Genomic Selection: Methodologies and procedures

Mario Calus

Animal Sciences Group – Wageningen UR, The Netherlands

Animal Breeding & Genomics Centre

ANIMAL SCIENCES GROUP
WAGENINGEN UR

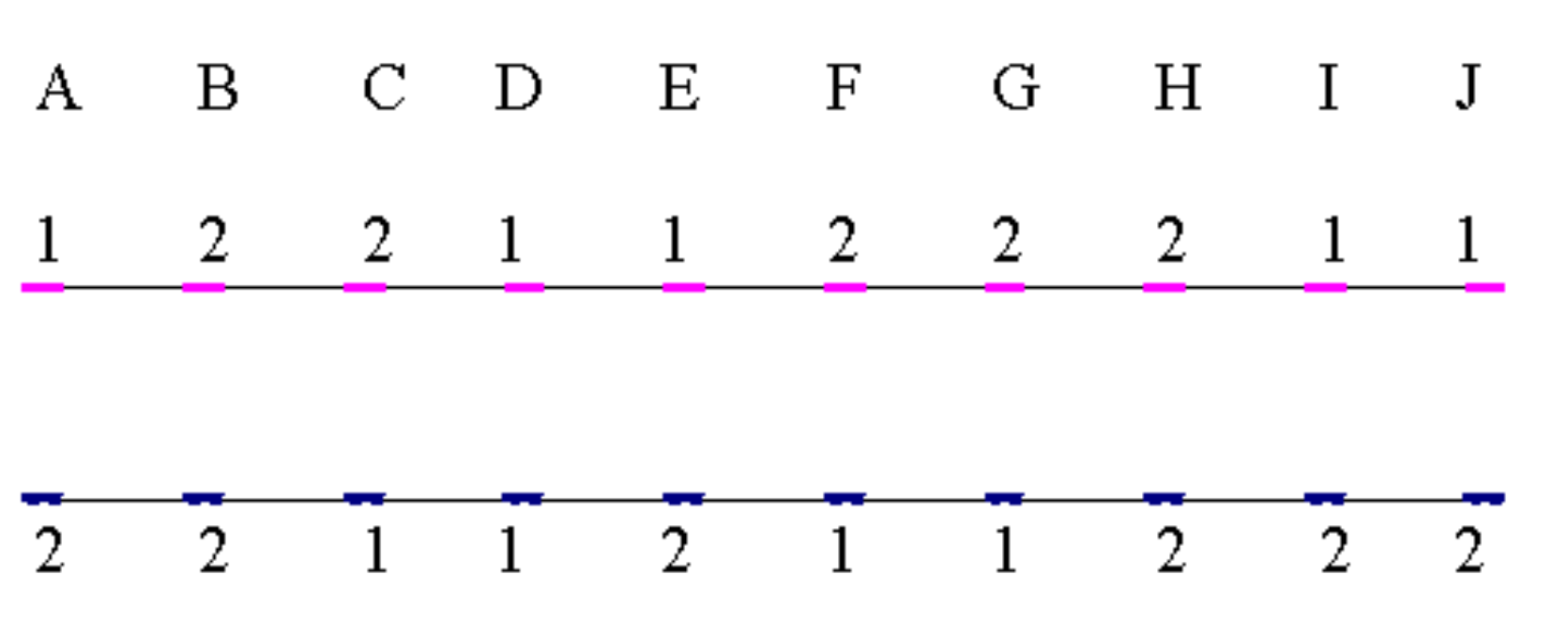
Animal Breeding & Genomics Centre

Objective of this presentation

- Principle of Genomic Selection (GS)
- Process of applying GS in a breeding program
- Estimation of Genomic Breeding Values (GEBVs)
- Accuracies of GEBVs

Introduction – Genomic Selection

- *Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001.*
- Genome of animal X (Markers A,B,..,J, possibly associated with QTL):



- Total breeding value animal X = A1 + A2 + B2 + B2 + … + J1 + J2

Genomic Selection – the process

Reference dataset:

1000+ animals with known
genotypes (SNPs) and reliable EBVs

Obtain EBVs for SNPs

Accurate EBVs young selection candidates

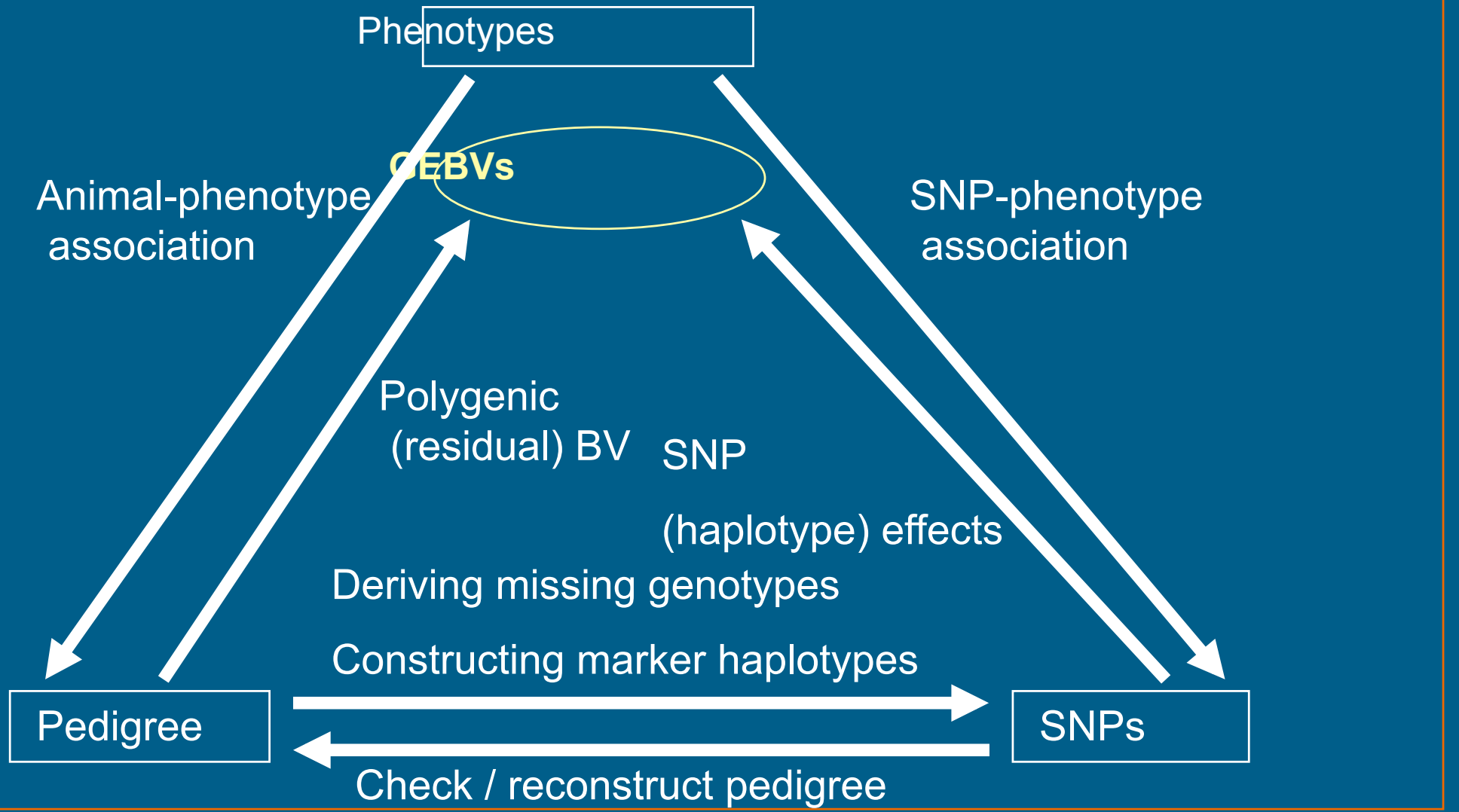
Young selection candidates with known genotypes (SNPs)
but WITHOUT performance records

Estimation of genomic breeding values (GEBVs)

- How to link different sources of data? (parameterization of the model)
- How to solve the model?

