NONSTATIONARY POPULATION GENETICS AND THE EVOLUTION OF HUMAN MOLECULAR DIVERSITY

by

Stephen Park Wooding

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Anthropology

The University of Utah

December 2001

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a dissertation submitted by

Stephen Park Wooding

This dissertation has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

	Chair: Alan R. Rogers
	Henry C. Harpending
	Dennis H. O'Rourke
	Lynn B. Jorde
	Jon A. Seger

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

FINAL READING APPROVAL

To the Graduate Council of the University of Utah:

I have read the dissertation of Stephen Park Wooding in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the Supervisory Committee and is ready for submission to The Graduate School.

Date

Alan R. Rogers
Chair, Supervisory Committee

Approved for the Major Department

Kristen Hawkes
Chair/Dean

Approved for the Graduate Council

David S. Chapman
Dean of The Graduate School

ABSTRACT

Population size change is common in natural populations, which nearly always vary over time. Well-known changes include expansions (black bears), declines (whales), brief bottlenecks (Argentine fire ants) and even cycles (lynx). Size changes should have a number of effects on population genetic diversity because of their effects on the shape of gene genealogies. However, the details of these effects are poorly understood because they are difficult to model. This thesis is an attempt to understand the effects of population size change in humans through a series of empirical, computational and theoretical studies.

"All models are wrong, but some are useful." - G. E. P. Box

CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I wish to acknowledge the help provided by my supervisory committe: Alan Rogers, Henry Harpending, Dennis O'Rourke, Lynn Jorde and Jon Seger. The chairman of my supervisory committee, Alan Rogers, provided particular support.

Chapters 3, 4, 5 and 6 are reprinted with the kind permission of *BioInformatics*, *Human Biology*, *The Proceedings of the National Academy of Sciences* and *Infection, Genetics and Evolution*, respectively.

The research in this thesis was supported in part by a Genome Informatics Training Grant to the University of Utah.

CHAPTER 1

INTRODUCTION AND OVERVIEW

1.1 Introduction

1.1.1 Population Size Change

Questions about the history of human population size underlie a number of fundamental problems in human biology. In anthropology, for example, much of the dispute about the origin of modern humans centers on disagreement over the effective size of archaic populations [1]. In genetic epidemiology, interpretations of molecular genetic diversity vary depending on assumptions about ancient population size [2]. Thus, the history of human population size is of interest for both academic and practical reasons.

Information about the history of human population size comes from several sources. Archaeology, paleoanthropology, linguistics and historical documentation are all important. Over the last 25 years, however, genetic evidence has risen to the forefront. By providing information inaccessible through traditional means, genetic data play a key role in inferences about the ancient human past. Central to this role are the theoretical tools of population genetics, which attempt to describe the relationship between demography and genetic diversity.

Early research in population genetics showed that variation in the rates of transmission of genes from parent to offspring, caused by factors like natural selection or population size change, has a number of effects on genetic variation. Subsequent research focused on predicting the pattern of genetic diversity expected under given population histories, or, conversely, inferring demographic conditions from extant patterns of variation. Over a span of 75 years, population genetic models ranging from the classical results of Wright [3] and Fisher [4] to the modern coalescent theories of Kingman [5], Tavaré [6] and Hudson [7] have been successful in identifying the influence of factors such as population size, mating system, subdivision, and genetic

recombination. An issue of definitive importance that remains difficult to address, however, is population size change [2, 8, 9, 10, 11, 12, 13, 14, 15].

Population size change is common in natural populations, which nearly always vary over time. Well-known changes include expansions (black bears), declines (whales) , brief bottlenecks (fire ants), and even cycles (lynx) [16, 17, 18, 19]. Such changes should have pervasive effects on genetic diversity because of their effects on the shape of gene genealogies [7, 20, 21]. However, the details of these effects are poorly understood because genealogical processes in populations with changing sizes, or non-stationary populations, are difficult to model. While models of genealogical processes in stationary populations can assume that the times between genealogical events are statistically independent, models of nonstationary populations must cope with the problem that the times between genealogical events are not independent. And because nonstationary populations are difficult to model, statistical methods useful for testing specific hypotheses about population history are limited. Many methods are available for testing the hypothesis that a population has been constant over time, for example, but few methods can discriminate between the hypotheses of, say, 10-fold growth from 100-fold growth over a specified time period [10, 22, 23].

The constraints faced by analyses unable to address size change are most obvious in humans. Demographic history has received more attention in *Homo sapiens* than in any other species—the term demography itself refers specifically to humans despite its generic usage (*demos* means people in ancient Greek)—and an abundance of data suggests to a human past characterized by large-scale expansion. Evidence from archeology and genetics implies that much of this growth occurred during the late Pleistocene, and that modern populations have not returned to genetic equilibrium [10]. Patterns of nucleotide substitution in mitochondrial DNA, for example, suggest that the modern human population is derived from a small original population that began growing rapidly around 80,000 years ago [10].

Population growth in humans has probably had many effects on genetic variation. It is known that population growth increases the proportion of low frequency genetic variants, and estimates of many statistics describing population genetic diversity are likely to be affected as a result [24]. Nonetheless, constant population size is

assumed in most studies of human genetic variation: inferences about allele age and the ascertainment of statistics like average heterozygosity often rely on the assumption of constant population size [25, 26]. The development of tools suitable for analyzing data from nonstationary populations is a clear goal for modern population genetics.

1.1.2 Coalescence Theory

Among tools used to study population genetic diversity, coalescent theory is widely regarded as the modern standard. Conceived originally by Kingman, the coalescent model focuses on the genealogical process [5, 27]. By defining the chances that pairs of individuals in any generation share a common ancestor in the previous one, Kingman's approach is able to provide a probabilistic description of ancestor-descendant relationships [7]. The superimposition of mutational processes onto this probabilistic description allows changes in genetic diversity to be modeled on a time scale. By incorporating information about genealogical characteristics, rather than just allele frequency, analytical methods based in coalescent theory are often able to extract more information from population genetic variation than traditional approaches.

Coalescence theory has been successful in addressing nonstationary processes in two main areas. First, coalescence theory provides a simple mechanism for simulating patterns of diversity under complicated demographic conditions. By working backwards in time and generating coalescent intervals one by one, simulations can generate instances of genealogies that are representative of specific demographic conditions [7]. Simulated genealogies can be aggregated and compared with observed genealogies to provide a basis for statistical analysis. Simulation-based tests, like Tajima's D, Fu's Fs and Rogers's Mismatch tests are common in studies of population genetic variation [10, 20, 21]. Second, several models are available for describing genealogical processes in nonstationary populations (e.g., Griffiths and Tavaré[9]). These models, however, have received little attention in empirical studies. One problem is that nonstationary coalescent models are usually complex. Another problem is that they are difficult to manipulate with computers. Although many potential applications for nonstationary coalescent analyses exist, few are used in practice.

1.2 Overview

This dissertation is composed of a series of studies that attempt to solve some of the theoretical and practical problems of studying genetic variation in nonstationary populations. The studies take a variety of approaches that address different problems. Chapter 2 describes a model of the genealogical process in nonstationary populations that uses a novel algebraic approach—a nonstationary Markov chain manipulated using matrices to simplify the calculation of expected coalescent intervals. Chapter 3 describes a computational resolution to the problem of analyzing exceptionally large or complex datasets—a parallel algorithm and distributed computer architecture used to speed up the analysis of DNA sequence mismatch distributions. Chapter 4 is a short paper that tests the hypothesis that human populations have been constant, using DNA sequence data from the X chromosome. Chapter 5 is a comparative analysis of evidence of population history in human and JC virus populations. Humans are the obligate host of the JC virus, which is largely benign, yet genetic evidence shows that the two species have distinctly different population histories. Chapter 6 is a synthetic analysis of population genetic variation in East Asia. This chapter shows that although several lines of evidence point to a north/south division of East Asian populations, genetic evidence does not support such a division. These five chapters are unified by an emphasis on the genealogical structure of human populations, and the difficulty of modeling genealogies in nonstationary populations. Together, the studies illustrate the diversity of applications of nonstationary coalescent theory, the difficulty of such applications, and several new directions for empirical and theoretical research.

CHAPTER 2

GENOME-WIDE SINGLE-NUCLEOTIDE POLYMORPHISMS: EVIDENCE FOR PURIFYING AND BALANCING NATURAL SELECTION IN HUMAN GENES

2.1 Abstract

To test the hypothesis that single-nucleotide polymorphisms (SNPs) in human genes show evidence of ancient human population growth, we analyzed the frequency spectrum of mutations in 158 SNPs from coding regions distributed broadly across the genome. A matrix representation of the coalescent process in populations that change in size over time was used to compare maximum likelihood estimates of demographic parameters. Observed patterns of allele frequency were not consistent with the hypothesis of explosive human population growth in the Late Pleistocene. In addition, demographic parameters estimated for synonymous SNPs and noncoding SNPs near genes were not significantly different from each other, but both differed significantly from parameters estimated for nonsynonymous SNPs. Whereas diversity patterns in nonsynonymous SNPs are consistent with population increase, diversity patterns in synonymous SNPs and noncoding SNPs near genes are not. These patterns may be explained by a combination of purifying and balancing natural selection in nuclear genes.

2.2 Introduction

The history of population size is a point of general interest in studies of biological variation. Among other things, population size changes can affect levels of heterozygosity, allele frequency, and the extent of linkage disequilibrium [14, 24]. In humans, these effects are an important consideration in problems ranging from evolutionary

biology to gene mapping. Information about long-term population size contributes both to the overall picture of ancient human history and to the understanding of modern human biology.

Several kinds of analysis can extract information about ancient population size from modern population genetic variation. However, applications of these analyses to humans yield ambiguous results [1, 28]. Some data, like haemoglobin amino acid sequences, mitochondrial DNA polymorphisms, Y chromosomal DNA sequences, and simple tandem repeats, support the hypothesis of a human population explosion at the end of the Pleistocene [10, 11, 29, 30, 31]. Other data, mainly from autosomal coding regions, do not [1, 28]. One explanation for the disparity among loci is that balancing selection has preserved ancient variation in genes, obscuring evidence of population growth in coding regions while leaving noncoding regions far from genes untouched [1].

To test whether multiple independently segregating sites within human genes show evidence of ancient human population growth, we used a matrix model of the coalescent process to analyze patterns of allele frequency in unlinked single-nucleotide polymorphisms. The matrix coalescent model provides a simple analytical method for calculating expected coalescence times under continuously varying population histories, which are difficult to evaluate otherwise [32]. These coalescence times can be used to calculate likelihoods under a variety of demographic conditions. Here, likelihood ratio tests were used to compare maximum likelihood estimates of demographic parameters within and between three SNP categories (coding nonsynonymous, coding synonymous and noncoding) under piecewise-constant population histories. Patterns of diversity in these SNPs show that although evidence for population growth is not present, it may be obscured by pervasive purifying and balancing natural selection.

2.3 Methods

2.3.1 The Matrix Coalescent

If time is measured backwards into the past, and a sample of k lineages is selected t generations before present from a population with size $N(t)$, then the probability that the k sampled lineages have $k-1$ distinct ancestors $t+1$ generations before present is approximately [7]

$$\alpha_k(t) = \frac{k(k-1)}{2N(t)}. \tag{2.1}$$

A sample of n lineages gathered at the present ($t = 0$ generations ago) will have a genealogy proceeding from the state of having n distinct lineages to the state of having $n-1$ lineages, and so on down to one lineage, at a rate determined by the transition probabilities $\alpha_n(t), \alpha_{n-1}(t), \ldots, \alpha_1(t)$. In general, the probability of observing k lineages t generations before present where $n \geq k \geq 1$ is described by a system of recurrence equations

$$p_k(t+1) = \begin{cases} p_k(t) \cdot (1 - \alpha_k(t)) + p_{k+1}(t) \cdot \alpha_{k+1}(t) & 1 \leq k < n \\ p_k(t) \cdot (1 - \alpha_k(t)) & k = n \end{cases} \tag{2.2}$$

with initial condition $p_n(0) = 1$, $p_{n-1}(0) = 0, \ldots, p_1(0) = 0$.

This system can be approximated in continuous time by

$$\frac{dp(t)}{dt} = Ap(t) \tag{2.3}$$

where $p(t)^T = [p_1(t), p_2(t), \ldots, p_n(t)]$ and A is the matrix of coefficients

$$A = \begin{bmatrix} -\alpha_1(t) & \alpha_2(t) & & \\ & -\alpha_2(t) & \ddots & \\ & & \ddots & \alpha_n(t) \\ & & & -\alpha_n(t) \end{bmatrix}.$$

2.3.1.1 Eigenvalues and Eigenvectors

Since A is a triangular matrix, its eigenvalues are equal to its diagonal entries: $-\alpha_1(t), \ldots, -\alpha_n(t)$. The column-eigenvectors of A are defined by the equation $Ac = c\lambda$, where λ is a scalar—one of the eigenvalues of A—and c a column-eigenvector. This equation can be re-expressed (suppressing t) as

$$c_{i+1} = c_i(\lambda + \alpha_i)/\alpha_{i+1} \tag{2.4}$$

where c_i is the ith entry in vector c. The jth eigenvector is calculated by setting $\lambda = -\alpha_j$, setting c_1 to an arbitrary constant, and then applying equation (2.4) repeatedly. When $i = j$, this equation becomes $c_{j+1} = c_j \times 0$. Consequently $c_i = 0$ for all $i > j$, and the matrix C of column eigenvectors is upper triangular.

Equation (2.4) also implies that the column-eigenvectors are time-invariant: substitute (2.1) into (2.4) for the jth column-eigenvector to obtain

$$c_{i+1} = c_i \frac{i(i-1) - j(j-1)}{i(i+1)}.$$

Since this expression does not depend on t, the matrix C of column-eigenvectors is time-invariant.

The row-eigenvectors of A are defined by $rA = \lambda r$, where λ is an eigenvalue of A and r is the corresponding row-eigenvector. This equation can be re–expressed as

$$r_{i-1} = r_i(\lambda + \alpha_i)/\alpha_i, \tag{2.5}$$

and row-eigenvectors can be calculated iteratively in the same way as column-eigenvectors. Like C, the matrix R of row eigenvectors will be upper triangular and time-invariant.

Solutions to equation (2.2) are given by the vector

$$p(t) = e^{\int_0^t A(z)dz} p(0) = CP(t)\, Rp(0) \tag{2.6}$$

where P is a diagonal matrix whose xth diagonal element is

$$P_x(t) = e^{-\int_0^t \alpha_x(\tau)d\tau} \tag{2.7}$$

and $p(0) = [0, 0, \ldots, 1]$ as described for equation (2.2). The kth element of $p(t)$ contains the probability of observing k distinct lineages t generations before present when population size is described by the function $N(t)$.

Since $p(t)$ contains the probability that the genealogy has k lineages t generations before present, it also contains the probability that generation t makes a contribution to the total amount of time that the genealogy is expected to have k lineages. Therefore, the vector of the expectations of the time that a genealogy has k lineages is given by m, the sum of all $p(t)$

$$m = \int_0^\infty p(t)\, dt = CERp(0) \tag{2.8}$$

where E is a diagonal matrix whose xth diagonal element is [33, Ch. 5]

$$E_x = \int_0^\infty e^{-\int_0^\tau \alpha_x(t)dt} d\tau. \tag{2.9}$$

The kth element of m contains the expectation of the length of time that the genealogy is composed of k lineages.

2.3.2 Piecewise–Constant Population Histories

The simplest nonstationary model of population history is that of two–epoch sudden change, which assumes that population size is constant apart from a single instantaneous growth or decline (Figure 2.1a). Two–epoch sudden change histories are defined by three parameters: N_1 (ancient population size), T (time of population size change) and N_0 (recent population size), and population history can be described by the function

$$N(t) = \begin{cases} N_0 & t \leq T \\ N_1 & t > T \end{cases}$$

where small values of t represent the recent past and large values of t represent the distant past. The two–epoch model provides a good approximation to more complex histories, such as exponential growth [34].

An algorithm for handling piecewise constant population histories with an arbitrary number of epochs is to divide the population's history into C epochs within each of which $N(t)$ is assumed constant. With large C, this model can approximate any history of population size. Even with small C, it seems likely to be realistic for populations whose sizes are ordinarily held constant by density–dependent population regulation.

The vector describing the probability that there are k distinct lineages $t_i + s$ generations before present (equation 2.6) within epoch i becomes

$$p(t_i + s) = e^{A_i s} p(t_i),$$

where t_i is the time (looking backwards from the present) at which epoch i begins, and A_i is the (constant) value of matrix A throughout epoch i. In calculations the "uniformization" algorithm [35, Ch. 8] can be used. This minimizes the numerical problem of calculating the matrix exponential [36].

Equation (2.8) becomes $m = \sum_{i=0}^{C-1} F_i p(0)$, where the integral $F_i = \int_{t_i}^{t_{i+1}} \exp[\int_0^t A_i t dz] dt$ spans epoch i, and where $t_C = \infty$. The contribution from epoch i is

$$F_i p(0) = \begin{cases} A_i^{-1}(\tilde{p}(t_{i+1}/N_i) - \tilde{p}(t_i/N_i)) & i < C \\ -A_C^{-1}\tilde{p}(t_{C-1}/N_{C-1}) & i = C \end{cases}$$

where $\tilde{p}(v)$ is the probability vector obtained by projecting from p from time 0 to time v while assuming that $N = 1$.

Figure 2.1a illustrates the relationship between population history and expected interval length for three two–epoch population histories: a population increase, a constant population size and a population decline.

