Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 9197–9201, August 1997
Genetics

Detecting immigration by using multilocus genotypes

BRUCE RANNALA* AND JOANNA L. MOUNTAIN

Department of Integrative Biology, University of California, Berkeley, CA 94720-3140

Communicated by L. L. Cavalli-Sforza, Stanford University School of Medicine, Stanford, CA, June 13, 1997 (received for review December 17, 1996)

ABSTRACT **Immigration is an important force shaping the social structure, evolution, and genetics of populations. A statistical method is presented that uses multilocus genotypes to identify individuals who are immigrants, or have recent immigrant ancestry. The method is appropriate for use with allozymes, microsatellites, or restriction fragment length polymorphisms (RFLPs) and assumes linkage equilibrium among loci. Potential applications include studies of dispersal among natural populations of animals and plants, human evolutionary studies, and typing zoo animals of unknown origin (for use in captive breeding programs). The method is illustrated by analyzing RFLP genotypes in samples of humans from Australian, Japanese, New Guinean, and Senegalese populations. The test has power to detect immigrant ancestors, for these data, up to two generations in the past even though the overall differentiation of allele frequencies among populations is low.**

Classical theory in population genetics has focused on the long term effects of immigration on allele frequency distributions in semi-isolated populations, concentrating on the stationary distribution resulting from a balance between forces of immigration, genetic drift, and mutation (1–4). Less theory exists addressing the effect of recent immigration among populations with low levels of genetic differentiation. A theory describing the effects of immigration on the genetic composition of individuals in populations that are not at genetic equilibrium is needed to interpret much of the data being generated using current genetic techniques.

In this paper we consider the multilocus genotypes that result when individuals are immigrants, or have recent immigrant ancestry. We propose a test that allows recent immigrants to be identified on the basis of their multilocus genotypes; the test has considerable power for detecting immigrant individuals even when the overall level of genetic differentiation among populations is low. Molecular genetic techniques that allow multilocus genotypes to be described from single individuals are relatively new, and much of the information contained in these types of data is not fully exploited by estimators of long term gene flow that are currently available (5–7). We provide an example of an application of the method to restriction fragment length polymorphism (RFLP) genotypes from human populations; the method may also be applied to analyze multilocus allozyme and microsatellite data.

Theory

A collection of I discrete populations of a diploid species exchange immigrants, with random mating among individuals within populations. Consider a set of l loci in linkage equilibrium, and let k_j be the number of alleles at the jth locus. Let $\mathbf{x} = \{x_{hji}\}$ be a matrix of the allele frequencies in each population, where x_{hji} is the frequency of the hth allele ($h = 1, 2, \ldots, k_j$) at the jth locus ($j = 1, 2, \ldots, l$) in the ith population ($i = 1, 2, \ldots, I$). A set of $\mathbf{n} = \{n_1, n_2, \ldots, n_I\}$ chromosomes are sampled from the I populations, where n_i is the number sampled from the ith population. Let $\mathbf{X} = \{X_{ijm}\}$ be the matrix of genotypes among the sampled individuals, where X_{ijm} is the genotype at the jth locus of the mth individual sampled from the ith population.

The population allele frequencies are generally unknown, and we therefore derive the probability density of allele frequencies in each population by using a Bayesian approach. The tests presented in this paper make use of the allele frequency distributions to calculate genotype probabilities. It is assumed that the total number of alleles at the jth locus in each population is identically k_j. The set of alleles observed in the collection of populations as a whole is used as an estimate of k_j. Without additional information, we initially assign an equal probability density to the frequencies of the alleles at the jth locus in the ith population. The prior probability density of allele frequencies (i.e., before sampling) is then (8)

$$\Pr(\mathbf{x}_{ji}) = \prod_{h=1}^{k_j} x_{hji}^{(1/k_j)-1}/\Gamma(1/k_j)\,. \qquad [1]$$

