Genomic Breeding Value Prediction
Lessons for and from simulations

Mario Calus

Workshop Genomic Selection, 7-8 September, Salzburg

Animal Breeding & Genomics Centre

Objectives of this presentation

- Assumptions of GS
- Reflection on simulations for GS
- Distribution of gene (QTL) effects
- Accuracy of GS / Breakdown of LD
- Implications for the analysis

GS versus QTL mapping

- GS uses effectively a ‘multiple QTL-model’
- How can GS work where QTL mapping failed?

Differences between GS and QTL mapping I

QTL mapping:

- Significant effect for an evaluated locus is required
- Estimate QTL effect may be biased, because only 1 QTL is fitted at the time

Genomic selection:

- All effects are estimated simultaneously
- If some SNP effects are overestimated, others must be underestimated (since y_i=sum(SNP))
- On average (across SNPs), bias may be limited

Differences between GS and QTL mapping II

- GS heavily depends on:
  - LD between marker-QTL, persistent across population
  - Dense marker maps

- Many QTL mapping studies sofar used:
  - Linkage analysis
  - Sparse marker maps

=> Implication for simulations for GS: generation of LD is important (i.e. r^2 between adjacent markers)

Simulations for GS - Introduction

Daniel Gianola's opinion about simulations:

- 'In short, they are like reading Playboy magazine: "what if" (the problem is the if...)'

Despite this, simulations are:

- Cheap to test:
  - Accuracy of GS
  - Accuracy of QTL mapping methods to detect and position QTL
- Useful to check models:
  - Technically
  - (Derivation from) model assumptions / sensitivity analysis

Still, it is important that simulated data reflect real life

Genomic Selection – the process

Reference dataset:

1000+ animals with known
genotypes (SNPs) and reliable phenotypes (e.g. EBVs)

