Genetic Population Structure in Pacific Islanders

Suyash Shringarpure and Eric Xing
Machine Learning Department, Carnegie Mellon University

March 15, 2010

Abstract

Analyzing genetic population structure is useful in gaining insight about the evolutionary history of human populations. It gives us an understanding of the genetic similarities and differences between different populations and helps study their isolation, migration and inter-mixing.

With modern genotype sequencing methods, new data sets about human population are becoming available. The Human Genome Diversity Project (HGDP) includes population samples from 52 worldwide populations genotyped at hundreds of polymorphic loci. However, the sampling of populations from across the world was not uniform, with some geographical regions such as the Indian subcontinent, the Americas and the Pacific islands severely under-sampled. Newer data sets have focused on genotyping samples from populations which were insufficiently sampled previously. The Pacific Islanders are of interest anthropologically due to their cultural diversity and evolutionarily due to their geographical location.

In this work, we analyze the genetic population structure in the Pacific islander populations using *mStruct*, a methodology based on mixed-membership models that we developed previously. We present results of the population structure analyses and our hypotheses about how the populations have evolved from ancient populations. Our results show that there is a large amount of genetic diversity present in the Pacific islanders, and that it varies across the various islands of the Pacific, largely determined by geography. There is also some variation in genetic profiles that correlates with local languages.

1 Introduction

Genetic population structure is the assignment of individuals to different groups or clusters based on a genetic measure of similarity. Human genetic population structure is of great interest for the study of human evolutionary history. Since DNA is inherited, a study of genetic similarity and differences gives us clues about human evolution after the origin of modern humans in Africa. Hypotheses about human migration into different parts of the world can be made by examining genetic population structure [1, 2]. Identifying population structure

has also shown been shown to be of importance in association studies, where it can cause false positive results [3].

In the past few years, a lot of human genetic data has become available through various international projects for scientific analysis. The HapMap project [4] and the Human Genome Diversity Project (HGDP) [5] are two such projects. The HapMap project has provided dense genotype data on small groups of individuals of African, European and Asian origin. The HGDP project has genotype data from 52 worldwide populations on 1056 individuals. These projects have allowed researchers to perform many studies on human genetic variation [6, 7]. They have enabled scientists to test various hypotheses about human evolution [1].

However, despite their utility, there have been some criticisms of the data sets, a major one being the presence of large regions of the world which are severely under-sampled, prominent examples being the Indian subcontinent, the Native American populations in North America, and the Pacific islander populations from Oceania. Each of these populations is interesting for various anthropological and evolutionary reasons. Pacific islander populations show distinctive cultural variation, which leads to questions about their evolution and migration into the various Pacific islands. These questions can be attempted to be answered by studying the genetic population structure present in these populations.

We analyzed the genetic population structure of Pacific islanders, using the *mStruct* methodology [8]. *mStruct* is a mixed-membership model that allows analysis of population structure in admixed populations while simultaneously allowing for allele mutations. *mStruct* is described in more detail later in Section 3.2.2. Our results show that there is a large amount of genetic diversity present in the Pacific islanders, and that it varies across the various islands of the Pacific according to geographical location and island size. Geography has a strong impact on genetic similarity and differences. There is also some variation in genetic profiles that correlates with local languages. We also performed model selection analysis to find the number of ancestral populations that gave rise to the modern Pacific islander populations.

2 Related Work

The earliest genetic analyses of Pacific islander populations have been performed by examining variation in mitochondrial DNA (mtDNA) [9, 10] and non-recombining Y-chromosome markers [10, 11]. Since Y-chromosomes are inherited paternally (by sons) and mtDNA is inherited maternally, these analyses provide an incomplete picture of evolutionary history. Autosomal microsatellite genotype data from the HGDP was also used to study worldwide populations, including a few Pacific islander populations [6, 12]. However the number of Pacific islander individuals genotyped in these studies has been very small ($\sim$50 individuals from two Melanesian populations). As a result, it has not been possible to study the genetic variation in Pacific islander populations in depth.

Friedlaender et al. [13] studied the genetic variation in 41 populations from islands in the Pacific using the *STRUCTURE* model by Pritchard et al. [14]. They found that genetic diversity within individual Pacific populations was low, but due to high differentiation among Melanasian groups, the overall genetic diversity of the region is very high. They observed that the amount of differentiation varied between islands, and was dependent on island size and topographical complexity. They found that patterns of differentiation loosely track language distinctions in the Pacific populations. A long standing question in the evolution of Pacific island populations has been about how fast Polynesian populations moved through Melanesia and the amount of intermixing between these two groups of populations [10, 15]. From their analysis, they concluded that the Polynesian migration through Melanesia was relatively rapid and there was only a modest amount of intermixing with the indigenous populations.

Our work is similar in methodology to the work by Friedlaender et al. [13]. In Shringarpure and Xing [8], we found that the *mStruct* model and *STRUCTURE* model often produce different summaries of population structure due to their modeling differences. In this re-analysis of the data in Friedlaender et al. [13], we found that our results mostly agree with Friedlaender et al. [13] with minor differences. We find that the amount of genetic diversity varies across different islands, with effects of both geography and language evident.

Recently, there has been also some work in using language phylogenies to study the exapnsion of humans into the Pacific islands [15]. These methods use information about language evolution and and similarity to address the question of the mechanisms of the settlement of Pacific islands.

3 Materials and Methods

We used data published in an earlier study by Friedlaender et al. [13] for our analysis. It is currently the largest set of genotype data for Pacific islander populations publicly available. For analyzing the data, we used *mStruct*, a method we developed previously in Shringarpure and Xing [8]. In the following text, we describe the data used for analysis and briefly describe the motivation and graphical model for *mStruct*.

3.1 The data

Friedlaender et al. [13] used data from 41 populations in the Pacific for their study. They used 687 autosomal microsatellites in 952 individuals from 41 Pacific populations. Due to restrictions on data release, the data published online does not contain individuals from Maori and Taiwanese populations (4 groups in total).

