

# *Continuing the transformation*



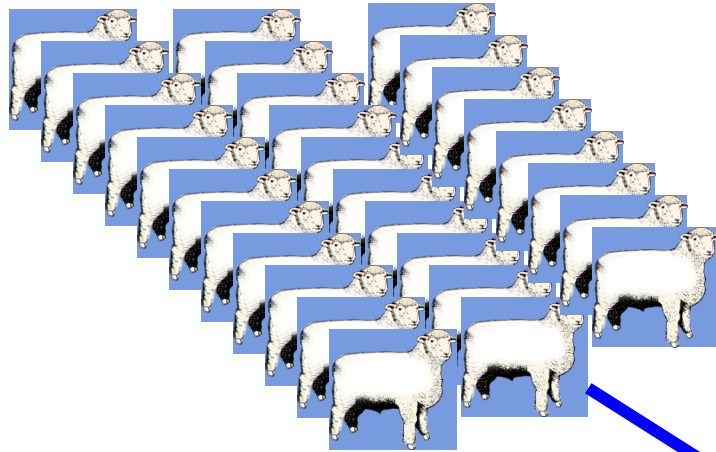
## Accuracy of Genomic Prediction

Julius van der Werf  
and Sang Hong Lee

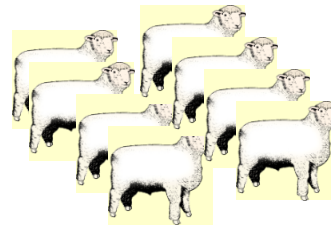
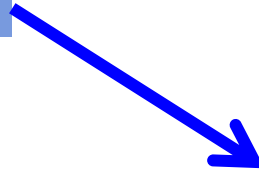
**UNE**  
University of  
New England



# Genomic Prediction: basic idea



Reference population  
measured and DNA tested



Target Set  
Only DNA tested

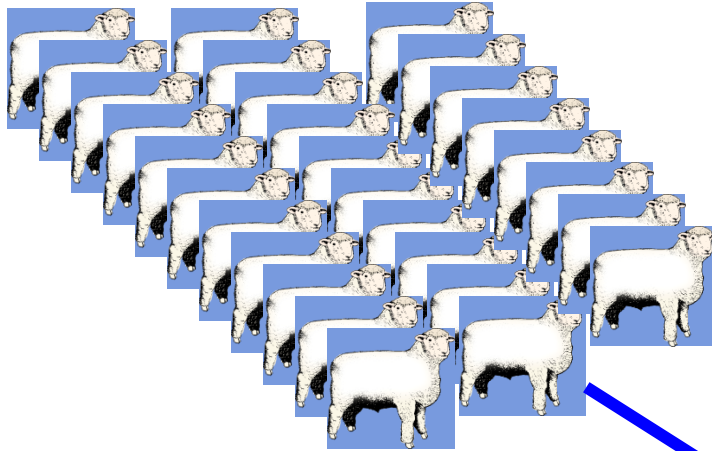
To predict a trait EBV at a young age,

good for:

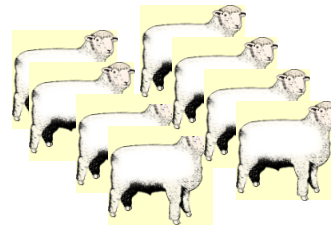
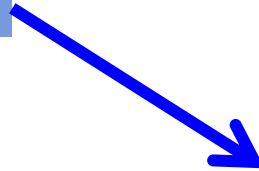
late traits

hard to measure traits

# Genomic Prediction: basic idea



Reference population  
measured and DNA tested



Target Set  
Only DNA tested

What if reference population is

- Another breed
- Multi-breed
- Crossbreds
- Small
- Less related
- Heterogeneous

?

