1st Leibniz-MMS-Days
January 27th-29th 2016, WIAS Berlin

Taking account for covariances in the Bayesian estimation of marker effects in backcross-experiments

Manuela Reichelt, Manfred Mayer, Friedrich Teuscher, Norbert Reinsch

Institut für Genetik und Biometrie,
Leibniz-Institut für Nutztierbiologie (FBN), 18196Dummerstorf

background

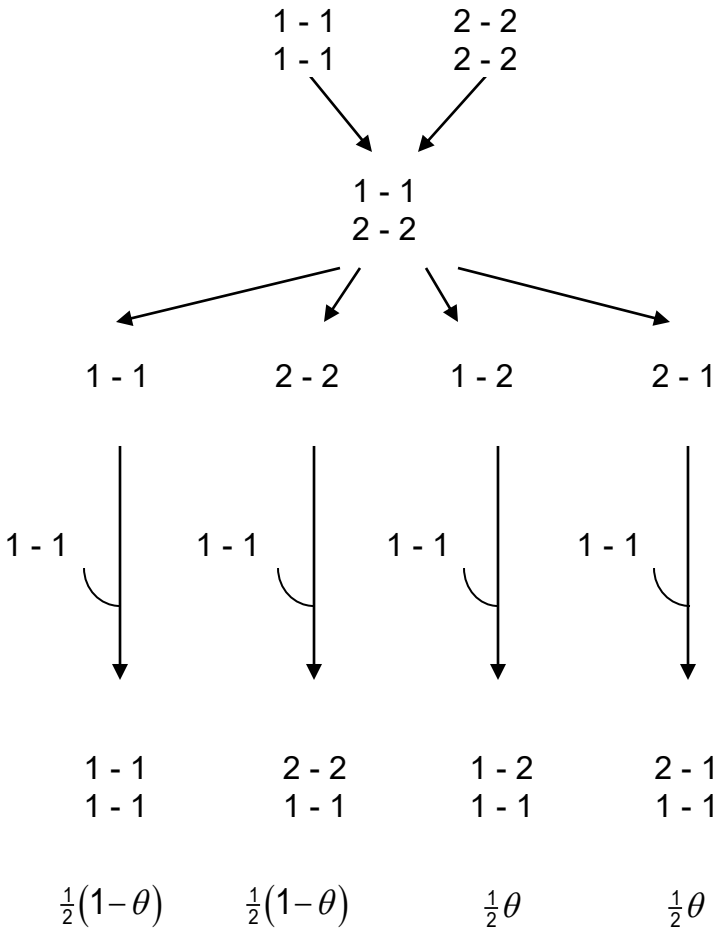
- high marker densities technically available at low cost => more similarities between ajacent marker genotypes
- covariance matrices between marker effects:
  - summarize knowledge on genetic maps (order and distance of markers)
  - are specific for type of experiment and type of effect
  - have been derived by our group
- question: are there **any benefits** from including such matices in QTL-analyses?
- treating marker effects as **independent** versus **correlated** compared by simulation
- simulations considered the case of a **backcross-experiment between inbred lines**

outline

- backcross-design
- covariance matrices between marker effects
- Baysian models: independent, dependent and „adaptively dependent“ marker effects
- simulated data
- results: comparisons between models