Microsatellites, or "short sequence tandem repeats" (sstr), are a class of genomic markers consisting of repeating sequences of small units of DNA, usually 1-6 base pairs in length. For example, a DNA sequence such as "CACACACACACACA" is a microsatellite with repeating unit "CA" and a repeat count of 7.

Microsatellites are highly polymorphic DNA markers due to their high mutation rate [16, 17] and variation is indicated by a change in the number of repeats of the unit. Each variant of a particular microsatellite is called an allele. Since microsatellites can have a very large number of alleles, they are very informative markers. Microsatellite loci have been used before in DNA fingerprinting [18], linkage analysis [19], and in the reconstruction of human phylogeny [2]. By applying theoretical models of microsatellite evolution to data, questions such as time of divergence of two populations can be attempted to be addressed [20, 21].

3.2 Admixture Models for Population Structure

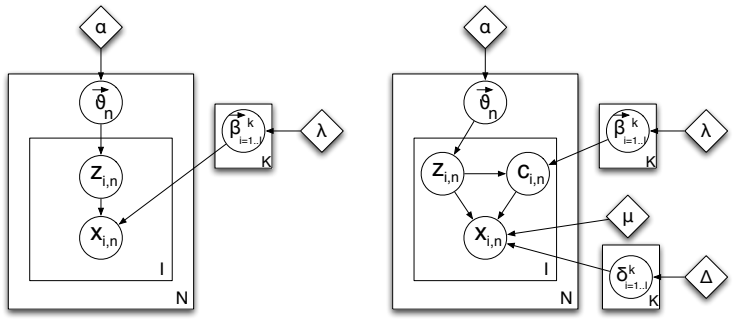
Admixtures are useful for modeling objects (e.g., human beings) each comprising multiple instances of some attributes (e.g., marker alleles), each of which comes from a (possibly different) source distribution $P_k(\cdot|\Theta^k)$, according to an individual-specific *admixing vector* (a.k.a. *ancestry vector*) $\vec{\theta}$. The *ancestry vector* represents the normalized contribution from each of the source distributions $\{P_k \ ; \ k = 1 : K\}$ to the object in question. For a single data set, all the *ancestry vectors* are assumed to be samples from an underlying structure prior with parameter α. We will represent individuals as a collection of alleles indexed by their locus, and also use an extra subscript $e \in \{1, 2\}$ to allow for the diploid nature of human chromosomes. Thus the allele at locus i in individual n, on chromosome copy e is indicated by x_{i,n_e}. Suppose, for every individual, the alleles at each locus may be inherited from founders in different ancestral populations, with each ancestral population represented by a unique distribution of founding alleles and parameters that determine the inheritance. Formally, this scenario can be captured in the following generative process:

1. For each individual n, draw the admixing vector: $\vec{\theta}_n \sim P(\cdot|\alpha)$, where $P(\cdot|\alpha)$ is a pre-chosen structure prior.
2. For the marker allele at locus $i, x_{i,n_e} \in \mathbf{x}_n$
   - 2.1: draw the latent *ancestral-population-origin* indicator $z_{i,n_e} \sim \text{Multinomial}(\cdot|\vec{\theta}_n)$
   - 2.2: draw the allele $x_{i,n_e}|z_{i,n_e} = k \sim P_k(\cdot|\Theta_i^k))$.

Depending on how ancestral populations and the way of inheritance of ancestral alleles are modeled, we can have different probability distributions for $P_k(\cdot|\Theta^k)$ in the last sampling step above, and thereby different admixtures of very different characteristics. Below we describe *STRUCTURE*, a popular admixture model for summarizing population structure. We then describe the motivation behind the *mStruct* model and the modeling, inference and parameter estimation procedures.

3.2.1 The *STRUCTURE* model

In *STRUCTURE*, the ancestral populations are represented by a set of population-specific allele frequency distributions. Thus the distribution $P_k(\cdot|\Theta^k)$ from which an observed allele can be sampled is a multinomial distribution defined by the frequencies of all observed alleles in the ancestral population, i.e.,



(a) *STRUCTURE* (b) *mStruct*

Figure 1: Graphical models with plate representation for *STRUCTURE* and *mStruct*. The subscript 'e' indicating chromosome copy is dropped for ease of notation.

$x_{i,n_e}|z_{i,n_e} = k \sim \text{Multinomial}(\cdot|\vec{\beta}_i^k)$. Thus in the *STRUCTURE*model, ancestral populations and modern populations have the same alleles, and inheritance of an ancestral allele to a modern allele is as a perfect copy. Figure 1(a) shows the plate representation of the *STRUCTURE* graphical model.

But a serious pitfall of using such a model, as pointed out by Excoffier and Hamilton [22], is that there is no mutation model for individual alleles with respect to the common prototypes, i.e., every unique allele measurement at a particular locus is assumed to correspond to a unique ancestral allele, rather than allowing the possibility of it just being derived from some common ancestral allele at that locus as a result of a mutation. This often results in an overestimation of the amount of genetic differences between populations.

3.2.2 The *mStruct* model

In Shringarpure and Xing [8], we proposed to represent each ancestral population by a set of population-specific mixtures of ancestral alleles (MAA). In a MAA for population k, for each locus (locus i for instance) we define a finite set of founders with prototypical alleles $\mu_i^k \equiv \{\mu_{i,1}^k, \ldots, \mu_{i,L_i}^k\}$ that can be different from the alleles observed in a modern population; each founder allele is associated with a unique frequency $\beta_{i,l}^k, l \in \{1, \cdots, \text{Ł}_i\}$, and a unique (if desired) mutation model P_m from the prototype allele parameterized by rate $\delta_{i,l}^k$. Under this representation, now the distribution $P_k(\cdot|\Theta_i^k)$ from which an observed allele can be sampled becomes a mixture of inheritance models each defined on a specific founder; and the ensuing sampling module that can be plugged into the general admixture scheme outlined above (to replace step 2.2) becomes a two-step generative process:

- 2.2a: draw the latent founder indicator $c_{i,n_e}|z_{i,n_e} = k \sim \text{Multinomial}(\cdot|\vec{\beta}_i^k)$;
- 2.2b: draw the allele $x_{i,n_e}|c_{i,n_e} = l, z_{i,n_e} = k \sim P_m(\cdot|\mu_{i,l}^k, \delta_{i,l}^k)$,

where $P_m()$ is a mutation model that can be flexibly defined based on the type of genetic marker being studied.