# How does genomic prediction work?

- Markers in LD with QTL?
- Genomic Relationships?
- We know that GBLUP is equivalent to SNP-BLUP
- We observe that SNP BLUP and Bayesian methods are pretty similar → “infinitesimal model”

# Genomic Prediction: GBLUP

## Example:

Data on sire 1, his sons (2 and 3) and an unrelated individual (4)

want to predict 5 (also a son of 1) ← no data

A-matrix (pedigree-based)

1	0.5	0.5	0	0.5
0.5	1	0.25	0	0.25
0.5	0.25	1	0	0.25
0	0	0	1	0
0.5	0.25	0.25	0	1

G-matrix (DNA-based)

1	0.5	0.5	0.02	0.5
0.5	1	0.20	0.015	0.20
0.5	0.20	1	0.025	0.30
0.02	0.015	0.025	1	0.025
0.5	0.20	0.30	0.025	1

Variation in  
relationship  
(animal 5 with 2  
and 3

Also a small  
relationship with  
'unrelated'

# Genomic Prediction: GBLUP

Example:

Data on sire 1, sons 2 and 3, 4  
unrelated, want to predict 5

A-matrix (pedigree-based)

1	0.5	0.5	0	0.5
0.5	1	0.25	0	0.25
0.5	0.25	1	0	0.25
0	0	0	1	0
0.5	0.25	0.25	0	1

G-matrix (DNA-based)

1	0.5	0.5	0.02	0.5
0.5	1	0.20	0.015	0.20
0.5	0.20	1	0.025	0.30
0.02	0.015	0.025	1	0.025
0.5	0.20	0.30	0.025	1

BLUP

$$\hat{u}_5 = 0.1136.y_1 + 0.0455.y_2 + 0.0455.y_3$$

GBLUP

$$\hat{g}_5 = 0.1135.y_1 + 0.0328.y_2 + 0.0591.y_3 + 0.00519.y_4$$

# Genomic Prediction: GBLUP

Example:

Data on sire 1, sons 2 and 3, 4  
unrelated, want to predict 5

A-matrix (pedigree-based)

1	0.5	0.5	0	0.5
0.5	1	0.25	0	0.25
0.5	0.25	1	0	0.25
0	0	0	1	0
0.5	0.25	0.25	0	1

BLUP uses: Family Info

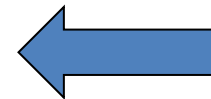
G-matrix (DNA-based)

1	0.5	0.5	0.02	0.5
0.5	1	0.20	0.015	0.20
0.5	0.20	1	0.025	0.30
0.02	0.015	0.025	1	0.025
0.5	0.20	0.30	0.025	1

GBLUP uses: Family Info  
Segregation within family  
Info on 'unrelated'

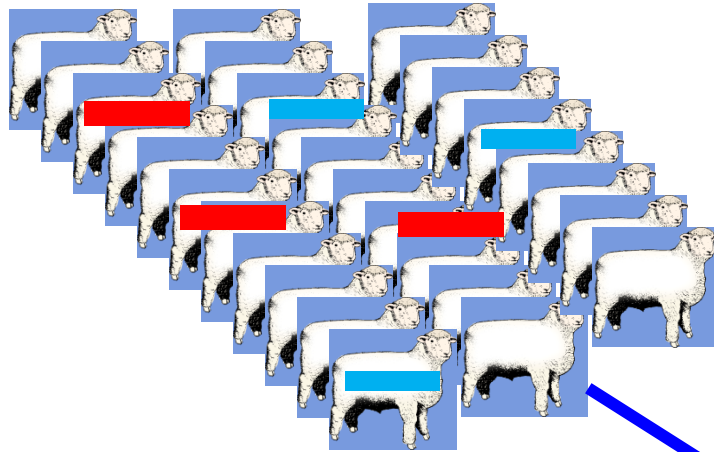
# Genomic prediction accuracy

- Derive from the model, e.g. PEV from GBLUP mixed model equations
- Validate with other EBVs or phenotypes
  - Validation population
  - Cross-validation
- Predict in advance based on theory and assumptions about population

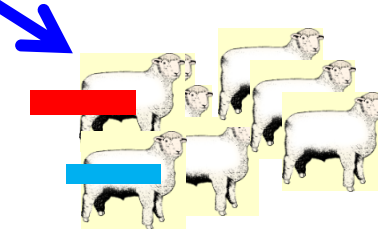




# Genomic Prediction: basic idea



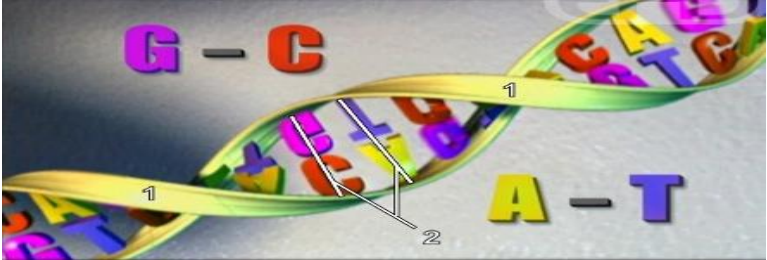
Reference population  
measured and DNA tested



