A Statistical Evaluation of Models for the Initial Settlement of the American Continent Emphasizes the Importance of Gene Flow with Asia

N. Ray,[1,2] D. Wegmann,[1,2] N.J.R. Fagundes,[3] S. Wang,[4] A. Ruiz-Linares,[4] and L. Excoffier*[,1,2]

[1]Computational and Molecular Population Genetics, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Department of Genetics, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

[4]The Galton Laboratory, Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

***Corresponding author:** E-mail: laurent.excoffier@iee.unibe.ch.

Associate editor: Beth Shapiro

Abstract

Although there is agreement in that the Bering Strait was the entry point for the initial colonization of the American continent, there is considerable uncertainty regarding the timing and pattern of human migration from Asia to America. In order to perform a statistical assessment of the relative probability of alternative migration scenarios and to estimate key demographic parameters associated with them, we used an approximate Bayesian computation framework to analyze a data set of 401 autosomal microsatellite loci typed in 29 native American populations. A major finding is that a single, discrete, wave of colonization is highly inconsistent with observed levels of genetic diversity. A scenario with two discrete migration waves is also not supported by the data. The current genetic diversity of Amerindian populations is best explained by a third model involving recurrent gene flow between Asia and America, after initial colonization. We estimate that this colonization involved about 100 individuals and occurred some 13,000 years ago, in agreement with well-established archeological data.

Key words: human settlement, colonization, Amerindians, approximate Bayesian computation, model choice.

Introduction

For decades, the initial colonization of the American continent has been a subject of investigation through a multitude of research approaches (Cavalli-Sforza et al. 1994; Crawford 1998). In a seminal publication integrating linguistic, dental, and genetic evidence, Greenberg et al. (1986) proposed that the American continent was settled by three distinct migration waves. According to this model, an initial migration wave would be at the origin of the well-established Clovis cultural complex (dated at about 13,000 years ago, Waters and Stafford 2007) and resulted in the dispersal of a single, large linguistic family (Amerind) across the continent. Two subsequent migrations would have been associated with the appearance of the Na Dene and Eskimo–Aleutian linguistic families, restricted to North America. Since its proposal, Greenberg's model has been a major reference point for the interpretation of novel data, including molecular genetic evidence. Two features of this model have been the subject of particular scrutiny: the time of initial colonization and the number of migratory waves. The discovery of archaeological sites predating Clovis (e.g., Dillehay 1997; Adovasio and Pedler 2004; Joyce 2006) has been particularly influential in stirring debates around the antiquity of the initial settlement of America. Analyses of genetic data have generally been found to be consistent with a pre-Clovis settlement (e.g., Bortolini et al. 2003; Fuselli et al. 2003; Zegura et al. 2004; Tamm et al. 2007; Fagundes et al. 2008), although with considerable uncertainty about exactly how much older. Several recent studies (particularly using genetic and craniometric data) have also questioned Greenberg's three-migration model and suggested alternative scenarios, particularly that one or perhaps two discrete migration waves (e.g., Neves and Pucciarelli 1991; Merriwether et al. 1995; Merriwether and Ferrell 1996; Bonatto and Salzano 1997; Santos et al. 1999; Lell et al. 2002; Bortolini et al. 2003; Gonzalez-Jose et al. 2005; Neves and Hubbe 2005; Powell 2005) or that gene flow across the Arctic (González-José et al. 2008) could be at the origin of the population diversity observed across the continent.

Recently, efficient statistical tools have been developed to contrast alternative demographic models using genetic data (Beaumont and Rannala 2004). Approximate Bayesian computation (ABC) methods (Beaumont et al. 2002) have been useful to explore the complex demographic history typical of natural populations (e.g., Miller et al. 2005; Fagundes et al. 2007; Pascual et al. 2007; Neuenschwander et al. 2008). In brief, ABC methods have been introduced

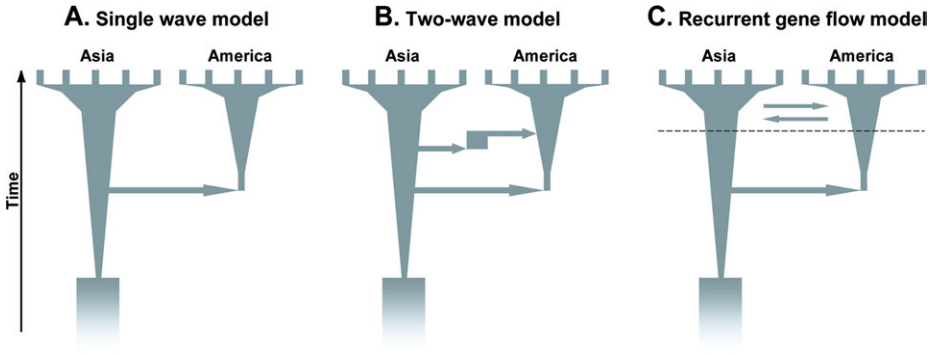

FIG. 1. Alternative models for the colonization of the Americas tested in this study.

to contrast evolutionary models and estimate their parameters when a likelihood cannot be computed but for which simulations can be performed (Tavare et al. 1997; Pritchard et al. 1999; Beaumont et al. 2002). Model parameters are drawn from specified priors and used to simulate data matching the observations in terms of sample size and number of loci. Simulated data are summarized with a set of summary statistics $\mathbf{S}$ and compared with the observed statistics $\mathbf{S}_o$. Simulations judged sufficiently close to the observed data (by means of an Euclidean distance $\delta_i = ||\mathbf{S}_i - \mathbf{S}_o||$) are retained for parameter estimation. This simple rejection scheme can also be used for model choice (Pritchard et al. 1999; Estoup et al. 2004; Fagundes et al. 2007). Here, we apply this ABC approach to the colonization of America. We contrast three scenarios that summarize the main controversies concerning the colonization of the continent. An outline of these three models is shown in figure 1. A detailed representation, including the demographic parameters defining the models, is shown in supplementary figure S1, Supplementary Material online.

