Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 32

Estimation of Selection Intensity under Overdominance by Bayesian Methods

Erkan Ozge Buzbas[*] Paul Joyce[†]
Zaid Abdo[‡]

[*] University of Idaho, buzbas@gmail.com
[†] University of Idaho, joyce@uidaho.edu
[‡] University of Idaho, zabdo@uidaho.edu

Estimation of Selection Intensity under Overdominance by Bayesian Methods*

Erkan Ozge Buzbas, Paul Joyce, and Zaid Abdo

Abstract

A balanced pattern in the allele frequencies of polymorphic loci is a potential sign of selection, particularly of overdominance. Although this type of selection is of some interest in population genetics, there exists no likelihood based approaches specifically tailored to make inference on selection intensity. To fill this gap, we present Bayesian methods to estimate selection intensity under *k*-allele models with overdominance. Our model allows for an arbitrary number of loci and alleles within a locus. The neutral and selected variability within each locus are modeled with corresponding *k*-allele models. To estimate the posterior distribution of the mean selection intensity in a multilocus region, a hierarchical setup between loci is used. The methods are demonstrated with data at the Human Leukocyte Antigen loci from world-wide populations.

KEYWORDS: overdominance, heterozygote advantage, balancing selection *k*-allele models, Bayesian inference

*Erkan Ozge Buzbas is supported both by NSF Grant DEB-0515738 and NIH Grant P20 RR016454 from the INBRE Program of the National Center for Research Resources. Paul Joyce is supported in part by NSF Grant DEB-0515738. Zaid Abdo is supported by COBRE/NCRR 1-P20 RR016448.

1 Introduction

Selection reshapes patterns of variation in the genome leaving its signature on allele frequencies. Different selective mechanisms produce a variety of patterns. A balancing selection pattern arises when heterozygous genotypes are favored, a mechanism known as overdominance (see [Maruyama, 1981] and references therein for background). Contrary to mechanisms which work by eliminating the genetic variability, overdominance actively maintains allelic polymorphism in populations. Consequently, exceptional levels of polymorphism are expected to mark prototypical loci under overdominance. In this paper, methods to estimate selection intensity in such genomic regions are presented.

While mathematically well grounded frameworks have been advanced to model overdominance, equally well grounded statistical methods, linking the observed patterns of variability to the estimates are yet to be developed. This inference problem has some unique aspects. A notable one is that both elevated mutation rates and selection promote genetic diversity. This fact implies that allele frequencies reshaped by both processes do not carry distinctive information about the respective parameters. In statistical terms, parameters representing mutation and selection are unidentifiable. Another point arises in estimation. A relatively detailed and well understood class of population genetic models that can accommodate overdominance is k-allele models with selection [Watterson, 1977, Wright, 1949]. Nevertheless, it has been shown that the maximum likelihood estimates cannot be reliably coupled with bootstrap to assess the error of estimates under k-allele models with selection [Buzbas and Joyce, 2009]. Intensive resampling of the allele frequency space creates a numerical instability, which causes the estimates to be both unreliable and inaccurate. Therefore, obtaining good interval estimates is a challenge. Further, polymorphic systems may span a number of genetic loci, sometimes with large number of alleles. This makes the scalability of computational methods an issue. To our knowledge, there exists no likelihood based methods that can handle data from multiple polymorphic loci to make inference on the strength of overdominance. Our main contribution is to provide such methods which overcome all the aforementioned problems.

We alleviate the problem of identifiability by defining two classes of “alleles” which capture two types of variation, “neutral” versus “selected”. We build two classes of k-allele models. One to identify plausible mutation rates using the neutral variation. Another to use this information and the selected variation to recover the signal due to overdominance. We solve the instability issue in estimation by taking a Bayesian view. Since posterior inference fixes the data and searches only the parameter space, it avoids pitfalls arising due to resampling of the data space. Our model can accommodate arbitrary number of loci and alleles. This flexibility allows us to obtain estimates of mean selection intensity for groups of loci using a hierarchical model setting.

As a real data application we consider the polymorphism in the Major Histocompatibility Complex (MHC) region of vertebrates. In humans, each major MHC locus has sufficient variability to be a good candidate for overdominance. However, there is extensive functional similarity and cooperation of molecular products encoded by different MHC loci. In such systems, handling information from the whole group of loci is a reasonable first approximation for an assessment of the intensity of selection in the region. Methods are demonstrated using Human Leukocyte Antigen data from world-wide populations. Data sets published in the Proceedings of the 13th International Histocompatibility Workshop and Conference (see [Meyer *et al.*, 2007] and references therein) are analyzed across loci for signals of overdominance.

2 Model

The k-allele model with symmetric overdominance is described using a Wright-Fisher population (see [Donnelly and Kurtz, 1999, Ethier and Griffiths, 1987, Ethier and Kurtz, 1993, Ewens, 2004] for background). It assumes a panmictic population of N diploid individuals at a non recombinant locus with non overlapping generations. There are k possible alleles. Each generation $2N$ genes are randomly paired to form N gene pairs or genotypes. Thus, (assuming large N) the genotype frequency of A_mA_l will be $2x_mx_l$. The genotypes are then sampled to form the next generation. The probability of sampling a genotype is proportional to its fitness, $w_{ml} = 1 + s_{ml}$, with $s_{ml} = 0$ if $m \neq l$ and $s_{ml} = -s$, $(s > 0)$ otherwise. A randomly chosen allele within each sampled genotype is subjected to mutation with probability u before it is included into the next generation's allele pool. This process is Markovian and there exists a stationary distribution of the allele frequencies, $\mathbf{x} = [x_1 \ldots x_k]$, given by

$$(1) \qquad f(\mathbf{x}|\theta,\sigma) = \frac{e^{-\sigma\sum_{i=1}^{k} x_i^2}}{c(\theta,\sigma)} \prod_{i=1}^{k} x_i^{(\theta/k-1)}$$

where $\theta = 4Nu$, $\sigma = 2Ns$ are mutation and selection parameters and

$$(2) \qquad c(\theta,\sigma) = \frac{\Gamma(\theta/k)^k}{\Gamma(\theta)} \int \cdots \int e^{-\sigma\sum_{i=1}^{k} x_i^2} f(\mathbf{x}|\theta) d\mathbf{x}$$

is the normalizing constant. Efficient numerical methods to compute $c(\theta,\sigma)$ are given in [Genz and Joyce, 2003] and [Joyce *et al.*, 2009]. Here, $f(\mathbf{x}|\theta)$ is the

DOI: 10.2202/1544-6115.1466

stationary distribution under neutrality which is appropriate when all genotypes have equal fitness. It can be obtained as a special case of equation 1 by setting $\sigma = 0$, which gives

$$f(\mathbf{x}|\theta) = \frac{\Gamma(\theta)}{\Gamma(\theta/k)^k}\prod_{i=1}^{k} x_i^{(\theta/k-1)}. \tag{3}$$

In the next two subsections, we describe our approach to jointly estimating θ, σ using both equations 1 and 3, with selected and neutral variability respectively.

