# Identifying genetic loci under selective pressure using a hierarchical Bayesian model with a posterior predictive p-value classifier

Toby Dylan Hocking
Mathieu Gautier
Jean-Louis Foulley

September 30, 2009

## Contents

**1 Introduction** **3**

**2 Methods** **4**
2.1 Model implementation . . . . . 4
2.2 PPP-value calculation . . . . . 4

**3 Results** **7**
3.1 Simulation verification . . . . . 7
3.2 Ancestral allele frequency estimates are sensitive to selection . 7
3.3 Population differentiation estimates are robust to selection . . 10
3.4 Simulation summaries using animations . . . . . 10
3.5 Prediction rates of the PPP-value classifier . . . . . 10

**4 Conclusions and future work** **15**

**5 Acknowledgements** **16**

# List of Figures


1 Final simulated blue allele frequency for 1000 loci and 12 populations is shown in a dotplot. Loci are ordered on the horizontal axis by ancestral allele frequency, and then divided into 3 panels by selection state. Note that loci under selection display no signs of selective pressure when in neutral color populations. Inversely, all loci which are not under selection behave similarly, regardless of population color. Also, the symmetry between blue and red alleles is clearly visible. . . . . 8

2 This grouped scatterplot illustrates that model estimates of ancestral allele frequency are not robust to selection. . . . . 9

3 Lineplots of differentiation parameter estimates $c_j$ evolving over time. The model was fit for 4 generations (50,100,150,200). Theoretical line shown in green is $t/N_j$. (FIXME: NEED BETTER FIGURE THAT SHOWS ROBUSTNESS TO SELECTION) . . . . . 11

4 Density estimates for PPP-values of each selection state, given data sets simulated with different selection strengths $s_i$. Note how it gets easier to distinguish selection as the selection strength parameter increases. . . . . 13

5 ROCs for several selection strengths, neutral allele concentrations, and population numbers. As shown in the densityplots, increasing selection strengths $s_i$ tend to increase the area under the curve. . . . . 14

# 1 Introduction

The recent explosion of molecular marker data in animal populations from technologies such as Single Nucleotide Polymorphism (SNP) assays opens new avenues of research for population genetics. These data allow testing of many aspects of our current models of population genetics, such as the evolutionary status of these markers relative to selective pressures. These data are usually investigated by either examining summary statistics and empirical distributions, or by using model-based approaches. We are more concerned with model-based approaches, with emphasis on detecting departures from the model.

In this study, we are interested in is the estimation of genetic differentiation between populations, and establishing methods for determining which markers and genomic regions have been under selective pressure. Genomic areas under selection are areas with probable functional significance, thus the goal is to develop methods we can use to identify functional genes and augment the current state of functional annotations for domestic animal genomes.

In the last several decades, the statistical tools of biologists and geneticists have evolved considerably. In particular, modern computers and stochastic methods such as Markov Chain Monte Carlo (MCMC) allow for estimation of the posterior distribution of parameters of Bayesian models of evolutionary systems. In this work, we investigate one such Bayesian model used to describe evolution of domestic animal populations, and extend it with a classifier for markers under selection.

The model of domestic animal evolution that we consider in this work is the hierarchical Bayesian model of Nicholson *et al.* [5], hereafter referred to as the Nicholson model. This model assumes that a single population gives rise to several subpopulations, which branch off at the same time, and begin independently evolving. The Nicholson model assumes all loci are affected only by genetic drift (not selection), and attempts to measure population differentiation in a manner which is analogous to the classical $F_{ST}$ of population genetics. In the process the model also yields estimates of ancestral allele frequency.

The new idea put forth in this work is a classifier for markers under selection, based on loci which do not fit well into the pure-drift Nicholson model. To quantify the probability that a locus fits the pure-drift model, we use Posterior Predictive P-values, or PPP-values [3, 2]. Essentially these PPP-values are the Bayesian analogue of the usual frequentist P-values, which indicate departures from the model hypotheses. The Nicholson model was designed for, and accurately models, independently evolving populations un-

der genetic drift. However, loci under selection in addition to genetic drift represent departures from the model hypotheses. Thus we use PPP-values estimated from the model to identify these aberrent loci.

Furthermore, to test the robustness of our classifier under different evolutionary conditions, we test it using extensive simulation of evolution by genetic drift and selection. The simulator has been implemented using the R programming language [6]. The model fitting has been implemented using compiled FORTRAN code dynamically linked to R. The simulations, analyses, and graphics discussed in this article can be reproduced by using the code published in the R package `nicholsonppp` on R-Forge [7]:

`http://nicholsonppp.r-forge.r-project.org/`

# 2 Methods

## 2.1 Model implementation

To simulate selection and genetic drift in several independent populations, we use a modified version of the simulator in [1]. We model these data using the pure-drift Nicholson model [5].