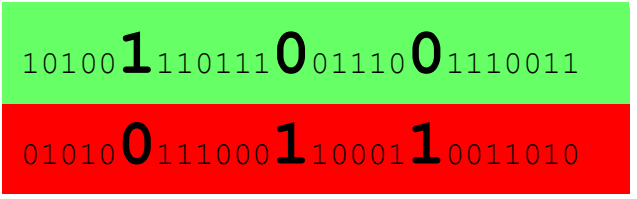
Target Set  
Only DNA tested

Illustrating (dis-)similarity of chromosome segments

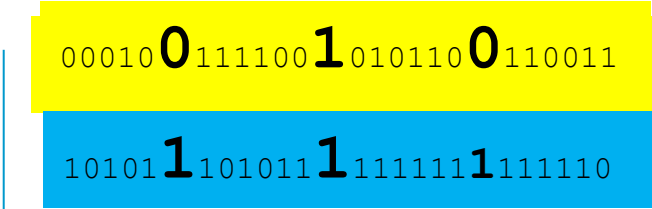
# Genotype information



Father



Mother



Progeny

*Chromosome segments are passed on*

# A whole population of haplotypes

Individual

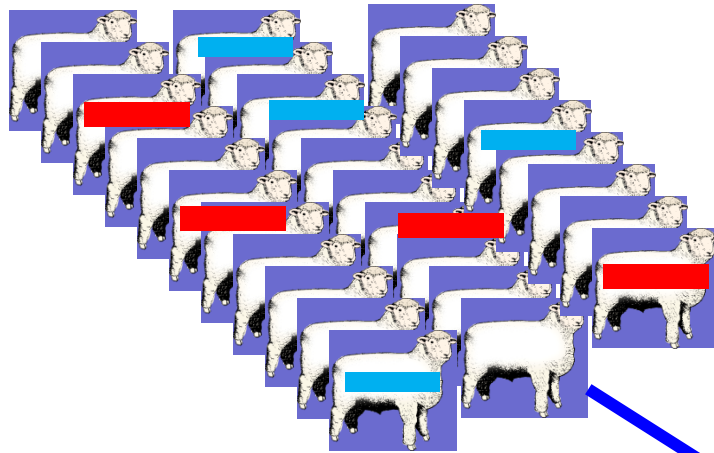
1	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	###	1541	1541	1541
	129	129	129	129	129	129	129	655	655	655	655	655	655	655	###	1129	1129	1129
2	1088	1088	1088	1088	1088	1088	1088	1088	1192	1192	1192	1192	1192	1192	###	1192	623	623
	178	655	891	891	891	891	891	891	891	1136	1136	1136	1136	1136	735	735	735	735
3	129	129	129	129	129	129	129	655	655	655	655	655	655	655	###	1038	1038	1038
	1192	1192	1192	1192	1192	1192	1192	1192	1192	1192	1192	1192	1192	1192	###	1192	1192	1043
4	424	424	424	424	424	424	424	424	503	503	503	503	503	503	503	503	503	503
	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	###	1541	1541	1541
5	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	###	1541	1541	1541
	1136	1136	1136	1136	1136	1136	1136	1136	178	178	178	178	178	178	178	1541	1541	1541
6	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	###	1043	1043	1043
	1478	1478	1478	1478	1478	1478	1478	1478	1478	1478	1478	1478	1478	1478	###	1478	1478	1478
7	129	129	129	129	1038	1038	199	199	129	129	129	129	129	129	129	129	129	129
	655	655	655	1358	1358	1358	1358	1358	1358	1358	342	342	342	342	342	342	342	1043
8	444	444	444	444	444	444	444	444	444	444	444	444	444	444	444	444	444	444
	1358	1358	1358	1358	1358	1358	1358	1358	1358	1358	342	342	342	342	342	342	342	342
9	1296	1296	1296	1296	1296	321	321	321	321	812	812	674	674	674	674	674	674	674
	891	891	891	210	210	210	210	1255	1262	1262	1478	1478	1478	###	1478	1478	1478	
10	178	655	655	655	655	210	210	1255	1262	1262	1478	1478	1478	###	1478	1478	1478	
	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	1541	###	1541	1541	1541

