Original Paper

Human Heredity

Hum Hered 2009;68:182–191
DOI: 10.1159/000224638

Received: December 11, 2008
Accepted after revision: February 12, 2009
Published online: June 11, 2009

Using Allele Sharing Distance for Detecting Human Population Stratification

Xiaoyi Gao[a] Eden R. Martin[b]

[a]Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Mo.,
[b]Center for Genetic Epidemiology and Statistical Genetics, Miami Institute for Human Genomics,
University of Miami Miller School of Medicine, Miami, Fla., USA

Key Words

Population stratification · Coancestry coefficient · Single nucleotide polymorphism · Allele sharing distance

Abstract

There is a long history of using allele sharing distance (ASD) and closely related metrics for population stratification analysis. However, the theory for this practical usage has not been reported. In this paper, we describe the theoretical background for using ASD on single nucleotide polymorphism (SNP) genetic data for human population stratification analysis. In showing the proof, we lay the ground work for a general distance-based method for classifying subpopulations using SNPs and ASD. We also show its relation to other closely related distance metrics using HapMap Phase I SNP data.

Introduction

With the advent of rapid single nucleotide polymorphism (SNP) genotyping and the availability of genome-wide SNPs there is a growing need for population stratification detection methods that can take advantage of large datasets and eliminate dependency on model parameter estimations. One particularly appealing approach is using the allele sharing distance (ASD) [1], or the similar metric identity by state (IBS) [2]. It does not require allele/genotype frequency estimation, which makes it still valid when sample sizes are small. It is also suitable for population outlier detection, is robust to high linkage disequilibrium (LD) among SNPs, and can be rapidly calculated.

There is a long history of using ASD and closely related metrics (either proportion of allele sharing or IBS) for population stratification analysis [1–8]. Despite the difference in the clustering approaches employed, the pair-wise ASD distance matrix is nearly the same for all the distance-based clustering methods. It seems that the pair-wise ASD distance matrix contains sufficient information for separation and is a basis for many distance-based clustering approaches. However, to date there has not been a rigorous theoretical foundation for using ASD for human population classification.

A.W.F. Edwards [9] showed that different haploid populations can be distinguished through the half matrix of pair-wise distances between individuals, but it is not clear how the rational applies to dipoid populations. For haploid populations, at a single biallelic locus there are only two possible categories, and they can be modeled by a binomial distribution [9]. However, human populations are diploid. If we consider biallelic SNPs, then there are three possible categories for the genotypes at each locus, and a simple binomial distribution no longer applies. Does Edwards's rational still apply to diploid populations? If so,

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2009 S. Karger AG, Basel
0001–5652/09/0683–0182$26.00/0

Accessible online at:
www.karger.com/hhe

Xiaoyi Gao, PhD
Division of Statistical Genomics
Washington University School of Medicine
St. Louis, MO 63108 (USA)
Tel. +1 314 362 3593, Fax +1 314 362 4227, E-Mail ray.x.gao@gmail.com

then human population stratification analysis can be greatly simplified since we can rely on the pair-wise distance matrix for the classification.

The work in this paper extends Edwards's idea [9] to diploid populations using a coancestry approach and lays the ground work for a general distance-based method for classifying subpopulations using SNPs and ASD. In this work, we derive the statistical background for using ASD on SNPs for population stratification detection using standard population genetics theory. We show that large numbers of random genome-wide SNPs provide sufficient information for accurate clustering. Diploid individuals from different subpopulations can be separated using the pair-wise distance matrix without estimation of allele frequencies.

Theory

Without loss of generality, we consider two subpopulations descended from a single reference ancestral population. Random mating is assumed within each subpopulation. The subpopulations have the same expected allele frequencies in the absence of disturbing forces, e.g. unequal selection in different subpopulations [10]. We denote θ as the probability that two alleles at a locus in the same subpopulation are identical by descent (IBD). Therefore, different alleles within a subpopulation are related to a magnitude θ. θ is also sometimes referred to as F_{ST} [11]. We assume that subpopulations have reached equilibrium so that θ is not changing over time [12].

Throughout the paper we consider only biallelic SNPs. We denote the population frequencies of alleles A_1 and A_2 at a locus as p_1 and p_2, respectively, and their values may be unknown in practice. The probability that an allele drawn from the subpopulation is of type A_1 is $P(A_1) = p_1$ and similarly $P(A_2) = p_2$ for A_2 and $p_1 + p_2 = 1$. Given that the coancestry coefficient between alleles at a locus is θ, the probability of observing two alleles both of type A_1 is $P(A_1A_1) = \theta p_1 + (1 - \theta)p_1^2$, and similarly $P(A_1A_2) = 2(1 - \theta)p_1p_2$. If $P(A_1^m A_2^n)$ is the probability of observing m copies of A_1 alleles and n copies of A_2 alleles in a random drawing of $m + n$ alleles, then it can be shown [13, 14] that

$$P\left(A_1^{m+1}A_2^n\right) = P\left(A_1^m A_2^n\right)\frac{m\theta + p_1\left(1-\theta\right)}{1+\left(m+n-1\right)\theta}. \tag{1}$$

The above formulas can also be derived through a Dirichlet approach [14–16] by realizing that the probability that the next allele sampled is of type A_i given that m_i copies of A_i have already been found in a set of m alleles is,

$$Pr\left(A_i | m_i \text{ copies of } A_i\right) = \frac{m_i\theta + \left(1-\theta\right)p_i}{1+\left(m-1\right)\theta}, \tag{2}$$

where $\{p_i\}$ are the population allele frequencies.