2.3.3 The Theoretical Frequency Spectrum of Mutations

The frequency spectrum is the distribution describing the relative abundance of alleles occurring $i = 1, 2, \ldots, n-1$ times in a sample of n homologous genes. Spectra from populations that have increased in size show an overabundance of rare variants relative to populations of constant size, but populations that have decreased show an underabundance [24, 37]. The sensitivity of the frequency spectrum to population size change is exploited in several statistical tests of stationarity or neutrality [20, 21].

A polymorphic nucleotide site is ordinarily present in only two states within a sample, one of which is ancestral and the other derived. The expected fraction, σ_k, of sites at which the derived allele occurs k times is given by

$$\sigma_k = \frac{\sum_{j=2}^{n} j m_j y(j,k,n)}{\sum_{j=2}^{n} j m_j}$$

where n is the number of DNA sequences in the sample, m_j is the expected length of the coalescent interval containing j distinct lineages, and $y(j,k,n)$ is the probability that a single lineage within coalescent interval j has k descendants in the sample.

The probability $y(j,k,n)$ is given by Polya's distribution:

$$y\,(j,k,n) = \frac{(j-1)\,(n-k-1)!\,(n-j)!}{(n-1)!\,(n-j-k+1)!}.$$

Frequency spectra for three populations are illustrated in Figure 2.1b. These spectra are consistent with simulations [37].

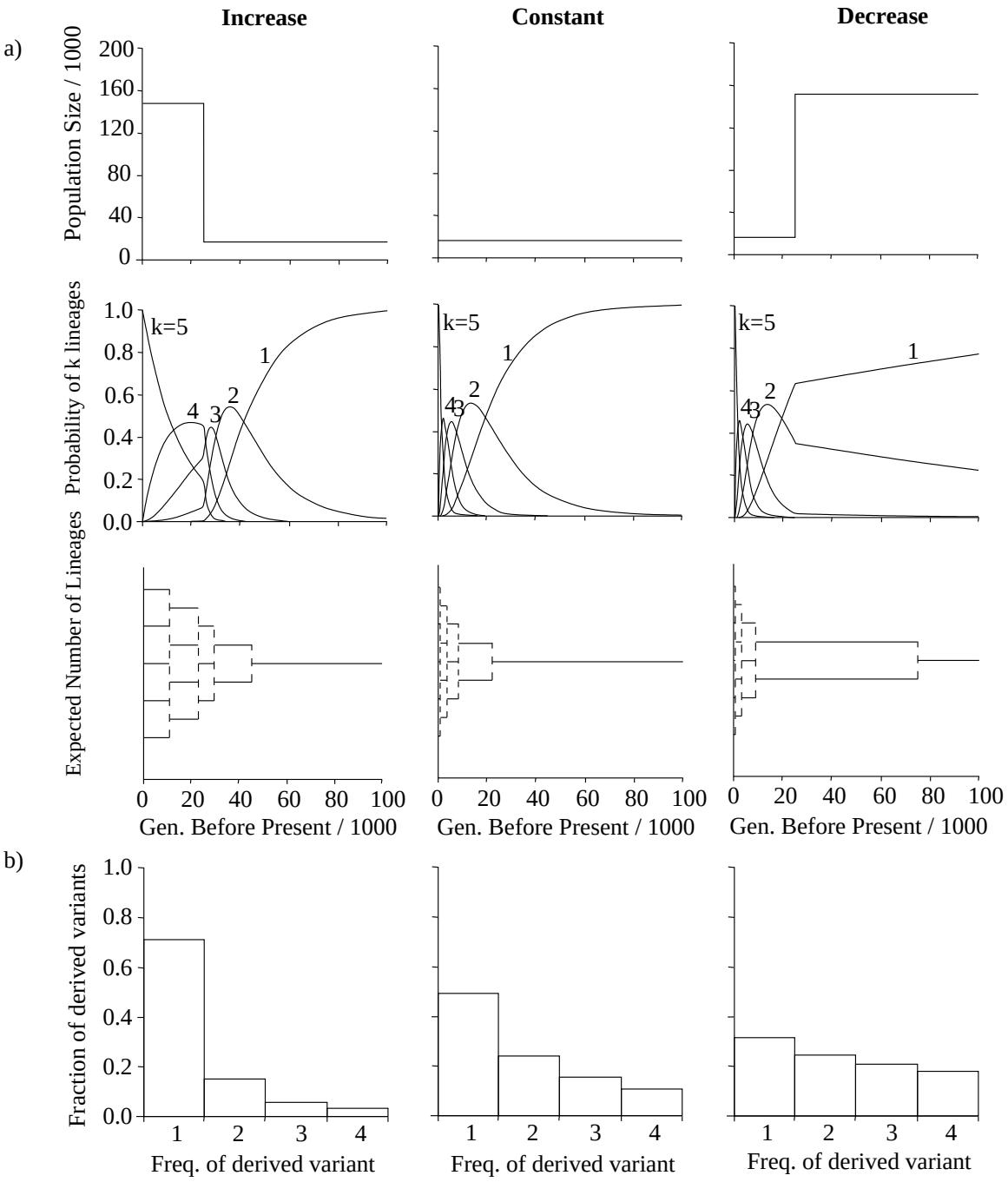

Figure 2.1. Theoretical expectations. The four panels show the relationship between population history, probability of coalescence, expected interval length and theoretical frequency spectrum under three population histories for samples of $n = 5$ lineages. a) top: population size over time; middle: the probability that there are still k distinct lineages in the genealogy t generations ago, given the population history at top; bottom: expected interval lengths given the population history at top. Dashed vertical lines indicate that no particular branching order is implied for the genealogies. b) Normalized frequency spectra for the genealogies represented in panel (a). These results were generated using the Maple 5.1 software package using default numerical precision.

2.3.4 A Likelihood Ratio Test

Under the assumption that the genealogies of unlinked sites are statistically independent, the log likelihood of an observed dataset (D) given a hypothetical population history (H) is

$$L(H) = \sum_{k=1}^{n-1} S_k \ln \sigma_k$$

where S_k is the number of sites occurring k times in the sample and σ_k is the conditional probability of a variant site occurring k times in the sample given H. If different sample sizes are used for different loci, σ_k changes from site to site. The ratios of likelihoods under different population histories can be compared using standard likelihood ratio tests [38, 39].

2.3.5 Application to SNPs

Cargill et al. [40] surveyed SNPs in 196.2 kb of nuclear DNA sequence in 20 Europeans, 14 Asians, 10 African Americans and 7 African Pygmies. Most of the sequence was from the coding portion of genes implicated in cardiovascular, endocrine and neuropsychiatric diseases, but some noncoding sequence was sequenced in flanking and intervening regions. Each amplified segment was screened by both DNA sequencing and denaturing high-performance liquid chromatography, and every putative SNP was verified by resequencing [40]. In total, 612 SNPs were identified in 106 genes.

The laboratory methodology used by Cargill et al. [40] avoided some problems like false positives, but two features of the SNP data made analysis difficult. First, the SNPs were a combination of linked and unlinked loci. Second, different SNP loci were assayed in different numbers of chromosomes. Some SNPs were sampled in 28 chromosomes, for example, whereas others were sampled in 114. Few analyses can cope with either of these problems, and in some analyses Cargill et al. [40] were forced to rely on assumptions that are known to be false. The matrix coalescent provides an alternative approach. It cannot accommodate sites with varying levels of linkage, but likelihood ratio tests can take varying sample sizes into account.

To take advantage of the informativeness of unlinked sites and to avoid the confounds associated with partial linkage, we resampled the original dataset randomly in three steps:

1. All of the SNPs reported by Cargill et al. [40] were divided into the three categories reported originally: coding nonsynonymous (cns), coding synonymous (cs) and noncoding (nc) sites near genes.

2. To minimize linkage between sampled sites, only one randomly chosen SNP from each category was scored for each reported gene. If no SNPs in a category were found in a given gene, then no SNP in that category was chosen from the gene.

3. The number of sites in each of the frequency categories reported in Cargill et al. [40] (0-5%, 5-15% and 15-50%) was tabulated for cns, cs and nc SNPs using the dbSNP database [41, 42].

A total of 60 cns loci, 68 cs loci and 30 nc loci were included in the randomized dataset, which was composed of sites from at least 19 different chromosomes. The sites within each category, which were always from different genes and often from different chromosomes, were assumed to be unlinked.

SNPs occurring k times could not be distinguished from SNPs occurring $n - k$ times for roughly half of the SNPs in the original dataset, so theoretical spectra were "folded" at frequency 0.5 in tests here, as described by Harpending et al. [24].

Likelihoods of hypotheses given the observed frequency spectra were generated for each dataset over a series of two-epoch sudden change histories. Two-epoch histories are ordinarily described using three parameters (N_0, T, and N_1), but the number of parameters describing a two-epoch history with indeterminate mutation rate is two, τ and ρ, where $\tau = \frac{T}{N_1}$, and $\rho = \frac{N_0}{N_1}$. Thus, ρ is a parameter representing the magnitude of population growth and τ is a parameter representing the time of population size change. Each parameter introduced one degree of freedom in likelihood ratio tests.

Maximum likelihood values for ρ and τ were estimated for each of the three SNP categories by iterating over the ρ-, τ-parameter space. The CLN-1.0.1 programming library was used to perform computations with a combination of high-precision (500 decimal-place) and rational numbers [43].

Five hypotheses were were tested for each SNP category. First, the maximum likelihood of each category was compared with the category's likelihood under the maximum likelihood parameters of the other two categories. Then the maximum

likelihood of each SNP category was tested against the category's likelihood under three alternatives hypotheses: a) stationarity, b) the most recent population expansion not excluded by Rogers [10] ($\tau = 4.7 \times 10^{-3}$, $\rho = 1000$) and c) the most ancient population expansion not excluded by Rogers [10] ($\tau = 2.1 \times 10^{-2}$, $\rho = 1000$) (Figure 2.3).

2.4 Results and Discussion

2.4.1 Contradictory Parameter Estimates

Cargill et al. [40] found that the frequency distribution of cs and nc SNPs differed significantly from that of cns SNPs, and that cns SNPs showed an excess of low frequency variants. Our parameter estimates confirm these results (Figure 2.2). The cns category had maximum likelihood parameters implying recent population growth under the assumption of selective neutrality ($\tau = 8.6 \times 10^{-6}$ and $\rho = 9900$), the cs and nc categories yielded estimates implying little or no change in population size ($\rho = 0.4$ for cs and 0.6 for nc).

Maximum likelihood estimates for the nc dataset were not rejected as an explanation for the cs dataset at the 0.05 level, but the maximum likelihood estimates for nc and cs datasets were rejected as an explanation for the cns dataset (Figure 2.3). Maximum likelihood estimates for the cns dataset were not rejected as an explanation for the nc dataset, but only marginally ($p < 0.08$). The nc and cs data could not be distinguished from each other, but both were distinguished from cns (Figure 2.3). In addition, the cns data showed an excess of low frequency variation relative to expectations under stationarity, as was observed previously [40].

The failure of likelihood ratio tests to distinguish between cs and nc categories is a result of their similar ρ estimates. When ρ is near 1 the time of population size change has little effect on the frequency spectrum, and confidence intervals around τ are broad. When ρ is exactly 1 they extend to infinity regardless of sample size.

Given the nearness of the nc SNPs to coding regions, the similarity of nc and cs frequency spectra is not surprising. The effects of hitchhiking on nc SNPs are likely to be strong if natural selection occurred in nearby coding regions. SNPs in genes (like cs and cns SNPs) or even near genes (like nc SNPs) are unlikely to be informative about human population history unless the effects of natural selection have been weak.

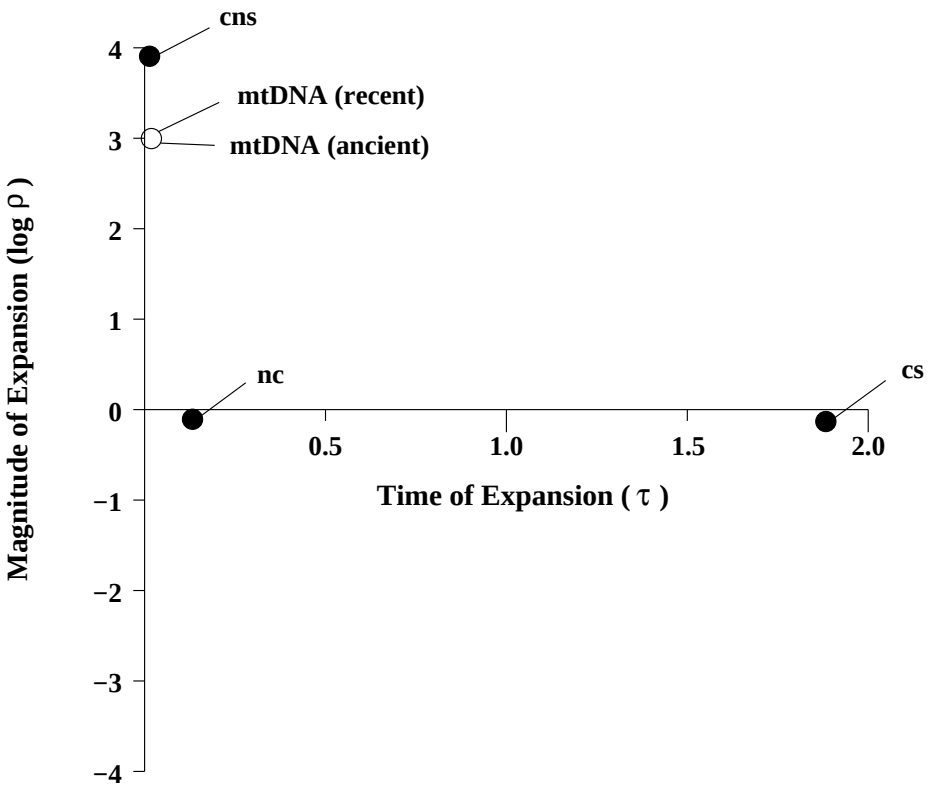

Figure 2.2. Maximum likelihood parameter estimates. Open circles show the parameters of two alternative hypotheses estimated from mitochondrial DNA (see Methods).

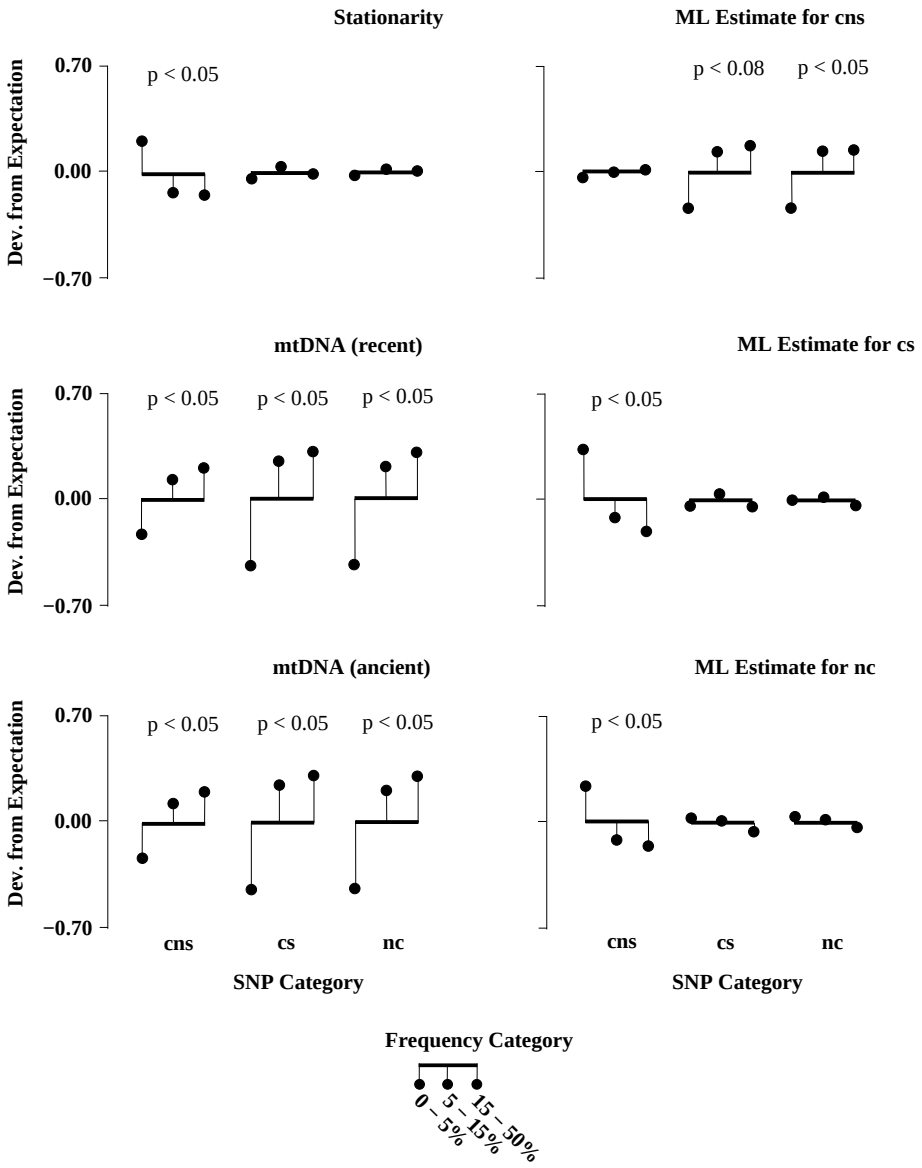

Figure 2.3. Pitch plot of deviations from expectation. Each panel shows the deviation of each frequency category (0-5%, 5-15% and 15-50%) from expectation in each SNP category (cns, cs and nc). Deviations given are the mean difference from expectation across all sample sizes within a SNP category, since not all loci were sampled in the same number of chromosomes. The hypotheses are: a) stationarity: constant population size and selective neutrality, b) mtDNA (recent): the most recent population expansion not rejected by Rogers [10] based on mtDNA polymorphism (see Methods), c) mtDNA (ancient): the most ancient population expansion not rejected by Rogers [10], d) the maximum likelihood parameters for coding nonsynonymous SNPs (cns), e) the maximum likelihood parameters for coding synonymous SNPs (cs), f) the maximum likelihood parameters for noncoding SNPs near genes (nc). The probability of the observed data given the history is indicated for hypotheses that were rejected.

