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

Stephen Wooding*,[1] and Alan Rogers†

Eccles Instititute 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.*, Takahata 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

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

[1]*Corresponding author:* Eccles Instititute of Human Genetics, University of Utah, 15 N. 2030 E., Salt Lake City, UT 84112-5330.
E-mail: swooding@genetics.utah.edu

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

$$p_k(t+1) = \begin{cases} p_k(t)\cdot(1-\alpha_k(t)) + p_{k+1}(t)\cdot\alpha_{k+1}(t), & 1 \leq k < n \\ p_k(t)\cdot(1-\alpha_k(t)), & k = n \end{cases} \tag{2}$$

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

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

$$p_1(t) = 1 - \sum_{i=2}^{n} p_i(t).$$

In matrix notation, Equation 2 becomes

$$p(t+1) = (\boldsymbol{I} + \boldsymbol{A}(t))p(t), \tag{3}$$

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

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

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

$$\frac{dp(t)}{dt} = \boldsymbol{A}(t)p(t), \tag{4}$$

which is solved by

$$p(t) = e^{\int_0^t \boldsymbol{A}(z)\,dz} p(0). \tag{5}$$

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

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

$$c_{i+1} = c_i(\lambda + \alpha_i)/\alpha_{i+1}, \tag{6}$$

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 (6) 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 $\boldsymbol{C}$ of column eigenvectors is upper triangular.

Equation 6 also implies that the column eigenvectors are time invariant: Substitute (1) into (6) 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 $\boldsymbol{C}$ of column eigenvectors is time invariant.

The row eigenvectors of $\boldsymbol{A}$ are defined by $r\boldsymbol{A} = \lambda r$, where λ is an eigenvalue of $\boldsymbol{A}$ and r is the corresponding row eigenvector. This equation can be reexpressed as

$$r_{i-1} = r_i(\lambda + \alpha_i)/\alpha_i, \tag{7}$$

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

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

$$p(t) = \boldsymbol{CP}(t)\boldsymbol{R}p(0), \tag{8}$$

where $\boldsymbol{P}(t)$ is a diagonal matrix whose xth diagonal element is

$$P_x(t) = e^{-\int_0^t \alpha_x(\tau)\,d\tau} \tag{9}$$

and $p(0) = [0, 0, \ldots, 1]$ as described for (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 $N(t)$.

The second row of plots in Figure 1a shows how $p_k(t)$ 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 kth 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 kth 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:

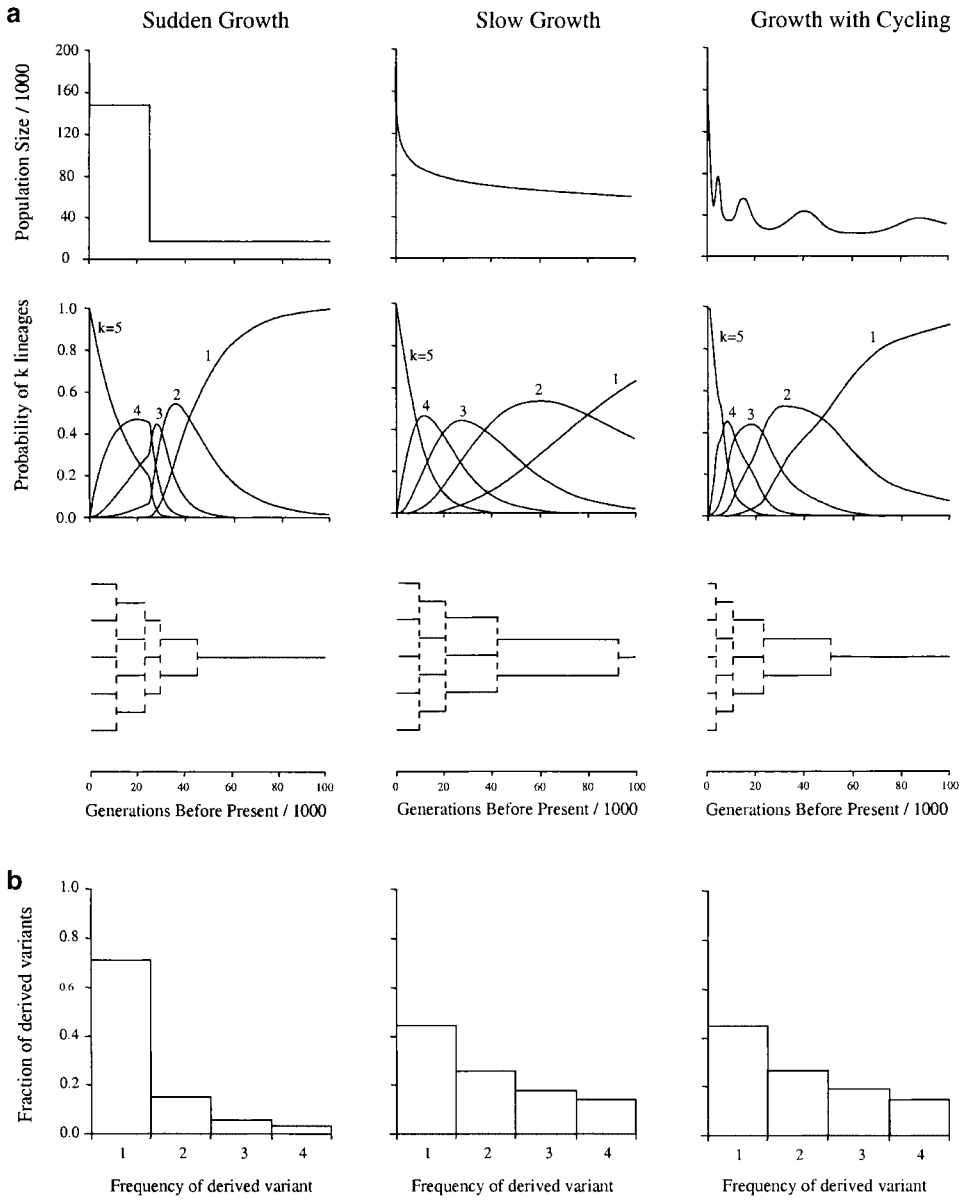

FIGURE 1.—Theoretical expectations. Shown are the relationships between population history, probability of coalescence, expected interval length, and theoretical frequency spectrum under three population histories for samples of $n = 5$ lineages. (a) Top, population size over time; middle, the probability 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 a. These results were generated using the Maple 5.1 software package using default numerical precision.

$$m = \int_0^\infty p(t)\,dt. \tag{10}$$

After substituting Equation 8, this becomes

$$m = \boldsymbol{CER}p(0), \tag{11}$$

where $\boldsymbol{E}$ is a diagonal matrix whose xth diagonal element is

$$E_x = \int_0^\infty e^{-\int_0^\tau \alpha_x(t)\,dt} d\tau \tag{12}$$

(Ross 1997, Chap. 5). The third row of plots in Figure 1a shows the expected length of each coalescent interval under several hypothetical population histories.

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, σ_k, of sites at which the derived allele occurs k times is given by

$$\sigma_k \approx \frac{\sum_{j=2}^{n} j m_j y(j,\ k,\ n)}{\sum_{j=2}^{n} j m_j}, \tag{13}$$

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 $\boldsymbol{C}$ 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 CLN-1.0.1 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 epoch 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

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

where S_k is the number of sites occurring k times in the sample and σ_k is the probability of a variant site occurring k times in the sample. 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 (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 accomodate 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 1