Within a population, members will share chromosome segments

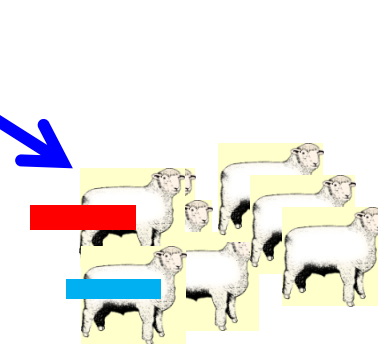
We can follow inheritance via SNPs

Degree of sharing can be represented in a genomic relationship (= observed based on SNPs)  
(similar to genetic relationship = expected based on pedigree)

# Genomic Prediction: basic idea



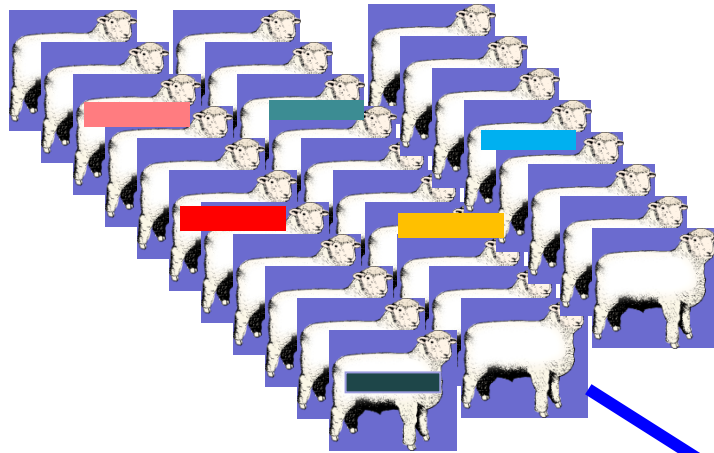
Reference population  
measured and DNA tested



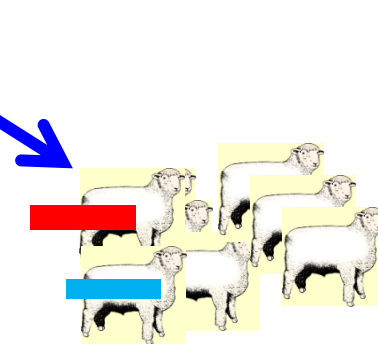
Target Set  
Only DNA tested

Small diversity of segments → more accuracy

# Genomic Prediction: basic idea



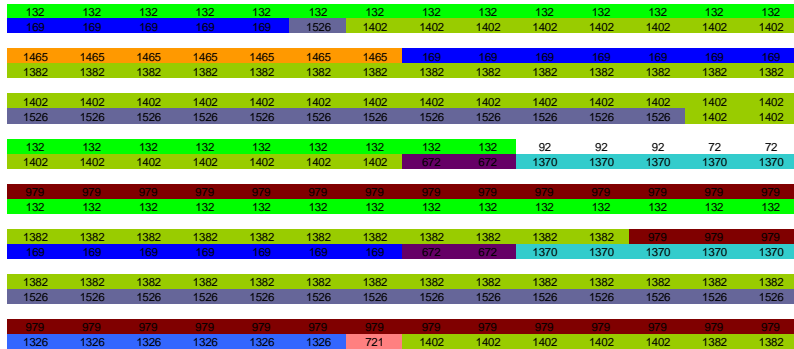
Reference population  
measured and DNA tested



Target Set  
Only DNA tested

Large diversity of segments  $\rightarrow$  less accuracy

# populations of haplotypes



Holstein Friesian, a pig/poultry nucleus

Limited diversity  
Long segment sharing

Smaller  $N_e$ , longer segment sharing, fewer “effective loci”

Merino sheep, humans

More diversity  
Short segment sharing  
Sub populations



SubPopA

SubPop B

Not only recent  $N_e$  but also historic  $N_e$  is relevant

## Design parameters for predicting GP accuracy

- Effective population size ( $N_e$ )
- Effective # chromosome segments ( $M_e$ )
- Sample size in reference data ( $n$ )
- Heritability ( $h^2$ )

# Genomic prediction accuracy Using Daetwyler et al, 2008

Accuracy<sup>2</sup> of estimating a random effect =  $n / (n + \lambda)$

$$\lambda = V_e / V_a$$

$n$  = nr obs'ns per effect

If genome exists of  $M_e$  independently segregating 'effective chromosome segments'