If evolutionary processes in SNPs are neutral, then the three categories should be indistinguishable, yet clearly they are not. The frequency spectrum in cns SNPs differs from that of nc and cs SNPs, and none of the observed spectra is consistent with hypotheses about human population growth inferred from other loci.

2.4.2 Natural Selection

One explanation for the difference between SNP categories is that human population size has been constant, but purifying selection has affected the cns data. Purifying selection removes new nonsynonymous variants from populations so their frequencies are usually low, but it has little effect on the frequencies of linked synonymous variants [44]. This effect would be most apparent in cns SNPs because they produce amino acid changes. Whereas purifying selection explains the excess of rare variants in cns SNPs, a history of constant population size explains the failure to reject stationarity in nc and cs SNPs.

The observed pattern could also have been produced by the combined actions of purifying selection, balancing selection, and population growth. In the absence of selection, population growth produces a genealogy without deep branches. Balancing selection has the opposite effect; it may maintain two or more allelic classes for a very long time. Since balancing selection and population growth affect genealogies in opposite ways, each tends to obscure the effect of the other. These countervailing effects, however, are not reflected equally in our three categories of data. Many mutations will occur on the long branches that separate allelic classes, but only the neutral mutations will survive long. Consequently, these long branches will contribute mainly to the SNPs in our cs and nc categories. This is of interest because mutations that occur on the deepest branches of the genealogy are more likely to have intermediate frequencies (i.e., far from 0 or 1) in the modern population. Thus, balancing selection inflates the count of loci with intermediate frequencies, but this effect is visible mainly in the the cs and nc categories. Since mutations on deep branches contribute less to the cns category, balancing selection is less likely to obscure the effect of population growth there. Thus, cns SNPs are more likely to show the elevated count of loci with extreme frequencies (near 0 or 1) that one associates

with a population expansion. The count of extreme-frequency cns SNPs should be additionally elevated by recent deleterious mutations that have not yet been removed by purifying selection. For both reasons, the count of extreme-frequency loci should be larger among cns SNPs than among cs or nc SNPs. This is exactly the pattern that we observe.

A criticism of the hypothesis of pervasive balancing selection is that it requires the explanation of potentially high genetic loads. For example, if balancing selection favors heterozygotes then large segregational loads will be present due to the presence of low-fitness individuals with many homozygous loci ([45]. This problem was addressed by Harpending and Rogers [1] who asked how many overdominant loci can be maintained by a given amount of genetic variance in fitness. They showed that millions of loci can be maintained even when coefficients of variation in fitness are modest. The possibility of widespread balancing selection in the human genome remains a clear possibility.

2.4.3 Practical Implications

SNP loci with alleles at intermediate frequencies are considered more valuable for disequilibrium mapping and disease association studies than loci with alleles near fixation [40, 46]. The relative overabundance of silent variants in the SNP sample described here suggests that more loci with intermediate allele frequencies may be present than predicted previously [46]. This overabundance of intermediate-frequency SNPs due to balancing selection could extend far into noncoding regions: the extent of linkage disequilibrium varies across the genome, but it is often detected at distances as far as 50kb [47]. The pervasive presence of balancing selection in the human genome may provide practical advantages in applications like gene mapping and the analysis of complex disease.

2.5 Acknowledgments

Henry Harpending, Jon Seger, Stewart Ethier, John Hawks, Pat Corneli, David Witherspoon, Josh Cherry, Pui-Yan Kwok, Brad Demarest, and Lara Carroll provided helpful comments and discussion. SW was supported by an NIH Genome Sciences

Training Grant (Genome Informatics) to the University of Utah. AR was supported by NIH grant GM-59290 to the University of Utah.

CHAPTER 3

PROMT: INFERRING DEMOGRAPHIC HISTORY FROM DNA SEQUENCES

3.1 Abstract

Summary: *I describe a parallel implementation of Rogers's mismatch algorithm, a method for making inferences about demographic history from DNA sequence data. The program is distributed on clusters of workstations, providing a substantial speedup and low execution times on large numbers of nodes.*

Availability: *Source code and documentation are available at http://mombasa.anthro.utah.edu/wooding/*

Contact: *stephen.wooding@anthro.utah.edu*

3.2 Methods, Results and Discussion

A common method for characterizing patterns of genetic variability in homologous DNA segments gathered from populations is to calculate the distribution of pairwise nucleotide differences between them — the mismatch distribution. That is, for a set of sequences, it is informative to examine the frequency distribution of pairs that differ at $i = 0, 1, 2, \ldots, k$ nucleotide positions. Simulations and mathematical models predict that mismatch distributions will be sensitive to both the timing and magnitude of population growth.

Rogers [10] developed a theoretical framework to allow the extraction of demographic information from extant patterns of genetic diversity, as well as an accompanying computer program for analyzing data, mmci. Rogers' approach constructs a confidence interval by subjecting a sample dataset to many comparisons, each constituting a separate statistical analysis. This tool has been informative in a number of contexts [10, 16]; however its use has been limited partly due to its high computational demands. To improve the speed of execution of the mismatch test,

and hence its power and flexibility, I have developed a parallel implementation of the algorithm called PRoMT (Parallel Rogers Mismatch Test).

PRoMT is a modification of Rogers's original algorithm that uses message passing to distribute the hypothesis testing process over an arbitrary number of networked workstations. The program is scalable and portable. Though designed and tested on a pair of desktop computers, it runs without modification on clusters with more complex organizations including 4- and 32-node systems. PRoMT's performance increases the rate at which Mismatch's subtests can be performed, improving the resolution and scope of the test as a whole while maintaining its availability to a broad spectrum of users.

PRoMT was adapted directly from the source code of mmci, Rogers' program for performing mismatch tests packaged with Mismatch version 4.2 [48]. Changes to mmci were minimized to allow future modifications to mmci to be incorporated easily into the PRoMT parallel framework. Message passing was accomplished using the MPICH 1.1 implementation of MPI (Message Passing Interface), a free, standardized interface for initiating and controlling processes on several machines simultaneously [49, 50].

The organization of Mismatch into a series of discrete statistical tests made parallelization straightforward. The task consisted mainly of orchestrating the distribution of small elements of the larger test to slave processes, and collating results at the end. Finally, PRoMT was compiled using the tools included with the MPICH distribution and the GNU C compiler (gcc).

PRoMT's performance was measured using wall clock run times, which give an indication of the program's practical usage. Speedup, which was calculated by dividing the wall clock run time of Rogers' unmodified mmci program on one of a cluster's slave nodes by the wall clock run time of PRoMT on a varying number of slave nodes, was used to measure the performance of PRoMT relative to mmci. Benchmarking measurements were made on a 32-node cluster composed of 350 MHz Pentium II computers. The example dataset included with Mismatch [48] was used in all benchmarks.

Speedup achieved with increasing cluster size was substantial. The ratio of speedup

to number of nodes was approximately 0.85 (Figure 3.1). This pattern is consistent with PRoMT's effective load distribution and the presence of a low communications overhead. The departure of PRoMT's speedup from linearity appears to be due to startup, which includes the addition of an approximately constant amount of time for each participating node.

The total execution time of PRoMT was reduced substantially by increasing node numbers ($6\frac{1}{2}$ minutes on 32 nodes versus $2\frac{1}{2}$ hours on one node). This level of performance is indicative of the program's usefulness: with a modest outlay of network and node preparation a high level of performance can be achieved. In actual usage, the program has provided a number of benefits. In addition to increases in test speed, the program has allowed an increase in project turnaround time. Adjustments to parameters used in individual analyses, as well as comparisons between different datasets, can be performed quickly to provide an improvement in result quality as well as speed.

3.3 Acknowledgments

The author wishes to thank Alan Rogers, Henry Harpending and Brian Haymore for helpful comments. This work was supported by a NIH Genome Sciences training grant and a grant of computer time by the Center for High Performance Computing at the University of Utah.

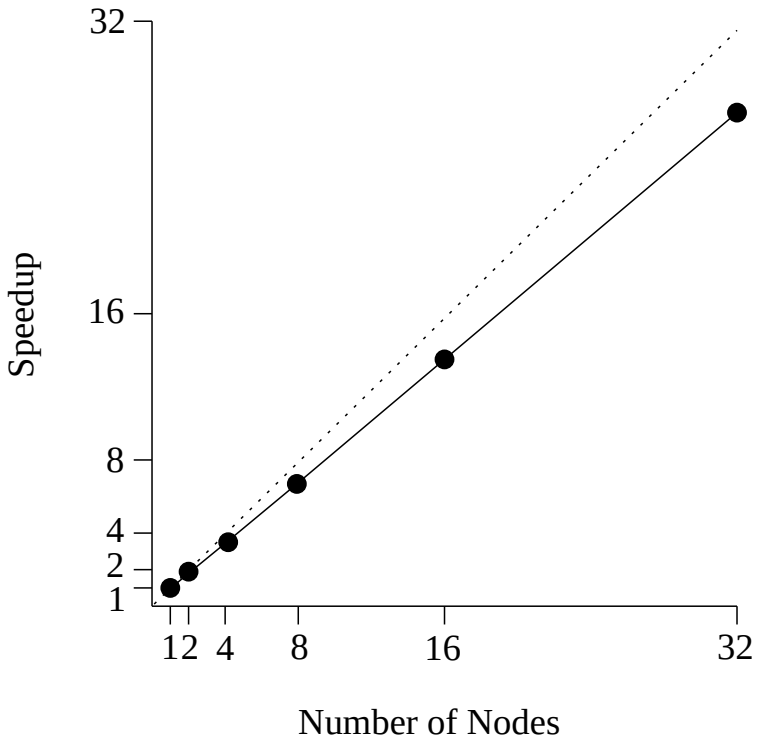

Figure 3.1. Relationship between cluster size and speedup in PRoMT on a 32-node cluster. Speedup was measured by dividing the run time of Rogers's unmodified mmci program running on a single node by the run time of PRoMT running on n nodes.

CHAPTER 4

A PLEISTOCENE POPULATION X-PLOSION?

In two recent papers, Kaessmann et al. presented DNA sequence data from the X chromosome (Xq13.3) of 30 chimpanzees and 69 humans [51, 52]. These data bear on two longstanding questions involving late Pleistocene demographic history: 1) whether the long-term demography of humans is characterized by explosive population growth, and 2) whether chimps show population growth coinciding temporally with growth in humans.

Some genetic loci, such as mitochondrial DNA sequences [10], autosomal microsatellites [11], and microsatellites on the Y chromosome [15] have been taken as support for an explosion in human population size; other data have not [28]. However, most of the data showing no evidence of a population increase have been gathered from nuclear sequences of functional importance, making them vulnerable to the effects of natural selection [51]. Xq13.3 provides the first substantial noncoding nuclear genetic sequence data useful for the evaluation of historical demography. The additional chimp data are pertinent to discussions of whether human demographic expansion was caused by generalized historical events affecting many species, such as global climate changes, or by human-specific factors such as technological innovations.

In their analysis of Xq13.3 diversity in humans, Kaessmann et al. found that a test of Tajima's D [20], which compares the total number of variable nucleotide positions in a sample of sequences with the mean pairwise difference, fails to reject the hypothesis of constant population size. However, two lines of evidence suggest that the data are consistent with a population expansion.

First, the application of Fu's F_s test [21], which compares the relative abundance of alleles at different frequencies in a sample and is more sensitive to population growth than D, yields a value of -7.570, rejecting the hypothesis of demographic

stationarity at the 5% significance level. The negative value of F_s indicates that more low-frequency variants are observed than are expected—a deviation reflected in a graphical comparison of observed data to theoretical expectations under constant size (Figure 4.1).

Second, the application of Rogers's method for estimating historical demographic parameters, which analyzes the distribution of pairwise nucleotide differences among sampled sequences, yields $\widehat{\theta_0} = 0.646$, and $\hat{\tau} = 2.423$ [10]. Under a two-epoch "sudden change" model of population history, which makes the simple assumption that a population changed instantaneously from an ancient population size to the modern one, $\widehat{\theta_0} = 2N_0\mu$ provides an estimate of ancient population size N_0 (measured as the number of X chromosomes) and $\hat{\tau} = 2\mu t$ provides an estimate of the length of time t (measured in generations) that has passed since the size change occurred [10]. N_0 and t can be obtained by letting μ equal the product of the nucleotide substitution rate (per site per year), the number of nucleotides in the sequence and the number of years per generation. Under the nucleotide substitution rate of 10^{-9} per site per year proposed by Kaessmann et al. [51], a sequence length of 10,000 nucleotides and a twenty year generation time, $\widehat{\theta_0}$, and $\hat{\tau}$ point to a rise in human populations from an initial effective size of around 1,600 X chromosomes (roughly 1,000 people), approximately 120,000 years before present. The confidence interval around both of these estimates is likely to be broad.

Xq13.3 sequences in chimps are more ambiguous. Fs $= -6.637$ for all chimps together, and -3.671 for West African chimps alone, failing marginally to reject stationarity. Neither constant size nor growth can be excluded.

In summary, Xq13.3 data from chimps are complex and ambiguous, but they do not appear to show evidence of a major Pleistocene expansion. However, F_s rejects the hypothesis that human population size has been constant, and $\widehat{\theta_0}$, and $\hat{\tau}$ are in general agreement with earlier findings of population growth based on autosomal microsatellites, microsatellites on the Y chromosome and mitochondrial DNA [11, 15, 10]. Xq13.3 sequences can be added to the list of loci supporting a late Pleistocene population explosion in humans.

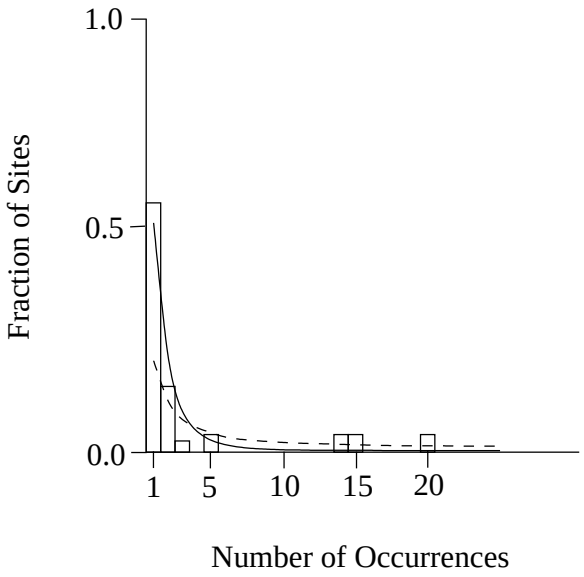

Figure 4.1. Frequency spectrum of mutations in human Xq13.3 sequences. Dashed line represents theoretical expectation under the hypothesis of no growth calculated using the method outlined in chapter 1. Solid line represents theoretical expectation under the hypothesis of 1,000-fold growth 120,000 years ago.

CHAPTER 5

POPULATION STRUCTURE AND HISTORY IN EAST ASIA

5.1 Abstract

Archaeological, anatomical, linguistic, and genetic data have suggested that there is an old and significant boundary between the populations of north and south China. We use three human genetic marker systems and one human-carried virus to examine the north-south distinction. We find no support for a major north/south division in these markers: rather the marker patterns suggest simple isolation by distance from south to north.

5.2 Introduction

Archaeological and genetic evidence has suggested that the human population experienced a dramatic expansion in the last 100,000 years, spreading rapidly to its current worldwide area of occupation. Much of the area inhabited by humans must have been reached via migration through East Asia. There is evidence of human occupation through central Asia to Beringia during the later Pleistocene [53, 54], and parallel evidence of early humans in Australia and southern Asia. One natural hypothesis is that the population of East Asia is a result of ancient contact and mixing between these northern and southern pincers of the modern human expansion. Alternatively, East Asia may have been reached by either a northern or southern route, followed by dispersal into nearby areas.

In separate analyses, patterns of dentition, archaeological assemblage composition, linguistics, familial surnames, and various low-resolution genetic systems have identified systematic differences between northern and southern groups [55, 56, 57, 58, 59, 60] (see Cavalli-Sforza et al. [61] for a review). These lines of evidence have been taken as support for a strong north/south distinction that would appear to support

the pincer model of the origin of East Asians. However, two recent high-resolution approaches to East Asian genetic diversity have come to different conclusions.

Chu et al. [62] and Su et al. [63] examined nuclear microsatellites and Y chromosome nucleotide polymorphisms respectively. Consistent with earlier work, patterns of diversity in microsatellite loci were found to fall into northern and southern clusters, with northern groups being polyphyletic. Y chromosome polymorphisms found in northern populations were a perfect subset of those in the south—every Y-chromosome haplotype observed in northern groups was observed in at least one southern group, but not every lineage observed in southern groups was observed in a northern group. Based on these findings, Chu et al. and Su et al. argued that northern East Asian populations are derived from southern East Asian populations.

In this paper we combine new evidence from mtDNA polymorphisms and five short tandem repeat (STR) loci, and previously published evidence from Y-chromosome polymorphisms, to reexamine the hypothesis that northern and southern China are distinct. We apply the same method to all loci, providing a simple basis for comparison, and contrast patterns of diversity in these markers with patterns in the distribution of JC virus, an asymptomatic urinary tract virus that is frequently transmitted from parent to child and may provide information about human migrations [64, 65].

