OPEN ACCESS Freely available online

PLoS one

Predictions of Native American Population Structure Using Linguistic Covariates in a Hidden Regression Framework

Flora Jay, Olivier François*, Michael G. B. Blum

Université Joseph Fourier, Grenoble, Centre National de la Recherche Scientifique, Laboratoire des Techniques de l'Ingénierie Médicale et de la Complexité, Equipe Biologie Computationnelle et Mathématique, Faculté de Médecine, La Tronche, France

Abstract

Background: The mainland of the Americas is home to a remarkable diversity of languages, and the relationships between genes and languages have attracted considerable attention in the past. Here we investigate to which extent geography and languages can predict the genetic structure of Native American populations.

Methodology/Principal Findings: Our approach is based on a Bayesian latent cluster regression model in which cluster membership is explained by geographic and linguistic covariates. After correcting for geographic effects, we find that the inclusion of linguistic information improves the prediction of individual membership to genetic clusters. We further compare the predictive power of Greenberg's and *The Ethnologue* classifications of Amerindian languages. We report that *The Ethnologue* classification provides a better genetic proxy than Greenberg's classification at the stock and at the group levels. Although high predictive values can be achieved from *The Ethnologue* classification, we nevertheless emphasize that Choco, Chibchan and Tupi linguistic families do not exhibit a univocal correspondence with genetic clusters.

Conclusions/Significance: The Bayesian latent class regression model described here is efficient at predicting population genetic structure using geographic and linguistic information in Native American populations.

Citation: Jay F, François O, Blum MGB (2011) Predictions of Native American Population Structure Using Linguistic Covariates in a Hidden Regression Framework. PLoS ONE 6(1): e16227. doi:10.1371/journal.pone.0016227

Editor: Thomas Mailund, Aarhus University, Denmark

Received October 27, 2010; **Accepted** December 17, 2010; **Published** January 31, 2011

Funding: OF received external fundings from the Institute of Complex Systems IXXI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: olivier.francois@imag.fr

Introduction

Comparing genetic and linguistic data provides information about various aspects of American prehistory, the process by which the Americas were originally colonized [1] or migration across linguistic barriers [2]. In addition to anthropological applications, evaluating the relationships between genes and languages has potential biomedical applications since language could be used as a *proxy* for genetic ancestry in various epidemiological contexts [3,4].

Previous analyses comparing genetic to linguistic differentiation in the Americas yielded equivocal results. Cavalli-Sforza *et al.* [5] reported that, prior to the publication of their book, three of seven studies supported congruence between genes and languages [6–12]. At that time, Ward *et al.* [13] found that rates of linguistic diversification are faster that rates of genetic differentiation in mtDNA, and concluded that there is little congruence between linguistic and genetic relationships in the Americas. In more recent studies also using mtDNA, the hypothesis that language classifications reflect the genetic structure of Native American populations was also rejected [2,14]. Lastly, an analysis of autosomal microsatellite markers in 28 Native American populations from the Human Genome Diversity Panel (HGDP) provided a qualitative correspondence between linguistic and genetic groupings [15]. However, tests of correlation were not significant for these data.

To investigate the relationships between genes and languages, the previous studies made use of tree-based or distance-based methods. Hunley and Long [2] and Hunley *et al.* [14] applied a test of treeness developed by Cavalli-Sforza and Piazza [16] to decide if a matrix of genetic distances is compatible with a language tree. These authors dealt with various hierarchical classifications of American languages, and they found that none of them were consistent with the mitochondrial genetic distances. Adopting another approach, Cavalli-Sforza *et al.* [17] found a high degree of association between linguistic and genetic trees using a consistency index. Alternatively, the association of genes and languages can be assessed by Mantel tests [18]. Mantel tests are used to reject the absence of correlation between a matrix of genetic distances and a matrix of linguistic distances, and do not require reconstructing population trees. Since a spurious association between genetic and linguistic distances may be detected when geography is not accounted for, more elaborate procedures called partial Mantel tests can be applied in order to control for geography [19]. Partial Mantel tests were applied to the HGDP and did not provide strong evidence of association in Native American populations [15].

By definition, the results obtained from tree-based and distance-based methods are influenced by specific choices of tree reconstruction methods or particular genetic and linguistic distances. The validity of population trees depends on the reliability of their reconstruction method and on the hypothesis that genetic differentiation results from population fission. Whereas trees are well-suited for describing evolutionary relationships of non-recombining sequences like mtDNA, they may be sensitive to distortion due to gene flow between populations when nuclear data are analyzed [20]. In addition, we still lack an evolutionary tree for languages as linguists have not yet reached a clear consensus on their classification [21], and even questioned the validity of branching trees as an adequate representation of linguistic patterns of divergence [22]. Finally, there are several pairwise measures of population differentiation or of linguistic divergence, and the choice of a specific measure can have a significant impact on Mantel tests [23]. Linguistic distances can, for instance, be based on a hierarchical linguistic classification [24], or they can be directly derived from structural linguistic features such as aspects of sound systems and grammar [25,26].

In this study, we introduce a novel method for investigating the relationships between genes and languages that avoids genetic and linguistic distances as well as tree reconstruction methods. We consider Bayesian *latent class regression* models [27] where we regress the unobserved genetic structure on linguistic and geographic variables. The principle of the method is to group individuals into genetic clusters at the same time as their latent cluster labels are regressed. To evaluate the predictive capacity of different sets of linguistic and geographic covariates, we also propose procedures of variable selection. Using this approach, the following questions are addressed. To what extent can geographic or linguistic origin explain individual membership to genetic clusters? Do languages contribute to a better prediction of cluster membership than geography alone? Among the classifications of Native American languages that have been proposed by linguists [28,29], which one is the best predictor of population genetic structure? Although some of these questions have received considerable attention in the context of evolutionary trees or evolutionary distance comparisons [2,23,30], examining their answers from a latent class individual-based model is new and potentially highly informative.

Methods

Several Bayesian model-based approaches have been proposed to assign individuals to genetic clusters [31–33]. To assess the effects of geographic and linguistic covariates on the assignment of individuals to genetic clusters, we considered a Bayesian latent class regression model [27,34,35]. This new model incorporates a hidden regression model within the framework proposed by Pritchard *et al.* [31] and implemented in the computer program structure.

Bayesian model

Consider a genotypic data set, X, for a sample of n diploid individuals genotyped at L loci, and assume that there are K clusters, each of which is characterized by a set of allele frequencies at each locus. Let $Z=(Z_1,\ldots,Z_n)$ be the vector of cluster labels of each individual in the sample, and let P be the set of allele frequencies. In addition, assume that a set of covariates is measured for each individual, and stored in a design matrix, $\tilde{X}$. The covariates represent the geographic and linguistic information that is available to build predictors of the population genetic structure that is encoded in vector Z. Regarding geography, predictors can be defined as linear or quadratic trend surfaces as proposed by Durand *et al.* [36]. Linear trend surfaces include two covariates, latitude and longitude, while quadratic surfaces also include squared and cross-product terms. Languages are coded as factors defined as binary dummy variables in the design matrix [37]. The factor levels will be dependent on the choice of the linguistic classifications considered further in this study. Remark that in regression models using factors, a linear constraint (or contrast) must be defined for identifiability reasons. In our study, we assumed that the sum of effects is null.

