Simulation and modeling genomic signatures of selection from SNP data

Toby Dylan Hocking

INRA GABI Jouy-en-Josas

5 June 2009

Outline

Simulating SNP data for model validation

Model estimation and results

Introduction

- Domestic cows (*Bos taurus*) have been selected over thousands of years for milk production, meat production, resistence to disease, etc.
- But how is this differential selection reflected in their genome?
- We can genotype a cow at 60,000 SNPs, and compare these genotypes between modern domestic populations.
- The question: can we derive a statistic – a "signature of selection" – that indicates a genomic region has been under selection?
- A possible answer: Nicholson *et al.*(2002) estimate ancestral population allele frequencies from SNP data.
- Can we extend the Nicholson model to inform about signatures of selection?
- Simulate the evolution using known evolution parameters, fit the model, then look for signatures of selection in the alleles we know were under selection.

A simple selection simulator, based on Beaumont, Balding (2004)

- Single ancestral population.
- Several subpopulations:
  - Initially with the same allele frequency but evolving independently.
  - Each has a different background color (blue, red, neutral).
- Several independent loci:
  - Two alleles (red, blue) to mimic the SNP data.
  - Each has a different selection coefficient $s \in \mathbb{R}^+$, but normally in reality $s < 1$.
  - Each has a different selection type (neutral, positive, or balancing).
- Evolution by drift and selection over several generations.

Simulation setup

- loci $i \in \{1, ..., L\}$
- 15 distinct selection parameters

$$s_i \in \{ \quad 0.0025, 0.0050, 0.0075, 0.0100, 0.0200, \\ 0.0300, 0.0400, 0.0500, 0.0750, 0.1000, \\ 0.5000, 0.7500, 1.0000, 1.5000, 2.0000 \quad \}$$

- 50 replications of each s value, for positive and balancing selection
- proportion of neutral loci p.neutral $\in \{1/3, 0.9\} \Rightarrow$ 750 or 13500 replications
- populations $j \in \{1, ..., P\}$, $P \in \{2, 4, 8\}$
- ancestral blue allele frequencies $\pi_1, ..., \pi_L \sim$ rbeta$(0.7, 0.7)$
- starting subpopulation blue allele frequencies $\alpha_{i1}(0) = ... = \alpha_{iP}(0) = \pi_i$
- effective subpopulation size popsize $\in \{100, 500, 1000\}$

Evolution equations

- genetic drift $\alpha^*_{ij}(t) = \text{rbinom}(\text{popsize}, \alpha_{ij}(t-1))$
- genotype frequency vector based on Hardy-Weinberg equilibrium

$$A_{ij}(t) = \begin{bmatrix} \alpha^*_{ij}(t)^2 & 2\alpha^*_{ij}(t)(1-\alpha^*_{ij}(t)) & (1-\alpha^*_{ij}(t))^2 \end{bmatrix}^T$$

- relative fitness of genotypes

w_{ij}^{BB}	w_{ij}^{BR}	w_{ij}^{RR}	i	j
1	$1+s_i/2$	$1+s_i$	positive	red
$1+s_i$	$1+s_i/2$	1	positive	blue
1	$1+s_i$	1	balancing	
1	1	1	neutral	

- allele frequency updated for selection:

$$\alpha_{ij}(t) = \frac{\begin{bmatrix} w_{ij}^{BB} & w_{ij}^{BR}/2 & 0 \end{bmatrix} \cdot A_{ij}(t)}{\begin{bmatrix} w_{ij}^{BB} & w_{ij}^{BR} & w_{ij}^{RR} \end{bmatrix} \cdot A_{ij}(t)}$$

