Inferring biological population membership: An exploration of the continuum of genetic relationships

by

Nicole M. Scott

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2010

Doctoral Committee:

Professor Jeffrey C. Long, Chair
Professor David T. Burke
Associate Professor Noah A. Rosenberg
Assistant Professor William C.L. Stewart
Assistant Professor Sebastian K. Zoellner

In loving memory of my Father, Terrence J. Scott

ACKNOWLEDGEMENTS

Firstly, I would like to thank my parents Terrence and Miriam, my sister Renée, and my grandmother Marie for their support and love. My father was always a strong influence on my intellectual pursuits, giving me the courage to seek different perspectives and think outside of the box. It was perhaps this beginning that led me to where I am today. As an undergraduate I was lucky to fall into a microbiology lab under the direction of Jeffrey Lawrence, who handled my dabbling and laboratory accidents with kindness and support. To Jeff I thank for the experience of knowing that I would not seek a career at the bench. Erstwhile, I was mentored intellectually by Stanley Shostak- who like my father, believed that unconventional thinking was not only necessary, but the only way to think. To Stan I owe an enormous gratitude, for my introduction to Philosophy of Science, and the encouragement to continue training to answer the longstanding questions I had about population variation. I also thank Mary Marazita, Kathy Neiswanger, and Brion Maher for their intellectual and emotional support and encouragement through my master's degree. To Charlie Sing and Tom Rea, I thank you for the inspiration and courage to pursue my intellectual questions in future - and the knowledge that environment and context are inescapable consequences of existence.

I thank my friends Mara Steinkamp and Colleen Currie, who in addition to serving as examples of strong successful women, always made time to encourage and support me and share the laughter of good times. Colleen, the diversions, motivation, and

laughter you provided in these last months has been indispensable-I could not have finished without you. To Lauren Lucas, I am forever indebted to you for your love and support, and ability to be strictly honest and frank. I am so lucky to have found such an amazing friend. To Andrea Butler, I thank you for your stories and laughter from Korea. I also would like to thank my friends Cris van Hout, Ali Shojaie, and Tim Connallon, with whom I spent countless hours of study group time, and have always served as both friends and intellectual colleagues. To Meghan Healey, I thank you for your kindness and friendship, and opening your home to me. To Xujishu and Jeff Gregory, your honesty and kindness, has helped me through the last year. To Barr Halevi, thank you for your kindness, support, and friendship through this stressful period. To my friends Asa Kolterud, Jonas Persson, Mattias and Rebeka Jakobsson, Michele Gornick, Kristen Stevens, Kanaan Shah, Chrysta Lienczewski, Felix Eichinger, David Britton, Simon Lunn, and Charlotte Jonsson, Zoe Prebble, to whom I am indebted for both intellectual and emotional support and an occasional diversion. My gratitude is also extended to Alice Brunner, whose encouragement, kindness, and patience has helped me to finish. I am lucky to have found such a patient and understanding mentor.

I also thank the Genome Sciences Training Grant for the opportunity to exist between a number of fields and three years of funding support. The Program in Biomedical Sciences for a year of support, and the Department of Human Genetics faculty and staff for their help and kindness. To my advisor Jeffrey Long, and my committee members David T. Burke, Noah A. Rosenberg, Sebastian K. Zoellner, and William C.L. Stewart I owe a debt of gratitude for your time and intellectual input.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

The concept of a population, composed of potentially interbreeding individuals, plays a central role in the pursuit to understand the distribution of genetic variation in time and space. Population genetics theory often begins with the idealized notion of panmixia in a population, which means that individuals choose mates at random, and that all potential mates have an equal probability of being chosen. Early in the twentieth century, Hardy, Weinberg, and others, showed that panmixia distributes alleles into diploid genotypes independently, and genotypes independently between unrelated individuals. Human mate choices approximate panmixia in many localities, but panmixia typically falls apart at the level of large regions, or ethnically and religiously diverse communities. Generally, individuals fall into approximately panmictic clusters in the space of mating probability, and this clustering creates non-random distributions of allele frequencies and DNA sequence variations.

It is difficult to know for any set of two or more humans whether or not they belong to the same panmictic population. Geography alone is not a sufficient condition, as people who belong to different populations because of religious or ethnic reasons may live in close proximity to each other. Neither is the act of mating a sufficient criterion, because most members of the same population will not mate randomly with each other, and there are occasional matings between people born into different populations.

In human genetics research it is often necessary that the samples of individuals represent the same population. Operational definitions are used as proxies for a panmictic population. However, given that geography or mating alone are not sufficient conditions for determining whether two individuals represent a population, it is challenging to test whether or not a sample represents a single population. Perhaps, the best way to find out if people belong to the same population is to test for non-random association of alleles into diploid genotypes.

This dissertation attempts to question how genetic samples in a collected sample relate to each other, and how we examine the collected sample's history and context. I begin by building on basic theory taken from Mendelian genetics and Hardy Weinberg Equilibrium (HWE) about a single random mating population. I use a simple measure of genotypic identity for a Mendelian population, homozygosity, and then devise a test for whether two individuals share more homozygosity than would be expected for a single random mating population. By observing either the genetic identity or the test statistic between all members of a sample, inferences about the relationships between any *a priori* population categories and the genetic data, relationships between individuals in the genetic data itself, or even more 'useful' categories may be drawn. It is easy to understand that any modifications at the data collection level (e.g. ascertainment, *a priori* population categories) alters the population of inference including its history and context, perhaps unbeknownst to the investigators.

In the second part of this thesis, I devise a method for simulating data that allows for population history and genetic marker ascertainment. I then use this method to study how population history and ascertainment both may alter the sampled population, through a number of methods designed to investigate population structure. Ultimately, as data has become more abundant, a better knowledge about where the boundaries are placed for population designation and whether the sample meets the investigators' criterion for the sampled population, is necessary to give us enough precision to make coherent inferences about genes and populations.

1.1 Hardy Weinberg Equilibrium

For a randomly mating (panmictic) population the genotype frequencies at a locus are easily computed from the allele frequencies. Let us assume that the population is of infinite size and is panmictic, there is no mutation, no migration, and no natural selection. According to the Hardy-Weinberg law, a population with genotypes $P(AA)$, $P(AB)$, and $P(BB)$ reaches equilibrium in one generation of panmixia, with genotype frequencies p^2, $2pq$, and q^2, respectively, where $p = \frac{1}{2}P(AB) + P(AA)$ is the allele frequency for 'A' , and $q = \frac{1}{2}P(AB) + P(BB)$ is the allele frequency for 'B'. This simple yet powerful law demonstrates that panmixia does not change allele frequencies from one generation to the next, and that genotype frequencies come to equilibrium in only one generation (Table 1.1). As a corollary, the law tells us that panmixia alone neither increases nor decreases variation in a population.

If there is mating between relatives, allele frequencies alone cannot predict genotype frequencies in the offspring. We must consider shared ancestors and paths of descent in the calculation of genotype probabilities. To facilitate the study of inbreeding, we denote the identity between two alleles as either identical by state (IBS), that is, identical in their DNA sequence or phenotype, or else identical by descent (IBD), identical in their DNA sequence through common descent from a recent ancestral sequence. Further, let allozygous denote that two alleles in a genotype are IBS but not IBD, and autozygous denote that two alleles in a genotype are IBD. We can then calculate genotype frequencies for offspring of relatives by summing the probabilities of allozygosity and autozygosity for a genotype. Let, F, be the probability of IBD (Weir & Cockerham, 1984) so that the probability of the genotypes are $P(AA) = pF + p^2(1 - F) = p^2 + pqF$, $P(AB) = 2pq(1 - F)$, and $P(BB) = qF + q^2(1 - F) = q^2 + pqF$. These genotype frequencies demonstrate that inbreeding caused by mating between relatives causes an increase in homozygotes relative to the homozygosity in a HWE population. If F is zero, then the frequencies reduce to HWE portions. If we let H_{rel} be the homozygosity for relatives, then for family members $H_{rel} = p^2 + 2pqF + q^2$ reflects the increase in homozygosity due to relative mating compared to the panmictic frequency for homozygosity: $H_{pan} = p^2 + q^2$.

1.2 Genetic drift

In a finite population allele frequencies change by chance alone. Such change is known as random drift, and results in the complete fixation of alleles over time. The rate of allele frequency change is inversely proportional to the size of the population. The dynamics of drift were studied extensively by R.A. Fisher and S. Wright using the following simplified model for sampling (Fisher, 1930; Wright, 1931). Let us say we have N individuals who produce an infinite pool of gametes, from which we sample N pairs at random to form the next generation. The probability of an allele is then dependent on the frequencies in the parent generation. The number of copies of a particular allelic type in the next generation is a binomial random variable. More formally, let $X(t)$ represent the number of copies of allele 'A' at time, t. If $X(t) = i$, the probability that $X(t + 1) = j$ is given by,

$$p_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(1-\frac{i}{2N}\right)^{2N-j}. \tag{1.1}$$

Basic probability shows that with the Wright-Fisher model (eq. 1.1) the expected allele frequency change is directionless. Assuming a panmictic finite population with no migration, no selection and no recombination, eventually fixation: $X(\cdot) = 0$ or $X(\cdot) = 2N$ will be reached. At this time all variation is lost.

A related consequence of finite population size is that it creates inbreeding. The probability of identity by descent from a previous generation $(t-1)$ for two alleles chosen at random in generation t from a population of size N is $(\frac{1}{2N})^2(2N) = \frac{1}{2N}$. That the two randomly chosen alleles were not identical in generation $t-1$ is $1-\frac{1}{2N}$, but these alleles may actually be identical (autozygous) from former generations with probability F measured at time $t-1$. Thus the total probability of identity by descent in generation t is $(\frac{1}{2N} + (1-\frac{1}{2N}))F_{t-1}$. We can write,

$$f_t = 1-(1-F_0)(1-\frac{1}{2N})^t. \tag{1.2}$$

This equation (eq. 1.2) is the probability of IBD from previous generations, but if we let F_0 be balanced by the effect of mutation, μ, we have one of the most well know equations in population genetics: $F = \frac{1}{1+4N_e\mu}$.

A diffusion model of genetic variation

The Wright-Fisher model (eq. 1.1) was the basis for understanding the mean and variance of the allele frequency distribution and developing the concept of effective population size, but the exact mathematical results for the theory could only be obtained for small populations. Developments in probability and stochastic processes in the mid-twentieth century by Kolomogorov, Mal$\acute{e}$cot, and Goldberg, and others aided in the development of an approximation of the Wright-Fisher model for drift in a large population (Kimura, 1955). To characterize changes in a small population, the binomial probability above was used to calculate the probability for the change in state between all possible allelic types. This discrete treatment of change occurs as a Markov chain, where the probability distribution at time t depends only on the previous generation. However, for large populations, the chain becomes large and computationally difficult. To deal with this problem, a continuous model was used to approximate the Markov chain (Wright, 1929; Fisher, 1930; Wright, 1931). This model, the Diffusion Approximation (also known as the Kolomogorov forward equation), describes the distribution of the probability density of an allele π at a locus at time t such that,

$$\frac{\delta\rho\left(\pi,t\right)}{\delta t}=-\frac{\delta}{\delta\pi}\left[M\left(\pi\right)\rho\left(\pi,t\right)-\frac{1}{2}\frac{\delta}{\delta\pi}V\left(\pi\right)\rho\left(\pi,t\right)\right], \tag{1.3}$$

where, $M(\pi)$ is the first moment and thus the average of the random variable $\rho(\pi,t)$, and $V(\pi)$ is the second moment or variance of the random variable $\rho(\pi,t)$. $M(\pi)$, the parameter to describe drift, and $V(\pi)$, the parameter to describe diffusion, were held constant by Wright (1937) to find the stationary or equilibrium distribution. The stationary distribution for the diffusion equation is given by

$$\rho\left(\pi\right)=\frac{c}{V\left(\pi\right)}\exp\left[2\int\frac{M\left(\pi\right)}{V\left(\pi\right)}d\pi\right] \tag{1.4}$$

(Wright, 1937). Work was still needed to understand the distribution of unfixed allele frequencies through time and space. Kimura (Kimura, 1955, 1957) solved the diffusion equation (eq. 1.3) for the probability of fixation at zero and one. The density of a 'drifted' allele frequency, p, that is, unfixed classes, as a function of the initial allele frequency, π, time, effective population size, N, and where F is the hypergeometric function, is

$$\begin{aligned} \phi\left(\pi,\,p,t\right) &= \sum_{j=1}^{\infty}\pi\left(1-\pi\right)j\left(2j+1\right)\left(1+j\right)F\left(1-j,\,j+2,2,\pi\right) \\ &\quad F\left(1-j,\,j+2,2,\,p\right)e^{\frac{-j(j+1)t}{4N}}. \end{aligned} \tag{1.5}$$

Thus, given the initial allele frequency and $t/2N$ it is possible to determine the allele frequency distribution for the unfixed classes. Figure 1.3 illustrates this for a range of t values.

1.3 Population structure

Large populations are rarely in panmixia. The isolation of local populations occurs spatially and by other means, e.g., cultural, linguistic, and geographic forces. Individuals within subpopulations are more likely to mate with each other. This and any other departure from panmixia is population structure. The structure of a large population is often organized into primary subdivisions, which may be split into a number of secondary subdivisions, and so on, until the groups are nearly panmictic. Thus, in a subdivided population, individuals chosen at random from the same subpopulation will have a higher probability of being more related and thus share more of the same alleles than we would expect if the population was in panmixia. Population structure plays a crucial role favoring adaptation to local niches, increasing variation in genotype fitness in different locations, and favoring an increase in new combinations of alleles (Cavalli-Sforza & Bodmer, 1971).

Walhund (1928) examined genetic variation in a subdivided population. Assuming that there is random mating within each subpopulation, he demonstrated that homozygosity increased with allele frequency variation among subpopulations, as compared with the homozygosity in the total population (Walhund's effect). Let us assume there are k subpopulations each with their own allele frequencies, p_k and $q_k = 1 - p_k$. The genotype frequencies are given by HWE: $P(AA) = p_k^2$, $P(AB) = 2p_k q_k$, and $P(BB) = q_k^2$. In the total population, the alleles' frequencies are:

$$\bar{p} = \frac{1}{k}\sum_{j=1}^{j} p_j, \tag{1.6}$$

and the average frequency of homozygous genotypes is:

$$\overline{p^2} = \frac{1}{k}\sum_{j=1}^{j} p_j^2. \tag{1.7}$$

HWE alone is insufficient to calculate the genotype frequencies because we cannot take the allele frequencies from the subpopulations to calculate the genotype frequencies in the total population as $\bar{p}^2 \neq \bar{p^2}$. We can, however, relate the average allele frequencies $\bar{p}$ to the frequency of the homozygotes in the total population $\bar{p^2}$, by the variance:

$$\sigma^2 = \frac{1}{k}\sum_{k=1}^{k} (\bar{p}_k - \bar{p})^2, \tag{1.8}$$

so that,

$$\overline{p^2} = \bar{p}^2 + \sigma^2. \tag{1.9}$$

A simple rearrangement gives $\overline{p^2} = \bar{p}_k^2 - \sigma^2$. From relating the average genotype frequencies in the total population to the genotype frequencies in the subpopulations we have,

$$\theta = \frac{\sigma^2}{\bar{p}\bar{q}}, \tag{1.10}$$

(Wright, 1951; Weir & Cockerham, 1984). θ is the increase in homozygosity in the total population that accompanies subdivision. The expected genotype frequencies in the total population are calculated by $P(AA) = p_k\theta + p_k^2(1-\theta) = p_k^2 + p_k q_k \theta$, $P(AB) = 2p_k q_k(1-\theta)$, and $P(BB) = q_k\theta + q_k^2(1-\theta) = q_k^2 + p_k q_k\theta$. If we take two individuals from different subpopulations, the genotype frequencies of their offspring are in HWE. In the subpopulation, the frequency of homozygous genotypes is then equal to the frequency in the total population as long as $\theta = 0$ in the subpopulation, however if $\theta > 0$ the frequency of homozygotes in the subpopulation relative to the total population is decreased. Letting H_{sub} be the homozygosity of the population with substructure, then $H_{sub} = p_k^2 + 2p_k q_k \theta + q_k^2$ reflects the increase in homozygosity due to substructure compared to the panmictic frequency for homozygosity: $H_{pan} = p_k^2 + q_k^2$.

Models of population structure

One of the simplest models of population structure is the island model. The model captures the probability of genotypes when the total population is no longer panmictic, but subdivided into subpopulations which are themselves panmictic (Wright, 1940, 1943). The model assumes that migration occurs at random to any of the subpopulations from a larger population of infinite size, or equivalently that migration between any two between subpopulations is equally likely. Figure 1.1 illustrates the island model.

In human populations immigrants come largely from neighboring groups, causing a relationship between genes and distance. Individuals spatially near each other have a higher probability of mating. This means that there is a decrease in IBD with geographic distance for which Wright (1943) coined the phrase "isolation by distance" to describe. Wright (1943), along with Mal*é*cot (1945), produced a continuous model to describe this phenomenon in which the density of individuals is clustered at nodes and frequency distribution of distance between the birth places of a parent and child describes the migration. A discrete circular version of this model was termed "the stepping stone model" (Kimura, 1953; Malecot, 1959; Kimura & Weiss, 1964). In the one-dimensional version, an infinite set of supopulations of size N are arranged linearly. Let m be the frequency of migrants per generation. From each subpopulation, $\frac{m}{2}$ migrate to the nearest left neighbor and $\frac{m}{2}$ migrate to the nearest right neighbor, thus $1 - m$ of the subpopulation contributes to itself in the next generation. Figure 1.2 illustrates the stepping stone model. Migration is 'local' in that subpopulations exchange mates only with their nearest neighbor. Using this model, both Mal*é*cot (1959) and Kimura and Weiss (1964), demonstrated that genetic similarity, as given by F, decays exponentially as the geographic distance increases between subpopulations.

Methods for determining population structure in a sample

Many methods have been designed to evaluate population structure (Beerli & Felsenstein, 1999, 2001; J. Pritchard, Stephens, & Donnelly, 2000; Patterson, Price, & Reich, 2006; Long, 2007; Alexander, Novembre, & Lange, 2009; Zhang, Niyogi, & McPeek, 2009). One popular method uses Bayesian clustering to infer membership of individuals into one of K hypothesized clusters. The program STRUCTURE employs this method (J. K. Pritchard & Rosenberg, 1999; J. Pritchard et al., 2000). STRUCTURE uses a Markov Chain Monte Carlo to approximate the allele frequency distribution and then infer the population of origin for the genotypes. Within each subpopulation, allele frequencies are assumed to be in HWE, and loci are assumed independent and unlinked. The goal, to evaluate relationships among individuals and in turn the the maximal number of clusters, or K, in a sample, is determined from the posterior probability distribution of the set of K tested, where, X is the sampled genotypes:

$$Pr\left(X|K_i\right) \approx \frac{exp^{lnPr(X|K_i)}}{\sum\limits^{I} exp^{lnPr(X|K_i)}}. \tag{1.11}$$

Pritchard et al. (2000) stress that this equation is at best an approximate value for the posterior distribution of K, and is an *ad hoc* method for supporting a K that is most consistent to the data. Support for a more appropriate K can also be given if the likelihood of the data fails to increase with a larger K, or no new clusters are produced with a larger K. Because Bayesian clustering methods classify individuals into discrete populations, the method depends on the ability of the researcher to choose the 'correct' value for K. Additionally, even with a number of modifications to the underlying computational algorithm to decrease computation time (Hubisz, Falush, Stephens, & Pritchard, 2009), the computation time is still rather long for large data sets. Other modifications to the Bayesian clustering implemented in STRUCTURE include a model for admixture (Falush, Stephens, & Pritchard, 2003). A number of other methods are also based on Bayesian clustering including PARTITION (Dawson & Belkhir, 2001), TESS (Chen, Durand, & Francois, 2007), and others (Franois, Ancelet, & Guillot, 2006).

Principal components methods were recently revived in a statistical test for population structure (Menozzi, Piazza, & Cavalli-Sforza, 1978; Price et al., 2006). For each variance component, a matrix of the number of alleles at each SNP (e.g. 0,1, or 2) is corrected by the mean of the genotype distribution of the sample. Then a variance/covariance matrix from the adjusted matrix and eigenvector with the largest eigenvalue is calculated. Each successive component is independent of any previous component. Representing the data by the first few components of the data reduces the dimensionality of the data, at the cost of losing some information. Each coordinate of the eigenvector denotes the covariance between an individual with others in the data set for that component. The largest eigenvector for each component approximately follows a Tracy-Widom distribution. From this convenient property, Price et al. (2006), formulate estimators for the number of components, variance and covariance of the eigenvalues, and a test statistic to determine whether population structure exists. Because this method applies a general linear model, there are many sample assumptions, including: linear relationships between variables, lack of multi-collinearity, multivariate normality for the variable distribution, and others. These assumptions may be met with large amounts of genetic data typed from each individual and many individuals used (a large N). The method is also relatively fast even for very large data sets. However, underlying cryptic relationships between individuals in the genetic data may make it difficult for the investigator to interpret results (Novembre et al., 2008).

Generalized hierarchial modeling is a simple but powerful method that relies on estimating gene identity (Nei, 1987) and testing a hierarchial model's fit to the data (Cavalli-Sforza & Piazza, 1975). Gene identity is the probability that two randomly chosen copies of a locus are identical in state. An unbiased estimate of gene identity from an individual is the proportion of homozygous loci in their genome and an estimate of gene identity between two diploid individuals is the probability of a homozygous “offspring”. A matrix is made from all possible pairwise estimates of gene identity, with estimates from within individuals on the diagonal and estimates between individuals on the off-diagonal. The gene identity matrix is then treated as a variance-covariance matrix, and a likelihood ratio test is used to statistically test the fit of an a *a priori* model to the sample variance/covariance matrix (Cavalli-Sforza & Piazza, 1975; Long & Kittles, 2003). A disadvantage to this method is that some structure must exist within the data to test the fit of a particular model. One could not use this method to test for presence or absence of population structure. However, in order to find evidence for *how* the population structure exists in the data, two (or more) *a priori* models of population "treeness" would need to be tested, one with no structure- that is, one which is star-like, e.g. the null model, and some other model of hierarchial relationships. The likelihood ratio statistics between models can be compared to infer a better fit of the model to the data. Steps for testing hierarchial models are elaborated in Generalized Hierarchical Modeling (GHM) (Long, 2007).

1.4 The statistical test that two individuals are from the same panmictic population

Much of the work in this dissertation is based on HWE. I begin by building on basic theory taken from Mendelian genetics and HWE about a single random mating population. As explained above in "Hardy Weinberg Equilibrium" we can see that after one generation of random mating we have HWE. Choosing an allele frequency at random from a random locus allows us to calculate the homozygosity (or inversely, the heterozygosity) at a locus for the population. If we take the average homozygosity over all loci in the population, we then have the proportion of homozygosity for an individual chosen from the population. Further, as long as two unrelated individuals are chosen from the same panmictic subpopulation or population, as demonstrated above, their homozygosity can be calculated using HWE.

Using these premises, I use a simple measure of gene identity for a panmictic

Mendelian population- homozygosity, and then devise a test for whether two individuals' estimates of homozygosity are not statistically different. As explained above, each individual should be an estimate of the population homozygosity. If they are members of the same population, any estimate of homozygosity from their genotypes or between them (an "offspring") should be the estimate of homozygosity of the population from which they come.

However, if we reject the null hypothesis that the two individuals are not random individuals from the same panmictic population, then we can explore the way in which they differ. If they share more homozygosity than would be expected for a panmictic population, that is, $H_{rel} > H_{pan}$ (where, H_{rel} and H_{pan} are defined above in "Hardy Weinberg Equilibrium"), then they are relatives. In particular, if two individuals have a higher than expected homozygosity in an expected offspring, then they are related individuals. If they share less homozygosity between them than would be expected for a panmictic population, that is, $H_{sub} > H_{pan}$ (where, and H_{sub} and H_{pan} are defined in "Population Structure" above), then they are not from the same panmictic population. In particular, if two individuals have a lower than expected homozygosity in an expected offspring, then they are either from different populations or inbred individuals. The method is presented in Chapter two of this dissertation.

1.5 A method for simulating genotypes from different with population history

I present a forward-in-time method for simulating genotypes from different populations that draws on diffusion theory in Chapter three. I model the ancestral population forward in time undergoing neutral processes and then ascertain allele frequencies in the descendent population. More specifically, I first take a sample of loci from a chosen ancestral allele frequency distribution and allow them to drift using Kimura's equation for the probability density of the distribution of allele frequencies under neutral processes (ϕ) given in equation 1.5 (Kimura, 1955). The only parameters needed are the the time to the ancestor and the effective population size of the descendent population. Descendent allele frequencies can be ascertained according to the user and used as desired.

1.6 Assessing ancestry and ascertainment on methods to detect population structure

In Chapter three of this thesis, I use the forward-in-time simulation method introduced above to simulate two underlying phylogenetic trees of data: one divergent and one less divergent. How the data is simulated can have an impact on the methods used to assess population structure. Therefore, I take concern in the simulation method I use because I want to realistically model data. Previous efforts to look at the impact of methods to assess population structure did not employ methods that may realistically represent the data (Latch, Dharmarajan, Glaubitz, & Rhodes, 2006; Schwartz & McKelvey, 2009). The simulation method presented here not only uses a population-genetic model, but affords us the opportunity to ascertain the data using different strategies because we simulate allele frequencies. This way, I can more accurately assess how the methods are affected by ascertainment and the underlying tree.

Mating		Mating frequency	Offspring frequencies		
Individual 1	Individual 2	*Pr (genotype pair)*	AA	AB	BB
	AA	x^2	1	0	0
AA	AB	$2xy$	$\frac{1}{2}$	$\frac{1}{2}$	0
	BB	$2xz$	0	1	0
AB	AB	y^2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	BB	$2yz$	0	$\frac{1}{2}$	$\frac{1}{2}$
BB	BB	z^2	0	0	1

Table 1.1: Frequencies of genotypes in a panmictic population.

Let a population with genotypes $P(AA)$, $P(AB)$, and $P(BB)$, have genotype frequencies x, y, and z, respectively, where $x+y+z = 1$. The frequency of allele A is then $p = \frac{1}{2}y+x$, and the frequency of B is then $q = \frac{1}{2}y+z$. For the next generation, the frequency of the mating is multiplied by the offspring frequency for each offspring type. In one generation of random mating the genotype frequencies come to equilibrium with frequencies, $P(AA) = p^2 = x\prime$, $P(AB) = 2pq = y\prime$ and $P(AA) = q^2 = z\prime$. The frequency of the A allele among the offspring is $p = P(AA) + \frac{1}{2}P(AB) = p^2 + pq$, and the frequency of B is $q = P(BB) + \frac{1}{2}P(AB) = q^2 + pq$.

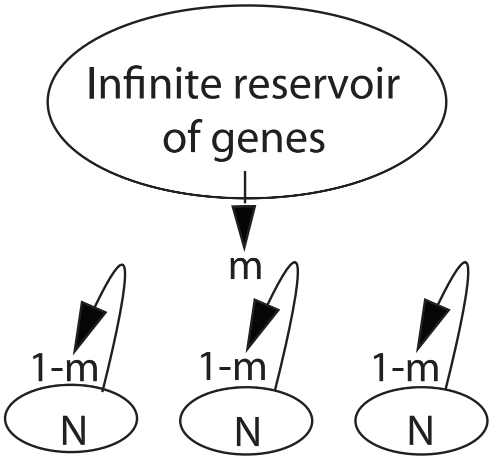

Figure 1.1: Illustration of the Island Model.

