Nonequilibrium Migration in Human History

John Wakeley

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138

Manuscript received May 6, 1999
Accepted for publication August 12, 1999

ABSTRACT

A nonequilibrium migration model is proposed and applied to genetic data from humans. The model assumes symmetric migration among all possible pairs of demes and that the number of demes is large. With these assumptions it is straightforward to allow for changes in demography, and here a single abrupt change is considered. Under the model this change is identical to a change in the ancestral effective population size and might be caused by changes in deme size, in the number of demes, or in the migration rate. Expressions for the expected numbers of sites segregating at particular frequencies in a multideme sample are derived. A maximum-likelihood analysis of independent polymorphic restriction sites in humans reveals a decrease in effective size. This is consistent with a change in the rates of migration among human subpopulations from ancient low levels to present high ones.

THE early reports of mitochondrial DNA (mtDNA) haplotype diversity among humans indicated a recent large increase in population size starting from a small number of founders (Cann *et al.* 1987; Vigilant *et al.* 1991). This pattern, which is marked by an increase in rare variants, has not subsequently been observed for nuclear loci. Instead, nuclear loci show an excess of polymorphic sites segregating at intermediate frequencies (Hey 1997). This is illustrated by Tajima's (1989) statistic, *D*, which is positive, although not significantly so, for four extensive nuclear DNA sequence data sets (Harding *et al.* 1997; Clark *et al.* 1998; Zietkiewicz *et al.* 1998; Harris and Hey 1999). In addition, single nucleotide polymorphism (SNP) data do not show evidence of population growth (Nielsen 2000). Here, I consider a subdivided population model, which allows for a single change in the effective size of the population, and apply the results to restriction fragment length polymorphism (RFLP) data from a worldwide sample of humans (Bowcock *et al.* 1987; Matullo *et al.* 1994; Poloni *et al.* 1995). The data provide strong evidence for a change in demography: a decrease in effective size. This is most likely due to a shift from a more ancient subdivided population to one with less structure today.

Numerous models of subdivision with migration have been proposed to explain patterns of genetic variation among local populations. The most important distinction among these is that some models, such as Kimura and Weiss' (1964) stepping-stone model, generate isolation by distance (Wright 1943), whereas others, such as Wright's (1931) island model, do not. The model considered here is of the latter type. Wakeley (1998) studied the ancestral genealogical process for samples from a *D*-deme version of the finite island model, in which the number of demes is assumed to be large. In the finite island model (Maruyama 1970; Latter 1973), all demes exchange migrants regardless of their geographic location. The single-generation probability of migration is assumed to be the same between each pair of demes. Specifically, m is the probability that a gene came from one of the other $D - 1$ demes to its present one in the previous generation. This model is realistic for species in which individuals choose either to stay put or to move to a location picked uniformly from the entire species range. In certain situations, it may also apply when individuals choose either to migrate a great distance or not at all.

When the number of demes is large, the genealogical history of a sample taken from such a population can be divided into two parts, which I will call the scattering phase and the collecting phase; see Figure 1. The scattering phase is the very recent history of the sample, during which coalescent and migration events bring it rapidly to the start of the second, and much longer, collecting phase of the history. The collecting phase begins when each ancestral lineage is in a separate deme and is characterized by migration events that move lineages from deme to deme, the great majority of which are unoccupied. Eventually, a migration event will place a pair of lineages into the same deme, at which time they can coalesce. The ability to break the genealogy into these two parts, and to consider them separately, depends only on the number of demes being large. When this is true, the time spent in the scattering phase can be ignored because the collecting phase dominates the history (Wakeley 1998). A similar separation of time

Address for correspondence: 288 Biological Laboratories, 16 Divinity Ave., Cambridge, MA 02138.
E-mail: jwakeley@oeb.harvard.edu

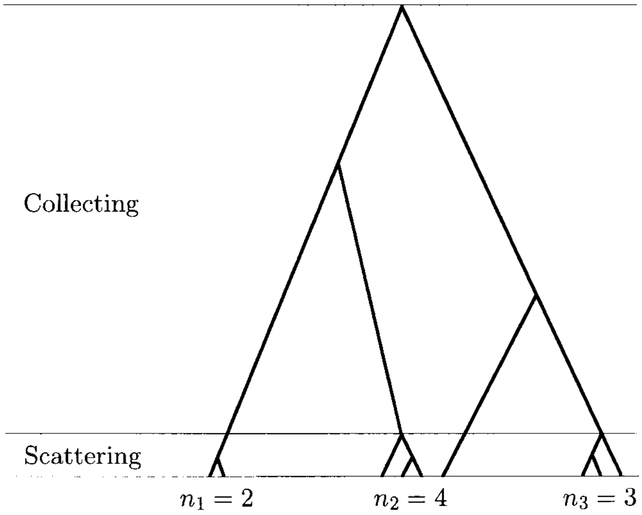

Figure 1.—A hypothetical genealogy of a sample from $d = 3$ demes under the model. In this case there was just one migration event during the scattering phase, and this was in the sample from deme 2. Thus $n_1' = 1$, $n_2' = 2$, $n_3' = 1$, and $n' = d + 1 = 4$.

scales has been found in the study of genealogies under partial selfing (Nordborg and Donnelly 1997).

Consider first the equilibrium version of this island model. We assume that each of D demes is of constant diploid size N, but any results will also apply to haploids if N is placed by $N/2$. This model is characterized by two parameters: $\theta = 4NDu$, in which u is the neutral mutation rate, and $M = 2NDm/(D - 1)$. When D is large, M becomes simply $2Nm$. A sample of n items, taken from d different demes, which are arbitrarily labeled 1 through d, is denoted $\boldsymbol{n} = (n_1, n_2, \ldots, n_d)$, where n_i is the sample size from the ith deme. Of course $n = \sum_{i=1}^{d} n_i$. This notation for the sample differs from that used in Wakeley (1998). We assume that θ is finite even though D and N are large. Thus, we have $D \gg n$ in addition to the usual coalescent condition, $N \gg n$. Simulations show that D does not have to be terribly big for the infinite-demes approximation to hold (Wakeley 1998).