The model was first implemented using WinBUGS [4]. To provide speed optimizations for the model fitting, a faster model-fitting program was written in FORTRAN. We compared parameter estimates of the two programs and found no significant differences.

The posterior distribution for each model parameter was sampled 1000 times using MCMC, giving us an approximate posterior distribution. However, to simplify the analyses, each distribution was summarized using the mean. Thus when we refer to model parameter estimates, we mean the sample mean of the values drawn from the posterior distribution.

## 2.2 PPP-value calculation

The PPP-value for locus $i$ is defined as

$$\mathrm{PPP}_i = P\left[T(y_{ij}^{\mathrm{rep}}, \theta_{ij}) > T(y_{ij}^{\mathrm{obs}}, \theta_{ij}) | y^{\mathrm{obs}}\right]$$

where $y_{ij} = x_{ij}/N_{ij}$ is the allele frequency for locus $i$ and population $j$, $\theta_{ij} = (\pi_i, c_j)$ is a vector of model parameters, and $T$ is a discrepancy criterion applied to replicated (rep) and observed (obs) data sets.

We need to choose a discrepancy criterion which depends on both data

and parameters. Here, we use a $\chi^2$-type criterion:

$$T(y_{ij}, \theta_{ij}) = \sum_{j=1}^{P} \frac{[y_{ij} - E(y_{ij}|\theta_{ij})]^2}{\text{Var}(y_{ij}|\theta_{ij})}$$

The derivation of the expectation and variance of $y_{ij}$ follows. First, we note that we can decompose the Binomial random variable

$$x_{ij} = \sum_{k=1}^{N_{ij}} x_{ijk}$$

into the sum of independent, identically distributed Bernoulli random variables $x_{ijk} \sim \text{Bernoulli}(\alpha_{ij})$. Then we can write

$$y_{ij}|\theta_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} x_{ijk}$$

To find the expectation and variance of $y_{ij}$ given the hyperparameters $\theta_{ij}$, we will first characterize the moments of $x_{ijk}$, using the fact that $\alpha_{ij} \sim N(\pi_i, c_j\pi_i(1-\pi_i))$:

$$E(x_{ijk}) = E_{\alpha_{ij}}\left[E(x_{ijk}|\alpha_{ij})\right] = E(\alpha_{ij}) = \pi_i$$

$$\begin{aligned}
\text{Var}(x_{ijk}) &= \text{Var}_{\alpha_{ij}}\left[E(x_{ijk}|\alpha_{ij})\right] + E_{\alpha_{ij}}\left[\text{Var}(x_{ijk}|\alpha_{ij})\right] \\
&= \text{Var}_{\alpha_{ij}}(\alpha_{ij}) + E_{\alpha_{ij}}(\alpha_{ij}) - E_{\alpha_{ij}}({\alpha_{ij}}^2) \\
&= E_{\alpha_{ij}}(\alpha_{ij}^2) - E_{\alpha_{ij}}(\alpha_{ij})^2 + E_{\alpha_{ij}}(\alpha_{ij}) - E_{\alpha_{ij}}(\alpha_{ij}^2) \\
&= E_{\alpha_{ij}}(\alpha_{ij})\left(1 - E_{\alpha_{ij}}(\alpha_{ij})\right) \\
&= \pi_i(1-\pi_i)
\end{aligned}$$

Using a similar derivation, we find that

$$\begin{aligned}
\text{Cov}(x_{ijk}, x_{ijk'}) &= E_{\alpha_{ij}}\big[\underbrace{\text{Cov}(x_{ijk}, x_{ijk'}|\alpha_{ij})}_{0}\big] + \text{Cov}\big[\underbrace{E(x_{ijk}|\alpha_{ij})}_{\alpha_{ij}}, \underbrace{E(x_{ijk'}|\alpha_{ij})}_{\alpha_{ij}}\big] \\
&= 0 + \text{Cov}(\alpha_{ij}, \alpha_{ij}) \\
&= \text{Var}(\alpha_{ij}) \\
&= c_j\pi_i(1-\pi_i)
\end{aligned}$$

