196.50</td>
<td>39.56</td>
<td>191.19</td>
</tr>
<tr>
<td>Triglycerides*</td>
<td>121.41</td>
<td>66.80</td>
<td>102.62</td>
</tr>
<tr>
<td>High Density Lipoproteins*</td>
<td>48.30</td>
<td>12.39</td>
<td>51.13</td>
</tr>
<tr>
<td>OGGT Glucose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At -10*</td>
<td>90.49</td>
<td>10.79</td>
<td>87.56</td>
</tr>
<tr>
<td>At 60*</td>
<td>168.65</td>
<td>44.32</td>
<td>149.84</td>
</tr>
<tr>
<td>At 120*</td>
<td>140.23</td>
<td>41.70</td>
<td>123.98</td>
</tr>
<tr>
<td>OGGT Insulin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At -10*</td>
<td>7.10</td>
<td>4.62</td>
<td>5.37</td>
</tr>
<tr>
<td>At 60*</td>
<td>73.87</td>
<td>60.65</td>
<td>57.14</td>
</tr>
<tr>
<td>At 120*</td>
<td>70.09</td>
<td>58.15</td>
<td>45.14</td>
</tr>
</tbody>
</table>

* Variables upon which clustering was based.

baseline and at one hour. For women, hypertensive status is also significantly different between clusters with p-value = 0.0015, but for men, the signal is less clear with p-value = 0.0939. For women who have not reached menopause, hypertensive status is significantly different between clusters with p-value = 0.0035, but women who have reached menopause, the signal, like for men, is not significant with p-value = 0.2408.

5. DISCUSSION

Elevated blood pressure is a major risk factor for human cardiovascular disease. Major research efforts around the world are focused on the identification of factors that contribute to blood pressure variation in an effort to decrease morbidity and mortality. This research has implicated a wide variety of genetic and environmental factors that contribute to variation in human blood pressure [Lifton:1996]. The results of "whole-genome" linkage analyses suggest that no single gene is likely to account for more than a small percentage of blood pressure variation. Recently, several genes accounting for a small fraction of the variance of human blood pressure have been identified by our group and others. However, even using the increased power of genetic association analysis as compared to genetic linkage analysis [Risch:1996], identification of such genes remains
Table 2. Means and standard deviations of variables that summarize physical status, glucose, and insulin, overall and by cluster membership, for males.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All Males</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Patients in Cohort</td>
<td>535</td>
<td>318</td>
<td>217</td>
</tr>
<tr>
<td>Physical Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>50.75</td>
<td>50.59</td>
<td>50.97</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>26.00</td>
<td>25.89</td>
<td>26.16</td>
</tr>
<tr>
<td>Hypertensive Status</td>
<td>0.77</td>
<td>0.74</td>
<td>0.81</td>
</tr>
<tr>
<td>Waist-Hip Ratio*</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Cholesterol*</td>
<td>189.19</td>
<td>186.28</td>
<td>193.46</td>
</tr>
<tr>
<td>Triglycerides*</td>
<td>143.35</td>
<td>122.66</td>
<td>173.68</td>
</tr>
<tr>
<td>High Density Lipoproteins*</td>
<td>39.65</td>
<td>41.83</td>
<td>36.47</td>
</tr>
<tr>
<td>OGTT Glucose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At -10*</td>
<td>93.05</td>
<td>90.11</td>
<td>97.34</td>
</tr>
<tr>
<td>At 60*</td>
<td>175.93</td>
<td>154.07</td>
<td>207.98</td>
</tr>
<tr>
<td>At 120*</td>
<td>133.65</td>
<td>111.32</td>
<td>166.38</td>
</tr>
<tr>
<td>OGTT Insulin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At -10*</td>
<td>7.57</td>
<td>5.98</td>
<td>9.98</td>
</tr>
<tr>
<td>At 60*</td>
<td>77.91</td>
<td>60.07</td>
<td>104.04</td>
</tr>
<tr>
<td>At 120*</td>
<td>64.48</td>
<td>36.64</td>
<td>105.28</td>
</tr>
</tbody>
</table>

* Variables upon which clustering was based.

a difficult task, requiring large sample sizes.

One approach to increase the power to identify genes contributing to human blood pressure variation is to define intermediate phenotypes that are associated with hypertension and that may be more impacted by single genes than hypertension itself. Numerous epidemiological studies have identified two major intermediate phenotypes associated with both hypertension and an increased risk of cardiovascular morbidity and mortality. The first is an abnormal plasma lipid profile of elevated triglycerides and decreased high-density lipoproteins; the second is an abnormality of glucose utilization, distinct from diabetes, manifesting as elevated fasting plasma insulin and an abnormal response to an oral glucose challenge [Chen:1997, Austin:1990]. Although there is evidence that genetic factors impact these intermediate phenotypes, the contribution of any single gene is unclear [Neel:1998]. Furthermore, the interaction of these phenotypes with each other and with hypertension is complex [Neel:1998]. Given this complexity, we have used k-means clustering to define intermediate phenotypes in siblings ascertained on the basis of either high or low blood pressure, that makes no prior assumptions regarding the interactions among the abnormal lipid profile, the abnormal glucose utilization, and hypertension.