In randomly mating subpopulations, DNA profiles are essentially samples of alleles [14, 15]. Therefore, the joint genotype probabilities for biallelic SNP markers for randomly drawn members of the same subpopulation with evolutionary relatedness can be obtained through successive implementation of equations (1) or (2) and are represented in the following series of equations:

$$P\left(A_1A_1, A_1A_1\right) = \frac{\left[3\theta + \left(1-\theta\right)p_1\right]\left[2\theta + \left(1-\theta\right)p_1\right]\left[\theta + \left(1-\theta\right)p_1\right]p_1}{\left(1+\theta\right)\left(1+2\theta\right)}, \tag{3}$$

$$P\left(A_2A_2, A_2A_2\right) = \frac{\left[3\theta + \left(1-\theta\right)p_2\right]\left[2\theta + \left(1-\theta\right)p_2\right]\left[\theta + \left(1-\theta\right)p_2\right]p_2}{\left(1+\theta\right)\left(1+2\theta\right)}, \tag{4}$$

$$P\left(A_1A_2, A_1A_2\right) = \frac{4\left(1-\theta\right)\left[\theta + \left(1-\theta\right)p_1\right]\left[\theta + \left(1-\theta\right)p_2\right]p_1p_2}{\left(1+\theta\right)\left(1+2\theta\right)}, \tag{5}$$

$$P\left(A_1A_1, A_1A_2\right) = \frac{4\left(1-\theta\right)\left[2\theta + \left(1-\theta\right)p_1\right]\left[\theta + \left(1-\theta\right)p_1\right]p_1p_2}{\left(1+\theta\right)\left(1+2\theta\right)}, \tag{6}$$

$$P\left(A_2A_2, A_1A_2\right) = \frac{4\left(1-\theta\right)\left[2\theta + \left(1-\theta\right)p_2\right]\left[\theta + \left(1-\theta\right)p_2\right]p_1p_2}{\left(1+\theta\right)\left(1+2\theta\right)}, \tag{7}$$

$$P\left(A_1A_1, A_2A_2\right) = \frac{2\left(1-\theta\right)\left[\theta + \left(1-\theta\right)p_1\right]\left[\theta + \left(1-\theta\right)p_2\right]p_1p_2}{\left(1+\theta\right)\left(1+2\theta\right)}, \tag{8}$$

where $p_2 = 1 - p_1$. The equations above already account for the different possible ordering of the genotypes and the two alleles within each genotype.

In order to quantify the distance among individuals and construct the pair-wise distance matrix, we need a distance measure. ASD, and closely related distance measures, are popular metrics in distance-based population structure analysis with a long history [1, 3–8]. In this work, we chose to use ASD to construct the distance matrix between all pairs of individuals.

The ASD distance between individuals i and j is defined as

$$D = \frac{1}{L}\sum_{l=1}^{L} d_l, \tag{9}$$

where

$$d_l = \begin{cases} 0, & \text{if individual } i \text{ and } j \text{ have two alleles in common at the } l\text{-th locus,} \\ 1, & \text{if individual } i \text{ and } j \text{ have only a single allele in common at the } l\text{-th locus,} \\ 2, & \text{if individual } i \text{ and } j \text{ have no allele in common at the } l\text{-th locus,} \end{cases} \tag{10}$$

and L is the number of SNP loci used.

At a SNP locus, there are nine possible genotype combinations between individuals i and j. Using the following notation, a genotype before a comma is from person i and a genotype after a comma is from person j, there are three situations where $d = 0$, i.e., (A_1A_1, A_1A_1), (A_1A_2, A_1A_2) and (A_2A_2, A_2A_2); four where $d = 1$, i.e., (A_1A_1, A_1A_2), (A_2A_2, A_1A_2), (A_1A_2, A_1A_1) and (A_1A_2, A_2A_2); and two where $d = 2$, i.e., (A_1A_1, A_2A_2) and (A_2A_2, A_1A_1). For a given θ, we denote d as d^θ. Therefore, the distribution of d^θ at a locus is

$$d^\theta = \begin{cases} 0, & \dfrac{\left[3\theta+(1-\theta)p_1\right]\left[2\theta+(1-\theta)p_1\right]\left[\theta+(1-\theta)p_1\right]p_1}{(1+\theta)(1+2\theta)} \\ & +\dfrac{\left[3\theta+(1-\theta)p_2\right]\left[2\theta+(1-\theta)p_2\right]\left[\theta+(1-\theta)p_2\right]p_2}{(1+\theta)(1+2\theta)} \\ & +\dfrac{4(1-\theta)\left[\theta+(1-\theta)p_1\right]\left[\theta+(1-\theta)p_2\right]p_1p_2}{(1+\theta)(1+2\theta)} \\ 1, & \dfrac{4(1-\theta)\left[2\theta+(1-\theta)p_1\right]\left[\theta+(1-\theta)p_1\right]p_1p_2}{(1+\theta)(1+2\theta)} \\ & +\dfrac{4(1-\theta)\left[2\theta+(1-\theta)p_2\right]\left[\theta+(1-\theta)p_2\right]p_1p_2}{(1+\theta)(1+2\theta)} \\ 2, & \dfrac{2(1-\theta)\left[\theta+(1-\theta)p_1\right]\left[\theta+(1-\theta)p_2\right]p_1p_2}{(1+\theta)(1+2\theta)} \end{cases} \tag{11}$$

After slightly tedious calculations, the expected value of d^θ is

$$E\left(d^\theta\right) = \frac{4(1-\theta)p_1p_2}{(1+\theta)(1+2\theta)}\left[2\theta^2+2\theta+1-(1-\theta)^2p_1p_2\right], \tag{12}$$

and the variance of d^θ is

$$Var\left(d^\theta\right) = 4(1-\theta)p_1p_2\left\{1-\frac{4(1-\theta)p_1p_2\left[2\theta^2+2\theta+1-(1-\theta)^2p_1p_2\right]^2}{(1+\theta)^2(1+2\theta)^2}\right\}. \tag{13}$$

