



University of Wisconsin-Madison

**A TWO-STEP METHOD FOR
DETECTING SELECTION SIGNATURES**

Daniel Gianola

(with Henner Simianer and Saber Qanbari)

Georg-August-Universität Göttingen



ACKNOWLEDGMENTS



Alexander von Humboldt

Stiftung / Foundation



- Wisconsin Agriculture Experiment Station
- NSF DMS-044371
- FUGATO (German Ministry of Education and Research)
- Lohmann Tierzucht GmbH
- H.Wilhelm Schaumann Stiftung, Hamburg

INTRODUCTION

- Genetics markers (SNPs) available (millions in humans)
- Variation in gene frequencies among groups can be used to assess “signatures” of forces such as selection
- Examples:
 - ➔ Low vs. high production breeds
 - ➔ Selection lines
 - ➔ Human populations
 - ➔ Cases (sick) vs. controls (healthy)

Our interests may include:

- Identify genomic regions associated with a trait
- Use such knowledge in marker-assisted breeding programs
- Find markers or genes associated with variation in disease traits + use this in individualized medicine (“personalized” medicine)
- Compare allelic frequencies between efficient and less efficient strains of animals (nutri-genomics)

$F_{ST}=\theta$ STATISTIC

(metric for measuring variation in allelic frequencies between populations)

old

- WRIGHT (1931, 1951)
- LEWONTIN AND KRAKAUER (1973)
- COCKERHAM (1969, 1973), NEI (1973)

new

- HOLSINGER AND WEIR (2009) review in *Nature Genetics Reviews*
- AKEY (2009) review in *Genome Research*

BRIEF TOUR OF F-STATISTICS

- Measure relatedness between alleles in a sub-population, relative to that in an undivided (e.g., ancestral) population

EQUIVALENTLY:

- Measure dispersion in gene frequencies among groups relative to variation expected in population from which such groups derived

Linear model formalism (Cockerham, 1969, 1973)

Notation:

$l=1,2,\dots,L$	<i>Denotes locus l</i>
$r=1,2,\dots,R$	<i>Denotes population or “replicate” r</i>
i	<i>Denotes individual</i>
j	<i>Denotes within-individual deviate</i>

$$x_{r\bar{i}j,l} = \begin{cases} 1 & \text{if an allele is } A_l \\ 0 & \text{otherwise.} \end{cases}$$

Bi-allelic locus (SNP):
in undivided population


$$p_l = \Pr(A_l)$$

$$1 - p_l = \Pr(a_l)$$

$$x_{rij,l} = p_l + a_{r,l} + b_{ri,l} + w_{rij,l},$$

where p_l is as before and $a_{r,l} \sim (0, \sigma_{a,l}^2)$, $b_{ri,l} \sim (0, \sigma_{b,l}^2)$, and $w_{rij,l} \sim (0, \sigma_{w,l}^2)$

uncorrelated



$$E(x_{rij,l}) = p_l,$$



$$Var(x_{rij,l}) = p_l(1 - p_l) = \sigma_{a,l}^2 + \sigma_{b,l}^2 + \sigma_{w,l}^2 = \sigma_l^2$$

Covariance structure between alleles

$$Cov(x_{rij,l}, x_{r'i'j',l}) = \begin{cases} \sigma_l^2 & \text{if } r = r', i = i', j = j' \\ \sigma_a^2 & \text{for alleles drawn from different individuals in the same replicate} \\ \sigma_{a,l}^2 + \sigma_{b,l}^2 & \text{for alleles of the same individual (over all replicates)} \\ Cov(a_r, a_{r'}) & \text{if replicates are correlated somehow.} \end{cases}$$

A standard assumption is $Cov(a_r, a_{r'}) = 0$. The following correlations (all positive) follow.

Correlation structure between alleles

- Pairs of alleles drawn at random from different individuals in the same group are correlated as

$$\rho_{a,l} = \frac{\sigma_{a,l}^2}{\sigma_{a,l}^2 + \sigma_{b,l}^2 + \sigma_{w,l}^2} = \theta_l = F_{ST,l}, \quad (2)$$

so $0 \leq \theta_l \leq 1$ for all l .

- Pairs of alleles drawn within individuals over all replicates bear a correlation equal to

$$\rho_{ab,l} = \frac{\sigma_{a,l}^2 + \sigma_{b,l}^2}{\sigma_{a,l}^2 + \sigma_{b,l}^2 + \sigma_{w,l}^2} = F_l = F_{IT,l}$$

where F is the total inbreeding coefficient, also known as F_{IT} (e.g., Weir and Hill, 2002).

- The correlation between alleles within individuals within the same replicate is

$$\rho_{b,l} = \frac{\sigma_{b,l}^2}{\sigma_{b,l}^2 + \sigma_{w,l}^2} = f_l = F_{IS,l}$$

which is the within sub-population inbreeding coefficient f .

Wright's F-statistics

Relationships between F values:

$$\begin{aligned} F_{IT,l} &= \frac{\sigma_{a,l}^2}{\sigma_{a,l}^2 + \sigma_{b,l}^2 + \sigma_{w,l}^2} + \frac{\sigma_{b,l}^2}{\sigma_{a,l}^2 + \sigma_{b,l}^2 + \sigma_{w,l}^2} \\ &= \theta_l + \frac{F_{IS,l} (\sigma_{b,l}^2 + \sigma_{w,l}^2)}{\sigma_{a,l}^2 + \sigma_{b,l}^2 + \sigma_{w,l}^2} \\ &= \theta_l + F_{IS,l} (1 - \theta_l), \end{aligned}$$

$$\theta_l = \frac{F_{IT,l} - F_{IS,l}}{1 - F_{IS,l}} = F_{ST,l}.$$

$$\underbrace{1 - F_{IT,l}}_{(a)} = \underbrace{(1 - F_{IS,l})}_{(c)} \underbrace{(1 - F_{ST,l})}_{(b)},$$

- (a) Total loss of heterozygosis
- (b) Loss due to population sub-division (Wahlund's)
- (c) Loss due to within population inbreeding

Important: note that

$$\theta_l = \frac{\sigma_{a,l}^2}{\sigma_{a,l}^2 + \sigma_{b,l}^2 + \sigma_{w,l}^2} = \frac{\sigma_{a,l}^2}{p_l (1 - p_l)}.$$

Consider a **given** realization of gene frequencies as in Nei (1973)



$$\theta_l = \frac{\frac{\sum_{r=1}^R (p_{r,l} - \bar{p}_l)^2}{R}}{\bar{p}_l (1 - \bar{p}_l)},$$

$$\bar{p}_l = \sum_{r=1}^R p_{r,l} / R$$



Making the parameter explicit in all unknown allelic frequencies

$$\theta_l = \frac{\sum_{r=1}^R p_{r,l}^2 - \frac{\left(\sum_{r=1}^R p_{r,l}\right)^2}{R}}{\left(\sum_{r=1}^R p_{r,l} - \frac{\left(\sum_{r=1}^R p_{r,l}\right)^2}{R}\right)},$$

IMPORTANT: This is a parametric definition

ILLUSTRATION

(how the concept is used in practice)

Genome Research
www.genome.org

2002

Interrogating a High-Density SNP Map for Signatures of Natural Selection

Joshua M. Akey,¹ Ge Zhang,¹ Kun Zhang,^{1,2} Li Jin,¹ and Mark D. Shriver^{3,4}

¹Center for Genome Information, University of Cincinnati, Cincinnati, Ohio, USA; ²Human Genetics Center, University of Texas-Houston, Houston, Texas 77225, USA; ³Department of Anthropology, Penn State University, University Park, Pennsylvania 16802, USA

In this work, we describe an analysis of 26,530 SNPs with allele frequencies that were determined in three populations: African-American, East Asian, and European-American.

Lightly colored bars: coalescent simulations under neutrality

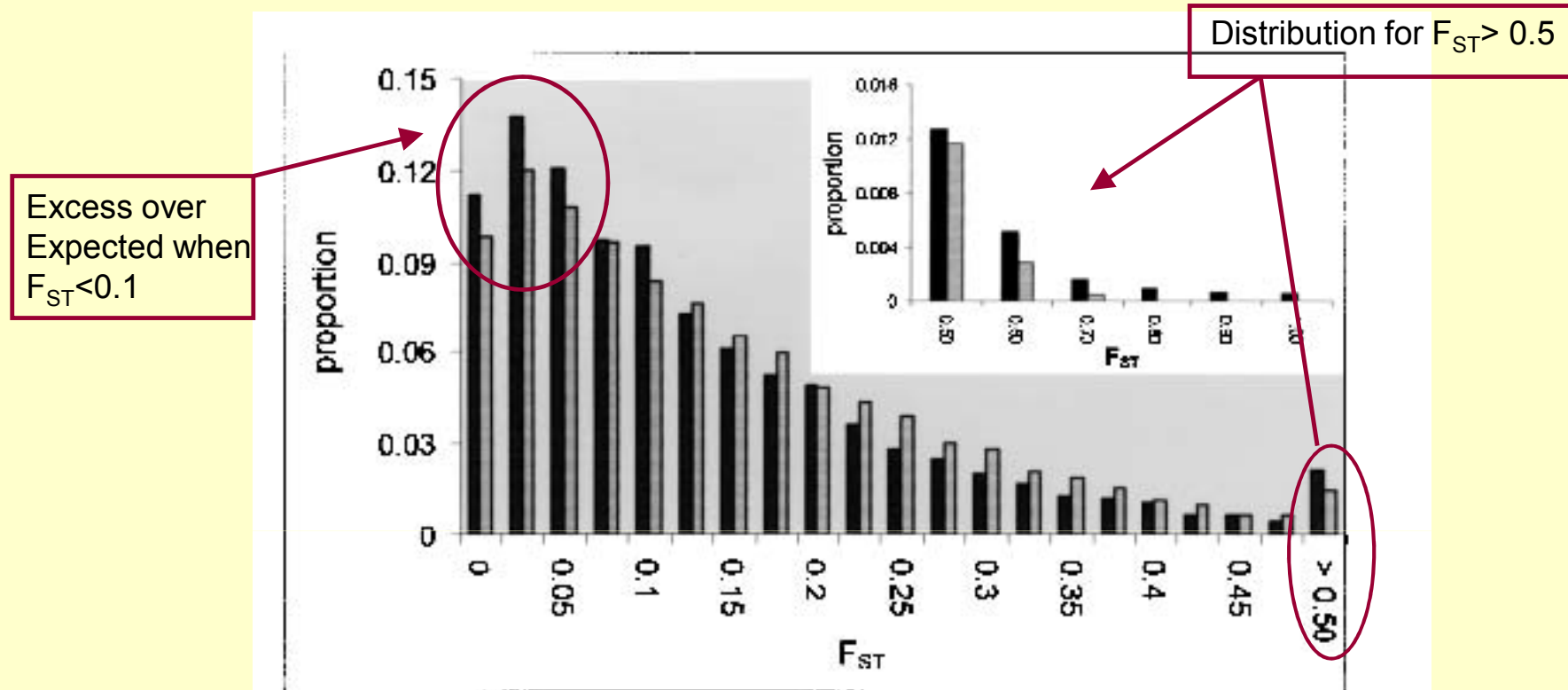
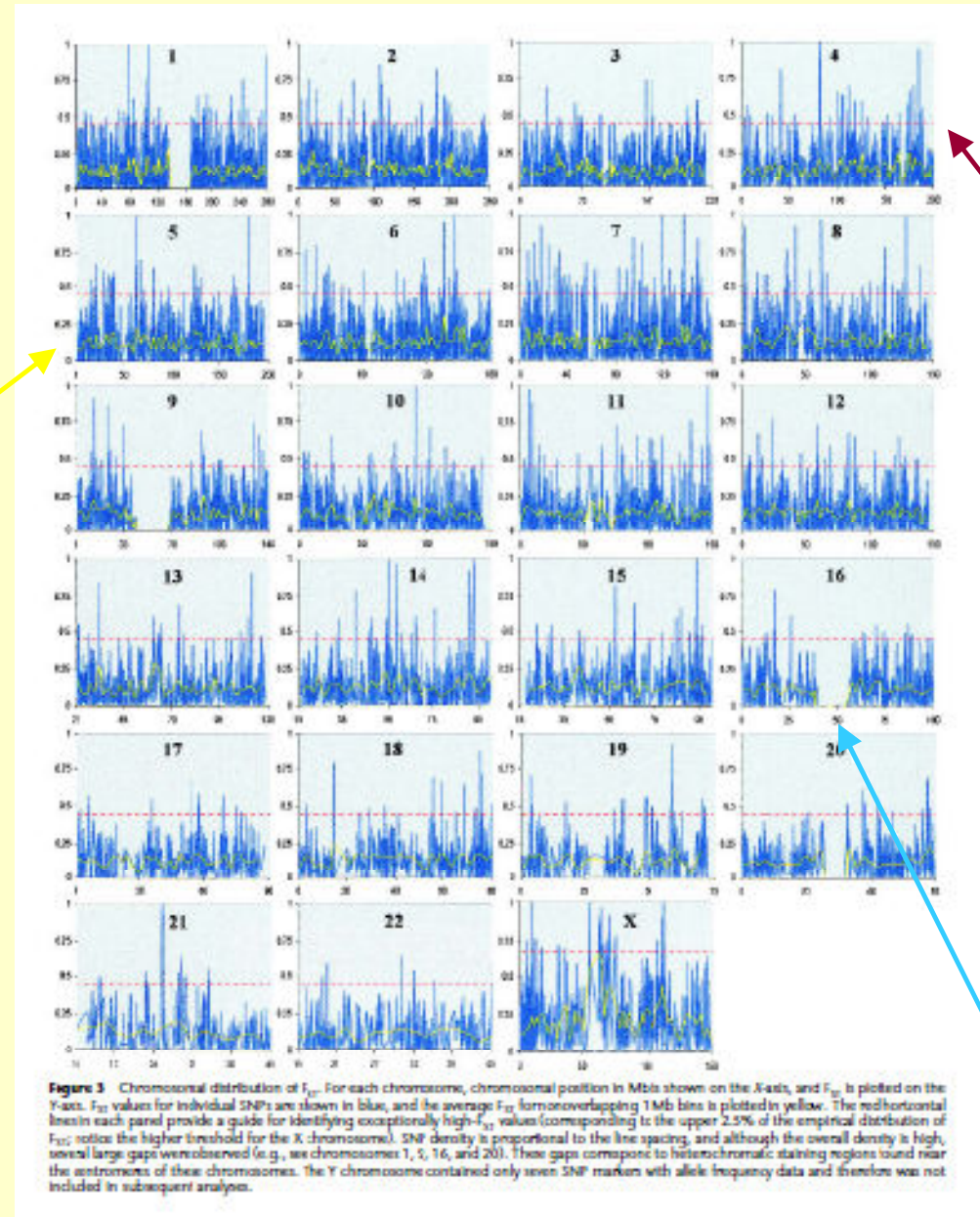