2.1 Allelic Variability

The data from single locus are summarized in Table 1. The first column identifies selectively distinct alleles, k of them in total. These alleles differ by a *non synonymous* mutation at least at one site in their sequence and constitute the "selected variation". Each line in the second column gives the frequency vector of neutral variants, for the corresponding selectively distinct allele. Elements of a vector differ from each other by *synonymous* substitutions only. For example, the j_i neutral variants associated with the i^{th} allele in the first column are denoted by $[x_{i1} \dots x_{ij_i}]$. These alleles are subject to the same selection and thus differ from each other by a "neutral" substitution. The third column gives the frequency of selected alleles which are then collected in the vector $\mathbf{x} = [x_1\, x_2 \dots x_k]$. Finally, in the last column are the normalized frequencies of neutral variants, to be used with the neutral model. These are collected in the vector of vectors $\mathbf{Y}$.

As we justify in section 5, if only $\mathbf{x}$ is available, θ and σ are statistically unidentifiable. This problem can be circumvented however, by observing that variation encapsulated in $\mathbf{Y}$ is reshaped by mutation only and it can be used to extract information about θ. The variation in $\mathbf{x}$ on the other hand, can be used to extract information about both parameters. The two types of data, $\mathbf{x}$ and $\mathbf{Y}$ are modeled as follows.

2.2 Single Locus Model

The neutral variation does not affect fitness, hence it is subject to equation 3. Assuming a constant mutation rate within a locus, Appendix A shows that the joint likelihood of the allele frequencies can be written as

$$P(\mathbf{Y}|\theta) \propto \prod_{i=1}^{k} f(\mathbf{y}_i|\theta). \tag{4}$$

Table 1: Partitioning the allelic variability at a locus.

Number of selected alleles	Frequency of neutral variants	Frequency of selected alleles	Normalized frequency of neutral variants
1	$[x_{11} \dots x_{1j_1}]$	$x_1 = \sum_{i=1}^{j_1} x_{1i}$	$\mathbf{y}_1 = [x_{11} \dots x_{1j_1}]/x_1$
2	$[x_{21} \dots x_{2j_2}]$	$x_2 = \sum_{i=1}^{j_2} x_{2i}$	$\mathbf{y}_2 = [x_{21} \dots x_{2j_2}]/x_2$
.	.	.	.
.	.	.	.
.	.	.	.
k	$[x_{k1} \dots x_{kj_k}]$	$x_k = \sum_{i=1}^{j_k} x_{ki}$	$\mathbf{y}_k = [x_{k1} \dots x_{kj_k}]/x_k$

Vector of selected allele frequencies $\mathbf{x} = [x_1 \; x_2 \dots x_k]$.
Vector of vectors for neutral allele frequencies $\mathbf{Y} = [\mathbf{y}_1 \; \mathbf{y}_2 \dots \mathbf{y}_k]$.

Using Bayes' rule we obtain

$$P(\theta|\mathbf{Y}) \propto P(\mathbf{Y}|\theta)P(\theta), \tag{5}$$

where $P(\theta)$ is a prior for the mutation parameter.

On the other hand, the selected variation results from fitness differences and under overdominance it follows equation 1. Denote the likelihood by $P(\mathbf{x}|\theta,\sigma)$, which is the density in equation 1 evaluated at the data $\mathbf{x}$ as a function of θ,σ. Assuming the prior independence of θ,σ and using Bayes' rule we link neutral and selected models by

$$P(\theta,\sigma|\mathbf{x},\mathbf{Y}) \propto P(\mathbf{x}|\theta,\sigma)P(\theta|\mathbf{Y})P(\sigma) \tag{6}$$

where $P(\sigma)$ is prior for the selection parameter. In this way, the information obtained using the density in equation 3 and the data $\mathbf{Y}$ are fused with that from equation 1 and $\mathbf{x}$. In practice, we first estimate the posterior distribution of the mutation parameter, $P(\theta|\mathbf{Y}) = P(\mathbf{Y}|\theta)P(\theta)$, then use this information to obtain the posterior in equation 6. The second analysis is still joint estimation, that is, our view of the mutation parameter from the first step is not fixed but subject to re-evaluation using the new evidence. Throughout we use proper uniform priors with diffuse bounds both for θ and σ.

2.3 Extension to Multiple Loci

To extend the single locus model to multiple loci we adopt a hierarchical normal model as a robust choice (see [Bustamante *et al.*, 2002] for an application in population genetics and [Gelman *et al.*, 2004] pp. 74-77 for a detailed treatment in general settings). Given $i = 1, 2, ..., m$ loci, we assume that selection parameters σ_i are normally distributed with group specific common mean μ and variance τ. Conditional on the normal distribution, the process at each locus is identical, corresponding to a (conditionally) independent Wright-Fisher population. From the Bayesian perspective, the normality has the interpretation of a prior on the selection parameters. Next, we derive posterior distributions of the mutation and selection parameters.

There are m conditionally independent posteriors, one for each selection parameter. Let $\mathbf{x}_i$ and $\mathbf{Y}_i$ denote the selected and the neutral frequencies for i^{th} locus. Appendix B shows that the conditional distribution of σ_i and θ_i are given respectively by

$$P(\sigma_i|\mathbf{x}_i, \theta_i, \mu, \tau) \propto P(\mathbf{x}_i|\theta_i, \sigma_i)P(\sigma_i|\mu, \tau), \tag{7}$$

$$P(\theta_i|\sigma_i, \mathbf{x}_i, \mathbf{Y}_i) \propto P(\mathbf{x}_i|\theta_i, \sigma_i)P(\mathbf{Y}_i|\theta_i)P(\theta_i). \tag{8}$$

We assign diffuse priors to hyperparameters. We use a uniform distribution on μ and an inverse chi-squared distribution on τ, which is conjugate for the normal model variance [Gelman *et al.*, 2004]. We write, $\tau \sim Inv - \chi^2(\nu_0, \tau_0)$ where $Inv - \chi^2$ denotes an inverse χ^2 distribution. Prior parameters are chosen so that this distribution is uninformative (i.e., large τ_0 and small ν_0). The joint posterior for (μ, τ) is given by

$$P(\mu, \tau|\boldsymbol{\sigma}) \propto \frac{1}{\tau^{(m+\nu_0+2)/2}} e^{-\frac{\Sigma_{i=1}^{m}(\sigma_i-\mu)^2-\nu_0\tau_0}{2\tau}}, \tag{9}$$

where $\boldsymbol{\sigma} = [\sigma_1 \ldots \sigma_m]$. The total number of parameters to be estimated is $2m + 2$, mutation and selection parameters for $2m$ loci in addition to μ, τ. In particular, our interest lies with μ, the group specific mean selection parameter.

The above treatment of the normal hierarchical model is a common one. However, the hard to compute integration constant in the stationary distribution of the allele frequencies creates difficulties in sampling the target posterior distributions. Below we present algorithms to accomplish this task.

3 Methods

3.1 Single Locus

The posterior distribution of θ under the neutral model and the joint posterior distribution of (θ, σ) under overdominance are obtained with the following (Metropolis-Hastings) algorithms respectively [Hastings, 1970, Metropolis *et al.*, 1953].

Algorithm 1

1. Start with an arbitrary initial value, $\theta^{(0)}$.
2. Generate θ^* independently, from $\theta^* \sim \text{Unif}(0, \theta_{\max})$ where $\theta_{\max}$ is a fixed constant.
3. Generate $U \sim \text{Unif}(0,1)$.
4. Set $\theta^{(1)} = \theta^*$ with probability $\alpha = \min\left\{1, \prod_{i=1}^{k} \frac{f(\mathbf{y}_i|\theta^*)}{f(\mathbf{y}_i|\theta^{(0)})}\right\}$.
5. Iterate from step 2.