5.3 Materials and Methods

5.3.1 Data

A dataset composed of four different marker types was assembled.

First, 473 comparable mtDNA RFLP profiles was collected. High-resolution restriction endonuclease mapping of 113 individuals from four ethnic groups in southwest China (Bai, Dai, Lisu and Yi) was performed according to the protocols of Torroni et al. [66], and analogous profiles from 360 individuals were collected from previously published papers to provide a basis for comparison (Table 5.1, Figure 5.1) . Data assembled from the literature included Ewenki [67], Korean [68], Malay Chinese [68], Nivikh [67], Taiwan Han [68], Tibetan [69], Udegey [67], Malay [68], Malay Aborigine [68], Sabah Aborigine [68] and Vietnamese [68] (Figure 5.1).

Second, five STR loci (12nt repeats in the Exo I region of Dopamine Receptor 4; 48nt repeats in the Exo III region of Dopamine Receptor 4; 120nt repeats in the

Table 5.1. Each column corresponds to one locus type (mtDNA, STR, Y-Chromosome or JC Virus) and each row corresponds to one population. Populations are designated by the locus abbreviation (M, S, Y, or J) and the ID number at left. For example the JCV sample from Seoul is designated J5. Sample sizes are in parentheses at the right side of each column.

	Marker			
	STR (n)	**mtDNA (n)**	**JCV (n)**	**Y−Chrom. (n)**
	s	**m**	**j**	**y**
Northern	Yi (92)	Lisu (32)	Beijing (10)	Buryat (4)
	Hani (56)	Yi (31)	Harbin (6)	Ewenki (8)
	Nu (18)	Tibetan (54)	Ishikawa (11)	Manchurian (18)
	Pumi (22)	Ewenki (51)	Okinawa (11)	Mongolian (24)
	Uygur (66)	Udegey (45)	Seoul (14)	Korean (7)
	Tibetan (36)	Nivikh (57)	Shenyang/Jinzhou (7)	Japanese (29)
	Liaoning Han (134)	Korean (13)	Tokyo (14)	Hui (20)
	Qingdao Han (70)	Taiwan Han (20)	Ulaanbataar (12)	Tibetan (8)
	Lisu (40)	Malay Chinese (14)	Chengdu (10)	Northern Han (82)
Southern	Buyi (56)	Bai (24)	Chiang Mai (11)	Southern Han (280)
	Dai (44)	Vietnamese (28)	Guangzhou (13)	Jingpo (5)
	Guangdong Han (86)	Sabah Aborigine (32)	Jakarta (17)	Tujia (10)
	Hong Kong Han (68)	Malay Aborigine (32)	Masai (14)	Yao Nandan (10)
	Wa (28)	Malay (14)	Pamalican Is. (8)	Yao Jinxiu (10)
		Dai (26)	Taipei (9)	Zhuang (28)
			Wuhan (10)	Dong (10)
			Yangon (15	Bulang (5)
				Lahu (5)
				Yi (14)
				She (11)
				Atayal (24)
				Yami (8)
				Paiwan (11)
				Ami (6)
				Li (11)
				Cambodian (26)
				Northeastern Thai (20)
				Malaysian (13)
				Batak (18)
				Javanese (11)

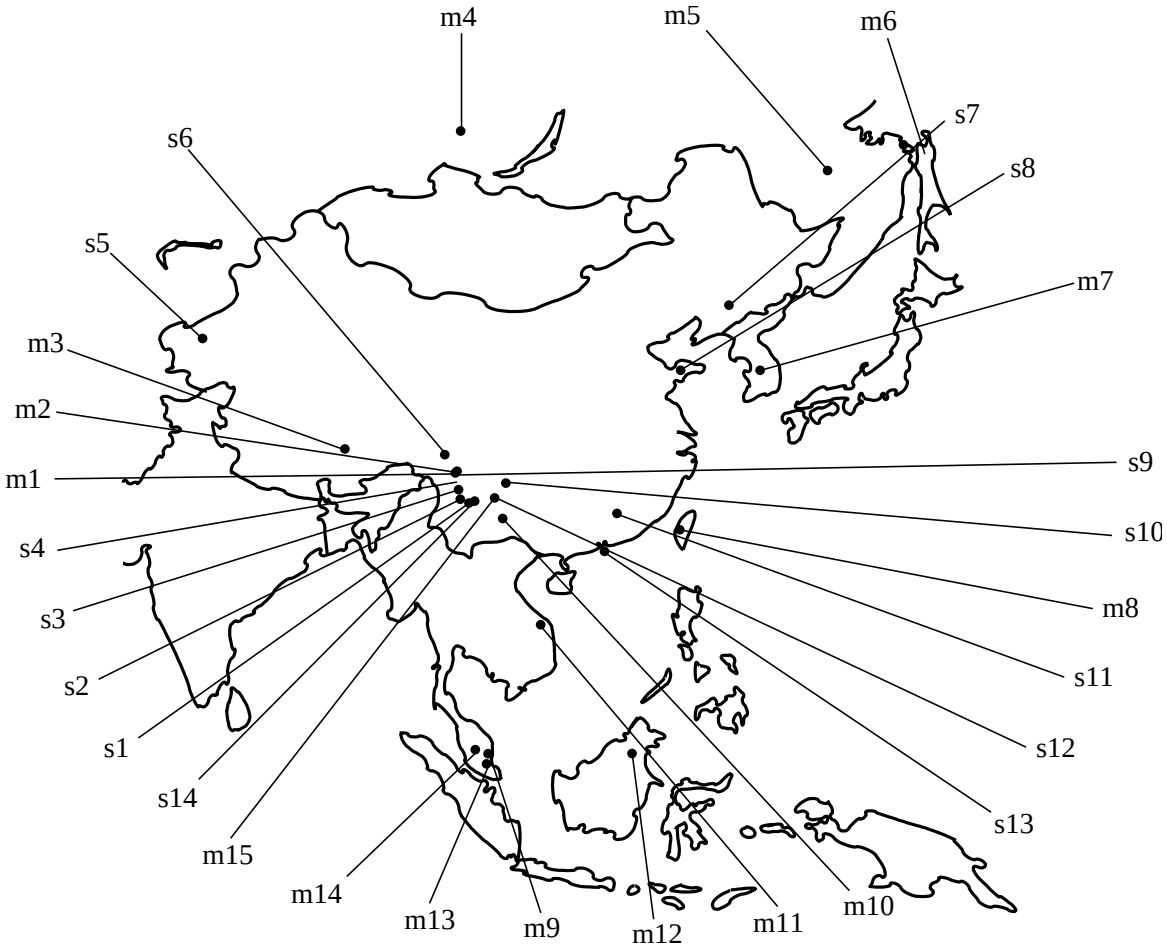

Figure 5.1. Points are populations sampled in surveys of mtDNA and STR loci. Data for other loci were assembled entirely from earlier studies (see Materials and Methods).

Pro region of Dopamine Receptor 4; 47nt repeats in the 7q subtelomere region; 49nt repeats in the 7q subtelomere region) were scored for repeat number in a total of 900 haploid genomes from 14 populations (Table 5.1).

Finally, 192 JC virus DNA sequences [65] from 17 populations and biallelic marker data from 836 Y chromosomes from 30 populations [63] were obtained from the literature (Table 5.1). Data on length polymorphisms in the Y-Chromosome were not included.

Populations were divided into northern and southern groups based on a consensus of geography and documented history. For example, historical documents indicate that among the groups sampled in southwest China, the Bai and Dai are originally from southeastern areas and the Yi and Lisu are from areas farther north, so the populations were divided accordingly, although all of them were sampled in the Yunnan province [70].

5.3.2 Analysis

We examined patterns of regional association by computing principal components of the population gene differences and plotting the first two principal components [71]. The pictures are the best, in the sense of least squares, two-dimensional representation of genetic distances among populations, where the squared distance between population x and y is the sum over all sites of

$$d_{xy} = \frac{(p_x - p_y)^2}{p(1-p)}$$

Here p_x and p_y are the frequencies of a genetic variant in populations x and y and p is the overall mean frequency of the variant.

Variation in mtDNA or the Y chromosome is often studied by computing an estimate of the whole genealogy of a sample of genes, a gene tree, and then interpreting the tree in terms of geography or population phylogeny. Details from these reconstructions may lead to appealing interpretations, but often little is known about statistical support for the interpretations. Gene trees are embedded in population histories but it is not so clear how to read the population history from the gene tree, nor how to predict a gene tree from a postulated population history. Instead we use

principal components, in an exploratory spirit, as a simple visual summary of patterns of population difference.

5.4 Results and Discussion

Recent investigations using molecular markers to study patterns of genetic diversity in East Asia have tried to address two main questions. First, are northern and southern East Asian populations genetically distinct? Second, are northern and southern East Asian populations descendants of the same, ancestral population or are they descended from different populations?

Chu et al. inferred a distinction between southern and northern Chinese populations, and a southern origin for northerners, by analyzing phylogenetic trees of populations constructed using microsatellite data [62]. Among trees constructed using the neighbor-joining method with bootstrapping, Chu et al. identified a "clear distinction between southern and northern Chinese populations" based on the presence of a paraphyletic northern group and an almost monophyletic southern clade. Su et al. inferred support for regional clusters based on an analysis of principal components of Y-chromosomal diversity and the observation of substantial lineage sharing among regions, also suggesting the possibility of a southern origin [63]. In a map of the first two principal components of variance, northern populations clustered together and southern populations clustered together. In contrast, we find support for neither a strong regional distinction nor a southern origin of northeast Asian populations.

Four panels in Figure 5.2 show the results from our four marker systems. In these maps, the first principal component accounted for 11%, 34%, 33% and 32% of the total dispersal in observed in mtDNA, STR, Y-chromosome and JCV respectively. The second principal component accounted for 10%, 19%, 23% and 14% of the variance in observed in mtDNA, STR, Y-chromosome and JCV respectively. These maps suggest an alternative explanation for diversity patterns in East Asia. Three features of the principal components maps in Figure 5.2 are most important.

First, Y and JC virus polymorphisms are geographically structured so that the principal components give a good portrayal of the underlying genetic distances: the first two principal components in these plots account for a substantial portion ($> 45\%$)

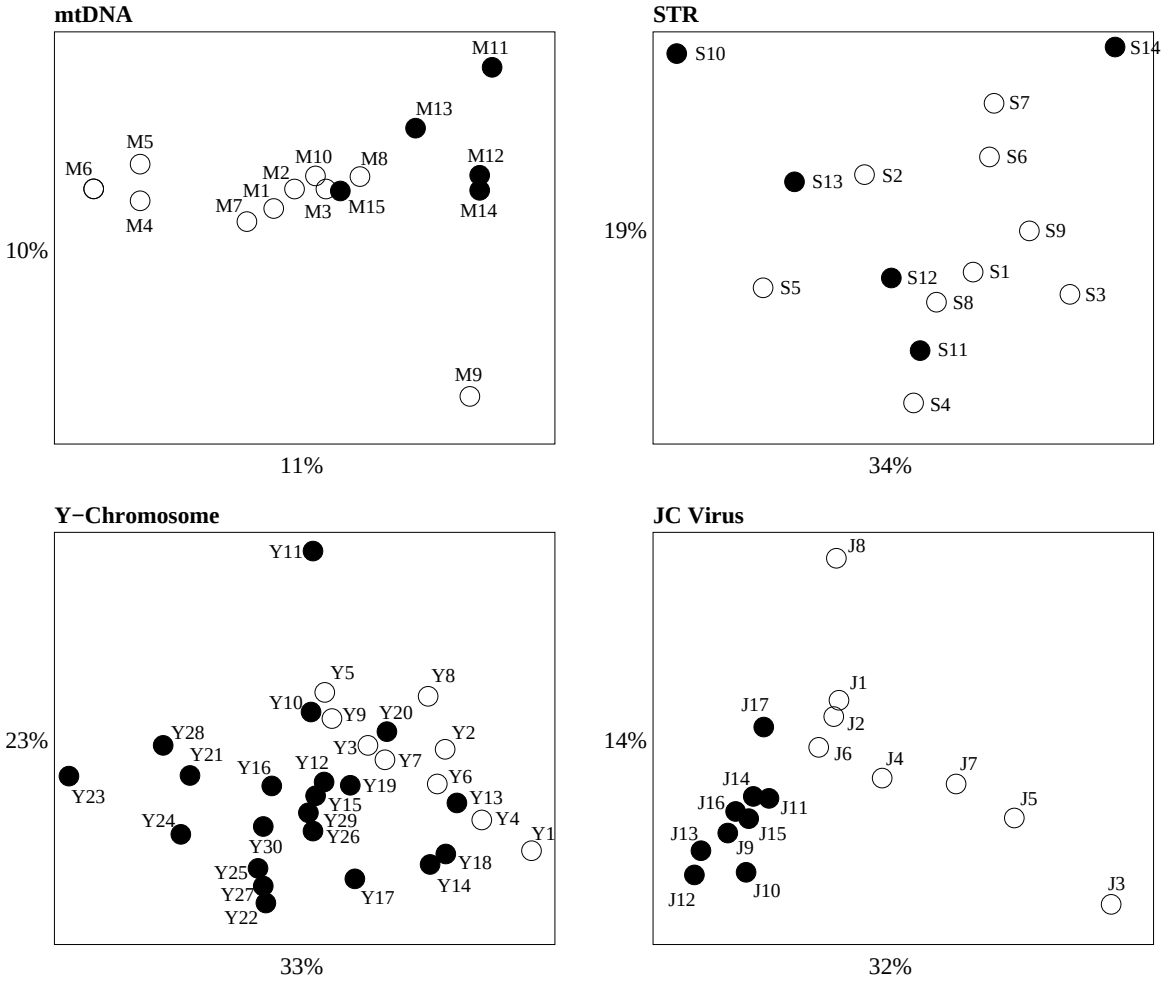

Figure 5.2. On each map, the x axis is the first principal component and the y axis is the second principal component. Northern populations are indicated by open circles and southern populations are indicated by closed circles. Axis labels indicate the percentage of variance accounted for by each component.

of the observed variance. And in spite of being drawn from just two genomic regions, more than 50 percent of the variance in the STRs was accounted for by the first two principal components—a pattern attributable to the large number of alleles. There is, on the other hand, almost no structure in the mtDNA differences among regions: each of the principal components accounts for roughly the same amount of variance, and the genetic distances among populations from mtDNA describe something like a high dimension sphere. Even though it is possible to generate an mtDNA phylogeny from the data, any such phylogeny could not reveal much of interest about population history here.

Second, although northern and southern populations generally fall into different regions of the principal components maps, the clusters are not distinct. For example in the map of diversity in mtDNA, a number of southern populations such as Bai and Dai are much more similar to other northern populations than they are to other southerners such as the Vietnamese or Malay Aborigines. The putative northern and southern clusters blend across a cline. There is no abrupt change.

Third, populations sampled from adjacent geographical areas tend to be near each other on the principal components maps. This finding is consistent with previously published evidence of genetic isolation by distance in China [72], and explains the lack of clustering in the principal components maps. Since populations are isolated by distance, and since they were sampled on a predominantly north/south axis, the principal components maps can be segregated into northern and southern groups delineated by any latitude. This feature explains the gradual, rather than sharp, divide across the arbitrarily chosen boundary between northern and southern populations.

The hypothesis that the differences between northern and southern East Asian populations are greater than expected under isolation by distance is difficult to test. Standard statistical tests for regional subdivision cannot take isolation by distance into account [73, 74]. New methods will be required. Nonetheless, close affinities between many northern and southern populations are evidence of extensive regional mixing. We assert that the genetic distance between northern and southern East Asian populations is no greater than expected by chance, given genetic isolation by distance.

The geographical organization of the principal components maps raises questions about the informativeness of human population "phylogenies." If populations that are isolated by distance are sampled along an axis, is artificial phylogenetic signal introduced? One possible explanation for the repeated identification of northern and southern clades in East Asia is that oversampling along a north/south axis has spuriously influenced phylogenetic inferences. Such a bias would be consistent with the low bootstrap values supporting northern and southern clades observed by Chu et al. [62].

It is of some interest that the clearest north/south distinction among the four principal components maps in Figure 5.2 is observed in JC virus. The JC virus is a nonpathogenic urinary tract virus that is thought to be largely vertically transmitted, and as such it might be a replicate of mtDNA. However, the true extent of horizontal transmission in JCV is unknown. Diversity patterns in JCV have been interpreted as evidence for a north/south distinction previously [75], and the fact that the virus displays a stronger regional distinctiveness than the human genes, and that the level of isolation by distance is different among viral subpopulations than among their human hosts.

The lack of regional clusters brings inferences about directional migration into question, as well. Su et al. suggest that the presence of every northern lineage in at least one southern population implies northward movement, but it is unreasonable to conclude that the northern population derived from the southern populations based on this evidence. One attractive alternative explanation for the asymmetry in lineage sharing is that northern and southern East Asian groups have had a long history of separation, but many lineages have migrated from north to south recently. Such asymmetric migration could easily generate regional differences in genetic diversity. However this explanation would predict a clear genetic difference between northern and southern groups. Another explanation for the asymmetry in lineage sharing is suggested by regional differences in demography. Whereas southern populations reside mostly in high-density areas, northern areas are sparsely populated [76]. Between-region migration, accompanied by high rates of genetic drift and lineage loss in northern groups, could account for an asymmetry in lineage composition

without causing appreciable between-region divergence.