The scattering phase of the genealogy takes the single-deme sample, n_i, through a series of migration and coalescent events to a new configuration where each remaining lineage is in a separate deme. The number of these recent ancestral lineages, n_i' $(1 \leq n_i' \leq n_i)$, depends on the sample size and the migration rate. It is important to note that each of these n_j' lineages is in a separate deme at the end of the scattering phase. The distribution of n_i' is given by

$$P[n_i'|n_i] = \frac{S_{n_i}^{(n_i')}\ (2M)^{n_i'}}{(2M)_{(n_i)}} \qquad (1)$$

(Wakeley 1998; Equation 32), where $x_{(r)} = x(x + 1) \ldots (x + r - 1)$, and $S_j^{(i)}$ is an unsigned Stirling number of the first kind. Abramowitz and Stegun (1964) define Stirling numbers and list many of their properties. The notation $|s(i, j)|$ was used instead of $S_j^{(i)}$ in Wakeley (1998). Slatkin's (1989) Equation 2 for the probability that n alleles are descended from nonimmigrants is a special case of Equation 1 above. (Note that there is a typographical error in Slatkin's equation; there should be a $2Nm$ in the numerator.)

Thus, the distribution of the number of ancestral lineages of each deme's sample at the start of the collecting phase is the same as the distribution of the number of alleles in the Ewens sampling formula (Ewens 1972; Karlin and McGregor 1972) but with mutation replaced by migration. There is one major difference: the n_i' items at the start of the collecting phase represent exactly that number of ancestral lineages, whereas the alleles counted by the Ewens sampling formula do not. Let $\boldsymbol{n}' = (n_1', n_2', \ldots, n_d')$ and $n' = \sum_{i=1}^{d} n_i'$. Because events in different demes are independent, the joint distribution of all the n_i' is given by

$$P[\boldsymbol{n}'|\boldsymbol{n}] = \prod_{i=1}^{d} P[n_i'|n_i]. \qquad (2)$$

The collecting phase of the history then begins with n' $(d \leq n' \leq n)$ lineages, and each of these is in a separate deme. Equation 2 above follows from Equation 36 in Wakeley (1998).

The genealogical history of these n' ancestral lineages during the collecting phase is a coalescent (Kingman 1982a, b) when time is measured in units of twice the effective population size

$$N_e = ND\left[1 + \frac{1}{2M}\right] \qquad (3)$$

(see appendix a). Nei and Takahata (1993) give a similar formula for the special case of two sampled sequences under the finite island model, which they noted was implicit in earlier work, and which quickly converges on Equation 3 as the number of demes increases. Thus, the duration of the collecting phase is at least of order ND generations, whereas the duration of the scattering phase is of order N generations. This is why the time of the scattering phase can be ignored when D is large.

Because the collecting phase is a coalescent, it is relatively simple to allow for changes in effective population size. Here I consider the possibility of a single abrupt change in demography at generation t in the past. Thus, the collecting phase itself is divided into two parts. Before time t, the population is assumed to have been of effective size $N_{eA} = N_A D_A[1 + 1/(2M_A)]$, which can be compared to Equation 3. Thus, the change in demography might have been caused by a change in deme size, in the number of demes, or in the migration rate. Part of the flexibility of the model is that the change might have been any combination of these. If the ancestral population was panmictic we have the demographic model that Takahata (1995) proposed for human history, but without extinction and recolonization.

This nonequilibrium model necessitates two parameters in addition to θ and M. The first is the time of the event, $T = t/(2ND)$, measured in units of $2ND$ generations. The second is $R = N_{eA}/(ND)$, the ratio of the ancestral effective size to the current total population size. Thus, values of $R/[1 + 1/(2M)]$ greater than one indicate that the ancestral effective size was larger than the current effective size. Given an estimate of R and assuming that $ND = kN_A D_A$, where k is some constant, we can retrieve $\hat{M}_A = 1/[2(k\hat{R} - 1)]$. This is of considerable interest below in the application of this model to human history. Further, given estimates of both R and M and assuming $M_A = kM$, we could estimate $(N_A D_A)/(ND)$ as $\hat{R}/[1 + 1/(2k\hat{M})]$, which in conjunction with $\hat{\theta}$ could be used to estimate θ_A.

Neutral mutations are modeled as a Poisson process along the branches of the genealogy (Hudson 1983, 1990). When a mutation happens it is assumed to occur at a previously unmutated site. This assumption is shared by Watterson's (1975) infinite sites model without recombination and Kimura's (1969) model, which assumes free recombination between sites. We apply these models to genetic data when the mutation rate per site is so small that multiple changes at any one site are unlikely. Watterson (1975) and others have studied the distribution of segregating sites under the neutral infinite sites model, and Sawyer and Hartl (1992) have done so for both selection and neutrality under Kimura's (1969) model. Under neutrality, both models make the same prediction about the expected number of segregating sites in a sample, either taken as a whole or broken into categories by frequency (Hudson 1983; Tajima 1989; Sawyer and Hartl 1992; Fu and Li 1993). In other words, expectations derived under particular assumptions about the recombination rate are valid for all rates of recombination. Here, I assume Watterson's (1975) model. The effect of linkage and recombination is on the variances, and these are largest under when the recombination rate is lower.