For algorithmic reasons, the latent regression model was implemented through a hidden *multinomial probit model* [38]. In the multinomial probit model, there are $K-1$ regression equations

$$\begin{aligned} &W_{i,k}=\tilde{X}_i\beta_k+\epsilon_{i,k}, i=1,\ldots,n, k=1,\ldots,K-1,\\ &\epsilon_i=(\epsilon_{i,1},\ldots,\epsilon_{i,K-1})\sim\mathcal{N}(0,\mathrm{Id}), \end{aligned} \tag{1}$$

each corresponding to a genetic cluster. The $(W_{i,k})$ are "augmented" continuous variables defined for each individual and each cluster, β_k is a column vector of regression coefficients, and Id denotes the identity matrix. For each individual i, a cluster label Z_i can be obtained from the augmented variables as follows

$$Z_i=\begin{cases} K & \text{if } \max_\ell W_{i,\ell}<0\\ k & \text{if } \max_\ell W_{i,\ell}>0 \text{ and } \max_\ell W_{i,\ell}=W_{i,k}. \end{cases} \tag{2}$$

In the multinomial probit model the role of the clusters is not symmetric. The estimates of the regression coefficients are defined with respect to the K^{th} cluster, called the *reference cluster*.

Given the above latent class model framework, we used a Markov Chain Monte Carlo (MCMC) algorithm based on Gibbs sampling to compute the joint posterior distribution on individual cluster labels, regression coefficients and allele frequencies

$$\Pr(Z,\beta,P|X)\propto\Pr(X|Z,P)\Pr(Z|\beta)\Pr(\beta)\Pr(P).$$

In this equation, the likelihood $\Pr(X|Z,P)$ and the prior distribution on allele frequencies $\Pr(P)$ are computed in the same way as in the model without admixture of the program structure (equations (2) and (4) in [31]). The distribution $\Pr(\beta)$ is a noninformative prior distribution (see Appendix A), and $\Pr(Z|\beta)$ corresponds to the distribution of cluster labels obtained from the multinomial probit model. The algorithm was implemented in the software POPS, and is described in more details in Appendix A.

For each subset of covariates, we additionally computed a matrix of posterior predictive membership probabilities using a Monte Carlo method. To perform the computations, we simulated cluster labels from the generative model described in equation (1) and (2) where the regression coefficients are sampled from their posterior distribution. To display predicted and inferred membership probabilities graphically, we used barplot representations. In these graphics, each individual is represented by K aligned colored segments, and the segment lengths are proportional to their estimated or predicted membership probabilities.

Variable selection

To investigate whether a particular subset of covariates is a suitable proxy for genetic assignment, we used two distinct measures. Both measures are based on the posterior of regression coefficients and cluster labels. The first measure is a Pearson

correlation coefficient, ρ. For a given subset of covariates, the ability of the model to predict genetic structure was evaluated by computing the correlation between the matrix of predicted membership coefficients and the matrix of estimated membership coefficients. The second measure is based on cross-validation, a technique used in the field of machine learning [39,40] and for latent class models [41]. In our analyses, a 2-fold cross-validation was implemented. More specifically, we divided the genotypic data set, X, into two non-overlapping data sets containing complementary subsets of loci. We considered one of these data sets as the training set, X^{training}, and the other one as the validation set, $X^{\text{validation}}$. The rationale of the cross-validation approach is that the demographic processes that shaped population genetic structure have affected all loci across the genome. Thus the training and validation sets are exchangeable, as they provide the same amount of information about population structure. We performed 500 runs of the Gibbs sampling algorithm using the training set, and retained the 50 runs having reached the highest likelihood values. For each of the retained runs, a predictive score was computed by averaging the log-probability of the validation set over the posterior distribution given the training set

$$\text{Predictive Score} = E\left[\log(\Pr(X^{\text{validation}}|Z))|X^{\text{training}}\right].$$

The computation of predictive scores is detailed in Appendix B. Another series of 50 scores was computed after exchanging the role of the validation and training sets, and a cross-validation score was obtained by averaging the resulting $2\times 50 = 100$ predictive scores.

Simulated data

We ran a first series of simulations using the generating model of the program POPS. Assuming three clusters, cluster labels of 300 individuals were simulated using the following regression equations

$$W_{i,1} = 1 + 3\tilde{X}_i^{\text{Lat}} + \epsilon_{i,1} \quad (3)$$

$$W_{i,2} = -4 + 12\tilde{X}_i^{\text{Lat}} + \epsilon_{i,2} \quad (4)$$

where $\epsilon_{i,k}$ is a standard Gaussian noise. The interpretation of the above linear trend model is that latitude is the only variable that influences individual cluster labels. Biallelic genotypes were simulated at $L = 20$, 40, 100 loci. Allele frequencies were dependent on the population of origin, and were equal to 30% and 70%, 70%−30% and 50%−50% in each population respectively. We implemented four hidden regression models: one model without covariates, one with latitude, one with longitude and one with both covariates.

In the second series of simulations, we extended the model by including a factor with five levels representing five languages. The hidden regression equations were defined as

$$W_{i,1} = 1 - 0.2\tilde{X}_i^{\text{Lat}} + 0.5L_i^1 + 1L_i^2 - 1.5L_i^3 - 2L_i^4 + 2L_i^5 + \epsilon_{i,1} \quad (5)$$

$$W_{i,2} = -3 + 9\tilde{X}_i^{\text{Lat}} + 6L_i^1 - 1.5L_i^2 + 3L_i^3 - 1.5L_i^4 - 6L_i^5 + \epsilon_{i,2} \quad (6)$$

where L_i^k is equal to 1 if individual i speaks the language k and is 0 otherwise. When running POPS to predict population genetic structure, we considered three linguistic classifications. The first classification contained five languages corresponding to the indicator variables used in the simulation. The second classification contained seven languages obtained after splitting the second and the third languages of the first classification into two sublanguages. The last classification contained three languages because we merged two pairs of unrelated languages from the first classification.

In the third series of experiments, we studied two previously published data sets simulated from a five-island model [42]. The simulated data represented one population structured into five subpopulations differentiated at F_{ST} levels equal to 0.03 and 0.04. Five hundred individuals (100 per subpopulation) were simulated using allele frequency distributions across 10 codominant unlinked loci. Spatial coordinates were simulated using Gaussian distributions. The subpopulations were adjacent to each other and arranged on a ring. We ran POPS using the spatial coordinates of each individual as covariates. In addition, we introduced a spurious noisy covariate independent on the subpopulation of origin. We considered the models defined by all the possible inclusions of those three covariates ($2^3 = 8$ models). These data enabled us to compare the performances of POPS to other programs using spatial covariates [42–44].

Native American data

We applied POPS to 512 Native American individuals from the Human Genome Diversity Panel (HGDP) data set [15]. Individuals from 28 populations were genotyped at 678 microsatellite loci. Fourteen Siberian individuals from the Tundra Nentsi population were also included in the study. In the regression models we considered three linguistic classifications. The first and second linguistic classifications corresponded to Greenberg's classification at the stock level and at the group level [28,45]. The third linguistic classification was given by the website *The Ethnologue* (www.ethnologue.com) [29,46]. The three linguistic classifications were encoded with factors having 8, 14 and 16 levels respectively (see Table S1). To account for geography, all models included quadratic trend surfaces. The combinations of geographic and linguistic variables resulted in the following four latent cluster regression models. Model A included geographic information only. Models B-D included geographic and linguistic information: Model B used Greenberg's classification at the stock level (8 levels), Model C used Greenberg's classification at the group level (14 levels), and Model D used *The Ethnologue* classification at the family level (16 levels).

MCMC parameters

For the simulated data, the runs of POPS used 2,000 sweeps with an initial burn-in period of 1,000 sweeps. For the human data, the runs used 5,000 sweeps with an initial burn-in period of 2,500 sweeps. These values ensured that the likelihoods stabilized around their stationary values. For the HGDP data and for each model, we ran a total of 500 MCMC runs. We retained the 50 runs with the largest likelihood values, and we averaged the resulting estimated and predicted membership coefficients using the computer program clumpp [47].