Algorithm 2

1. Start with arbitrary initial values $(\sigma^{(0)}, \theta^{(0)})$.
2. Generate (θ^*, σ^*) independently, from $\theta^* \sim P(\theta|\mathbf{Y})$ and $\sigma^* \sim \text{Unif}(-\sigma_{\max}, \sigma_{\max})$ where $\sigma_{\max} > 0$ is a fixed constant.
3. Generate $U \sim \text{Unif}(0,1)$.
4. Set $\sigma^{(1)} = \sigma^*,\ \theta^{(1)} = \theta^*$ with probability $\alpha = \min\left\{1, \frac{f(\sigma^*, \theta^*|\mathbf{x}, \mathbf{Y})}{f(\sigma^{(0)}, \theta^{(0)}|\mathbf{x}, \mathbf{Y})}\right\}$.
5. Iterate from step 2.

If diffuse limits for $\theta_{\max}$ and $\sigma_{\max}$ are used, the priors will be uninformative.

3.2 Multiple Loci

Under the hierarchical model of section 2.3, a relatively easy strategy to simulate from the joint posterior distribution of the parameters is as follows [Bustamante *et al.*, 2002, Gelman *et al.*, 2004]:

1. Simulate τ from its marginal posterior distribution and μ from its conditional distribution given τ.
2. Given the values of the hyperparameters and the data, generate the selection and mutation parameters for each locus from their conditional distributions respectively.

To perform Step 1 we exploit

$$P(\mu,\tau|\boldsymbol{\sigma}) = P(\mu|\boldsymbol{\sigma},\tau)P(\tau|\boldsymbol{\sigma}). \tag{10}$$

The posterior distribution of the mean selection parameter conditional on τ is given by $(\mu|\tau,\boldsymbol{\sigma}) \sim N(\bar{\sigma},\tau/m)$, where $\bar{\sigma}$ denotes the mean of selection parameters and m is the number of loci. The marginal distribution of the variance is given by

$$(\tau|\boldsymbol{\sigma}) \sim Inv-\chi^2(\nu_0+m, \frac{\nu_0\tau_0+(m-1)s_{\boldsymbol{\sigma}}}{\nu_0+m}),$$

where $s_{\boldsymbol{\sigma}}$ is the sample variance of selection parameters.

The posterior distributions of θ and σ to be sampled in Step 2 do not have familiar forms and simulating directly from these conditionals is not possible with standard methods. One way to sample them is via the inverse method using empirical cumulative distribution functions evaluated on a grid, but this is computationally expensive under k-allele models. In the rest of this section we describe efficient methods to sample from these distributions. These methods are embedded in a Gibbs sampler including all the parameters and hyperparameters.

We start with the selection parameter. Using equation 7 without subscripts for notational convenience, the conditional distribution of σ can be written as

$$P(\sigma|\mathbf{x},\theta,\mu,\tau) \propto \frac{e^{-\sigma F}G^{(\theta/k-1)}}{c(\theta,\sigma)}e^{-(\sigma-\mu)^2/2\tau} \tag{11}$$

where $F=\sum_{i=1}^{k} x_i^2, G=\prod_{i=1}^{k} x_i$. Completing the square and collecting the exponential terms we get

$$P(\sigma|\mathbf{x},\theta,\mu,\tau) \propto \left[\frac{1}{\sqrt{2\pi\tau}}e^{-\frac{(\sigma-(\mu-\tau F))^2}{2\tau}}\right]\left[\frac{1}{c(\theta,\sigma)}\right]. \tag{12}$$

The first term on the right is the familiar normal density with mean $\mu-\tau F$ and variance τ. The second term is a normalizing constant from the stationary distribution under selection which we know how to compute numerically [Genz and Joyce, 2003, Joyce *et al.*, 2009]. To sample the density in equation 12 we use a result due to Damien et al. [Damien *et al.*, 1999]. Put in our context, the result states that if $c(\theta,\sigma)$ is invertible and non-negative, then there exists a Gibbs sampler for $P(\sigma|\mathbf{x},\theta,\mu,\tau)$. The condition is satisfied since $c(\theta,\sigma)>0$ for all θ,σ and it follows from equation 2 that $c(\theta,\sigma)$ is a decreasing function in σ for fixed θ. To build the Gibbs sampler, an auxiliary variable U is introduced such that the conditional distribution $U|\sigma$ is uniform on $(0,c(\sigma,\theta)^{-1})$. On the other hand, the conditional distribution $\sigma|U$ is the normal distribution given in equation 12, restricted to the set

$B_\sigma = \{\sigma : c(\theta,\sigma)^{-1} > u\}$. Hence, the conditional of σ is a truncated normal with truncation point updated at each iteration of the Gibbs sampler. The truncation is from the left and for strong selection the set B_σ consists of large values of σ. This involves drawing from extreme tails of a normal distribution, a difficult feat using the standard inverse method due to the cumulative distribution function approaching to unity. We circumvent this problem using an accept-reject algorithm that is efficient for hard to draw values from truncated normal distributions [Geweke, 1991]. The choice of the instrumental distribution depends on the truncation point. For the hard to draw region described above a truncated exponential distribution with rate $\min(B_\sigma)$ on $\sigma > \min(B_\sigma)$ is used, where $\min(B_\sigma)$ is the truncation point.

The posterior distribution of the mutation parameter, θ, is given by

$$P(\theta|\mathbf{x},\mathbf{Y},\sigma) \propto \frac{G^{(\theta/k-1)}}{c(\theta,\sigma)} P(\mathbf{Y}|\theta)P(\theta). \tag{13}$$

To obtain a workable form we fit a gamma distribution (with parameters λ,α), to $P(\theta|\mathbf{Y}) = P(\mathbf{Y}|\theta)P(\theta)$. After some algebra we get

$$P(\theta|\mathbf{x},\sigma,\lambda,\alpha) \propto \left[\frac{(\lambda - \log G/k)^\alpha \theta^{\alpha-1} e^{-(\lambda-\log G/k)\theta}}{\Gamma(\alpha)}\right] \left[\frac{1}{c(\theta,\sigma)}\right]. \tag{14}$$

The first term on the right is a gamma density with parameters $\lambda - \log G/k$ and α, whereas the second term is again the normalizing constant of the density under selection. Similar to the case of σ, one way to sample the posterior distribution of θ is by first finding the restriction set for θ and then drawing from the truncated version of the gamma given by the first term of equation 14. There exist efficient accept-reject algorithms to sample a truncated gamma density to obtain draws in this way [Dagpunar, 1978, Phillippe, 1997]. Here we opt for an alternative, the inverse cumulative distribution function method coupled with a truncated exponential density [Damien and Walker, 2001]. There is little difference between the two methods from computational point of view. We let $B_\theta = \{\theta : \theta^{\alpha-1}c(\theta,\sigma)^{-1} > u\}$. Now, generating from the posterior distribution of θ is equivalent to generate from $P(\theta|\lambda,\mathbf{x}) = (\lambda - \log G/k)e^{-(\lambda-\log G/k)\theta} I(\theta > \min(B_\theta))$, which is a truncated exponential distribution with parameter $(\lambda - \log G/k)$. We use the inverse method to generate θ by

$$\theta = -\log(U/(\lambda - \log G/k))/(\lambda - \log G/k) + \min(B_\theta),$$

where $U \sim \text{Unif}(0,1)$.

