



Non-parametric whole genome linkage/linkage disequilibrium mapping - simulation results and an application to Canadian Holstein data

Henner SIMIANER¹, Eduardo da Cruz Gouveia PIMENTEL¹,
Mehdi SARGOLZAEI² and Flavio SCHENKEL²

¹ Animal Breeding and Genetics Group
Georg-August-University, Goettingen, Germany

² Department of Animal and Poultry Science
University of Guelph, Canada





New situation in farm animal genetics

- lots (10'000s) of SNP genotypes
- for many (1'000s) individuals
- to combine with performance information for a number (10+) of yield and functional traits

New challenges:

- estimation of genomic breeding values (see e.g. Meuwissen, Hayes and Goddard, 2001 – not my topic today)
- fine mapping of quantitative trait loci (QTL)



QTL fine mapping – two basic principles:

1. Linkage mapping

- **Positional resolution** and **statistical power** in a given design is limited (> 10 cM)
- statistically significant (non-false positive) linkage is **‘true’ linkage**

2. Association mapping

- **High power** in the immediate vicinity (< 1 cM) of a QTL
- Highly susceptible to **false positives** (c.f. Simpson’s paradox)

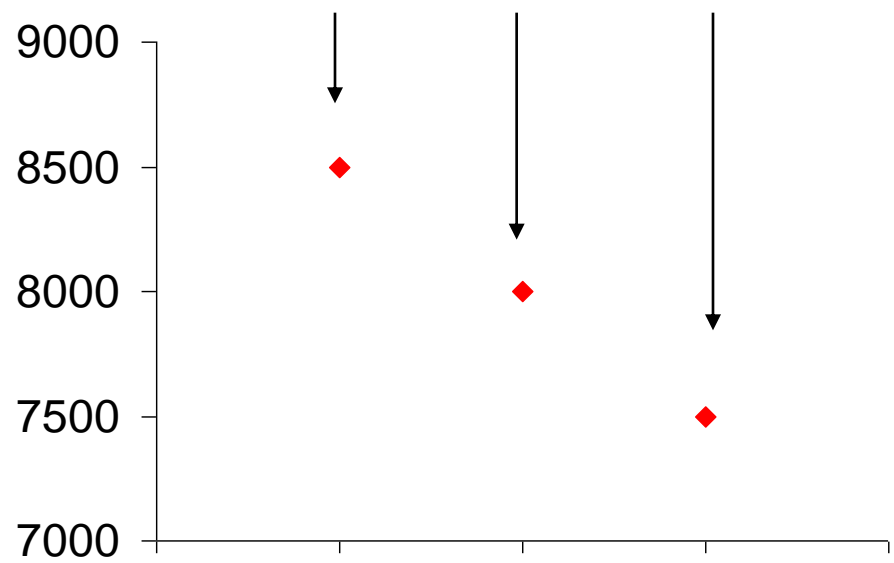


	SNP-genotype	Breeding value
Sire 1:	AA	+1000kg
Sire 2:	aa	-1000kg
Cow population:	$P(A) = 0.5$	\emptyset breeding value = 0 \emptyset performance = 8000



	SNP-genotype	Breeding value
Sire 1:	AA	+1000kg
Sire 2:	aa	-1000kg
Cow population:	$P(A) = 0.5$	\emptyset breeding value = 0 \emptyset performance = 8000

	AA	Aa	aa	\emptyset performance
100 offspring sire 1:	50	50	0	$\mu + 500$
100 offspring sire 2:	0	50	50	$\mu - 500$





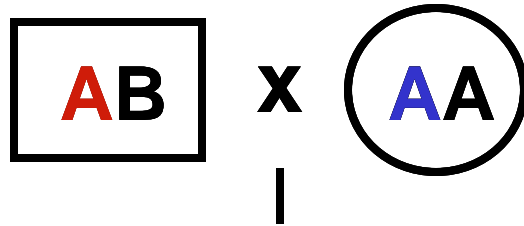
A good idea: get the best of both worlds

Combine linkage and linkage disequilibrium mapping

- Have the high power and the high positional resolution of LD-mapping, and
- be sure to have true linkage due to linkage mapping



The transmission disequilibrium test (TDT, Spielman et al. 1993)



- Trios with one affected offspring
- for one allele at a time



		Non-transmitted allele	
		A	not A
transmitted allele	A	+1	+1
	not A		



The transmission disequilibrium test (TDT, Spielman et al. 1993)

- Trios with one affected offspring
- for one allele at a time

McNemar-Test

$$N_{A-} \neq N_{-A}$$

$$H_0: \delta(1 - 2\theta) = 0$$

Positive result only if
 $\delta > 0$ and $\theta < 0.5$

		Non-transmitted allele	
		A	not A
transmitted allele	A		N_{A-}
	not A	N_{-A}	



The transmission disequilibrium test (TDT, Spielman et al. 1993)

- combines linkage and linkage disequilibrium mapping
- the ‚non-transmitted‘ allele is used as a control
- disequilibrium which is not due to close linkage will affect both transmitted and non-transmitted alleles in the same fashion, therefore the contrast is corrected for this effect
- originally designed for qualitative traits, extensions to
 - use of multiple alleles and haplotypes (Clayton, 1999)
 - general pedigrees (Rabinowitz and Laird, 2000)
 - analysis of quantitative traits (Szyda et al., 1998)
- see Laird and Lange (2006) for a comprehensive review



Which phenotype information to use

- raw or corrected individual phenotypes
- family (e.g. offspring group) means
- estimated breeding values
- de-regressed proofs, DYDs
- Mendelian sampling terms

$$u_i = 0.5(u_s + u_d) + m_i$$

- estimates readily computable from EBVs
- corrected for all environmental and parental effects
- mixture of normals with inhomogeneous variances
- first suggested to be used in association studies for candidate genes for beef traits by Morsci et al. (2006)



A good idea: get the best of both worlds

- The transmission disequilibrium test (TDT, Spielman et al. 1993) is based on associations that segregate within families (parent → offspring)
- Mendelian sampling is the deviation of offspring from the parent mean

Apply a quantitative TDT to the (estimated) Mendelian sampling term





Methodology

1. **BLUP** breeding value estimation (standard in most animal breeding programs for many traits) $\rightarrow \hat{u}_i$
2. For an individual i with parents s and d estimate the **Mendelian sampling** term as:

$$\hat{m}_i = \hat{u}_i - \frac{\hat{u}_s + \hat{u}_d}{2}$$

- $\hat{m}_i = 0$ if the offspring has no own or progeny information
- $\hat{m}_i > 0$ if a positive sample of parental alleles was obtained
- $\hat{m}_i < 0$ if a negative sample of parental alleles was obtained



Methodology

3. Sum over all families the contrast of estimated Mendelian sampling terms between:
 - progeny that have obtained allele 1 but not 2, and
 - progeny that have obtained allele 2 but not 1



Case	Unordered Genotype			Paternal Allele		Maternal Allele	
	Sire	Dam	Offspring	transmitted	not transmitted	transmitted	not transmitted
1	{1,1}	{1,1}	{1,1}				
2	{1,1}	{1,2}	{1,1}				
3	{1,1}	{1,2}	{1,2}				
4	{1,1}	{2,2}	{1,2}				
→ 5	{1,2}	{1,1}	{1,1}				
6	{1,2}	{1,1}	{1,2}				
7	{1,2}	{1,2}	{1,1}				
8	{1,2}	{1,2}	{1,2}				
9	{1,2}	{1,2}	{2,2}				
10	{1,2}	{2,2}	{1,2}				
11	{1,2}	{2,2}	{2,2}				
12	{2,2}	{1,1}	{1,2}				
13	{2,2}	{1,2}	{1,2}				
14	{2,2}	{1,2}	{2,2}				
15	{2,2}	{2,2}	{2,2}				



Case	Unordered Genotype			Paternal Allele		Maternal Allele	
	Sire	Dam	Offspring	transmitted	not transmitted	transmitted	not transmitted
1	{1,1}	{1,1}	{1,1}				
2	{1,1}	{1,2}	{1,1}				
3	{1,1}	{1,2}	{1,2}				
4	{1,1}	{2,2}	{1,2}				
→ 5	{1,2}	{1,1}	{1,1}	1	2	1	1
6	{1,2}	{1,1}	{1,2}				
7	{1,2}	{1,2}	{1,1}				
8	{1,2}	{1,2}	{1,2}				
9	{1,2}	{1,2}	{2,2}				
10	{1,2}	{2,2}	{1,2}				
11	{1,2}	{2,2}	{2,2}				
12	{2,2}	{1,1}	{1,2}				
13	{2,2}	{1,2}	{1,2}				
14	{2,2}	{1,2}	{2,2}				
15	{2,2}	{2,2}	{2,2}				