The posterior probability density of the allele frequencies at the jth locus, conditioned on the alleles observed in a sample from population i, is now determined. Let the vector $\mathbf{n}_{ji} = \{n_{1ji}, \ldots, n_{k_j ji}\}$, where n_{hji} is the observed number of copies of the hth allele at the jth locus in a sample from the ith population. The posterior probability density of allele frequencies is then

$$\Pr(\mathbf{x}_{ji}|\mathbf{n}_{ji}) = \frac{\Pr(\mathbf{n}_{ji}|\mathbf{x}_{ji})\Pr(\mathbf{x}_{ji})}{\Pr(\mathbf{n}_{ji})}, \qquad [2]$$

where

$$\Pr(\mathbf{n}_{ji}|\mathbf{x}_{ji}) = \binom{n_{ji}}{n_{1ji}, \ldots, n_{k_j ji}} \prod_{h=1}^{k_j} x_{hji}^{n_{hji}}, \qquad [3]$$

and we define $n_{ji} = \sum_{h=1}^{k_j} n_{hji}$. The marginal distribution of $\mathbf{n}_{ji}$ (conditional on n_{ji}) is

$$\Pr(\mathbf{n}_{ji}) = \prod_{h=1}^{k_j} \frac{\Gamma(n_{hji} + 1/k_j)}{\Gamma(n_{hji} + 1)\Gamma(1/k_j)}\,. \qquad [4]$$

Eq. **2** then simplifies to

$$\Pr(\mathbf{x}_{ji}|\mathbf{n}_{ji}) = \Gamma(\theta) \prod_{h=1}^{k_j} \frac{x_{hji}^{\theta a_h - 1}}{\Gamma(\theta a_h)}, \qquad [5]$$

where $\theta = n_{ji} + 1$ and

$$a_h = \frac{n_{hji} + 1/k_j}{n_{ji} + 1}. \qquad [6]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: RFLP, restriction fragment length polymorphism.

*To whom reprint requests should be addressed. e-mail: bruce@mws4.biol.berkeley.edu.

Eq. **5** is a Dirichlet probability density function (8) with parameters θ and a_h, where $h = 1, 2, \ldots, k_j$.

Genotype Probabilities

If individual m is born to nonimmigrant parents in population i, the probability that the individual has genotype X_{ijm} at the jth locus, assuming random mating, is

$$\Pr(X_{ijm}|\mathbf{x}_{ji}) = \begin{cases} x_{hji}^2 & \text{if} \quad X_{ijm} = hh \\ 2\, x_{hji}\, x_{gji} & \text{if} \quad X_{ijm} = hg \end{cases} \qquad [7]$$

for all $h = 1, 2, \ldots, k_j$ and $g = 1, 2, \ldots, k_j$ where $g \neq h$. The actual allele frequencies in population i are unknown and we therefore consider the probability of the genotype for individual m, conditional on the sample of alleles from the ith population, denoted as $\Pr(X_{ijm}|\mathbf{n}_{ji})$. The genotype of individual m is created by sampling two alleles at random from population i. Since the allele frequencies are not known we use the Dirichlet density of Eq. **5** to describe the probability density of allele frequencies and integrate over all possible allele frequencies

$$\Pr(X_{ijm}|\mathbf{n}_{ij}) = \int \Pr(X_{ijm}|\mathbf{x}_{ji})\Pr(\mathbf{x}_{ji}|\mathbf{n}_{ji})d\mathbf{x}_{ji}. \qquad [8]$$

The marginal probability of the observed genotype, obtained in this way, is equal to the probability of sampling two alleles from a compound multinomial-Dirichlet distribution (7)

$$\Pr(X_{ijm}|\mathbf{n}_{ji}) = \begin{cases} \dfrac{(n_{hji} + 1/k_j + 1)(n_{hji} + 1/k_j)}{(n_{ji} + 2)(n_{ji} + 1)} & \text{if } X_{ijm} = hh \\ \dfrac{2(n_{hji} + 1/k_j)(n_{gji} + 1/k_j)}{(n_{ji} + 2)(n_{ji} + 1)} & \text{if } X_{ijm} = hg. \end{cases} \qquad [9]$$