Migration into the subpopulations comes from an infinite genetic reservoir (or 'mainland'), at m migrants per generation. $1 - m$ migrants come from each subpopulation itself, and contributes to the next generation for that subpopulation.

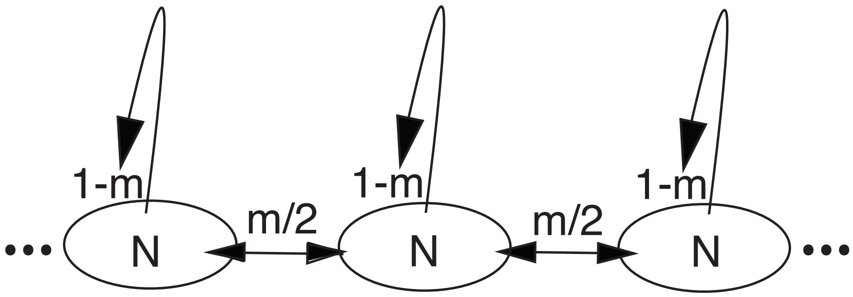

Figure 1.2: Illustration of the Stepping Stone model.

Migration into a subpopulation comes from each nearest neighboring population at $\frac{m}{2}$ migrants per generation. $1 - m$ migrants come from each subpopulation itself, and contributes to the next generation for that subpopulation.

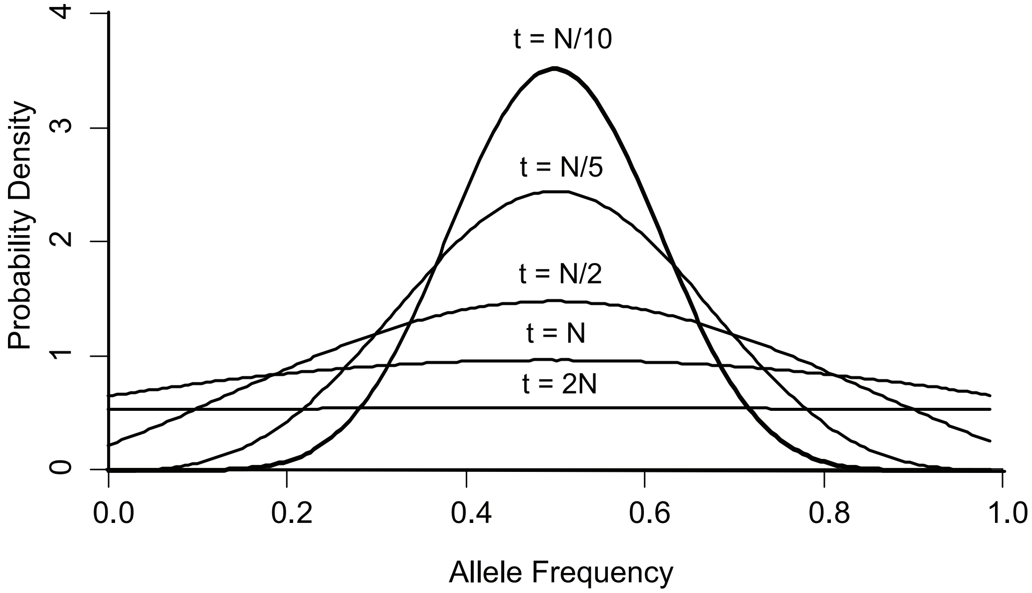

Figure 1.3: Stationary distribution for the probability density of unfixed classes for an initial allele frequency of 0.5 and various $t/2N$ generated from ϕ (eq. 1.5).

CHAPTER II

A test that two individuals are from the same Hardy Weinberg Equilibrium (HWE) population

2.1 Abstract

Here we demonstrate a method to test the homogeneity of three estimates of population homozygosity made from two individuals. We make the three estimates: (1) the proportion of homozygous loci from the first individual and (2) the proportion of homozygous loci from the second individual, and (3) the expected proportion of homozygous loci if the pair was to produce an offspring. If the two individuals are unrelated members of the same population, the three homozygosity estimates are all unbiased estimates of the same population parameters. By contrast, if the two individuals are family members, are from different populations, or are inbred, the estimates will differ in expected ways. We compare the three estimates in a chi-square test statistic with two degrees of freedom, and distinguish between alternative hypotheses using orthogonal contrasts of the three estimates. Using simulated data, we show that our test has the correct Type I error, and demonstrate its power under various alternative hypotheses. We apply our test to HapMap SNP data, and demonstrate its ability to identify pairs that are unlikely to belong to the same population. Significantly, our test does not require any prior information about allele frequency or populations to which the samples may belong. Potential applications

include (1) identifying population heterogeneity in association studies, (2) gaining insight into forensic cases in which only molecular data are available, (3) informing breeding programs for endangered species, and (4) examining population structure in genetics.

2.2 Introduction

Unknown to the research investigator, genetic samples collected from the same population may contain individuals that are non-members or cryptically related. Such samples can obscure or skew any analysis based on these data. In this paper, we introduce a statistical test between pairs of individuals using only DNA sequence typing information. We test the hypothesis that a pair of individuals are unrelated members of the same randomly mating population. To perform this test, we devise estimators of homozygosity for within and between individuals. We base our estimators on simple expectations for a large randomly mating population is in Hardy Weinberg Equilibrium (HWE). Our first estimator is simply the proportion of homozygous loci from an individual. Thus from two individuals, we take one estimate from the first individual, and one estimate from the second individual. From HWE we also know that neither homozygosity for a locus, nor the proportion of homozygosity in an individual's genome changes from one generation to the next. These principles motivate an estimator of population homozygosity between a pair of individuals. Our second estimator is the probability of a homozygous genotype when a randomly chosen allele from the locus of the first individual is identical in state with a randomly chosen allele from the locus of the second individual. This second estimator could also be interpreted as the expected homozygosity if the two individuals were to produce an offspring. For multiple loci, we take an average of this estimator across loci.

From the pair of individuals and two ways to estimate homozygosity we have three estimates: one homozygosity estimate from the first individual, one homozygosity estimate from the second individual, and one homozygosity estimate between the pair of individuals. To compare the three estimates, we construct a chi-square test statistic with two degrees of freedom. If we reject the null hypothesis, we distinguish between alternative hypotheses by constructing orthogonal contrasts of the three estimates. In this manner we can test for a directionality of the estimates and the alternative hypotheses that (A) individuals are related, or (B) that individuals are from different populations or are inbred. If the two individuals are unrelated members of the same population, the three homozygosity estimates are all unbiased estimates of the same population homozygosity. By contrast, if the two individuals are family members, the homozygosity estimates from each individual are unbiased estimates of population homozygosity, while the homozygosity between the two will be in excess. If the two individuals are from different populations, the homozygosity estimates from each individual are unbiased estimates of the population homozygosity to which they belong, while the homozygosity between the two individuals will be less than at least one of the individual estimates.

In this paper, we give the details of this method for sequence data from two individuals. We also demonstrate the validity and utility of the method on HapMap data (The International HapMapConsortium, 2007). The significance of this novel method is that the homogeneity test uses only genotype data from two individuals and does not require any knowledge about allele frequencies, linkage disequilibrium, or population membership.

2.3 Methods

In deriving our estimators we will assume that there is random mating within a population, the loci are autosomal, bi-allelic, statistically independent, selectively neutral, and have the same mutation rate, and that the individuals are diploid, e.g. Hardy Weinberg Equilibrium (HWE). We begin by letting the average population homozygosity, H, represent a population, such that,

$$H = \int_{\frac{1}{2}}^{1} h\varphi\left(h\right) dh, \tag{2.1}$$

where $\varphi\left(h\right)$ is the genome wide probability density of homozygosity. In terms of allele frequency p, let h be the probability of homozygosity or $h = p^2 + q^2$, where $q = 1 - p$ is the frequency of the other allele. It is possible to estimate H (eq. 2.1) using a randomly chosen single locus or randomly chosen multiple loci from a single individual. In an individual i (where i is either 1 or 2 for the pair) if a single locus, k, is homozygous let $\hat{w}_{ik} = 1$, and if otherwise let $\hat{w}_{ik} = 0$. The estimator of genome-wide homozygosity for an individual is the proportion of homozygous loci:

$$\hat{\hat{w}}_i = \frac{1}{L}\sum_{k=1}^{L}\hat{w}_{ik}, \tag{2.2}$$

where L is the number of loci in the random set chosen for analysis.

We also estimate H using a pair of individuals. From alleles between two individuals at a locus k, let $\hat{b}_k$ be the estimator of homozygosity if the pair was to produce a child. The probability of homozygosity from the same locus in a sample of two different individuals is derived explicitly in Table 2.1. The estimator of genome-wide homozygosity between the pair is the expected proportion of homozygous loci if the pair was to produce a child:

$$\hat{\bar{b}} = \frac{1}{L}\sum_{k=1}^{L}\hat{b}_k. \tag{2.3}$$

We use these three estimates calculated from two individuals, $\hat{\bar{w}}_1$ and $\hat{\bar{w}}_2$ from each individual, and $\hat{\bar{b}}$ between the two individuals in a homogeneity test statistic.

We first define the vector $\hat{\mathbf{X}} = \left[(\hat{\bar{w}}_1 - \hat{\bar{w}}_2), (\hat{\bar{b}} - \hat{\bar{w}})\right]$, where $\hat{\bar{w}} = \frac{\hat{\bar{w}}_1 + \hat{\bar{w}}_2}{2}$. From first principles for the variance of the linear combination of two random variables, let

$$\widehat{var}\left(\hat{\bar{w}}_1 - \hat{\bar{w}}_2\right) = \frac{1}{L\left(L-1\right)}\omega, \tag{2.4}$$

and

$$\widehat{var}\left(\hat{\bar{b}} - \hat{\bar{w}}\right) = \frac{1}{L\left(L-1\right)}\left[\sum_{k=1}^{L}\left(\hat{b}_k - \hat{\bar{b}}\right)^2 + \frac{1}{4}\omega - \eta_1 - \eta_2\right], \tag{2.5}$$

where

$$\omega = \sum_{k=1}^{L}\left[\left(\hat{w}_{1k} - \hat{\bar{w}}_1\right)^2 + \left(\hat{w}_{2k} - \hat{\bar{w}}_2\right)^2\right], \tag{2.6}$$

and

$$\eta_i = \sum_{k=1}^{L} \left[\left(\hat{b}_k - \hat{\hat{b}} \right) \left(\hat{w}_{ik} - \hat{\hat{w}}_i \right) \right]. \tag{2.7}$$

We define the homogeneity test statistic as,

$$T = \hat{\mathbf{X}} \hat{\Sigma}^{-1} \hat{\mathbf{X}}, \tag{2.8}$$

where,

$$\hat{\Sigma} = \begin{bmatrix} \widehat{var} \left(\hat{\hat{w}}_1 - \hat{\hat{w}}_2 \right) & 0 \\ 0 & \widehat{var} \left(\hat{\hat{b}} - \hat{\hat{w}} \right) \end{bmatrix}. \tag{2.9}$$

From Normal Distribution theory, under the null hypothesis T is distributed as χ^2_2. We reject the null at α level of significance if T is greater than the $1 - \alpha$ quantile of χ^2_2.

Planned comparisons of alternative hypotheses

In the event that we reject the null hypothesis, we propose a method to reveal the relationship between the pair. We test two different alternative hypotheses: (A) the pair are relatives, or (B), the pair are from different populations or inbred members from the same population. To do this, we first test the null hypothesis: $\hat{\hat{w}}_1 = \hat{\hat{w}}_2$,

$$M_{H_{0_1}} = \frac{\hat{\hat{w}}_1 - \hat{\hat{w}}_2}{\sqrt{\hat{\sigma}_{(\hat{\hat{w}}_1 - \hat{\hat{w}}_2)}}}. \tag{2.10}$$

This hypothesis tests if two individuals are randomly chosen from the same panmictic population. Under HWE we expect that estimates from randomly chosen individuals from the same panmictic population should be similar. If this is not the case then the two are either from different populations or are inbred. Otherwise, if we fail to reject $\hat{\hat{w}}_1 = \hat{\hat{w}}_2$, then the two individuals may be either from the same population (this case would have been caught by the global test statistic T, eq. 2.8, explained above), *or* have similar values of homozygosity but are from different populations (or are inbred from the same population). In this case we further test $\hat{\hat{b}} = \hat{\hat{w}}$, by

$$M_{H_{0_2}} = \frac{\hat{\hat{b}} - \hat{\hat{w}}}{\sqrt{\hat{\sigma}_{(\hat{\hat{b}} - \hat{\hat{w}})}}}. \tag{2.11}$$

If $\hat{\hat{w}} < \hat{\hat{b}}$ then we find support for the alternative hypothesis (A), the pair are relatives. Otherwise, if $\hat{\hat{w}} > \hat{\hat{b}}$ we find support for the alternative (B), the pair are either from different populations or are both inbred from the same population.

Moments and distributions under the null hypothesis

Both estimators converge in probability (pr) to H (eq. 2.1), that is $\hat{\hat{w}}_i = \frac{1}{L}\sum_{k=1}^{L} \hat{w}_{ik} \xrightarrow{pr} H$ and $\hat{\hat{b}} = \frac{1}{L}\sum_{k=1}^{L} \hat{b}_k \xrightarrow{pr} H$. The estimators are also unbiased as: $E\left[\hat{w}_{ik}\right] = E\left[E\left[\hat{w}_{ik}|p\right]\right] = E\left[\hat{\hat{w}}_i\right] = H$, and $E\left[\hat{b}_k\right] = E\left[E\left[\hat{b}_k|p\right]\right] = E\left[\hat{\hat{b}}\right] = H$, while the $E\left[\hat{\mathbf{X}}\right] = [0,0]$. From above we see that $\hat{\hat{w}}_i$ is the sum of the identical, independently distributed binomial random variables, or $\frac{1}{L}\sum_{k=1}^{L} \hat{w}_{ik}$ (eq. 2.2), thus each $\hat{w}_{ik}$ has $E[\hat{w}_{ik}] = H$ and

$$\sigma_{\hat{w}_{ik}} = p_k(1 - p_k). \tag{2.12}$$

Thus, by the Central Limit Theorem (CLT) (Casella & Berger, 2002), for a large

L, $\hat{\bar{w}}_i \underset{CLT}{\longrightarrow} N\left(H, \sigma^2_{\hat{\bar{w}}_i}\right)$, where

$$\sigma_{\hat{\bar{w}}_i} = \sqrt{\frac{\sum\limits_{k=1}^{L} E\left[h_k^2\right] - \left(E\left[h_k\right]\right)^2}{L}}. \tag{2.13}$$

Analogously, since $\hat{\bar{b}}$ is the average of identical independently distributed random variables, $\hat{\bar{b}} = \frac{1}{L}\sum\limits_{k=1}^{L} \hat{b}_k$ (eq. 2.3), each with expectation equal to H (eq. 2.1) and

$$\sigma_{\hat{b}_k} = \frac{1}{2}h(1-h) - \frac{1}{4}(1-h)^2, \tag{2.14}$$

where $h = p^2 + q^2$ as above, then by the CLT $\bar{b} \underset{CLT}{\longrightarrow} N\left(H, \sigma^2_{\hat{\bar{b}}}\right)$ where,

$$\sigma_{\hat{\bar{b}}} = \sqrt{\frac{\frac{1}{2}\sigma_{\hat{\bar{w}}_i} - \frac{1}{4}E\left[(1-h_k)^2\right]}{L}}. \tag{2.15}$$

It then follows that $(\hat{\bar{w}}_1 - \hat{\bar{w}}_2)$ and $(\hat{\bar{b}} - \hat{\bar{w}})$ are also distributed normally, each with expectation equal to zero and $\sigma_{\bar{w}_1 - \bar{w}_2} = \sigma_{\bar{w}_1} + \sigma_{\hat{\bar{w}}_1}$, and $\sigma_{\bar{b}-\bar{w}} = \sigma_{\bar{b}} + \frac{1}{4}\sigma_{\bar{w}_1} + \frac{1}{4}\sigma_{\bar{w}_1}$. By the Multivariate Central Limit Theorem (MCLT), the vector $\hat{\mathbf{X}} \underset{MCLT}{\longrightarrow}$ Bivariate Normal $([0,0], \sum)$, where, $\hat{\Sigma}$ converges asymptotically to:

$$\Sigma = \begin{bmatrix} \sigma_{\bar{w}_1 - \bar{w}_2} & 0 \\ 0 & \sigma_{\bar{b}-\bar{w}} \end{bmatrix} \tag{2.16}$$

(Rencher, 2002). Conveniently, the covariance of $\mathbf{X}_1$ and $\mathbf{X}_2$ in the matrix Σ is zero.

Further, because of the asymptotically normality of the individual and between homozygosity estimators, we can take advantage of the properties of normally distributed random variables. The sum of squared normal random variables follows a chi-square distribution. Likewise, as we defined our test statistic, it is the sum of squared normal random variables and thus follows a chi-square distribution. This is also analytically proven using Slutsky's theorem (Casella & Berger, 2002). Since $\hat{\hat{w}}_i$ and $\hat{\hat{b}}$ converge in probability to H, and since they are asymptotically equivalent, as vectors of $\hat{\mathbf{X}}$, the sum of the squared vectors will both converge in distribution to the χ^2_2 under the null hypothesis, or $T \xrightarrow{d} \chi^2_2$. Under the alternative, however, the vectors of $\hat{\mathbf{X}}$ are no longer asymptotically equivalent, then $f(\hat{X}_1, \hat{X}_2) \xrightarrow{d} f(X_1, X_2)$, where d denotes convergence in distribution .

For the contrast statistics, M, we can see that $\hat{\hat{w}}_1 - \hat{\hat{w}}_2$ and $\hat{\hat{b}} - \hat{\hat{w}}$ are distributed normally, then by the Central Limit theorem, $M_i \underset{CLT}{\longrightarrow} N\left(0, 1\right)$. By Fisher's least significant difference (LSD) we do not need to correct for multiple tests. LSD states that when only significant cases are tested in further statistical tests, those additional tests will have an $\alpha < 0.05$ ((Casella & Berger, 2002)). Likewise here, we do not need to correct for multiple tests because only pairs that have a significant global homogeneity test and thus an $\alpha < 0.05$ are then tested through the two planned comparisons of alternative hypotheses, thus guaranteeing a Type I error rate of less than α.

Simulations of multi-locus SNP genotypes

To test the validity and power of the homogeneity statistic under the null hypothesis and different alternatives, we simulate population genetic SNP data. Our simulations are based on drawing multi-locus SNP genotypes for a pair of individuals from an allele frequency distribution that is chosen to represent a population. A summary of the methods are given in Figure 2.1.

Under the null hypothesis: We let a beta distribution represent a distribution of allele frequencies across loci for a population with drift and mutation, in the absence of selection (Wright, 1931). The parameters for the beta are $4Nv$ and $4Nu$, where N is the effective population size, v is the mutation rate, and u is the reversion rate. Here we choose to use estimates of 10,000 for N, 1.2×10^{-8} for v, and 0.8×10^{-8} for u. From this distribution, we take a sample of L loci and let p_k and $q_k = 1 - p_k$, be the allele frequencies at the kth locus. Only p_k between 5% and 95% are used. We generate the genotype for the first individual, using p_k^2, $2p_kq_k$, and q_k^2. We then repeat this for each $k = 1 \ldots L$ from the sample of L loci. The genotype of the second individual is then generated similarly from the sample of L loci.

Under alternative H_{0_1} - relative pairs from the same populations: Again we take a sample of L loci from a beta distribution to generate the genotype of the first individual. We then generate the multi-locus genotype of their relative at each locus as below.

A) For a parent-offspring pair, we let the first individual be the parent. For the offspring (the second individual) at each locus, we take one allele randomly from the parent and one allele from the population. We do this for the sample of all L loci.

B) For a full-sibling pair, we let the first individual be a sibling. For the second sibling at each locus with $Pr\left(\frac{1}{4}\right)$ we take the genotype from the population (the pair share no alleles), with $Pr\left(\frac{1}{2}\right)$ we follow the parent-offspring pair above (the pair share 1 allele), and with $Pr\left(\frac{1}{4}\right)$ we use the genotype from the first sibling (the pair share 2 alleles). This is repeated for the sample of all L loci.

C) For general unilineal relatives, let ϕ be the kinship coefficient of the desired relative pair. For the other relative (second individual) at each locus with $Pr(\phi)$ we follow the parent-offspring pair above and with $Pr(1-\phi)$ we take both alleles from the population. This is again repeated for the sample of all L loci.

Under H_{0_2} - pairs from different populations: For pairs from different populations, we simulate descendent SNP allele frequencies based on a forward-time model whereby a sample of ancestral allele frequencies drift according to the time to the ancestor from the descendant and the effective population size of the descendent population. This method has been described in detail elsewhere (in Chapter three of this thesis). For the ancestral distribution we again use a beta distribution with the parameters as above. The time to the ancestor and effective population size parameters were garnered from STR data (Hunley, Jie, Lewis, Malhi, & Long, n.d.) to create 'European', 'Chinese', and 'Japanese' distributions of allele frequencies. SNPs were ascertained if polymorphic in both populations and genotypes were then generated assuming HWE.

Simulations using empirical allele frequencies. As an alternative to using a beta distribution to represent a population, and then taking a sample of L loci from that distribution to be the allele frequencies, we use extant empirical allele frequencies calculated from HapMap. For simulations under the null hypothesis, we take the sample of L loci allele frequencies calculated from the HapMap populations. We then generate the genotypes using p_k^2, $2p_kq_k$, and q_k^2, repeating this for each $k = 1 \ldots L$ from the sample of L loci.

HapMap Data. We used the International HapMap Project (The International HapMap Consortium 2003) data release #23 in which samples were typed using the Affymetrix 500k chip. The 270 samples included in the project are 30 trios of Yoruba from Ibadan, Nigeria (YRI), 45 unrelated Japanese from Tokyo, Japan (JPT), 45 unrelated Han Chinese from Beijing, China (CHB), and 30 trios from the CEPH, Utah residents with ancestry from northern and western Europe (CEU). The data are completely de-identified. Our inclusion criteria for SNPs included: a minor allele frequency greater than 5% across all populations, a chi-squared test for HWE within each population with a p-value of greater than 10^{-6}, and no more than 20% missing data across all populations. From these SNP data, we calculated the empirical allele frequencies for the Han Chinese and Japanese populations. For the application of the method, we chose approximately 100,000 equidistantly spaced SNPs. The data from the HapMap project are an ideal public data set to explore our method, because it contains population samples of unrelated individuals and trios of related individuals.

2.4 Results

We can obtain unbiased estimates of homozygosity and standard errors of our estimates of H (eq. 2.1) from genotypes without any knowledge of allele frequencies. The variances of the estimators, for a single locus across a range of allele frequencies, are shown in Figure 2.2. The between homozygosity estimator variance (eq. 2.14) has a local minimum at the allele frequency 0.5 due to the increased probability of at least one of the pair being heterozygous. It might seem more intuitive to average the individual estimators to create a shared estimator for homozygosity. However, the small variance of our homozygosity estimator between the individuals demonstrates that averaging the two individual estimates ($\hat{w}_2$ and $\hat{w}_2$) would have a larger variance across a range of allele frequencies. The shape of the variance of the estimator between the individuals (eq. 2.14) and the covariance at a single locus (or $cov(\hat{w}_{ik}, \hat{b}_k)$) is due to the quartic nature of the functions. As is seen from both the standard deviation and covariance of the estimators, the performance of the estimators and test statistic depends on the probability density of the homozygosity integrated across all allele frequencies or $\varphi\left(h\right)$, which is unknown.

We show the distribution of the estimates $\hat{\hat{w}}_i$ and $\hat{\hat{b}}$ calculated from simulated data in Figure 2.3. The simulated estimates approximate a normal distribution even if the estimates are calculated from as few as 100 loci (Figure 2.3). The estimates are also normally distributed when genotypes are simulated under a wide range of underlying allelic site-frequency distributions (data not shown). As we expect, the T statistic (eq. 2.8) follows a χ_2^2 as demonstrated in Figure 2.4.

Power of the test statistics

The power of the homogeneity test statistic is demonstrated in Table 2.2. Importantly, pairs simulated under the null hypothesis, have a very low Type I error (where $\alpha 0.05$) for both the underlying beta distribution and the empirical allele frequencies. As might be expected, the simulations of relative pairs reflecting closer genetic relationships, such as parent-child or full-siblings, enable us to detect a difference between the individuals even for a small number of loci (Table 2.2). For more distantly related pairs, such as second cousins, as many as 100K loci may be required. This power, however, is dependent on the underlying allele frequency distribution from which the alleles for relative pairs are chosen. For pairs simulated from different populations, the power to reject the null hypothesis increases for more distantly related populations.

The contrast test statistics, $M_{H_{0_1}}$ and $M_{H_{0_2}}$ (eqs. 2.10 and 2.11) have high power even for a small number of loci. Further, by the time the number of loci is increased to achieve appropriate power for the homogeneity test there is full power for the contrast test statistics.

Application of the method to HapMap data

Figure 2.5 shows the homozygosity estimates for the HapMap individuals for

100,000 SNPs. Along the diagonal are the homozygosity estimates for each individual, while the off-diagonals are homozygosity estimates between pairs of individuals. Just off-diagonal in pink are the estimates between parent and offspring pairs from the trios in the CEU and YRI populations. Homogeneity test statistics from the data set are located in Figure 2.6. On the diagonal, statistics are set to zero, as no statistic was calculated, and on the off-diagonals are the test statistics calculated from pairs of individuals. The CEU and YRI populations contain trios. For pairs of these related individuals, we reject the global homogeneity test statistic. The identification of the related individuals in the trios demonstrates the validity and utility of the test. The homogeneity test also identifies a number of outliers in this data set (Figures 2.7- 2.16. The trios are further identified using the contrast tests in the CEU (Figure 2.7) and YRI populations (Figure 2.16). The contrast tests also identifies a number of outliers in the data set (for example, Figure 2.14 for JPT/JPT outliers).

2.5 Discussion

We have devised a method to evaluate the null hypothesis that two individuals are randomly drawn from the same population. To do this, we derived two unbiased estimators for homozygosity. The resulting estimates are precise, valid, and normally distributed. From these estimates, we construct a statistical test of homogeneity, in which the statistic is distributed as a chi-square with two degrees of freedom. If the null hypothesis is rejected, we then test for either (A) that the pair are relatives or (B) that they are from different populations or are both inbred from the same population.

This method, based on population-genetic theory and principles, evaluates the relationship between a pair of individuals. Since our method only requires sequence data and does not require prior knowledge about the allele frequencies or population from which the two samples came, it is will be especially useful in the context of identification in forensics, paternity, and relatedness. Given an unknown forensic DNA sample, the test could be reliably used to include or exclude potential suspects or to determine parent-offspring or other relative pairs. In cases of paternity, the pairwise test has particularly high power (Table 2.2) to detect closely related pairs. A number of methods currently exist to identify pairs of familial related individuals (Thompson, 1975; Queller & Goodnight, 1989; Ritland, 1996; Weir, Anderson, & Hepler, 2006; J. Pritchard et al., 2000; Lynch, 1988; Lynch & Ritland, 1999; Epstein, Duren, & Boehnke, 2000). Our method is computationally simple, easy to interpret, and only requires two genetic samples. This makes our method advantageous in certain circumstances where an exact relationship is not needed, only the knowledge that a familial relationship exists. We demonstrated the ability to pick out pairs of related individuals using the method on HapMap data. We were clearly able to discern the trios with very few loci, and we were even able to find more distantly familial-related individuals *between* trios in the HapMap data.