Under these mutations-at-unique-sites models, each segregating site will divide the sample into two groups: one that has inherited the mutant base and one that still shows the ancestral base. Let $\boldsymbol{i} = (i_1, i_2, \ldots, i_d)$ be the counts of mutant bases at a single polymorphic site in the sample $\boldsymbol{n} = (n_1, n_2, \ldots, n_d)$. Further, let z_i be the number of sites in the sample that show the pattern $\boldsymbol{i} = (i_1, i_2, \ldots, i_d)$. In this article, I calculate

$$L(M, R, T; \boldsymbol{i}) = \Pr(\boldsymbol{i}|\text{site is variable}; M, R, T), \quad (4)$$

the likelihood of a particular mutant site pattern given that a site is polymorphic. I do this by first deriving $E(z_i)$, the expected number of sites that show the pattern $\boldsymbol{i} = (i_1, i_2, \ldots, i_d)$, for a sample of sequences under Watterson's infinite sites model. As discussed above, this expectation will hold for arbitrary rates of recombination. Sawyer and Hartl (1992) showed that with free recombination between sites, z_i is Poisson distributed with parameter $E(z_i)$, so

$$L(M, R, T; \boldsymbol{i}) = \frac{E(z_i)}{\sum_i E(z_i)}. \quad (5)$$

The denominator in Equation 5 is simply equal to the expected number of segregating sites in the sample. As $E(z_i)$ is linear in θ, the likelihood depends only on the demographic parameters, M, R, and T, and not on the mutation parameter, θ. Unless an outgroup sequence is available, it is impossible to distinguish the mutant from the ancestral base. Thus, except for the special case of $i_j = n_j/2$ for all j, the likelihood of a particular site pattern is $L(M, R, T; \boldsymbol{i}) + L(M, R, T; \boldsymbol{i}^c)$, where $\boldsymbol{i}^c = (n_1 - i_1, n_2 - i_2, \ldots, n_d - i_d)$. With this modification, Equation 5 is suitable for the analysis of SNP and RFLP data. If different sites can be assumed to be independent of each other, as they can for the human RFLP data analyzed below, the likelihood of the entire data is the product of the likelihoods for individual sites.

THEORY

The approach taken to deriving $E(z_i)$ is to condition on the numbers of lineages and the mutant counts at the start of the collecting phase and at the time of the abrupt change in demography. Thus, we have

$$\begin{aligned} E(z_i) = &\sum_{n'_1=1}^{n_1}\sum_{n'_2=1}^{n_2} \cdots \sum_{n'_d=1}^{n_d} P[\boldsymbol{n}'|\boldsymbol{n}] \\ &\times \sum_{i'=\max(1,i-n+n')}^{\min(n'-1,i)} \Pr[\boldsymbol{i}|i', \boldsymbol{n}']E_{n'}(z_{i'}), \quad (6) \end{aligned}$$

in which $P[\boldsymbol{n}'|\boldsymbol{n}]$ is given by Equation 2, and where $\max(i, j)$ and $\min(i, j)$ refer to the larger and smaller of i and j, respectively. The conditioning on the number of lineages and mutant counts at the time of the abrupt change in demography is imbedded in $E_{n'}(z_{i'})$—see below—which is the expected number of sites where the mutant is in i' copies in the collecting-phase sample, n'. This is multiplied by $\Pr[\boldsymbol{i}|i', \boldsymbol{n}']$, the probability that such a site will appear in counts $(i_1, i_2, \ldots, i_d)$ in the original sample $(n_1, n_2, \ldots, n_d)$, and the result is averaged over all possible collecting-phase samples and mutant counts.

Consider the genealogy of the n' lineages during the collecting phase of the history. Because this is a coalescent, the demographic change in the model is identical to a single, instantaneous change in effective population size from $N_{eA} = N_A D_A[1 + 1/(2M_A)]$ to N_e of Equation 3. Wakeley and Hey (1997) found expressions for the expected site frequencies under this size-change model by summing the numbers expected to have occurred before and after the change. Their Equation 20 is adapted to the present situation by replacing their θ_1 with $\theta[1 + 1/(2M)]$, and their θ_A with θR, and by rescaling time according to Equation 3. This gives

$$E_{n'}(z_{i'}) = \theta\left(1 + \frac{1}{2M}\right) \times \left\{\frac{1}{i'} + \left[\frac{R}{1 + 1/2M} - 1\right] \times \sum_{n''=2}^{n'} g_{n'n''}(\tau_m) \sum_{i''=\max(1,i'-n'+n'')}^{\min(n''-1,i')} P_c(i'|i'', n'')\frac{1}{i''}\right\}. \tag{7}$$

The function $g_{n'n''}(\tau_m)$ is the probability that there are n'' ancestral lineages, of the n' that began the collecting phase, still present at time $\tau_m = T/[1 + 1/2M]$, measured in units of twice the effective size (Equation 3). This is given by Equations 6.1 and 6.2 in Tavaré (1984),

$$g_{ij}(\tau) = \sum_{k=j}^{i} \frac{e^{-k(k-1)\tau/2}(2k - 1)(-i)^{k-j} j_{(k-1)} i_{[k]}}{j!(k - j)! i_{(k)}}, \quad 2 \leq j \leq i \tag{8}$$

$$g_{i1}(\tau) = 1 - \sum_{k=2}^{i} \frac{e^{-k(k-1)\tau/2}(2k - 1)(-i)^{k} i_{[k]}}{i_{(k)}}, \tag{9}$$

in which $i_{(k)} = i(i + 1) \ldots (i + k - 1)$ and $i_{[k]} = i(i - 1) \ldots (i - k + 1)$. An alternative form of Equations 8 and 9 was derived by Takahata and Nei (1985). When the migration rate is small, τ_m is small, which means that the change in demography appears more recent, and when the migration rate is high, τ_m approaches T.

$P_c(i'|i'', n'')$ in Equation 7 is given by expression 18 in Wakeley and Hey (1997), or equivalently by

$$P_c(i'|i'', n'') = \frac{\binom{n' - i' - 1}{n'' - i'' - 1}\binom{i' - 1}{i'' - 1}}{\binom{n' - 1}{n'' - 1}} \tag{10}$$

(Slatkin 1996). This is the probability that a mutant in i'' copies among n'' lineages (at time T) grows to i' copies, where $i'' \leq i' \leq i'' + (n' - n'')$, among the n' lineages at the start of the collecting phase. The subscript in $P_c(i'|i'', n'')$ stands for coalescent, and is necessary to distinguish it from a related probability derived below.