Figure 2 Genome-wide distribution of F_{ST} . Solid bars show the observed distribution of F_{ST} for 25,549 autosomal SNPs. The X chromosome was not included in this analysis because it has a different effective population size compared with that of autosomal markers. Lightly shaded bars represent the simulated distribution of F_{ST} . The inset figure shows the observed and simulated distributions of F_{ST} for values ≥ 0.5 .



3. Yellow lines:
Average F_{ST} for
Non-overlapping
1 Mb bins

1. Upper 2.5% of
empirical
Distribution of
 F -values

2. Gaps: heterochromatic
regions near centromeres

Distribution of F_{ST} by chromosome
(F -values for adjacent SNPs are correlated)

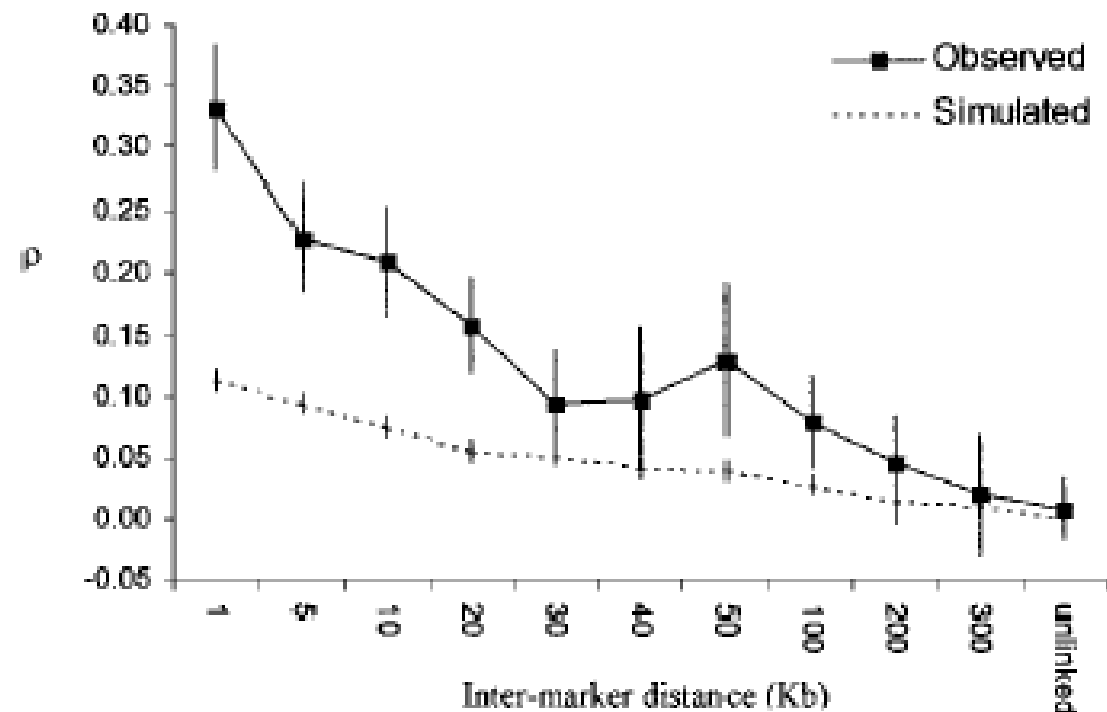


Figure 4 Correlation between F_{ST} values as a function of physical distance. Intermarker distance was calculated between adjacent SNPs across the genome. Marker pairs were then separated into various bins (shown on the X-axis) according to their intermarker distance, and ρ was calculated for each bin. In the observed data, ρ was calculated for unlinked markers by comparing F_{ST} values on different chromosomes. Vertical bars represent 95% confidence intervals.

Simulations under the coalescent understate correlation between F-statistics for linked SNPs

Table 2. Average F_{ST} as a Function of SNP Category

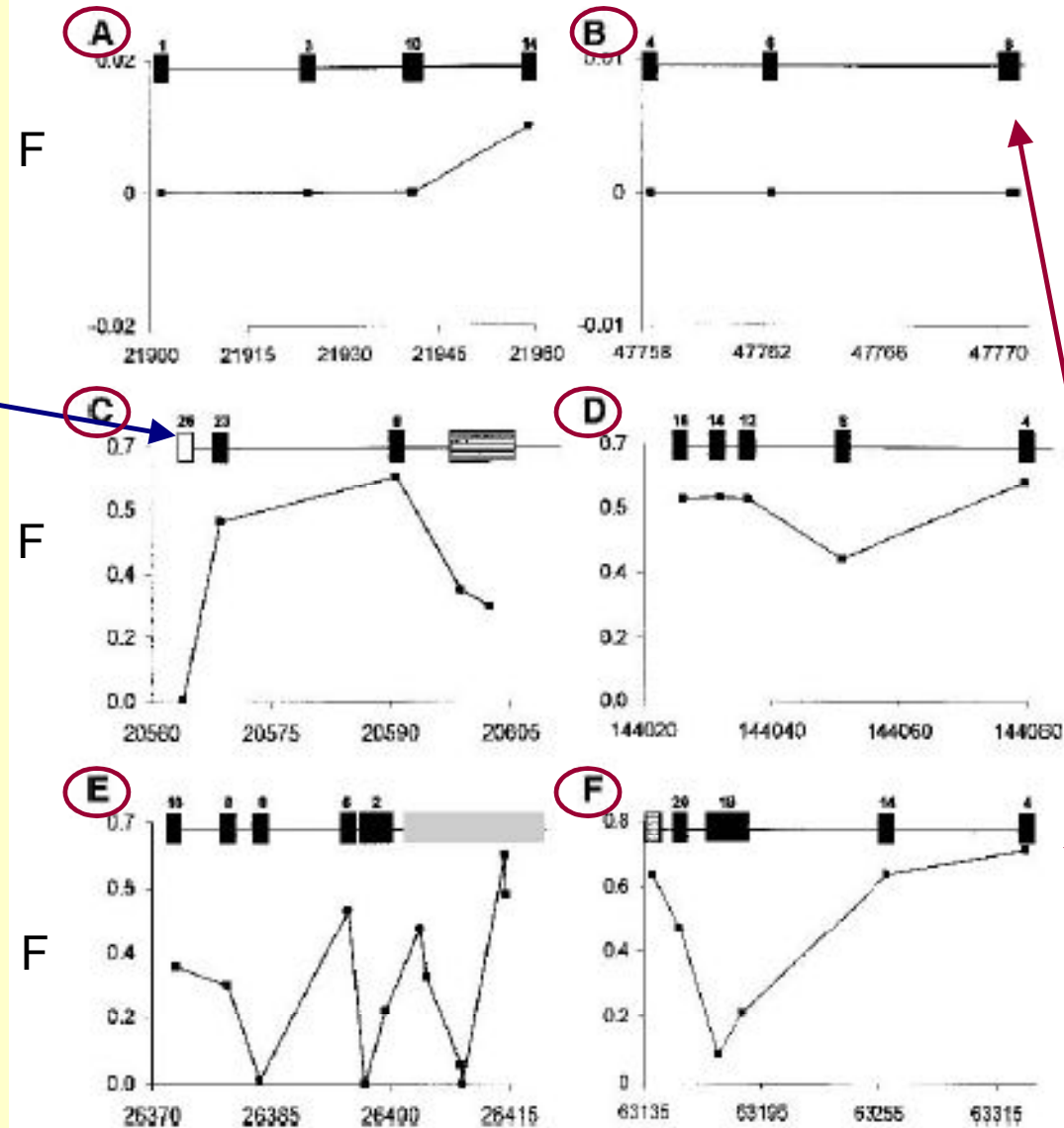
Category	No.	Average F_{ST}	SE	Significance of difference in average F_{ST} ^a	
				Coding	Intronic
Coding	238	0.107	0.008	–	–
Intronic	5,455	0.118	0.002	0.094	–
Noncoding	13,615	0.123	0.001	0.024	0.008

^aEmpirical P values were determined by randomly permuting F_{ST} values between SNP categories 10,000 times and then counting the number of permutations with difference in average F_{ST} equal to or greater than the original difference.

Intron: region within gene not translated into protein

Non-coding: no instructions for making protein

Exon



Low

High

Introns

High+low

Figure 5 F_{ST} profiles for six genes showing signatures of natural selection. For each gene, F_{ST} is plotted on the Y-axis, and chromosomal position in Kb is plotted on the X-axis. The genes shown here include guanine nucleotide exchange factor for Rap1 (*GFR*; (A)), tropomodulin 3 (*TMOD3*; (B)), apolipoprotein B (*APOB*; (C)), phosphoinositide-3-kinase, catalytic, β -polypeptide (*PK3CB*; (D)), cytidine monophosphate-N-acetylneuraminic acid hydroxylase (*CMAH*; (E)), and oligophrenin 1 (*OPHN1*; (F)). The location of SNPs within each gene is denoted as boxes: Introns (black), exons (open), 5' UTR (grey), 5' upstream (vertically striped), and 3' downstream (hatched). Intron and exon numbers are noted within each box where appropriate.

