On the Number of New World Founders: A Population Genetic Portrait of the Peopling of the Americas

Jody Hey

Department of Genetics, Rutgers, the State University of New Jersey, Piscataway, New Jersey, United States of America

The founding of New World populations by Asian peoples is the focus of considerable archaeological and genetic research, and there persist important questions on when and how these events occurred. Genetic data offer great potential for the study of human population history, but there are significant challenges in discerning distinct demographic processes. A new method for the study of diverging populations was applied to questions on the founding and history of Amerind-speaking Native American populations. The model permits estimation of founding population sizes, changes in population size, time of population formation, and gene flow. Analyses of data from nine loci are consistent with the general portrait that has emerged from archaeological and other kinds of evidence. The estimated effective size of the founding population for the New World is fewer than 80 individuals, approximately 1% of the effective size of the estimated ancestral Asian population. By adding a splitting parameter to population divergence models it becomes possible to develop detailed portraits of human demographic history. Analyses of Asian and New World data support a model of a recent founding of the New World by a population of quite small effective size.

Citation: Hey J (2005) On the number of New World founders: A population genetic portrait of the peopling of the Americas. PLoS Biol 3(6): e193.

Introduction

Archeological evidence, as well as anatomical, linguistic, and genetic evidence, have shown that the original human inhabitants of the Western Hemisphere arrived from Asia during the Late Pleistocene [1–4]. However, there persists uncertainty on the source, within Asia, of peoples who migrated to the New World [5], on the timing of the earliest migration [6–10], and on whether there have been multiple migrations [3,11–13].

For complex historical subjects such as the colonization of the Americas, there are many ways that models can be constructed, examined, and compared. One approach is to develop a portrait based on a particular kind of data, such as linguistic [6], skeletal [14], or archaeological [15] data, or on DNA sequence data from a particular portion of the human genome such as the mitochondria [4,16–19] or the Y chromosome [9]. Yet each source of data has unique sources of variation. In the case of genetic data there occurs a large stochastic variance of the coalescent history among genes that causes different loci to vary widely in levels of genetic variation and in apparent patterns of relationships among populations [20–22]. This stochastic variance is sometimes overlooked, for example in discussions of the histories of the individual DNA sequence haplotypes [18], and it is easy to underestimate the many possible histories that are consistent with a finding that haplotypes are shared by different populations [23–25].

To accommodate the stochastic variance among loci, population geneticists have turned in recent years to Bayesian and likelihood methods that explicitly take into account the range of possible gene tree histories that are consistent with a given dataset [26–30]. For questions on population divergence, the focus has been on models of population splitting in which an ancestral population divides into two descendant populations, after which there may be gene flow between the descendant populations. These "isolation with migration" (IM) models can have a large number of parameters, and they offer the possibility of capturing many of the dynamics that occur in the early stages of population divergence or speciation [30–33].

Figure 1A shows the basic IM model, in which the ancestral and descendant populations each have a constant size. Each of the terms in the model is explained in Table 1. Basic limitations of this model are that it cannot provide details on how descendant populations were founded or whether population sizes have changed. Certainly for human populations there is considerable genetic evidence that population sizes have grown [34–37], and it would be helpful if it were possible to capture information on the sizes of descendant populations as they are formed. For example, if one descendant population formed as a small founder population that later grew to a large size, such dynamics would not be revealed in the fitting of the basic IM model. To allow the study of such histories, an additional parameter has been added to the IM model. Figure 1B shows a model in which an ancestral population splits in two, with the relative sizes of those two new populations reflected in the parameter

Received July 12, 2004; Accepted March 25, 2005; Published May 24, 2005
DOI: 10.1371/journal.pbio.0030193

Abbreviation: IM, isolation with migration

Academic Editor: Andy G. Clark, Cornell University, United States of America

E-mail: hey@biology.rutgers.edu

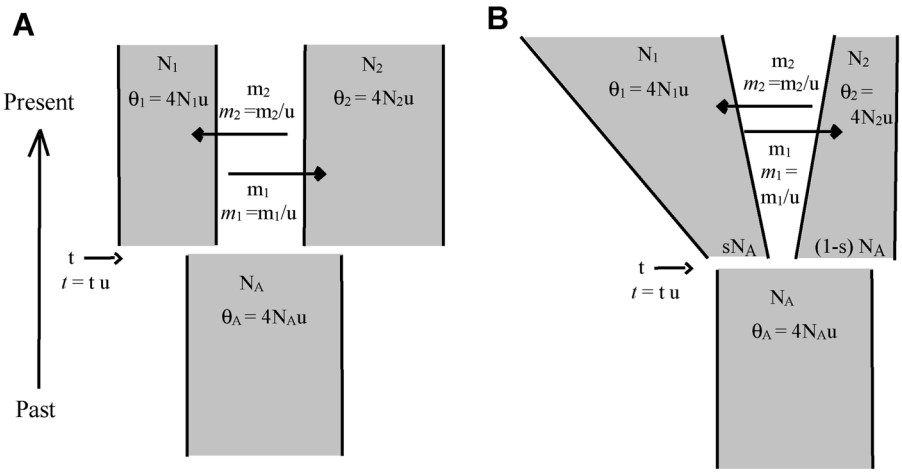


Figure 1. Isolation with Migration Models

(A) The basic IM model. The demographic terms are effective population sizes (N_1, N_2, and N_A), gene flow rates (m_1 and m_2), and population splitting time (t). Also shown are parameters scaled by the neutral mutation rate (u), as they are actually used in the model fitting. Terms for basic demographic parameters, including N, m, t, and u, are not italicized. Note that the migration parameters are identified by the source of migrants as time goes backward in the coalescent. In other words, the migration rate from population 1 to population 2 (i.e., m_1) actually corresponds to the movement of genes from population 2 to population 1 as time moves forward.

(B) The IM model with changing population size. An additional parameter, *s*, is the fraction of N_A that forms N_1 (i.e., the fraction $1 - s$ gives rise to N_2)

DOI: 10.1371/journal.pbio.0030193.g001

s, where $0 < s < 1$. At the time of the split, descendant population 1 has size sN_A from which it moves to size N_1 at the time of sampling. Similarly, population 2 begins with size $(1 - s)N_A$ from which it moves to size N_2 at the time of sampling. Figure 1B depicts one population growing and the other shrinking, but in fact either population is free to either grow or shrink under this model.

These models were applied to questions on the founding of New World populations from Asia. A total of nine DNA sequence datasets that included Asian and Native American (Amerind-speaking) samples were drawn from the literature (Figure 2 and Table 1) and analyzed jointly using a procedure that provides posterior probability distributions for each of the model parameters [30,33]. The stochastic variance among loci is clearly evident in the variation of F_{ST} values (between Asian and New World samples) observed among the loci. Of the nine loci included in the present study, three have fairly high F_{ST} values, while the remainder are either negative (undefined) or near zero (Table 1).

Asian samples were arbitrarily designated as being from

Table 1. Parameter Summary and Description

Parameter	Description
N_1	Effective size of population 1 (present-day Asia)
N_2	Effective size of population 2 (present-day New World)
N_A	Effective size of the ancestral Asian population
m_1	Probability of migration from the New World to Asia, per gene copy per generation
m_2	Probability of migration from Asia to the New World, per gene copy per generation
t	The time since the founding of the New World from Asia
s	The fraction of the ancestral population that did not found the New World population
$1 - s$	The fraction of the ancestral population that founded the New World population
u	The neutral mutation rate (for the entire sequence, not per base pair) per generation; for multiple loci, this is the geometric mean of the mutation rates per generation
$\theta_1 = 4N_1u$	The population mutation rate for population 1 (present-day Asia)
$\theta_2 = 4N_2u$	The population mutation rate for population 2 (present-day New World)
$m_1 = m_1/u$	The migration rate, per mutation, from the New World to Asia
$m_2 = m_2/u$	The migration rate, per mutation, from Asia to the New World
$t = t\ u$	The number of mutations since the time of founding of the New World
$2N_1m_1$	The effective number of gene migrants into Asia, per generation
$2N_2m_2$	The effective number of gene migrants into the New World, per generation

