Genome-wide association studies using a Bayesian dominance model

Jörn Bennewitz, Theo Meuwissen & Robin Wellmann

Institute of Animal Science

University Hohenheim, Germany

Institute of Animal and Aquacultural Science

Norwegian University of Life Science

Single-marker GWAS

- **One SNP at a time, mixed models with fixed SNP substitution effect, simple & fast calculations (ASReml, GCTA, PLINK, …).**

- **Produces a ‚*p-value*‘, convienient to use for post-GWAS calculations (e.g. Bonferroni, FDR, meta-analysis).**

- **Many associations, but explained variance by mapped QTL is small due to imperfect LD & small QTL effects.**

- **Neighboured SNP may explain jointly much more QTL variance than any SNP by itself.**

Multi-marker GWAS

- **GS-methods: fitting all markers simultaneoulsy. Population structure is well approximated (even in admixed populations).**
- **Marginal marker effects (effects not explained by other markers).**
- **Window approach: explained variance of markers within a windows (e.g. 1 cM in size).**
- **BayesC and BayesR probably most used gs-methods for GWAS.**

Aim of the study: BayesC vs. BayesD for GWAS

- **BayesC (Verbyla et al. 2009) uses priors about e.g. the distribution of additive effects and the proprotion of important markers, but dominance is not considered.**

- **BayesD (W. & B. 2012) is an extension of BayesC towards accounting für dominance effects.**

Aim of the study:

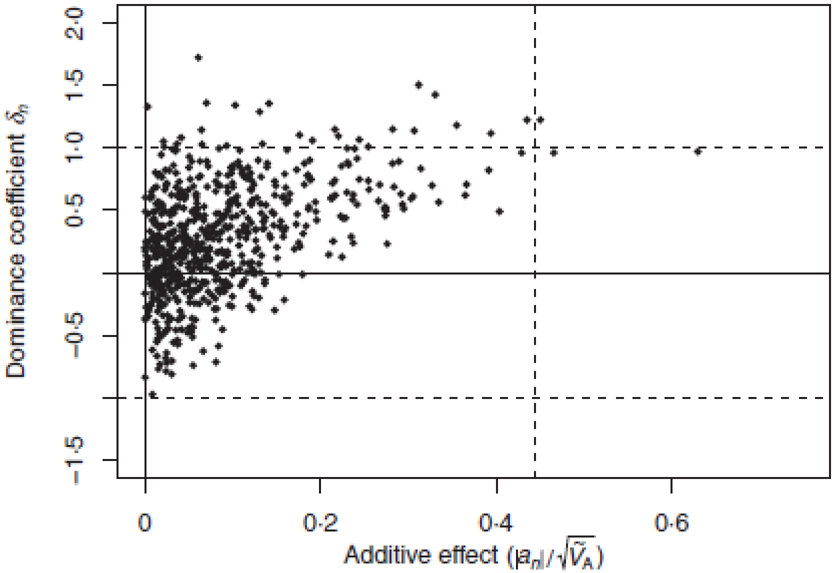
Can we improve power and precision of QTL mapping when using BayesD compared to BayesC?

Simulation protocol

- **Fischer-Wright populations, various marker densities & full sequence data**

- **In the last gen 15 SNP/chr randomly selected to become a QTL**

- **QTL additive and dominance coefficients (delta) sampled based on what is known about their dependencies**



$$\tilde{\delta}_n \sim \mathcal{N}(0{\cdot}2, 0{\cdot}3^2)$$

$$\tilde{a}_n|\tilde{\delta}_n \sim \mathcal{N}(0, \exp(3\tilde{\delta}_n))$$

Simulation protocol

- **Calculation of breeding values and of dominance deviations of the individuals using standard notations.**

- **Residuals sampled in order to obtain narrow sense h_2=0.3.**

- **Sampling of additive and dominance QTL effects results in average d_2=VD/VP=0.1 (range:0.01-0.29),**

- **this range fits nicely to cattle literature reports (Bolormaa et al. 2015).**

The BayesD-model

We consider a linear regression model of the form

$$y = X\beta + Z_A a + Z_D d + Zu + E,$$

where

y phenotypic observations

β vector of fixed effects that includes the overal mean

a vector of additive effects of the markers

d vector of dominance effects of the markers

u vector of other normally distributed random effects

E vector of normally distributed errors

X, Z_A, Z_D, Z design matrices.