And each segment has variance  $V_a / M_e$ , then accuracy<sup>2</sup> of estimating each segment

$$\frac{n}{n + V_e / (V_a / M_e)} = \frac{nV_a}{nV_a + \underbrace{V_e}_{V_e \cong V_p} M_e} = \frac{h^2}{h^2 + M_e / n}$$

$$r_{g, \hat{g}} = \sqrt{\frac{h^2}{h^2 + M_e / n}}$$

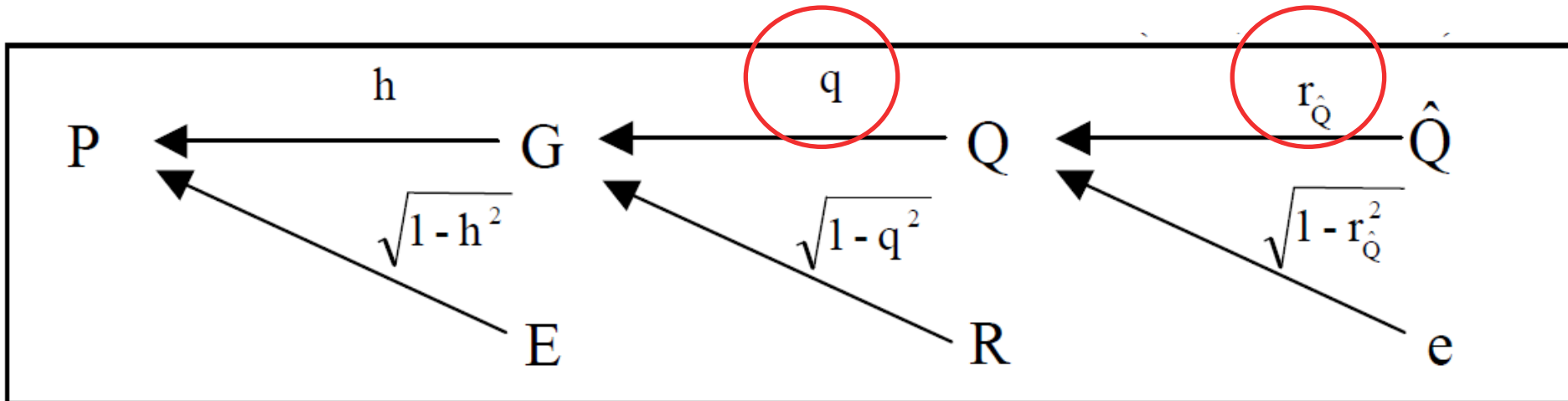
$n$  = nr observations

$M_e$  = effective nr loci

Valid if "all genetic variance is captured by markers"



# Dekkers 2007 (Path coefficient method)



Trait heritability =  $h^2$

G = total BV

Q = genetic effects captured by marker(s)

R = residual polygenic effects

Model for phenotype:  $P = G + E$

Model for BV:  $G = Q + R$

$h$ ,  $q$ , and  $r_{\hat{Q}}$  etc. are correlations

# Genomic prediction accuracy *Using Goddard et al, 2011*

Depends on

i) Proportion of genetic variance at QTL captured by markers  $q^2$

ii) Reliability of estimating marker effects  $r^2_{Qhat}$

$$\begin{aligned} \text{Accuracy} &= \sqrt{q^2 \cdot r^2_{Qhat}} \\ &= q \cdot r_{Qhat} \end{aligned}$$



# Genomic prediction accuracy Using Goddard et al, 2011

Depends on

i) Proportion of genetic variance at QTL captured by markers

$$q^2 = M / (M_e + M)$$

↳ Depends on marker-QTL LD

↳ Depends on

M = # markers

$M_e$  = 'effective number of chromosome segments'

i) Accuracy of estimating marker effects

# Genomic prediction accuracy Using Goddard et al, 2011

Depends on

- i) Proportion of genetic variance at QTL captured by markers  $q^2 = M/(M_e + M)$

↳ Depends on marker-QTL LD

↳ Depends on  $M = \# \text{ markers}$   $M_e = \text{'effective number of chromosome segments'}$