Using the previous results, we can derive the expectation and variance of $y_{ij}$, which we need to calculate the PPP-values:

$$E(y_{ij}|\theta_{ij}) = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} E(x_{ijk}) = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} \pi_i = \pi_i$$

$$\begin{aligned}
\mathrm{Var}(y_{ij}|\theta_{ij}) &= \mathrm{Var}\left( \sum_{k=1}^{N_{ij}} \frac{x_{ijk}}{N_{ij}} \right) \\
&= \frac{1}{N_{ij}^2} \left[ \sum_{k=1}^{N_{ij}} \mathrm{Var}(x_{ijk}) + 2 \sum_{k=1}^{N_{ij}} \sum_{k'=k+1}^{N_{ij}} \mathrm{Cov}(x_{ijk}, x_{ijk'}) \right] \\
&= \frac{1}{N_{ij}^2} \left[ \sum_{k=1}^{N_{ij}} \pi_i(1-\pi_i) + 2 \sum_{k=1}^{N_{ij}} \sum_{k'=k+1}^{N_{ij}} c_j \pi_i (1-\pi_i) \right] \\
&= \frac{N_{ij}\pi_i(1-\pi_i) + N_{ij}(N_{ij}-1)c_j\pi_i(1-\pi_i)}{N_{ij}^2} \\
&= \frac{1}{N_{ij}} \pi_i(1-\pi_i) \left[ 1 + (N_{ij}-1)c_j \right]
\end{aligned}$$

In practice, we will calculate the discrepancy criterion $T$ for the replicated (rep) and observed (obs) data at each iteration in the Markov chain. Thus, using $^t$ to denote the value of a variable in the $t$-th iteration through the Markov chain,

$$\begin{aligned}
T(y_{ij}^{\mathrm{rep},t}, \theta_{ij}^t) &= \frac{\left( \mathrm{Bin}(N_{ij}, \alpha_{ij}^t)/N_{ij} - \pi_i^t \right)^2}{\pi_i^t(1-\pi_i^t)\left[1 + (N_{ij}-1)c_j^t\right]/N_{ij}} \\
T(y_{ij}^{\mathrm{obs},t}, \theta_{ij}^t) &= \frac{\left( Y_{ij}/N_{ij} - \pi_i^t \right)^2}{\pi_i^t(1-\pi_i^t)\left[1 + (N_{ij}-1)c_j^t\right]/N_{ij}}
\end{aligned}$$

where $\mathrm{Bin}(\cdot,\cdot)$ represents a randomly generated number from the binomial distribution.

Next, we define the indicator variable

$$\mathrm{PPP}_i^t = \begin{cases} 1 & \text{if } \sum_{j=1}^{P} \left[ T(y_{ij}^{\mathrm{rep},t}, \theta_{ij}^t) - T(y_{ij}^{\mathrm{obs},t}, \theta_{ij}^t) \right] > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $P$ is the number of populations. We sum over populations since we need to have a criterion for each locus at each iteration.

Finally, the PPP-value for locus $i$ is calculated at the end of the $N_{\text{iter}}$ iterations of Markov chain sampling from the posterior distribution:

$$\text{PPP}_i = \frac{1}{N_{\text{iter}}} \sum_{t=1}^{N_{\text{iter}}} \text{PPP}_i^t$$

# 3 Results

## 3.1 Simulation verification

To verify that all loci behave according to expectations, we used dotplots of final allele frequency (Figure 1).

This dotplot clearly shows that all loci under selection display no signs of selective pressure when in neutral color populations. Inversely, all loci which are not under selection behave similarly, regardless of population color. Also, the symmetry between blue and red alleles is clearly visible. This dotplot efficiently shows that the allele frequencies evolved according to the simulator hypotheses.

## 3.2 Ancestral allele frequency estimates are sensitive to selection

We are simulating loci under selection, and analyzing them using the pure-drift Nicholson model. Thus we expect that the loci under selection will not fit the model well.

To diagnose how selection state influences model fit, we plot ancestral allele frequency estimates for each loci versus actual values from the simulation (Figure 2).

This scatterplot clearly shows that neutral loci are well estimated by the model, but loci under balancing and positive selection are not well estimated. This result is sensible in view of the fact that the Nicholson model was designed with only genetic drift in mind. Thus it is expected that model parameters for loci under selection are not estimated well.