Not only is it useful to understand the relationship between a pair of individuals, but it is also useful to understand the genetic relationship between individuals composing a sample. Perhaps the investigator wants to know if a set of genetic samples from individuals meets the expectations for the population that was intended to be collected, e.g., that were was no population structure or relatedness in the sample. This is important when analyses are dependent on assumptions of panmixia in a sample. Population structure caused by deviations from panmixia creates groups of individuals within a sample that are more related to each other, meaning that the presence or absence of certain alleles in a sample may actually be due to fact that certain individuals are more ancestrally related to each other, rather than the difference being due to a disease or trait. Thus, cryptic and fine-scale departures from panmixia can cause spurious findings in association analyses for traits. As genetic data sets become larger, differences of this type become more crucial to find the true associations.

To deal with this problem investigators have employed principal components analysis (PCA) methods to parse individuals into groups without having to create *a priori* boundaries on the clusters. However, the axes of variation for PCA are inscrutable. Biological inferences are not easily made. Other investigators have used Bayesian clustering methods to determine the number of genetic clusters in the data with some success (Rosenberg et al., 2002; Rosenberg, Li, Ward, & Pritchard, 2003; Parra et al., 2003; Tero, Aspi, Siikamki, Jklniemi, & Tuomi, 2003; Falush, Stephens, & Pritchard, 2003; Barr et al., 2008; A.C. et al., 2008; Milot, Weimerskirch, & Bernatchez, 2008; Shtir et al., 2009). But *a priori* information is necessary for this method, and recapitulating the number of population categories designated *a priori* may not be useful in a data set that actually contains cryptic structure.

Our method also has a broader use when it is applied to all pairs of individuals in a data set. This biologically coherent method offers a way to explore the data without having to place boundaries on the data *a priori.* To do this, we simply apply the test to all pairs in a data set. Looking at homogeneity estimates (Figure 2.5) and test statistics (Figure 2.6) from all pairs of individuals and in this manner, the method identifies outliers. We can start to make inferences about individual and population relationships within the sample. Further, we can also test for whether the collected sample is a panmictic representative sample of the population. We can also test for cryptic relatedness and structure in the sample which may or may not be desired in the sample by the research investigators. Interestingly, the method also demonstrates that HapMap populations do not represent natural populations very well. There are related individuals between trios in the CEPH and Yoruba populations. There are a number of very significant outliers in the data set as well.

Conclusion

We have presented a valid method for investigating relationships between a single pair of individuals or a sample of individuals. Since our method relies on population genetics principles and bases population membership on the underlying allele frequency distribution, it offers a new way to investigate population structure in a biological context. From genetic information from only two individuals we can test whether the samples are randomly chosen from the same panmictic population. It is also possible to understand more about these outliers by clustering similar homozygosity estimates and/or test statistics. Thus, applying the method across a data set can aid in identifying cryptically related individuals and investigating population structure. Inbreeding, subject misidentification, or loss of heterozygosity can create outliers. This method is an unbiased way to investigate structure and population membership, without *a priori* information from the genetic samples. The understanding of genetic relationships in a sample of individuals is motivated by their use in linkage studies, conservation genetics, forensics, studies of selection, and population structure.

Using the method on HapMap data demonstrates that although the homozygosity estimates, homogeneity test statistics, and contrast statistics exhibit clusters of individuals that tend to fall along the *a priori* designated population categories, the clusters are not strict and are in some cases somewhat continuous. Socially constructed population categories do not perfectly fully reflect the complexity of the underlying genetic relationships.

Genotype Pair					
Individual 1	Individual 2	$Pr\,(genotype\ pair)$	$\hat{\hat{w}}_{1k}$	$\hat{\hat{w}}_{2k}$	$\hat{\hat{b}}_k$
AA	AA	p^4	1	1	1
	Aa	$2p^3q$	1	0	$\frac{1}{2}$
	aa	p^2q^2	1	0	0
Aa	AA	$2p^3q$	0	1	$\frac{1}{2}$
	Aa	$4p^2q^2$	0	0	$\frac{1}{2}$
	aa	$2pq^3$	0	1	$\frac{1}{2}$
aa	AA	p^2q^2	1	1	0
	Aa	$2pq^3$	1	0	$\frac{1}{2}$
	aa	q^4	1	1	1

Table 2.1: Probability estimates for the homozygosity estimators for one locus.

Here, p and q are the allele frequencies at the locus. $\hat{\hat{w}}_{1k}$ is the genome-wide homozygosity estimate for the first individual, and $\hat{\hat{w}}_{2k}$ is the genome-wide homozygosity estimate for the second individual. $\hat{\hat{b}}_k$ is the genome-wide homozygosity estimate from between the two individuals.

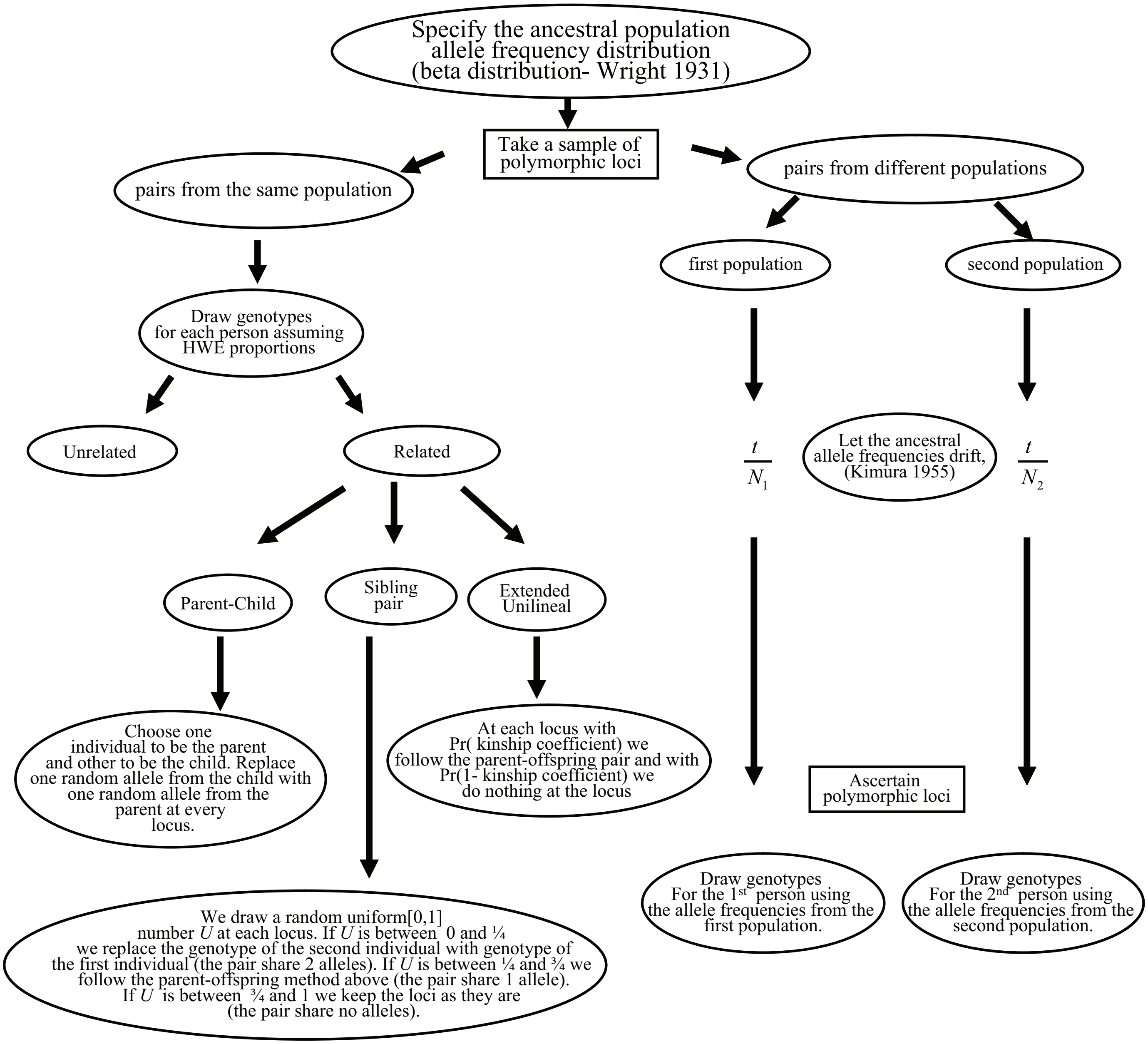
Specify the ancestral population allele frequency distribution (beta distribution- Wright 1931)
Take a sample of polymorphic loci
pairs from the same population
pairs from different populations
Draw genotypes for each person assuming HWE proportions
Unrelated
Related
Parent-Child
Sibling pair
Extended Unilineal
Choose one individual to be the parent and other to be the child. Replace one random allele from the child with one random allele from the parent at every locus.
At each locus with Pr(kinship coefficient) we follow the parent-offspring pair and with Pr(1- kinship coefficient) we do nothing at the locus
We draw a random uniform[0,1] number U at each locus. If U is between 0 and ¼ we replace the genotype of the second individual with genotype of the first individual (the pair share 2 alleles). If U is between ¼ and ¾ we follow the parent-offspring method above (the pair share 1 allele). If U is between ¾ and 1 we keep the loci as they are (the pair share no alleles).
first population
second population
$\frac{t}{N_1}$
Let the ancestral allele frequencies drift, (Kimura 1955)
$\frac{t}{N_2}$
Ascertain polymorphic loci
Draw genotypes For the 1st person using the allele frequencies from the first population.
Draw genotypes For the 2nd person using the allele frequencies from the second population.

Figure 2.1: Overview of the muti-locus genotype simulations.

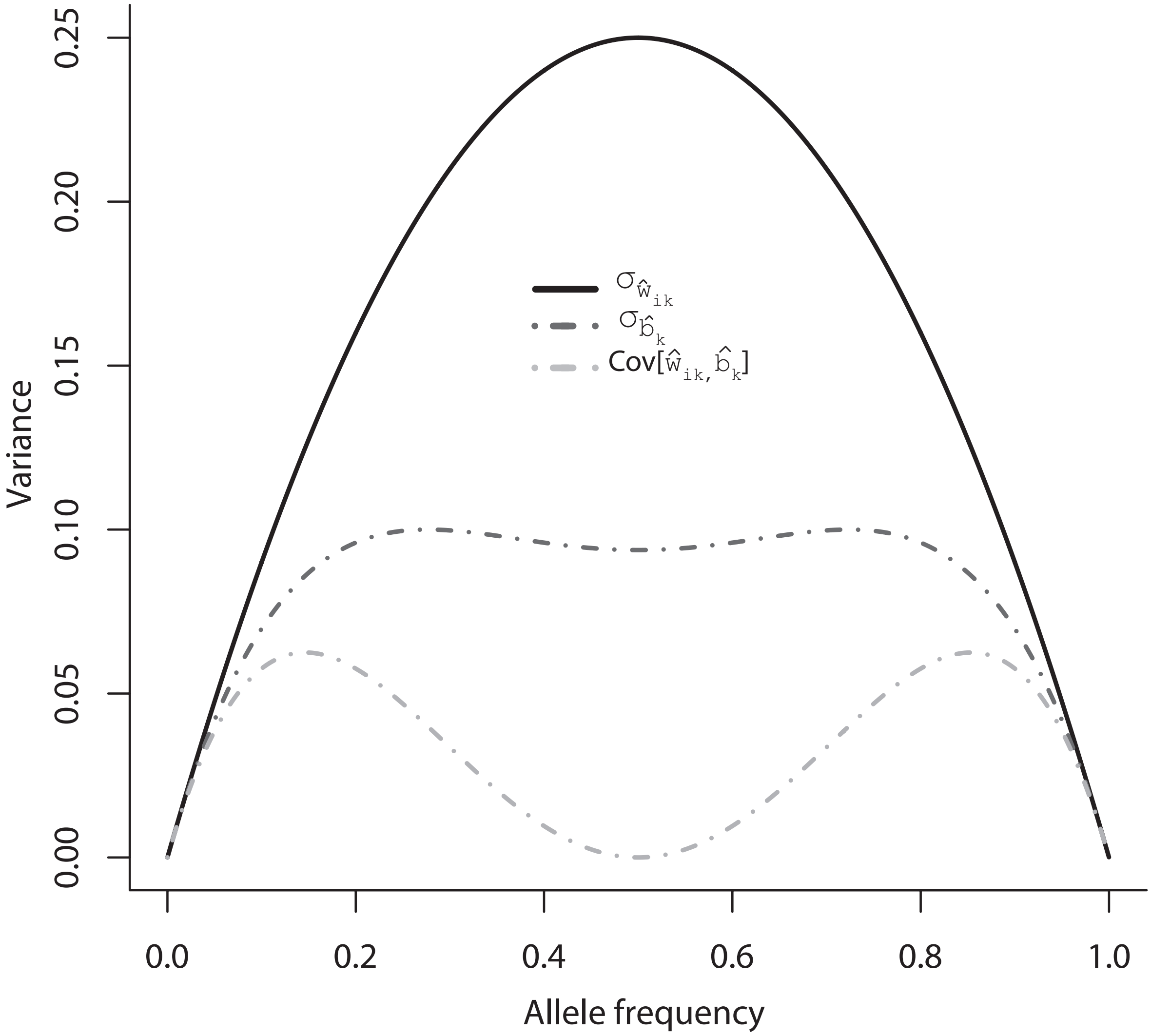

Figure 2.2: Variance and covariance of the estimators. Here w_{ik} is the homozygosity for individual i, at the kth locus. b_k is the homozygosity from between the two individuals at the kth locus. These variances are a function of the allele frequency, as given in 2.12 and 2.14.

	Loci need to achieve power of:	
Simulated Pair Type	80%	100%
Parent offspring	$\leq 1,000$	$\leq 1,000$
Full-sibling pair	$\leq 1,000$	$\leq 1,000$
First-cousin pair	10,000	20,000
Second-cousin pair	100,000	125,000
CEU-CHB like [a]	$\leq 1,000$	$\leq 1,000$
CEU-JPT like [a]	$\leq 1,000$	$\leq 1,000$
CEU-YRI like [a]	$\leq 1,000$	$\leq 1,000$
CHB-JPT like [a]	90,000	125,000
CHB-YRI like [a]	$\leq 1,000$	$\leq 1,000$
JPT-YRI like [a]	$\leq 1,000$	$\leq 1,000$
CEU-CHB empirical [b]	$\leq 1,000$	$\leq 1,000$
CEU-JPT empirical [b]	$\leq 1,000$	$\leq 1,000$
CEU-YRI empirical [b]	$\leq 1,000$	$\leq 1,000$
CHB-JPT empirical [b]	25,000	35,000
CHB-YRI empirical [b]	$\leq 1,000$	$\leq 1,000$
JPT-YRI empirical [b]	$\leq 1,000$	$\leq 1,000$

[a] Simulated as demonstrated in Chapter four:
$4Nu$ and $4Nv$ for the ancestral distribution were 0.00048 and 0.00082, respectively.
[b] Simulated from empirical HapMap frequencies.

Table 2.2: Power of the homogeneity test calculated from simulated data.

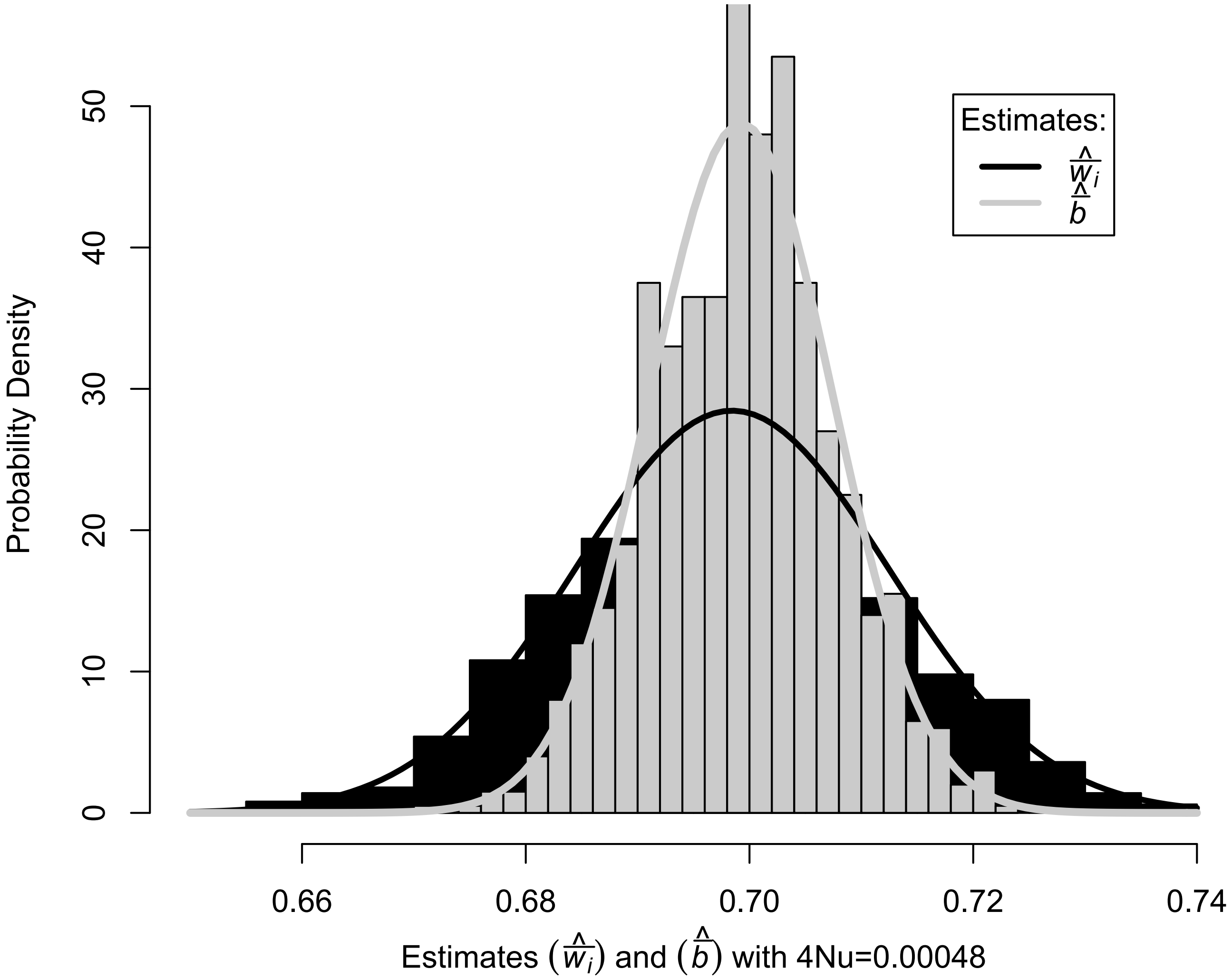

Figure 2.3: Histogram of the simulated homozygosity estimates.

We simulated genotypes at 100 loci for 1000 individuals from an equilibrium allele frequency distribution where $4Nu$=0.00048, and then estimated homozygosity for each individual ($\hat{w}_{ik}$) and homozygosity between each pair of individuals $\hat{\hat{b}}_k$. The two statistics follow the normal distribution (non significant Kolmogorov-Smirnov test). We obtained similar results for a wide range of $4Nu$.

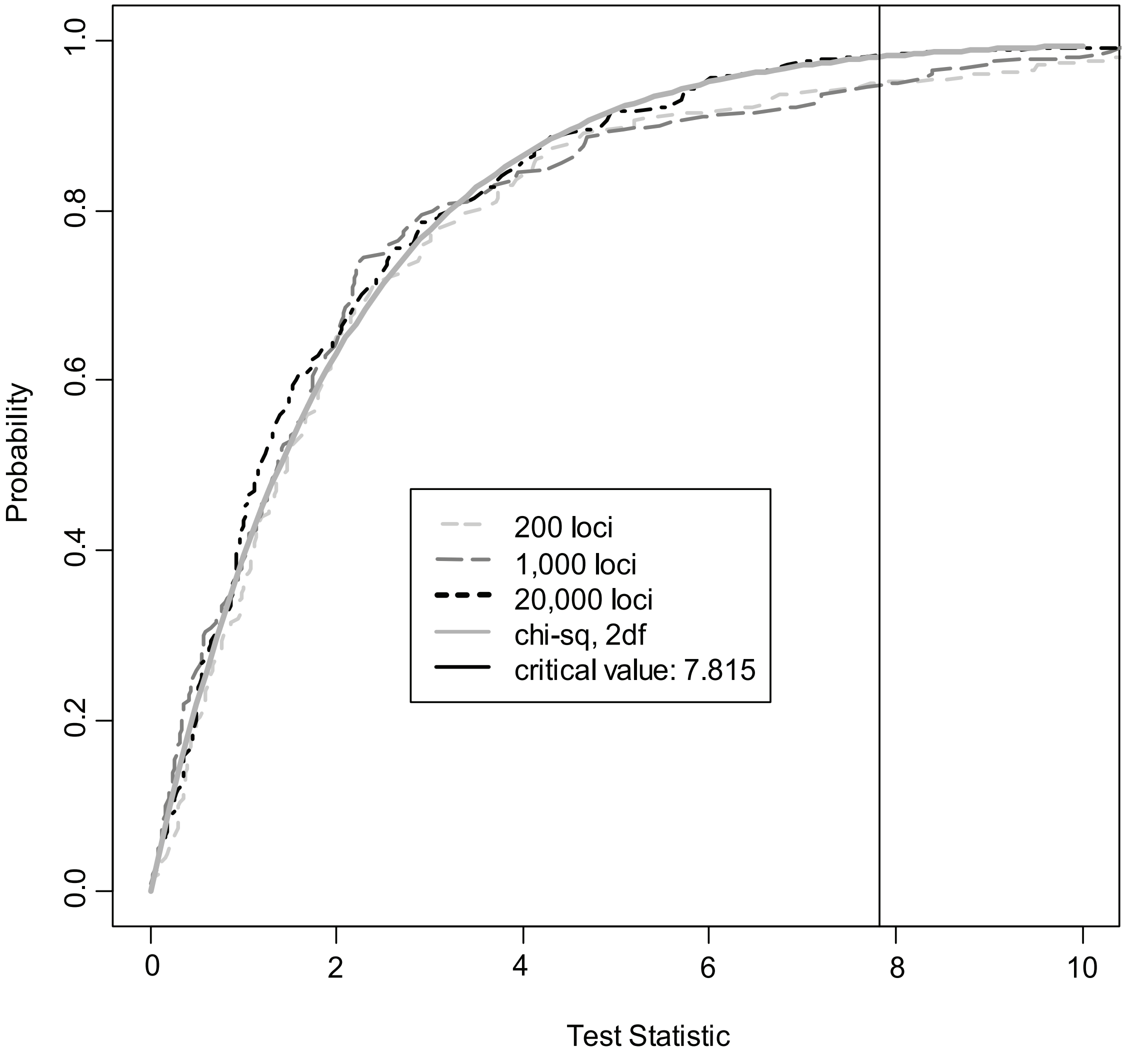

Figure 2.4: Cumulative distribution of the test statistics.

T follows a chi-square distribution with two degrees of freedom, as predicted. The homogeneity statistic is calculated from individuals simulated from an equilibrium allele distribution with $4Nu$=0.00048.

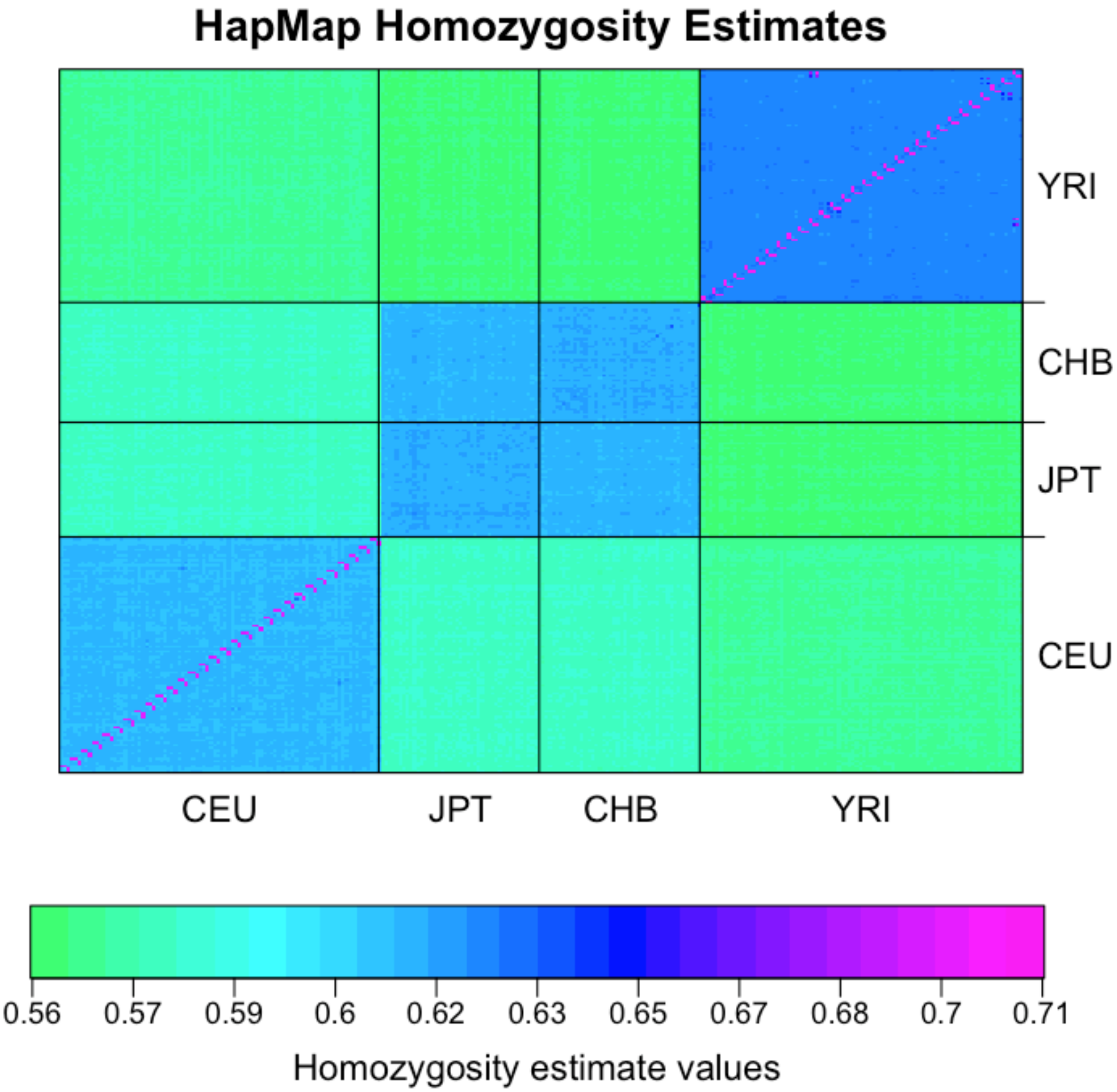

Figure 2.5: Homozygosity estimates calculated from 100,000 SNPS from all pairs of HapMap individuals.

