Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
BIOLOGIA EVOLUTIONAISTICA E AMBIENTALE

CICLO XXIII°

COORDINATORE Prof. Guido Barbujani

ON THE STUDY OF GENETIC STRUCTURE IN HUMAN POPULATIONS AND THE EFFECTS OF DEMOGRAPHIC HISTORY, CONSANGUINITY, AND STUDY DESIGN ON DETECTION, WITH INVESTIGATIONS OF HUMAN EVOLUTIONARY MODELS, ARCHAIC INTROGRESSION, AND NATURAL SELECTION

Settore Scientifico Disciplinare BIO/18

Dottorando	**Tutore**
Dott. Ferrucci Ronald Robert	Prof. Barbujani Guido
_______________________________	_______________________________
(firma)	*(firma)*

Anni 2008/2010

Università degli Studi di Ferrara

DOCTOR OF PHILOSOPHY IN
EVOLUTIONARY AND ENVIRONMENTAL BIOLOGY

23rd CYCLE

COORDINATOR Prof. Guido Barbujani

ON THE STUDY OF GENETIC STRUCTURE IN HUMAN POPULATIONS AND THE EFFECTS OF DEMOGRAPHIC HISTORY, CONSANGUINITY, AND STUDY DESIGN ON DETECTION, WITH INVESTIGATIONS OF HUMAN EVOLUTIONARY MODELS, ARCHAIC INTROGRESSION, AND NATURAL SELECTION

Settore Scientifico Disciplinare BIO/18

PhD Student
Mr. Ronald R. Ferrucci

(firma)

Tutor
Prof. Guido Barbujani

(firma)

Years 2008/2010

UNIFE FORMAZIONE POSTLAUREA

Your E-Mail Address
ronald.ferrucci@gmail.com
Subject
Genetics
Io sottoscritto Dott. (Cognome e Nome)
Ferrucci Ronald Robert
nato a
New Haven, Connecticut, United States
Provincia
Connecticut
il giorno
8 January 1972
avendo frequentato il corso di Dottorato di Ricerca in:
Evolutionary and Environmental Biology
Ciclo di Dottorato
23
Titolo della tesi in Italiano
SULLO STUDIO DELLA STRUTTURA GENETICA DI POPOLAZIONI UMANE E LE EFFETTI DELLA STORIA DEMOGRAFICA, LA CONSANGUINEITÀ, E DESIGN STUDIO SUL RILEVAMENTO, CON LE INDAGINI DI HUMAN EVOLUTIVA MODELS, ARCAICA INTROGRESSIONE, E LA SELEZIONE NATURALE
Titolo della tesi in Inglese
ON THE STUDY OF GENETIC STRUCTURE IN HUMAN POPULATIONS AND THE EFFECTS OF DEMOGRAPHIC HISTORY, CONSANGUINITY, AND STUDY DESIGN ON DETECTION, WITH INVESTIGATIONS OF HUMAN EVOLUTIONARY MODELS, ARCHAIC INTROGRESSION, AND NATURAL SELECTION
Titolo della tesi in altra Lingua Straniera
Tutore - Prof:
Guido Barbujani
Settore Scientifico Disciplinare (SSD)
BIO/18
Parole chiave (max 10)
population genetics, evolution, Neanderthal, structure, CEPH HGDP, Cilento, genetic structure, conservation genetics, microsatellites, model-based clustering
Consapevole - Dichiara
CONSAPEVOLE --- 1) del fatto che in caso di dichiarazioni mendaci, oltre alle sanzioni previste dal codice penale e dalle Leggi speciali per l'ipotesi di falsità in atti ed uso di atti falsi, decade fin dall'inizio e senza necessità di alcuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni; -- 2) dell'obbligo per l'Università di provvedere al deposito di legge delle tesi di dottorato al fine di assicurarne la conservazione e la consultabilità da parte di terzi; -- 3) della procedura adottata

dall'Università di Ferrara ove si richiede che la tesi sia consegnata dal dottorando in 4 copie di cui una in formato cartaceo e tre in formato .pdf, non modificabile su idonei supporti (CD-ROM, DVD) secondo le istruzioni pubblicate sul sito : http://www.unife.it/dottorati/dottorati.htm alla voce ESAME FINALE – disposizioni e modulistica; -- 4) del fatto che l'Università sulla base dei dati forniti, archivierà e renderà consultabile in rete il testo completo della tesi di dottorato di cui alla presente dichiarazione attraverso l'Archivio istituzionale ad accesso aperto "EPRINTS.unife.it" oltre che attraverso i Cataloghi delle Biblioteche Nazionali Centrali di Roma e Firenze. --- DICHIARO SOTTO LA MIA RESPONSABILITA' --- 1) che la copia della tesi depositata presso l'Università di Ferrara in formato cartaceo, è del tutto identica a quelle presentate in formato elettronico (CD-ROM, DVD), a quelle da inviare ai Commissari di esame finale e alla copia che produrrò in seduta d'esame finale. Di conseguenza va esclusa qualsiasi responsabilità dell'Ateneo stesso per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi; -- 2) di prendere atto che la tesi in formato cartaceo è l'unica alla quale farà riferimento l'Università per rilasciare, a mia richiesta, la dichiarazione di conformità di eventuali copie; -- 3) che il contenuto e l'organizzazione della tesi è opera originale da me realizzata e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto l'Università è in ogni caso esente da responsabilità di qualsivoglia natura civile, amministrativa o penale e sarà da me tenuta indenne da qualsiasi richiesta o rivendicazione da parte di terzi; -- 4) che la tesi di dottorato non è il risultato di attività rientranti nella normativa sulla proprietà industriale, non è stata prodotta nell'ambito di progetti finanziati da soggetti pubblici o privati con vincoli alla divulgazione dei risultati, non è oggetto di eventuali registrazioni di tipo brevettale o di tutela. --- PER ACCETTAZIONE DI QUANTO SOPRA RIPORTATO

Firma Dottorando

Ferrara, lì ____________________________

Firma del Dottorando __________________________________

Firma Tutore

Visto: Il Tutore Si approva Firma del Tutore

____

INDEX

1 Introduction

Charles Darwin ignited a revolution in biological thought with the publication of *On the Origin of Species* in 1859. The idea that plants and animals, as well as other life forms, were not fixed in form and had, in fact, evolved from common ancestors to fill various niches to which they were supremely adapted was a momentous one; one that would shake up the scientific world in the years to come. Though Darwin did not address human origins until publication of *The Descent of Man* in 1874, he portended what would come, writing that "[l]ight will be thrown on the origin of man and his history." Just as Darwin's book had earlier revolutionized biological thought, the advent of genetics and molecular technologies at the end of the last century has revolutionized research into human origins. Genetic studies have shown that humans share a common ancestor with chimpanzees, that all human populations are descended from an ancestral population in East Africa, and have enabled the reconstruction of a great deal of the evolutionary history of the human species.

One of the more prolific areas of recent research in human population genetics has been on the structure of human populations, which has seen great strides made in the last few years due to extensive collection of DNA samples from worldwide populations, both at the global [1,2], continental [3–5], and fine scale or geographically limited [6,7] levels, and from the generation of large amounts of data using high-throughput technologies [8–10]. Studies of genetic structure, the observation that populations are not genetically homogenous, are important in disease-gene association studies, conservation genetics, and anthropological research.

1.1 Genetic Variation

1.1.1 Genetic Structure

The study of genetic variation has done much to inform us of human origins and relationships. Furthermore, interest in studies of genetic variation have proven applicable to the medical field, particularly in association studies, where cryptic population structure may flummox attempts at discovering genetic variants associated with susceptibility to complex disease. Genetic structure is the observation that populations are not genetically homogenous, that is, that they can be divided, genetically, into subpopulations or groups of subpopulations. This results when some sort of barrier (*e.g.*, geographic, linguistic, cultural), or adequate distance, results in isolation between a set of populations [11]. Consequently, isolation creates drift within populations, resulting in the differential distribution of alleles between and among populations, which, over time, results in differences in allele frequencies among them [12]. The subsequent divergence between populations is referred to as structure.

1.1.2 Genetic Drift

Genetic drift is a function of finite sampling: random gametes are sampled from a limited gene pool during reproduction to be represented in the next generation. As a result, some alleles tend to be overrepresented in different populations in the next generation, some underrepresented, and some may disappear completely. Over time, this continuous, cumulative random selection of gametes, known as genetic drift, results in differential distribution of alleles between and among populations, which is seen as differences in gene frequencies. Genetic drift, and thus genetic structure, is affected by a number of important evolutionary processes, including effective size, divergence time, and gene flow. Effective size (N_e) refers to the size (*i.e.*, number of breeding individuals) of the optimum population showing allelic frequencies similar to the one being considered with the same degree of inbreeding, which is in Hardy-Weinberg equilibrium. Effective size is often smaller than census size. Decreasing the effective size within populations and increasing the divergence time between populations results in increased genetic drift within and between populations, thus resulting in greater genetic differentiation. To the contrary, gene flow counteracts the effect of drift by breaking down differentiation between populations, through transmission of new alleles into populations.

1.1.3 Coalescence

The coalescent is a stochastic process providing a backward-in-time approach for reconstructing the evolutionary history of a set of DNA sequences or genomes [13]. Going back in time, lineages are randomly chosen and merged at each generation until converging at the *most recent common ancestor* (*MRCA*) of all lineages [13]. Coalescent based simulations have proven useful in population genetic studies for their ability to reconstruct complex demographic scenarios, including bottlenecks, founder effects, and population fissions and fusions, among others. Their usefulness extends to testing hypotheses of human evolutionary models and differentiating natural selection from neutral evolution [13]. Figure 1 shows a sample coalescence for N=10 haploid individuals, without recombination or mutation. In the absence of recombination, coalescent history will often be constructed first and then mutations are placed along lineages descending from the *MRCA*. With recombination in coalescent simulations, lineages bifurcate as well as coalesce. As such, different pieces of DNA or different genomes will have slightly different topologies. Topologies are essentially phylogenetic representations of evolutionary history. Thus, with recombination different evolutionary histories result from different chromosomes or chromosomal pieces (see Figure 2). In a coalescent simulation, diploid individuals will be modeled following *2N* haploid individuals. Prior to discovery of the coalescent, researchers used forward in time approaches, which were very time and computationally intensive. The Coalescent is more efficient because it only has to keep track of lineages that survive [13].

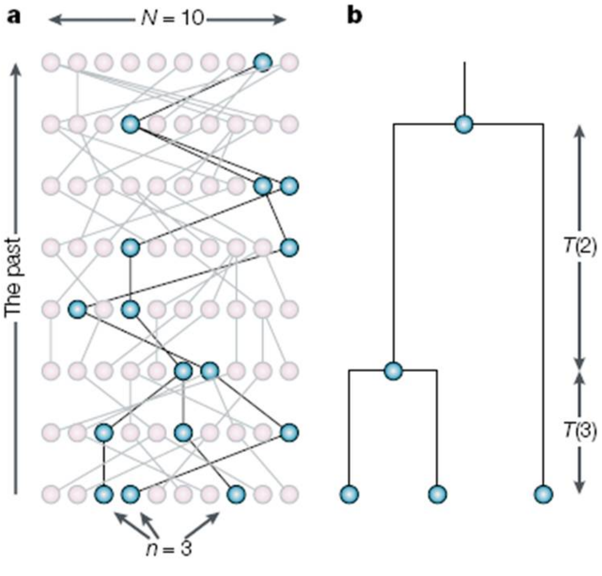


Figure 1: (a) Sample coalescent for 10 haploid individuals tracing back 8 generations. The sub-genealogy for three single genomes that coalesce back to a single genome are shown in (b), with blacks lines showing the coalescent lineages. Coalescent times for the first and second coalescent events are indicated by T(3) and T(2), respectively [13].

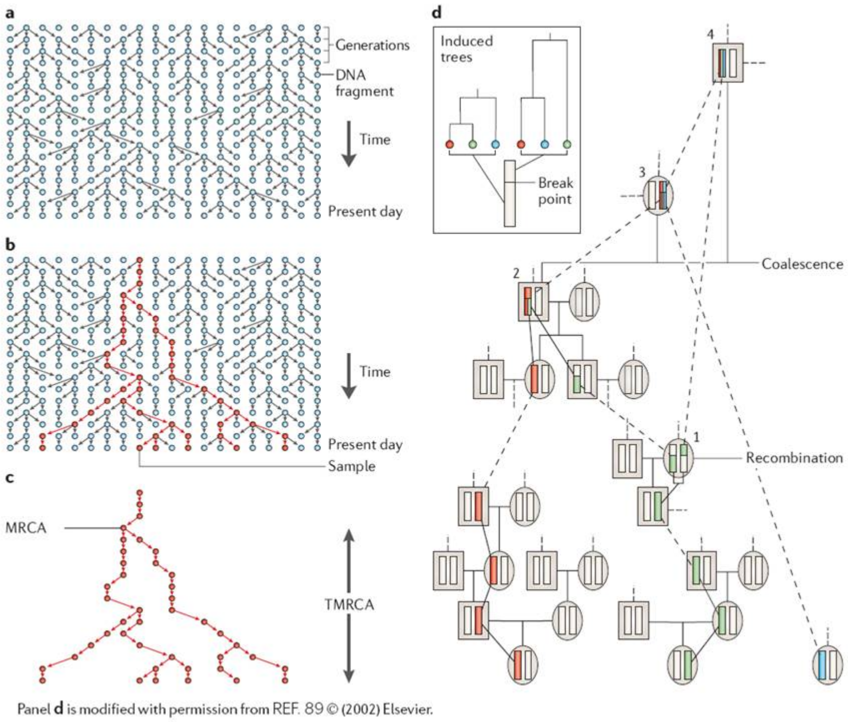


Figure 2: (a) Coalescent history for 20 haploid individuals tracing back 17 generations, with (b) and (c) red lines showing the coalescent history for six individual genomes tracing back to a single genome. (d) shows the coalescent history for a set of chromosomes with recombination [14].

1.1.4 Markers

Genetic variation is studied using a number of genetic markers located in non-coding (*i.e.*, neutral) genomic regions. This is important because it ensures that differences in allele frequencies among populations are due solely to the effects of drift, mutation, and gene flow, and not to selective pressures. *Single-nucleotide polymorphisms* (SNPs) are markers that are based on variation in the state of nucleotides at a particular site and are often, though not always, biallelic. More than two states may exist, but it is the exception not the rule. With SNPs, there will often be a sequence of DNA that is mostly non-variant, that is, the DNA sequence at most sites is the same in all individuals. However, at certain sites, approximately once every 1500 nucleotides, there will be differences in the state of the nucleotide among individuals. Though often occurring in neutral areas of the genome, SNPs may also appear in genes and can affect regulation, if occurring in the promoter region, or protein structure, if appearing in exons. *Microsatellites*, or *short tandem repeat polymorphisms* (STRs or STRPs) are markers that show variation in the number or length of repeats of a two to six letter motif. These often have higher mutation rates than SNPs, and are therefore useful in more recently diverged species or populations, such as in the human species. As such, they were the marker of choice for the CEPH HGDP [1,2], as well as extensions of the panel in Native American [5], South Pacific [3], African [4], and South Asian populations [15]. SNP markers, however, are more amendable to high-throughput genotyping and have been the marker of choice for more recent analyses [6, 7, 10], including the CEPH HGDP [16]. A number of models of microsatellite mutation have been proposed. The more generally accepted one is the Single Stepwise Mutation Model (SSM, or SMM), where addition or deletion of repeats is expected to occur in single-steps [17]. Other models include a Generalized Stepwise Mutation Model, GSM [18], whereby addition or deletion of repeats may be more than one repeat according to a geometric probability distribution, and the Infinite Alleles Model (IAM), where any repeat number can change to any other repeat number. That is, any number of repeats may be added or deleted [19, 20]. The IAM is also the accepted mutational model for SNPs and genes.

1.1.5 Metrics

F-statistics. Wrights F-statistics [21] is one of the more widely used methods devised for analyses of population structure. F-statistics are measures of genetic variation within and between populations and regions. Essentially, they describe the degree to which genetic variation amongst a set of populations may be attributed to variation within a local population (that is, between individuals within a local population), between populations within a region, and between regions. Another way to look at them is as a measure of correlation between alleles randomly chosen from one of these levels of apportionment [22]. One of the more widely used measures is F_{ST}, which is the correlation between alleles chosen randomly from within a subpopulation compared to the population as a whole [22]. It may also be seen as the proportion of genetic variation that is attributed to genetic differences between subpopulations [22]. As such, it may be used to measure differentiation between or among populations, and can be considered a measure of genetic distance between pairs of populations [22]. F_{ST} can be seen as a function of migration, effective size, and divergence times. As divergence times increase and effective population sizes decrease, genetic drift increases within populations, thus increasing F_{ST} between and amongst them [11]. As well, reduced migration rates also increase the effect of genetic drift, thus increasing F_{ST}. F_{ST} ranges from 0, when subpopulations are completely identical, to 1, when all subpopulations are fixed for different alleles [11].

Gene Identity [23]. Gene identity is the probability that two randomly sampled copies of an allele, drawn from the same or different populations, are identical by state. Gene identity is also a measure of heterozygosity. As gene identity increases, heterozygosity decreases, and vice versa. We refer to gene identity using the notation $J_{k,l}$, where $k = l$ indicates gene identity taken from within the same local populations, and $k \neq l$ indicates gene identity taken from between two different local populations. Gene identity is calculated from gene frequencies. If calculating gene identity from within the same local population, we use the equation,

$$J_{k,k} = \sum_{u} p_{ku}^2$$

where p_{ku} is the frequency of the u^{th} allele, with summation over all alleles.

Gene identity at a locus between two different local populations is calculated using the equation,

$$J_{k,l} = \sum_{u} p_{ku} p_{lu}$$

where p_{ku} is the frequency of the u^{th} allele in the k^{th} population and p_{lu} is the frequency of the u^{th} allele in the l^{th} population. As above, summation occurs over all alleles. Other statistics, such as genetic distances or fixation indices, used in population genetic studies can be seen as functions of gene identities. As such, these statistics can be estimated from gene identities. However, the reverse is not necessarily true.

In analyses, it is helpful to organize sets of gene identities for pairs of populations in a matrix, which we label J. Values on the diagonal correspond to gene identities within populations, while values on the off diagonal correspond to gene identities between populations. As long as populations in the matrix are organized according to their tree of descent, gene identities in these matrices will take a block like pattern where populations sharing the same *MRCA* will be grouped together. In addition, gene identity between a pair of population will be equal to the gene identity between their $MRCA$.

Kinship. Another way to measure relatedness between individuals is through coefficients of relatedness measuring the probability of *identity by descent*, such as estimated from pedigree data. This is usually seen as the proportion of ancestry shared between pairs of individuals. There are a number of approaches for calculating kinship indices. The first, devised by Sewall Wright in 1922 [24], is the path counting approach, which traces all possible pathways shared by two individuals through common ancestors. Here, it is useful to think of relatedness between individuals as a set of loops connecting them through common relatives, and as a function of a hypothetical inbreeding coefficient in possible descendents, which we would determine using the equation

$$\Phi_{jk} = F_i = (\frac{1}{2})^K$$

with Φ_{jk} being the kinship between individuals j and k, and F_i being the inbreeding coefficient in their hypothetical offspring, i, and where K is the number of ancestors in the loop connecting one allele in the “offspring” to the other. The calculation of kinship and inbreeding is additive across pathways. The coefficient of kinship is the probability that two individuals share alleles identical by descent, while the coefficient of inbreeding is the probability that two alleles within an individual are identical by descent. More recently, recursive approaches that are more computationally efficient have been devised [25] and implemented [26, 27], which are useful when considering large pedigrees.

Clusteredness. Model-based clustering programs output membership coefficients that quantify a sample's probability of belonging to a cluster or population (another way of looking at it, is as the proportion of ancestry from each cluster). These can be visually represented using a bar graph display as produced from the *distruct* program [28] or implemented in the gui version of *structure* [29]. Though this approach may be adequate when considering low numbers of experiments, a quantitative approach for summarizing the data is needed for large numbers of experiments. We chose clusteredness [30] as a metric for determining significant differentiation between populations. Essentially, clusteredness is the average membership coefficient for all sample individuals across all clusters, across all individuals, standardized by the number of clusters K, as per the following equation,

$$G = \frac{1}{I} \sum_{i=1}^{I} \sqrt{\frac{K}{K-1} \sum_{k=1}^{K} (q_{ik} - 1/K)^2}$$

where I = number of individuals, K = number of clusters, and q_{ik} = the membership coefficient of the i^{th} individual to the k^{th} cluster. Standardizing by the number of clusters is important because it allows for comparison across different numbers of K and is more intuitive to understand. Instead of average membership coefficients running from $1/K$, as an indication of no clustering, to 1.0, as an indication of complete clustering, G runs from 0.0 to 1.0. Thus, also, one can easily identify a minimum clusteredness as indicative of structure that applies to all levels of K. For example, 0.5 corresponds to an average membership coefficient of 0.75 (or 0.9, to 0.95).

Symmetric Similarity Coefficient. Model-based clustering algorithms may sometimes produce different outputs from the same data when using different starting points (*i.e.*, different random numbers). It is useful to consider the similarity between runs to determine whether results are reliable. The following equation, implemented in the *clumpp* software, has been devised for quantifying the extent of similarity between pairs of runs,

$$SCC(Q_i, Q_j) = 1 - \frac{\min \|Q_i - P(Q_j\|_F}{\sqrt{\|Q_i - S\|_F \|Q_j - S\|_F}}$$

where Q_i and Q_j are I x K matrices of membership coefficients for runs i and j and, with columns (K) corresponding to clusters and rows (I) corresponding to individuals [31]. Each element is the membership coefficient of each individual to each cluster. P is a permutation of the columns, with the minimum taken over all permutations, for K permutations. F is the Frobenius matrix norm and S is a probability matrix of K = (number of clusters) columns, with all elements = $1/K$ [2, 31]. Use of this equation assumes that comparisons are across runs that have the same number of clusters, i. Rosenberg et al. [2], supplemental materials, provides suggestions for the interpretation of results: values of 0.85-1.00, nearly all individuals have nearly identical membership coefficients between runs; 0.4-0.85, most, but not all, individuals have nearly identical membership coefficients between runs; 0.1-0.4, some clusters have the same individuals but other clusters differ between runs; and <0.1, little to no similarities between clusters.

Genetic Distance. Another measure of difference between populations is genetic distance, such as Nei's *minimum genetic distance.* Nei's distance measure is a function of allele frequencies and can be calculated from the gene identity within and between local populations using the equation,

$$D_{kl}^2 = \sum (p_{ki} - p_{li})^2 = J_{kk} + J_{ll} - 2J_{kl}$$

where p_{ki} and p_{li} are the i^{th} allele frequencies in the k^{th} and l^{th} populations, respectively, and J_{kk}, J_{ll}, are gene identities within the k^{th} and l^{th} populations, respectively, and J_{kl}, between the k^{th} and l^{th} populations.

1.1.6 Approaches to Studies of Genetic Structure

Traditional. Traditional approaches to the study of genetic structure often involve apportioning genetic variation into three categories: between individuals within local populations, between populations within a specific region or grouping, and between regions (or other grouping) within the dataset as a whole. The proportion of variation attributed to each category gives an indication of the genetic structure between populations [12]. A similar approach, termed Analysis of Molecular Variance (AMOVA), essentially an extension of ANOVA, was proposed for analysis of molecular data [32]. AMOVA has given some statistical standing to the study of genetic apportioning by providing a method for determining significance. Most estimates of genetic structure in human populations provide that 85% of genetic variation is found within a population, with 15% left between populations, either within the region or between regions [12]. The traditional approach contains an inherent flaw in that it assumes division of populations neatly into regions, and without regard to evolutionary or demographic history. As Long et al. [33] show, discrete compartmentalization of human populations may be flawed in light of human evolutionary history.

Model-Based Clustering. The development of model-based clustering methods in recent years has revolutionized the study of genetic structure in human populations. Previously, analyses of genetic structure required pre-defined assignment of samples into populations. This required the assumption that populations from which individuals were sampled correspond to the actual population from which they originated. While this may sometimes be a valid assumption, particularly in the case of isolated populations, large increases in migration between populations have occurred over the last 100 years, even in isolated populations (*i.e.*, breakdown of isolates [34]). In our own sample populations from southern Italy, we have found evidence of individuals being sampled from populations other than the one into they were born. This was evidenced through comparison with parents, or other relative, that were also sampled in our study. Also, some sample individuals are descended from parents from two populations. Model-based clustering approaches allow researchers to overcome limitations such as these by using Bayesian [35], expectation-maximization [36], or maximum likelihood [37] methods to determine the optimum distribution of sample individuals into a set of clusters based on statistically optimized allele frequency distributions. In addition, these methods can also group populations with other populations to which they are more closely aligned genetically and, in the presence of admixture can determine the probability of an individual belonging to a cluster or the proportion of their genome having membership to a cluster.

In addition, output from model-based clustering programs, such as *structure* [35], can help to identify the most likely number of clusters in the sample. One may use the original approach suggested by Jonathan Pritchard of analyzing posterior priors output by his program, *structure*, for a set of K number of clusters, from 1, 2, 3,. . . ,N, with N being the maximum number of clusters estimated [29]. However, this approach sometimes fails to provide a definitive answer. Another approach, suggested by Evanno [38], looks at the change in log likelihood as a guide. Finally, an additional *ad hoc* approach more recently suggested by Pritchard is to examine the membership coefficients estimated for the samples [29]. When individuals have a tendency to be placed into a single cluster as opposed to being distributed across a number of clusters, this may be considered to be the correct K.

The structure algorithm is a Markov Chain Monte Carlo method, which uses burnin length and number of iterations following burnin to minimize the effect of the starting configuration and optimize estimation of parameters, respectively [29]. Burnin length and iterations following burnin are determined by the user. Adequate length of burnin and number of iterations following burnin are important in ensuring convergence of the MCMC chain [29]. These should be chosen so that parameters converge, that is, reach an equilibrium, before data is collected.

Generalized Hierarchical Modeling [23]. While clustering populations using model-based clustering programs can be useful for identifying clusters of closely-related populations, cryptic population structure, and inter-individual relatedness, it may be useful to turn to other methods to test hypotheses on population structure and relatedness, methods that allow testing of complex scenarios of structure. One such method is *generalized hierarchical modeling* (*ghm*), which may also be termed *generalized analyses of molecular variance*, which provides for a structured approach to testing hypotheses. Generalized hierarchical modeling uses a system of equations developed by Anderson to fit models to data [39]. Application of these systems of equations to genetic data were first adapted by Cavalli-Sforza and Piazza in 1975 [40]. Models may include simple models, such as an island model, whereby all populations originate from a single ancestral population at one time and evolve independently, or more complex hierarchical models, whereby each population or set of populations branches off from earlier populations. The former model may also be called an independent regions model, while the latter models may also be called, and more easily-understood as, ‘nested’ models. These hierarchical models are assumed to be strictly nested, where the previous entry is a superset and the next entry is a subset [41].