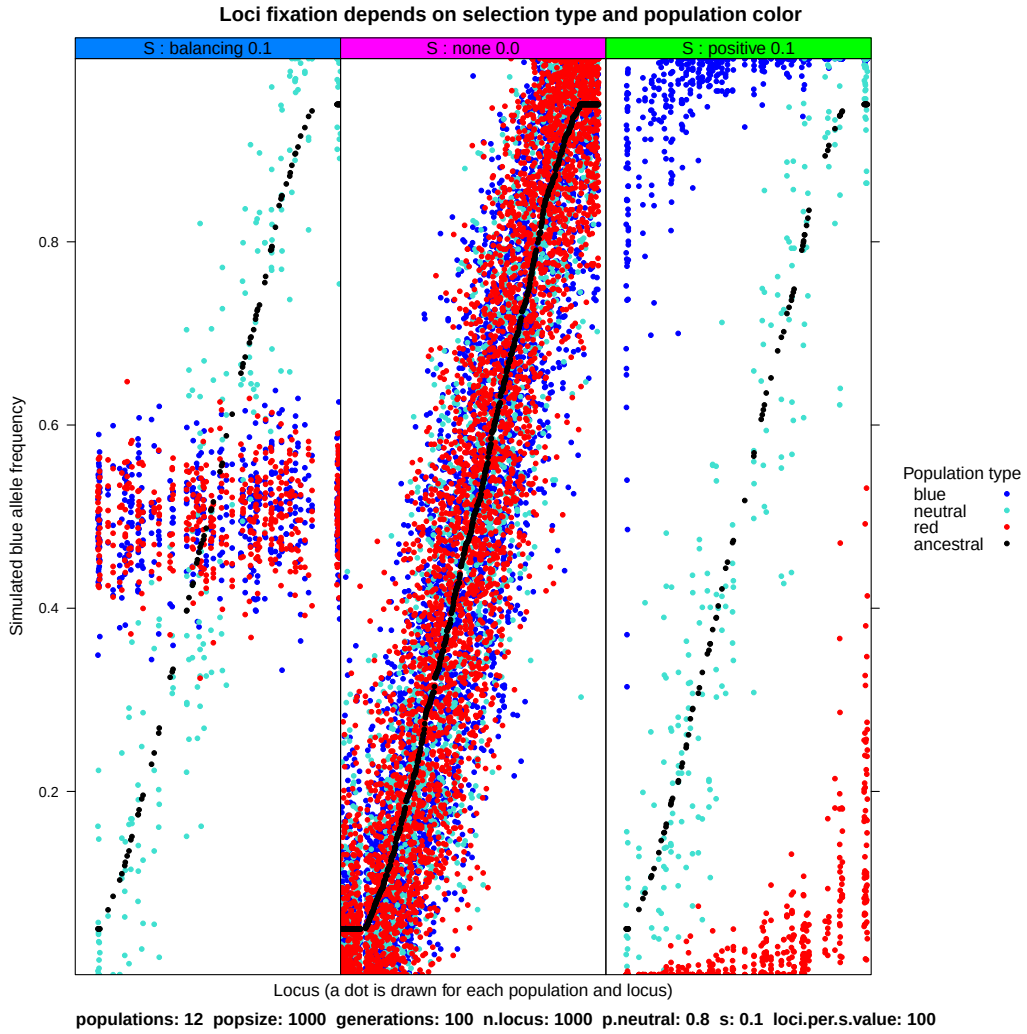


Figure 1: Final simulated blue allele frequency for 1000 loci and 12 populations is shown in a dotplot. Loci are ordered on the horizontal axis by ancestral allele frequency, and then divided into 3 panels by selection state. Note that loci under selection display no signs of selective pressure when in neutral color populations. Inversely, all loci which are not under selection behave similarly, regardless of population color. Also, the symmetry between blue and red alleles is clearly visible.