Homozygosity estimates from within (diagonal) individuals and between pairs of individuals (off-diagonal) are demonstrated for 100k loci from the HapMap data. The matrix shown is symmetrical. The populations included here are the Yoruba from Ibadan, Nigeria (YRI), 45 unrelated Japanese from Tokyo, Japan (JPT), 45 unrelated Han Chinese from Beijing, China (CHB), and the CEPH, Utah residents with ancestry from northern and western Europe (CEU). These estimates clearly identify mother-father-offspring trios (pink dots). Note that the JPT and CHB populations are more closely-related to each other than they are other populations included in the sample (blue block of the between homozygosity estimates as compared with green blocks of comparisons between either one of the Asian populations, JPT or CHB, and the CEU or YRI). Additionally, there are several outlier individuals in each population (dots of very different colors located within the blocks of the between homozygosity estimates from the same population).

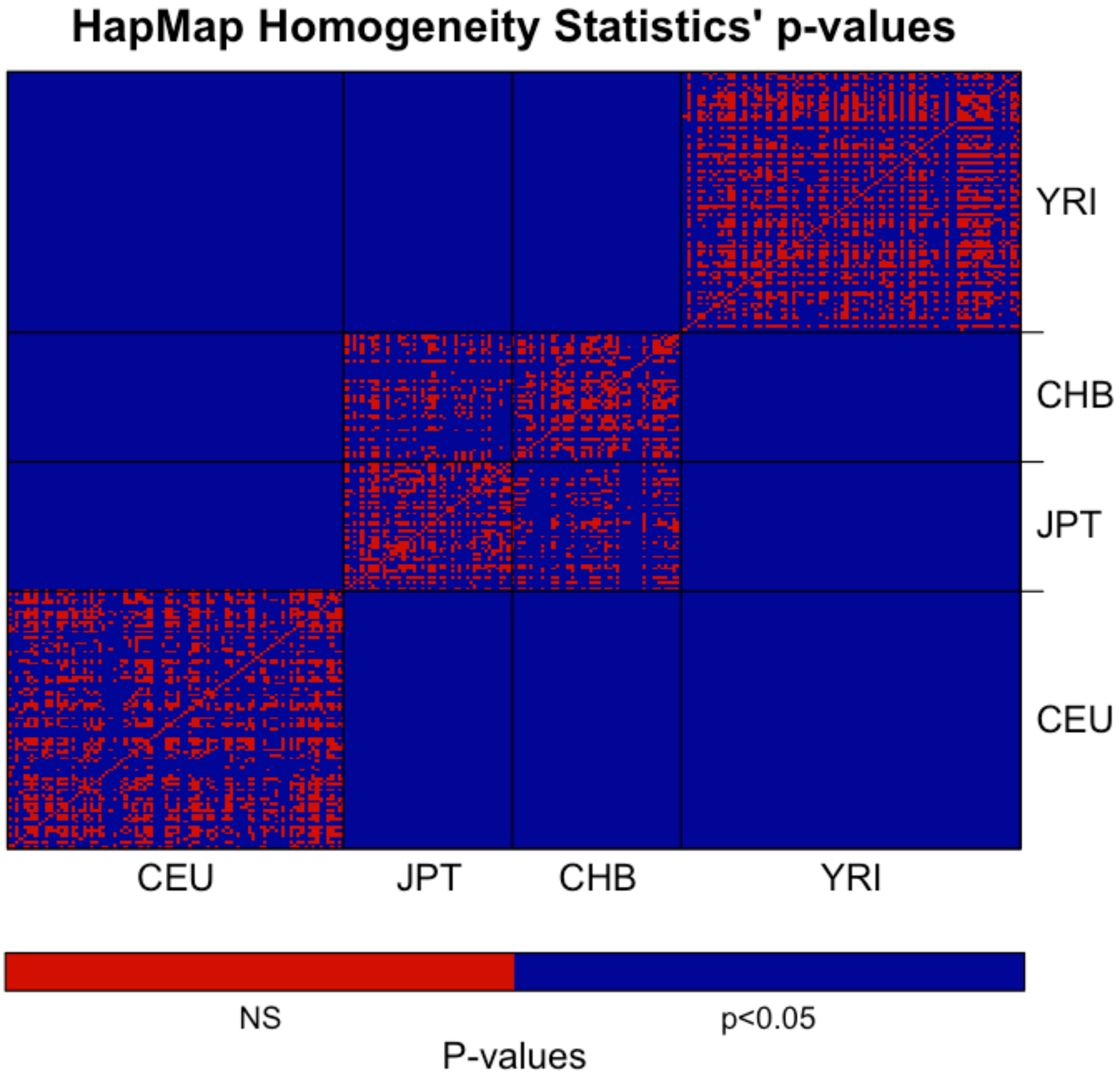

Figure 2.6: Homozygosity test statistics (T) taken from 100,000 SNPS from all pairs of HapMap individuals.

Colors represent p-values of the test statistics. Again the matrix is symmetrical. The SNP data is from HapMap; Yoruba from Ibadan, Nigeria (YRI), 45 unrelated Japanese from Tokyo, Japan (JPT), 45 unrelated Han Chinese from Beijing, China (CHB), and the CEPH, Utah residents with ancestry from northern and western Europe (CEU). Test statistics from pairs from the same population tend to be non-significant, whereas test statistics from pairs of different populations are significant. However, test statistics calculated from between the JPT and CHB pairs often fail to reject the test. The test statistic also accurately identifies known relatives with the CEU and YRI populations. A number of outliers appear in each population as blue dots within the red blocks. Inbreeding, subject misidentification, or loss of heterozygosity can create outliers.

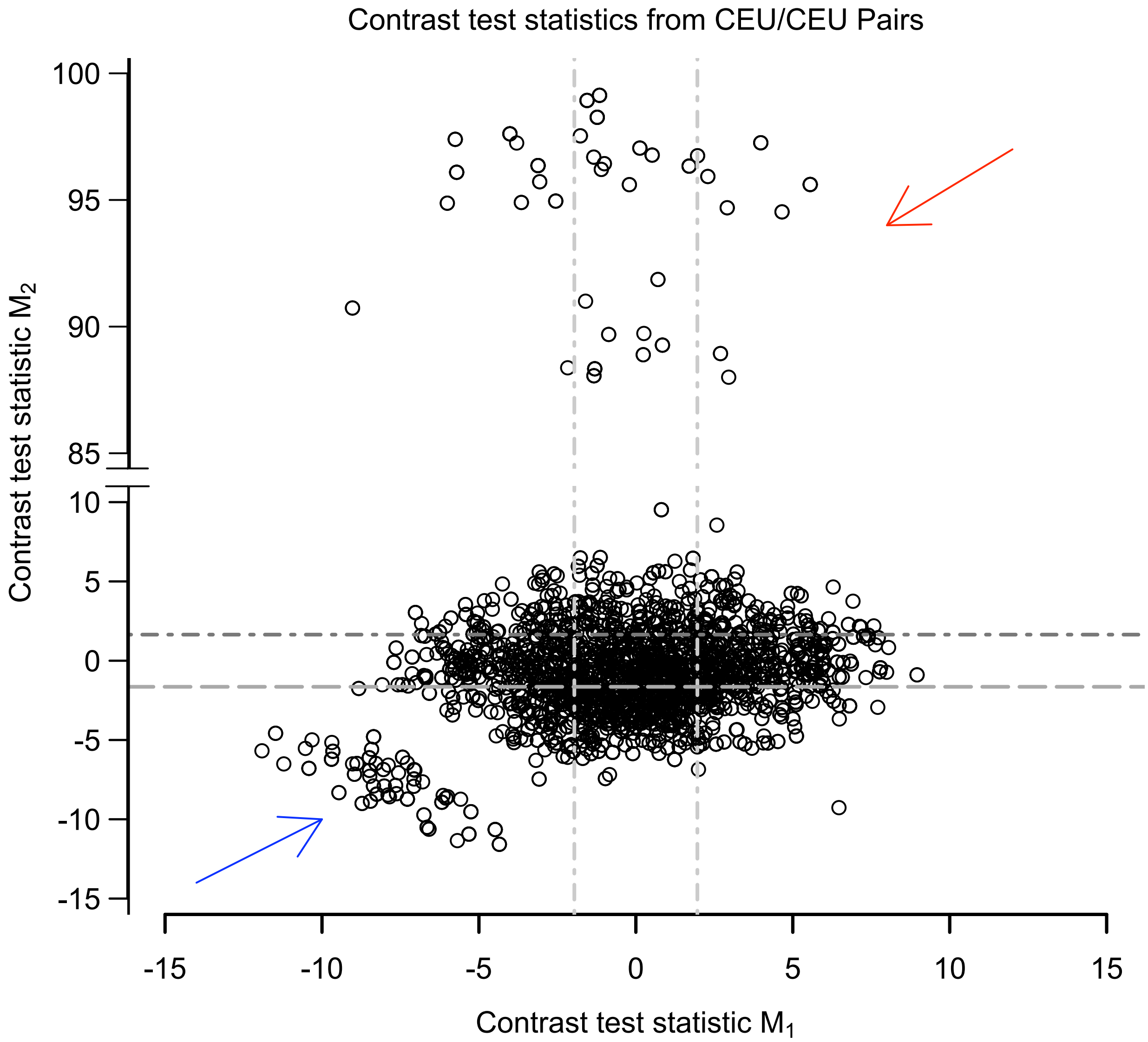

Figure 2.7: Contrast test statistics from CEPH European (CEU/CEU) pairs.

Contrast test statistics for the pairs from the HapMap. Grey vertical lines represent the critical values (±1.96) for $H_{0_1} : \hat{\bar{w}}_1 = \hat{\bar{w}}_2$, while the lower light grey line horizontal line is the critical value (-1.645) for the one sided test $\hat{\bar{b}} < \hat{\bar{w}}$ (different populations or inbred from the same populations) and the upper light grey line horizontal line is the critical value (1.645) for the one sided test $\hat{\bar{b}} > \hat{\bar{w}}$ (the pair are related). Test statistics from CEU/CEU pairs show two clusters of related individuals (the top of the graph). The contrasts statistics easily distinguish the trios in the CEPH European HapMap data (red arrow points to the two clusters). There are also a number of outliers in the data set (blue arrow).

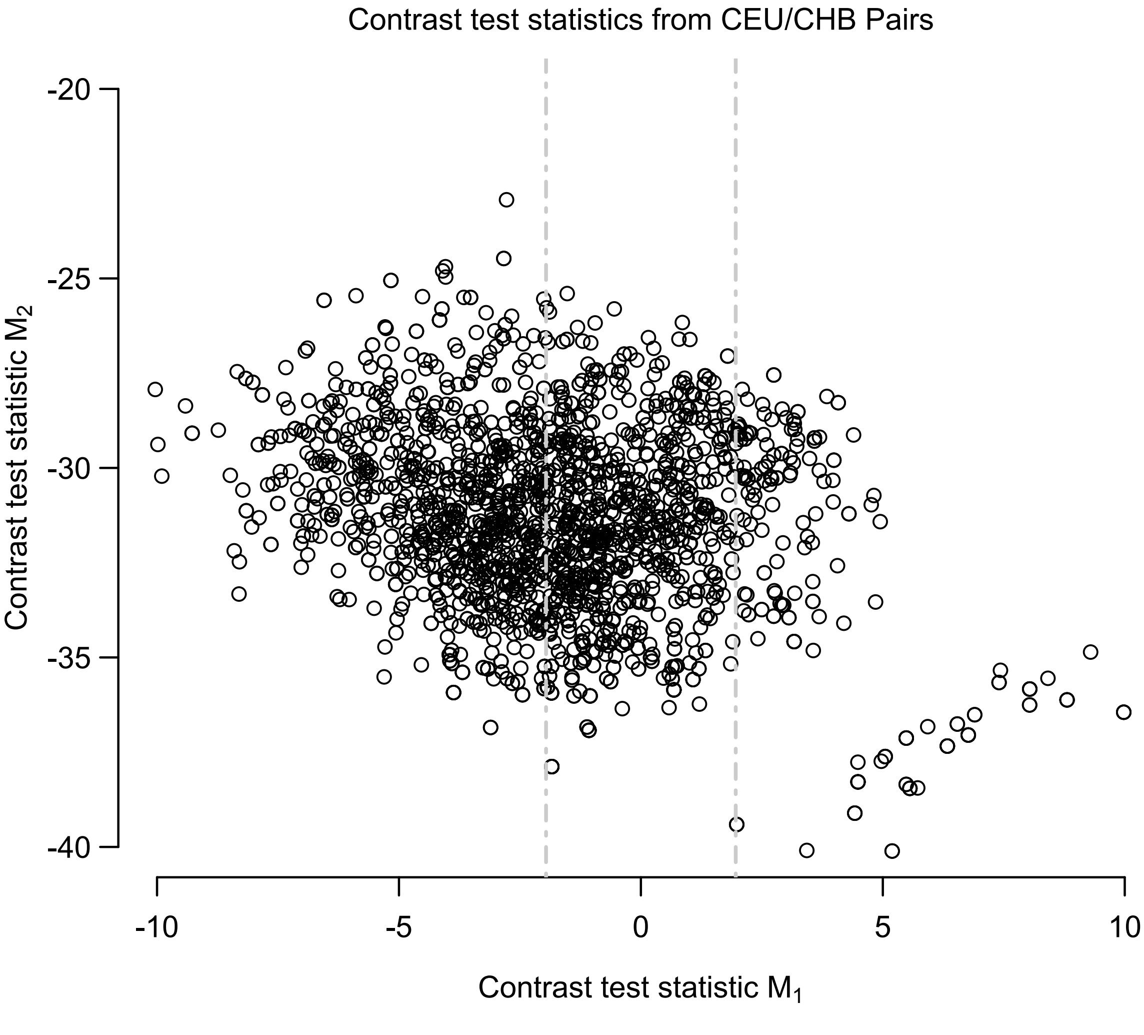

Figure 2.8: Contrast test statistics from CEPH European and Han Chinese (CEU/CHB) pairs.

Contrast test statistics for the pairs from the HapMap. Grey vertical lines represent the critical values (±1.96) for $H_{0_1} : \hat{\bar{w}}_1 = \hat{\bar{w}}_2$. Test statistics from CEU/CHB pairs distinguish between populations.

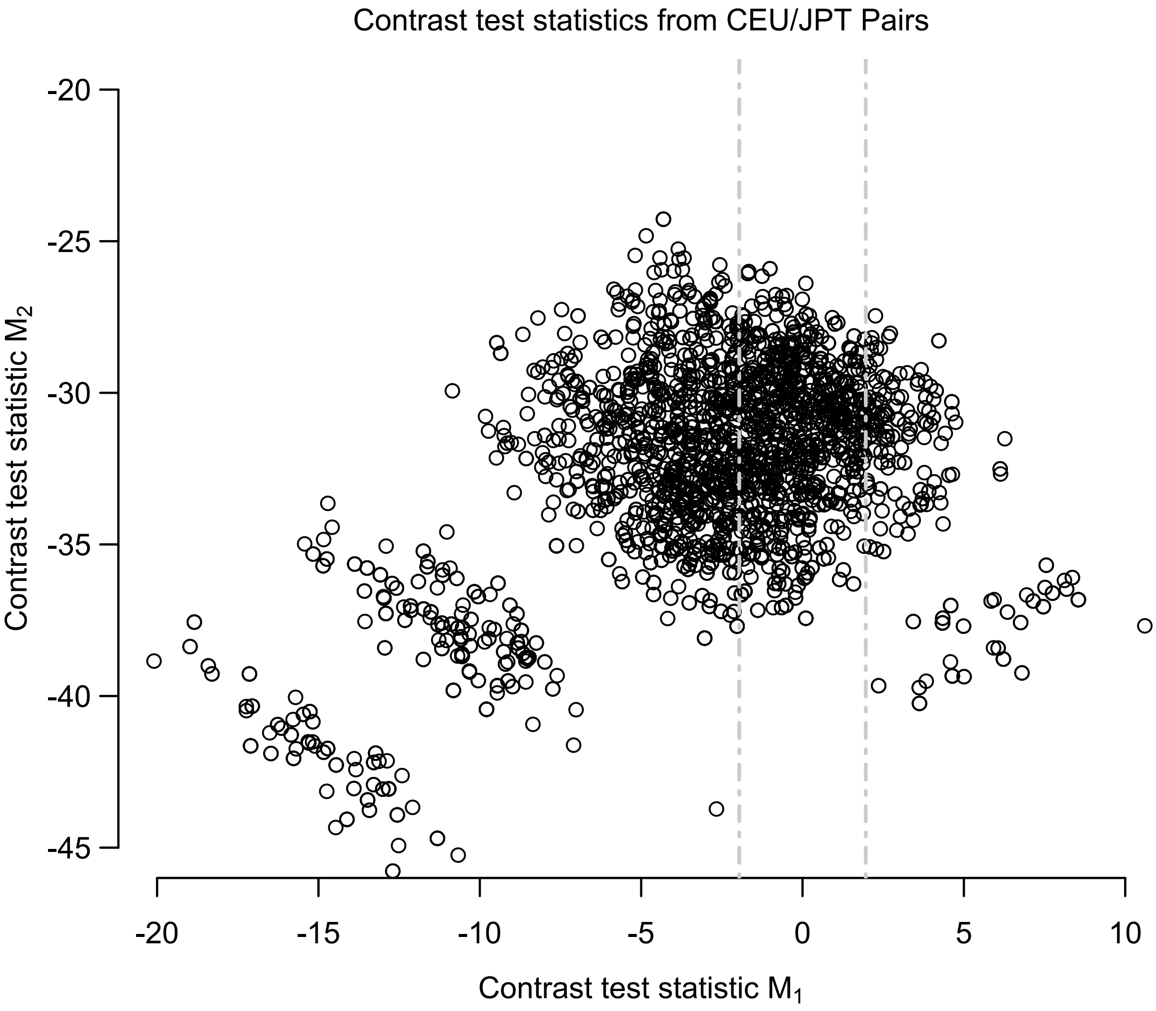

Figure 2.9: Contrast test statistics from CEPH European and Japanese (CEU/JPT) pairs.

Contrast test statistics for the pairs from the HapMap. Grey vertical lines represent the critical values (±1.96) for $H_{0_1} : \hat{\bar{w}}_1 = \hat{\bar{w}}_2$. Test statistics from CEU/JPT pairs distinguish between the populations, as the statistics are all well below the critical value of 1.645 for $\bar{b} < \bar{w}$. There are several groups of outlier statistics from pairs of individuals. Given the outliers in each of these populations separately, we might conclude that once these individuals are paired their statistics might be larger giving the clusters of outlier statistics.

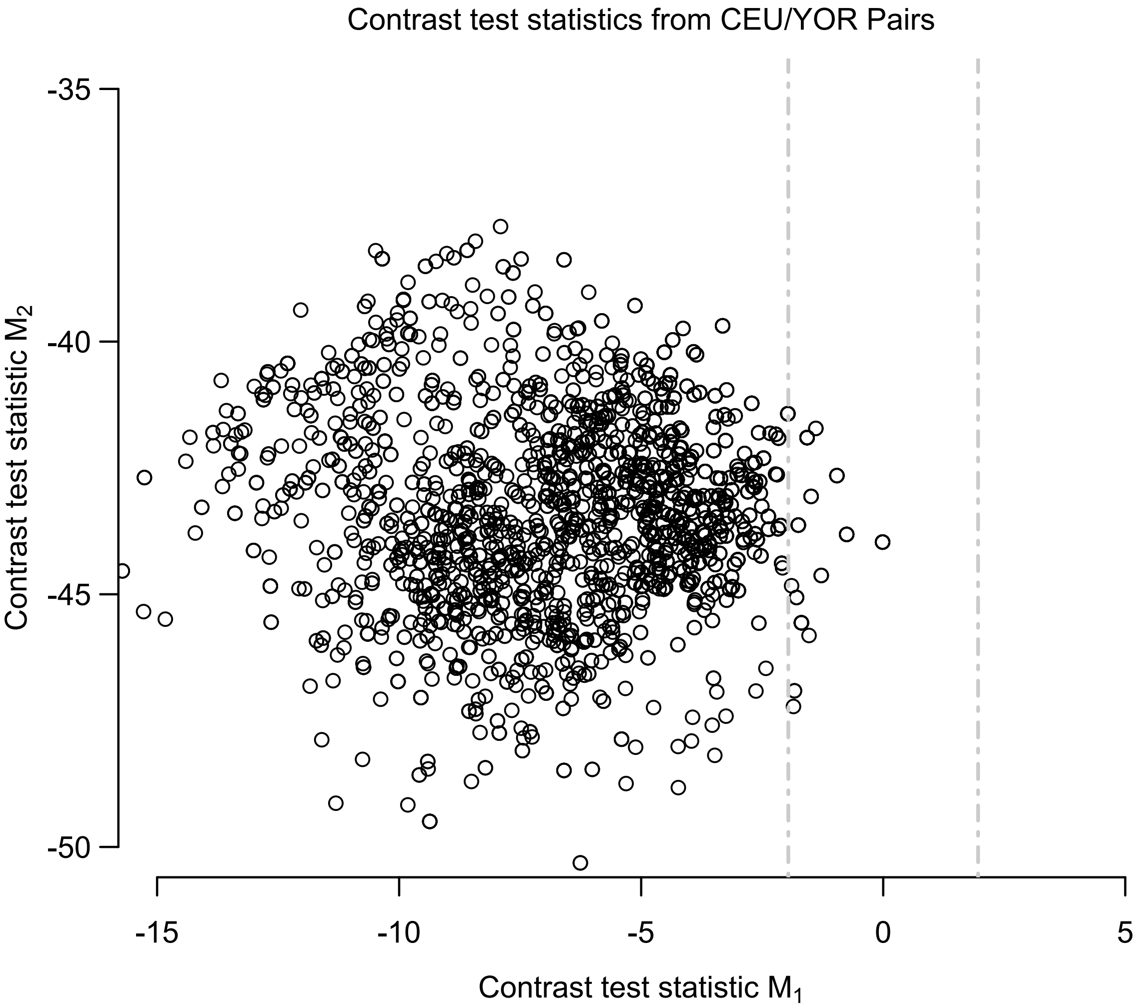

Figure 2.10: Contrast test statistics from CEPH European and Yoruba (CEU/YRI) pairs.

Contrast test statistics for the pairs from the HapMap. Grey vertical lines represent the critical values (±1.96) for $H_{0_1} : \hat{\hat{w}}_1 = \hat{\hat{w}}_2$. For test statistics from CEU/YRI pairs, we find that one sided test $\bar{b} < \bar{w}$ are all significant, so that all the statistics are far below the critical value of -1.645.

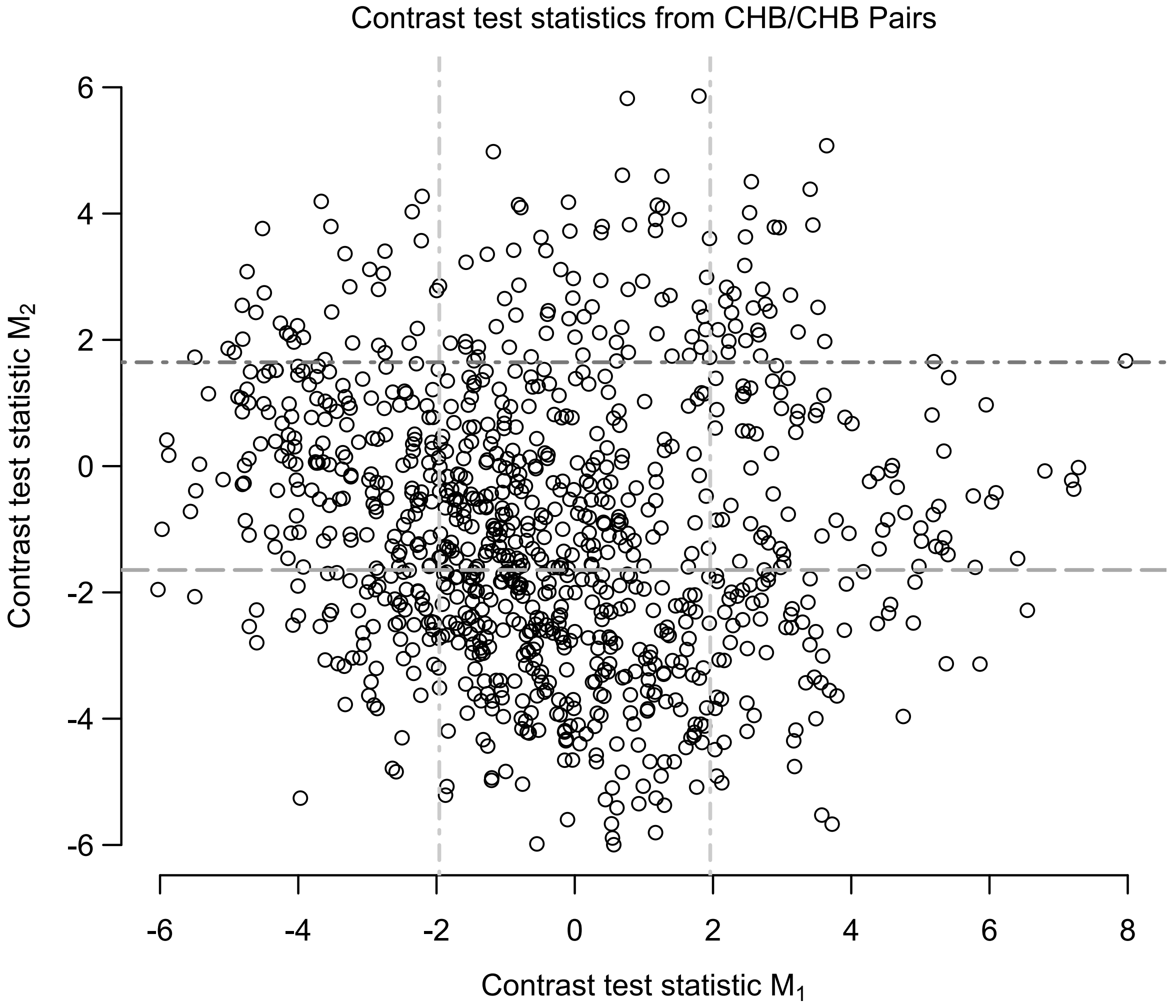

Figure 2.11: Contrast test statistics from Han Chinese (CHB/CHB) pairs.

Contrast test statistics for the pairs from the HapMap. Grey vertical lines represent the critical values (±1.96) for $H_{0_1} : \hat{\hat{w}}_1 = \hat{\hat{w}}_2$, while the lower light grey line horizontal line is the critical value (-1.645) for the one sided test $\hat{\hat{b}} < \hat{\hat{w}}$ (different populations or inbred from the same populations) and the upper light grey line horizontal line is the critical value (1.645) for the one sided test $\hat{\hat{b}} > \hat{\hat{w}}$ (the pair are related). For test statistics from the CHB/CHB pairs, we find that we fail to reject either hypothesis for many of the pairs (the middle box).