Case	Unordered Genotype			Paternal Allele		Maternal Allele		
	Sire	Dam	Offspring	transmitted	not transmitted	transmitted	not transmitted	
1	{1,1}	{1,1}	{1,1}	1	1	1	1	
2	{1,1}	{1,2}	{1,1}	1	1	1	2	+1
3	{1,1}	{1,2}	{1,2}	1	1	2	1	-1
4	{1,1}	{2,2}	{1,2}	1	1	2	2	
5	{1,2}	{1,1}	{1,1}	1	2	1	1	+1
6	{1,2}	{1,1}	{1,2}	2	1	1	1	-1
7	{1,2}	{1,2}	{1,1}	1	2	1	2	+2
8	{1,2}	{1,2}	{1,2}	unknown				
9	{1,2}	{1,2}	{2,2}	2	1	2	1	-2
10	{1,2}	{2,2}	{1,2}	1	2	2	2	+1
11	{1,2}	{2,2}	{2,2}	2	1	2	2	-1
12	{2,2}	{1,1}	{1,2}	2	2	1	1	
13	{2,2}	{1,2}	{1,2}	2	2	1	2	+1
14	{2,2}	{1,2}	{2,2}	2	2	2	1	-1
15	{2,2}	{2,2}	{2,2}	2	2	2	2	



Case	Unordered Genotype			Paternal Allele		Maternal Allele	
	Sire	Dam	Offspring	transmitted	not transmitted	transmitted	not transmitted
s1	{1,2}	-	{1,1}	1	2	1	unknown
s2	{1,2}	-	{2,2}	2	1	2	unknown
d1	-	{1,2}	{1,1}	1	unknown	1	2
d2	-	{1,2}	{2,2}	2	unknown	2	1



Methodology

3. Sum over all families the contrast of estimated Mendelian Sampling terms between:

→ progeny that have obtained allele 1 but not 2, and

→ progeny that have obtained allele 2 but not 1

$$\delta = \sum_{\forall [1\sqrt{2}]} \hat{m}_i - \sum_{\forall [2\sqrt{1}]} \hat{m}_i$$

4. Test if the contrast δ is different from zero (two-sided t-test)

δ is zero if:

→ there is no co-segregation within families or

→ the SNP and the trait are not in LD (or both)



Simulated data

Data set from the 12th QTL MAS workshop (May 2008, Uppsala)

- 6 chromosomes à 1 M, 1000 equidistant SNPs per chromosome
- 48 QTLs simulated in base population
- 50 generations random mating with $N_e = 100$
- in generation 50: 15♂ ♂ and 150♀ ♀ as parents → 1500 offspring
- in generation 51 to 54: dito
- parents in generation 50 and all individuals in generation 51 to 54 ($\Sigma 4665$ individuals) genotyped and phenotyped ($h^2 = 0.3$)
- 15 QTL explained $> 1\%$ (2.1% - 18.1%) of the additive genetic variance in generation 50 to 54



Analysis

1. Variance component estimation

Animal model, VCE 4.2.5 \rightarrow $\text{Var}(u) = 1.36$, $\text{Var}(e) = 3.12$, $h^2 = 0.304$

2. BLUP breeding value estimation

\rightarrow estimates of u_i , m_i , $\text{Var}(m) = 0.123$

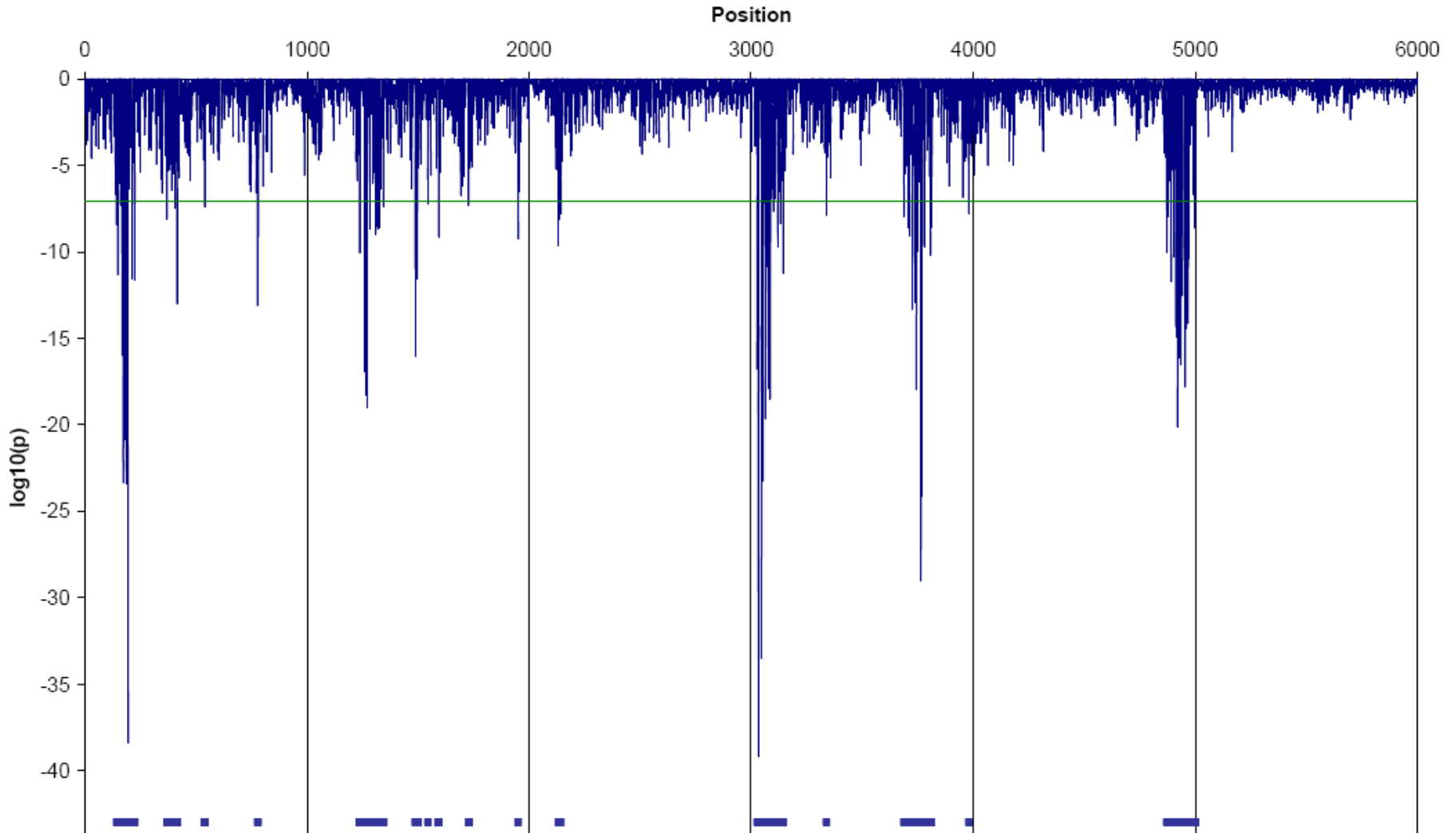
3. Analyse Mendelian sampling for 4500 complete trios with QTDT

\rightarrow δ , two-sided t-test on the genome-wide $\alpha = 0.001$ error level

$$p_c = 0.0005/6000 \quad \log_{10}(p_c) = -7.079$$

Decadic log of error probabilities

- critical value for genome-wide $p=0.001$ with Bonferroni correction
- █ genome segments with significant signals