While fitting models to the data, *ghm* estimates two sets of the researchers chosen metric, gene identity for example, for each model to be evaluated: expected and realized. Realized gene identities are probably the most intuitive, they are gene identities as measured from the data without regard to a particular model. We may also call them raw gene identities. Expected gene identities are those estimated from a given hierarchical model fitted to the gene identity matrix. To test the fit of models, we use a likelihood ratio statistic, Cavalli-Sforza and Piazza’s treeness statistic

$$\Lambda_{0\Omega} = v \cdot (ln|det(\hat{\Sigma}_0| - ln|det(\hat{J}| + tr\hat{J}\hat{\Sigma}_0^{-1} - r)$$

which is distributed as a χ^2 statistic with degrees of freedoms equal to $(r(r+1)/2)$ minus the number of parameters needed to fit the tree, where r is the number of populations, to determine the fit of the model to the data. v is the number of independent observations. An observation is an allele at a locus, and the number of independent observations is equal to (number of alleles at all loci - number of loci - 1). This factor is most appropriate when allele frequencies are equal for each marker. Ongoing research hints that a more appropriate determinant for independent observations, when this requirement does not hold, is the number of effective alleles. The number of effective alleles is the inverse of the homozygosity, subtracting one and multiplying by the number of loci. In any case, the treeness statistic is still useful for testing the fit of a model and comparing the fit between models. Models can be ranked by their χ^2 values, with lower values corresponding to better fitting models [42]. $\hat{\Sigma}_0$ is the matrix of expected gene identities determined by the model and $\hat{J}$ is the matrix of observed gene identities. If the model fits perfectly, that is, if the observed and expected gene identities differ by no more than would be expected from genetic and statistical sampling, Λ will be equal to the number of degrees of freedom [42]. Oftentimes, researchers will want to test different models to determine which one fits the data the best. These may begin as a parameter rich model that is subsequently reduced, as parameter poor models to which higher level groupings are subsequently added, or as models with different hierarchical structures that are independent of one another (such as models of Native American language evolution, see [42]).

1.2 Isolated Populations

Due to their expected genetic homogeneity and common environmental background, resulting from isolation, geographical isolates are prime subjects for investigating complex genetic traits and identifying common alleles involved in susceptibility to complex diseases [43–46]. However, even in populations considered to be particularly homogenous [47], the presence of undetected population stratification may result in the presence of groups of closely-related individuals, considered to be a major confounding factor in disease-gene association studies [48–50]. Unfortunately, lacking reconstruction of genealogical relationships, detection of population stratification, from the presence of inter-individual relatedness or cryptic structure, is most difficult [51,52]. Most reconstructed pedigrees have been limited in completeness, spanning at most a few generations. As such, researchers must rely on indirect evidence of genealogical relatedness, such as through studies of genetic variation. We felt it fruitful, therefore, to explore comparisons between genetic and genealogical data.

For isolated populations, in our work we used a set of closely-related populations from the Cilento National Park in southern Italy, Gioi and Cardile (Figure 3). Historical sources document that the village of Gioi was settled first in the 9^{th} century by Greek immigrants, with a secondary settling of Cardile in the 18^{th} century through an exodus of Gioi residents. Though located approximately 6 km apart, the villages of Gioi and Cardile experienced high levels of reproductive isolation until the 20^{th} century. As in the case of many isolated villages around Europe, a breakdown of isolates occurred following World War II that saw large scale migration from the Cilento region.

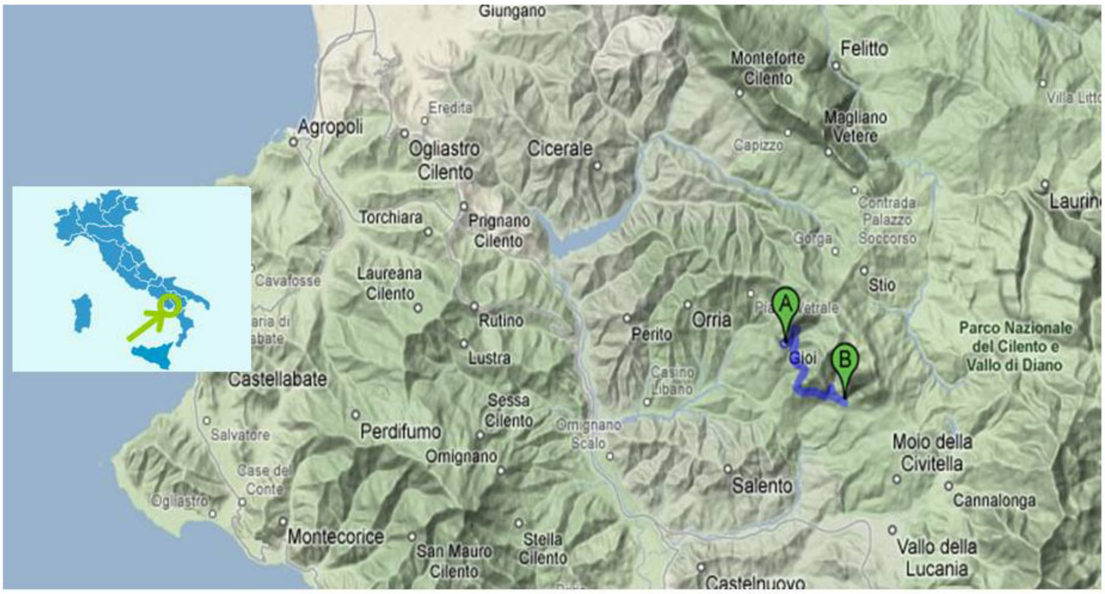


Figure 3: Map showing location of Gioi and Cardile within the Cilento National Park in southern Italy.

1.3 Study Design

1.3.1 Scantily-Differentiated Populations

Bayesian clustering algorithms have shown to be effective at identifying genetic clusters in human populations [35,36,53]. Detection of differentiation and clustering with these methods may be affected by a number of factors, such as number of markers considered and sample sizes [30,54], mutation rate [55], and geographical dispersal of sample populations [30, 56]. The usefulness of model-based clustering methods in describing genetic structure has been demonstrated in studies of globally-distributed, genetically well-differentiated populations [2, 30, 54], as well as in more closely-related, geographically-limited, but still genetically well-differentiated [3–5], ones. However, their efficacy in highly closely-related, scantily-differentiated populations has been limited to few studies involving real populations [6, 57], or to simulations of scantily-differentiated populations [58]. Nearly all other previous studies have concentrated on populations among which genetic differences are substantial. As such, their efficacy for the analysis of scantily-differentiated populations is still an open question.

We used simulations to study the behavior of *structure*, one of the more extensively used programs, in the presence of limited genetic differentiation. Here, we varied effective size and divergence times, along with differences in sample sizes and markers numbers. See Figure 4 for our model.

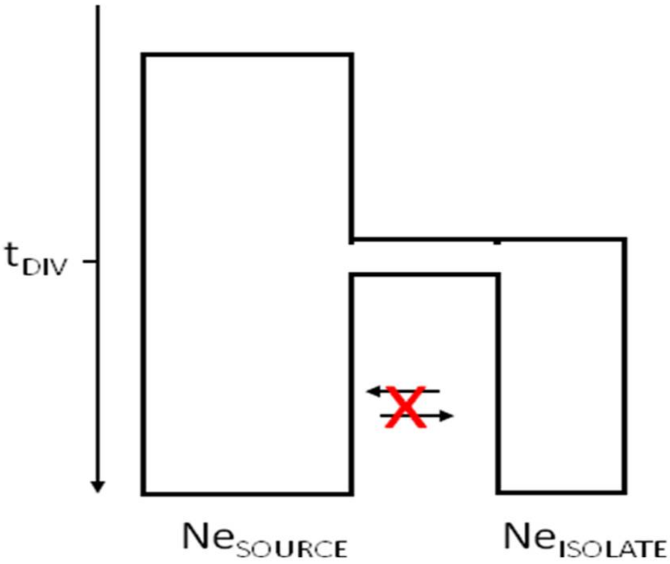


Figure 4: Diagram of model showing isolated population diverging from its original source population.

1.3.2 Relatedness

Increased consanguinity is common in scantily-differentiated populations. Though considered to be an important factor influencing inferences of genetic structure [59], the effect of related individuals on the performance of model-based clustering methods is not well documented, regardless of the existence of methods for estimation [60,61]. In fact, a number of studies in human populations have taken care to avoid including related individuals in order to limit the potential confounding effect of consanguinity [4, 5, 15, 62]. However, the performance of model-based clustering methods, and the effects of study design, is largely unknown for consanguineous populations. To the best of our knowledge, the only other study to investigate the effect of consanguinity on clustering analyses was in rainbow trout, a species that differs from human populations in being polyandrous, and showing high fecundity and variance in reproductive success [63]. Further, their study consisted of a single-family group consisting of siblings and half siblings plus otherwise completely unrelated, or at least not obviously related, individuals, and their simulations modeled similarly structured populations, rather than a number of groups of related individuals with complex networks and varying degrees of relatedness that we see in human populations. Here, we test model-based clustering approaches in a set of consanguineous populations controlling for different levels of relatedness. We make use of an extensive genealogical dataset dating back three centuries to reconstruct genealogical links between sample individuals. In this part of the study, we identified and removed consanguineous individuals to investigate the effect of reducing relatedness in a sample on the performance of *structure*.

1.3.3 Effect of Marker Numbers on the Detection of Structure

While the study by Vitart [57] showed differentiation amongst closely-related populations among the Dalmatian islands, the observation of differentiation is weak and mainly between villages that have approximately 0.02 or greater F_{ST} (paired villages with lower F_{ST} values tend to cluster with other populations and do not differentiate separately). Appropriately, Latch [58] showed that population identification by Bayesian methods breaks down amongst populations with F_{ST} below 0.02. However, conclusions from both these studies were based on just a handful of markers—26 in the former and 10 in the latter—, which we have found may be too low to detect clustering in scantily-differentiated populations. Analyses of European populations, a lowly-differentiated group of populations with F_{ST} <0.007 [2] documents a recognizable structure with 377 markers [2]. One may conclude, therefore, that it is difficult to judge whether failure to identify structure in scantily-differentiated populations is due to either the prevalence of the effects of gene flow over those of genetic drift, or to an inadequate number of markers in the analysis. However, even rather large numbers of markers may not necessarily be adequate, as in the case of linguistically-differentiated populations in India [15].

The question thus remains, is there a specific lower level of differentiation beyond which, barring highly related populations, genetic structure is undetectable through model-based clustering methods? Or, is the detection of structure with different levels of F_{ST} dependent on the number of markers available? Essentially, would increasing the number of markers analyzed increase the possibility of observing structure in populations with lower differentiation? Bamshad et al. [54] shows that the accuracy at which structure infers group membership for large-scale (*i.e.*, continental or geographical) groupings is indeed affected by marker numbers, whereby increasing the numbers of markers analyzed increases correct predictions of individuals into their sampled continental populations. Further, Rosenberg et al. [30] shows a marker number effect on clusteredness, that is, the degree to which populations cluster with one specific group, which we may also refer to as the ability to detect structure, and demonstrated a contribution of marker numbers to clustering success in the case of chicken breeds [64].

Recently, Morin [65] showed that the power to detect structure increases with increasing number of markers analyzed in a data set, and that this observation was affected by differentiation level. Both moderately ($F_{ST} = 0.01$) and scantily differentiated populations ($F_{ST} = 0.0025$) show increasing power up to 75 markers. However, high F_{ST} populations achieve high power approaching 75 markers (power >0.9), while low F_{ST} populations do not even reach 50%. Increasing the number of SNPs analyzed might have increased the observed power in scantily differentiated populations. These authors also show that minor allele frequencies seem to have no effect on results, indicating that SNP choice is not an issue. Although the marker system studied here were SNPs, we expect a similar relationship with microsatellites, our marker system of choice.

1.3.4 SNP Markers and the Detection of Structure

We focus on microsatellites in this study, as they have been one of the most commonly used markers systems, and are more useful in relatively recently diverged populations (*i.e.*, most human populations) because of their higher mutation rate.

However, it may also be of interest to ask how marker choice affects our ability to detect structure considering different numbers of markers. Lao et al. [8] suggests that more STRP than SNP markers would be needed for detection of population structure because of their high mutation rate. However, this is counter-intuitive; one would assume that given the low mutation rate of SNPs, STRPs would be more advantageous given the recent separation and shared ancestry of human populations (particularly for closely-related populations). In fact, STRPs have historically been used in studies of closely related populations because of their high mutation rate and high degree of variability. Though, Lao et al. [8] was referring to the case of using carefully ascertained SNPs.

1.3.5 Effect of Marker Choice on the Detection of Structure

Finally, Rosenberg et al. [66] showed that choice of markers, that is, choosing markers that are more informative, can affect identification of population clustering, whereby datasets of markers chosen to be informative are more useful in identifying genetic structure than are datasets of randomly chosen markers, reducing the numbers of markers needed for structure analyses. As an aside, it would be of interest to know whether informativeness of markers influences the ability to detect differentiation between pairs of populations.

1.4 ALDH2

Acetaldehyde dehydrogenase 2 (ALDH2) is an enzyme involved in the alcohol metabolism pathway, specifically converting acetaldehyde to acetate (see Figure 5, top). A broken copy of the gene (see Figure 5, bottom), referred to as ALDH2*2, has been identified that causes accumulation of acetaldehyde in carriers [67], resulting in a flushing reaction in the face [68]. In addition to this flushing reaction, which may reduce alcoholism because of its unpleasantness, accumulation of acetaldehyde is also toxic and carcinogenic [69]. As a dominant acting allele heterozygotes are also affected. Interestingly, this allele is found only in East Asian populations, and is a common allele in those populations [70]. We concern ourselves that this allele may have been the subject of recent selection on the East Asian branch. First, the high frequency of the allele indicates that it would have to be an old allele, but old alleles tend to be more dispersed globally, and second, the negative effect of the allele would imply some counter advantage to it. Further, genetic loci that are common in one population tend to also be found dispersed amongst populations either because they are shared by descent or because they are transferred between populations through gene flow. In addition, a second allele is found only in populations on the OOA branch. Here, we explore the evidence for natural selection of ALDH2*2 through simulation of genes with features similar to ALDH2, comparing the distribution of alleles with frequencies similar to those observed in ALDH2*2 and the OOA limited allele.

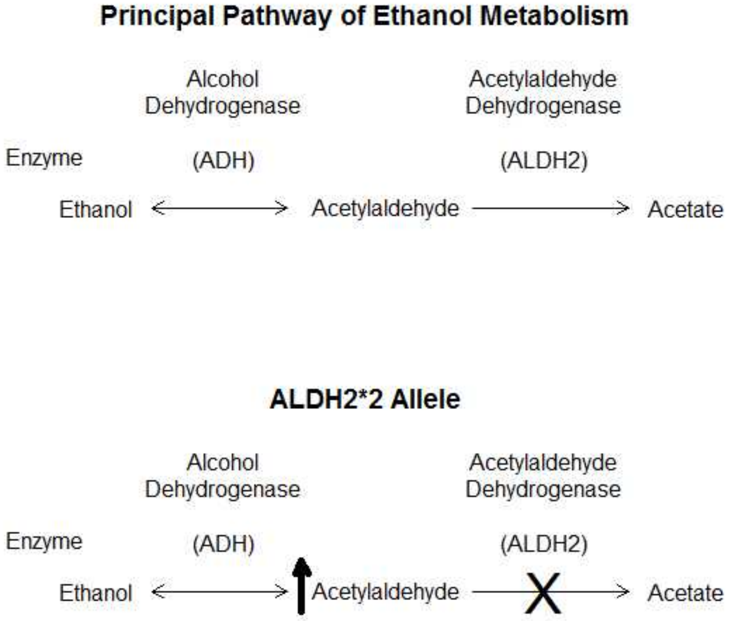


Figure 5: Ethanol metabolism pathway, showing the breakdown of ethanol to acetate, through an acetylaldehyde intermediate (top). The ALDH2*2 allele causes the accumulation of acetylaldehyde (bottom).

1.5 Models of Human Evolution

A number of models of human origins and evolutionary have been presented. These can easily be tested against predictions of gene identity patterns.

Independent Regions. Under the independent regions model, a set of modern human populations split from a single ancestral population, with little to no subsequent gene flow between or among them, allowing them to evolve independently of one another. This is analogous to the multi-regional model of human origins, but with modern human populations splitting from an ancestral modern population rather than an ancestral archaic population. This may also be referred to as an island model. We can also consider a nested independent regions model, where each geographical region splits from a single ancestral population, and all populations within the geographical region originate from that geographical population.This model predicts highest gene identities within local populations, lower ones between populations within geographical regions, and lowest between populations in different regions. Human genetic variation has been shown to be inconsistent with the independent regions model [71].

Isolation by distance. Isolation by distance is a function of the interacting effects of drift and long-term gene flow with neighboring populations [11]. This occurs when individuals have greater likelihood of mating with individuals in neighboring populations than they do with ones located father away and results in individuals having a greater probability of sharing relatives in neighboring populations and lower probability with populations located at greater distances, which can be seen as a gradual decline in genetic relatedness with distance. Under this model, we expect declining gene identities with geographical distance between populations. Human genetic variation has been shown to be consistent with a model of long-range and local gene flow amongst Eurasian populations [71].

Serial Founder Effects. Previous studies have shown human genetic diversity to be consistent with a serial founder effects (*SFE*) model [72]. Under the serial founder effects model, we see a set of population fissions, whereby each new population originates from a previous founding population and subsequently gives rise to future populations. However, subsequent investigations show that human genetic variation is better explained by a nested version of SFE model, where a series of major founder effects marked by bottlenecks occurs in major geographical regions, followed by a series of founder effects within regions [71]. Under the serial founder effects model, we expect to see: 1) lowest gene identities in the original, ancestral population (essentially, the root of the tree), 2) a sequential increase in gene identities between regions with each subsequent founder population, 3) a layered pattern of gene identity variation between regions and between populations within regions [71], and 4) equal or similar gene identity between all pairs of populations that share the same *MRCA*. In addition, we expect a tree of descent according to the pattern of fissions and length of branches on the tree that are proportional to the ratio of evolutionary time to effective size.

Though testing of hierarchical models show that human genetic variation is, generally, consistent with predictions from the SFE model, there are still some deviations from the model that need to be accounted for. Specifically, non-African populations show greater diversity than expected under the SFE, resulting in greater than expected genetic distances between African and non-African populations. We predict that early modern humans leaving Africa interbred and admixed with archaic populations that they encountered along the way, and that this injection of new genetic variants resulted in the increased variation that we observe.

1.6 Archaic Admixture

One of the most fundamental and contentious issues in human evolutionary studies has been the conflict between competing models of human origin. For years, the two reigning theories have been Multiregional Evolution (MRE) and the Out-of-Africa (OOA), or replacement, theories [73]. The MRE states that modern *Homo sapiens* originated, from early hominids, separately in Europe, Africa, and Asia and evolved independently in these regions, with subsequent gene flow between regional populations [74]. In contrast to the MRE, the OOA theory posits that all modern humans evolved from a single population in Africa, approximately 200,000 years ago, and then replaced existing *Homo* species in the rest of the Old World as they left Africa [75]. Much of the genetic evidence has favored the OOA theory to the detriment of the MRE theory. However, other theories, compromises between MRE and OOA, have been presented in recent years. Many of these new ideas are modified versions of OOA, with allowance for admixture [76]. Thus, the question being asked now is not whether humans originated in one location (OOA) or independently in the three main regions of the old world separately (MRE), but whether there was admixture between the population of modern humans leaving Africa 60,000 years ago and populations of archaic humans that they encountered along the way.

Studies of Neanderthal mitochondrial DNA (mtDNA) unequivocally show no evidence of admixture with modern humans [56, 77]. Comparisons of mtDNA sequences between modern human and Neanderthal mtDNA sequences showed that the genetic differences between Neanderthal and current modern humans were significantly different than that between current modern human populations such that there could have been no admixture [77]. Further, a comparison of Neanderthal and early modern human sequences found that the early modern human mitochondria contained zero Neanderthal-like DNA [56]. Finally, mtDNA from Cro-magnon fossils shows variation within the range of current modern humans but distinct from that of Neanderthals [78]. Recent simulations also support an African replacement model for human evolution [79].

However, mitochondrial DNA is only one locus, and therefore can only tell part of the story [80]. If gene flow from Neanderthals into early modern humans were purely paternal, it would not leave a signature on mtDNA. Further, even if gene flow from Neanderthal were maternal, genetic drift could have removed all trace of admixture from the mitochondrial portion of the human genome [81]. In addition, we may expect that given the high mutation rate of mtDNA and length of time since possible admixture events, mutation could have erased signals of admixture as well. Of course, any possible signature on the modern human genome would be dependent on the admixture rate [82]. If Neanderthals made a tiny genetic contribution, there is a greater chance that those genes would be lost through drift. Estimates of rates of admixture range from 15 [83, 84] to 25% [56], to less than 0.1% [85]. Thus, if admixture rates were on the low end of the estimated admixture spectrum (*i.e.*, <0.1%), it is not likely that we would see its signature in the modern human genome.

Recent technological advances have allowed amplification and sequencing of Neanderthal autosomal DNA, analyses of which, in comparison with modern human DNA, have revealed evidence of admixture between archaic and modern humans [86]. Here, we show that evidence for archaic admixture with early modern humans can be detected in modern human genomic diversity, which may explain deviations from the SFE model.

1.7 Autocorrelation

Analyses of spatial autocorrelation, the dependence of values of a variable with values at different, usually adjoining, locations [87], are informative of demographic and evolutionary processes. Initially conceived by geographers and statisticians, spatial autocorrelation methods were readily adopted by biologists for the study of genetic [88], morphometric [89], and ecological data [90]. However, the usage of spatial autocorrelation was limited to allele frequencies of individual markers or polymorphisms. Subsequently, Bertorelle and Barbujani [91] further adapted these methods for use with molecular data, such as DNA sequences, which they termed Autocorrelation Indices for DNA Analysis or AIDA. AIDA statistics II and cc, modified forms of the spatial autocorrelation indices Moran's I and Geary's c, respectively, measure sequence or haplotype similarity with distance [92]. Non-AIDA spatial autocorrelation analyses for the study of genetic variation has been likened a multivariable approach, because results from multiple analyses were often compared, in contrast to multivariate approaches such as principal components analysis (PCA) [93]; thus, AIDA can be viewed as an approach transforming spatial autocorrelation analysis from a multivariable to a multivariate approach.

Demographic and evolutionary processes, such as genetic drift, gene flow, and natural selection affect genetic variation generated by the action of mutation. While the effects of selection may sometimes be of interest to researchers, it is the patterns created by the opposing forces of drift and gene flow that are of interest in the case of spatial autocorrelation analyses. While genetic variation between and among populations is often thought of in discrete terms, and in fact genetic variation may be observed as discontinuous and can therefore be studied as such [35, 94], differences between and among populations may sometimes be more subtle, exhibiting clinal or continuous, often geographically-based, variation [87, 95]. Further, some evolutionary processes, such as isolation by distance, may require approaches that take into account subtle genetic variation, methods that do not require populations to be measured as completely distinct entities. Methods such as spatial autocorrelation are useful in this regard. Indeed, since not all genetic variation is discontinuous it would be inappropriate to treat all data as such. In addition, while clinal patterns of variation are easily observed by locating gene or haplotype frequencies on a map (using some graphical visualization), the statistical significance of these patterns may be open to question. Autocorrelation analyses may add statistical weight to these observed patterns.

In human populations, AIDA statistics have been mainly used to study Y-chromosome single-nucleotide polymorphisms [96–103] and mtDNA [77, 92, 104–107] data. In addition, limited usage has been done on X-chromosomal data [108] and coding sequences [109].

While an independent extension to AIDA for microsatellite data has been available for a number of years [110], autocorrelation analyses on microsatellites in humans have been limited to a few studies, using alleles at single markers [111], haplotype frequencies associated with single, specific haplogroups [112], or individual loci separately with AIDA [113], or haplogroups with [114] or without [115] AIDA. As such, the full power of genetic variability associated with the rapid mutation rate of Y chromosome microsatellites has not yet been realized. During our recent update of AIDA into the python programming language, which was done to increase the computational limit of AIDA (*i.e.*, to free AIDA from limitations on number and length of sequences associated with previous versions), we decided to incorporate an option for microsatellite analyses into the new version. Along with this decision, it was determined that autocorrelation analyses of Y-chromosome microsatellites in European, both continental and regional, populations was long overdue. This also allowed us to test our new version.

2 SCOPE OF THE THESIS

a) To show that genealogical relationships within and between populations are expressed in genetic data

b) To show how demographic history can affect the detection of clustering in scantily-differentiated populations

c) To show how consanguinity affects the signal of structure in closely-related populations

d) To show that the detection of differentiation between pairs of populations is dependent on sample size and marker numbers considered

e) To show that the number of markers needed to identify differentiation between pairs of populations is dependent on their level of divergence

f) To develop a new version of AIDA with increased computational power and extended use for analyzing microsatellites

g) To show that modern human genetic variation is consistent with a serial founder effect model, long-range gene flow between populations in different regions of Eurasia, and introgression of archaic DNA into modern humans on the OOA branch

h) To show that the limited distribution of a detrimental allele in Asian populations is not consistent with neutral expectations

3 MATERIALS AND METHODS

3.1 Datasets

3.1.1 Cilento

From southern Italy, we have data available on 1356 individuals from two villages, Gioi (n = 882) and Cardile (n = 474), corresponding to nearly all current residents, located within the Cilento National Park. Two sets of data were collected for these samples: genealogical and genetic. Our genealogical data is composed of 20,383 birth records, also documenting mortality and parental relationships, spanning the past four centuries, collected from registry office and parish archives. These data allowed us to reconstruct genealogical relationships between all pairs of extant individuals. For genetic data, we obtained 1122 microsatellites, with average marker spacing of 3.6 cM and mean marker heterozygosity of 0.70, from a genome-wide scan performed by the DeCode genotyping service on DNA extracted from peripheral blood. Genetic data was obtained from all study samples.