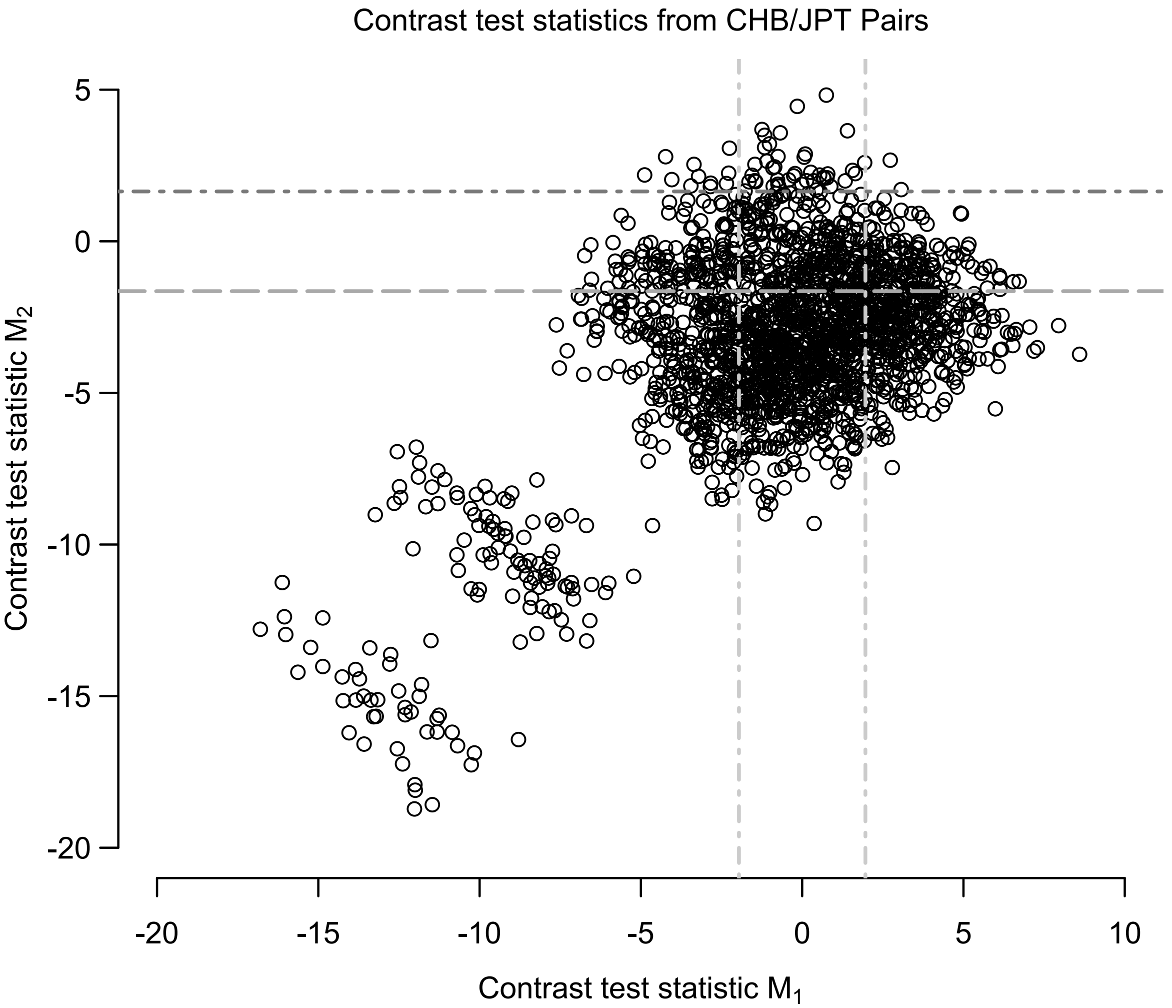

Figure 2.12: Contrast test statistics from Han Chinese and Japanese (CHB/JPT) pairs.

Contrast test statistics for the pairs from the HapMap. Grey vertical lines represent the critical values (±1.96) for $H_{0_1} : \hat{\hat{w}}_1 = \hat{\hat{w}}_2$, while the lower light grey line horizontal line is the critical value (-1.645) for the one sided test $\hat{\hat{b}} < \hat{\hat{w}}$ (different populations or inbred from the same populations) and the upper light grey line horizontal line is the critical value (1.645) for the one sided test $\hat{\hat{b}} > \hat{\hat{w}}$ (the pair are related). The test fails to detect differences for many CHB/JPT pairs. CHB /JPT pairs show a cluster of individuals that would be distinguished as being from different populations. There are some many clear outlier statistics in the data set showing that some individuals are outliers in the sample.

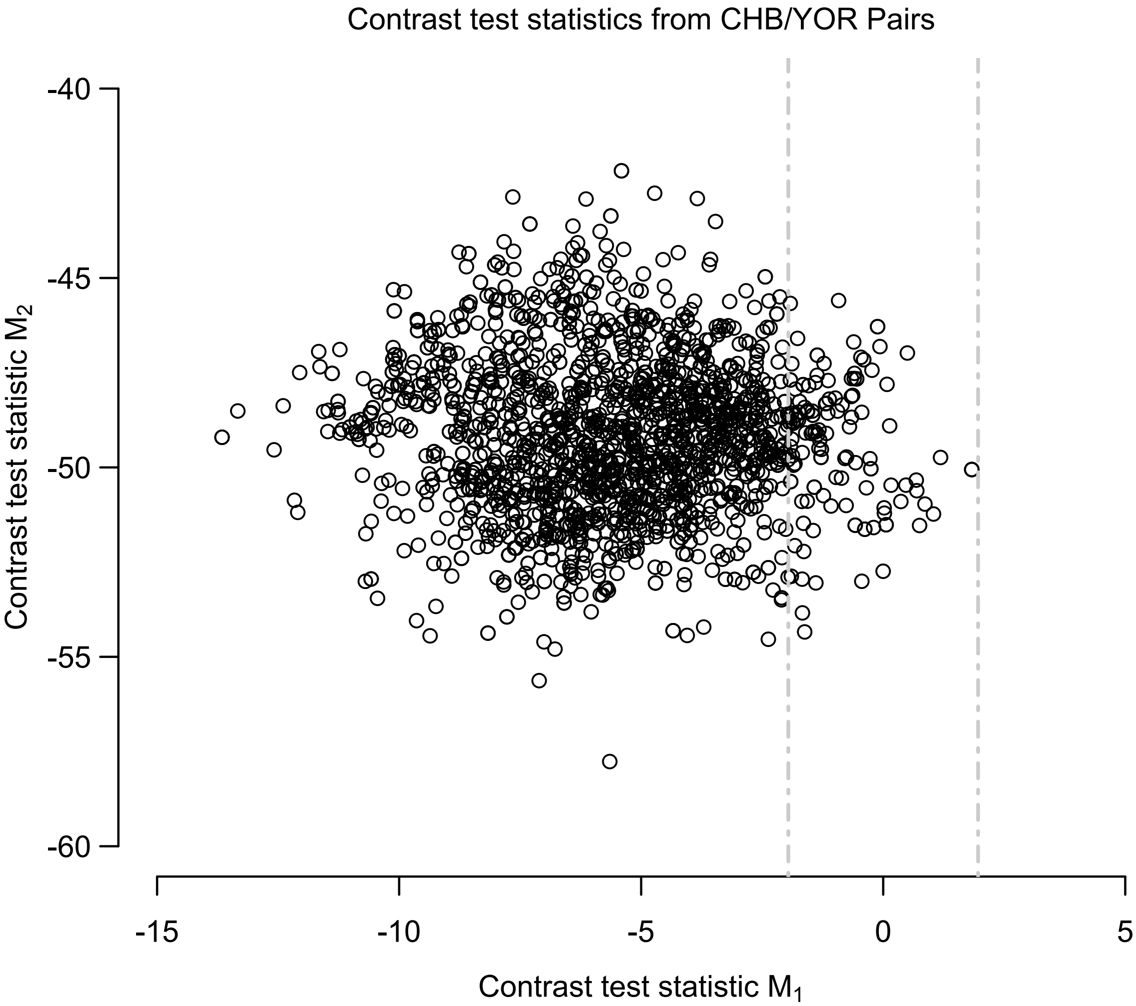

Figure 2.13: Contrast test statistics from Han Chinese and Yoruba (CHB/YRI) pairs.

Contrast test statistics for the pairs from the HapMap. Grey vertical lines represent the critical values (±1.96) for $H_{0_1} : \hat{\bar{w}}_1 = \hat{\bar{w}}_2$. Here we see that the one sided test $\bar{b} < \bar{w}$ statistics are all significant, so that all the statistics are far below the critical value of -1.645. Additionally we reject $\hat{\bar{w}}_1 = \hat{\bar{w}}_2$ for many of the pairs. The test easily distinguishes between Chinese and Yoruban individuals.

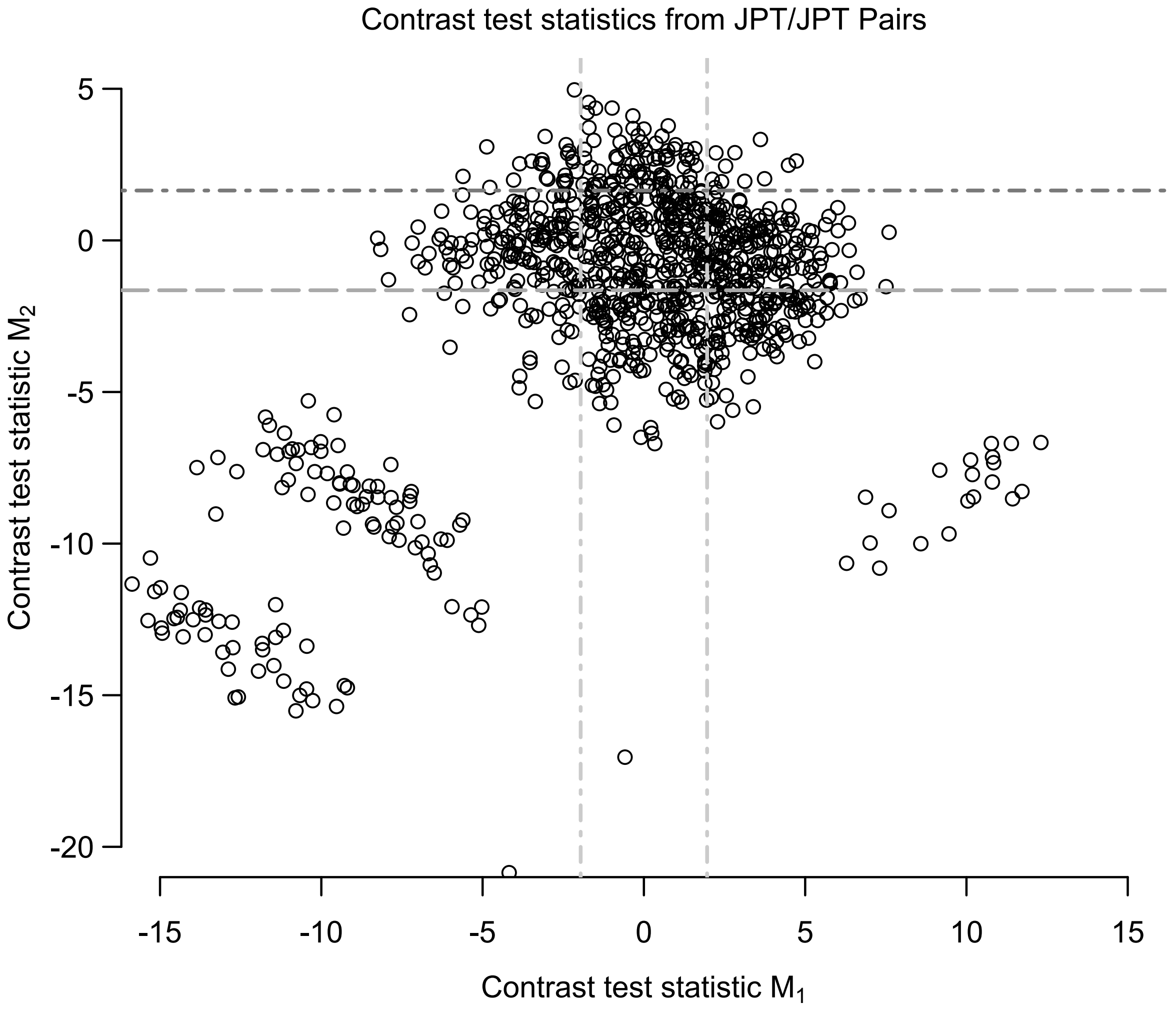

Figure 2.14: Contrast test statistics from Japanese (JPT/JPT) pairs.

Contrast test statistics for the pairs from the HapMap. Grey vertical lines represent the critical values (±1.96) for $H_{0_1} : \hat{\hat{w}}_1 = \hat{\hat{w}}_2$, while the lower light grey line horizontal line is the critical value (-1.645) for the one sided test $\hat{\hat{b}} < \hat{\hat{w}}$ (different populations or inbred from the same populations) and the upper light grey line horizontal line is the critical value (1.645) for the one sided test $\hat{\hat{b}} > \hat{\hat{w}}$ (the pair are related). Test statistics from JPT/JPT pairs demonstrate outliers in the data set.

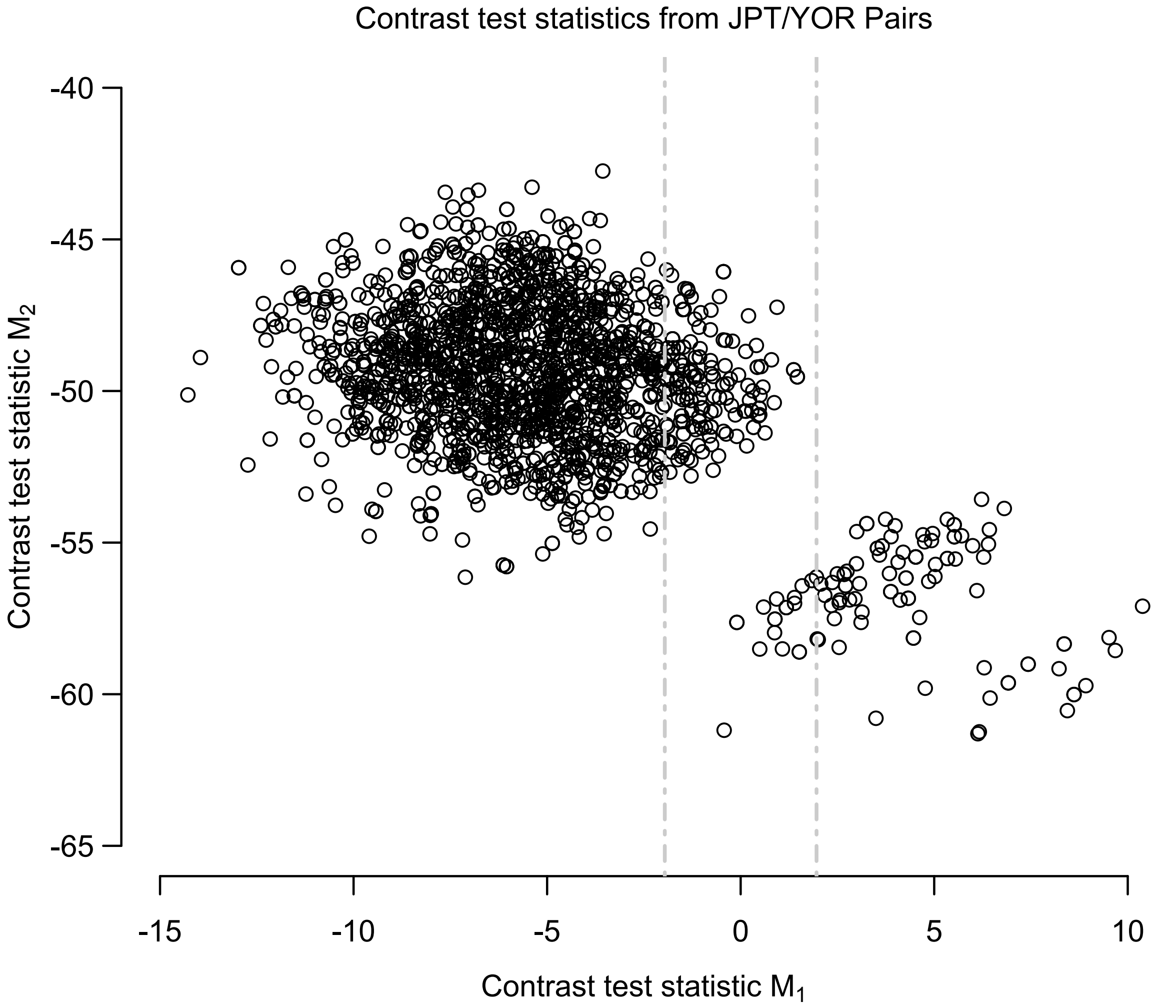

Figure 2.15: Contrast test statistics from Japanese and Yoruba (JPT/YRI) pairs.

Contrast test statistics for the pairs from the HapMap for JPT/YRI pairs. Grey vertical lines represent the critical values for $H_{0_1} : \hat{\hat{w}}_1 = \hat{\hat{w}}_2$. The test statistics are are all much less than the critical value of 1.645 for the one sided test $\hat{\hat{b}} < \hat{w}$ (different populations or inbred from the same populations). Test statistics from JPT/YRI pairs demonstrate individuals in the pair are from different populations or inbred from the same population. We also see outlier statistics.

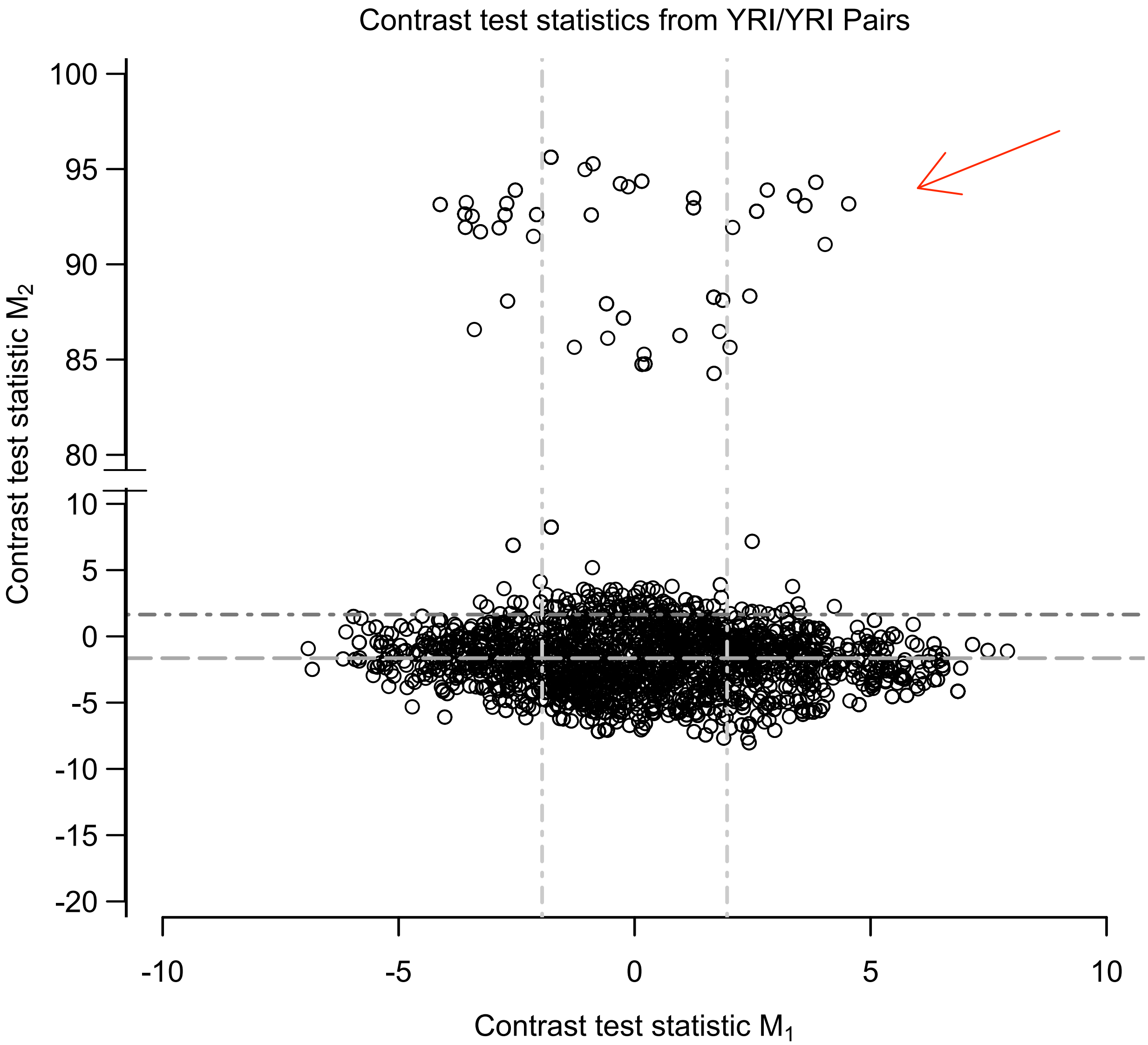

Figure 2.16: Contrast test statistics from Yoruba (YRI/YRI) pairs.

Contrast test statistics for the pairs from the HapMap for YRI/YRI pairs. Grey vertical lines represent the critical values for $H_{0_1} : \hat{\hat{w}}_1 = \hat{\hat{w}}_2$. The test statistics are are all much less than the critical value of -1.645 for the one sided test $\hat{\hat{b}} < \hat{\hat{w}}$ (different populations or inbred from the same populations). Test statistics from YRI/YRI pairs demonstrate the relative pairs (the topmost group of statistics denoted by the red arrow). The contrast$\left(\hat{\hat{b}} - \hat{\hat{w}}\right)$ easily distinguishes relatives.

CHAPTER III

The effects of ancestry and ascertainment on methods of assessing population structure

3.1 Abstract

Here using data simulated from a known ancestral tree, we evaluate the effects of the underlying ancestral relationships, ascertainment of loci, and misspecification of the models on methods designed to assess population structure. We first present a novel forward-time strategy for simulating allele frequencies from different populations using a population genetic model that draws on methods from diffusion theory. The simulation method uses population genetics theory to model genetic drift on allele frequencies in a descendent population given an ancestral population. To do this we first take a sample of loci from a chosen ancestral allele frequency distribution. Then we allow the resulting descendent allele frequencies to drift from the ancestral sample of loci using Kimura's equation for the probability density of allele frequencies under neutral genetic processes (ϕ) (Kimura, 1955) to make a hierarchically structured population. We also input estimates of the time in generations to the ancestor, and the effective population size of the descendent population. After the hierarchically structured population has been simulated, the resulting descendent allele frequencies can be ascertained according to different schemes that mimic available genetic data. We then use simulated data sets with a method that tests for

a phylogenetic structure, GHM (generalized hierarchical modeling), and one which does not test for a phylogenetic structure, but searches for clusters in the data, using the STRUCTURE program. We also investigate the effects of different methods of ascertainment on the simulated data.

3.2 Introduction

Genetic variation is shared among all populations in a species. However, geographical and cultural factors cause non-random mating and in turn create subdivisions within a species. This phenomenon is known as population structure. Individuals chosen from the same subdivision are typically more related to each other and thus share more of the same alleles than we would expect if we chose individuals randomly from the population. Population structure creates spurious associations between genes and phenotypes, because the genetic marker may be marking the factor that caused the population structure rather than a causitive genetic factor for the phenotype. Correctly identifying population structure is of interest because it can help avoid spurious findings in clinical studies, and it may reveal unknown aspects of evolutionary history.

One such method to examine population structure, is a Bayesian model-based clustering implemented in the STRUCTURE program (J. Pritchard et al., 2000; Falush, Stephens, & Pritchard, 2003; Falush, Wirth, et al., 2003; Hubisz et al., 2009), has become a valuable tool for assessing population structure under a number of varying modeling conditions and has been used to identify structure in clinical samples, make inferences about population history, landscape features, and breeding structure (Seldin et al., 2006; Bauchet et al., 2007; Latch, Scognamillo, Fike, Chamberlain, & Rhodes, 2008; Coop et al., 2009; Latch, Heffelfinger, Fike, & Rhodes, 2009; Schwartz & McKelvey, 2009; Song et al., 2009). Though this method is statistically model-based, it tests one very simple model, which may or may not represent the data. That is, it does not test for a particular underlying phylogenetic model of population relationships.

Another example of a method to assess the ancestry and population structure underlying a sample is hierarchical modeling. In this method, the data is tested for fit to an *a priori* model of population relationships from a user supplied phylogenetic model (tree) of the genetic relationships (Cavalli-Sforza & Piazza, 1975). This method is applied using the Generalized Hierarchical Modeling program (GHM) (Long, 2007).

There have been a number studies focusing on the effects of migration and spatial structure on STRUCTURE's ability to detect population structure (Evanno, Regnaut, & Goudet, 2005; Latch et al., 2006; Schwartz & McKelvey, 2009). Using θ, or the reduction in heterozygosity in the total population due to subdivision (Weir & Cockerham, 1984) as a measure of population differentiation, Latch et al. (2006) suggested that STRUCTURE could correctly identify the number of true populations in a simulated data set with 97% accuracy with θ as low as 0.05 between simulated populations. The simulation scheme in Latch et al. (2006) consisted of iteratively drawing founder genotypes from a uniform allele frequency distribution until a particular θ was reached. Random genotypes were then drawn from each 'population' with a particular θ. Unfortunately, because this method draws its founders from a uniform distribution and then separately 'drifts' each subpopulation, there is no correlation between the resulting populations. However, in another simulation study, Evanno et al. (2005) could not easily infer the correct K for simulated data and found it necessary to devise another method to correctly infer the K clusters. The simulation scheme in Evanno et al. (2005) used three models: island, hierarchical, and contact zone with a K allele model for mutation, resulting in θ between groups from 0.16 to 0.4. Rosenberg et al. (2005) investigated sampling, sample size, and the number of loci using STRUCTURE on a worldwide genetic short tandem repeat (STR) data set. Sample size and the number of loci influenced how well STRUCTURE could fully classify individuals into a cluster. Interestingly, samples taken from differeing locations seemed to neagtively affect STRUCTURE's ability to classify individuals into distinct clusters, yet it was able to recapitulate the population catagories rather consistently.

In order to better evaluate the effect of population structure in a continuous population and sampling schemes of individuals on STRUCTURE clustering, Schwartz and McKelvey (2009) used a grid scheme to simulate genotypes. Their results demonstrated that using STRUCTURE to make inferences about the geographic and landscape features of the data may lead to erroneous results. However, their simulations attempted to create "autocorrelations" in the data by using grid and neighbor mating over 20 generations to produce substructure and enable different sampling schemes. The simulations had short ancestral times (rather than deep time) and assumed a linear relationship between geography and genetic variation. They also used only 15 loci with 7 alleles. Even with the scenarios tested from multiple studies, there is still a need for more investigations into the number of loci, SNPs, and additional sampling schemes (Schwartz & McKelvey, 2009).

Generalized hierarchical modeling is a simple but powerful method that relies on estimating gene identity (Nei, 1987) and testing a hierarchical model's fit to the data (Cavalli-Sforza & Piazza, 1975). Gene identity is the probability that two randomly chosen copies of a locus are identical in state. An unbiased estimate of gene identity from an individual is the proportion of homozygous loci in their genome and an estimate of gene identity between two individuals is the probability of a homozygous "offspring". A matrix is made from all possible pairwise estimates of gene identity, with estimates from within individuals on the diagonal and estimates between individuals on the off-diagonal. The gene identity matrix is then treated as a variance-covariance matrix, and a likelihood ratio test is used to statistically test the fit of an *a priori* model to the data (Cavalli-Sforza & Piazza, 1975; Long & Kittles, 2003). Urbanek et al. (1996) investigated the effect of differeing hiearchial models and ascertainment on estimated gene diversity using a STR data set from a number of populations. Their results suggested that ascertainment could affect the relative genetic differentiation between populations, and thus possibly alter the results from models that are evaluated. In general as a method to investigate population structure, GHM only tests the data against a user provided model, and thus is limited to the models that are supplied. Further, a higher likelihood for a model amongst a number of other models, does not exclude the possibility of an untested model having a higher likelihood. Evaluations of the impact of spatial sampling and ascertainment of loci on the generalized hierarchical model has yet to be accomplished.

