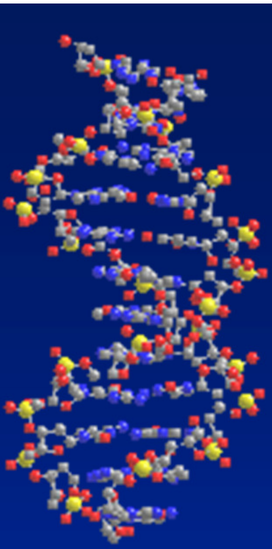
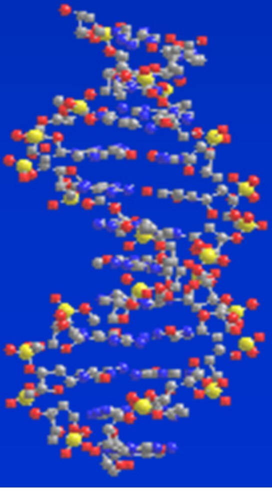


Genome wide assocation and Genomic Selection in the era of Genome sequencing



Course overview

- Day 1
  - Quantitative traits
  - Linkage disequilibrium
  - Genome wide association studies
- Day 2 and 3
  - Genomic prediction - BLUP and GBLUP
  - Genomic prediction – Bayesian methods
- Day 4
  - Validation of genomic predictions
  - Optimal breeding program design with genomic selection
- Day 5
  - Imputation and whole genome sequencing for genomic selection

Day 4

- Design of reference populations for Genomic selection
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
- Validation – traps for young players!

Reference populations for GS

- Also called “training sets”
- Two principles for design
- 1) Make it large -> QTL effects are small!
- 2) Make it close to candidates for selection

Reference populations for GS

- How large?

Reference populations for GS

- Parameters affecting accuracy of genomic breeding values
  - N Size of reference population
  - h^2 Heritability of trait
  - q Number of loci affecting the trait
    - Daetwyler et al. (2008), Goddard (2008)

$$r = \sqrt{Nh^2 / (Nh^2 + q)}$$

Reference populations for GS

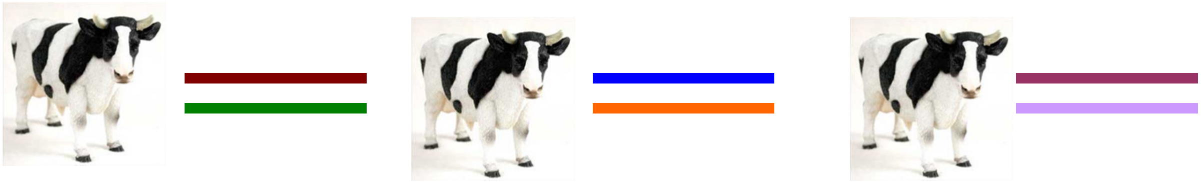
- Parameters affecting accuracy of genomic breeding values
  - N Size of reference population
  - h^2 Heritability of trait
  - q Number of loci affecting the trait ??

Reference populations for GS

- Number of loci affecting the trait
  - Conservative assumption - quantitative traits affected by very large number of loci, normal distribution of effects
  - = number of independent chromosome segments

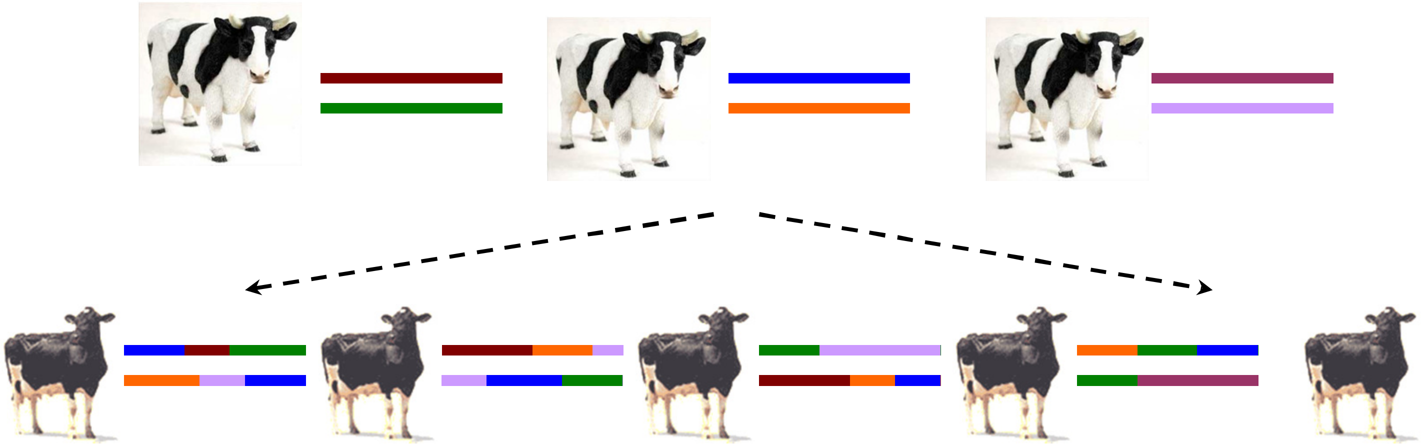
Reference populations for GS

- Number of loci affecting the trait
  - Conservative assumption - quantitative traits affected by very large number of loci, normal distribution of effects
  - = number of independent chromosome segments



Reference populations for GS

- Number of loci affecting the trait
  - Conservative assumption - quantitative traits affected by very large number of loci, normal distribution of effects
  - = number of independent chromosome segments



  - $q = 2N_eL$
    - N_e = effective population size, L is genome length in Morgans