CANDIDATE GENES IDENTIFIED ON F_{ST} VALUES SUGGESTIVE OF SELECTION

Table 3. Molecular Function of Candidate Selection Genes

Gene ontology term	High F_{ST}	Low F_{ST}
Total number terms	183	31
Apoptosis regulator	1 (0.5%)	0 (0.0%)
Cell adhesion molecule	4 (2.2%)	0 (0.0%)
Cell growth and/or maintenance	2 (1.1%)	0 (0.0%)
Chaperone	2 (1.1%)	0 (0.0%)
Defense/immunity protein	3 (1.6%)	2 (6.5%)
Enzyme	50 (27.3%)	5 (16.1%)
Hydrolase	11 (6.0%)	3 (9.7%)
Kinase	11 (6.0%)	0 (0.0%)
Transferase	12 (6.6%)	1 (3.2%)
Enzyme regulator	4 (2.2%)	0 (0.0%)
Ligand binding or carrier	57 (31.1%)	9 (29.0%)
Calcium binding	7 (3.8%)	0 (0.0%)
Nucleic acid binding	23 (12.6%)	1 (3.2%)
Protein binding	3 (1.6%)	6 (19.4%)
Motor	0 (0.0%)	1 (3.2%)
Signal transducer	27 (14.8%)	10 (32.3%)
Ligand	4 (2.2%)	1 (3.2%)
Receptor	14 (7.7%)	7 (22.6%)
Structural molecule	6 (3.3%)	2 (6.5%)
Transcriptional regulator	9 (4.9%)	1 (3.2%)
Transporter	18 (9.8%)	1 (3.2%)

In the Gene Ontology (GO) classification system, a parent term can have multiple subcategories, or children terms (indented text). For instance, hydrolase, kinase, and transferase are the children of the parent term enzyme. A single gene can have multiple parent and children terms (see Ashburner et al. 2000 for more specific information). Note that percentages sum to 100% for parent terms only.

Table 4. Biological Processes of Candidate Selection Genes

Gene ontology term	High F_{ST}	Low F_{ST}
Total number of terms	123	39
Behavior	2 (1.6%)	1 (2.6%)
Cell communication	38 (30.9%)	15 (38.5%)
Cell adhesion	6 (4.9%)	2 (5.1%)
Cell-cell signaling	2 (1.6%)	1 (2.6%)
Response to external stimulus	7 (5.7%)	3 (7.7%)
Immune response	1 (0.8%)	2 (5.1%)
Perception of external stimulus	6 (4.9%)	0 (0.0%)
Signal transduction	21 (17.1%)	7 (18.0%)
Cell growth and/or maintenance	69 (56.1%)	11 (28.2%)
Metabolism	43 (35.0%)	6 (15.4%)
Protein metabolism and modification	15 (12.2%)	0 (0.0%)
Transcription	9 (7.3%)	2 (5.1%)
Transport	12 (9.8%)	0 (0.0%)
Death	1 (0.8%)	2 (5.1%)
Developmental processes	10 (8.1%)	5 (12.8%)
Embryogenesis and morphogenesis	3 (2.4%)	5 (12.8%)
Epigenetic control of gene expression	2 (1.6%)	0 (0.0%)
Reproduction	1 (0.8%)	0 (0.0%)
Physiological processes	3 (2.4%)	5 (12.8%)
Pregnancy	1 (0.8%)	0 (0.0%)

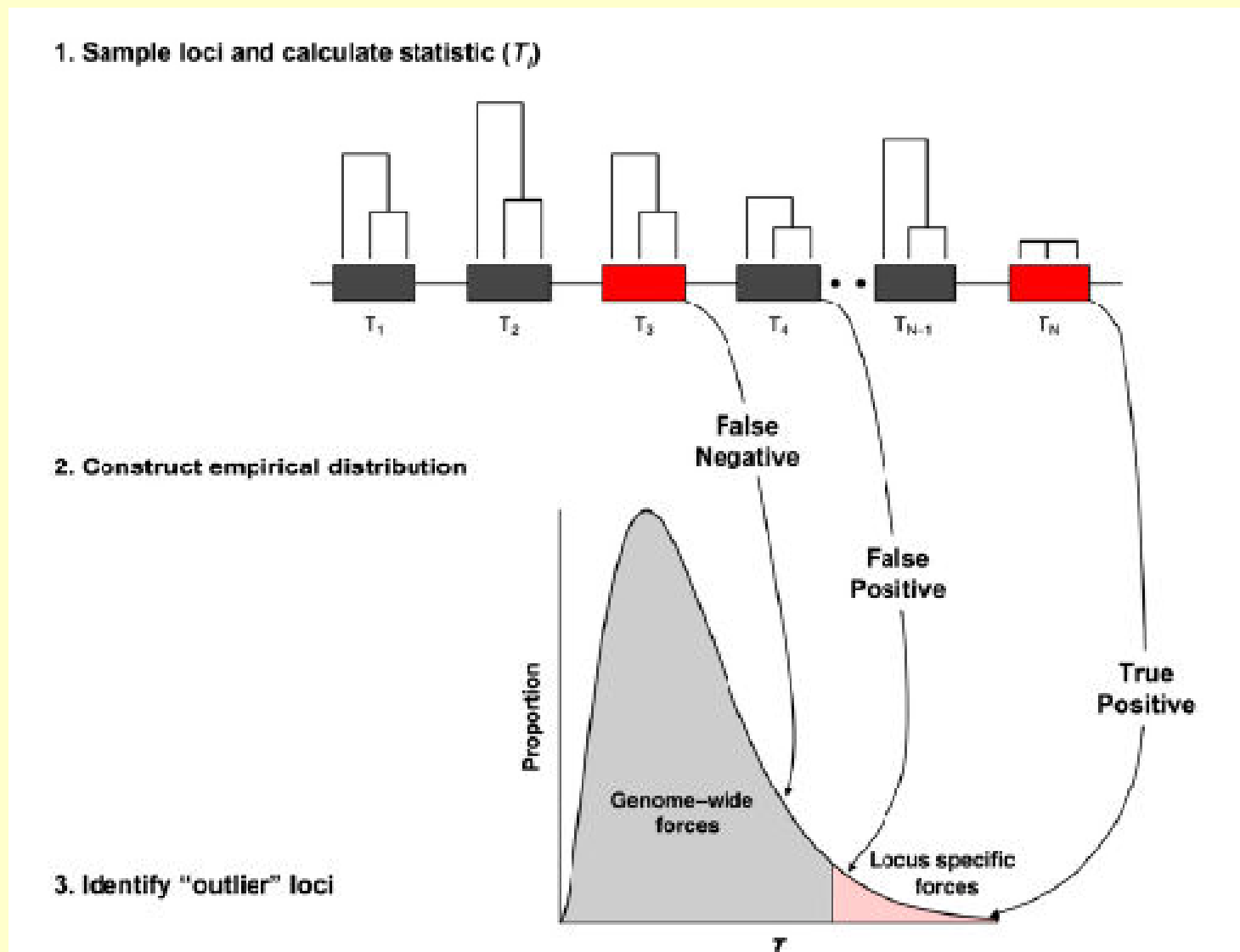
See notes to Table 3.

STATISTICAL INFERENCE

- Given a set of loci, assume that all follow same demographic history and patterns of migration
- If all loci are neutral and have same mutation rates, can be viewed as realizations of the same evolutionary process
- Under selective neutrality, distribution of F-values determined entirely by drift
- Outliers regarded as “**selection signatures**”
 - ➔ Low values: balancing selection (Cavalli-Sforza, 1966)
 - ➔ High values: selection favor some alleles in some populations (milk yield: Holsteins; mastitis: NRF)

VERY ARBITRARY!

Population genomics standard design for selection signatures (Akey, 2009)



Meta-map of selection signatures (Akey, 2009): “overlap is underwhelming”

Based on
9 scans

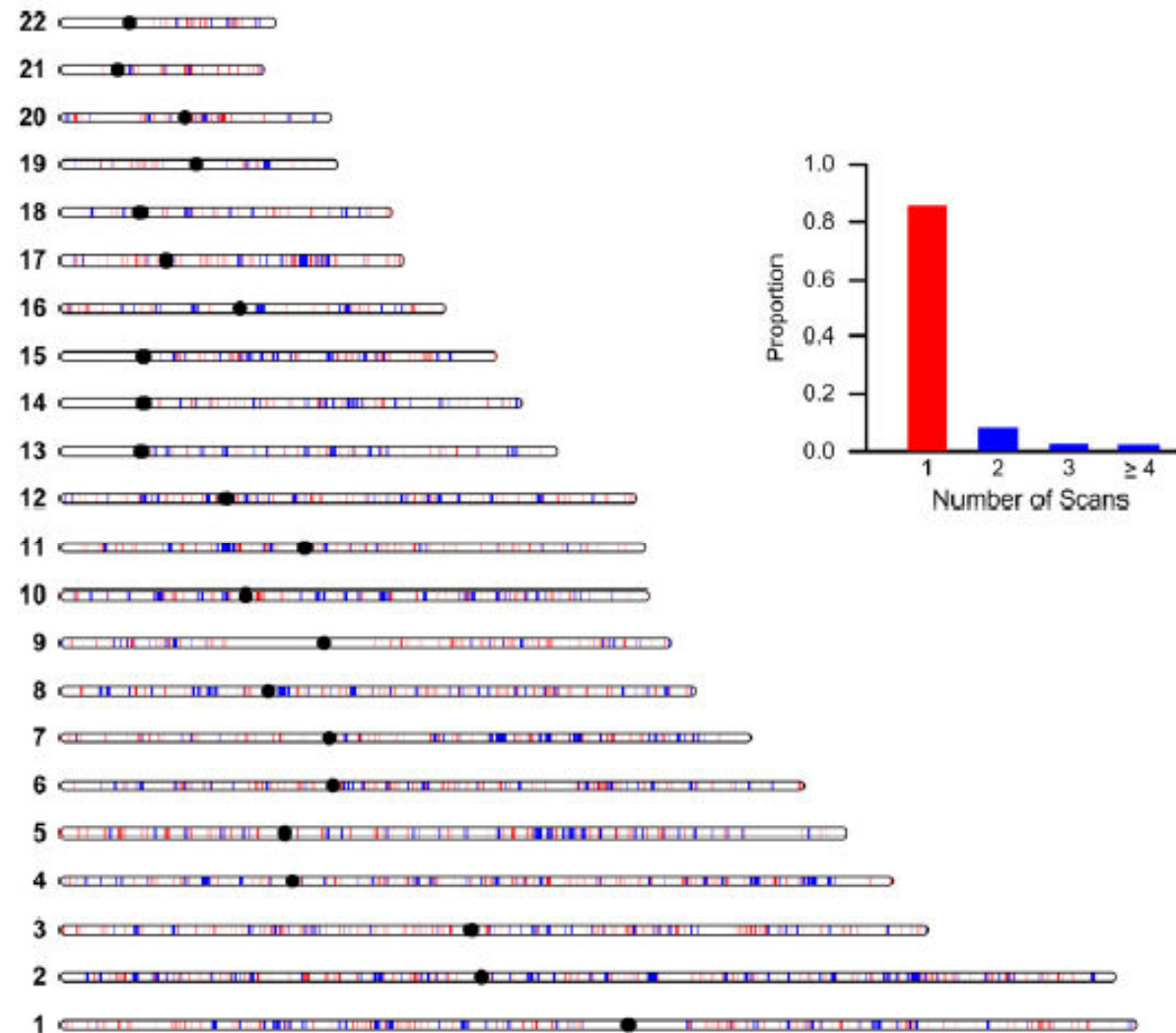


Figure 2. Integrated genomic map of positive selection. Vertical red lines on each autosome indicate loci that were identified in a single genome-wide scan, and blue lines denote regions identified in two or more studies. The histogram shows the proportion of putatively selected loci (y-axis) as a function of the number of genome-wide scans in which they were identified (x-axis).

METHODS OF INFERENCE

- Moments (ANOVA type): crudest, widely used
- Maximum likelihood (asymptotic properties)
- Bayesian: exact finite sample inference (at the expense of priors)

Example of Bayesian method (Holsinger, 1999)

111 Let $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_R)'$ be an $RL \times 1$ vector of allelic frequencies for all R groups,
112 where $\mathbf{p}_r = (p_{r,1}, p_{r,2}, \dots, p_{r,L})'$ has order $L \times 1$. Under the mutual independence assumption, the
113 likelihood conferred by the observed number of copies of alleles to the gene frequencies is

114
$$l(\mathbf{p}|DATA) = \prod_{r=1}^R \prod_{l=1}^L p_{r,l}^{n_{r,A_l}} (1 - p_{r,l})^{n_{r,a_l}} . \quad (5)$$

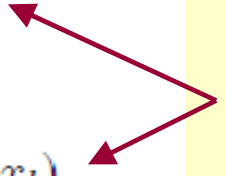
The maximum likelihood estimator of $p_{r,l}$ is $\hat{p}_{r,l} = \frac{n_{r,A_l}}{2n_r}$ and its empirical variance is $\widehat{Var}(\hat{p}_{r,l}) = \frac{\hat{p}_{r,l}(1-\hat{p}_{r,l})}{2n_r}$. The maximum likelihood estimator is unbiased but unstable, and may take values at the edges of the parameter space in small samples.