In addition to time, spatial structure, and migration as factors that affect the population structure in empirical data, ascertainment of loci also adds complications to understanding population structure. Because we define a biological population as a group of interbreeding individuals that creates an underlying distribution of allele frequencies, the ascertainment of specific extant loci with biased characteristics may also be of concern for the analyst of population structure. Ascertainment may skew the actual distribution of allele frequencies, and thus distort our perception of the relationship between populations. Of course, in using single nucleotide polymorphisms

(SNPs) from SNP typing, we cannot avoid ascertained loci. As is often the case, the sites chosen for projects such as HapMap (The International HapMapConsortium, 2007) were based on cost efficiency, technologies, and changed over time (Clark, Hubisz, Bustamante, Williamson, & Nielsen, 2005). Thus, with the complications of ancestral relationships between individuals and ascertainment in extant samples, our ability of detect the actual population structure is uncertain.

We use a population-genetic model to simulate data from 1) a high-divergence tree: patterned after indigenous populations from Africa, Europe, Asia, and the Americas; and 2) a low-divergence tree: patterned after Europeans, and East Asians. However, simulating allele frequencies and genotypes from different populations is complicated by the hierarchical structure of human populations (Figure 3.1). In addition, simulating SNPs is particularly challenging, because the population structure must be assumed or inferred.

A number of methods to simulate data with population structure are based on choosing allele frequencies and then following those allele frequencies forward in time. The disadvantage to these methods is their slow speed, because all loci are followed forward in time including those lost or fixed. However, the algorithms tend to be rather simple- based on the random uniting of gametes, so that the addition of parameters is easy (FGP, (Hey, 1999), SimuPOP, (Peng & Kimmel, 2005; Peng, Amos, & Kimmel, 2007)). Some of these methods are based on the K alleles model for mutation, with mutation and selection (EASYPOP, (Balloux, 2001)). The K alleles model is a finite mutation model where there are k to $k-1$ possible mutation states for an allele, each with equal probability of occurring (Kimura & Weiss, 1964; Tajima, 1995). Though these methods are well-suited to simulate alleles at loci, the simulation of allele frequencies requires an additional step of calculating genotype frequencies from a simulated population.

To simulate allele frequencies for simulated populations, other methods begin with an allele frequency distribution and the probability of identity by descent to calculate the allele frequency forward in time (Price et al., 2006). For this method the authors choose ancestral allele frequencies from a uniform distribution and then used a beta distribution to choose the descendent frequencies (with parameters $p(1-\theta_1)$ and $(1-p)(1-\theta_1)$, where $\theta_1 = 4Nm$, N is the effective population size, and m is the mutation rate) (Balding & Nichols, 1994, 1995). The problem with this relatively simple method, is that each subpopulation is assumed to be a completely discrete entity, e.g. the covariance between populations is zero.

A backward-in-time approach to simulating data involves creating a genealogical tree (topology) and then overlaying the sequence differences (mutations) using coalescent methods (Kingman, 1982; Hudson, 1990; Excoffier, Novembre, & Schneider, 2000; Hudson, 2002). To create the genealogy the coalescent traditionally uses geometrically distributed random variables, so that sets of conditional probabilities up to k sequences for any sample are computed. An easier method involves approximating the geometric distribution by an exponential distribution. In this manner, the geometric random variable X is then exponentially distributed with $\lambda = \frac{k(k-1)}{4N}$. To simulate the genealogy, a generation back is drawn from an exponential distribution with expectation of $\frac{4N}{k(k-1)}$. Two randomly chosen sequences are then combined so that they share a common ancestor sequence. Since on average one coalescence occurs every generation, k is then decreased by one. This logic is repeated until $k = 1$, and thus only the founding lineage is left (Hein, Schierup, & Wiuf, 2005).

Mutations in a coalescent simulation are applied to the tree under the constant-rate neutral model (Watterson, 1975) which supposes that mutations may occur at any site along a lineage at a rate that is independent of demographic and selection pressures. Each mutation has a small probability of occurring in any one lineage, so that the number of mutations is a Poisson distributed random variable, S, with $\lambda = \frac{\theta_1}{2}$. To calculate the number of mutations that are placed on any given tree, both the expected total time of the tree and mutation rate of the sequence are utilized. The advantage to coalescent based methods are that simulating alleles at loci is fast. However if allele frequencies are desired the efficiency is lost because of the need to simulate a sample of individuals from the population and then calculate allele frequencies. The coalescent with an additional migration parameter can also be used to simulate genotypes with population structure (Hudson, 1990; Excoffier et al., 2000; Excoffier, Laval, & Schneider, 2005; Buendia & Narasimhan, 2006).

Here we present a forward-in-time method using population genetics theory to model the allele frequencies of populations descended from a single ancestral population undergoing neutral processes. The basic method relies on determining descendent allele frequencies by letting ancestral allele frequencies drift using Kimura's continuous approximation to Fisher-Wright sampling, given the ancestral allele frequency, time since divergence from the ancestor, and effective population size (Kimura, 1955). The resulting descendent allele frequencies can be ascertained by flexible criteria to determine the sensitivity of analytical results to differing ascertainment schemes. We demonstrate the validity of the method and compare simulated allele frequencies with parameters estimated from European and Asian data to CEPH and Asian HapMap samples.

We ascertain the allele frequencies using two methods: A.) where loci are chosen because their allele frequencies are polymorphic across all populations, or B.) where loci are chosen because their allele frequencies are polymorphic in a particular population. Because we know both the ancestry and ascertainment of the simulated sample, we can examine the outcomes of using STRUCTURE and GHM methods to detect population structure and ancestry.

3.3 Methods

Simulating populations descended from an ancestral population

The general strategy is take a sample of loci from an ancestral distribution, let the allele frequencies drift, and then ascertain the loci as the descendent population. The sampled loci represent the sampling of loci from an entire genome. The simulation method only requires parameters for the time to the common ancestor and the effective population size of each descendent population. An overview of the method is demonstrated in Figure 3.2 for two populations.

To begin, let us randomly choose allele frequencies for a set of ancestral loci from a beta distribution with parameters Nu and Nv, where N is the effective population size, u is the mutation rate, and v is the reversion rate. Both theory and empirical data demonstrates the beta distribution to be a good choice for the distribution of allele frequencies at a locus. The beta distribution is a special case of the diffusion approximation that specifies the equilibrium allele frequency distribution under only neutral forces, that is, with mutation, reversion, and drift (Wright, 1931). The estimates for parameters we use are $N = 10,000$ for effective population size, $v = 1.2$ x 10^{-8} for the mutation rate, and $u = 0.8$ x 10^{-8} for the reversion rate (Figure 3.3). Now, let us assume that all loci we choose from the ancestral population are independent and identically distributed and that the allele frequencies across loci follow a beta distribution. We can then take a sample of L polymorphic loci to represent the ancestral population, letting π_k and $1 - \pi_k$ be the ancestral allele frequencies at the kth locus. We let the ancestral allele frequency at each locus drift, so that in the descendent population it has 1.) a finite probability of fixation at zero or one, and 2.) a probability density of polymorphism. Let N be the effective population size of the descendent population, and t be the time to the ancestor. The probability of fixation is:

$$f\left(1,\, t\right) = \pi_k + \sum_{j=1}^{\infty} \left(2j+1\right) \pi_k \left(1-\pi_k\right) \left(-1\right)^j F\left(1-j,\, j+2, 2, \pi_k\right) e^{\frac{-j(j+1)t_i}{4N}}, \tag{3.1}$$

where F is the hypergeometric series (Kimura, 1955). The probability of loss is $f\left(0,\, t\right)$ and is similar to equation 3.1 above except that π_k is substituted by $1-\pi_k$ (Kimura, 1955). The probability density of polymorphism for the allele frequency, p, in the descendent population at the kth locus is:

$$\begin{aligned} \phi\left(\pi,\, p, t\right) &= \sum_{j=1}^{\infty} \pi\left(1-\pi\right) j \left(2j+1\right)\left(1+j\right) F\left(1-j,\, j+2, 2, \pi\right) \\ & \quad F\left(1-j,\, j+2, 2,\, p\right) e^{\frac{-j(j+1)t}{4N}} \end{aligned} \tag{3.2}$$

(Kimura, 1955). In practice, equation 3.2 cannot be computed exactly, as it is a limit of sum of infinite sums, but can be evaluated numerically. To implement this method we use follow a several steps to simulate an allele frequency at each locus:

1. For allele frequencies in the ancestral population, we take a sample of polymorphic loci from a beta distribution with parameters for effective population size, N, mutation rate, v, and reversion rate, u.

2. Using the ratio of the time from the descendent population back to the ancestor in generations to the effective population size of a descendent population, or $\frac{t}{2N}$,

we calculate the probability of fixation or loss for each ancestral allele frequency using $f(0,t)$ and $f(1,t)$. Then using those probabilities, we randomly determine if each locus is lost, fixed, or polymorphic in the descendent population. If the allele is fixed or lost, skip step 3), otherwise go to step 4).

3. If the locus remains polymorphic in the descendant, we use rejection sampling scheme. We choose a random number from the uniform distribution [0, 1] to evaluate equation 3.2 . If the calculated probability is equal to or less than a random number from the uniform distribution [0, *max*], (where the maximum is determined by the value of $\frac{t}{2N}$), we use the calculated probability as the descendant allele frequency. Otherwise we choose another random number from the uniform distribution [0, 1] again and repeat the rejection sampling.

4. Ascertain the allele frequencies.

We follow these steps for a set of L loci in the first population, and then again for the second population using different values of $\frac{t}{2N}$.

With this method we can ascertain SNPs according to a desired strategy. Possible strategies include: ascertaining polymorphic loci across both descendent populations or based on polymorphism in one particular population. It is easy to extend this method to simulate several different populations. In this case, we follow the algorithm above, except that for each additional descendent population another $\frac{t}{2N}$ is used.

Validity of the simulation method

We demonstrate our simulation method using parameters garnered from a phylogenetic tree constructed from 580 STRs from 16 populations (Hunley, Cabana, Merriwether, & Long, 2007). We simulated allele frequencies for European-like (CEU), and Asian-like distributions of allele frequencies. For the simulated European population we used an estimate of 0.034 for $\frac{t}{2N}$ and for the simulated Asian population we used an estimate of 0.15 for $\frac{t}{2N}$. These estimates are garnered from a hierarchical model fitted to 580 STR loci for 16 populations (figure 3.4) (Hunley et al., 2007).

Simulating populations from a divergent and less divergent tree

To investigate the effect of the underlying phylogenetic tree and ascertainment, we simulate data under a known phylogenetic tree using the method above. We take all our parameter estimates from a hierarchical model fitted to 580 STR loci for 16 populations (figure 3.4) (Hunley et al., 2007). For the first tree we simulate divergent world-wide populations. We pattern the tree after indigenous populations from Africa, Europe, Asia and the Americas with a common ancestor that arose prior to the Out-of-Africa migration. Figure 3.5A. shows the estimates for $\frac{t}{2N}$ and an illustration of the divergent tree. For the second tree, we simulate a low-divergence tree. We pattern the data after Europeans and a number of East Asian populations including those of Cambodia, China, and Japan. We estimate from the common ancestor of the European and Asian divergence where all non-Africans are one population for the root node. Figure 3.5B. shows the estimates for $\frac{t}{2N}$ and the second tree. The simulated allele frequencies are then: 1.) ascertained as polymorphic across all populations, 2.) ascertained as polymorphic in the European-like population, or 3.) not ascertained. From each ascertainment and population simulated, data sets of 50 individuals from each population of 500 and 2000 SNPs were created and prepared for STRUCTURE and GHM programs.

Methods of determining population structure

The algorithm underlying STRUCTURE, assigns individuals to one of K a-priori clusters each with a characteristic allele frequency distribution.The maximal value of K is inferred from the posterior probability of the distribution of K. This means that the most appropriate value of K for the data set is supported when the likelihood of the data fails to increase with a larger K, or no new clusters are produced with a larger K. Using the STRUCTURE program, each tree and ascertainment was run using a parameter set with no admixture, correlated allele frequencies between populations, and a fixed value for lambda. A burn in of 10,000 and 100,000 reps were used for each (Latch et al., 2006; Schwartz & McKelvey, 2009). The program was run five times for each K, one through six.

The GHM program works by comparing the estimated variance covariance matrix of gene identity (Nei, 1987) of the observed data to the expected variance covariance matrix generated from a specified *a priori* model. Here we test the fit of the data to three hierarchical models: a true model, a null (or island model), and a model in which there is an outgroup. The models are shown in Figure 3.6 and 3.7. The inability of a particular model to fit the data is measured by the likelihood ratio statistic $\Lambda = \nu(\ln|\hat{\Sigma}_0 - ln\hat{J} + tr\hat{J}\hat{\Sigma}_0^{-1} - r)$, where ν is the number of independent observations in the data, Σ is the expected gene identity matrix, $\hat{J}$ is the observed gene identity matrix, and r is the number of populations sampled. This likelihood ratio statistic is distributed as a chi-square statistic with $(r(r+1)/2) - s$, where, and s, is the number of parameters in the *a priori* model tree (Cavalli-Sforza & Piazza, 1975).

HapMap Data

We use the International HapMap Project (The International HapMap Consortium 2003) data release #23 in which samples were typed using the Affymetrix 500k chip. The project included four geographically distinct populations. To demonstrate correlations in figure 3.1 we used 45 unrelated Japanese from Tokyo, Japan (JPT), 45 unrelated Han Chinese from Beijing, China (CHB), and 20 unrelated individuals from the CEPH, Utah residents with ancestry from northern and western Europe (CEU). Inclusion criteria for SNPs included here are: a minor allele frequency greater than 5%, a chi-square test for HWE with a p-value of greater than 10^{-6}, and no more than 20% missing data. From these SNP data, we calculated the empirical allele frequencies.

3.4 Results and Discussion

Validation of the population-genetic simulation method

For allele frequencies that are polymorphic, our rejection sampling scheme closely approximates the density of the Kimura probability for a wide range of ancestral allele frequencies. We show the distribution of sampled allele frequencies for a somewhat extreme parameter set in Figure 3.8 (non-significant by the KolmogorovSmirnov goodness-of-fit test). It must be noted that the Kimura equations are in themselves an approximation of a continuous process, the approximation is not caculatable, that is, does not converge for ancestral allele frequency values of less than 0.012 or greater than 0.482 and values of $\frac{t}{2N} > 0.01$. If descendent allele frequencies are desired for ancestral allele frequency values less than 0.012 or greater than 0.482 or $\frac{t}{2N} < 0.01$, two methods can easily be employed, either: 1.) use more polymorphic ancestral allele frequencies, or 2.) use a normal approximation where $p_{ik} \sim N\left(\pi_k, \sigma^2_{p_{ik}}\right)$, with $m = \pi_k$ and $\sigma^2_{p_{ik}} = \pi_k\left(1-\pi_k\right)\left[1-e^{-t_i 2N_i}\right]$. In this normal approximation case, for p_{ik} that are greater than one, we set them to one, and for frequencies less than zero, we set them to zero.

Our method of simulating allele frequencies from different populations allows for flexibility in ascertainment. We show simulated allele frequencies ascertained based on polymorphism in both populations in Figure 3.9 or based on the frequencies in the simulated CEU population Figure 3.10. The correlation coefficient decreases when loci are ascertained based on polymorphism in the simulated CEU (compare 3.9 with 3.10 , where $r=$ 0.857 and 0.80, respectively). This is because more loci in the simulated Asian population are more likely to be fixed because of the larger value of $\frac{t}{2N}$, decreasing the correlation. The larger $\frac{t}{2N}$ reflects the longer time back to the common ancestor with the CEU indicative of a flatter probability density of polymorphism and a larger probability of being fixed. Comparing the simulated allele frequencies to HapMap data, we find that the correlation coefficients from the simulated data using either ascertainment method is in the range of the values from HapMap. Given that the ascertainment of HapMap is very complicated with the scheme changing as the project progressed, it is unlikely that the ascertainment scheme can be fully reconstructed (for a thorough explanation see Clark et al., 2005). Because of this, we do not expect our simulated data to match the HapMap allele frequencies exactly. Nor would we expect our method and underlying model to be corroborated by the HapMap data exactly. That the simulated frequencies are within the range that was demonstrated by HapMap does however represent a general proof of principle of the method.

The gene identity calculated from simulated allele frequencies for each tree and ascertainment is shown in Table 3.1. The simulated data gene identities are relatively close to the expected gene identities (Table 3.1).

Effect of the underlying tree and ascertainment on the methods to evaluate population structure

There are several things of notice from the STRUCTURE results. We first discuss the trends in the "proportion of membership of each pre-defined population in each of the K clusters", illustrated in Tables 3.2 and 3.3 for 2000 loci. These results are for the correlated allele frequency model. For the divergent tree, STRUCTURE always split the Europeans evenly amongst the “Inferred clusters” for each value of $K > 1$, for both the ascertainment across all populations and using the Europeans (Table 3.2A. and B). For the other three simulated populations, Africans, Asians, and Native Americans, it tended to cluster them together, with an increasingly large proportion of membership in the other cluster with an increase in K (Table 3.2). For the data set without ascertainment, the Europeans were clustered as a separate population (for $K > 1$, Table 3.2C.), whereas the other three populations had similar proportions of membership in a second cluster.

For the correlated allele frequency model, the low-divergence tree is approximately split between the inferred clusters for $K > 1$ with slightly less membership for one of the clusters for all ascertainment strategies (Table 3.3). The Asian populations-Han Chinese, Japanese, and Cambodians demonstrate some shared ancestry with each other by sharing slightly more in a cluster with each other (Table 3.3), and some shared ancestry with the Europeans, by sharing a proportion in a cluster with them.

Using STRUCTURE, the results of the likelihood analysis are quite inconclusive. The largest likelihood values are found for a $K = 1$ for the ascertainment across all populations and the ascertainment using the Europeans. For the simulated data sets with no ascertainment, the likelihood is larger for $K = 5$. For both scenarios, the data sets without ascertainment gave huge variances for the log likelihood for all values of K, on the order of 10^{11} to 10^{14}. Our problems with STRUCTURE may reflect the use of a small SNP data set, with a N=500 or 2000 SNPs. Some previous studies had some success in using STRUCTURE with few loci and shallow time (Latch et al., 2006), but others did not (Schwartz & McKelvey, 2009), or had to find new methods to find support for a correct K (Evanno et al., 2005). Based on these previous studies, the number of loci used here should have been sufficient but the combination of closely ancestrally related populations and few loci used in this study, may have led to a severely underpowered scenario for the correlated allele frequencies model.

The results for each given model tested in GHM for the divergent tree are shown in Table 3.4, and also for the low-divergence tree in Table 3.5. For the divergent tree, the chi-square statistics from the true and African outgroup models are rejected less readily, for all ascertainment strategies. Perhaps somewhat troublingly, the ascertainment for polymorphic loci and the ascertainment using the Europeans, in the 500 locus data set, make it difficult to discern between models. The R^2 values for the models tested show that the proportion of variation accounted for by the model being tested is relatively high, except for the 2000 SNP data set. With 2000 SNPs the R^2 values are 0.2 to 0.5 for the ascertainment using the European population data set and the data set without ascertainment. Even with some low R^2 values, there is still higher values for the true and African outgroup models as compared with the null model in all ascertainment strategies. The ascertainment using the Europeans makes results slightly less clear, but generally the trends in the chi-square and R^2 are towards supporting the fit of true model for the divergent data. These trends in the divergent data set are not as clear as in the less divergent data set. Interestingly, for the low-divergence tree, chi-square statistics from GHM more clearly discern the true model in the data without ascertainment even for the 500 SNP data set (e.g. χ^2 values are 13.703, 76.266, and 17.846 for the true, null and grouped Asian model, respectively, Table 3.5). It could even slightly distinguish between the true model and a model of grouped Asians for the other ascertainment strategies in the low-divergence simulated data. The R^2 values, and in turn, correlation coefficients, also lend support for the true model under all the ascertainment strategies even with only 500 SNPs. Ascertainment seems to have less of an effect on the low-divergence data set. It is interesting that the GHM method finds support for the correct models in the data set with larger overlap in allele frequencies (low-divergence) than for a data set with more time to separate lineages from the common ancestor (divergent). Perhaps this is because the deeper time and larger separation between ancestor populations could yield plausible support for a number of possible trees. It would be interesting in the future to test simulated data sets with less time to the common ancestor to evaluate this claim.

3.5 Conclusions

It might be predicted that STRUCTURE would be better at separating populations in the divergent data set. Using a correlated allele frequency model, support for the correct number of populations was unclear. The correlated allele frequencies model is recommended for subtle population structure to investigate population structure (Falush, Stephens, & Pritchard, 2003). However, like the results of Schwartz and McKelvey (2009), we were not able recover the correct value for K using this model. Additionally, STRUCTURE lacks the ability to test between possible models. Since there are other models using STRUCTURE that were not tested, such as the independent allele frequencies model, it is possible that another model may be more informative than the correlated allele frequencies model used here. Using other models and parameters in STRUCTURE with this simulated data would be important for future work.

The hierarchical modeling procedure (GHM) was able to correctly identify the underlying model for both the divergent and low-divergence data sets. It was even able to identify the correct model for the data set in which individuals had a very close ancestral history with as few as 500 SNPs. However, the number of tree models for a data set increases factorially with each additional population. Thus, using this method requires a good deal of *a priori* knowledge about the populations sampled to select the most plausible models. Of course, knowledge about the population structure in empirical samples may be relatively unknown, presenting a dilemma for investigators.

For both methods to detect the population structure, ascertainment seemed to affect the divergent data more than the less divergent data. Perhaps, for the divergent tree, since the allele frequency distributions are rather separate, when loci are ascertained the degree of polymorphism between populations is made less dramatic. In the low-divergence tree ascertainment is less of an issue because the Europeans are the obvious outgroup. When the simulated European population is used to ascertain polymorphic loci in the low-divergence tree, there it is more likely that the true underlying tree is preserved. Whereas, for the divergent tree the data was further distorted by ascertaining loci using the Europeans, which was not the outgroup. When an empirical sample is collected, the underlying true population relationships may be unknown and further distorted by ascertainment.

This work demonstrates the importance of background information about possible hypotheses for testing. For either method, a great deal of *a priori* knowledge about the populations sampled, garnered from oral history, the archeological record, demography etc., would be necessary to formulate any plausible inferences of either values of K for STRUCTURE or models to test in GHM.

An investigator, must define *a priori* populations from which to sample, sample those populations, and then test for population structure. This creates a dilemma:

in finding population structure, the investigator is merely finding evidence for recapitulating those same population categories rather than the genetic relationships in the data. Perhaps these categories serve as useful proxies for environmental exposures or epigenetic factors, or perhaps they add uncertainty to the underlying genetic relationships.

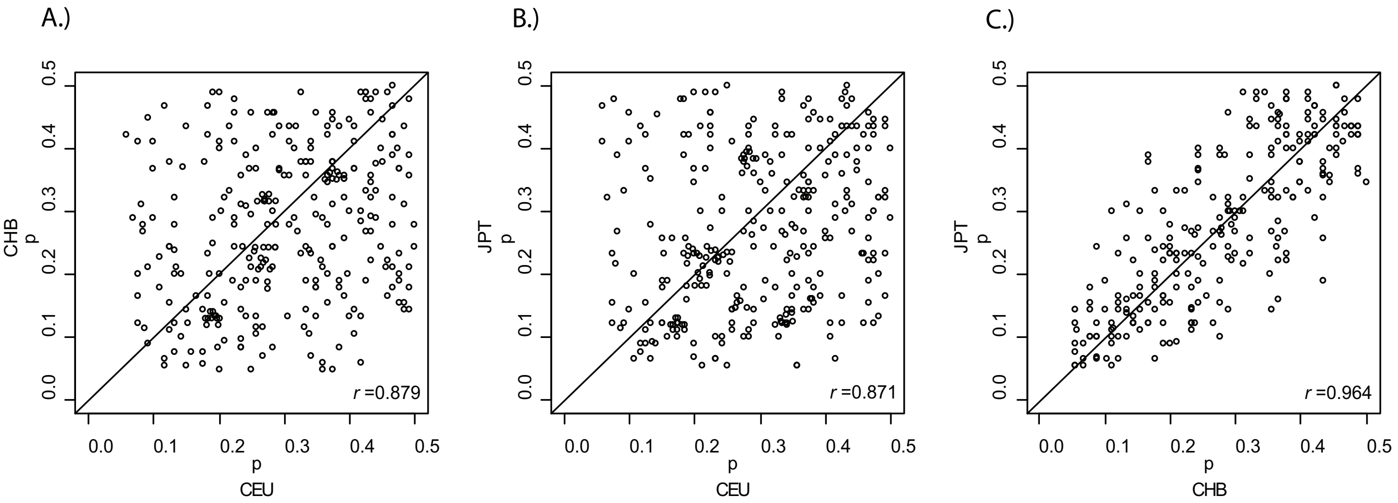

Figure 3.1: Linear regression of HapMap allele frequencies from 260 loci from each population.

260 randomly chosen loci from unrelated HapMap populations are included in the linear regressions. A.) CEPH (CEU) and Chinese (CHB), B.) CEPH and Japanese (JPT), and C.) Chinese and Japanese. Chinese and Japanese allele frequencies each display about the same correlation coefficient,r, with the CEPH population (0.879 and 0.871 respectively). However, the Chinese and Japanese have a much higher correlation coefficient (r = 0.964), reflecting their close ancestral relationship, as compared with either population with the CEPH.

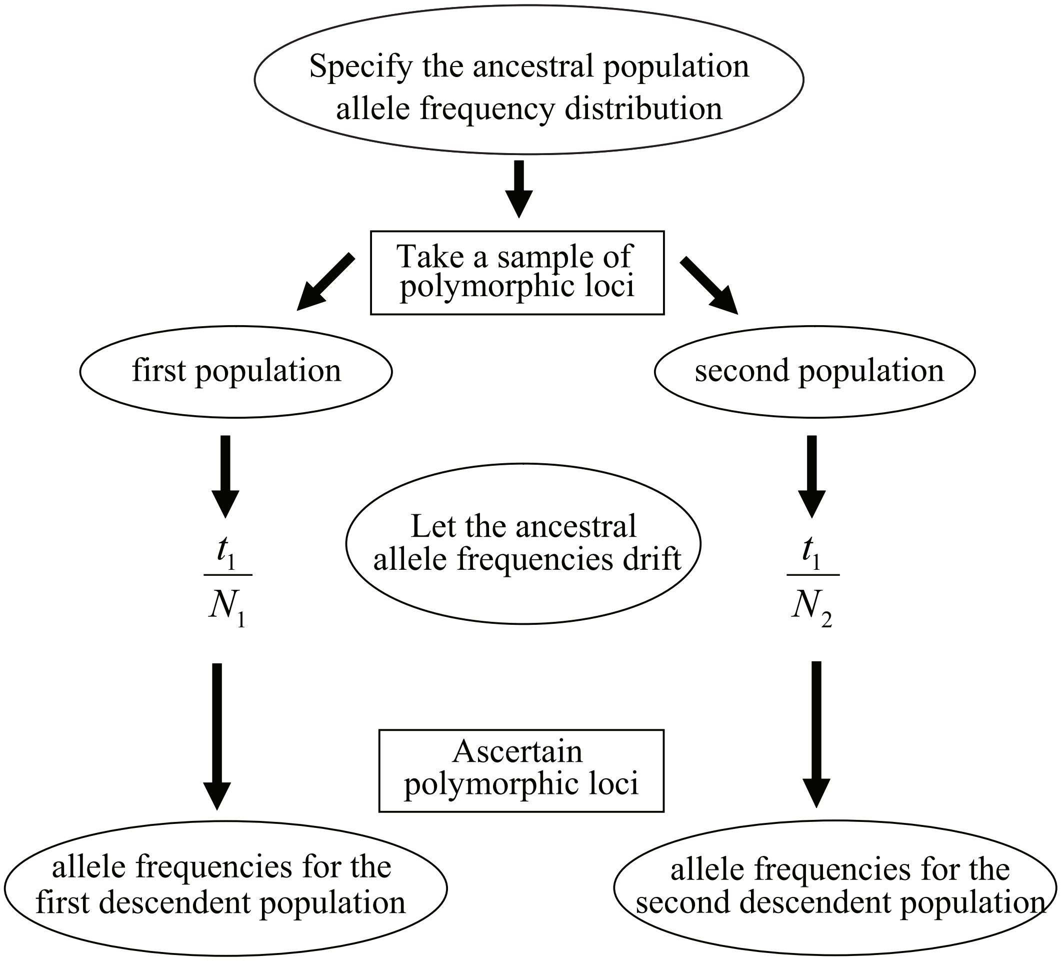
Specify the ancestral population allele frequency distribution
Take a sample of polymorphic loci
first population
second population
Let the ancestral allele frequencies drift
$\frac{t_1}{N_1}$
$\frac{t_1}{N_2}$
Ascertain polymorphic loci
allele frequencies for the first descendent population
allele frequencies for the second descendent population

Figure 3.2: Overview of our simulation method using population-genetic theory.