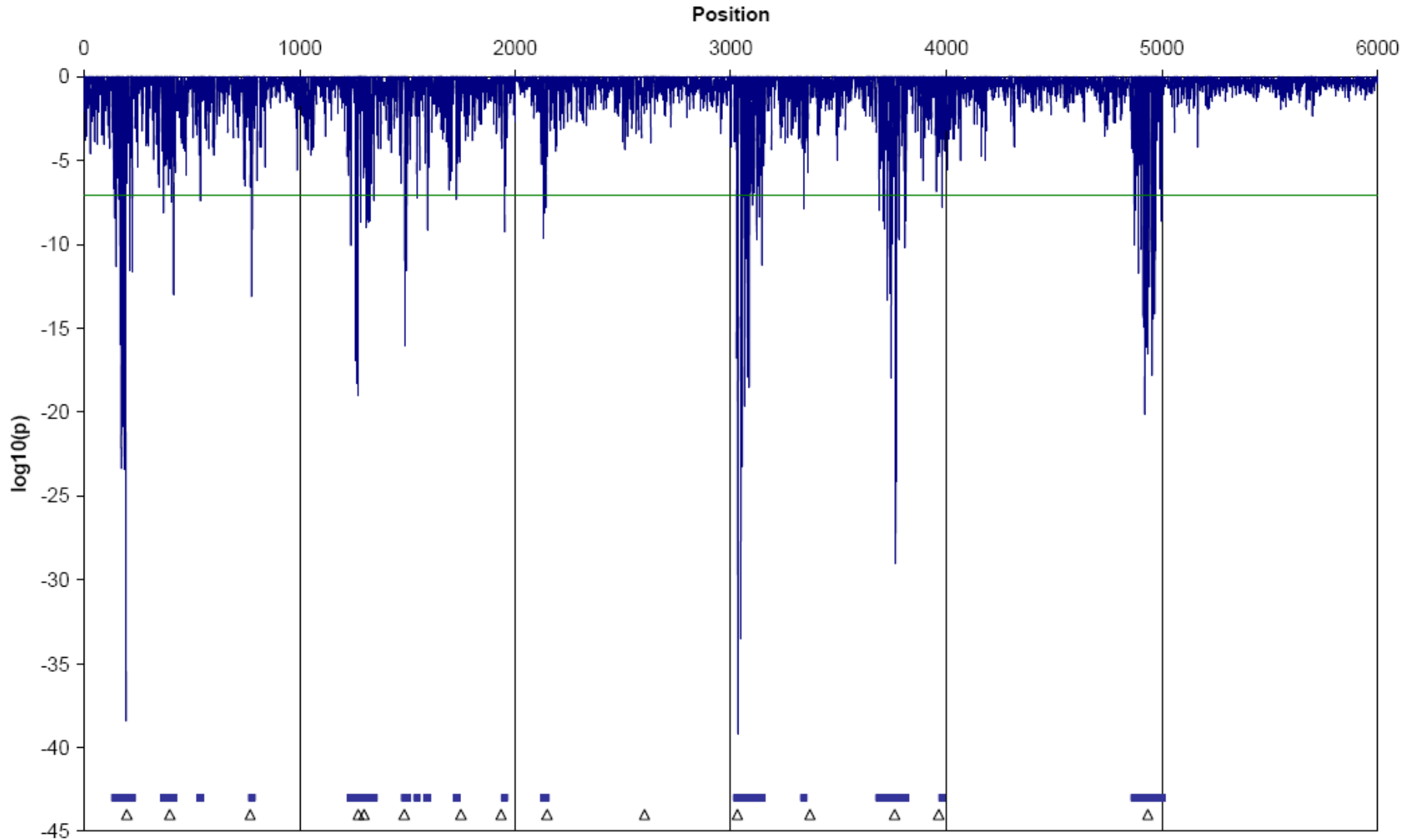


Decadic log of error probabilities

— critical value for genome-wide $p=0.001$ with Bonferroni correction

■ genome segments with significant signals

△ true position of QTLs explaining > 1 per cent of the genetic variance





Results

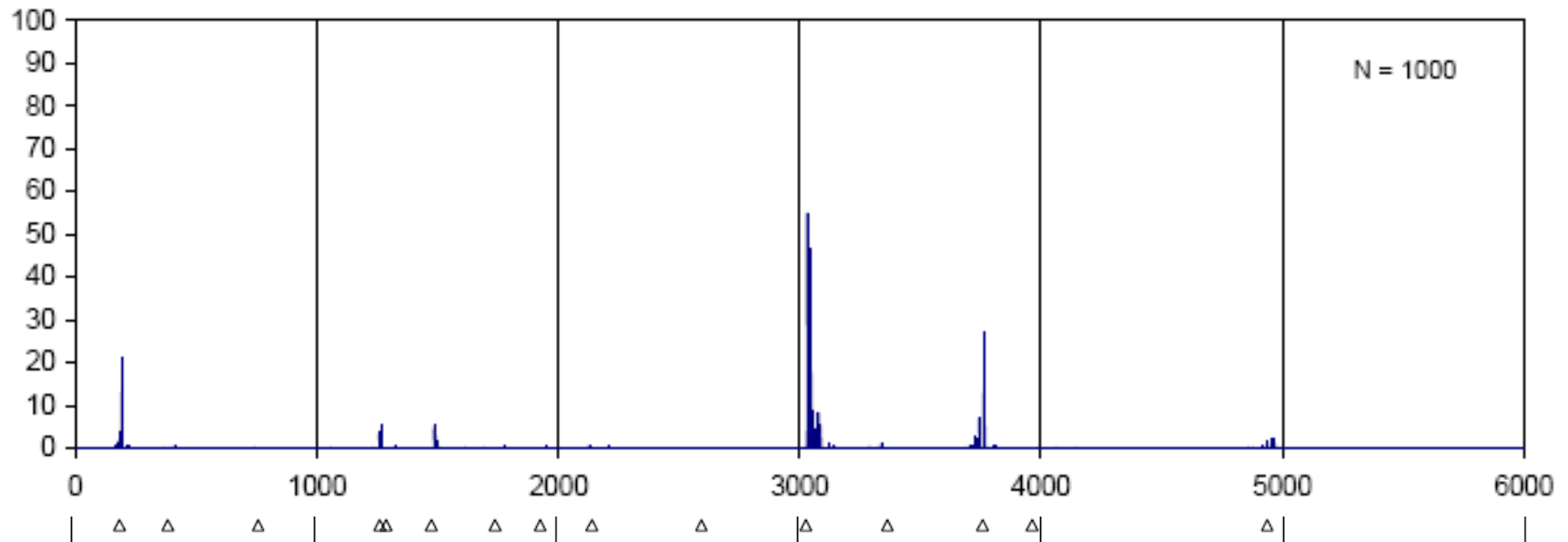
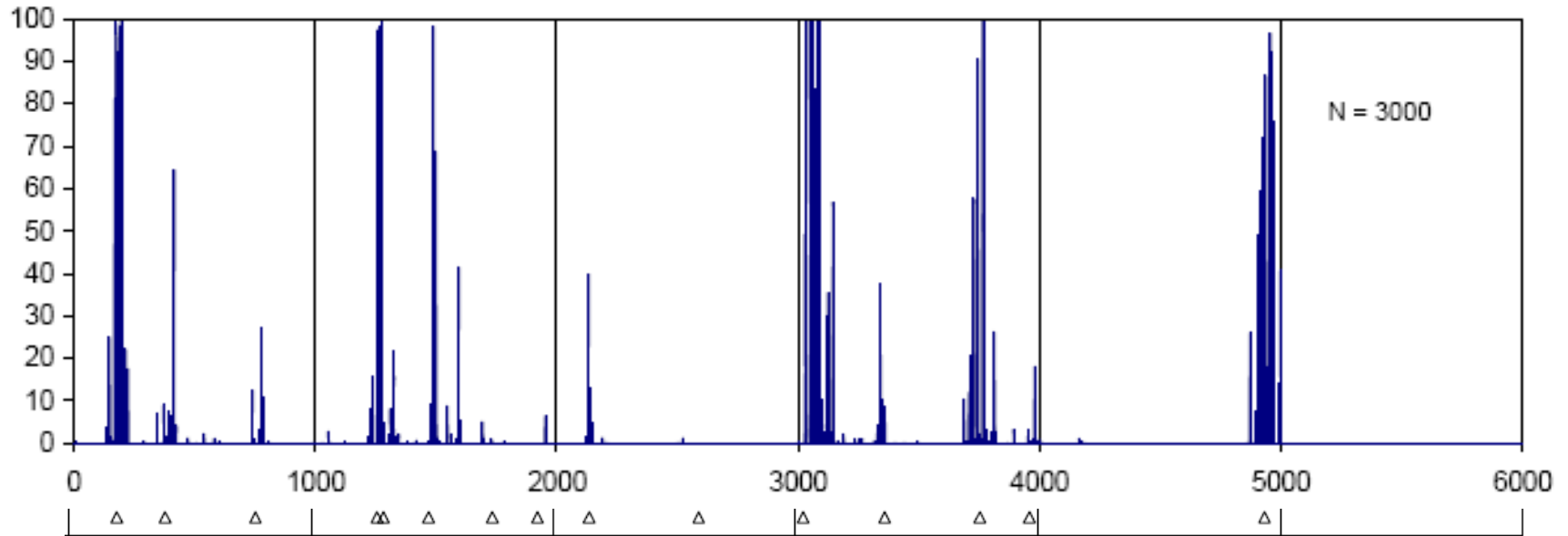
- 13 of 15 QTL mapped within ± 1 cM of the true position, one 2.8 cM from the true position, one completely missed
- Few false positives
- Computations are linearly proportional to #markers x #trios x #traits
- therefore computationally highly efficient (67.9 seconds on a 533 MHz double processor workstation for 6000 markers/4500 trios/1 trait)



