The Matrix Coalescent and an Application to Human Single-Nucleotide Polymorphisms

Stephen Wooding* and Alan Rogers†

*Eccles Institute of Human Genetics, University of Utah, Salt Lake City, Utah 84112-5330 and †Department of Anthropology, University of Utah, Salt Lake City, Utah 84112-0060

Manuscript received July 5, 2001
Accepted May 10, 2002

ABSTRACT

The “matrix coalescent” is a reformulation of the familiar coalescent process of population genetics. It ignores the topology of the gene tree and treats the coalescent as a Markov process describing the decay in the number of ancestors of a sample of genes as one proceeds backward in time. The matrix formulation of this process is convenient when the population changes in size, because such changes affect only the eigenvalues of the transition matrix, not the eigenvectors. The model is used here to calculate the expectation of the site frequency spectrum under various assumptions about population history. To illustrate how this method can be used with data, we then use it in conjunction with a set of SNPs to test hypotheses about the history of human population size.

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 (Harpending et al. 1998; Terwilliger et al. 1998). In humans, these effects are an important consideration in problems ranging from evolutionary biology to gene mapping. Thus, information about long-term population size contributes to the understanding of both ancient human history and modern human biology.

Information about the history of human population size comes from a variety of 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. Among these tools, models of the coalescent process have distinguished themselves as a way to extract information about past patterns of population size change from present patterns of genetic variation (Fu and Li 2001).

The coalescent process (Kingman 1982a; Hudson 1990) describes the ancestry of a sample of genes. As we trace the ancestry of each modern gene backward from ancestor to ancestor, we occasionally encounter common ancestors—genes whose descendants include more than one gene in the modern sample. Each time this happens, the number of ancestors decreases by one. Eventually, we reach the gene that is ancestral to the entire modern sample, and the process ends. This process provides a natural description of genetic variation, which can be described both in terms of the topological (or genealogical) relationships among genetic lineages and the genetic distances (or coalescence times) between them.

The coalescent process is an example of a Markov process—a stochastic process in which the probability of moving from one state to another depends only on the state you are in, not on the states you have previously visited. In previous literature, attention has focused on the Markov chain that governs not only the lengths of the intervals between coalescent events but also the topology of the resulting gene genealogy (e.g., Takehata 1988). In this article, we introduce a reduced version of the Markov chain that ignores topology and deals only with the lengths of intervals. Our procedure has a number of advantages, especially in dealing with variation in population size. After introducing the model, we use it to study a set of human single-nucleotide polymorphisms (SNPs).

MODEL

The matrix coalescent: If time is measured backward into the past, and a sample of $k$ lineages is selected $t$ generations before present from a haploid 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

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

(1)
(Hudson 1990). For diploid populations, $2N(t)$ can be replaced by $4N(t)$.

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 $\alpha_1(t)$, $\alpha_{n-1}(t)$, ..., $\alpha_2(t)$. In general, the probability, $p_i(t)$, of observing $k$ lineages $t$ generations before present where $n \geq k \geq 1$ is described by a system of recurrence equations

$$p_k(t) = p_k(t) \cdot (1 - \alpha_k(t)) + p_{k+1}(t) \cdot \alpha_{k+1}(t), \quad 1 \leq k < n$$

$$p_n(t) = p_n(t) \cdot (1 - \alpha_n(t)), \quad k = n$$

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

In calculations, we exclude terms for the absorbing state, in which there is just a single lineage. This is not restrictive, since we can always calculate

$$p_i(t) = 1 - \sum_{i=2}^{n} p_i(t).$$

In matrix notation, Equation 2 becomes

$$p(t + 1) = (I + A(t))p(t),$$

where $p(t)$ is a column vector with entries $p_2(t), p_3(t), ..., p_n(t)$, where $I$ is the identity matrix, and where

$$A(t) = \begin{bmatrix}
-\alpha_2(t) & \alpha_3(t) & \cdots & \cdots & \alpha_n(t) & -\alpha_1(t)
\end{bmatrix}$$

is the transition rate matrix. Equation 3 can be approximated in continuous time by an ordinary differential equation

$$\frac{dp(t)}{dt} = A(t)p(t),$$

which is solved by

$$p(t) = e^{A(t)\cdot t}p(0).$$

The entries, $p_i(0)$, of the initial vector $p(0)$ are defined above.