Figure 1(b) shows a graphical model representation of the *mStruct* model. For simplicity of presentation, in the model described above, we assume that for a particular individual, the genetic markers at each locus are conditionally *iid* samples (given the ancestry proportion) from a set of population-specific fixed-dimensional mixture of inheritance models, and that the set of founder alleles (but not their frequencies) at a particular locus is the same for all ancestral populations (i.e., $\mu_i^k \equiv \mu_i$). We shall also assume that the mutation parameters for each population at any locus are independent of the alleles at that locus (i.e., $\delta_{i,l}^k \equiv \delta_i^k$). Also, our model assumes Hardy-Weinberg equilibrium within populations. The simplifying assumptions of *unlinked loci and no linkage disequilibrium between loci within populations* can be easily removed by incorporating Markovian dependencies over ancestral indicators Z_{i,n_e} and Z_{i+1,n_e} of adjacent loci, and over other parameters such as the allele frequencies $\vec{\beta}_i^k$ in exactly the same way as in $STRUCTURE$.

3.2.3 Other Modeling Issues

Apart from the graphical model, there are certain other issues that need to be discussed before inference and estimation. Below we discuss the issue of what mutation model to use, how to decide what the ancestral alleles are and how to set the mutation parameters of the ancestral populations.

Microsatellite mutation model The choice of a suitable microsatellite mutation model is important, for both computational and interpretation purposes. Below we discuss the mutation model that we use and the biological interpretation of the parameters of the mutation model. We begin with a stepwise mutation model for microsatellites widely used in forensic analysis [23].

This model defines a conditional distribution of a progeny allele b given its progenitor allele a, both of which take continuous values:

$$p(b|a) = \frac{1}{2}\xi(1-\delta)\delta^{|b-a|-1}, \tag{1}$$

where ξ is the mutation rate (probability of any mutation), and δ is the factor by which mutation decreases as distance between the two alleles increases. Although this mutation distribution is not stationary (i.e., it does not ensure allele frequencies to be constant over the generations), it is commonly used in forensic inference due to its simplicity. To some degree δ can be regarded as a parameter that controls the probability of unit-distance mutation, as can be seen from the following identity: $p(b+1|a)/p(b|a) = \delta$.

In practice, the alleles for almost all microsatellites are represented by discrete counts. The two-parameter stepwise mutation model described above complicates the inference procedure. We propose a discrete microsatellite mutation model that is a simplification of Eq. 1, but captures its main idea. We posit that: $P(b|a) \propto \delta^{|b-a|}$. Since $b \in [1, \infty)$, the normalization constant of this distribution is:

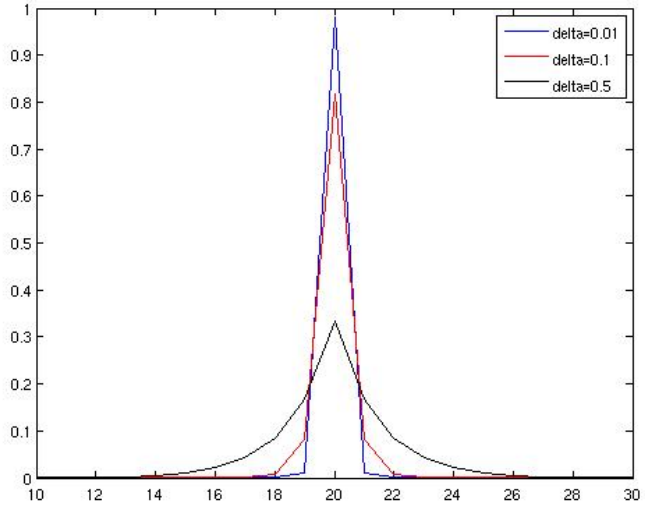


Figure 2: Probability distribution for various values of parameter δ

$$\begin{aligned}\sum_{b=1}^{\infty} P(b|a) &= \sum_{b=1}^{a} \delta^{a-b} + \sum_{b=a+1}^{\infty} \delta^{b-a} \\ &= \frac{1-\delta^a}{1-\delta} + \frac{\delta}{1-\delta} \\ &= \frac{1+\delta-\delta^a}{1-\delta},\end{aligned}$$

which gives the mutation model as

$$P(b|a) = \frac{1-\delta}{1-\delta^a+\delta}\delta^{|b-a|}. \tag{2}$$

We can interpret δ as a variance parameter, the factor by which probability drops as a fuction of the distance between the mutated version b of the allele a. Figure 2 shows the discrete pdf for various values of δ.

Determination of founder set at each locus: According to our model assumptions, there can be a different number of founder alleles at each locus. This number is typically smaller than the number of alleles observed at each marker since the founder alleles are "ancestral". In principle, some alleles may be lost when modern populations are derived from ancestral populations. However, since such alleles will never be observed in the data, we will choose not to model

them. To estimate the appropriate number and allele states of founders, we fit finite mixtures (of fixed size, corresponding to the desired number of ancestral alleles) of microsatellite mutation models over all the measurements at a particular marker for all individuals. We use the Bayesian Information Criterion (BIC) [24] to determine the best number and states of founder alleles to use at each locus, since information criteria tend to favor smaller number of founder alleles which fit the observed data well.

For each locus, we fit many different finite-sized mixtures of mutation distributions, with the size varying from 1 to the number of observed alleles at the locus. For each mixture size, the likelihood is optimized and a BIC value is computed. The number of founder alleles is chosen to be the size of the mixture that has the best (minimum) BIC value. We can do this as a pre-processing step before the actual inference or estimation procedures. This is possible since we assumed that the set of founder alleles at each locus was the same for all populations.

Choice of mutation prior: In our model, the δ parameter, as explained earlier, is a population-specific parameter that controls the probability of stepwise mutations. Being a parameter that controls the variance of the mutation distribution, there is a possibility that inference on the model will encourage higher values of δ to improve the log-likelihood, in the absence of any prior distribution on δ. To avoid this situation, and to allow more meaningful and realistic results to emerge from the inference process, we impose on δ a beta prior that will be biased towards smaller values of δ. The beta prior will be a fixed one and will not be among the parameters we estimate.