With $E(d^\theta)$ and $Var(d^\theta)$ derived, we can now examine the effects of different values for θ. In the simplest case the subpopulations have been isolated since the ancestral population and alleles in each subpopulation are not related to those in the other. Therefore, alleles between subpopulations are independent ($\theta = 0$) [10]. Then

$$E(d^0) = 4p_1p_2(1 - p_1p_2), \tag{14}$$

and

$$Var(d^0) = 4p_1p_2[1 - 4p_1p_2(1 - p_1p_2)^2]. \tag{15}$$

The expected allele sharing distance difference between individuals from different subpopulations and individuals from the same subpopulation is

$$E\left(d^0\right)-E\left(d^\theta\right) = \frac{4\theta p_1p_2\left\{\theta^2\left[2-p_1p_2\right]+\theta\left[2+p_1p_2\right]+2\left[1-3p_1p_2\right]\right\}}{(1+\theta)(1+2\theta)}, \tag{16}$$

which is >0 ∀ $p_1 \in (0, 1)$ since $p_1p_2 \le 0.25$.

It is also possible that alleles in different subpopulations are related by a nonzero coancestry, e.g. when there is migration between subpopulations [10]. However, it can be shown that the partial derivative of the expected value of d^θ with respect to θ is

$$\frac{\partial}{\partial\theta}E\left(d^\theta\right) = -\frac{8p_1p_2\begin{pmatrix}1-3p_1p_2+2\theta+p_1p_2\theta+4\theta^2+\\ 6p_1p_2\theta^2+6\theta^3-3p_1p_2\theta^3+2\theta^4-p_1p_2\theta^4\end{pmatrix}}{(1+\theta)^2(1+2\theta)^2}, \tag{17}$$

which is <0 ∀ $p_1 \in (0, 1)$ since $p_1p_2 \le 0.25$. Therefore, the expected value of d_{ij}^θ is a decreasing function over θ, which means that the expected ASD is always greater between than within subpopulations even when the coancestry between subpopulations is nonzero since the coancestry within is greater than between subpopulations. This relationship holds for all allele frequencies.

Now suppose that there are L independent SNP loci (the independent assumption is valid given the number of loci required to separate populations is much lower than the number of SNPs in the human genome), the expected value and variance of allele sharing distance D over these loci for a given θ are

$$E(D) = E\left(\frac{1}{L}\sum_{l=1}^{L}d_l^\theta\right) = \frac{1}{L}\sum_{l=1}^{L}E\left(d_l^\theta\right). \tag{18}$$

$$Var(D) = Var\left(\frac{1}{L}\sum_{l=1}^{L}d_l^\theta\right) = \frac{1}{L^2}\sum_{l=1}^{L}Var\left(d_l^\theta\right). \tag{19}$$

It is not difficult to show that the maximum value of the variance of d^θ is 0.4375. Therefore, the variance of D decreases as more SNPs are used because $Var(d_l^\theta)$ has the upper bound 0.4375 and L^2 increases much faster than the increase of $\Sigma_{l=1}^{L} Var(d_l^\theta)$, $Var(D) \to 0$ as $L \to$ ••. Increasing the number of SNP loci used makes it more apparent that there are significant ASD differences between subpopulations, as compared with the differences within subpopulations, and the expectation of D between individuals from the same subpopulation will hardly overlap with individuals from different subpopulations. Therefore it is possible to differentiate subpopulations from the half-matrix of pair-wise distances without explicitly estimating allele frequencies for each subpopulation. The population stratification problem is thus reduced to contrasting the ASD means of different groups. It is through the accumulated effect of many SNP loci and the coancestry among individuals within subpopulations that population stratification can be identified.

Numerical Examples

Given the ASD properties that we have just derived, we can use numerical examples to explore it. Specifically, for a given coancestry coefficient we can determine the number of SNPs required to separate subpopulations. In calculating the variance of ASD, we will use the upper bound of variance (0.4375) for all SNP loci. Thus, $Var(D) \approx 0.4375/L$. This approach of variance approximation is simple, though slightly conservative. In order for the expected ASD of between $(E(D^{\theta_b}))$ and within subpopulations $(E(D^{\theta_w}))$ to be x standard deviations apart, i.e. $(E(D^{\theta_b}) - (E(D^{\theta_w}) \ge x\sqrt{0.4375/L}$, the L has to be at least

x^2 0.4375/$(E(D^{\theta_b}) - E(D^{\theta_w}))^2$. Also, while allele frequencies are discrete in nature, it is reasonable to approximate them as continuous values [17–19]. Therefore, we model the distribution of allele frequencies by two kinds of distributions: Uniform and Beta.

In the first example we consider p_1 as the minor allele frequency (MAF) and model it by a Uniform [0.1, 0.5] distribution (consider only common SNPs). This approximation is reasonable, though biased given the distribution of allele frequencies for genome-wide autosomal SNPs appear to be 'flat' in the Hapmap Phase I dataset under predisposed ascertainment [20]. If we treat $E(d^{\theta})$ as a function of p_1 and denote $f(p_1) = E(d^{\theta})$, the expected ASD between individuals can be derived as

$$E(D)=\int_{0.1}^{0.5} f(p_1)\frac{1}{0.5-0.1}\,dp_1 .$$

Suppose $\theta = 0.15$ and $\theta = 0$ for individuals within subpopulations and between subpopulations, then the corresponding values for $E(D)$ are 0.534 and 0.623, respectively. For a set of 497 independent loci, the means ± SD are 0.534 ± 0.03 and 0.623 ± 0.03. Thus, the difference between means, 0.089, is about 3 SD away. In cases when alleles in different subpopulations are related by a nonzero coancestry, we assume $\theta = 0.05$, and thus, $E(D) =$ 0.593. For a set of 1,131 independent loci, the means ± SD for individuals within subpopulations and between subpopulations are 0.534 ± 0.020 and 0.593 ± 0.020, respectively. Thus, the difference between means, 0.059, is about 3 SD away. This demonstrates that it requires more SNP loci to reach similar separation between ASD means when different subpopulations are related by a nonzero coancestry. With enough SNP loci, the entries of the triangular-matrix of pair-wise distances will divide into two separate groups with hardly any overlap.