**Eigenvalues and eigenvectors:** Since $A(t)$ is a triangular matrix, its eigenvalues are equal to its diagonal entries: $-\alpha_2(t), ..., -\alpha_n(t)$. The column eigenvectors of $A(t)$ are defined by the equation $A(t)e = \lambda e$, where $\lambda$ is a scalar—one of the eigenvalues of $A$—and $e$ a column eigenvector. This equation can be reexpressed (suppressing $t$) as

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

where $c_i$ is the $i$th entry in vector $c$. The $j$th eigenvector is calculated by setting $\lambda = -\alpha_j$, setting $c_i$ to an arbitrary constant, and then applying (6) repeatedly. When $i = j$, this equation becomes $c_{i+1} = c_i \times 0$. Consequently $c_i = 0$ for all $i > j$, and the matrix $C$ of column eigenvectors is upper triangular.

Equation 6 also implies that the column eigenvectors are time invariant: Substitute (1) into (6) for the $j$th column eigenvector to obtain

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

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 $\lambda$ is an eigenvalue of $A$ and $r$ is the corresponding row eigenvector. This equation can be reexpressed as

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

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.

Before these eigenvectors can be used, they must be normalized so that $RC = I$, where $I$ is the identity matrix. Since both matrices are upper triangular, this requires only that, for eigenvector $j$, we ensure that $r_{ij} = 1$. Our computer program normalizes the eigenvectors by setting $r_j = c_j = 1$.

By expanding the matrix exponential in Equation 5 in diagonal form, we obtain

$$p(t) = CP(t)R\cdot p(0),$$

where $P(t)$ is a diagonal matrix whose $i$th diagonal element is

$$P_i(t) = e^{-\int_{0}^{t}\frac{1}{\alpha_i}dt}$$

and $p(0) = [0, 0, ..., 1]$ as described for (2). The $j$th element of $p(t)$ contains the probability of observing $k$ distinct lineages $t$ generations before present when population size is described by the function $N(t)$.

The second row of plots in Figure 1a shows how $p_i(t)$ varies with $t$ for several different values of $k$ and under three population histories: a sudden population increase, a gradual increase, and a gradual increase with periodic cycling.

**Expected lengths of coalescent intervals:** Let $m$ denote the vector whose $k$th entry, $m_k$, is the expected duration in generations of the interval during which the process contains $k$ lineages. There is a close relationship between $m$ and $p$: The $k$th entry in $p(t)$ is the probability that generation $t$ makes a contribution to the interval during which the process has $k$ lineages. To calculate $m$, we integrate across $p$: 

$$m_k = \int_0^\infty p_k(t) dt.$$
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 (Harpending et al. 1998; Wooding 1999). The sensitivity of the frequency spectrum to population size change is exploited in several statistical tests of stationarity or neutrality (Tajima 1989; Fu 1997).

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, \( \sigma_k \), of sites at which the derived allele occurs \( k \) times is given by

\[
\sigma_k \approx \frac{\sum_{j=2}^{n} \sum_{k=0}^{n-j} jm_k y(j, k, n)}{\sum_{j=1}^{n} j m_j},
\]

where \( n \) is the number of DNA sequences in the sample, \( m_j \) is the expected length of the coalescent interval con-
taining \( j \) distinct lineages, and \( y(j, k, n) \) is the probability that a single lineage within coalescent interval \( j \) has \( k \) descendants in a sample of size \( n \). This equation is derived in Appendix A.

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

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

(Felsenstein 1992; Sherry et al. 1997; see also Equation 21 in Fu 1995). Figure 1b shows the theoretical frequency spectrum under several assumptions about population history.

**Numerical methods:** Equation 5 contains a matrix exponential, and these are notoriously difficult to evaluate numerically (Moler and Loan 1978). It does not help to expand the exponential in terms of eigenvalues and eigenvectors. When the sample size is much over 50, the two eigenvector matrices in Equation 8 will contain very large numbers as well as small ones. Even worse, the entries of each row of \( C \) alternate in sign, leading to severe cancellation errors in the matrix product on the right-hand side of Equation 8. With samples of even moderate size, straightforward evaluation of these equations can produce results without any significant digits.

We deal with these problems in two different ways. For population histories of arbitrary complexity, we resort to brute force and use the CLN1.0.1 programming library (Haible 2000) to perform computations either with high-precision floating-point numbers or with rational numbers. By varying the precision, it is possible to determine how many digits of precision are needed. Some of our calculations were performed with floating-point numbers using 500 decimal digits of precision.