3.3 Inference and Parameter Estimation

For notational convenience, we will ignore the diploid nature of observations in the analysis that follows. With the understanding that the analysis is carried out for an arbitrary n^{th} individual, we will drop the subscript n. Also, we overload the indicator variables z_i and c_i to be both, arrays with only one element equal to 1 and the rest equal to 0, as well as scalars with a value equal to the index at which the array forms have 1s. We will let K be the number of populations, and N be the number of individuals. In other words: $z_i \in \{1, \ldots, K\}$ or $z_i = [z_{i,1}, \ldots, z_{i,K}]$, where $z_{i,k} = \mathcal{I}[z_i = k]$, and $\mathcal{I}[\cdot]$ denotes an indicator function that equals to 1 when the predicate argument is true and 0 otherwise. A similar overloading is also assumed for the c_i variables. For generalization across different types of markers, we shall use $f(x_i|\mu_{i,c_i}, \delta_{i,z_i})$ to denote $P(x_i|c_i, z_i, \mu_i, \delta_i)$. Different mutation models can be used in *mStruct* by varying the form of the function $f()$.

The joint probability distribution of the the data and the relevant variables under the *mStruct* model can then be written as:

$$
\begin{aligned}
& P\left(\mathbf{x}, \mathbf{z}, \mathbf{c}, \vec{\theta} | \alpha, \boldsymbol{\beta}, \mu, \delta\right) \\
= \; & p\left(\vec{\theta}|\alpha\right) \prod_{i=1}^{I} P\left(z_i|\vec{\theta}\right) P\left(c_i|z_i, \vec{\beta}_i^{k=1:K}\right) P\left(x_i|c_i, z_i, \mu_i, \delta_i^{k=1:K}\right).
\end{aligned}
$$

The marginal likelihood of the data can be computed by summing/integrating out the latent variables. However, a closed-form solution to the summation/integration is not possible, and indeed exact inference on hidden variables such as the ancestry proportions $\vec{\theta}$, and estimation of model parameters such as the mutation parameters δ under *mStruct* is intractable. Pritchard et al. [14] developed an MCMC algorithm for approximate inference for their admixture model underlying *Structure*. While it is straightforward to implement a similar MCMC scheme for *mStruct*, we choose to apply a computationally more efficient approximate inference method known as variational inference [25].

3.3.1 Variational Inference

We use a mean-field approximation for performing inference on the model. This approximation method approximates an intractable joint posterior $p()$ of all the hidden variables in the model by a product of marginal distributions $q() = \prod q_i()$, each over only a single hidden variable. The optimal parameterization of $q_i()$ for each variable is obtained by minimizing the Kullback-Leibler divergence between the variational approximation q and the true joint posterior p. Using results from the Generalised Mean Field theory [26], we can write the variational distributions of the latent variables in *mStruct* as follows:

$$
\begin{aligned}
q(\vec{\theta}) &\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + \sum_{i=1}^{I} \langle z_{i,k} \rangle} \\
q(c_i) &\propto \prod_{l=1}^{L} \left(\prod_{k=1}^{K} \left(\beta_{i,l}^k f(x_i|\mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle} \right)^{c_{i,l}} \\
q(z_i) &\propto \prod_{k=1}^{K} \left(e^{\langle \log(\theta_k) \rangle} \left(\prod_{l=1}^{L} \beta_{i,l}^k f(x_i|\mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right) \right)^{z_{i,k}}.
\end{aligned}
$$

In the distributions above, the '$\langle \cdot \rangle$' are used to indicate the expected values of the enclosed random variables. A close inspection of the above formulas reveals that these variational distributions have the form $q(\vec{\theta}) \sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_K)$, $q(z_i) \sim \text{Multinomial}(\rho_{i,1}, \ldots, \rho_{i,K})$, and $q(c_i) \sim \text{Multinomial}(\xi_{i,1}, \ldots, \xi_{i,L})$, respectively, of which the parameters $\gamma_k, \rho_{i,k}$ and $\xi_{i,l}$ are given by the following

equations:

$$\begin{aligned}
\gamma_k &= \alpha_k + \sum_{i=1}^{I} \langle z_{i,k} \rangle \\
\rho_{i,k} &= \frac{e^{\langle \log(\theta_k) \rangle} \left(\prod_{l=1}^{L} \beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right)}{\sum_{k=1}^{K} \left(e^{\langle \log(\theta_k) \rangle} \left(\prod_{l=1}^{L} \beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right) \right)} \\
\xi_{i,l} &= \frac{\prod_{k=1}^{K} \left(\beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle}}{\sum_{k=1}^{K} \left(\prod_{k=1}^{K} \left(\beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle} \right)}
\end{aligned}$$

and they have the properties: $\langle \log(\theta_k) \rangle = \psi(\gamma_k) - \psi(\sum_k \gamma_k)$, $\langle z_{i,k} \rangle = \rho_{i,k}$ and $\langle c_{i,l} \rangle = \xi_{i,l}$, which suggest that they can be computed via fixed point iterations. (The digamma function $\psi()$ used above is the first derivative of the logarithm of the gamma function $\Gamma()$.) It can be shown that this iteration will converge to a local optimum, similar to what happens in an EM algorithm. Empirically, a near global optimal can be obtained by multiple random restarts of the fixed point iteration. Typically, such a mean-field variational inference converges much faster than sampling [26]. Upon convergence, we can easily compute an estimate of the map vector $\vec{\theta}$ for each individual from $q(\vec{\theta})$.