In the second example we adopt Wright's two-state mutation model [18] for the stationary pdf of allele frequencies, q, as

$$\varphi(q)=\frac{\Gamma\left(4Nu+4Nv\right)}{\Gamma(4nu)\Gamma(4Nv)}\,q^{4Nv-1}\left(1-q\right)^{4Nu-1},$$

where N is the effective population size, v is the mutation rate from allele a to A and rate u for A to a and $\Gamma()$ denotes the gamma function. Assuming N is 10^6 and the SNP mutation rate equals 10^{-8}, we approximate MAF as a Beta(0.04, 0.04) distribution ($4 \times 10^6 \times 10^{-8} = 0.04$). Again, we only consider common SNPs with MAF $\in$ [0.1, 0.5]. As shown in the previous example, we can derive the $E(D)$ by integration for $\theta = 0.15$ and $\theta = 0$ and get the corresponding means of ASD as 0.500 and 0.587, respectively. For a set of 520 independent loci, the means ± SD are 0.500 ± 0.029 and 0.587 ± 0.029, respectively. Thus, the difference between means, 0.087, is about 3 SD away. When alleles in different subpopulations are related by a nonzero coancestry, we assume $\theta = 0.05$, and thus, $E(D)$ = 0.557. For a set of 1,212 independent loci, the means ± SD for individuals within subpopulations and between subpopulations are 0.500 ± 0.019 and 0.557 ± 0.019. Thus, the difference between means, 0.057, is then about 3 SD away. Again, we see that with enough SNP loci, the entries of the triangular-matrix of pair-wise distances will divide into two separate groups with hardly any overlap. Individuals from different subpopulations are separable based on the ASD matrix without having to calculate the allele frequencies.

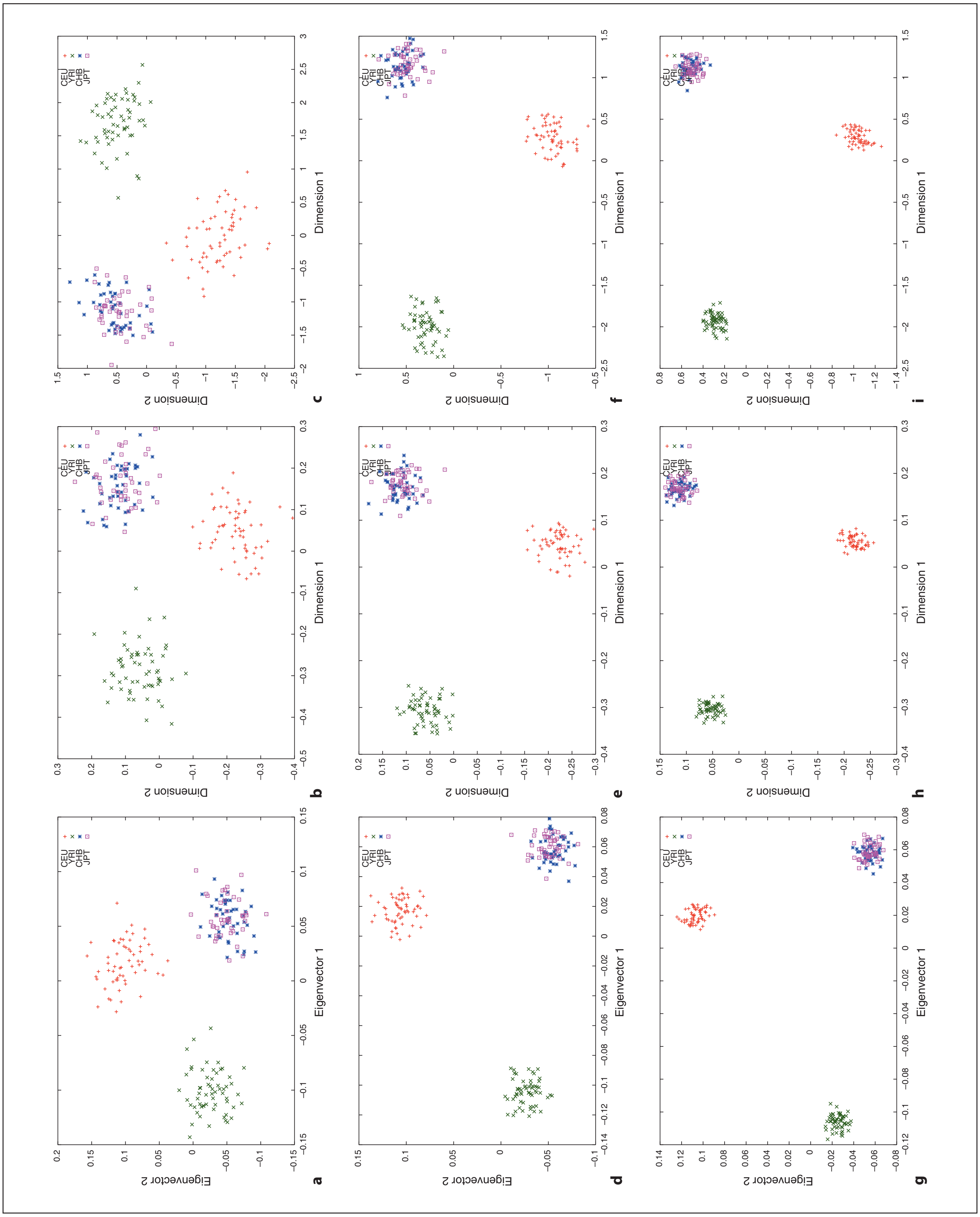
Relation to Other Work

Recently, principal component analysis (PCA)-based approaches were proposed to address population stratification issues [21, 22], which use covariance among normalized genotype scores of individuals. ASD and the covariance-based approaches are similar in nature. In the ASD-based approaches, individuals with small ASD tend to cluster together. While in PCA-based approaches, large eigenvalues (nonrandom population stratification) are consequences of high correlation among vectors of genotype scores. Intuitively, small ASD among individuals will correspond to high correlation among genotype vectors. Both small ASD and high correlation should be due to the result of evolutionary relatedness, coancestry, which can be captured by genome-wide random SNPs.