This is the posterior probability that the genotype X_{ijm} is observed for the jth locus when the individual is a nonimmigrant from population i. If the allele frequencies are independent among loci (i.e., there is no linkage disequilibrium), the genotype probabilities at other loci are calculated similarly. The probability of the multilocus genotype $\mathbf{X}_{im} = \{X_{i1m}, \ldots, X_{ilm}\}$ of individual m is then a product over the probabilities of the observed allelic configurations for each locus

$$\Pr(\mathbf{X}_{im}|\mathbf{n}_i) = \prod_{j=1}^{l} \Pr(X_{ijm}|\mathbf{n}_{ji}). \qquad [10]$$

We now consider situations in which one parent is a resident of population i, and the other is an immigrant born to nonimmigrant parents in population i'. In this case, one allele copy is of immigrant origin. There is generally no prior information regarding the source of an individual's alleles (chromosomes), and each copy at a locus is therefore equally likely to have been derived from the immigrant source. If we consider the genotype of individual m, born in population i, and denote an immigrant allele from population i' using a prime symbol, the probability of the mixed genotype $X_{(i,i')jm}$ at the jth locus is

$$\Pr(X_{(i,i')jm}) = \frac{1}{2}\{\Pr(X_{[i',i]jm}) + \Pr(X_{[i,i']jm})\}, \qquad [11]$$

where parentheses in the subscripts indicate that the alleles making up the genotype are averaged with respect to possible source populations, and brackets indicate that the alleles making up the genotype are labeled according to their source population. If an individual has alleles h and g at a particular locus, for example, these possibilities are $X_{[i,i']jm} = hg'$ and $X_{[i',i]jm} = h'g$, where h' indicates that allele h is derived from population i' and h indicates that it is derived from population i. Conditional on the allele frequencies, the probabilities of the genotypes, labeled according to source population, are

$$\Pr(X_{[i,i']jm}|\mathbf{x}_{ji'}, \mathbf{x}_{ji}) = x_{hji}\, x_{gji'}$$
$$\Pr(X_{[i',i]jm}|\mathbf{x}_{ji'}, \mathbf{x}_{ji}) = x_{hji'}\, x_{gji}. \qquad [12]$$

Sampling a single allele for each population from a multinomial-Dirichlet density with the appropriate population parameters, we obtain

$$\Pr(X_{(i',i)jm}|\mathbf{n}_{ji}, \mathbf{n}_{ji'}) = \begin{cases} \dfrac{(n_{hji'} + 1/k_j)(n_{hji} + 1/k_j)}{(n_{ji'} + 1)(n_{ji} + 1)} & \text{if } X_{(i',i)jm} = hh \\ \dfrac{(n_{hji'} + 1/k_j)(n_{gji} + 1/k_j) + (n_{hji} + 1/k_j)(n_{gji'} + 1/k_j)}{(n_{ji'} + 1)(n_{ji} + 1)} & \text{if } X_{(i',i)jm} = hg \end{cases} \qquad [13]$$

for all $h = 1, 2, \ldots, k_j$ and $g = 1, 2, \ldots, k_j$ where $g \neq h$. This is the posterior probability that the genotype $X_{(i',i)jm}$ is observed for the jth locus when the individual is of mixed ancestry from populations i and i'. The probability of the multilocus genotype $\mathbf{X}_{(i',i)m} = \{X_{(i',i)1m}, \ldots, X_{(i',i)lm}\}$ is then

$$\Pr(\mathbf{X}_{(i,i')m}|\mathbf{n}_i, \mathbf{n}_{i'}) = \prod_{j=1}^{l} \Pr(X_{(i,i')jm}|\mathbf{n}_{ji}, \mathbf{n}_{ji'}). \qquad [14]$$