3.3.2 Parameter Estimation

The parameters of our model are the ancestral alleles μ, the mutation parameters δ, the ancestral allele frequency distributions $\boldsymbol{\beta}$, and the Dirichlet hyperparameter that is the prior on ancestral populations, α. For the hyperparameter estimation, we perform empirical Bayes estimation using the variational Expectation Maximization algorithm described in [27]. The variational inference described in Section 3.3.1 provides us with a tractable lower bound on the log-likelihood as a function of the current values of the hyperparameters. We can thus maximize it with respect to the hyperparameters. If we alternately carry out variational inference with fixed hyperparameters, followed by a maximization of the lower bound with respect to the hyperparameters for fixed values of the variational parameters, we can get an empirical Bayes estimate of the hyperparameters. The derivation, details of which we will not show here, leads to the following iterative algorithm:

1. (*E-step*) For each individual, find the optimizing values of the variational parameters $(\gamma_n, \rho_n, \xi_n; n \in 1, \ldots, N)$ using the variational updates described above.

2. (*M-step*) Maximize the resulting variational lower bound on the likelihood with respect to the model parameters, namely $\alpha, \beta, \mu, \delta$.

The two steps are repeated until the lower bound on the log-likelihood converges.

3.4 Experiments

We ran *mStruct* on the dataset for values of K (the number of ancestral populations) ranging from K=2 to K=10. For each value of K, 20 runs were started from random values of the initial parameters to account for local optima. For each value of K, the run with the maximum likelihood was chosen for the final analysis of population structure and model selection.

4 Results and Discussions

Figure 3 shows the inferred population structure for the Pacific islander populations. Figure 4 shows the map of the Melanasian islands to enable us to understand the influence of geography and languages on population structure. In the figure, the stippled regions are Oceanic-speaking groups and Papuan-speaking regions have a grid or stripes. Papuan-speaking populations are also labeled in bold italics. The orange dots indicate inland groups while shore locations are yellow dots. The pies alongside each population label show the average ancestry proportion for that population, with 6 ancestral populations. We will first address the question of model selection and then analyze the population structure as inferred by models with different number of ancestral populations.

4.1 Model selection

The question of model selection occurs in most statistical analyses. In the case of population structure analyses, it takes the form of deciding the correct number of ancestral populations that gave rise to the modern populations we now observe. *mStruct* infers ancestry proportions for a given sample for a user-defined number of ancestral populations K. The question of what value of K is optimal must therefore be solved outside of the *mStruct* analysis. In this case, we use the Bayesian Information Criterion (BIC) [24] to decide the optimal number of ancestral populations. It is important to note, however, that we use the variational lower bound to the log-likelihood to compute BIC.

Figure 5 shows the plot of BIC vs K for values of K ranging from 2 to 10. We observe that the value of BIC is minimum for K=3. The best run for K=3 shows an ancestry map with three main components. The brown component denotes Baining ancestry. The green component is dominant in the Ata and Mamusi populations and the blue component is largely present in populations from New Ireland and Bougainville, together with a few populations from New Britain. This suggests that the population of the Melanasian islands was due to three groups of individuals in different islands. This is in agreement with the "Pause and Pulse" scenario of Pacific settlement described in Gray et al. [15], rather than the gradual drifting of a single group of individuals to settle all islands as in the "Slow Boat" scenario described by Kayser et al. [10].

