Simulation and modeling of genomic signatures of selection in order to identify functional genes in domestic animals

Toby Dylan Hocking

INRA Jouy-en-Josas, Génétique Animale et Biologie Intégrative

17 June 2009

Outline

Education, work experience, qualifications, and interests

Genomic differentiation and selection of domestic animals

Education and internships 2002-2006

- Bachelor of Arts, University of California, Berkeley, May 2006
  - Major in Molecular and Cell Biology, emphasis Genetics and Development
  - Major in Statistics, with honors thesis for research with Dr. Terry Speed on chromosomal copy number analysis on Single Nucleotide Polymorphism (SNP) microarray data
- Research internships
  - Expression microarray analysis, Lawrence Berkeley National Laboratory (LBNL), with Dr. Saira Mian, 2003-2004
  - Human tissue culture and molecular biology, LBNL, with Dr. Chris Patil, 2004
  - Fungal genetics, Genencor International, with Dr. Huaming Wang, 2005

Work experience 2006-2008

- Research assistant at Sangamo BioSciences, with Dr. Jeff Miller
  - Biochemistry experiments to determine DNA-binding sequnces of zinc finger proteins
  - Doyon Y, McCammon JM, Miller JC, Faraji F, Ngo C, Katibah GE, Amora R, **Hocking TD**, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Amacher SL. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases, *Nature Biotechnology* 26, 702-708 (2008).
  - Designed linker sequences for chimeric nucleases and tested their specificity and activity using a yeast reporter system
  - Implemented an interactive database/webserver for statistical analysis and visualization (open sourced a plotting framework)

Skills, current work 2008-2009

- Language skills: English (mother tongue) and French (spoken since 2007, living in Paris since August 2008)
- Programming skills: C, Perl, Python, R, SAS, HTML, SQL, PHP, CSS, LaTeX, Subversion
- Master 2 Statistics at University of Pierre and Marie Curie, Paris 6, director Paul Deheuvels
- Currently doing a research internship with Drs. Mathieu Gautier, Jean-Louis Foulley, and Gilles Celeux at INRA/INRIA

Outline

Education, work experience, qualifications, and interests

Genomic differentiation and selection of domestic animals

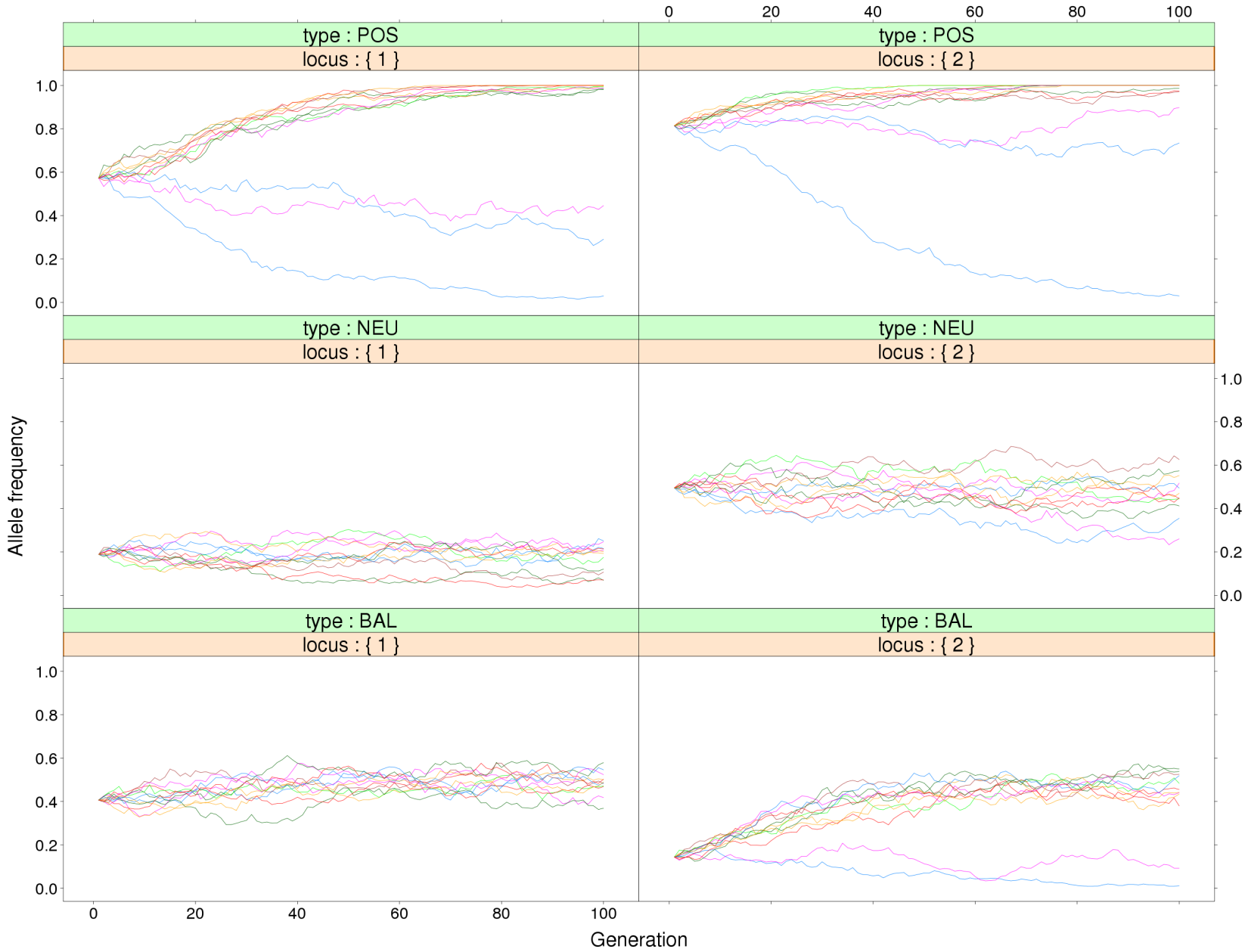
Need to identify functional genes motivates study of domestic animals