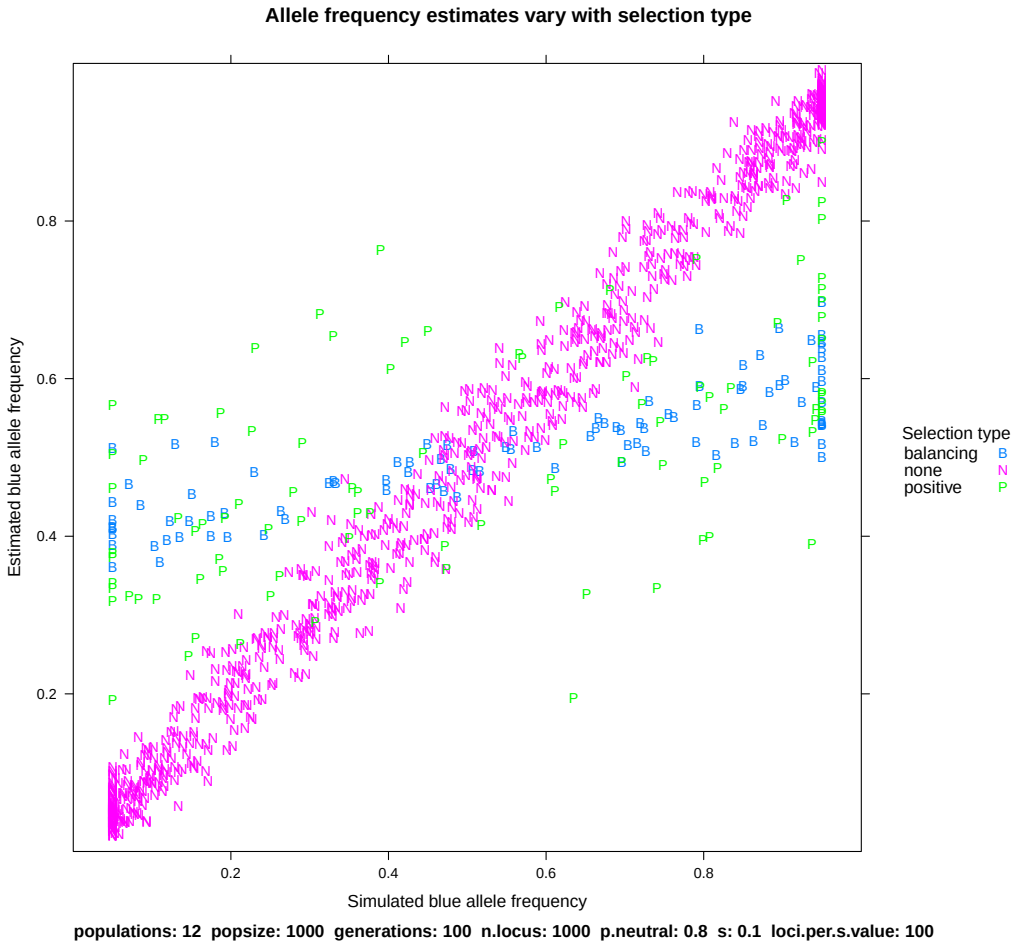


Figure 2: This grouped scatterplot illustrates that model estimates of ancestral allele frequency are not robust to selection.

## 3.3 Population differentiation estimates are robust to selection

To investigate the estimates of the population differentiation parameters $c_j$, we first note that from population genetics, we expect that

$$c_j = 1 - (1 - \frac{1}{N_j})^t \approx t/N_j \quad (1)$$

where $t$ is number of generations and $N_j$ is effective population size of population $j$.

Note that this approximation implies that we expect estimates of $c_j$ to increase linearly over time. To visualize this linear trend, we used lineplots of the estimate differentiation parameter $c_j$ over time $t$ (Figure 3).

Note the linear behavior of the model estimates, as expected. However, the slopes of the lines do not always match the expected theoretical slopes, which can be attributed to approximation errors in Equation 1.

## 3.4 Simulation summaries using animations

To visualize how the allele frequencies change over time, we made plots of allele frequency time series, ancestral estimates, and dotplots. Then we used to the animation package [8] to create a series of these plots, one for each generation. These images are put together and viewed in sequence to form a statistical animation that reveals how the populations evolve during the simulation. The animations can be viewed on the accompanying website:

`http://nicholsonppp.r-forge.r-project.org/`

The link between the plots, combined with the animation over several generations, has proven to be a powerful pedagogical device that encourages rapid understanding of the simulator and model hypotheses. Multivariate statistical animations such as this can be useful as teaching tools for students of statistical population genetics.

## 3.5 Prediction rates of the PPP-value classifier

To evaluate the sensitivity and specificity of the PPP-value classifier, we fit the model on 3 sets of 5 simulations with different parameter values:

| Set | Populations | Loci |
|---|---|---|
| usual | 12 | 1000 |
| few populations | 4 | 1000 |
| many neutral loci | 12 | 19999 |