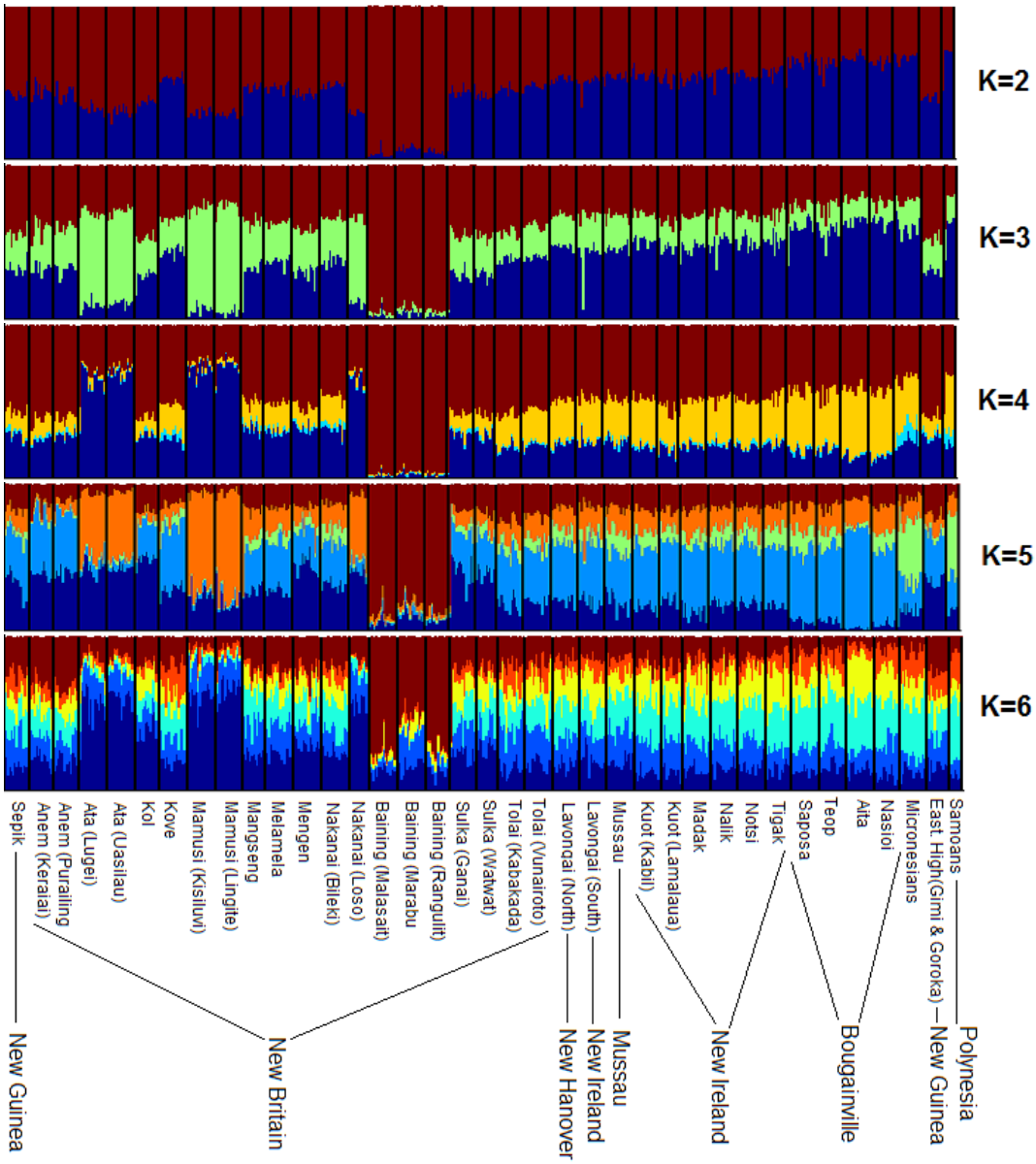


Figure 3: Population Structure of Pacific Islanders for K=2 to K=6, with the value of K indicated by the side. The black lines separate the populations from each other. Below each population name is the island it is located on. For values of K=7 and larger, the model quickly finds a local optimum and no noticeable population structure is inferred, so no results are presented.

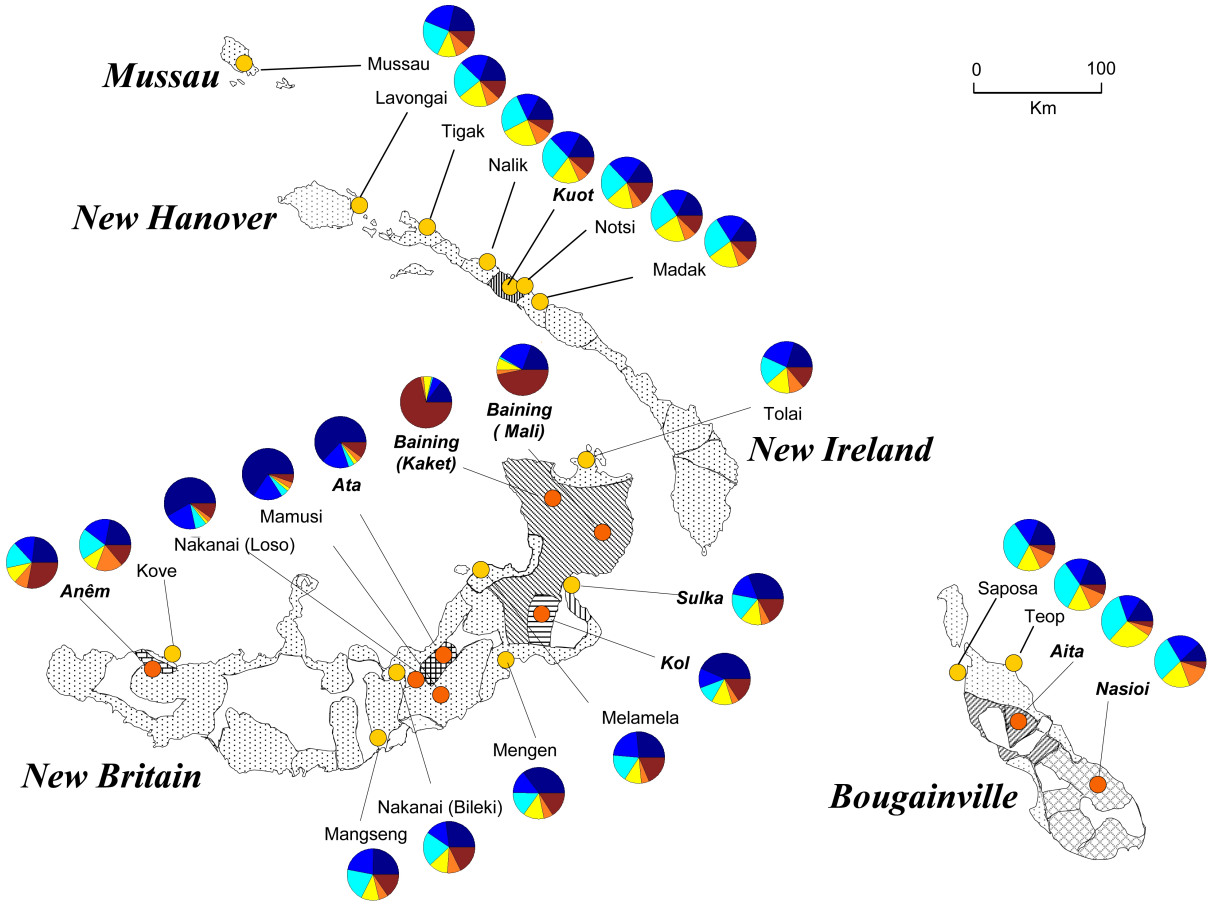


Figure 4: Map of the Melanasian Islands. The stippled regions are Oceanic-speaking; Papuan-speaking regions have a grid or stripes. Papuan-speaking populations are labeled in bold italics. The orange dots indicate inland groups while shore locations are yellow dots. The pies alongside each population label show the average ancestry proportion for that population, with 6 ancestral populations.

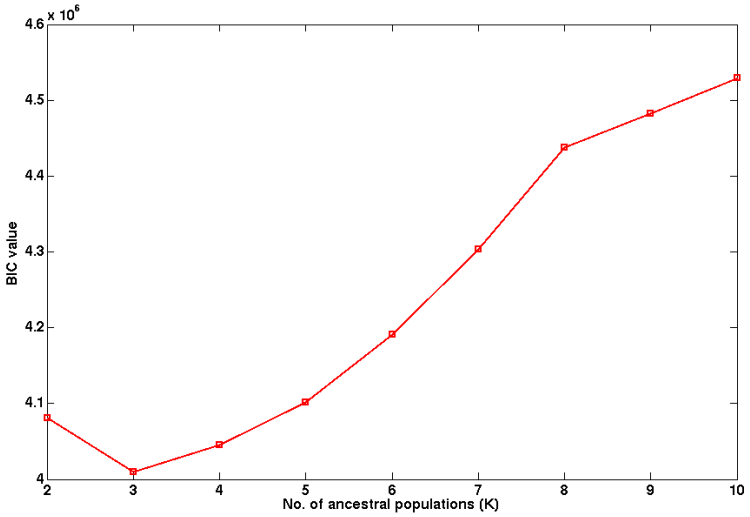


Figure 5: Bayesian Information Criterion (BIC) plot for different values of K, the number of ancestral populations. The value of K with minimum BIC is the best one according to the criterion.