The potentially important role of Central Asia in questions about the genetic composition of East Asia is emphasized by patterns of mtDNA diversity in the region. In a comparison of mtDNA sequences from Europe, Central Asia and East Asia, Comas et al. [77] found a closer affiliation between Mongols and Talas Kirghiz populations than between the Talas Kirghiz and Sarytash Kirghiz. No southern East Asians were included in Comas et al.'s study, and their relationships to Central Asians are unknown. However the similarity of some northern East Asian populations to Central Asians indicates that the large migrations associated with trade along the Silk Road and during later times may have had an influence on diversity in the far East. The inclusion of Central Asian samples in future studies including both northern and southern East Asians will be important in answering more detailed questions about East Asian origins.

The existence of a genetic distinction between northern and southern East Asia is not well supported. Patterns of genetic diversity in the area are more consistent with the notion that local gene flow since the end of Pleistocene have erased old human population differences over much of the world at neutral marker loci [78], and that much of the differentiation in the region is attributable to simple isolation by distance [72]. Such erasure is expected to happen even if migration rates are relatively low [79]. The lack of strong north/south differentiation in East Asia suggests that many of the anthropological trends previously held to define pervasive regional distinction are strictly cultural phenomena with no implications for genetic differentiation. This finding is itself interesting—regional cultural trends in East Asia seem to have persisted for long time periods in spite of evident genetic continuity.

5.5 Acknowledgments

We are indebted to Dr. Antonio Torroni for his detailed explanation of both experimental procedures and statistical methods. We thank Dr. Wen Wang and Long Nie for their intellectual contribution and support in sample collection. We also thank Xue-Mei Lu and Jing Luo for their comments on an early version of thispaper. This

work was supported by the Chinese Academy of Sciences, Natural Science Foundation of Yunnan Province and NSFC.

CHAPTER 6

DO HUMAN AND JC VIRUS GENES SHOW EVIDENCE OF HOST-PARASITE CO-DEMOGRAPHY?

6.1 Abstract

Information about similarities and differences in the demographic history of host and parasite populations is potentially useful for making inferences about a variety of evolutionary processes. However, it is difficult to observe the historical demographic properties of natural populations directly. Here, the extent of demographic similarity in a host and its parasite was examined indirectly by inferring long-term population history from patterns of genetic variation. Nucleotide sequence diversity in human and JC virus (JCV) DNA is consistent with a long-term demographic connection between the two species: both show evidence of large-scale population expansion. However, genetic data also suggest that the two species have different patterns of population substructure. These similarities and differences have implications for adaptive evolution in JCV that are not evident when the two species are considered separately.

6.2 Introduction

Host-parasite interactions cause correlations between host and parasite populations that can be used to learn about evolutionary processes [80]. For example, studies of evolutionary "arms races" use correlations between host and parasite adaptations to learn about evolutionary trajectories [81], and phylogenetic studies can use correlations in the long-term population history of hosts and parasites to make inferences about evolutionary rates [82].

Models of population dynamics play an important role in the study of host-parasite interactions by describing their demographic consequences [83, 84]. Under assumptions about parasite transmission rate, virulence, host reproduction rate and other parameters, population dynamic models make predictions about the properties of host-parasite interactions that can be compared against natural observations [85]. Nonetheless, empirical studies of host-parasite population dynamics are uncommon [86]. One problem is that evolutionary processes are difficult to observe directly on short time scales [87].

An alternative to the use of direct observation in studies of population dynamics is the use of molecular genetic data to make inferences. Population genetic analyses can use existing patterns of genetic variation to infer historical processes such as population size, population subdivision and natural selection [10, 16, 88]. It should be possible to test hypotheses about specific host-parasite relationships by using the tools of population genetics to look for demographic correlations involving host-parasite pairs. I present here a population genetic analysis of the parasite JC virus (JCV) and its obligate host, *Homo sapiens*.

JCV is a small, circular DNA virus that resides with little pathological effect in the urinary tract of seventy to ninety percent of adults worldwide [89]. JCV is transmitted horizontally, but it is passed most often from parent to child—an effectively vertical pattern of inheritance—and established infections persist in the host for extended periods [64].

Evidence that the pathogenicity of JCV is low and that its pattern of transmission is largely vertical have led to the suggestion that the virus may be a good surrogate for human samples in studies of migration and population affiliation [65, 75, 90, 91, 92]. This suggestion is supported by some empirical findings, which show that JCV and human populations do share certain demographic properties. The relationship between phylogenetic relatedness and geographical distribution is similar in the two species, for example [65, 75]. The extent to which patterns of genetic variation in JCV are representative of patterns in humans is still questionable, however. Superficial similarities in phylogeny and geographical distribution might be observed even if deeper differences are present. Factors like isolation by distance could lead to similarities

on a broad scale even if factors like population subdivision or natural selection cause between-population divergence.

To investigate the extent to which human and JC virus population histories are similar, I compared DNA sequences from Africa, Asia and Europe with a focus on their implications for demography. Human population history has been studied extensively and genetic diversity in humans has several distinctive features that provide a good basis for comparison. Under the null hypothesis that JCV and human demographies have historically been identical, JCV should show evidence of the same distinctive features as humans. Departures from similarity are potentially informative about the relevance of underlying factors such as natural selection and regional differentiation.

6.3 Materials and Methods

6.3.1 Data

A sample of 379 nucleotide sequences from the V-T intergenic region of JCV was assembled from Sugimoto et al. [65]). The sequences were from 42 populations on three continents (Table 6.1, and see Figure 1 in Sugimoto et al. [65]. The sequences were 612 nucleotides long and spanned portions of two coding regions (capsid VP1 and Large T antigen) and a noncoding segment (intergenic region) of the JCV genome [89].

Sequence variation in the sample was composed of both nucleotide substitutions and insertions/deletions. One sequence from Italy (accession number AB004477 in Genbank; type IT-4 in Sugimoto et al. [65] contained a long sequence gap of unknown origin, so it was removed from all analyses, leaving a total of 378 nucleotide sequences, with 11 from Italy, in this paper.

Sequences were aligned by eye, since the overall differences between them were low, and were then partitioned into two main datasets. Amino acid sequences were inferred for the coding portions of the dataset using the translation code designated by Genbank (standard). DNA sequences were modified by removing four nucleotide positions at which an insertion or deletion was observed, leaving a total of 608 nucleotide positions.

Computer files containing the JCV datasets used in this paper are available from the author.

Table 6.1. Sampled locations. Locations, symbols and sample sizes of data assembled from Sugimoto et al. [65]. (* only 11 of 12 sequences from Italy reported by Sugimoto et al. were used in this study. Genbank ID AB004477, type IT-4, was excluded because it has a long gap of unknown origin.)

	Location	Symbol	n
Africa	Accra, Ghana	GH	4
	Addas Abbaba, Ethiopia	ET	8
	Bangui, Central African Republic	CA	11
	Fes/Ifane, Morocco	MR	21
	Khartoum, Sudan	SU	9
	Lusaka, Zambia	ZA	5
	Nairobi, Kenya	KE	8
	Nouakchott, Mauritania	MA	10
	Port Louis, Mauritius	MU	8
	Tessaoua, Niger	NG	8
	Welkom, South Africa	SO	6
Asia	Ankara, Turkey	TU	15
	Beijing, China	CB	10
	Chengdu, China	CD	10
	Chiang Mai, Thailand	TL	11
	Colombo, Sri Lanka	SL	5
	Guangzhou, China	GZ	13
	Harbin, China	HB	6
	Ishikawa, Japan	IK	11
	Jakarta, Indonesia	ID	17
	Masai, Malaysia	ML	14
	Okinawa, Japan	ON	11
	Pamalican Is., Phillipines	PH	8
	Riyadh, Saudi Arabia	SA	20
	Seoul, South Korea	SK	14
	Shenyang/Jinzhou, China	SJ	7
	Taipei, China	TP	9
	Tokyo, Japan	TY	14
	Ulaanbaatar, Mongolia	MO	12
	Varanasi, India	IN	17
	Wuhan, China	CW	10
	Yangon, Myanmar	MN	15
Europe	Athens, Greece	GR	20
	Barcelona, Spain	SP	13
	Budapest, Hungary	HU	17
	Deventer, Netherlands	N	12
	Illertissen, Germany	G	8
	London, United Kingdom	UK	6
	Novosibirsk, Russia	RS	14
	Prague, Czech Republic	CR	18
	Rome, Italy	IT	11*
	Stockholm, Sweden	SW	18

6.3.2 Analyses

Five analyses that are standard in studies of human genetic diversity were applied to patterns of diversity in JCV. These were chosen to focus on characteristics of human data that are regarded as hallmark sources of evidence about human demographic history.

First, patterns of differentiation among individual DNA sequences and amino acid sequences, and among populations of DNA sequences, were examined using principal components analysis, which gives a summary description of genetic differences [71]. Second, Tajima's D test was applied to the nucleotide sequence data to test the hypothesis that the JCV population has been stationary (i.e., variation is selectively neutral and population size has been constant) [20]. Third, Rogers's method of moments was applied to the distribution of pairwise differences among sequences (or mismatch distribution) to estimate historical demographic parameters [10, 48]. Fourth, levels of genetic diversity within continents were compared using π, the per nucleotide difference between sequences [93]. Fifth, the ratio of synonymous substitutions per synonymous nucleotide position and nonsynonymous substitutions per nonsynonymous nucleotide position (i.e., K_s/K_a ratios) were calculated using the tools of Yang [94], and analyzed in pairwise comparisons as described by Nei and Kumar [95].

The results of these analyses were compared with similar analyses of human genetic data in the published literature [1, 10, 61, 96, 97, 98].

6.4 Results and Discussion

Patterns of genetic diversity in the JC virus have been used as a surrogate for human genetic diversity in several studies of population structure and migration [65, 75]. Such studies have relied on an apparent large-scale correlation between JCV and human population structure: on average, distantly related human populations contain distantly related JCV lineages, and closely related human populations contain closely related JCV lineages [65]. However, a variety of factors could lead to superficial similarities between human and JC virus populations even if deeper differences are present. If isolation by distance occurs in both populations, for example, then

human and JCV populations might show similar patterns of regional differentiation superimposed on differences in population size trends or subdivision.

Given the rising availability of human genetic data, a more effective approach is to treat the historical relationship between human and JCV as an unknown. Human history is exceptionally well studied and information about humans should provide a good basis for comparison with other species. Evidence for low levels of pathogenicity and largely vertical patterns of transmission in JCV suggest the null hypothesis that human and JCV populations have strongly correlated demographic histories. Inferred differences may be useful for learning about long-term demographic and evolutionary processes in the virus.

Human genetic diversity in a number of genetic systems (including Y chromosome DNA sequences [30], nuclear DNA sequences [31], mitochondrial DNA polymorphisms [12] and nuclear simple tandem repeats [11]) is characterized by four hallmark trends that are relevant to the investigation of long-term demographic processes in JCV [1]. First, phylogenies of human genes tend to be comb-shaped, with most terminal branches being nearly the same length [1, 99]. Second, the distribution of pairwise differences among linked polymorphisms, or mismatch distribution, is a smooth unimodal wave in most human populations [1, 12, 30, 31]. Third, genetic differences between human populations tend to be correlated with geographical distance, and genetic differences between continents tend to be relatively uniform [1, 61, 96]. Fourth, human populations within Africa tend to be more genetically diverse than populations on other continents [97]. These patterns are consistent with Late Pleistocene population growth and range expansion originating in Africa [1].

Among surveys of genetic diversity in JCV, the sample of Sugimoto et al. [65] is most informative in a demographic context. It includes a large number of sample populations distributed over three continents and is comparable in both size and geographical distribution to a number of human genetic datasets [1, 61] (Figure 6.1). In earlier phylogeographic analyses of the sample, three main trends were identified. Phylogenetic trees relating DNA sequences showed evidence of several clades with relatively limited geographical distributions, including a distinct, basal European clade (designated Eu) (Figure 1 in Sugimoto et al. [65]). Graphical displays of

the data supported the existence of regional differences in lineage composition and relatedness [65, 75]. And populations with known histories of human migration and admixture showed evidence of migration and admixture with respect to viral diversity [100, 65].

Earlier findings in JCV are consistent with some observations based on human genes [1], but there are important differences. The tendency of human genetic lineages to group geographically and the distinctness of migrant lineages in admixed populations are well established [61]. However the presence of a distinct, basal clade of European lineages is unexpected. In human gene genealogies the basal position of lineages from Africa is often interpreted as evidence for an African origin of modern humans [61, 99, 101, 102]. The phylogenetic position of European lineages in JCV is inconsistent with the argument that JCV infected humans prior to their hypothesized expansion out of Africa.

The distinctness of the Eu clade recognized by Sugimoto et al. [65] is also unexpected. In a principal components analysis of differences among V-T intergenic region nucleotide sequences based on synonymous nucleotide substitutions only, the first two principal components of variation account for nine and five percent of the total variance respectively and divide JCV lineages into two distinct clusters (Figure 6.1). A large cluster (A) contains lineages found predominantly in Asia and Africa, and a smaller cluster (B) contains thee Eu clade identified previously. Clusters are often observed in principal components maps of human genetic diversity, however the clusters are usually not so distinctly separated. Some forms of balancing selection could cause clustering to occur, but such selection would explain only the presence of clusters, not their different geographical distributions [103].

A principal components analysis of JCV subpopulations (Figure 6.1) based on DNA sequence variation is consistent with the hypothesis of regional subdivision, as well. Like JCV nucleotide sequences, JCV subpopulations are divided into two main clusters defined by geography (Figure 6.2). These population clusters are not as distinct as the DNA sequence clusters. Although African and Asian populations are composed mostly of lineages from cluster A and European populations are composed mostly of lineages from cluster B, the geographical division is not strict. Lineage

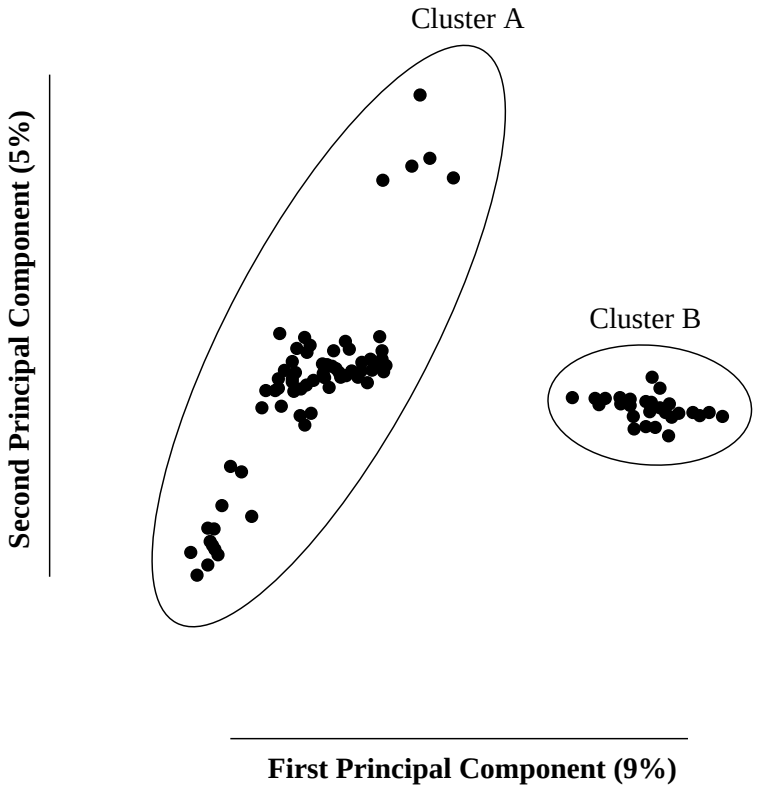

Figure 6.1. Principal components map of JCV DNA sequences. Percent values on the axes indicate the proportion of variance accounted for by each component, but otherwise the axes are dimensionless.

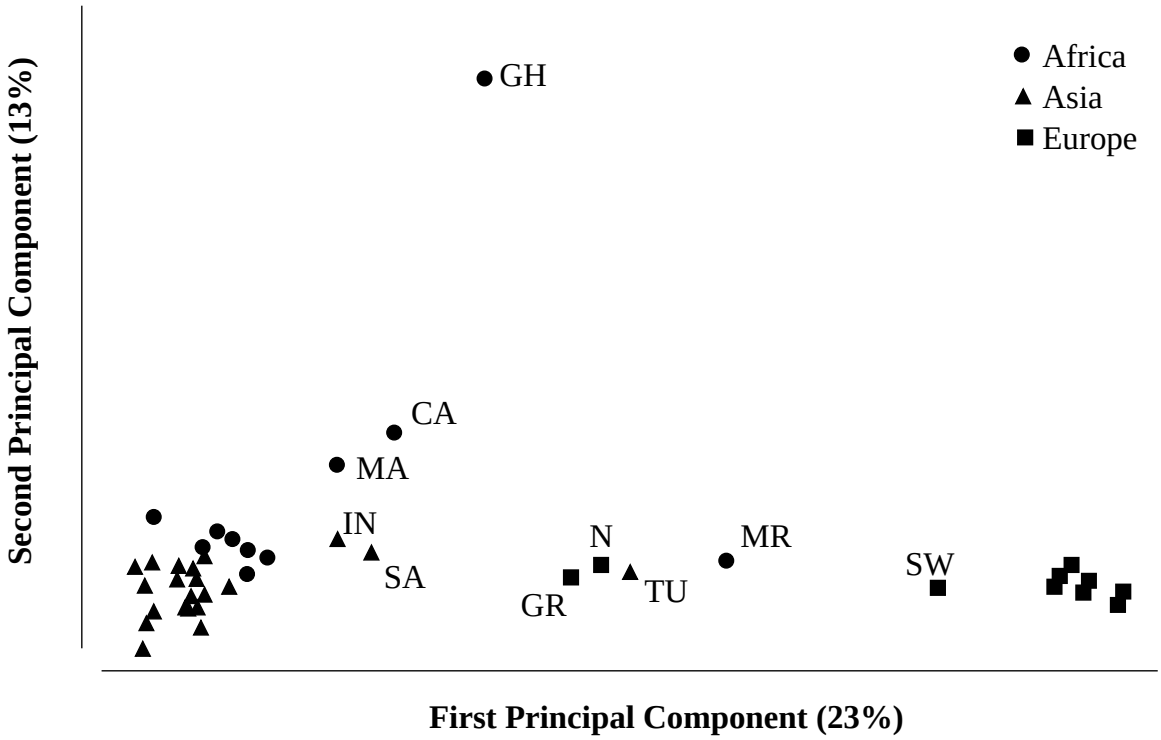

Figure 6.2. Principal components map of JCV subpopulations. Distances are based on DNA sequence variation. Percent values on the axes indicate the proportion of variance accounted for by each component, but otherwise the axes are dimensionless. A key for population symbols is in Figure 6.1.

clusters A and B are both found on all three continents.