To sample posterior distributions of all the parameters and hyperparameters we setup a Gibbs sampler as follows.

Algorithm 3

1. Start with initial values for the parameter vector

$$\mu^{(0)}, \tau^{(0)}, \boldsymbol{\sigma}^{(0)} = [\sigma_1^{(0)}, \cdots, \sigma_k^{(0)}], \boldsymbol{\theta}^{(0)} = [\theta_1^{(0)}, \cdots, \theta_k^{(0)}].$$

2. Iterate for all i :

   (a) Draw $u = U$ from $\text{Unif}(0,1)$ and find the set

   $$B_\sigma = \{\sigma : c(\theta_i^{(0)}, \sigma)^{-1} > uc(\theta_i^{(0)}, \sigma_i^{(0)})^{-1}\}.$$

   (b) Sample $\sigma_i^{(1)} \sim TN\left((\mu^{(0)} - \tau^{(0)} F_i), \tau^{(0)}\right)$ where TN denotes a truncated normal distribution with truncation point given by $\min(B_\sigma)$.

   (c) Draw $u = U$ from $\text{Unif}(0,1)$ and find the set

   $$B_\theta = \{\theta : \theta_i^{\alpha-1} c(\theta_i, \sigma_i^{(0)})^{-1} > u(\theta_i^{(0)})^{\alpha-1} c(\theta_i^{(0)}, \sigma_i^{(0)})^{-1}\}.$$

   (d) Draw $u = U$ from $\text{Unif}(0,1)$ and find $\theta_i^{(1)} = -\log(u/(\lambda - \log G/k))/(\lambda - \log G/k) + \min(B_\theta)$.

3. Given $\boldsymbol{\sigma}^{(1)}$, sample from $\tau \sim Inv\chi^2(m-1, s_{\boldsymbol{\sigma}^{(1)}})$.

4. Given $\tau^{(1)}$, sample from $\mu \sim N(\bar{\sigma}^{(1)}, \tau^{(1)}/m)$, where $\bar{\sigma}^{(1)}$ is the sample mean of $\boldsymbol{\sigma}^{(1)}$.

5. Iterate from Step 2.

Before moving to specific examples let us recapitulate the above procedures. Given a set of loci, considered as a group for purposes of estimating the intensity of selection under overdominance, we adopt the following strategy.

1. Identify neutral and selected frequencies at each locus as defined in table 1.

2. Use neutral variation and *Algorithm 1* to construct a prior view of the mutation parameter at each locus.

3. Use this prior, selected variation and *Algorithm 3* to sample the posterior distribution for the mean and variance of the selection parameter for each group.

4 Method Validation and Examples

4.1 Simulations

The focus of our simulations is two fold. First, we explore the effect of the distribution of information among loci on the estimates of μ. Second, we assess the amount of data required to obtain reasonable error bounds on μ under appreciable selection. In the same context, we also analyze the simulated data with fixed θ, to assess the quality of joint estimation with respect to the known mutation parameter case.

Consider data sets with

$$(m,k) = \{(3,32)\ ,\ (4,24)\ ,\ (6,16)\ ,\ (8,12)\ ,\ (12,8)\ ,\ (16,6)\ ,\ (24,4)\ ,\ (32,3)\}$$

all generated under $\sigma = 100$. Each data set has 96 allele frequencies in total, however, the organization of information is quite different. We simulated thirty replicates under each parameter combination and obtained posterior distributions as explained in *Algorithm 3*. Credible intervals for μ show that data sets with fewer alleles distributed in many loci yield smaller variance estimates in comparison to those with large number of alleles distributed over a few loci (figure 1). This result is not totally unexpected and it can be interpreted as a realization of the fact that data from different loci are treated as (conditionally) independent. In other words, an allele at a new locus has more information (due to independence) than an additional allele within a locus (where the frequencies are correlated).

As can be deduced from the analysis just presented, single locus data are not expected to provide precise estimates. Note that, a multivariate **x** is actually sample of size 1 for each locus. Combined information from multiple loci on the other hand, is expected to improve the precision considerably. Accordingly, we now fix k and turn to answer how many loci yield a reasonable precision on the mean selection parameter. The data have $m = 5,\ 10,\ 20,\ 25,\ 30,\ k = 10$ and $\sigma = 100$. The improvement in precision with the number of loci are illustrated as the mean of thirty replicates in terms of 95% credible intervals and coefficient of variation in μ. A comparison of these intervals with corresponding intervals from the analysis with fixed θ show very little difference (figure 2). Hence, prior distributions of θ obtained using neutral variation are pretty informative and do an excellent job in recovering information about individual mutation parameters.

4.2 Human Leukocyte Antigen

As an application of the methods we consider data from the Human Leukocyte Antigen (HLA) loci, a most intensively studied region of the human genome. HLA

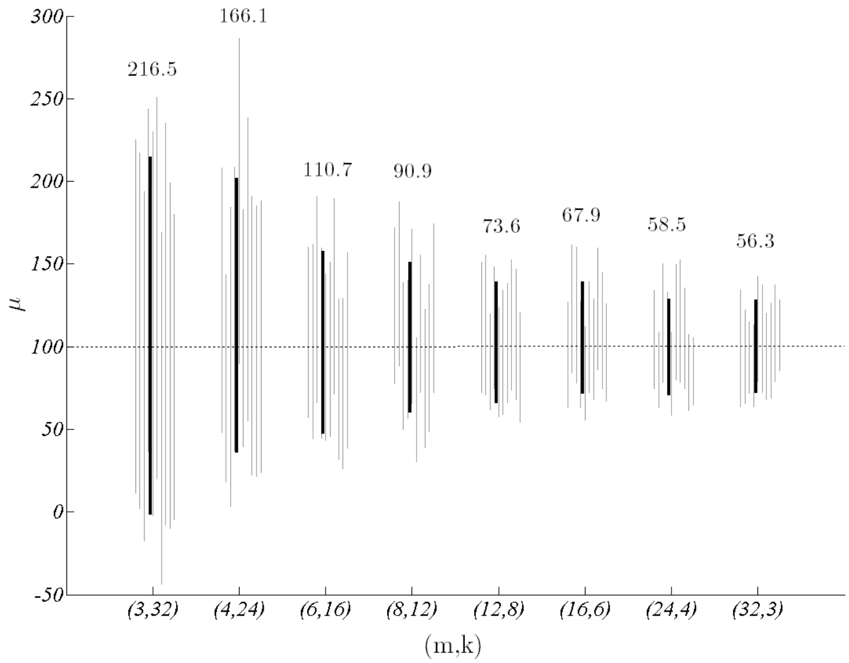

Figure 1: 95% credible intervals of μ for data sets with $m = 3$, 4, 6, 8, 12, 16, 24, 32 loci and $k = 32$, 24, 16, 12, 8, 6, 4, 3 alleles respectively, all generated under $\sigma = 100$. For each parameter vector the mean credible interval based on thirty replicates (black) and ten individual intervals are shown (shades). Interval lengths for means are given on top.

genes code for molecular products that regulate and control immune system functions. High levels of genetic variability is observed in these genes. Complex adaptive processes that resulted in diversification of functional regions such as HLA are yet to be resolved [Black and Hedrick, 1997] and the astounding[1] variability in these regions has many implications. A widely stated hypothesis is that the high genetic variability at HLA is a result of overdominance. The molecular products coded by HLA help detect foreign agents such as pathogens, bacteria, virus etc. Briefly, these molecular products are attached on the cell surface and they either remain inactive if they recognize an agent as “self” (i.e., produced by the body itself) or signal to the immune system if they recognize it as “non-self”. Higher genetic variability is promoted since the production of different molecules gives an opportunity to recognize a wider range of non-self pathogenic agents. Therefore, heterozygous individuals are hypothesized to have a selective advantage over homozygous ones.