Equation 7 applies to the unobserved ancestral sample at the start of the collecting phase. This is related to the observable sample in Equation 6 by the probability that a mutant found in i' copies in the collecting phase sample, n', will appear in $(i_1, i_2, \ldots, i_d)$ copies from each deme in the multideme sample $(n_1, n_2, \ldots, n_d)$. This probability, $\Pr[\boldsymbol{i}|i', \boldsymbol{n}']$, is derived as follows. First, the distribution of the i' mutant copies among the sequences that make up the ancestral sample, $(n'_1, \ldots, n'_d)$ is given by the multivariate hypergeometric distribution,

$$\Pr[\boldsymbol{i}'|i', \boldsymbol{n}'] = \frac{\prod_{j=1}^{d} \binom{n'_j}{i'_j}}{\binom{n'}{i'}}, \tag{11}$$

in which $i' = \Sigma_{j=1}^{d} i'_j$ and $n' = \Sigma_{j=1}^{d} n'_j$. This is a straightforward extension of Wakeley and Hey's (1997) Equation 23. Note that $0 \leq i'_j \leq n'_j$; the ancestral samples of some demes may not be polymorphic.

Next, consider how mutants with counts $(i'_1, \ldots, i'_d)$ in the collecting phase sample, $(n'_1, \ldots, n'_d)$, become ones with counts $(i_1, \ldots, i_d)$ in the original sample, $(n_1, \ldots, n_d)$. As the scattering phase occurs independently for each deme—*e.g.*, see Equation 2—this probability, which is called $P_s(i_j|i'_j, n'_j)$, is calculated separately for each deme and the results multiplied together. appendix b shows that

$$P_s(i_j|i'_j, n'_j) = \frac{\binom{n_j}{i_j} S_{i_j}^{(i'_j)} S_{n_j - i_j}^{(n'_j - i'_j)}}{\binom{n'_j}{i'_j} S_{n_j}^{(n'_j)}}. \tag{12}$$

The subscript in $P_s(i_j|i'_j, n'_j)$ stands for scattering and serves to distinguish this probability from $P_c(i'|i'', n'')$ of Equation 10 above.

Thus, we have the probability that a mutant in i' copies in the collecting phase sample is distributed as $(i_1, \ldots, i_d)$ in the starting sample, $(n_1, \ldots, n_d)$,

$$\Pr[\boldsymbol{i}|i', \boldsymbol{n}'] = \sum_{\boldsymbol{i}'} \Pr[\boldsymbol{i}'|i', \boldsymbol{n}'] \prod_{j=1}^{d} P_s(i_j|i'_j, n'_j), \tag{13}$$

where the sum is taken over all patterns, $\boldsymbol{i}'$, such that $i' = \Sigma_{j=1}^{d} i'_j$ and $\max(0, i_j - n_j + n'_j) \leq i'_j \leq \min(i_j, n'_j)$ for all j. The sum of Equation 13 over all $\boldsymbol{i}$ is equal to one. We are now able to compute, using Equation 6, the expected number of sites that show the mutant pattern $\boldsymbol{i} = (i_1, i_2, \ldots, i_d)$. Finally, the likelihood of a particular mutant pattern at a single site is given by Equation 5, from which θ drops out. Again, when we cannot distinguish the mutant from the ancestral state, the likelihood is just the sum of the likelihoods of the mutant pattern, $\boldsymbol{i}$, and its complement, $\boldsymbol{i}^c = \boldsymbol{n} - \boldsymbol{i}$.

Equation 6 becomes computationally prohibitive as the sample size per deme and the number of demes sampled increase. A great deal of time is saved by noting that the denominator of Equation 5 is equivalent to the expected number of segregating sites in the sample, and that is more efficiently computed as

$$E(S) = \sum_{n'_1=1}^{n_1} \sum_{n'_2=1}^{n_2} \ldots \sum_{n'_d=1}^{n_d} P[\boldsymbol{n}'|\boldsymbol{n}] E_{n'}(S), \tag{14}$$

where $n' = \Sigma_{j=1}^{d} n'_j$ and

TABLE 1

The subsampled data

Locus	i_1	i_2	i_3	i_4	Locus	i_1	i_2	i_3	i_4
1	3	7	5	2	23	5	1	4	4
2	4	5	2	1	24	7	6	6	4
3	2	0	1	2	25	3	6	5	6
4	0	0	0	1	26	6	7	7	7
5	4	6	7	4	27	3	1	6	3
6	0	3	7	2	28	1	4	4	3
7	3	3	7	7	29	7	7	6	7
8	6	3	3	4	30	3	1	4	6
9	2	0	1	0	31	1	6	2	4
10	1	7	7	5	32	1	4	5	1
11	7	7	3	5	33	6	0	7	4
12	3	2	6	3	34	5	5	7	7
13	7	6	6	7	35	4	7	7	4
14	1	0	1	0	36	7	3	5	5
15	5	4	7	6	37	4	6	7	1
16	6	7	7	6	38	4	3	4	4
17	3	0	7	5	39	3	4	6	6
18	6	7	4	3	40	3	7	0	1
19	7	7	7	6	41	4	7	7	5
20	5	7	7	5	42	3	3	6	6
21	0	0	1	0	43	2	7	3	6
22	7	7	7	4	44	3	2	4	3

For all loci and all demes, $n_j = 7$. The demes are numbered 1, Japan; 2, New Guinea; 3, Senegal; and 4, Italy.

$$E_{n'}(S) = \theta\left(1 + \frac{1}{2M}\right) \times \left\{\sum_{i'=1}^{n'-1}\frac{1}{i'} + \left[\frac{R}{1 + 1/2M} - 1\right] \times \sum_{n''=2}^{n'} g_{n'n''}(\tau_{\mathrm{m}})\sum_{i''=1}^{n''-1}\frac{1}{i''}\right\}, \qquad (15)$$

which is the sum of expression (7) over all possible mutant counts.

