A framework to incorporate knowledge on gene interaction into genomic relationship

H. Simianer [1] J.W.R.Martini [1] M. Erbe [2] V. Wimmer [3]

[1]Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, Goettingen, Germany

[2]Institute of Animal Breeding, Bavarian State Research Centre for Agriculture, Grub, Germany

[3]KWS SAAT SE, Einbeck, Germany

Common prediction approach: A linear marker effect model

Linear mixed model

$$y = \mu + M\beta + \epsilon,$$

where y denotes the vector of phenotypes, which is assumed to be a linear function of SNP effects $\beta_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_\beta)$ and ϵ a random error term ($\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$).

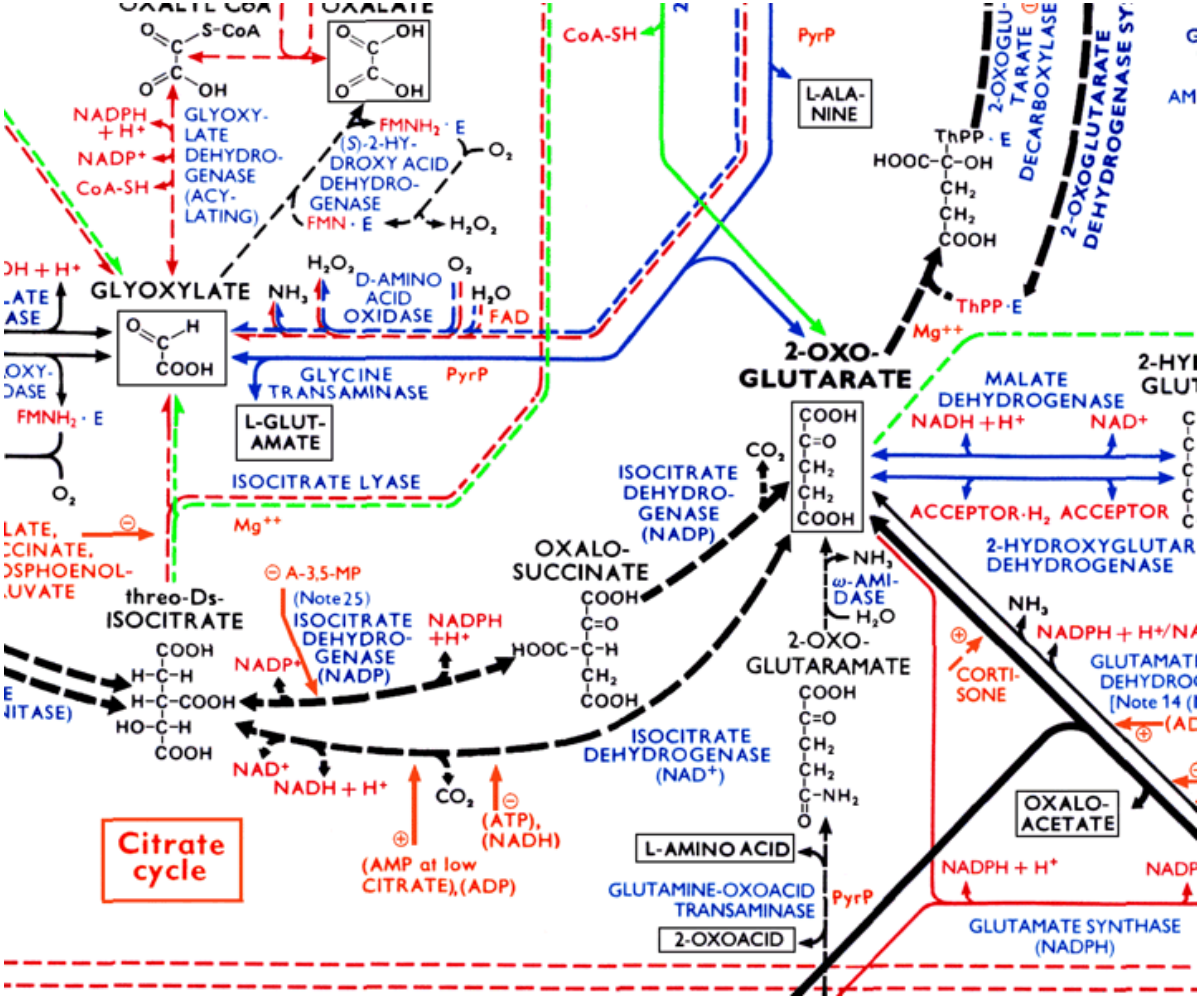
Common prediction approach: A linear marker effect model