=> Application of GS in animal breeding is a 'number-crunching' issue

GEBVs: Sources of data (1000+ animals)



General model

$y_i = \mu + animal_i + sum(SNP_{ijk}) + e_i$

- y_i may be phenotypes, national EBVs, DYD's, etc.
- $animal_i$ is polygenic effect
- $sum(SNP_{ijk})$ is sum of SNP effects, summed across all loci

- 1000+ animals & 50,000 SNPs

Problem: #SNP effects >>> #phenotypes

=> How to solve the model?

Dealing with #SNP effects >>> #phenotypes

BLUP (Meuwissen et al. 2001):

- Assume equal contributions of SNPs (genes) to the genetic variance across the genome

- However, distribution of gene effects implies (Hayes et al. 2001):
  - many loci of small (near zero) effect
  - few loci with large effect

Þ How can we eliminate loci with (near) zero effect?

Model distribution of gene effects more closely

- Select reduced set of explaining loci
  - Tag-SNPs: select SNP based on mutual LD

- Select only loci with effect on trait

Before the analysis:

  - Implicitly considering SNP-phenotype associations (Long et al., 2007)

In the model:

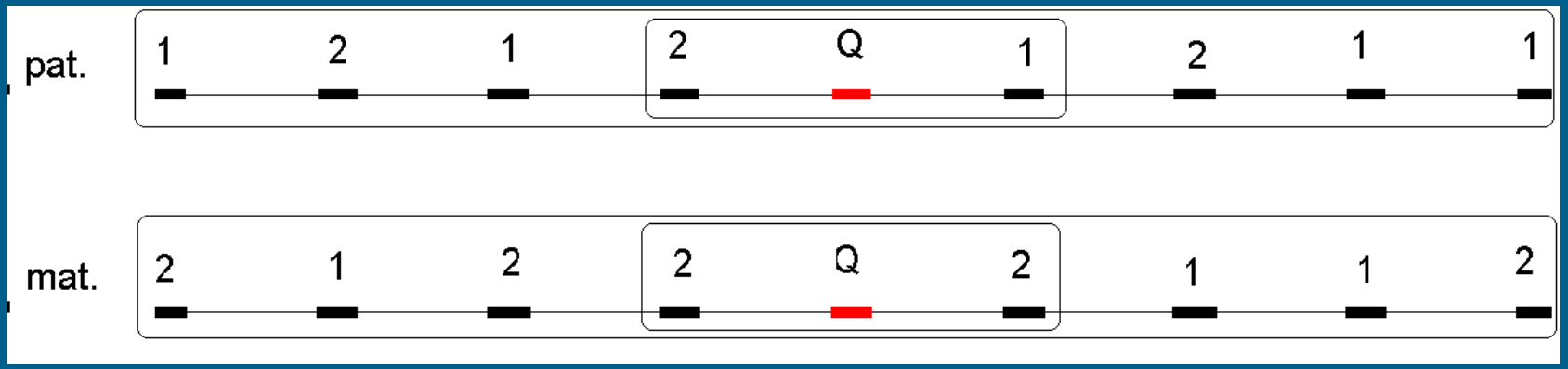
- BayesB (Meuwissen et al. 2001):
  - Association of loci to phenotype (0 / 1) is sampled in model
- Gibbs sampling (derived from BayesB; Meuwissen et al., 2004; Calus et al., 2008):
  - Similar to BayesB, but avoids Metropolis-Hastings step

Alternative models

- Regression with forward / backward elimination (Habier et al., 2007)
- Kernel regression techniques (Gianola et al., 2006)
- Principal component analysis (PCA), Partial least squares (PLS), etc. (Solberg et al., 2008; Moser et al., 2008)