APPLICATION TO HUMAN RFLP DATA

The method was applied to some data derived from the extensive RFLP surveys of Bowcock *et al.* (1987), Matullo *et al.* (1994), and Poloni *et al.* (1995). These data overlap with those analyzed by Nielsen *et al.* (1998). I considered data from four localities, which I call demes: Japan, New Guinea, Senegal, and Italy. Joanna Mountain kindly supplied these data, which consisted of $n_j = 20$, for $j = 1, 2, 3, 4$, at 45 independent loci. Due to the computational burden of calculating the likelihood 6, I made a smaller data set of $n_j = 7$, for all j, by randomly sampling haplotypes within demes (without replacement). One locus was lost because it was not polymorphic in the subsample. This left 44 independent loci for the analysis, and these data are shown in Table 1. I also performed all of the analyses described below on subsamples of size $n_j = 3$ and $n_j = 5$, and found nearly identical results. The full data sent to me by Joanna Mountain included two more African sample locations: Central African Republic and Zaire. I excluded these two samples for reasons discussed below. However, I also analyzed just the three African samples and the results were identical in overall pattern to what is reported here for the four widely separated samples. The biggest difference was that the estimate of the current migration rate was almost five times larger.

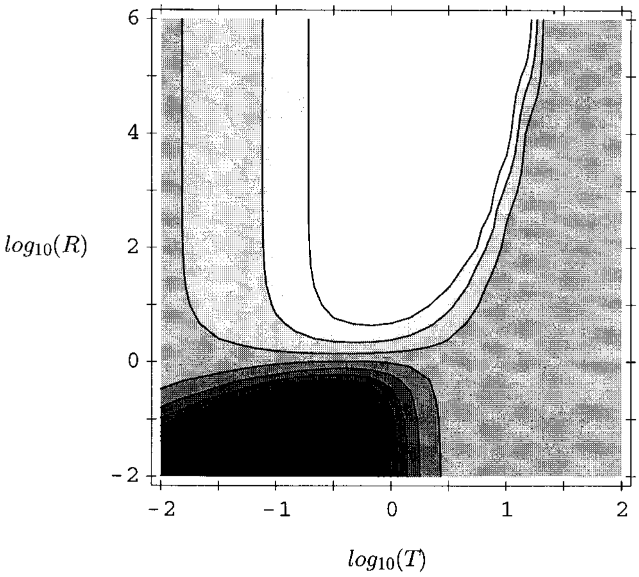

Figure 2.—A contour plot of the likelihood of the data in Table 1 over a range of R and T, when $M = 1.83$ (see text). Any waviness in the surface results from the fact that the likelihood was calculated at a finite number (25 × 25 = 625) of points.

For the data in Table 1, the maximum-likelihood estimate of the migration rate was $\hat{M} = 1.80$, with a 95% confidence interval of (1.22, 2.83). Thus, current levels of gene flow among human subpopulations appear to be relatively high. Estimates of M varied only slightly over the entire range of R and T. For example, the equilibrium migration model—*i.e.*, setting R equal to $1 + 1/(2M)$—gave $\hat{M} = 1.83$. Therefore, to save time, a single value of $M = 1.83$ was used to compute the likelihood surface for R and T shown in Figure 2. I confirmed that this had little effect on the results by separately maximizing the likelihood, over M, at nine evenly distributed points on the surface, and comparing these results with those obtained using $M = 1.83$. Figure 2 is a contour plot of the likelihood of the data in Table 1 over a broad range of both R and T. Each contour in Figure 2 represents a 2-unit change in log likelihood, so that the approximate joint 95% confidence region for R and T is enclosed within the first contour.

A number of conclusions can be drawn from this figure. First, it is clear that small values of R and T (lower left in Figure 2), which could represent recent

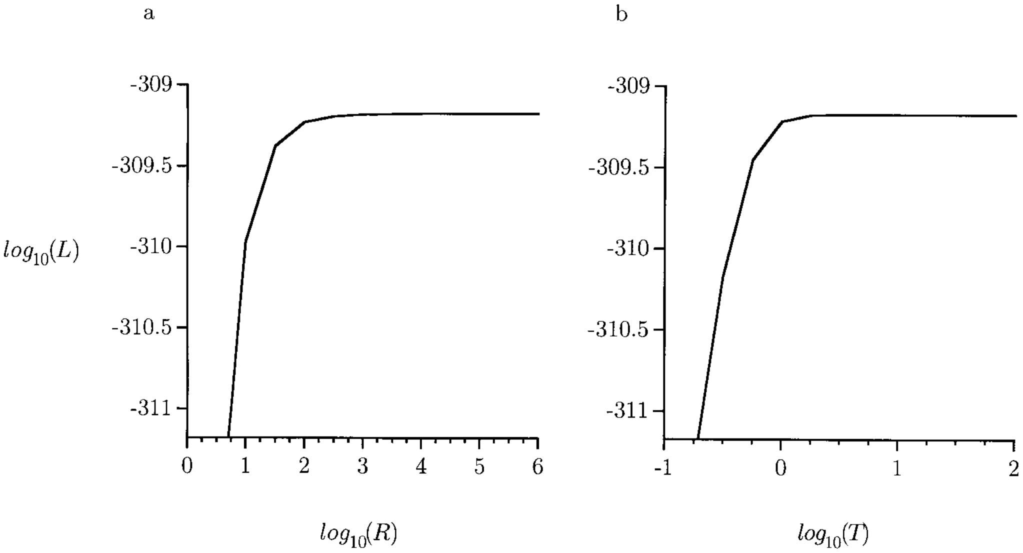

Figure 3.—The likelihood surfaces (a) for R and (b) for T when $R = \infty$. Only values within two log-likelihood units of the maximum are shown. Thus, a traces the ridge of the surface shown in Figure 2, and b shows a slice of the surface but for R much greater than any in Figure 2.

population expansion, are strongly rejected. Next, the present model reduces to an equilibrium model when $R = 1 + 1/(2M)$, that is when the effective size was the same before and after T. For $M = 1.83$, this gives $R = 1.27$ for all T, a value that falls between the third and the fourth contours in Figure 2. These contours define a broad flat area more than 6 log-likelihood units below the maximum, which occupies much of the graph. Thus, equilibrium migration (*i.e.*, no change in demography) is also strongly rejected. As T either increases or decreases for a given finite value of R, the likelihood returns to this broad flat area, which is consistent with an equilibrium model. This makes intuitive sense because the change in demography will have little effect if it is either extremely recent or extremely ancient.