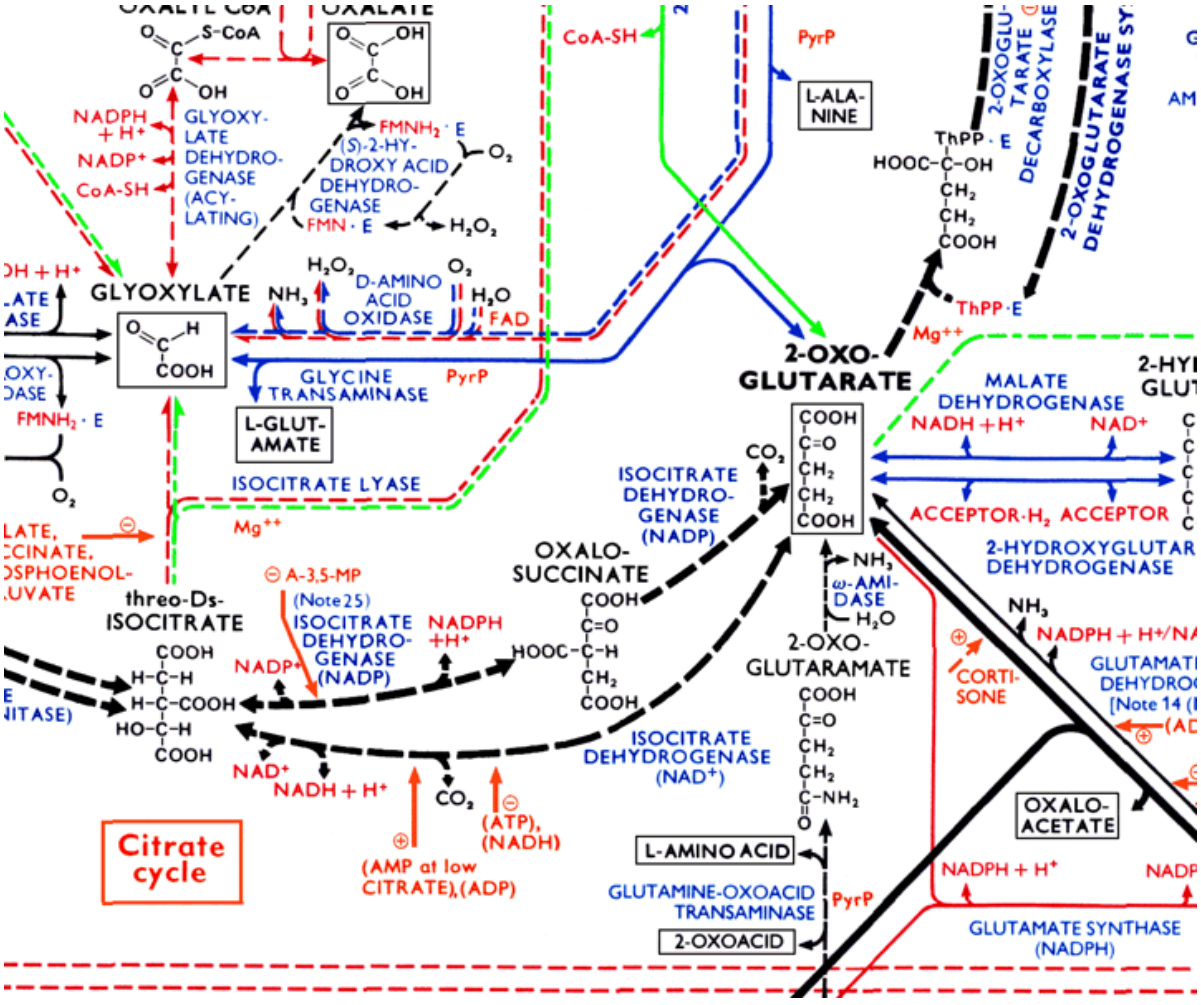
Linear mixed model

$$y = \mu + M\beta + \epsilon,$$

where y denotes the vector of phenotypes, which is assumed to be a linear function of SNP effects $\beta_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_\beta)$ and ϵ a random error term ($\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$).