Backcross-experiment

marker coding indicates line origin



two inbred lines

F_1-cross: **parents**

gametes:
variability due to recombination

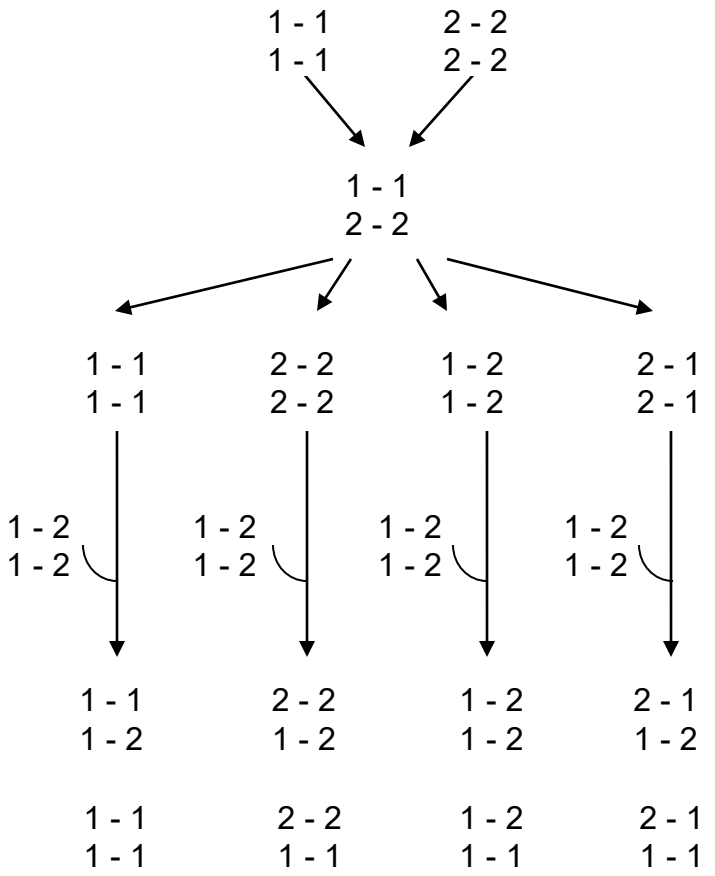
backcross to founder line: no variation between gametes

progeny: all variation from F1-parents

probabilities

backcross-effect at each marker : difference between 1,1 and 1,2

Equivalent Design with doubled Haploids (DH)



two DH lines

F_1-cross

double haploids

DH tester line (example)

final progeny

equivalent back-cross genotype (F_1 back-crossed to first founder line)

Backcross-experiment

marker coding indicates line origin

genotype	x_1	x_2	
1-1 1-1	-1	-1	frequent: $1-\theta,\ \ 0 \le \theta \le \frac{1}{2}$
2-2 1-1	+1	+1	concordant marker genotypes are frequent at adjacent loci
1-2 1-1	-1	+1	rare: θ
2-1 1-1	+1	-1	discordant marker genotypes are rare at adjacent loci

m_1, m_2 : backcross-effects at first and second marker : difference between 1,1 and 1,2 genotypes

x_1, x_2 : regressor variables

$$Var\begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = Var\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & exp(-2d) \\ exp(-2d) & 1 \end{bmatrix} = \mathbf{R}$$