↓

Obtain EBVs for SNPs

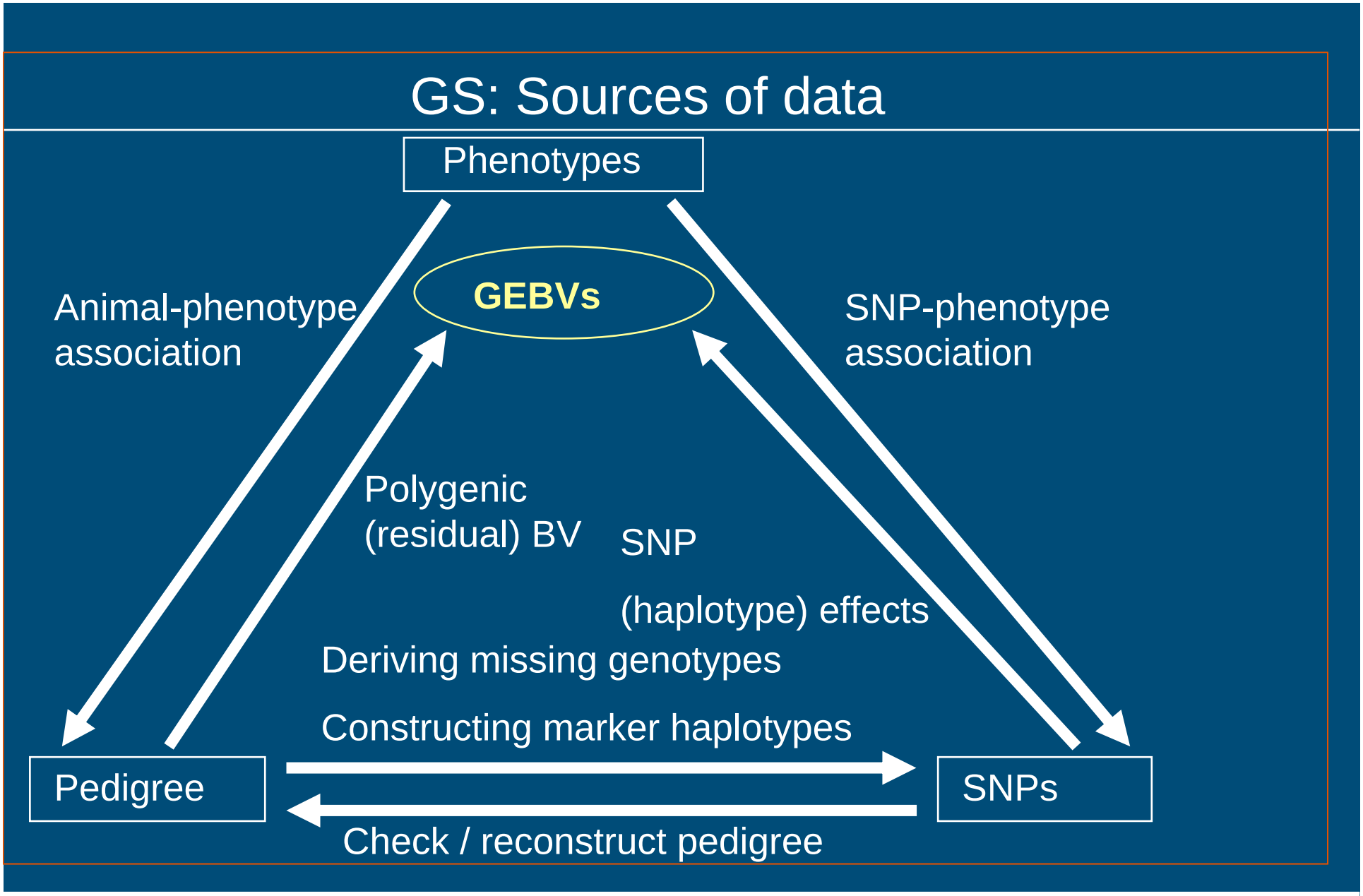
↓

Accurate EBVs young selection candidates

↑

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

GS: Sources of data



Simulation of marker (and QTL) data

LD between loci:

- Simulate coalescence (gene drop) process across many generations
  - pedigree evolves simultaneously
- Sample generation of animals with segregating loci directly from (known) distribution
  - no pedigree directly available
  - Pedigree can be generated by (random) mating for some generations

Simulation of pedigree

Important issues:

- Mating
  - Random or selection?
- Effective population size (Ne)
  - Constant across generations?
  - Strongly affects genetic drift / LD

Simulation of LD

Coalescence:

- Simulate 100+ generations:
  - Monomorphic or segregation loci in generation 0
  - Mutations throughout generations

⇒ LD due to drift, selection, migration,…

Directly from distribution:

- Draw alleles at first locus, using some distr. of allele frequencies
- Draw r (r2) between alleles on two loci
- Draw alleles at second locus, conditional on r2 and alleles at first locus

How to avoid these issues?

Use real data with known genotypes & pedigree:

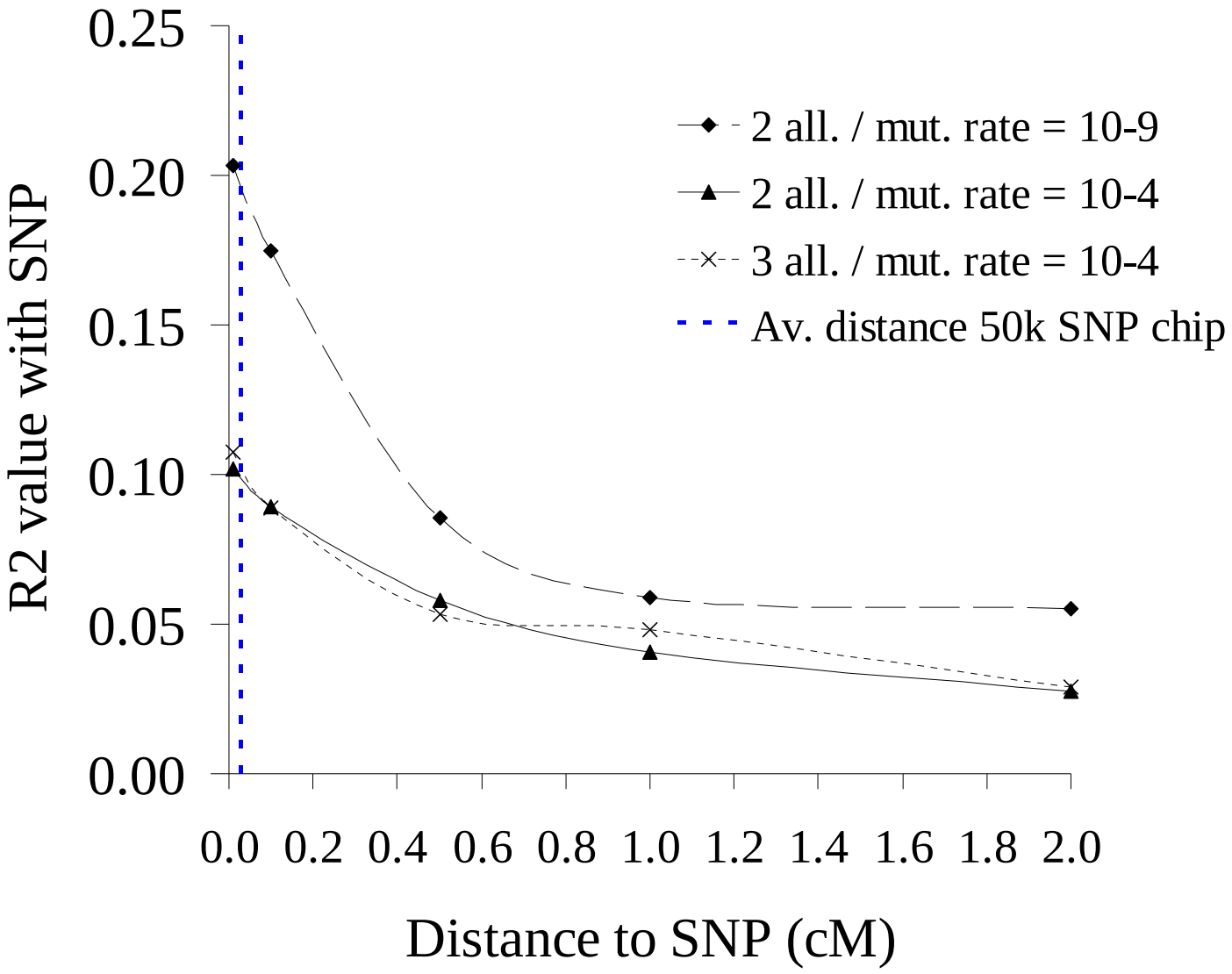
- Draw some marker loci to be QTL
- Simulate QTL effect for those loci
- Remove ‘QTL’ loci from marker data

Still, the following assumptions are made:

- QTL have the same characteristics as SNP
  - Mutation rate / number of alleles / LD with surrounding SNP
- Distribution of QTL effect is known

Characteristics of QTL – LD with SNP

- Effect of mutation rate on LD (Calus, De Koning & Haley, unp. data):



Distribution of QTL effect

- Important for analysis:
  - Which model to use?
  - Prior information in Bayesian
- Only a few QTN are detected until now (perhaps only a few really exist?)
- Simulating QTL effects from Gamma (or normal) distribution may be too optimistic?

=> Make sure number of large QTL is not too big

Implications from analysis of real data

- Results on real data indicate that sampling variance SNP effects from one distribution may be sufficient (e.g. Janss et al., 2008;):
  - Roughly equal contributed variance for all SNPs
  - Close to ‘BLUP’ implementation Meuwissen et al., (2001)
- What does this tell us about the distribution of SNP (QTL) effects?
  - SNP effects are roughly equal
  - What about the (true) QTL effects?
  - What is the relation between estimated SNP effects and real QTL effects?