3.1.2 CEPH HGDP

Our primary dataset of worldwide populations is the CEPH Human Genome Diversity Panel (HGDP), which was initially a set of 1050 individuals distributed amongst 52 world-wide populations [1, 2] originally typed on 377 microsatellites, and subsequently extended to 783 microsatellites [30]. Subsequently, additional samples and populations have been typed for many of these same markers in Native American [5], South Asian [15], African [4], and South Pacific [3] populations. We term this the ‘extended’ CEPH HGDP (see Figure 6 for locations of populations, black dots correspond to waypoints). This dataset contains 5179 individuals from 245 populations genotyped at 619 common microsatellite markers. The original CEPH populations have also been typed for SNP markers [116]. For our first project using the Cilento populations, comparing results from genealogical and genetic analyses, we use European populations from the original CEPH HGDP dataset, using a subset of 36 markers shared between that dataset and our populations from Cilento. For the ALDH2 project, we use 16 populations from the original CEPH dataset. For our project analyzing marker numbers needed to observe differentiation between populations with different levels of divergence, we use the Wang et al. [5] dataset of original CEPH data plus additional Native American populations. For the archaic admixture project, we use subsamples taken from the 'extended' CEPH HGDP. Geographical distances between populations are great circle distances calculated using the indicated waypoints taking into account inferred directions of dispersal across land.

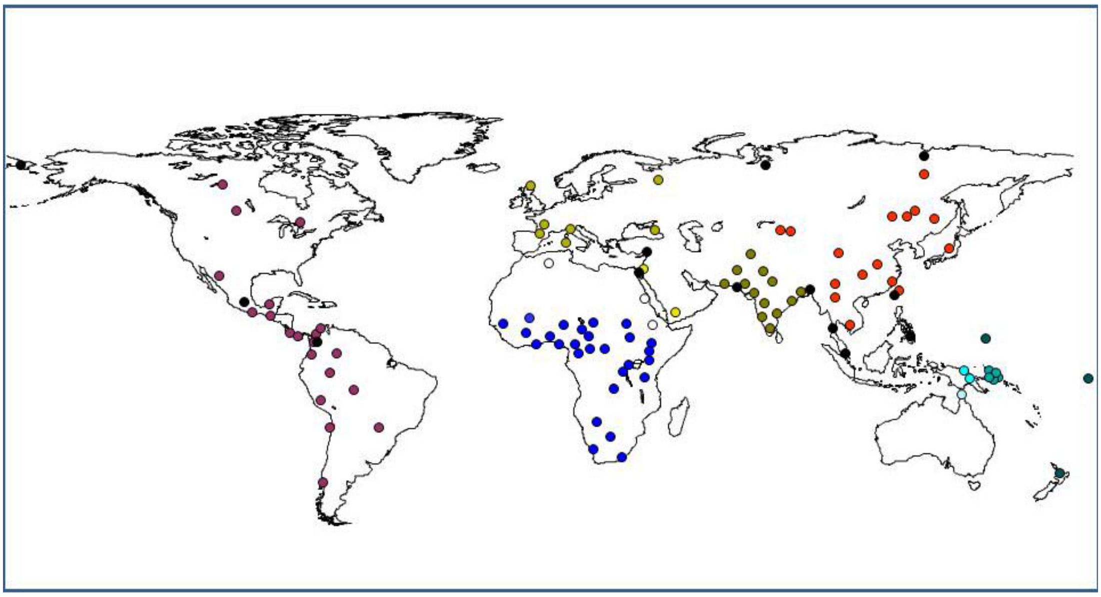


Figure 6: Map of world with locations of populations from extended CEPH HGDP. Black dots indicate waypoints.

3.1.3 ALDH2

We obtained DNA samples from four populations each from four geographical regions: Africa (Mbuti, 5; Biaka, 5; Nigerian, 8; and Kenyan, 8), Europe (Iberian, 9; Southern Eussian, 9; Italian, 10; and Russians from Moscow, 10), Asia (Japanese, 10; Han Chinese, 10; Aboriginal Tiawanese, 10; and Southeast Asians, 10), and America (Mexican Indians, 5; Mayans, 4; Surui, 5; and Karitiana, 5). These DNA samples were obtained from The Coriell Institute for Biomedical Research in Camden, NJ. Following PCR amplification, these 123 samples were sequenced for 5387 base pairs in the ALDH2 gene. Primers for PCR and sequencing were designed using the human references sequence from the July 2003 build (NCBI Build 34), accessed on the UCSC genome browser (http://genome.uscs.edu). Sequencing was done using dye-terminator sequencing at the University of Michigan DNA sequencing core facility with Applied Biosystems Big-Dye reagents on an Applied Biosystems automated sequencer. Fourteen variable sites were identified. The Seqman module in the DNASTAR package was used to analyze chromatograms and perform alignments, and to assemble the ALDH2 segment for all individuals.

3.1.4 Y-STRP data

Four datasets were used in this study for testing this new version of AIDA: one Europe-wide dataset [115] and three country-limited ones. Of the country-limited datasets, Italy [117] contains samples found within the larger European dataset though with some additional markers, while the other two, Finland [118] and the United Kingdom [119], are completely independent. This gives us the opportunity to test whether nested populations share similar patterns (that is, does the pattern of the more locally focused population mirror that of the larger population from which it is derived), plus two other independent populations for comparison. Actually, the YBase dataset contains a single Finnish and three UK populations, but the samples are completely different from the ones we study separately. In addition, an additional dataset of mtDNA sequences from the UK from Sykes [119] were included in this study since it includes the same populations as the Y-STRP dataset and therefore allows us to examine autocorrelation patterns of Y-STRP with an additional uniparental genetic system typed in the same populations, and Y-SNPs from the same Finnish individuals were also analyzed to compare the results of two uniparental genetic systems from the same chromosomes, but with different mutation rates.

For the European sample, 90 out of 91 available populations were considered in our analyses. One population, Turks from Bulgaria, were removed from analyses because identification of the sampling location would be difficult. As such, 12,666 samples were analyzed here. Population samples are dispersed across Europe, including those sampled from: Portugal (4), Spain (10), Netherlands (5), Germany (14), Italy (10), Sweden (8), Norway (6), Poland (6), Austria (3), Estonia (2), Russia (2), Switzerland (2), and France (3). The number of populations from each country are in parentheses. Single populations are available from Albania, Greece, Hungary, Bulgaria, Denmark, Finland, the Ukraine, Slovenia, England, Latvia, Romania, Ireland, Lithuania, Croatia, Belgium, Belorussia, and Turkey. In addition, a reduced dataset of 7710 was considered, reducing population samples to 100 individuals or less to enable calculation of confidence intervals for Europe as the full dataset crashes the memory on one cluster we use, and takes too much time to calculate on the other cluster (the cluster has a restriction on the length of time each job can be run). Analyzing a reduced population also allows us to reduce the potential effect of uneven sample sizes from some populations containing larger samples (for example, some German populations contain 500+ samples, while some smaller ones only contain 40). Seven Y-STRs were available for the European YBase dataset: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS393, and DYS392.

The Italian dataset contains 1175 samples from the 10 Italian populations considered in the European dataset. Sampling locations are distributed around Italy: the Marches, Puglia, Tuscany, Liguria, Sicily, Lombardy, Emilia Romagna, Veneto, Latvia, and Umbria. Nine Y-STRs were considered in our analyses of Italy: the seven found in the larger European dataset as well as DYS385a and b. Data for nine Finnish provinces were contained within the Finnish dataset: Southern Ostrobothnia, Häme, South-Western Finland, Swedish-Speaking Ostrobothnia, Satakunta, Northern Karelia, Southern Karelia, Northern Savo, and Northern Ostrobothnia. Ten Y-STRs were considered for this analysis: The seven considered in the European dataset plus DYS385a and b, and DYS388. In addition, 12 biallelic SNPS were considered. For the United Kingdom analyses, 2425 samples from 18 populations sampled from around Great Britain were considered. Populations were sampled from: Argyll, the Borders, Central England, East Anglia, Grampian, the Hebrides, the Highlands, Ireland, Isle of Man, London, North England, the Northern Isles, the Orkney Islands, Northumbria, South England, South-west England, Strathclyde, Tayside & Fife, and Wales. The seven Y-STRs found in the European dataset were considered. Sykes also considered 3 additional markers, which we do not use here because not all individuals were typed for those markers. As well, 3685 mtDNA sequences were also considered.

All analyses were done using both Infinite Alleles (IAM) and Stepwise Mutation (SMM) models. For all populations, latitude and longitude data were taken from publicly available data. For the Italian populations, exact sampling locations were not given so individual major cities were chosen for coordinate locations. Sampling locations reported for individuals in Sykes dataset were birthplaces of

Table 1: Distribution of pedigrees in the genealogical dataset and kinship values of sampled individuals for the whole genealogical dataset and largest pedigree in the dataset [120]

		Gioi-Cardile	**Gioi**	**Cardile**
WHOLE DATASET	individuals	5272	4190	2384
	pedigrees	63	45	19
	sampled individuals	1356	882	474
	mean kinship for sampled individuals (± s.d.)	0.0030 ± 0.0147 (max=0.292)	0.0040 ± 0.0177 (max=0.291)	0.0086 ± 0.0235 (max=0.292)
	median kinship (25%-75% quartiles)	0.000061 (0.0000-0.0015)	0.0065 (0.0000-0.0023)	0.0035 (0.0006-0.0082)
LARGEST PEDIGREE	individuals	5165	4113	2354
	pedigrees	1	1	1
	sampled individuals	1274	828	446
	mean kinship for sampled individuals (± s.d.)	0.0034 ± 0.0155 (max=0.293)	0.0045 ± 0.0188 (max=0.291)	0.0097 ± 0.0246 (max=0.293)
	median kinship (25%-75% quartiles)	0.000197 (0.0000-0.0019)	0.000862 (0.0000-0.0026)	0.000862 (0.0009-0.0088)

the paternal grandfather and maternal grandmother of the sampled individual for Y and mtDNA data, respectively [119].

3.2 Genealogical analyses

From our genealogical pedigree data, we constructed pedigrees spanning 350 years (15-17 generations), from which we calculated kinship coefficients Φ_{ij} between pairs of individuals, i and j, using Karigl's method [25] implemented in the KinInBCoeff module of the CC-QLS package [26]. Pedigrees for the whole dataset and largest single pedigree are summarized in Table 1. We also used modules from the PyPedal pedigree analysis package [27] to construct pedigrees and calculate pedigree-based relationship coefficients, which were halved to obtain kinship coefficients [24].

3.3 Simulation and Sampling Schemes

All coalescent simulations were performed with SimCoal2 [121], version 2.1.1. For all our simulations, we assume a generation time of 25 years [122], an instantaneous growth model, that is, immediate change in population size, and 100% merging of populations unless indicated.

3.3.1 Scantily-Differentiated Populations

In our simulations, we modeled an isolated population of varying effective size, N_e = 500, 1000, or 2000 individuals diverging from a source population of fixed effective size N_e = 20,000, at varying times in the past, t_{div} = 250, 500, 750, 1000, or 1250 years (see Figure 4). Here, we used a generalized stepwise mutational model [18] with a mutation rate of $\mu = 7x10^{-4}$ [123]. For each set of divergence time and effective size, we simulated 1000 microsatellite loci. Our sample output for each population from SimCoal2 was equal to the effective size of the isolate. From these master datasets, we assembled diploid datasets for further analyses by randomly sampling with replacement pairs of haploid genotypes at m markers (m=20, 40, 80, or 200) for n diploid individuals (n=10, 20, or 50) for each population.

3.3.2 Family group analyses

Using inferred genealogical relationships, we created 8 kin-groups of samples defined by restricted relatedness (Table 2). Kin-group zero is the most inclusive, including all pairs of individuals, while kin-group 7 is the most exclusive, including only unrelated, or apparently unrelated, individuals. That is, estimated kinship between all pairs of individuals in this kin-group are zero. For all other kin-groups, we restrict access to the group according to relatedness thresholds as defined in Table 2. Here, we analyzed all pairs of individuals and removed one element of the pair when their kinship is greater than the degree of allowed relatedness.

3.3.3 Marker Numbers

Our approach for studying the effect of numbers of markers on the performance of *structure* was to create a set of natural experiments by analyzing pairs of pop-

Table 2: Features and number of individuals for kin-groups

KIN-GROUP	ALLOWED RELATEDNESS	GIOI	CARDILE
0	All	897	492
1	Up to Sibling	450	227
2	Up to Half-Sibling	245	118
3	Up to 1^{st} Cousin	187	88
4	Up to 2^{nd} Cousin	137	58
5	Up to 3^{rd} Cousin	134	51
6	Up to 4^{th} Cousin	127	51
7	None	129	50

ulations from publicly-available data sets of human populations. For the initial analysis, we considered the dataset of Wang et al. [5], which is comprised of the original CEPH HGDP populations with an additional 24 Native American populations, herein referred to as the 'STR678' dataset. In total, this dataset contains 78 populations, giving us a possible 3003 population pairs, typed at 678 autosomal microsatellites. For comparison to another commonly used genetic marker system, we used the dataset of Conrad et al. [116], herein referred to as the 'SNP' dataset. This dataset contains 53 (reported as 52, but Han is divided into "Han" and "Han-NChina") populations typed at 2834 single-nucleotide polymorphisms, giving us 1378 possible population pairs. For the testing of informativeness of markers, we considered the original 377 autosomal microsatellite dataset of Rosenberg et al. [2], herein referred to as the 'CEPH377' dataset. This dataset was used because it is the dataset in which Rosenberg [66] identified informativeness of markers. Here we have 52 populations, giving us 1326 possible population pairs. Datasets are summarized in Table 3.

Table 3: Datasets included in this study, showing number of populations (Npops), number of markers (Nmarkers), marker types (MarkerType), and number of possible population pairs considered (PopPairs)

Dataset	Npops	Nmarkers	MarkerType	PopPairs	Citation
Wang	78	678	Microsat	3003	Wang et al. (2007)
CEPH377	52	377	Microsat	1326	Rosenberg et al. (2003)
SNP	43	2834	SNP	1378	Conrad et al. (2006)

Pairs of populations in each dataset were divided according to their pairwise

F_{ST} into mildly- to highly-differentiated ($F_{ST} \geq 0.01$) and scantily differentiated (F_{ST} <0.01) populations. Though F_{ST} determined by structure output was used in clustering analyses, we used *Arlequin* [124] for initial calculation of F_{ST}.

3.4 Clustering analyses

Genetic structure was analyzed using the *structure* software package [35], under assumptions of admixture, correlated allele frequencies (the F model), and no prior population information [35, 125]. We use the F model because it has improved capability of differentiating similar, that is, with very low F_{ST}, but distinct populations. This was useful for us, because we were either investigating a set of populations with known low F_{ST}, or studying a range of populations, including ones with low F_{ST}. Each study followed different study designs for structure as follows:

3.4.1 Comparison of Genealogical versus Genetic data

For K clusters from 1 to 8, we performed 50 runs of structure with a burnin length of 20,000 followed by 10,000 iterations. For each K, we determined the posterior probability of clustering using the average logarithmic probability of data across runs. However, this approach was inconclusive in identifying the number of clusters in our data. Thus, we used Evanno's [38] method of using the second order rate of change to identify the optimal number of clusters in the data. Finally, we input resulting matrices of membership coefficients into *clumpp* [126], using the LargeKGreedy algorithm, to look for possible multimodality or label switching. We find no obvious multimodality among runs, with average similarity (G' values) of 0.99, 0.79, and 0.89 for K=2, 3, and 4 respectively. *Distruct* [28] was used to graph membership coefficients quantifying the probability of each sample individual belonging to each cluster. We performed two sets of runs using the above conditions. First, we analyzed Gioi and Cardile alone at 239 loci for all 1356 individuals. Second, we analyzed Gioi and Cardile with the 161 European samples available in the CEPH HGDP, using subsets of 37 and 22 individuals from Gioi and Cardile, respectively, at 36 markers shared between the two datasets.

3.4.2 Scantily-differentiated populations

For this project, we used *structure*, version 2.3 [127]. Usually, researchers run structure for a set K clusters and then determine the optimum number of clusters in the data from comparisons across results from different K values. However, since we already knew how many independent populations were in our data, we only used K=2 for our analyses. As in the previous study, we used a burnin length of 20,000 followed by 10,000 iterations. In addition, we assessed convergence by comparing results from different runs under the same conditions, using higher numbers of iterations following burnin. We found no difference in our results. From our simulated populations, we randomly sampled n=10, 20, or 50 diploid individuals, at m=20, 40, 80, or 200 markers, for all combinations of $tdiv$ and N_e for clustering analyses. We conducted 100 replicates of each simulation, giving us 18,000 experiments across all combinations of variables.

3.4.3 Consanguineous populations

For each of our kin-groups defined in Table 3, we randomly sampled n=10, 25, or 50 individuals from each village, considering their genotypes at m=20, 40, 80, or 200 randomly chosen loci, from the 1122 microsatellite markers available in our total dataset, for clustering analyses. For each kin-group and each n, all individuals are considered at the same randomly chosen markers. Here, we conducted 50 replicates for each of 96 combinations of kin-group, sample size, and number of markers, giving us a total of 4200 *structure* runs. As shown in Table 3, the numbers of individuals in each kin-group diminishes with increasing threshold of relatedness. Hence, we have an increased chance of sampling the same individuals since the pools we are sampling from decrease in size with increasing relatedness.

3.4.4 Marker Numbers

For all analyses, 10 runs were performed using a burnin length of 5000 followed by 1000 iterations. For a limited number of experiments to determine the adequacy of burnin length and iterations, membership coefficients generated by *structure* were input into *clumpp* and analysed using the LargeKGreedy algorithm [126]. Output from *clump* provides estimation of similarity between runs.

As per Rosenberg et al. [30], we tested a few subsamples of population pairs for different burnin length and iteration combinations to determine whether increased burnin length and/or iterations following would have a significant effect on results, using similarity coefficients (SCP) calculated with *clump* [30]. Considering Rosenberg's suggestion for similarity of membership coefficients between runs of 0.85 - 1.0 indicating high similarity ([2], supplemental material), we find no significant difference using burnin lengths of 5000 and iterations following burnin of 1000, and any other increased burnin/iteration combinations that we tried. This indicated to us that a 5000 burnin length and 1000 iterations for each run were adequate for convergence. This was important to us because of the large number of experiments that we were considering; a shorter running length makes the experiments faster. For example, for the Wang populations we ran 229,830 structure runs, and so choosing a burnin length of 1000 instead of 5000 reduced by almost one billion the number of simulation cycles we had to run.

For clustering analyses, we chose to analyze subsets of the data comprising from five to the maximum number of markers available, in increments of five, for mildly- to highly-differentiated populations, and from 25 to the maximum number of markers, in increments of 25, for scantily-differentiated populations. The maximum number of markers for each dataset was considered as the number of markers in the full dataset that is a multiple of five (*e.g.*, a maximum of 675 markers is found among 678 markers in the STR678 dataset).

For the Wang and SNP datasets, markers were randomly chosen, so some markers may not be shared across marker sets, but all population pairs were analyzed with the same markers for all marker number sets at which they are analyzed. To test for informativeness, markers were chosen as the top (*i.e.*, most informative) m, with m=5, 10, 15,. . . ,M, where M = maximum number of markers in the dataset that is a multiple of 5, informative markers in the World-52 dataset [66]. As such, markers found within lower marker number datasets are also found in upper marker number datasets (that is, lower marker number datasets are nested within upper marker number datasets).

3.5 Analytical Approaches

3.5.1 Genealogical versus genetic data

In addition to cluster inference and determination of the optimum placement of samples into inferred clusters, *structure* and other model-based clustering algorithms also output, when using admixture models, membership coefficients of individuals to each inferred cluster to which the individual has some affinity. Membership coefficients may be seen as the probability that an individual belongs to a particular cluster, or the proportion of their genome derived from each cluster. First, we took a majority takes all approach, where we placed samples into clusters with which they share 50% of their membership. This is referred to as the 0.50 threshold to the cluster. Then, we compared the average membership of individuals with their average kinship with all other members of the cluster. Membership coefficient is for each individual, versus the average kinship of each individual to other individuals placed within that cluster by the 0.50 threshold requirement. This was done only for K=2. Pearson correlations were performed within each cluster. Also, for K=2, 3, and 4, we grouped individuals within clusters according to increasingly stringent threshold requirements of T=50, 75, 90, and 95% membership. Here, we tested the average kinships within each cluster and F_{ST} between clusters for each of these threshold requirements.

3.5.2 Scantily-differentiated Populations

While the visual approach implemented in *structure* (the gui version) or *distuct* [28] may be adequate when considering one or just a few experiments, a quantitative approach for summarizing the data is needed for large numbers of experiments. Here, we use the clusteredness statistic (G), defined in the introduction, to quantify the extent to which individuals have their ancestry distributed solely within one cluster, even across all clusters, or some intermediate degree between clusters. Thus, we measure the change in differentiation between clusters when using different effective sizes, divergence times, marker numbers, and sample sizes. We also used clusteredness in our family group analyses.

3.5.3 Marker Numbers

In addition to using it to determine differences in differentiation between clusters, we use clusteredness [30] as a metric for determining significant differentiation between populations. Our choice in identifying sufficient numbers for identifying structure for each pair of populations was to use an initial clusteredness (G) threshold of 0.5, herein G50, followed by a more stringent threshold of 0.9, herein G90. These clusteredness thresholds chosen for indication of significant clusteredness are based on Rosenberg's standard of 75% membership coefficient for cluster assignment [64], and for 95% confident assignment as per [128]. We also determine a maximum clusteredness for population pairs, *i.e.*, the clusteredness at the maximum number of markers (G_{max}). This allows us to identify populations that do not cluster at all in our analyses, even at the maximum number of markers.

We performed clustering analyses between population pairs at 5, 10, 15,. . . ,N, where N is the maximum number of markers that is a multiple of 5 (or 25 for F_{ST} <0.01), until they achieved clusteredness values at or above 0.9 for population pairs that had a $G_{max} \geq 0.9$, and 0.5 for all other population pairs with $G_{max} \geq$ 0.5 but G_{max} <0.9. Population pairs that have maximum clusteredness <0.5 were not considered in our analyses. We took this approach because of the large number of runs that we anticipated. For example, running 10 runs each for all pairs of populations just for marker numbers from five to 100 in five marker increments would need 600,000 runs. Marker numbers needed for differentiating population pairs were determined to be the minimum number at which a population pair had clusteredness values at or above the threshold levels of 0.5 or 0.9. In this regard, we required that at least 60% of runs meet these thresholds with a range of clusteredness values of 0.1 or less for ensuring consistency among runs, similar to Tishkoff et al.'s [4] 60% standard for high stability of clustering across runs.

3.6 Generalized Hierarchical Modeling

Generalized hierarchical modeling [41] was used to fit a serial founder effects model to observed microsatellite data from the extended CEPH HGDP. For our initial analyses of human genetic variation, where we find evidence for archaic admixture in the human genome, we fit a serial founder effects model to a subsample of 100 worldwide populations. We then compared expected and realized, that is, predicted and observed, respectively, gene identity coefficients using a *ghm* plot (R code, available from Jeff Long). See Figure 7 for a hierarchical plot showing these populations. Branch lengths are proportional to the change in gene identity along the branch. Our initial tree was determined using the *neighbor joining* algorithm (NJ) [129] implemented in *phylip* [130] on genetic distances calculated from gene identities between local populations. Subsequently, new internal nodes were added and *ghm* was used to determine whether the addition of new internal nodes improved the fit of our model to the data. We used *MEGA* version 4.0 [131, 132] for visualization of phylogenetic trees. Our metric here is Nei's gene identity [133]. We constructed matrices of gene identity averages across loci, with the gene identity within local populations represented on the diagonal, and gene identity between local populations on the off diagonals. Gene identities at nodes are estimated from the gene identities between populations sharing a MRCA represented by the node. The root of our tree was determined to be the node with the lowest gene identity calculated between populations showing a MRCA represented by the node. Ghm uses numerical approximation procedures to identify maximum likelihood solutions for fitting models to observed data [134]. To determine fit, we use the likelihood ratio statistic discussed in the introductory section. Fit can also be inspected visually using a scatter plot of expected versus realized genetic distances, calculated from the gene identities, between pairs of populations. These would be expected to assume a cigar-shaped distribution [42]. However, this visual approach does not provide for a useful procedure for deciding between better fitting models. Better fitting models are expected to produce a lower Λ. One surprising observation on the tree in Figure 2 is that a model placing Oceanian populations with Cambodia, determined by the NJ algorithm, provides a better fit than when placed with other South Pacific island populations. An incorrect tree can results when deviations from treeness are present [134], which is predicted to occur amongst South Pacific island populations.

For our ALDH2 and archaic admixture simulations, *ghm* was also used to find the most likely demographic history for a subset of 16 populations from this dataset. We chose four populations from each of four geographical regions, Asia, America, Europe, and Africa. This larger dataset includes South Pacific populations [3], but they were not included in our simulations. See Figure 8 for the phylogenetic tree showing these populations, with an included archaic outgroup.

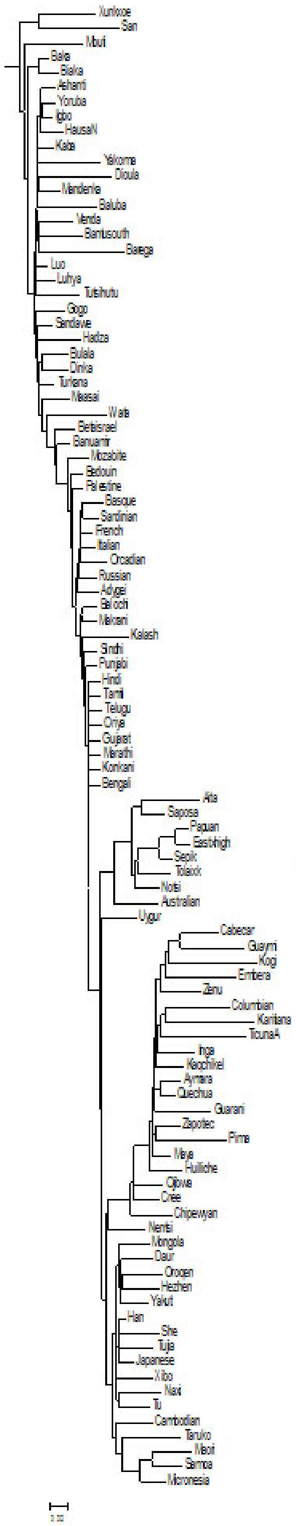


Figure 7: Hierarchical plot fit to the data from 614 STR loci for 100 populations. The x axis scale corresponds to gene identity, which is proportional to t/N_e, where t is the generation time along the branch (one generation = 25 years) and N_e is the effective size along the branch.