d: distance between Markers in Morgan

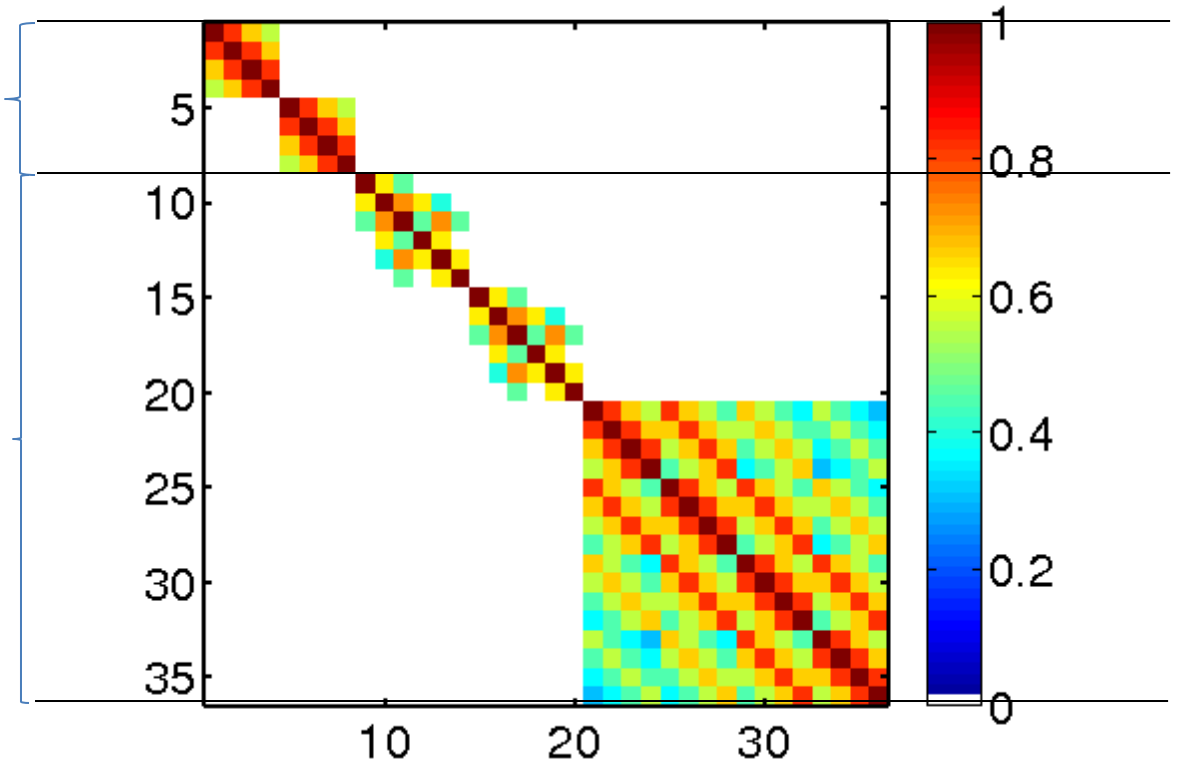
covariance of marker effects

backcross-effects

2 cromosomes a´ 4 markers
1 block per chromosome

pairwise interactions

with 2 2hromosomes: 3 blocks



is a block-diagonal matrix (when effects are suitably ordered)

without interactions: one block per chromosome

can directly be set up from a marker map

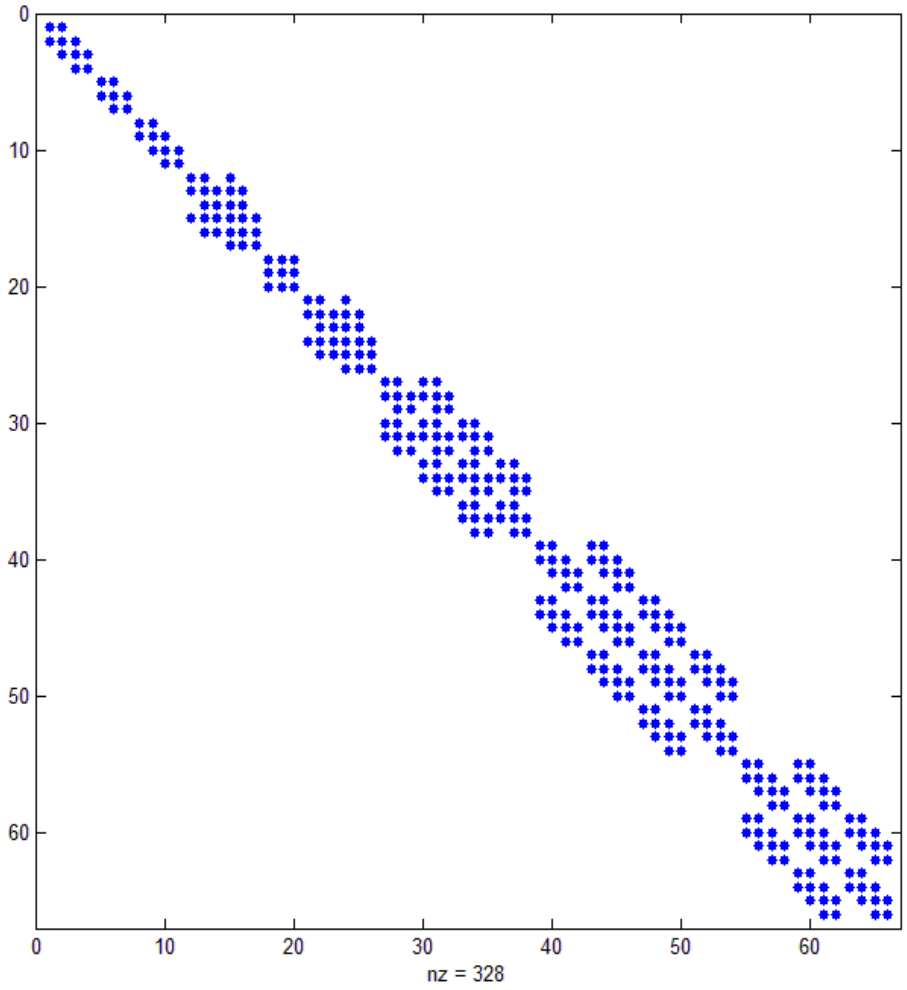
represents prior knowledge on the genetic map of a certain species