DOI: 10.1371/journal.pbio.0030193.t001

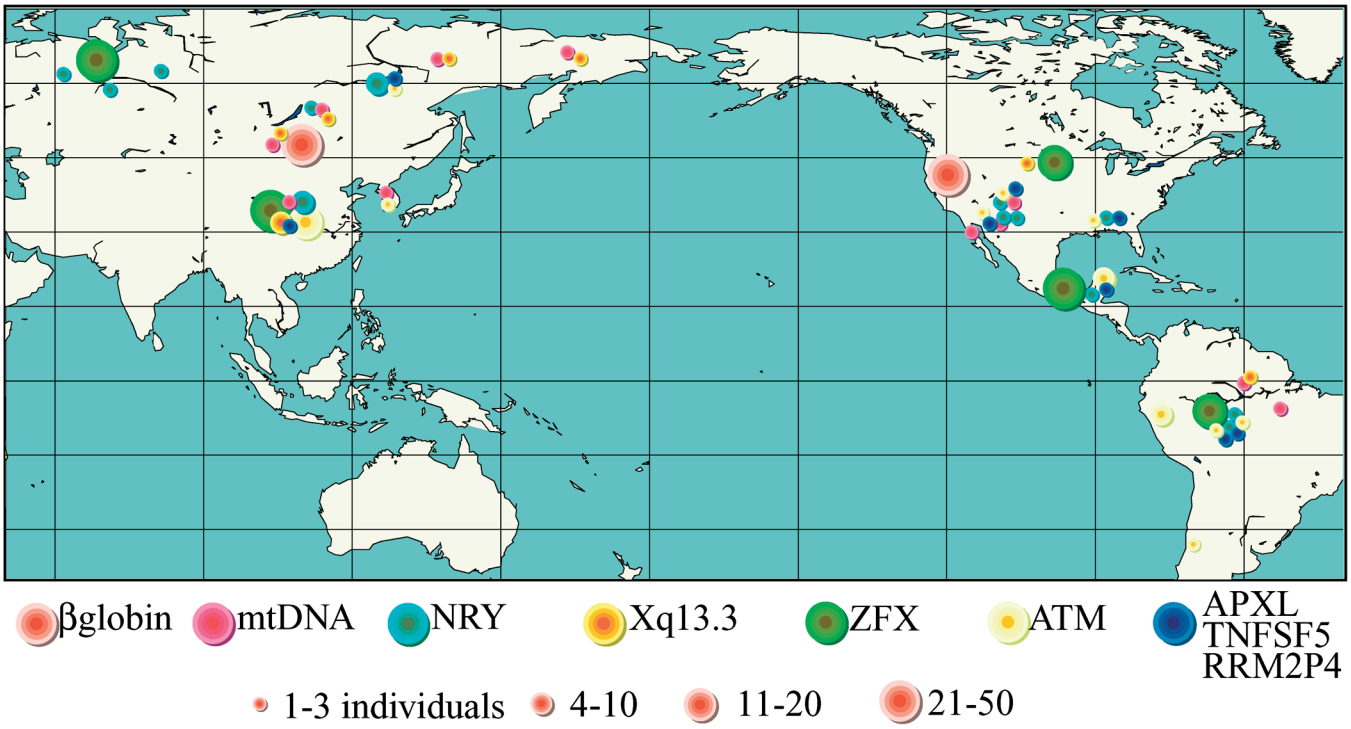


Figure 2. Approximate Geographic Locations, and Sample Sizes per location, for Each Locus Listed in Table 1

In some cases locations are based on actual geographic locations, in other cases the locations are the approximate center of the geographic region occupied by ethnic groups identified in the original references (Table 1).

DOI: 10.1371/journal.pbio.0030193.g002

population 1 and the New World samples from population 2. In this case, $1 - s$ is the fraction of the ancestral population that founded the New World population. The analyses also require that prior distributions be specified for each model parameter. It was assumed that the New World was founded by a minority of the ancestral Asian population, corresponding to a specified uniform prior distribution for s between 0.5 and 1. For the other parameters, flat prior distributions were selected that would span the entire range of the posterior densities (i.e., uninformative priors) [30]. However, in some cases the posterior distributions were quite flat over the highest portions of parameter ranges. In these cases the choice of the upper bound on the prior distribution does affect the posterior distribution, and we are not able to use an uninformative prior distribution. However, parameters can still be estimated on the basis of the locations of peaks in the parameter regions that can be assessed, and the effect of altering the prior distribution on these estimates can be determined.

The overall picture that emerges is one in which the New World was very recently founded by a small number of individuals (effective size of about 70), and then grew by a factor of about 10. The data do suggest that there has been gene exchange between Asia and the New World since that time; however, the likelihood surfaces are quite flat, so confidence in gene flow estimates is low.

Results

The method assumes that the loci have not been subject to recombination or to directional or balancing selection. For recombination, we used only those loci that showed no evidence of recombination by the four-gamete test [38]. It is possible that this has missed some recombination since the time of common ancestry. Regarding natural selection, the study was limited to loci that had not individually been reported to show evidence of directional or balancing selection. However, it is possible that when considered together, and polymorphism and divergence from chimpanzees are considered under a common neutral model, that there is evidence of selection. An HKA test [39] of the eight loci with estimates of divergence from chimpanzees (Table 2) yielded a p value of 0.054, which is nearly statistically significant. This test assumes, as do the models analyzed in this study, that all loci are sampled from the same panmictic population [39], and it is possible that the differing geographic sources of the loci included in the study may have contributed some variation.

The estimated posterior distributions are shown in Figure 3. For the initial analysis, allowing for exponential population size changes, the posterior distribution for t yielded both a major and a minor peak (the curve for t with a high t_{upper}, Figure 3D). Given the mutation rate estimates (see Table 1), the location of the major peak ($t = 0.032$) corresponds to 7,130 y, whereas the location of the minor peak ($t = 0.27$) corresponds to 44,400 y. Given the remote possibility of such an ancient time as the latter, analyses were also done with a smaller upper bound on t of 0.2 (identified as "low t_{upper}" in Figure 3), which corresponds to 33,000 y. Analyses were done with this reduced upper limit for t for both models in Figure 1, allowing for population size change and for the case of fixed population sizes. In the case of constant population

Table 2. Information on Loci Used in the Study

Locus	Scalar[a]	Sample Sizes		Length (Basepairs)	D%[b]	F_{ST}[c]
		Asia	New World			
β-globin[d]	1.0	24	48	1,643	1.16	0.037
mtDNA[e]	0.25	8	7	15,440	0.81	Undef [c]
NRY[f]	0.25	13	13	26,500	1.66	0.505
Xq13.3[g]	0.75	9	3	10,138	0.94	Undef
ZFX[h]	0.75	50	58	1,134	1.50	0.02
ATM[i]	1.0	20	20	—[i]	1.6	Undef
APXL[j]	0.75	5	10	4,638	1.47	0.167
TNFSF5[j]	0.75	5	10	5,239	0.67	Undef
RRM2P4[j]	0.75	5	10	2,385	1.01	0.405

See Dataset S1 and Protocol S1 for more detail.

[a] The inheritance scalar was set to reflect the expected effective population size experienced by a locus relative to an autosome, assuming equal sex ratios and variance in reproductive success: autosomal loci, 1.0; X-linked loci, 0.75; maternally or paternally inherited loci, 0.25.

[b] The percentage of basepairs that differ between a human and a chimpanzee sequence.

[c] F_{ST} is the proportion of variation that lies between samples pooled for Asia and the New World for each locus [70,71]. When divergence is low, calculation may yield a negative value (Undef).

[d] Data from [72]. The β-*globin* locus falls near a recombination hotspot [73]. Of the 3,011 bases of a large population genetic study of the β-*globin* region [72], the 5′ half shows ample evidence of historical recombination by the four-gamete criterion [38], whereas the 3′ half that was used for this study showed no evidence of historical recombination. Divergence from chimpanzees was measured over this region from the available chimpanzee sequence [74].

[e] Full-length mtDNA sequences were used [75,76]. Because of the need for an absence of homoplasy by the computer program fitting the model, control region sequences were removed and only transversion differences were used.