Better alternatives are available when the population’s history is piecewise constant. By this we mean that the history is divided into a series of epochs within each of which \( N(t) \) is constant. If the number of epochs is large, the piecewise constant model can approximate any history of population size. Even with only a few epochs, it is probably realistic for populations whose sizes are ordinarily held constant by density-dependent population regulation.

For such histories, we use the “uniformization” algorithm of Stewart (1994, Chap. 8), to evaluate equation 5 across a single epoch of the population’s history. This makes it possible to project the vector \( p \) backward in time by epoch. With this method, double-precision floating-point calculations are able to deal with problems involving samples of at least 1000.

This method for projecting \( p \) backward in time also makes it easy to calculate \( m \). Details are given in Appendix B.

**Statistical methods:** Under the assumption that the genealogies of unlinked sites are statistically independent, the log-likelihood of an observed data set \((D)\) given a hypothetical population history \((H)\) is

\[
L(D|H) = \sum_{i=1}^{n-1} S_i \ln \sigma_i,
\]

where \( S_i \) is the number of sites occurring \( k \) times in the sample and \( \sigma_i \) is the probability of a variant site occurring \( k \) times in the sample. If different sample sizes are used for different loci, \( \sigma_i \) changes from site to site. The ratios of likelihoods under different population histories can be compared using standard likelihood-ratio tests (Bulmer 1979; Edwards 1992).

**Application to human SNPs:** SNPs are a potentially valuable source of information about population history: They are abundant, they are spread widely across the genome, and they are relatively inexpensive to assay. Most studies of SNPs are focused on their potential epidemiological applications (e.g., Cargill et al. 1999; Halushka et al. 1999). SNPs have also been exploited as a source of information about the process of natural selection (Sunyaev et al. 2000; Fay et al. 2001). We focus here on human population history, although we include some discussion of selective processes.

Cargill et al. (1999) 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 (Cargill et al. 1999). In total, 612 SNPs were identified in 106 genes.

The laboratory methodology used by Cargill et al. (1999) avoided some problems such as 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, while others were sampled in 114. To cope with these problems, Cargill et al. (1999) were forced in some analyses to rely on doubtful assumptions. 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 data set randomly in three steps:

1. All of the SNPs reported by Cargill et al. (1999) were divided into the three categories reported originally: coding nonsynonymous (cns) and 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
<table>
<thead>
<tr>
<th>WIAF</th>
<th>Gene</th>
<th>Location</th>
<th>p</th>
<th>n</th>
<th>WIAF</th>
<th>Gene</th>
<th>Location</th>
<th>p</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>10547</td>
<td>CYP21</td>
<td>6p21</td>
<td>1</td>
<td>82</td>
<td>10547</td>
<td>CYP21</td>
<td>6p21</td>
<td>1</td>
<td>82</td>
</tr>
<tr>
<td>10549</td>
<td>FSHR</td>
<td>chr2</td>
<td>1</td>
<td>82</td>
<td>10549</td>
<td>FSHR</td>
<td>chr2</td>
<td>1</td>
<td>84</td>
</tr>
<tr>
<td>10554</td>
<td>GNRHR</td>
<td>8p21</td>
<td>3</td>
<td>86</td>
<td>10554</td>
<td>GNRHR</td>
<td>8p21</td>
<td>3</td>
<td>86</td>
</tr>
<tr>
<td>10557</td>
<td>DRL</td>
<td>5q31</td>
<td>1</td>
<td>86</td>
<td>10557</td>
<td>DRL</td>
<td>5q31</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10568</td>
<td>CYP11B1</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
<td>10568</td>
<td>CYP11B1</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10591</td>
<td>GH1</td>
<td>17q21</td>
<td>1</td>
<td>86</td>
<td>10591</td>
<td>GH1</td>
<td>17q21</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10605</td>
<td>GHR</td>
<td>5p13</td>
<td>1</td>
<td>86</td>
<td>10605</td>
<td>GHR</td>
<td>5p13</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10606</td>
<td>CYP11A</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
<td>10606</td>
<td>CYP11A</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10607</td>
<td>FSH</td>
<td>11p13</td>
<td>3</td>
<td>86</td>
<td>10607</td>
<td>FSH</td>
<td>11p13</td>
<td>3</td>
<td>86</td>
</tr>
<tr>
<td>10638</td>
<td>CYP11B2</td>
<td>8q21</td>
<td>3</td>
<td>86</td>
<td>10638</td>
<td>CYP11B2</td>
<td>8q21</td>
<td>3</td>
<td>86</td>
</tr>
<tr>
<td>10651</td>
<td>IF1</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
<td>10651</td>
<td>IF1</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10657</td>
<td>SHBG</td>
<td>17q13</td>
<td>2</td>
<td>96</td>
<td>10657</td>
<td>SHBG</td>
<td>17q13</td>
<td>2</td>
<td>96</td>
</tr>
<tr>
<td>10726</td>
<td>HSDSB2</td>
<td>1p13</td>
<td>1</td>
<td>86</td>
<td>10726</td>
<td>HSDSB2</td>
<td>1p13</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10753</td>
<td>DNB</td>
<td>1p13</td>
<td>3</td>
<td>72</td>
<td>10753</td>
<td>DNB</td>
<td>1p13</td>
<td>3</td>
<td>72</td>
</tr>
<tr>
<td>10780</td>
<td>DRD3</td>
<td>3q13</td>
<td>7</td>
<td>86</td>
<td>10780</td>
<td>DRD3</td>
<td>3q13</td>
<td>7</td>
<td>86</td>
</tr>
<tr>
<td>10792</td>
<td>COMT</td>
<td>2q11</td>
<td>3</td>
<td>86</td>
<td>10792</td>
<td>COMT</td>
<td>2q11</td>
<td>3</td>
<td>86</td>
</tr>
<tr>
<td>10846</td>
<td>HTR1D</td>
<td>1p36</td>
<td>1</td>
<td>78</td>
<td>10846</td>
<td>HTR1D</td>
<td>1p36</td>
<td>1</td>
<td>78</td>
</tr>
<tr>
<td>10862</td>
<td>SLCOa4</td>
<td>Unk.</td>
<td>1</td>
<td>74</td>
<td>10862</td>
<td>SLCOa4</td>
<td>Unk.</td>
<td>1</td>
<td>74</td>
</tr>
<tr>
<td>10865</td>
<td>TH</td>
<td>Unk.</td>
<td>3</td>
<td>74</td>
<td>10865</td>
<td>TH</td>
<td>Unk.</td>
<td>3</td>
<td>74</td>
</tr>
<tr>
<td>10870</td>
<td>HTR1E</td>
<td>6q14</td>
<td>3</td>
<td>74</td>
<td>10870</td>
<td>HTR1E</td>
<td>6q14</td>
<td>3</td>
<td>74</td>
</tr>
<tr>
<td>10898</td>
<td>HTR2A</td>
<td>Unk.</td>
<td>2</td>
<td>74</td>
<td>10898</td>
<td>HTR2A</td>
<td>Unk.</td>
<td>2</td>
<td>74</td>
</tr>
<tr>
<td>10914</td>
<td>CETF</td>
<td>Unk.</td>
<td>3</td>
<td>114</td>
<td>10914</td>
<td>CETF</td>
<td>Unk.</td>
<td>3</td>
<td>114</td>
</tr>
<tr>
<td>10952</td>
<td>F</td>
<td>11p11</td>
<td>1</td>
<td>86</td>
<td>10952</td>
<td>F</td>
<td>11p11</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10958</td>
<td>F2R</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
<td>10958</td>
<td>F2R</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10960</td>
<td>F3</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
<td>10960</td>
<td>F3</td>
<td>Unk.</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td>10971</td>
<td>HCF2</td>
<td>22q11</td>
<td>1</td>
<td>106</td>
<td>10971</td>
<td>HCF2</td>
<td>22q11</td>
<td>1</td>