BayesD: Method 2 from Wellmann (W. u. B. 2012)

- **Extension of BayesC towards accounting for dominance.**
- **Prior distribution of additive effects: Mixture of two t-distributions, which differ by a scaling factor.**
- **Prior prob that a marker is important (belongs to the distribution with larger variance): pLD.**
- **Prior assumption: independence of |a| and delta = d / abs(a).**
- **Small prob that d is much larger than a (i.e. overdominance is a rare but not neglible event)**

Bayes for GWAS

- **Sliding window approach (size: 0.25, 0.5 and 1 cM).**

- **Window variance of estimated genomic values of individuals calculated using standard notations.**

- **'Test-statistic': Window Posterior Probability of Association, controls Proportion of False Positives (WPPA, R. Fernando, 2014)**

Calculation of power and precision

- 10 Populations and 5 traits per population (50 replicates) simulated and analysed.

- A QTL is mapped if at least one window around the true QTL position shows a WPPA above a defined threshold.

- Power = #(mapped QTL) / #(number of QTL).
- Power_large = #(mapped large QTL) / #(number of large QTL).

- Mapping precision is measured as the size around the QTL with significant windows in cM.

Results from simulations: window size 0.5 cM

Marker density	WPPA	BayesC		BayesD	
		Power_large	Precision	Power_large	Precision
0.5K	0.85	0.55	1.00	0.54	1.02
	0.95	0.44	0.99	0.44	0.98
	0.99	0.27	0.97	0.36	0.88
1K	0.85	0.58	0.94	0.62	0.95
	0.95	0.43	0.93	0.51	0.91
	0.99	0.37	0.92	0.38	0.93
2K	0.85	0.60	0.90	0.66	0.88
	0.95	0.50	0.91	0.51	0.88
	0.99	0.43	0.92	0.41	0.89

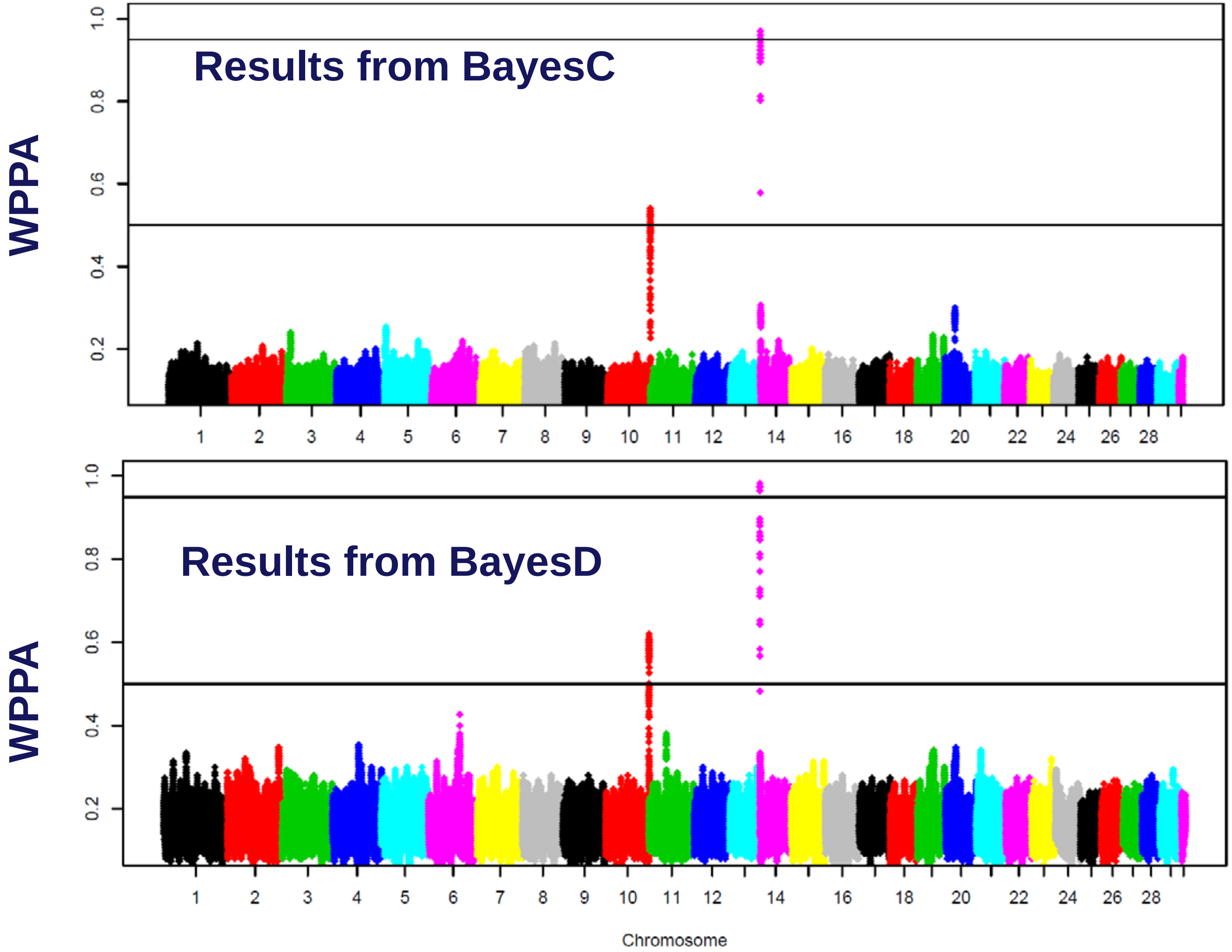
Results from simulations: window size 1 cM