Reference populations for GS

- accuracy of genomic breeding values

$$r = \sqrt{Nh^2 / (Nh^2 + q)}$$

- Number of loci affecting the trait
  - $q = 2N_eL$

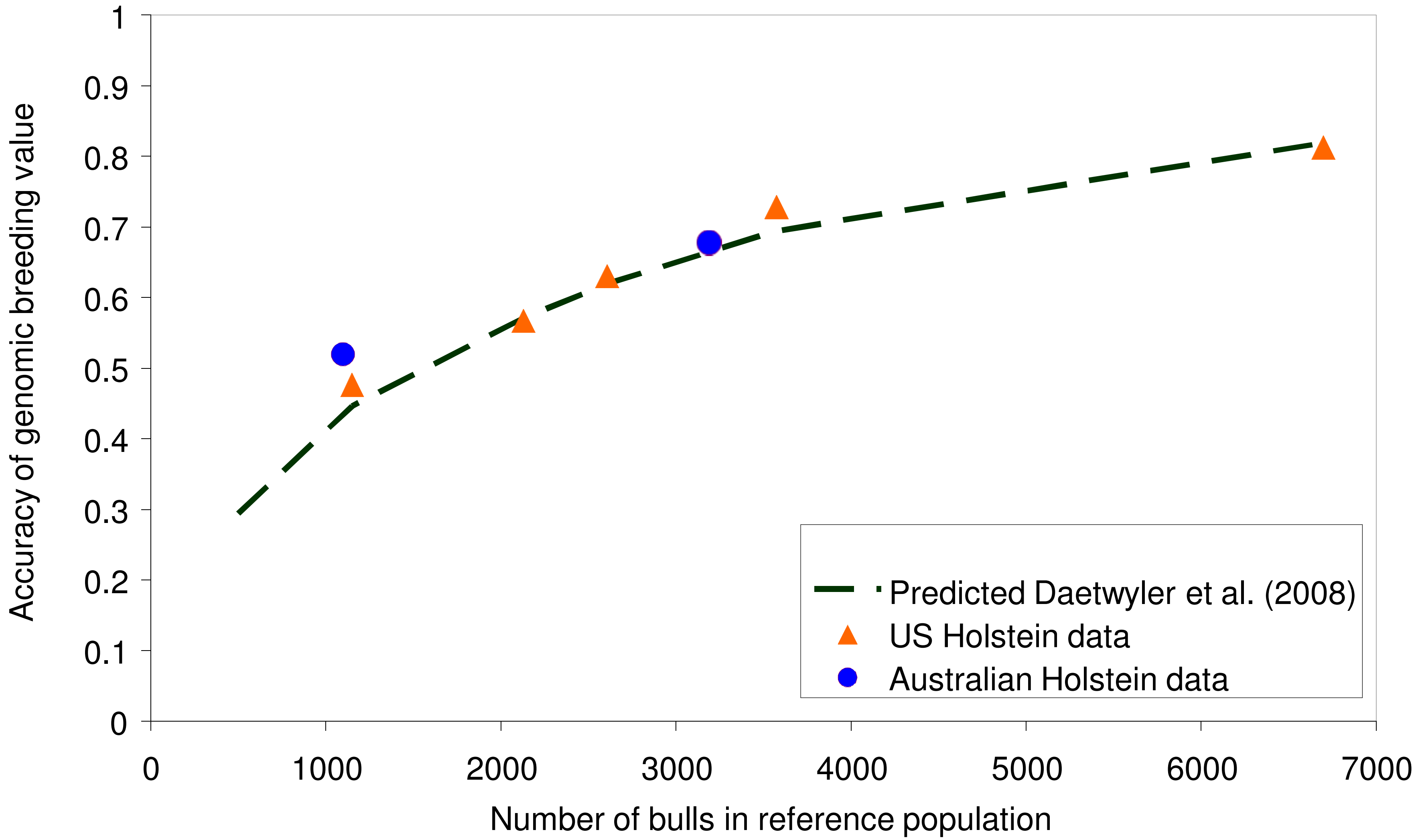
- Test against real data

Real Data

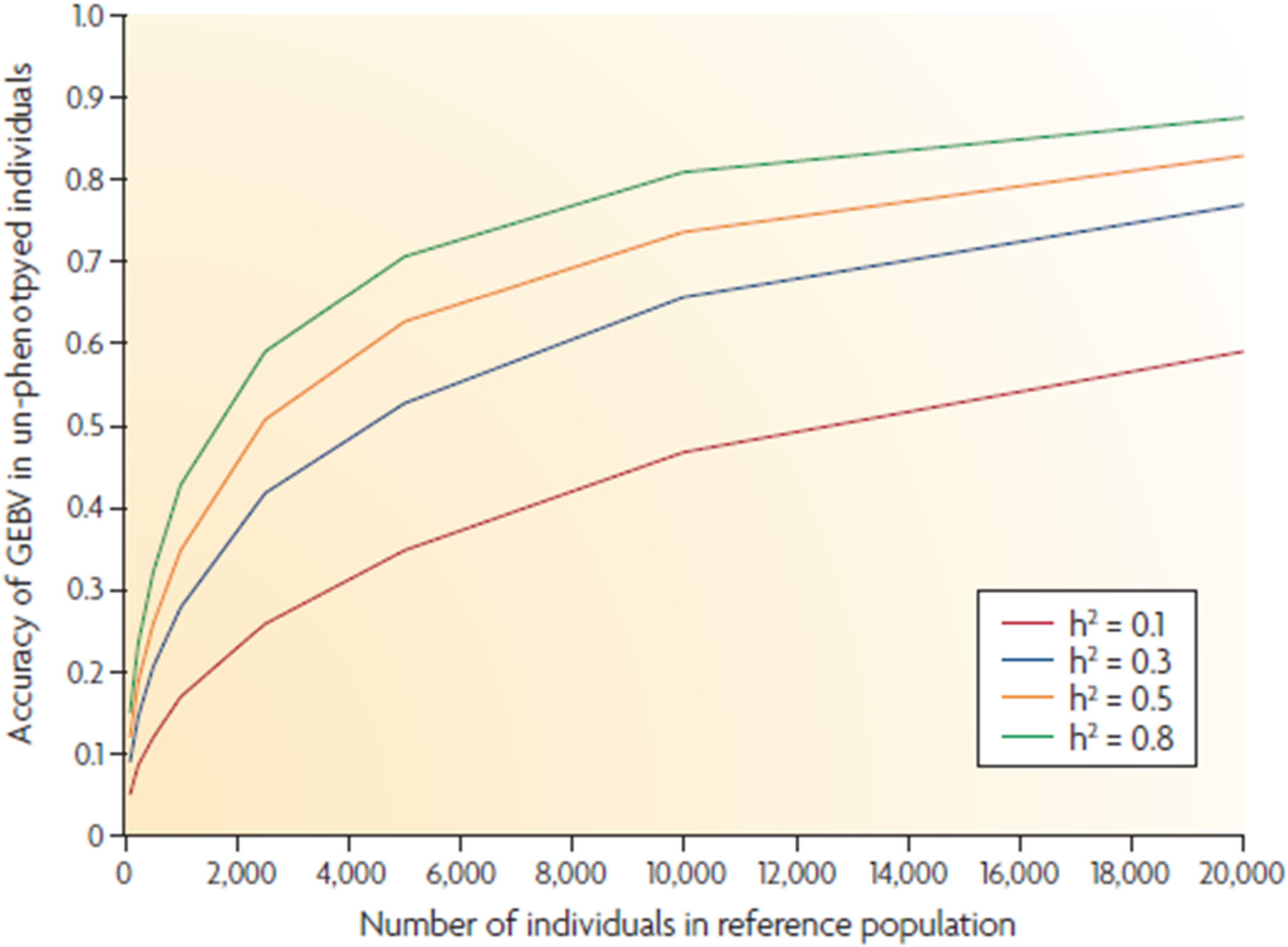
- Dairy cattle (Holsteins)
  - USA results (N=1000-6700) for Net Merit Index (VanRaden et al. 2009)
  - Australian results (N=1100-3300) for Australian Profit Ranking
  - $h^2=0.9$
  - $N_e = 100$