(For figures see next pages.)

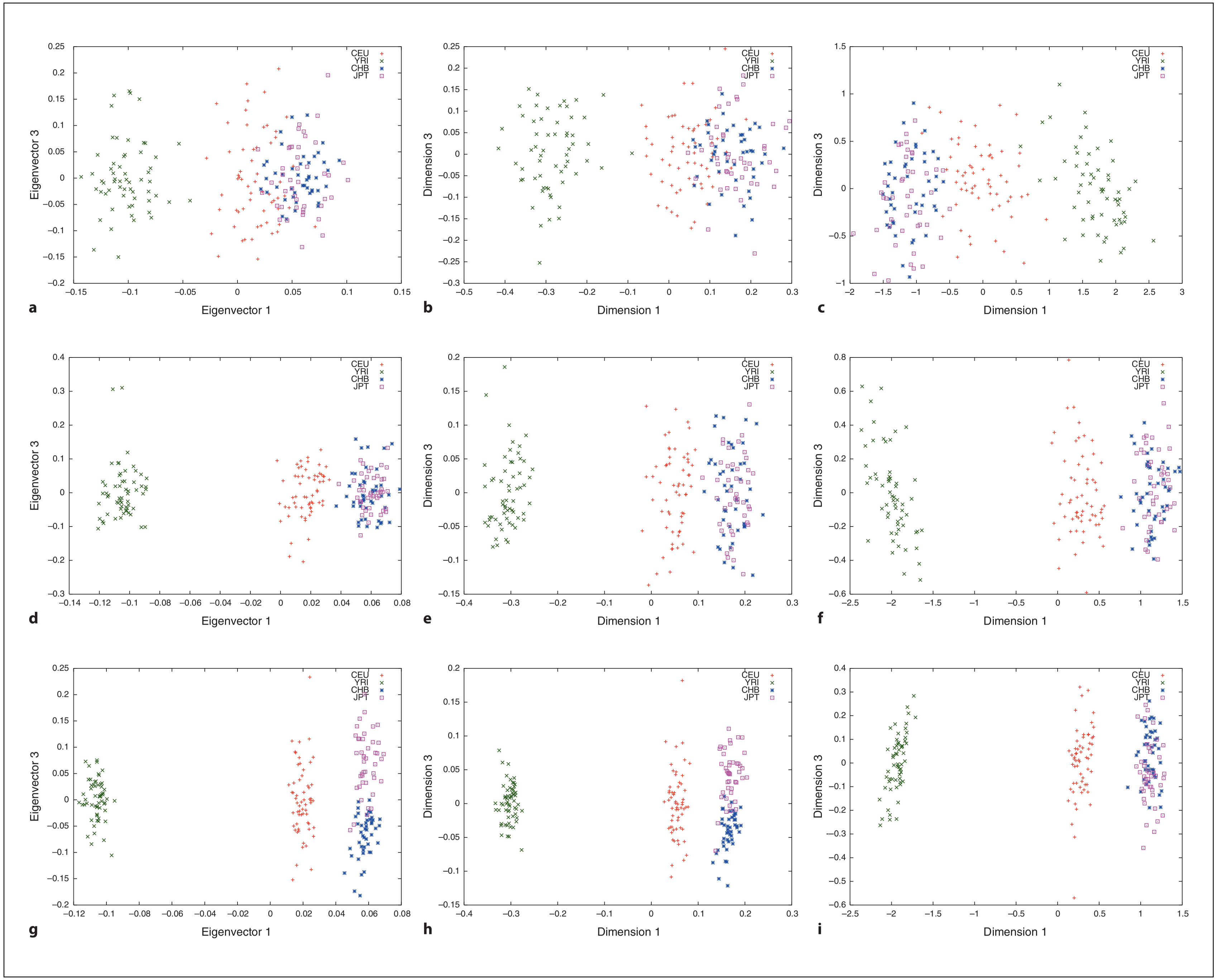
Fig. 1. Cluster results for the CEU, YRI and CHB + JPT individuals (Dimension 1 vs. 2). This figure shows the clustering results using different distance metrics and clustering methods with a number of genome-wide random autosomal SNP loci. In the left column, (**a**), (**d**) and (**g**) correspond to the clusters generated using the covariance matrix and PCA. In the middle column, (**b**), (**e**) and (**h**) correspond to the clusters generated using the ASD matrix and MDS. In the right column, (**c**), (**f**) and (**i**) correspond to the clusters generated using the correlation matrix and MDS. In the top row, (**a**), (**b**) and (**c**), 100 SNPs are used. In the middle row, (**d**), (**e**) and (**f**), 500 SNPs are used. In the bottom row, (**g**), (**h**) and (**i**), 2,000 SNPs are used. Abbreviation: SNP, single nucleotide polymorphism; PCA, principal component analysis; ASD, allele sharing distance; MDS, multidimensional scaling.

Fig. 2. Cluster results for the CEU, YRI and CHB + JPT individuals (Dimension 1 vs. 3). The figure legend is the same as that in figure 1.



1

Color version available online



Color version available online

2

While ASD and correlation-based methods are related, they lend themselves to different types of analysis and this makes it difficult to make a definitive statement about the superiority of one method over another. Depending on the starting matrix, different ordination or clustering methods can be used to inspect the ethnic relationship among individuals. If a dissimilarity matrix like ASD is used, multidimensional scaling (MDS) and Ward's minimum variance method can be used [1, 2, 23]. If the starting matrix is a correlation matrix, PCA can be used instead [21, 22] as well as MDS.