[1] As of May 2008, 2128 Class I, 954 Class II HLA alleles [Robinson *et al.*, 2003]. IMGT/HLA database reports

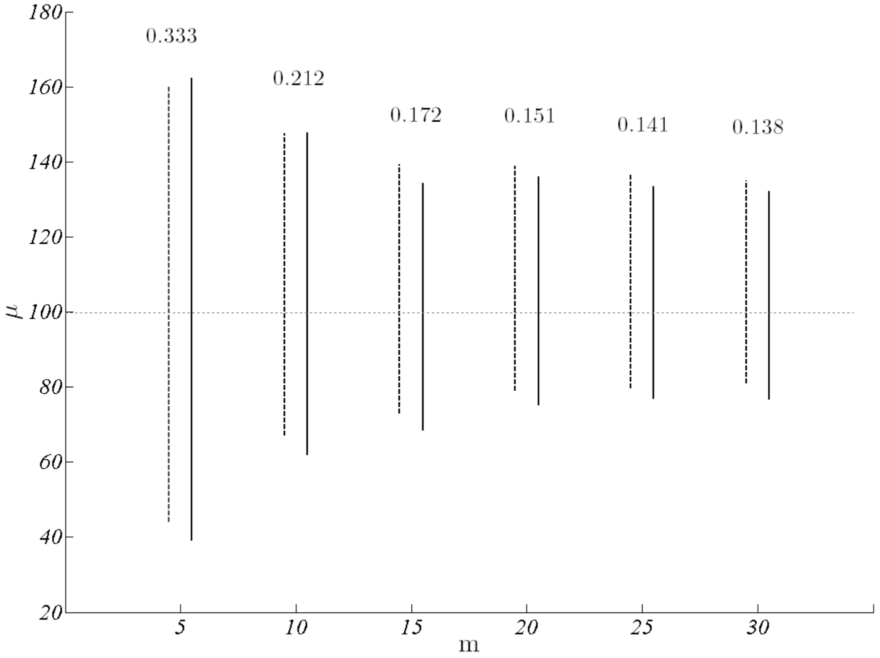

Figure 2: The decrease in 95% credible intervals of μ with m. For all data $k = 10$ and $\sigma = 100$. Interval estimates from joint estimation (solid lines) and fixed θ estimation (dotted lines) show little difference. Coefficient of variation for each μ from joint estimation is given on top of the corresponding interval.

There exist analyses of single locus HLA data sets in the literature aiming at estimating the selection intensity from allele frequency data. For example, Muirhead and Slatkin [2000] analyzed large data sets of HLA from world wide populations for signals of balancing selection. Their model is essentially the same as the single locus model presented in this paper, however, they used simpler estimators of selection intensity since current methods and computational power were not available at that time. Here, we use some of the data on HLA loci provided by the Proceedings of the 13^{th} International Histocompatibility Workshop and Conference [Meyer *et al.*, 2007] to illustrate our methods. We use data from three Class I loci (i.e., A, B and C) and two Class II loci (i.e., DRB and DQB) consisting of three geographical populations: European, South American and Sub-Sahara African. The most polymorphic of these is the Sub-Sahara African population, with an exceptional variability both in neutral and selected variation, whereas the least polymorphic is the South American population (Table 2). We analyze the data using the hierarchical model with the goal of estimating the mean selection intensity in the HLA region of

DOI: 10.2202/1544-6115.1466

Table 2: The HLA data (modified from 13^{th} Histocompatibility workshop proceedings [Meyer *et al.*, 2007]). Only loci that provide sufficient variation at synonymous level are used. Populations/Loci not used for the analysis are indicated by (*). The first figure in each cell is the number of serologically differing alleles (selected variability). The neutral variation is given parenthetically: a^n denotes that there are n groups with a neutral alleles each.

Pop./ Locus	A	B	C	DRB	DQB
European	17,(2,3)	28,$(2^4,3)$	13,$(2^4,3)$	13,(2^7)	5,(2^2)
South American	9,(2)	10,(2^3)	7,(2^3)	10,(2)	*
Sub-Sahara African	21,$(2^2,3^2)$	30,$(2^5,3^2)$	14,$(2^5,3,4)$	13,$(2^8,3)$	5,$(2^5,3)$

the genome, for each of these populations. Since there exists sufficient variability at serotype groups, we assume that selection acts at the antigen level and identify the selected alleles accordingly. High variability in the neutral data for each locus provided informative prior distributions of θ. Posterior samples of μ for European and Sub-Sahara African populations (figure 3), indicate that overdominance might be a plausible hypothesis for the HLA loci in these populations, since 95% credible intervals do not include zero, the neutral case. On the other hand, South American population frequencies are not inconsistent with neutrality. For this population, one would fail to conclude a signal for overdominance. Note that this result is consistent with our observation that this is the least polymorphic of the three populations.

5 Discussion

In this paper, we presented an overdominance model that can accommodate data from multiple loci with multiple alleles and likelihood based methods to estimate the selection intensity under this model. Our methods use two types of genetic variability: neutral and selected. The necessity of neutral variability as part of the data has been mentioned several times hitherto. Its crucial role is to restrict the mutation parameter to a range consistent with the data such that unrealistically large values are not considered consistent with the data. If only selected variation is used, large mutation rates are able to account for the variation actually produced by selection. Consequently, selection intensities would be biased towards lower values leading to erroneous inference. In fact, under the k-allele model, mutation and selection parameters turn out to be statistically unidentifiable if the data consist only of allele frequencies reshaped by both forces (figure 4). Therefore, some external data limiting the mutation rates are necessary for meaningful inference.

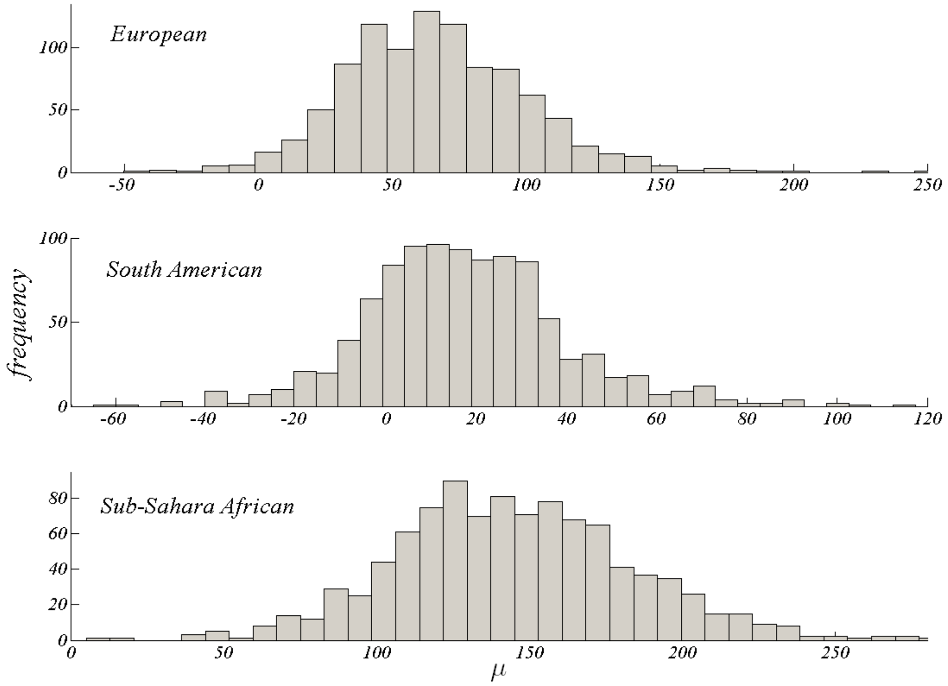

Figure 3: Posterior samples for the mean selection intensity, μ, under selective overdominance model for European, South American and Sub-Sahara African populations.