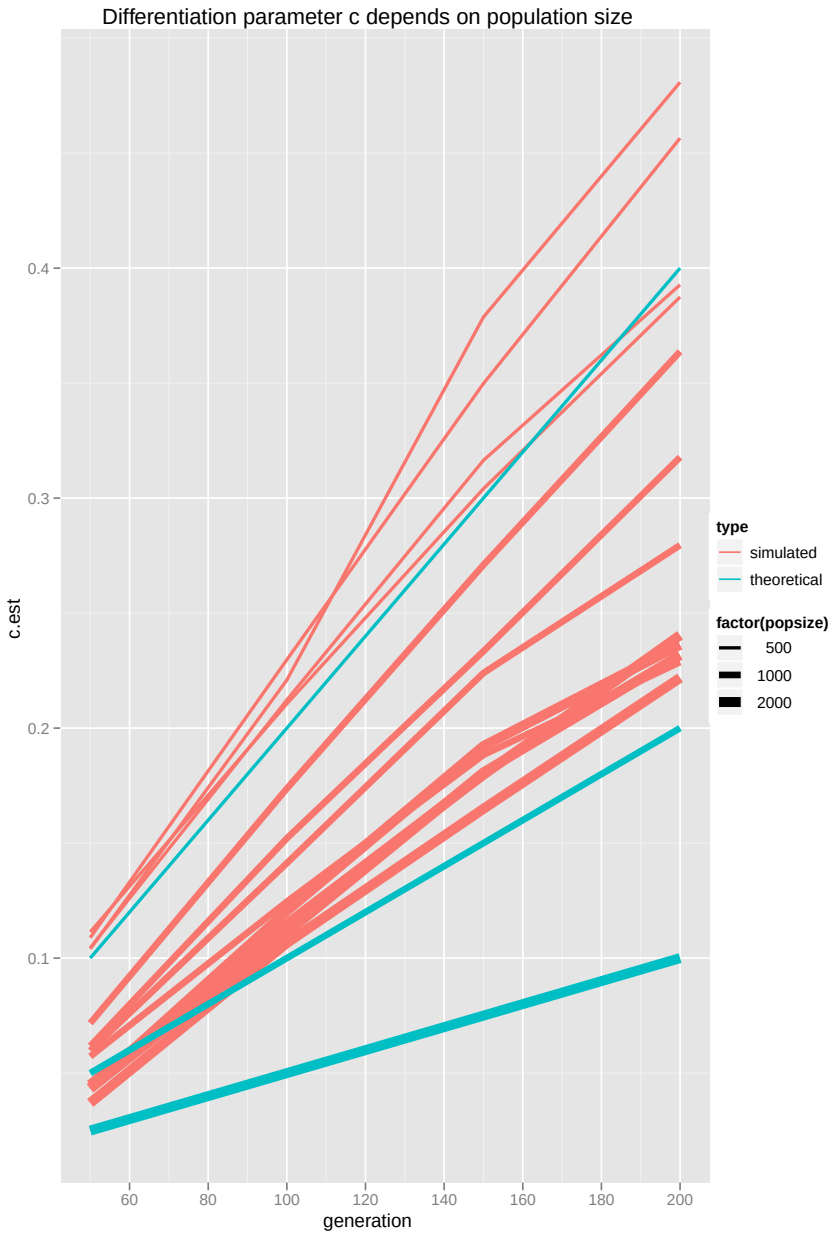


Figure 3: Lineplots of differentiation parameter estimates $c_j$ evolving over time. The model was fit for 4 generations (50,100,150,200). Theoretical line shown in green is $t/N_j$. (FIXME: NEED BETTER FIGURE THAT SHOWS ROBUSTNESS TO SELECTION)

For each of the above parameter sets, we fixed constant parameter values of population size 1000 and 100 generations of evolution. Then we did 5 different simulations with 100 loci each of

$$s_i \in \{0.001, 0.01, 0.032, 0.1, 1\} = \{10^{-3}, 10^{-2}, 10^{-1.5}, 10^{-1}, 10^{0}\}$$

These values were chosen since 1 is a value higher than usually observed in nature, and 0.001 is as weak as if there was no selection at all.

For each of these sets we first made density plots of PPP-value conditional on selection state for each $s_i$ value. The results for the usual simulation set are shown in Figure 4.

These density plots clearly visualize the different distributions of PPP-values for different selection types. These plots suggest that low PPP-values can be used to indicate positive selection, and high PPP-values can be used to indicate balancing selection.

For the purposes of this work, we will limit ourselves to the classification of a locus as positive or not positive, ignoring balancing selection.

The density plots suggest the following classifier for positive loci:

$$\hat{S}_i(h) = \text{state of locus } i \text{ given threshhold } h = \begin{cases} \text{positive} & \text{if PPP}_i < h \\ \text{none} & \text{if PPP}_i \geq h \end{cases}$$

where $h$ is a PPP-value threshhold that will determine the false positive/false negative rates.

These density plots also clearly show that increasing the value of the selection coefficient $s_i$ tends to increase the separation of distributions of PPP-values.

We found that with fewer populations it is more difficult to distinguish the behavior of loci under selection. When we examined the simulation with very many neutral loci, we saw that the densities of the different selection states are clearly distinguishable. However, the sheer number of neutral loci makes for a large false positive rate.