represents prior knowledge on parental diplotypes - depends on type of experiment

depends on type of experiment and type of effect

inverse covariance matrix

assuming Haldane's map function $\theta = \frac{1}{2}[1 - exp(-2d / 100)]$



is sparse

can directly be set up from a marker map

represents prior knowledge about genetic map of a species ….

inverse for backcross-effects:
autoregressive covariance structure

decomposition of inverse covariance matrix

$$R^{-1} = r^{11} \cdot \mathbf{LL}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\rho & 1 & 0 & 0 \\ 0 & -\rho & 1 & 0 \\ 0 & 0 & -\rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} 1 & -\rho & 0 & 0 \\ 0 & 1 & -\rho & 0 \\ 0 & 0 & 1 & -\rho \\ 0 & 0 & 0 & \sqrt{1-\rho^2} \end{bmatrix} \cdot r^{11}$$

$$= \begin{bmatrix} 1 & -\rho & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & -\rho & 1 \end{bmatrix} \cdot r^{11} \qquad \text{für M=4}$$

L can also be computed from the genetic map

multiple regression model

$\mathbf{m} = [\mathbf{m}_1, ..., \mathbf{m}_k, ...\mathbf{m}_C]$

vector of marker effects with subvectors m_k

each subvector contains all marker effects of a chromosome in the order of their position

each subvector has a corrsponding block $\mathbf{R}_k$ in $\mathbf{R}$

C: number of chromosomes

$\mathbf{y} = \mathbf{Xb} + \mathbf{Zm} + \mathbf{e}$

$\mathbf{y}$	vector of observations
$\mathbf{b}$	general mean
$\mathbf{m}$	vector of marker effects
$\mathbf{e}$	residual
$\mathbf{X}, \mathbf{Z}$	design matrices

three kinds of prior (in-)dependence

IN changing (marker-specific) variances on each chromosome, independent marker effects

$\mathbf{m}_k : N(0, \mathbf{D}_k)$ $\quad \mathbf{D} = diag\left\{\tau^2_{k,j}\right\}$

k, j: indicate chromosome and marker within chromosome

as described by Xu, Genetics 2003

DE constant variance across all chromosomes, dependent marker effects

$\mathbf{m}_k : N(0, \mathbf{R}_k\tau^2)$

general variance of marker effects τ^2

$\mathbf{R}_k$ is the kth diagonal block in **R**

Bayesian mixed model

AD changing variance on each chromosome, dependent marker effects

$\mathbf{m}_k : N(0, \mathbf{L}_k\mathbf{G}_k\mathbf{L}'_k\tau^2_k)$ uses Choloesky-decomposition $\mathbf{R}_k = \mathbf{L}_k\mathbf{L}'_k$

$\mathbf{G}_k = diag\left\{\gamma_{k,j}\right\}$

similar to Lang, Fronk, and Fahrmeir , 2001

Gibbs sampler

- marker effects drawn from normal distributions (e.g. Wang et al., GSE 1994)
- block sampling necessary for convergence (-> as known from Gaussian Markov Random Fields)
- variance parameters inverse Chi-square distributions in all cases, including residuals

IN $\tau^2_{k,j} \left| others : \frac{\hat{m}^2_{k,j}}{\chi^2_1} \right.$ as variance for each marker effect

DE $\tau^2 \left| \text{others} : \frac{\hat{\mathbf{m}}' \mathbf{R}^{-1} \hat{\mathbf{m}}}{\chi^2_{df}} \right.$ df: number of markers

AD $\tau^2_k \left| \text{others} : \frac{\hat{\mathbf{m}}'_k \mathbf{R}^{-1}_k \hat{\mathbf{m}}_k}{\chi^2_{df}} \right.$ df: number of markers on chromosome k