Materials and Methods

Scenarios of Settlement

The first scenario is a single-wave (SW) model, which posits that all native American diversity stems from a single migration event from Asia occurring T_{W1} generations ago, without any subsequent gene flow between the two continents. The main feature of this model is that it allows colonization times that are younger or older than the earliest known archaeological remains in the Americas.

The second scenario is a two-wave (2W) model that allows for a second migration from Asia to have occurred more recently (T_{W2} generations ago). During this second wave, the subpopulation originating in Asia was isolated for a short period (10 generations) before migrating to the Americas. Note that the impact of this second wave on current genetic diversity is a parameter of the model. This parameter is modeled by the probability (M_{P2}, randomly chosen between 0 and 1) that current gene lineages originate from this second wave. If $M_{P2} = 0$, the 2W model converges to the SW model, and if $M_{P2} = 1$, the 2W model converges to a case of complete and late replacement. The introduction of a fixed isolation time (10 generations) for the second wave from Asia allows for a separate bottleneck. Because the strength of this bottleneck depends on the ratio of population size and its duration, we allow for the bottleneck population size Nb_{AM2} to determine its intensity.

The third scenario is identical to the SW model but allows for asymmetric and recurrent gene flow (RGF) between Asia and the Americas after the initial colonization. The prior distribution for the onset of this gene flow is identical to that of the time of the second wave in the 2W model (see supplementary table S1, Supplementary Material online). Note that we did not study a three-wave model because the third wave proposed by Greenberg et al. (1986) would have given rise to Eskimo–Aleut populations, which are not represented in our data set.

Overview of Demographic and Genetic Modeling

In order to allow for recent coalescent events to occur within each population/tribe, we considered that the populations have been isolated for a short period of time (T_{POP}, identical in Asia and in the Americas). The population sizes of these isolated populations were assumed to be gamma distributed as Gamma(10, 10/X), where X is either the average current size of Asian (P_{AS}) or Amerindian (P_{AM}) populations (see supplementary table S1, Supplementary Material online). With this parameterization, we allow samples to have different population sizes and thus to have different rates of drift and therefore to show variable levels of genetic diversity (Wang et al. 2007). Going backward in time, the isolation of the populations ends, and all remaining lineages are brought together in a large continental metapopulation of effective size N_{AS} and N_{AM} in Asia and America, respectively. The size of this continental population then decreases exponentially backward in time until reaching the initial bottleneck size (Nb_{AS} or Nb_{AM}). This modeling of a single unstructured population for the Americas and Asia is justified by theoretical work showing that the structure of a coalescent in a metapopulation is identical to that of a single stable population, only that adjusted effective population size allows a rescaling of the coalescent times (Wakeley 2001; Wakeley and Aliacar 2001).

Although our focus is to compare scenarios of the settlement of the Americas, we also had to model the genetic

diversity of current Asian populations. We chose a simple model of exponential growth of the Asian metapopulation after an initial bottleneck, which could represent the out-of-Africa expansion (see, e.g., Fagundes et al. 2007). This model is consistent with studies suggesting that Asian populations went through a recent period of exponential growth (e.g., Pritchard et al. 1999; Hamilton et al. 2005; Fagundes et al. 2007). An important assumption of this model is that the Asian populations included in the analyses belong to the same metapopulation as the populations that colonized the Americas.

Genetic Data

We used part of the worldwide data set of Wang et al. (2007) comprising 78 populations typed at 678 autosomal microsatellite loci. We retained all 29 Amerindian populations from this data set. Because the place of origin in Asia of the initial migrants to America is not well established, we considered (as representative of the Asian source "metapopulation") either a set of the 10 East Asian populations geographically closest to the American continent (called here the Asian10S data set) or a set of the two Central Siberian populations for which data are available (the Asian2S data set) because Central Southern Siberia has often been considered a major potential source for the settlement of the Americas (Cavalli-Sforza et al. 1994; Karafet et al. 1999; Santos et al. 1999; Lell et al. 2002; Bortolini et al. 2003; Wang et al. 2007). The Asian2S data set included the Tundra Nentsi and Yakut populations. The Asian10S data set included the following additional eight populations: Oroqen, Hezhen, Daur, Japanese, Mongolia, Han, She, and Tujia. The complete data set therefore includes 691 individuals from 39 populations. We first re-coded all microsatellite alleles into number of repeats, to allow comparison with our simulations, performed under a strict stepwise mutation model (SMM). Alleles not fitting a pure SMM were coded as missing data. We then used Arlequin ver. 3.1 (Excoffier, Laval, and Schneider 2005) to detect and remove loci with more than 5% missing data over all individuals, leaving 407 loci. Finally, we removed all (six) dinucleotide repeat loci because of their high mutation rate (Zhivotovsky et al. 2003). The final data set consisted of 401 loci, including 75 tri-, 320 tetra-, and 6 pentanucleotide repeat loci.