As Figure 2 implies, there is no upper bound for R. The 95% confidence interval for R is $(5.0, \infty)$, and in fact the likelihood of the data continues to increase as R grows. As R increases, $E_{n'}(z_{i'})$ and $E_{n'}(S)$—see Equations 7 and 15—become linear in R as well as θ. For instance, $E_{n'}(S)$ converges on

$$\theta R \sum_{n''=2}^{n'} g_{n'n''}(\tau_{\mathrm{m}}) \sum_{i''=1}^{n''-1} \frac{1}{i''} \qquad (16)$$

as R goes to infinity. Both θ and R then drop out of the likelihood. When R is large, our power to say anything about its exact value diminishes. Using the large-R limiting expressions gives a maximum log-likelihood value of -309.17535. Figure 3a, which traces the ridge of maximum likelihood on the surface in Figure 2, shows that the likelihood is within 1% of its maximum value by the time R reaches 10^3. Thus, while we cannot reject huge values of R, we might argue that the evidence for R being greater than about 10^3 is not strong. In sum, there is clear evidence for at least a fivefold difference in recent *vs.* ancient effective population size in humans. If we assume that $N_A D_A = ND$, then from the confidence interval for R we find that the ancestral migration rate among human demes, M_A, must have been 0.125 or smaller.

The timing of this change in demography is also difficult to establish precisely. Figure 2 shows that the lower bound of the 95% confidence interval for T does not depend much on R, but that the upper bound is positively correlated with it. If we were to draw a line along the ridge of the maximum-likelihood surface in Figure 2, it would run from $T = 0.81$ at $R = 10$ to $T = 3.42$ at $R = 10^6$. The more ancient the event was, the more extreme it has to have been to explain the data. Using the limiting expressions for $E_{n'}(z_{i'})$ and $E_{n'}(S)$ gives a 95% confidence interval for T as $(0.2, \infty)$. Figure 3b shows that the likelihood increases with increasing T, but that the surface becomes quite flat above $T = 2.0$. Translating the confidence interval for T into years requires values for the recent effective size and the generation time for humans. Some typical values are $N_e = 10^4$ and $g = 15$ yr for the generation time (Takahata *et al.* 1995; Mountain 1998). Assuming that N_e is given by Equation 3, that $M = 1.8$, and using the lower bound of 0.2 for T, we estimate that this demographic event could have occurred as recently as $2N_eTg/[1 + 1/(2M)] \approx 47{,}000$ years ago, but might have been much more ancient.

DISCUSSION

The subdivided population coalescent model considered here and in Wakeley (1998) is flexible and easy to analyze. The genealogical history of a sample is characterized by a scattering phase and a collecting phase. The scattering phase depends on the current number of demes, the current deme size, and the current migration rate. The collecting phase is a coalescent on a time scale of $2N_e$ generations, where N_e is given by Equation

3. Because of this it is straightforward to incorporate changes in the ancestral numbers of demes, deme sizes, and migration rates, by considering their influence on the effective population size. We may write

$$N_e(t) = N(t)D(t)\left[1 + \frac{1}{2M(t)}\right], \tag{17}$$

where t is the time back into the past, measured in units of $2N(0)D(0)$ generations. While Equation 17 displays the flexibility of the model, it also shows that we cannot, in general, distinguish whether changes in N_e have been caused by changes in N, D, or M. In the case of humans, we can be sure that the total population number, $N(t)D(t)$, is larger now than it was in the past. The results presented here show that, in spite of this known increase in number, the effective population size of humans was greater in the past than it is now. Using a quite different method, Takahata *et al.* (1995) also conclude that past effective sizes were larger than present ones, and Takahata (1995) suggests that this may be due to frequent extinction and recolonization of local populations. Under the present model, the decrease in effective size is attributable to a recent increase in migration rates from lower historic to current high levels.

The results presented here are preliminary, as some of the assumptions of the model are violated for humans (see below). The purpose of this work has been to suggest that changes in migration rates over time may have shaped patterns of genetic polymorphism in humans. Because these RFLP data span the human genome (Bowcock *et al.* 1987; Matullo *et al.* 1994; Poloni *et al.* 1995), we are probably correct in attributing the patterns they show to population structure rather than to natural selection. This is because population-level processes affect all loci in a similar way while selection acts on particular sites or regions. The contrasting pattern seen in mtDNA may very well be the result of selection on the tightly linked mitochondrial genome, as suggested by Hey (1997). Alternatively, because of the fourfold difference in effective size between nuclear and mtDNA, the mitochondrial data might reflect more recent events and the nuclear data more ancient ones (Fay and Wu 1999; Hey and Harris 1999). The present work suggests that historical patterns of migration could also be important in accounting for differences between nuclear and mitochondrial data, for instance if the male and female migration rates vary in different ways over time.

One assumption of the model that is clearly violated for humans is that migration is equal and symmetric among every possible pair of demes, regardless of geographic location. Poloni *et al.* (1995), for example, found significant isolation by distance in a larger sample of demes that included those considered here. In the hope of approximating an island model, the four samples used in this study were chosen because they are relatively equidistant and widely separated. Slatkin and Barton (1989) found that samples taken from a two-dimensional stepping-stone model will behave roughly, at least in terms of F_{ST}, like samples from an island model if the distances between them are much greater than the characteristic scale of variation. Other assumptions of the model that might not hold include the assumption that all demes are equal in size and that only a single abrupt demographic event occurred. Clearly, these issues deserve to be explored further. The last of them would be straightforward to address. Using the approach of Wakeley and Hey (1997), multiple demographic events could be included—as, for instance, Hey and Harris (1999) have recently done for population bottleneck—and this could be used to approximate a continuous change in effective size. However, a continuous change that occurs rapidly will be approximated well by a single abrupt event—see, for example, Reich *et al.* (1999)—so this might not be a serious problem with the present analysis.

