

Bayesian Variable Selection for Detecting Adaptive Genomic Differences among Populations

Andrea Riebler

University of Zurich

Bern, September 2007

Joint work with Wolfgang Stephan and Leonhard Held.

Outline

- 1 Introduction
- 2 Method: Hierarchical Bayesian model
 - without Bayesian variable selection
 - with Bayesian variable selection
- 3 Applications
 - Real dataset
 - Simulated datasets
- 4 Summary

1. Introduction

Key problem:

Detection of DNA regions affected by selection

Considered classes of selection within a population

- Balancing selection: Stable frequencies of alleles
- Directional selection: Greater fitness of one allele

The F_{st} -coefficient

F_{st} -coefficient

Quantification of the amount of genetic differentiation, $F_{st} \in [0, 1]$

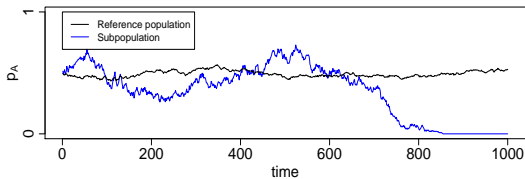
Low F_{st} \Rightarrow balancing selection

High F_{st} \Rightarrow directional selection

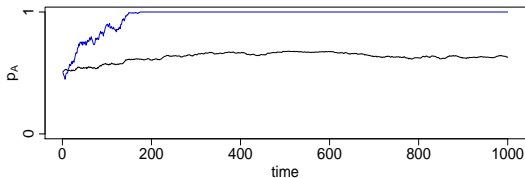
Interpretation of F_{st}^{ij}

Measure for allele frequency difference at a locus i between subpopulation j and a reference population

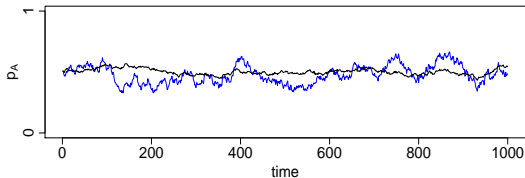
Interpretation of the F_{st} -coefficient



Random genetic drift
 \Rightarrow slowly increasing F_{st}^{ij}



Directional selection
 $p_{A_j} \not\approx p_{A_{ref}} \Rightarrow$ high F_{st}^{ij}



Balancing selection
 $p_{A_j} \approx p_{A_{ref}} \Rightarrow$ low F_{st}^{ij}

2. Method: Hierarchical Bayesian model

Basic method:

Hierarchical Bayesian model by Beaumont and Balding*

Two levels:

- 1 Expression of the likelihood for the allele frequency counts as function of F_{st}
- 2 Definition of the F_{st} -values

* Beaumont, M. A. and Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans, *Molecular Ecology* **13**: 969-980

Multinomial-Dirichlet likelihood for allele counts

1. Level

Multinomial-Dirichlet likelihood L_{ij} for the allele counts, so that

$$a_{ij1}, \dots, a_{ijK_i} \sim \text{MultDir} \left(\frac{1}{F_{st}^{ij}} - 1, x_{i1}, \dots, x_{iK_i} \right)$$

a_{ijk} : Count of allele k in population j at locus i

x_{ik} : Frequency of allele k at locus i

Joint likelihood:
$$L = \prod_{i=1}^I \prod_{j=1}^J L_{ij}$$

(with $I = \#\text{loci}$, $J = \#\text{populations}$, $K_i = \#\text{alleles at locus } i$)

Combining information across loci and populations

2. Level

Incorporation of locus-specific and population-specific effects in F_{st}^{ij}

$$F_{st}^{ij} = \frac{\exp(\alpha_i + \beta_j + \gamma_{ij})}{1 + \exp(\alpha_i + \beta_j + \gamma_{ij})},$$

with

α_i : locus-specific random effect

β_j : population-specific random effect

γ_{ij} : locus-by-population-specific random effect

Implementation

Goal

Obtaining values from the posterior distribution

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x} | \mathbf{a}) \propto \underbrace{P(\mathbf{a} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})}_{L = \prod_{i=1}^I \prod_{j=1}^J L_{ij}} \cdot f(\boldsymbol{\alpha}) f(\boldsymbol{\beta}) f(\boldsymbol{\gamma}) f(\mathbf{x})$$

using the prior distributions given in Beaumont and Balding (2004)