$\gamma_{k,j} \left| \text{others} : \frac{\left(\hat{m}'_{k,j} - \rho_{k,j;k,j+1} \hat{m}_{k,j+1} \right)^2 \tau_k^{-2}}{\chi^2_1} \right.$ $\rho_{k,j;k,j+1}$ is the regression of a marker effect on it's neighbour (from matrix **L**)

simulated QTL

CHROMOSOME 1:
position 21, Effekt 1.0
position 34, Effekt 1.0
position 96, Effekt 0.5

CHROMOSOME 2:
position 6, Effekt 0.5
position 51, Effekt 0.5
position 60, Effekt 0.25
position 67, Effekt 0.5

CHROMOSOME 3:
position 40, Effekt 0.1
position 67, Effekt 0.25
position 81, Effekt 0.1

CHROMOSOME 4:
position 34, Effekt -0.25
position 51, Effekt 0.25

12 linked QTL on 4 chromosomes, or

12 unlinked QTL on 12 chromosomes

number of QTL and size of their individual effects are equal in all cases

genetic variance differs due to linkage

total number of chromosomes: 20

18 different simulated scenarios

scenario				
h^2	chromosomes	markerspacing (CM)	σ_g^2	σ_e^2
0.17	12	1	3.27	16.0
	4	1	6.5364	31.93
0.29	12	1	3.27	8.01
	4	10	6.5364	16.0
		5	6.5364	16.0
		1	6.5364	16.0
0.70	12	1	3.27	1.4
	4	1	6.5364	2.8

heritabilities of 17%, 29% and 70%

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

500 observations per experiment, normal errors

marker spacings:

10 cM – 210 effects

5 cM – 420 effects

1 cM – 2020 effects

corresponds to:

p << n

p < n

p >> n

20 chromosomes of equal length (100cM), 8 (16) of them empty

200 experiments per scenario

criteria of comparison

- **average (true) variance of estimated genetic effects** (ability to quantify genetic variability)

$$\frac{1}{200}\sum_{j=1}^{200}\hat{\mathbf{m}}'_j\mathbf{R}\hat{\mathbf{m}}_j = \sigma^2_{\hat{g}}$$

- **evidence for a non-zero marker effects = evidence for a QTL** (ability of QTL detection)

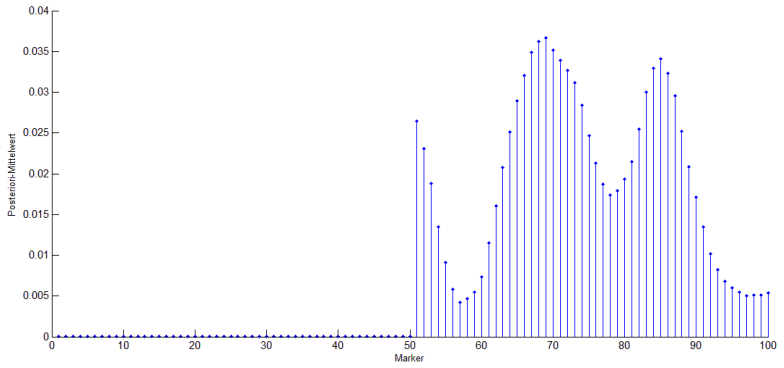
$p\left(\hat{m}_{k,j} \geq 0\right) > 0.95$ or

$p\left(\hat{m}_{k,j} \geq 0\right) < 0.05$

posterior probability of a positive marker effect exceeds 95% or is below 5%, evaluated at each marker

picture of estimated marker effects

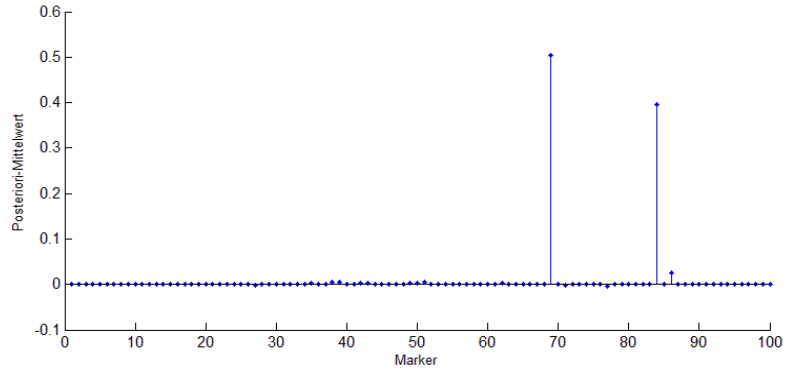
with covariance matrix: curve of marker effcts along the chromosome **DE** **AD**