However, as noted earlier, this conclusion is not a high-confidence one for two reasons - firstly that we use a variational lower bound as an approximation to the log-likelihood in computing BIC and secondly that the BIC criterion is a purely statistical tool for model choice which does not take any prior knowledge about anthropology, geography or biology into account. For analyzing population structure, we will not focus on just the model chosen by the BIC criterion, but use other models too since the models with larger value of K allow for easier visual interpretation. We shall show later that the claims we make about genetic diversity based on inferred ancestry vectors can also be quantitatively substantiated.

4.2 Population Structure

From the population structure graph in Figure 3, we can make various inferences about the genetic diversity based on the geography of the region and the languages in the region indicated in Figure 4. First, the most prominent signature of population structure visible in all the maps is a distinct population component for the Baining populations. As shown by heterozygosity analyses of the Pacific islanders in [13] and studies of mtDNA, X and Y chromosomes [28], the Baining populations are highly differentiated from the rest of the Pacific islander populations. The three Baining subpopulations are groups speaking different dialects, but their language (of Papuan origin) is not shared by another Pacific islander population nearby. As a result, there are no other populations genetically identical to the Baining but most Oceanic speaking populations on New Britain show a significant portion of Baining ancestry. This suggests that the

Baining populations might be one of the older populations of Melanesia.

Figure 3 contains the population labels for all the populations together with their island labels. When we examine the population structure at higher values of K (K=5 and 6), we see that the amount of genetic diversity in an island varies according to the size of the island. New Britain, which is the largest of the Melanasian Islands has the maximum genetic diversity in terms of the number of different profiles of ancestry proportions. However, this conclusion is based only on visual inspection. We can try to examine this statistically using the low-dimensional representation provided by the ancestry vectors. Table 1 shows the average between-population distance for each island, computing using a euclidean distance measure. The average between-population distance is one way of measuring the genetic diversity present within a set of population (where populations are defined by geographic labels). For different numbers of ancestral populations, we can see that the largest island, New Britain, has the largest average between-population distance. Bougainville has the second largest between-population distance (except when K=2, when the distance is larger for New Guinea). This agrees with the conclusions made by inspection of the population structure.

Island (No. of groups)/K→	2	3	4	5	6
Bougainville (4)	0.029	0.055	0.059	0.135	0.121
Micronesia (1)	-	-	-	-	-
Mussau (1)	-	-	-	-	-
New Britain (20)	0.226	0.383	0.376	0.469	0.344
New Guinea (2)	0.041	0.051	0.047	0.045	0.037
New Hanover (2)	0.020	0.049	0.040	0.038	0.062
New Ireland (6)	0.027	0.034	0.030	0.049	0.065
Polynesia (1)	-	-	-	-	-

Table 1: The variation of average between-population distance across islands for different numbers of ancestral populations. The numbers in parentheses next to the island names indicate the number of population groups present on each island.

Within islands, genetic profiles (in terms of the ancestry proportions) vary according to geography and language. This effect can be most prominently seen in the genetic profiles observed in the island of New Britain. In general, geographical proximity determines the amount of similarity in ancestry proportions. Within New Britain, the Oceanic-speaking Kove and Papuan-speaking Anem, which are geographically adjacent, have considerably similar genetic profiles despite their language differences. Of particular interest are the Ata and the Mamusi populations, which are genetically almost identical despite the fact that they speak different languages. This might be explained by the hypothesis that the Mamusi are originally a Papuan-speaking population who later adopted an Oceanian language [13]. Of the Nakanai populations, the Bileki subpopulation, which is geographically close to the Ata and the Mamusi but speaks a

different language, has a genetic profile similar to its neighbours while the genetic profile of the Loso subpopulation matches more closely that of the other Oceanic-speaking groups. This is not observed in the *STRUCTURE* analysis by Friedlaender et al. [13] but is in general agreement with the high correlation between geographic adjacency and genetic similarity.

The high correlation between geographic adjacency and genetic similarity is more evident in the smaller islands of New Ireland and Bougainville, where the population groups are not geographically far apart. Among the various populations on New Ireland we see only a gradual variation between genetic profiles due to their geographic adjacency. We also see that despite New Hannover being a distinct island, its proximity to New Ireland means that genetically its inabitants are quite similar to the population groups in New Ireland. A similar gradual change in genetic profiles is seen in the populations of the Bougainville island despite the presence of linguistic diversity on the island. Thus we see that both geography and language have an effect on population structure but geography has a much stronger footprint on genetic profiles than linguistic similarity.

However, the analysis does show certain caveats that must be taken into account in a study of population structure using *mStruct*. One possible problem is when one of the inferred ancestral populations out of K populations has very little contribution to the genetic profiles. This happened for the case of K=4 in our analysis in Figure 3.

Another problem we observe is that for values of K higher than 6, the *mStruct* model quickly falls into a local optimum and none of the 20 runs produce any detectable stratification. This phenomenon is likely due the the high degree of freedom the model is allowed in fitting its parameters. Since *mStruct* hypothesizes the presence of populations which are truly "ancestral" (in terms of having lesser genetic diversity than modern populations), it is not advisable to set the number of ancestral populations to a very high number. A possible tradeoff might be to allow only small amounts of mutation when setting the number of ancestral populations to a higher value. In this scenario, *mStruct* will infer ancestral populations which are relatively "young" and have genetic diversity intermdiate that of the truly ancestral populations and the modern populations. However, in the current experiements, we have set the prior on the mutation parameters so that that mutation parameters have non-zero values. Setting the mutation priors to encourage low values usually results in the *mStruct* model collapsing to *STRUCTURE* due to the construction of the model.

An interesting question that we were unable to address in this work due to scarcity of data is the migration of Polynesian populations through Melanasia. The version of data publicly available has only 11 Samoan individuals from Polynesia and any hypotheses made from such a small sample are likely to be inaccurate.