- ii) Accuracy of estimating marker effects

$$r^2_{\text{Qhat}} = V_{\text{qhat}}/V_q = n/(n + \lambda) = h^2 / (h^2 + M_e/(q^2n))$$

as  $\lambda = V_e / (q^2V_a/M_e) = M_e/(q^2h^2)$

Accuracy of genomic prediction is  
 $= \sqrt{q^2 \cdot r^2_{\text{Qhat}}}$   
 $= q \cdot r_{\text{Qhat}}$



# Comparing

*Daetwyler et al, 2008*

*Goddard et al, 2011*

## With very many markers

- i) Proportion of genetic variance at QTL captured by markers  $q^2 = M/(M_e + M)$

$$q^2 = 1$$



- i) Accuracy of estimating marker effects

$$r^2_{\text{Qhat}} = V_{\text{qhat}}/V_q = n/(n + \lambda) = h^2 / (h^2 + M_e/n)$$

$$\lambda = M_e/h^2$$

same as Daetwyler

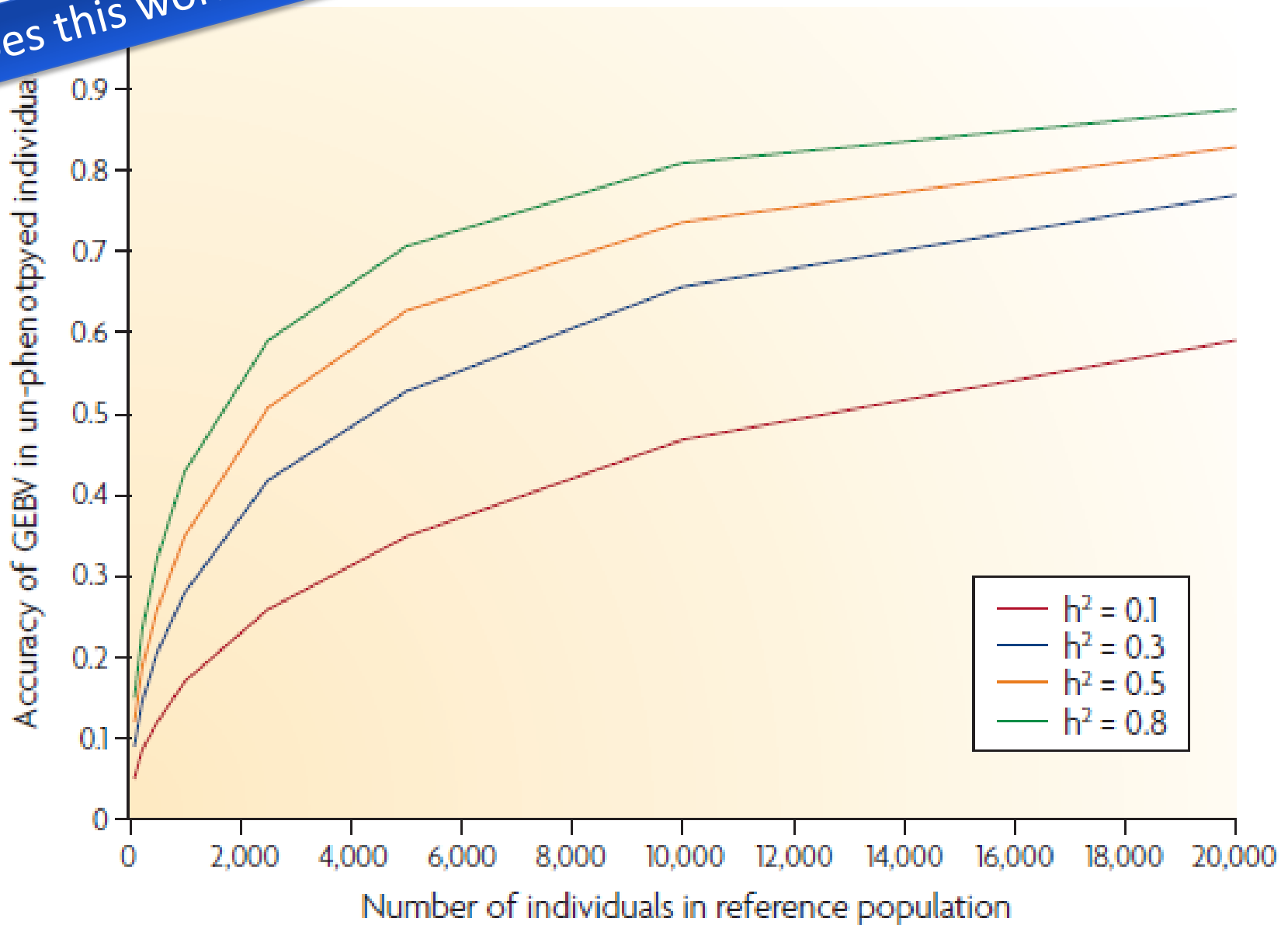
$$\text{Accuracy} = \sqrt{r^2_{\text{Qhat}}}$$