Empirical power

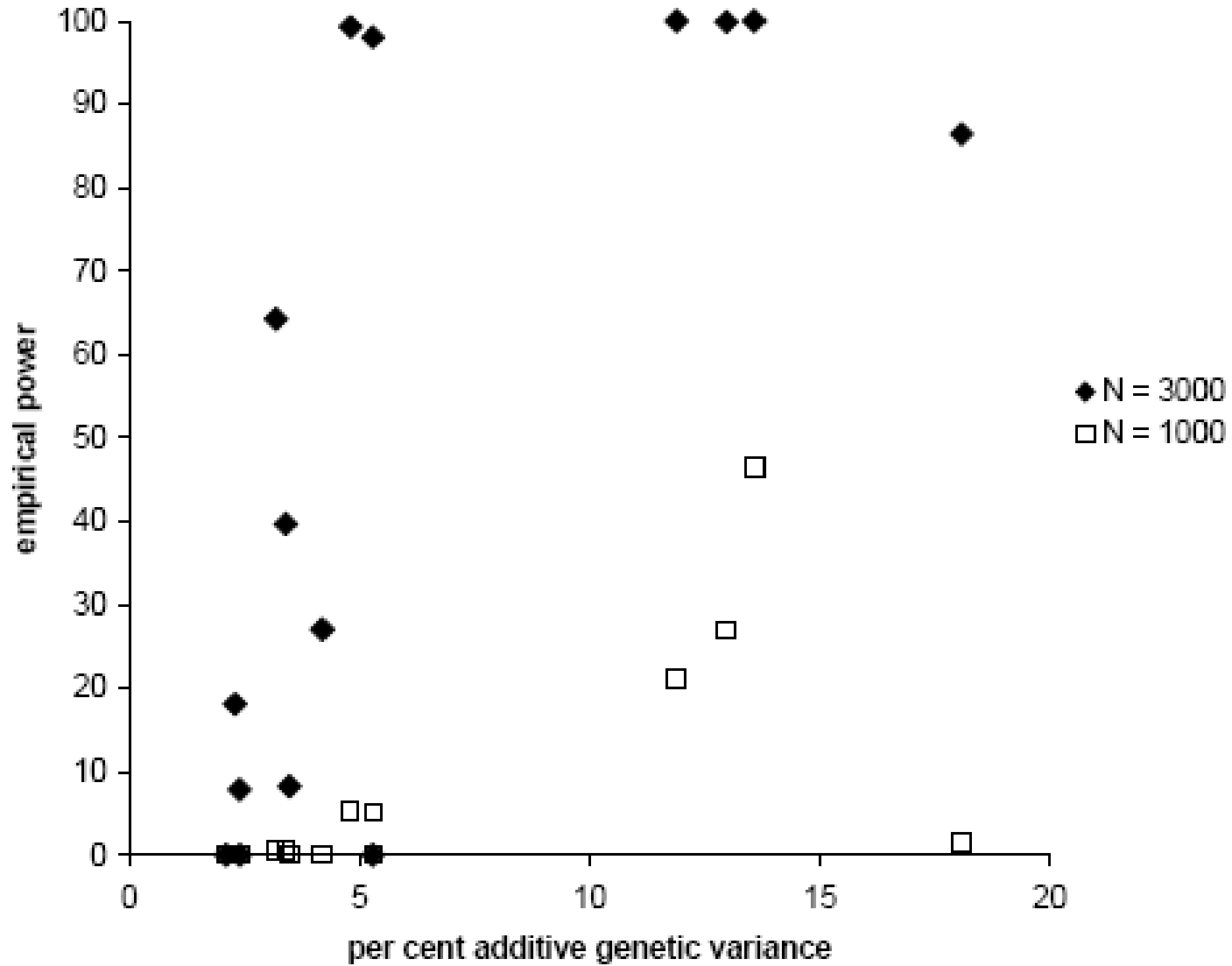
- Random samples of 3000 and 1000 genotyped individuals from the complete simulated data set (individuals sampled in trios)
- Estimated Mendelian sampling terms from the complete data
- 1000 replicates per sample size
- Significance level: 1 per cent genome-wide, two-sided, with Bonferroni correction
- Empirical power = proportion of significant results within ± 1 cM of the true QTL position

Empirical power





Empirical power





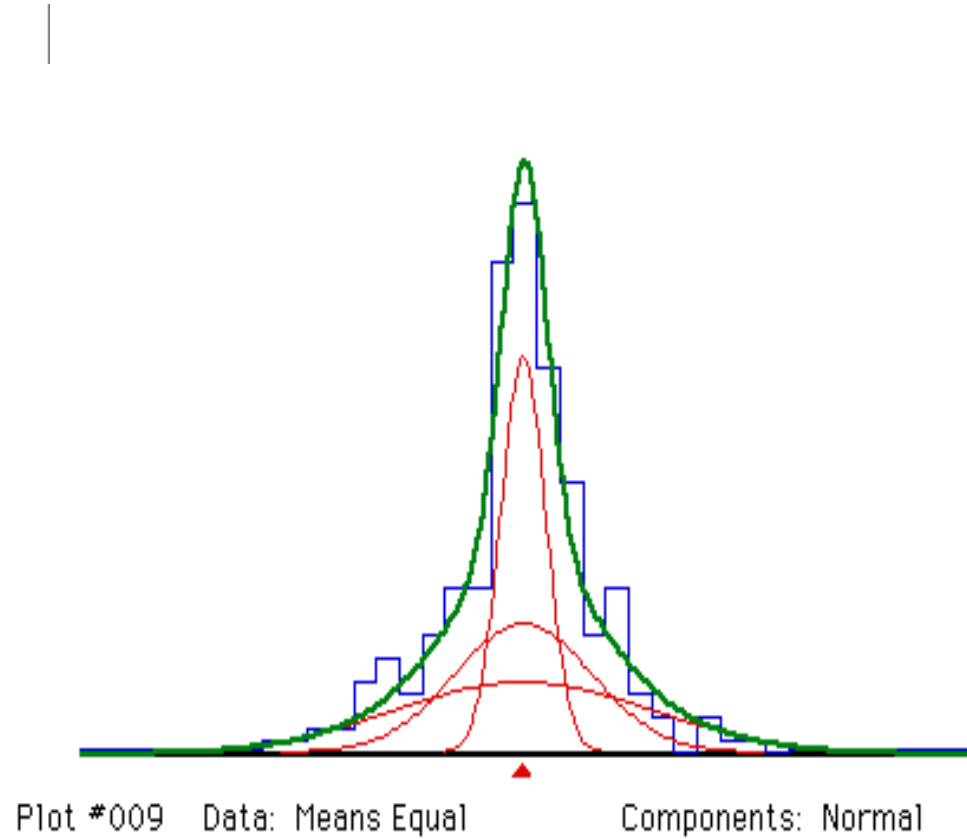
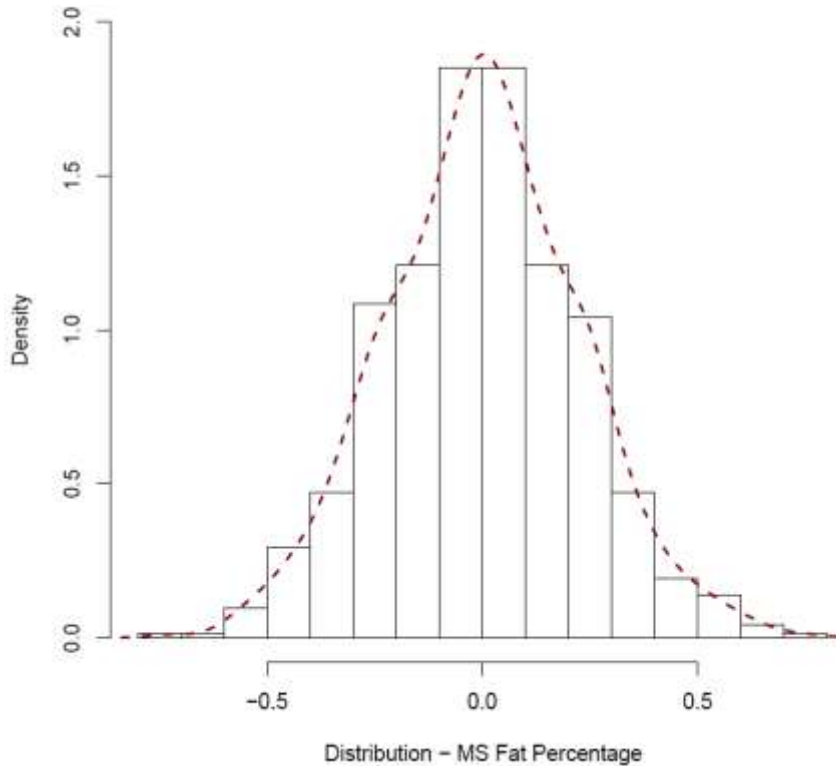
Application to real data