- Accuracies r(GEBV,EBV) in validation data sets

Deterministic prediction vs. Holstein data



Deterministic prediction

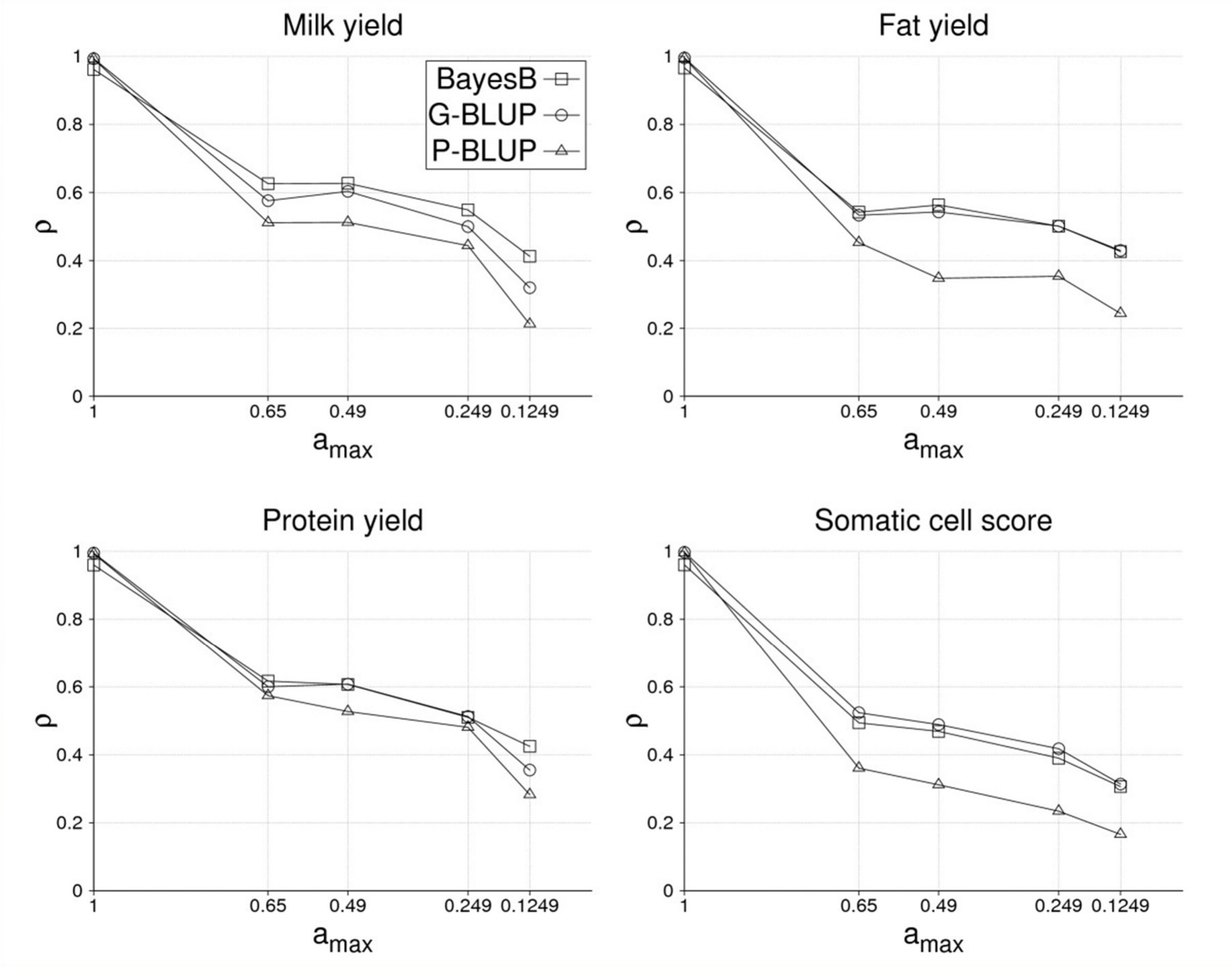


Reference populations for GS

- Which individuals/lines?

- The relationship of the reference population to the selection candidates affects accuracy of GEBV

Habier et al. 2010 Gen. Sel. Evol. 42:5



Reference populations for GS

- Which individuals/lines?

- The relationship of the reference population to the selection candidates affects accuracy of GEBV

- Need individuals close to those being predicted in reference
- At the same time, as diverse as possible so that many individuals/lines can be accurately predicted

Day 4

- Design of reference populations for Genomic selection
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
- Validation – traps for young players!

How many markers?

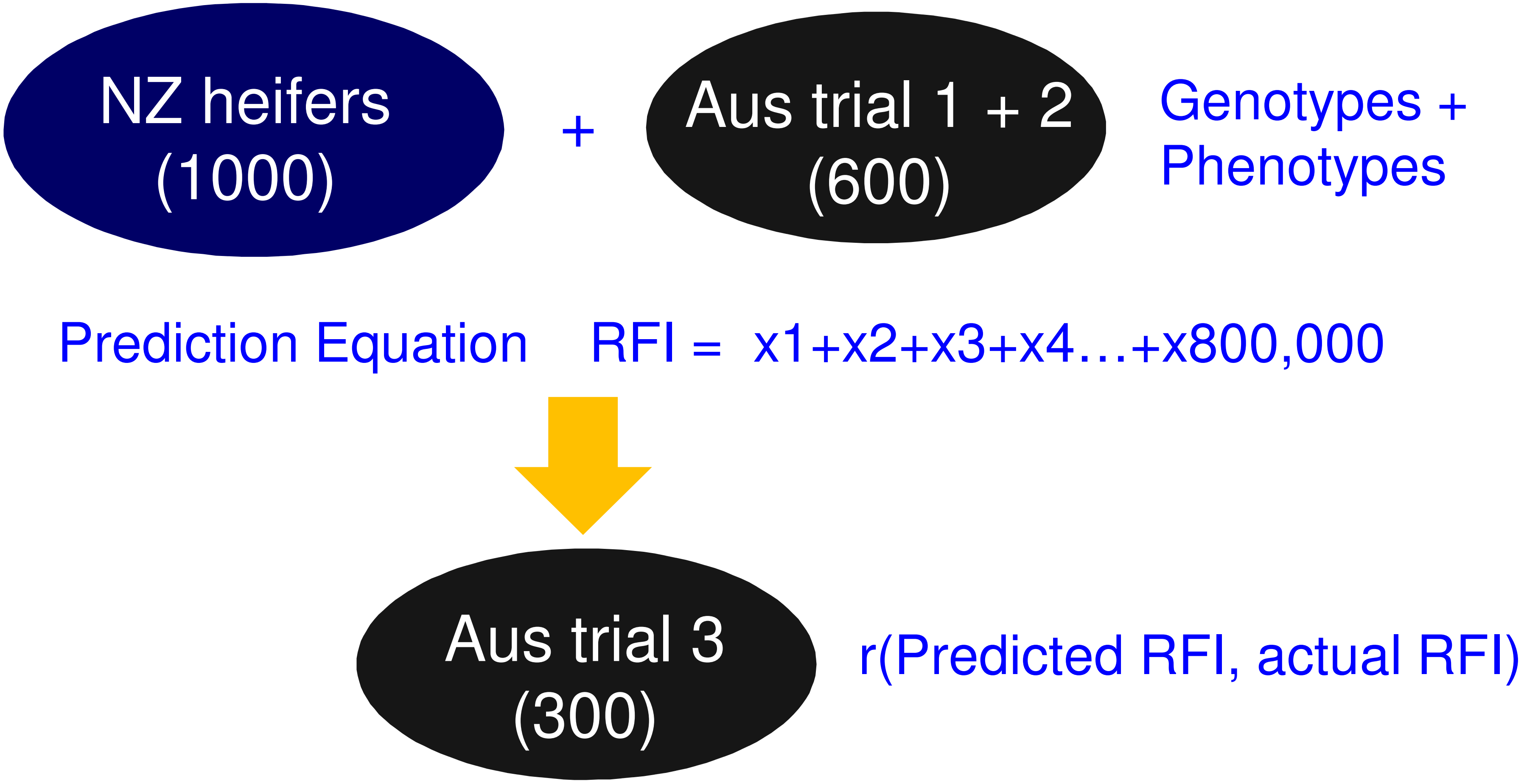
- 10*Ne*L
  - Ne = effective population size
  - L = length of genome (Morgans)
  - Meuwissen et al. (2009) Gen. Sel Evol. Jun 11;41:35

  - Eg. Holsteins
  - 10*100*30 = 30,000

Genomic Predictions Residual Feed Intake

- Collaboration DPI Vic, Livestock Improvement Corporation and Dairy NZ (Richard Spelman, Kevin MacDonald, et al.)
- 1000 heifers each
- Genotyped 800,000 SNPs (Illumina Bovine HD)

Genomic predictions



NZ heifers (1000) + Aus trial 1 + 2 (600)

Genotypes + Phenotypes

Prediction Equation RFI = x1+x2+x3+x4…+x800,000

Aus trial 3 (300)

r(Predicted RFI, actual RFI)

Genomic Predictions Residual Feed Intake

- To derive prediction equation
- GBLUP -> all markers have small, non zero effect
- BayesR -> proportion of markers have zero effect, rest have small to moderate effects

Accuracy GEBV Residual Feed Intake

Trait	Marker Panel	GBLUP	BayesR
Liveweight	50K	0.35	0.35
	800K	0.38	0.40
Residual Feed Intake	50K	0.29	0.39
	800K	0.29	0.41

Day 4

- Design of reference populations for Genomic selection
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
- Validation – traps for young players!

Genomic selection

- How often to re-estimate SNP effects?
  - If the markers used in genomic selection were actually the underlying mutations causing the QTL effects, the estimation of SNP could be performed once in the reference population.
  - GEBVs for all subsequent generations could be predicted using these effects.

Genomic selection

- How often to re-estimate SNP effects?
  - In practise will be markers with low to moderate levels of r^2 with the underlying mutations (QTL)
  - Do not capture all of QTL variance
  - Over time, recombination between the markers and QTL will reduce the accuracy of the GEBV using SNP effects from the original reference population.
  - We need to re-estimate SNP effects
  - How often?

Genomic selection

- How often to re-estimate SNP effects?