We used HapMap Phase I SNP data [20, 24] to explore the relation between different distance metrics, i.e. ASD vs. correlation/covariance, using corresponding clustering methods, MDS as implemented in R (http://www.r-project.org/) and PCA as implemented in the software, EigenSoft (http://genepath.med.harvard.edu/reich/Software.htm). In the HapMap Phase I SNP data, about 1.1 million SNPs were genotyped genome-wide in 269 individuals from four ethnic populations: Yoruba in Ibadan (YRI), CEPH in Utah residents with ancestry from northern and western Europe (CEU), Han Chinese from Beijing, China (CHB), and Japanese from Tokyo, Japan (JPT), among which there are 209 unrelated individuals: 60 CEU, 60 YRI, 45 CHB and 44 JPT. For this study, we only used unrelated individuals and genomewide random autosomal SNPs. The cluster results are plotted in figures 1–4. Figures 1, 2 and 3 show the clustering results (Dimension 1 vs. 2, Dimension 1 vs. 3 and Dimension 2 vs. 3, respectively) for the CEU, YRI and CHB + JPT individuals using different distance metrics and clustering methods with a number of genome-wide autosomal SNP loci. 100, 500 and 2,000 SNPs are used in the top, middle and bottom panel, respectively. Figure 4 shows the clustering results for the CHB and JPT individuals using different distance metrics and clustering methods with a number of genome-wide autosomal SNP loci. 1,000, 5,000 and 20,000 SNPs are used in the top, middle and bottom panel, respectively. In figures 1–4, clusters in the left, middle and right column are generated using the covariance matrix and PCA, the ASD matrix and MDS, the correlation matrix and MDS, respectively.

Figures 1–4 show the same pattern: with the increasing number of SNPs used, the within subpopulation individuals are closer and closer to each other while the between subpopulation individuals are farther and farther away from each other. Moreover, different distance metrics, ASD and correlation/covariance, with different clustering methods, MDS and PCA, identified very similar clusters in these data sets except that covariance + PCA and ASD + MDS showed better separation on CHB and JPT individuals than correlation + MDS did when enough SNPs were used, e.g. 2000 SNPs, in figures 2 and 3 g and h versus i. Multiple runs also gave similar results (data not shown).

Discussion

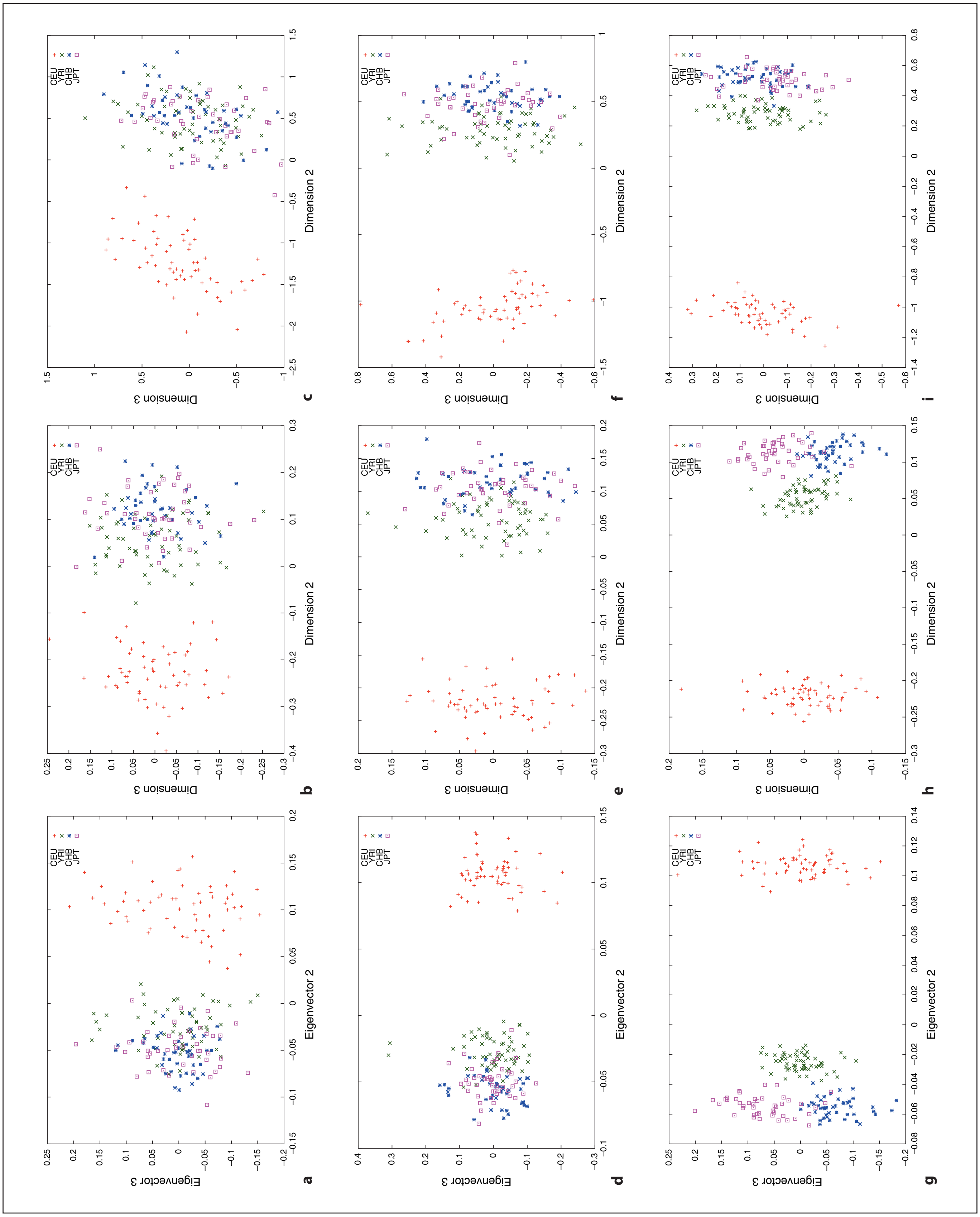
The importance of this work is to provide theoretical support for the observation that human populations can be separated simply through a pair-wise distance matrix, which has been shown in a large empirical study using Hapmap [20, 24] and Perlegen [25] SNP data sets [1]. Moreover, through the derivation of the distance method, the population stratification problem is reduced to contrasting the means of different clusters. We focus on explaining why ASD works in population stratification analysis rather than how to cluster genetic data using ASD. Based on the ASD matrix, standard statistical clustering algorithms, e.g. Ward's minimum variance and MDS methods, can be used to further inspect the ethnic relationship among individuals. We concentrate on biallelic SNPs in this study. The multiallelic genetic markers, microsatellites, are also effective in human evolutionary studies using the pairwise distance matrix of proportion of allele sharing [3] and should follow similar reasons: individuals within subpopulations have a higher proportion of allele sharing than between subpopulations since the match probabilities within is greater than between subpopulations due to coancestry.