In order to evaluate whether different colonization events impacted North and South America, we carried out separate analyses for North/Central American populations (Chipewyans, Cree, Ojibwa, Cabecar, Guaymi, Quichean, Maya, Mixe, Mixtec, Pima, and Zapotec) and South American populations (Arhuaco, Aymara, Embera, Huilliche, Ingano, Kogi, Quechua, Waunana, Wayuu, Zenu, Ache, Guarani, Kaingang, Karitiana, Piapoco, Surui, Ticuna North, and Ticuna South). The statistical analyses were thus performed on the six possible combinations of the different Asian and Amerindian data sets: Asian2S/All America, Asian2S/North America, Asian2S/South America, Asian10S/All America, Asian10S/North America, and Asian10S/South America.

Approximate Bayesian Computation

We briefly outline below the ABC procedure (for details, see Beaumont et al. 2002; Excoffier, Estoup, and Cornuet 2005). For each model, we used the program SIMCOAL ver. 2.0 (Laval and Excoffier 2004) to perform coalescent simulations. For each simulation, we drew model parameters from their prior distributions listed in supplementary table S1, Supplementary Material online. These parameters were used to write an input file used by SIMCOAL to simulate the genetic diversity of samples having the same properties than those observed (i.e., same sample size and same number of loci). Microsatellite diversity was generated under a strict SMM. Summary statistics ($\mathbf{S}$) identical to those computed on the observed data ($\mathbf{S}_o$) were then calculated for the simulated data set. Note that we simulated haploid individuals and we therefore report population sizes in haploid number of genes. Following Beaumont et al. (2002), an Euclidean distance $\delta = ||\mathbf{S} - \mathbf{S}_o||$ was calculated between observed and simulated summary statistics (which were previously normalized, see below). The prior distributions of the parameters of the three models examined are shown in supplementary table S1, Supplementary Material online. We always used the same prior distribution for parameters common to all models. We used different average mutation rates $(\bar{\mu})$ for each category of microsatellite defined by their repeat lengths, as estimated by Zhivotovsky et al. (2003) under a pure SMM mode. We used $\bar{\mu} = 7.1 \times 10^{-4}$ per generation for trinucleotide and $\bar{\mu} = 6.4 \times 10^{-4}$ per generation for tetra- and pentanucleotide loci. Individual locus mutation rates μ_i were then sampled from a gamma distribution Gamma(α,$\alpha/\bar{\mu}$), where α is a hyperparameter drawn from a uniform distribution [1–20]. The mean of the distribution of individual μ_i is thus equal to $\bar{\mu}$, but the distribution allows values varying by several orders of magnitudes to be sampled depending on α values. With $\alpha = 1$, the 95% highest probability density (HPD) interval for this distribution is [1.0×10^{-9}–2.1×10^{-3}], and the corresponding HPD interval for $\alpha = 20$ is [4.2×10^{-4}–1.0×10^{-3}].

Summary Statistics

Summary statistics were calculated using the program Arlequin ver. 3.1 (Excoffier, Laval, and Schneider 2005). Four summary statistics were considered to describe the genetic diversity within Amerindian populations: the number of alleles (K), the heterozygosity (H), the allelic range (R), and the modified Garza–Williamson GW* statistic (Garza and Williamson 2001), defined here for the ith locus in the jth population as $GW^*_{ij} = K_{ij}/(R_{TOT\,i} + 1)$, where K_{ij} is the number of alleles and $R_{TOT\,i}$ is the total allelic range observed for the ith locus over the whole metapopulation (in this case, Asia and the Americas). For each of these statistics, we computed their average over loci and population and their average coefficients of variation (c.v.) over populations (where population-specific c.v. are computed over loci), providing us with eight distinct summary statistics about Amerindian diversity. We also computed

two statistics F_{ST} (Weir and Cockerham 1984) and $(\delta\mu)^2$ (Goldstein et al. 1995) that are informative about the extent of differentiation between Asia and America. These 10 statistics were used to compare the three scenarios of the settlement of the Americas because they summarize the genetic diversity of Amerindian populations as well as their relationship with Asian populations and thus are most sensitive to differences between the scenarios we envisioned. For completeness, another set of analyses considered six additional statistics to account for Asian diversity (average number of alleles in Asia K_{Asia}, average total number of alleles over Asia and the Americas K_{TOT}, average heterozygosity in Asia H_{Asia}, average GW* in Asia, average allelic range R in Asia R_{Asia}, and R_{TOT} defined above), leading to a total of 16 summary statistics. However, because we modeled a very simple scenario for Asia, Asian-based summary statistics are likely to dominate the overall Euclidean distances computed between observed and simulated data and prevent a proper discrimination between settlement scenarios for the Americas. We focus therefore, on the analyses based on the first set of 10 summary statistics, which is the most informative regarding scenarios of the settlement of the Americas.

Model Choice Procedures

We performed a total of one million simulations of the genetic diversity at 401 microsatellite loci in our samples for each of the three models and data set combination. Two model choice procedures were used. The first one is based on the simple rejection procedure proposed by Pritchard et al. (1999) for which model posterior probabilities were computed as follows: we retained, for each model, the 5,000 simulations with smallest associated Euclidean distances between the simulated and the observed summary statistics; the selected 15,000 simulations were ordered by ascending Euclidean distances recomputed on summary statistics standardized with common mean and standard deviation; and the posterior probability of a given model was then simply computed as the proportion of simulations done under a given model included among n-smallest distances, where n is an arbitrary number (usually ranging between 100 and 1,000) (Estoup et al. 2004). In the second procedure proposed by Beaumont (2008), we computed the posterior probability for each model, again using only the 5,000 simulations closest to the observed data, following a weighted multinomial logistic regression procedure (Beaumont 2008). This latter procedure is an extension of conventional logistic regression to more than two categories and has previously been applied in a human evolutionary context (Fagundes et al. 2007).