One notable feature of the principal components map of JCV subpopulations is the similarity of Asian and African subpopulations, which differ from European subpopulations. Analyses of human genetic diversity ordinarily reflect more uniform between-continent differences [96, 104]. Another notable feature is that subpopulations located near the geographical intersection of continents are located near the intersection of continents on the principal components map, as well. Populations from Greece, Turkey, Morocco and Saudi Arabia, for example, fall between populations from Europe and Africa, and Europe and Asia. This pattern is consistent with the argument that isolation by distance may be the source of similarity in the large-scale genetic structure of human and JCV populations. Details of the pattern may be due to factors like migration.

Levels of diversity within continents differ from observations in humans as well. Measures based on the mean number of nucleotide differences between human genes consistently find more genetic variability in Africa than in Asia and Europe. This trend is generally interpreted as evidence for African origins [97]. In a comparison of DNA sequences from chromosome 22, for example, Zhao et al. [98] found π values in Europe, Asia and Africa of 0.0077, 0.0075 and 0.0085. In JCV, diversity is higher in Europe ($\pi = 0.0185$) than in Asia ($\pi = 0.0171$) or Africa ($\pi = 0.0169$).

The phylogenetic differences between lineage clusters A and B and overall differences in diversity between the European and African/Asian clusters suggest that the JCV subpopulations found in Europe are demographically distinct from JCV populations elsewhere. Geographically defined demographic units may be present. An effective method for comparing the demographic properties of populations is to compare the distribution of pairwise differences among sequences, or mismatch distribution. The mismatch distribution is a well-established tool for studying populations that have been subdivided or have changed in size over time. Whereas single, panmictic populations tend to show moderately rough mismatch distributions, strongly subdivided populations tend to show two or more modes [105]. And whereas populations that have expanded exhibit smooth mismatch distributions with a single mode correlated with the time of expansion, populations that have not expanded

exhibit rough mismatch distributions [10, 12].

The mismatch distribution in JCV DNA sequences shows evidence of both expansion and subdivision (Figure 6.3a). Unlike the mismatch distribution for human mtDNA, which is smooth and unimodal, the mismatch distribution of JCV sequences is smooth, but strongly bimodal, with peaks at 11 and 21 nucleotide differences. Such patterns are unusual, but have been described in other cases population subdivision followed by population expansion [16]. No methods are available to estimate demographic parameters for subdivided, expanding populations, but the distinctness of JCV clusters A and B indicates that it is appropriate to analyze them separately.

Separate analyses of the lineages in cluster A and cluster B yield results expected under population growth, without subdivision. A test of the frequency spectrum of mutations in each sample using Tajima's D statistic, which compares the total number of variable nucleotide positions in a sample of sequences with the mean pairwise difference, rejects the hypothesis of demographic stationarity for both groups at the 95% confidence level (D $= -1.76$ for cluster A, and -1.94 for cluster B) [20]. The mismatch distributions of the two clusters also match expectations under population expansion when analyzed separately. Both are smooth and unimodal. However, the modal difference within cluster A is nine nucleotides, and the modal difference within cluster B is four nucleotides (Figure 6.3b).

The application of Rogers's method of moments yields $\widehat{\theta_0} = 3.524$, $\widehat{\theta_1} = 47.32$ and $\hat{\tau} = 6.822$ for cluster A and $\widehat{\theta_0} = 0.01$, $\widehat{\theta_1} = 118.20$ and $\hat{\tau} = 5.352$ for cluster B [10]. The estimate of $\hat{\tau}$ for cluster A seems slightly too low, possibly due to substructure within the cluster. Under a two-epoch "sudden change" model of population history, which makes the assumption that a population changed instantaneously from an ancient population size to the modern one, $\widehat{\theta_0} = 2N_0\mu$ provides an estimate of ancient population size N_0 (the number of haploid genomes in the population), where μ is the per nucleotide substitution rate times the number of nucleotides in the sequence under consideration. Similarly, $\hat{\tau} = 2\mu t$ provides an estimate of the length of time since the population size changed [10].

Under the synonymous nucleotide substitution rate for primate polyomaviruses proposed by Yasunaga and Miyata [106] (3.8×10^{-8} synonymous substitutions per

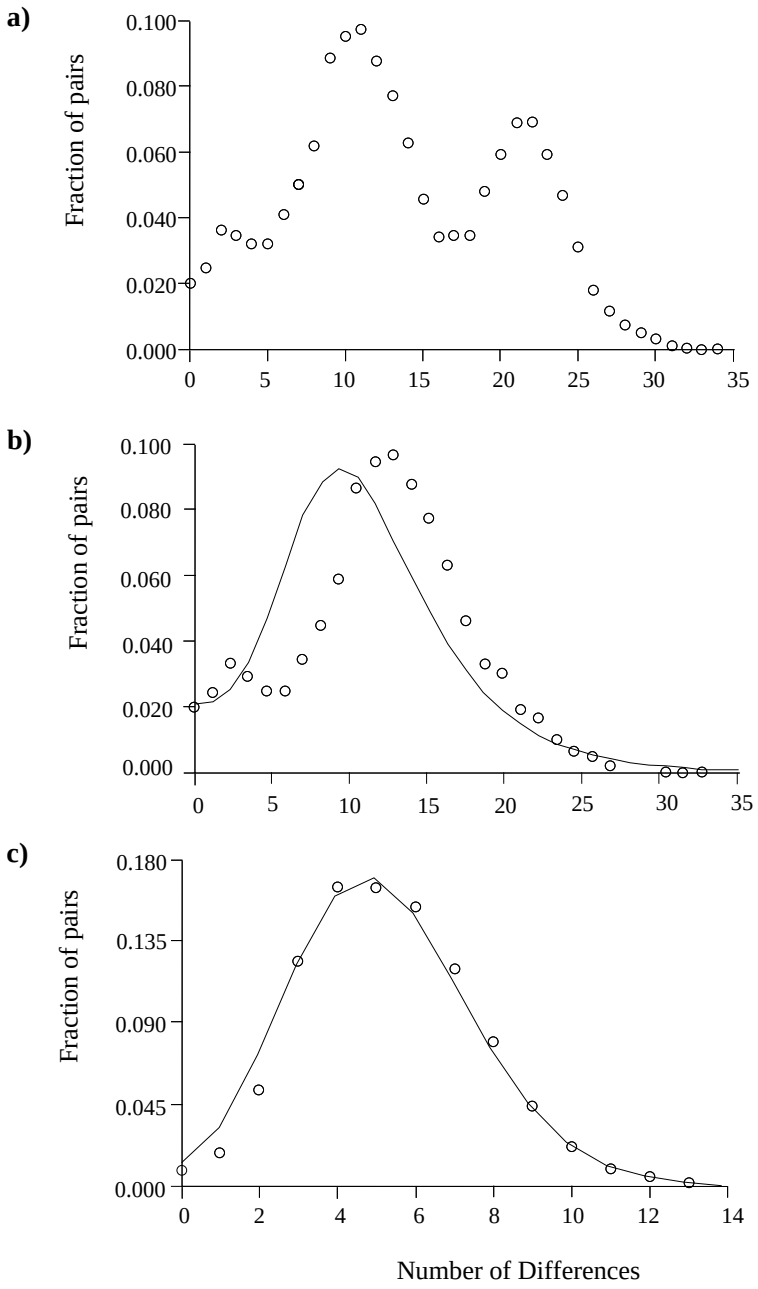

Figure 6.3. Mismatch distributions. a) Mismatch distribution of all JCV nucleotide sequences. Circles represent observed differences. b) Mismatch distribution of JCV nucleotide sequences from cluster A. Circles represent observed differences. Solid line represents the mismatch distribution expected for the parameters estimated for cluster A using Rogers's method of moments ($\widehat{\theta_0} = 3.524$, $\widehat{\theta_1} = 47.32$ and $\hat{\tau} = 6.822$). c) Mismatch distribution of JCV nucleotide sequences from cluster B. Circles represent observed differences. Solid line represents the mismatch distribution expected for the parameters estimated for cluster B using Rogers's method of moments ($\widehat{\theta_0} = 0.01$, $\widehat{\theta_1} = 118.20$ and $\hat{\tau} = 5.352$) .

site per year), $\mu = 2.3 \times 10^{-5}$ and estimates of τ obtained using the method of moments imply an expansion time of 160,000 years before present for cluster A and 150,000 years before present for cluster B. These expansion times are similar to the expansion time of 66,000 – 150,000 years before present for humans inferred from mitochondrial DNA polymorphism [10]. However, the time of expansion for cluster B under this substitution rate predates the permanent occupation of Europe by humans, which occurred after the last glacial maximum around 18,000 years ago [107]. Such an early expansion time implies that this substitution rate is too low.

The rate of 3.8×10^{-7} synonymous substitutions per site per year proposed for JCV by Hatwell and Sharp [92] yields different results. It implies an expansion time of around 14,000 years before present for cluster A and 11,000 years before present for cluster B. However, this rate rests on the assumption that the earliest divergences in JCV correspond to those found in humans. It is circular to use Hatwell and Sharp's estimate for comparisons between JCV and human diversity here since the substitution rates of the two species were not calculated independently.

The signature of population expansion and subdivision in JCV could be due to a variety of factors. One possibility is that JCV populations have been divided into demes corresponding to isolated human subpopulations. Another possibility is that natural selection is causing between-continent differentiation. Under the latter hypothesis, variation should be present in the phenotypes of JCV variants. These might appear in the form of amino acid substitutions in the V-T intergenic region.

Translations of the DNA sequences reported by Sugimoto et al. [65] show that amino acid sequence variation is present in both of the coding regions reported. Thirty-seven different sequences defined by 27 variable amino acid positions are observed (Figure 6.4). Twenty-three of the observed variants occur only once in the sample, with the remaining 14 occurring two or more times.

The structure of a parsimony network generated using Sneath's method [108] shows some of the structure suggested by the principal components analysis of JCV DNA sequences in Figures 6.1 and 6.2: amino acid sequences in clusters A and B fall into different parts of the network. Europe and Africa each contain fewer amino acid variants than Asia does, and every lineage occurring more than once in the sample

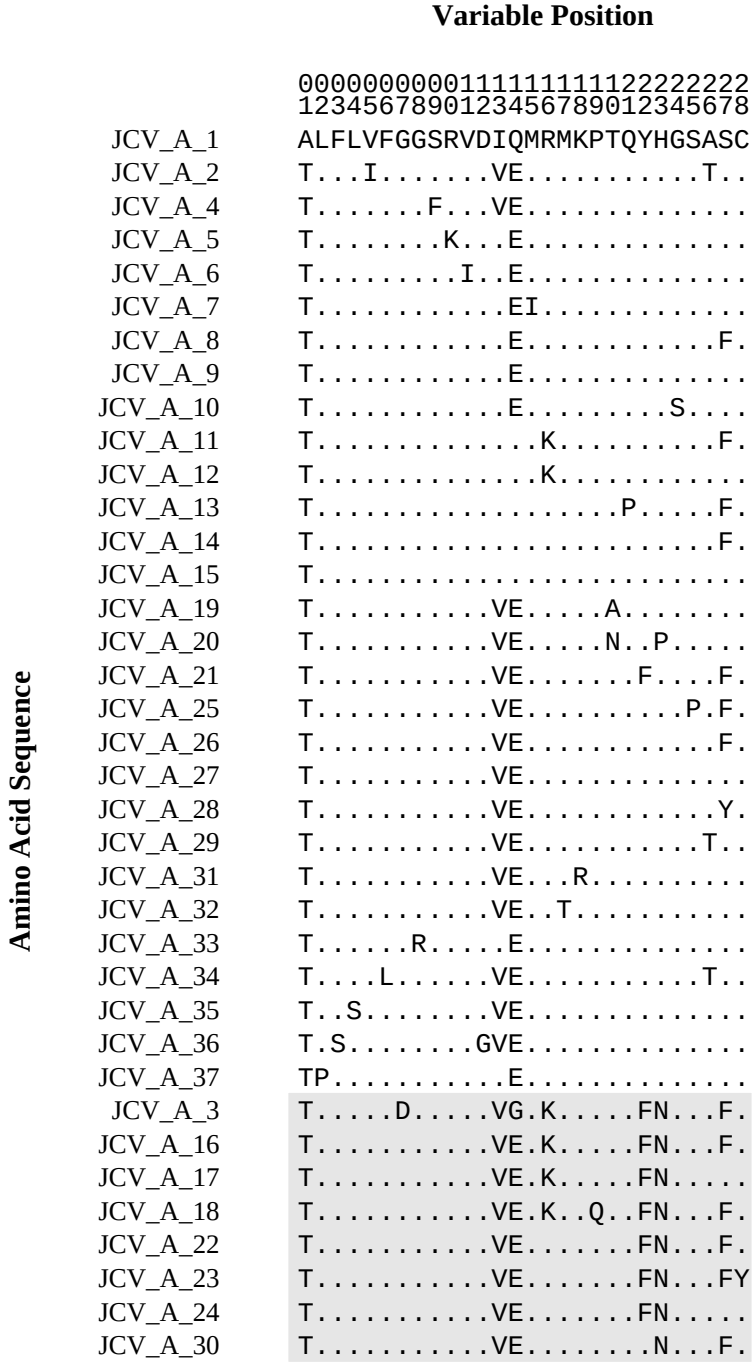

Figure 6.4. Summary of variable amino acid positions. Variable sites are listed and numbered in the order they occur in the sequence data. Sequence JCV_A_1 is the reference sequence. Below it, letters indicate substitutions and dots indicate identity. Unshaded sequences are from cluster A and shaded sequences are from cluster B.

is found in Asia. Also, as suggested by the principal components analysis of JCV subpopulations in Figure 6.2, the lineages endemic to Europe are relatively distinct from those found in Africa and Asia, although lineage compositions of the continents overlap (Figure 6.5).

The population structure suggested by the relative abundances of amino acid lineages in each region is reflected in a principal components map of JCV subpopulations based on amino acid sequence variation (Figure 6.6). In this map, European subpopulations are distinct from African and Asian subpopulations, but African and Asian subpopulations are indistinguishable from each other and form a single nebulous cluster (Figure 6.6). The extent of overlap among African and Asian subpopulations to the exclusion of Europeans is remarkable given that the amino acid sequence variation is a subset of DNA sequence variation distinguishing the three continents. The simplest explanation for this pattern is that natural selection has occurred, resulting in convergent evolution.

The frequency of nucleotide substitutions leading to amino acid substitutions in the sample is consistent with the presence of purifying natural selection, which tends to remove new variants from the population. In expectation, the number of synonymous substitutions per synonymous site (K_s) should equal the number of nonsynonymous substitutions per nonsynonymous nucleotide position (K_a) if substitutions are selectively neutral [93]. K_s/K_a ratios in neutrally evolving populations should be around 1. The mean K_s/K_a ratio observed in pairwise comparisons of all the different JCV sequences is 16.8. Fewer amino acid substitutions are observed than would be expected under neutrality.