The number of clusters was set to $K = 9$ [15]. Among these nine clusters, there were eight Native American clusters plus the reference cluster. For Native American population samples, we chose the Siberian population (Tundra-Nentsi) to represent the reference group. Individuals in the reference cluster were not allowed to switch to other clusters during the MCMC runs.

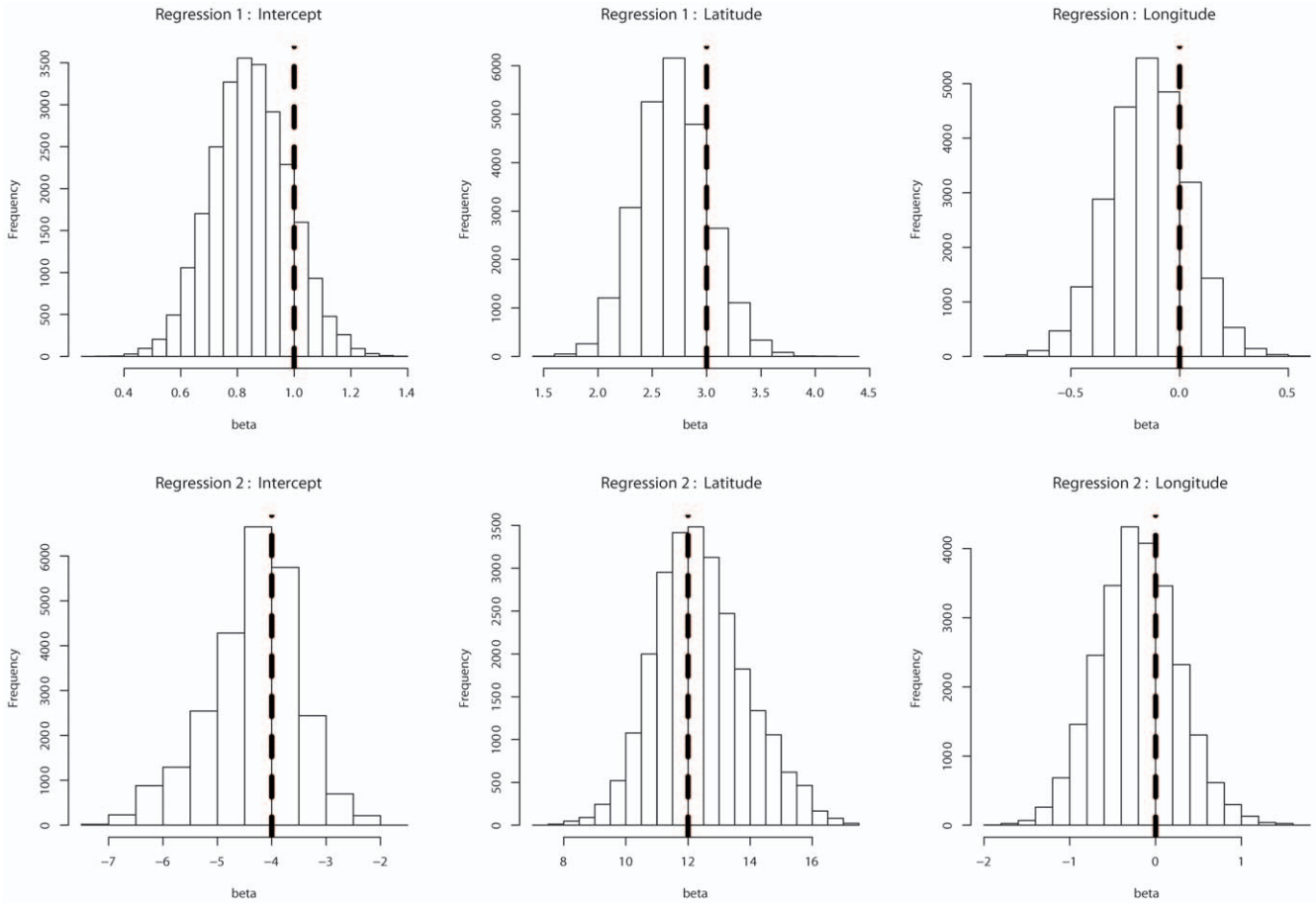

Figure 1. Posterior distributions of the regression coefficients for a data set simulated with the hidden regression model ($K=3$). The dashed vertical lines correspond to the regression coefficients used for generating the data. Two spatial covariates (latitude and longitude) are included in the regression model but only the first one influences genetic structure.
doi:10.1371/journal.pone.0016227.g001

Results

Simulation results

Using simulated data sets, we investigated whether including geographic and linguistic covariates can improve the estimation of membership probabilities or not, and we evaluated which subsets of variables best predict the estimated population genetic structure.

For the simulations where latitude was influential (equations (3) and (4)), we found that the true values of the regression coefficients were close to the mode of the posterior distributions (Figure 1). The influence of each covariate was thus correctly ascertained by POPS when the data were generated under its underlying statistical model. To further evaluate if missing the true set of covariates modifies the inference and the prediction of membership coefficients, we evaluated the performances of POPS using various hidden regression models. For all models, the misclassification rates were less than 4%. The upper bound was obtained under a model without covariates and for the smallest number of loci ($L=20$, Figure 2A). The misclassification rates never increased when we included a spurious longitude variable. With $L=20$ loci, the misclassification rate decreased to 2% when the correct covariate (latitude) was used. With $L=40$ loci, the misclassification rates were less than 1% for all hidden models. All individuals were perfectly assigned to their population of origin when latitude was included. For $L=100$, the misclassification rate was equal to 0% for all models. In the second series of simulations, linguistic covariates were added to the generating model (equations (5) and (6)). The misclassification rates were less than 30%, a value obtained for $L=20$ loci in a model without covariates (Figure 2B). With $L=20$, the misclassification rate decreased to 5% when including latitude and a linguistic variable with five levels. With $L=100$ loci, the misclassification rate of the model without covariates was around 1%. We conclude that when the data are generated from a hidden regression model, including covariates in POPS increases the performances of the program. This is particularly true when the number of loci is relatively small.

Finally, we studied the variable selection criteria for the data where latitude was influential (equations (3) and (4)) as well as linguistic covariates (equations (5) and (6)). Whatever the number of loci we considered, the increase of the correlation coefficient was larger when including latitude rather than longitude in the regression model. Figure 3 shows that the correlation coefficient and the cross-validation score reach a plateau when the true predictors are included in the hidden regression model. This plateau was found when latitude was the sole determinant of genetic structure and when linguistic covariates had an additional contribution to genetic differentiation.