Beta prior with parameters

$$a_l = \frac{1 - \theta}{\theta} x_l,$$

$$b_l = \frac{1 - \theta}{\theta} (1 - x_l).$$

Allele frequencies
In original pop.




$$\frac{\text{Var}(p_l)}{E(p_l)[1 - E(p_l)]} = \frac{\frac{a_l b_l}{(a_l + b_l)^2 (a_l + b_l + 1)}}{\frac{a_l}{a_l + b_l} \cdot \frac{b_l}{a_l + b_l}} = \theta.$$

Joint posterior

$$g(\mathbf{p}, \theta, \mathbf{x} | DATA) \propto \prod_{r=1}^R \prod_{l=1}^L p_{r,l}^{n_{r,A_l} + \frac{1-\theta}{\theta} x_l - 1} (1 - p_{r,l})^{n_{r,a_l} + \frac{1-\theta}{\theta} (1-x_l) - 1} g(\theta) g(\mathbf{x}).$$

Beta(2,1)



Uniform



FIND MARGINAL POSTERIOR OF θ TO ESTABLISH NULL PROCESS
(MARKOV CHAIN MONTE CARLO SAMPLING NEEDED)

A TWO-STEP PROCEDURE

- **First:** infer θ locus by locus. Bayesian model with minimally informative prior assigned to allelic frequencies
 - **Second:** feed posterior means or transformations thereof (or entire collection of samples) to mixture model
- ➔ Use mixture model to construct clusters of θ -values
- ➔ Interpret clusters in the light of available biological knowledge

Step 1

a) Prior for allelic frequency of A at each locus
(Jeffreys , maximum entropy, reference prior)

$$Beta\left(\frac{1}{2}, \frac{1}{2}\right)$$

b) Likelihood function of all allelic frequencies (assuming linkage equilibrium)

$$l(\mathbf{p}|DATA) = \prod_{r=1}^R \prod_{l=1}^L p_{r,l}^{n_{r,A_l}} (1 - p_{r,l})^{n_{r,a_l}} .$$

c) Joint posterior distribution of allelic frequencies

$$\begin{aligned} g(\mathbf{p}|DATA) &\propto \prod_{r=1}^R \prod_{l=1}^L p_{r,l}^{n_{r,A_l} + \frac{1}{2} - 1} (1 - p_{r,l})^{n_{r,a_l} + \frac{1}{2} - 1} \\ &= \prod_{r=1}^R \prod_{l=1}^L Beta\left(n_{r,A_l} + \frac{1}{2}, n_{r,a_l} + \frac{1}{2}\right) . \end{aligned}$$

- d) Draw S samples from posterior distribution of θ_l by evaluation of samples from posterior distributions of allelic frequencies

Sample $s \rightarrow$

$$\theta_l^{(s)} = \frac{\sum_{r=1}^R \left(p_{r,l}^{(s)} \right)^2 - \frac{\left(\sum_{r=1}^R p_{r,l}^{(s)} \right)^2}{R}}{\left(\frac{R \sum_{r=1}^R p_{r,l}^{(s)} - \left(\sum_{r=1}^R p_{r,l}^{(s)} \right)^2}{R} \right)},$$

(

- e) From samples, estimate posterior mean, SD, density, distribution function for each locus. Vector of posterior means is of order $L \times 1$
- f) Check whether or not the θ 's observed depart from what would be expected by chance. If not, sample lacks power to address the question of whether or not the locus has been affected by selection

EXAMPLE

- Hypothetical population **M**: 100 individuals sampled. $\#(A_i)=199$ $\#(a_i)=1$
- Hypothetical population **N**: 30 individuals sampled. $\#(A_i)=10$ $\#(a_i)=50$
- If the draws had been from a **single** population: 130 individuals.
 $\#(A_i)=209$ $\#(a_i)=51$
- Draw $S=1000$ samples from posterior distribution of allelic frequencies and θ

1

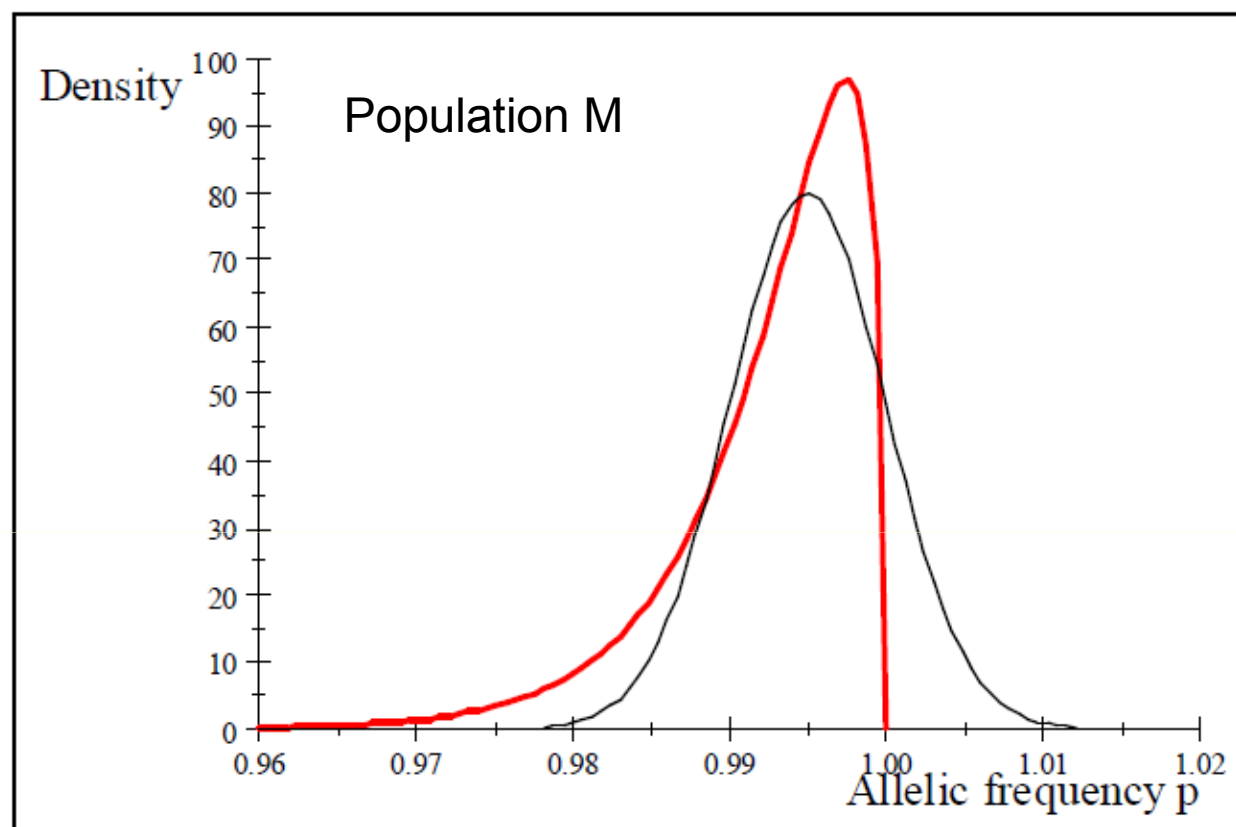


Figure 1. Posterior density (thick line) of the allelic frequency p at a locus for which 199 copies have been observed out of 200 alleles counted in hypothetical population M ; the posterior distribution is $Beta\left(199 + \frac{1}{2}, 1 + \frac{1}{2}\right)$. The thin line is the density of a normal approximation to the sampling distribution of the maximum likelihood estimator.

2

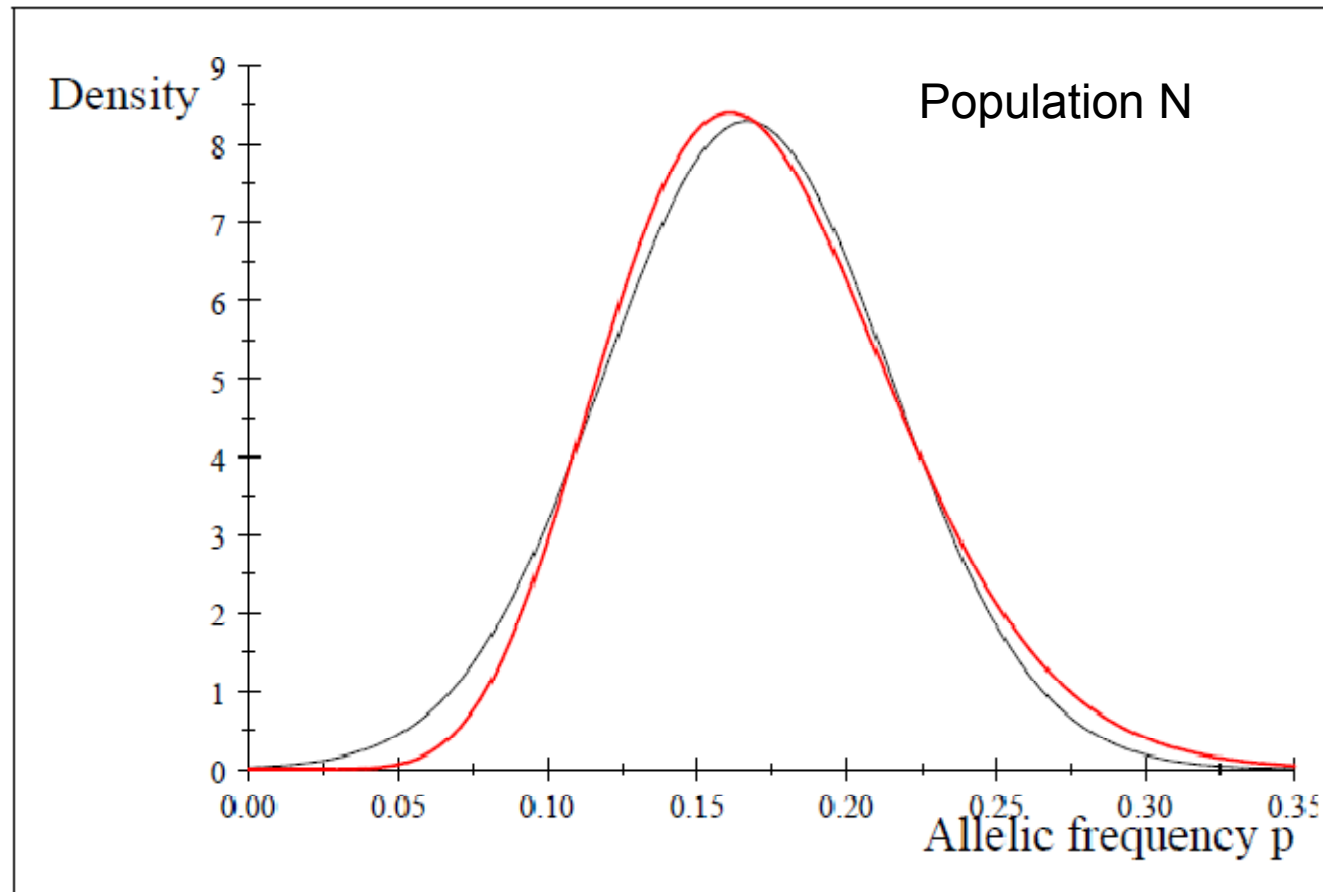


Figure 2. Posterior density (thick line) of the allelic frequency p at a locus for which 10 copies have been observed out of 60 alleles counted in hypothetical population N ; the posterior distribution is $Beta(10 + \frac{1}{2}, 50 + \frac{1}{2})$. The thin line is the density of a normal approximation to the sampling distribution of the maximum likelihood estimator.