[f] Concatenated data from several noncoding regions of the nonrecombining portion of the Y chromosomes (NRY) [48]. Human-chimpanzee divergence for the NRY was estimated from 4,758 noncoding basepairs of the *SMCY* locus [77].

[g] Data from [78,79].

[h] Data from [80,81].

[i] Haplotypes were determined over multiple points across this locus [82]. A data summary was provided by Yvonne Thorstenson. The region used for this analysis included pieces scattered over 96 kilobasepairs that showed no evidence of recombination in Asian and New World samples. This locus was not included in the estimate of mutation rate per year because of length ambiguity of the sampled sequence and uncertainty over human-chimpanzee divergence.

[j] Data from [83].

DOI: 10.1371/journal.pbio.0030193.t002

Table 3. Model Parameter Estimates

Parameter	Population Size Change High t_{upper}	Population Size Change Low t_{upper}	Constant Population Low t_{upper}
θ_1	2.55	6.72	1.92
θ_2	0.17 (0.095–17.7)	0.29 (0.086–23.10)	0.27 (0.102–1.10)
θ_A	3.26 (1.98–10.21)	3.17 (2.21–4.78)	3.44 (2.42–8.72)
t	0.032 (0.05–0.71)	0.028 (0.006–0.20)	0.20[a] (0.038)
m_1	9.27	3.12	3.33
m_2	10.08	9.68	16.63
s	0.992	0.992	NA

Parameter estimates are shown for three models described in the text. For those parameters in which the complete posterior distribution appeared to be estimated, the 90% highest posterior density interval was also determined and given as a range (in parentheses). This range is the shortest interval that contains 90% of the probability.

[a] The location of the highest value of *t* is at the right margin of the distribution. The location of the secondary peak is also given in parentheses.

NA, not applicable

DOI: 10.1371/journal.pbio.0030193.t003

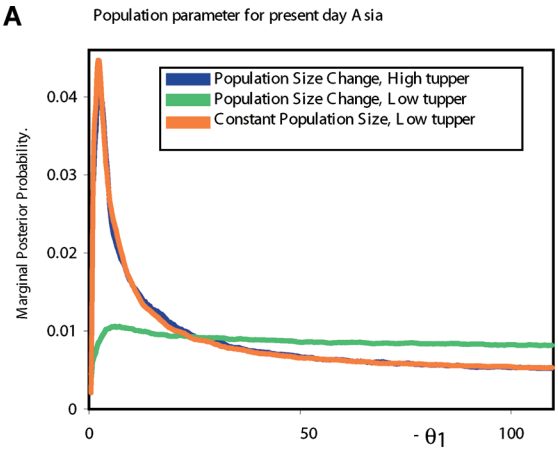
sizes, the distribution for *t* shows a peak ($t = 0.038$) very near those for the analyses under population size change; however, the highest posterior density is found at the upper limit of *t*. When the constant population size model was run with a higher upper limit on *t*, the posterior distribution showed the same low value peak as well as a steadily rising curve for higher values of *t* (unpublished data).

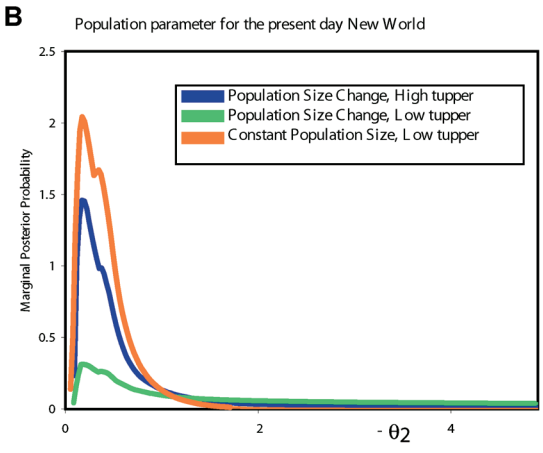
The archaeological portrait of early New World populations has largely centered around widespread Clovis sites that have an earliest estimated age of about 13,000 y before the present [15,40,41]. The oldest generally agreed upon New World archaeological date is from the non-Clovis Monte Verde site in Southern Chile, which has been dated to about 14,000 y before the present [10,42,43]. Clearly the time points associated with our estimates of *t* are more recent than expected, given the archaeological estimates. However, these distributions do span the time periods that have been most discussed. For example, a time of 14,000 y has a relatively high probability in each of the analyses (Figure 3E). Given that people have lived in the New World probably for only several hundred generations, it is noteworthy both that the posterior densities for *t* do show clear peaks in the expected time

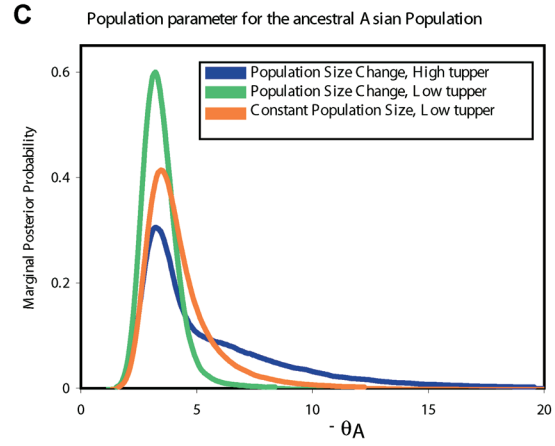
Figure 3. Marginal Posterior Probability Densities

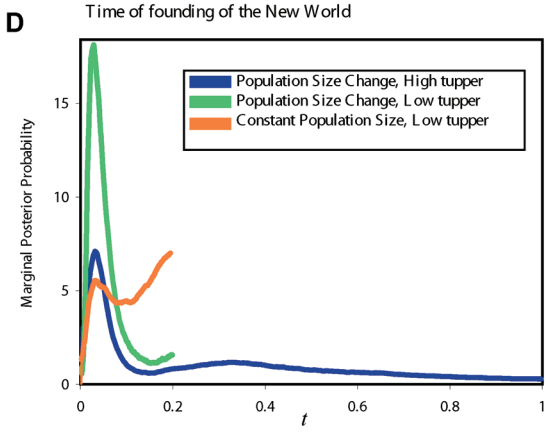
Probability densities for each of the parameters described in Figure 1 are shown, as follows: (A) θ_1; (B) θ_2; (C) θ_A; (D) *t* (i.e., t/u); (E) t shown on a scale of years over the range corresponding to a maximum *t* value of 0.2; (F) *s;* (G) m_1; and (H) m_2. The analysis in which a high upper limit on the prior distribution for *t* was used is identified as "high t_{upper}," while those analyses with a smaller upper limit on the prior distribution of *t* are identified as "low t_{upper}." Each curve is based upon the results of multiple simulations over millions of Markov chain updates (see Materials and Methods), and is plotted over the specified prior range of that parameter.

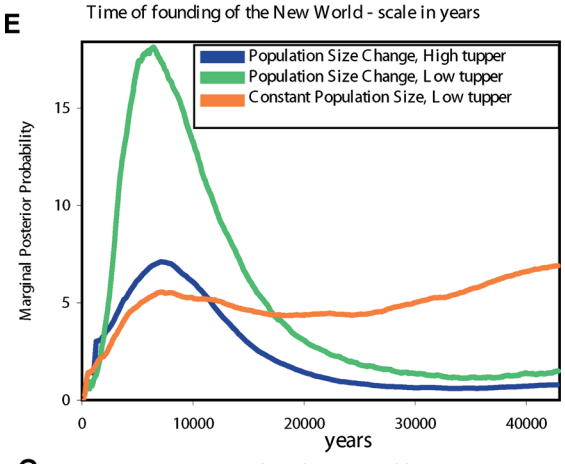
DOI: 10.1371/journal.pbio.0030193.g003

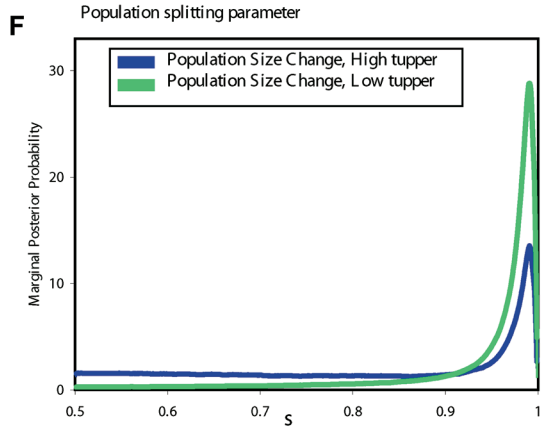

A
Population parameter for present day Asia
Population Size Change, High tupper
Population Size Change, Low tupper
Constant Population Size, Low tupper
Marginal Posterior Probability.
0.04
0.03
0.02
0.01
0
0
50
- θ1
100