From the ASD derivation, we see that when sufficient SNP loci are used in the analysis, the distribution of with-

(For figures see next pages.)

Fig. 3. Cluster results for the CEU, YRI and CHB + JPT individuals (Dimension 2 vs. 3). The figure legend is the same as that in figure 1.

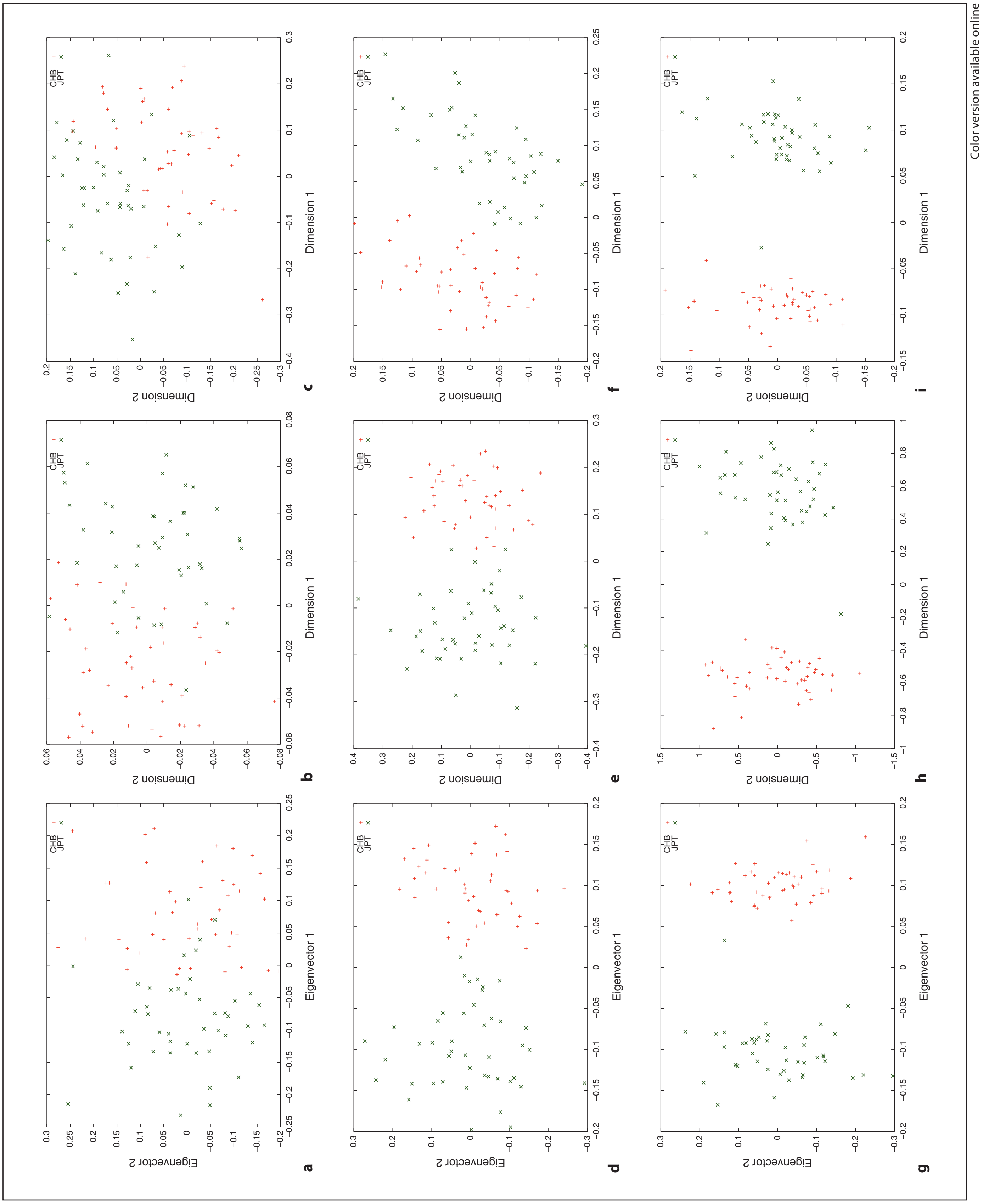
Fig. 4. Cluster results for the CHB and JPT individuals. This figure shows the clustering results using different distance metrics and clustering methods with a number of genome-wide random autosomal SNP loci. In the left column, (**a**), (**d**) and (**g**) correspond to the clusters generated using the covariance matrix and PCA. In the middle column, (**b**), (**e**) and (**h**) correspond to the clusters generated using the ASD matrix and MDS. In the right column, (**c**), (**f**) and (**i**) correspond to the clusters generated using the correlation matrix and MDS. In the top row, (**a**), (**b**) and (**c**), 1000 SNPs are used. In the middle row, (**d**), (**e**) and (**f**), 5,000 SNPs are used. In the bottom row, (**g**), (**h**) and (**i**), 20,000 SNPs are used. Abbreviation: SNP = single nucleotide polymorphism; PCA = principal component analysis; ASD = allele sharing distance; MDS = multidimensional scaling.