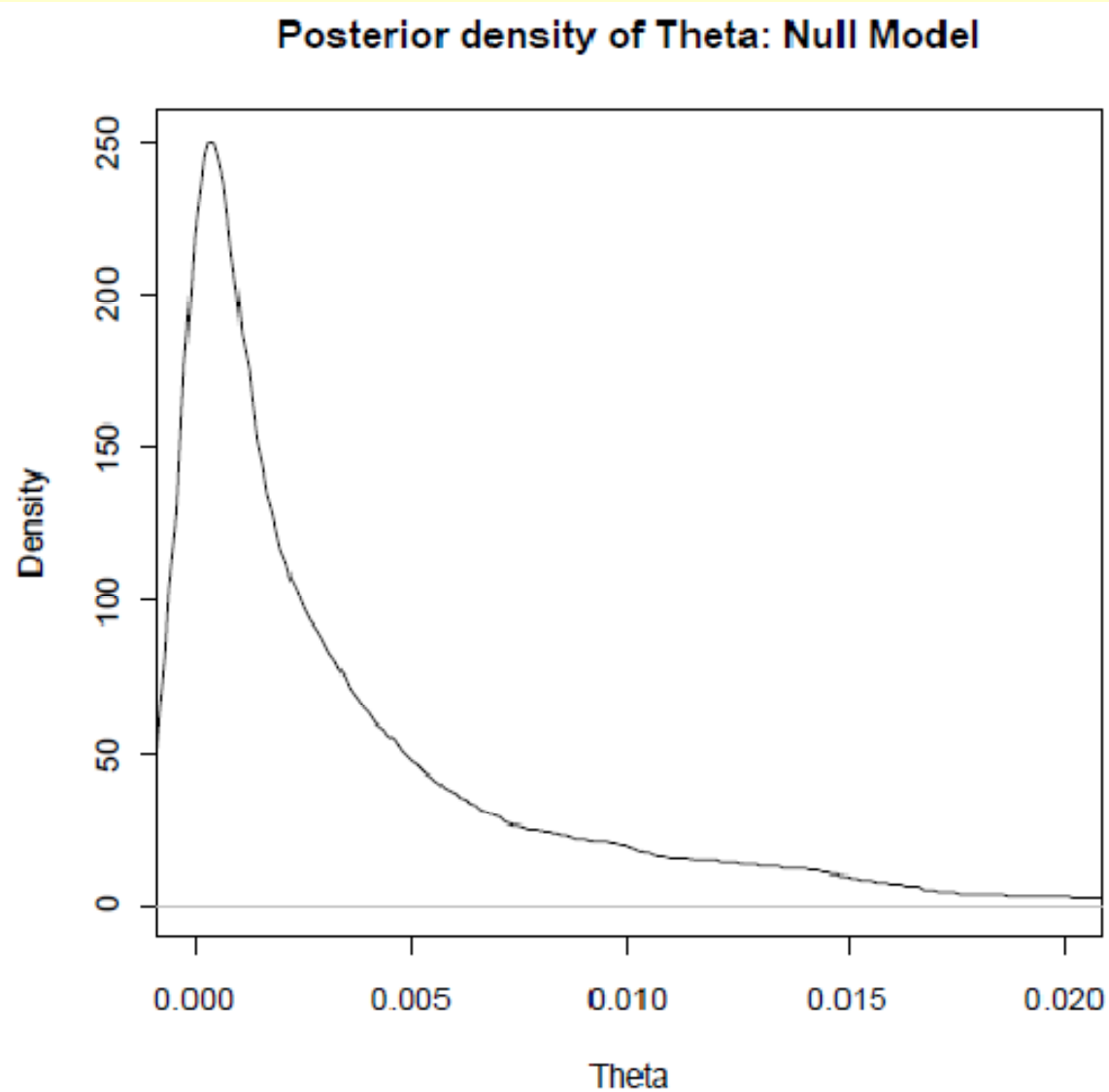


Figure 5. Posterior density of θ_i under the null model for the hypothetical example of populations M and N .

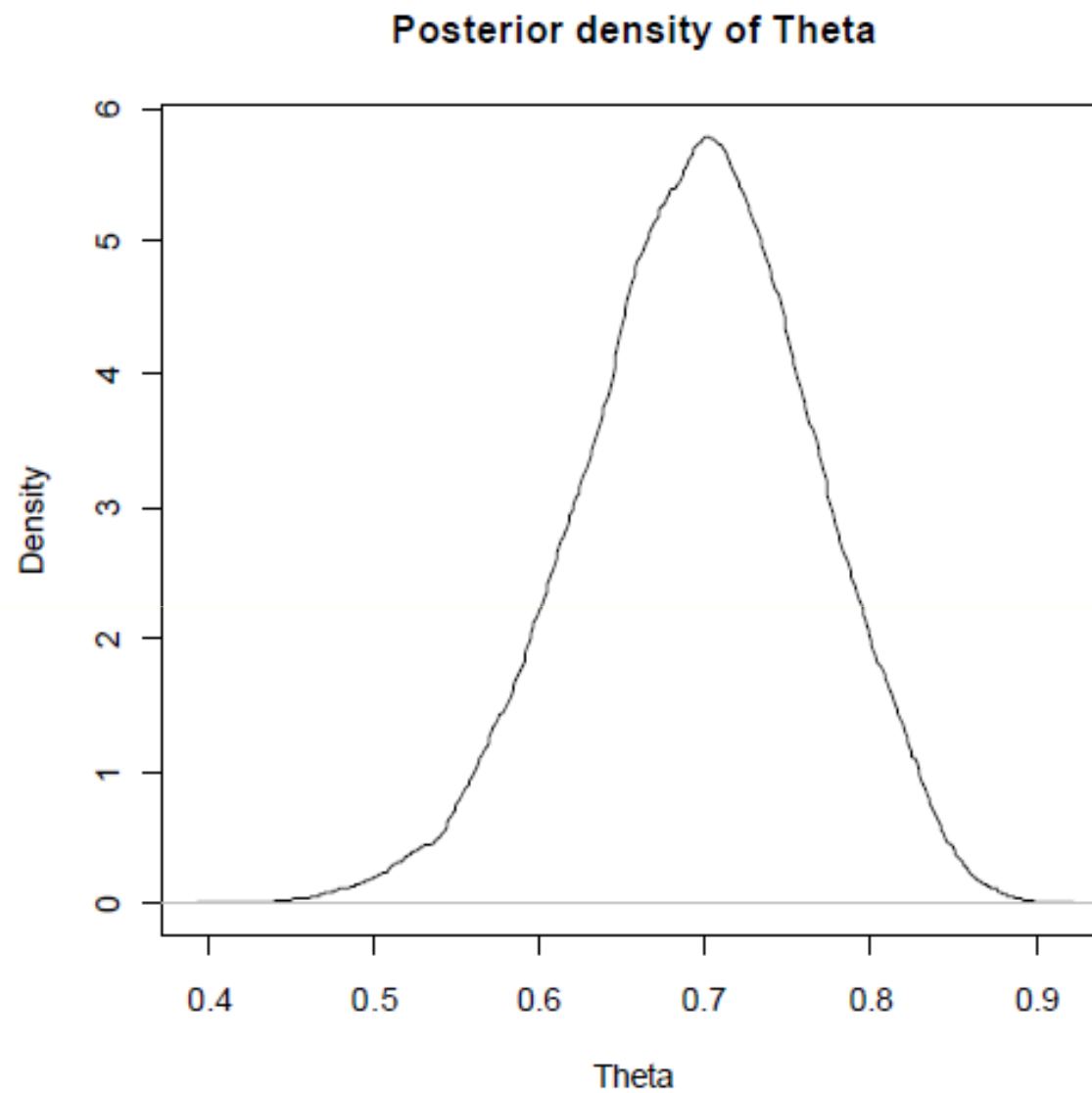


Figure 2. Posterior density of θ_t for the hypothetical example of populations M and N .

Step 2

- TREAT POSTERIOR MEANS AS RESPONSE VARIABLES
- FIT NORMAL MIXTURE MODELS TO POSTERIOR MEANS OR TRANSFORMS THEREOF (e.g., MAXIMUM LIKELIHOOD, EM algorithm, *FlexMix* in R)
- FIND BEST FITTING MODEL (AIC, BIC)
- CLUSTER LOCI ACCORDING TO θ VALUES
- INTERPRET CLUSTERS ACCORDING TO AVAILABLE BIOLOGICAL KNOWLEDGE

MIXTURE MODEL

components (clusters)

$$\bar{\theta}_l \text{ or } \log \left(\frac{\bar{\theta}_l}{1 - \bar{\theta}_l} \right) \text{ or } -\log(-\log(\bar{\theta}_l)) \sim \sum_{k=1}^K \pi_k N(\bar{\theta}_l | \mu_k, \sigma_k^2),$$

(Logit)

(Gompit)

Probability of membership

POSTERIOR (given parameter estimates) PROBABILITIES OF MEMBERSHIP

$$\Pr(\text{locus } l \in \text{cluster } k | \text{parameter estimates}) = \frac{\hat{\pi}_k N(\bar{\theta}_l | \hat{\mu}_k, \hat{\sigma}_k^2)}{\sum_{k=1}^K \hat{\pi}_k N(\bar{\theta}_l | \hat{\mu}_k, \hat{\sigma}_k^2)}.$$

CHOOSE K YIELDING SMALLEST AIC

$$AIC(k) = 2 \left[p_k - \sum_{l=1}^{\#loci} \log \left(\sum_{k=1}^K \hat{\pi}_k N(\bar{\theta}_l | \hat{\mu}_k, \hat{\sigma}_k^2) \right) \right].$$

THE CONCEPT OF PENALTY FOR NUMBER OF PARAMETERS: AKAIKE'S INFORMATION CRITERION (choose models with smallest value)

$$AIC(\text{Model } k) = 2 \sum_{i=1}^N \log f_k(y_i | \theta_k) + 2p$$

Deviance (decreases with # parameters)

#parameters in model k

Example: regression model

$$y_i \sim N(\mu_i, \sigma^2) \quad [\text{independence assumed}]$$

$$\sigma^2 \sim \text{Inv-}\chi^2_p$$

ML estimates



$$\hat{\theta}^2 = \frac{1}{n} \sum_{i=1}^N \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right)^2$$

Maximized log-likelihood

$$\begin{aligned}\log L(\hat{\beta}, \hat{\sigma}^2) &= -\frac{N}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^N (y_i - \mathbf{x}_i' \hat{\beta})^2 \\ &= -\frac{N}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} N\hat{\sigma}^2 \\ &= -\frac{1}{2} \left[N \log(2\pi\hat{\sigma}^2) + N \right]\end{aligned}$$

$$AIC(p \text{ regressions}) = -2 \log L(\hat{\beta}, \hat{\sigma}^2) = N \log(2\pi\hat{\sigma}^2) + N$$

$$= \text{constant} + N \log(\hat{\sigma}^2) + 2p$$

Model with more predictors decrease deviance but have more complexity (p)

TREE DATA FROM PETIT et al. (1998)

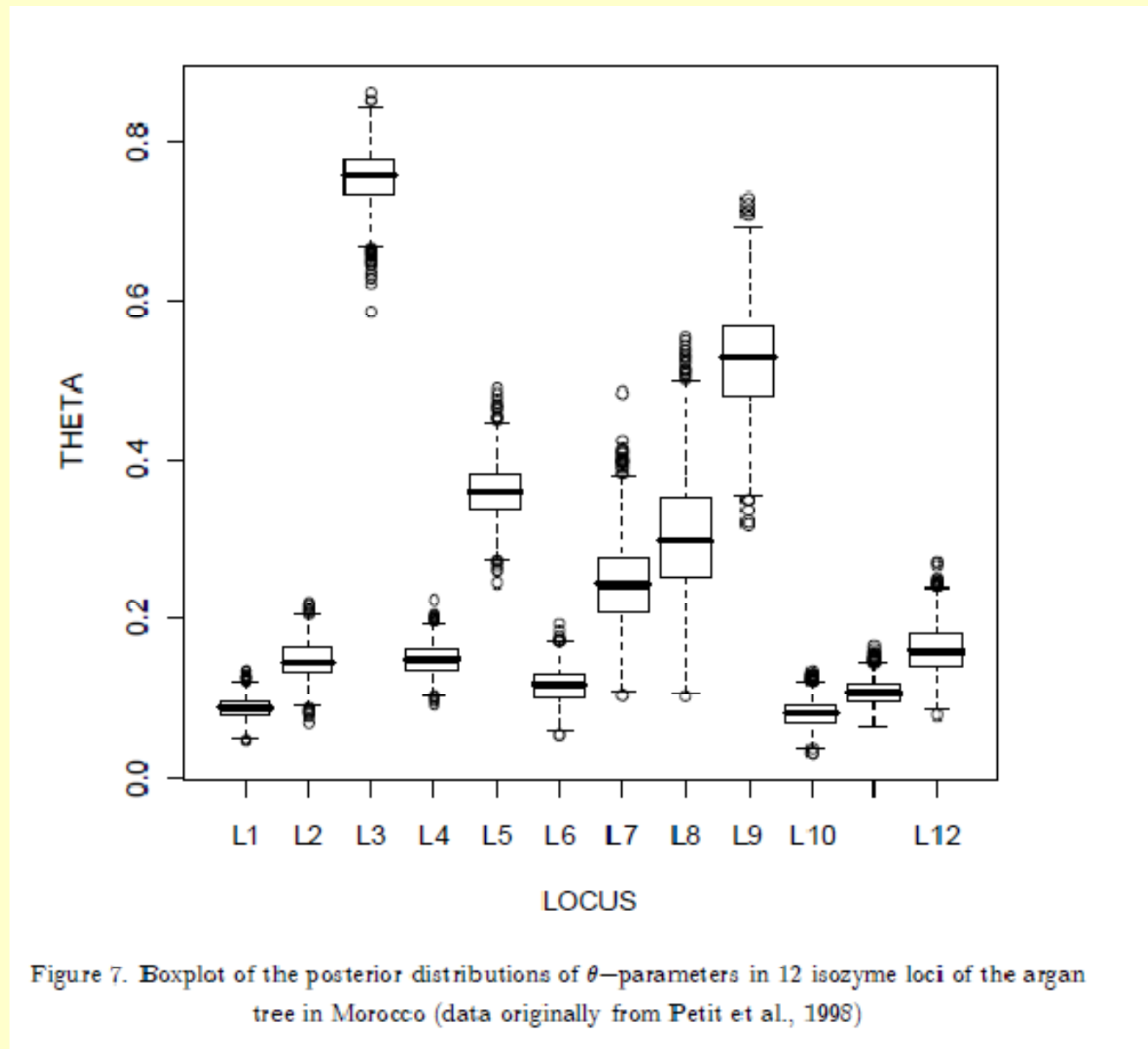
Table 1. Allelic frequencies at 12 isozyme loci in each of 12 Argan tree populations, adapted from Petit et al. (1998) by making all loci bi-allelic. A1-A12 represent frequencies of the "A" allele at loci 1-12; No. A1-No. A12 are the observed number of copies of the alleles. The number of "a" alleles can be calculated from the number of individuals samples and the number of "A" alleles observed.