Sampled SNPs

Coding nonsynonymous SNPs (cns)					Coding synonymous SNPs (cs)					Noncoding SNPs near genes (nc)				
WIAF	Gene	Location	*p*	*n*	WIAF	Gene	Location	*p*	*n*	WIAF	Gene	Location	*p*	*n*
10547	CYP21	6p21	1	82	10522	AHC	Unk.	2	86	10561	GRL	Unk.	1	86
10549	FSHR	chr2	3	82	10525	AR	chr2	1	84	10562	PTH	Unk.	3	86
10554	GNRHR	8p21	3	86	10529	CYP11B1	Unk.	3	84	10620	GH1	17q22	1	86
10557	DRL	5q31	1	86	10540	CYP17	Unk.	3	80	10650	HSD3B2	Unk.	2	82
10568	CYP11B1	Unk.	1	86	10548	FSH	11p13	3	86	10656	IGF1	Unk.	1	84
10591	GH1	17q22	1	86	10552	GHR	Unk.	3	86	10657	IGF2	11p15	1	80
10605	GHR	5p13	1	86	10555	GNRHR	8p21	1	86	10660	PC1	Unk.	3	82
10624	CYP11A	Unk.	1	86	10560	GRL	5q31	3	78	10695	PACE	15q25	3	82
10625	FSH	Unk.	1	86	10563	PTH	Unk.	2	86	10700	PTHLH	Unk.	3	86
10638	CYP11B2	8q21	3	86	10566	CGA	Unk.	2	82	10762	DRD2	11q23	3	74
10651	IGF1	Unk.	1	86	10582	CYP21	6p21	1	86	10770	NGFB	1p13	3	74
10667	SHBG	17p13	2	86	10614	FSHR	chr2	1	86	10778	COMT	22q11	3	70
10726	HSD3B2	1p13	1	86	10637	CYP11B2	8q21	3	86	10843	HTR1A	Unk.	3	78
10753	BDNF	11p13	3	72	10658	PACE	15q25	3	80	10849	HTR1DB	6q13	2	78
10780	DRD3	3q13	3	74	10668	SHBG	17p13	1	86	10857	SLC6A1	Unk.	3	72
10791	COMT	22q11	3	70	10724	PRL	6p22	1	86	10864	SLC6A4	Unk.	3	72
10793	DBH	9q34	2	70	10733	PC1	Unk.	1	86	10900	HTR2A	Unk.	1	74
10800	NGFB	1p13	3	74	10759	DRD2	11q23	3	74	10972	HCF2	22q11	2	104
10801	ADORA2	22q11	1	74	10766	DRD5	4p16	3	74	11029	HMGCR	Unk.	1	114
10826	DRD5	4p16	2	74	10768	GRIN1	Unk.	3	74	11078	ANX3	Unk.	2	104
10842	NTRK1	Unk.	1	74	10773	NTRK1	Unk.	3	74	11220	PAI2	18q21	1	104
10846	HTR1D	1p36	1	78	10792	COMT	22q11	3	70	11237	LIPC	Unk.	1	106
10862	SLC6A4	Unk.	1	74	10803	DRD1	5q35	1	74	11346	F5	1q23	1	108
10865	TH	Unk.	3	74	10827	GAP43	Unk.	1	70	11438	GABRB1	Unk.	2	38
10870	HTR1E	6q14	1	74	10848	HTR1DB	6q13	3	74	11490	TBXAS1	Unk.	2	34
10879	CNTF	11q12	1	74	10853	HTR2A	Unk.	3	74	11566	THP0	3q27	3	28
10898	HTR2A	Unk.	2	76	10855	HTR5A	7q36	2	78	11576	F10	Unk.	1	38