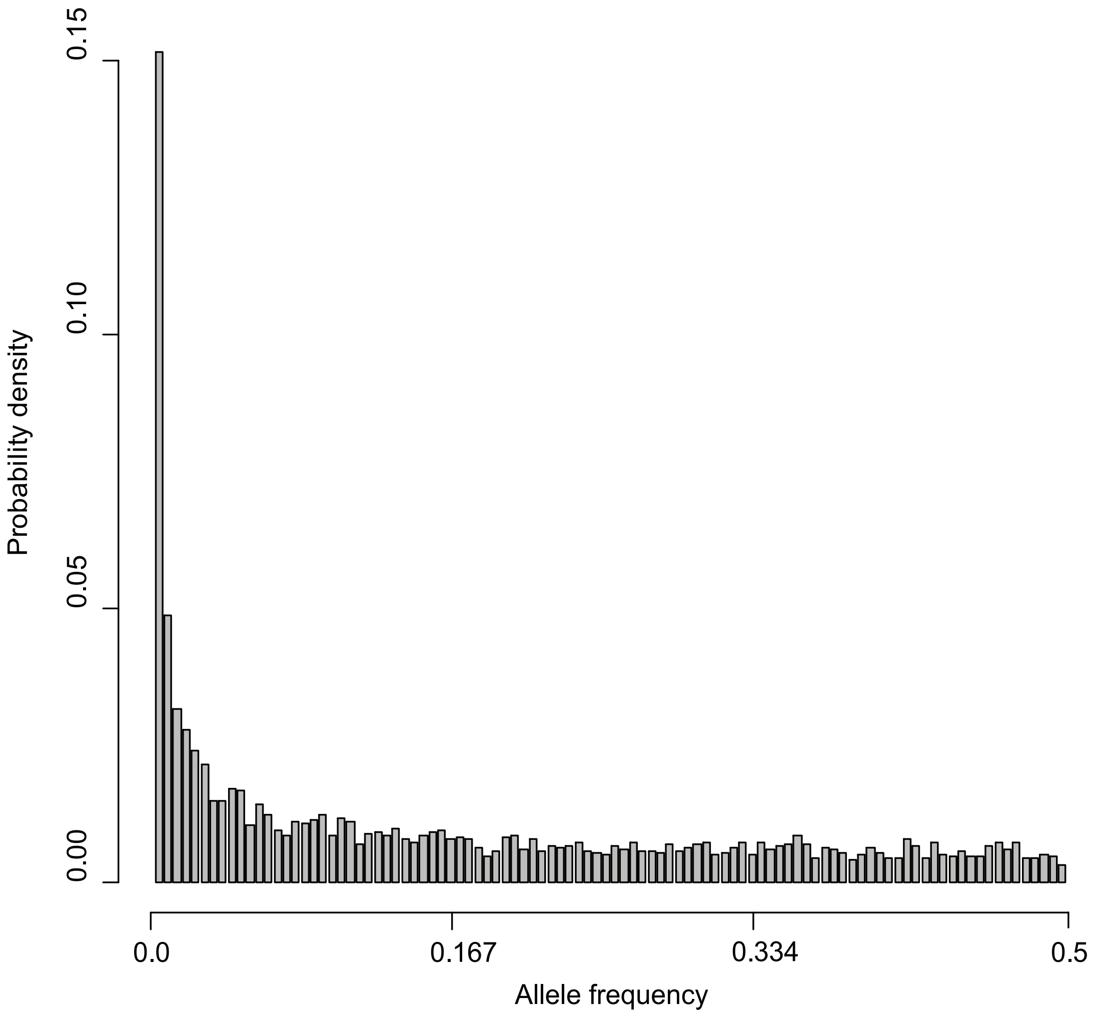

Figure 3.3: Simulation of the beta distribution.

Simulated minor allele frequencies from the beta distribution with parameters $4Nu = 0.00048$ and $4Nv = 0.0008$.

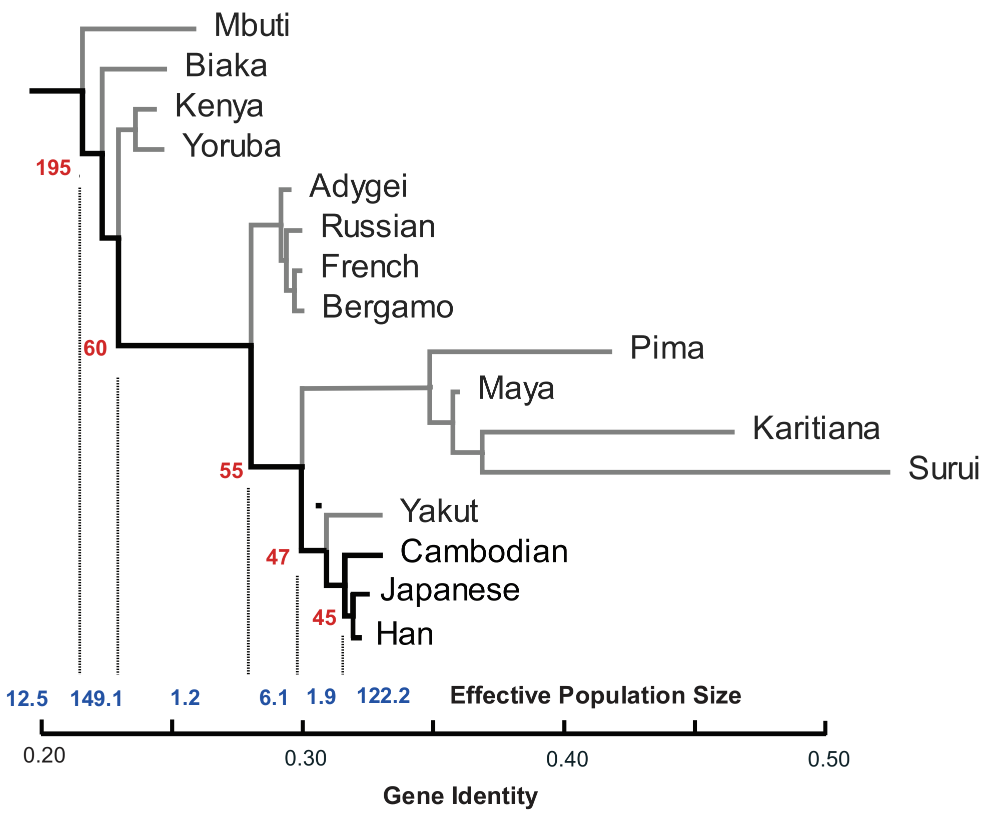

Figure 3.4: Hierarchical model of STR loci from 580 STR loci genotyped in 16 populations.

The best-fit hierarchical model fitted using GHM to 580 STR loci genotyped for 16 populations taken from (Hunley et al., n.d.). On the x-axis-scale is the effective population size (in thousands of individuals) at a node and gene identity. On each internal node, a number represents an approximate date for the split (in thousands of years). The Han Chinese history is divided into five epochs delimited by dotted vertical lines: (i) prior to Out-of-Africa migration, (ii) the non-Africans are one population, (iii) the East Asians and Native Americans are one population, (iv) the East Asians are one population, and (v) the Han Chinese are one population.

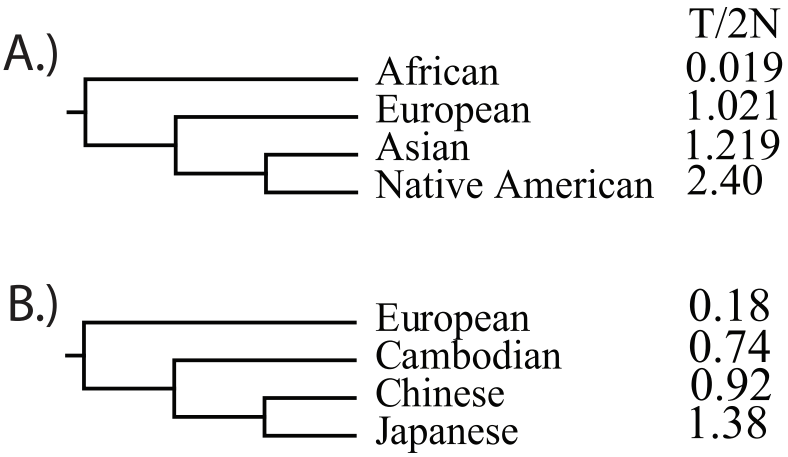

Figure 3.5: Cladograms and values for $t/2N$ used to simulate data.

The cladograms used to simulate data for the divergent and low-divergence trees. A.) The divergent tree and values for $t/2N$, and B.) low-divergence tree and values for $t/2N$.

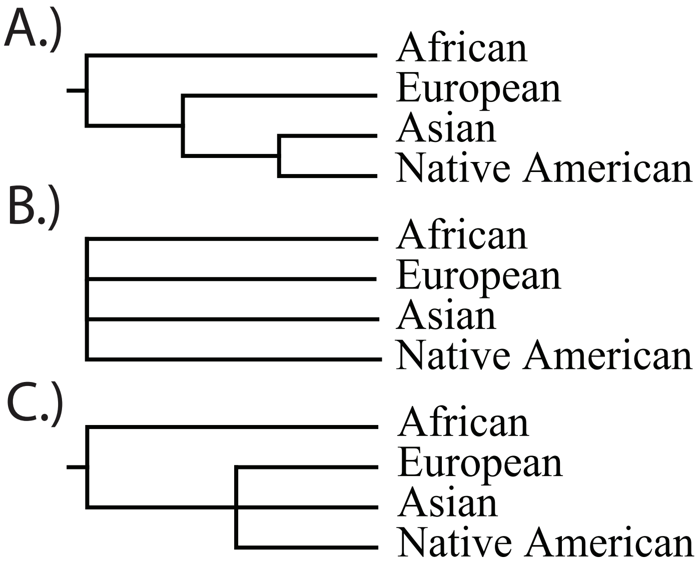

Figure 3.6: Cladograms tested in GHM for the divergent tree.

The three models tested in GHM. A.)The 'true' tree model, B.) the null model, testing treeness of the data, and C.) the African population used as an outgroup (African outgroup model).

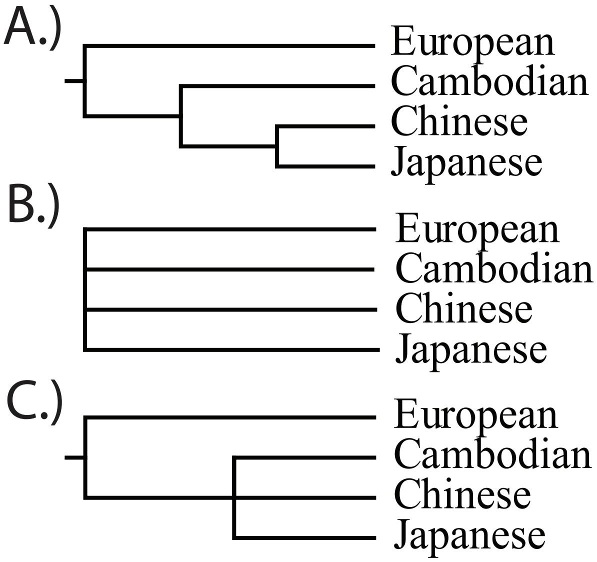

Figure 3.7: Cladograms tested in GHM for the low-divergence tree.

The three models tested in GHM. A.)The 'true' tree model tested in GHM, B.) the null model, testing treeness of the data, and C.) the European population used as an outgroup (Grouped Asians model).

Tree	Ascertainment	Population	Gene identity from simulated population
Divergent	Across	European	0.634
		African	0.605
		Asian	0.634
		Americas	0.671
	Based on CEU	European	0.631
		African	0.650
		Asian	0.721
		Americas	0.917
	None	European	0.852
		African	0.654
		Asian	0.876
		Americas	0.964
Low divergence	Across	European	0.610
		Han Chinese	0.636
		Japanese	0.632
		Cambodian	0.628
	Based on CEU	European	0.611
		Han Chinese	0.827
		Japanese	0.885
		Cambodian	0.787
	None	European	0.636
		Han Chinese	0.833
		Japanese	0.889
		Cambodian	0.794

Table 3.1: Gene identity calculated from simulated populations and given ascertainment.

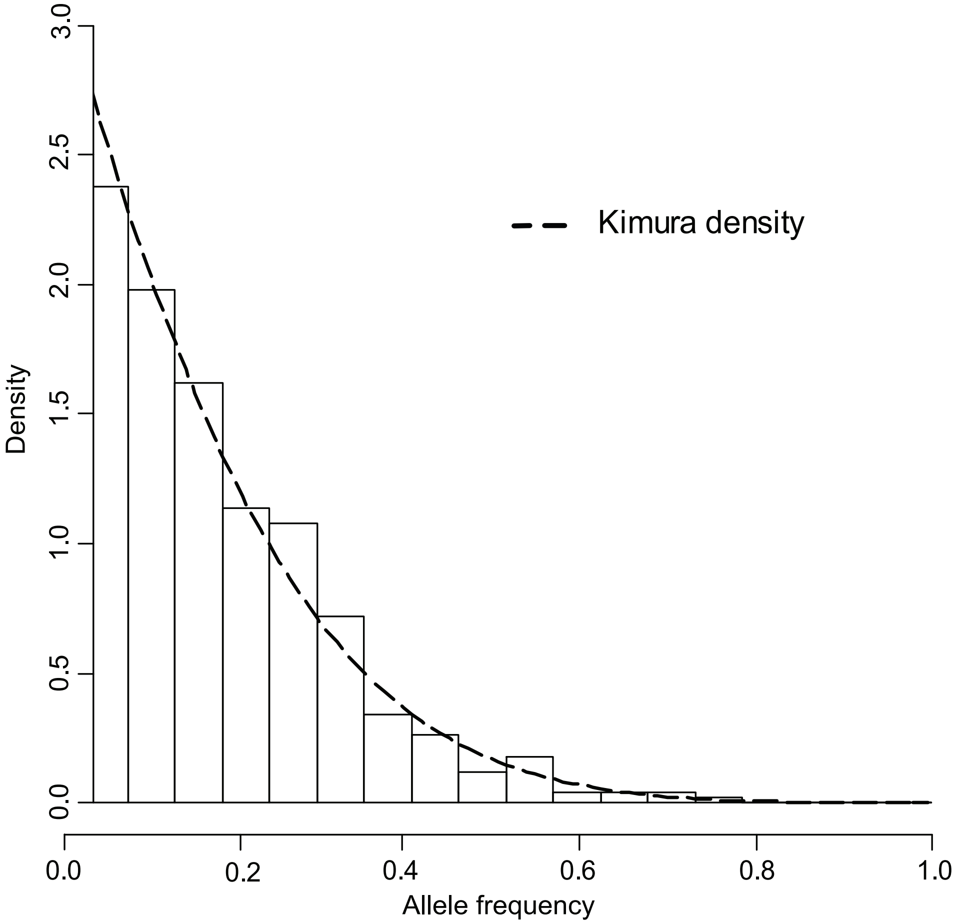

Figure 3.8: Histogram of the simulated data from the rejection sampling scheme compared with the actual Kimura probability density curve for the same parameters.

If the locus remains polymorphic in the descendant, we use a rejection sampling scheme. We choose a random number from the uniform distribution [0,1] to evaluate equation 3.2 . If the calculated probability is equal to or less than a random number from the uniform distribution [0, *max*], (where the maximum is determined by the value of $\frac{t}{2N}$), we use the calculated probability as the descendant allele frequency. Otherwise we choose another random number from the uniform distribution [0,1] again and repeat the steps. Rejection sampled allele frequencies displayed in the histogram were simulated with $\frac{t}{2N}$ equal to 0.15 and ancestral allele frequency fixed to 0.1. The fit of the Kimura probability density with the same parameters is very close (non-significant KolmogorovSmirnov goodness-of-fit test).

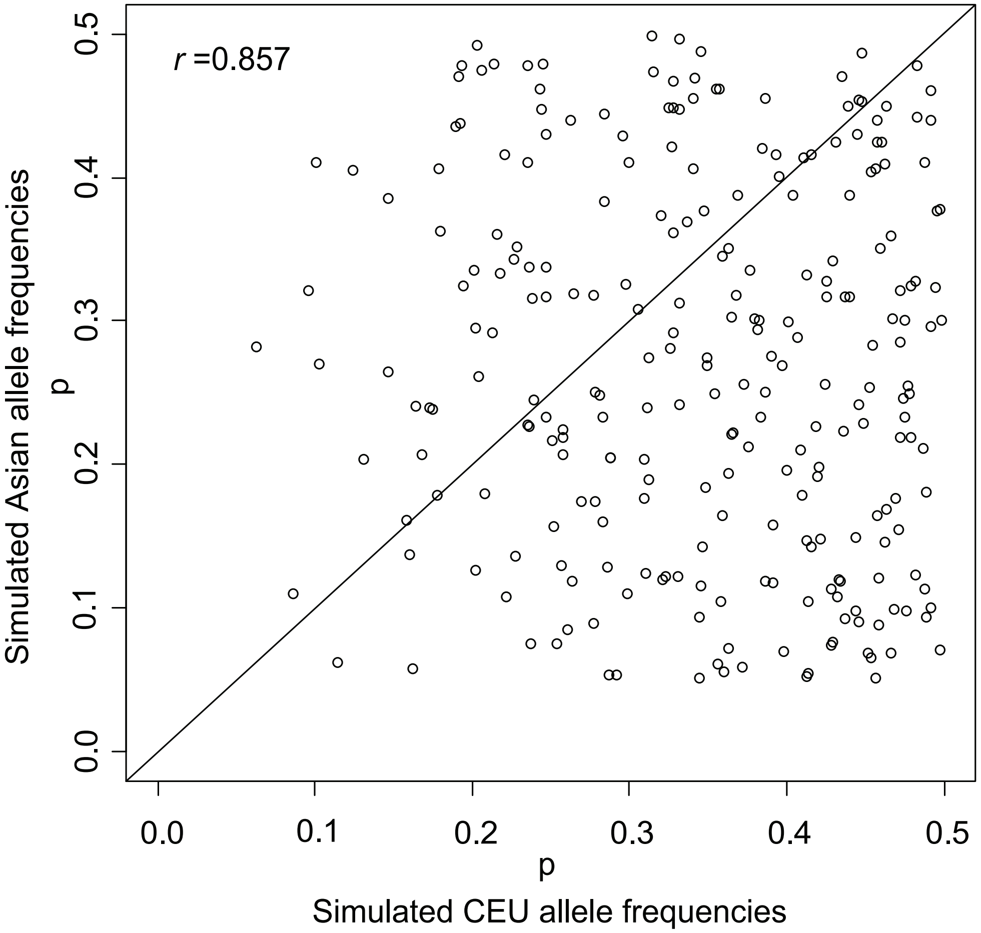

Figure 3.9: Simulated folded allele frequencies for CEU and Asian like populations ascertained for polymorphism in both populations.

Simulated folded allele frequencies for the CEU (x-axis) and Asian (y-axis) populations. Parameters for the simulation were $\frac{t}{2N} = 0.034$ and $\frac{t}{2N} = 0.15$ for the simulated CEU and Asian allele frequencies, respectively. Loci were then chosen if they were polymorphic in both populations.

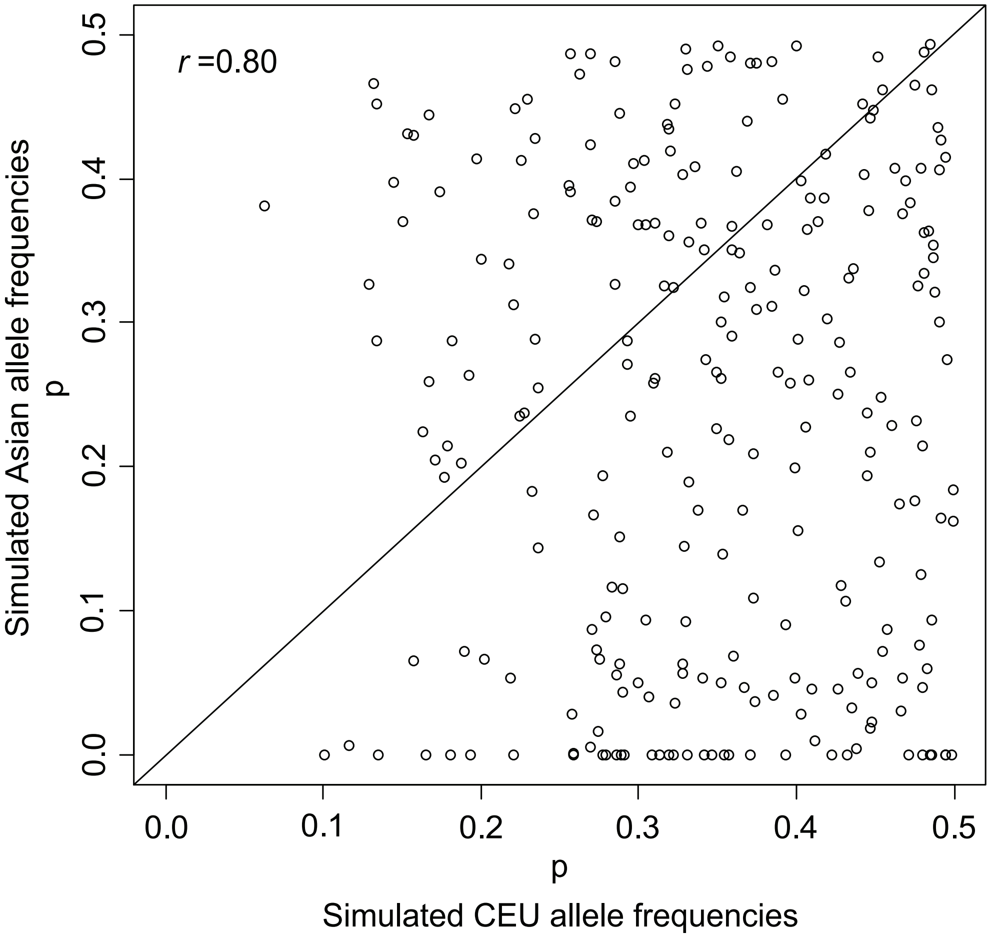

Figure 3.10: Simulated folded allele frequencies for CEU and Asian like populations ascertained by polymorphism in the simulated CEU allele frequencies.

Simulated folded allele frequencies for the CEU (x-axis) and Asian (y-axis) populations with the same parameters as Figure 3.9. Loci were chosen based on polymorphism in the CEU. More loci are fixed in the Asian population by this ascertainment strategy, decreasing the correlation.

		K=1	K=2		K=3			K=4				K=5				
	Given population	1	1	2	1	2	3	1	2	3	4	1	2	3	4	5
A.)	European	1.0	0.518	0.482	0.313	0.331	0.355	0.293	0.251	0.215	0.241	0.189	0.220	0.181	0.184	0.227
	African	1.0	0.623	0.377	0.301	0.327	0.372	0.273	0.257	0.230	0.240	0.165	0.273	0.163	0.175	0.224
	Asian	1.0	0.674	0.326	0.304	0.307	0.389	0.298	0.225	0.226	0.251	0.187	0.259	0.177	0.174	0.203
	Americas	1.0	0.501	0.499	0.307	0.335	0.358	0.277	0.267	0.239	0.218	0.170	0.241	0.157	0.204	0.228
B.)	European	1.0	0.535	0.465	0.338	0.334	0.328	0.252	0.252	0.247	0.249	0.258	0.204	0.185	0.196	0.157
	African	1.0	0.584	0.416	0.258	0.346	0.296	0.280	0.242	0.213	0.265	0.205	0.221	0.196	0.222	0.156
	Asian	1.0	0.492	0.508	0.335	0.299	0.366	0.235	0.248	0.282	0234	0.283	0.180	0.188	0.180	0.169
	Americas	1.0	0.594	0.406	0.356	0.363	0.281	0.235	0.245	0.214	0.306	0.203	0.226	0.199	0.212	0.160
C.)	European	1.0	0.020	0.980	0.021	0.967	0.012	0.020	0.973	0.000	0.007	0.020	0.000	0.000	0.980	0.000
	African	1.0	0.508	0.492	0.337	0.330	0.333	0.251	0.243	0.259	0.248	0.215	0.202	0.194	0.194	0.196
	Asian	1.0	0.317	0.683	0.238	0.549	0.213	0.188	0.511	0.150	0.151	0.165	0.125	0.123	0.462	0.125
	Americas	1.0	0.475	0.525	0.335	0.330	0.334	0.258	0.253	0.244	0.245	0.197	0.201	0.202	0.205	0.196

Table 3.2: Proportion of membership in the clusters for each K from STRUCTURE with a correlated allele frequency model.

Using the STRUCTURE program the proportion of membership in each K cluster. A.) The divergent 2000 polymorphic SNP data set, B.) the divergent 2000 SNP data set ascertained in the European data set, and C.) the divergent 2000 SNP data set without any ascertainment.