For the five-island data with a level of differentiation of $F_{ST}=0.04$, the misclassification rates were less than 5% (Figure 2C). The worst performances were obtained for a model without covariates. When latitude (or longitude) was included in the hidden regression model, the misclassification rate decreased

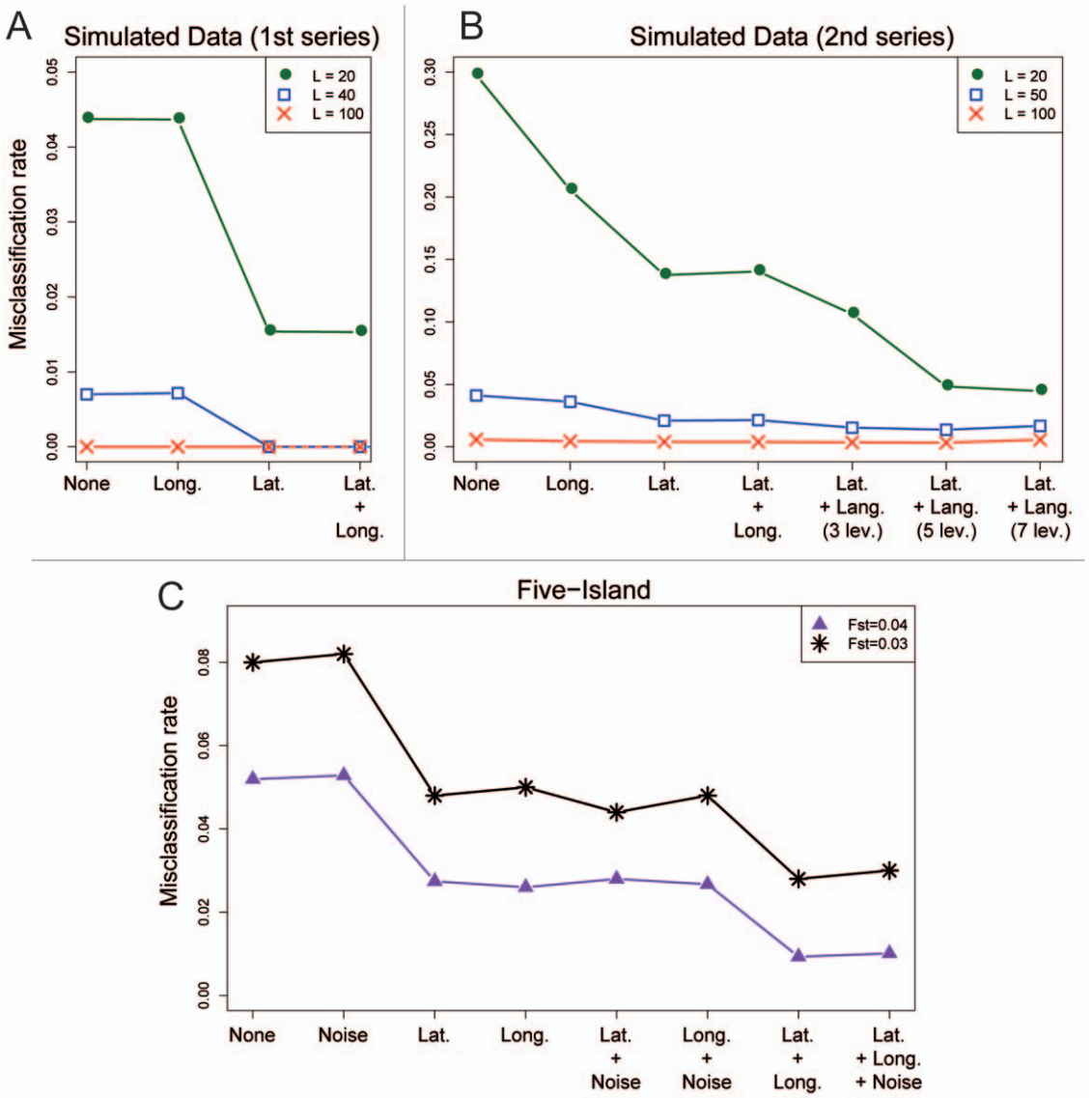

Figure 2. Misclassification rates for simulated data as a function of the covariates included in the clustering algorithm. A. The cluster memberships are influenced by latitude but not by longitude. B. The data are generated using latitude and a 5-level linguistic classification. C. The data are generated in a five-island model for which $F_{ST} = 0.03$ or 0.04.
doi:10.1371/journal.pone.0016227.g002

to 3%. When both latitude and longitude were included in the model, the misclassification rate decreased to 1%. The addition of a spurious noisy covariate did not impact the performance of the program. Regarding variable selection, Figure 3C shows that the correlation coefficients and the validation scores reach a plateau when longitude and latitude are included in the hidden regression model. For the five-island data with a level of differentiation of $F_{ST} = 0.03$, a model including latitude and longitude was also selected. In this case, the misclassification rate was equal to 2.8%. For these data, POPS compared favorably to the spatial versions of BAPS (misclassification rate = 3.9%) and TESS (misclassification rate = 4.4%) [42–44].

Native American HGDP data

To investigate the relationships between geography, languages and genes in Native American populations, we applied POPS to a multilocus genotype data set including 512 individuals from the HGDP. We compared the posterior membership coefficients predicted by four different models that use distinct linguistic classifications and we computed two variable selection criteria in order to discriminate among models (see Material and Methods).

The four clustering models resulted in highly similar patterns of estimated membership coefficients, and these patterns were also similar to the pattern found with structure (Figure 4, Figure S1, Wang *et al.* [15]). As we used a large number of microsatellite loci, these results are not surprising, and they warrant that the predictive power of the three linguistic classifications will be ascertained consistently.

Using a quadratic trend surface to correct for geographic effects, we compared the predictions of a model without languages (Model A) to the predictions of a model using Greenberg's classification at the stock level (Model B), a model using Greenberg's classification at the group level (Model C), and a model using *The Ethnologue* classification (Model D). Figure 4A compares the predictions of