A further concern is the effect of ascertainment bias in the data, which could produce spuriously high values of R. These RFLP data were originally selected for study because they were polymorphic among Europeans. If there were a detailed record of how this was done, we could use Nielsen's (2000) method to correct for the bias, after adjusting the method to take population subdivision into account. There appears to be no such detailed record, but simulations have been carried out to measure the possible effect of ascertainment bias (Mountain and Cavalli-Sforza 1994). These indicate that the bias will be strongest for European samples and ones genetically close to them, and that ascertainment bias alone does not account for the genetic signal in the data. In an attempt to gauge the importance of ascertainment bias to this work, I excluded the European sample and redid the analysis. This did not change the results appreciably. As mentioned above, I also looked at just three African samples and found similar results. Because other nuclear data—both DNA sequences (Harding *et al.* 1997; Clark *et al.* 1998; Zietkiewicz *et al.* 1998; Harris and Hey 1999) and SNPs (Nielsen 2000)—appear to show a similar pattern, it seems likely that these RFLP data reflect past human demography rather than, or at least in addition to, biased sampling.

The conclusions reached here stand in contrast to most models of human history that assume a panmictic ancestral population; for a recent review of these see Mountain (1998). Here it was found that subdivision was stronger among ancestral human populations than it is now. This may be easy to accept given what we know about recent human demography, but it brings into focus another very important point. We can probably reject models of the isolation and subsequent divergence of human populations without gene flow. These would be characterized by very small values of M, probably small values of R, and a value of T corresponding to the time of isolation. Figure 2 shows that such values are strongly rejected. Some caution is needed here because strict isolation models are not simple limiting cases of corresponding migration models (Nath and Griffiths 1993). However, the data indicate that humans are a web of demes interconnected by gene flow rather than a tree of isolated populations.

A C program that performs the above analyses is available from the author.

I thank Rasmus Nielsen, Joanna Mountain, Monty Slatkin, and Carl Morris for helpful discussions of this material. Jody Hey and Warren Ewens provided comments on an earlier version. I thank Naoyuki Takahata and an anonymous referee for helpful reviews. Joanna Mountain kindly provided the RFLP data. This work was supported by a grant from the National Science Foundation (DEB-9815367).

LITERATURE CITED

Abramowitz, M., and I. A. Stegun, 1964 *Handbook of Mathematical Functions*. Dover, New York.

Arratia, R., A. D. Barbour and S. Tavaré, 1992 Poisson process approximations for the Ewens sampling formula. Ann. Appl. Probab. **2:** 519–535.

Bowcock, A. M., C. Bucci, J. M. Hebert, J. R. Kidd, K. K. Kidd *et al.*, 1987 Study of 47 DNA markers in five populations from four continents. Gene Geog. **1:** 47–64.

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

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595–612.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

Fay, J. C., and C.-I. Wu, 1999 A human population bottleneck can account for the discordance between patterns of mitochondrial and nuclear DNA variation. Mol. Biol. Evol. **16:** 1003–1005.

Fu, X.-Y., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox *et al.*, 1997 Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772–789.

Harris, E. E., and J. Hey, 1999 X chromosome evidence for ancient human histories. Proc. Natl. Acad. Sci. USA **96:** 3320–3324.

Hey, J., 1997 Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol. Biol. Evol. **14:** 166–172.

Hey, J., and E. Harris, 1999 Population bottlenecks and patterns of human polymorphism. Mol. Biol. Evol. **16:** 1423–1426.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. J. Futuyma and J. Antonovics. Oxford University Press, Oxford.

Karlin, S., and J. McGregor, 1972 Addendum to a paper of W. Ewens. Theor. Popul. Biol. **3:** 113–116.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to the steady flux of mutations. Genetics **61:** 893–903.

Kimura, M., and G. H. Weiss, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49:** 561–576.

Kingman, J. F. C., 1982a The coalescent. Stochastic Process. Appl. **13:** 235–248.

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

Latter, B. D. H., 1973 The island model of population differentiation: a general solution. Genetics **73:** 147–157.

Maruyama, T., 1970 Effective number of alleles in a subdivided population. Theor. Popul. Biol. **1:** 273–306.

Matullo, G., R. M. Griffo, J. L. Mountain, A. Piazza and L. L. Cavalli-Sforza, 1994 RFLP analysis on a sample from northern Italy. Gene Geog. **8:** 25–34.

Mountain, J. L., 1998 Molecular evolution and modern human origins. Evol. Anthropol. **7:** 21–37.

Mountain, J. L., and L. L. Cavalli-Sforza, 1994 Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc. Natl. Acad. Sci. USA **91:** 6515–6519.

Nath, H. B., and R. C. Griffiths, 1993 The coalescent in two colonies with symmetric migration. J. Math. Biol. **31:** 841–852.

Nei, M., and N. Takahata, 1993 Effective population size, genetic diversity, and coalescence time in subdivided populations. J. Mol. Evol. **37:** 240–244.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms (SNP's). Genetics (in press).

Nielsen, R., J. L. Mountain, J. P. Huelsenbeck and M. Slatkin, 1998 Maximum likelihood estimation of population divergence times and population phylogeny in models without mutation. Evolution **52:** 669–677.

Nordborg, M., and P. Donnelly, 1997 The coalescent process with selfing. Genetics **146:** 1185–1195.

Poloni, E. S., L. Excoffier, J. L. Mountain, A. Langaney and L. L. Cavalli-Sforza, 1995 Nuclear DNA polymorphism in a Mandenka population from Senegal: comparison with eight other human populations. Ann. Hum. Genet. **59:** 53–61.

Reich, D. E., M. W. Feldman and D. B. Goldstein, 1999 Statistical properties of two tests that use multilocus data sets to detect population expansions. Mol. Biol. Evol. **16:** 453–466.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161–1176.

Slatkin, M., 1989 Detecting small amounts of gene flow from phylogenies of alleles. Genetics **121:** 609–612.

Slatkin, M., 1996 Gene genealogies within mutant allelic classes. Genetics **143:** 579–587.

Slatkin, M., and N. H. Barton, 1989 A comparison of three indirect measures for estimating average levels of gene flow. Evolution **43:** 1349–1368.

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

Takahata, N., 1995 A genetic perspective on the origin and history of humans. Annu. Rev. Ecol. Syst. **26:** 343–372.