Parameter Estimation

Parameters were estimated under a conventional ABC framework (Beaumont et al. 2002) after transformation of the original summary statistics using a partial least-squares (PLS) approach (e.g., Tenehaus et al. 1995; Boulesteix and Strimmer 2007), as detailed in Wegmann et al. (2009). The PLS approach is similar to a principal component analysis, but here principal components in the summary statistics space are also chosen such as to maximally explain the variability of the parameters. This procedure is advantageous because it allows us to 1) reduce the dimensionality of the problem and 2) get a set of uncorrelated transformed statistics. We used the R package "pls" to compute PLS components and to define the optimal set of components (Mevik and Wehrens 2007). In short, the optimal number of PLS component is chosen such as the addition of more components would not reduce the root mean square error (RMSE) of the parameters predicted from these components. The RMSE of each predicted parameter is computed using a leave-one-out procedure for an increasing number of components. A visual inspection then judges when additional components do not improve on the quality of the predictions (Mevik and Wehrens 2007). In our case, the eight largest PLS components were selected by this leave-one-out procedure. For a given model, we then retained the 5,000 simulations with smallest associated Euclidean distances between PLS-transformed observed and simulated statistics. Note that the use of PLS components has been shown to lead to improved posterior distributions in an ABC framework, most noticeably improving their coverage properties (see Wegmann et al. 2009).

The posterior distributions of the parameters were then conventionally obtained by performing a locally weighted multivariate regression (Beaumont et al. 2002). Parameters (x) were transformed as $z = \log[\tan(x)^{-1}]$ before regression to keep posteriors within prior ranges (Hamilton et al. 2005). We chose to report the mode of the posterior distribution as a point estimator. The quality of the estimated parameters (i.e., the potential for a parameter to be correctly estimated) was assessed using the coefficient of determination R^2 (i.e., the proportion of parameter variance explained by the summary statistics) computed across all simulations (Lefebvre 1983). Previous studies have shown that parameters for which R^2 is smaller than 5–10% are usually difficult to estimate (Neuenschwander et al. 2008).

Results and Discussion

Rejection of a Single Early Wave Model

Table 1 shows the relative probabilities of the three demographic models considered (detailed results are given in supplementary figs. S4 and S5, Supplementary Material online). Both model choice procedures show very strong support for the RGF model, with posterior probabilities larger than 0.97 for the six possible combinations of two Asian and three American data sets. The SW model shows the poorest fit to the data, with posterior probabilities much lower than the 2W model. Similar results were obtained with the North/Central and South American data subsets, suggesting that these regions have been similarly impacted by the colonization from Asia. A closer examination of the distribution of the summary statistics generated under the three models (supplementary fig. S2, Supplementary Material online) shows that the RGF model provides a much better fit than the other two models especially for three

Table 1. Relative Probabilities of the Three Models of the Colonization of the Americas for Each of the Six Combinations of Asian and American Population Samples.

	Asian10S data set			Asian2S data set		
Model	All Americas	North America	South America	All Americas	North America	South America
SW	0 (1.48×10^{-14})	0.01 (1.97×10^{-09})	0 (4.69×10^{-13})	0 (9.19×10^{-15})	0.01 (5.43×10^{-10})	0 (4.93×10^{-13})
2W	0 (6.55×10^{-08})	0 (2.16×10^{-05})	0 (1.86×10^{-07})	0 (6.98×10^{-08})	0.015 (5.97×10^{-05})	0 (3.77×10^{-07})
RGF	1 (0.999)	0.99 (0.999)	1 (0.999)	1 (0.999)	0.975 (0.999)	1 (0.999)

NOTE.—Relative probabilities obtained with the approach of Beaumont (2008) are given in parentheses.

statistics calculated in the native American populations: the observed number of alleles *K*, the heterozygosity *H*, and the allele size range *R*. It seems therefore that gene flow is primarily required to fit the observed levels of native American diversity, which are higher than expected under the other two models, given the amount of divergence between Asia and America, which is itself well accounted for by all models (see distributions for F_{ST} and $(\delta\mu)^2$ in supplementary fig. S2, Supplementary Material online). Note that the better fit for the RGF model is not due to a bad choice of priors for the other two models because each model is able to reproduce each observed summary statistics (see supplementary fig. S3, Supplementary Material online).

The results of the model choice procedure based on the set of 16 summary statistics (including the six additional statistics computed on Asian samples) also globally favor the RGF model, especially when the Asia metapopulation is represented by the two Siberian samples (Asian2S data set). In that case, posterior probabilities for the RGF model are 0.76, 0.54, and 0.59, when American diversity is, respectively, computed on all the Amerindian samples, on North America only, and on South America only (see supplementary fig. S7, Supplementary Material online). When the larger Asian10S data set is used (supplementary fig. S6, Supplementary Material online), the RGF model is still the dominant model, but only marginally so, with posterior probabilities of 0.60, 0.38, and 0.40 for the All Americas, North America, and South America data sets, respectively. For the Asian2S data sets, the second best supported model is the 2W model when Amerindian summary statistics are computed on North American samples, whereas the single early wave model is the second best model when only South American samples are used. These less clear-cut results confirm that the addition of summary statistics computed on Asian samples make it more difficult to distinguish between models of the settlement of the Americas, even though the use of the two Siberian populations as representative of the Asian metapopulation from which the Americas would have been settled also provides a very strong support for the RGF model.

Below, we discuss more fully some caveats regarding the analysis of the three demographic models evaluated here under an ABC framework.