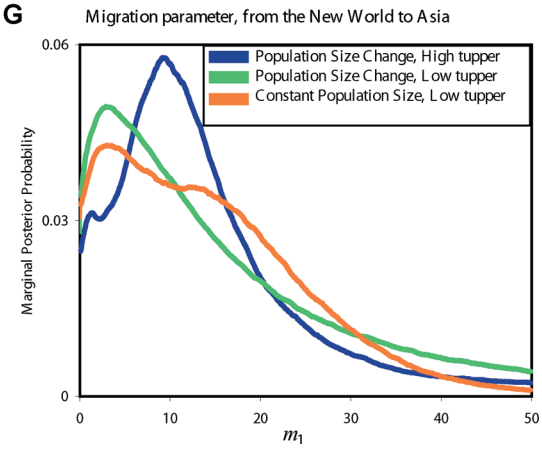

B
Population parameter for the present day New World
Population Size Change, High tupper
Population Size Change, Low tupper
Constant Population Size, Low tupper
Marginal Posterior Probability
2.5
2
1.5
1
0.5
0
0
2
- θ2
4


C
Population parameter for the ancestral Asian Population
Population Size Change, High tupper
Population Size Change, Low tupper
Constant Population Size, Low tupper
Marginal Posterior Probability
0.6
0.4
0.2
0
0
5
10
15
20
- θA


D
Time of founding of the New World
Population Size Change, High tupper
Population Size Change, Low tupper
Constant Population Size, Low tupper
Marginal Posterior Probability
15
10
5
0
0
0.2
0.4
0.6
0.8
1
t


E
Time of founding of the New World - scale in years
Population Size Change, High tupper
Population Size Change, Low tupper
Constant Population Size, Low tupper
Marginal Posterior Probability
15
10
5
0
0
10000
20000
30000
40000
years


F
Population splitting parameter
Population Size Change, High tupper
Population Size Change, Low tupper
Marginal Posterior Probability
30
20
10
0
0.5
0.6
0.7
0.8
0.9
1
s


G
Migration parameter, from the New World to Asia
Population Size Change, High tupper
Population Size Change, Low tupper
Constant Population Size, Low tupper
Marginal Posterior Probability
0.06
0.03
0
0
10
20
30
40
50
m1

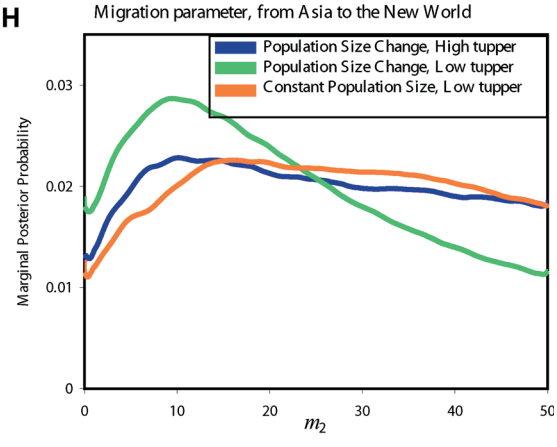

H
Migration parameter, from Asia to the New World
Population Size Change, High tupper
Population Size Change, Low tupper
Constant Population Size, Low tupper
Marginal Posterior Probability
0.03
0.02
0.01
0
0
10
20
30
40
50
m2

Table 4. Estimates of Demographic Quantities

Demographic Term	Population Size Change High t_{upper}	Population Size Change Low t_{upper}	Constant Population Size Low t_{upper}
N_1	7,190	19,200	5,394
N_2	480	830	770
N_A	9,180	9,040	9,640
s N_A	9,100	8,970	Not applicable
(1-s) N_A	76	70	Not applicable
t (years)	7,130	6,350	44,400[a] (7,900)
$2N_1m_1 = \theta_1 m_1 /2$	11.8	10.5	3.2
$2N_2m_2 = \theta_2 m_2 /2$	0.9	1.4	2.3

The conversion of model parameters to demographic terms is described in "Analyses" in Materials and Methods.
[a] The estimated time associated with the highest value of t which is at the right margin of the distribution. The estimated time associated with the secondary peak is given in parentheses.
DOI: 10.1371/journal.pbio.0030193.t004

period and that the probability estimates drop to zero as t approaches zero. In other words, the data contain a clear signal of a nonzero, albeit recent, founding time of New World populations.

With regard to migration, each of the three analyses show nonzero peaks for both directions of gene flow. These may well reflect the occurrence of more than one episode of migration to the New World. For example, it has been suggested on the basis of mitochondrial DNA haplotypes and glaciation history that an initial migration along a coastal route may have been followed later by another migration, possibly through an ice-free noncoastal corridor [13]. However, the posterior distributions shown here have little resolution, as all of the curves for m_1 and m_2 are broad, and all have high probability at the lower limit of resolution, indicating that zero gene flow is nearly as well supported by the data as are nonzero gene flow levels.

The ancestral population parameter, θ_A, shows a relatively narrow distribution with a very consistent peak location across the three analyses. These attributes are partly to be expected, given that the very large majority of the variation in the samples is older than t. In effect, more information is available for θ_A than for the other parameters. The estimated effective size of the ancestral population is about 9,000 (Table 3), which is roughly consistent with previous estimates for Asian samples [44]. The current Asian population parameter (θ_1) revealed broad distributions and estimates that are near those for the ancestral population. Although the estimates of current effective size in Asia vary among the analyses (Table 3), they are all fairly close to the ancestral size estimates, suggesting that there has not been much population growth in Asia since t. Also consistent with the apparent constancy of population size is the distribution of s, the splitting parameter, which shows a peak at 0.992, signifying that only a small portion (less than 1%) of the ancestral Asian population left to found the New World population.

In contrast to the Asian population, the New World population parameter (θ_2) is much smaller, and suggests a recent New World effective population size of less than 1,000 (Table 3). However, given the estimate of the effective size of the founding New World population (about 70; Table 4), the overall picture is of a nearly 10-fold growth in the New World effective size since t.

In order to gain a sense of how consistent the data actually are with the model and the parameter estimates, 500 simulated datasets were generated under the model in Figure 1B, with sample sizes and true parameter values (see Table 2, column 3) that were the same as for the actual data. From each simulated dataset, the average number of pairwise differences between sequences were calculated within each population (Asia and the New World) and between these populations. The average of these values from the 500 simulated datasets, and the observed values from the actual data, are shown in Table 5. In general, the observed and

Table 5. Contrasting Observed and Expected Levels of Variation

Locus	Within Asia		Within New World		Between	
	Obs	Exp	Obs	Exp	Obs	Exp
β-globin	3.0	1.5	2.4	0.6	2.84	2.2
mtDNA	1.9	2.2	2.3	1.1	1.89	3.0
NRY	1.5	2.4	1.0	1.3	2.54	3.5
Xq13.3	1.2	2.9	4.7	1.9	2.70	4.2
ZFX	1.0	1.1	0.8	0.3	0.95	1.6
ATM	3.6	2.4	3.7	1.3	3.52	3.6
APXL	1.0	1.7	3.2	0.7	2.52	2.4
TNFSF5	1.2	2.3	2.3	1.1	1.72	3.0
RRM2P4	2.8	2.9	1.9	1.6	3.92	3.9

Shown, both within and between populations, are the values of the average number of differences between pairs of sequences.
Exp, expected; Obs, observed
DOI: 10.1371/journal.pbio.0030193.t005

expected values are similar; however, one consistent pattern of departure is that the data from the New World, for most loci, show more variation by this measure than were found in the simulated data.

Discussion

The method described is one of several new approaches that can glean information about ancestral population sizes [30,45–47]. By including a new parameter for population splitting, it is possible to generate estimates not only of the size of the ancestral population, but also of the founding size of each founder population.