Takahata, N., and M. Nei, 1985 Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110:** 325–344.

Takahata, N., Y. Satta and J. Klein, 1995 Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. **48:** 198–221.

Tavaré, S., 1984 Lines-of-descent and genealogical processes, and their application in population genetic models. Theor. Popul. Biol. **26:** 119–164.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes and A. C. Wislon, 1991 African populations and the evolution of human mitochondrial DNA. Science **253:** 1503–1507.

Wakeley, J., 1998 Segregating sites in Wright's island model. Theor. Popul. Biol. **53:** 166–175.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

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

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

Wright, S., 1943 Isolation by distance. Genetics **28:** 114–138.

Zietkiewicz, E., V. Yotova, M. Jarnick, M. Korab-Laskowska, K. K. Kidd *et al.*, 1998 Genetic structure of the ancestral population of modern humans. J. Mol. Evol. **47:** 146–155.

Communicating editor: N. Takahata

APPENDIX A

The result given in Equation 3 is implicit in Wakeley (1998), where the expectation and variance of the total length of the genealogy during the collecting phase were derived. Under the assumption that the number of demes is large ($D \gg n$), the time to a common ancestor of n' sequences is a two-step process. For two members of the sample to coalesce, they must first be in the same deme. When time is measured in units of $2ND$ generations, the waiting time to this event is exponentially distributed with parameter $2\mathrm{M}\binom{n}{2}'$; *e.g.*, see Equation 24 in Wakeley (1998). When $D \gg n$, the two items then either coalesce or one of them migrates to an unoccupied deme. The probability that they coalesce is $1/(2M + 1)$, and the probability that one of them migrates is $2M/(2M + 1)$; see Equation 25 in Wakeley (1998). If one of them migrates, there is another exponential waiting time before they again have the chance to coalesce, and the probability of this is still $1/(2M + 1)$.

Let k be the number of times that two of the sequences have the chance to coalesce before they actually do. Then k is geometrically distributed:

$$P(k) = \frac{1}{2M + 1}\left(\frac{2M}{2M + 1}\right)^{k-1}. \tag{18}$$

The distribution of the time to coalescence, given k, is the sum of k exponential$(2M\binom{n'}{2})$ times, which is gamma distributed:

$$f(t|k) = \frac{[2M\binom{n'}{2}]^k}{\Gamma(k)} t^{k-1} e^{-2M\binom{n'}{2}t}. \tag{19}$$

The unconditional distribution of the total waiting time to coalescence is calculated as $\Sigma_{k=1}^{\infty} f(t|k)P(k)$ and is found to be exponentially distributed with parameter

$$2M\binom{n'}{2} \times \frac{1}{2M + 1} = \frac{\binom{n'}{2}}{1 + 1/2M}. \tag{20}$$

When two of the lineages do finally coalesce, which they do with probability one, n' is decremented by one and the process starts anew. Thus, we have Kingman's coalescent on a time scale of $2ND[1 + 1/(2M)]$ generations.

APPENDIX B

The scattering process that leads to Equation 1 is identical to what Arratia *et al.* (1992) call the Feller coupling. This is one of several stochastic processes known to generate the Ewens sampling formula. It follows, then, that not only does the (marginal) distribution in Equation 1 hold, but also that the distribution of the numbers of descendents of these lineages is given by the full Ewens sampling formula. Consider the single-deme sample, n_j, and to simplify notation, for the moment let $r = n_j$ and $k = n'_j$. The distribution of the numbers of descendents—r_1, r_2, . . . , r_k—of these k lineages is

$$P(r_{1,}\, r_{2,}, \ldots, r_k|k) = \frac{r!}{S_r^{(k)} k! r_1 r_2 \ldots r_k} \tag{21}$$

using Ewens (1972) Equation A5. $P(r_1, r_2, \ldots, r_k|k)$ forms a probability distribution on $(r_1, r_2, \ldots, r_k)$ such that $\Sigma_{i=1}^{k} r_i = r$. In other words,

$$S_r^{(k)} = \sum_{\{r_i:\Sigma_i r_i = r\}} \frac{r!}{k! r_1 r_2 \ldots r_k}. \tag{22}$$

Now suppose that a of the k lineages can be distinguished from the other $k - a$. Then Equation 21 becomes

$$P(l_1, \ldots, l_a, m_1, \ldots, m_{k-a}|a, k) = \frac{r!}{S_r^{(k)} k! l_1 \ldots l_a m_1 \ldots m_{k-a}}, \tag{23}$$

which for given a sums to one over $(l_1, \ldots, l_a, m_1, \ldots, m_{k-a})$ such that $\Sigma_{i=1}^{a} l_i + \Sigma_{j=1}^{k-a} m_j = r$. We are interested in the probability that $\Sigma_{i=1}^{a} l_i = b$, and this is gotten by summing (23) appropriately:

$$\begin{aligned} P(b|a, k) &= \sum_{\{l_i:\Sigma_{i=1}^{a} l_i = b\}} \sum_{\{m_j:\Sigma_{j=1}^{k-a} m_j = r-b\}} \frac{r!}{S_r^{(k)} k! l_1 \ldots l_\mathrm{a} m_1 \ldots m_{k-a}} \\ &= \frac{r! a! (k-a)!}{S_r^{(k)} k! b! (r-b)!} \left[\sum_{\{l_i\}} \frac{b!}{a! l_1 \ldots l_a}\right] \\ &\quad \times \left[\sum_{\{m_j\}} \frac{(r-b)!}{(k-a)! m_1 \ldots m_{k-a}}\right] = \frac{\binom{r}{b} S_b^{(\mathrm{a})} S_{r-b}^{(k-a)}}{\binom{k}{a} S_r^{(k)}}. \end{aligned} \tag{24}$$

As required, $P(b|a, k)$ forms a probability distribution on b such that $a \leq b \leq a + (r - k)$; see section 24.1.3IIA in Abramowitz and Stegun (1964). Putting $r = n_j$, $k = n'_j$, $a = i'_j$, and $b = i_j$ gives Equation 12 in the text.