10949	CETP	Unk.	3	114	10856	NT3	12p13	3	72	13040	CYP21	6p21	1	82
10952	F2	11p11	1	128	10859	SLC6A3	Unk.	1	58	13068	DYP11B2	8q21	3	56
10958	F2R	Unk.	1	106	10866	TH	Unk.	2	78	13073	CYP11B1	Unk.	2	70
10960	F3	Unk.	1	86	10869	HTR1E	6q14	1	74					
10971	HCF2	22q11	1	106	10888	SLC6A1	Unk.	1	78					
10975	HMGCR	Unk.	1	114	10895	HTR1D	Unk.	1	74					
11004	TFPI	Unk.	1	100	10897	HTR1EL	3p12	1	74					
11020	CLanalog	Unk.	1	76	10904	HTR6	Unk.	1	72					
11030	ITGA2B	17q21	1	112	10945	AT3	1q23	3	106					
11033	ITGB3	Unk.	1	106	10962	F5	1q23	2	126					
11035	LDLR	Unk.	1	126	10970	HCF2	22q11	1	106					
11036	LPL	8p22	1	126	10981	LCAT	16q22	1	128					
11041	PTAFR	Unk.	1	114	10996	LDLR	Unk.	3	126					
11062	F5	1q23	1	108	10997	LPL	8p22	2	126					
11070	FGA	Unk.	3	108	10998	PROC	2q13	3	108					
11071	FGB	Unk.	3	108	11003	TBXA2R	19p13	3	112					
11074	PCI	chr14	2	102	11009	ITGB3	Unk.	3	106					
11082	APOD	3q26	1	108	11017	CETP	Unk.	1	114					
11085	F13A1	Unk.	1	108	11022	CLanalog	Unk.	2	76					
11102	F11	Unk.	1	108	11025	F2	11p11	1	126					
11117	F7	1q31	1	104	11027	F2R	Unk.	1	106					
11179	F13B	1q31	1	108	11031	ITGA2B	17q21	2	112					
11200	ANX3	Unk.	1	108	11045	F10	Unk.	3	104					
11203	F10	Unk.	1	108	11048	F11	Unk.	3	106					
11215	LIPC	Unk.	3	104	11051	F13A1	Unk.	2	108					
11227	SELP	1q23	1	106	11056	F13B	1q31	3	106					
11236	TBXAS1	Unk.	1	106	11067	F7	13q34	2	108					
11294	VLDLR	Unk.	1	106	11079	ANX3	Unk.	1	108					
11301	PAI2	Unk.	1	96	11099	APOD	3q26	1	108					
11304	GP1BA	17p12	1	98	11193	FGA	4q28	1	108					
11311	MPL	Unk.	1	104	11207	FGB	Unk.	1	106					
11325	CD36	Unk.	1	90	11216	LIPC	Unk.	3	106					
11339	PAI1	7q21	1	106	11218	PAI1	7q21	1	106					
					11224	PROS1	3p11	3	106					
					11232	SELP	1q23	3	98					
					11241	VLDLR	Unk.	1	106					
					11244	PAI2	18q21	1	98					
					11326	GP9	3q21	1	86					
					11332	TBXAS1	7q32	1	106					
					11431	GABRB1	Unk.	1	40					
					11449	HTR1A	5q11	2	28					