A large number of loci might not always be available for the biological system of interest. For example the HLA system has six major Class I and Class II genes. Data consisting of 32 loci such as ones considered in the simulations seem overly optimistic. When the number of loci is small, interval estimates can be too wide to be useful even if there is appreciable signal for selection.

Importantly, when m, k (i.e., the number of loci and allele frequencies in the data) are small to moderate, the estimates of σ and μ will be sensitive to the prior parameter τ_0 for the variance. This effect is due to the fact that τ is a parameter in the posterior distributions of σ and μ. In such cases, τ_0 should be chosen carefully to minimize its effect on inference. For the conjugate prior considered above, the $Inv-\chi^2$ prior is uninformative when τ_0 is large. However, for a given pair of m, k, a too large τ_0 introduces a negative bias in the estimates, whereas a too small τ_0 creates a positive bias. To minimize the effect in either direction, τ_0 can be optimized based on the number of loci and allele frequencies for the loci of interest. Note that, these quantities are constants in the model and not part of the data, legitimizing their role to optimize the prior parameter. An optimal τ_0 for a given m, k pair has on average a minimal biasing effect. A good way to determine such a τ_0 is as follows.

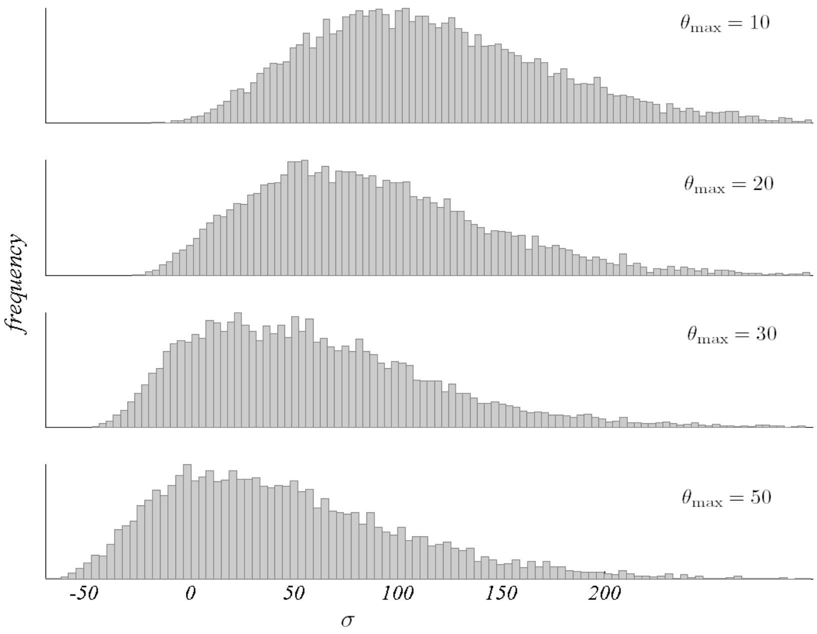

Figure 4: The effect of θ, on the posterior distribution of σ illustrating the identifiability problem.

First, we generate a large sample of data sets with the m and k of the actual data. (The actual value of μ under which the data are generated, has a negligible effect on the bias.) We choose an arbitrary τ_0 and obtain the posterior distribution of μ. If the mode is above the true value, we decrease, otherwise increase τ_0 and reiterate until an acceptable level of bias is achieved. The effect of prior parameter τ_0 on the estimates is shown by mean estimates of μ, based on 10 independent simulated data sets for each m, k combination and a range of τ_0 values (figure 5 and 6).

A certain amount of genetic linkage is expected in a set of loci located on the same chromosome. Linkage affects a multi locus system by creating correlation in the allele frequencies between loci. In this case, the joint distribution of data across loci currently given by equations 18 through 21 in Appendix A, would no longer be the product of the likelihoods, but would be more complex. To derive such a likelihood one would have to know specific information concerning the linkage between loci. However, if these adjustments are possible, the rest of the hierarchical Bayesian setup would remain unchanged. Yet, this strategy is not useful unless a realistic correlation structure about the system is available. However, if locus i and locus j are linked, one would expect the estimated selection coefficients to be closer than if they were unlinked. From an estimation perspective, the consequence

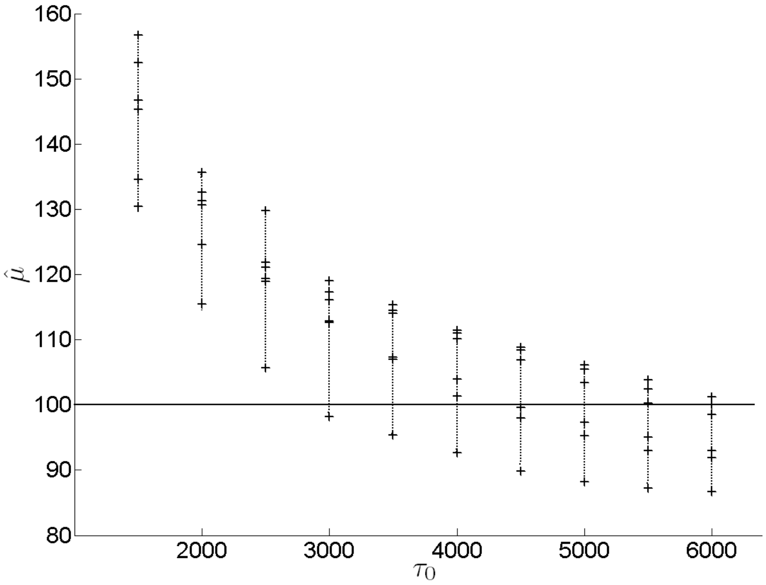

Figure 5: Mean $\hat{\mu}$ for a range of τ_0 for data with increasing number of loci ($k = 10$ constant). For each τ_0, marks indicate estimates for (from top to bottom) $m = 5, 10, 15, 20, 25, 30$. Each estimate is the mean of 10 data sets. The unbiased point estimates corresponding to the given number of loci are obtained respectively at optimal $\tau_0 = 6100, 6000, 5500, 4450, 4100, 2850$ (not shown).

of correlation between allele frequencies is decreased variance of the estimates of σ; which in turn decreases τ.