3

Color version available online

4


a
b
c
d
e
f
g
h
i
CHB
JPT
Eigenvector 1
Eigenvector 2
Dimension 1
Dimension 2
Color version available online

in-subpopulation ASD and between-subpopulation ASD hardly overlap with each other and the subpopulations are separable. One of the advantages of the distance method is that there is no need to specify the allele frequencies explicitly. Therefore, population allele frequencies do not have to be approximated by sample allele frequencies. This is important because sample allele frequencies may give biased results due to high variability. The allele frequencies and coancestry information are embedded in the pair-wise distance matrix over a large number of random SNP loci. In contrast, model-based methods, like, STRUCTURE [26] and L-POP [27], which use sample allele frequencies as surrogates for subpopulation allele frequencies, need relatively large data sets for each subpopulation in order to estimate allele frequencies reliably, which may not be feasible in practice. Another advantage of the distance method is it is easy to calculate with no decrease in accuracy and is also suitable for population outlier detection (it is unlikely to know the allele frequencies for the outlier individuals).

In summary, we have shown the theoretical foundation for using ASD for human population stratification analysis. The ASD method combined with SNP markers have considerable power in population stratification analysis and it is not necessary to estimate allele frequencies to separate individuals with different ethnic backgrounds. The correlation/coancestry among individuals within subpopulations, which can be captured by the ASD, contributes to the classification. Diploid individuals from different subpopulations can thus be separated from half-matrix of pair-wise distances.

Acknowledgments

This work was supported in part by NIH grants R01HL08770002S1, U01AG02374605, GM45344 and NS39764. Dr. Bruce Weir provided initial guidance for population genetics theory. Dr. Joshua Starmer gave advice in manuscript preparation. We thank anonymous reviewers for providing constructive feedback for this work.

References

1 Gao X, Starmer J: Human population structure detection via multilocus genotype clustering. BMC Genet 2007;8:34. doi:10.1186/1471-2156-8-34.

2 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: Plink: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.

3 Bowcock A, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd J, Cavalli-Sforza L: High resolution of human evolutionary trees with polymorphic microsatellites. Nature 1994; 368:455–457.

4 Mountain J, Cavalli-Sforza L: Multilocus genotypes, a tree of individuals, and human evolutionary history. Am J Hum Genet 1997; 61:705–718.

5 Nakamura T, Shoji A, Fujisawa H, Kamatani N: Cluster analysis and association study of structured multilocus genotype data. J Hum Genet 2005;50:53–61.

6 Shriver M, Kennedy G, Parra E, Lawson H, Sonpar V, Huang J, Akey J, Jones K: The genomic distribution of population substructure in four populations using 8,525 autosomal snps. Hum Genomics 2004;1:274–286.

7 Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, Schurr TG, Zhadanov SI, Osipova LP, Brutsaert TD, Friedlaender J, Jorde LB, Watkins WS, Bamshad MJ, Gutierrez G, Loi H, Matsuzaki H, Kittles RA, Argyropoulos G, Fernandez JR, Akey JM, Jones KW: Large-scale snp analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2005;2:8189.

8 Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD: Measuring European population stratification with microarray genotype data. Am J Hum Genet 2007;80: 948–956.

9 Edwards AWF: Human genetic diversity: Lewontin's fallacy. BioEssays 2003;25:798–801.

10 Weir BS: Genetic Data Analysis II. MA, USA: Sinauer Associates Inc., ed 2, 1996.

11 Weir BS, Hill WG: Estimating F-statistics. Annu Rev Genet 2002;36:721–750.

12 Anderson AD, Weir BS: A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics 2007;176:421–440.

13 Balding DJ, Nichols RA: DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 1994;64:125–140.

14 Balding DJ, Nichols RA: A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 1995;96:3–12.

15 Ayres KL: Relatedness testing in subdivided populations. Forensic Sci Int 2000;114:107–115.

16 Weir BS, Anderson AD, Hepler AB: Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet 2006;7:771–780.

17 Fisher RA: The Genetical Theory of Natural Selection. Oxford, UK: Oxford University Press, 1930.

18 Wright S: Evolution in medelian populations. Genetics 1931;16:97–159.

19 Kimura M: Solution of a process of random genetic drift with a continuous model. Proc Natl Acad Sci USA 1955;41:144–150.

20 The international HapMap consortium: A haplotype map of the human genome. Nature 2005;437:1299–1320.

21 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. PLoS Genetics 2006;2:e190. doi:10.1371/journal.pgen.0020190.

22 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909.

23 Gao X, Starmer J: AWclust: point-and-click software for non-parametric population structure analysis. BMC Bioinformatics 2008;9:77. doi:10.1186/1471-2105-9-77.

24 The international HapMap consortium: The international HapMap project. Nature 2003; 426:789–796.

25 Hinds D, Stuve L, Nilsen G, Halperin E, Eskin E, Ballinger D, Frazer K, Cox D: Whole genome patterns of common DNA variation in three human populations. Science 2005; 307:1072–1079.

26 Pritchard JK, Stephens M, Donelly P: Inference of population structure using multilocus genotype data. Am J Hum Genet 2000; 67:945–959.

27 Purcell S, Sham P: Properties of structured association approaches to detecting population stratification. Hum Hered 2004;58:93–107.