τ_1^2 klein τ_2^2 groß

$\Rightarrow$ series of smaller effects

independent marker effects: „needles“ **IN**

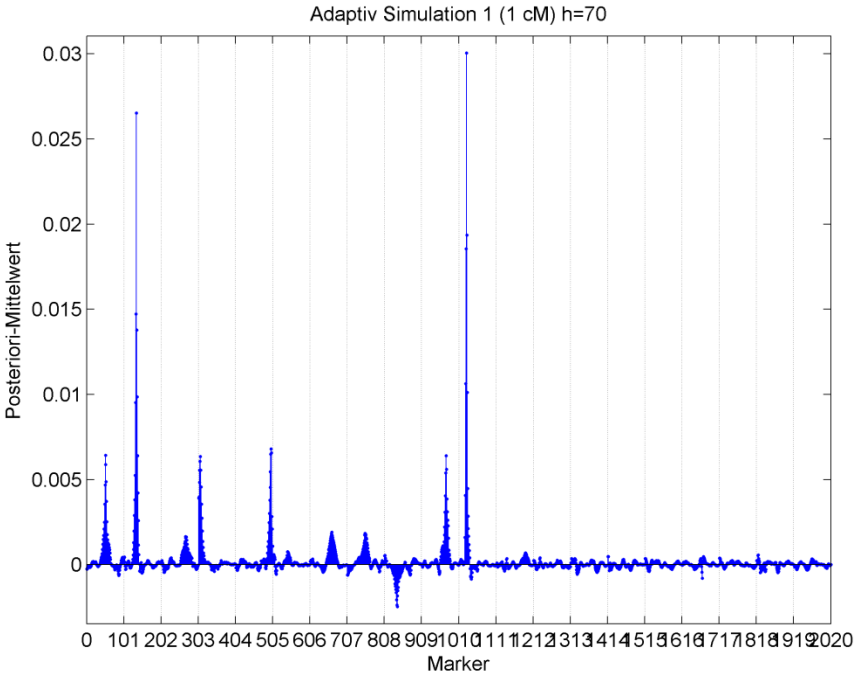


$\Rightarrow$ „needles“ indicate QTL

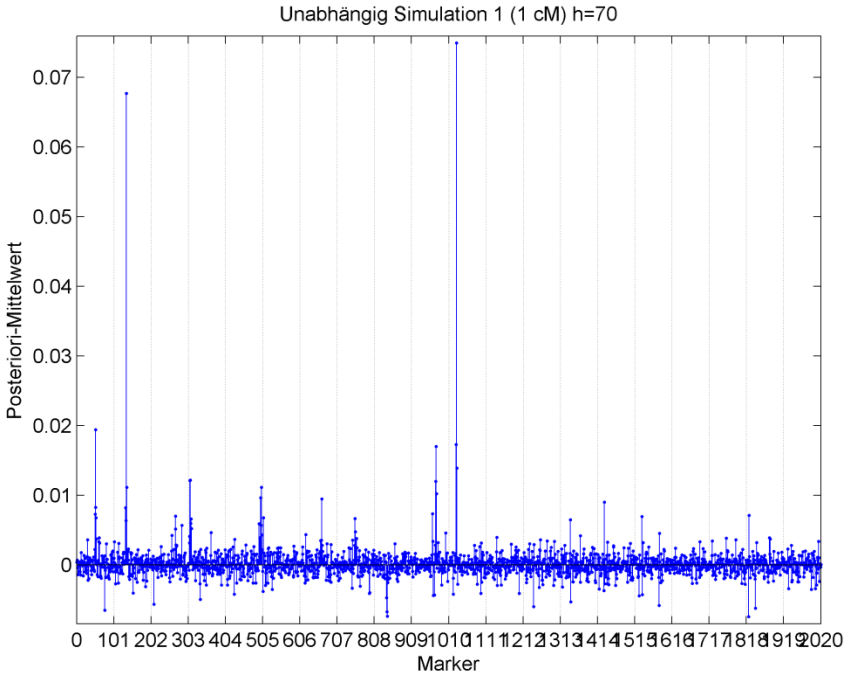
picture of estimated marker effects

posterior means averaged over experiments

AD



IN



Evidence for QTL

Scenario				Chromosome A			Chromosome B			Chromosome C			Chromosome D		
nr.	# chr	h^2	cM	**IN**	**DE**	**AD**	**IN**	**DE**	**AD**	**IN**	**DE**	**AD**	**IN**	**DE**	**AD**
1	12	0.17	10	0.31	0.87	0.94	0.63	0.93	0.95	0.04	0.09	0.21	0.03	0.14	0.08
2			5	0.11	0.88	0.91	0.20	0.90	0.88	0.01	0.13	0.18	0.02	0.10	0.07
3			1	0.01	0.91	0.36	0.00	0.90	0.30	0.01	0.04	0.03	0.01	0.18	0.02
4		0.29	10	0.59	1.00	0.99	0.91	1.00	1.00	0.19	0.41	0.59	0.09	0.38	0.22
5			5	0.26	1.00	0.98	0.51	1.00	1.00	0.03	0.49	0.52	0.05	0.31	0.18
6			1	0.01	1.00	0.73	0.01	1.00	0.70	0.01	0.29	0.08	0.01	0.46	0.04
7		0.70	10	1.00	1.00	1.00	1.00	1.00	1.00	0.94	1.00	1.00	0.40	1.00	0.94
8			5	0.73	1.00	0.98	1.00	1.00	1.00	0.67	1.00	0.99	0.73	1.00	0.99
9			1	0.30	1.00	0.98	0.31	1.00	0.99	0.01	1.00	0.81	0.02	1.00	0.70