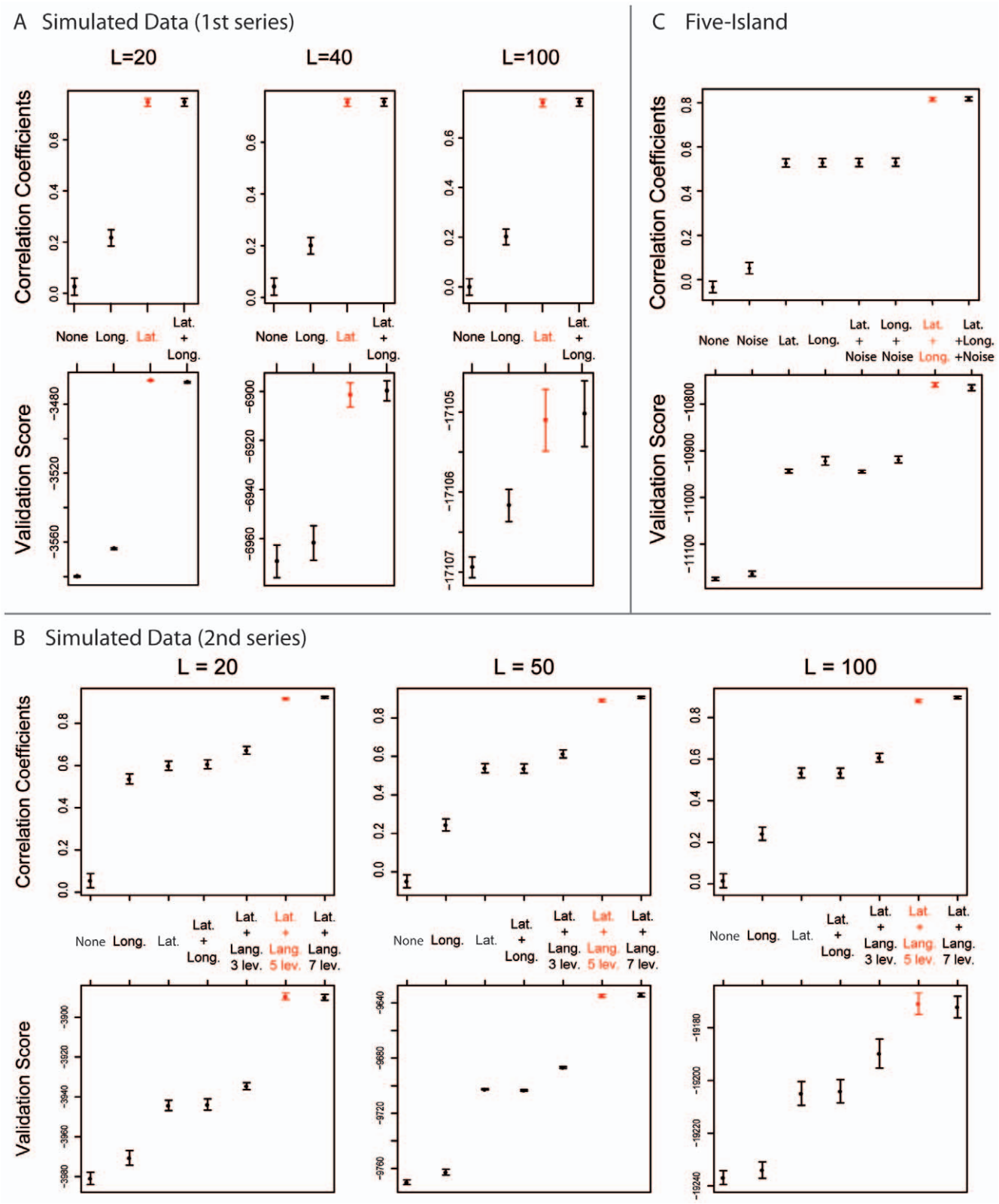

Figure 3. Variable selection for simulated data. The correlation coefficients ρ correspond to the correlations between the estimated and predicted membership probabilities. Confidence intervals of the correlation coefficients are estimated by assuming that the Fisher's transform $\operatorname{arctanh}(\rho)$ follows a Gaussian distribution [65]. The validation scores are estimated with the 2-fold cross-validation method. Their standard deviations are estimated by using a non-parametric bootstrap method. A. The cluster memberships are influenced by latitude but not by longitude. B. The data are generated using latitude and a 5-level linguistic classification. C. The data are generated in a five-island model for which $F_{ST} = 0.04$.
doi:10.1371/journal.pone.0016227.g003

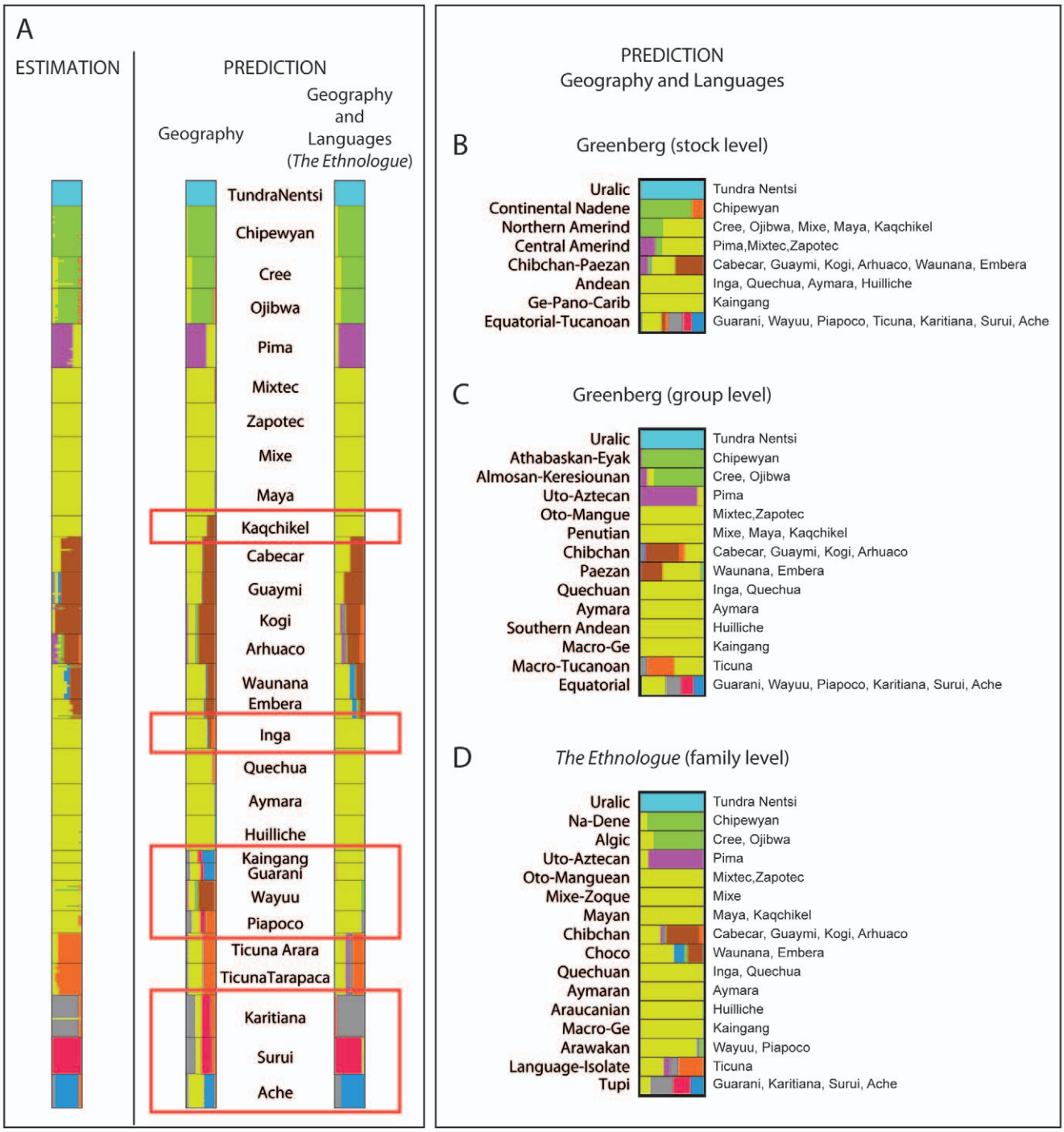

Figure 4. Estimated and predicted population genetic structure for 28 Native American populations. A. The membership coefficients are estimated in a model that includes spatial information (longitude, latitude). Inference of genetic structure is unchanged when we include additional linguistic covariates (Supporting Information Figure S1). The main differences between predictions obtained with or without linguistic information are framed in red. B-D. Membership coefficients predicted by Models B–D. The membership coefficients are averaged over individuals within the same linguistic unit.
doi:10.1371/journal.pone.0016227.g004

Model A and Model D. For many population samples, the membership probabilities predicted by Model A were close to the estimated coefficients ($\rho = 81\%$, Figure 5A). The predictions of Model A for every geographic location in the American mainland are displayed in Figure 6. The value of the correlation coefficient and the map of predicted membership coefficients confirmed that geography is a good predictor of genetic structure in Native American populations. When including linguistic covariates (Models B–D), the predictions of cluster membership were closer to the estimates of the MCMC algorithm than those obtained without languages (Model A) except for the Pima. The correlation coefficient increased from $\rho = 0.81$ to $\rho = 0.94 - 0.98$ (Figure 5A),