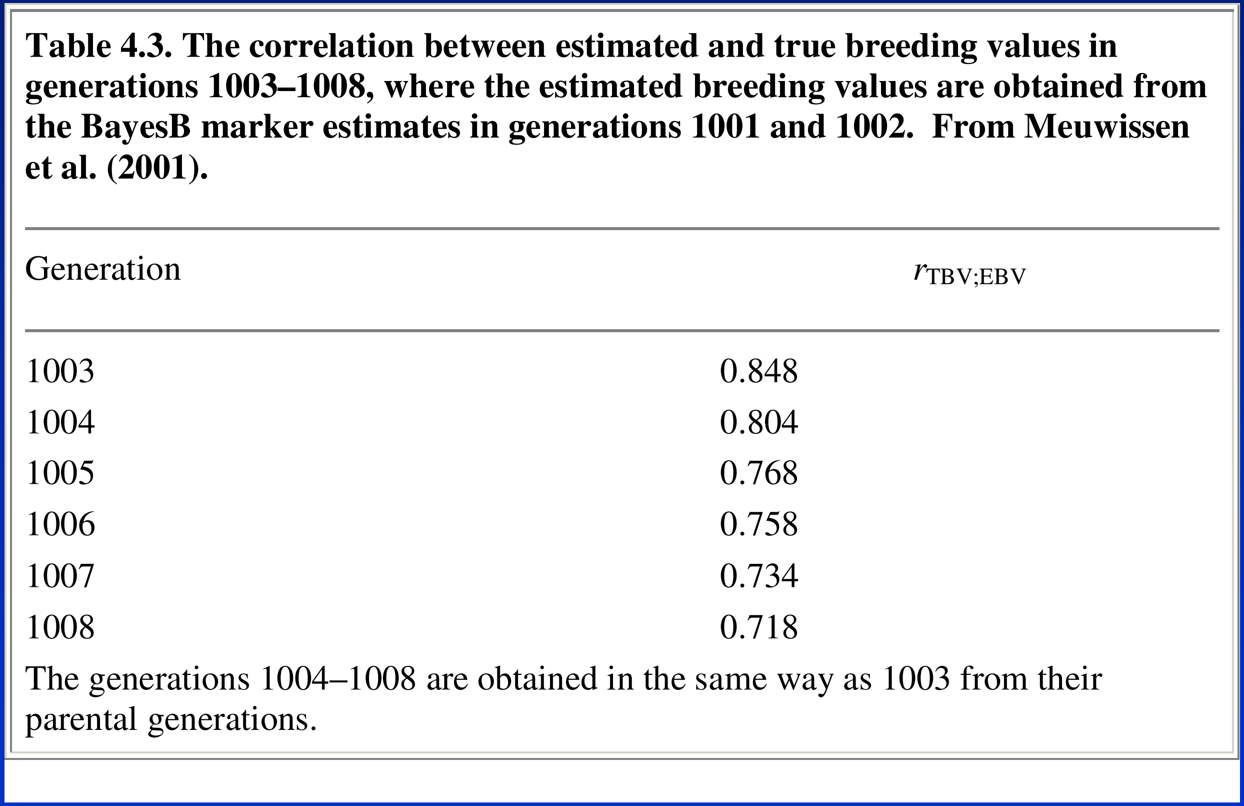


Table 4.3. The correlation between estimated and true breeding values in generations 1003–1008, where the estimated breeding values are obtained from the BayesB marker estimates in generations 1001 and 1002. From Meuwissen et al. (2001).

Generation	$r_{TBV;EBV}$
1003	0.848
1004	0.804
1005	0.768
1006	0.758
1007	0.734
1008	0.718

The generations 1004–1008 are obtained in the same way as 1003 from their parental generations.

Genomic selection

- However decay of accuracy is dependant on genomic selection method…..

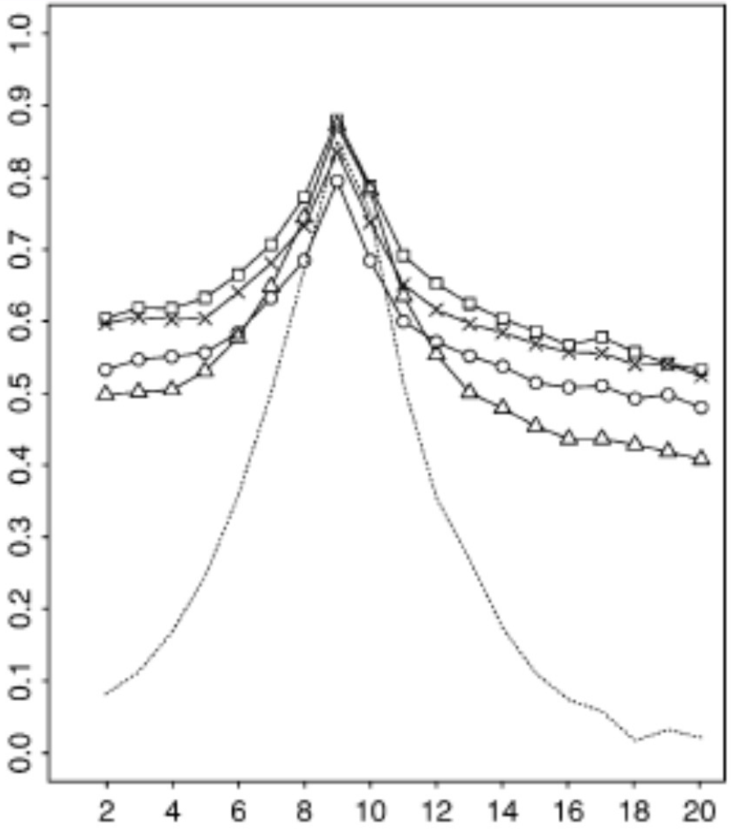
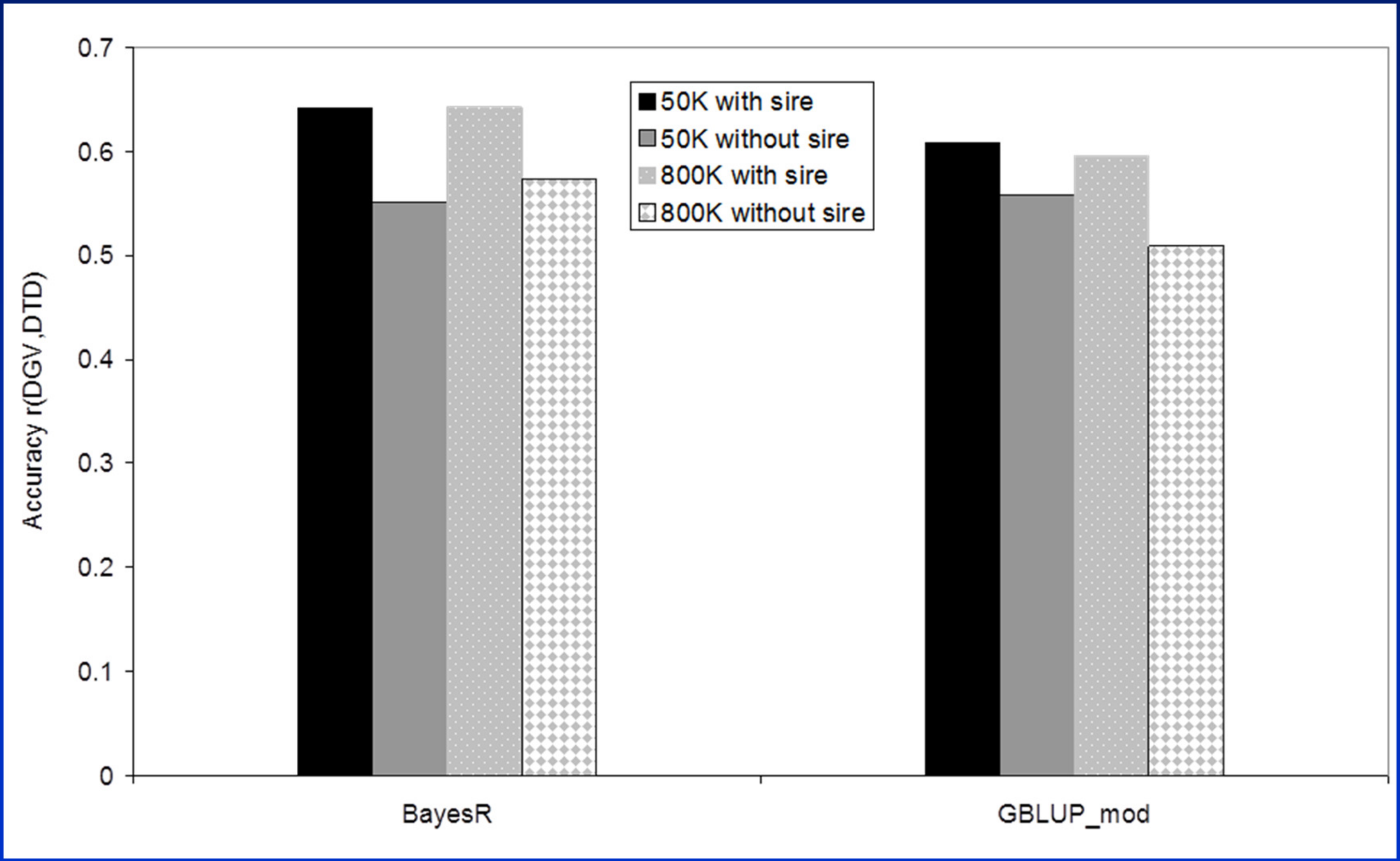