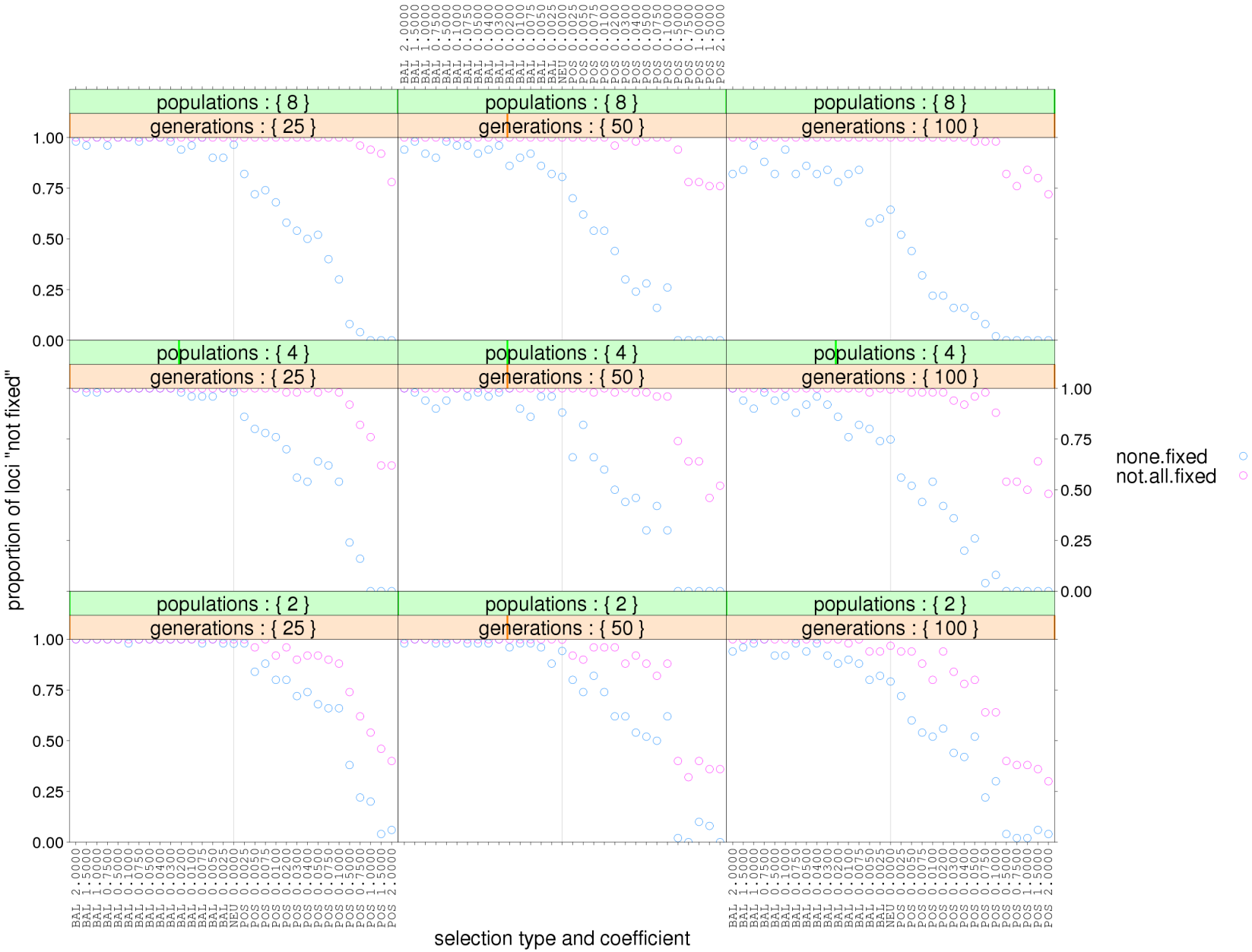
Loci fixation?

- Some alleles can be fixed at the end of the simulation, $\alpha_{ij}(t) \in \{0, 1\}$
- Since we are modeling Single Nucleotide Polymorphisms the data are probably not fixed.
- We can consider non-fixed loci only, for the model fitting later.
- Criteria:

  not.all.fixed Throw out the locus if all subpopulations fixed.

  none.fixed Throw out the locus if one or more subpopulations fixed.

Loci fixation depends on selection type and coefficient

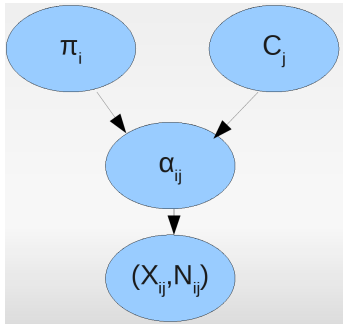


popsize : { 1000 } loci : { 50 } fixed : { none.fixed, not.all.fixed } p.neutral : { 0.333333333333333 }