- The Bovine Genome Sequencing and Analysis Consortium, Christine G. Elsik, Ross L. Tellam, Kim C. Worley, et al. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution, *Science*, 2009 Apr 24; 324:522-528.
- Domestic animals (not just cows) have been subjected to domestication and natural selection for thousands of years.
- Selection of favorable phenotypes: resistance to disease, milk production, meat production, etc.
- How can we characterize the genomic regions and genes that correspond to this selection?
- Example: climate change may introduce an African disease to Europe, we would like to identify which genes in African cattle are responsible for resistance.

Toward a signature of selection

- We can inexpensively genotype a cow at 60,000 SNPs using microarrays, 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?

Allele frequency variance over time can be used to distinguish different selection types



Models of evolution and genetic differentiation

- 4 evolutive forces: drift, selection, migration, and mutation.
- Several types of selection:
- Positive selection acts to favor homozygotes, thus increasing the allele frequency.
- Balancing selection favors heterozygotes, thus tending to maintain several favorable alleles, evolving allele frequencies towards $1/2$.
- The "signature of selection" will enable identification of loci under selection and estimation of the strength of selection.

Bayesian statistical models of evolution

- Hierarchical Bayesian models offer a robust framework for studying genetics.
- Models can be fit using Markov Chain Monte Carlo (MCMC) techniques.
- Some current models work well but only consider drift (Nicholson *et al.* 2002).
- Examine which loci do not fit, to infer which loci are not consistent with the pure-drift model.
- Some other models attempt to estimate selection coefficients (Beaumont and Balding 2004).
- With more complex genetic models of evolution (selection, migration), our statistical models become less tractable.
- In very complex genetic models (Kimura equations) it is impossible to write the likelihood.
- But we can still simulate the data, so potentially can use Approximate Bayesian Computation (ABC).

Conclusion: steps toward a signature of selection

- Introduce parameters for selection into a hierarchical Bayesian model.
- Design a per-locus "signature of selection" for behavior different than a neutral allele.
- Most current models assume independence of loci, which is false.
- Use Approximate Bayesian Computation (ABC) to estimate selection parameters when it is impossible to write the likelihood.
- Model the ascertainment bias inherent in microarray design.

Merci pour votre attention

- Questions?

- Supplementary slides follow

Why Bayesian models?

- Structured way to model different sources of variation
- Posterior distributions (and credible interval) for parameters, rather than point estimates
- Hierarchical structure for shrinkage of parameters towards zero (= sparsity assumption)
- Can be computationally efficient through empirical Bayesian approach
- Incorporate biological information into priors
- Avoid overfitting that occurs with ML methods

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

- Simulate the evolution using known evolution parameters, fit the model, then look for signatures of selection in the alleles we know were under selection.
- 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.

Evolution equations

- locus i, population j, time t
- blue allele frequency $\alpha_{ij}(t)$
- genetic drift $\alpha^*_{ij}(t) = \text{rbinom}(\text{popsize}, \alpha_{ij}(t-1))/\text{popsize}$
- relative fitness of genotypes

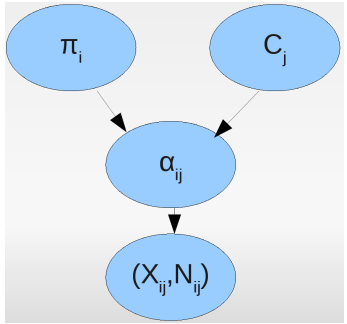
w_{ij}^{BB}	w_{ij}^{BR}	w_{ij}^{RR}	i	j
1	$1 + s_i/2$	$1 + s_i$	positive	red
$1 + s_i$	$1 + s_i/2$	1	positive	blue
1	$1 + s_i$	1	balancing	
1	1	1	neutral	