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.

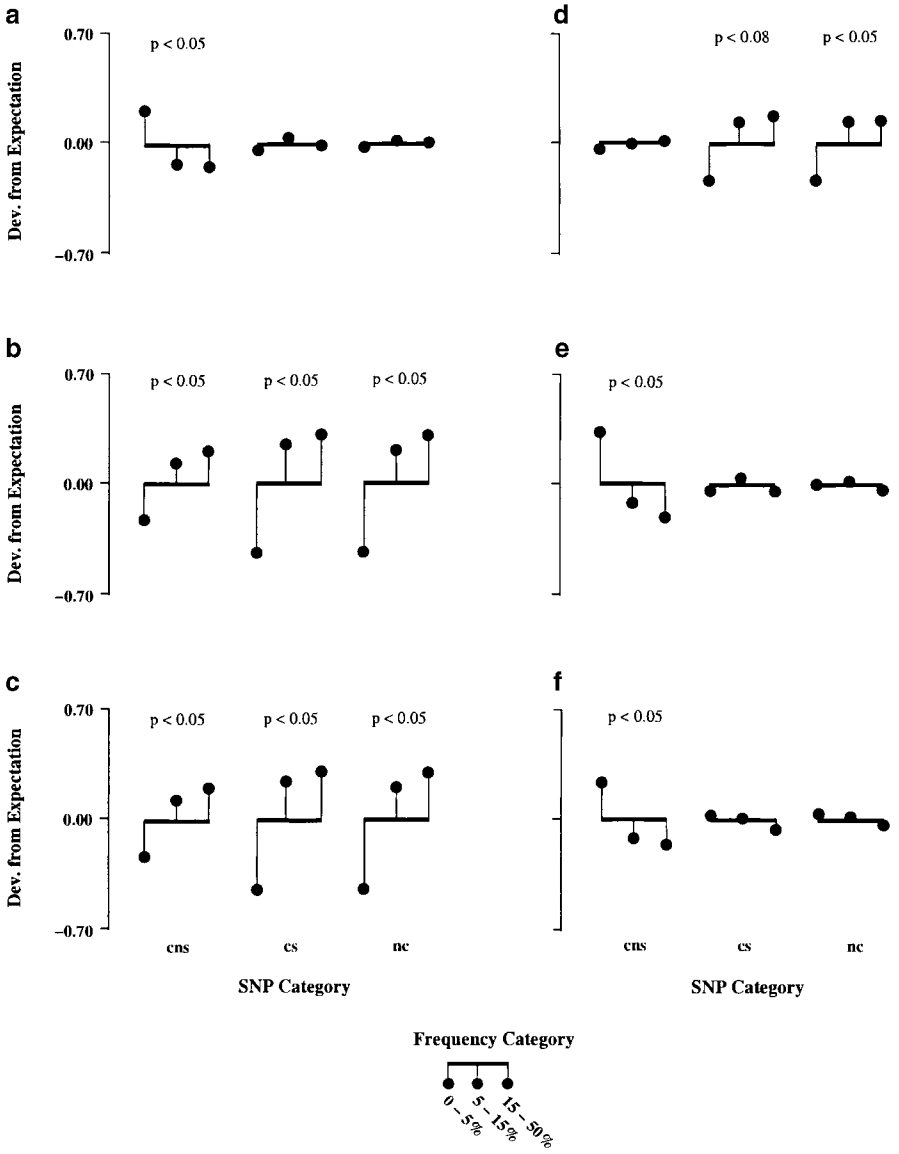

FIGURE 2.—Pitch plot of deviations from expectation. The deviation of each frequency category (0–5%, 5–15%, and 15–50%) from expectation in each SNP category (cns, cs, and nc is shown). Deviations given are the mean difference from expectation across all sample sizes within a SNP category, since not 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

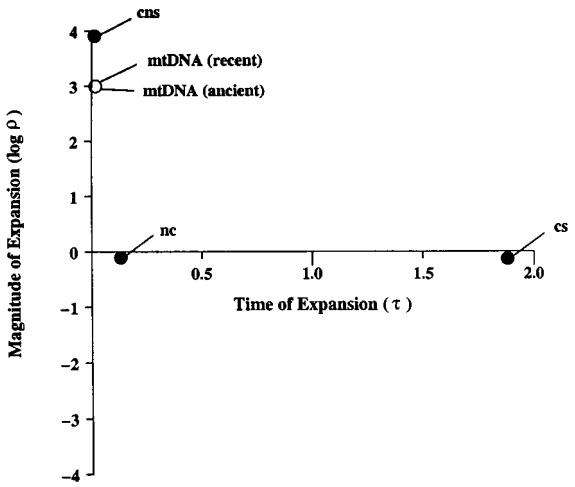

FIGURE 3.—Maximum-likelihood parameter estimates. Open circles show the parameters of two alternative hypotheses estimated from mitochondrial DNA (see MODEL).

three parameters just defined, we work instead with two: $\tau = T/N_0$ and $\rho = N_0/N_1$. Here, 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 1 d.f. in likelihood-ratio tests. Maximum-likelihood estimates of ρ and τ were obtained for each SNP category by iterating over a series of values of ρ and τ.

Five hypotheses 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 (Figure 2). Then the maximum likelihood of each SNP category was tested against the category's likelihood of three alternatives: (a) stationarity, (b) the most recent population expansion not excluded by ROGERS (1995; $\tau = 4.7 \times 10^{-3}$, $\rho = 1000$), and (c) the most ancient population expansion not excluded by ROGERS (1995; $\tau = 2.1 \times 10^{-2}$, $\rho = 1000$); (see Figure 2).