- 4916 SNPs
- for most traits 707 genotyped duos (707 sons of 66 genotyped sires) with estimated Mendelian sampling terms
- 11 traits
 - MY = Milk yield
 - FY = Fat yield
 - PY = Protein yield
 - FP = Fat percentage
 - PP = Protein percentage
 - LP = Lactation persistency
 - SCS = Somatic cell score
 - DCA = Daughter calving ability
 - DF = Daughter fertility
 - C = Conformation
 - LPI = Lifetime profit index



Application to real data

Distribution of estimated Mendelian sampling terms



Textbook example



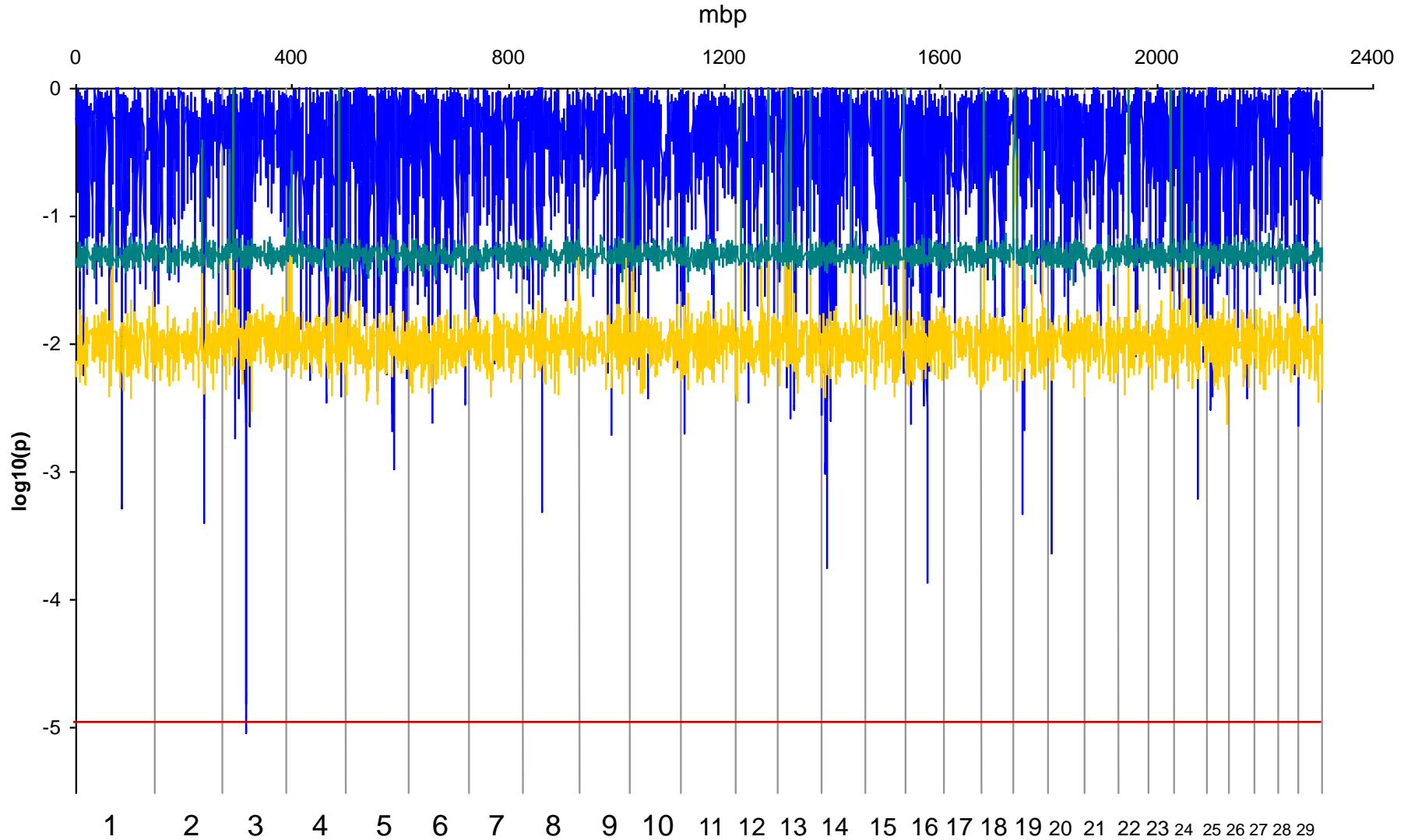
Application to real data

- Estimation of error probability (two sided t-test)
- Empirical thresholds via **permutation test** (Churchill and Doerge, 1994), 1000 permutations, for point-wise test with $\alpha = 0.05$ ($\log_{10}(\alpha) = -1.3$) and $\alpha = 0.01$ ($\log_{10}(\alpha) = -2$)
- Genome-wide theoretical significance threshold with $\alpha = 0.05$
 - $\log_{10}(\alpha/4916) = -4.99$