- Limitation: only additive effects accounted for


Citrate cycle
GLYOXYLATE
threo-Ds-ISOCITRATE
2-OXO-GLUTARATE
OXALO-SUCCINATE
2-OXO-GLUTARAMATE
GLYOXYLATE DEHYDROGENASE (ACYLATING)
NADPH + H+
NADP+
CoA-SH
(S)-2-HYDROXY ACID DEHYDROGENASE
FMN · E
FMNH2 · E
D-AMINO ACID OXIDASE
FAD
GLYCINE TRANSAMINASE
PyrP
L-GLUTAMATE
ISOCITRATE LYASE
Mg++
A-3,5-MP
(Note 25)
ISOCITRATE DEHYDROGENASE (NADP)
ISOCITRATE DEHYDROGENASE (NAD+)
NAD+
NADH + H+
(AMP at low CITRATE),(ADP)
(ATP), (NADH)
L-ALANINE
2-OXOGLUTARATE DECARBOXYLASE
2-OXOGLUTARATE DEHYDROGENASE
ThPP · E
MALATE DEHYDROGENASE
ACCEPTOR·H2
ACCEPTOR
2-HYDROXYGLUTARATE DEHYDROGENASE
ω-AMIDASE
CORTISONE
GLUTAMATE DEHYDROGENASE
OXALOACETATE
L-AMINO ACID
GLUTAMINE-OXOACID TRANSAMINASE
2-OXOACID
GLUTAMATE SYNTHASE (NADPH)



How to incorporate interactions in the model?

The epistasis model

A generalization to a model with polynomials of degree two:

$$y_i = \mu + M_{i\bullet}\beta + \sum_{k=1,\ldots,p;l=k+1,\ldots,p} M_{ik}M_{il}h_{kl} + \epsilon_i$$

two-locus epistasis model

The epistasis model

A generalization to a model with polynomials of degree two:

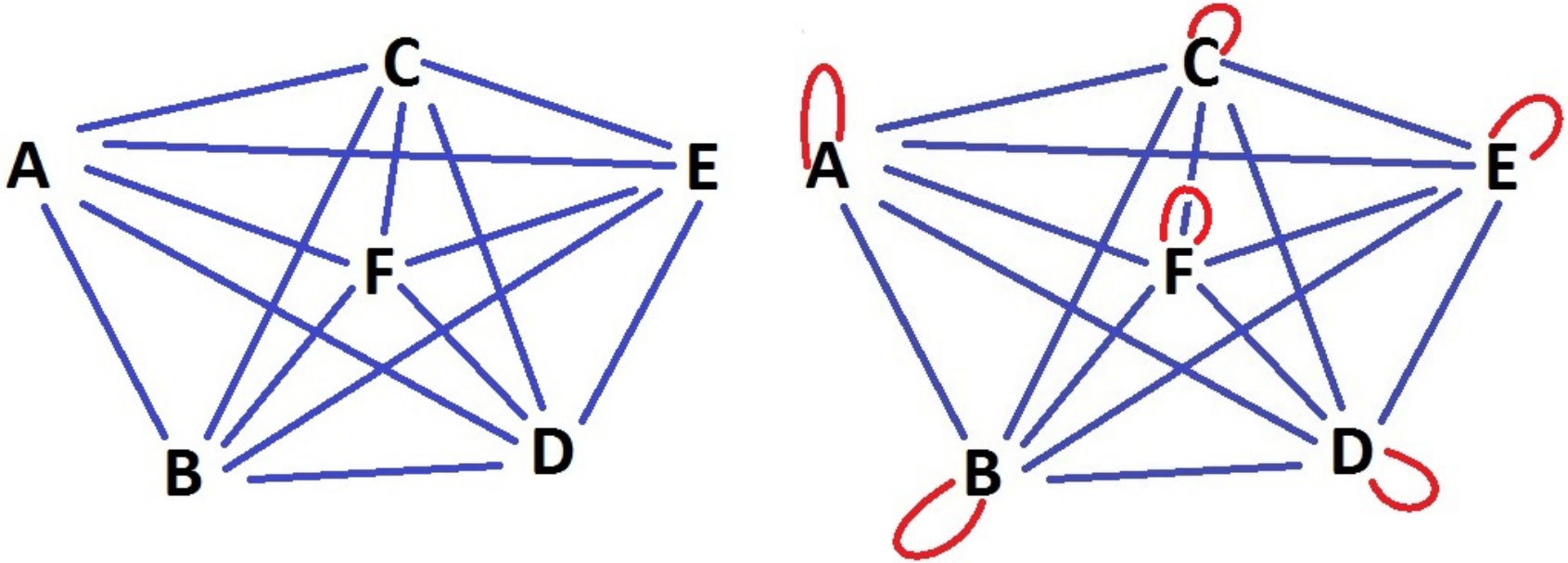
$$y_i = \mu + M_{i\bullet}\beta + \sum_{k=1,\ldots,p;l=k+1,\ldots,p} M_{ik}M_{il}h_{kl} + \epsilon_i$$