Outline

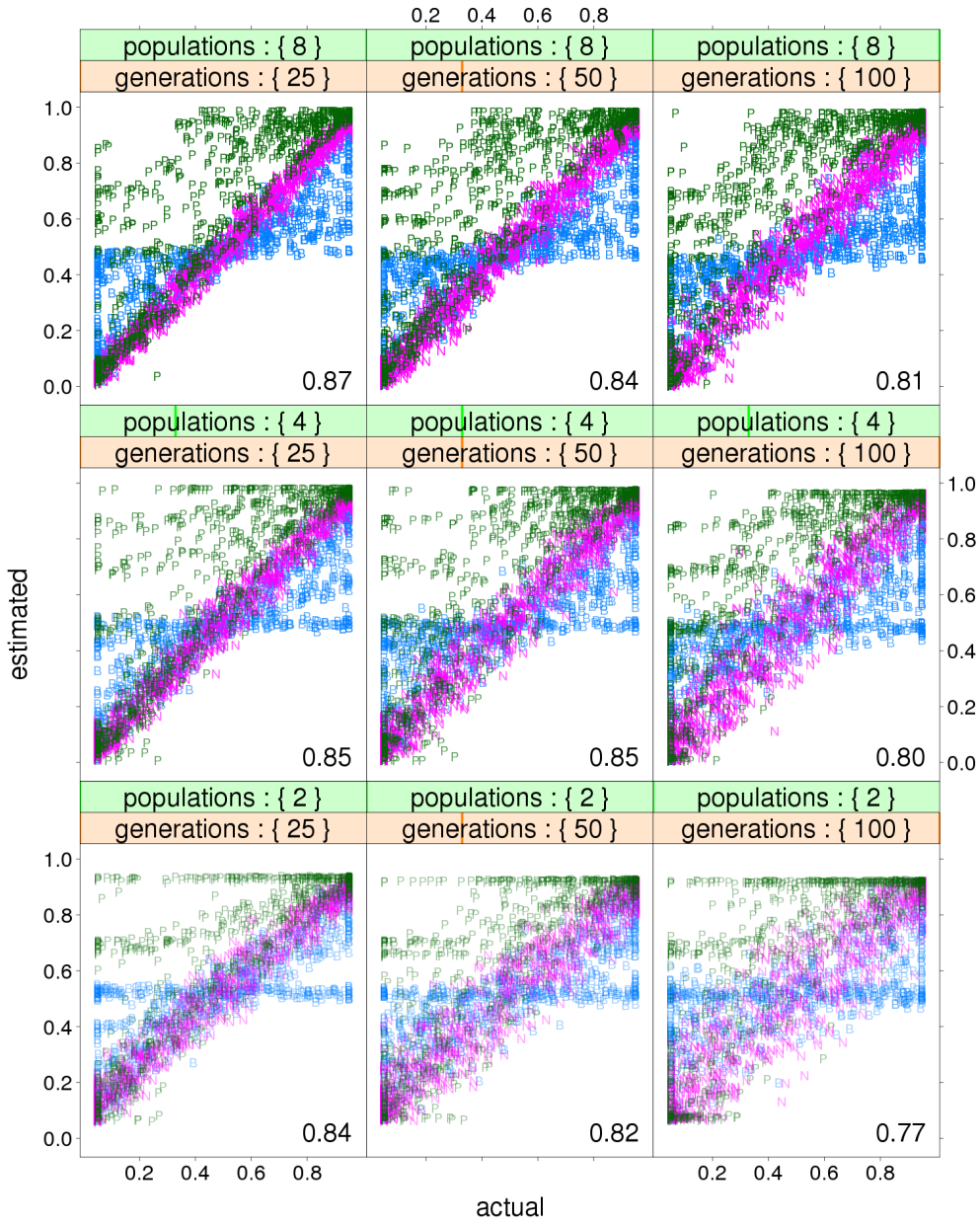
The hierarchical bayesian Nicholson model

- number of alleles $x_{ij} \sim$ Binomial(popsize, α_{ij})
- subpopulation allele frequency $\alpha_{ij} \sim N(\pi_i, c_j\pi_i(1-\pi_i))$
- ancestral allele frequency prior $\pi_i \sim \beta(a, a)$
- population divergence prior $c_j \sim U[0, 1]$