FIGURE 3.—Accuracies of GEBVs obtained by fixed regression–least squares (FR–LS), random regression–BLUP (RR–BLUP), Bayes-B1, and Bayes-B2 in lines 1 and 2 in comparison to the accuracies of EBVs obtained by trait-pedigree–BLUP (TP–BLUP) using 1000 individuals in generation 10 each with a trait phenotype and 1000 SNP markers (160 replicates).

- Habier et al. (Genetics 177:2389)

Genomic selection

- How often to re-estimate SNP effects?



  – Denser markers >> generations between re-estimation of effects

Day 4

- Design of reference populations for Genomic selection
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
- Validation – traps for young players!

Optimal breeding program design

- With genomic selection, we can potentially predict GEBV with an accuracy of 0.8 for selection candidates at birth
- How does this change the optimal breeding program design?

Optimal breeding program design

- With genomic selection, we can potentially predict GEBV with an accuracy of 0.8 for selection candidates at birth
- How does this change the optimal breeding program design?

- Breed from animals as early as possible

Optimal breeding program design

- In dairy cattle current structure is
  - Each year select a team of calves to form a progeny test team
  - At two years of age these bulls are mated to random cows from the population
  - At four years of age the daughters of the bulls start lactating

Optimal breeding program design

- In dairy cattle current structure is
  - Each year select a team of calves to form a progeny test team
  - At two years of age these bulls are mated to random cows from the population
  - At four years of age the daughters of the bulls start lactating
  - At five years of age the bulls receive a progeny test “proof” based on the performance of their daughters
  - The bulls are then selected on the basis of these proofs to be “breeding bulls”
    - Semen sold to commercial farmers

Optimal breeding program design

- In dairy cattle with genomic selection..
  - Genotype a large number of bull calves from the population
  - Calculate GEBVs for these calves
    - Accuracy = 0.8 = accuracy of progeny test
  - Select team based on GEBV
  - Sell semen from these bulls as soon as they can produce it

Optimal breeding program design

- In dairy cattle with genomic selection..
  - Genotype a large number of bull calves from the population
  - Calculate GEBVs for these calves
    - Accuracy = 0.8 = accuracy of progeny test
  - Select team based on GEBV
  - Sell semen from these bulls as soon as they can produce it
  - Generation interval reduced from ~4 yrs to ~ 2 yrs
    - $\Delta G = ir\sigma_g/L$
  - Double rate of genetic gain

Optimal breeding program design

- In dairy cattle with genomic selection..
  - Genotype a large number of bull calves from the population
  - Calculate GEBVs for these calves
    - Accuracy = 0.8 = accuracy of progeny test
  - Select team based on GEBV
  - Sell semen from these bulls as soon as they can produce it

  - Generation interval reduced from ~4 yrs to ~ 2 yrs
    - $\Delta G = ir\sigma_g/L$
  - Double rate of genetic gain
  - Save the cost of progeny testing!
    - Reduce costs by 92% (Schaeffer et al. 2006)

Optimal breeding program design

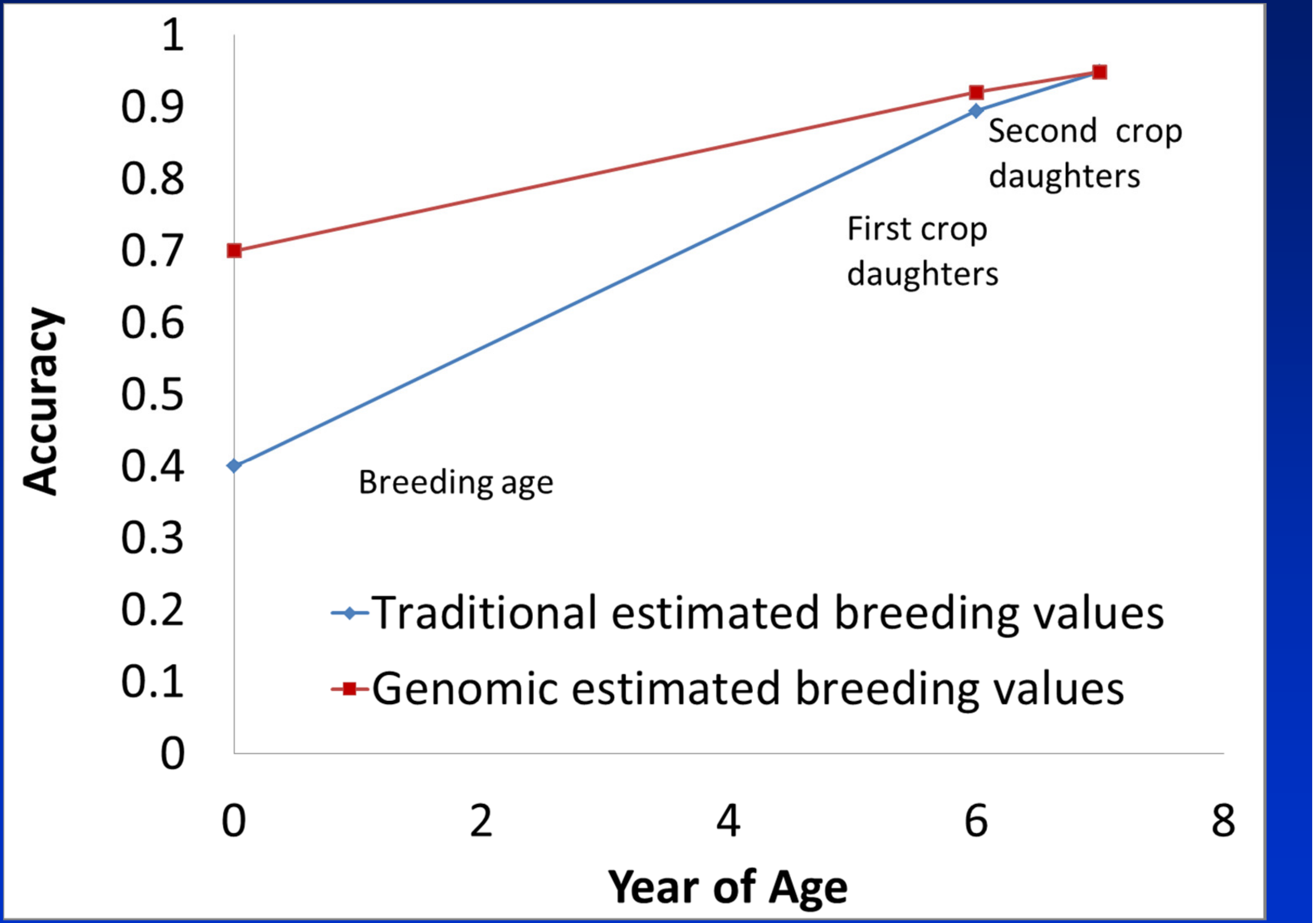
- In pigs
  - Currently EBV for traits like meat quality, sow fertility, disease resistance based on performance of relatives
  - Exploits between family variance, not within
  - Feed conversion efficiency = expensive