10	4	0.17	10	0.33	0.99	0.99	0.08	0.77	0.81	0.01	0.01	0.08	0.00	0.00	0.01
11			5	0.07	1.00	0.97	0.02	0.73	0.74	0.01	0.02	0.05	0.00	0.00	0.02
12			1	0.01	1.00	0.26	0.01	0.79	0.07	0.01	0.02	0.02	0.01	0.01	0.02
13		0.29	10	0.57	1.00	0.99	0.18	0.98	0.98	0.01	0.05	0.12	0.00	0.00	0.02
14			5	0.23	1.00	0.98	0.03	0.95	0.97	0.01	0.07	0.10	0.00	0.01	0.02
15			1	0.02	1.00	0.59	0.01	0.98	0.20	0.01	0.04	0.01	0.01	0.01	0.02
16		0.70	10	1.00	1.00	1.00	0.65	1.00	1.00	0.06	0.79	0.88	0.03	0.01	0.05
17			5	0.98	1.00	1.00	0.37	1.00	1.00	0.01	0.86	0.85	0.01	0.02	0.05
18			1	0.15	1.00	0.98	0.03	1.00	0.81	0.01	0.88	0.09	0.01	0.02	0.04

Table 3:: Maximum frequency of detected QTLs on several chromosomes. For the first part (scenario 1 to 9) Chromosome A presents chromosome 2, Chromosome B stands for chromosome 11, Chromosome C for chromosome 1 and Chromosome D for chromosome 5. The second part presents chromosome 1 to 4 for scenario 10 to 18 with dependent QTLs.

Comparison of methods: evidence for a QTL

High marker density (p>>n) in combination with models assuming independence (**IN**) leads to a breakdown of the ability to detect QTL

Same phenomenon, but weaker for adaptive models (**AD**)