$$= r_{\text{Qhat}}$$



# Deterministic prediction

Does this work?



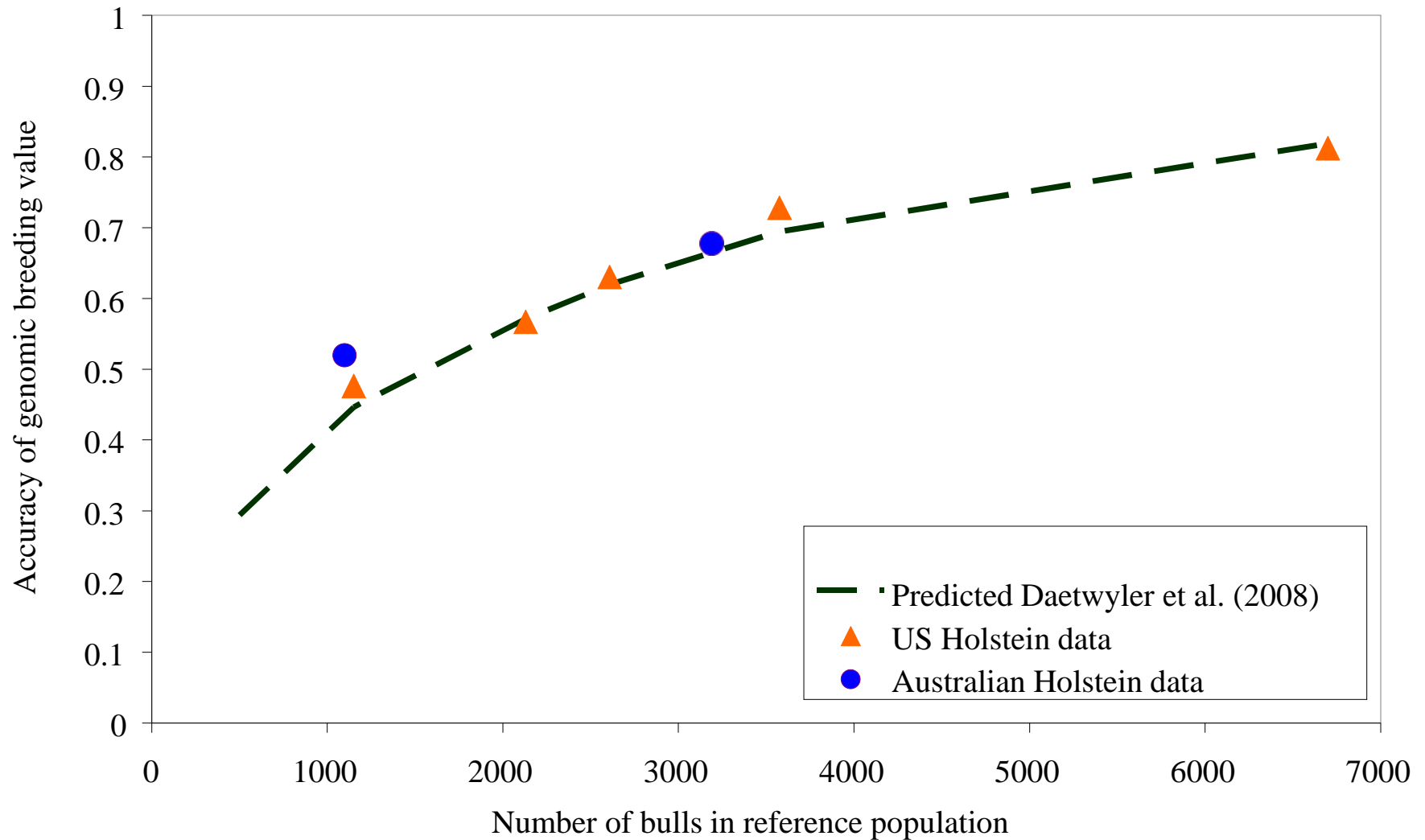
# Real Data

## ■ 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
- $h^2=0.9$  (heritability of progeny means)