Using ten different variables for clustering, we found that these variables aggregated into two distinct clusters in both males and females. The significance of these clusters was validated by graphical bootstrapping of the cross-validated distortion curve. In both sexes, Cluster 2 consisted
Table 3. Attained significance of permutation tests with 100 iterations for the null hypothesis of “no familiality,” for variables upon which clustering was based, separately for females and males.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All Females</th>
<th>All Males</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-values</td>
<td>P-values</td>
</tr>
<tr>
<td>No. of Patients in Cohort</td>
<td>597</td>
<td>535</td>
</tr>
<tr>
<td>Physical Staus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waist-Hip Ratio</td>
<td>0.0099</td>
<td>0.0099</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>0.0099</td>
<td>0.1584</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>0.0099</td>
<td>0.0297</td>
</tr>
<tr>
<td>High Density Lipoproteins</td>
<td>0.0099</td>
<td>0.0099</td>
</tr>
<tr>
<td>OGTT Glucose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At -10</td>
<td>0.0099</td>
<td>0.0099</td>
</tr>
<tr>
<td>At 60</td>
<td>0.0099</td>
<td>0.2673</td>
</tr>
<tr>
<td>At 120</td>
<td>0.0099</td>
<td>0.0198</td>
</tr>
<tr>
<td>OGTT Insulin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At -10</td>
<td>0.0099</td>
<td>0.3366</td>
</tr>
<tr>
<td>At 60</td>
<td>0.0099</td>
<td>0.0891</td>
</tr>
<tr>
<td>At 120</td>
<td>0.0099</td>
<td>0.0099</td>
</tr>
</tbody>
</table>

of individuals with significantly altered plasma lipid profiles and abnormal glucose utilization, as compared to individuals in Cluster 1, suggesting that these intermediate phenotypes are significantly associated in our study population. Numerous epidemiological studies demonstrating significant association of the abnormal lipid and glucose utilization phenotypes in many but not all populations are consistent with our clustering results [Neel:1998]. However, our finding that females, but not males, have significantly increased blood pressure in Cluster 2 compared to Cluster 1 is novel. Interestingly, all of this effect can be attributed to premenopausal women in Cluster 2 (see Results). The fact that younger men show no such effect strongly suggests a hormonal basis for the association of elevated blood pressure with the abnormal lipid profile and abnormal glucose utilization in premenopausal women in Cluster 2. Our results, coupled with the recent finding that estrogen supplementation does not significantly reduce the risk of adverse cardiovascular outcomes in postmenopausal women, suggest a more complex relationship between hormonal status and cardiovascular risk than was appreciated previously [Nabulsi:1993].

Our data suggest that stratification based on cluster, sex, and menopausal status may provide a useful approach for identifying novel genes that play a role in blood pressure variation. Further studies analyzing the association of genetic variation in cardiovascular candidate genes with our study population will test this hypothesis.
6. ACKNOWLEDGEMENTS

We thank patients for participating in this study. We also thank Stephen Mockrin and Susan Old of the National Heart Lung and Blood Institute and other members of the SAPPHIRE project (listed on the attached page) for their help.

This work was supported by grants U01 HL54527-0151 from the National Heart Lung and Blood Institute, NIH (USA); NSC-85-2331-B002-350Y, NSC-86-2314-B002-345Y, and NSC-87-2312-B002-021Y from the National Science Council, Taiwan, ROC; and NIH grant 2 R01 CA59039.

REFERENCES


APPENDIX

Testing Familiality of Clusters

It would be consistent with our clusters being genuine intermediate phenotypes were they familial. Family homogeneity in a cluster can be used as a proxy for how familial a cluster is. The question is if we were to allocate siblings at random to the clusters, how unlikely would it be to observe homogeneity to the extent we saw or more so than what we found in our data. We tested this hypothesis by combining men and women.

Suppose we have \( k \) clusters; there are \( f \) families; the \( i^{th} \) family has \( n_i \) siblings; and the \( j^{th} \) cluster has \( h_{ij} \) sibs from the \( i^{th} \) family. Then, the probability of what we observe in the \( j^{th} \) cluster is:

\[
\frac{\prod_{i=1}^{f} \left( \frac{n_i}{n_i} \right)}{\left( \sum_{i=1}^{f} n_i \right) \left( \sum_{i=1}^{f} h_{ij} \right)}
\]

Therefore, the exact p-value is:

\[
\sum_{j=1}^{h_{ij}} \sum_{i=1}^{f} \prod_{i=1}^{f} \left( \frac{n_i}{n_i} \right) \leq \frac{\prod_{i=1}^{f} \left( \frac{n_i}{n_i} \right)}{\left( \sum_{i=1}^{f} n_i \right) \left( \sum_{i=1}^{f} h_{ij} \right)}
\]

We submit that finding the exact p-value is not computationally feasible for our problem. We can estimate the p-value using permutation testing [Efron:1993] by using the following algorithm:
1. For our original data, calculate the probability of observing the distribution of sibs among families for each cluster configuration as

\[
\frac{\prod_{i=1}^{f} \binom{n_i}{h_{ij}}}{\sum_{i=1}^{f} n_i \choose h_{ij}}
\]

2. For \( t = 1 \) to \( trial \), assign individuals randomly to clusters keeping the size of the clusters the same. For each iteration, calculate the probability of observing this configuration for each cluster.

3. The p-value for the familiality of a cluster is the sum of all the probabilities that we calculated from the iterations that are less than or equal to that of our observed data divided by the sum of all the probabilities that we calculated from the iterations for that cluster.

One might ask why we combined women and men to test for the familiality of the clusters even though other comparisons were done separately by gender. For one, each gender's data separately led to exactly two clusters. In both cases one was clearly the insulin resistant cluster and the other not. In both cases hypertension was more prevalent in the insulin resistant cluster. Thus, it was easy to assign each member of a sibship to being in the insulin resistant group or not. Obviously families can be of mixed gender, and unless insulin resistant status is patently sex-linked (for which we do not have any information), it seemed better to us to preserve the full family structure than otherwise. Had we done the computation of familiality by cluster separately by gender, there would have been many sibships with only one sib, thereby reducing power.