		K=1	K=2		K=3			K=4				K=5				
	Given population	1	1	2	1	2	3	1	2	3	4	1	2	3	4	5
	European	1.0	0.498	0.502	0.337	0.338	0.325	0.258	0.250	0.248	0.244	0.209	0.182	0.200	0.198	0.211
A.)	Han Chinese	1.0	0.494	0.506	0.336	0.337	0.327	0.263	0.248	0.248	0.241	0.234	0.193	0.192	0.176	0.206
	Japanese	1.0	0.511	0.498	0.355	0.342	0.304	0.260	0.260	0.234	0.246	0.191	0.214	0.214	0.185	0.196
	Cambodian	1.0	0.494	0.506	0.328	0.344	0.327	0.276	0.242	0.249	0.233	0.214	0.213	0.189	0.178	0.206
	European	1.0	0.515	0.485	0.342	0.323	0.335	0.285	0.256	0.218	0.240	0.195	0.182	0.199	0.214	0.210
B.)	Han Chinese	1.0	0.527	0.473	0.360	0.304	0.336	0.281	0.209	0.199	0.311	0.176	0.188	0.219	0.217	0.199
	Japanese	1.0	0.598	0.402	0.409	0.296	0.295	0.208	0.282	0.229	0.281	0.193	0.188	0.231	0.226	0.164
	Cambodian	1.0	0.539	0.461	0.374	0.299	0.327	0.277	0.222	0.204	0.297	0.194	0.181	0.197	0.245	0.184
	European	1.0	0.477	0.523	0.342	0.330	0.328	0.247	0.269	0.252	0.232	0.221	0.206	0.168	0.196	0.209
C.)	Han Chinese	1.0	0.4997	0.503	0.337	0.336	0.327	0.261	0.247	0.231	0.261	0.198	0.205	0.201	0.198	0.197
	Japanese	1.0	0.500	0.500	0.334	0.336	0.331	0.254	0.242	0.242	0.263	0.204	0.203	0.200	0.200	0.193
	Cambodian	1.0	0.490	0.510	0.337	0.335	0.328	0.268	0.247	0.238	0.247	0.197	0.198	0.199	0.203	0.203

Table 3.3: Proportion of membership in the clusters for each K from STRUCTURE with a correlated allele frequency model.

Using the STRUCTURE program the proportion of membership in each K cluster. A.) The low-divergence 2000 polymorphic SNP data set, B.) the low-divergence 2000 SNP ascertained in the European data set, and C.) the low-divergence 2000 SNP data set without any ascertainment.

Loci	Ascertainment	Tested Model	R^2	χ^2	p-value
500	Across	True	0.627	9.228	0.026
		Null	0.605	9.599	0.087
		African outgroup	0.625	9.364	0.053
	Based on CEU	True	0.880	41.473	< 0.00001
		Null	0.824	51.497	< 0.00001
		African outgroup	0.889	42.749	< 0.00001
	None	True	0.691	34.922	< 0.00001
		Null	0.165	189.603	< 0.00001
		African outgroup	0.693	35.032	< 0.00001
2000	Across	True	0.909	11.724	0.0084
		Null	0.894	23.797	0.002
		African outgroup	0.894	13.115	0.0107
	Based on CEU	True	0.888	157.392	< 0.00001
		Null	0.815	227.223	< 0.00001
		African outgroup	0.903	175.029	0.001
	None	True	0.771	133.737	< 0.0001
		Null	0.220	857.725	< 0.00001
		African outgroup	0.771	133.738	< 0.00001

Table 3.4: Results for the divergent tree for the three models tested using GHM.

For data simulated under the divergent tree, the R^2 from the observed vs. expected values, χ^2 values, and p-value for each ascertainment and tested model used in GHM are shown with their corresponding results. The R^2 values are calculated from the observed and expected gene identity matrix, where the expected gene identity matrix is calculated based on the particular model that is being tested. The ascertainment is accomplished by choosing polymorphic loci across all populations (polymorphic), choosing polymorphic loci in the CEU population (based on CEU), or choosing a random set of loci (no ascertainment). The models tested in GHM for the divergent tree are shown in Figure 3.6.

Loci	Ascertainment	Tested Model	R^2	χ^2	p-value
500	Across	True	0.901	3.906	0.272
		Null	0.915	13.437	0.019
		Grouped Asians	.901	3.910	0.418
	Based on CEU	True	0.918	19.594	0.0002
		Null	0.668	92.771	< 0.00001
		Grouped Asians	0.892	21.842	0.0002
	None	True	0.909	13.703	0.003
		Null	0.657	76.266	< 0.0001
		Grouped Asians	0.854	17.846	0.0013
2000	Across	True	0.939	5.341	0.148
		Null	0.966	13.500	0.019
		Grouped Asians	0.930	5.956	0.203
	Based on CEU	True	0.911	43.014	< 0.00001
		Null	0.504	447.607	< 0.00001
		Grouped Asians	0.846	59.098	< 0.00001
	None	True	0.907	39.865	< 0.00001
		Null	0.401	421.960	< 0.00001
		Grouped Asians	0.838	54.464	< 0.00001

Table 3.5: Results for the low-divergence tree for the three models tested using GHM.

For data simulated under the low-divergence tree, the R^2 from the observed vs. expected values, χ^2 values, and p-value for each ascertainment and tested model used in GHM are shown with their corresponding results. The R^2 values are calculated from the observed and expected gene identity matrix, where expected gene identity matrix is calculated based on the particular model that is being tested. The ascertainment was accomplished by choosing polymorphic loci across all populations (polymorphic), choosing polymorphic loci in the CEU population (based on CEU), or choosing a random set of loci (no ascertainment). The models tested in GHM for the low-divergence tree are shown in Figure 3.7.

CHAPTER IV

Conclusion

The purpose of this dissertation was twofold: 1) to explore the relationships between individuals that exist in a collected sample for ancestry or population structure because we are interested in the structure or and relationships themselves, and 2) to know if the sample we have collected is a random mating sample of a population of interest, so that we can apply certain assumptions and in turn methods to the data. Both of these purposes are interconnected: we cannot understand the relationships between individuals that exist in a sample if we do not make some assumptions about a population beforehand, but neither can we understand a population without some assumptions about the relationships between individuals within it.

The approach to these goals was to begin with a definition of population based on the probabilities of genotypes calculated from allele frequencies - simply, Hardy Weinberg Equilibrium. Though the assumptions underlying HWE are rarely strictly met, HWE gives values for which reasonable expectations of population variation can be based. HWE demonstrates that the mechanism of inheritance does not decrease variation in a population over time. Thus, the initial value of an allele frequency will continue to represent a panmictic population though time and its descendent progeny. Across loci a single individual is an estimate of homozygosity for their population. This is somewhat profound. With a single individual we have an estimate of population homozygosity. However, with two individuals we have a dilemma: we know that we have a population, but could we have two populations? This problem continues with more individuals. This dissertation presented a method to statistically approach this problem and test whether a sample of just two individuals represents the same single panmictic population. With two data points we can define populations. By observing the gene identity or the test statistic between all members of a sample, inferences about the relationships between any *a priori* population categories and the genetic data can be drawn *and* relationships within the genetic sequences themselves can be examined. I validated the method and demonstrated its power to detect a difference in homozygosity between pairs of individuals. Moreover, the method does not need *a priori* information about an individual's genetic relationship with another individual, nor an individual's genetic identity to a population.

To throughly address both of the goals above, I also needed to further explore models of the population structure and ancestral population history. The underlying history due to stochastic and non-stochastic processes creates relationships between individuals, between sub-populations, and between populations. The relationships disturb panmixia. This was demonstrated in Chapter two with the statistical test to determine if two individuals are from the same panmictic population. In particular, in the assessment of power, the method has a high power to test for relatives; this is because relatives, though they are from the same population they share more alleles identical by descent, and are not from a panmictic population. The test has the highest power to test for more closely related individuals, such as parent-offspring and full-sibling pairs who on average share the most alleles. I also have high power to test for individuals that are distantly ancestrally related populations because their underlying allele frequency distributions are more different from each other.

To explore population structure- I looked to the history of population genetics and its simplest models of population structure. Historically, explorations of stochastic processes in a finite population led to the Wright-Fisher model (Fisher, 1930; Wright, 1931). This model was crucial to understanding the change in the allele frequency distribution population over time. Without the establishment of HWE and the Wright-Fisher model, diffusion methods would not have had a theoretical and biological basis, after all the diffusion approximation is merely an approximation of a continuous version of the Wright-Fisher model for a large population. Diffusion methods (Wright, 1929; Fisher, 1930; Wright, 1931; Kimura, 1955) were the basis for the method of simulating allele frequencies from different populations. I demonstrated the validity of this method for a range of initial values in Chapter three. Though it is somewhat time intensive for a forward-in-time simulation method, it is able to accurately capture population history and the underlying correlated allele frequencies between populations. For those values that are not calculatable for the diffusion model, I use a normal approximation. In general, the method is unique, because it simulates allele frequencies that can be ascertained. This property makes it convenient for simulating SNPs.

Lastly, my exploration of population structure and ancestry led me to look to methods that detect and test models of population structure in the data. Evolutionary ancestral history has left a deep indelible structure on genetic data (Cavalli-Sforza & Piazza, 1975; Cavalli-Sforza, Piazza, Menozzi, & Mountain, 1989; Wang et al., 2007; Ramachandran et al., 2005; Rosenberg et al., 2005; Hunley & Long, 2005; Hunley, Healy, & Long, 2009), this process has led to correlations between allele frequencies (Figure 3.1) which must be considered in simulating genetic data and when considering methods to evaluate population structure on sample data. I do this in simulating allele frequencies through the simulation method above. However, I am also interested in methods that detect structure in the data. In Chapter three I looked at the effects of ascertainment and underlying population history on methods to detect population structure. Interestingly, both STRUCTURE and GHM are influenced more by the underlying history than the ascertainment. This is perhaps because the underlying population history creates a correlation between allele frequencies that is stronger than the ascertainment. Perhaps with more loci, ascertainment might have a stronger effect on the simulated data. Few loci were used in the evaluation, because many of the previous studies used few loci, and it seemed to make sense to test the minimum number of loci that could be used to investigate these models. It would be interesting in the future to use more loci and simulate some models of more recent admixture and investigate them in STRUCTURE and GHM.

4.1 Final words

This dissertation work demonstrates the importance of exploring the population(s) and individuals in a sample, and the method used to make those inferences. This is not a simple task. Individuals fall into approximately panmictic clusters in the space of mating probability, and this clustering creates nodes of non-random distributions of allele frequencies and DNA sequence variations. As the nodes form subpopulations, their relationships to each other are retained through the structure that random and non-random processes have shaped.

With the method presented in Chapter two, the identity estimates and test statistics viewed from all pairs offer a way to view the nodes without having to place boundaries on the data *a priori*. The method is biologically coherent. We can start to make inferences about individual and population relationships within the sample. Further, we can also test for whether the collected sample is a panmictic representative sample of the population. We can also test for cryptic relatedness and structure in the sample, which may or may not be desired in the sample by the research investigators. Interestingly, the method also demonstrates that HapMap populations do not represent natural populations very well. There are related individuals between trios in the CEPH and Yoruba populations. There are a number of very significant outliers in the data set, as well. Both cryptical relatedness and population structure can be confounding factors in genetic samples (Voight & Pritchard, 2005). That many studies have relied upon this data set as representative populations (Mengel-From, Wong, Morling, Rees, & Jackson, 2009; Makoff et al., 2009; Shyn et al., 2009; Ge, Budowle, Planz, & Chakraborty, 2009; Shriner et al., 2009) is somewhat alarming from our results here.

The second project discussed in Chapter three demonstrates that a great deal of *a priori* knowledge about the populations sampled is necessary to formulate any plausible inferences about the sample. However, the investigator must be careful that the *a priori* desired sample information is appropriate for the research question that is being asked. If the *a priori* boundary for the population is not accurate to the genetic structure of that population, or representative of the non-genetic factors, the research question may not be accurately tested. Further, if the study seeks to find genetic contributions to traits then the investigator must be careful that they are not merely finding evidence for recapitulating *a priori* population categories, which may not represent natural populations. With genetic data becoming more abundant, more precision in the knowledge about where the boundaries are placed and whether the sample meets those criterion for the population of inference, is crucial to make coherent inferences about genes and populations.

References

A.C., Fitzpatrick, J., Bowman, R., Stith, B., Makarewich, C., Stenzler, L., et al. (2008, Apr). Congruent population structure inferred from dispersal behaviour and intensive genetic surveys of the threatened Florida scrub-jay (*Aphelocoma coerulescens*). *Mol Ecol*, *17*(7), 1685–1701. Available from `http://dx.doi.org/10.1111/j.1365-294X.2008.03705.x`

Alexander, D., Novembre, J., & Lange, K. (2009, Sep). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, *19*(9), 1655–1664. Available from `http://dx.doi.org/10.1101/gr.094052.109`

Balding, D. J., & Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, *96*(1-2), 3-12.

Balding, D. J., & Nichols, R. A. (1994, Feb). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci Int*, *64*(2-3), 125–140.

Balloux, F. (2001). EASYPOP (version 1.7): a computer program for population genetics simulations. *J Hered*, *92*(3), 301–302.

Barr, K., Lindsay, D., Athrey, G., Lance, R., Hayden, T., Tweddale, S., et al. (2008, Aug). Population structure in an endangered songbird: maintenance of genetic differentiation despite high vagility and significant population recovery. *Mol Ecol*, *17*(16), 3628–3639. Available from `http://dx.doi.org/10.1111/j.1365-294X.2008.03868.x`

Bauchet, M., McEvoy, B., Pearson, L., Quillen, E., Sarkisian, T., Hovhannesyan, K., et al. (2007, May). Measuring European population stratification with microarray genotype data. *Am J Hum Genet*, *80*(5), 948–956. Available from `http://dx.doi.org/10.1086/513477`

Beerli, P., & Felsenstein, J. (1999, Jun). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, *152*(2), 763–773.

Beerli, P., & Felsenstein, J. (2001, Apr). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci USA*, *98*(8), 4563–4568. Available from `http://dx.doi.org/10.1073/pnas.081068098`

Buendia, P., & Narasimhan, G. (2006, Sep). Serial NetEvolve: a flexible utility for generating serially-sampled sequences along a tree or recombinant network. *Bioinformatics*, *22*(18), 2313–2314. Available from `http://dx.doi.org/10.1093/bioinformatics/btl387`

Casella, G., & Berger, R. (2002). *Statistical Inference.* Thomson Learning.

Cavalli-Sforza, L., & Bodmer, W. (1971). *The genetics of human populations.* W.H. Freeman and Co.

Cavalli-Sforza, L., & Piazza, A. (1975, Oct). Analysis of evolution: evolutionary rates, independence and treeness. *Theor Popul Biol*, *8*(2), 127–165.

Cavalli-Sforza, L., Piazza, A., Menozzi, P., & Mountain, J. (1989, Jun). Genetic and linguistic evolution. *Science*, *244*(4909), 1128–1129.

Chen, C., Durand, E., & Francois, O. (2007). Bayesian clustering algorithms ascertaining spacial population structure: a new computer program and a comparison study. *Mol Ecol Notes*, *7*, 747-756.

Clark, A., Hubisz, M., Bustamante, C., Williamson, S., & Nielsen, R. (2005, Nov). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*, *15*(11), 1496–1502. Available from `http://dx.doi.org/10.1101/gr.4107905`

Coop, G., Pickrell, J., Novembre, J., Kudaravalli, S., Li, J., Absher, D., et al. (2009, Jun). The role of geography in human adaptation. *PLoS Genet*, *5*(6), e1000500. Available from `http://dx.doi.org/10.1371/journal.pgen.1000500`

Dawson, K., & Belkhir, K. (2001, Aug). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet Res*, *78*(1), 59–77.

Epstein, M., Duren, W., & Boehnke, M. (2000, Nov). Improved inference of relationship for pairs of individuals. *Am J Hum Genet*, *67*(5), 1219-1231. Available from `http://dx.doi.org/10.1016/S0002-9297(07)62952-8`

Evanno, G., Regnaut, S., & Goudet, J. (2005, Jul). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*, *14*(8), 2611–2620. Available from `http://dx.doi.org/10.1111/j.1365-294X.2005.02553.x`

Excoffier, L., Laval, G., & Schneider, S. (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol Bioinform Online*, *1*, 47–50.

Excoffier, L., Novembre, J., & Schneider, S. (2000). SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J Hered*, *91*(6), 506–509.

Falush, D., Stephens, M., & Pritchard, J. (2003, Aug). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567–1587.

Falush, D., Wirth, T., Linz, B., Pritchard, J., Stephens, M., Kidd, M., et al. (2003, Mar). Traces of human migrations in *Helicobacter pylori* populations. *Science*, *299*(5612), 1582–1585. Available from `http://dx.doi.org/10.1126/science.1080857`

Fisher, R. (1930). *The genetical theory of natural selection* (A Complete Variorum Edition Re-Print 1999 ed.). Oxford University Press.

Franois, O., Ancelet, S., & Guillot, G. (2006, Oct). Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, *174*(2), 805–816. Available from `http://dx.doi.org/10.1534/genetics.106.059923`

Ge, J., Budowle, B., Planz, J., & Chakraborty, R. (2009, Dec). Haplotype block: a new type of forensic DNA markers. *Int J Legal Med*. Available from `http://dx.doi.org/10.1007/s00414-009-0400-5`

Hein, J., Schierup, M., & Wiuf, C. (2005). *Gene genealogoes, variation and evolution: A primer in coalescent theory.* Oxford University Press, Oxford, UK.

Hey, J. (1999). FPG (a computer program for Forward Population Genetic simulation) [Computer software manual].

Hubisz, M., Falush, D., Stephens, M., & Pritchard, J. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, *9*, 1322-1332.

Hudson, R. (1990). Gene genealogies and the coalescent processes. *Oxford Surveys in Evolution, Biology*, *7*, 1-43.

Hudson, R. (2002, Feb). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338.

Hunley, K., Cabana, G., Merriwether, D., & Long, J. (2007, Apr). A formal test of linguistic and genetic coevolution in Native Central and South America. *Am J Phys Anthropol*, *132*(4), 622–631. Available from `http://dx.doi.org/10.1002/ajpa.20542`

Hunley, K., Healy, M., & Long, J. (2009, May). The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race. *Am J Phys Anthropol*, *139*(1), 35–46. Available from `http://dx.doi.org/10.1002/ajpa.20932`

Hunley, K., Jie, L., Lewis, C., Malhi, R., & Long, J. (n.d.). *Low intraallelic variation and ancient coalescent time reject neutral evolution for ALDH2*2.*

Hunley, K., & Long, J. (2005, Feb). Gene flow across linguistic boundaries in Native North American populations. *Proc Natl Acad Sci USA*, *102*(5), 1312–1317. Available from `http://dx.doi.org/10.1073/pnas.0409301102`

Kimura, M. (1953). "Stepping Stone" model of population. *Ann rept Nat Inst Genetics Japan*, *3*, 62-63.

Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc Natl Acad Sci U S A*, *41*(3), 144–50.

Kimura, M. (1957). Some problems of stochastic processes in genetics. *Annals of Mathematical Statistics*, *28*, 882-901.

Kimura, M., & Weiss, G. (1964, Apr). The Stepping Stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, *49*(4), 561–576.

Kingman, J. (1982). The Coalescent. *Stochastic processes and their applications*, *13*, 235-248.

Latch, E., Dharmarajan, G., Glaubitz, J., & Rhodes, O. (2006). Relative performance of Bayesian clustering software for inferring substructure and individual assignment at low levels of population differentiation. *Conserv Genet*, *7*, 295-302.

Latch, E., Heffelfinger, R., Fike, J., & Rhodes, O. (2009, Apr). Species-wide phylogeography of North American mule deer (*Odocoileus hemionus*): cryptic glacial refugia and postglacial recolonization. *Mol Ecol*, *18*(8), 1730–1745. Available from `http://dx.doi.org/10.1111/j.1365-294X.2009.04153.x`

Latch, E., Scognamillo, D., Fike, J., Chamberlain, M., & Rhodes, O. (2008). Deciphering ecological barriers to North American river otter (*Lontra canadensis*) gene flow in the Louisiana landscape. *J Hered*, *99*(3), 265–274. Available from `http://dx.doi.org/10.1093/jhered/esn009`

Long, J. (2007). Generalized hierarchical modeling (ghm) program (2.1 ed.) [Computer software manual].

Long, J., & Kittles, R. (2003, Aug). Human genetic diversity and the nonexistence of biological races. *Hum Biol*, *75*(4), 449–471.

Lynch, M. (1988). Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution*, *5*(5), 584–599.

Lynch, M., & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, *152*(4), 1753–66.

Makoff, A., Lai, T., Barratt, C., Valentin, A., Moran, N., Asherson, P., et al. (2009, Dec). High-density SNP screen of sodium channel genes by haplotype tagging and DNA pooling for association with idiopathic generalized epilepsy. *Epilepsia*. Available from `http://dx.doi.org/10.1111/j.1528-1167.2009.02473.x`

Malecot, G. (1959). Remarks on the decrease of relationship with distance, following paper by M. Kimura. In *Cold spring harbor symp. quant bio.* (Vol. 20, p. 52-53).

Mengel-From, J., Wong, T. H., Morling, N., Rees, J. L., & Jackson, I. J. (2009, Dec). Genetic determinants of hair and eye colour in the Scottish and Danish populations. *BMC Genet*, *10*(1), 88. Available from `http://dx.doi.org/10.1186/1471-2156-10-88`

Menozzi, P., Piazza, A., & Cavalli-Sforza, L. (1978, Sep). Synthetic maps of human gene frequencies in Europeans. *Science*, *201*(4358), 786–792.

Milot, E., Weimerskirch, H., & Bernatchez, L. (2008, Apr). The seabird paradox: dispersal, genetic structure and population dynamics in a highly mobile, but philopatric albatross species. *Mol Ecol*, *17*(7), 1658–1673. Available from `http://dx.doi.org/10.1111/j.1365-294X.2008.03700.x`

Nei, M. (1987). *Molecular evolutionary genetics*. New York: Columbia University Press.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., et al. (2008, Nov). Genes mirror geography within europe. *Nature*, *456*(7218), 98–101. Available from `http://dx.doi.org/10.1038/nature07331`

Parra, F. C., Amado, R. C., Lambertucci, J. R., Rocha, J., Antunes, C. M., & Pena, S. D. J. (2003, Jan). Color and genomic ancestry in Brazilians. *Proc Natl Acad Sci U S A*, *100*(1), 177–182. Available from `http://dx.doi.org/10.1073/pnas.0126614100`

Patterson, N., Price, A. L., & Reich, D. (2006, Dec). Population structure and eigenanalysis. *PLoS Genet*, *2*(12), e190. Available from `http://dx.doi.org/10.1371/journal.pgen.0020190`

Peng, B., Amos, C., & Kimmel, M. (2007, Mar). Forward-time simulations of human populations with complex diseases. *PLoS Genet*, *3*(3), e47. Available from http://dx.doi.org/10.1371/journal.pgen.0030047

Peng, B., & Kimmel, M. (2005, Sep). simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, *21*(18), 3686–3687. Available from http://dx.doi.org/10.1093/bioinformatics/bti584

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., & Reich, D. (2006, Aug). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, *38*(8), 904–909. Available from http://dx.doi.org/10.1038/ng1847

Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–59.

Pritchard, J. K., & Rosenberg, N. (1999, Jul). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, *65*(1), 220–228. Available from http://dx.doi.org/10.1086/302449

Queller, D., & Goodnight, K. (1989). Estimating relatedness using genetic-markers. *Evolution*, *43*(2), 258–275. (English)

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005, Nov). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A*, *102*(44), 15942–15947. Available from http://dx.doi.org/10.1073/pnas.0507611102

Rencher, A. (2002). *Methods of Multivariate Analysis*. John Wiley & Sons, Inc.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, *67*(2), 175–185.

Rosenberg, N., Li, L., Ward, R., & Pritchard, J. (2003, Dec). Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet*, *73*(6), 1402–1422. Available from http://dx.doi.org/10.1086/380416

Rosenberg, N., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J., & Feldman, M. (2005, Dec). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet*, *1*(6), e70. Available from http://dx.doi.org/10.1371/journal.pgen.0010070

Rosenberg, N., Pritchard, J., Weber, J., Cann, H., Kidd, K., Zhivotovsky, L., et al. (2002, Dec). Genetic structure of human populations. *Science*, *298*(5602), 2381–2385. Available from http://dx.doi.org/10.1126/science.1078311

Schwartz, M., & McKelvey, K. (2009). Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. *Conserv Genet*, *10*, 441-452.

Seldin, M., Shigeta, R., Villoslada, P., Selmi, C., Tuomilehto, J., Silva, G., et al. (2006, Sep). European population substructure: clustering of Northern and Southern populations. *PLoS Genet*, *2*(9), e143. Available from http://dx.doi.org//journal.pgen.0020143

Shriner, D., Adeyemo, A., Gerry, N., Herbert, A., Chen, G., Doumatey, A., et al. (2009). Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. *PLoS One*, *4*(12), e8398. Available from http://dx.doi.org/10.1371/journal.pone.0008398

Shtir, C., Marjoram, P., Azen, S., Conti, D., Marchand, L., Haiman, C., et al. (2009). Variation in genetic admixture and population structure among Latinos: the Los Angeles Latino eye study (LALES). *BMC Genet*, *10*, 71. Available from http://dx.doi.org/10.1186/1471-2156-10-71

Shyn, S., J.Shi, Kraft, J., Potash, J., Knowles, J., Weissman, M., et al. (2009, Dec). Novel loci for major depression identified by genome-wide association study of sequenced treatment alternatives to relieve depression and meta-analysis of three studies. *Mol Psychiatry*. Available from http://dx.doi.org/10.1038/mp.2009.125

Song, B., Windsor, A., Schmid, K., Ramos-Onsins, S., Schranz, M., Heidel, A., et al. (2009, Mar). Multilocus patterns of nucleotide diversity, population structure and linkage disequilibrium in *Boechera stricta*, a wild relative of *Arabidopsis*. *Genetics*, *181*(3), 1021–1033. Available from http://dx.doi.org/10.1534/genetics.108.095364

Tajima, F. (1995, Dec). Effect of non-random sampling on the estimation of parameters in population genetics. *Genet Res*, *66*(3), 267–276.

Tero, N., Aspi, J., Siikamki, P., Jklniemi, A., & Tuomi, J. (2003, Aug). Genetic structure and gene flow in a metapopulation of an endangered plant species, *Silene tatarica*. *Mol Ecol*, *12*(8), 2073–2085.

The International HapMapConsortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–912.

Thompson, E. (1975). Estimation of pairwise relationships. *Annals of Human Genetics*, *39*(Oct), 173–188.

Urbanek, M., Goldman, D., & Long, J. (1996, Sep). The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: a new approach to measuring gene identity reveals asymmetric patterns of divergence. *Mol Biol Evol*, *13*(7), 943–953.

Voight, B., & Pritchard, J. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*, *1*(3), e32.

Walhund, S. (1928). Zuzammensetzung von populationen und korreltioserscheinungen vom standpunkt der vererbungslehre aus betrachet. *Hereditas*, *11*, 65106.

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., et al. (2007, Nov). Genetic variation and population structure in Native Americans. *PLoS Genet*, *3*(11), e185. Available from `http://dx.doi.org/10.1371/journal.pgen.0030185`

Watterson, G. (1975, Apr). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, *7*(2), 256–276.

Weir, B., Anderson, A., & Hepler, A. (2006). Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet*, *7*(10), 771–80.

Weir, B., & Cockerham, C. (1984). Estimating F-statistics for the analysis of poulation structure. *Evolution*, *38*, 1358-1370.

Wright, S. (1929). Evolution of dominance. *American Naturalist*, *63*, 556-561.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97-159.

Wright, S. (1937). The distribution of gene frequencies in populations. *Proc Natl Acad Sci U S A*, *23*, 307-320.

Wright, S. (1940). Breeding structure of populations in relation to speciation. *Am. Naturalist*, *74*, 232-248.

Wright, S. (1943, Mar). Isolation by distance. *Genetics*, *28*(2), 114–138.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*(4), 323–354.

Zhang, J., Niyogi, P., & McPeek, M. (2009). Laplacian eigenfunctions learn population structure. *PLoS One*, *4*(12), e7928. Available from `http://dx.doi.org/10.1371/journal.pone.0007928`