two-locus epistasis model or

$$y_i = \mu + M_{i\bullet}\beta + \sum_{k,l=1,\ldots,p} M_{ik}M_{il}h_{kl} + \epsilon_i$$

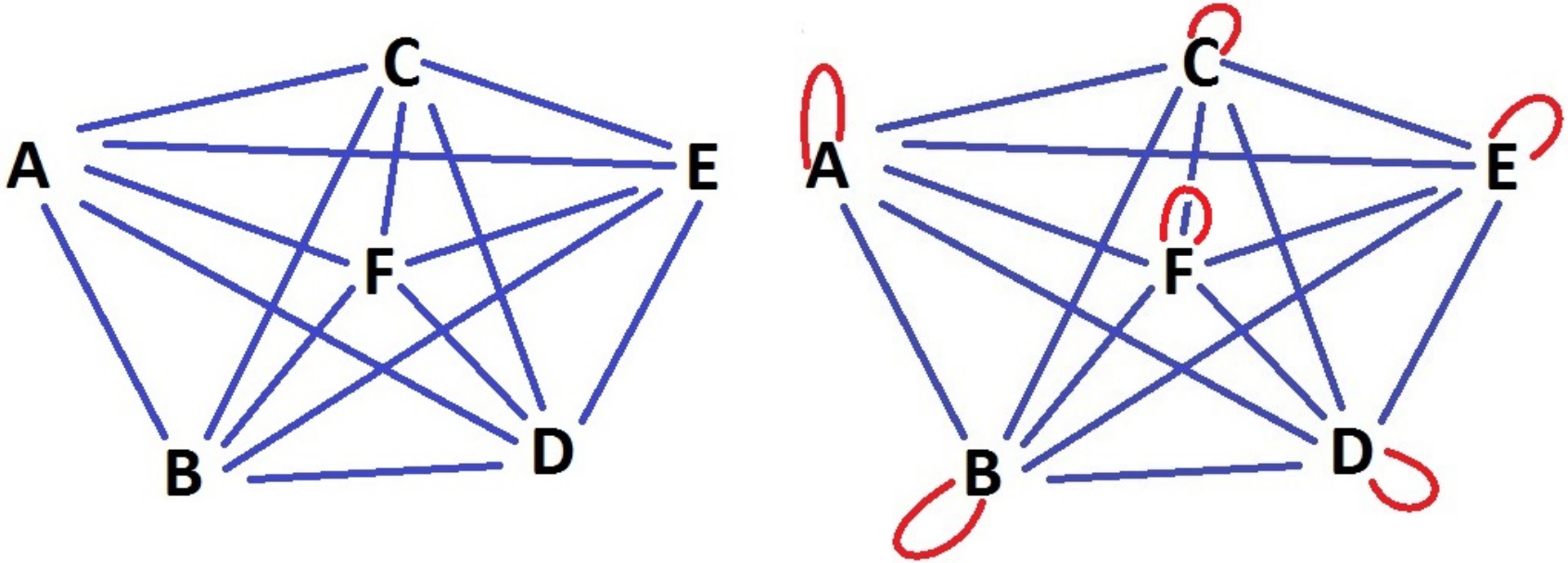
dominance + *two-locus* epistasis model with $h_{kl} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_h^2)$.

The epistasis model



- Obvious problem: Parametric inflation of the model.

The epistasis model



- Obvious problem: Parametric inflation of the model.
- Do we need to estimate all effects to use the model for genomic prediction?