Comparison of Models with Different Numbers of Parameters

The model posterior probabilities obtained are simple extensions of Bayes's factors when more than two models are compared. Bayes's factor adequately compares models with unequal number of parameters because they are ratios of marginal likelihood (integrals of likelihoods weighted by prior distributions) that remove dependencies on the number of parameters used by each model by integrating over all parameter values of these models. They also include an automatic penalty for models depending on a higher number of parameters (see, e.g., MacKay 2003, p. 349, on Occam's razor). The penalty for models with additional parameters is simple to understand: for a fixed number of simulations, the parameter space of models with more parameters will be less well explored than for simpler models. If the additional parameters are not informative, there will be fewer simulations close to the observations, and the model with extra parameters will have a lower posterior probability. Therefore, the higher support for the model with RGF is not due to its additional parameters as compared with the other models tested.

Importance of Priors

Because Bayesian model choice procedures compare posterior distributions, priors are implicitly (in ABC) or explicitly (in likelihood-based methods) incorporated in the computation of Bayes's factors or model posterior probabilities. When data are not very informative, priors may dominate the outcome. One should, therefore, check that the final results do not overly depend on the chosen priors, for instance, if models are based on very different sets of priors. In our case, we have used identical priors for all parameters shared by the different models. Furthermore, posterior distributions are informative for most parameters (see, for instance, priors and posteriors in supplementary fig. S8, Supplementary Material online, for the RGF model). Posterior probabilities could, however, be sensitive to parameters specific to the 2W and RGF models, which can be considered as simple extensions of the SW model.

Under the 2W model, the duration of the bottleneck for the second wave is fixed to 10 generations, but the bottleneck size Nb_{AM2} can vary between 2 and 1,000 genes, allowing for a very strong bottleneck (when $Nb_{AM2} = 2$) or no bottleneck (when $Nb_{AM2} = 1{,}000$). The age of this second wave was allowed to range between 200 and 400 generations (5,000–10,000 years) based on archaeological and anthropological evidence suggesting the apparition of derived craniometric traits around 7,000–8,000 years ago in both Asia and the Americas (González-José et al. 2008 and references therein). This range is also consistent with the estimated date for the Na Dene language around

Table 2. Point Estimates (mode) of Demographic and Mutation Parameters Obtained under the RGF Model.

Parameters	Priors	All Americas	North America	South America	R^{2a}
P_{AS}	[50–1,000]	900 [341–1,000]	928 [362–1,000]	907 [316–1,000]	0.25
P_{AM}	[50–1,000]	904 [158–1,000]	789 [179–1,000]	751 [135–988]	0.16
N_{AS}	[10,000–100,000]	87,640 [38,290–99,945]	92,928 [40,858–99,955]	92,033 [32,868–99,954]	0.17
N_{AM}	[10,000–100,000]	89,972 [19,352–99,919]	89,990 [17,811–99,491]	24,525 [13,221–96,081]	0.14
Nb_{AM}	[2–1,000]	173 [83–280]	250 [98–528]	174 [93–301]	0.63
Nb_{AS}	[2–1,000]	225 [40–856]	184 [2–770]	194 [3–795]	0.52
$N_{A\text{-}AS}$	[1,000–50,000]	9,936 [1,035–45,610]	6,939 [1,031–45,351]	7,436 [1,040–45,544]	0.28
T_{POP}	[4–20]	5 [4–14]	5 [4–14]	5 [4–16]	0.04
T_{W1}	[400–1,200]	528 [400–956]	610 [400–1,043]	562 [400–997]	0.08
T_{GF}	[200–400]	390 [298–400]	388 [267–400]	384 [259–400]	0.02
T_{AS}	[1,600–8,000]	3407 [1,983–7,475]	3344 [1,767–6,725]	4445 [2,337–7,610]	0.39
$M_{AS\to AM}$ ($\times 10^4$)	[0.1–10]	6.59 [3.37–9.99]	5.99 [1.46–9.99]	3.91 [1.12–9.99]	0.34
$M_{AM\to AS}$ ($\times 10^4$)	[0.1–10]	9.70 [8.26–9.99]	9.69 [7.47–9.99$^-$]	9.60 [7.41–9.99]	0.13
Gamma	[1–20]	9.58 [4.77–16.31]	10.16 [5.07–16.54]	8.35 [4.36–14.88]	0.66
$2Nm_{As\to Am}$[b]	—	20.9	19.8	14.0	NA
$2Nm_{Am\to As}$[b]	—	9.4	11.4	3.4	NA

NOTE.—These estimates were obtained considering the two Siberian populations as representative of the source Asian populations (Asian2S data set). The figures reported in the table represent the mode of the marginal posterior density distribution, followed in brackets by the 95% HPD interval. For the prior distributions, we report their ranges and note that they were all set as uniform over the specified range.

[a] R^2 is the average coefficient of determination of the parameter by the summary statistics, computed over the three sets of Amerindian populations.

[b] Mean 2*Nm* estimates were obtained by averaging the population size during the gene flow period. We used the estimated values of parameters for the population growth model, which is why we do not report priors and HPD intervals for these estimates.

P_{AS}: current average size of Asian populations; P_{AM}: current average size of Amerindian populations; N_{AS}: Asian size prior to subdivision; N_{AM}: Amerindian size prior to subdivision; Nb_{AM}: size of the bottleneck when entering America; Nb_{AS}: size of the bottleneck when entering Asia; NA_{AS}: ancestral population size, T_{POP}; T_{W1}: time of initial migration wave; T_{W2}: onset of gene flow; T_{AS}: time of speciation/out of Africa; $M_{AS\to AM}$: migration rate per generation from Asia to America; $M_{AM\to AS}$: migration rate per generation from America to Asia; Gamma: shape parameter of the Gamma distribution of locus-specific mutation rates; $2Nm_{As\to Am}$: mean number of diploid migrants (toward America) during period of gene flow; $2Nm_{Am\to As}$: mean number of diploid migrants (toward Asia) during period of gene flow. Population sizes are reported in haploid number of genes.