12 populations

Population	AB	AD	AR	BS	GO	MI	OG	SI	TA	TE	TM	TT
No. Individuals	20	40	20	30	32	20	30	20	30	20	20	50
A1	0.525	0.512	0.475	0.467	0.047	0.475	0.517	0.575	0.517	0.425	0.55	0.52
No. A1	21	41	19	28	3	19	31	23	31	17	22	52
A2	0.4	0.438	0.55	0.917	0.688	0.525	0.467	0.825	0.483	0.925	0.475	0.51
No. A2	16	35	22	55	44	21	28	33	29	37	19	51
A3	1	1	1	0	1	1	1	1	0.75	1	1	1
No. A3	40	80	40	0	64	40	60	40	45	40	40	100
A4	0.525	0.375	0.45	0.517	0.922	0.525	1	0.7	0.467	0.575	0.5	0.52
No. A4	21	30	18	31	59	21	60	28	28	23	20	52
A5	0.475	0.463	0.475	1	1	1	1	1	0.817	1	1	0.51
No. A5	19	37	19	60	64	40	60	40	49	40	40	51
A6	0.85	0.538	0.9	0.533	0.922	0.575	0.55	0.75	0.517	0.525	0.55	0.53
No. A6	34	43	36	32	59	23	33	30	31	21	22	53
A7	1	1	1	0.567	0.922	0.9	1	1	0.967	1	1	1
No. A7	40	80	40	34	59	36	60	40	58	40	40	100
A8	1	1	1	1	1	1	1	1	1	1	0.575	0.97
No. A8	40	80	40	60	64	40	60	40	60	40	23	97
A9	1	0.937	1	1	0.312	1	1	1	1	1	1	1
No. A9	40	75	40	60	20	40	60	40	60	40	40	100
A10	0.925	0.5	0.525	0.625	0.475	0.5	0.55	0.4	0.575	0.5	0.475	0.5
No. A10	37	40	21	38	30	20	33	16	35	20	19	50
A11	0.6	0.7	0.575	0.5	0.6	0.525	1	0.375	0.625	0.475	0.55	0.47
No. A11	24	56	23	30	38	21	60	15	38	19	22	47
A12	1	1	0.85	0.6	0.875	0.775	1	0.875	1	1	1	0.87
No. A12	40	80	34	36	56	31	60	35	60	40	40	87

12 loci

Box plots of posterior distributions of θ for each of the 12 loci (2000 samples per locus)



Put the 24000 samples in bag and estimate the density of the resulting distribution

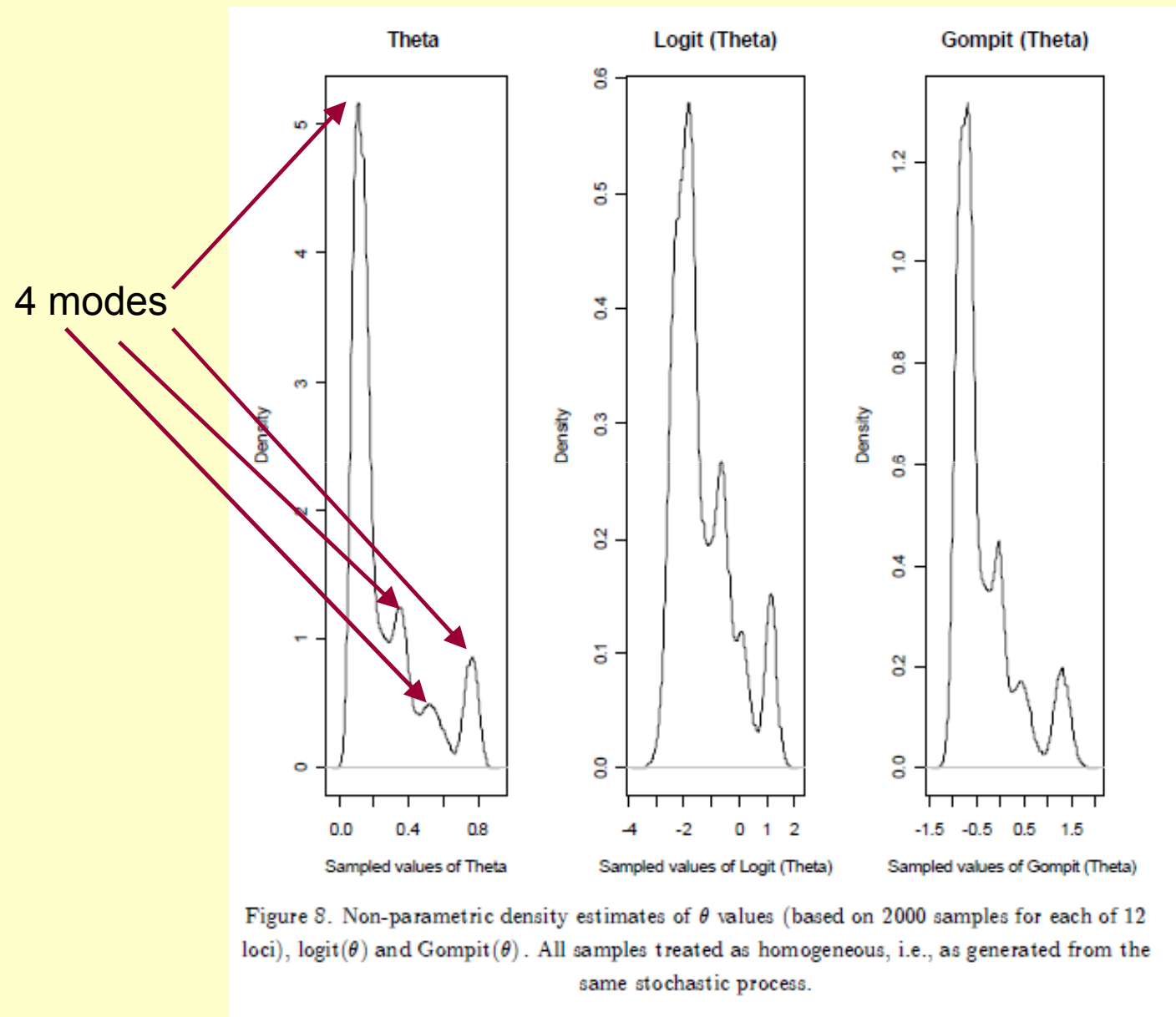


Table 2. Comparison of mixture models with 2, 3 or 4 components fitted to the 12 posterior means of θ -parameters and their logit or Gompit transforms in the argan tree data of Petit et al. (1998). AIC: Akaike's information criterion (models with smallest values are favored and indicated in boldface)

Variable	No. components (k)	Iterations to convergence	AIC
θ	k=1	2	-0.651
	k=2	16	-6.299
	k=3	36	-2.921
	k=4	39	3.079
$\log \frac{\theta}{1-\theta}$	k=1	2	39.100
	k=2	28	40.102
	k=3	77	44.392
	k=4	94	50.392
$-\log[-\log(-\theta)]$	k=1	2	26.909
	k=2	36	24.328
	k=3	41	27.742
	k=4	48	33.742

Analysis supports no more than 2 clusters of Θ values

Table 3. Conditional probabilities of membership to one of two clusters for mixture models fitted to the posterior means of θ for the 12 loci in the argan tree, and their logit, $\log\left(\frac{\theta}{1-\theta}\right)$, and Gompit, $-\log(-\log(\theta))$, transformations (boldfaced probability indicates the cluster with largest probability of membership).

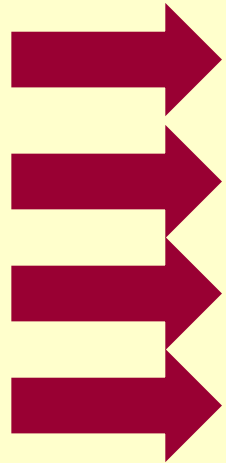
	θ means	logit(θ)		Gompit(θ)		
Locus	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	0.93	0.07	0.91	0.09	0.91	0.09
2	0.92	0.08	0.83	0.17	0.89	0.11
3	0.00	1.00	0.00	1.00	0.00	1.00
4	0.92	0.08	0.82	0.18	0.88	0.12
5	0.00	1.00	0.00	1.00	0.00	1.00
6	0.95	0.05	0.91	0.09	0.93	0.07
7	0.00	1.00	0.08	0.92	0.04	0.96
8	0.00	1.00	0.00	1.00	0.00	1.00
9	0.00	1.00	0.00	1.00	0.00	1.00
10	0.92	0.08	0.89	0.11	0.89	0.11
11	0.95	0.05	0.92	0.08	0.93	0.07
12	0.87	0.13	0.76	0.24	0.83	0.17
Cluster Mean	0.12	0.41	-2.03	-0.92	-0.11	0.76
Cluster standard deviation	0.03	0.21	0.32	1.02	0.67	0.13

SAME CLUSTERS ARRIVED AT IRRESPECTIVE OF TRANSFORMATION

Detecting selection signatures in cattle

Qanbari, Gianola, Hayes, Schenkel,
Miller, Moore, Thaller, Simianer

Description of samples



Breed	Code	Data set		Sample size (n)	Country	Purpose
Holstein	HS	I	II	2091	Germany	Dairy
Brown Swiss	BS	I	II	277	Germany	Dairy
Simmental	SI	I	II	462	Germany	Dual-purpose
Canadian Angus	CA	-	II	103	Canada	Beef
Piedemontese	PI	-	II	43	Canada	Beef
Australian Angus	AA	I	-	232	Australia	Beef
Brahman	BR	I	-	80	Australia	Beef
Belmond Red	BE	I	-	166	Australia	Beef
Hereford	HR	I	-	158	Australia	Beef
Murray Gray	MG	I	-	57	Australia	Beef
Santa Gertrudis	SG	I	-	126	Australia	Beef
Shorthorns	SH	I	-	81	Australia	Beef

Genome wide summary of marker statistics for the breeds used in a LD analysis (data set I).