Identifying Immigrant Genotypes

In this section, we describe a test for detecting individuals born in a population other than the one from which they are sampled; these individuals are first-generation immigrants. Consider an individual m randomly sampled from population i. The probability of the observed multilocus genotype for the individual, given that the individual was born in population i and has no recent immigrant ancestry, is calculated using Eq. **10** above as $\Pr(\mathbf{X}_{im}|\mathbf{n}_i)$. If individual m was instead born in population i' to parents with no recent immigrant ancestry and subsequently immigrated to population i, then the probability of observing the multilocus genotype of the individual is calculated using Eq. **10** above as $\Pr(\mathbf{X}_{im}|\mathbf{n}_{i'})$. The relative probability that the individual was born to parents with no recent immigrant ancestry in population i, rather than population i', is therefore given by the ratio of the probabilities

$$\Lambda = \frac{\Pr(\mathbf{X}_{im}|\mathbf{n}_i)}{\Pr(\mathbf{X}_{im}|\mathbf{n}_{i'})}. \qquad [15]$$

In practice, we take logarithms and use the equivalent form

$$\ln \Lambda = \ln[\Pr(\mathbf{X}_{im}|\mathbf{n}_i)] - \ln[\Pr(\mathbf{X}_{im}|\mathbf{n}_{i'})]. \qquad [16]$$

Positive values of $\ln \Lambda$ indicate that the null hypothesis (that the individual is not an immigrant) is favored, while negative values indicate that the alternative hypothesis (that the individual is an immigrant) is favored. A value of $\ln \Lambda = \ln(10) = 2.30$, for example, indicates that individual m is 10 times *more* likely to have arisen in population i, while a value of $\ln \Lambda = -2.30$ indicates the individual is 10 times *less* likely to have arisen in i than i'. The distribution of the statistic Λ under the null hypothesis (that the individual is not an immigrant) was examined using Monte Carlo simulation (see below).

We now describe a test for detecting an individual with a single parent that is an immigrant, or is descended from an immigrant. In this case, one allele copy at each locus is of possible immigrant origin and the other is of local origin. For

l independent loci, the probability of observing the genotype of individual m, given that the individual was born in population i and has an immigrant parent from population i', is calculated using Eq. **14** as $\Pr(\mathbf{X}_{(i',i)m}|\mathbf{n}_i, \mathbf{n}_{i'})$.

The individual might instead have an ancestor d generations in the past that was an immigrant. The probability of the observed genotype $X_{(i',i)jm}$ at the jth locus under this hypothesis is

$$\Pr^{(d)}(\mathbf{X}_{(i',i)jm}|\mathbf{n}_{ji}, \mathbf{n}_{ji'}) = \frac{\Pr(\mathbf{X}_{(i',i)jm}|\mathbf{n}_{ji}, \mathbf{n}_{ji'}) - \Pr(\mathbf{X}_{ijm}|\mathbf{n}_{ji})}{2^{d-1}} + \Pr(\mathbf{X}_{ijm}|\mathbf{n}_{ji}). \quad [17]$$

For l independent loci, the posterior probability of observing the genotype of individual m, given that this individual has an immigrant ancestor d generations removed from population i', is

$$\Pr^{(d)}(\mathbf{X}_{(i',i)m}|\mathbf{n}_i, \mathbf{n}_{i'}) = \prod_{j=1}^{l} \Pr^{(d)}(X_{(i,i')jm}|\mathbf{n}_{ji}, \mathbf{n}_{ji'}). \quad [18]$$

The relative probability that individual m, born in population i, did not have an immigrant ancestor from population i' at generation d in the past is

$$\Lambda_d = \frac{\Pr(\mathbf{X}_{im}|\mathbf{n}_i)}{\Pr^{(d)}(\mathbf{X}_{(i,i')m}|\mathbf{n}_i, \mathbf{n}_{i'})}. \quad [19]$$

We again use logarithms to calculate this statistic as $\ln \Lambda_d$. The analysis can be extended to consider individuals of mixed parentage over several generations, but the number of possible outcomes makes an exhaustive analysis difficult. In certain cases, when fewer ancestral immigration patterns are possible, based on prior information, the method outlined above might be extended to decide among the possible alternatives.