9,000 years bp (Greenberg et al. 1986). Because the impact of this second wave is difficult to assess, we used a very wide prior for the proportion of current lineages stemming from this second wave, which was allowed to vary between zero (where the 2W model tends to the EW model) and one (corresponding to a single recent wave). If the actual contribution of the second wave was very small, this wide prior would penalize this model, but it is difficult to justify imposing a low upper limit (e.g., like 5%) for this parameter.

For the RGF model, the prior for the onset of gene flow was made similar to the occurrence of the second wave in the 2W model, for the reasons mentioned above. A possible restriction of this model is that it disallows a very recent onset of gene flow (in the last 200 generations). However, we would expect that patterns of diversity would be similar in the case of limited gene flow over a long time period and when strong gene flow occurs recently, so that this restriction should not be too penalizing for this model. Finally, gene flow was allowed to be asymmetric and ranging from very low to relatively high migration rates.

A parameter common to all models that imposes a strong constraint on genetic diversity is the divergence time from Asia T_{W1}, which needs to be at least 400 generations or approximately 10,000 years in all three models. Supplementary figure S2, Supplementary Material online, shows that F_{ST} is well estimated for all models but that the EW and 2W models result in a relatively low number of alleles *K*, low allelic range size *R*, and high heterozygosity *H*. It seems that gene flow under the RGF model can bring additional alleles from Asia and therefore increase both *K* and *R* while not affecting *H* too severely explaining the better fit of this model. We note that recent admixture in the sampled Amerindian populations could also lead to increased *K* and *R* in Amerindian samples. However, with the exception of the Chipewyan, Cree, and Ojibwa, the Amerindian samples examined here show very little evidence of non-native admixture (see fig. 5 in Wang et al. 2007, an observation supported by more recent mitochondrial DNA [mtDNA] and Y-chromosome analyses; Yang NN, personal communication). Moreover, we would expect that if the RGF model was favored due to recent non-native admixture, the estimation of the onset of the gene flow T_{GF} would point toward recent times, which is the opposite of what we observe (table 2 and supplementary fig. S8, Supplementary Material online). This suggests that a prolonged period of gene flow is necessary to account for the observed genetic diversity.

For computational reasons, it was not possible to fully explore the effect of additional sets of priors on model choice. However, we believe that the distinction between models is due to their parameterization rather than to our choice of priors for models with additional parameters. It is a general problem that one needs to define a finite (and plausible) set of models to explore, realizing that the priors are part of the model definition. We are aware that other models and alternative sets of sensible priors and could be examined, but we believe that the three models capture the major current controversies regarding the settlement of the American continent.

Demographic Parameters Estimated under the RGF Model

Demographic parameter estimates obtained with the Siberian (Asian2S) data set under the RGF model are shown in table 2 (comparable results were obtained when considering the 10 Asian populations included in the Asian10S data set; supplementary table S2, Supplementary Material online). The time estimate for initial entry into the American continent T_{W1} (528, 556, and 610 generations for Asian2S/AllAmericas, Asian2S/SouthAmerica, and Asian2S/NorthAmerica data sets, respectively) corresponds to a range of about 13,200–15,250 years bp assuming a generation time of 25 years. These dates are consistent with the oldest archaeological sites in the Americas (Dillehay 1997; Goebel et al. 2008; Rothhammer and Dillehay 2009) and point toward a late Pleistocene colonization, considerably later than the last glacial maximum (~22,000 years ago). Our date estimate ties within the range of previous genetic estimates being older than other two dates based on nuclear loci (~7,000 years ago, Hey 2005; ~10,000 years ago, Fagundes et al. 2007) and in better agreement with entry dates obtained recently with mtDNA (Fagundes et al. 2008; Ho and Endicott 2008) and Y chromosome (Bortolini et al. 2003; Zegura et al. 2004) that are closer to 15,000 years. Our modal values have a relatively wide 95% credible interval (400–956 generations for the Asian2S/AllAmericas data set), reflecting the difficulty in estimating this parameter from summary statistics that explain only 8% of its overall variability. However, the examination of the posterior distribution (supplementary fig. S8, Supplementary Material online) suggests that there is information about this parameter and that it points to a relatively recent age for the colonization. Interestingly, the fact that the estimated colonization time is similar for the two subsets of Amerindian populations is consistent with the classic view of a colonization of South America soon after initial entry into North America (Martin 1973; Cavalli-Sforza et al. 1994; Crawford 1998; Fagundes et al. 2008).