Application to real data



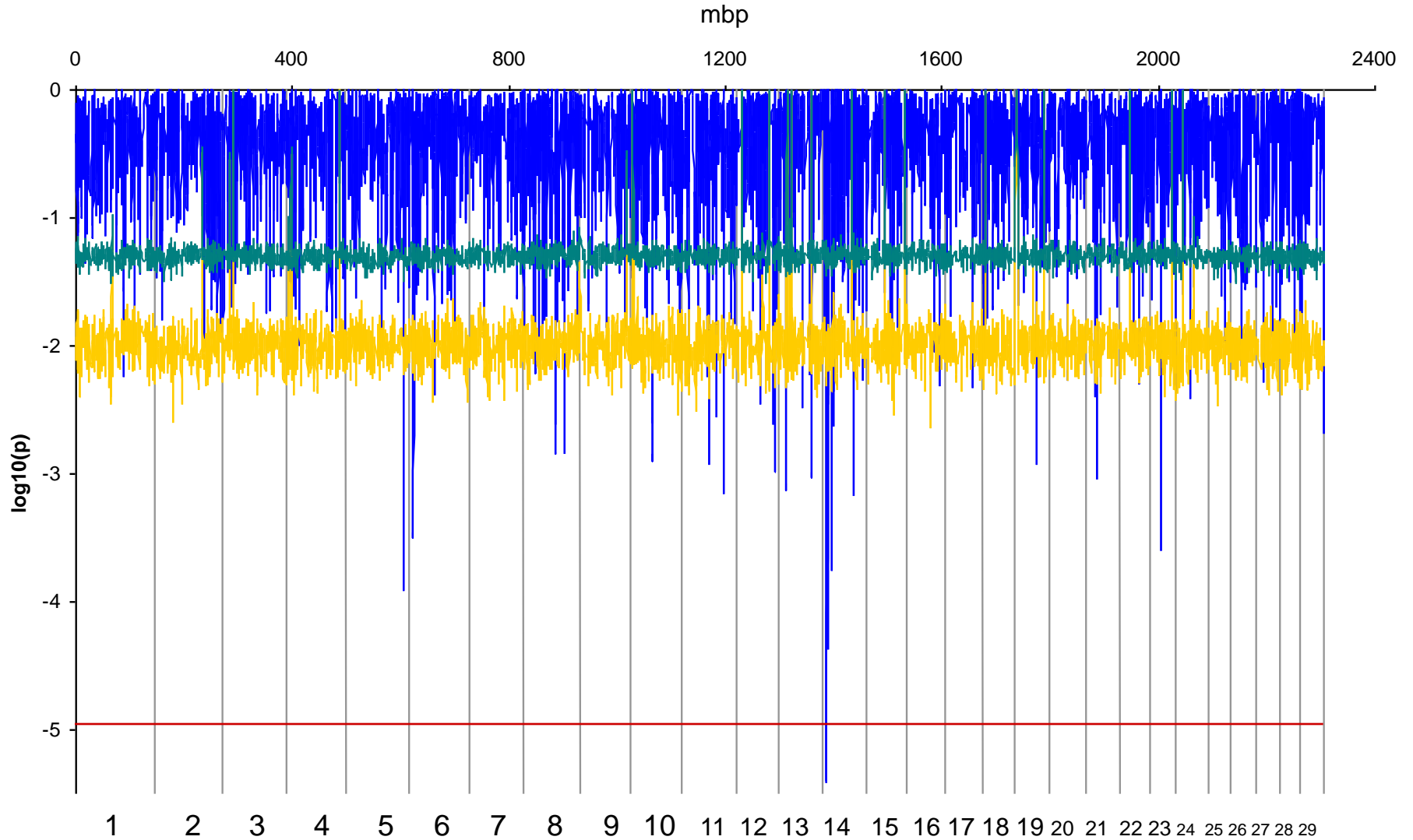
Milk yield



Application to real data



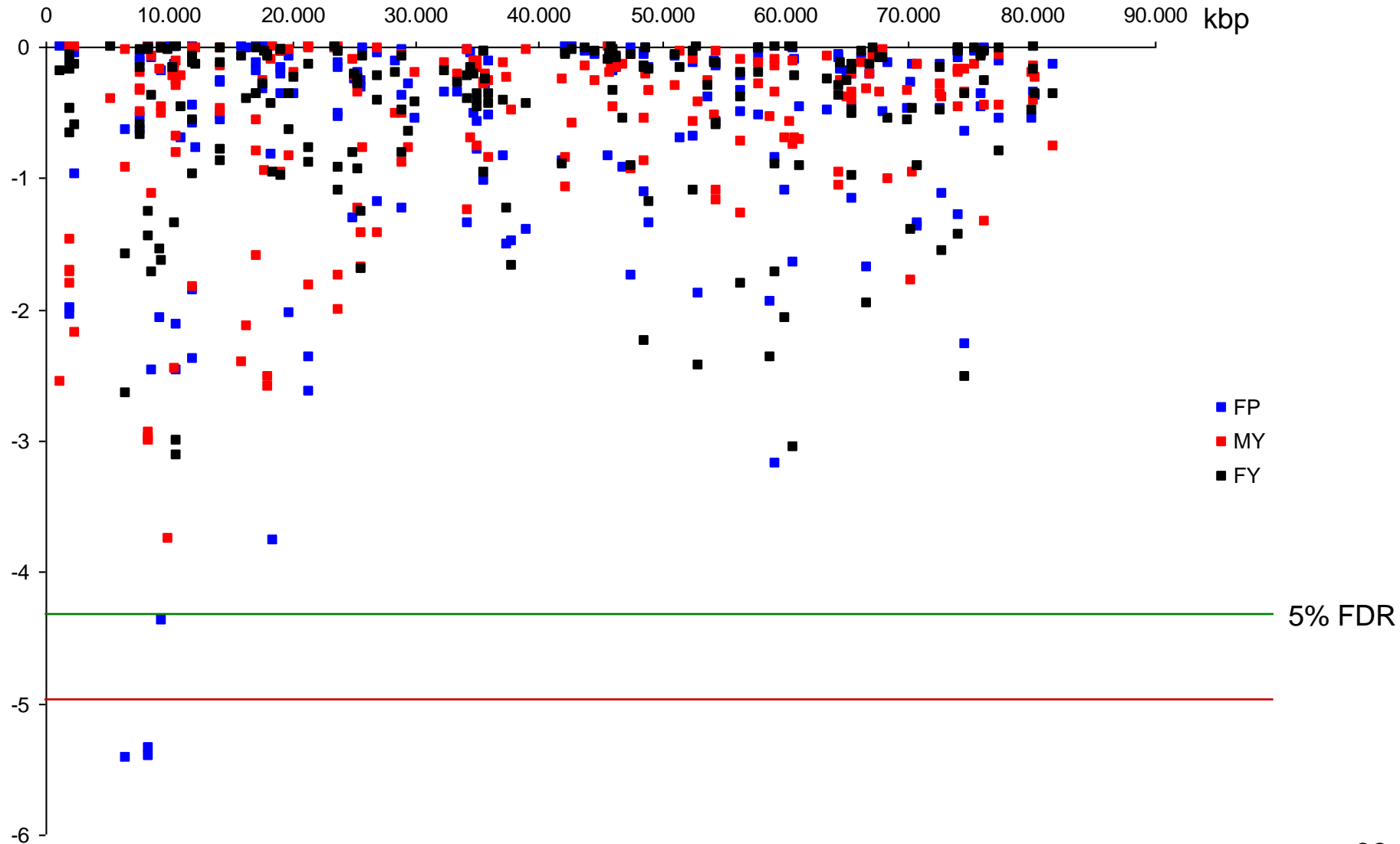
Fat percent

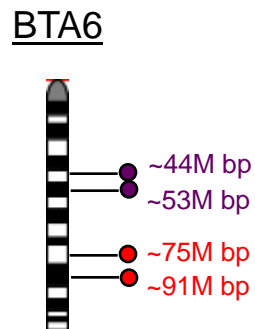
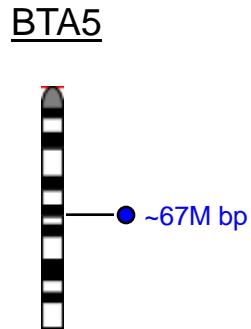
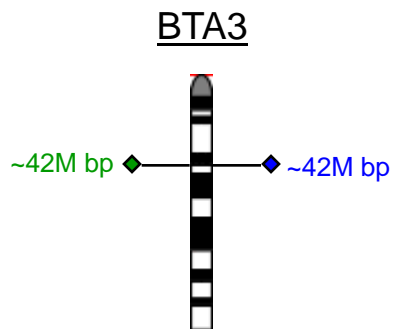


Application to real data



Chromosome BTA14





..... $\text{Log}_{10}(p) < -5.0$

— $\text{Log}_{10}(p) < -4.0$

◆ [MY]

◆ [PY]

◆ [FY]

◆ [FP]

◆ [PP]

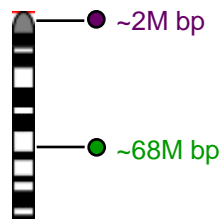
● [LP]

● [DCA]

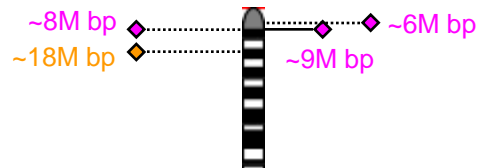
● [SCS]

● [LPI]

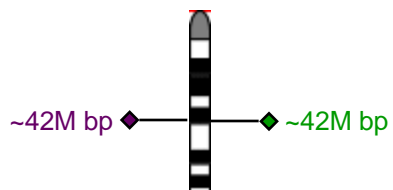
BTA8



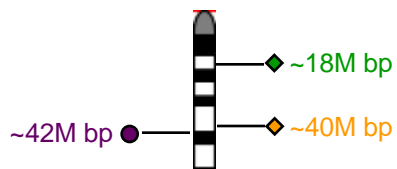
BTA14



BTA16



BTA19



BTA28





Discussion

The approach is **non-parametric**, δ is not an estimate of allele substitution effect a , but (under additivity):