Taken together, the analyses in this study suggest a recent founding of the New World Amerind-speaking peoples by a small population of effective size near 70, followed by population growth in the New World. It is interesting that the analyses do not suggest much population size change in Asia since the time of the founding of the New World population. Given the very broad distributions for θ_1, it is possible that the true value of this parameter is much higher than suggested by the peak location, and that there has been considerable population growth in Asia. The analyses reveal very broad distributions for migration parameters, and although the peak locations suggest that gene flow has been fairly high (2Nm values greater than 1; see Table 3), the estimated probabilities of migration rates having been zero are also high (Figure 3G and 3H). Also, because Eskimo-Aleut and Na Déne speakers were not included in this study, the question of separate migrations for these groups has not been addressed [3].

As parameter-rich as the method is, neither this nor any mathematical model can be expected to fully represent the complex history of two related populations. However, the same is essentially true of narrative models, as investigators are always constrained by limited data and the need to keep explanations as simple as possible given their data. In this light, the IM model provides a fairly complete framework for some oft-debated questions on human history. With the addition of a new parameter, the IM framework can now also be used to address questions about the founding size of populations and of population size change.

In the context of human demographic history, the most problematic assumption under the IM model is that each population is panmictic. Certainly this is not the case today, and it is likely to have even been less true in times past. This raises the general and important question of how local patterns of population structure affect regional or global estimates of diversity [44,48,49]. Although this question cannot be answered here, the analyses do suggest that some kinds of departures from panmixia have not occurred. For example, if the New World had been founded by a local population that had long been separated from other Asian populations, then the estimate of t would be expected to reflect this older population structure, rather than the founding of the New World. Our generally low estimates of t argue against this scenario. Similarly, if the sampled Asian populations had been highly structured, with many long-separated local populations, then this would have inflated the estimates of N_A and N_1, respectively. However, the generally low estimates of effective population size argue against this particular kind of population structure.

The analyses presented here share with some other genetic studies estimated dates for the peopling of the Americas that are more recent than archeologically based estimates [8,9,16]. However, the difficulty of estimating such recent events using genetic data alone should not be overestimated [18]. When considering human populations within the past few tens of thousands of years, two gene copies that share the same haplotype will often have had a common ancestor far longer ago than any of the dates in question. Similarly, genetic evidence on the peopling of the Americas has been interpreted both as consistent with multiple migrations [12] and as indicating just a single founder event [16,19,50]. Divergent interpretations are understandable, given that a finding of two populations that share sequence haplotypes at a locus can be taken as evidence of two quite different models: (1) a recent population separation; or (2) gene exchange between populations.

The available data do not yet allow precise estimates of founding time nor of whether there has been gene flow between the New World and Asia following the initial founding event. However, the new method implements a parameter-rich model of divergence and has the potential to recover the history of complex divergence processes. The method can also be applied to a large number of loci, with large sample sizes, and in the future can be expected to provide increasingly detailed portraits of human population divergence.

Materials and Methods

Selected loci and samples. Given the prevailing model of the founding of New World populations via a Bering land bridge, the descendant populations were defined as the Amerind speakers of the New World and the peoples of northeastern Asia. Greenberg et al. [3] proposed that New World populations include three linguistic groups (Eskimo-Aleut, Na Déne, and Amerind), each associated with a separate episode or period of migration. Because of the limited number of published comparative DNA sequence studies that include samples from Eskimo-Aleut and Na Déne group, the present study was limited to samples from Amerind-speaking populations. Asian samples were limited to those from China, Mongolia, Korea, and Siberia. These are partly arbitrary boundaries selected as a balance between the need to include as many loci as possible and uncertainty about the present locations of descendants of those Asian populations that gave rise to the founders of the New World.

The model fitting requires data from loci that do not show evidence of recombination and that do not show clear evidence of directional or balancing selection. All available datasets from the literature that met these criteria and that had multiple DNA sequences from both of the designated sample regions were selected. The selected loci are listed in Table 1. The input data file is provided in Dataset S1, and a list of sample locations is provided in Protocol S1.

Model development. At the center of the method for estimating the parameters is an expression for the posterior probability distribution of model parameters Θ, given the data. For the case of multiple loci

$$f(\Theta|X_1, X_2, ..., X_n) = cf(\Theta)\prod_{i=1}^{n}\int_{G_i} f(X_i|G_i)f(G_i|\Theta)dG_i \qquad (1)$$

where Θ refers to the vector of parameters of the model, X_i refers to the data for locus i, and G_i is the genealogy for locus i [33]. With n loci, the full set of parameters includes six or seven demographic parameters, depending on the inclusion of s, as well as n locus-specific mutation rate scalars [33]. A genealogy includes the topology of an ultrametric tree, the associated coalescence times, and the times of migrations on each branch of the tree [30]. For a given locus i, the probability $f(X_i|G_i)$ is calculated using the mutation model for that locus and the branch lengths in the genealogy. The probability $f(G_i|\Theta)$

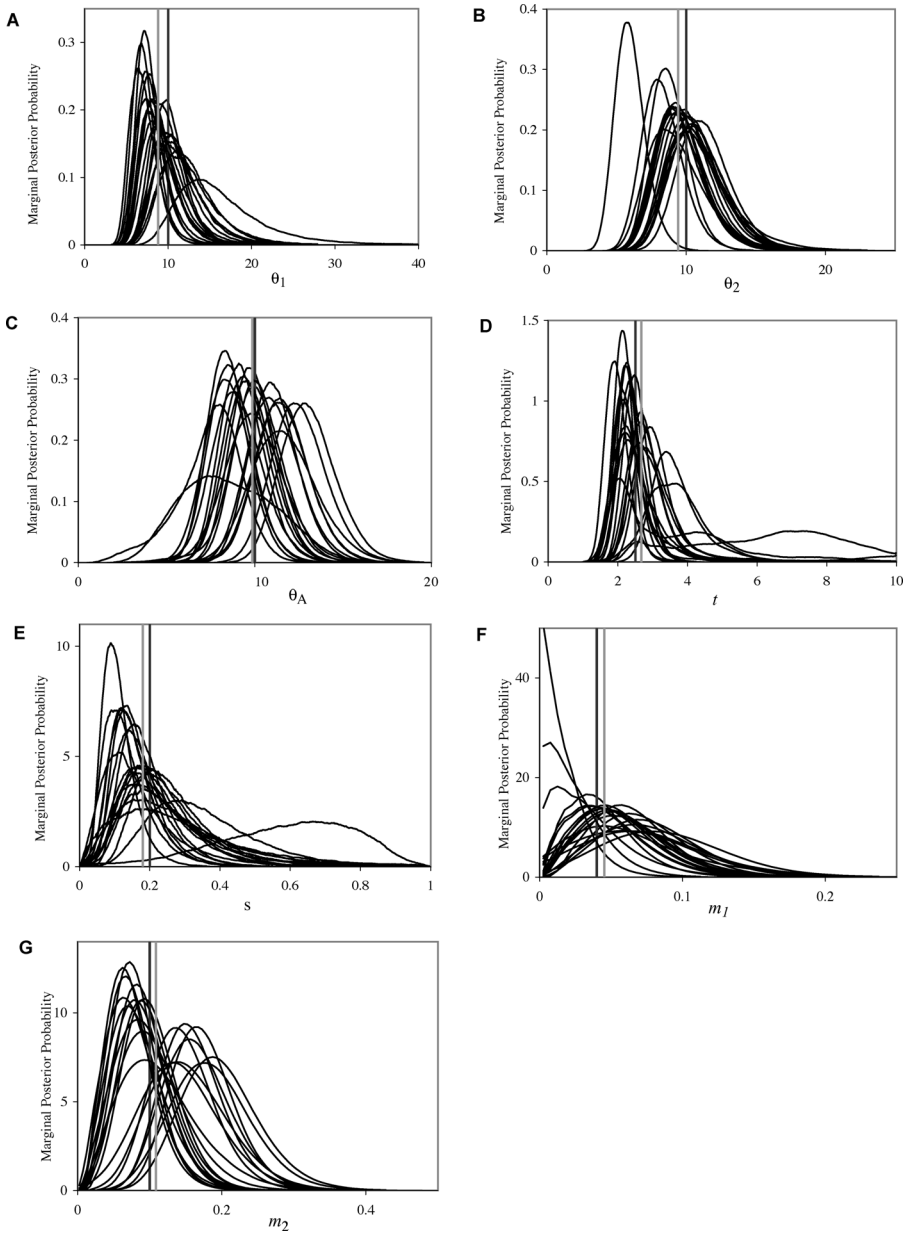


Figure 4. The Marginal Densities Obtained by Fitting the Model with Population Size Change to Simulated Data