The pattern of amino acid substitution in the parsimony network is consistent with the hypothesis that convergent evolution has occurred in JCV proteins (Figure 6.7). Of the eight amino acid substitutions informative in a parsimony analysis (i.e., substitutions in which at least two variants are observed in two or more samples), five appear to have occurred only once in JCV's history, but three must have occurred more than once. Two (16 R-K and 22 Y-F) can be explained by one convergent amino acid substitution each. One substitution (27 S-F) requires at least six convergent amino acid substitutions to explain. For example, if the 27 S-F substitution occurred

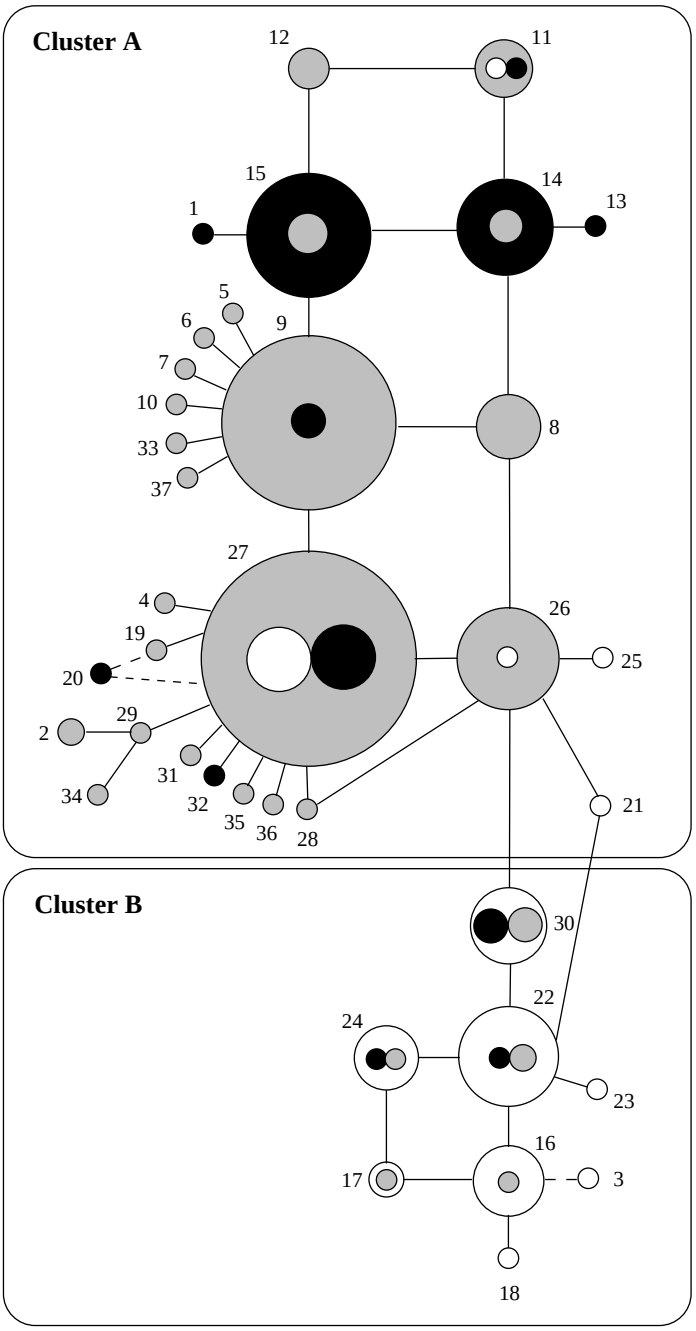

Figure 6.5. Parsimony network relating amino acid variants. Node sizes are proportional to lineage frequencies. Within each node, shading indicates the relative abundance of the lineage in each of the three continents: black represents Africa, gray represents Asia and white represents Europe.

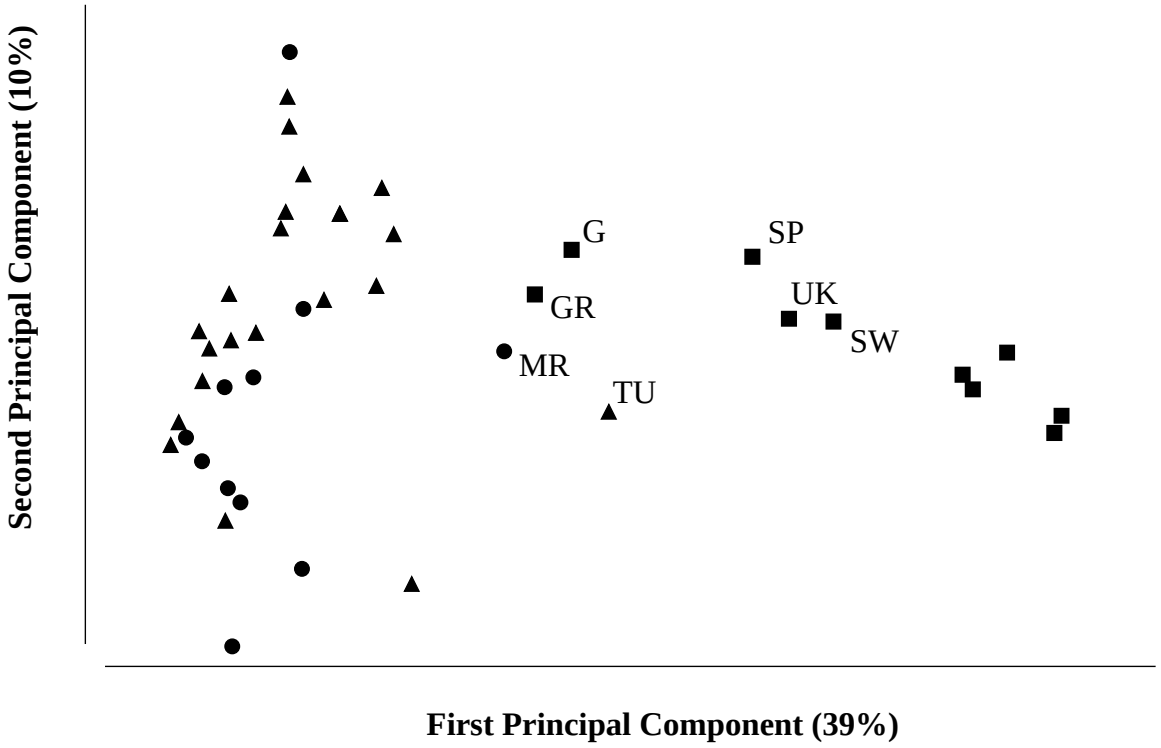

Figure 6.6. Principal components map of JCV subpopulations based on amino acid sequence variation. Percent values on the axes indicate the proportion of variance accounted for by each component, but otherwise the axes are dimensionless. A key for population symbols is in Figure 6.1.

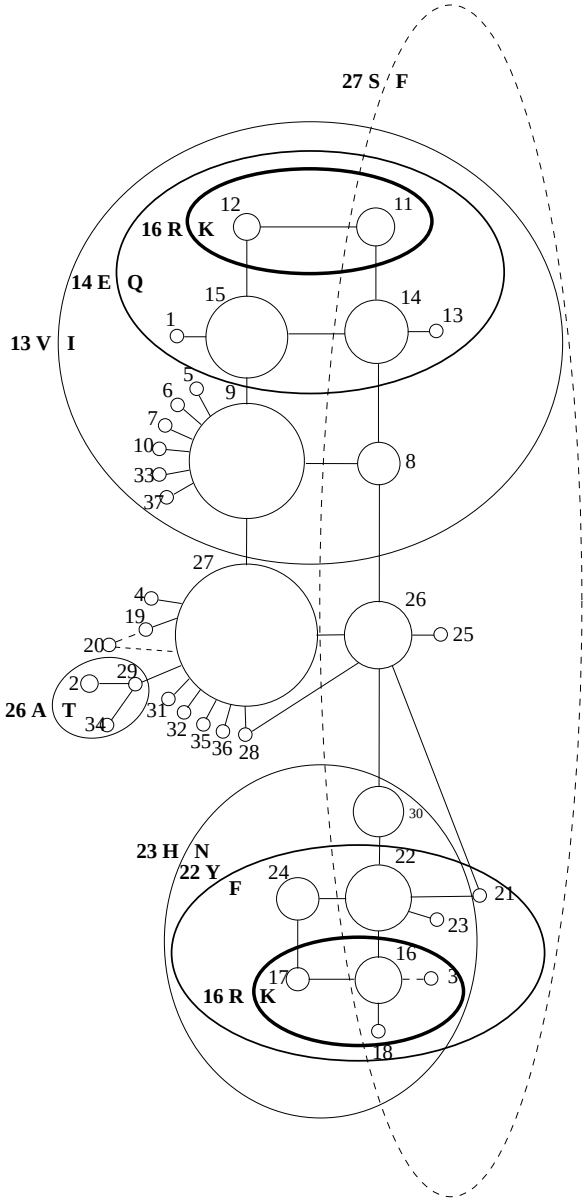

Figure 6.7. Parsimony network showing amino acid substitutions. Contours define the state of amino acid substitutions in which each variant is found in two or more samples (i.e., substitutions informative in a parsimony analysis). The number-letter pairs on each contour indicate the substitution and its alternate states. The number is the substitution number as defined in Figure 6.4. The number inside a contour indicates the residue found in lineages inside the contour. The letter outside a contour indicates the residue found in lineages outside the contour. The dashed contour indicates the 27 S-F substitution.

once between lineage 11 and lineage 12, then the similarity of lineage 14 to lineage 15 would require a parallel mutation, the similarity of lineage 8 to lineage 9 would require another parallel mutation and so on.

An alternative explanation for the recurrence of 27 S-F substitutions in JCV haplotypes is genetic recombination. Recombination has been regarded in much of the published literature as an insignificant factor in natural JCV populations, but the pattern of observed here would be easily explained by the presence of a recombinational hot spot. Unfortunately, an analysis using the method of Jakobsen and Easteal [109], which tests for recombination by identifying cladistic incompatibilities among variable sites, shows that the 27 S-F substitution is the last cladistically informative substitution in the data set. It is impossible to tell whether this position is cladistically incompatible with substitutions both upstream and downstream, which would point to repeated amino acid substitution, or if it only incompatible with substitutions that are upstream, which would point to recombination.

One way to distinguish the hypothesis of multiple substitution from recombination would be to examine the synonymous nucleotide changes in the vicinity of the amino acid substitution. If recombination has occurred, then synonymous substitutions upstream from the 27 S-F substitution should be incompatible with those downstream. Unfortunately, not only is the amino acid substitution of interest the last amino acid substitution in the sequence, the nonsynonymous change that causes the substitution is the third to last informative nucleotide substitution in the sequence. No convincing pattern is found in the final two informative positions. Comparisons of more JCV DNA sequence data will be required to answer the question of whether the 27 S-F changes have occurred through parallel substitution or recombination.

The bimodality of JCV's mismatch distribution, the basal position of the European clade in a phylogeny, the excessive genetic diversity in Europe and the predominantly European origins of cluster B imply that although JCV and humans have a tightly coupled biological relationship, their demographic histories differ. JCV DNA sequences are most likely a poor source of data for studies of human demography outside of the low-resolution applications reported so far. The use of genetic diversity in the JCV parasite to study human populations may be possible at low scales of resolution,

but differences between JCV and human demography indicate that inferences about human populations based on diversity in the virus should be treated with caution.

Given the abundant and rapidly increasing base of human genetic data and the depth of detailed investigations into human population history based on genetic, linguistic, morphometric and archaeological evidence [61], an alternative approach to analyzing genetic diversity in JCV is justified. Information about human demographic history should be used as a context for understanding the demographic history of the virus, rather than the other way around. The incorporation of well-supported prior findings about either host or parasite populations into studies of host-parasite coevolution is likely to provide a useful foundation for identifying otherwise unobservable trends the evolutionary process. In the case of the human-JCV relationship, similarities and differences in patterns of genetic diversity have a variety of implications that are not evident when the two species are considered separately.

The most obvious pattern of diversity in JCV that is placed in context by information about human population history is evidence for population growth. Whereas evidence for a population expansion in JCV in the absence of an expansion in humans would be clear evidence of an epidemiological sweep, evidence for an expansion in both species allows the possibility that they expanded simultaneously. The hypothesis that human and JCV populations expanded simultaneously will be testable if more accurate estimates of nucleotide substitution rates in JCV can be obtained.

Information about human history is important for interpreting phylogenetic topologies and diversity levels in JCV populations as well. These lines of evidence suggest that JCV's relationship with humans began in Europe. While not statistically testable, the basal position of a European clade of JCV lineages, and the relatively high level of diversity (π) levels there, are both more consistent with European origins than they are with prior evidence for human origins in Africa, which are supported by analogous patterns in human genes.

Patterns of between-continent divergence in JCV are similar to patterns found in humans in that they both show evidence for isolation by distance. But the distinct difference between European and Africa/Asian JCV subpopulations, and the marked similarity between subpopulations from Asia and Africa, indicate that human popula-

tion structure alone cannot account for diversity patterns in the virus. The population dynamics of JCV must include the influence of factors such as natural selection or life history traits, such as horizontal transmission, that could enforce between-region similarities and differences independent of those found in humans. High levels of amino acid diversity and evidence for purifying selection imply that natural selection is an important underlying force that may be important in the regional differentiation of viral and human populations. Further examination of differences between regional subpopulations will be required to define fine-scale trends in the adaptive evolution of JCV. The specific comparison of European and non-European populations should be particularly informative in understanding these trends.

6.5 Acknowledgments

Helpful comments and discussion were provided by John Hawks, Henry Harpending, Lynn Jorde, Dennis O'Rourke, Alan Rogers, Jon Seger and four anonymous reviewers.

This project was supported partly by a NIH Genome Sciences Training Grant to the University of Utah.

CHAPTER 7

CONCLUSIONS

Population growth over the last 100,000 years has had pervasive effects on the patterns of genetic diversity found in humans today. But these effects are not uniform. Some parts of our genome, like the nonrecombining portion of the Y chromosome, show the low levels of diversity expected under growth. Others, like SNPs from chromosomal coding regions, do not. This variation presents unique opportunities. On the one hand, the human genome contains information about history that can be used to learn about the emergence and diversification of modern humans. On the other, individual genes contain information about natural selection that can be used to learn about the mechanisms of molecular function and adaptation. But how can this information be extracted? The studies described in this dissertation have outlined a number of methods for analyzing human genetic variation. They also suggest some new, unexplored directions.

The problem of modeling nonstationary populations is clearly fundamental. Abundant evidence indicates that the human population has expanded greatly in recent times, but the expected effects of this growth have been difficult to quantify. Without explicit predictions about diversity patterns under growth, observed batterns of human genetic variation are difficult to interpret. The matrix coalescent model, outlined in Chapter 2, shows that the effects of population size on patterns of human genetic diversity are probably strong: rare alleles should be common, and common alleles should be rare. Empirical observations, however, are mixed.

The abundance of rare alleles caused by population growth is well-documented in many genomic regions (e.g., Chapter 4). However, not all regions are affected. In a survey of diversity restricted to autosomal coding regions, Chapter 2 describes a more unusual pattern: rare alleles are rare, and common alleles are common. This result is best explained by a combination of purifying and balancing natural selection. The

surprising aspect of this result is not that natural selection is observed, but rather that balancing selection appears to be pervasive.

In the 1970s, at the height of electrophoretic enzyme assays, an academic industry emerged that predicted that balancing selection, driven by heterozygote advantage, would be ubiquitous. Aside from a few well documented examples, however, heterozygote advantage was hard to find. Enthusiasm for the theory waned. The results in Chapter 2 suggest that the spectre heterozygote advantage has not disappeared altogether. Tests for Hardy–Weinberg equilibrium—an old standby in population genetics—will certainly play an important role in genome–wide studies of human genetic variation.

The success of the matrix coalescent in modelling populations with changing sizes belies the difficulty of the problem in general. Whereas the matrix coalescent is capable of handling populations with essentially any history of population size change, the computational problems of analyzing nonstationary populations can be overwhelming. Many of the results in Chapter 2, for example, took several days to calculate, and larger sample sizes would have been intractable. Even when good models are available, they are not always feasibly applied to real data.

Chapter 3 describes a parallel approach that works well for many kinds of computationally intensive analysis. By exploiting the availability of numerous, powerful, personal computers, parallel computations can greatly reduce the time required for some analyses. Even simple modifications to existing computer programs can give improved results. In spite of the obvious benefits of this approach in principle, however, its complexity in practice is a problem. A clear goal for future work on parallel programs is to guarantee their accessibility to naive users. The cheap availability of computer networks on university campuses provides an obvious way to achieve this accessibility.

One of the main consequences of human population growth is that humans from around the world are largely alike, having descended from a recent common ancestor. This can be a problem. Whereas population genetic analyses in some species can rely on modest sampling, for example studies of human populations often require intensive sampling. However, some consequences of human population growth can be exploited.

Chapters 5 and 6 show both sides of the coin.

Patterns of human genetic variation in East Asia continue to fuel dispute about the origins of humans in that region. Some interpretations of the East Asian data infer a distinct difference between northern and southern populations. Others do not. Chapter 5 demonstrates that questions about north/south distinctions are difficult to resolve largely because humans have only recently dispersed. Unlike, say, black bears in which sufficient time has passed to allow populations to diverge greatly, East Asian human populations are only recently derived, and genetic isolation by distance in the region is difficult to separate from population subdivision. More intensive population sampling and more sophisticated methods of analysis will be required.

Chapter 6 shows that the patterns of genetic variation so problematic in studying humans can provide a valuable tool for studying other populations. One of the key consequences of population growth is that the signature of growth is stable in the sense that it has similar affects across neutral genes. This provides a basis for comparison with other species whose histories may correlate with that of humans. The JC virus (JCV) is one such species. A substantial virological literature has relied over the last five years on the assumption of a tight correlation between viral and human population histories. However, even simple comparisons of diversity patterns affected most by population growth show this assumption to be false. Whereas the signature of human population growth in humans should manifest itself as a unimodal wave in the mismatch distribution of JCV DNA sequences, the mismatch distribution in JCV is strongly bimodal. Clearly, different forces have shaped diversity in human and JCV populations. Future comparisons of patterns of genetic diversity in human parasites will profit from treating human signatures of growth as a basis for comparison, rather than as the basis for assumption.

Although many basic questions about the relationship between history and genetics have been answered, many more remain. Population size change will remain central to the study of both human origins and human biology for a long time to come.

REFERENCES

[1] H. C. Harpending and A. R. Rogers. Genetic perspectives on human origins and differentiation. *Ann. Rev. Genomics Hum. Gen.*, 1:361–385, 2000.

[2] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Gen.*, 22:139–144, 1999.

[3] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

[4] R. A. Fisher. *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford, first edition, 1930.

[5] J. F. C. Kingman. The coalescent. *Stoc. Proc. Appl.*, 13:235–248, 1982.

[6] S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Pop. Biol.*, 26:119–164, 1984.

[7] R. R. Hudson. Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics, editors, *Oxford Series in Evolutionary Biology*, volume 7, pages 1–44, Oxford, UK, 1990. Oxford University Press.