We initially chose populations for which we had adequate knowledge of ALDH2*2 frequency to match to populations that were typed for ALDH2. These populations were subsequently used in our admixture simulations. We included the same number of populations from each geographical region to avoid an unbalanced representation of genetic variation. Output from *ghm* provided us with estimates of gene identities within and between populations, which we use to estimate demographic parameters. As for the 100 population subsample, here we first determined the most likely branching pattern using the neighbor joining algorithm [129] as implemented in *phylip* [130] on genetic distances calculated from gene identities between local populations, and then fit this hierarchy to the data using generalized hierarchical modeling [41, 71]. Again, the root of our tree was determined to be the node with the lowest gene identity calculated between populations showing a MRCA represented by the node.

3.7 Estimation of Demographic Parameters

Taking a given evolutionary tree for a set of populations, we set out to estimate the parameters t/N_e (for 2N - 1 branches, with N = the number of populations) and mutation rate, μ, that would generate this tree in simulations. We assume that evolution is independent on each branch in the tree. From this assumption, we can further assume that gene identity between pairs of populations sharing a MRCA is the same for all pairs. In the absence of mutation, the gene identity between pairs of populations is the gene identity within the population of the MRCA; with mutation, the gene identity between populations is the gene identity within the population of the MRCA increased by mutations along branches separating the populations. Thus, we assume that $J_{MRCA,X\&Y} \leq J_{XY}$. Branch lengths are proportional to t/N_e in the absence of mutation. However, mutation has the effect of decreasing branch length. Thus, we assume that branch lengths ≤ t/Ne.

We obtained our initial estimate of the stepwise mutation rate, μ, of $2.2 * 10^{-4}$ from the mutation/drift equilibrium formula [135],

$$J_{\infty} = 1/\sqrt{1 + 8N\mu}$$

with a basal effective population sizes for modern humans, N_e, of 12500 and gene

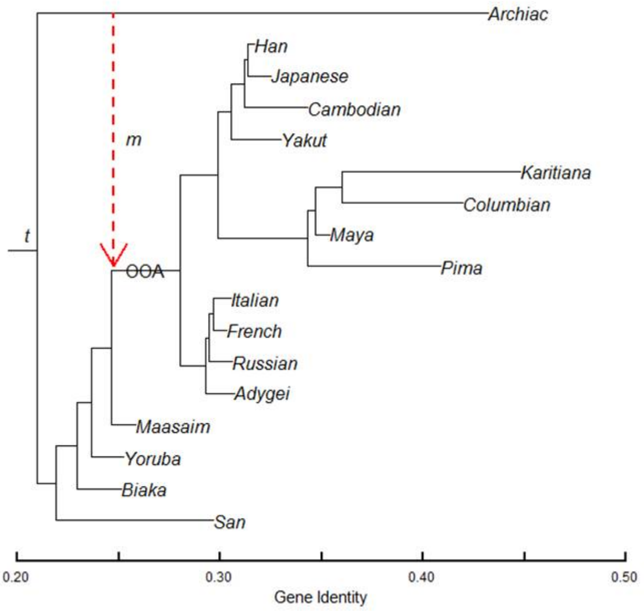


Figure 8: Hierarchical plot fit to the data from 614 STR loci for 16 populations, plus an archaic outgroup. The x axis scale corresponds to gene identities, which are proportional to t/N_e, where t is the generation time along the branch (one generation = 25 years) and N_e is the effective size along the branch. The red dotted line and arrow indicates the point of archaic contribution to the Out-of-Africa (OOA) branch, with m indicating the proportion of the modern human genome on the OOA branch contributed by the archaic lineage. t at the base indicates divergence times between archaic and modern humans.

identity at the root of 0.209—which was estimated from the data. However, in our simulations this mutation rate resulted in a gene identity coefficient at the base that was lower than expected ($J_{[E]} = 0.209$) with an effective size of 12500 individuals. Decreasing the mutation rate to $9 * 10^{-5}$ returned our expected gene identity at the basal node as estimated from the data.

Initial effective population sizes of internal and terminal branches were estimated using gene identity coefficients obtained from *ghm* along with rough divergence times, using the equation,

$$2N_e = -t/\log(\frac{1-J_t}{1-J_0})$$

where t is the number of generations between the beginning and end of branches, where one generation was considered to be 25 years, J_0 is the gene identity at the beginning of the branch, and J_t is the gene identity at the end of the branch. $2N_e$ is the haploid effective size. Effective size estimates ranged from a few hundred in the Brazilian populations to over 100,000 on some internal branches.

As in the case of the basal node, simulated gene identity coefficients on internal nodes and tips were also lower than those estimated from the data. In addition, differences between observed and simulated gene identity coefficients were generally greater in nodes further from the root. This resulted in shorter than expected branches. This was likely the result of cumulative deviations from lower branches. Since branch length is influenced by the parameter t/N_e, we adjusted gene identity coefficients on nodes at the ends of branches in the equation above to estimate new effective population sizes to calibrate simulations to our observed tree (that is, the tree fitted to our data). Inserting these new, adjusted effective sizes into our simulations returned gene identity coefficients and branch lengths more closely approximating what we observe in the tree fitted to the data. Keeping divergence time as set, we increased gene identity coefficients on nodes at the ends of branches by 0.01, thereby decreasing our estimated effective size along the branch, until simulated and observed gene identity coefficients differed by 0.005 or lower. This was done successively beginning with the basal node and continued until all nodes returned their gene identities as estimated from the data. Changes on lower level nodes also affected gene identities on upper level ones. As effective sizes on lower nodes were adjusted to return proper gene identities, higher nodes returned gene identities closer to those estimated from the data. Internal nodes leading to all regions needed adjustment, but nodes within regions did not. In addition, nodes in extant OOA populations required no changes, while a few in extant African populations did. Adjusting gene identities on nodes to those estimated from the data returned estimated branch lengths as well. We feel our estimates of effective size calibrated by crude archaeological estimates are robust to the true history of these populations. Any change in divergence times will result in a concomitant change in effective size to maintain a level of drift to produce the observed gene identity at each node. In Table 4 we present adjusted effective sizes and divergence times for terminal populations and internal nodes.

Table 4: Demographic parameters for 16 population simulation. Branch and node refers to the numbers of the populations (*i.e.*, terminal end of the branch) and connecting nodes (*i.e.* internal node of the branch), respectively, on the phylogenetic tree, and PopulationName refers to either the name of the population, in the case of extant populations (first 16 are extant populations) or populations joined by the connecting node. $AdjN_e$ are effective sizes calculated from rough divergence times (in generations), t_{Div}, adjusted to return gene identities estimated for the node.

Branch	Node	AdjNe	tDiv	PopulationName
1	17	31812	7800	San
2	18	89377	4000	Biaka
3	19	111827	3000	Yoruba
4	20	111827	2400	Maasai
5	22	91869	1960	Adygei
6	23	23610	400	Russian
7	24	20524	200	French
8	24	15341	200	Italian
9	26	7355	800	Pima
10	27	36233	400	Maya
11	28	398	40	Colombian
12	28	262	40	Karitiana
13	29	48467	1800	Yakut
14	30	12494	600	Cambodian
15	31	17768	300	Japanese
16	31	61592	300	Han
Internal nodes				
17	root	25000	7800	root
18	17	99340	4000	Biaka
19	18	41431	3000	Yoruba
20	19	22126	2400	Massia
21	20	2619	2200	OOA
22	21	7434	1960	Europe
23	22	635398	400	Russian
24	23	59008	200	French/Italian
25	21	8029	1880	Asia/America
26	25	11096	800	America
27	26	61389	400	Maya
28	27	16978	40	Columbian/Karitiana
29	25	3271	1800	Asia
30	29	126376	600	Cambodian
31	30	111700	300	Japanese/Han

3.8 Archaic Admixture Simulation and Analysis

We simulated 1000 completely unlinked microsatellites, using a single step mutation model [136] (SMM), that is, insertion or deletion of one repeat unit per mutational event, and limited the number of alleles to 35 for each marker. After setting up and calibrating our tree with extant human populations, we added an archaic admixture event on the OOA branch (see Figure 7). Effective sizes of archaic and ancestral (that is, the MRCA of archaic and modern humans) humans were also set to 12,500 individuals. From our simulations, we sampled 20 haploid genomes for all populations except archaics. At each historical event, 100% of the source population was merged into the sink population and the effective size of the sink population was converted to the effective size of the ancestral population on the branch leading to the shared node. We chose an instantaneous growth model for our populations for the change in effective population size at each historical event. That is, all changes in population sizes occurred instantaneously at the time of the historical event. In general, no gene flow is allowed between populations.

For our simulations with archaic admixture, we varied divergence times of archaic and modern humans from 300,000 to 700,000, in 25,000 year intervals, and archaic contribution to modern humans from 0 to 50% in 2.5% intervals. Archaic contributions occurred in one shot at the beginning of the OOA branch (2400 generations). Since *SimCoal2* is a backward in-time simulator, the movement of migrants is actually from the OOA branch to the archaic branch, and the contribution refers to the proportion of haploid genomes moved from the OOA branch to the archaic branch.

For analyses, we calculated the differences between expected and realized gene identities (that is $J_d = J_e - J_r$). Our observed difference calculated from the data is 0.02, which is the value to which we compare our results. Gene identity differences were regressed on archaic contribution within divergence times using quadratic regression and plotted on a contour map with archaic contribution and divergence time on the X and Y axes, respectively. R [137] was used for production of contour maps.

3.9 ALDH2 Simulation and Analysis

We simulated 20,000 chromosomes with features similar to ALDH2, DNA sequence of length 5387 using an infinite sites model with a mutation rate of $1.2 * 10^{-9}$, following a demographic scenario of 16 populations, four each from four geographic regions. We used the demographic history, that is, effective sizes and divergence times, as determined from the microsatellite data in our simulations. We chose 16 populations from the CEPH HGDP that are the same or closely match those from which we obtained ALDH2 sequences: Mbuti, Biaka, Kenya, and Yoruba from Africa; Adygei, Russian, French, and Italian from Europe; Pima, Maya, Karitiana, and Surui from America; and Yakut, Cambodian, Japanese, and Han from Asia. Samples sizes of simulated populations match those of sequenced data. Following simulations, we sampled alleles that had frequencies matching the ALDH2*2 allele, 6.9% worldwide and 22% in Asia. We also sampled alleles having frequencies matching the OOA restricted allele, 33% worldwide and 42% on the OOA branch.

3.10 AIDA

As with the original version of AIDA, we calculate *II* and *cc*, analogous to Moran's *I* and Geary's *c*, respectively. However, the current version also provides the option of choosing either *II* or *cc* to be calculated, which is an advantage for large data sets as it reduces calculation time by approximately 50%. Indices are calculated per distance class, with comparisons between individuals located within a pair of defined distances distance falling within the distance class.

Statistics

The revised AIDA equations are:

$$II = \frac{n \sum_{k=1}^{S} Wij \sum_{i=1}^{n-1} \sum_{j>1}^{n} (p_{ik} - \bar{p_k}(p_{jk} - \bar{p_k}}{W \sum_{i=1}^{n} \sum_{k=1}^{S} (p_{ik} - \bar{p_k}^2}$$

and

$$cc = \frac{(n-1)\sum_{k=1}^{S} Wij \sum_{i=1}^{n-1} \sum_{j>1}^{n}(p_{ik} - \bar{p_k}^2}{2W \sum_{i=1}^{n} \sum_{k=1}^{S}(p_{ik} - \bar{p_k}^2}$$

where n is the sample size, W is the number of pairwise comparisons in the distance class of interest (essentially the sum of the matrix of weights), p_{ik} and p_{jk} are alleles at the k^{th} site for the i^{th} and j^{th} individuals, respectively, and p_k is the average p value across all individuals for the k^{th} site (or the k^{th} value in the average vector). w_{ik} are weights: 1 if a pairwise comparison is found in the distance class, 0 if otherwise. *Wij* is the matrix of weights. All weight values are placed in a matrix with values on the i^{th} row and j^{th} column correspond to comparison between the i^{th} and j^{th} individuals. Comparisons on the diagonal of the matrix of weights (i.e., where $i = j$ are always 0 and are never considered. Products, for *II*, and squared differences, for *cc*, respectively, for all pairwise comparisons for each k are calculated and weighted according to *wi*j, the weighting value for the comparison between individuals i and j, then summed over n individuals in the sample within k, and finally summed across S sites. Our calculations of *II* and *cc* here differ from the old version of AIDA. We use NumPy [138] matrices for calculations; as such, differences are calculated across individuals first for all segregating sites prior to across segregating sites, as in the original version [91]. This change does not affect calculation of indices.

Significance

For significance testing, AIDA calculates empirical confidence limits based on pseudo-indices generated by permutation of the dataset. Sequences are randomly distributed among geographical sites, which each geographical site maintaining its size, N1 number of times, pseudo-indices obtained on the randomized data, and confidence limits defined for 95, 99, and 99.5 percent limits. Confidence limits are the upper and lower values that define the confidence boundaries. Users may choose for no confidence intervals, per distance-class, or one set for all distance classes. The latter is recommended for large datasets as computation time is reduced, though the confidence intervals are not as robust, while the first may be recommended for exploratory purposes.

In the case of per distance class confidence intervals, sample sizes per distance class are retained per distance class. However, for a single set of confidence intervals, sample size is set at the number (N2) of the distance class showing the lowest number of comparisons. A set of N2 sequences are randomly chosen from the initial set N1 times.

Implementation

Our updated Python-based version of AIDA makes extensive use of the NumPy [138] package for the Python programming language. Linux and Windows versions are available, with the Windows version available in command-line or GUI versions. Python is a high-level dynamic programming language, and unlike Pascal, the original language of AIDA, python does not require pre-compilation declaration of vectors and therefore is not constrained by sample size or number of sites.

Data input may be Excel or *Arlequin* file format, and may be diploid, haploid, or population frequency data. Also, DNA sequence, SNP, or RFLP data is accommodated, along with the addition of microsatellite capability. Diploid data is recommended only in the case that the possibility of recombination has been eliminated as it can obscure the signal of autocorrelation. AIDA tests for sequence similarity between sequences at different locations, therefore recombination may obscure the assessment of similarity [91]. For microsatellite analyses, users have the option of choosing Infinite Alleles (IAM) or Stepwise Mutation (SMM) Models. In addition, the original AIDA program only allowed for biallelic data, so SNP markers with more than three states would have had to be excluded from analyses. In the new AIDA version, these markers can be accommodated by using the Infinite Alleles (IAM) microsatellite option. Three options are available for distance-classes: equal intervals (classes are equal width), equal frequencies (approximately equal number of comparisons in each distance class), and user-defined. AIDA takes as input (*i.e.*, polar) or geographical (*i.e.*, latitude, longitude), from which distances between populations are determined. In the case of Cartesian coordinates, distances are Euclidean, whereas great circle distances are used for geographical coordinates. Users also have the option of inputting a user-defined distance matrix. Unfortunately, the usage of great circle or Euclidean distances does not take into account geographical barriers, so this option provides for this possibility through user-inclusion of waypoints in distance calculations. In addition, users may want to use non-geographical distances (*e.g.*, linguistic).

3.11 Correlation and Other Statistical Analyses

Correlation analyses were done using the R computing language, version 2.10.1 [137]. Fst values were calculated using the *Arlequin* software package, version 3.1 [124] and the *arlsumstat* module of *Arlequin* version 3.5 [139]. Other calculations were performed using in-house scripts written in the *python* programming language [140]. All figures except *distruct* output were produced in R, version 2.10.1 [137].

4 RESULTS

4.1 Comparison of Genealogical and Genetic Data

4.1.1 Genetic clustering

When analyzed in the absence of the CEPH HGDP European populations, Gioi and Cardile appear as two clusters roughly corresponding to the two villages, albeit with a couple of outliers (see Figure 9). We define outliers to be individuals that are sampled from one population but which cluster with another population. In addition, we also observe samples that show some mixture between populations. Since we sampled nearly all members of the villages, we were able to compare villages of sampling between offspring and parents, which show, for some outlier individuals, individuals that cluster with the population from which their parents were sampled rather than the population from which they themselves were sampled. Results from analyzing the distribution of logarithmic probability of the data were inconclusive, showing no obvious peaks between consecutive values of K (Figure 10a). As such, we did not compute the posterior probability of the data to determine the most statistically likely number of clusters as initially suggested by Pritchard [59] and instead used Evanno's rate of change method [38]. This showed the most likely number of clusters in our data to be K=2 (see Figure 10b). Further, most individuals are clearly assigned to one of the two clusters, with 78% showing membership coefficients 0.75, 55%, 0.9, and 37%, 0.95, fulfilling Prichard's more recent suggestion of looking at the distribution of individuals into clusters for a guide in determining the number of clusters present in the data.

Analyzing our Cilento populations with the CEPH HGDP European populations, at 36 markers shared between datasets, erases the observed clustering from our analyses (Figure 11). This was expected given the limited genetic differentiation between our populations from Cilento and other populations within Europe. Also, the number of markers shared between our dataset and the CEPH HGDP are likely to be too low to enable proper differentiation amongst European populations.

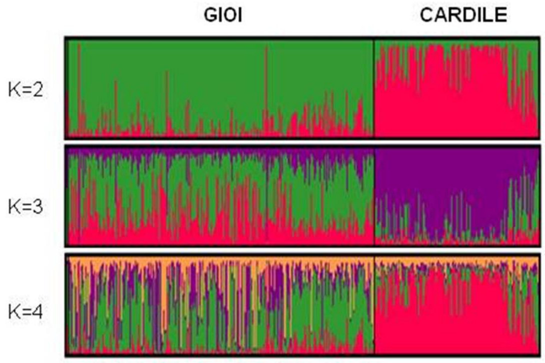


Figure 9: *Distruct* graph showing distribution of samples into clusters according to membership coefficients for K=2, 3, and 4 [120].

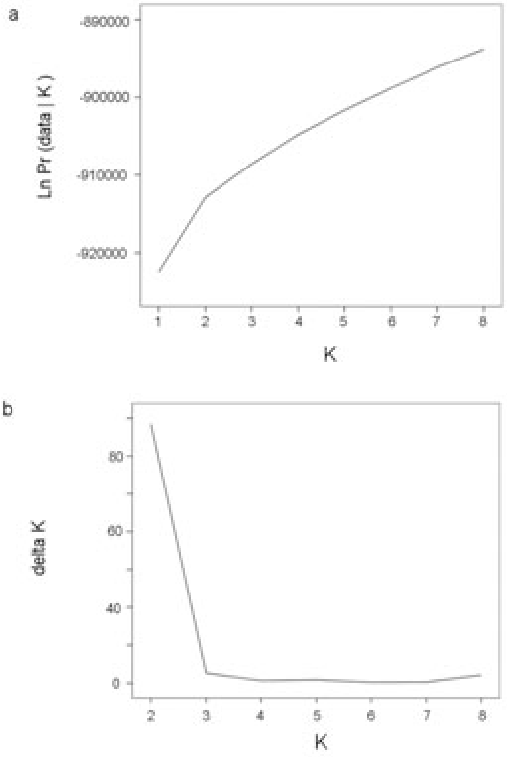


Figure 10: Graphs showing (a) estimated logarithm of the probability of K clusters given the data and (b) the modal value of the second order rate of change of the likelihood function given the number of clusters [120].

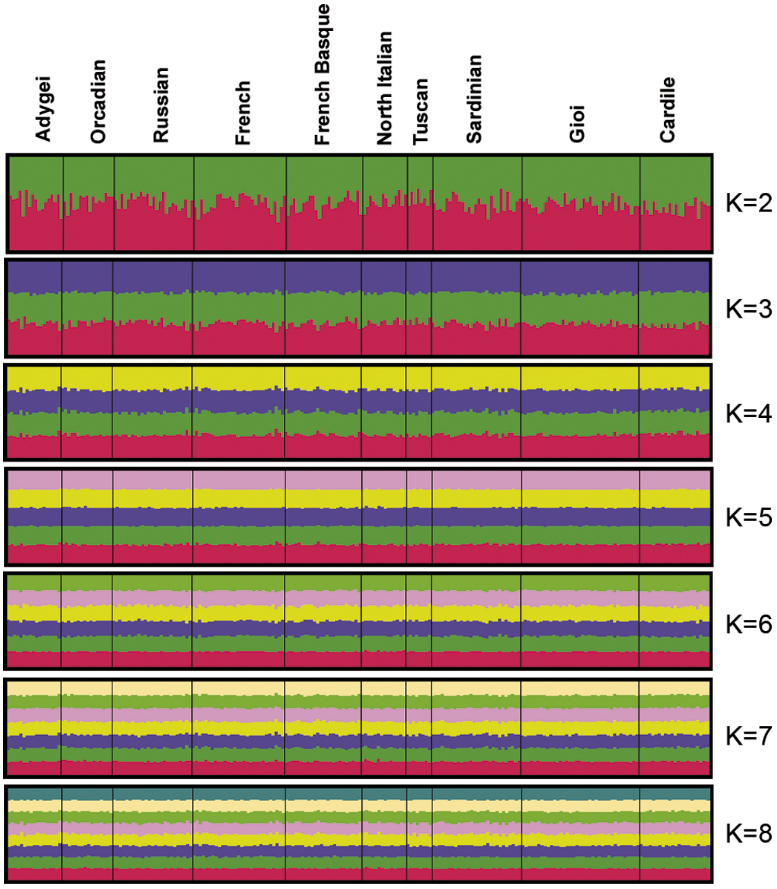


Figure 11: Distruct graph showing distribution of samples into clusters according to membership coefficients for K=2 to 8, including European populations from the CEPH HGDP, inferred from 36 loci [120].

4.1.2 Validation of genetic clustering using genealogical data

Beginning from our extant sample of individuals, we backwards reconstructed pedigrees using genealogical data. As can be seen in Table 1, nearly all extant individuals, in both populations separately and in the combined population, are found within a single pedigree. This confirms that the two villages have a shared history, and shows the complexity of relatedness within and between the populations. We quantified within and between village relatedness using pairwise kinship coefficients calculated from pedigree data. In Table 5, we report summary statistics of these values along with similar data from other isolated populations obtained from the literature. The average values of kinship between all individuals in the populations of Gioi and Cardile are equal to that between third cousins (Φ_{ij} = 0.004) and approaching that between second cousins, respectively (Φ_{ij} = 0.015), confirming a high degree of consanguinity in both villages.

Table 5: Kinship (ϕ_{ij}) summary statistics of Gio and Cardile compared with those of those of other isolated populations from the literature [120].

	Sample Size	Average ± SD	Median	25-75 percentiles
Gioi-Cardile	1356	0.003 ± 0.0015	0.001	0.000-0.002
Gioi	882	0.004 ± 0.0018	0.001	0.000-0.002
Cardile	474	0.009 ± 0.0024	0.004	0.001-0.008
Perdasdefogu[a]	821	NA	0.007	0.004-0.011
Talana[a]	875	NA	0.0014	0.009-0.0021
S-leut Hutterites[b]	806	0.0042 ± 0.0031	NA	NA
Iceland[c]				
1925-1949 cohort	37762	0.008 ± NA	NA	0.000-0.001
1950-1965 cohort	38336	0.005 ± NA	NA	0.000-0.001

[c] Estimates based on pairs of married couples

[a] Falchi *et al.* 2004

[b] Abney *et al.* 2002

[c] Helgason *et al.* 2008

4.1.3 Cluster Membership and Relatedness

Using membership coefficients estimated from *structure*, we placed individuals into clusters according to various threshold requirements. First, we took a majority takes all approach and placed samples into the cluster to which they had more than 50% membership. Here, we estimated the average kinship of each individual to all other members of the cluster, Φ_{Ci}, where C corresponds to the cluster (C = 1 for the green cluster, C = 2 for the red) and i represents individuals. We thus compared these values for each individual with their membership coefficient to the cluster, using Pearson correlation analysis and graphic visualization (Figure 12). We found significant correlations for both the green and red clusters, with correlation values r = 0.74 (p $<10^{-10}$, N = 934) and r = 0.082 (p $<10^{-10}$, N = 423), respectively.

For further evaluation of the relationship between relatedness within a cluster and membership to the cluster, we restricted cluster membership with increasingly stringent membership threshold requirements, of $T \geq$ 0.50, 0.75, 0.90, 0.95, and 0.99 (Figure 13), $T \geq 0.99$ not shown) for K=2. Here, we estimated the average kinship within each cluster amongst all individuals. As can be seen in the figure, restricting access to the clusters to individuals that meet the increasingly stringent membership requirement increases the average kinship within the cluster. These

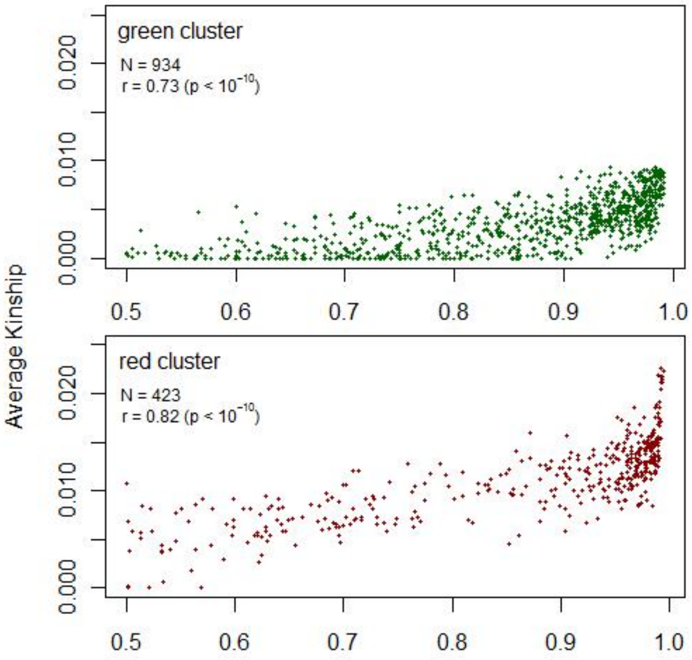


Figure 12: Average kinship of sample individuals with all other members of the cluster versus membership coefficient to the cluster for the red and green clusters, with Pearson correlation coefficients.

are nested clusters, as individuals within higher threshold clusters are also found within lower ones.

4.1.4 F_{st} and kinship

F_{ST} calculated between the two sampled villages using 239 unlinked markers is low ($F_{ST} = 0.008$). We used the clusters determined above to see whether F_{ST} calculations are affected by restricting analyses to subsets of individuals with increasing relatedness within clusters (since we know that increased threshold restrictions for access to clusters results in increased relatedness within them). Pairwise F_{ST} values between clusters for $\geq$ 0.50, 0.75, and 0.90 are shown in Figure 14. Similarly to what we saw in the case of kinship above, restricting cluster membership to increasingly stringent thresholds of membership results in increasing F_{ST}, and thus differentiation, between clusters.