Marker density	WPPA	BayesC		BayesD	
		Power_large	Precision	Power_large	Precision
0.5K	0.85	0.59	1.75	0.61	1.76
	0.95	0.45	1.70	0.49	1.73
	0.99	0.34	1.66	0.39	1.64
1K	0.85	0.64	1.73	0.69	1.74
	0.95	0.53	1.68	0.58	1.73
	0.99	0.45	1.68	0.46	1.67
2K	0.85	0.68	1.77	0.73	1.70
	0.95	0.60	1.69	0.61	1.69
	0.99	0.50	1.70	0.50	1.67

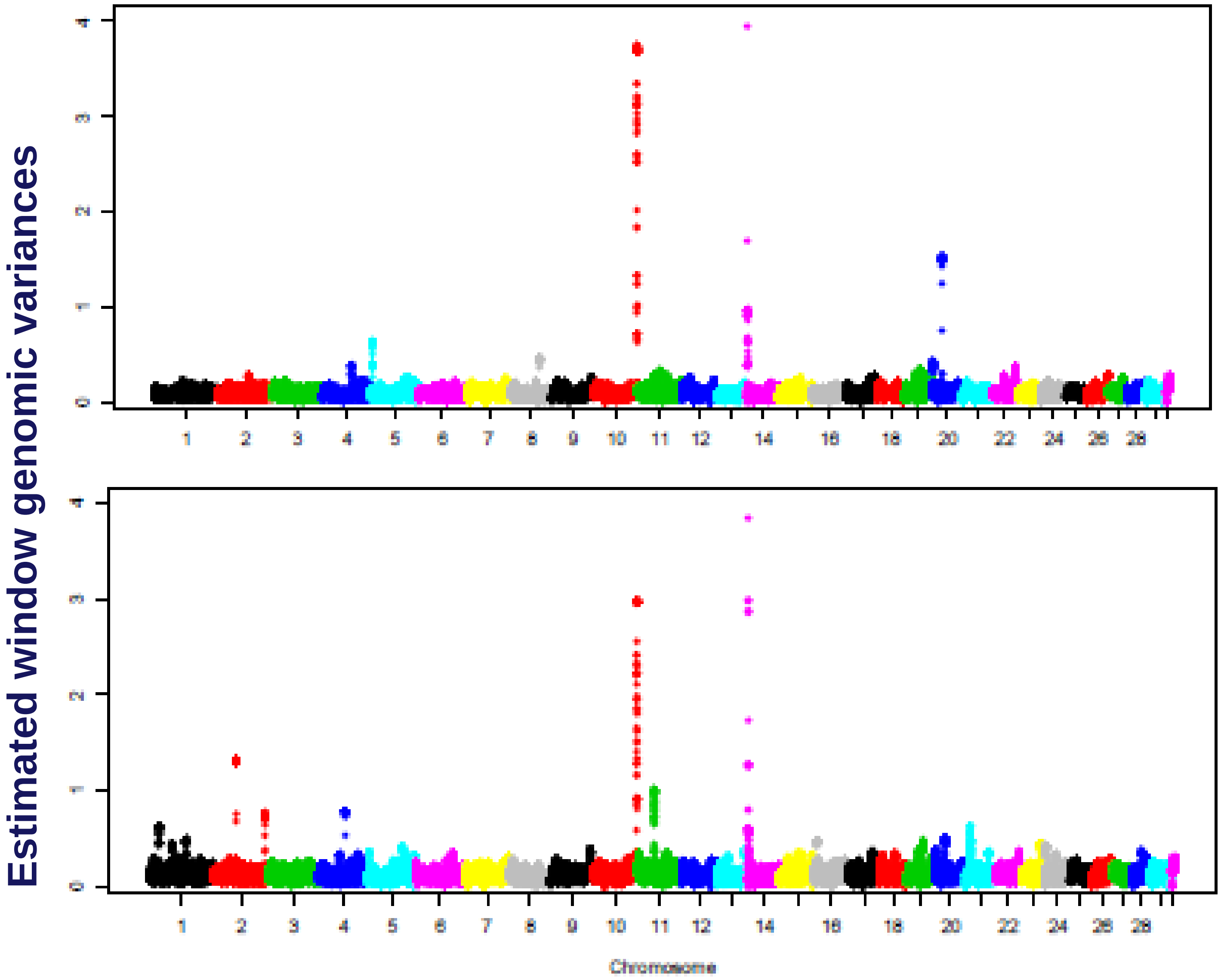
Application to a Fleckvieh cattle data set (Ertl et al. 2014)