Method

Markov chain Monte Carlo (MCMC) algorithm

Interpretation

Primary interest towards the posterior distribution of the α_i :

Directional selection

$$P(\alpha_i < 0) \leq 0.05$$

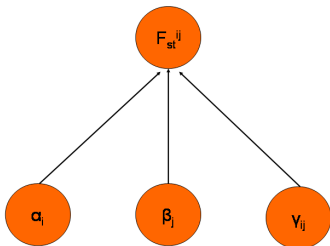
Balancing selection

$$P(\alpha_i < 0) \geq 0.95$$

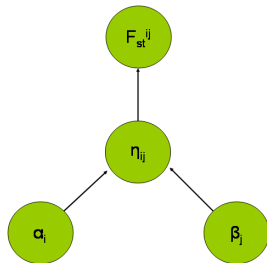
Model extension

Reparameterization

Introduction of a new variable $\eta_{ij} = \log \left(\frac{F_{st}^{ij}}{1 - F_{st}^{ij}} \right) = \alpha_i + \beta_j + \gamma_{ij}$



(a) Current definition



(b) Reparameterization

Reparameterization

Prior distribution : $\eta_{ij} | \alpha_i, \beta_j \sim N(\alpha_i + \beta_j + \mu_\gamma, \sigma_\gamma^2)$

Advantage

Full-conditional of α_i and β_j become normal distributions

⇒ Gibbs-sampling possible

Bayesian variable selection

Idea

Introduction of a Bayesian auxiliary variable δ_i for every α_i :

$$\eta_{ij} = \delta_i \cdot \alpha_i + \beta_j + \gamma_{ij} \quad \text{with} \quad \delta_i | p \sim \text{Bernoulli}(p), \quad p \sim U(0, 1)$$

to indicate loci subject to selection.

Equivalent to a hierarchical prior distribution:

- Number of included α_i is uniformly distributed.
- Given the number, the included α_i 's are a random sample.

The δ_i are dependent in this approach.

Bayesian variable selection: Interpretation

Before

Classification using Bayesian P-values

Now

Classification using $P(\delta_i = 1|\text{data})$:

$$P(\delta_i = 1|\text{data}) \geq 0.9 \quad \Rightarrow \quad \text{Locus } i \text{ subject to selection}$$

3. Applications

Real dataset:

Fruit fly allozyme dataset

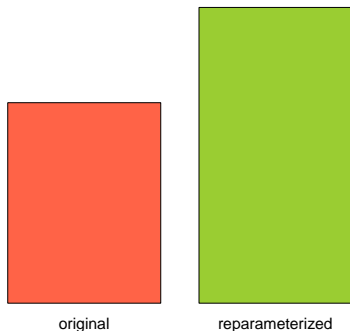
- $I = 61$ loci
- $J = 15$ populations
- sample size: 36 – 120
- Loci mostly di- or tri-allelic

Performance measure: original vs. reparameterized method

Effective sample size (ESS)

Estimate on the information content of the MCMC samples in terms of an equivalent number of independent samples

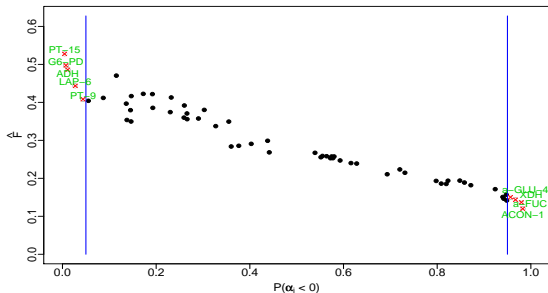
Relation of ESS standardised for CPU run time