Critical Region and Power of Tests

The critical (rejection) region for the test statistic calculated using the methods described in the preceding sections contains all values of the statistic such that $\Lambda < C$, where C is chosen to satisfy $\Pr(\Lambda < C) = \alpha$ under the null hypothesis. For a specified value of C, the value of α is given by

$$\alpha = \sum_{h=1}^{G} \mathcal{I}_{\Lambda(\mathbf{X}_{ih},\mathbf{n}_i,\mathbf{n}_{i'})} \Pr(\mathbf{X}_{ih}|\mathbf{n}_i), \quad [20]$$

where

$$\mathcal{I}_{\Lambda(\mathbf{X}_{ih},\mathbf{n}_i,\mathbf{n}_{i'})} = \begin{cases} 1 \text{ if } \Lambda(\mathbf{X}_{ih},\mathbf{n}_i,\mathbf{n}_{i'}) < C \\ 0 \text{ otherwise,} \end{cases} \quad [21]$$

and the sum is over the total number of possible genotype configurations $G = \prod_{j=1}^{l}(k_j + 1)k_j/2$. A Monte Carlo estimator of α is

$$\alpha \approx \frac{1}{R} \sum_{r=1}^{R} \mathcal{I}_{\Lambda(\mathbf{X}_i(r),\mathbf{n}_i,\mathbf{n}_{i'})}, \quad [22]$$

where $\mathbf{X}_i(r)$ is the rth simulated genotype with R genotypes simulated in total from the posterior probability distribution $\Pr(\mathbf{X}_{ih}|\mathbf{n}_i)$. The random variables $\mathbf{X}_i(r)$ can be generated using the following procedure: for the jth locus, generate the first allele by assigning to allele type h the probability

$$\frac{n_{hji} + 1/k_j}{n_{ji} + 1}. \quad [23]$$

If the first allele is of type h, generate the second allele by assigning to allele type h the probability

$$\frac{n_{hji} + 1/k_j + 1}{n_{ji} + 2}, \quad [24]$$

or to allele type $g \neq h$ the probability

$$\frac{n_{gji} + 1/k_j}{n_{ji} + 2}. \quad [25]$$

It is also possible to determine C by generating a set of genotypes as outlined above and considering the value of the test statistic that falls below $1 - \alpha$ percent of the values for the simulated genotypes (see Fig. 1). The power of the test to reject the null hypothesis when it is false, for a specified critical region α, is

$$\beta = \sum_{h=1}^{G} \mathcal{I}_{\Lambda(\mathbf{X}_{i'h},\mathbf{n}_i,\mathbf{n}_{i'})} \Pr(\mathbf{X}_{i'h}|\mathbf{n}_{i'}), \quad [26]$$

where

$$\mathcal{I}_{\Lambda(\mathbf{X}_{i'h},\mathbf{n}_i,\mathbf{n}_{i'})} = \begin{cases} 1 \text{ if } \Lambda(\mathbf{X}_{i'h},\mathbf{n}_i,\mathbf{n}_{i'}) < C(\alpha) \\ 0 \text{ otherwise,} \end{cases} \quad [27]$$

where $C(\alpha)$ is the value of C that specifies the critical region with probability α determined using Eq. **20**. A Monte Carlo estimator of the power β is

$$\beta = \sum_{r=1}^{R} \mathcal{I}_{\Lambda(\mathbf{X}_{i'}(r),\mathbf{n}_i,\mathbf{n}_{i'})}, \quad [28]$$

where $\mathbf{X}_{i'}(r)$ is the rth simulated genotype, with R genotypes simulated in total from the posterior probability distribution $\Pr(\mathbf{X}_{i'h}|\mathbf{n}_{i'})$. The power of the test is illustrated graphically as the overlap between the distributions of the statistic generated by simulating genotypes under the null and alternative hypotheses (see Fig. 2).