Our model assumes no population demography. Under balancing selection, migration turns out to be one of the confounding demographic effects [Hudson, 1991]. In the diffusion approximation to Wright-Fisher model, if included, the effect of migration on the allele frequencies is similar to that of mutation. Both migration and mutation affect the allele frequencies linearly, whereas selection affects the allele frequencies non-linearly. In particular, under our model, symmetric migration can be incorporated by replacing the mutation parameter θ by $\theta + M$ where $M = 4Nm$ is the population scaled migration parameter and m is the migration rate. It is also possible to introduce asymmetric migration structures where the rate of migration is different for each allele, although such a model will be data hungry since the number of parameters increases. On the other hand, if there is some migration ignoring it will inflate the mutation parameter estimates. The selection parameter estimates however, will not be affected unless migration is much stronger than mutation or it has a highly asymmetric structure.

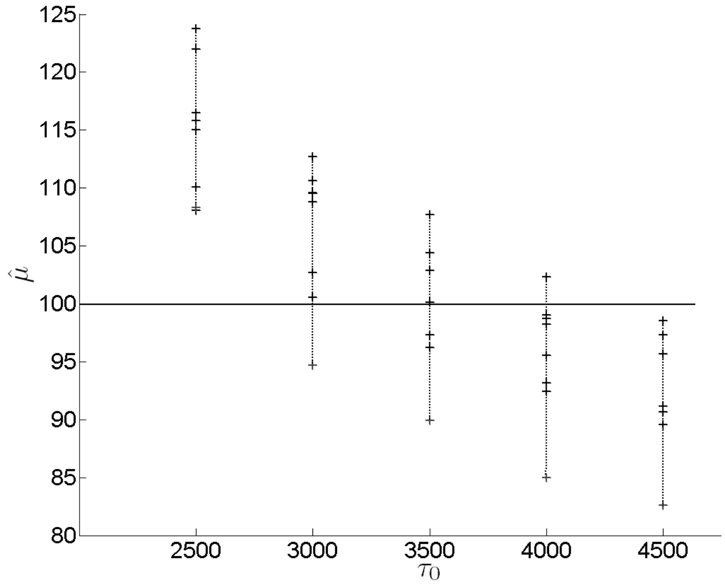

Figure 6: Mean $\hat{\mu}$ for a range of τ_0 for data with different (m,k) combinations. From top to bottom: $(m,k) = \{(3,32),(4,24),(6,16),(8,12),(12,8),(16,6),(24,4),(32,3)\}$ The unbiased point estimates corresponding to the given number of loci are obtained respectively at $\tau_0 = 4350, 3900, 3850, 3800, 3800, 3500, 3250, 3050, 2700$. Note that, in the given range of (m,k) values, all but very extreme pairs have similar optimal τ_0 (approximately 3500-4000). That is, when both m and k are relatively large, optimal τ_0 stays constant.

The hierarchical model is a first pass to make inference from a multiple loci system with a high number of alleles. As such, it does not take into account allele frequency changes due to epistatic interactions between loci. This is not surprising for a model of additive effects. A model capturing more biological detail such as epistasis is necessarily more complicated than the one presented in this paper. Under the framework presented here such a model implies a more intricate hierarchy. From an inference perspective, epistasis might be easier to handle with a model able to treat the multi locus allele frequencies without appealing to a hierarchical structure. For example, the mathematical machinery of the multiple loci k-allele diffusion can be also established in the presence of epistasis, as shown by new theoretical developments [Fearnhead, 2006]. Fearnhead extended the theory of single locus diffusion approximations to the multiple loci case, where a group of genes act as a single unit of selection. Using this result, overdominance can be modeled in the presence of epistatic interactions between loci. This avenue is particularly appealing for inference on loci related to immune system, where there is evidence for epistatic interactions between loci. Therefore, a future direction is to develop statistical methods for such models, a problem which we tackle in a subsequent paper.

Appendix A

We are interested in the joint density of θ, σ given the two data sets, $\mathbf{x}, \mathbf{Y}$. Apply Bayes' rule

$$\begin{aligned}
(15) \qquad P(\theta, \sigma | \mathbf{x}, \mathbf{Y}) &\propto P(\mathbf{x}, \mathbf{Y} | \theta, \sigma) P(\sigma, \theta) \\
(16) \qquad &= P(\mathbf{Y} | \theta, \mathbf{x}) P(\mathbf{x} | \theta, \sigma) P(\theta, \sigma) \\
(17) \qquad &= \left(\prod_{i=1}^{k} P(\mathbf{y}_i | \mathbf{x}_i, \theta) \right) P(\mathbf{x} | \theta, \sigma) P(\theta, \sigma).
\end{aligned}$$

The last equality follows from

$$\begin{aligned}
(18) \qquad P(\mathbf{Y} | \theta, \mathbf{x}) &= P(\mathbf{Y}, \mathbf{x} | \theta) / P(\mathbf{x} | \theta) = \prod_{i=1}^{k} P(\mathbf{y}_i, \mathbf{x}_i | \theta) / P(\mathbf{x} | \theta) \\
(19) \qquad &= \prod_{i=1}^{k} P(\mathbf{y}_i | \mathbf{x}_i, \theta) P(\mathbf{x}_i | \theta) / P(\mathbf{x} | \theta) \\
(20) \qquad &= \prod_{i=1}^{k} P(\mathbf{y}_i | \mathbf{x}_i, \theta) P(\mathbf{x} | \theta) / P(\mathbf{x} | \theta) \\
(21) \qquad &= \prod_{i=1}^{k} P(\mathbf{y}_i | \mathbf{x}_i, \theta).
\end{aligned}$$

The second equality in equation 18 follows from the fact that given the i^{th} sum, $\mathbf{x}_i$, the synonymous allele frequencies for that class $\mathbf{y}_i$ are independent of the other sums. The result in equation 17 says that the joint posterior of θ, σ has three pieces: The joint likelihood under neutrality, where only the synonymous data are used, the joint likelihood under selection, where only the non-synonymous data are used and the prior. This formulation is equivalent to use the posterior of θ from the synonymous data analysis as the prior of θ for the analysis under selection. This can be seen by expressing equation 17 as

$$\begin{aligned}
(22) \qquad & \prod_{i=1}^{k} P(\mathbf{y}_i|\mathbf{x}_i,\theta)P(\mathbf{x}|\theta,\sigma)P(\theta)P(\sigma) \\
(23) \qquad = \; & \prod_{i=1}^{k} P(\mathbf{y}_i|\mathbf{x}_i,\theta)P(\theta)P(\mathbf{x}|\theta,\sigma)P(\sigma) \\
(24) \qquad \propto \; & P(\theta|\mathbf{x},\mathbf{Y})P(\mathbf{x}|\theta,\sigma)P(\sigma).
\end{aligned}$$

Appendix B

The joint distribution of all the parameters for ith locus is given by

$$\begin{aligned}
(25) \qquad f(\sigma_i,\theta_i,\mathbf{x}_i,\mu,\tau) &= f(\sigma_i|\theta_i,\mathbf{x}_i,\mu,\tau)f(\theta_i,\mathbf{x}_i,\mu,\tau) \\
(26) \qquad &= f(\mathbf{x}_i|\sigma_i,\theta_i,\mu,\tau)f(\sigma_i,\theta_i,\mu,\tau).
\end{aligned}$$

By the second equality we get

$$(27) \qquad f(\sigma_i|\theta_i,\mathbf{x}_i,\mu,\tau) = \frac{f(\mathbf{x}_i|\sigma_i,\theta_i,\mu,\tau)f(\sigma_i,\theta_i,\mu,\tau)}{f(\theta_i,\mathbf{x}_i,\mu,\tau)}.$$