Breed	SNP (n)	MAF (%)	ObsHET (%)	Inter-marker distance (kb)	Max gap (kb)
Holstein	39474	28.2±13	37.2±12	64.45±62.5	2081.4
Brown Swiss	35226	27.7±13	36.6±13	72.26±72.8	2081.4
Simmental	37976	27.5±13	37.0±12	67.06±69.8	2145.7
Australian Angus	44938	24.3±15	32.3±16	56.70±52.4	2081.5
Brahman	45173	16.4±14	23.7±17	56.40±51.3	1677.8
Belmond Red	47416	24.1±15	32.3±16	53.74±47.9	1677.8
Hereford	45322	25.5±15	34.1±16	56.22±52.1	2081.5
Murray Gray	41369	24.4±15	33.3±17	61.52±59.0	2081.5
Santa Gertrudis	46809	23.6±15	31.7±17	54.44±48.9	1677.8
Shorthorns	42280	21.7±15	28.5±16	60.26±56.9	2081.5

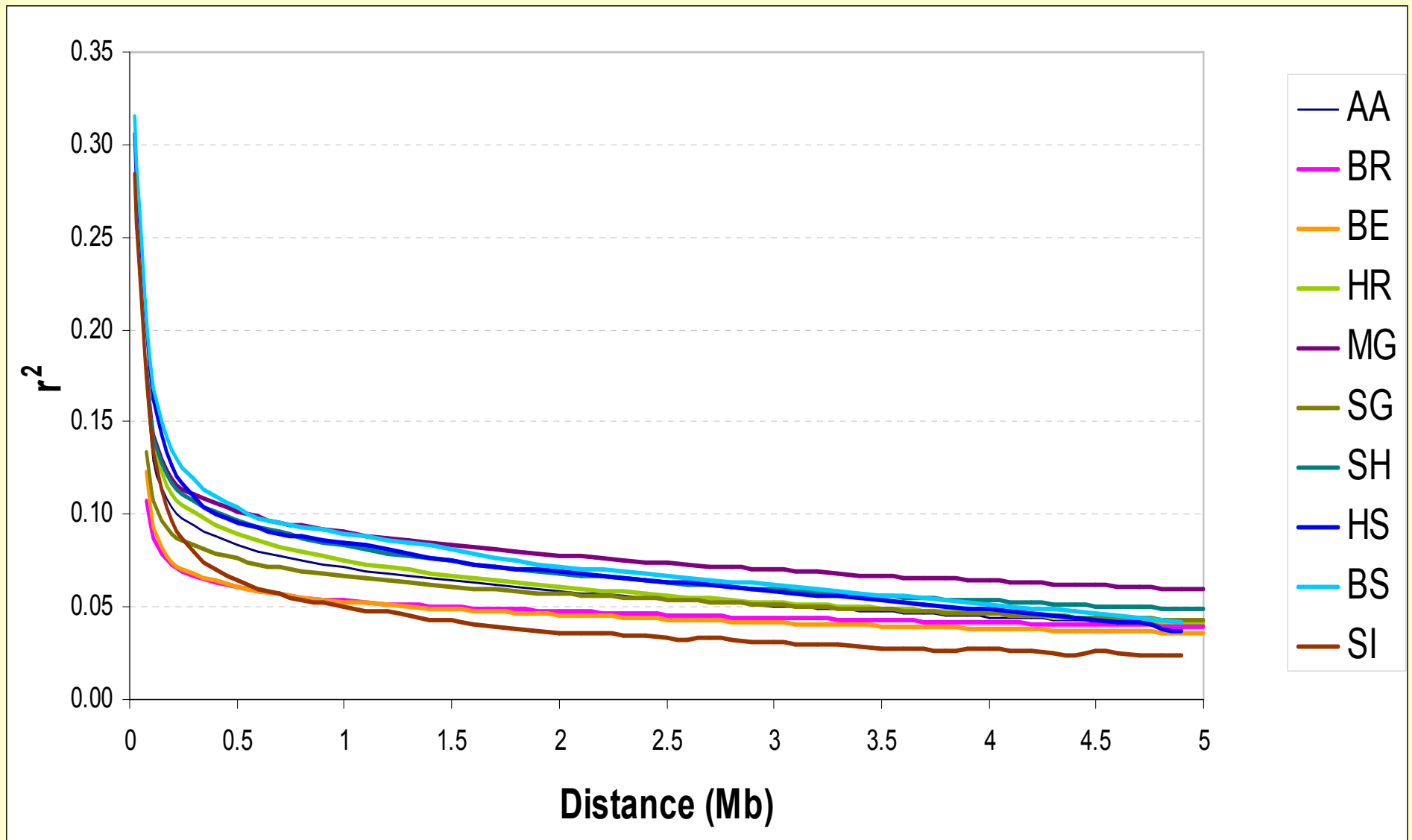
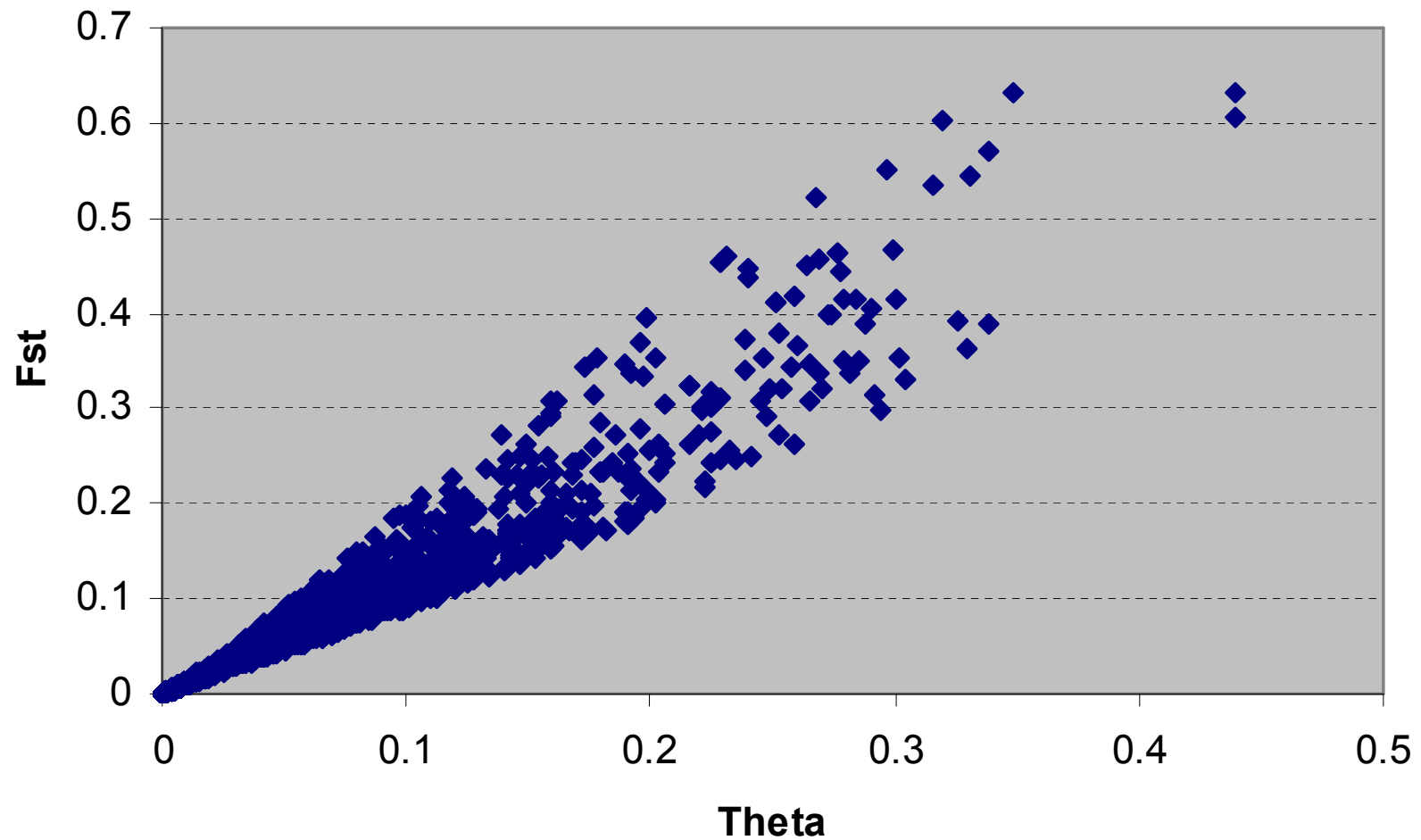


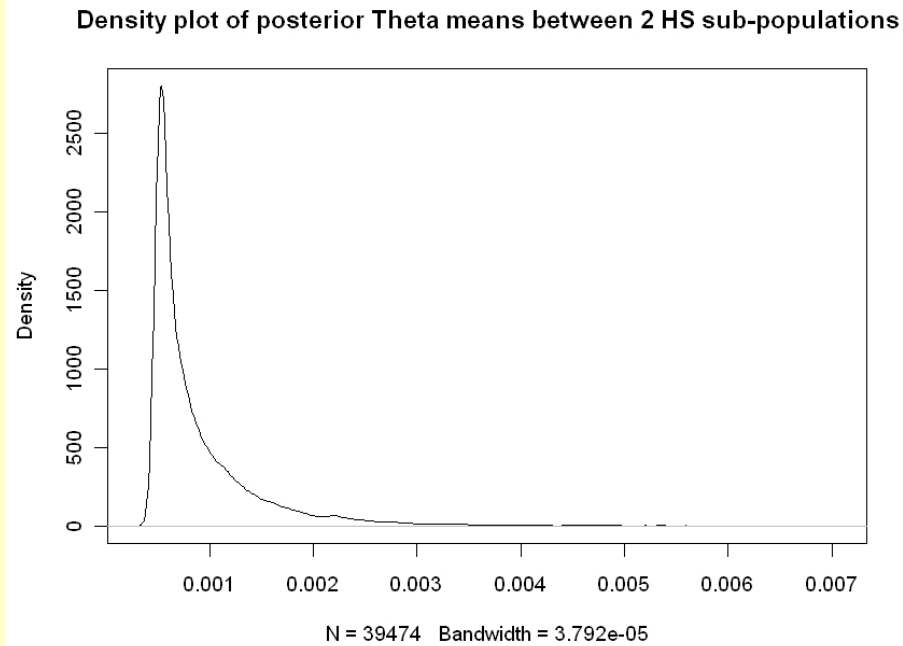
Figure 1. Decay of LD as a function of inter-marker distance in dairy and beef breeds