Application

We have applied our method to a set of 12 individuals from each of four human populations. We chose to compare two

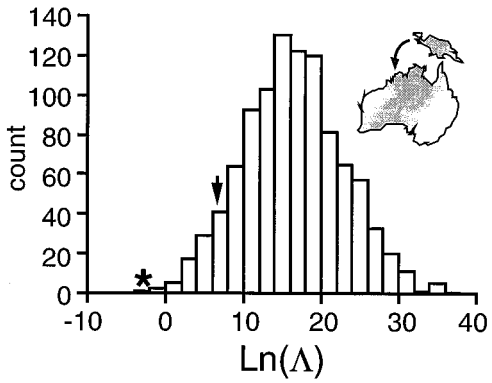

FIG. 1. Illustration of Monte Carlo method for examining significance of test statistic $\ln \Lambda$ for comparison of Australian (sample) and New Guinean (potential source) populations. Histogram of 1,000 values of the ln-probability difference generated by simulating genotypes given the allele counts observed for the Australian (sample) population. A total of 72 markers, for which the individual Australian 1 has been typed, were used to generate the distribution. See Eqs. **23–25**. The critical region for the test statistic ($\alpha = 0.05$) is that portion of the distribution to the left of the arrow. The posterior probability ratio ($\ln \Lambda = -2.76$) for the individual Australian 1 is indicated by an asterisk.

population samples with quite low genetic differentiation (9) and two population samples with quite high genetic differentiation from a set of population samples studied previously (10). The samples with low differentiation are from an Australian population and a New Guinean population (F_{ST} distance = 0.056). The samples with high differentiation are from a Japanese population and a Senegalese population (F_{ST} distance = 0.232). The Australian sample was collected from a coastal region of Australia, and the New Guinea sample from the highland region of New Guinea. The Japanese sample consists of individuals born in Japan and was collected in the San Francisco Bay Area (11). The Senegalese sample consists of Niokolonke individuals of the Mandenka population collected in southeastern Senegal (10). These 48 individuals have been typed at approximately 50 loci (12) by using RFLPs. The physical locations of the loci suggest that most are unlinked. Multiple restriction enzymes were used to type several of the loci so that the total number of genetic markers was approximately 75. The procedures employed in the sampling and the genetic analysis are described in detail elsewhere (11).

The power of the test to detect immigrants depends on the extent of differentiation between the populations compared (Table 1) as well as the number of loci examined and the number of individuals sampled (unpublished observations). A test of the hypothesis that an individual is an immigrant has high power in all the population comparisons. A test of the hypothesis that an individual has an immigrant parent has lower power for a comparison of individuals from the Australian and New Guinean samples than for a comparison of individuals from the Japanese and Senegalese samples. The test has power to detect an immigrant ancestor through the grandparent generation for a comparison of individuals from Japan and Senegal.

The distribution of the statistic under Monte Carlo simulation (Fig. 2) illustrates the power of the tests. In Fig. 2*a*, individuals sampled in Australia are postulated to have immigrated from New Guinea. There is little overlap between the distribution of the test statistic generated by Monte Carlo simulation under the null hypothesis that an individual was born in the Australian population (at right of Fig. 2*a*) and that under the alternative hypothesis that an individual is an immigrant from the New Guinea population (at left of Fig. 2*a*). In Fig. 2*b*, individuals in the Australian sample have a single parent that is an immigrant from New Guinea under the alternative hypothesis. In this case there is more overlap between the distributions generated under the null and alternative hypotheses, indicating that the test has reduced power by comparison with the test for detecting first-generation immigrants (i.e., Fig. 2*a*).

We applied the test to predict whether individuals sampled in Australia have New Guinean ancestry, and *vice versa*, and whether individuals sampled in Japan have African ancestry, and *vice versa*. A total of four individuals from the complete set of 48 comparisons produced significant test statistics at some level of ancestry (Table 2). Three of the four individuals (Australia 1, Australia 2, and Australia 3) who appeared to be immigrants, or descended from immigrants, were drawn from the Australian population, which appears likely to have experienced recent exchanges of immigrants (11). In the case of three individuals (Australia 1, Australia 3, and Japanese 1) it appears possible that an ancestor two or more generations removed was an immigrant, whereas in the case of one individual (Australia 2) it appears most probable that the individual is a first-generation immigrant. Given these results, one might consider excluding individual Australia 1, for example, from the Australian population sample for evolutionary studies, as it is quite probable that this individual has recent immigrant ancestry.