<td>106</td>
</tr>
<tr>
<td>10975</td>
<td>HMGCer</td>
<td>Unk.</td>
<td>1</td>
<td>114</td>
<td>10975</td>
<td>HMGCer</td>
<td>Unk.</td>
<td>1</td>
<td>114</td>
</tr>
<tr>
<td>11004</td>
<td>TPPI</td>
<td>Unk.</td>
<td>1</td>
<td>100</td>
<td>11004</td>
<td>TPPI</td>
<td>Unk.</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>11020</td>
<td>CLanalog</td>
<td>Unk.</td>
<td>1</td>
<td>76</td>
<td>11020</td>
<td>CLanalog</td>
<td>Unk.</td>
<td>1</td>
<td>76</td>
</tr>
<tr>
<td>11030</td>
<td>ITG92B</td>
<td>17q11</td>
<td>1</td>
<td>112</td>
<td>11030</td>
<td>ITG92B</td>
<td>17q11</td>
<td>1</td>
<td>112</td>
</tr>
<tr>
<td>11033</td>
<td>ITG92C</td>
<td>Unk.</td>
<td>1</td>
<td>106</td>
<td>11033</td>
<td>ITG92C</td>
<td>Unk.</td>
<td>1</td>
<td>106</td>
</tr>
<tr>
<td>11035</td>
<td>LDLR</td>
<td>Unk.</td>
<td>1</td>
<td>126</td>
<td>11035</td>
<td>LDLR</td>
<td>Unk.</td>
<td>1</td>
<td>126</td>
</tr>
<tr>
<td>11037</td>
<td>LPL</td>
<td>8p22</td>
<td>1</td>
<td>126</td>
<td>11037</td>
<td>LPL</td>
<td>8p22</td>
<td>1</td>
<td>126</td>
</tr>
<tr>
<td>11041</td>
<td>F1</td>
<td>1p36</td>
<td>1</td>
<td>108</td>
<td>11041</td>
<td>F1</td>
<td>1p36</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>11050</td>
<td>G</td>
<td>10p11</td>
<td>1</td>
<td>108</td>
<td>11050</td>
<td>G</td>
<td>10p11</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>11070</td>
<td>FGA</td>
<td>Unk.</td>
<td>5</td>
<td>108</td>
<td>11070</td>
<td>FGA</td>
<td>Unk.</td>
<td>5</td>
<td>108</td>
</tr>
<tr>
<td>11071</td>
<td>FGB</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
<td>11071</td>
<td>FGB</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>11074</td>
<td>PCIF</td>
<td>1q14</td>
<td>1</td>
<td>102</td>
<td>11074</td>
<td>PCIF</td>
<td>1q14</td>
<td>1</td>
<td>102</td>
</tr>
<tr>
<td>11080</td>
<td>APOD</td>
<td>3q26</td>
<td>1</td>
<td>108</td>
<td>11080</td>
<td>APOD</td>
<td>3q26</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>11085</td>
<td>F13A1</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
<td>11085</td>
<td>F13A1</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>11092</td>
<td>F11</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
<td>11092</td>
<td>F11</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>11114</td>
<td>F7</td>
<td>1q31</td>
<td>1</td>
<td>104</td>
<td>11114</td>
<td>F7</td>
<td>1q31</td>
<td>1</td>
<td>104</td>
</tr>
<tr>
<td>11170</td>
<td>F13B</td>
<td>1q31</td>
<td>1</td>
<td>103</td>
<td>11170</td>
<td>F13B</td>
<td>1q31</td>
<td>1</td>
<td>103</td>
</tr>
<tr>
<td>11200</td>
<td>ANX3</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
<td>11200</td>
<td>ANX3</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>11203</td>
<td>F10</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
<td>11203</td>
<td>F10</td>
<td>Unk.</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>11215</td>
<td>LIPC</td>
<td>Unk.</td>
<td>3</td>
<td>104</td>
<td>11215</td>
<td>LIPC</td>
<td>Unk.</td>
<td>3</td>
<td>104</td>
</tr>
<tr>
<td>11227</td>
<td>SELIP Unk.</td>
<td>1</td>
<td>106</td>
<td>11227</td>
<td>SELIP Unk.</td>
<td>1</td>
<td>106</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11226</td>
<td>TBXAS1</td>
<td>Unk.</td>
<td>1</td>
<td>106</td>
<td>11226</td>
<td>TBXAS1</td>
<td>Unk.</td>
<td>1</td>
<td>106</td>
</tr>
<tr>
<td>11229</td>
<td>NLRD</td>
<td>Unk.</td>
<td>1</td>
<td>106</td>
<td>11229</td>
<td>NLRD</td>
<td>Unk.</td>
<td>1</td>
<td>106</td>
</tr>
<tr>
<td>11301</td>
<td>PAI2</td>
<td>Unk.</td>
<td>1</td>
<td>96</td>
<td>11301</td>
<td>PAI2</td>
<td>Unk.</td>
<td>1</td>
<td>96</td>
</tr>
<tr>
<td>11304</td>
<td>GP1BA</td>
<td>17q12</td>
<td>3</td>
<td>99</td>
<td>11304</td>
<td>GP1BA</td>
<td>17q12</td>
<td>3</td>
<td>99</td>
</tr>
<tr>
<td>11311</td>
<td>MPL</td>
<td>Unk.</td>
<td>1</td>
<td>104</td>
<td>11311</td>
<td>MPL</td>
<td>Unk.</td>
<td>1</td>
<td>104</td>
</tr>
<tr>
<td>11325</td>
<td>CD36</td>
<td>Unk.</td>
<td>1</td>
<td>90</td>
<td>11325</td>
<td>CD36</td>
<td>Unk.</td>
<td>1</td>
<td>90</td>
</tr>
<tr>
<td>11339</td>
<td>PAI1</td>
<td>7q21</td>
<td>1</td>
<td>106</td>
<td>11339</td>
<td>PAI1</td>
<td>7q21</td>
<td>1</td>
<td>106</td>
</tr>
</tbody>
</table>