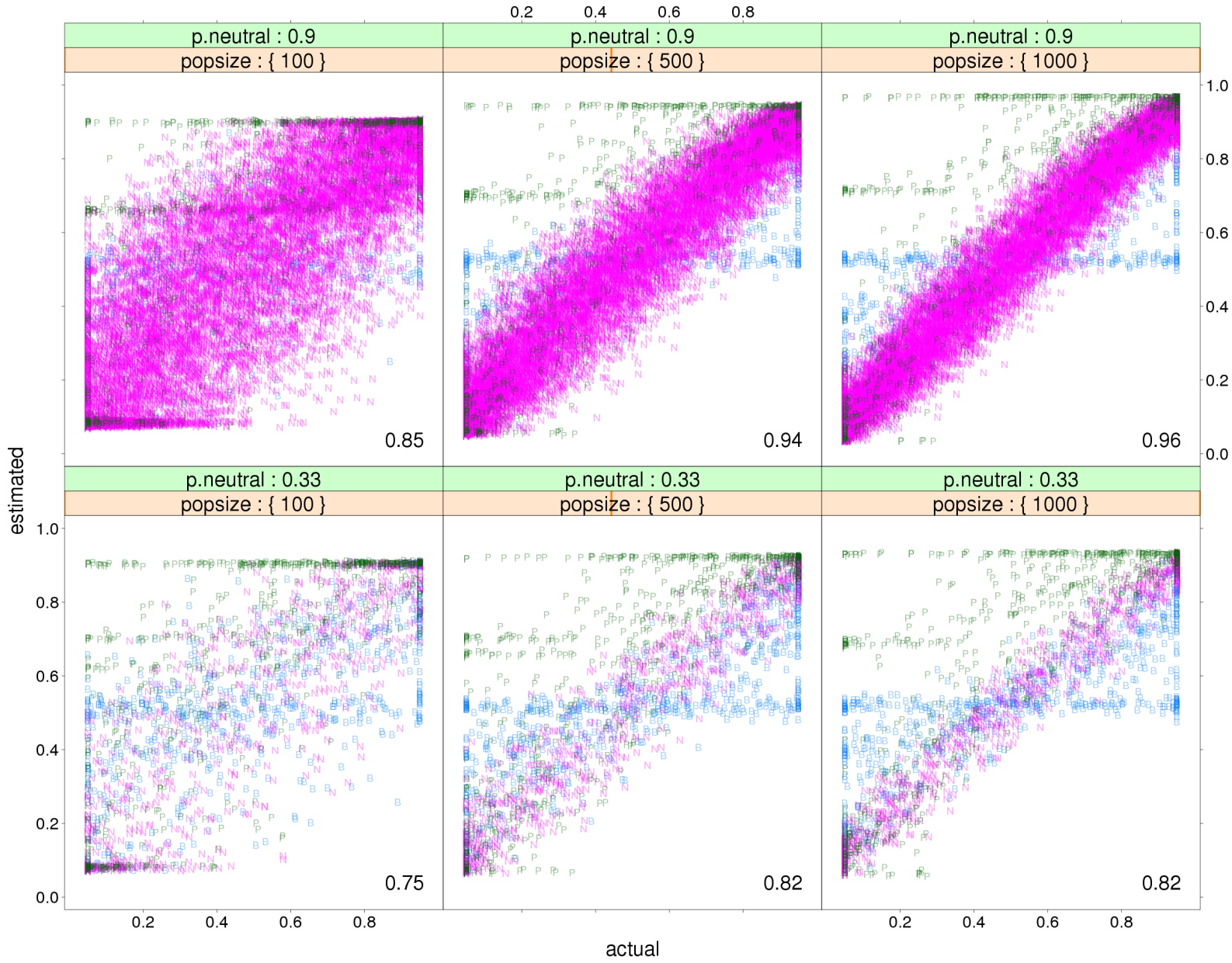


- MCMC sampling:
  1. $\alpha^t = P(\alpha|c^{t-1}, \pi^{t-1}, x)$
  2. $\pi^t = P(\pi|c^{t-1}, \alpha^t, a)$
  3. $c^t = P(c|\pi^t, \alpha^t)$
- Implemented using a Gibbs sampler in a FORTRAN program.
- Run on each simulation, and each subset of loci (all, not.all.fixed, none.fixed), independently.