Equivalence of models

Well-known:

$$y = \mu + M\beta + \epsilon \;\hat{=}\; y = \mu + g + \epsilon$$

with $g \sim \mathcal{N}_n(0, \sigma^2_\beta MM')$.

Equivalence of models

Well-known:

$$y = \mu + M\beta + \epsilon \;\hat{=}\; y = \mu + g + \epsilon$$

with $g \sim \mathcal{N}_n(0, \sigma^2_\beta MM')$.

- GBLUP model has advantages since the number of markers p usually exceeds the number of genotyped individuals.

Equivalence of models

Well-known:

$$y = \mu + M\beta + \epsilon \;\hat{=}\; y = \mu + g + \epsilon$$

with $g \sim \mathcal{N}_n(0, \sigma^2_\beta MM')$.

- GBLUP model has advantages since the number of markers p usually exceeds the number of genotyped individuals.
- Is there an analogous equivalence for the epistasis models?

Equivalence of models

For the epistatic models, we have

$$y_i = \mu + M_{i\bullet}\beta + \sum_{k,l=1,\ldots,p} M_{ik} M_{il} h_{kl} + \epsilon_i \;\hat{=}\; y = \mu + g + a + \epsilon$$

with $a \sim \mathcal{N}_n(0, \sigma_h^2 MM' \circ MM')$ for the model incorporating dominance or

Equivalence of models

For the epistatic models, we have

$$y_i = \mu + M_{i\bullet}\beta + \sum_{k,l=1,\ldots,p} M_{ik}M_{il}h_{kl} + \epsilon_i \;\hat{=}\; y = \mu + g + a + \epsilon$$

with $a \sim \mathcal{N}_n(0, \sigma_h^2 MM' \circ MM')$ for the model incorporating dominance or

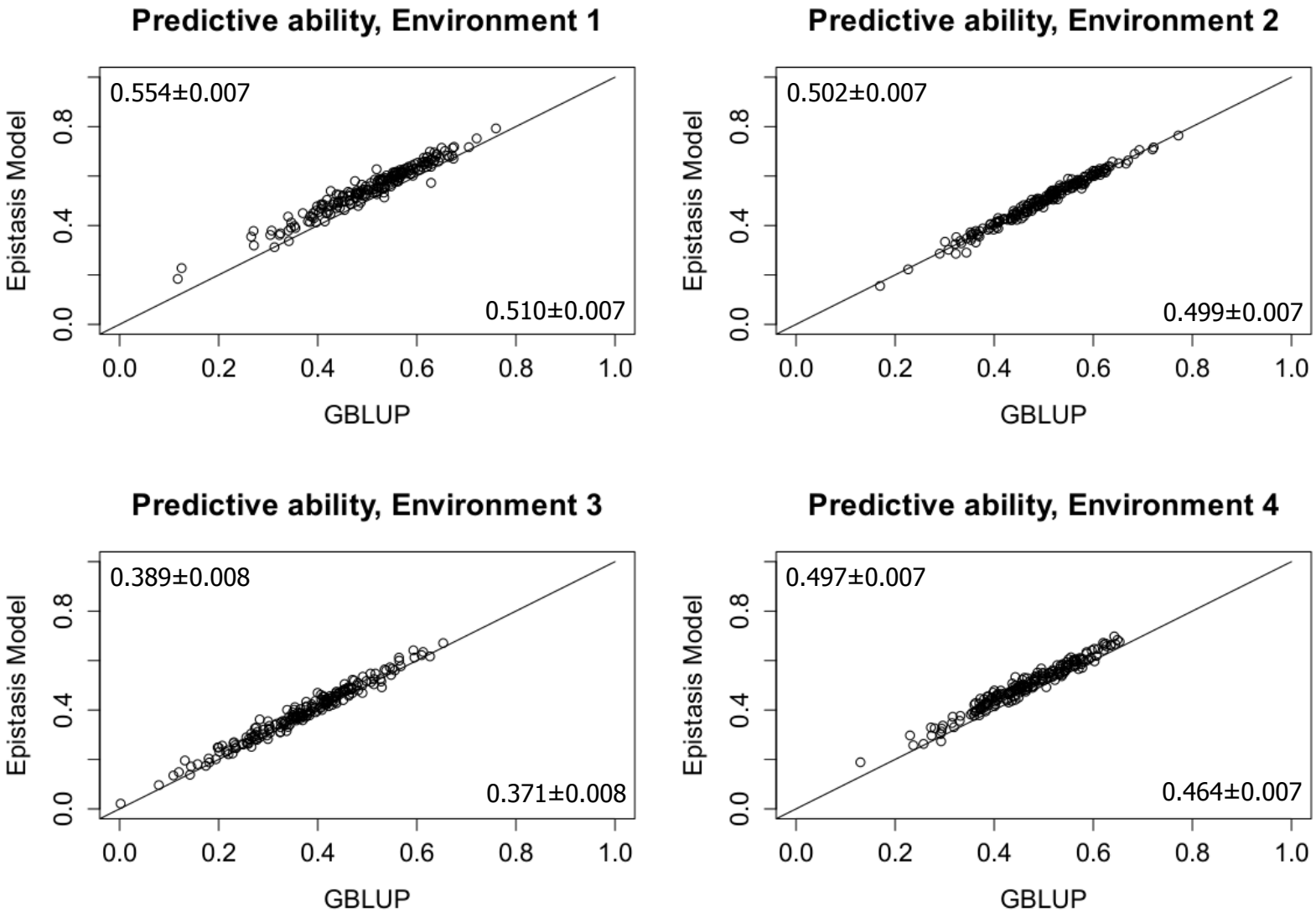
$$a \sim \mathcal{N}_n(0, \sigma_h^2 \left((MM' \circ MM') - (M \circ M)(M \circ M)'\right))$$