The three major columns represent cns, cs, and nc sites, respectively. The five minor columns represent, from left to right, WIAF, Whitehead Institute/Affymetrix identifier dbSNP ID; gene name; genomic location if known; p, frequency category (1 = 0–5%, 2 = 5–15%, 3 = 15–50%); and n, number of chromosomes surveyed. Unk., unknown.
Figure 2.—Pitch plot of deviations from expectation. The deviation of each frequency category (0–5%, 5–15%, and 15–50%) from expectation in each SNP category (cns, cs, and nc) is shown. 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 as follows: (a) stationarity, constant population size, and selective neutrality; (b) mtDNA (recent), the most recent population expansion not rejected by Rogers (1995) on the basis of mtDNA polymorphism (see model); (c) mtDNA (ancient), the most ancient population expansion not rejected by Rogers (1995); (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.

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. (1999; 0–5%, 5–15%, and 15–50%) was tabulated for cns, cs, and nc SNPs using the dbSNP database (Sherry et al. 2000, 2001).

Totals of 60 cns loci, 68 cs loci, and 30 nc loci were included in the randomized data set, which was composed of sites from at least 19 different chromosomes (Table 1). 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 one-half of the SNPs in the original data set, so theoretical spectra were “folded” at frequency 0.5 in tests here, as described by Harpending et al. (1998).

Likelihoods of hypotheses given the observed frequency spectra were generated for each data set over a series of hypothetical population histories. Although the matrix coalescent can cope with very complicated models of history, it is doubtful that we could estimate more than a few parameters with the data at hand. We have therefore limited our analysis to piecewise-constant population histories containing two history epochs. We define these histories using three parameters: $N_0$ is the population size during the most recent epoch (epoch 0), $N_1$ is that in the earlier epoch (epoch 1), and $T$ is the duration of epoch 0 in generations. Epoch 1 is assumed to have infinite duration. Our analysis loses a degree of freedom because the data are a collection of polymorphic sites and do not inform us about the fraction of sites that are polymorphic within the region of the genome under study. Thus, instead of working directly with the
three alternatives: (a) stationarity, (b) the most recent population sizes, as well: One simply uses $1/\tau$. Then the maximum likelihood of each SNP nities for the study of populations that vary in size.

maximum-likelihood 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 data set were not rejected as an explanation for the cs data set at the 0.05 level, but the maximum-likelihood estimates for nc and cs data sets were rejected as an explanation for the cns data set. The maximum-likelihood parameters for the cns data set were almost (but not quite) rejected as an explanation for the nc ($p < 0.08$). The nc and cs data were indistinguishable, but both could be distinguished from cns (Figure 2). In addition, the cns data showed an excess of low frequency variation relative to expectations under stationarity, as Cargill et al. (1999) also observed.

The failure of likelihood-ratio tests to distinguish between cs and nc categories is a result of their similar $\rho$ estimates. When $\rho$ is near 1 the time of population size change has little effect on the frequency spectrum, and confidence intervals around $\tau$ are broad. When $\rho$ 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 consistent.

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 mtDNA.

DISCUSSION

The model introduced here differs from the coalescent theory introduced by Kingman (1982a) in that it ignores the topology of the gene genealogy. This simplified theory has a smaller state space than the classical theory, and it is easy to apply the elementary methods of the theory of Markov chains. This opens up opportunities for the study of populations that vary in size.