Table 1. Power of posterior probability ratio tests for recent immigration, with $\alpha = 0.05$

Sample population	Potential source	Power at d				
		0	1	2	3	4
Australian	New Guinean	1.00	0.83	0.37	0.17	0.09
New Guinean	Australia	1.00	0.94	0.60	0.25	0.14
Senegalese	Japanese	1.00	1.00	0.78	0.37	0.16
Japanese	Senegalese	1.00	1.00	0.76	0.41	0.20

If $d = 0$, the individual under consideration immigrated from source population; $d = 1$, one parent of the individual immigrated; if $d = 2$, one grandparent of the individual immigrated; if $d = 3$, one great-grandparent of the individual immigrated; if $d = 4$, one great-great-grandparent of the individual immigrated from source population.

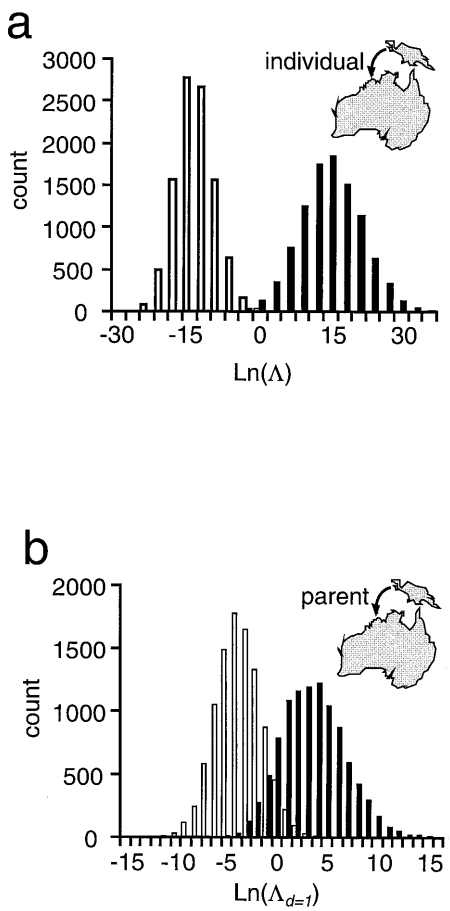

FIG. 2. Histograms indicating the power of the immigration tests for two cases. (*a*) The hypothesis that an Australian individual is an immigrant ($d = 0$) from New Guinea is considered. The shaded columns represent the distribution of ln Λ generated given the alleles observed for the Australian sample, while the unshaded columns represent the distribution of ln Λ generated given the alleles observed for the New Guinean sample. (*b*) The hypothesis that one parent of an Australian individual was an immigrant ($d = 1$) from New Guinea is considered. The shaded columns represent the distribution of ln Λ generated given the alleles observed for the Australian sample, while the unshaded columns represent the distribution of ln Λ generated given the alleles observed for the Australian and New Guinean samples and assuming that the individual received one allele at each locus from each population.

Discussion

The test for detecting recent immigration developed in this paper provides information relevant to a wide range of problems in population biology and human genetics. In the area of human genetics, for example, the method may be used to identify individuals whose genomes are not typical of the populations in which they currently live, or of their ethnic

Table 2. Power of the posterior probability ratio test to detect immigrant ancestry: Four individuals with posterior probability ratios indicating possible immigration ($\alpha < 0.05$)

Individual	Potential source	No. of markers	Value	Hypothetical immigrant ancestor: Individual ($d = 0$)	Parent ($d = 1$)	Grandparent ($d = 2$)	Great-grandparent ($d = 3$)
AUS1	NGN	76	$\ln \Lambda$	−2.76	−2.89	−1.65	−0.89
			α	0.000	0.009	0.022	0.037
			Power	1.000	0.821	0.347	0.197
AUS2	NGN	73	$\ln \Lambda$	4.48	0.87	−0.37	−0.11
			α	0.032	0.179	0.244	0.288
			Power	1.000	0.828	0.332	0.136
AUS3	NGN	82	$\ln \Lambda$	5.23	−0.50	−0.90	−0.56
			α	0.032	0.049	0.064	0.092
			Power	1.000	0.862	0.375	0.149
JPN1	SEN	69	$\ln \Lambda$	17.80	1.52	−1.26	−1.10
			α	0.021	0.014	0.029	0.045
			Power	1.000	0.999	0.771	0.431