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 of re-estimating SNP breeding values

Select young animals based on GEBV

↓

Use in the population

↓

Record own phenotypes and / or from relatives

Replace GEBV by EBV

=> 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)

Lessons from analyzing simulated data:

Parametrization of the model

- *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*.

- Aim of this study:

Compare effect of definition of haplotypes (based on 1 or more markers) and the relationships between haplotypes at the same locus, on accuracy of GEBVs

General model

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

- animal is polygenic effect
- $sum(haplotype_{ijk})$ is sum of paternal and maternal haplotype effects, summed across all loci

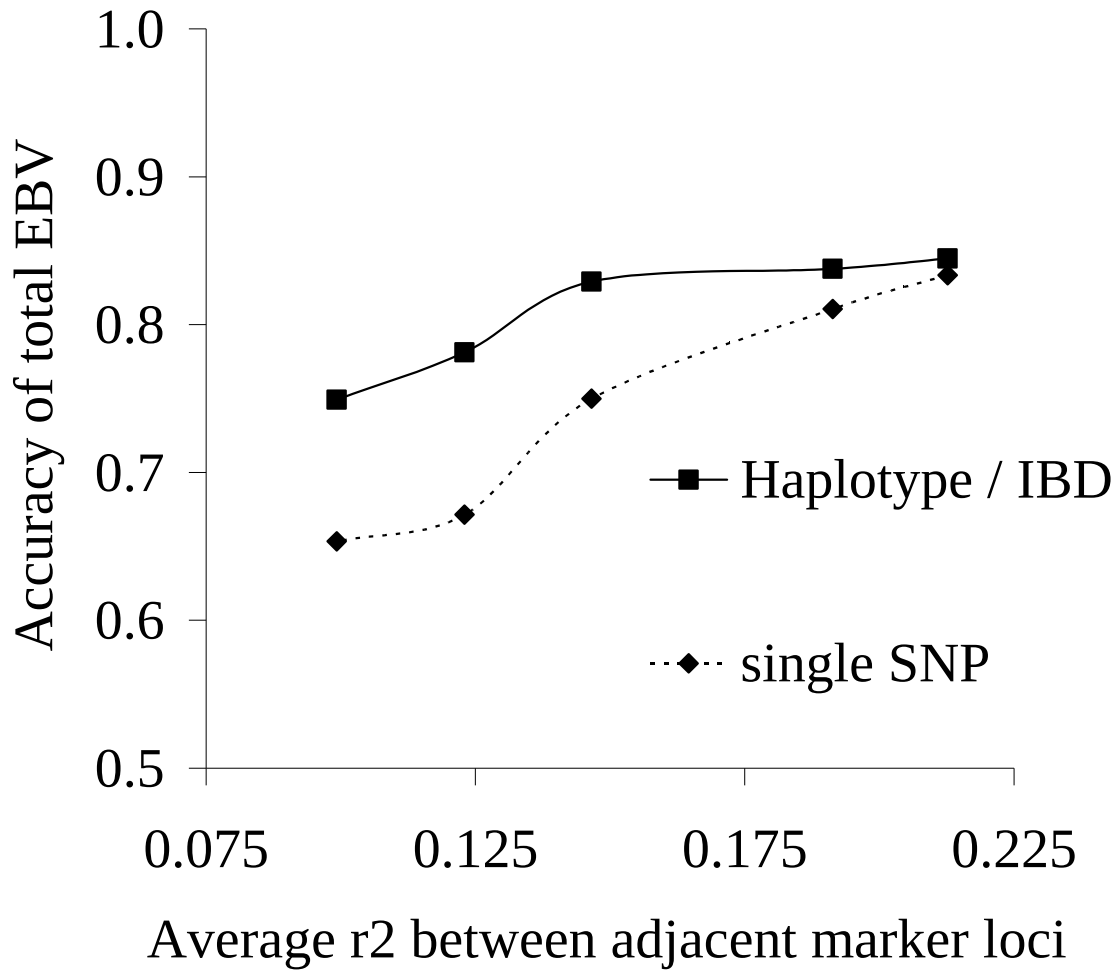
- Solved using Gibbs sampling, avoiding the Metropolis-Hastings step