CARGILL *et al.* (1999) 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. FAY *et al.* (2001) found differences between the frequency spectra of synonymous and nonsynonymous changes in the CARGILL *et al.* (1999) data set as well. Our parameter estimates confirm these results (Figure 3). The cns category had maximum-likelihood parameters implying recent population growth under the assumption of selective neutrality ($\tau = 8.6 \times 10^{-6}$ and $\rho = 9900$), and 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 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 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 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

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

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

BULMER, M. G., 1979 *Principles of Statistics.* Dover Publications, New York.

CARGILL, M., D. ALTSHULER, J. IRELAND, P. SKLAR, K. ARDLIE *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231–238.

EDWARDS, A. W. F., 1992 *Likelihood.* The Johns Hopkins University Press, Baltimore.

FAY, J. C., G. J. WYCKOFF and C.-I WU, 2001 Positive and negative selection on the human genome. Genetics **158:** 1227–1234.

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59:** 139–147.

FU, Y.-X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172–197.

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

FU, Y.-X., and W.-H. LI, 2001 Coalescing into the 21st century: an overview and prospects of coalescent theory. Theor. Popul. Biol. **56:** 1–10.

HAIBLE, B., 2000 *CLN: Class Library for Numbers Version 1.0.1.* Computer program distributed by the author, http://clisp.cons.org/~haible/packages-cln.html.

HALUSHKA, M. K., J.-B. FAN, K. BENTLEY, L. HSIE, N. SHEN *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239–247.

HARPENDING, H. C., and A. R. ROGERS, 2000 Genetic perspectives on human origins and differentiation. Annu. Rev. Genomics Hum. Gen. **1:** 361–385.

HARPENDING, H. C., M. A. BATZER, M. GURVEN, L. B. JORDE and A. R. ROGERS, 1998 Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95:** 1961–1967.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Series in Evolutionary Biology*, Vol. 7, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford.

KINGMAN, J. F. C., 1982a *The Coalescent.* Stoc. Proc. Appl. **13:** 235–248.

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

MOLER, C. B., and C. F. V. LOAN, 1978 Nineteen dubious ways to compute the exponential of a matrix. SIAM Rev. **20:** 801–836.

PRZEWORSKI, M., B. CHARLESWORTH and J. D. WALL, 1999 Genealogies and weak purifying selection. Mol. Biol. Evol. **16:** 246–252.

Rogers, A. R., 1995 Genetic evidence for a Pleistocene population explosion. Evolution **49:** 608–615.
Rogers, A. R., 2001 Order emerging from chaos in human evolutionary genetics. Proc. Natl. Acad. Sci. USA **98:** 779–780.
Ross, S., 1997 *A First Course in Probability*, Ed. 5. Prentice Hall, Upper Saddle River, NJ.
Sherry, S. T., H. C. Harpending, M. A. Batzer and M. Stoneking, 1997 Alu evolution in human populations: Using the coalescent to estimate effective population size. Genetics **147:** 1977–1982.
Sherry, S. T., M. Ward and K. Sirotkin, 2000 Use of molecular variation in the NCBI dbSNP database. Hum. Mutat. **15:** 68–75.
Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan *et al.*, 2001 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. **29:** 308–311.
Stewart, W. J., 1994 *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, Princeton, NJ.
Sunyaev, S. R., W. Lathe, V. E. Ramensky and P. Bork, 2000 SNP frequencies in human genes: an excess of rare alleles and differing modes of selection. Trends Genet. **16:** 335–337.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.
Takahata, N., 1988 The coalescent in two partially isolated diffusion populations. Genet. Res. **52:** 213–222.
Terwilliger, J. D., S. Zöllner, M. Laan and S. Pääbo, 1998 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.
Wooding, S., 1999 *TreeToy Coalescent Simulation Version 1.0b*. Computer program distributed by the author, http://www.anthro.utah.edu/popgen/programs/TreeToy