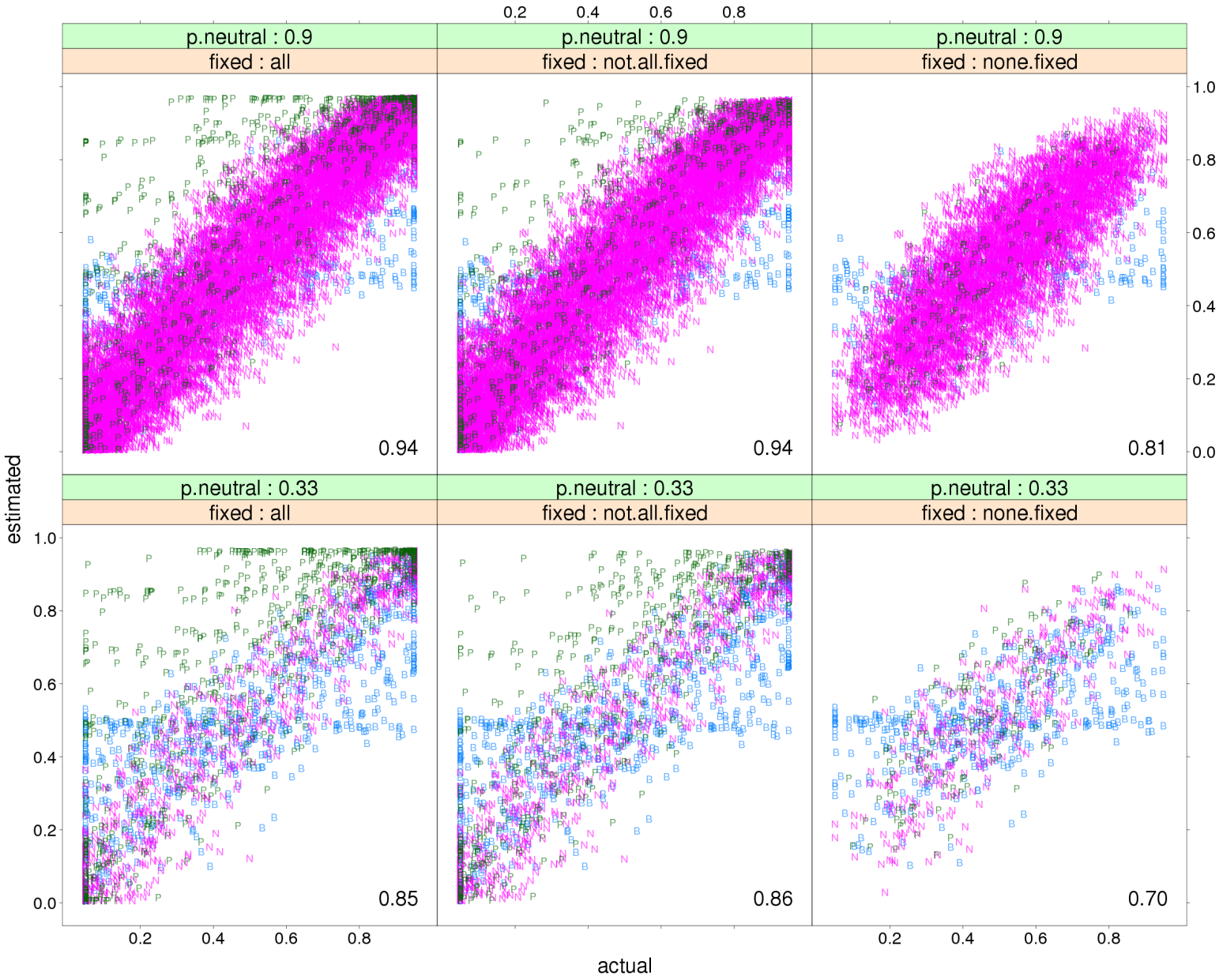
Ancestral allele frequency estimate depends on number of generations and populations