Parameterization of the model

=> Linking SNPs to (putative) QTL alleles

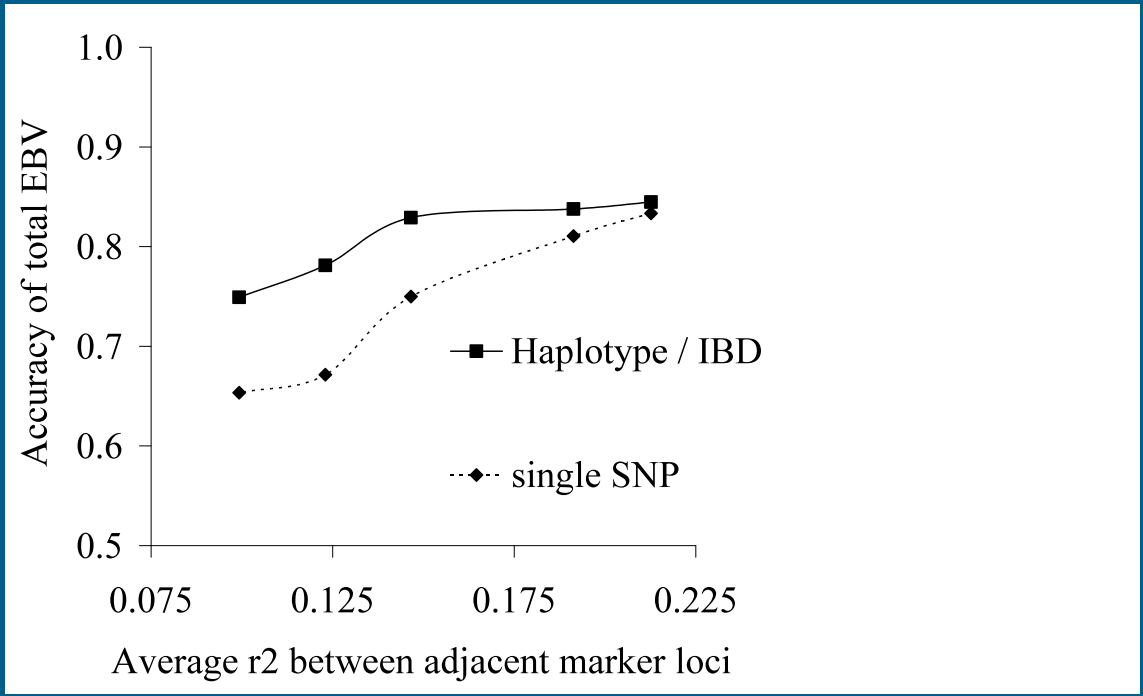


Parameterizations differ by:

- Definition of SNP effects:
  - 1 or more marker alleles combined to haplotypes
- Assumed relation between haplotypes:
  - 0 / 1; the same or not (linkage disequilibrium; LD)
  - Continuous scale: 0 – 1; based on identity-by-descent (IBD; combined LD & linkage analysis)

Accuracy using SNP alleles / haplotypes

- Haplotypes / IBD have higher accuracy at low marker density



[1]*Calus M.P.L., Meuwissen T.H.E., De Roos A.P.W., Veerkamp R.F., Accuracy of genomic selection using different methods to define haplotypes, Genetics 178 (2008) 553–561.*