Receiver operating characteristics (ROCs) were also traced, to compare all 15 simulations at the same time (Figure 5). In ROCs, we plot lines of sensitivity against $1 -$ specificity for every possible threshhold value of the classifier, where sensitivity is the true positive rate and specificity is the false positive rate.

The ROCs in Figure 5 show that increasing selection strengths $s_i$ tend to increase the area under the curve. This indicates that loci under stronger selection are easier to distinguish from loci not under selection. Also, note how the data sets with 4 populations generate decision rules which are noticeably less powerful than those in the other 2 simulations.

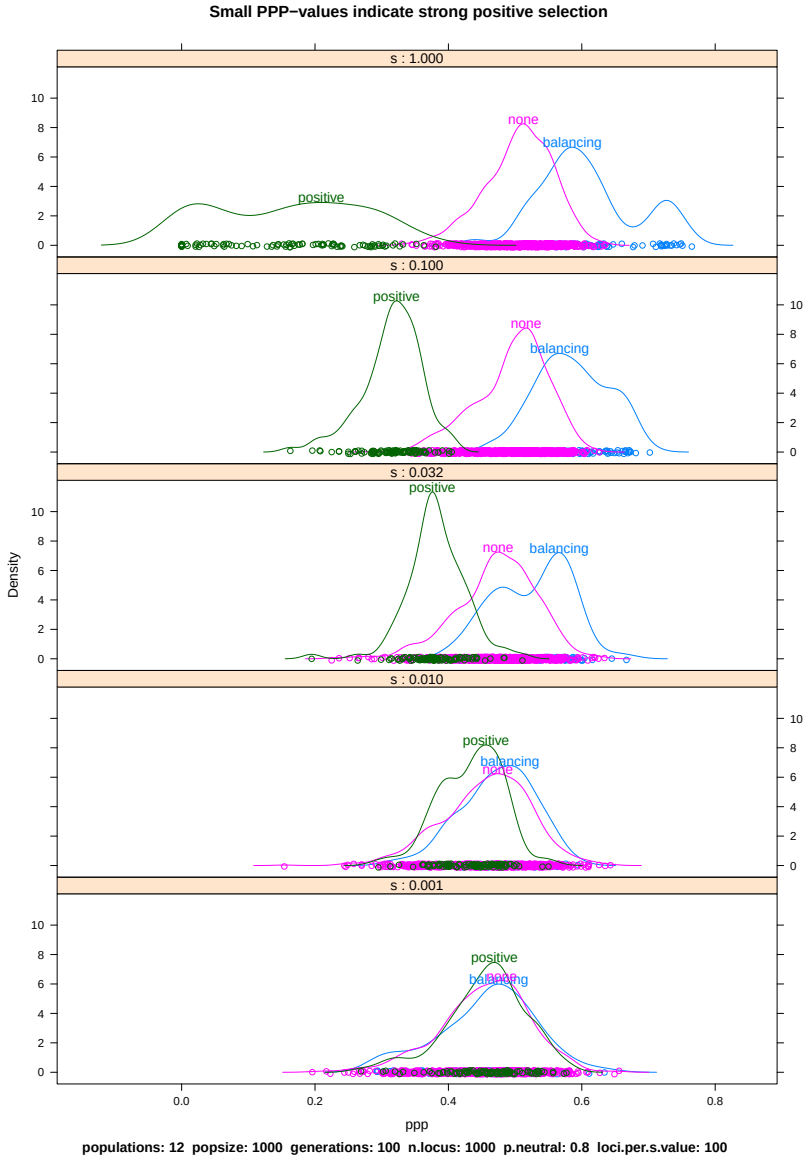


Figure 4: Density estimates for PPP-values of each selection state, given data sets simulated with different selection strengths $s_i$. Note how it gets easier to distinguish selection as the selection strength parameter increases.