Communicating editor: Y.-X. Fu

APPENDIX A: THE EXPECTED SITE FREQUENCY SPECTRUM

We assume that mutations are rare enough that the possibility of multiple mutations in a single gene genealogy can be ignored. Let A_j denote the event that exactly one mutation occurs within the portion of the genealogy containing j lineages, B the event that exactly one mutation occurs within the genealogy 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

$$\sigma_k = \sum_{j=2}^{n} \Pr[A_j|B]\, y(j, k, n). \tag{14}$$

Using Bayes' rule,

$$\Pr[A_j|B] = \frac{\Pr[B|A_j]\Pr[A_j]}{\Pr[B]}. \tag{15}$$

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

To calculate the unconditional probability of A_j, let L_j denote the length of the jth coalescent interval in a random gene tree. Then jL_j is the total branch length associated with that coalescent interval. The conditional probability, given L_j, that a single mutation occurs within this interval is

$$\Pr[A_j|L_j] = \mu j L_j e^{-\mu j L_j} \approx \mu j L_j,$$

where μ is the mutation rate, and we assume a Poisson distribution of mutations. The approximation here assumes that μ^2 is negligible in comparison to μ. The unconditional probability of A_j is

$$\Pr[A_j] \approx E[\mu j L_j] = \mu j m_j,$$

where m_j is the expected value of L_j.

A similar argument gives

$$\Pr[B] \approx E\left[\mu \sum_{j=2}^{n} j L_j\right] = \mu \sum_{j=2}^{n} j m_j,$$

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

$$\beta_i = N(t)\cdot\alpha_i(t) = i(i-1)/2$$

and

$$\boldsymbol{B} = \begin{pmatrix} -\beta_2 & \beta_3 & & \\ & -\beta_3 & \ddots & \\ & & \ddots & \beta_n \\ & & & -\beta_n \end{pmatrix}.$$

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

$$\begin{aligned} m &= \int_0^\infty p(t)\, dt \\ &= \int_0^\infty e^{\boldsymbol{B}\int_0^t N^{-1}(z)\, dz} dt\; p(0) \\ &= \int_0^\infty N(v)\, e^{\boldsymbol{B}v} dv\; p(0) \\ &= F(0, \infty)\, p(0), \end{aligned} \tag{16}$$

where $v = \int_0^t N^{-1}(z)\, dz$ and

$$F(a, b) = \int_a^b N(v)\, e^{\boldsymbol{B}v} dv.$$

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 v that is encompassed by history epoch 0, (v_1, v_2) is that encompassed by epoch 1, and (v_K, ∞) is that encompassed by epoch K. Within the interval between v_i and v_{i+1}, the population size is a constant, N_i. Consequently, these integrals can be evaluated directly. For epochs of finite length,

$$F(v_i, v_{i+1}) = N_i \int_{v_i}^{v_{i+1}} e^{\mathbf{B}v} dv$$
$$= N_i \mathbf{B}^{-1} (e^{\mathbf{B}v_{i+1}} - e^{\mathbf{B}v_i}).$$

For the final epoch, which has infinite length, this becomes

$$F(v_K, \infty) = N_K \int_{v_K}^{\infty} e^{\mathbf{B}v} dv$$
$$= -N_K \mathbf{B}^{-1} e^{\mathbf{B}v_K}.$$

To recover m from Equation 16, we must right multiply each of the F's by $p(0)$, a process that yields

$$F(v_i, v_{i+1}) p(0) = N_i \mathbf{B}^{-1} (\tilde{p}(v_{i+1}) - \tilde{p}(v_i))$$
$$F(v_K, \infty) p(0) = -N_K \mathbf{B}^{-1} \tilde{p}(v_K),$$

where

$$\tilde{p}(v) = e^{\mathbf{B}v} p(0)$$

is the result of projecting the initial vector, $p(0)$, 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 $\tilde{p}(v)$, then subtracts pairs of vectors, and finally applies $N_i \mathbf{B}^{-1}$. This last step is easy. For example, if $n = 4$,

$$\mathbf{B}^{-1} \begin{pmatrix} u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} \dfrac{u_2 + u_3 + u_4}{-\beta_2} \\ \dfrac{u_3 + u_4}{-\beta_3} \\ \dfrac{u_4}{-\beta_4} \end{pmatrix},$$

where $-\beta_2$, $-\beta_3$, and $-\beta_4$ are the diagonal entries of $\mathbf{B}$.