- allele frequency updated for selection based on Hardy-Weinberg equilibrium:

$$\alpha_{ij}(t) = \frac{w_{ij}^{BB}\alpha^*_{ij}(t)^2 + w_{ij}^{BR}\alpha^*_{ij}(t)[1-\alpha^*_{ij}(t)]/2}{w_{ij}^{BB}\alpha^*_{ij}(t)^2 + w_{ij}^{BR}\alpha^*_{ij}(t)[1-\alpha^*_{ij}(t)] + w_{ij}^{RR}[1-\alpha^*_{ij}(t)]^2}$$

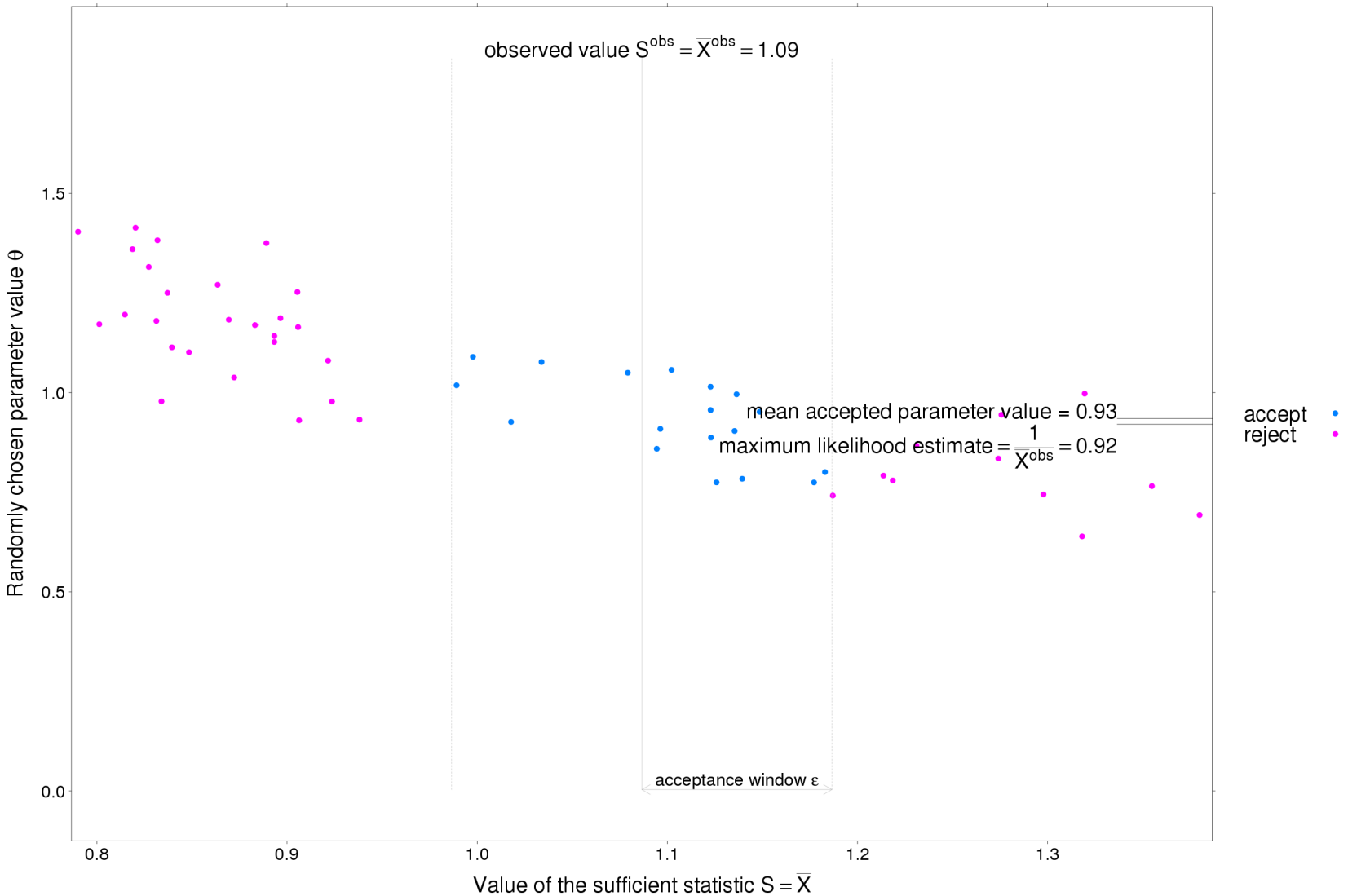
The hierarchical bayesian Nicholson model

- 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 differentiation prior** $c_j \sim U[0, 1]$



- Monte Carlo Markov Chain 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.

Approximate Bayesian computation yields a posterior parameter distribution by rejecting distant simulated parameter values



1. Generate a parameter value $\theta \sim \pi(.) = U[0, 1.84]$
2. Generate a dataset $Y^{sim} \sim f(.|\theta) = Exp(\theta)$ using this parameter value
3. Compute difference $L(S^{sim}, S^{obs})$ between sufficient statistics of simulated and observed data
4. If the difference $L < \varepsilon$ is small enough, accept this parameter value

(original sample: 100 observations taken from a Exp(1) distribution)