Optimal breeding program design

- In pigs with genomic selection
  - Accurate GEBVs for meat quality, sow fertility, disease resistance based on own marker genotype
  - Exploits between and within family variance
  - Feed conversion efficiency GEBV?
  - Will accelerate genetic gain for these traits
  - Reverse declines in meat quality for example

Genomic selection: Dairy cattle

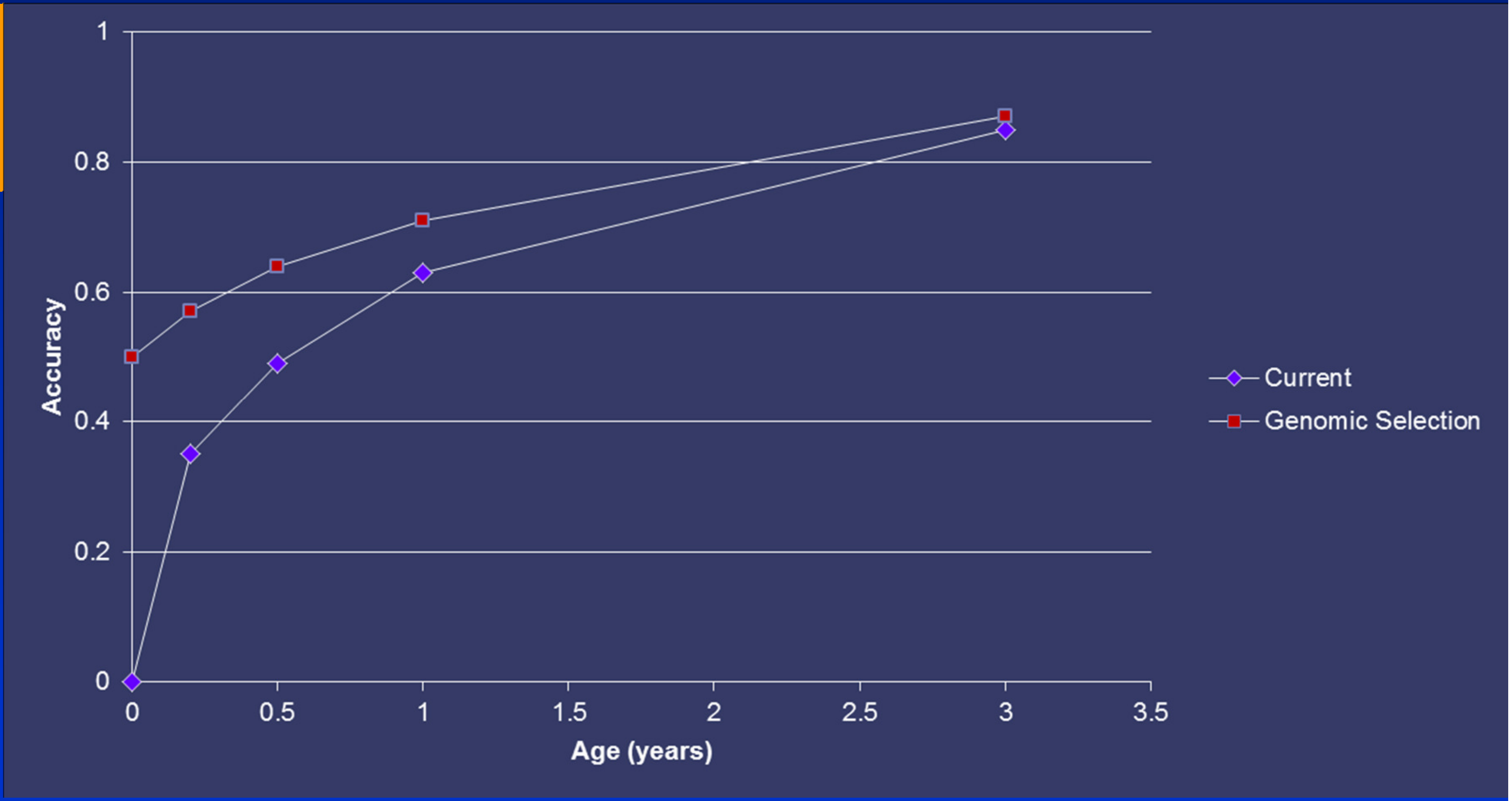
$$\Delta G = \frac{i r \sigma_g}{L}$$



Genomic selection: Meat sheep

But gains to be made by selection for breeding objective traits directly, eg. Lean meat yield vs. scanned eye muscle area

$$\Delta G = \frac{ir\sigma_g}{L}$$



Increased genetic gain from genomic selection

Industry	Potential increase
Dairy Cattle	60-120% (Pryce et al. 2011)
Meat sheep	21% (van der Werf 2011)
Wool sheep	38% (van der Werf 2011)
Beef cattle	29-158% Van Eenennaam 2011
Layers	40% (Dekkers et al 2009)
Broilers	20% (Dekkers et al. 2009)

Optimal breeding program design

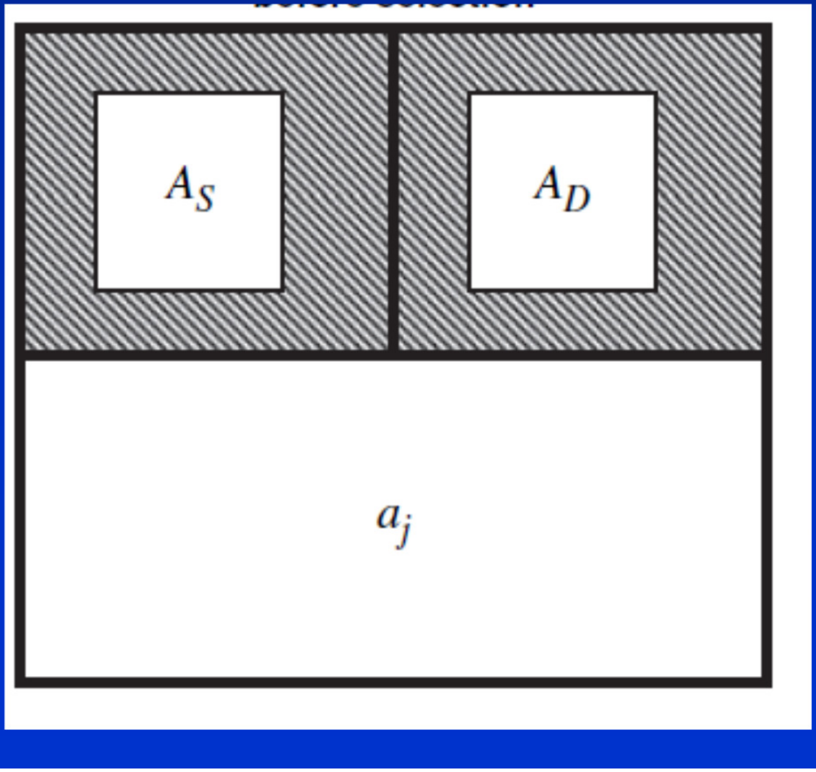
- Synergy with reproductive technologies
- If we can predict genetic gain accurately at birth, genetic gain depends on generation interval
- Reproductive technologies to reduce this
  - Juvenile in-vitro embryo transfer?
  - Extreme technologies like in-vitro meosis
- Must manage inbreeding!!

Day 4

- Design of reference populations for Genomic selection
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
- Validation – traps for young players!