Weir's unbiased F_{ST} vs Bayesian posterior mean of Theta

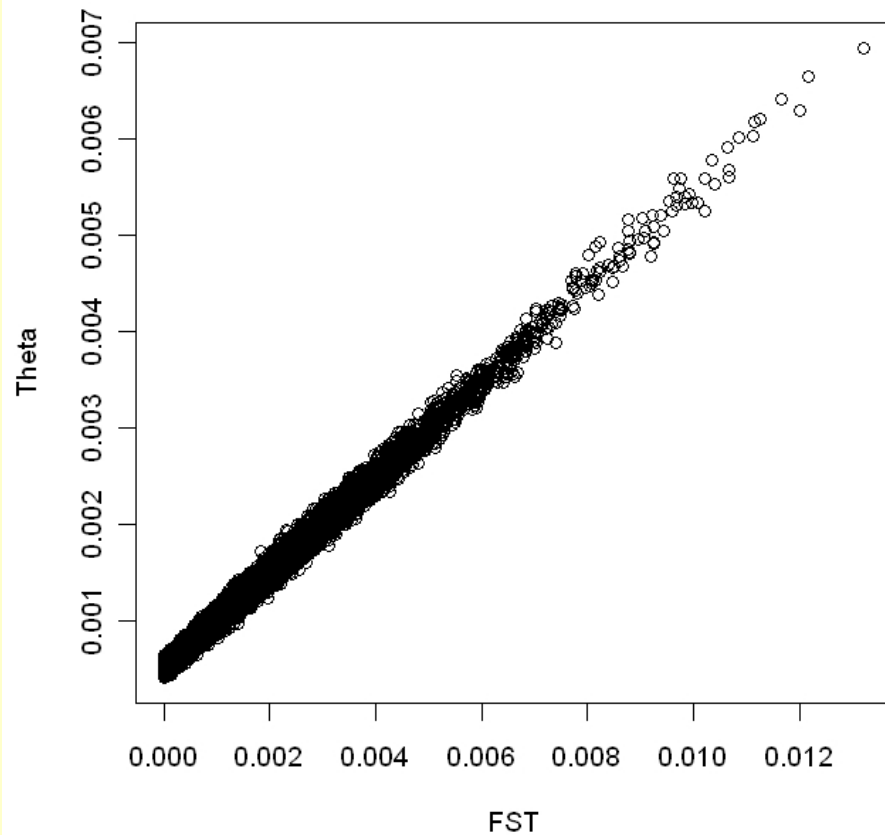


Disagreement at “high” differentiation: 1) shrinkage; 2) division by $R-1$

HOLSTEIN VS HOLSTEIN: 2 RANDOM PARTITIONS



No differentiation



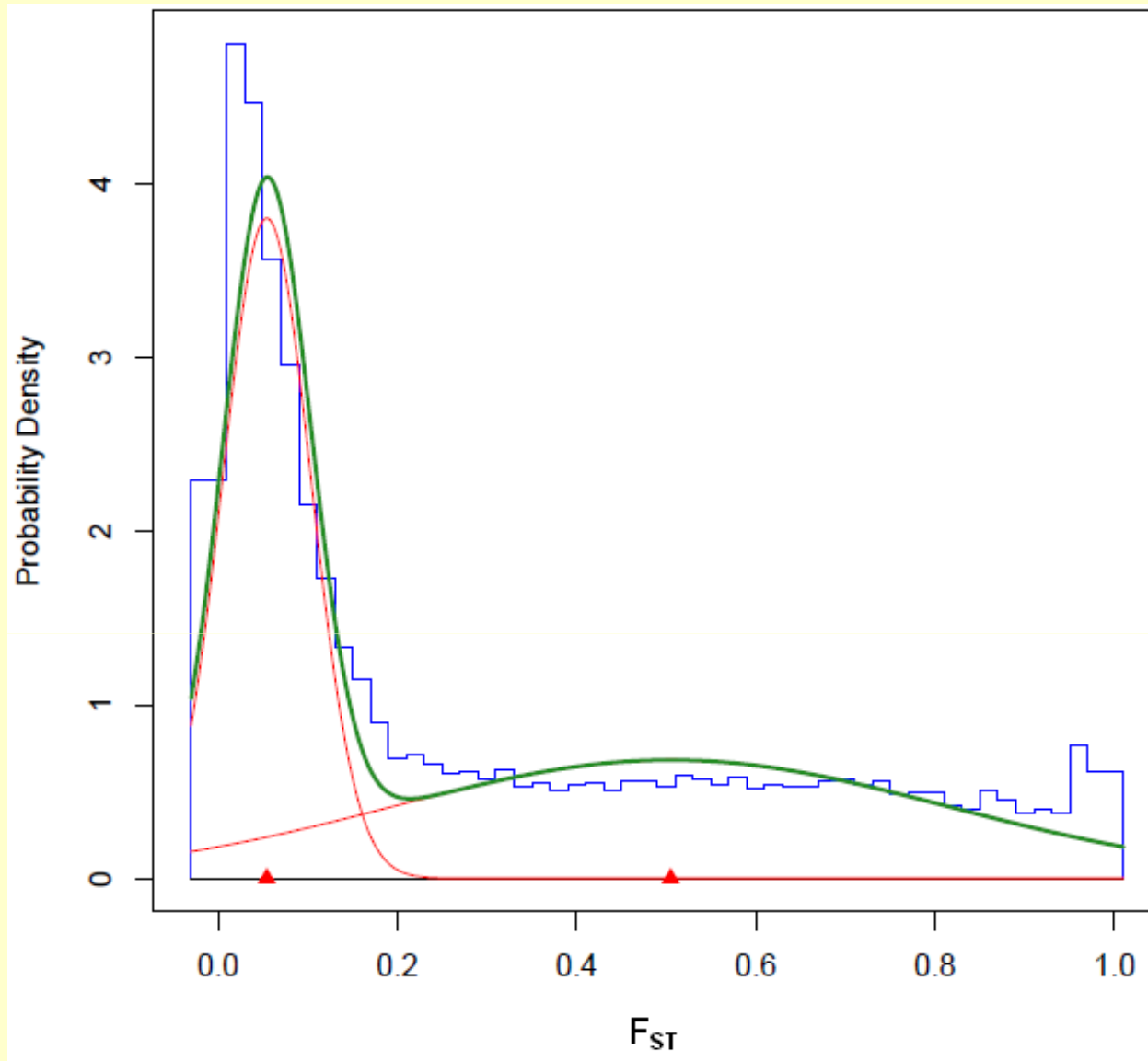
Good agreement with unbiased estimator
in the absence of differentiation₄₇

Summary statistics of pair-wise estimates of F_{ST} and clustering information

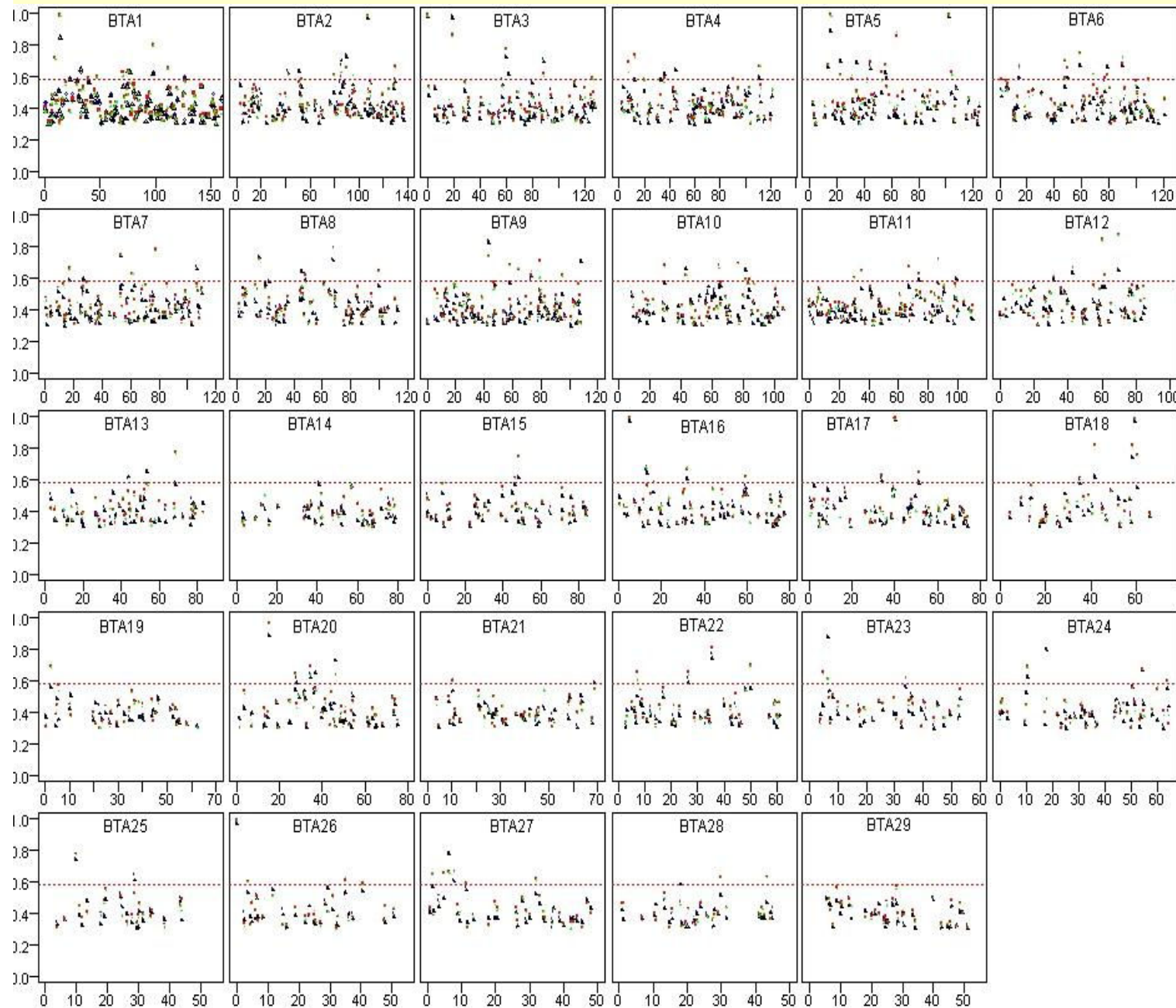
	HS			BS			SI			AN		
	θ	K ¹	L ²	θ	K	L	θ	K	L	θ	K	L
BS	0.05	5	4878									
SI	0.04	4	7796	0.04	5	7691						
AN	0.27	3	12106	0.29	4	5571	0.28	3	10882			
PI	0.27	3	19442	0.28	3	18637	0.27	3	8867	0.02	7	2247

¹ K= Number of clusters; ² L= Number of SNPs with largest θ values representing the first cluster of loci

Dairy vs Beef: HS and AN (mean of posterior means) → 0.27 ± 0.01
Beef vs Beef: AN and PI → 0.02 ± 0.01
Dairy vs Dairy: HS and BS → 0.05 ± 0.01 .



Distribution (blue) of posterior means over loci of θ values using two dairy (HS, BS) and two beef breeds (CA and PI) and densities of the underlying mixture of two normals (green) and the respective components (red). The data support a 2-component mixture.



Dots:
 HS-AN
 HS-PI
 BS-AN
 BS-PI

-29% of
 Windows overlapped
 -BTA 9 80 windows
 covering 0.35 of
 Chromosome
 -BTA 25 23 windows
 covering 0.26 of
 chromosome

Chromosomal position (Mb)

Windows with $F_{ST} > 0.3$, indicating genomic position of the most diverse regions of dairy vs. beef breeds.

Dashed lines → upper 2.5% of the distribution of posterior means.

GENOME ANOTATION WITH *iHS* (“integrated haplotype score”) or F_{ST}

Chr	Position (Mbp)	<i>iHS</i> or F_{ST} *	Breed	Gene/EST (n)	Candidate Gene	Function
18	57.25-57.75	2.2, <i>0.78</i>	HS	30	SIGLEC5,8,10	Sialic acid binding Ig-like lectin 5, 8, 10
16	19.75-20.25	2.6	HS	2	SPATA17	Spermatogenesis associated 17
6	61.75-62.75	3.41	BS	13	UGDH APBB2	UDP-glucose dehydrogenase Amyloid beta (A4) precursor protein-binding, family B, member 2 (Fe65-like)
13	30.5-31.5	2.68	BS	8	TRDMT1	Cysteine and methionine metabolism
1	79-81.5	2.10	HR	6	SST	Somatostatin
2	34.5-36	2.26	HR	6	GCG FAP	Glucagon Fibroblast activation protein, alpha
6	80-83		HR	9	SRD5A2L2	Lipid metabolism
7	39-41	1.9	AN	15	COL23A1 MGAT1	Collagen, type XXIII, alpha 1 Fertilization and early development of the embryos
12	36-38	2.03	AN	19	ATP12A	ATPase activity
14	64-65	2.02	AN	6	MATN2	Developing cartilage rudiments
16	39-40	1.98	AN	14	NMNAT1	Methylenetetrahydrofolate reductase (NADPH) activity
17	31-32.5	2.05	AN/HR	15	PGRMC2	Progesterone receptor membrane component 2
2	70-73	2.06	MG/BE/ SI/BR	5	-	-
10	29-31	2.24	BE/SII	8	ACTC1	Actinin, Involved in the formation of filaments
1	12-13	0.92	-	0	-	-
2	111.5-112	0.98	-	11	ABCB6 GLB1L	ATP-binding cassette, sub-family B (MDR/TAP), member 6 Galactosidase, beta 1-like
3	119.2-119.7	0.92	-	11	SMCP	Sperm mitochondria-associated cysteine-rich protein
7	53.25-53.75	0.74	-	4	FGF1	A growth factor which stimulates growth or differentiation, key role in embryonic development
9	42-43	0.78	-	12	LACE1 PPII.6	Lactation elevated 1 Peptidylprolyl isomerase (cyclophilin)-like 6
13	53.5-54	0.98	-	7	SIRPA	Signal-regulatory protein
16	4.75-5.25	0.98	-	5	-	-
17	39.5-40.5	0.98	-	4	-	-
18	58.25-58.75	0.98	-	15	-	-
20	15.25-15.75	0.92	-	8	ADAMTS6	-
22	35.25-35.75	0.77	-	3	-	-

“Gene desert”
-regulatory regions?
-non-coding DNA fixed by drift?

* F_{ST} values are in italic

CONCLUSIONS

- F-statistics used for detecting signatures of selection
- Several Bayesian methods available (with and without MCMC)
- Simple 2-step procedure proposed
- Mixture model can be enriched by placing more structure on means (e.g., chromosome, coding vs. non-coding)
- Generalization to multiple alleles
- EM algorithm breaks down if entire set of posterior samples is fed (can use, e.g., medians and upper and lower percentiles)
- Main challenge: accommodate linked and LD loci, e.g. introduce kernel structure in mixture model