Twelve individuals from each of four populations were included. Australians (AUS) were considered as possible immigrants, or descendants of immigrants, from New Guinea (NGN), and *vice versa.* Japanese (JPN) were considered as possible immigrants, or descendants of immigrants, from the Senegalese (SEN) population, and *vice versa.* Values of $\ln \Lambda$ or $\ln \Lambda_d$ are given in the first row for each individual. Values in the second row are significance levels (α values) approximated using the Monte Carlo approach (1,000 iterations per test). Values in the third row are the power of the test for this individual ($\alpha < 0.05$).

group. This may be helpful in genetic counselling. In the area of evolutionary biology, it is often important to identify immigrant individuals to study their behavior and interactions with resident individuals. It may also be important to quantify the amount of recent immigration in populations that are not at genetic equilibrium. In the field of conservation genetics, this test may be useful for identifying the population of origin for zoo animals whose history is poorly known to implement successful captive breeding programs.

At least three potentially misleading results may arise when applying the method considered here. First, the failure to reject the hypothesis that an individual was an immigrant, or descended from immigrants, may simply reflect the fact that the appropriate populations for comparison were not included in the analysis. Second, an individual might incorrectly appear to have originated in a particular population other than the one from which it was sampled. This might be due to similarities in allele frequencies, due to long-term gene flow, between that population and a third population from which the individual actually originated, but which was not included in the sample of populations. Third, the fact that many pairwise comparisons between populations are performed for each of a large number of individuals means that some individuals will appear to be immigrants purely by chance. This can be corrected for by using smaller values for α.

The analyses of human populations presented in this paper show that, even with a sample of only 60 independent loci, the method we have proposed has power to detect immigrant ancestry up to two generations in the past. This is despite our conservative correction for uncertainties of allele frequencies. A larger number of loci will increase the power and could allow a single immigrant great-grandparent (out of 8 total), or a single immigrant great-great-grandparent (out of 16 total), to be identified. The precise number of loci needed to obtain a given level of power depends on the degree of genetic differentiation between populations; with greater differentiation, fewer loci are needed to obtain the same level of power. Computer simulations should prove useful in exploring the statistical performance of the method more generally.

Program availability. A program written in the C computer language for performing the calculations described in this paper is available by anonymous ftp from mw511.biol.berkeley.edu in directory /pub, or on the World-Wide Web at site http://mw511.biol.berkeley.edu/homepage.html.

This research was supported, in part, by a National Institutes of Health Grant (GM40282) to Montgomery Slatkin and by a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada to B.R.

1. Wright, S. (1931) *Genetics* **16,** 97–159.
2. Kimura, M. (1953) *Annu. Rep. Natl. Inst. Genet.* **3,** 63.
3. Maruyama, T. (1970) *Theor. Pop. Biol.* **1,** 273–306.
4. Slatkin, M. (1985) *Annu. Rev. Ecol. System.* **16,** 393–430.
5. Slatkin, M. & Barton, N. H. (1989) *Evolution* **43,** 1349–1368.
6. Weir, B. S. & Cockerham, C. C. (1984) *Evolution* **38,** 1358–1370.
7. Rannala, B. & Hartigan, J. A. (1996) *Genet. Res.* **67,** 147–158.
8. Johnson, N. L. & Kotz, S. (1970) *Distributions in Statistics: Continuous Multivariate Distributions* (Wiley, New York).
9. Reynolds, J, Weir, B. S. & Cockerham, C. C. (1983) *Genetics* **105,** 767–779.
10. Poloni, E. S., Excoffier, L., Mountain, J. L., Langaney, A. & Cavalli-Sforza, L. L. (1995) *Ann. Hum. Genet.* **59,** 43–61.
11. Lin, A. A., Hebert, J. M., Mountain, J. L. & Cavalli-Sforza, L. L. (1994) *Gene Geography* **8,** 191–214.
12. Mountain, J. L. (1994) Ph.D. thesis (Stanford University, Stanford, CA).