$$E(\delta) = (2\pi - 1)(1 - 2\theta)\rho a$$

where

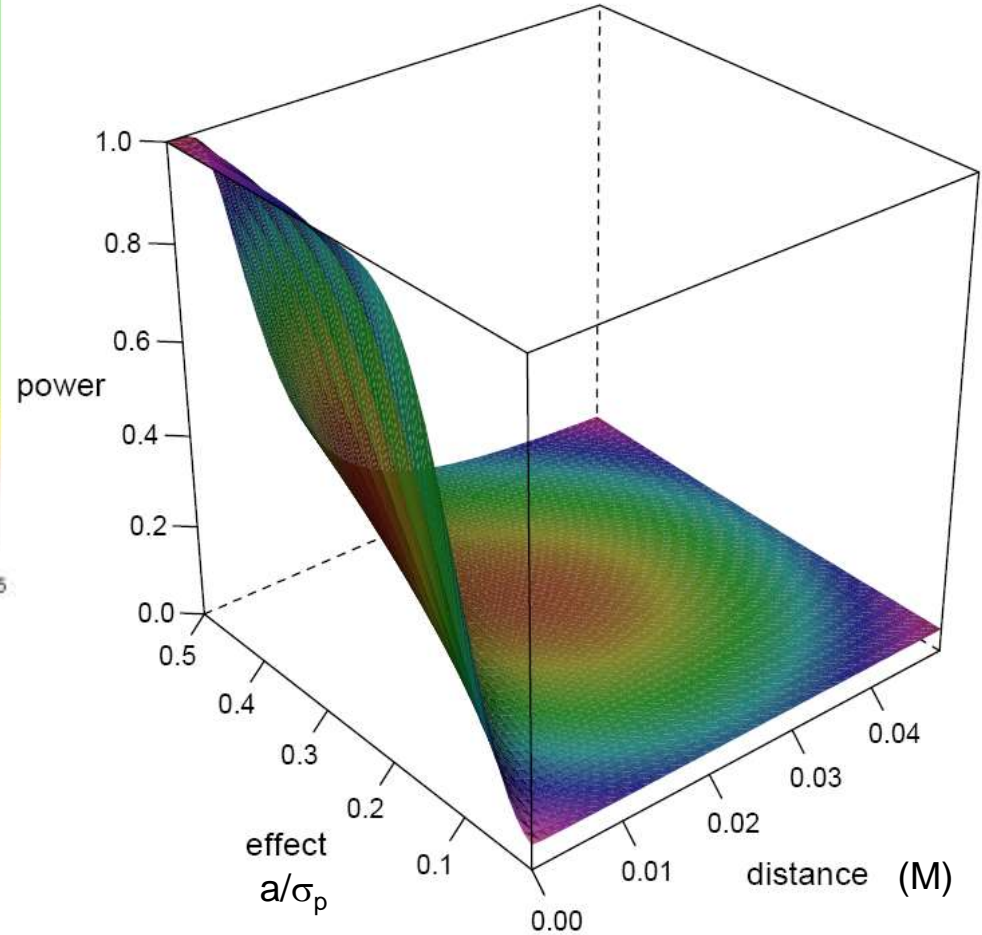
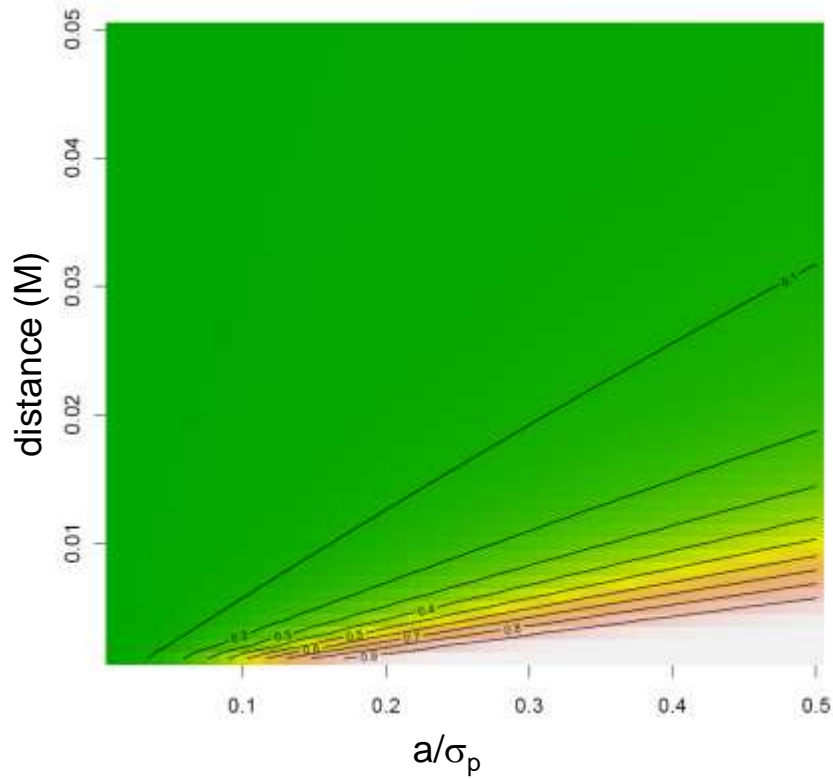
θ is the recombination rate between SNP and QTL [0 ≤ θ ≤ 0.5]

π is a disequilibrium parameter $\pi = P(Q | 1)$ [0 ≤ π ≤ 1]

ρ is the average accuracy of estimated Mendelian sampling [0 ≤ ρ ≤ 1]
(in simulated data $\rho = 0.45$, for real data/milk yield $\rho = 0.81$)



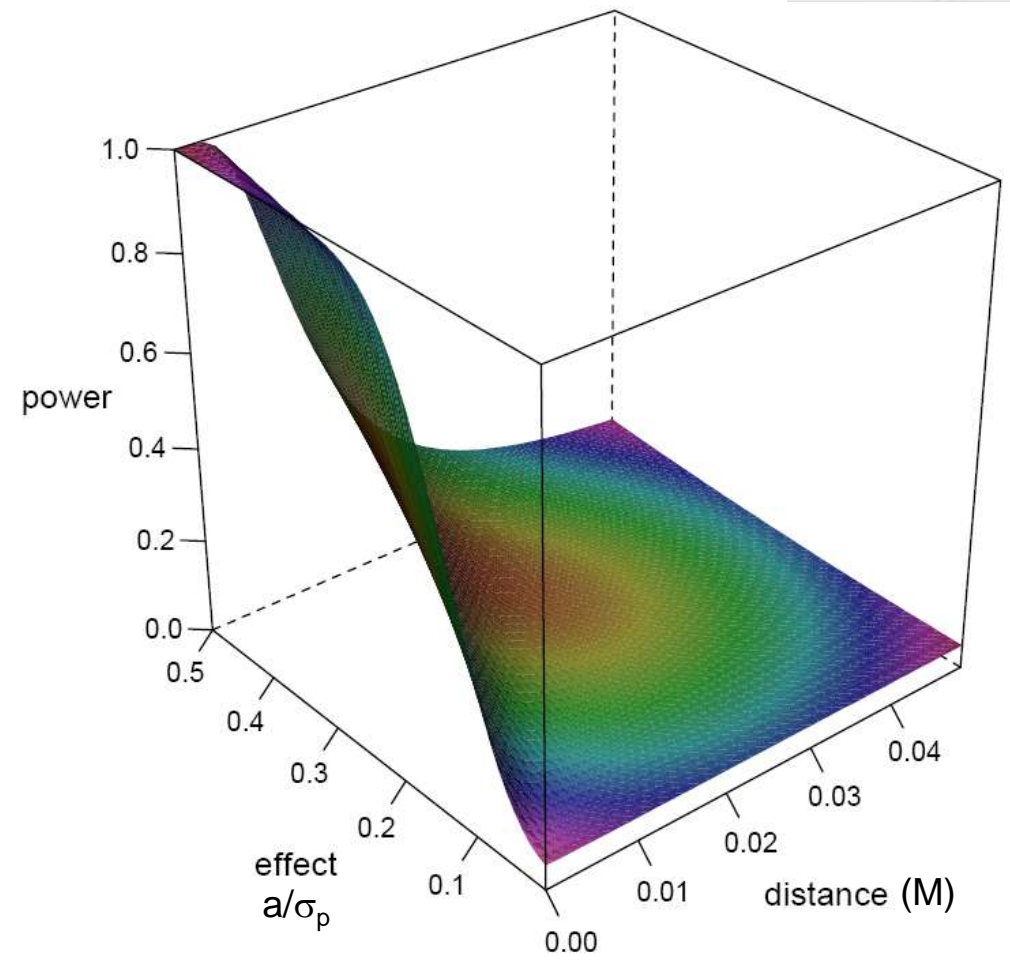
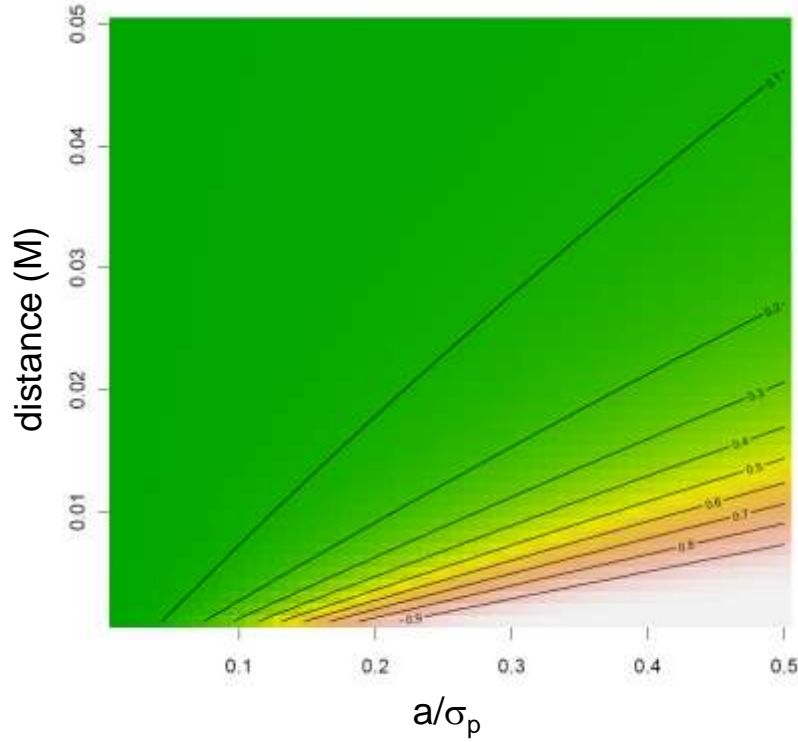
Power calculations