popsize : { 500 } loci : { 50 } fixed : { all } p.neutral : { 0.33 }

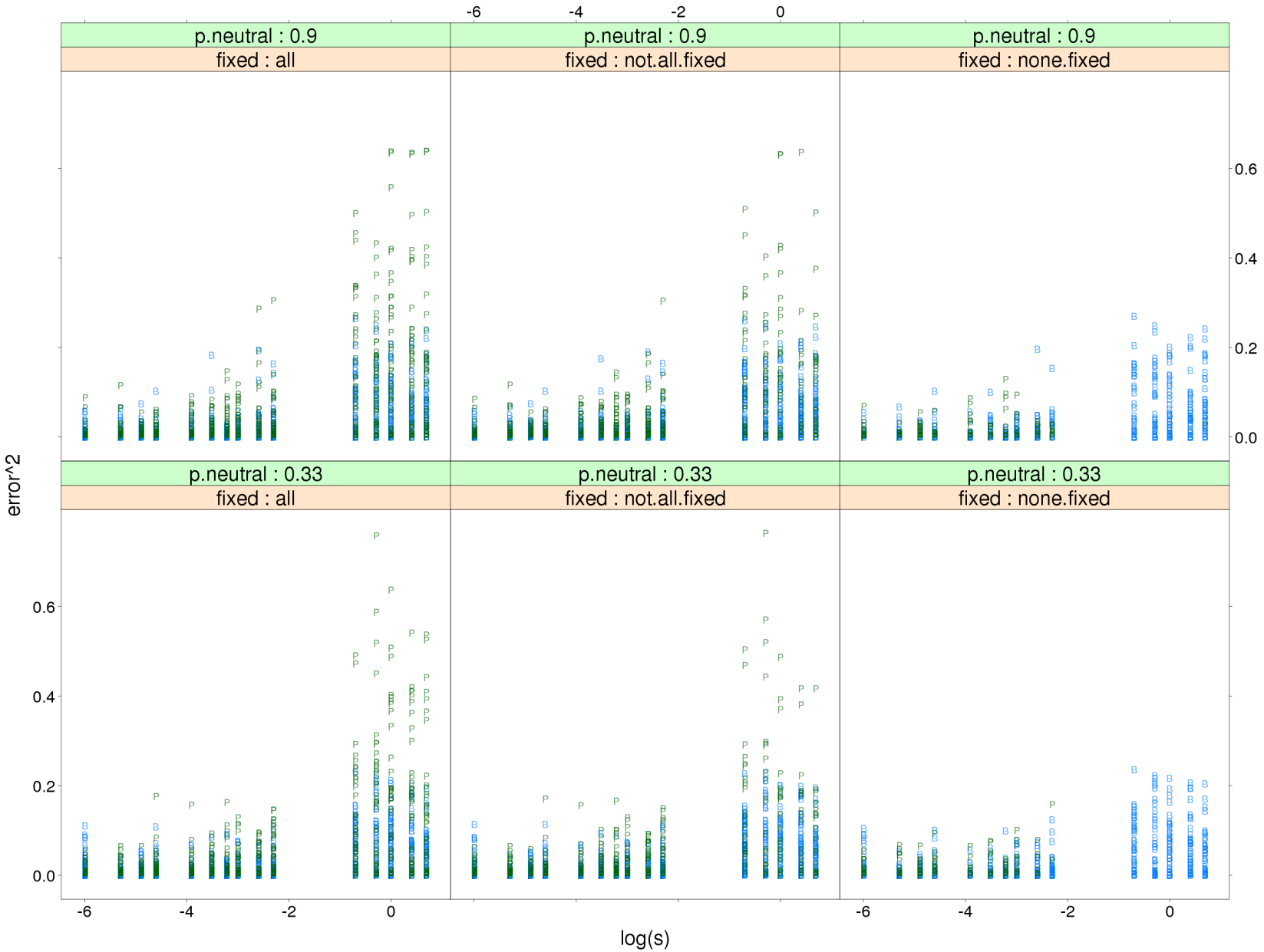

Ancestral allele frequency estimate depends on population size
p.neutral : 0.9
popsize : { 100 }
p.neutral : 0.9
popsize : { 500 }
p.neutral : 0.9
popsize : { 1000 }
p.neutral : 0.33
popsize : { 100 }
p.neutral : 0.33
popsize : { 500 }
p.neutral : 0.33
popsize : { 1000 }
0.85
0.94
0.96
0.75
0.82
0.82
estimated
actual
0.0
0.2
0.4
0.6
0.8
1.0
populations : { 2 } generations : { 50 } loci : { 50 } fixed : { all }