Models including covariances (**DE** und **AD**) in average (over all simulated experiments) exhibit a higher QTL detection probability compared to models without (**IN**)

comparison: estimated genetic variance

scenario						mean estimate: σ_g^2 $(rMSE)$			mean estimate: $\hat{\sigma}_e^2$ $(rMSE)$		
nr.	# chr	h^2	cM	σ_g^2	$\hat{\sigma}_e^2$	**IN**	**DE**	**AD**	**IN**	**DE**	**AD**
1	12	0.17	10	3.27	16.00	3.33 (0.05)	1.92 (0.11)	2.57 (0.08)	15.60 (1.13)	16.90 (1.40)	16.51 (1.16)
2			5	3.27	16.00	4.23 (0.06)	1.96 (0.11)	3.10 (0.06)	14.82 (1.57)	16.94 (1.42)	16.17 (1.06)
3			1	3.27	16.00	12.66 (0.47)	1.98 (0.11)	6.22 (0.15)	7.17 (8.91)	16.84 (1.80)	14.64 (2.02)
4		0.29	10	3.27	8.01	3.16 (0.03)	2.24 (0.08)	2.66 (0.06)	7.93 (0.55)	8.70 (0.90)	8.48 (0.72)
5			5	3.27	8.01	3.68 (0.03)	2.28 (0.08)	2.98 (0.04)	7.47 (0.75)	8.70 (0.90)	8.25 (0.60)
6			1	3.27	8.01	8.27 (0.25)	2.30 (0.08)	4.70 (0.08)	3.41 (4.64)	8.69 (0.90)	7.45 (0.79)
7		0.70	10	3.27	1.40	3.09 (0.02)	2.81 (0.04)	2.97 (0.03)	1.52 (0.17)	1.69 (0.31)	1.63 (0.26)
8			5	3.27	1.40	3.28 (0.01)	2.85 (0.03)	3.10 (0.02)	1.34 (0.11)	1.65 (0.28)	1.51 (0.16)
9			1	3.27	1.40	4.45 (0.07)	2.86 (0.03)	3.46 (0.01)	0.48 (0.93)	1.65 (0.28)	1.33 (0.12)
10	4	0.17	10	6.54	31.93	7.15 (0.13)	5.68 (0.14)	6.23 (0.14)	30.72 (2.41)	32.22 (2.07)	32.03 (2.05)
11			5	6.54	31.93	8.67 (0.15)	5.70 (0.14)	6.97 (0.12)	29.39 (3.26)	32.30 (2.06)	31.61 (2.08)
12			1	6.54	31.93	27.35 (1.18)	5.66 (0.15)	12.61 (0.31)	14.99 (17.08)	32.27 (2.06)	29.18 (3.50)
13		0.29	10	6.54	16.00	6.66 (0.09)	5.90 (0.10)	6.13 (0.11)	15.55 (1.16)	16.32 (1.09)	16.28 (1.09)
14			5	6.54	16.00	7.50 (0.09)	5.91 (0.10)	6.55 (0.09)	14.80 (1.58)	16.35 (1.09)	16.02 (1.05)
15			1	6.54	16.00	17.57 (0.62)	5.88 (0.11)	9.52 (0.16)	7.19 (8.88)	16.35 (1.10)	14.71 (1.69)
16		0.70	10	6.54	2.80	6.37 (0.05)	6.22 (0.05)	6.24 (0.05)	2.85 (0.21)	3.00 (0.28)	3.01 (0.29)
17			5	6.54	2.80	6.63 (0.03)	6.25 (0.05)	6.37 (0.04)	2.62 (0.26)	2.96 (0.25)	2.91 (0.22)
18			1	6.54	2.80	8.73 (0.13)	6.25 (0.04)	6.97 (0.04)	1.05 (1.76)	2.96 (0.25)	2.61 (0.27)

Table 2:: Simulated and estimated genetic variance, residual variance and root mean squared error $(rMSE)$ of all scenarios.

comparison: estimated genetic variance

Independent marker effects (**IN**) in the case of p>>n leads to a heavy overestimation of the genetic variance

Same tendency in the case of p>>n with method **AD**, but less severe

Estimates of genetic variability (and residual variance) hardly affected with method **DE**, tendency for a weak underestimation

Inclusion of covariances (**DE** or **AD**) in case of (p>>n) always results in more realistic estimates of the genetic variance, irrespective of the localisation of QTL

conclusions and outlook

more realistic estimates of genetic variation when covariances are included

superior ability of QTL detection when covariances are part of the model

covariance-models obviously better cope with a growing number of parameters

further research:

- **other types of families**
- **other types of effects (dominance, interactions)**
- **fine-tuning of models**

Thank you for listening !