for the two-locus model, where $\circ$ represents the Hadamard product.

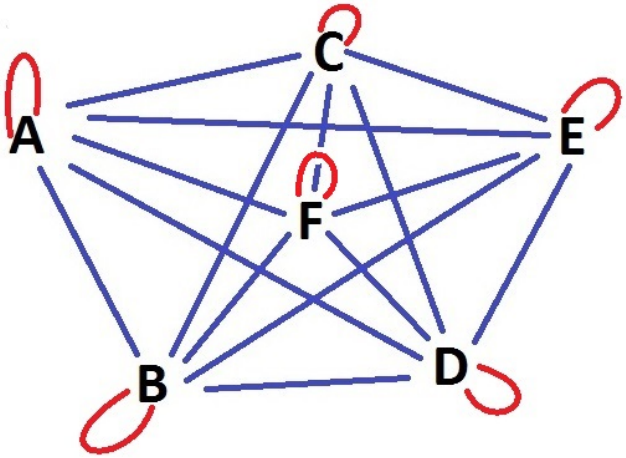
Test on a wheat dataset (Crossa et al., 2010)

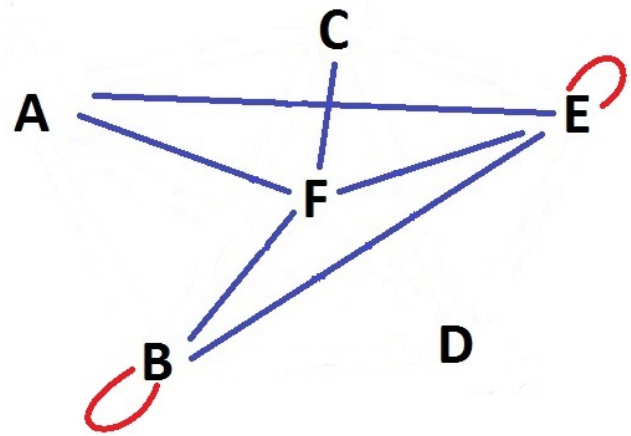
(599 lines, 1279 DArT markers encoded by $\{0, 1\}$, four different environments, trait: grain yield)

GBLUP vs. Epistasis model $y = \mu + a + \epsilon$ with $a \sim \mathcal{N}_n(0, \sigma_h^2 MM' \circ MM')$