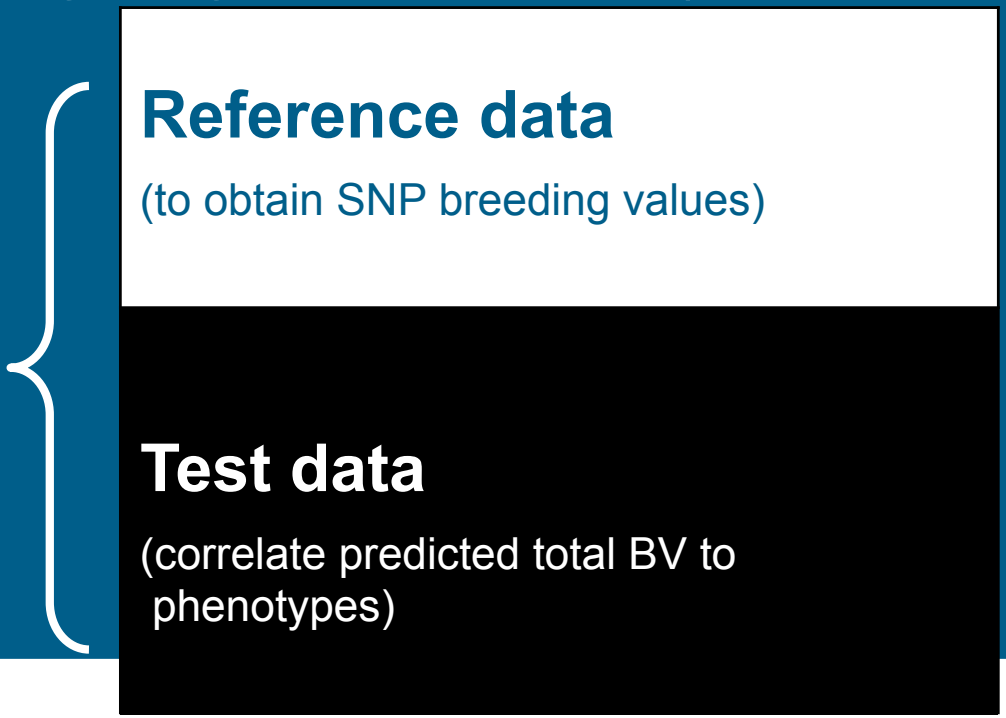
Accuracy (r) of GEBVs

Accuracies can be predicted by:

- Simulation study
  - How close is the simulated data to real data?

- Cross-validation (e.g. Legarra et al. 2007):

Full data (genotyped / phenotyped)



Accuracy (r) of GEBVs

Accuracy of GEBVs depends on (Goddard, 2007):

- Number and size of QTL
- Accuracy of estimated (QTL) effects; size reference data:
  - Number of animals (i.e. phenotypes)
  - Number of markers (LD (r^2) between QTL and marker)
- Reference data may increase in time:
  - Number of animals increases (accuracy GEBVs ↑)
  - LD between QTL and markers may change (accuracy GEBVs ↓)

=> In time GEBVs need to be re-estimated, but how often??

Frequency re-estimation GEBVs

Frequency of re-estimating SNP breeding values:

- What is the desired frequency from the perspective of the breeding program?
  - Re-estimation is possible when phenotypes of GS -selected animals can be added to reference data

  => Time to obtain phenotypes determines time frame for re -estimation

- What frequency is required to ensure accurate selection?
  - Depends on break-down LD between SNP and QTL