Conventional coalescent theory can deal with varying population sizes, as well: One simply uses $1/N(t)$ as the unit of time in generation $t$. [This procedure was suggested by Kingman (1982b, p. 31) and has been used by many later authors; we use it in appendix b of this article.] However, this procedure is awkward when mutations are introduced, because mutations occur at a constant rate on the normal (not the rescaled) timescale. This has complicated efforts to calculate quantities like the expected site frequency spectrum under models of varying population size. Such problems are easier under the formulation introduced here.

The results of this study clearly reject the hypothesis that the cns, cs, and nc SNP data were produced by drift and mutation alone under a model of recent population expansion. The simplest explanation for the present results, taken in isolation, is that human population size has been constant, but some form of selection has affected the cns data. The preponderance of low-frequency polymorphisms in those data is consistent either with purifying selection acting on linked sites or with a selective sweep (Braverman et al. 1995; Fu 1997). Fay

![Figure 3.—Maximum-likelihood parameter estimates. Open circles show the parameters of two alternative hypotheses estimated from mitochondrial DNA (see model).](https://academic.oup.com/genetics/article/161/4/1641/6049862)
et al. (2001), for example, found evidence for purifying selection in an analysis of the ratios of synonymous and nonsynonymous variants in different frequency categories in the Cargill et al. (1999) data set. Similar patterns of variation have been attributed to weak purifying selection elsewhere (Przeworski et al. 1999). Yet the present results should probably not be taken in isolation. Genetic data from substantial human samples involving a variety of genetic systems are now published. These can be divided into two categories: noncoding regions that on a priori grounds ought to be selectively neutral and coding regions (or closely linked introns) that on a priori grounds are more likely to be selected. The presumably neutral systems all show evidence either of population growth or of a selective sweep. (We cannot tell the difference.) The presumably selected systems are all consistent either with neutral evolution under constant population size or with weak balancing selection. To account for this strange pattern, Harpending and Rogers (2000) suggested that population growth did in fact occur during the Late Pleistocene, but that its signature has been obscured in the coding portions of the human genome by pervasive balancing selection. Additional data sets that have appeared since then have been consistent with this hypothesis (Rogers 2001). Thus, it is natural to wonder whether the present data set is also consistent with this hypothesis. Let us consider, then, the possibility that the present data reflect the simultaneous effects of population growth and selection.

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, would not be reflected equally in our three categories of data. Many mutations would occur on the long branches that separate allelic classes, but only the neutral mutations would survive long. Consequently, these long branches would 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 can have intermediate frequencies (i.e., far from 0 or 1). 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 alleles 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.

There are undoubtedly other ways to explain these data, and there is no good reason for confidence in the hypothesis we just proposed. Our point is merely that the present data are consistent with the view that the human population underwent an expansion whose effects are visible in data from neutral loci but are hidden by balancing selection at protein-coding loci.

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. Nelson Beebe provided helpful advice on numerical methods. Yun-Xin Fu and two anonymous reviewers provided helpful comments. S.W. was supported by a National Institutes of Health (NIH) Genome Sciences Training Grant (Genome Informatics) to the University of Utah. A.R. was supported by NIH grant GM-59290 to the University of Utah. Software developed for this project is available at http://www.anthro.utah.edu/~rogers/src.

LITERATURE CITED


APPENDIX A: THE EXPECTED SITE FREQUENCY SPECTRUM

We assume that mutations are rare enough that the possibility of multiple mutations in a single gene
ancestry can be ignored. Let \( A_j \) denote the event that exactly one mutation occurs within the portion of the
ancestry containing \( j \) lineages, \( B \) the event that exactly one mutation occurs within the ancestry as a whole, and \( Pr[A_j|B] \)
the conditional probability of \( A_j \) given \( B \). The conditional probability that the mutant site will appear \( k \)
times within a sample, given \( B \), is

\[
\sigma_k = \sum_{j=2}^{n} Pr[A_j|B] \gamma(j, k, n).
\]

Using Bayes’ rule,

\[
Pr[A_j|B] = \frac{Pr[B|A_j]Pr[A_j]}{Pr[B]}.
\]

Here, \( Pr[B|A_j] = 1 \) because event \( B \) occurs whenever \( A_j \) does.