Whereas forward migration rates from America to Asia ($M_{AM\rightarrow AS}$) are found larger ($9.6–9.7 \times 10^{-4}$) than those from Asia to America ($M_{AS\rightarrow AM}$) ($3.9–6.6 \times 10^{-4}$) (table 2), the average number of immigrants received from Asia ($2Nm_{As\rightarrow Am}$) during the period of gene flow is larger than the number of emigrants send to Asia ($2Nm_{Am\rightarrow As}$) (see table 2) because the Asian population size is larger during this period. In keeping with the notion that gene flow occurs preferentially across the Bering Strait, the estimated effective number of diploid migrants from Asia to North America is found higher than that to South America (~20 vs. ~14, respectively). We also find evidence for gene flow in the opposite direction, with a smaller average of approximately nine diploid migrants from America to Asia per generation (11 migrants for North America and 3 for South America). The higher number of migrants exchanged between Asia and North America than between Asia and South America is also consistent with the observation of mtDNA (e.g., haplogroup X2a) and Y-chromosome variants (e.g., haplogroups M3 and RPSY) restricted to North America and extreme Northeastern Siberia (e.g., Brown et al. 1998; Lell et al. 2002). The onset of this bidirectional gene flow is estimated to be relatively ancient, about 390 generations (9,750 years for generation time of 25 years). Note however that summary statistics have little explanatory power for this parameter ($R^2 = 2.2\%$), and therefore this estimate should be taken with caution. Previous Bayesian analysis of Asian and Amerindian population diversity also found evidence of bidirectional gene flow between the two continents (Hey 2005), but a precise comparison with our results is difficult because the previous study assumed no subdivision within Asia and America and used a mixture of haploid and diploid markers genotyped in different individuals and different populations.

Consistent with our results, a recent reanalysis of genetic and morphologic variation across Americas suggests that gene flow in the Arctic could have influenced the patterns observed for both crania and genes (González-José et al. 2008). Finally, we estimate that the Amerindian settlement occurred after a bottleneck through ~173 gene lineages (95% HPD 83–280), corresponding to about 87 effective diploid founders. A similar estimate was obtained by Hey (2005) and suggests a relatively important population contraction at the origin of Amerindians, in agreement with their considerably reduced diversity relative to other continental populations (Wang et al. 2007).

In conclusion, our analyses strongly reject the settlement of the Americas by a single, discrete, colonization wave from Asia and underline the importance of gene flow between Asia and America during the evolution of native American populations. We estimate that the initial settlement very likely occurred after the last glacial maximum, perhaps around the time of the deglaciation of the Pacific coastal corridor (Dyke 2004), in keeping with the recent results based on autosomal and Y-chromosome diversity (Bortolini et al. 2003; Seielstad et al. 2003; Fagundes et al. 2007) and with some analyses of mtDNA (Tamm et al. 2007; Fagundes et al. 2008; Ho and Endicott 2008). As a next step, more detailed, spatially explicit simulations (e.g., Currat et al. 2004; Ray et al. 2005) could be envisaged in order to better characterize the ancestral American gene pool and define the nature of gene flow with Northeast Asia.

Supplementary Material

Supplementary tables S1–S2 and Supplementary figures S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Damian Labuda for useful comments on an earlier version of the manuscript and Matthieu Foll for help with computational issues. This work was partly supported by Swiss National Science Foundation grants 3100A0-112072 and 3100A0-126074 to L.E.

References

Adovasio JM, Pedler D. 2004. Pre-Clovis sites and their implications for human occupation before the last glacial maximum. In: Madsen DB, editor. Entering America: northeast Asia and Beringia before the last glacial maximum. Salt Lake City (UT): University of Utah Press. p. 139–158.

Beaumont MA. 2008. Joint determination of topology, divergence time, and immigration in population trees. In: Matsumura S, Forster P, Renfrew C, editors. Simulations, genetics and human prehistory. Cambridge (UK): McDonald Institute Monographs. p. 135–154.

Beaumont MA, Rannala B. 2004. The Bayesian revolution in genetics. *Nat Rev Genet*. 5:251–261.

Beaumont MA, Zhang WY, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics*. 162:2025–2035.

Bonatto SL, Salzano FM. 1997. A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. *Proc Natl Acad Sci USA*. 94:1866–1871.

Bortolini MC, Salzano FM, Thomas MG, et al. (21 co-authors). 2003. Y-chromosome evidence for differing ancient demographic histories in the Americas. *Am J Hum Genet*. 73:524–539.

Boulesteix AL, Strimmer K. 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform*. 8:32–44.

Brown MD, Hosseini SH, Torroni A, Bandelt HJ, Allen JC, Schurr TG, Scozzari R, Cruciani F, Wallace DC. 1998. mtDNA haplogroup X: an ancient link between Europe/Western Asia and North America? *Am J Hum Genet*. 63:1852–1861.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton (NJ): Princeton University Press.

Crawford MH. 1998. The origins of native Americans: evidence from anthropological genetics. Cambridge (UK): Cambridge University Press.

Currat M, Ray N, Excoffier L. 2004. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol Ecol Notes*. 4:139–142.

Dillehay TD. 1997. Monte Verde: a late Pleistocene settlement in Chile. The archaeological context and interpretation. Vol. 2. Washington (DC): Smithsonian Institution Press.

Dyke AS. 2004. An outline of North American deglaciation with emphasis on central and northern Canada. In: Ehlers J, Gibbard PL, editors. Quaternary glaciations—extent and chronology, part II: North America. Amsterdam: Elsevier. p. 373–424.

Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet JM. 2004. Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, Bufo marinus. *Evolution*. 58:2021–2036.

Excoffier L, Estoup A, Cornuet JM. 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*. 169:1727–1738.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online*. 1:47–50.

Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci USA*. 104:17614–17619.

Fagundes NJR, Kanitz R, Eckert R, et al. (12 co-authors). 2008. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am J Hum Genet*. 82:583–592.

Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, Pettener D. 2003. Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. *Mol Biol Evol*. 20:1682–1691.

Garza JC, Williamson EG. 2001. Detection of reduction in population size using data from microsatellite loci. *Mol Ecol*. 10:305–318.

Goebel T, Waters MR, O'Rourke DH. 2008. The late Pleistocene dispersal of modern humans in the Americas. *Science*. 319:1497–1502.

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci USA*. 92:6723–6727.