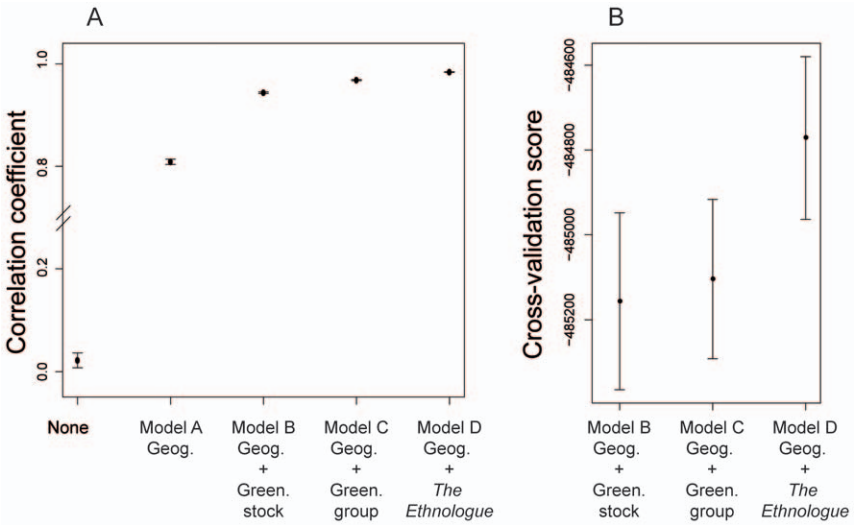

Figure 5. Variable selection for the Native American HGDP data. Geographic information includes longitude and latitude. Green. stands for Greenberg and Geog. stands for geography. The best model uses *The Ethnologue* linguistic classification.
doi:10.1371/journal.pone.0016227.g005

and the predicted genetic structure changed substantially (Figure 4A and Figure S1). For several populations the predictions obtained from linguistic covariates (Models B–D) differed from the predictions obtained with the geographic covariates only: Model A predicted that the Kaqchikel and the Wayuu samples shared substantial ancestry with a group comprising Cabecar, Guaymi, Kogi, Arhuaco, Waunana and Embera populations; Model A also predicted that the Kaingang and Guarani samples clustered with the Ache population, and that the Inga and Piapoco samples were grouped with the Ticuna sample.

Figure 4 B–D displays the membership coefficients predicted by POPS using Greenberg's and *The Ethnologue* classifications (Models B–D), grouping populations with the same linguistic taxon. At the exception of the Andean and Ge-Pano-Carib stocks, Greenberg's linguistic stocks were associated with multiple clusters (Figure 4B). Refining Greenberg's classification at the group level improved the characterization of genetic clusters by linguistic taxa (Model C, Figure 4C). At the group level, the Northern Amerind stock split into Almosan-Keresiouan and Penutian groups that correspond to genetically divergent clusters. Similarly, the Central Amerind stock split into Uto-Aztecan and Oto-Mangue groups which are also genetically divergent. However, the split of the Equatorial-Tucanoan stock into the Macro-Tucanoan and Equatorial groups, and the split of the Chibchan-Paezan stock into the Chibchan and Paezan groups, did not improve the prediction of genetic clusters. In *The Ethnologue* classification (Model D), the Equatorial group split into the Arawakan and Tupi families. This separation improved the prediction of genetic clusters since the Arawakan family was associated with a unique genetic cluster. In contrast, the separation of the Penutian group into the Mixe-Zoque and Mayan families did not improve the characterization of genetic groups. Overall *The Ethnologue* classification provided better predictions of genetic groups than Greenberg's classification. Among the 16 families of *The Ethnologue* classification, only the Tupi, Choco and Chibchan families were not associated to a unique genetic cluster (Figure 4D). Supporting these comparisons, Figure 5B shows that the cross-validation score increases when using *The Ethnologue* (Model D). The values of the cross-validation scores are approximately equal to $-485{,}100$ for Models B and C, and around $-484{,}750$ for Model D. These scores provide quantitative evidence that the classification of *The Ethnologue* leads to better predictions of genetic structure than Greenberg's classification at the stock or group levels.

Discussion

We proposed a Bayesian latent class regression model to investigate to which extent geographic and linguistic information can predict population genetic structure in Native American populations. The originality of this approach was to model individual responses, i.e., the unobserved genetic cluster labels for each individual, using spatial and linguistic variables.

Our simulation study provided evidence that a hidden regression layer can improve the inference of genetic structure in addition to allowing their predictions from covariates. We also tested two criteria of variable selection based on correlation coefficients and cross-validation scores and found that these statistical indices reached a plateau when the true set of covariates was included in the POPS model. With small numbers of loci, the use of covariates decreased the misclassification rates of the clustering program significantly. For large numbers of loci, the estimation performances were hard to improve, especially when the likelihood dominated the prior distribution. However, using large numbers of loci made predictions and the use of the variable selection criteria reliable.

Using 678 microsatellite markers from the HGDP data set, we evaluated the suitability of geographic and linguistic predictors for Native American population genetic structure. Geography predicted genetic clusters rather accurately. However considering linguistic origin in addition to geographic origin improved the prediction of genetic structure. After correcting for geographic effects, we evaluated the predictive capabilities of three linguistic classifications: Greenberg's classification at two distinct levels and *The Ethnologue* classification. We did not consider Greenberg's tripartite classification (Amerind, Na-Dene, and Eskimo-Aleut) because, in addition to being controversial [48], all Native American HGDP populations, except the Chipewyan, belong to the Amerind family. We rather focused our analysis on taxonomically lower levels of Greenberg's classification: linguistic stocks and groups. Considering those refined levels, *The Ethnologue*