The input parameters for the simulations were as follows: (A) $\theta_1 = 10$; (B) $\theta_2 = 10$; (C) $\theta_A = 10$; (D) $t = 2.5$, (E) $s = 0.2$, (F) $m_1 = 0.04$; (G) $m_2 = 0.2$; and $t = 5$ ($t/2N_A = 0.5$). For each simulated dataset, coalescent simulations were done for each of 20 loci with identical mutation rates under an infinite sites mutation model, each with sample sizes of 10 for each of the two populations. Each simulated dataset was analyzed using wide uniform prior distributions for each parameter. Each analysis began with a burn-in period of 300,000 steps followed by a primary chain of 3 million to 10 million steps. The curves for parameters θ_1 through m_2 are shown in (A) through (G), respectively. For each figure, the true parameter value used in the simulations is shown as a black vertical bar, and the mean of the estimates for the 20 simulations (based on peak locations) is shown as a gray vertical bar.

DOI: 10.1371/journal.pbio.0030193.g004

is calculated using expressions from basic coalescent theory [30,51–55]. By integrating over all possible genealogies that are consistent with the data, the results obtained are not conditioned on any particular estimate of the genealogy, and they necessarily incorporate all of the stochastic variance that arises among independent loci under the model.

The integration in Equation 1 cannot be solved directly for any but the simplest of models, but it can be approximated using a Markov chain simulation [56]. This approach was originally applied to the IM model by Nielsen and Wakeley [30], and then augmented to include multiple loci [33] and additional mutation models [32,57].

Over the course of a simulation the genealogy for a given locus varies for topology, branch lengths, and migration times. However, the probability of the data for a locus given a particular genealogy, $f(X_i|G_i)$, depends only upon the branch lengths and the mutation model for that locus [30]. Although inclusion of s will affect the genealogies that arise in the course of the simulation, there will be no effect on the calculation of the probability of the data for a given genealogy (i.e., $f(X_i|G_i)$ is not a function of s), and thus including s has no effect on the applicability of the method to diverse mutation models. In contrast, the probability of a genealogy given a set of parameter values, $f(G_i|\Theta)$, depends strongly on s because the probability of individual coalescent and migration times are functions of population size.

The calculation of $f(G_i|\Theta)$ is most directly done by taking the product of the probabilities of each of the coalescent and migration events that occur in the genealogy. Griffiths and Tavare [55] developed the general theory for the probability distribution of coalescence times when the population size is changing. Given a function $v(\tau) = N_\tau/N_0$ of the population size at time τ, relative to that at time 0, they provide a general expression for the distribution of coalescent times. For population 1, the effective size goes from N_1 at time zero, to sN_A at time t. If it is assumed that the size change is exponential over this period, then for population 1,

$$v(\tau) = \left(\frac{sN_A}{N_1}\right)^{\tau/t} \quad (2)$$

and for population 2,

$$v(\tau) = \left(\frac{(1-s)N_A}{N_2}\right)^{\tau/t} \quad (3)$$

One additional complication that arises is that when the population is growing exponentially back into the past (decreasing in size as time moves forward), there is a finite probability that the time to coalescence will be infinity [58]. Thus, for population 1 when sN_A is less than N_1, it is necessary to calculate the probability of coalescence time conditioned on there being a coalescent event.

Migration under an exponentially changing population size can also be incorporated under this same framework with two changes. First, unlike coalescence, where the rate is inversely proportional to population size, the rate of migration is directly proportional to population size. Second, as time goes backward in the coalescent, the migration rate from population 1 to population 2 (i.e., m_1) actually corresponds to the movement of genes from population 2 to population 1 as time moves forward. This means that in the coalescent under changing population size, we expect that the migration rate from population 1 to 2 will vary with the size of population 2. Thus the corresponding relative rate function for migration from population 2 to population 1 is

$$v(\tau) = \left(\frac{N_2}{(1-s)N_A}\right)^{\tau/t} \quad (4)$$

and for migration in the reverse direction it is

$$v(\tau) = \left(\frac{N_1}{sN_A}\right)^{\tau/t} \quad (5)$$

These intensity functions for coalescence and migration were used to develop an expression for $f(G_i|\Theta)$ that includes s, and that could be directly incorporated into the update criteria for all of the demographic, mutation, and inheritance scalars described in Hey and Nielsen [33]. Also needed, in order to allow for changing population size, are the update criteria for s and the update criteria for the genealogies. For s, updates are drawn from a uniform distribution over the user-specified prior range (e.g., in the current study, an interval within the range of 0.5 to 1). An update from s to s^* will affect the probability of all genealogies and thus has an acceptance probability, under the Metropolis Hastings criterion, of

$$\min\left\{1, \frac{f(s^*)q(s^* \to s)}{f(s)q(s \to s^*)} \prod_{i=1}^{n} \frac{f(G_i|s^*)}{f(G_i|s)}\right\} \quad (6)$$

where n is the number of loci and G_i is the current genealogy for locus i (see Equation 3 in Hey and Nielsen [33]). If we assume a uniform prior distribution for s, such that the prior probability of s, $f(s)$, is constant for all s, and if we choose updates such that the $q(s^* \to s) = q(s \to s^*)$ [30], then this simplifies to

$$\min\left\{1, \prod_{i=1}^{n} \frac{f(G_i * |s*)}{f(G_i|s)}\right\} \quad (7)$$

For genealogy updates the same proposal distribution of genealogies that was used in the case without s was retained, and then this proposal distribution was incorporated into the update criteria [59]. If $f(G_i|s)$ denotes the probability of the genealogy for locus i, given the other parameters including s, and $f(G_i)$ is the Hastings term for the proposal probability of the genealogy for locus i, given the other parameters1 excluding s, then the update criteria for the genealogy for locus i is

$$\frac{f(X_i|G_i{}^*)f(G_i{}^*|s)f(G_i)}{f(X_i|G_i)f(G_i|s)f(G_i{}^*)} \quad (8)$$

Performance. The IM computer program [33] was modified to include the additional parameter. The program is available from http://lifesci.rutgers.edu~heylab/HeylabSoftware.htm#IM. For the Markov chain simulation that is implemented by the program, it is difficult to assess how well the method works, because of the need to generate large numbers of simulated datasets and because of the long run times required [33]. To conduct testing, a program was written to generate simulated datasets under the models in Figure 1. Datasets were simulated in groups of 10 or 20, each having 10–20 loci, for a given set of parameter values, and for a range of parameter values. Figure 4 shows the marginal posterior densities estimated from each of 20 independent simulations for a case of modest population growth with the following parameter values. $\theta_1 = 10$; θ_2, $= 10$; $\theta_A = 10$; $t = 2.5$; $s = 0.2$; $m_1 = 0.04$; and $m_2 = 0.1$. For each parameter, the mean of the 20 estimates is shown, and in general these are fairly close to the true value, though there is considerable variance for the peak locations in individual runs. To test whether the locations of these distributions are consistent with the true values of the parameters (i.e., the values used in the simulations), probabilities were combined by treating each simulation as an independent test of the same hypothesis [60]. For each posterior density p_i, $i = 1,\ldots,20$, is the chance that a parameter value is more extreme (i.e., departs more from the mean of the distribution) than is the actual true value. That is, if x is the area of the curve to the left of the true value then $p_i = 2x$ if $x < 0.5$ and $p_i = 2(1-x)$ if $x > 0.5$. If the p_i's are uniformly distributed, then the quantity

$$z = -2\sum_{i=1}^{20} Log(p_i) \quad (9)$$

is χ^2 distributed with 40 degrees of freedom (i.e., two times the number of densities). The z values were as follows: θ_1, 35.5; θ_2, 26.4; θ_A, 41.7; t, 41.1; s, 26.4; m_1, 29.9; m_2, 44.1; and the mean of the seven values was 35.0. In the corresponding χ^2 distribution, 90% of the probability mass falls above 29.05; 50% falls above 39.3; and 10% falls above 51.8 [61]. Clearly these values are not entirely independent of each other, but they all fall in the middle of the χ^2 distribution with a mean (35.0) close to the 50% point of the χ^2 distribution (39.3).