A relationhip matrix based on a subset of interactions



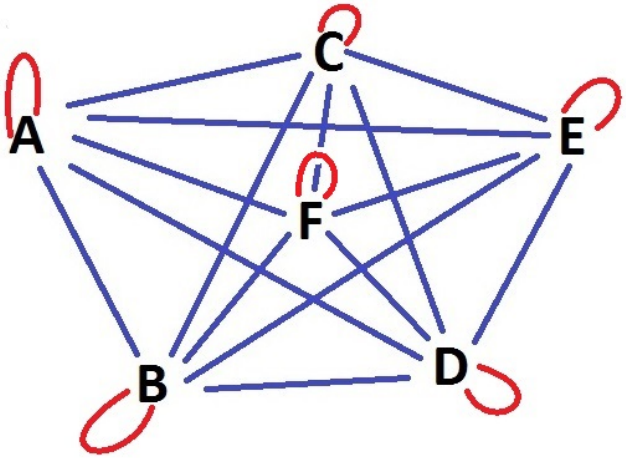


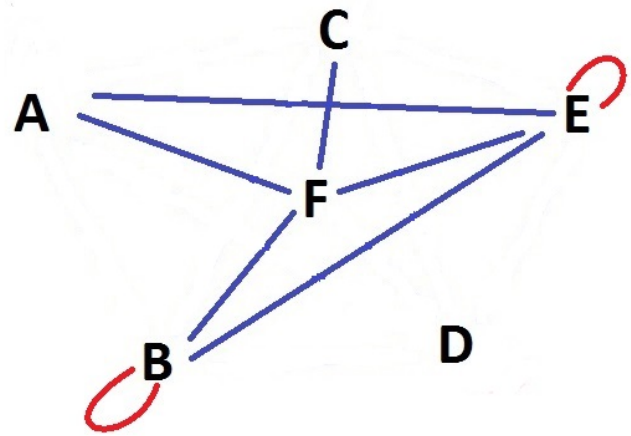
The interaction of marker k and j translates into the matrix

$$\left(M_{\bullet k} M'_{\bullet k}\right) \circ \left(M_{\bullet j} M'_{\bullet j}\right) \sigma_h^2$$

where $M_{\bullet k}$ denotes the kth column of M reflecting the marker realizations of all genotypes at position k.

A relationhip matrix based on a subset of interactions





The interaction of marker k and j translates into the matrix

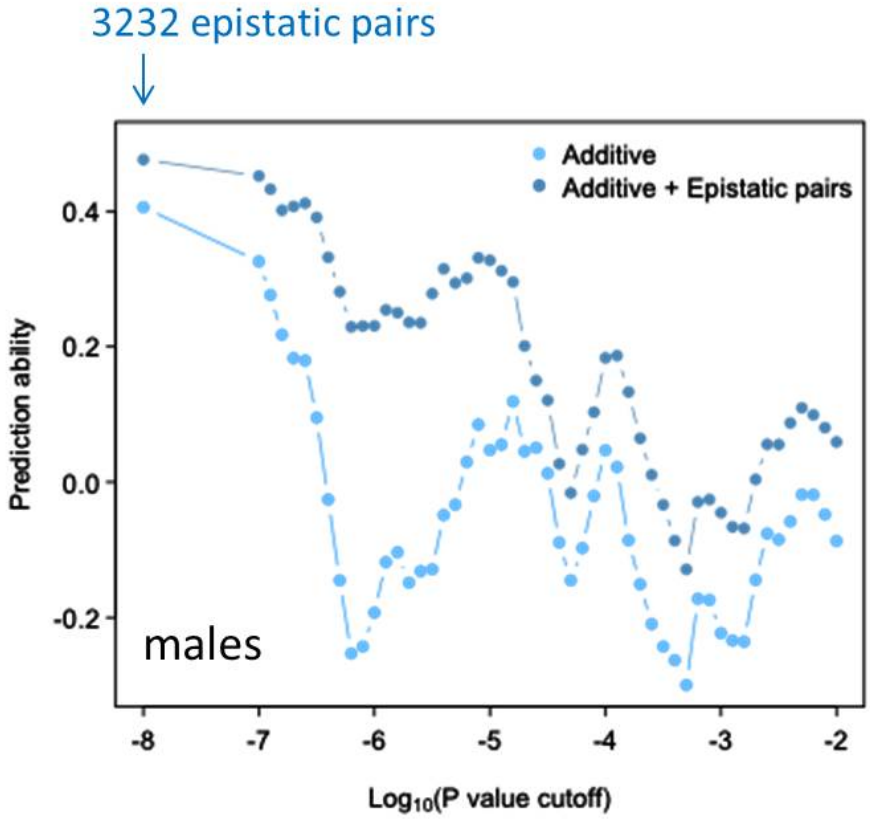
$$\left(M_{\bullet k} M'_{\bullet k}\right) \circ \left(M_{\bullet j} M'_{\bullet j}\right) \sigma_h^2$$

where $M_{\bullet k}$ denotes the kth column of M reflecting the marker realizations of all genotypes at position k.