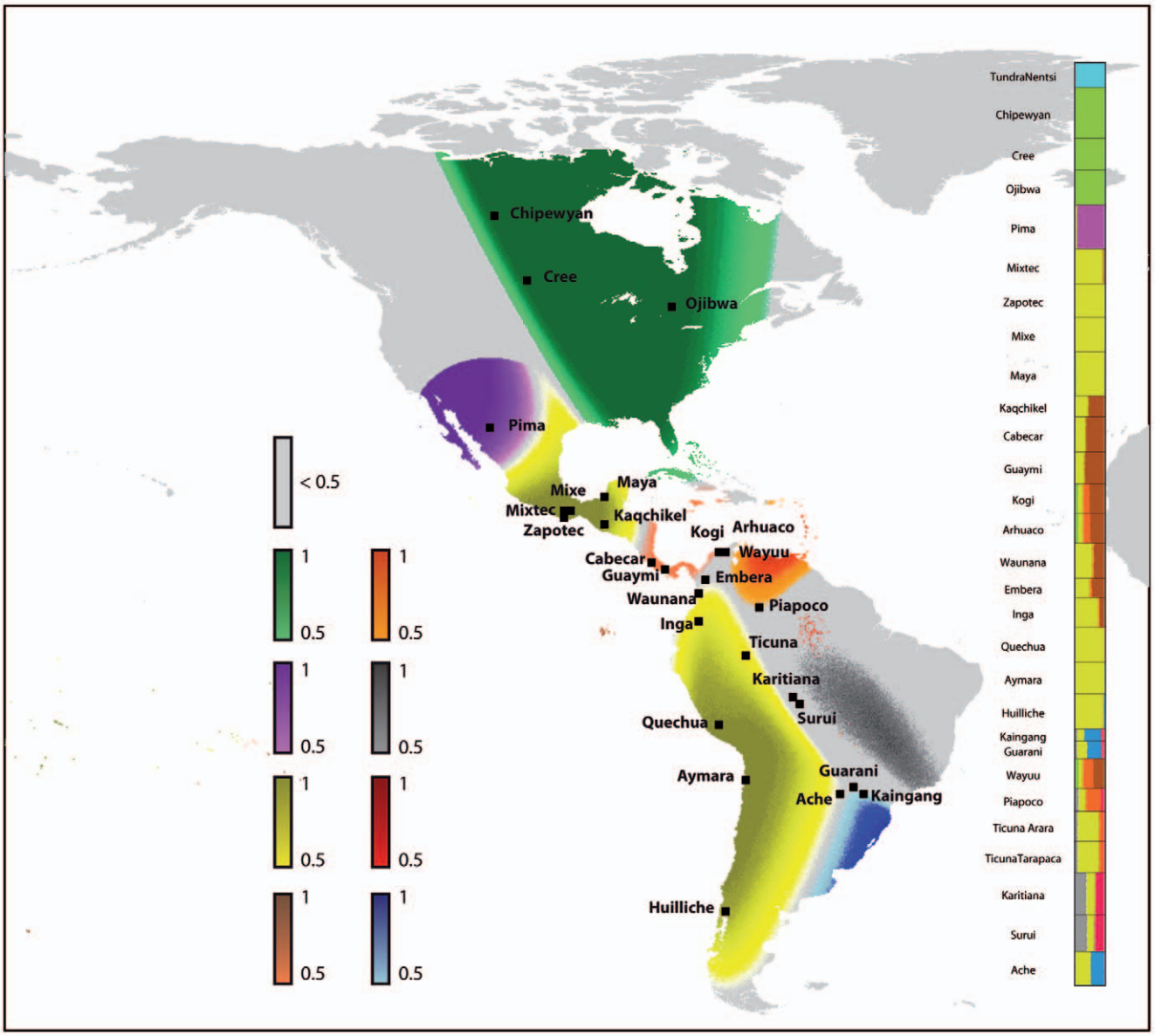

Figure 6. Genetic structure of Native American populations as predicted by geographical covariates. Geographical covariates include latitude, longitude, quadratic terms and an interaction term. Locations for which there is a cluster with a predicted membership coefficient larger than 0.5 are colored with the cluster color. Locations for which there is no cluster that reaches the 0.5 threshold or that are too distant from a sampled population are colored in grey. The barplot displays the membership probabilities as predicted by geographical covariates.
doi:10.1371/journal.pone.0016227.g006

provided better predictions of population genetic structure than Greenberg's classification.

Though *The Ethnologue* classification provided a better genetic proxy than Greenberg's classification, some linguistic families were not perfectly characterized in terms of genetic clustering. The Chibchan and Choco families were grouped in a Chibchan-Paezan stock by Greenberg [28]. These populations shared genetic ancestry with northern Mesoamerican populations (Mixtec, Zapotec, Mixe, Maya and Kaqchikel) and with southern Andean populations (Inga, Quechua, Aymara and Huilliche) (Figure 4A). Based on mtDNA data, Melton *et al.* [49] also found genetic relationships between Chibchan speakers and a Mayan population from Mesoamerica. To explain these relationships, it has been argued that Chibchan and Mesoamerican languages were all interrelated at one time into a larger Proto-Mesoamerican linguistic group that subsequently splintered into different language families after the intensification of agriculture in Mesoamerica [50,51]. The shared genetic relationships between Mesoamerican populations and Chibchan-Choco populations would result from their shared common history. Another family lacking genetic characterization was the Tupi. The Tupi family encompasses approximately 41 languages that spread throughout eastern South America several millennia ago [52,53]. Since the Tupi expansion involved language replacement, it may have blurred the relationships between genes and languages. Additionally, the Surui and Ache are populations with Tupi languages and small effective population sizes [15]. The 'genetic patchwork' of the Tupi would then result from genetic drift essentially.

Despite the intrinsic difference between methods, our analysis confirmed previous findings that a sizeable correspondence between

genetic and linguistic differentiation may exist only below a certain level of linguistic differentiation. The tests of treeness indicated that language classifications provide the best fit to mitochondrial data when they included external features of language classification trees and no deeper internal relationships between languages [14]. Using partial Mantel tests, Wang *et al.* [15] found a low partial correlation ($r = 0.01$) between linguistic (Greenberg's stock level) and genetic dissimilarities, but the correlation increased to $r = 0.40$ when the authors considered pairs of populations within stocks. Our analysis revealed that the congruence between genetic and linguistic diversification is more evident when considering a finer grain of linguistic differentiation than the stock level.

To further investigate potential scale effects, we applied POPS to 77 world-wide population samples from the HGDP data set excluding two language isolates (Basque and Burushaski) and grouping the sub-Saharian samples in a reference cluster (Table S2). The genetic clusters detected by POPS agreed with those detected by structure (Figure S2) [15,54]. The geographic predictions of a quadratic trend surface model were highly correlated to the estimated membership coefficients ($\rho = 0.97$). The high value of the correlation coefficient confirmed that geography is a good predictor of genetic structure at the world-wide scale [55–61]. Adding the linguistic covariates taken from *The Ethnologue* classification increased the correlation coefficient from $\rho = 0.97$ to $\rho = 0.98$. Thus it improved the prediction of genetic structure only marginally. These results provided evidence that the effects of language on the prediction of genetic structure are dependent on the scale considered. The results of POPS were also comparable to those obtained by Belle and Barbujani [23] reporting that languages have a small effect on the pattern of molecular variation at the world-wide scale. At the global scale, the patterns of genetic population structure are likely to reflect ancient demographic events, such as population divergence associated with the colonization of major geographic regions of the world [25]. At the continental scale, cultural traits contribute to the mediation of gene flow between human groups [62]. The predictive power provided by languages in the Americas could thus result from preferential mating within linguistic groups.

The examination of linguistic and genetic relationships in the Americas would obviously benefit from a more extensive sampling from the Na-Dene linguistic stock and from the inclusion of the Eskimo-Aleut stock. In a regression framework, a large dispersion of the explanatory variables is preferable. Though the sampling design of the HGDP was not optimal in our framework, our approach provided evidence that linguistic proxies improved the prediction of Native American population genetic structure. As human genomic data expand in genetic and geographic coverage [61,63,64], the use of latent class regression models could result in a more detailed picture of the role of geography and cultural factors in shaping human genetic variation.

Supporting Information

Figure S1 Estimated and predicted genetic structure of Native American populations, with $K = 9$ clusters, using different set of covariates in the probit model (Model A–D).
(TIF)

Figure S2 Genetic structure at a worldwide scale as predicted by geographical covariates when K = 7. Geographical covariates include latitude, longitude and distance to the Addis Abeba, which is computed by included five obligatory waypoints. The three barplots correspond to 1) the genetic structure as inferred with genetic data and both spatial and linguistic covariates, 2) the structure as predicted with spatial information and 3) the structure as predicted with spatial and linguistic information. The linguistic variable is a qualitative variable corresponding to The Ethnologue classification.
(TIF)

Table S1 Coordinates and linguistic entities of 28 Native American populations from the Human Genome Diversity Panel.
(PDF)

Table S2 Coordinates, distance to Addis-Abeba, and linguistic families of 77 worldwide populations from the Human Genome Diversity Panel.
(PDF)

Appendix S1 Gibbs sampler.
(PDF)

Appendix S2 Computation of the predictive score for cross-validation.
(PDF)

Acknowledgments

The software POPS implementing the algorithm described in this article is available at www-timc.imag.fr/Olivier.Francois/tess.html. We thank Eric Durand for his comments at various stages of this work. Simulations were run on the the UJF-CIMENT cluster of computers (http://healthphy.grenoble.cnrs.fr/).

Author Contributions

Conceived and designed the experiments: FJ OF MB. Performed the experiments: FJ. Analyzed the data: FJ OF MB. Contributed reagents/materials/analysis tools: FJ OF MB. Wrote the paper: FJ OF MB.