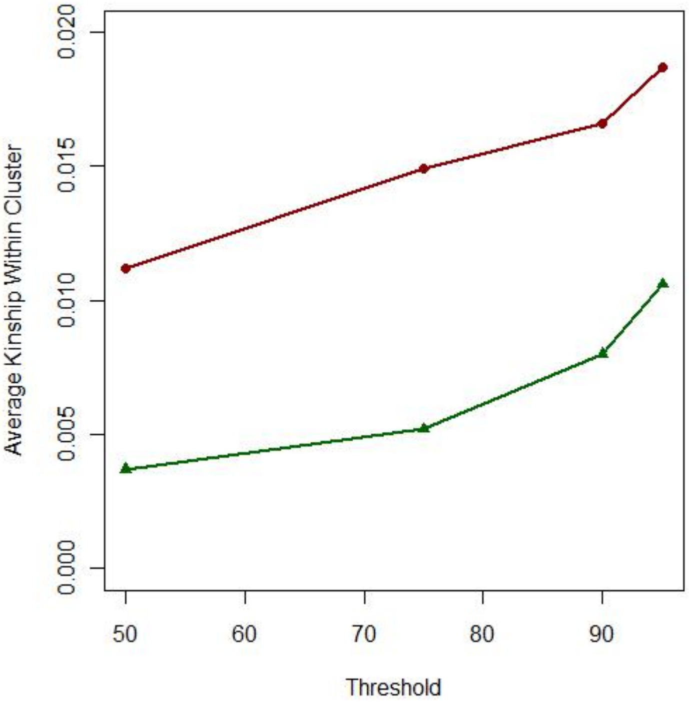


Figure 13: kinship within each cluster at given membership threshold levels for K=2.

4.2 Scantily-differentiated populations

4.2.1 Effect of Demographic Parameters on Clustering Methods

In our simulations, we modelled a scenario with an isolated population instantaneously splitting from a source population at a given time point, t_{div}. We assume no gene flow between populations or from external populations, and constant population size. Figure 15 displays mean clusteredness values and 95% confidence intervals, across 50 replicates, according to isolate population size and divergence time, with varying numbers of markers and sample sizes.

We see a few trends in the data in Figure 15. First, with decreasing effective size and increasing divergence times, our ability to measure differentiation increases. We also see an interaction effect, such that decreasing effective size and increasing divergence times together increases our ability to detect structure more so than either one separately. We predict that this is the combined effect of both of these parameters increasing the effects of genetic drift, which increases the divergence between the two populations. Increasing divergence predicts increasing ability to

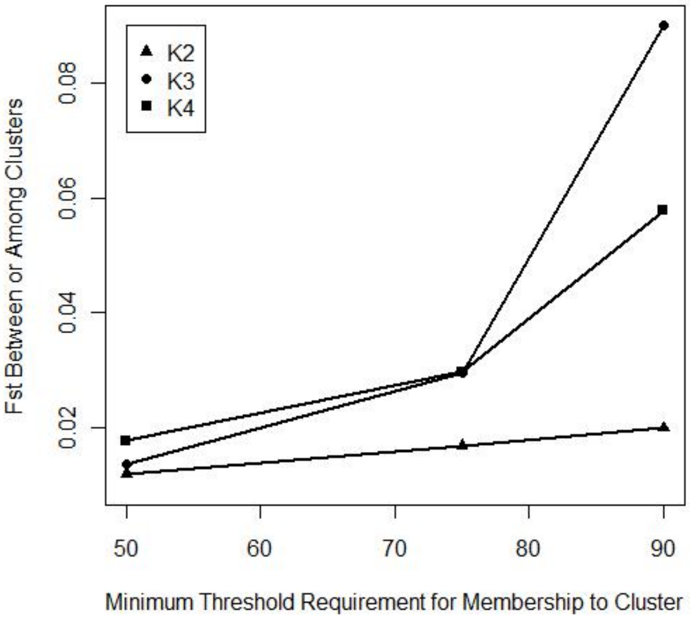


Figure 14: Pairwise F_{ST} between clusters at given membership threshold levels, for T 0.50, 0.75, and 0.90, for K=2, 3, and 4.

observe differentiation/clustering. We also see that increasing sample sizes and number of markers considered increases our ability to detect structure as well. These two aspects of study design also appear to interact, such that increasing sample size and marker numbers tends to increase our ability to observe structure more so than either one separately. From here, one could see a compensatory effect of marker numbers on sample size, and vice versa. Increasing marker numbers reduces the size of the sample required to observe structure, and vice versa. We also see that at different levels of differentiation, gauged by differences in effective size and divergence times, different sample sizes and marker numbers are needed to observe differentiation. Essentially, as populations become more divergent, lower numbers of markers and sample sizes are needed to observe differentiation.

4.2.2 Effect of Consanguinity on Clustering Methods

Figure 16 shows that when relatedness is restricted to individuals related up to the level of second cousin, all detection of structure is erased from our analyses, even

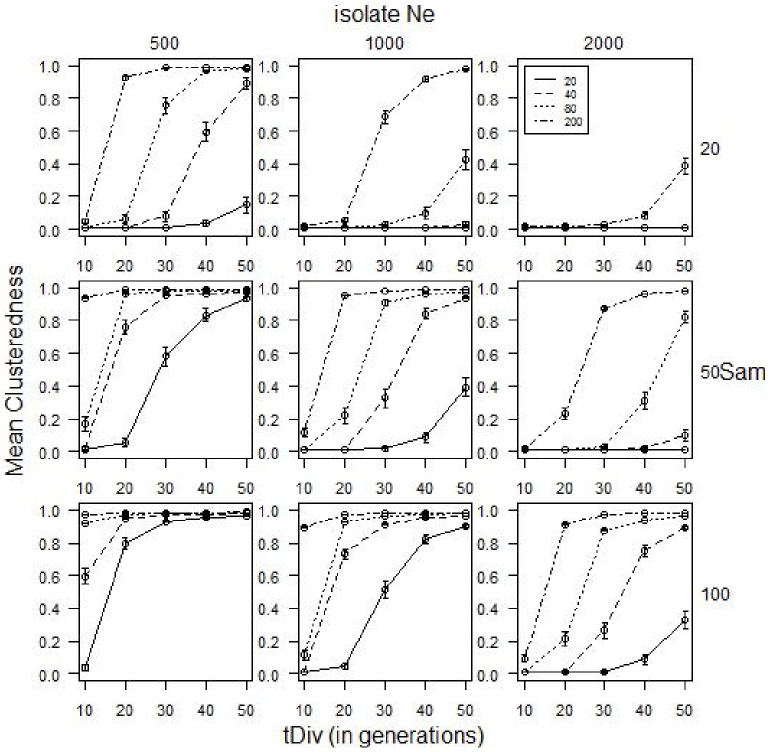


Figure 15: Average clusteredness values and 95% confidence intervals (C.I.) across 50 replicates for all parameter sets.

when using large sample sizes (n=100) and numbers of markers (m=200). Further, in the case of n=200 and m=80, we see a decreasing trend of clusteredness between samples where restriction is up to half-siblings, and where samples are restricted up to second cousin. We also see a decline in observations of differentiation as sample sizes and marker numbers decrease.

The analyses considering groups of individuals at different levels of relatedness (Figure 16) show that our observations of clustering using these methods can largely be regarded as an effect of consanguinity. When adequate numbers of markers and sample size are considered, genetic structure is detectable either when individuals are chosen at random (kin-group = 0) or when samples contain highly-related individuals (kin groups 1 and 2). Conversely, no structure is detected in kin groups 4 through 7, namely in samples from which individuals are unrelated at the level of 2^{nd} cousin or higher; removal of consanguineous individuals leads to the disappearance of population structure. The structure observed between the two villages seems, for the most part, to depend on the presence of

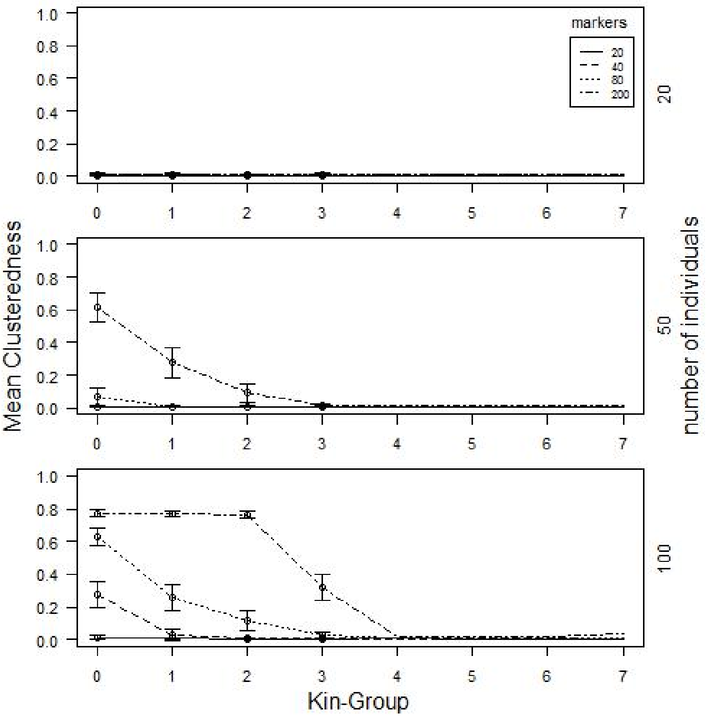


Figure 16: Average clusteredness values and 95% confidence intervals across 50 replicates. Consanguinity decreasing from kingroup 0 to 7.

recent genealogical relationships, at least to the level of first cousins. Sample sizes are also important in the observation of structure, with no structure whatsoever emerging from analyses of 20 individuals, and little structuring with samples of 50 individuals when using the largest number of markers. This confirms previous results [120] suggesting that clustering is observed when a random sample (kin-group = 0 in Figure 16 panel C), adequate numbers of markers (239 and 200 in previous and current study respectively), and appropriate numbers of sample individuals are considered. This also explains our results with the CEPH HGDP European populations analyzed with Cilento.

4.3 Marker Numbers

We first performed clustering analyses using the full marker sets. In the STR678 dataset, nearly all population pairs show some significant degree of differentiation at the maximum number of markers, with 98% having clusteredness of at least 0.5, 83%, 0.9, and 65%, 0.95, corresponding to average membership coefficients of

0.75, 0.95, and 0.975, respectively. In the SNP dataset, 1338 population pairs were analyzed for the G50 analysis, with 445 of them analyzed for the G90 analysis, meaning that 97.1% of population pairs had Gmax values $\geq$ 0.5 and 33.3% had G_{max} values greater than or equal to 0.9. However, many of these latter population pairs did not meet the G $\geq$ 0.5 in subsequent analyses with lower numbers of markers. In fact, in subsequent analyses only 31% of population pairs reached a G $\geq$ 0.9 using SNP markers.

STR678. For mid to high F_{ST} values (Figure 17), and using a threshold of $G \geq 0.50$ for detection of structure, there does not appear to be a distinct lower limit to the number of markers needed to detect population structure given specific levels of differentiation. Given F_{ST} greater than, or equal to, 0.01, one does not necessarily need increased numbers of markers to detect structure at lower levels of differentiation; however, in some cases with $F_{ST} \geq 0.01$, up to a hundred markers (or more) may be needed. What we actually see is a range of marker numbers that may be needed to observe structure given specific levels of differentiation in a range of populations, so that genetic structure between different population pairs at given levels of differentiation may be recognized with different numbers of markers. However, although we do not observe a general trend of decreasing lower limit of markers with increasing F_{ST}, we do observe a decreasing trend in the upper limit to the number of markers required. As populations become more differentiated, the maximum number of markers needed for the detection of structure decreases, as would be expected, for all differentiation levels, with highly-differentiated populations needing 20 or fewer markers for the detection of structure. At F_{ST} values below 0.01 (see insert, Figure 17), marker numbers in the hundreds are generally required. Correlations between markers number and F_{ST} for mid to high F_{ST} and low F_{ST} are -0.64 and -0.50, respectively.

On the other hand, when an increased threshold for determination of adequate structure is required, that is, $G \geq 0.9$, we see a similar pattern as before for the upper limit of markers needed for detecting structure (Figure 18); however, a more distinct pattern emerges for the lower limit of markers needed to detect structure. We also see some odd behavior for datasets analyzed with 20 markers, in that 20 markers is still adequate for detecting structure at lower F_{ST} levels that may require 40 or more markers for detection (*indicated by the observed pattern of points from scantily to highly differentiated population pairs*). Here,

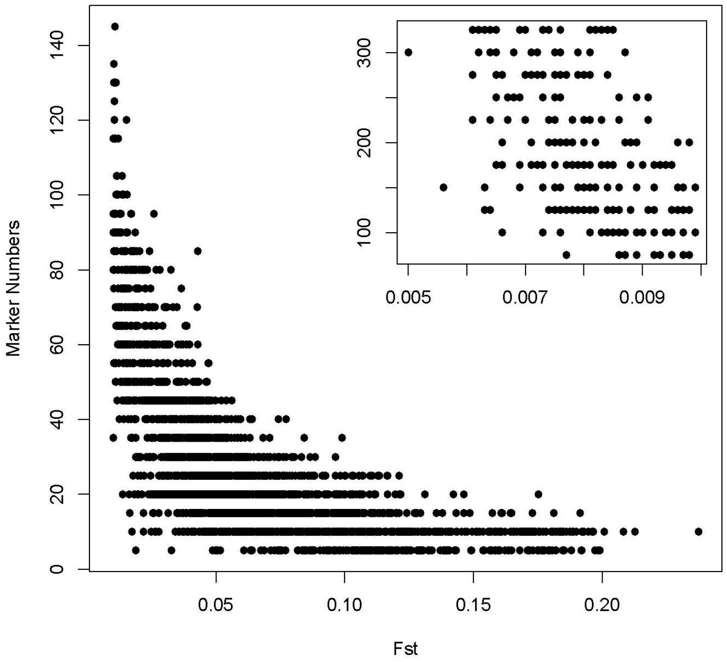


Figure 17: Number of randomly chosen microsatellite markers needed to detect structure in pairs of populations with a given level of differentiation (F_{ST}) and a clusteredness threshold of 0.5. Each data point is a pair of populations, each represented only once. On the X axis is their F_{ST} measured at the maximum number of markers, and at the Y axis is the lowest number of markers at which structure is observed determined by the given threshold. The embedded graph contains results for low F_{ST} population pairs.

we may expect that by chance we chose markers that were more informative in the population pairs that still showed differentiation at lower F_{ST} levels with 20 markers, even when using a higher standard of differentiation. As Rosenberg [66] showed, while some markers are informative in populations from all geographic backgrounds (save Oceania and America), some markers may only be informative, or more informative, in populations from specific, limited geographic backgrounds. Correlations between markers number and F_{ST} for mid to high F_{ST} and low F_{ST} are -0.65 and -0.52, respectively, which we see is a slight increase over the G50 analysis.

At this stage, it seems that that two variables may be contributing to differences in marker numbers needed for observing structure at certain degrees of differentiation (*basically, to explain the variation*): different sample sizes and informativeness of markers. As mentioned above, Rosenberg et al. [66] showed that

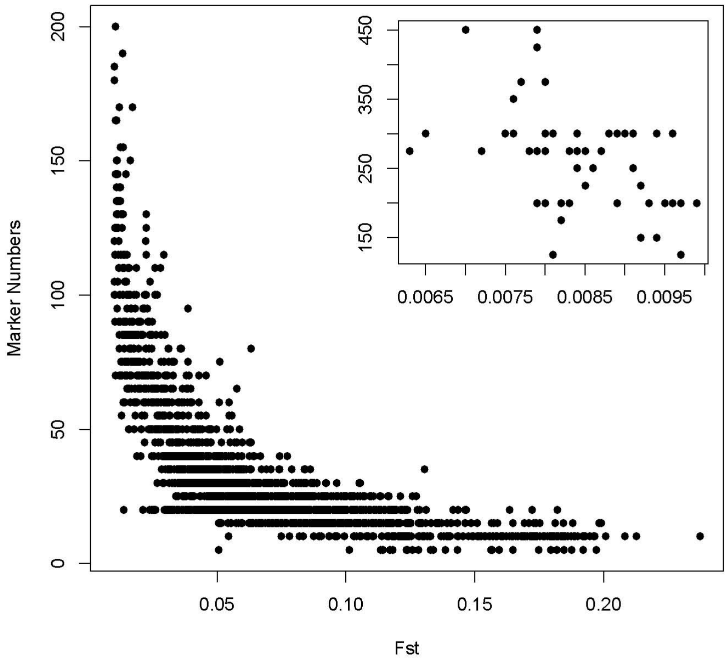


Figure 18: Number of randomly chosen microsatellite markers needed to detect structure in pairs of populations with a given level of differentiation (F_{ST}) and a clusteredness threshold of 0.9. Each data point is a pair of populations, each represented only once. On the X axis is their F_{ST} measured at the maximum number of markers, and at the Y axis is the lowest number of markers at which structure is observed determined by the given threshold. The embedded graph contains results for low F_{ST} population pairs.

though some markers are informative in all populations, there are some markers that are more informative in certain geographical groups, and by chance we may have randomly chosen markers that are informative in specific populations. For example, when pairs from those populations, or where one population in the pair, are from a group of populations in which the markers are informative, we may expect to see lower numbers of markers needed to differentiate them than, say, where the markers are not informative in either population. Also, we may expect to see decreasing numbers of markers needed with increasing numbers of populations in which markers are informative (*i.e.*, 0>1>2). This is not as big a concern with higher numbers of markers, as you will be sampling markers with a range of informativeness, including more highly informative ones. Further, Rosenberg et al. [66] showed that, while some markers were informative across most populations, this did not hold for American and Oceanian populations. Could it be possible that populations in pairs tending to need more markers at specific differentiation levels come from these regions? A glance at a table of population pairs and marker numbers needed to observe structure shows that this is not necessarily the case. A number of population pairs that show no clustering with up to 375 markers or higher numbers of markers are found to contain neither American or Oceanian populations, and a number of population pairs demonstrating structure given low numbers of markers are found to contain either Oceanian or American populations. We may conclude, then, that if informativeness of markers between population pairs is one culprit here, it does so regardless of the geographical region from which the population pairs were sampled. That is, the markers may be informative between certain populations, but randomly informative.

In addition, the horizontal variation in the STR678 G50 graph (Figure 17) is likely to be due, at least partly, to differences in the extent of clusteredness. For a given number of markers, those from higher F_{ST} pairs are likely to show a clusteredness of 0.9 or 0.95, while those at the lower F_{ST} range may show clusteredness of 0.5 or 0.6. If analysis is restricted to marker numbers where population pairs show clusteredness of 0.9, we do start seeing more of a trend of lower limit to numbers of markers needed (as observed in Figure 18). Thus, we observe less variability in number of markers needed to observe structure at specific levels of differentiation. Even here, though, observations at 20 markers still pretty much show a range from low F_{ST} to high, adding to our prediction of informativeness of these markers.

In order to test whether differences in sample sizes may be contributing to the variation in marker numbers needed to observe differentiation, we performed Pearson correlation analyses on sample sizes and marker numbers (Table 6). Correlations between sample sizes and marker numbers needed to observe differentiation for the G50 analysis for mid to high F_{ST} and low F_{ST} are -0.22 and -0.53, respectively, and for G90 analysis for mid to high F_{ST} and low F_{ST} are -0.13 and -0.52, respectively. Interestingly, sample sizes appear to be more important in determining marker numbers needed to observe differentiation for low F_{ST} population pairs when using the G50 threshold. We see a significant correlation between marker numbers and sample size, indicating that some of the variation in marker numbers needed to identify clustering at given F_{ST} levels may be due to differences in sample sizes.

Table 6: Correlations between F_{ST} (F_{ST} vs m) and sample size (n vs m) and marker numbers needed to observe differentiation at the specified threshold level (T), and partial correlation between F_{ST} and marker numbers controlling for sample size ($pcor$), for N number of population pairs.

	T	Fst	Fst vs m	n vs m	pcor	N
Wang	G50	≥ 0.01	-0.64	-0.21	-0.66	2732
CEPH377	G50	≥ 0.01	-0.52	-0.19	-0.53	1064
SNP	G50	≥ 0.01	-0.55	-0.22	-0.58	1337
Wang	G50	< 0.01	-0.51	-0.53	-0.44	218
CEPH377	G50	< 0.01	-0.31	-0.49	-0.25	194
SNP	G50	< 0.01	NA	NA	NA	NA
Wang	G90	≥ 0.01	-0.65	-0.13	-0.65	2263
CEPH377	G90	≥ 0.01	-0.47	-0.10	-0.46	972
SNP	G90	≥ 0.01	-0.62	-0.22	-0.62	444
Wang	G90	< 0.01	-0.52	-0.20	-0.53	52
CEPH377	G90	< 0.01	-0.51	-0.12	-0.51	70
SNP	G90	< 0.01	NA	NA	NA	NA

SNPS. SNP markers show a similar pattern to microsatellites (Figure 19), that is, no distinct lower limit to the number of markers needed to detect structure at different levels of differentiation, and a decreasing upper limit as differentiation increases, for the G50 analysis. G90, however, loses a lot of data points between in the middle range of F_{ST}. That is, data points exist for F_{ST} around and greater than 0.1 and for F_{ST} around and lower than 0.025, but none in the middle. Also, all data points for SNP markers, both in the G50 and G90 analyses have F_{ST} $\geq$ 0.01. Some of the variation in marker numbers that we explained in the case of microsatellites holds for SNP markers as well. We see a correlation between sample sizes and marker numbers needed to identify structuring.

We may attribute the large drop in samples from the G50 to G90 SNP datasets to the fact that a threshold of $G \geq 0.9$ is a much harder standard to meet. On the average, individuals in a pair of populations must have 95% membership to that cluster. We may also be concerned about the low numbers of burnins and iterations used in our analyses, as SNPs tend to require longer MCMC chains. We may be concerned that a failure of convergence due to lower numbers of burnins and iterations may result in a lowered estimation of clusteredness. We reanalyzed all pairs of populations with increased lengths of burnins and iterations, 20,000 and 10,000, respectively, using the full marker set to test whether a difference in chain length would affect clusteredness. 6.9% of population pairs show a 0.1

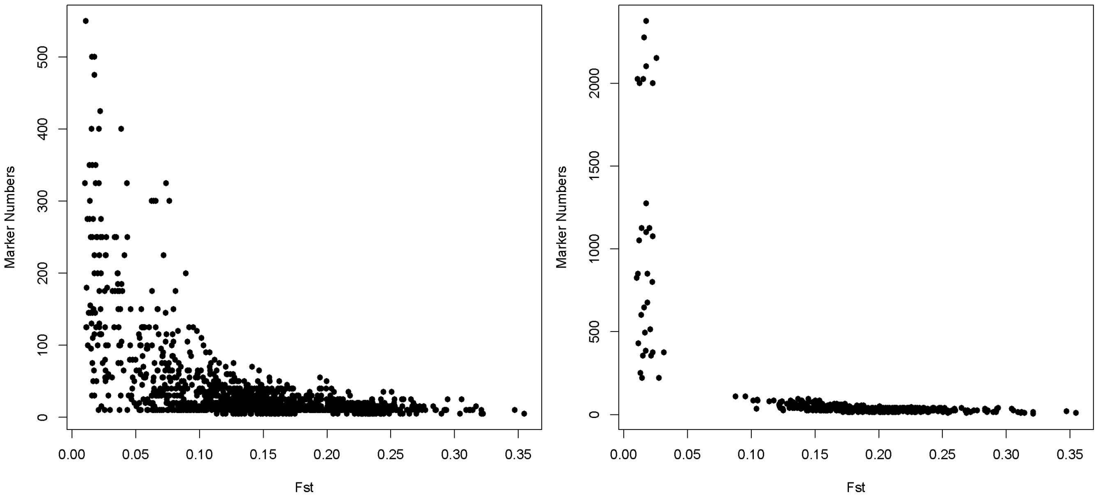


Figure 19: Number of SNP markers needed to detect structure in pairs of populations with a given level of differentiation (F_{ST}) and a clusteredness threshold of 0.5 (left) and 0.9 (right). Each data point is a pair of populations, each represented only once. On the X axis is their F_{ST} measured at the maximum number of markers, and at the Y axis is the lowest number of markers at which structure is observed determined by the given threshold.

point difference in the new analyses, with no population pairs showing increased clusteredness above 0.9, and only 3.7% of new observations decreasing below 0.9. We can say, therefore, that length of chain does not significantly affect results. Specifically, increasing the length of the chain does not increase or significantly decrease the number of populations analyzed in the G90 analysis.

We also performed correlation analyses for all other datasets as well (Table 6). As one would expect, all correlations are negative. Essentially, as F_{ST} or sample sizes increase, the number of markers needed to observe differentiation decreases. As in the case of the STR678 dataset, change in magnitude of correlations from simple to partial correlation is minimal for nearly all datasets. We also see that correlations of sample sizes to markers for G50 analyses for low F_{ST} are much greater than for G90, and in one case greater than that of F_{ST} versus marker numbers, which are usually greater. These are also the cases where we see a drop in magnitude of correlation from simple to partial. Given the decrease in magnitude of correlation when using the higher threshold standard of G90, it appears that sample size has a more significant effect on the number of markers needed to observe differentiation when using the lower standard of G50.

Microsatellites versus SNPS. To test which marker system required more markers for detection of differentiation, we compared results up to 675 markers, the maximum number of STRPs considered, for both SNPs and STRPs. For STRP and SNPs, we find that 94.87 and 97.10%, respectively of population pairs demonstrate differentiation at the 50% threshold level, and 78.58 and 31.93% at the 90% threshold level. To consider this analysis further, and to include informativeness of microsatellites as well, we compared results at different marker numbers up to 100 markers (see Table 7, Figure 20). In the case of the STR678 dataset, we refer in this table as "Random" to indicate that the markers were randomly chosen in contrast to "Informative" ones. "Overall" refers to the frequency at the maximum number of markers. As we can see, the likelihood of observing structure at lower numbers of markers is greater when considering informative markers. Interestingly, the probability of observing structure at most numbers of markers, up to 100, is greater for randomly chosen SNPS than for randomly chosen microsatellites when considering a standard level of assignment; however, the observation is reversed when using a confident (*i.e.*, G90) level of assignment. For this analysis, only those population pairs shared among the three datasets were considered (1326 population pairs).

We can say, then, that informative markers are more efficient for detecting genetic structure between pairs of populations, but our conclusion regarding efficacy of SNPS versus microsatellites is not conclusive. While SNPs and microsatellites have similar abilities overall to detect structure at the standard assignment level, if one uses the higher standard of confident assignment, it would appear that microsatellites are more powerful. This makes sense, since the attractiveness of microsatellites has been that their higher mutation rate is more appropriate for shorter time scales, such as observed for the separation between human populations.