Simulation example:

- $N = 2250$ contrasts
- $N_e = 100$
- $\rho = 0.4$
- point-wise $\alpha = 0.05$ error level

Power calculations

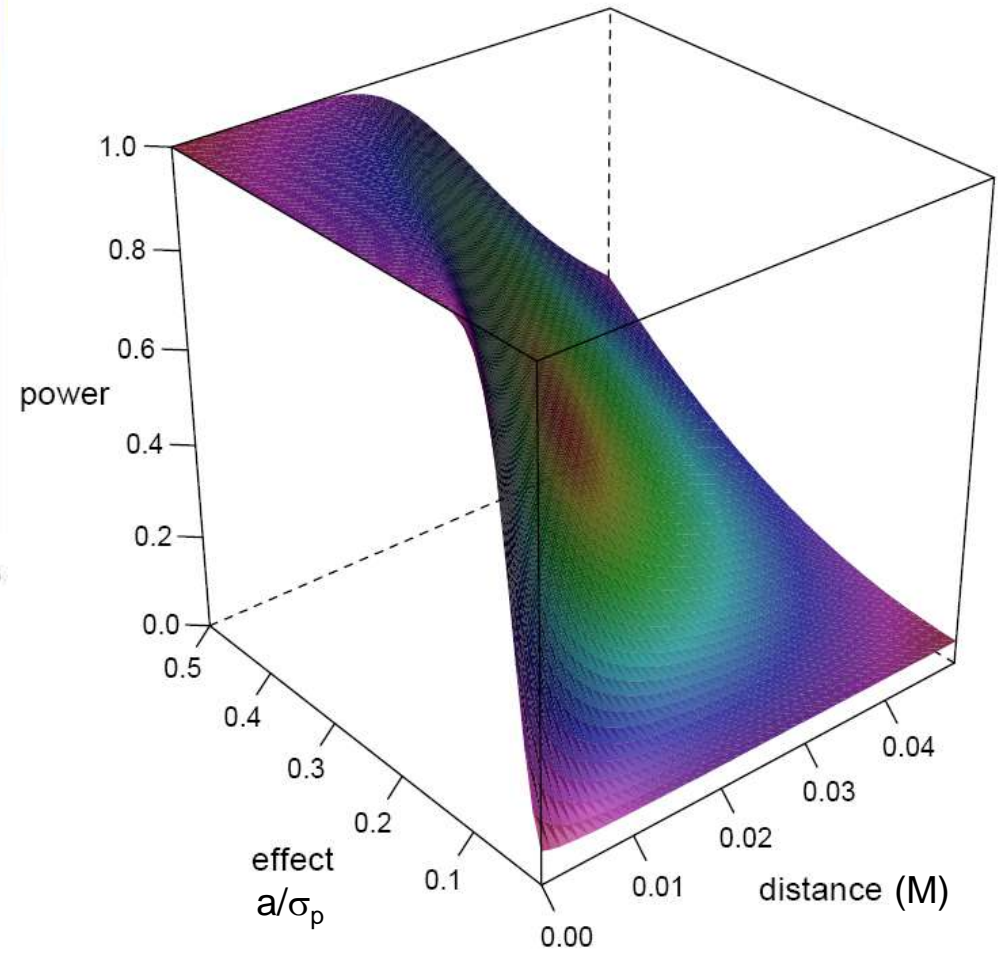
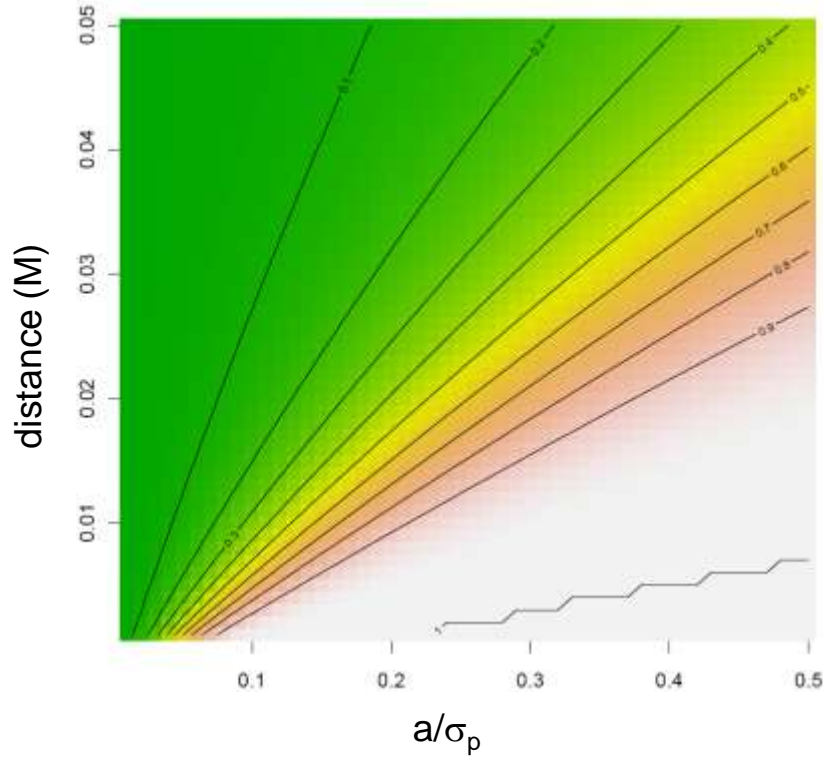


Dairy example:

- $N = 350$ contrasts
- $N_e = 50$
- $\rho = 0.8$
- point-wise $\alpha = 0.05$ error level



Power calculations



Dairy example - future

- $N = 3000$ contrasts
- $N_e = 50$
- $\rho = 0.8$
- point-wise $\alpha = 0.05$ error level



Discussion

But: is the genetic background of quantitative traits predominantly additive?

Model-free interpretation: The approach will pick up a signal if the accumulated genetic effects in linkage and linkage disequilibrium with the analysed marker are large enough to yield a significant deviation of δ from 0.



Summary and conclusions

- The suggested QTDT for Mendelian sampling terms combines efficiently **linkage and linkage disequilibrium** information and provides high mapping power and high positional resolution
- Estimated breeding values are available for many traits in most farm animal breeding systems, therefore estimated **Mendelian sampling terms** are easy to obtain
- The approach is **fast** enough to scan whole genomes with high marker density
- The approach is **non-parametric** and **model-free**, it only indicates QTL positions, extensions to the unbiased estimation of QTL-effects are in preparation
- Model-based **parametric approaches** (e.g. Meuwissen et al. 2001) can be used to further inspect the found QTL regions
- The approach has been successfully applied to **real data** (11 traits in a dairy population)



Summary and conclusions

- The TDT-concept provides a **safeguard** against **spurious associations** which are not due to linkage, but to population stratifications, which are expected to be of considerable relevance in selected farm animal populations
- The price for this safeguard is **reduced power** by not including all contrasts (e.g. homozygous parents)
- The approaches suggested for genomic breeding value estimation do not avoid this source of bias/noise – **room for improvement?**



ACKNOWLEDGEMENT

This study is part of the project FUGATO-plus GenoTrack and was financially supported by the German Ministry of Education and Research, BMBF, the Förderverein Biotechnologieforschung e.V. (FBF), Bonn, and Lohmann Tierzucht GmbH, Cuxhaven.



Bundesministerium
für Bildung
und Forschung



LOHMANN
TIERZUCHT