Ancestral allele frequency is harder to estimate for fixed loci

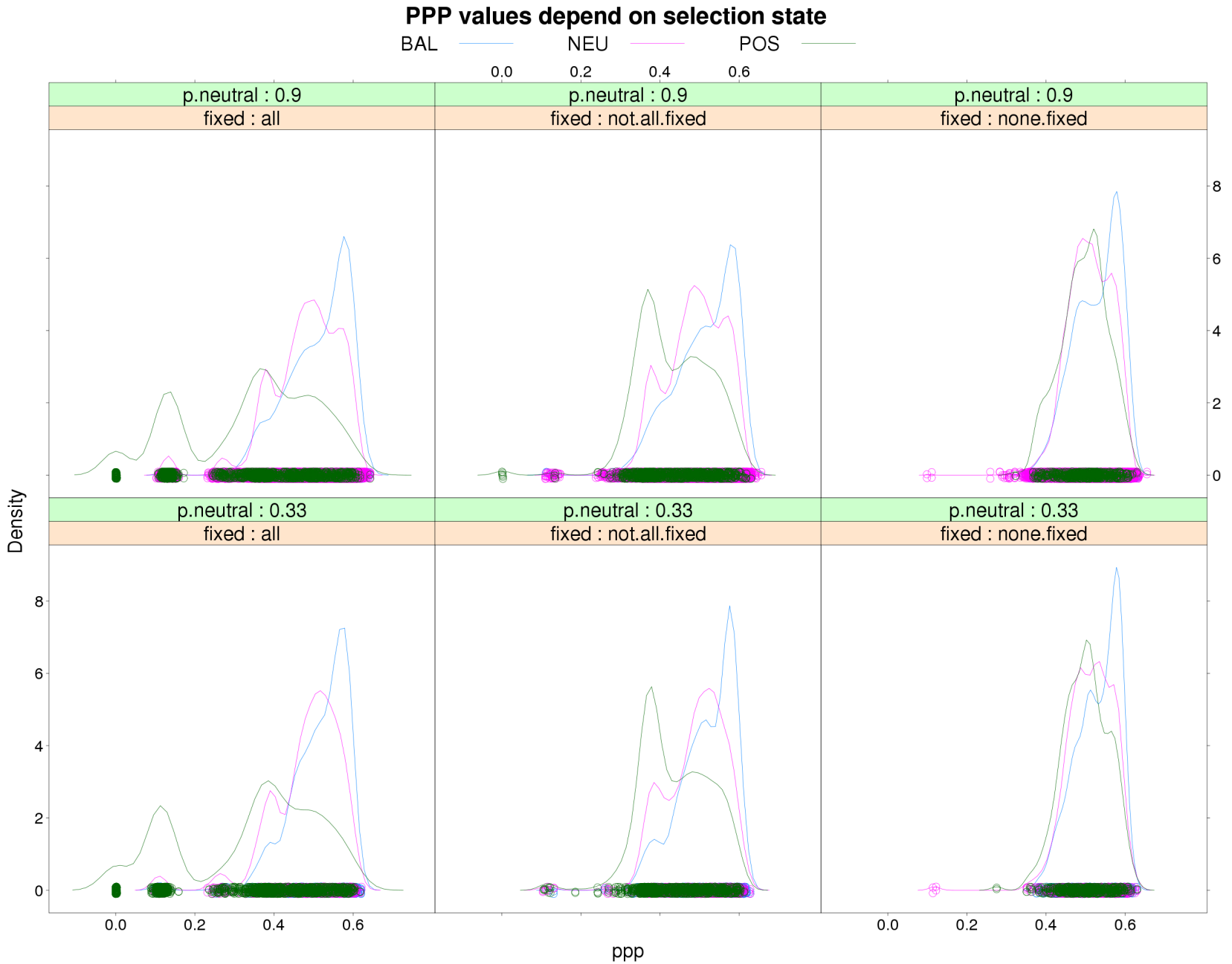


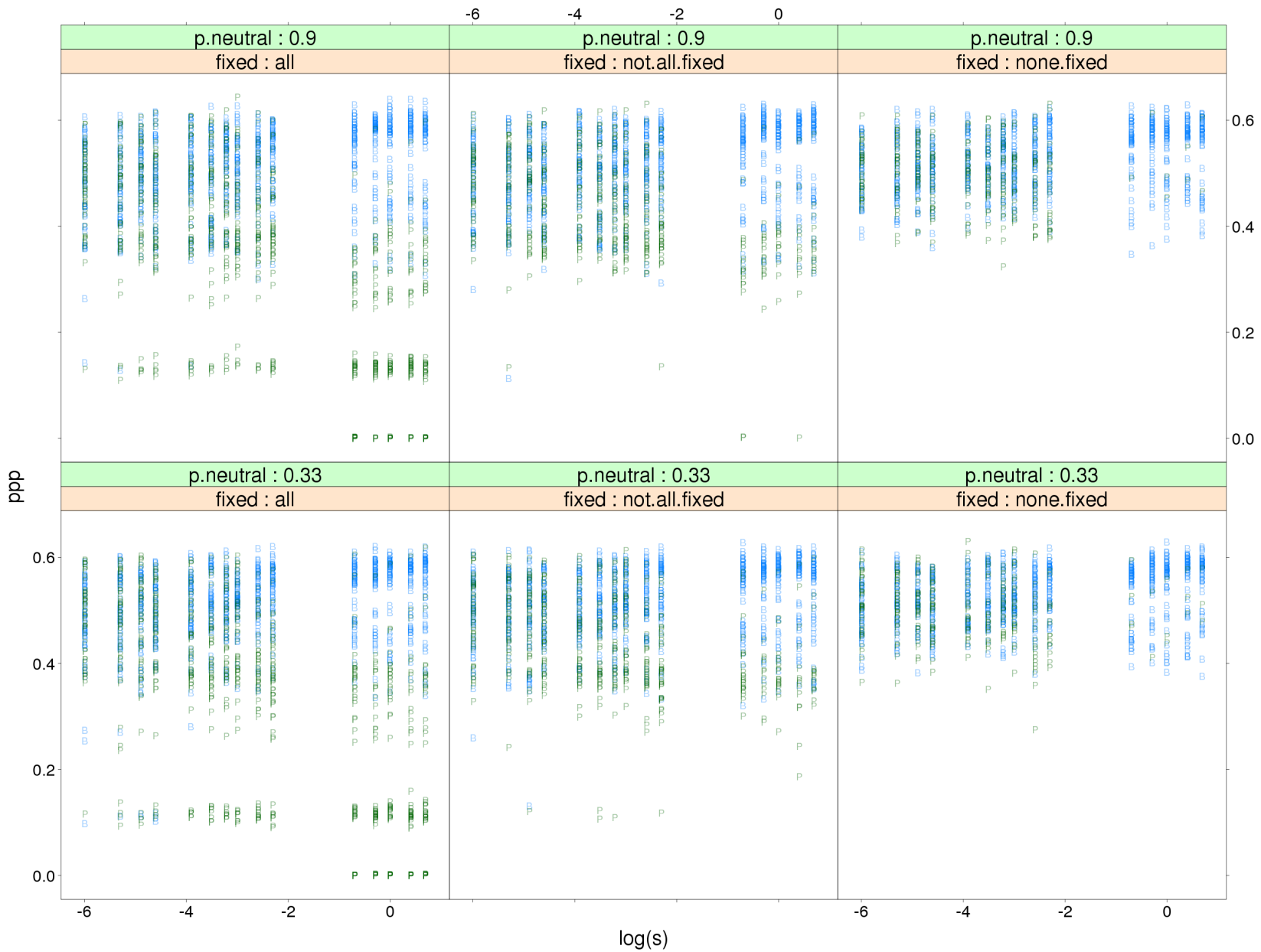
populations : { 4 } generations : { 25 } popsize : { 100 } loci : { 50 }

Error in estimating ancestral allele frequency does not depend on proportion of neutral alleles



populations : { 4 } generations : { 25 } popsize : { 100 } loci : { 50 }


PPP values depend on selection state
BAL
NEU
POS
p.neutral : 0.9
fixed : all
p.neutral : 0.9
fixed : not.all.fixed
p.neutral : 0.9
fixed : none.fixed
p.neutral : 0.33
fixed : all
p.neutral : 0.33
fixed : not.all.fixed
p.neutral : 0.33
fixed : none.fixed
Density
ppp
0.0
0.2
0.4
0.6
0
2
4
6
8
populations : { 2 } generations : { 50 } popsize : { 500 } loci : { 50 }


PPP values depend on selection parameter
p.neutral : 0.9
fixed : all
p.neutral : 0.9
fixed : not.all.fixed
p.neutral : 0.9
fixed : none.fixed
p.neutral : 0.33
fixed : all
p.neutral : 0.33
fixed : not.all.fixed
p.neutral : 0.33
fixed : none.fixed
ppp
log(s)
populations : { 2 } generations : { 50 } popsize : { 500 } loci : { 50 }

Next steps

- Try more focused simulations, smaller number of distinct s values, more replications.
- Examine distribution of $\hat{c}_j$.
- Examine distribution of end allele frequencies and fixation, conditional on initial population allele frequencies and selection parameter s_i.
- Trace false positive and false negative curves using a simple ppp-value cutoff rule.
- Introduce model parameters for selection (Nicholson model assumes only drift).