González-José R, Bortolini MC, Santos FR, Bonatto SL. 2008. The peopling of America: craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. *Am J Phys Anthropol*. 137:175–187.

Gonzalez-Jose R, Neves W, Lahr MM, Gonzalez S, Pucciarelli H, Martinez MH, Correal G. 2005. Late Pleistocene/Holocene craniofacial morphology in Mesoamerican Paleoindians: implications for the peopling of the New World. *Am J Phys Anthropol*. 128:772–780.

Greenberg JH, Ii CGT, Zegura SL. 1986. The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence. *Curr Anthropol*. 27:477.

Hamilton G, Stoneking M, Excoffier L. 2005. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc Natl Acad Sci USA*. 102:7476–7480.

Hey J. 2005. On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biol*. 3:e193.

Ho SY, Endicott P. 2008. The crucial role of calibration in molecular date estimates for the peopling of the Americas. *Am J Hum Genet*. 83:142–146; author reply 146–147.

Joyce DJ. 2006. Chronology and new research on the Schaefer mammoth (?*Mammuthus primigenius*) site, Kenosha County, Wisconsin, USA. *Q Int*. 142:44–57.

Karafet TM, Zegura SL, Posukh O, et al. (14 co-authors). 1999. Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am J Hum Genet*. 64:817–831.

Laval G, Excoffier L. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* 20: 2485–2487.

Lefebvre J. 1983. Introduction aux analyses statistiques multidimensionnelles. Masson, Paris. p. 102–104.

Lell JT, Sukernik RI, Starikovskaya YB, Su B, Jin L, Schurr TG, Underhill PA, Wallace DC. 2002. The dual origin and Siberian affinities of native American Y chromosomes. *Am J Hum Genet*. 70:192–206.

MacKay DJC. 2003. Information theory, inference and learning algorithms. Cambridge (UK): Cambridge University Press.

Martin PS. 1973. The discovery of America: the first Americans may have swept the Western Hemisphere and decimated its fauna within 1000 years. *Science* 179:969–974.

Merriwether DA, Ferrell RE. 1996. The four founding lineage hypothesis for the New World: a critical reevaluation. *Mol Phylogenet Evol*. 5:241–246.

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.

Mevik BH, Wehrens R. 2007. The pls package: principal component and partial least squares regression in R. *J Stat Softw*. 18:1–28.

Miller N, Estoup A, Toepfer S, Bourguet D, Lapchin L, Derridj S, Kim KS, Reynaud P, Furlan L, Guillemaud T. 2005. Multiple transatlantic introductions of the western corn rootworm. *Science* 310:992.

Neuenschwander S, Largiader CR, Ray N, Currat M, Vonlanthen P, Excoffier L. 2008. Colonization history of the Swiss Rhine basin

by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. *Mol Ecol*. 17:757–772.

Neves WA, Hubbe M. 2005. Cranial morphology of early Americans from Lagoa Santa, Brazil: implications for the settlement of the New World. *Proc Nat Acad Sci USA*. 102:18309–18314.

Neves WA, Pucciarelli HM. 1991. Morphological affinities of the first Americans: an exploratory analysis based on early South American human remains. *J Hum Evol*. 21:261–273.

Pascual M, Chapuis MP, Mestres F, Balanya J, Huey RB, Gilchrist GW, Serra L, Estoup A. 2007. Introduction history of Drosophila subobscura in the New World: a microsatellite-based survey using ABC methods. *Mol Ecol*. 16:3069–3083.

Powell JF. 2005. The first Americans. Cambridge (UK): Cambridge University Press.

Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol*. 16:1791–1798.

Ray N, Currat M, Berthier P, Excoffier L. 2005. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res*. 15:1161–1167.

Rothhammer F, Dillehay TD. 2009. The late Pleistocene colonization of South America: an interdisciplinary perspective. *Ann Hum Genet*. 73:540–549.

Santos FR, Pandya A, Tyler-Smith C, Pena SD, Schanfield M, Leonard WR, Osipova L, Crawford MH, Mitchell RJ. 1999. The central Siberian origin for native American Y chromosomes. *Am J Hum Genet*. 64:619–628.

Seielstad M, Yuldasheva N, Singh N, Underhill P, Oefner P, Shen P, Wells RS. 2003. A novel Y-chromosome variant puts an upper limit on the timing of first entry into the Americas. *Am J Hum Genet*. 73:700–705.

Tamm E, Kivisild T, Reidla M, et al. (21 co-authors). 2007. Beringian standstill and spread of native American founders. *PLoS One*. 2:e829.

Tavare S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. *Genetics*. 145:505–518.

Tenehaus M, Gauchi J-P, Ménardo C. 1995. Régression PLS et applications. *Rev Stat Appl*. 43:7–64.

Wakeley J. 2001. The coalescent in an island model of population subdivision with variation among demes. *Theor Popul Biol*. 59:133–144.

Wakeley J, Aliacar N. 2001. Gene genealogies in a metapopulation. *Genetics* 159:893–905.

Wang S, Jakobsson M, Lewis JCM, et al. (26 co-authors). 2007. Genetic variation and population structure in native Americans. *PLoS Genet*. 3:e185.

Waters MR, Stafford TW Jr. 2007. Redefining the age of Clovis: implications for the peopling of the Americas. *Science* 315:1122–1126.

Wegmann D, Leuenberger C, Excoffier L. 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* 182:1207–1218.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.

Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF. 2004. High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of native American Y chromosomes into the Americas. *Mol Biol Evol*. 21:164–175.

Zhivotovsky LA, Rosenberg NA, Feldman MW. 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am J Hum Genet*. 72:1171–1186.