Using the fact that the hyperparameters affect the data only through the parameters and conditioning we have

$$(28) \qquad f(\sigma_i|\theta_i,\mathbf{x}_i,\mu,\tau) = \frac{f(\mathbf{x}_i|\sigma_i,\theta_i)f(\sigma_i|\theta_i,\mu,\tau)f(\theta_i,\mu,\tau)}{f(\mathbf{x}_i|\theta_i,\mu,\tau)f(\theta_i,\mu,\tau)}.$$

Noting that μ and τ fully specify the distribution of σ_i, we write

$$(29) \qquad f(\sigma_i|\theta_i,\mathbf{x}_i,\mu,\tau) = \frac{f(\mathbf{x}_i|\sigma_i,\theta_i)f(\sigma_i|\mu,\tau)}{f(\mathbf{x}_i|\theta_i,\mu,\tau)}$$

and finally

$$(30) \qquad f(\sigma_i|\theta_i,\mathbf{x}_i,\mu,\tau) \propto f(\mathbf{x}_i|\sigma_i,\theta_i)f(\sigma_i|\mu,\tau).$$

The first term is the stationary distribution of the allele frequencies under selection from the single locus model and the second term is the normal density.

Similarly for θ_i we get

$$(31) \qquad f(\theta_i|\sigma_i,\mathbf{x}_i,\mathbf{Y}_i,\mu,\tau)f(\theta_i,\mathbf{x}_i,\mathbf{Y}_i,\mu,\tau) = f(\mathbf{x}_i|\sigma_i,\theta_i,\mu,\tau)f(\sigma_i,\mathbf{Y}_i,\theta_i,\mu,\tau).$$

$$f(\theta_i|\sigma_i,\mathbf{x}_i,\mathbf{Y}_i,\mu,\tau) = \frac{f(\mathbf{x}_i|\sigma_i,\theta_i)f(\theta_i|\sigma_i,\mathbf{Y}_i,\mu,\tau)f(\sigma_i,\mu,\tau)}{f(\mathbf{x}_i|\sigma_i,\mu,\tau)f(\sigma_i,\mu,\tau)}. \tag{32}$$

$$f(\theta_i|\sigma_i,\mathbf{x}_i,\mathbf{Y}_i,\mu,\tau) = \frac{f(\mathbf{x}_i|\sigma_i,\theta_i)f(\mathbf{Y}_i|\theta)f(\theta_i)}{f(\mathbf{x}_i|\sigma_i,\mu,\tau)}. \tag{33}$$

So we have

$$f(\theta_i|\sigma_i,\mathbf{x}_i,\mathbf{Y}_i) \propto f(\mathbf{x}_i|\sigma_i,\theta_i)f(\mathbf{Y}_i|\theta)f(\theta_i). \tag{34}$$

Again, the first term is the stationary distribution of the allele frequencies under selection from the single locus model and the second term is the prior distribution of θ_i which is the posterior obtained from the neutral model analysis for ith locus.

References

Black, F.L., Hedrick, P.W. (1997). Strong balancing selection at hla loci: Evidence from segregation in south amerindian families. *Proc. Natl. Acad. Sci. USA* 94(23), 12452–12456.

Bustamante, C.D., Nielsen, R., Hartl, D.L. (2002). Maximum likelihood and bayesian methods for estimating the distribution of selective effects among classes of mutations using dna polymorphism data. *Theor. Popul. Biol.* 63, 91–103.

Buzbas, E.O., Joyce, P. (2009). Maximum likelihood estimates under *k*-allele models with selection can be numerically unstable. *Ann. Appl. Stat.* (In press).

Dagpunar, J.S., (1978). Sampling of variates from a truncated gamma distribution. *Journal of Statistical Computation and Simulation* 8, 59–64.

Damien, P., Wakefield, J., Walker, S.G. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Statist. Soc. (Series B).* 159, 853-867.

Damien, P., Walker, S.G. (2001). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of Computational and Graphical Statistics* Vol.10, No.2, 206-215.

Donnelly, P., Kurtz, T. (1999). A countable representation of the Fleming-Viot measure-valued diffusion. *Ann. Prob.* 24, 743–760.

Donnelly, P., Nordborg, M., Joyce, P. (2001). Likelihood and simulation methods for a class of nonneutral population genetics models. *Genetics* 159, 853-867.

Ethier, S.N., Griffiths, R.C. (1987). The infinitely-many-sites model as a measure-valued diffusion. *Ann. Prob.* 15, 515–545.

Ethier, S.N. and Kurtz, T.G. (1993). Fleming-viot processes in population genetics. *SIAM Journal of Control and Optimization* 31, 345–386.

Ewens, W. (2004). *Mathematical Population Genetics: I, Second ed. Interdisciplinary Applied Mathematics, Vol. 27. Springer-Verlag, New York. Theoretical Introduction.*

Fearnhead, P. (2006). The stationary distribution of allele frequencies when selection acts at unlinked loci. *Theor. Popul. Biol.* 70, 376–386.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D. (2004). *Bayesian Data Analysis. Second Ed.* Chapman and Hall, Boca Raton, FL.

Genz, A., Joyce, P. (2003). Computation of the normalization constant for exponentially weighted Dirichlet Distribution. *Computing Science and Statistics*, 35, 557-563.

Geweke, J. (1991). Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints. In E.M. Keramidas (Ed.), *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface,* pp. 571-578, Fairfax: Interface Foundation of North America, Inc.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their application. *Biometrika* 57, 97–109.

Hudson R. (1991). Gene genealogies and the coalescent process. In Futuyma D. and Antanovics J. (Eds.) *Oxford Evolutionary Surveys Vol.7, pp.1-44.* Oxford University Press.

Joyce, P., Genz, A., Buzbas, E.O. Efficient simulation methods for a class of nonneutral population genetics models. *Theor. Popul. Biol.* (Under Review).

Maruyama, T. and Nei, M. (1981). Genetic variability maintained by mutation and over-dominant selection in finite populations. *Genetics* 98, 441-459.

Metropolis, N., Rosenbluth, A.E., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092.

Meyer, D., Single, R. M., Mack, S. J., Lancaster, A. K., Nelson, M. P., Erlich, H. A., Fernandez-Vina, M, Thomson, G. (2007). Single locus polymorphism of classical HLA genes. In Hansen, J.A., (Ed.), *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference,* Vol. I, pp. 653–704, Seattle, WA: IHWG Press.

Muirhead, C., Slatkin, M. (2000). A Method for Estimating the Intensity of Overdominant Selection From the Distribution of Allele Frequencies. *Genetics* 156, 2119-2126.

Philippe, A. (1997). Simulation of right and left truncated gamma distributions by mixtures. *Statistics and Computing* 7, 173-181.

Robinson, J., Waller, M.J., Parham, P, de Groot, N., Bontrop, R., Kennedy, L.J., Stoehr, P., Marsh, S.G.E. (2003). Imgt/hla and imgt/mhc: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Research* 31, 311–314.

Watterson, G.A. (1977) Heterosis or neutrality? *Genetics* 85, 789–814.

Watterson, G.A. (1978). The homozygosity test of neutrality. *Genetics* 88, 405–417.

Wright, S. (1949). Adaptation and selection. In Jepson, G.L., Simpson, G.G., Mayr, E., (Ed.), *Genetics, Paleontology, and Evolution, p.383. Princeton Univ. Press, Princeton, NJ.*

DOI: 10.2202/1544-6115.1466