To calculate the unconditional probability of \( A_j \), let \( L_i \) be the length of the \( j \)th coalescent interval in a
random gene tree. Then \( \mathbb{E}(L_i) \) is the total branch length associated with that coalescent interval. The conditional
probability, given \( L_i \), that a single mutation occurs within this interval is

\[
Pr[A_j|L_i] = \mu L_i e^{-\mu L_i} = \mu j L_i,
\]

where \( \mu \) is the mutation rate, and we assume a Poisson distribution of mutations. The approximation here as-
sumes that \( \mu^2 \) is negligible in comparison to \( \mu \). The unconditional probability of \( A_j \) is

\[
Pr[A_j] = \mathbb{E}[\mu j L_i] = \mu j m_j,
\]

where \( m_j \) is the expected value of \( L_i \). A similar argument gives

\[
Pr[B] = \mathbb{E}\left[ \mu \sum_{j=2}^{n} j L_i \right] = \mu \sum_{j=2}^{n} j m_j,
\]

where the sum on the right is the expected length of the gene tree as a whole. Substituting these results back
into Equations 15 and 14 gives Equation 13.

APPENDIX B: CALCULATING EXPECTED INTERVAL LENGTHS UNDER PIECEWISE CONSTANT POPULATION HISTORIES

Our goal in this section is to calculate the vector \( m \), which contains the expected lengths of the intervals
between coalescent events. To simplify the problem, we first separate \( N(t) \) from \( A(t) \) by defining

\[
\beta_i = N(t) - \alpha_i(t) = i(i - 1)/2
\]

and

\[
B = \begin{pmatrix} -\beta_2 & \beta_3 \\ -\beta_3 & \ddots \\ \vdots & \ddots & \ddots \\ -\beta_n & \ddots & -\beta_n \end{pmatrix}
\]

With these definitions, substitution of (5) into (10) gives

\[
m = \int_0^\infty p(t) \, dt = \int_0^\infty e^{B \mathbb{E}(N^{-1}(0) dV)} \mathbb{E}(V) p(0) = F(0, \infty) p(0),
\]

where \( v = \int_{0}^{\infty} \mathbb{E}(N^{-1}) (z) \, dz \) and

\[
F(a, b) = \int_{a}^{b} \mathbb{E}(N^{-1}) (z) \, dz.
\]

Suppose now that the population’s history is divided into \( K + 1 \) epochs within each of which \( N(t) \) is constant. We can reexpress \( F(0, \infty) \) as a sum of contributions from these epochs:

\[
F(0, \infty) = F(0, v_1) + F(v_1, v_2) + \ldots + F(v_K, \infty).
\]

Here, \( (0, v_1) \) is the interval of variable \( \nu \) that is encompassed by history epoch 0, \( (v_1, v_2) \) is that encompassed by epoch 1, and \( (v_K, \infty) \) is that encompassed by epoch \( K \). Within the interval between \( v_i \) and \( v_{i+1} \), the population size is a constant, \( N \). Consequently, these integrals can be evaluated directly. For epochs of finite length,
\[ F(v_i, v_{i+1}) = N \int_{v_i}^{v_{i+1}} e^{Bu} dv \]
\[ = N_i B^{-1} (e^{Bu_{i+1}} - e^{Bu_i}) . \]

For the final epoch, which has infinite length, this becomes
\[ F(v_i, \infty) = N_i \int_{v_i}^{\infty} e^{Bu} dv \]
\[ = -N_i B^{-1} e^{Bu_i} . \]

To recover \( m \) from Equation 16, we must right multiply each of the \( F \)'s by \( \hat{p}(0) \), a process that yields
\[ F(v_i, v_{i+1})\hat{p}(0) = N_i B^{-1} (\hat{p}(v_{i+1}) - \hat{p}(v_i)) \]
\[ F(v_i, \infty)\hat{p}(0) = -N_i B^{-1} \hat{p}(v_i) , \]
where

\[ \hat{p}(v) = e^{Bu} \hat{p}(0) \]

is the result of projecting the initial vector, \( \hat{p}(0) \), backward by \( v \) units of time under the assumption that \( N(t) = 1 \), a constant.

Our computer program uses the projection methods discussed previously to calculate the probability vectors \( \hat{p}(v) \), then subtracts pairs of vectors, and finally applies \( N_i B^{-1} \). This last step is easy. For example, if \( n = 4 \),

\[
\begin{pmatrix}
  u_2 \\
  u_3 \\
  u_4
\end{pmatrix} =
\begin{pmatrix}
  \frac{u_2 + u_3 + u_4}{-\beta_2} \\
  \frac{u_3 + u_4}{-\beta_3} \\
  \frac{u_4}{-\beta_4}
\end{pmatrix},
\]

where \(-\beta_2, -\beta_3, \) and \(-\beta_4\) are the diagonal entries of \( B \).