From these simulations, and many others (additional results provided in Protocol S1), it is clear that sample sizes do need to be large for the posterior distributions to be informative. With data from fewer than five loci or fewer than ten individuals per population per locus, it is often the case that distributions are very flat or that there are multiple peaks. There is a tradeoff in sampling effort required for different kinds of histories. When t is small, sampling effort should be shifted to larger sample sizes per locus, whereas when t is large, sampling effort should be shifted toward more loci. This tradeoff is a byproduct of the fact that the stochastic variance among loci, that is associated with coalescent and migration events in genealogies at times near t, goes up as t increases. Another tradeoff that arises is between s and the migration rate parameters. Just as the frequency of polymorphic sites can be used to estimate changes in population size [62], it can also be appreciated that the information for s must reside in the distribution of times of node intervals in the descendant populations. Migration can have dramatic effects on node interval times within populations. In practice, via simulation, the method does not resolve a sharp peak for s for populations that have had more than moderate amounts of migration (e.g., 2Nm values are greater than 0.5; see Protocol S1).

Analyses. Each of the three analyses were done using at least three independent runs, with ten or more independent chains under Metropolis coupling [33] as described by Geyer [63]. Each chain was initiated with a burn-in period of 100,000 updates, and the total run length of each analysis was between 10 million and 30 million updates. The mixing properties of individual runs were monitored by measuring the autocorrelation of individual parameters over the course of the run, and by estimating the effective sample size for each of the parameters as a function of the autocorrelation estimates (see p. 499 in [64]). Analyses were taken to have converged upon the stationary distribution if independent runs generated similar distributions, with each having a lowest effective sample size of 50 for the time parameter (the parameter to show the slowest rate of mixing).

To convert estimates of parameters that include the mutation rate to more easily interpreted units, a value of 6 million y since the splitting of human and chimpanzee lineages was used [65–69]. The

geometric mean of the human-chimpanzee DNA sequence divergence of each locus, except *ATM* (see Table 2), was calculated and then used as a molecular clock calibration for converting the estimate of the time parameter, *t*, to divergence in years. The geometric mean mutation rate across these loci was estimated to be 4.66×10^{-6} mutations per year. The geometric mean is used rather than an arithmetic mean, because under the multilocus model, the mutation rate by which demographic parameters are scaled is the geometric mean of the individual locus-specific mutation rates [33].

To convert the estimates of the population mutation rate parameters (θ_1, θ_2, and θ_A) to estimates of effective population size (N_1, N_2, and N_A, respectively) a measure of mutation rate on a scale of generations is needed. Thus, an assumption was made of 20 y per generation, and the geometric mean divergence between humans and chimpanzees for each species contrast was divided by 12 million y then multiplied by 20 y per generation. These calculations yielded a geometric mean value of 9.32×10^{-5} mutations per generation. These mutation rate values were then used to convert individual θ estimates to effective population size estimates (i.e., $\theta = 4Nu$, and $N = \theta/4u$).

Migration parameters in the model can be used to obtain population migration rate estimates (i.e., $M = 2Nm$, the product of the effective number of gene copies and the per gene copy migration rate) using an estimate of the population mutation rate ($\theta = 4Nu$). Thus $\theta \times m/2 = (4Nu \times m/u)/2 = 2Nm$ [32].

Supporting Information

Dataset S1. Peopling of Americas Data File: Nine Loci

This is the input file that contains all of the data and that was analyzed using the IM computer program.

DOI: 10.1371/journal.pbio.0030193.sd001 (582 KB TXT).

Protocol S1. Additional Simulations and List of Sample Locations

DOI: 10.1371/journal.pbio.0030193.sd002 (92 KB DOC).

Acknowledgments

John Wakeley, Tad Schurr, and David Meltzer provided input on an early draft of the paper. Rasmus Nielsen provided some helpful suggestions on parameter updating. Thanks also to three reviewers for very helpful suggestions and critique.

Competing interests. The author has declared that no competing interests exist.

Author contributions. JH conceived and designed the model and analyses, selected the datasets, wrote the computer programs, performed the analyses, and wrote the paper. ∎

References

1. Brace CL, Nelson AR, Seguchi N, Oe H, Sering L, et al. (2001) Old World sources of the first New World human inhabitants: A comparative craniofacial view. Proc Natl Acad Sci U S A 98: 10017–10022.
2. Goebel T (1999) Pleistocene human colonization of Siberia and peopling of the Americas: An ecological approach. Evol Anthropol 8: 208–227.
3. Greenberg JH, Turner CG, Zegura SL (1986) The settlement of the Americas: A comparison of the linguistic, dental and genetic evidence. Curr Anthropol 27: 477–497.
4. Wallace DC, Garrison K, Knowler WC (1985) Dramatic founder effects in Amerindian mitochondrial DNAs. Am J Phys Anthropol 68: 149–155.
5. Goebel T, Waters MR, Dikova M (2003) The archaeology of Ushki Lake, Kamchatka, and the Pleistocene peopling of the Americas. Science 301: 501–505.
6. Nichols J (1990) Linguistic diversity and the first settlement of the New World. Language 66: 475–521.
7. Nettle D (1999) Linguistic diversity of the Americas can be reconciled with a recent colonization. Proc Natl Acad Sci U S A 96: 3325–3329.
8. Starikovskaya YB, Sukernik RI, Schurr TG, Kogelnik AM, Wallace DC (1998) mtDNA diversity in Chukchi and Siberian Eskimos: Implications for the genetic history of Ancient Beringia and the peopling of the New World. Am J Hum Genet 63: 1473–1491.
9. Lell JT, Brown MD, Schurr TG, Sukernik RI, Starikovskaya YB, et al. (1997) Y chromosome polymorphisms in native American and Siberian populations: Identification of native American Y chromosome haplotypes. Hum Genet 100: 536–543.
10. Fiedel SJ (2000) The peopling of the New World: Present evidence, new theories, and future directions. J Archaeol Res 8: 39–103.
11. Tarazona-Santos E, Santos FR (2002) The peopling of the americas: A second major migration? Am J Hum Genet 70: 1377–1380.
12. Lell JT, Sukernik RI, Starikovskaya YB, Su B, Jin L, et al. (2002) The dual origin and Siberian affinities of Native American Y chromosomes. Am J Hum Genet 70: 192–206.
13. Schurr TG, Sherry ST (2004) Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: Evolutionary and demographic evidence. Am J Human Biol 16: 420–439.
14. Gonzalez-Jose R, Dahinten SL, Luis MA, Hernandez M, Pucciarelli HM (2001) Craniometric variation and the settlement of the Americas: Testing hypotheses by means of R-matrix and matrix correlation analyses. Am J Phys Anthropol 116: 154–165.
15. Taylor RE, Haynes CV, Jr, Stuiver M (1996) Clovis and Folsom age estimates: Stratigraphic context and radiocarbon calibration. Antiquity 70: 515–525.
16. Bonatto SL, Salzano FM (1997) Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the New World. Am J Hum Genet 61: 1413–1423.
17. Merriwether DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, et al. (1991) The structure of human mitochondrial DNA variation. J Mol Evol 33: 543–555.
18. Malhi RS, Eshleman JA, Greenberg JA, Weiss DA, Schultz Shook BA, et al. (2002) The structure of diversity within New World mitochondrial DNA haplogroups: Implications for the prehistory of North America. Am J Hum Genet 70: 905–919.
19. Silva WA, Jr, Bonatto SL, Holanda AJ, Ribeiro-Dos-Santos AK, Paixao BM, et al. (2002) Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. Am J Hum Genet 71: 187–192.
20. Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110: 325–344.
21. Hudson RR (1990) Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J, editors. Oxford Surveys in Evolutionary Biology. New York: Oxford University Press. pp. 1–44.
22. Hudson RR, Turelli M (2003) Stochasticity overrules the "three-times rule": Genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57: 182–190.
23. Knowles LL, Maddison WP (2002) Statistical phylogeography. Mol Ecol 11: 2623–2635.
24. Templeton AR (2004) Statistical phylogeography: Methods of evaluating and minimizing inference errors. Mol Ecol 13: 789–809.
25. Hey J, Machado CA (2003) The study of structured populations—New hope for a difficult and divided science. Nat Rev Genet 4: 535–543.
26. Felsenstein J (1992) Estimating effective population size from samples of sequences: A bootstrap Monte Carlo integration method. Genet Res 60: 209–220.
27. Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140: 1421–1430.
28. Griffiths RC, Tavare S (1994) Ancestral inference in population genetics. Stat Sci 9: 307–319.
29. Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152: 763–773.
30. Nielsen R, Wakeley J (2001) Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158: 885–896.
31. Wakeley J, Hey J (1998) Testing speciation models with DNA sequence data. In: DeSalle R, Schierwater B, editors. Molecular Approaches to Ecology and Evolution. Basel: Birkhäuser Verlag. pp. 157–175.
32. Hey J, Won Y-J, Sivasundar A, Nielsen R, Markert JA (2004) Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. Mol Ecol 13: 909–919.
33. Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167: 747–760.
34. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: A study of Y chromosome microsatellites. Mol Biol Evol 16: 1791–1798.
35. Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, et al. (1998) Genetic traces of ancient demography. Proc Natl Acad Sci U S A 95: 1961–1967.
36. Sherry ST, Rogers AR, Harpending H, Soodyall H, Jenkins T, et al. (1994) Mismatch distributions of mtDNA reveal recent human population expansions. Hum Biol 66: 761–775.
37. Watkins WS, Ricker CE, Bamshad MJ, Carroll ML, Nguyen SV, et al. (2001) Patterns of ancestral human diversity: An analysis of alu-insertion and restriction-site polymorphisms. Am J Hum Genet 68: 738–752.
38. Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147–164.
39. Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.
40. Meltzer DJ (1995) Clocking the first Americans. Annu Rev Anthropol 24: 21–45.
41. Anderson DG, Faught MK (2000) Palaeoindian artefact distributions: Evidence and implications. Antiquity 74: 507–513.