- **1996 FV cows, genotyped with Illumina HD-SNP chip, ~630k SNPs.**
- **Milk fat yield, because Wellmann et al. (2014) showed increase in prediction accuracy for this trait with BayesD.**
- **Dominance is important for this trait in this data set (Ertl et al. 2014).**

Plots of WPPA, Results from FV cattle data set



Estimated window genomic variances: FV data set



Results from BayesC

Results from BayesD

Conclusions: Simulation & Bayes methods for GWAS

- **Simulation protocol: As realistic as possible (we hope so).**
- **Multi-marker GWAS by MCMC-based gs methods: Some nice properties. WPPA controls PFP (Fernando et al. 2014) & easy to calculate fro MCMC samples.**
- **Care must be taken when choosing the input parameter: pLD, df, window size, MCMC chain length, threshold q_w (needed for WPPA), …**
- **Single-marker GWAS: Straightforward implementation.**

Conclusions: Considering dominance by BayesD

- **Power increased and precision decreased with larger window sizes: trade off. Better definition of window boundaries needed (e.g. Beissinger et al. 2015)**

- **Considering dominance improves power (shift in power: -2 - 9 %),**

- **Shift in power due to the use of the additional genetic variance source (dominance variance) by diplotype marker information.**

- **A diplotype (matched haplotype pairs) breaks down fastly as distances increases: Improved precision was expected as well, …**

- **but only observed for low marker densities.**

Thanks for providing the Fleckvieh cattle data set to

Christian Edel (Institute of Animal Breeding, Bavarian State Research Center for Agriculture)

Ruedi Fries (Chair of Animal Breeding, Technical University of Munich)

The study was supported by a grant form the German Research Foundation, DFG

WPPA (Fernando et al. 2014)

- **C: Number of MCMC samples in which the window genetic variance exceeds a threshold q_w. WPPA = C / #samples.**
- **Choice of q_w is critical. Here: chosen under the assumption of an equal distribution of the genetic variance across the genome.**
- **WPPA of 0.85, 0.95 and 0.99 are used as thresholds, results in controlling proportion of false positive (PFP) of <0.15, <0.05 and <0.01 (see Fernando et al. 2014).**

Simulation protocol

- Fischer-Wright populations. 1 M & 1 chromosome genomes.
- N_e-pattern that is observed in cattle breeds (Villa-Angulo et al. 2009), fast decrease from 1000 to 100 within few generations. N=1500 in last gen.
- Expected number of mutations per individual: 4. Results in approx. 7K SNPs (with MAF > 0.01). ds)
- Scaling argument from gs theory (Meuwissen 2009):
  - 30M genome with N=45 000 and Ne=100 or
  - 30M genome with N=450 000 and Ne=1000 (across breeds)

Simulation protocol

- 3 marker panels based on distances and MAF generated: 2k, 1k and 0.5k.
- LD is a function of 4Ne*d (d is the distance between loci) -> allows to scale the simulated genomes towards different Ne.
- Corresponds to marker densities with same LD structure:
  - 60k, 30k and 15k in a 30M genome with Ne=100 or
  - 600k, 300k and 150k in a 30M genome with Ne=1000