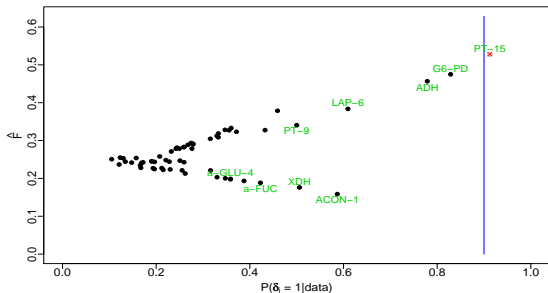
about 50% improvement in
the effective sample size

Results for the fruit fly allozyme dataset

without
variable selection:



with
variable selection:



Generation of simulated datasets

General settings:

- 10 subpopulations
- 900 neutral loci
- 50 loci subject to balancing/directional selection
- 3 alleles for all loci

Wright-Fisher model

Seven simulations using different selection coefficients and a sample size of 100 individuals per population

Generation of simulated datasets

Simulation from a Wright-Fisher model with migration

- Chromosomes in the current generation are replaced with immigrants
- The next generation is sampled according to a specified selection coefficient s .

The algorithm is repeated for 1000 generations.

Afterwards chromosomes are sampled with replacement.

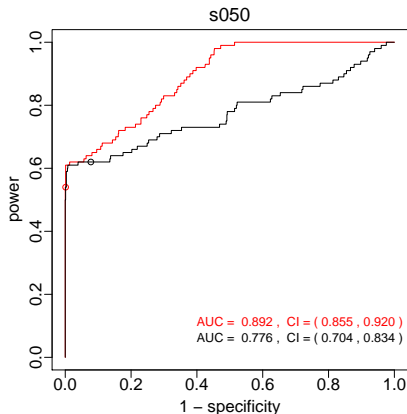
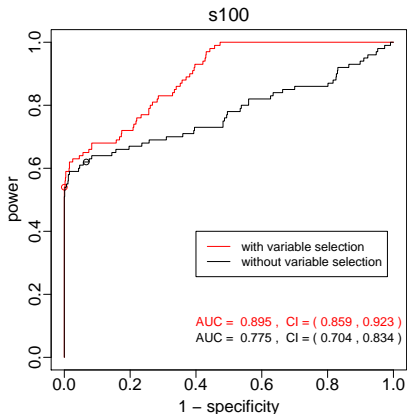
Summary of the results for the Wright-Fisher model

	without var. selection	with var. selection
Total false positives	492	1
Total dir. detected	350	350
Total bal. detected	65	19

with a total of:

- 6300 neutral loci
- 350 loci subject to balancing selection
- 350 loci subject to directional selection

ROC-curve: Wright-Fisher model



ROC-analysis: Wright-Fisher model

Dataset	$\widehat{\Delta AUC}$	$\widehat{\text{var}}(\widehat{\Delta AUC})$	lower CI	upper CI	$\frac{\widehat{\Delta AUC}}{\text{se}(\widehat{\Delta AUC})}$
s100	0.119	0.00033	0.084	0.155	6.531
s050	0.116	0.00032	0.081	0.151	6.490
s020	0.111	0.00027	0.079	0.143	6.736
s100-Fb	0.087	0.00024	0.057	0.118	5.655
s050-Fb	0.107	0.00029	0.073	0.140	6.238
s020-Fb	0.118	0.00027	0.085	0.150	7.098
s100-Fb-40	0.104	0.00022	0.075	0.133	7.026

Fb: flexible immigration rate among populations

Fb-40: flexible immigration rate among populations, sample size 40

Summary

Summary

Presentation of a Bayesian variable selection framework for detecting loci being subject to selection using Bayesian variable selection.

Results

The ROC analysis clearly favored the model including Bayesian variable selection in the case of the simulations from the Wright-Fisher model.

Thank you for your attention!



Any questions?