4.4 ALDH2

Out of 20,000 simulated chromosomes, 6023 alleles had frequencies of 6.9% in the world, and 5261 had frequencies of 22% in Asia for the Asian-limited allele. In addition, for the OOA-limited allele 305 had frequencies of 33.3% in the world, and 1355 had frequencies of 42.3% on the OOA branch. When we sampled alleles that

Table 7: Cumulative percentage of population pairs that showing differentiation given a specific number of markers for informative and randomly chosen microsatellites, and randomly chosen SNPS for standard and confident assignment levels. Overall percentages apply to whole datasets, that is, observations up to and including the maximum number of markers.

	Informative		Random		SNPS	
Markers	G50	G90	G50	G90	G50	G90
5	10.41	4.61	4.93	0.79	4.77	0.08
10	29.73	20.43	17.09	5.72	29.33	0.40
15	47.46	37.84	25.83	11.13	49.60	3.82
20	59.86	49.52	41.73	25.12	59.86	5.33
25	67.97	55.80	47.69	31.64	65.82	12.00
30	74.01	62.24	52.38	38.55	72.42	13.28
35	77.58	65.58	56.36	43.00	77.27	15.26
40	79.81	68.52	59.86	48.17	81.80	22.89
45	81.32	70.11	66.30	48.89	83.15	23.45
50	83.15	71.54	69.32	51.43	85.06	24.17
55	84.82	72.50	71.38	52.54	86.72	25.20
60	85.69	73.37	74.80	55.09	87.68	25.83
65	86.57	73.93	76.63	57.07	89.19	26.23
70	86.88	74.40	78.62	59.54	89.67	26.63
75	87.84	74.48	80.92	61.29	90.62	27.27
80	89.11	74.96	82.35	61.84	91.10	27.42
85	90.30	75.36	83.55	64.71	91.65	27.82
90	91.57	75.76	84.50	65.90	91.97	27.90
95	93.24	76.07	84.90	66.06	92.29	27.98
100	94.20	76.39	87.04	67.33	97.10	27.98
overall	94.87	78.58	96.38	70.97	97.10	31.93

had a 6.9% frequency in the worldwide population, that is an overall frequency in the simulated data, we find that it only occurs with a 0.0092 probability in a single population (see Figure 21). Further, in the simulated Asian population, it only occurs with a 0.003 probability. When we sampled alleles with a frequency of 22% in Asia (that is, in the simulated Asian population) we find a 0.9966 probability of it being found elsewhere (see Figure 22). For alleles with frequencies matching the OOA-restricted alleles, we find that it has a probability of 0.0024 of only being found on the OOA branch (that is, anywhere but Africa) when its frequency is 33% globally and probability of being found in Africa of 0.9985 when it occurs at 42% on the OOA branch (results not shown).

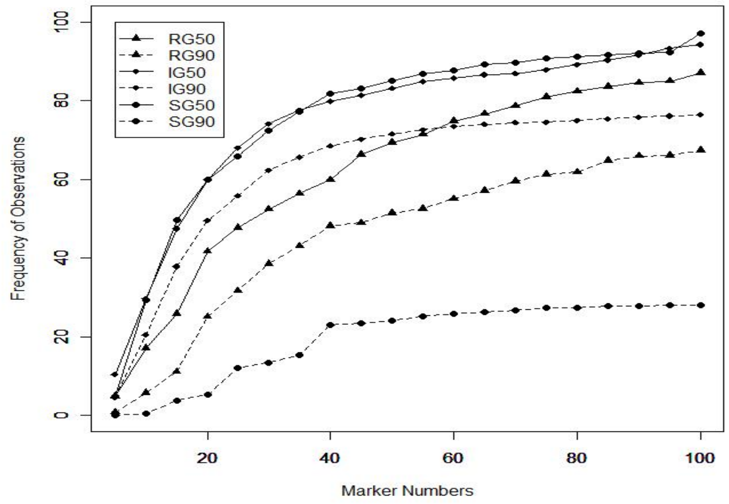
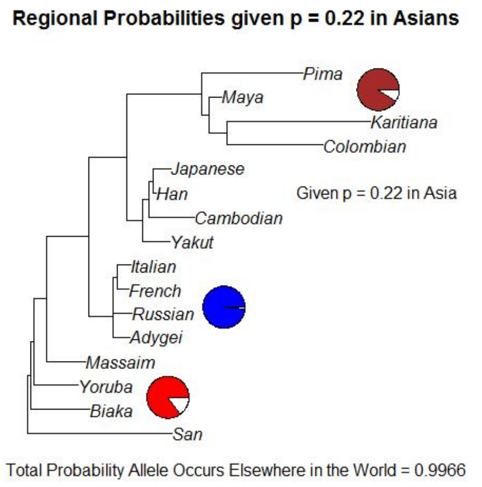


Figure 20: Comparison of numbers of markers needed to identity differentiation between pairs of populations for random (R) and informative (I) microsatellite, and SNP (S) markers for $G \geq 0.50$ (G50) and $G \geq 0.90$ (G90). X axis is marker numbers and Y is cumulative frequency of observations where differentiation is observed at that number of markers.



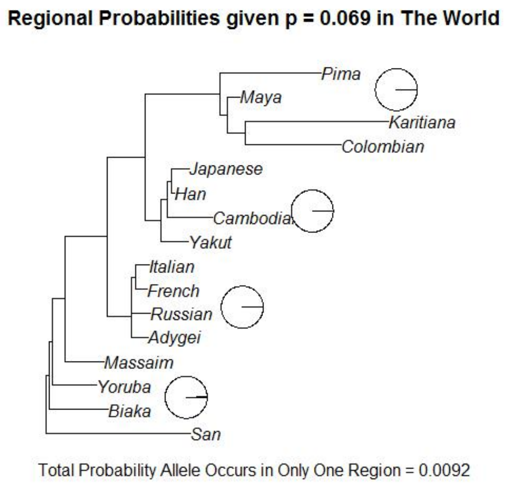

Regional Probabilities given p = 0.069 in The World
Pima
Maya
Karitiana
Colombian
Japanese
Han
Cambodia
Yakut
Italian
French
Russian
Adygei
Massaim
Yoruba
Biaka
San
Total Probability Allele Occurs in Only One Region = 0.0092

Figure 21: Probability of simulated alleles with frequencies in Asia of 0.22 appearing in other geographical populations (left).

Figure 22: Probability of simulated alleles with global frequencies 0.069 appearing in different geographical populations (rights).

4.5 Archaic Admixture

We used generalized hierarchical modeling [41] to fit a serial founder effects model to our data of 100 worldwide populations. In Out-of-Africa (OOA) populations, we observe a decrease in realized (that is, estimated from the data) gene identities relative to expected gene identities (Figure 23), with no difference observed in African populations, and a decrease between African and non-African populations. We suspect that input of new, unique mutations from a separate lineage, such as an archaic one, into modern humans leaving Africa increased heterozygosity in the populations on that branch. Hence, we decided to simulate a set of populations following an inferred human demographic history with and without archaic admixture to determine whether observed deviations can be attributed to archaic admixture on the OOA branch. *Ghm* plots in Figures 24 and 25 illustrate what we actually observe in the data and what we expect to observe sans admixture, respectively, for the subset of 16 populations. We see in Figure 25 that realized and expected gene identities match. Also, though the expected and realized genetic distances correlate pretty closely (r = 0.99) in the first case, showing that the data fits the model well, we see a closer fit in the case of our second figure (r = 0.999), indicating that a SFE model without admixture improves the fit to the data.

The observed difference between expected and realized gene identity coefficients in our subsample of 16 populations is similar that what we observe in the larger sample of 100 populations (Figure 23) from the extended CEPH HGDP. Therefore, we felt justified in only simulating and analyzing a 16 population subsample. When we simulate a model with no admixture, we see no difference between expected and realized gene identity coefficients in the ghm plot (Figure 25). This shows what we expect from a pure SFE model with no archaic admixture.

We also compared gene identities between populations within one geographical region, or at one node marking a major founder effect, and populations in other geographical regions against geographical distance. Results are shown in Figures 26. In these analyses, only 99 out of the 100 populations are shown. In the comparison of Native American populations versus populations in other regions, we see a decrease of realized gene identity compared to that expected from the model. Interestingly, for East Asian and European populations versus populations in other

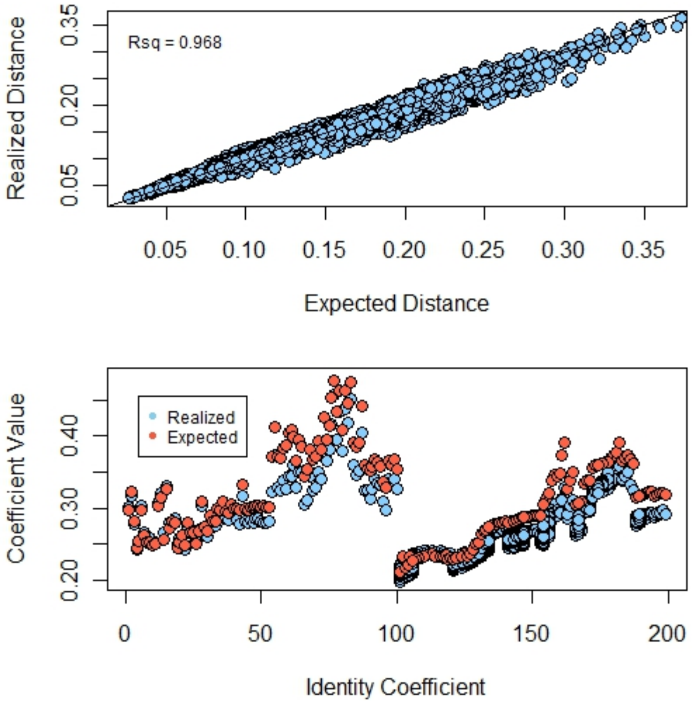


Figure 23: Scatter plot showing goodness of fit of the hierarchical model (top) and *ghm* plot (bottom) from observed results for 100 populations.

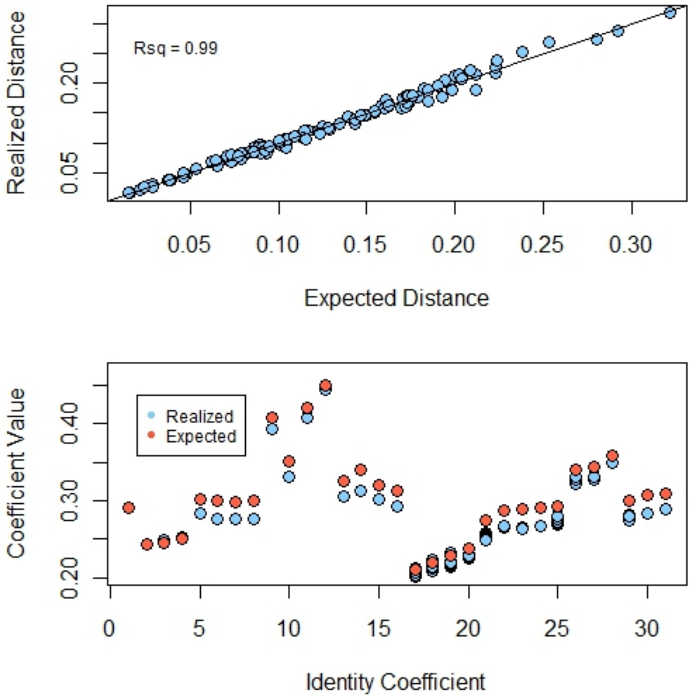


Figure 24: Scatter plot showing goodness of fit of the hierarchical model (top) and *ghm* plot (bottom) from observed results for 16 populations.

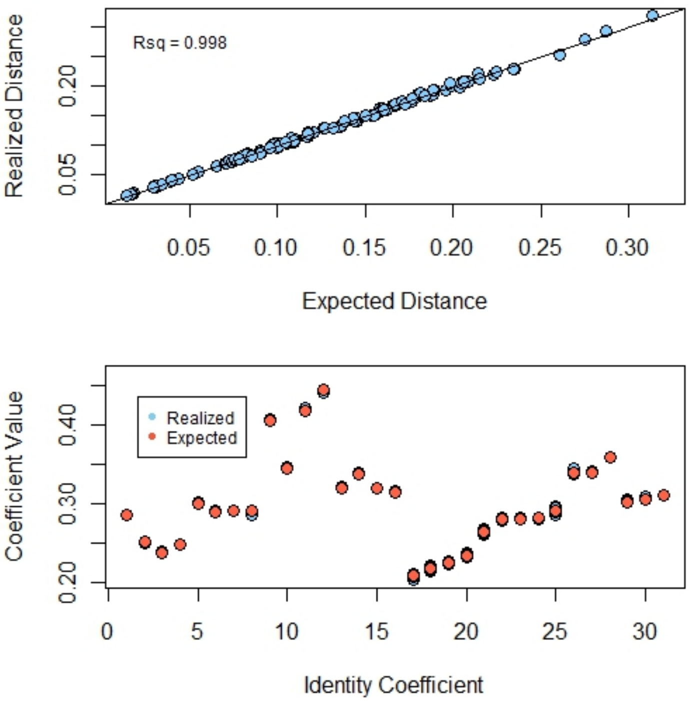


Figure 25: Scatter plot showing goodness of fit of the hierarchical model (top) and *ghm* plot (bottom) expected in the absence of archaic admixture for the subset of 16 populations.

regions, we see some interesting patterns. In addition to the decrease in realized gene identity, we also see decreasing gene identity with increasing distance between East Asian, and European and South Asian populations; and between European, and East Asian and South Asian populations. Finally, between populations separated at the root node (node 100), we see a decrease in realized gene identity between African and non-African populations, and little to no decrease in realized gene identity between African and other African populations.

Our simulations show that observed differences between realized and expected gene identities may be explained by archaic admixture with modern humans. In addition, we see that higher admixture rates results in increased divergence between realized and expected gene identities (see Figure 27). Further, we see that increasing divergence time increases the effect of archaic admixture on gene identity differences, or to put it another way increasing divergence times lowers the archaic contribution needed to produce a given difference, for example our observed difference of 0.02. A comparison of observed to simulated results shows

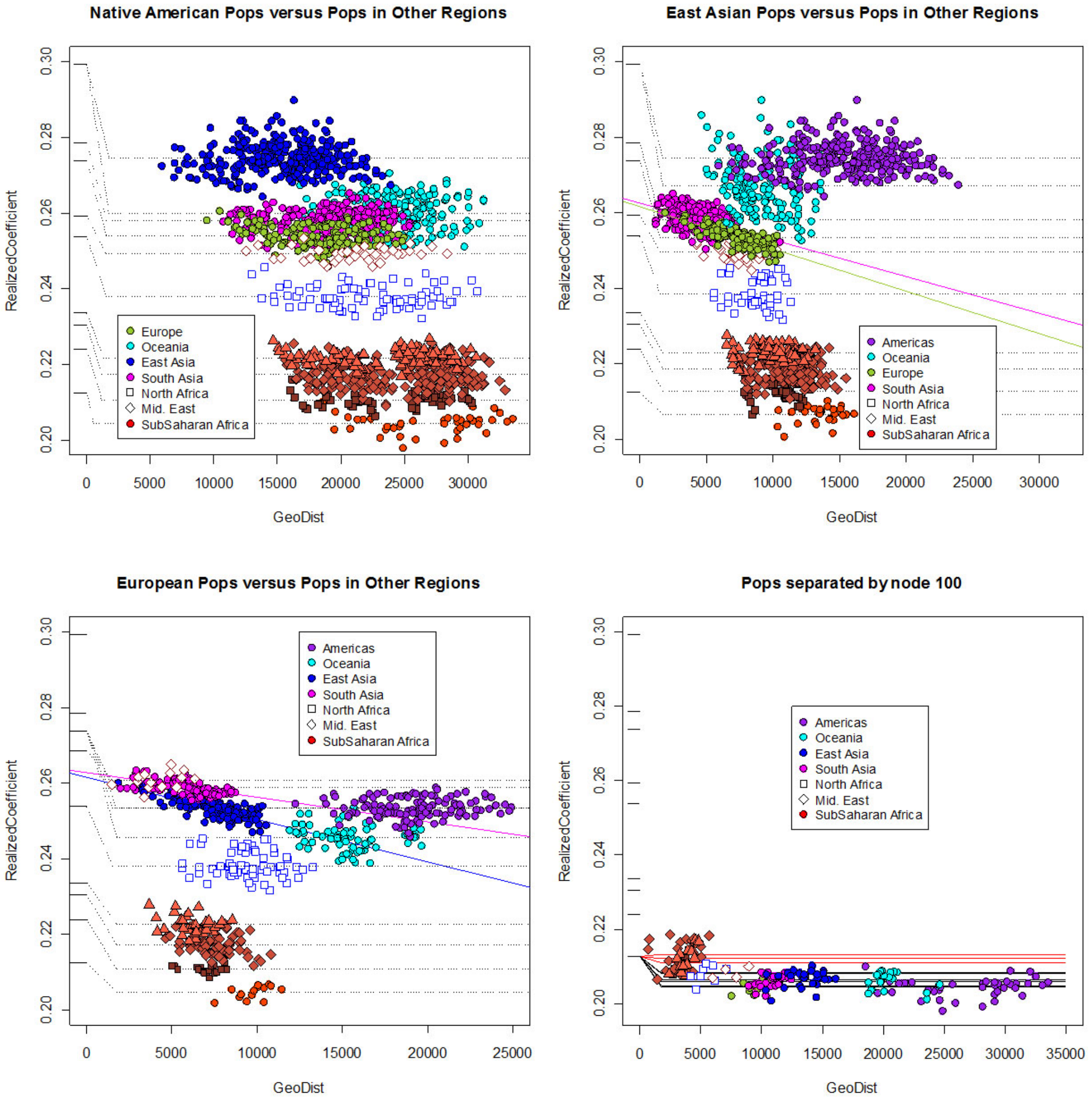


Figure 26: Gene identities between populations in different regions of the world against geogrpahic distance (GeoDeo). Bold lines on the Y axis indicate expected gene identities between populations at the shared nodes marking each geographical region, while dotted lines indicate realized gene identities between the populations.

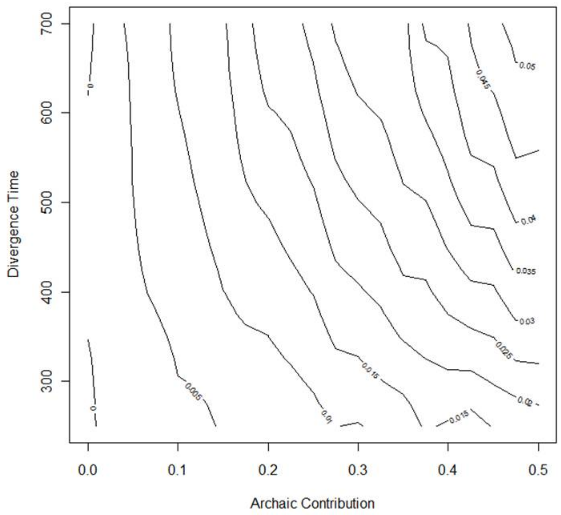


Figure 27: Contour plot of divergence between expected and realized gene identities for SMM. Gene identity difference is regressed on archaic contribution within each level of divergence time, using quadratic regression.

that the patterns in the observed are consistent with an admixture event on the OOA branch.

Simcoal2 also implements an option for microsatellite mutation following the generalized stepwise mutation (GSM) model [18]. We decided to test the effect of using different geometric probability distribution parameters for the GSM model on estimation of archaic contribution on the OOA branch. Setting the GSM parameter to 1 increases, slightly, the archaic contribution needed to produce observed differences at given divergence times (results not shown). However, setting the GSW parameter to 0.5 or 0.9, to model a Generalized Stepwise Mutation model [18], considerably decreases the archaic contribution needed to produce these differences. Using GSW = 0.5, the archaic contribution needed to approximate the observed difference is approximately 10 and 20% at 700 and 450k years (results not shown), respectively, while using GSW = 0.9 requires archaic contributions of approximately 10, and 20% at 650 and 325k years, respectively (Figure 28). We needed to change mutation rates to $4.75 * 10^{-5}$ and $4.00 * 10^{-5}$ to match our observed gene identity at the base when using GSW = 0.5 and 0.9, respectively. Contour plot for GSW = 0.9 shown in Figure 28.

We also tested our models using a range of effective population sizes, holding divergence time constant at 400 years, to test whether Green [141] or Huffs [142]

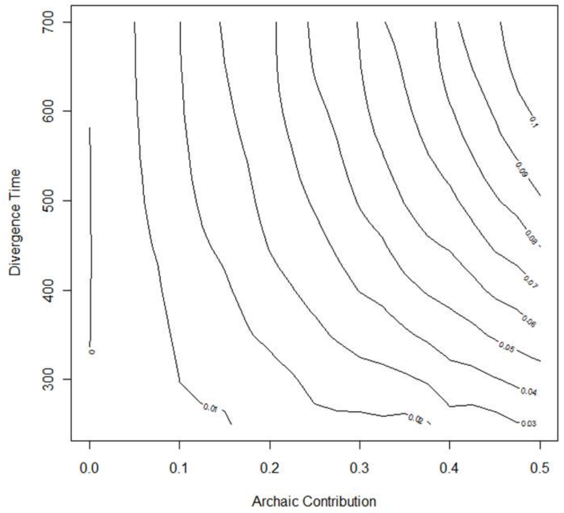


Figure 28: Contour plot of divergence between expected and realized gene identities for GSW = 0.9. Gene identity difference is regressed on archaic contribution within each level of divergence time, using quadratic regression.

estimates of 3000 and 18,500 individuals had any effect on results. We find no effect of effective population size on gene identity differences in the presence of archaic admixture.

4.6 AIDA

Of note, we refer to the first distance class as the zero distance class and the more distant classes numbered positively from one, such that what may be considered the second distance class is actually distance class one. Figure 29 contains cladograms for all datasets. We observe clinal structuring from isolation-by-distance on a continent-wide scale, as observed when analyzing the YBase dataset, as well as at a more local, regional scale, as observed when Finnish, Italian, and UK populations are analyzed separately. Some confusion abounds regarding what is and what isnt a pattern indicative of isolation by distance, particularly as visualized in a correlogram. Isolation by distance occurs when there is a tendency for mate choice to occur within ones own population or between closely located populations, which would be observed as a steady, or approximately steady, decline in similarity or autocorrelation. The case where you have significant positive autocorrelation at zero distance followed by non-significant or nearly zero auto-

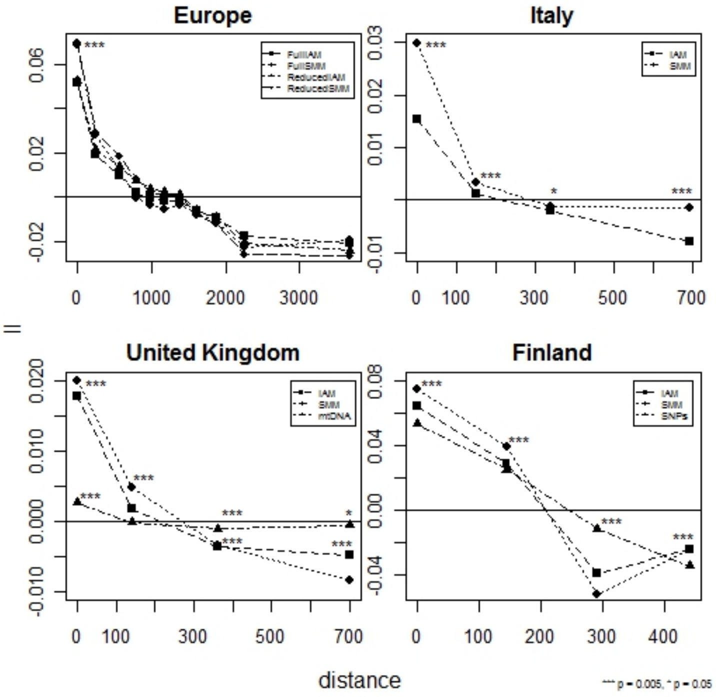


Figure 29: Correlograms for European and regional datasets. Distances except for the 0 distance class are the average of the distance boundaries for the distance class. The symbols * and *** indicate significance at the 0.005 and 0.05 levels, respectively. For the European, that is, the YBase, dataset, II values are significant for all distance classes, for both IAM and SMM models for the reduced dataset. For Italy all distance classes for IAM and SMM models are significant at the 0.005 level for except the third one (distance class 2), which is significant at the 0.05 level. UK populations have significant II values for IAM and SMM models for all distance classes, and for the mtDNA for all distance classes except the second. All significance is at 0.005 level except for the final UK mtDNA distance class, where significance is at 0.05 level. The Finns have significant II values for IAM and SMM models, and for Y-SNPs for all distance classes.

correlation for other distance classes or a distinct, steep drop in autocorrelation following the zero distance class, is more indicative of discrete structuring in the dataset.

In Europe (that is, the Ybase dataset), the pattern observed for both SMM and IAM models mirror, to some extent, that found in Roewer [115] using haplogroup frequencies. However, while their results show an exponential pattern indicative of what some would refer to as isolation by distance [143], and though ours initially appears to show an exponential isolation by distance pattern, most stark between the 0 and 1 distance classes, removal of the 0 and final distance classes reveals a distinct clinal pattern through increasing distance classes (as observed in the insert in the European correlogram in Figure 29. The pattern observed in Europe appears to be a result of structure amongst populations creating a large decline in autocorrelation from the zero distance class, with minor sharing of haplotypes with populations in close proximity resulting in a clinal pattern across non-zero distance classes. Results from the reduced European dataset approximately mirror those of the full dataset, showing that there is little to no effect of uneven sampling on results.

In Italy, under the SMM model autocorrelation shows a structured pattern of similarity for haplotypes in the zero distance class and insignificant or close to zero similarity at other distance classes, while IAM appears to be more clinal (though the steps between all distance classes but 0 and the first are slight). The Finns show probably one of the more distinctly clinal patterns (at least from 0 distance class to 1 and 2), but there is also a depression between distance classes 1 and 3. Autocorrelation on Y-SNPs in Finnish populations show a similarly distinct cline, but with no depression as observed for Y-STRPs. Also, *II* values are pretty similar for both mutational models at all distance classes, and also for SNPS at all distance classes except the second. The UK populations also display a distinct clinal pattern, with the SMM model showing a slightly steeper cline. mtDNA for the UK populations, however, clearly evidences a structured pattern of similarity of haplotypes within populations and zero-similarity of haplotypes with other populations. For European, Finnish, and UK datasets, and Italian dataset for all but distance class 2, all II values for YSTR results and YSNP results for Finnish populations are significant at the 0.005 level. For mtDNA, results in the zero distance class and distance class 2 are significant at the 0.005 level and distance class 3 at the 0.05 level.