3 Components of a breeding value

- Breeding values of sire and dam
- Mendelian sampling term
  - Deviation due to sampling of alleles from parents



Using genetic markers

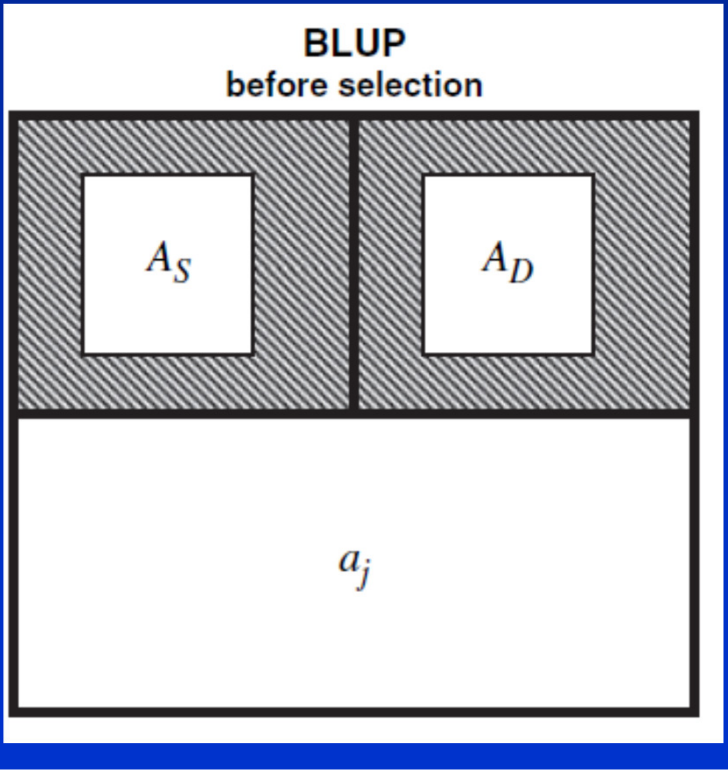
- Markers explain some within family variance
  - Give information on Mendelian sampling term

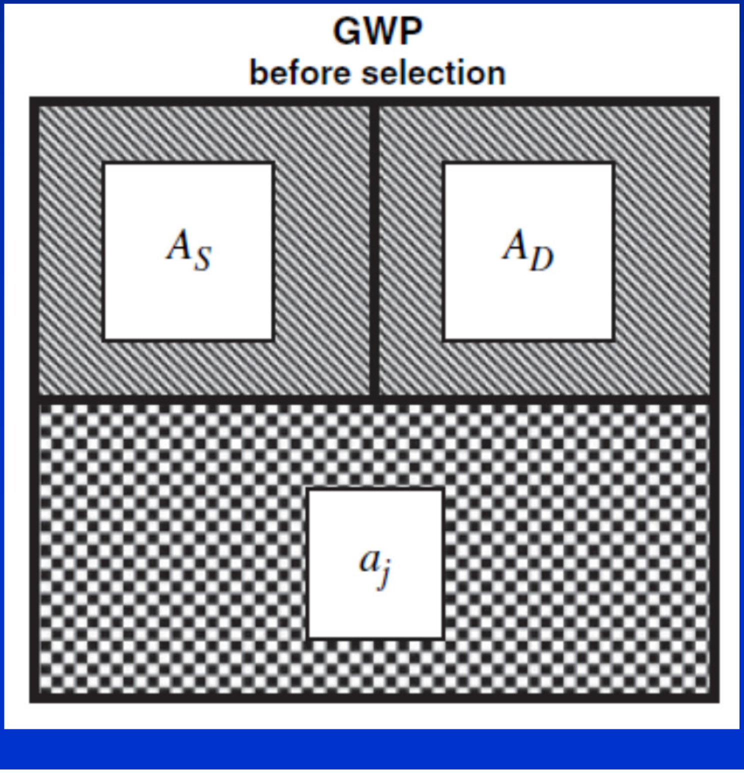
- Genomic selection can estimate Mendelian sampling term very accurately
  - Due to many markers
  - Expected versus ‘realised’ relationships

Meuwissen and van Arendonk 1992, MacKinnon and Georges, 1998, Woolliams et al 2002

Methods’ utilisation of components

- Assume a juvenile without phenotypes
- Genomic selection uses Mendelian sampling term
- Selection can act on whole breeding value!





Measures of inbreeding

- Rate of inbreeding per generation
  - Appropriate for comparing methods
  - Counteracting forces also occur per generation
    - Mutation
- Rate of inbreeding per year
  - Relevant for breeding programs

Effect of genomic selection on inbreeding

- Example: 4 young elite full brothers
  - Pedigree breeding values (BLUP) are the parent average → the same for all 4 bulls
    - Select all 4
  - GEBV will be different for all 4
    - Only select best

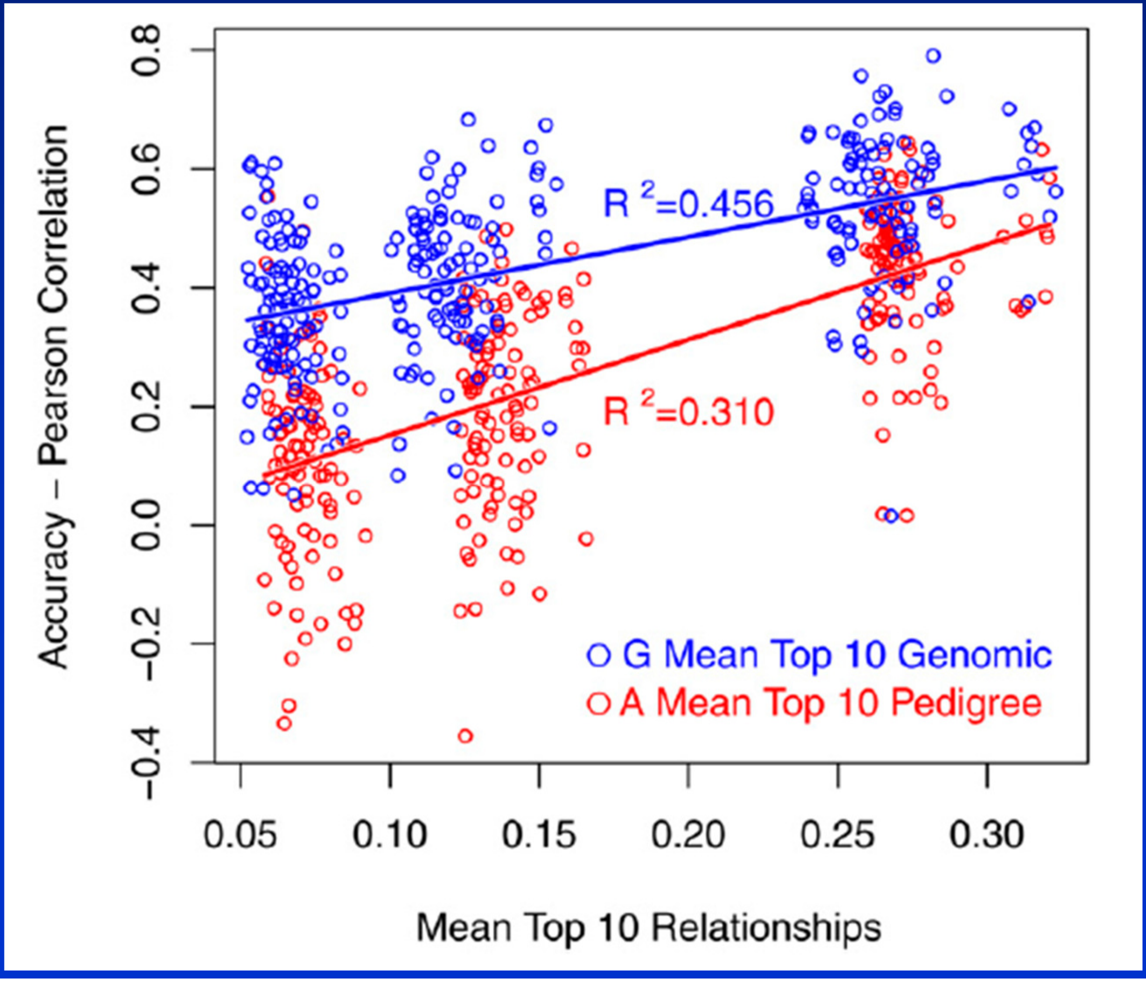
- Genomic selection results in less inbreeding per generation than BLUP
  - Selection on Mendelian sampling terms
  - reduced co-selection
  - Breeding values of sibs less correlated