42. Dillehay TD, editor (1996) Monte Verde: A late Pleistocene settlement in Chile. Vol 2: The archaeological context and interpretation. Washington, D C: Smithsonian Insititution Press. 1071 p..
43. Meltzer DJ (1997) Monte Verde and the Pleistocene peopling of the Americas. Science 276: 754–755.
44. Tishkoff SA, Verrelli BC (2003) Patterns of human genetic diversity: Implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4: 293–340.
45. Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645–1656.
46. Rosenberg NA, Feldman MW (2002) The relationship between coalescence times and population divergence times. In: Slatkin M, Veuille M, editors. Modern developments in theoretical population genetics. Oxford: Oxford University Press. pp. 130–164.
47. Takahata N, Lee SH, Satta Y (2001) Testing multiregionality of modern human origins. Mol Biol Evol 18: 172–183.
48. Hammer MF, Blackmer F, Garrigan D, Nachman MW, Wilder JA (2003) Human population structure and its effects on sampling y chromosome sequence variation. Genetics 164: 1495–1509.
49. Zietkiewicz E, Yotova V, Gehl D, Wambach T, Arrieta I, et al. (2003) Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. Am J Hum Genet 73: 994–1015.
50. Merriwether DA, Rothhammer F, Ferrell RE (1995) Distribution of the four founding lineage haplotypes in Native Americans suggests a single wave of migration for the New World. Am J Phys Anthropol 98: 411–430.
51. Kingman JFC (1982) The coalescent. Stochastic Processes Appl 13: 235–248.
52. Kingman JFC (1982) On the genealogy of large populations. J Appl Probab 19A: 27–43.
53. Tavare S (1984) Line-of-Descent and genealogical processes, and their applications in population genetics models. Theor Popul Biol 26: 119–164.
54. Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23: 183–201.
55. Griffiths RC, Tavare S (1994) Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci 344: 403–410.
56. Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo in practice. Boca Raton, FL: Chapman and Hall. 486 p.
57. Palsbøll PJ, Berube M, Aguilar A, Notarbartolo-Di-Sciara G, Nielsen R (2004) Discerning between recurrent gene flow and recent divergence under a finite-site mutation model applied to North Atlantic and Mediterranean Sea fin whale *(Balaenoptera physalus)* populations. Evolution Int J Org Evolution 58: 670–675.
58. Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149: 429–434.
59. Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97–109.
60. Fisher RA (1954) Statistical methods for research workers. Edinburgh: Oliver and Boyd. 360 p.
61. Rohlf FJ, Sokal RR (1981) Statistical tables. San Francisco: W. H. Freeman and Company. 192 p.
62. Tajima F (1989) The effect of change in population size on DNA polymorphism. Genetics 123: 597–601.
63. Geyer CJ (1991) Markov chain Monte Carlo maximum likelihood. In: Keramidas EM, editor. Computing science and statistics. Proceedings of the 23rd Symposium on the Interface; April 21–24, 1991. Seattle, Washington: Interface Foundation of North America. pp. 156–163.
64. Robert CP, Casella G (2004) Monte Carlo statistical methods. New York, New York: Springer. 645 p.
65. Glazko GV, Nei M (2003) Estimation of divergence times for major lineages of primate species. Mol Biol Evol 20: 424–434.
66. Vignaud P, Duringer P, Mackaye HT, Likius A, Blondel C, et al. (2002) Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. Nature 418: 152–155.
67. Brunet M, Guy F, Pilbeam D, Mackaye HT, Likius A, et al. (2002) A new hominid from the Upper Miocene of Chad, Central Africa. Nature 418: 145–151.
68. Wildman DE, Uddin M, Liu G, Grossman LI, Goodman M (2003) Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: Enlarging genus *Homo*. Proc Natl Acad Sci U S A 100: 7181–7188.
69. Chen F-C, Li W-H (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68: 444–456.
70. Wright S (1951) The genetical structure of populations. Ann Eugenics 15: 323–354.
71. Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.
72. Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, et al. (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60: 772–789.
73. Chakravarti A, Buetow KH, Antonarakis SE, Waber PG, Boehm CD, et al. (1984) Nonuniform recombination within the human β-globin gene cluster. Am J Hum Genet 36: 1239–1258.
74. Savatier P, Trabuchet G, Faure C, Chebloune Y, Gouy M, et al. (1985) Evolution of the primate beta-globin gene region. High rate of variation in CpG dinucleotides and in short repeated sequences between man and chimpanzee. J Mol Biol 182: 21–29.
75. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, et al. (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100: 171–176.
76. Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408: 708–713.
77. Shen P, Wang F, Underhill PA, Franco C, Yang WH, et al. (2000) Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci U S A 97: 7354–7359.
78. Kaessmann H, Heissig F, von Haeseler A, Paabo S (1999) DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet 22: 78–81.
79. Kaessmann H, Wiebe V, Paabo S (1999) Extensive nuclear DNA sequence diversity among chimpanzees. Science 286: 1159–1162.
80. Jaruzelska J, Zietkiewicz E, Batzer M, Cole DE, Moisan JP, et al. (1999) Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron. Analysis of the haplotype structure and genealogy. Genetics 152: 1091–1101.
81. Jaruzelska J, Zietkiewicz E, Labuda D (1999) Is selection responsible for the low level of variation in the last intron of the ZFY locus? Mol Biol Evol 16: 1633–1640.
82. Thorstenson YR, Shen P, Tusher VG, Wayne TL, Davis RW, et al. (2001) Global analysis of ATM polymorphism reveals significant functional constraint. Am J Hum Genet 69: 396–412.
83. Hammer MF, Garrigan D, Wood E, Wilder JA, Mobasher Z, et al. (2004) Heterogeneous patterns of variation among multiple human X-linked loci: The possible role of diversity-reducing selection in non-Africans. Genetics 167: 1841–1853.