5 DISCUSSION

5.1 Comparison of Genetic and Genealogical Data

Historical evidence indicates a common origin for the villages of Gioi and Cardile, separating approximately 1000 years ago. We were interested in determining whether recently separated populations with a shared origin, limited geographical distance between villages, and small population sizes could be differentiable using model-based clustering methods. Divergence between villages is low (F_{ST} = 0.008), but approximating what has been estimated in European populations. Rosenberg et al. [2] found differentiation amongst European populations with a lower F_{ST} of 0.007 using 377 microsatellites. Our observed high level of consanguinity within villages indicates that these populations do in fact represent genetic isolates. As such, they may be useful in disease-gene association studies. Though the degree of consanguinity is lower in the combined villages than in each village separately, it is still high enough to provide additional evidence of a common origin for the two villages.

Access to genealogical records dating to the 17^{th} century allowed extensive reconstruction of pedigrees, from which we calculated kinship coefficients. Though kinship estimates estimated through genealogy path counting are not considered as accurate as those estimated from genomic data, particularly in the presence of high consanguinity [144, 145], we needed an alternative measure of relatedness, independent from data used to infer genetic clusters, against which we could compare our genomic results. We find a large and significant correlation between the two descriptors of population structure, one based on genealogical pedigrees, and the other estimated from neutral genetic polymorphisms. This shows that non-random mate choice and limited population size are reflected in the distribution of allele frequencies.

Latch et al. [58] found 97% accuracy in the performance of model-based clustering methods in assigning individuals to clusters when F_{ST} is greater than 5%. To the contrary, we were able to identify two clusters, roughly corresponding to the two villages, and nearly consistently assigned individuals to the cluster corresponding to their village of sampling, despite our lower F_{ST} of 0.008. One reason for the reduced performance of clustering methods in their study with lower levels of F_{ST}, and the inconsistency with our findings, may be due to the lower numbers they considered (*e.g.*, 10) and the fact that they simulated co-dominant markers, which do not have the resolving power of microsatellites and therefore may be considered inappropriate for studying scantily-differentiated populations. Supporting our results, we show (1) increased relatedness of individuals within inferred clusters and a correlation of membership to a cluster with relatedness to other members of the cluster (Figure 12), (2) increased kinship within a cluster with increased mean membership coefficient to the cluster (Figure 13), and (3) increased differentiation between and among clusters with increased relatedness within each cluster (shown by increase in F_{ST} with increase in average kinship within clusters as membership within clusters is restricted to increasingly stringent threshold requirements, Figure 14). We find evidence that individuals that cluster with the village opposite from which they were sampled are likely the result of recent migration from the other village.

5.2 Scantily-differentiated populations

15% of global variation among worldwide human populations has been found to be attributed to variation between populations, either within or between regions, with the rest accounted for by genetic differences between members of the same local population [9,11,146]. Advances in technology for genotyping and genetic analyses have made possible fine scale analysis of human population structure [10,147], but shortage of markers and inadequate sample sizes may reduce power for detecting structure between and among populations with lower levels of differentiation. We were interested in quantifying the effect of limited marker numbers and sample sizes on the detection of structure in scantily-differentiated populations. Here, we explored the behavior of microsatellite markers since they have been the most commonly used marker systems for investigating genetic variation and demographic history in closely-related populations, particularly in humans.

Our results confirm those of earlier studies using microsatellite and other marker systems. Bamshad et al. [54] demonstrated a 90% or greater mean accuracy of assignment to continent of origin when at least 60 alu insertion or microsatellite markers were used, which increased to greater than 99% mean accuracy with 160 markers. Similarly, Waples and Gaggiotti [55] failed to identify population structure when using 20 low mutation rate ($\mu = 5 * 10^{-7}$) markers and 50 individuals from four populations with $N_e = 50$ with scantily-differentiated populations (F_{ST} = 1%), but correctly assigned individuals with 80% accuracy with $F_{ST} \geq 5\%$.

We show that the observed genetic structure in our isolated populations from southern Italy is largely the result of the high consanguinity present within the villages. In fact, removal of individuals related to the level of first cousins completely removes all trace of structure between these populations. When observations of genetic structure are solely the result of consanguinity, genetic differences may only affect a subset of loci in a portion of the population [60]; however, when it results from reproductive isolation, genetic differences may systemically affect the whole genome in nearly all individuals. This latter observation may be of concern for health and forensic science applications. Therefore, it is important to identify the extensiveness of consanguinity in study populations. That is, one should determine whether a mere subset of study subjects is affected, or whether it is more widespread, affecting most or all of the population. In the former scenario, removal of consanguineous samples allows estimation of structure without bias caused by relatedness, but in the latter, removal of consanguineous individuals may remove all trace of structure from the population, as observed when we removed individuals related to the level of first cousins. In human populations, genetic relationships between pairs of individuals within a genealogy may be seen as loops composed of ancestors connecting them, with the width of the loop proportional to the probability of sharing genomic regions identical by descent. Along with these loops connecting specific pairs of individuals, ancestors within one loop may be connected to other loops. Thus, removal of closely related individuals may result in loss of genealogical information about other, more distant, relationships. Though we intended on only removing individuals related up to a specific lower degree of relatedness, in the process we probably removed a number of individuals with more distant degrees of relatedness as well. For example, removing one individual from a pair related at the 1^{st} cousin level (or removing all individuals related to one particular individual related at the 1^{st} cousin level) will result in a loss of a number of individuals that are related at the 6^{th} or 7^{th} degree level. Not only did removing relatedness to the first cousin level eliminate all observation of structure, it also removed all detection of kinship as well. As such, we created a pair of populations where individuals within one village were just as likely to not be related to other individuals with the village as they would be to individuals in the other village. Essentially, we have two villages where individuals are completely unrelated to each other in their own or the other village.

When dealing with consanguineous populations, it may be useful to generate random samples from the dataset for consideration in model-based clustering methods. However, this is only applicable in populations with diffuse consanguinity. It is important to note that family-based clustering is a form of genetic structuring appropriately identified by clustering methods. In the absence of genealogical information, likelihood based methods for joint estimation of consanguinity and isolation from allele frequencies are available for estimating the contribution of consanguinity to genetic structure [60, 61].

We find here evidence of population structure despite the recent separation and geographical closeness of the two villages. In addition, we find that limited numbers of markers may be sufficient to detect structure even at low levels of differentiation, and that results of genetic analyses mirror measures of structure obtained from genealogical analyses. Thus, when reconstruction of genealogical pedigrees is impossible due to lack of records, or limited access to records, one may be able to observe the effects of kinship by analyzing a few hundred genetic polymorphisms. However, observed structure in scantily-differentiated may be the result of close familial relationships amongst members within villages, which should be taken into consideration when designing studies involving closely-related geographical isolates. Care should be taken to reduce the effect of related individuals in populations, but we must be aware that this may reduce completely the observation of structure. In addition, it is important to note that limited numbers of markers and small sample sizes may prevent observations of structure when it does in fact exist, particularly at lower levels of differentiation.

5.3 Marker Numbers

We used clusteredness as the metric for our analyses rather than the accuracy of placement approach of Bamshad et al. [54] because clusteredness depends little on where populations are sampled from and therefore is free from concerns that populations may actually share more genetic affinity with the opposite population rather than their sampled population. However, we also observe in our initial results on the STR678 dataset that, except in the case of low F_{ST} population pairs, most populations tend to have similar population and individual level clusteredness. Population level-clusteredness is an indication of the extent that individuals in the population tend to cluster more with the population to which they were sampled from (for example, if a population has high individual-level clusteredness but significantly lower population-level, it would indicate that individuals are not necessarily clustering with their sampled population).

Considering the case of randomly chosen markers, that is, ignoring the informativeness of markers which may decrease the number of markers needed to observe structuring, we see that a mere handful of markers ($m \leq 20$) may be needed to detect structure in well-differentiated populations. However, and predictably, as differentiation between populations decreases, greater numbers of markers are needed to identify population structure. However, and perhaps less predictably, even in poorly-differentiated populations structure may still be detected with low numbers of markers, although not always. For populations with F_{ST} of less than 0.01, marker numbers in the hundreds are likely to be required. Strictly speaking, this study shows that these results apply to the case where you are identifying two clusters with a given F_{ST} between populations; we are not sure how these results will play out when considering more than two populations, but we expect a similar relationship.

We also see that even in well-differentiated populations, 10 markers, as is still common in molecular ecology studies [148–151] may not be sufficient to detect structure. Though Latch (2006) seemed to have had no difficulty in this regard with her simulated populations, she also used large numbers of samples per population (*i.e.*, 100). We found that, based on the correlation of sample size with number of markers needed to observe structuring, some of the variation in marker numbers needed to observe structuring at similar F_{ST} levels appears to be a result of differences in sample sizes. Therefore, we may conclude that increasing sample size may also increase the likelihood of observing structure, particularly in scantily-differentiated populations, if structure exists. Also, increasing sample sizes may decrease the number of markers needed in analyses. However, often it is much easier to type more markers than to sample more individuals, particularly in genetic isolates where many individuals are genealogically related [63, 152].

Also, although our sample sizes are generally smaller than those in Latch's simulated populations, so our results may not be directly comparable, her 100 individuals per sample may not be realistic for most studies, particularly for human populations, and especially if one is careful to avoid related individuals. To examine this further, we looked at sample sizes of 245 human populations from the extended CEPH HGDP: the original CEPH HGDP [1,2], plus additional populations from South and North America [5], Africa [4], India [15], and Pacific Islanders [3]. Here, the maximum number of individuals in a single population was 63 and the mean number of individuals per sample was 21. The mode was only slightly larger at 25 (median = 48). Other human population genetic studies tend to show smaller sample sizes as well.

For an examination of sample sizes in typical molecular ecology studies, one need only look at some of Latch's own work, where most sample sizes were lower, often much lower, than her 100 simulated samples per population. While a few populations in her studies had $\geq$ 100 samples, this was not often the case, and even in these studies mean sample sizes were quite lower: 44 [153] and 62 [149]. Other studies show mean sample sizes of 37 with maximum size of 47 [151], 37 with a maximum size of 40 [150], and 36 with a maximum size of 58 [58]. Recent publications from other groups showed small sample sizes as well [148, 154–157]. While sample sizes such as these groups used may be sufficient to address specific questions of ecological or evolutionary relevance, they may be inadequate for identifying structure using model-based clustering algorithms (*i.e.*, stochastic) unless adequate numbers of markers are utilized.

While human population genetics studies have been increasingly using high-throughput SNP genotypes, many studies still use limited numbers of markers in their analyses, particularly in molecular ecology [38, 149], but also in human population genetics [57, 158]. Conclusions on genetic structure inferred from limited numbers of markers, typically that no substructure exists among populations (or limited structuring), should be approached with caution, because increasing the number of markers considered may reveal otherwise undetected structuring. In addition, we need to consider that results from recent simulation studies in molecular ecology which have used low numbers of markers, but large, unrealistic sample sizes in testing clustering algorithms [38, 55, 58] may need to be considered with caution. Their conclusions/results may be affected if sample sizes are adjusted to reflect common sample sizes.

Above and beyond the effects of sample size and number of markers considered, the informativeness of the markers also affects our ability to differentiate between populations pairs. At any number of markers, for G50 or G90, informative markers perform better at differentiating pairs of populations. The story with SNP markers is a little bit more complex; contrary to the conclusions of Lao et al. [8], it is not necessarily the case that fewer SNPs than STRPs are needed for detecting structure. While SNP markers perform better, though barely, at the G50 standard given different numbers of markers considered, STRP markers perform better at the G90 level. That is, at the G50 level more populations are differentiated with lower numbers of SNP markers, but at the G90 levels more populations are differentiated with lower numbers of STRP markers. Whichever marker system requires fewer markers may be dependent on the standard chosen for clustering. Of note, Lao et al. [8] concluded that fewer SNPs would be needed than STRPs when using carefully chosen markers, which we did not investigate here. SNPs may very well be shown to outperform STRPs if we specifically chose informative markers. Also, while we are testing for pairwise differentiation, Lao and colleagues was more concerned with observing geographical structure. Informative SNPs may work better at grouping geographical populations, where the divergence was more ancient, while STRPs may perform better at identifying differentiation between more recently diverged populations.

High-throughput datasets have become *de rigueur* in recent years [9, 16, 159]. While the increase in data may allow more detailed analyses of human populations, we may ask ourselves whether it is too much. We show that for many populations, at least when considering population pairs, most show structure with much smaller datasets. Though our analyses only constituted pairs of populations, it is reasonable to assume that our findings are applicable to sets of populations with complex patterns of clustering. However, we also have to take into account that one of the purposes of these large-scale studies, or interesting results to be gleaned from them, is not necessarily that of discrete structuring, but more fine scale structuring as per [10], and for particularly closely-related populations. While smaller datasets may be appropriate for identifying genetic structure, larger datasets may yet be important for higher resolution analyses of ancestry and admixture. Nevertheless, the question still stands, how much data is enough?

5.4 ALDH2

For both the Asia and OOA restricted alleles, observed distributions are inconsistent with expectations for neutral variants. Therefore, we can conclude that both alleles may have been subjected to historical selection in East Asia and in non-African populations, respectively. Additional evidence from this project supports the finding that natural selection is responsible for the restricted distribution of ALDH2*2. First, simulated alleles with frequencies matching those of ALDH2*2 have ages that encompass the OOA migration [160]. As such, these alleles would be expected to show a more widespread distribution. Also, neutrality is rejected using the Slatkin-Bertorelle intraallelic variability test [161], both for an intraallelic nucleotide substitution in ALDH2 and for a closely linked STR marker [160].

5.5 Archaic Admixture

It has been speculated that archaic admixture occurred on the population leaving Africa, but not within Africa [86]. This is consistent with our observations of reduced gene identity in OOA populations, but not in African populations, and between OOA and African populations. A part of this prediction is that input of new alleles from an archaic lineage has the effect of increasing heterozygosity on the OOA branch, thereby reducing gene identity. These predictions are borne out by our simulations. Input from a simulated archaic branch into the OOA branch results in reduction in gene identity in OOA populations, not in African populations, and between OOA and African populations. In addition, our results give support for isolation by distance between Eurasian populations in different regions, as shown by the decreasing gene identity between East Asian and European populations with increasing geographic distance in Figures 27 and 28, respectively.

In our simulations, the archaic contribution needed to produce the observed difference between expected and realized gene identities is higher than that estimated by Green [86]. To achieve our observed difference at 650k years, near the high end of divergence time estimates, we would need an archaic contribution of 20%, while a 15% contribution would require a divergence time of 850k years (not shown), outside the range of divergence time estimates. Using divergence time of 400k, slightly higher than the mean estimate of 370k years would require a 30% archaic contribution. However, choosing a GSM mutational model with geometric probability distribution parameter of 0.9 reduces the divergence time at which archaic contributions of 10 and 20% give us our observed difference between expected and realized gene identities. We do not observe the 5% archaic contribution estimated by Green [86] within our range of divergence times, even using the estimated divergence time of Homo erectus of 1 million years (not shown).

Another point to consider is that although our estimates are higher than those of Green, they are actually perfectly within the 0 to 20% CI range estimated by Noonan [162], and Wall et al. Wall [163] found 14% (CI 8 20%) and 1.5% (CI 0.5 - 2.5%) admixture estimates in European and East Asian samples, respectively.

Here, we develop a model for the effect of archaic introgression on modern human genetic variation. We start with a baseline model of pure phylogenetic radiation, with assumptions that 1) genetic drift reduces genetic variation, 2) mutation is the only process that adds genetic variation, and 3) evolution along each branch is independent of one another. Under these assumptions, the expected gene identity between any two contemporary populations will be less than or equal to the gene identity within their most recent common ancestor (*mrca*, and the lowest gene identity corresponds to that estimated between pairs of populations that share a most recent common ancestor defined by the root of the tree. In modern humans, this would correspond to the common ancestor of all contemporary modern humans, which we identify as *MCA* (*modern human common ancestor*).

To our baseline model, we add an archaic population, which persisted after the split from modern humans but which is currently extinct, and an introgression event on the OOA lineage. Our assumptions are that the introgression event is from the archaic into the modern population, though there is nothing to preclude introgression in the other direction. Though not violating assumptions one and three, this event does violate assumption two by allowing gene flow as well as mutation to increase variation in a lineage. This results in a more complex pattern of gene identities in the human population. A few considerations are in order: 1) the ancient common ancestor of modern and archaic humans (ACA), 2) the gene identity along the branch leading to the extinct archaic population (J_A), 3) the point on the modern human lineage at which introgression occurred, and 4) the percentage of genes in the modern human population descended from the archaic population (m). We note that the admixture event divides the tree into two moieties: hybrid descendents, populations and lineages on the out-of-Africa branch, and hybrid non-descendents, populations and lineages on African branches.

We label the branch on which admixture occurred b, and the populations before and after admixture B and H, respectively. At the point of admixture, population B becomes population H with gene identity decreasing according to the equation,

$$J_H = m^2 J_A + (1-m)^2 J_B + 2m(1-m) J_{ACA}$$

where J_A, J_B , J_H, and J_{ACA} are gene identities in the extinct archaic population, in the admixed modern human population prior to admixture, in the admixed modern human population following admixture, and in the archaic common ancestor, respectively. Expectedly, $J_{ACA} \leq min(J_A, J_B)$, and therefore $J_H \leq J_B$. We label the common ancestor of all populations descended from H, the $HRCA$ (hybrid radiation common ancestor), and we expect that $J_{HRCA} \geq J_H$. It is possible that admixture occurred just before radiation of populations from the OOA node, in which case $H = HRCA$ and $J_H = J_{HRCA}$, but not necessarily. In the event that admixture did not occur just before radiation, we would expect that $J_{HRCA}{>}J_H$ as a result of genetic drift after admixture.

Finally, and more amenable to analysis, we consider the gene identity between extant populations of the two moieties. Gene identity between hybrid non-descendent populations is unaffected by the admixture event, and these populations maintain a perfect tree-like structure with gene identity J_{MCA} at the base. Similarly, hybrid descendent populations maintain a perfect tree-like structure, though it differs from what we assume without admixture only by having a slightly lower gene identity at the root (*i.e.*, J_{HRCA}, as above). However, the gene identity between pairs of populations from each of these moieties is affected by hybridization as such,

$$J_{X_i Y_j} = m J_{ACA} + (1-m) J_{MCA}$$

where X_i and Y_j are populations from the hybrid descendent and hybrid non-descendent moieties, respectively. It is expected that $J_{X_i Y_j}$ will be less affected by the admixture event because $(1-m){>}(1-m)^2$.

Though we do not know the gene identity of the archaic common ancestor in reality, we can estimate the gene identity of simulated archaic common ancestors.

This allows us to test our second derived equation for estimating the gene identity in hybrid descendent/non-descendent common ancestors. Here, we considered the gene identity at nodes within Africa not considering the final African node. We took as the gene identity before admixture to be the gene identity between African populations sharing that node, and the gene identity between African and non-African populations sharing that node to represent the gene identity at the node after admixture. For all combinations of archaic contribution and divergence times, we compared results of $J_{X_iY_j}$ calculated using the equation and estimated from the data. Calculated results refer to those where $J_{X_iY_j}$ was determined by plugging in the gene identities between African populations sharing the node, while estimated results refer to those where $J_{X_iY_j}$ was estimated by the gene identities between African and non-African populations sharing the node. Differences between these results for all comparisons were similar to within 0.005.

To estimate the archaic contribution, we used our 100 population subsample, chosen to include 10 nodes within Africa. We then ran a regression on estimated gene identities before and after admixture, as the X and Y terms, respectively. The term $(1 - m)$ was considered to be the slope and mJ_{ACA} to be the intercept. We thus estimated the archaic contribution to be 11.1%, corresponding to a divergence time of 500k years from our simulations using a generalized stepwise mutation model with a geometric parameter of 0.9 (Figure 28). This estimation is lower than that estimated by Wall et al. [163] for archaic contribution to the modern human genome.

5.6 AIDA

Previous AIDA analyses on other uniparentally inherited markers have shown varied results. In European populations, Y-Chromosomal SNPs show a similar clinal pattern [102], while those in East Asian populations follows a quasi-clinal pattern [100], as seen in Finland. While no autocorrelation is seen for Portugal when observed separately [97], a more geographically dispersed analysis of the Iberian Penninsula shows a clinal autocorrelation, though only for the first three distance classes, with negative autocorrelation for the most distant distance class and the two other distance classes being not significant [99]. A previous analysis of Italy with Y-SNPs shows significant positive and negative autocorrelation for the zero and most distant distance classes, respectively, but no pattern or other significant values between them [98]. In addition, Y-chromosomal haplogroups in Italy and Greece, individually, display quasi-clinal patterns [114].

Siberian populations show positive autocorrelation at close distances, but little autocorrelation at further distances; however, when compared with other populations with possible historical contact with Siberian populations a clinal (or nearly clinal) autocorrelation pattern appears [101]. In addition, Northwest Siberian populations show a clinal pattern when compared with North Eurasian populations [164] for mtDNA and Y haplogroups. In North Africa, there is apparent autocorrelation, though with no clinal pattern (slightly declining positive autocorrelation at close distance classes, negative or no auto correlation at more distant distance classes), while the addition of the Middle East results in significant II values for the first 3 distance classes (where a slight decrease is seen before) to a clinal pattern at more distant distance classes [96]. Finally, Levant Y haplogroups show a quasi-clinal pattern and Central Asia a clinal pattern with some deviation [165].

MtDNA in European populations show spatial autocorrelation, but not in a clinal pattern as observed here [107], while Nile River populations do show the clinal pattern as we observe for Y microsatellites [77]. In South America, autocorrelation is clinal for close distance classes, but then gets lost in the more distant distance classes. When divided into eastern and western populations, eastern populations show a clinal pattern, while western ones show no autocorrelation [106]. MtDNA for Italy (with and without Sardinians) show quasi-clinal patterns, though a bit choppy, and some non-significance at some distance classes [92]. Aleutian islands show a clinal pattern [105]. A lack of autocorrelation is observed in Tuscany [104].

In the case of Sweden, for both mtDNA and Y chromosome SNPs autocorrelation is observed, but no real discernible patterns. Only significance given is for 99%, which only occurs, negatively, in the most distant distance class [166]. Some significant autocorrelation, but no distinct patterns, for 15 European population samples; however, when analyzed separately Mediterranean populations show a quasi-clinal pattern with two non-significant mid-distance classes [111].

A number of microsatellite-specific mutational model (*i.e.*, Stepwise Mutational Model, or SMM) specific statistics have been formulated for analyses of microsatellite data under the assumption that statistics based on the mutational scheme of microsatellites would better fit the data. However, at least in the case of spatial autocorrelation and European populations, mutational model choice does not affect results to a significant extent. We see no major difference between results when considering a more realistic stepwise model (SMM) then when considering an independent allele model (IAM). However, Italy does show some difference, with one model showing more clinal variation of differences indicate of isolation by distance, and the other appearing to demonstrate a clearly structured population. In addition, we also considered a limited number of datasets, and analyses of additional datasets with different demographic histories may reveal differences in results when considering different models.

BIBLIOGRAPHY

[1] Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A Human Genome Diversity Cell Line Panel. Science *296*, 261–262.

[2] Rosenberg, N., Pritchard, J., Weber, J., Cann, H., Kidd, K., Zhivotovsky, L., and Feldman, M. (2002). Genetic structure of human populations. Science *298*, 2381–2385.

[3] Friedlaender, J. S., Friedlaender, F. R., Reed, F. A., Kidd, K. K., Kidd, J. R., Chambers, G. K., Lea, R. A., Loo, J.-H., Koki, G., Hodgson, J. A., et al. (2008). The genetic structure of Pacific Islanders. PLoS Genet *4*, e19.

[4] Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.

[5] Wang, S., Lewis, C., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M., Molina, J., Gallo, C., et al. (2007). Genetic variation and population structure in native Americans. PLoS Genetics *3*, 2045–2067.

[6] Pistis, G., Piras, I., Pirastu, N., Persico, I., Sassu, A., Picciau, A., Prodi, D., Fraumene, C., Mocci, E., Manias, M. T., et al. (2009). High differentiation among eight villages in a secluded area of Sardinia revealed by genome-wide high density SNPs analysis. PLoS ONE *4*, e4654.

[7] Price, A. L., Helgason, A., Palsson, S., Stefansson, H., St. Clair, D., Andreassen, O. A., Reich, D., Kong, A., and Stefansson, K. (2009). The impact of divergence time on the nature of population structure: An example from Iceland. PLoS Genet *5*, e1000505.

[8] Lao, O., van Duijn, K., Kersbergen, P., de Knijff, P., and Kayser, M. (2006). Proportioning whole-genome single-nucleotide-polymorphism diversity for

the identification of geographic population structure and genetic ancestry. American Journal of Human Genetics *78*, 680–690.

[9] Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

[10] Xing, J., Watkins, W. S., Witherspoon, D. J., Zhang, Y., Guthery, S. L., Thara, R., Mowry, B. J., Bulayeva, K., Weiss, R. B., and Jorde, L. B. (2009). Fine-scaled human genetic structure revealed by SNP microarrays. Genome Research *19*, 815–825.

[11] Barbujani, G. and Colonna, V. (2010). Human genome diversity: frequently asked questions. Trends in Genetics *26*, 285–295.

[12] Barbujani, G. (2005). Human races: Classifying people vs understanding diversity. Current Genomics *6*, 215–226.

[13] Rosenberg, N. A. and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews in Genetics *3*, 380–90.

[14] Marjoram, P. and Tavare, S. (2006). Modern computational approaches for analysing molecular genetic variation data. Nature Reviews in Genetics *7*, 759–770.

[15] Rosenberg, N. A., Mahajan, S., Gonzalez-Quevedo, C., Blum, M. G. B., Nino-Rosales, L., Ninis, V., Das, P., Hegde, M., Molinari, L., Zapata, G., et al. (2006). Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet *2*, e215.

[16] Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*, 998–1003.

[17] Valdes, A. M., Slatkin, M., and Freimer, N. B. (1993). Allele frequencies at microsatellite loci: The stepwise mutation model revisited. Genetics *133*, 737–749.

[18] Kimmel, M. and Chakraborty, R. (1996). Measures of variation at DNA repeat loci under a general stepwise mutation model. Theoretical Population Biology *50*, 345–367.

[19] Crow, J. F. and Maruyama, T. (1971). The number of neutral alleles maintained in a finite, geographically structured population. Theoretical Population Biology *2*, 437–453.