References

1. Greenberg J, Turner CI, Zegura S (1986) The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence. Curr Anthropol 27: 477–97.
2. Hunley K, Long JC (2005) Gene flow across linguistic boundaries in Native North American populations. Proc Natl Acad Sci USA 102: 1312–1317.
3. Bamshad M, Wooding S, Watkins W, Ostler C, Batzer MA, et al. (2003) Human population genetic structure and inference of group membership. Am J Hum Genet 72: 578–89.
4. Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. Nature Genetics 36: S21–S27.
5. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. Princeton University Press.
6. Spuhler J (1972) Genetic, linguistic and geographical distances in Native North America. In: Weiner JS, Huizinga J, eds. The Assessment of Population Affinities in Man. Oxford: Clarendon Press. pp 73–95.
7. Spielman RS, Migliazza EC, Neel JV (1974) Regional linguistic and genetic differences among Yanomama indians. Science 184: 637–644.
8. Chakraborty R, Blanco R, Rothhammer F, Llop E (1976) Genetic variability in Chilean Indian populations and its association with geography, language, and culture. Soc Biol 23: 73–81.
9. Murillo F, Rothhammer F, Llop E (1977) The Chipaya of Bolivia: dermatoglyphics and ethnic relationships. Am J Phys Anthropol 46: 45–50.
10. Salzano FM, Neel JV, Gershowitz H, Migliazza EC (1977) Intra and intertribal genetic variation within a linguistic group: the Ge-speaking indians of Brazil. Am J Phys Anthropol 42: 337–347.
11. Spuhler J (1979) Genetic distance, trees, and maps of North American Indians. In: Laughlin WS, Harper AB, eds. The First Americans: Origins, Affinities, and Adaptations. New York: Gustav Fischer. pp 135–183.
12. Barrantes R, Smouse PE, Mohrenweiser HW, Gershowitz H, Azofeifa J, et al. (1990) Microevolution in lower Central America: genetic characterization of the Chibcha-speaking groups of Costa Rica and Panama, and a consensus taxonomy based on genetic and linguistic affinity. Am J Hum Genet 46: 63–84.
13. Ward RH, Redd A, Valencia D, Frazier B, Pääbo S (1993) Genetic and linguistic differentiation in the Americas. Proc Natl Acad Sci USA 90: 10663–10667.

14. Hunley KL, Cabana GS, Merriwether DA, Long JC (2007) A formal test of linguistic and genetic coevolution in native Central and South America. Am J Phys Anthropol 132: 622–631.
15. Wang S, Lewis CM, Jr., Jakobsson M, Ramachandran S, Ray N, et al. (2007) Genetic variation and population structure in Native Americans. PLoS Genet 3: e185.
16. Cavalli-Sforza LL, Piazza A (1975) Analysis of evolution: evolutionary rates, independence and treeness. Theor Popul Biol 8: 127–165.
17. Cavalli-Sforza LL, Minch E, Mountain JL (1992) Coevolution of genes and languages revisited. Proc Natl Acad Sci USA 89: 5620–5624.
18. Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27: 209–220.
19. Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Biol 35: 627–632.
20. Ayub Q, Mansoor A, Ismail M, Khaliq S, Mohyuddin A, et al. (2003) Reconstruction of human evolutionary tree using polymorphic autosomal microsatellites. Am J Phys Anthropol 122: 259–68.
21. Campbell L (2006) Long-range comparison: methodological disputes. In: Brown K, ed. Encyclopedia of Language and Linguistics. Oxford: Elsevier. 2nd edition. pp 324–331.
22. Heggarty P, Maguire W, McMahon A (2010) Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. Philosophical Transactions of the Royal Society B: Biological Sciences 365: 3829–3843.
23. Belle E, Barbujani G (2007) Worldwide analysis of multiple microsatellites: language diversity has a detectable in inuence on DNA diversity. Am J Phys Anthropol 133: 1137–1146.
24. Excoffier L, Harding R, Sokal R, Pellegrini B, Sanchez-Mazas A (1991) Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. Hum Biol 63: 273–297.
25. Hunley K, Dunn M, Lindström E, Reesink G, Terrill A, et al. (2008) Genetic and linguistic coevolution in northern island Melanesia. PLoS Genet 4: e1000239.
26. Colonna V, Boattini A, Guardiano C, Longobardi G, Pettener D, et al. (2010) Long-range comparisons between genes and languages based on syntactic differences. Hum Hered In press.
27. Bandeen-Roche K, Miglioretti D, Zeger S, Rathouz P (1997) Latent variable regression for multiple discrete outcomes. J Am Stat Assoc 92: 1375–1386.
28. Greenberg J (1987) Language in the Americas. Stanford: Stanford University Press.
29. Gordon RG (2005) Ethnologue: Languages of the World. Dallas: SIL International, fifteenth edition. 533 p. URL www.ethnologue.com.
30. Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci USA 85: 6002–6006.
31. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
32. Dawson KJ, Belkhir K (2001) A bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet Res 78: 59–77.
33. Corander J, Waldmann P, Sillanpaa MJ (2003) Bayesian analysis of genetic differentiation between populations. Genetics 163: 367–374.
34. DeSarbo W, Cron W (1988) A maximum likelihood methodology for clusterwise linear regression. J Classif 5: 249–282.
35. Chung H, Flaherty B, Schafer J (2006) Latent class logistic regression: application to marijuana use and attitudes among high-school seniors. J R Stat Soc Ser A 169: 723–743.
36. Durand E, Jay F, Gaggiotti OE, François O (2009) Spatial inference of admixture proportions and secondary contact zones. Mol Biol Evol 26: 1963–1973.
37. Suits DB (1957) Use of dummy variables in regression equations. J Am Stat Assoc 52: 548–551.
38. Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88: 669–679.
39. Ripley BD (1996) Pattern Recognition and Neural Networks. Cambridge University Press.
40. Hastie T, Tibshirani R, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag, second edition. 533 p.
41. Smyth P (2000) Model selection for probabilistic clustering using cross-validated likelihood. Stat Comput 10: 63–72.
42. Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Mol Ecol Notes 7: 747–756.
43. Corander J, Sirn J, Arjas E (2008) Bayesian spatial modeling of genetic population structure. Comput Stat 23: 111–129.
44. François O, Durand E (2010) Spatially explicit Bayesian clustering models in population genetics. Mol Ecol Resour 10: 773–784.
45. Ruhlen M (1991) A Guide to the World's Languages. Volume 1: Classification. Stanford University Press.
46. Campbell L (1997) American Indian Languages: The Historical Linguistics of Native America. New York: Oxford University Press.
47. Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801–1806.
48. Lewin R (1988) American Indian language dispute. Science 242: 1632–1633.
49. Melton P, Briceno I, Gomez A, Devor E, Bernal J, et al. (2007) Biological relationship between Central and South American Chibchan speaking populations: evidence from mtDNA. Am J Phys Anthropol 133: 753–770.
50. Witkowski S, Brown C (1981) Mesoamerican historical linguistics and distant genetic relationship. Am Anthropol 83: 905–911.
51. Bellwood PS (2005) The First Farmers: The Origins of Agricultural Societies. Oxford: Blackwell.
52. Noelli FS (1998) The Tupi: explaining origin and expansions in terms of archaeology and of historical linguistics. Antiquity 72: 648–663.
53. Noelli F (2008) The Tupi expansion. In: Silverman H, Isbell WH, eds. The Handbook of South American Archaeology. New York: Springer. pp 400–401.
54. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) Genetic structure of human populations. Science 298: 2381–2385.
55. Dugoujon JM, Hazout S, Loirat F, Mourrieras B, Crouau-Roy B, et al. (2004) GM haplotype diversity of 82 populations over the world suggests a centrifugal model of human migrations. Am J Phys Anthropol 125: 175–192.
56. Manica A, Prugnolle F, Balloux F (2005) Geography is a better determinant of human genetic differentiation than ethnicity. Hum Genet 118: 366–371.
57. Prugnolle F, Manica A, Balloux F (2005) Geography predicts neutral genetic diversity of human populations. Curr Biol 15: R159–R160.
58. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 102: 15942–15947.
59. Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of populations. Genetics 174: 875–891.
60. Handley LJL, Manica A, Goudet J, Balloux F (2007) Going the distance: human population genetics in a clinal world. Trends Genet 23: 432–439.
61. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98–101.
62. Premo LS, Hublin JJ (2009) Culture, population structure, and low genetic diversity in Pleistocene hominins. Proc Natl Acad Sci USA 106: 33–37.
63. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998–1003.
64. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.
65. Fisher R (1915) Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika 10: 507–521.