- Trait-specific, network based relationship matrices can be built by summing over the matrices of the individual interactions.

Adding selected epistatic interactions to an additive model can substantially improve predictive ability of genomic prediction

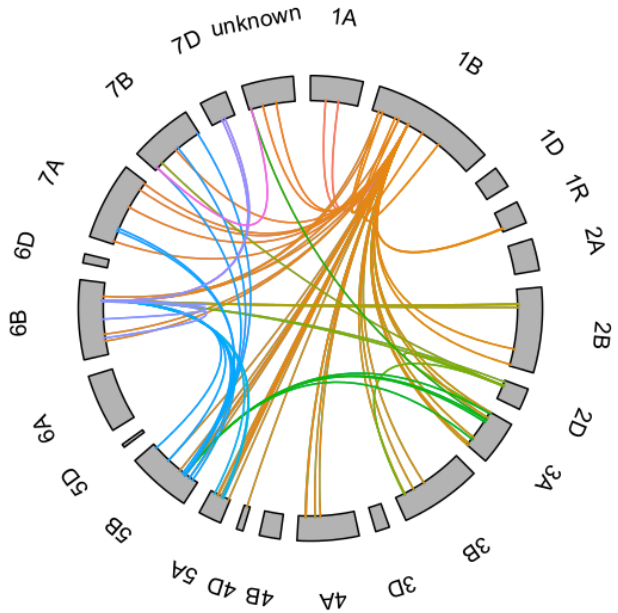
Adding selected epistatic interactions to an additive model can substantially improve predictive ability of genomic prediction

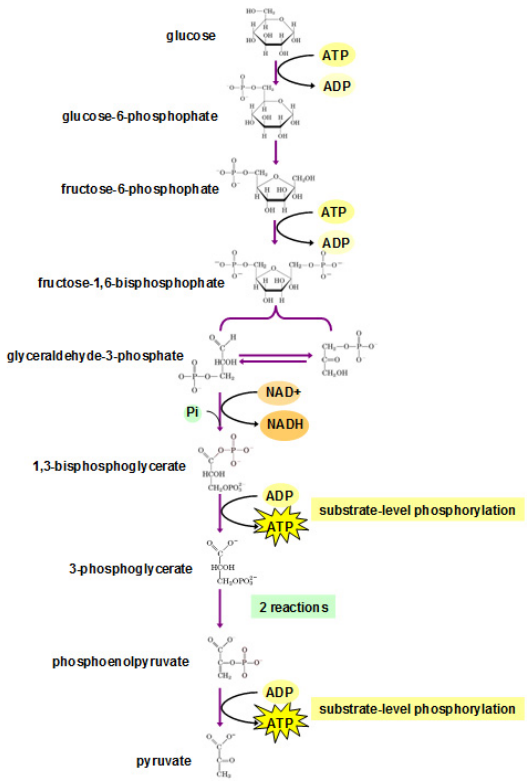


Example: Chill coma recovery in the Drosophila Genome Reference Panel (DGRP) (Ober et al., 2015).

Outlook

- Goal: Translating available informations on gene interaction and biochemical pathways into interaction within the epistasis model.





Acknowledgements

KWS SAAT SE is gratefully acknowledged for funding this research

Johannes W.R. Martini

References

Crossa, J., G. de Los Campos, P. Pérez, D. Gianola, J. Burgueño, J. L. Araus, D. Makumbi, R. P. Singh, S. Dreisigacker, J. Yan, et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics 186*(2), 713–724.

Martini, J. W. R., V. Wimmer, M. Erbe, and H. Simianer (2015). Epistasis and covariance: How gene interaction translates into genomic relationship. *Theoretical and Applied Genetics*, submitted.

Ober, U., W. Huang, M. Magwire, M. Schlather, H. Simianer, and T. F. C. Mackay (2015). Accounting for genetic architecture improves sequence based genomic prediction for a drosophila fitness trait. *PLOS ONE* (10), e0126880.