Models

- SNP1: marker locus is putative QTL locus with two haplotypes (1 and 2)
- HAP_IBD10: midpoint of window of 10 marker loci is putative QTL locus with many haplotypes depending on P(IBD)

Accuracy using SNP alleles / haplotypes

- Haplotypes / IBD have higher accuracy at low marker density



QTL-MASXII workshop – May 2008; Uppsala Sweden

Simulated data:

- 14 medium-size QTL; 36 small QTL (Gamma distributed)

Results:

- Medium-sized QTL were (nearly) all found doing QTL-mapping or GS
- **NONE** of the small QTL was detected

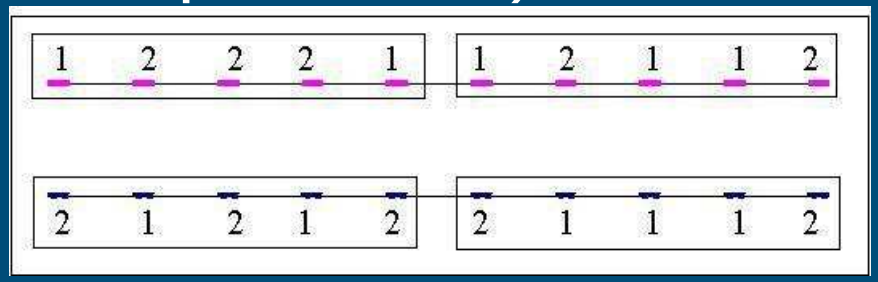
QTL-MASXII workshop – May 2008; Uppsala Sweden II

High accuracies for animals with no phenotypic performance:

- 0.92 (Villumsen et al., using IBS-haplotypes)
- 0.87 (Calus et al., using single SNP approach)

What causes difference IBS-haplotypes vs. single SNP?

- Non-overlapping IBS haplotypes (treating it as a locus with multiple alleles)



- IBS haplotypes may be better able to track QTL than single SNP approach, when a number of SNP are in moderate LD with the QTL

Results based on different haplotype lengths

Haplotype length	Number of 'loci'	Total number of haplo's	Accuracy young animals
HAP_IBD (20)	5994	366,959	0.84
1	6000	11925	0.87
2	3000	11630	0.89
5	1200	21607	0.90
10	600	41419	0.87
20	300	50572	0.82

⇒Optimal haplotype length probably resembles number of SNP that are on average in 'reasonable' LD with QTL

⇒Additional SNPs (i.e. increasing haplotype length) adds 'noise' and therefore reduces accuracy

Conclusions

Simulations in GS:

- Useful for hypothesis testing
- Be careful with assumptions about number and distribution of QTL!!

Parametrization of the model may help to:

- Fine-tune the model
- Make inferences about the data:
  - QTL-SNP LD
  - Distribution of QTL effects

QTL-MAS XIII Workshop
20-21 April, 2009
Wageningen
The Netherlands

ANIMAL SCIENCES GROUP
WAGENINGEN UR

Animal Breeding &
Genomics Centre

Acknowledgements

- Involved companies: Hendrix Genetics, CRV (HG)
- Netherlands Organisation for Scientific Research (NWO – Casimir)
- EU-projects:

ROBUSTMILK

SABRE CUTTING EDGE GENOMICS FOR SUSTAINABLE ANIMAL BREEDING