[20] Kimura, M. and Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. Genetics *49*, 725–738.

[21] Wright, S. (1951). The genetical structure of human populations. Annals of Eugenics *15*, 323–354.

[22] Holsinger, K. E. and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nature Reviews in Genetics *10*, 639–650.

[23] Long, J. C. (2007). GHM: Generalized Hierarchical Modeling. Department of Anthropology, University of New Mexico.

[24] Wright, S. (1922). Coefficients of inbreeding and relationship. The American Naturalist *56*, pp. 330–338.

[25] Karigl, G. (1981). A recursive algorithm for the calculation of identity coefficients. Annals of Human Genetics *45*.

[26] Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., and McPeek, M. S. (2003). Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. American Journal of Human Genetics *73*, 612–626.

[27] Cole, J. B. (2007). PyPedal: A computer program for pedigree analysis. Computers and Electronics in Agriculture *57*, 107–113.

[28] Rosenberg, N. (2004). Distruct: a program for the graphical display of population structure. Molecular Ecology Notes *4*, 137–138.

[29] Pritchard, J. K., Wen, X., and Falush, D. (2010). Documentation for structure software: Version 2.3. Technical report Department of Human Genetics, University of Chicago Chicago.

[30] Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., and Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet *1*, e70.

[31] Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., et al. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol *3*, e196.

[32] Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics *131*, 479–491.

[33] Long, J. C., Li, J., and Healy, M. E. (2009). Human DNA sequences: more variation and less race. American Journal of Physical Anthropology *139*, 23–34.

[34] Cavalli-Sforza, L. L., Moroni, A., and Zei, G. (2004). Consanguinity, Inbreeding, and Genetic Drift in Italy. (Princeton, NJ: Princeton University Press).

[35] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.

[36] Chen, C., Durand, E., Forbes, F., and FranOis, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Molecular Ecology Notes *7*, 747–756.

[37] Wu, B., Liu, N., and Zhao, H. (2006). PSMIX: an R package for population structure inference via maximum likelihood method. BMC Bioinformatics *7*, 317.

[38] Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology *14*, 2611–2620.

[39] Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. The Annals of Statistics *1*, 135–141.

[40] Cavalli-Sforza, L. L. and Piazza, A. (1975). Analysis of evolution: Evolutionary rates, independence and treeness. Theoretical Population Biology *8*, 127–165.

[41] Lewis, C. M. and Long, J. C. (2008). Native South American genetic structure and prehistory inferred from hierarchical modeling of mtDNA. Molecular Biology and Evolution *25*, 478–486.

[42] Hunley, K., Cabana, G. S., Merriwether, D., and Long, J. C. (2007). A formal test of linguistic and genetic coevolution in native Central and South America. American Journal of Physical Anthropology *132*, 622–631.

[43] Kristiansson, K., Naukkarinen, J., and Peltonen, L. (2008). Isolated populations and complex disease gene identification. Genome Biology *9*, 109.

[44] Peltonen, L. (2000). Positional cloning of disease genes: advantages of genetic isolates. Human Heredity *50*, 66–75.

[45] Shifman, S. and Darvasi, A. (2001). The value of isolated populations. Nature Genetics *28*, 309–10.

[46] Wright, A. F., Carothers, A. D., and Pirastu, M. (1999). Population choice in mapping genes for complex diseases. Nature Genetics *23*, 397–404.

[47] Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J., and Stefansson, K. (2005). An Icelandic example of the impact of population structure on association studies. Nature Genetics *37*, 90–95.

[48] Abney, M., Ober, C., and McPeek, M. S. (2002). Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. American Journal of Human Genetics *70*, 920–934.

[49] Bourgain, C. and Genin, E. (2005). Complex trait mapping in isolated populations: Are specific statistical methods required? European Journal of Human Genetics *13*, 698–706.

[50] Newman, D. L., Abney, M., McPeek, M. S., Ober, C., and Cox, N. J. (2001). The importance of genealogy in determining genetic associations with complex traits. American Journal of Human Genetics *69*, 1146–1148.

[51] Helgason, A., Plsson, S., Gubjartsson, D. F., Kristjnsson, r., and Stefnsson, K. (2008). An association between the kinship and fertility of human couples. Science *319*, 813–816.

[52] Madrigal, L. and Melendez-Obando, M. (2008). Grandmothers' longevity negatively affects daughters' fertility. American Journal of Physical Anthropology *136*, 223–229.

[53] Corander, J., Marttinen, P., Siren, J., and Tang, J. (2008). Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics *9*, 539.

[54] Bamshad, M. J., Wooding, S., Watkins, W. S., Ostler, C. T., Batzer, M. A., and Jorde, L. B. (2003). Human population genetic structure and inference of group membership. American Journal of Human Genetics *72*, 578–589.

[55] Waples, R. S. and Gaggiotti, O. (2006). What is a population? an empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology *15*, 1419–39.

[56] Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., Hofreiter, M., Possnert, G., and Pbo, S. (2004). No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol *2*, e57.

[57] Vitart, V., Biloglav, Z., Hayward, C., Janicijevic, B., Smolej-Narancic, N., Barac, L., Pericic, M., Klaric, I. M., Skaric-Juric, T., Barbalic, M., et al. (2006). 3000 years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia. Eur J Hum Genet *14*, 478–87.

[58] Latch, E., Dharmarajan, G., Glaubitz, J., and Rhodes, O. (2006). Relative performance of bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics *7*, 295–302.

[59] Pritchard, J. (2004). Documentation for structure software: Version 2. Technical report Department of Human Genetics, University of Chicago Chicago.

[60] Overall, A. D., Ahmad, M., Thomas, M. G., and Nichols, R. A. (2003). An analysis of consanguinity and social structure within the UK Asian population using microsatellite data. Annals of Human Genetics *67*.

[61] Overall, A. D. J. and Nichols, R. A. (2001). A method for distinguishing consanguinity and population substructure using multilocus genotype data. Molecular Biology and Evolution *18*, 2048–2056.

[62] Rosenberg, N. A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Annals of Human Genetics *70*.

[63] Anderson, E. C. and Dunham, K. K. (2008). The influence of family groups on inferences made with the program Structure. Molecular Ecology Resources *8*, 1219–1229.

[64] Rosenberg, N. A., Burke, T., Elo, K., Feldman, M. W., Freidlin, P. J., Groenen, M. A. M., Hillel, J., Maki-Tanila, A., Tixier-Boichard, M., Vignal, A., et al. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics *159*, 699–713.

[65] Morin, P. A., Martien, K. K., and Taylor, B. L. (2009). Assessing statistical power of SNPs for population structure and conservation studies. Molecular Ecology Resources *9*, 66–73.

[66] Rosenberg, N. A., Li, L. M., Ward, R., and Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics *73*, 1402–22.

[67] Inoue, K., Fukunaga, M., Kiriyama, T., and Komura, S. (1984). Accumulation of acetaldehyde in alcohol-sensitive Japanese: relation to ethanol

and acetaldehyde oxidizing capacity. Alcoholism, Clinical and Experimental Research *8*, 319–22.

[68] Goedde, H. W., Harada, S., and Agarwal, D. P. (1979). Racial differences in alcohol sensitivity: A new hypothesis. Human Genetics *51*, 331–334.

[69] Eriksson, C. J. (2001). The role of acetaldehyde in the actions of alcohol (update 2000). Alcoholism, Clinical and Experimental Research *25*, 15S–32S.

[70] Li, H., Borinskaya, S., Yoshimura, K., Kal'ina, N., Marusin, A., Stepanov, V. A., Qin, Z., Khaliq, S., Lee, M. Y., Yang, Y., et al. (2009). Refined geographic distribution of the oriental ALDH2*504Lys (nee 487Lys) variant. Annals of Human Genetics *73*.

[71] Hunley, K. L., Healy, M. E., and Long, J. C. (2009). The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race. American Journal of Physical Anthropology *139*, 35–46.

[72] Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America *102*, 15942–15947.

[73] Jobling, M., Hurles, M., and Tyler-Smith, C. (2004). Human Evolutionary Genetics: Origins, People & Disease. (New York: Garland Science).

[74] Wolpoff, M. (December 2004). Why not the Neandertals? World Archaeology *36*, 527–546.

[75] Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. Nature *325*, 31–36.

[76] Eswaran, V. (2002). A diffusion wave out of Africa. Current Anthropology *43*, 749–774.

[77] Krings, M., Geisert, H., Schmitz, R. W., Krainitzki, H., and Pbo, S. (1999). DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. Proceedings of the National Academy of Sciences of the United States of America *96*, 5581–5585.

[78] Caramelli, D., Lalueza-Fox, C., Vernesi, C., Lari, M., Casoli, A., Mallegni, F., Chiarelli, B., Dupanloup, I., Bertranpetit, J., Barbujani, G., et al. (2003). Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proceedings of the National Academy of Sciences of the United States of America *100*, 6593–6597.

[79] Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., and Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. Proceedings of the National Academy of Sciences *104*, 17614–17619.

[80] Wall, J. D. (2000). Detecting ancient admixture in humans using sequence polymorphism data. Genetics *154*, 1271–1279.

[81] Relethford, J. H. (2001). Absence of regional affinities of Neandertal DNA with living humans does not reject multiregional evolution. American Journal of Physical Anthropology *115*, 95–8.

[82] Takahata, N., Lee, S.-H., and Satta, Y. (2001). Testing multiregionality of modern human origins. Molecular Biology and Evolution *18*, 172–183.

[83] Cyran, K. A. and Kimmel, M. (2005). Interactions of Neanderthals and modern humans: what can be inferred from mitochondrial DNA? Mathematical Biosciences and Engineering *2*, 487–498.

[84] Cyran, K. A. and Kimmel, M. (2010). Alternatives to the Wright-Fisher model: The robustness of mitochondrial Eve dating. Theoretical Population Biology *78*, 165–172.

[85] Currat, M. and Excoffier, L. (2004). Modern humans did not admix with Neanderthals during their range expansion into Europe. PLoS Biol *2*, e421.

[86] Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., et al. (2010). A Draft Sequence of the Neandertal Genome. Science *328*, 710–722.

[87] Sokal, R. R. and Oden, N. L. (1978). Spatial autocorrelation analysis in biology. 1. methodology. Biological Journal of the Linnean Society *10*, 199–228.

[88] Sokal, R. R., Oden, N. L., and Barker, J. S. F. (1987). Spatial structure in Drosophila buzzatii populations: Simple and directional spatial autocorrelation. The American Naturalist *129*, pp. 122–142.

[89] Sokal, R. R. and Friedlaender, J. (1982). Spatial autocorrelation analysis of biological variation on Bougainville Island Spatial autocorrelation analysis of biological variation on Bougainville Island. (Plenum Publishing Corporation).

[90] Sokal, R. R. and Oden, N. L. (1978). Spatial autocorrelation in biology 2. some biological implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society *10*, 229–249.

[91] Bertorelle, G. and Barbujani, G. (1995). Analysis of DNA diversity by spatial autocorrelation. Genetics *140*, 811–819.

[92] Barbujani, G., Bertorelle, G., Capitani, G., and Scozzari, R. (1995). Geographical structuring in the mtDNA of Italians. Proceedings of the National Academy of Sciences of the United States of America *92*, 9171–9175.

[93] Barbujani, G. (1987). Autocorrelation of gene frequencies under isolation by distance. Genetics *117*, 777–782.

[94] Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis: Models and estimation procedures. Evolution *21*, 550–570.

[95] Barbujani, G., Oden, N. L., and Sokal, R. R. (1989). Detecting regions of abrupt change in maps of biological variables. Systematic Biology *38*, 376–389.

[96] Arredi, B., Poloni, E. S., Paracchini, S., Zerjal, T., Fathallah, D. M., Makrelouf, M., Pascali, V. L., Novelletto, A., and Tyler-Smith, C. (2004). A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. American Journal of Human Genetics *75*, 338–345.

[97] Beleza, S., Gusmo, L., Lopes, A., Alves, C., Gomes, I., Giouzeli, M., Calafell, F., Carracedo, A., and Amorim, A. (2005). Micro-phylogeographic and demographic history of Portuguese male lineages. Annals of Human Genetics *70*, 181–194.

[98] Capelli, C., Brisighelli, F., Scarnicci, F., Arredi, B., Caglia', A., Vetrugno, G., Tofanelli, S., Onofri, V., Tagliabracci, A., Paoli, G., et al. (2007). Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. Molecular Phylogenetics and Evolution *44*, 228–239.

[99] Flores, C., Maca-Meyer, N., Gonzalez, A. M., Oefner, P. J., Shen, P., Perez, J. A., Rojas, A., Larruga, J. M., and Underhill, P. A. (2004). Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: implications for population demography. European Journal of Human Genetics *12*.

[100] Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R. S., Redd, A. J., Zegura, S. L., and Hammer, M. F. (2001). Paternal population history of East Asia: Sources, patterns, and microevolutionary processes. American Journal of Human Genetics *69*, 615–628.

[101] Karafet, T. M., Osipova, L. P., Gubina, M. A., Posukh, O. L., Zegura, S. L., and Hammer, M. F. (2002). High levels of Y-chromosome differentiation among Native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. Human Biology *74*, 761–789.

[102] Rosser, Z. H., Zerjal, T., Hurles, M. E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., et al. (2000). Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. American Journal of Human Genetics *67*, 1526–1543.

[103] Xue, Y., Zerjal, T., Bao, W., Zhu, S., Shu, Q., Xu, J., Du, R., Fu, S., Li, P., Hurles, M. E., et al. (2006). Male demography in East Asia: A north-south contrast in human population expansion times. Genetics *172*, 2431–2439.

[104] Bertorelle, G., Calafell, F., Francalacci, P., Bertranpetit, J., and Barbujani, G. (1996). Geographic homogeneity and non-equilibrium patterns of mtDNA sequences in Tuscany, Italy. Human Genetics *98*, 145–150.

[105] Crawford, M. (2007). Genetic structure of circumpolar populations: a synthesis. American Journal of Human Genetics *19*, 203–217.

[106] Fuselli, S., Tarazona-Santos, E., Dupanloup, I., Soto, A., Luiselli, D., and Pettener, D. (2003). Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. Molecular Biology and Evolution *20*, 1682–1691.

[107] Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J., and Barbujani, G. (2000). Geographic patterns of mtDNA diversity in Europe. American Journal of Human Genetics *66*, 262–278.

[108] Xiao, F.-X., Yotova, V., Zietkiewicz, E., Lovell, A., Gehl, D., Bourgeois, S., Moreau, C., Spanaki, C., Plaitakis, A., Moisan, J.-P., et al. (2003). Human X-chromosomal lineages in Europe reveal Middle Eastern and Asiatic contacts. European Journal of Human Genetics *2004*, 301–311.

[109] Fuselli, S., Gilman, R. H., Chanock, S. J., Bonatto, S. L., Stefano, G. D., Evans, C. A., Labuda, D., Luiselli, D., Salzano, F. M., Soto, G., et al. (2007). Analysis of nucleotide diversity of NAT2 coding region reveals homogeneity across Native American populations and high intra-population diversity. The Pharmacogenomics Journal *7*, 144–152.

[110] Pálsson, S. (2004). Isolation by distance, based on microsatellite data, tested with spatial autocorrelation (spaida) and assignment test (spassign). Molecular Ecology Notes *4*.

[111] Quintana-Murci, L., Veitia, R., Fellous, M., Semino, O., and Poloni, E. S. (2003). Genetic structure of Mediterranean populations revealed by Y-

chromosome haplotype analysis. American Journal of Physical Anthropology *121*, 157–171.

[112] Tarazona-Santos, E., Carvalho-Silva, D. R., Pettener, D., Luiselli, D., Stefano, G. F. D., Labarga, C. M., Rickards, O., Tyler-Smith, C., Pena, S. D., and Santos, F. R. (2001). Genetic differentiation in South Amerindians is related to environmental and cultural diversity: Evidence from the Y chromosome. American Journal of Human Genetics *68*, 1485–1496.

[113] Casalottif, R., Simoni, L., Belledi, M., and Barbujani, G. (1999). Y-chromosome polymorphisms and the origins of the European gene pool. Proceedings of the Royal Society of London. Series B: Biological Sciences *266*, 1959–1965.

[114] Giacomo, F. D., Luca, F., Anagnou, N., Ciavarella, G., Corbo, R., Cresta, M., Cucci, F., Stasi, L. D., Agostiano, V., Giparaki, M., et al. (2003). Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. Molecular Phylogenetics and Evolution *28*, 387–395. Special Issue: Papers presented at the Mammalian Phylogeny symposium during the 2002 Annual Meeting of the Society for Molecular Biology and Evolution, Sorrento, Italy, June 13-16, 2002.

[115] Roewer, L., Croucher, P. J. P., Willuweit, S., Lu, T. T., Kayser, M., Lessig, R., de Knijff, P., Jobling, M. A., Tyler-Smith, C., and Krawczak, M. (2005). Signature of recent historical events in the european y-chromosomal str haplotype distribution. Human Genetics *116*, 279–291.

[116] Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., and Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature Genetics *38*, 1251–1260.

[117] Presciuttini, S., Cagli, A., Al, M., Asmundo, A., Buscemi, L., Caenazzo, L., Carnevali, E., Carra, E., Battisti, Z. D., Stefano, F. D., et al. (2001). Y-chromosome haplotypes in Italy: the GEFI collaborative database. Forensic Science International *122*, 184–188.

[118] Lappalainen, T., Koivumki, S., Salmela, E., Huoponen, K., Sistonen, P., Savontaus, M.-L., and Lahermo, P. (2006). Regional differences among the Finns: A Y-chromosomal perspective. Gene *376*, 207 – 215.

[119] Sykes, B. (2006). Blood of the Isles: exploring the genetic roots of our tribal history. (London: Bantam).

[120] Colonna, V., Nutile, T., Ferrucci, R. R., Fardella, G., Aversano, M., Barbujani, G., and Ciullo, M. (2009). Comparing population structure as inferred from genealogical versus genetic information. European Journal of Human Genetics *17*, 1635–1641.

[121] Laval, G. and Excoffier, L. (2004). SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics *20*, 2485–2487.

[122] Fenner, J. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology *128*, 415–423.

[123] Zhivotovsky, L. A., Rosenberg, N. A., and Feldman, M. W. (2003). Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. American Journal of Human Genetics *72*, 1171–86.

[124] Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online *1*, 47–50.

[125] Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics *164*, 1567–1587.

[126] Jakobsson, M. and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics *23*, 1801–1806.

[127] Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources *9*, 1322–1332.

[128] Hauser, L., Seamons, T. R., Dauer, M., Naish, K. A., and Quinn, T. P. (2006). An empirical verification of population assignment methods by marking and parentage data: hatchery and wild steelhead (*Oncorhynchus mykiss*) in Forks Creek, Washington, USA. Molecular Ecology *15*, 3157–73.

[129] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution *4*, 406–425.

[130] Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle. Seattle.

[131] Kumar, S., Nei, M., Dudley, J., and Tamura, K. (2008). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in Bioinformatics *9*, 299–306.

[132] Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution *24*, 1596–1599.

[133] Nei, M. (1987). Molecular Evolutionary Genetics. (New York: Columbia University Press).

[134] Lewis, C. M. (2009). Hierarchical modeling of genome-wide Short Tandem Repeat (STR) markers infers native American prehistory. American Journal of Physical Anthropology *141*, 281–289.

[135] Kimura, M. The Neutral Theory of Evolution.).

[136] Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. Genetics *139*, 457–462.

[137] R Development Core Team. (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.

[138] Oliphant, T. E. (2006). Guide to NumPy. (Trelgol Publishing).

[139] Excoffier, L. and Lischer, H. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources *10*, 564–567.

[140] van Rossum, G. (1995). Python tutorial, Technical Report CS-R9526. Technical report CWI (Centre for Mathematics and Computer Science) Amsterdam, The Netherlands.

[141] Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., Rothberg, J. M., Paunovic, M., et al. (2006). Analysis of one million base pairs of Neanderthal DNA. Nature *444*, 330–336.

[142] Huff, C. D., Xing, J., Rogers, A. R., Witherspoon, D., and Jorde, L. B. (2010). Mobile elements reveal small population size in the ancient ancestors of Homo sapiens.

[143] Relethford, J. (2004). Global patterns of isolation by distance based on genetic and morphological data. Human Biology *76*, 499–513.

[144] Leutenegger, A. L., Prum, B., Genin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E. A. (2003). Estimation of the inbreeding coefficient through use of genomic data. American Journal of Human Genetics *73*, 516–23.

[145] Liu, F., Elefante, S., van Duijn, C. M., and Aulchenko, Y. S. (2006). Ignoring distant genealogic loops leads to false-positives in homozygosity mapping. Annals of Human Genetics *70*, 965–70.

[146] Barbujani, G., Magagni, A., Minch, E., and Cavalli-Sforza, L. L. (1997). An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences of the United States of America *94*, 4516–4519.

[147] Witherspoon, D. J., Wooding, S., Rogers, A. R., Marchani, E. E., Watkins, W. S., Batzer, M. A., and Jorde, L. B. (2007). Genetic similarities within and between human populations. Genetics *176*, 351–359.

[148] Bryja, J., Charbonnel, N., Berthier, K., Galan, M., and Cosson, J. F. (2007). Density-related changes in selection pattern for major histocompatibility

complex genes in fluctuating populations of voles. Molecular Ecology *16*, 5084–5097.

[149] Drauch, A., Fisher, B., Latch, E., Fike, J., and Rhodes, O. (2008). Evaluation of a remnant lake sturgeon populations utility as a source for reintroductions in the Ohio River system. Conservation Genetics *9*, 1195–1209. 10.1007/s10592-007-9441-9.

[150] Latch, E. and Rhodes, O. (2005). The effects of gene flow and population isolation on the genetic structure of reintroduced wild turkey populations: Are genetic signatures of source populations retained? Conservation Genetics *6*, 981–997.

[151] Latch, E. K., Scognamillo, D. G., Fike, J. A., Chamberlain, M. J., and Rhodes, O. E. (2008). Deciphering ecological barriers to North American river otter (*lontra canadensis*) gene flow in the Louisiana landscape. Journal of Heredity *99*, 265–274.

[152] Colonna, V., Ferrucci, R., M.Ciullo, M., and Barbujani, G. (2010). Detection of genetic structure in scantily differentiated populations: effects of consanguinity, divergence time, and effective population size. manuscript in preparation.

[153] Latch, E., Heffelfinger, J., Fike, J., and Rhodes, O. (2009). Species-wide phylogeography of North American mule deer: cryptic glacial refugia and postglacial recolonization. Molecular Ecology pp. 1730–1745.

[154] Aspi, J., Roininen, E., Kiiskil, J., Ruokonen, M., Kojola, I., Bljudnik, L., Danilov, P., Heikkinen, S., and Pulliainen, E. (2009). Genetic structure of the northwestern Russian wolf populations and gene flow between Russia and Finland. Conservation Genetics *10*, 815–826.

[155] Kennedy, L. J., Angles, J. M., Barnes, A., Carmichael, L. E., Radford, A. D., Ollier, W. E., and Happ, G. M. (2007). DLA-DRB1, DQA1, and DQB1 alleles and haplotypes in North American gray wolves. Journal of Heredity *98*, 491–499.

[156] Wheeldon, T. and White, B. N. (2009). Genetic analysis of historic western Great Lakes region wolf samples reveals early Canis lupus/lycaon hybridization. Biology Letters *5*, 101–104.

[157] Wilson, P. J., Grewal, S. K., Mallory, F. F., and White, B. N. (2009). Genetic characterization of hybrid wolves across Ontario. Journal of Heredity *100*, S80–S89.

[158] Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Thery, S., Froment, A., Bomin, S. L., Gessain, A., Hombert, J. M., et al. (2009). Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. Current Biology *19*, 312–8.

[159] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.

[160] Hunley, K., Li, J., Lewis, T. C., Malhi, R. S., and Long, J. C. (2010). Low intrallelic variation and ancient coalescent time reject neutral evolution for ALDH2*2. manuscript in preparation.

[161] Slatkin, M. and Bertorelle, G. (2001). The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics *158*, 865–874.

[162] Noonan, J. P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Pbo, S., Pritchard, J. K., et al. (2006). Sequencing and analysis of Neanderthal genomic DNA. Science *314*, 1113–1118.

[163] Wall, J. D., Lohmueller, K. E., and Plagnol, V. (2009). Detecting ancient admixture and estimating demographic parameters in multiple human populations. Molecular Biology and Evolution *26*, 1823–7.

[164] Pimenoff, V. N., Comas, D., Palo, J. U., Vershubsky, G., Kozlov, A., and Sajantila, A. (2008). Northwest Siberian Khanty and Mansi in the junction of West and East Eurasian gene pools as revealed by uniparental markers. European Journal of Human Genetics *16*, 1254–1264.

[165] Zerjal, T., Wells, R. S., Yuldasheva, N., Ruzibakiev, R., and Tyler-Smith, C. (2002). A genetic landscape reshaped by recent events: Y-Chromosomal insights into Central Asia. American Journal of Human Genetics *71*, 466–482.

[166] Lappalainen, T., Hannelius, U., Salmela, E., von Dbeln, U., Lindgren, C. M., Huoponen, K., Savontaus, M.-L., Kere, J., and Lahermo, P. (2008). Population Structure in Contemporary Sweden-A Y-Chromosomal and Mitochondrial DNA Analysis. Annals of Human Genetics *73*, 61–73.

ACKNOWLEDGEMENTS

I would like to thank Guido Barbujani for allowing me the opportunity to do research with him for my PhD, Vincenza Colonna for working with and helping me on projects during my PhD period, and Jeff Long for giving me the opportunity to work with him during my period abroad, as well as for giving me my first post-postgrad job. I would also like to thank Steve Wooding for introducing me to the python programming language, Henry Harpending for guidance during my transition from the University of Utah to Ferrara, Lynn Jorde for helping me to finish my Master of Science in Human Genetics degree from the University of Utah, and David Grunwald for allowing me to spend time in his lab before moving from the University of Utah to the University of Ferrara. Finally, I would like to acknowledge the Population Genetics and Genetic Epidemiology group at University of Ferrara and the Genetic Anthropology group at the University of New Mexico.

A very special thanks to my parents, Ronald and Barbara Ferrucci, for their support all these years as I traveled the world in search of education and adventure, my twin brother Ralph, with whom I usually stay when I am back in Connecticut, and my little sister Joann, who put up with Ralph and me for many years. I should also mention the many friends I made in my 2 years in Ferrara; their companionship made it easier being so far from my country for so long. In this regard, I would like to particularly thank Marco Marzocchi and Samantha Ragazzi, who have been babysitting my bicycle in their garage while I was back in the United States for a year.