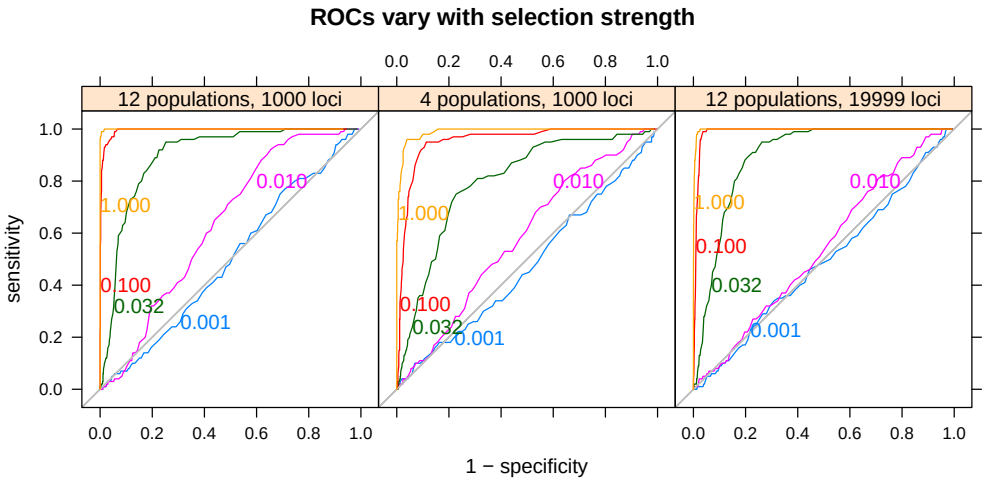


Figure 5: ROCs for several selection strengths, neutral allele concentrations, and population numbers. As shown in the densityplots, increasing selection strengths $s_i$ tend to increase the area under the curve.

# 4 Conclusions and future work

The Nicholson model investigated in this work has been thoroughly tested by exhaustive simulation of genetic drift and selection. The Nicholson model is a simple, robust, and useful Bayesian framework for modeling population divergence under genetic drift from a recent common ancestor.

To extend the Nicholson model to account for loci under selection, we looked for loci which did not fit the Nicholson model well. Our extension of the Nicholson model used PPP-values to classify the selection state of each locus.

The PPP-values are an effective classifier for selection state when the selection coefficient of the locus is sufficiently strong ($s_i > 0.01$). The false positive rate of the classifier drops as the selection strength of the loci increases. But if there are an overwhelming majority of loci not under selection, there will inevitably be high false positive rates.

Several graphical methods were used to visualize the data, including use of statistical animations to understand the behavior of all loci in the simulations. Such methods can be adapted as illustrative teaching tools to facilitate rapid comprehension of these multivariate data.

In this study, we did simulations and sensitivity analysis on every possible threshhold to characterize the classifier. However, with real data sets, we will need a more concrete criterion for choosing the PPP-value threshhold for the classifier.

A hypothesis of the Nicholson model is that all loci are independent. Genetic markers are found on linear chromosomes, so some loci are closer than others. Some markers may even be in the same gene. Thus the hypothesis of loci independence is clearly false, suggesting that some model of correlation between the loci could be introduced into the model.

The Nicholson model supposes that all populations diverge from their common ancestor at the same time, which is false. Thus another beneficial model complication would be to introduce some parameters that model the more tree-like structure of real genetic histories.

To come up with a useful genome annotation method, we would need a method of synthesizing classification of loci into classification of genomic regions. Such methods for combining classifier predictions exist, but would need to be adapted for this particular use with PPP-values.

More work needs to be done to characterize the expected number of false positives and false negatives in a real dataset. In particular, a comparison of this classifier with other existing models of locus selection state should be done.

The model we used only accounts for genetic drift, and detects selection

as aberrations from the model. An enhanced selection state classifier could result if we introduced a more complicated Bayesian model, with parameters for the selection state and/or coefficient.

Finally, we should apply this model to several well-characterized empirical datasets. Due to the portability of the accompanying R package, this should not be a difficult task. The results of our simulations suggest that we will be able to accurately identify loci under strong selection using this method.

# 5 Acknowledgements

Thanks to Gilles Celeux of the SELECT laboratory of the Institut National de Recherche en Informatique et en Automatique (INRIA), who provided funding for this project.

# References

[1] Mark A. Beaumont and David J. Balding. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13:969–980, 2004.

[2] Jean-Louis Foulley and Mathieu Gautier. Detecting selected loci via hierarchical bayesian models and kindred methodology with an application to SNPs in cattle. In *Statistical Methods for Post-genomic Data workshop, SMPGD'09*, AgroParisTech, Paris, January 2009.

[3] A. Gelman, X. L. Meng, and H. S. Stern. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica*, 6:733–760, 1996.

[4] W. R. Gilks, A. Thomas, and D. J. Spiegelhalter. A language and program for complex Bayesian modeling. *The Statistician*, 43:169–178, 1994.

[5] George Nicholson, Albert V. Smith, Frosti Jónsson, Ómar Gústafsson, Kári Stefánsson, and Peter Donnelly. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Statist. Soc.*, 64:695–715, 2002.

[6] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[7] Stefan Theußl and Achim Zeileis. Collaborative Software Development Using R-Forge. *The R Journal*, 1(1):9–14, May 2009.

[8] Yihui Xie. *animation: Demonstrate Animations in Statistics*, 2009. R package version 1.0-4.