5 Conclusions and Future work

We analyzed data from 37 populations in the Pacific islands. Our analysis revealed considerable genetic diversity in the region. We also found that the diversity varied across islands depending on island size and geography. We also found that language had a small but noticeable effect on the population structure of the region while geography was a major determinant of genetic similarity.

With more data from the Polynesian populations, we might be able to address the question of migration of the Polynesian populations and their interaction with other populations in the Pacific. Geographical data regarding the populations will also help in answering this question using *mStruct*.

The hierarchical labeling of populations by regions and islands suggests that a more accurate representation of the populations would be in the form of a tree structure. Such a representation, while computationally expensive, would be more informative than current models, and would enable us to represent relationships between populations more naturally and accurately.

It is important to note, however, that graphical models in population genetics have been largely used as a tool for exploratory analyses of population data. They are usually validated on simulated population data due to lack of ground truth. Criterion such as perplexity or held-out likelihood, which are commonly used in statistics and machine learning, are usually not applicable to problems in which real genetic data is analyzed. Therefore, conclusions made from analyses using these models cannot be accepted as truth on their own, and must be supported by knowledge from history, archaeology and anthropology.

References

[1] M F Hammer, T Karafet, A Rasanayagam, E T Wood, T K Altheide, T Jenkins, R C Griffiths, A R Templeton, and S L Zegura. Out of africa and back again: nested cladistic analysis of human y chromosome variation. *Mol Biol Evol*, 15(4):427–441, Apr 1998.

[2] A.M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L. Cavalli-Sforza. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368 (6470):455–457, 1994.

[3] K. Roeder, M. Escoar, J.B. Kadane, and I. Balazs. Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika*, 85(2):269, 1998.

[4] C. HapMap. The International HapMap Project. *Nature*, 426(6968):789–96, 2003.

[5] L.L. Cavalli-Sforza. The human genome diversity project: past, present and future. *Nat Rev Genet*, 6(4):333–340, 2005.

[6] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. Genetic structure of human populations. *Science*, 298(5602):2381, 2002.

[7] H.M. Cann, C. de Toma, L. Cazes, M.F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, et al. A Human Genome Diversity Cell Line Panel. *Science*, 296(5566):261–262, 2002.

[8] S. Shringarpure and E. Xing. mStruct: Inference of Population Structure in Light of Both Genetic Admixing and Allele Mutations. *Genetics*, 2009.

[9] M.J. Pierson, R. Martinez-Arias, B.R. Holland, N.J. Gemmell, M.E. Hurles, and D. Penny. Deciphering past human population movements in Oceania: provably optimal trees of 127 mtDNA genomes. *Molecular biology and evolution*, 23(10):1966, 2006.

[10] M. Kayser, S. Brauer, R. Cordaux, A. Casto, O. Lao, L.A. Zhivotovsky, C. Moyse-Faurie, R.B. Rutledge, W. Schiefenhoevel, D. Gil, et al. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Molecular biology and evolution*, 23 (11):2234, 2006.

[11] L. Scheinfeldt, F. Friedlaender, J. Friedlaender, K. Latham, G. Koki, T. Karafet, M. Hammer, and J. Lorenz. Unexpected NRY chromosome variation in Northern Island Melanesia. *Molecular biology and evolution*, 23(8):1628, 2006.

[12] D.F. Conrad, M. Jakobsson, G. Coop, X. Wen, J.D. Wall, N.A. Rosenberg, and J.K. Pritchard. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature genetics*, 38(11):1251–1260, 2006.

[13] J.S. Friedlaender, F.R. Friedlaender, F.A. Reed, K.K. Kidd, J.R. Kidd, G.K. Chambers, R.A. Lea, J.H. Loo, G. Koki, J.A. Hodgson, et al. The genetic structure of Pacific Islanders. *PLoS Genet*, 4(1):e19, 2008.

[14] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 2000.

[15] RD Gray, AJ Drummond, and SJ Greenhill. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *science*, 323(5913):479, 2009.

[16] R. Kelly, M. Gibbs, A. Collick, and A.J. Jeffreys. Spontaneous Mutation at the Hypervariable Mouse Minisatellite Locus Ms6-hm: Flanking DNA Sequence and Analysis of Germline and Early Somatic Mutation Events. *Proceedings: Biological Sciences*, 245(1314):235–245, 1991.

[17] S.T. Henderson and T.D. Petes. Instability of simple sequence DNA in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 12(6):2749–2757, 1992.

[18] D.C. Queller, J.E. Strassmann, and C.R. Hughes. Microsatellites and kinship. *Trends in Ecology & Evolution*, 8(8):285–288, 1993.

[19] W. Dietrich, H. Katz, S.E. Lincoln, H.S. Shin, J. Friedman, N.L. Dracopoli, and E.S. Lander. A Genetic Map of the Mouse Suitable for Typing Intraspecific Crosses. *Genetics*, 131(2): 423–447, 1992.

[20] D. Pisani, L.L. Poling, M. Lyons-Weiler, and S.B. Hedges. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biology*, 2004.

[21] L.A. Zhivotovsky, P.A. Underhill, C. Cinnioglu, M. Kayser, B. Morar, T. Kivisild, R. Scozzari, F. Cruciani, G. Destro-bisol, G. Spedini, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *American journal of human genetics*, 74(1):50–61, 2004.

[22] L. Excoffier and G. Hamilton. Comment on Genetic Structure of Human Populations. *Science*, 300(5627):1877–1877, 2003.

[23] AM Valdes, M. Slatkin, and NB Freimer. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics*, 133(3):737, 1993.

[24] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[25] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999.

[26] E.P. Xing, M.I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. *Uncertainty in Artificial Intelligence (UAI2003). Morgan Kaufmann Publishers*, 2003.

[27] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(5):993–1022, 2003.

[28] JA Wilder and MF Hammer. Extraordinary population structure among the Baining of New Britain. *Genes, Language, and Culture History in the Southwest Pacific, edited by JS Friedlaender. Oxford University Press, Oxford*, pages 199–207, 2007.