Inbreeding of genomic selection breeding programs

- Genomic selection can drastically reduce generation intervals
- Increases genetic gain per year but also increases rate of inbreeding per year
- Need for controlling inbreeding:
  - Mate selection for less inbred progeny
  - Optimum contribution selection
    - Maximise genetic gain at a given level of inbreeding
  - Use a diverse set of sires
  - Highly likely that elite sires change every year

Influence of relationships on GS accuracy

- Relationship of validation to reference important contributor to accuracy



Influence of relationships on GS accuracy

- Relationship of validation to reference important contributor to accuracy

- Closely related individuals have GEBV with higher accuracy
  - Higher accuracy → less regression towards mean → greater chance to have high GEBV → greater chance to be selected → greater inbreeding?

- However, also an issue with BLUP
  - Stronger trend with genomic selection?
  - Implications for reference populations

Day 4

- Design of reference populations for Genomic selection
- How many markers?
- How often to re-estimate SNP effects?
- Optimal breeding program design with genomic selection
- Genomic selection and inbreeding
- Validation – traps for young players!

Validation of genomic selection

- Aim of genomic selection
  - predict (young) selection candidates without phenotypes
- How to test or validate predictions?
- Test predictions in a population sample that is similar to selection candidates
- Key principle of validation
  - Independence of reference and validation populations

Accuracy and bias

- Most commonly used:
  - r=correlation(GEBV,phenotypes)
  - Gives accuracy of a group of individuals
- Individual accuracy
  - Calculated using the prediction error variance from the diagonal of the coefficient matrix (GBLUP)
- Regression of phenotypes (y) on GEBV (x)
  - Deviation from expectation of the slope
  - Expectation is usually 1
  - If not close to expectation → then biased

Standard error of a correlation

- Correlations have a standard error which depends on sample size and the magnitude of the correlation

- An approximation of the standard error was given by Fisher (see Fisher z transform)
  - SE ~ 1/sqrt(N-3)

- In our practical examples
  - 31 individuals
  - SE = 1/sqrt(31-3) = 0.189

Two main ways to (cross)-validate

- 1st way: Highly accurate individuals
  - Dairy bull progeny test (e.g. Daughter trait deviations)
  - Very large progeny groups or many clones (plants)
  - Step1: Estimate marker effects in reference population
  - Step2: Predict highly accurate individuals and calculate accuracy

- 2nd way: ‘Classic’ cross-validation
  - Step 1: Divide dataset into n subsets of individuals
  - Step 2: Predict each subset using all other subsets
  - Step 3: Calculate accuracy in each subset and take mean across all subsets

Approximating the accuracy of breeding value

- The upper limit of genomic selection accuracy is given by the accuracy of observations

- Divide by accuracy of observation to approximate accuracy of additive genetic component (i.e. breeding value)

- If using DYD
  - r/accuracy(DYD)
- If using phenotypes
  - r/sqrt(h^2)

Validation - Independence

- Always ask question:
  - If the validation individuals were selection candidates what data would be available?
  - Then only use that data for reference!
- Independence of ‘data’ not independence in relationship

Independence

- Validation individuals are not used in the reference pop

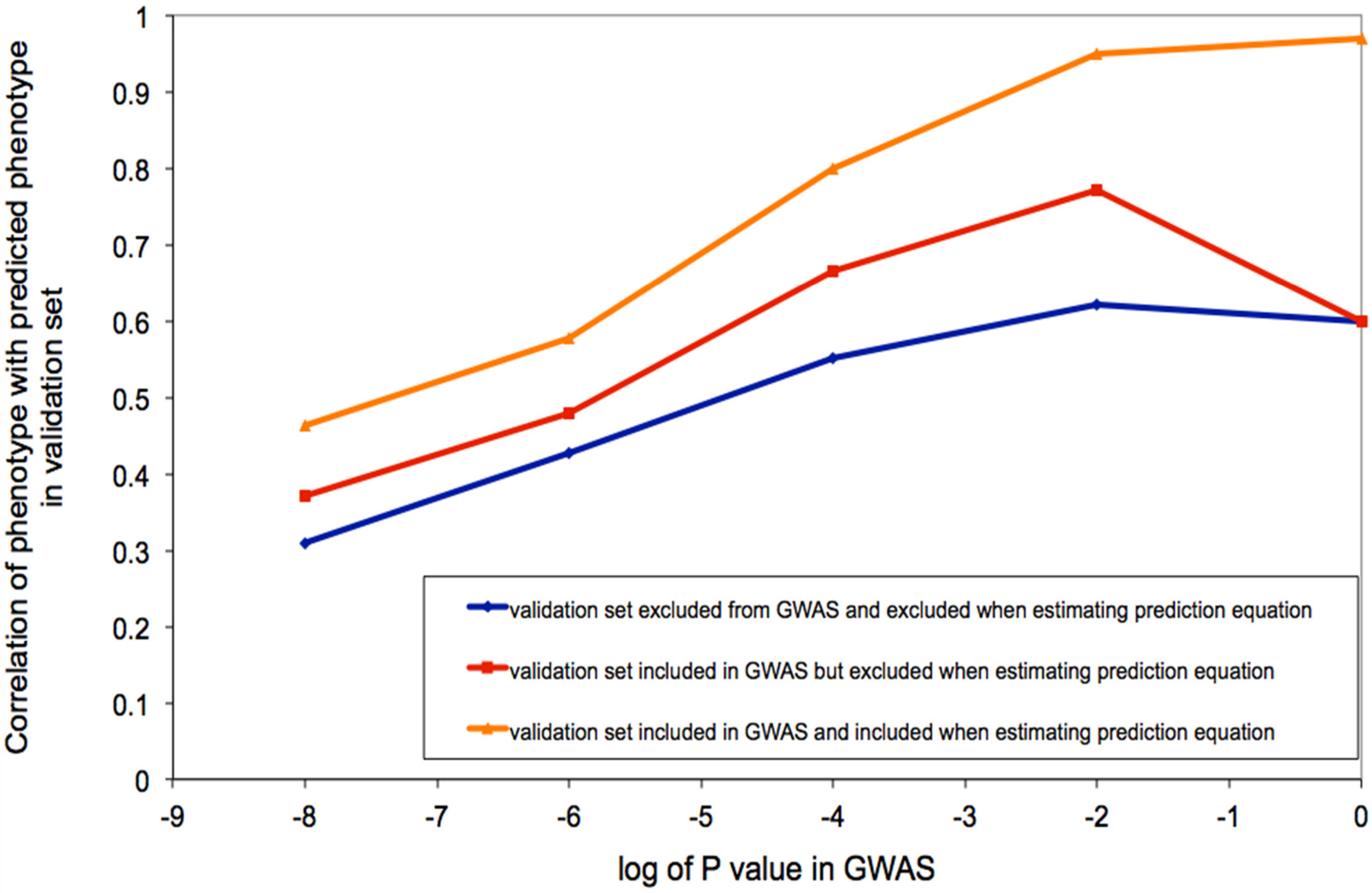
- Validation phenotypes do not contribute to observed variables of reference pop
  - E.g. excluded when calculating estimated breeding values

- Validation individuals do not have contemporaries of same age in reference

Independence

- Choosing a subset of SNP with a GWAS
  - Only use reference population to choose SNP
  - If validation is used for GWAS then you are overfitting (upward bias in accuracy)

Independence



Target of prediction

- Validation population should be similar to selection candidates

- Similar relationship to reference as sel. cand.
  - Same number of generations removed
  - Same breeds
  - Same population

- Same SNP density
  - Consider imputation error