Breakdown of LD between SNP and QTL

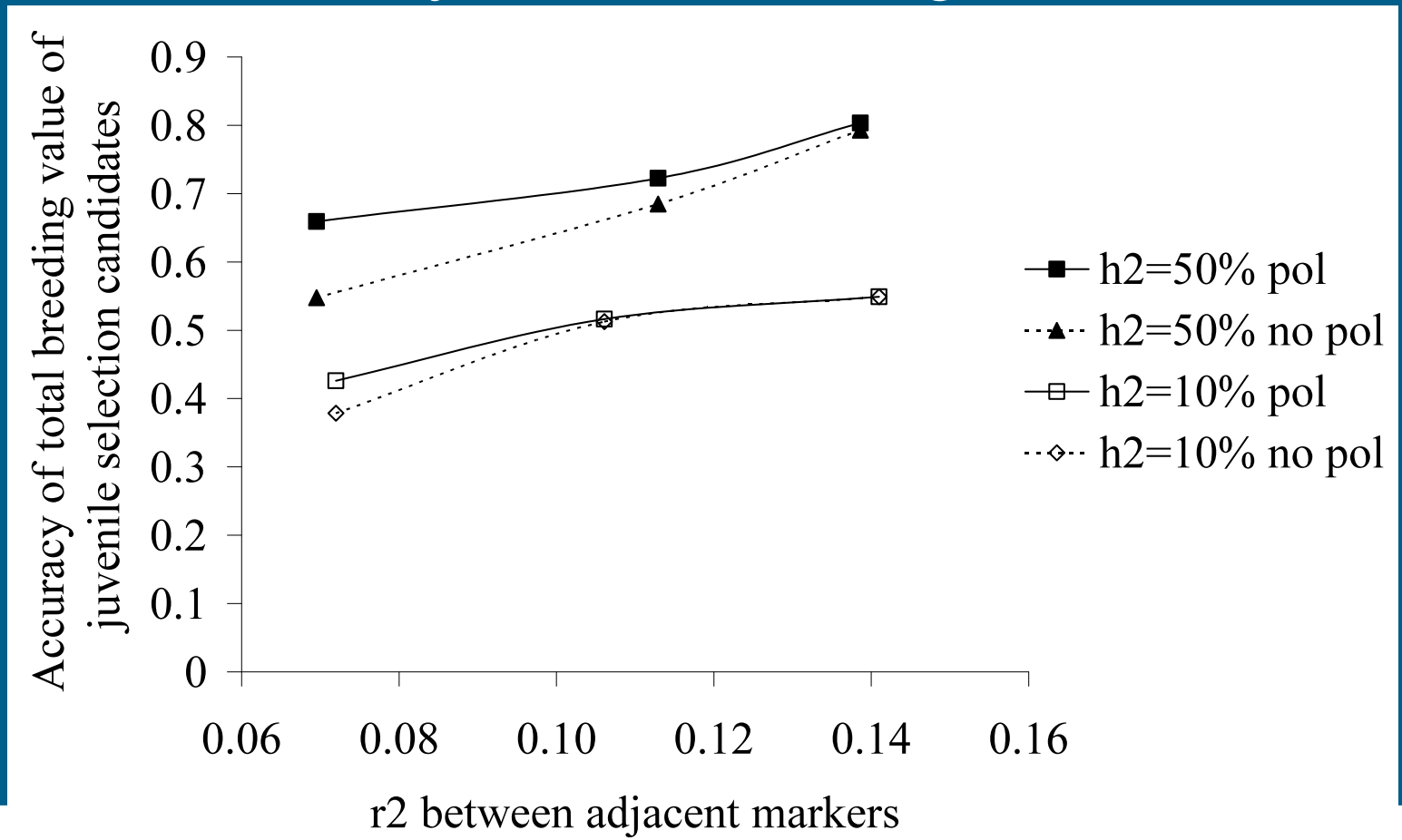
- LD between loci can be changed by selection
  - Due to change in allele frequencies
  - Accuracy of GS ↓
- Reported results (from simulation):
  - Slow decrease when mating is random (Meuwissen et al., 2001; Solberg et al., 2008)
  - Rapid decrease under selection (Habier et al., 2008; Muir, 2008)

Effect on accuracy forward prediction

- Accuracy forward prediction (across generations) using:
  - SNPs
  - polygenic effects
- Habier et al., (2008): SNPs may ‘absorb’ genetic (pedigree) relationship
- Likely depends on:
  - Association SNP-phenotype (LD-based or spurious)
  - Number of generations in reference data

Including polygenic BVs in the model

- Calus & Veerkamp (2008): Higher accuracy at low marker density, no effect at high marker density



Future perspectives

Are more markers needed (i.e. higher marker-QTL LD), depending on the objective?

- Increasing accuracy of GS:
  - More phenotypes may have a greater impact (Meuwissen et al., 2001)
- Within or across breed GS:
  - In cattle, 50k SNPs sufficient within a breed; ~300k required across breeds (De Roos et al., 2008)
- When fine-mapping is an additional goal?

Future perspectives

- Use of low density SNPs to ‘pre-screen’ populations (Habier et al., 2008)
- Parents genotyped using high density SNPs
- Combine low & high density, to ‘derive’ high density genotypes for selection candidates

Conclusion

- Reference data is key in application of GS
- Obtaining of GEBVs is challenging
- Existence and breakdown of LD between SNP and QTL are crucial issues
- Available marker density may be sufficient within breeds, not across breeds

Acknowledgements

- Involved companies:

Hendrix Genetics, CRV (HG)

- Netherlands Organisation for Scientific Research (NWO – Casimir)