[8] M. Slatkin and R. R. Hudson. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129:555–562, 1991.

[9] R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Phil. Trans. Roy. Soc. Lond., Ser. B*, 29:403–410, 1994.

[10] A. R. Rogers. Genetic evidence for a Pleistocene population explosion. *Evolution*, 49:608–615, 1995.

[11] M. Kimmel, R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins, and L. B. Jorde. Signatures of population expansion in microsatellite repeat data. *Genetics*, 148:1921–1930, 1998.

[12] A. R. Rogers and H. C. Harpending. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.*, 9:552–569, 1992.

[13] M. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, 149:429–434, 1998.

[14] J. D. Terwilliger, S. Zöllner, M. Laan, and S. Pääbo. Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'Drift Mapping' in small populations with no demographic expansion. *Hum. Hered.*, 48:138–154, 1998.

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

[16] S. Wooding and R. Ward. Phylogeography and Pleistocene evolution in the North American black bear. *Mol. Biol. Evol.*, 14:1096–1105, 1997.

[17] A. P. Rooney, R. L. Honeycutt, and J. N. Derr. Historical population size change of bowhead whales inferred from DNA sequence polymorphism data. *Evolution*, 55:1678–1685, 2001.

[18] N. D. Tsutsui, A. V. Suarez, D. A. Holway, and T. J. Case. Reduced genetic variation and the success of an invasive species. *Proc. Nat. Acad. Sci. USA*, pages 5948–5953, 2000.

[19] E. R. Pianka. *Evolutionary Ecology.* Addison–Wesley Publishing Co., New York, fifth edition, 1995.

[20] F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123:585–595, 1989.

[21] Y.-X. Fu. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147:915–925, 1997.

[22] S. Schneider and L. Excoffier. Estimation of past demographic parameters from the distribution of pairwise differences when mutation rates vary among sites: application to human mitochondrial DNA. *Genetics*, 152:1079–1089, 1999.

[23] S. Schneider and L. Excoffier. Why hunter–gatherer populations do not show signs of Pleistocene demographic expansions. *Proc. Nat. Acad. Sci. USA*, 96:10597–10602, 1999.

[24] H. C. Harpending, M. A. Batzer, M. Gurven, L. B. Jorde, and A. R. Rogers. Genetic traces of ancient demography. *Proc. Nat. Acad. Sci. USA*, 95:1961–1967, 1998.

[25] R. Nielsen and D. M. Weinreich. The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. *Genetics*, 153:497–506, 1999.

[26] A. R. Rogers and L. B. Jorde. Ascertainment bias in estimates of average heterozygosity. *Am. J. Hum. Genet.*, 58:1033–1041, 1996.

[27] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19a:27–43, 1982.

[28] E. E. Harris and J. Hey. Human demography in the Pleistocene: do mitochondrial and nuclear genes tell the same story? *Evol. Anth.*, 8:81–86, 1999.

[29] J. Haigh and J. Maynard Smith. Population size and protein variation in man. *Genet. Res. Camb.*, 19:73–89, 1972.

[30] S. Wooding and A. Rogers. A Pleistocene population X-plosion? *Human Biology*, 72:693–695, 2000.

[31] S. Alonso and J. A. L. Armour. A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. *Proc. Nat. Acad. Sci. USA*, 98:864–869, 2001.

[32] R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154:931–942, 2000.

[33] S. Ross. *A First Course in Probability.* Prentice Hall, Upper Saddle River, NJ, fifth edition, 1997.

[34] A. R. Rogers. Population structure and modern human origins. In P. J. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution*, pages 55–79, New York, 1997. Springer–Verlag.

[35] W. J. Stewart. *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, Princeton, New Jersey, 1994.

[36] C. B. Moler and C. F. Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20:801–836, 1978.

[37] S. Wooding. *TreeToy coalescent simulation version 1.0b.* Computer program distributed by the author, http://mombasa.anthro.utah.edu/wooding/TreeToy, 1999.

[38] M. G. Bulmer. *Principles of Statistics.* Dover Publications Inc., New York, 1979.

[39] A. W. F. Edwards. *Likelihood.* The Johns Hopkins University Press, Baltimore, 1992.

[40] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Gen.*, 22:231–238, 1999.

[41] S. T. Sherry, M. Ward, and K. Sirotkin. Use of molecular variation in the NCBI dbSNP database. *Hum. Mutat.*, 15:68–75, 2000.

[42] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29:308–311, 2001.

[43] B. Haible. *CLN: Class Library for Numbers version 1.0.1.* Computer program distributed by the author, http://clisp.cons.org/ haible/packages-cln.html, 2000.

[44] B. Charlesworth, M. T. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134:1289–1303, 1993.

[45] R. C. Lewontin and J. L. Hubby. A molecular approach to the study of genetic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, 54:595–609, 1966.

[46] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Gen.*, 22:139–144, 1999.

[47] L. B. Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome Research*, 10:1435–1444, 2000.

[48] A. R. Rogers. *Mismatch version 4.3*. Computer program distributed by the author, Department of Anthropology, University of Utah, 1997.

[49] W. Gropp, E. Lusk, and A. Skjellum. *A high-performance, portable implementation of the MPI Message Passing Interface standard*. Computer program distributed by the authors, http://www-unix.mcs.anl.gov/mpi/mpich, 1994.

[50] Message Passing Interface Forum. MPI: A message-passing interface standard. *International Journal of Supercomputing Applications*, 8(3), 1994.

[51] H. Kaessmann, F. Heißig, A. von Haeseler, and Svante Pääbo. DNA sequence variation in a non–coding region of low recombination on the human X chromosome. *Nature Genetics*, 22:78–81, 1999.

[52] H. Kaessmann, V. Wiebe, and Svante Pääbo. Extensive nuclear DNA sequence diversity among chimpanzees. *Science*, 286:1159–1162, 1999.

[53] T. Goebel. Pleistocene human colonization of Siberia and peopling of the Americas: an ecological approach. *Evol. Anth.*, 8:208–227, 1999.

[54] R. S. Davis and V. A. Ranov. Recent work on the Paleolithic of Central Asia. *Evol. Anth.*, 8:186–193, 1999.

[55] C. G. II Turner. Late Pleistocene and Holocene population history of East Asia based on dental variation. *Am. J. Phys. Anth.*, pages 305–321, 1987.

[56] G. L. Barnes. *The Rise of Civilization in East Asia: the Archaeology of China, Korea and Japan*. Thames and Hudson, Ltd., London, 1999.

[57] M. Ruhlen. *The Origin of Language*. John Wiley and Sons, 1994.

[58] R. Du, Y. Yuan, J. Huang, J. Mountain, and L. L. Cavalli-Sforza. Chinese surnames and the genetic differences between North and South china. Number 7 in Journal of Chinese Linguistics Monograph Series, Oxford, UK, 1992. Oxford University Press.

[59] H. Matsumoto. Characteristics of mongoloid and neighboring populations based on the genetic markers of human immunoglobulins. *Hum. Genet.*, 80:207–218, 1988.

[60] T. M. Zhao and T. D. Lee. Gm and Km allotypes in 74 Chinese populations: a hypothesis of the origin of the Chinese nation. *Hum. Genet.*, 83:101–110, 1989.

[61] L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The History and Geography of Human Genes*. Princeton Univ. Press, Princeton, NJ, 1994.

[62] J. Y. Chu, W. Huang, S. Q. Kuang J. M. Wang, J. J. Xu, Z. T. Chu, Z. Q. Yang, K. Q. Lin, P. Li, M. Wu, Z. C. Geng, , C. C. Tan, R. F. Du, and L. Jin. Genetic relationship of populations in China. *Proc. Nat. Acad. Sci. USA*, 95:11763–11768, 1998.

[63] B. Su, J. Xiao, P. Underhill, R. Deka, W. Zhang, J. Akey, W. Huang, D. Shen, D. Lu, J. Luo, J. Chu, J. Tan, P. Shen, R. Davis, L. Cavalli-Sforza, R. Chakraborty, M. Xiong, R. Du, P. Oefner, Z. Chen, and L. Jin. Y-chromosome evidence for a northward migration of modern humans into eastern Asia during the last ice age. *Am. J. Hum. Genet.*, 65:1718–1724, 1999.

[64] T. Kunitake, T. Kitamura, J. Guo, F. Taguchi, K. Kawabe, and Y. Yogo. Parent-to-child transmission is relatively common in the spread of the human polyomavirus JC virus. *J. Clin. Micro.*, 33:1448–1451, 1995.

[65] C. Sugimoto, T. Kitamura, J. Guo, M. N. Al-Ahdal, S. N. Shchelkunov, B. Otova, P. Ondrejka, J. Y. Chollet, S. El-Safi, M. Ettayebi, G. Gresenguet, T. Kocagoz, S. Chaiyarasamee, K. Z. Thant, Thein, K. Moe, N. Kobayashi, F. Taguchi, and Y. Yogo. Typing of urinary JC virus DNA offers a novel means of tracing human migrations. *Proc. Nat. Acad. Sci. USA*, 94:9191–9196, 1997.

[66] A. Torroni, T. G. Schurr, C. C. Yang, E. J. Szathmary, R. C. Williams, M. S. Schanfield, G. A. Troup, W. C. Knowler, D. N. Lawrence, K. M. Weiss, and D. C. Wallace. Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics*, 130:153–162, 1992.

[67] A. Torroni, R. I. Sukernik, T. G. Schurr Y. B. Starikorskaya, M. F. Cabell, M. H. Crawford, A. G. Comuzzie, and D. C. Wallace. MtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. *Am. J. Hum. Genet.*, 53:591–608, 1993.

[68] S. W. Ballinger, T. G. Schurr, A. Torroni, Y. Y. Gan, J. A. Hodge, K. Hassan, K. H. Chen, and D. C. Wallace. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. *Genetics*, 130:139–152, 1992.

[69] A. Torroni, J. A. Miller, L. G. Moore, S. Zamudio, J. Zhuang, T. Droma, and D. C. Wallace. Mitochondrial DNA analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude. *Am. J. Phys. Anth.*, 93:189–199, 1994.

[70] R. Du and F. Y. Vincent. *Ethnic Groups in China*. Science Press, Beijing, 1993.

[71] H. C. Harpending and T. Jenkins. Genetic distance among southern African populations. In M. Crawford and P. Workman, editors, *Method and Theory in Anthropological Genetics*, pages 177–199, Albuquerque, 1973. University of New Mexico Press.

[72] K.–h. Chen. Genetic findings and Mongoloid population migration in China. *Bulletin of the Ethnography Academia Sinica*, 73:209–232, 1992.

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

[74] R. R. Hudson, D. D. Boos, and N. L. Kaplan. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.*, 9:138–151, 1992.

[75] J. Guo, C. Sugimoto, T. Kitamura, H. Ebihara, A. Kato, Z. Guo, J. Liu, S. P. Zheng, Y. L. Wang, Y. Q. Na, M. Suzuki, F. Taguchi, and Y. Yogo. Four geographically distinct genotypes of JC virus are prevalent in China and Mongolia: implications for the racial composition of modern China. *J. Gen. Virol.*, 79:2499–2505, 1998.

[76] C.–m. Hsieh and J. K. Hsieh. *China: a provincial atlas.* Macmillan Publishing, New York, 1995.

[77] D. Comas, F. Calafell, E. Mateu, A. Pérez-Lezaun, E. Bosch, R. Martínez-Arias, J. Clarimon, F. Facchini, G. Fiori, D. Luiselli, D. Pettener, and J. Bertranpetit. Trading genes along the Silk Road: mtDNA sequences and the origin of Central Asian populations. *Am. J. Hum. Genet.*, 63:1824–1838, 1998.

[78] H. C. Harpending and A. R. Rogers. Genetic perspectives on human origins and differentiation. *Ann. Rev. Genomics Hum. Gen.*, 1:361–385, 2000.

[79] D. L. Hartl and A. G. Clark. *Principles of Population Genetics.* Sinauer Associates, Sunderland, MA, third edition, 1997.

[80] R. D. M. Page and M. A. Charleston. Trees within trees: phylogeny and historical association. *TREE*, 13:356–359, 1998.

[81] W. H. Hamilton. Haploid dynamical polymorphism in a host with matching parasites: effects of mutation/subdivision, linkage and patterns of selection. *J. Heredity*, 84:328–338, 1993.

[82] H. Ochman, S. Elwyn, and N. A. Moran. Calibrating bacterial evolution. *Proc. Nat. Acad. Sci. USA*, 96:12638–12643, 1999.

[83] R. Poulin. *Evolutionary Ecology of Parasites: from individuals to communities.* Chapman & Hall, New York, 1998.

[84] M. Boots and A. Sasaki. 'Small worlds' and the evolution of virulence: infection occurs locally and at a distance. *Proc. Roy. Soc. Lond. Ser. B*, 266:1933–1938, 1999.

[85] R. M. May. The dynamics and genetics of host–parasite associations. In C. A. Toft, A. Aeschlimann, and L. Bolis, editors, *Parasite–host associations: coexistence or conflict?*, pages 102–128, Oxford, 1991. Oxford Science Publications.

[86] K. Berthier, M. Langlais, Pierre Auger, and D. Pontier. Dynamics of a feline virus with two transmission modes within exponentially growing host populations. *Proc. Roy. Soc. Lond. Ser. B*, 267:2049–2056, 2000.

[87] R. M. Anderson and R. May. Population biology of infectious diseases: Part I. *Nature*, 280:361–367, 1979.

[88] P. M. Zanotto, E. G. Kallas, R. F. de Souza, and E. C. Holmes. Positive selection in the nef gene of HIV–1. *Genetics*, 153:1077–1089, 1999.

[89] G. S. Ault and G. L. Stoner. Two major types of JC virus defined in progressive multifocal leukoencephalopathy brain by early and late coding region DNA sequences. *J. Gen. Virol.*, 73:2669–2678, 1992.

[90] S. C. Chima, C. F. Ryschkewitsch, K. J. Fan, and G. L. Stoner. Polyomavirus JC genotypes in an urban United States population reflect the history of African origin and genetic admixture in modern African Americans. *Hum. Biol.*, 72:837–850, 2000.

[91] C. F. Ryschkewitsch, J. S. Friedlaender, C. S. Mgone, D.V. D. V. Jobes, H. T. Agostini, S. C. Chima, M. P. Alpers, G. Koki, R. Yanagihara, and G. L Stoner. Human polyomavirus JC variants in Papua New Guinea and Guam reflect ancient population settlement and viral evolution. *Microbes and Infection*, 2:987–996, 2000.

[92] J. N. Hatwell and P. M. Sharp. Evolution of human polyomavirus JC. *J. Gen. Virol.*, 81:1191–1200, 2000.

[93] W.-H. Li. *Molecular Evolution*. Sinauer Associates, 1997.

[94] Z. Yang. *Phylogenetic Analysis by Maximum Likelihood (PAML) version 3.0c*. Computer program distributed by the author, University College London, London, England, 2000.

[95] M. Nei and S. Kumar. *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford, England, 2000.

[96] E. Eller. Population substructure and isolation by distance in three continental regions. *Am. J. Phys. Anth.*, 108:147–159, 1999.

[97] M. Stoneking. Recent African oringin of human mitochondrial DNA: Review of the evidence and current status of the hypothesis. In P. J. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Evolution*, pages 1–14, New York, 1997. Springer–Verlag.

[98] Z. Zhao, L. Jin, Y.-X Fu, M. Ramsay, T. Jenkins, E. Leskinen, P. Pamilo, M. Trexler, L. Patthy, L. B. Jorde, S. Ramos-Onsins, N. Yu, and W.-H. Li. Worldwide dna sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proc. Nat. Acad. Sci. USA*, 97:11354–11358, 2000.

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

[100] A. Kato, T. Kitamura, C. Sugimoto, Y. Ogawa, K. Nakazato, K. Nagashima, W. W. Hall, K. Kawabe, and Y. Yogo. Lack of evidence for the transmission of JC polyomavirus between human populations. *Arch. Virol.*, 142:875–882, 1997.

[101] S. Horai, K. Hayasaka, R. Kondo, K. Tsugane, and N. Takahata. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Nat. Acad. Sci. USA*, 92:532–536, 1995.

[102] R. Thompson, J. K. Pritchard, P. Shen, P. J. Oefner, and M. W. Feldman. Recent common ancestry of human Y chromosomes: evidence form DNA sequence data. *Proc. Nat. Acad. Sci. USA*, pages 7360–7365, 2000.

[103] R. D. M. Page and E. C. Holmes. *Molecular evolution: a phylogenetic approach.* Blackwell Science Ltd., London, 1998.

[104] Y.-C. Ding, S. Wooding, H. Harpending, H.-C Chi, H.-P Li, Y.-X. Fu, J.-F. Pang, Y.-G. Yao, J.-G. Xiang Yu, R. Moyzis, and Y.-P. Zhang. Population structure and history in East Asia. *Proc. Nat. Acad. Sci. USA*, 25:14003–14006, 2000.

[105] P. Marjoram and P. Donnelly. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics*, 136:673–683, 1994.

[106] T. Yasunaga and T. Miyata. Evolutionary changes of nucleotide sequences of papova viruses BKV and SV40: they are possibly hybrids. *J. Mol. Evol.*, 19:72–92, 1982.

[107] R. G. Klein. *The Human Career: Human Biological and Cultural Origins.* University of Chicago Press, Chicago, 1999.

[108] P. H. A. Sneath and R. R Sokal. *Numerical taxonomy : the principles and practice of numerical classification.* W. H. Freeman, San Francisco, 1973.

[109] I. B. Jakobsen and S. Easteal. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comp. Appl. Biosci.*, 12:291–295, 1996.