- $N_e = 100$

## ■ Accuracies $r(\text{GEBV}, \text{EBV})$ in validation data sets

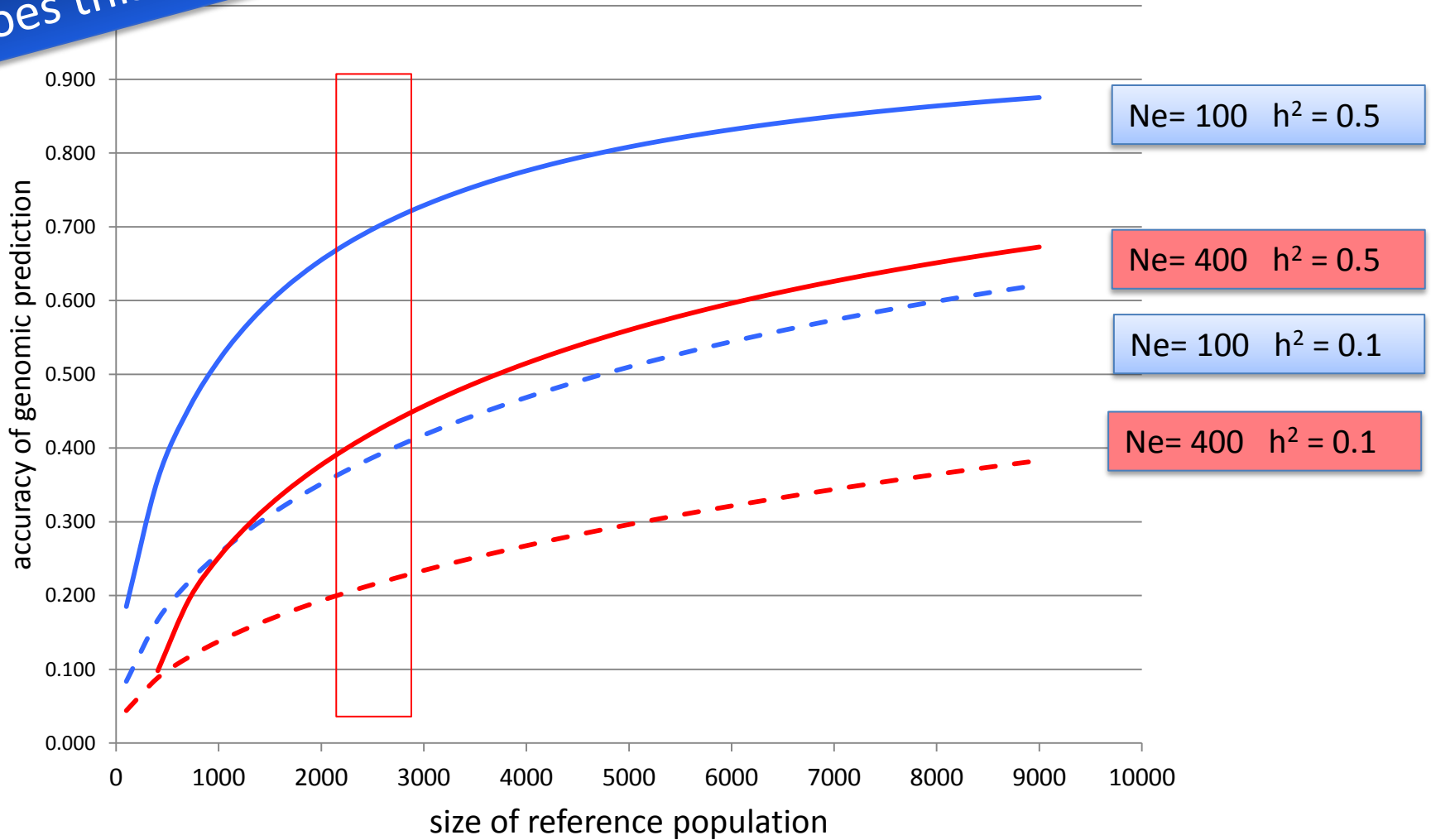
# Deterministic prediction vs. Holstein data





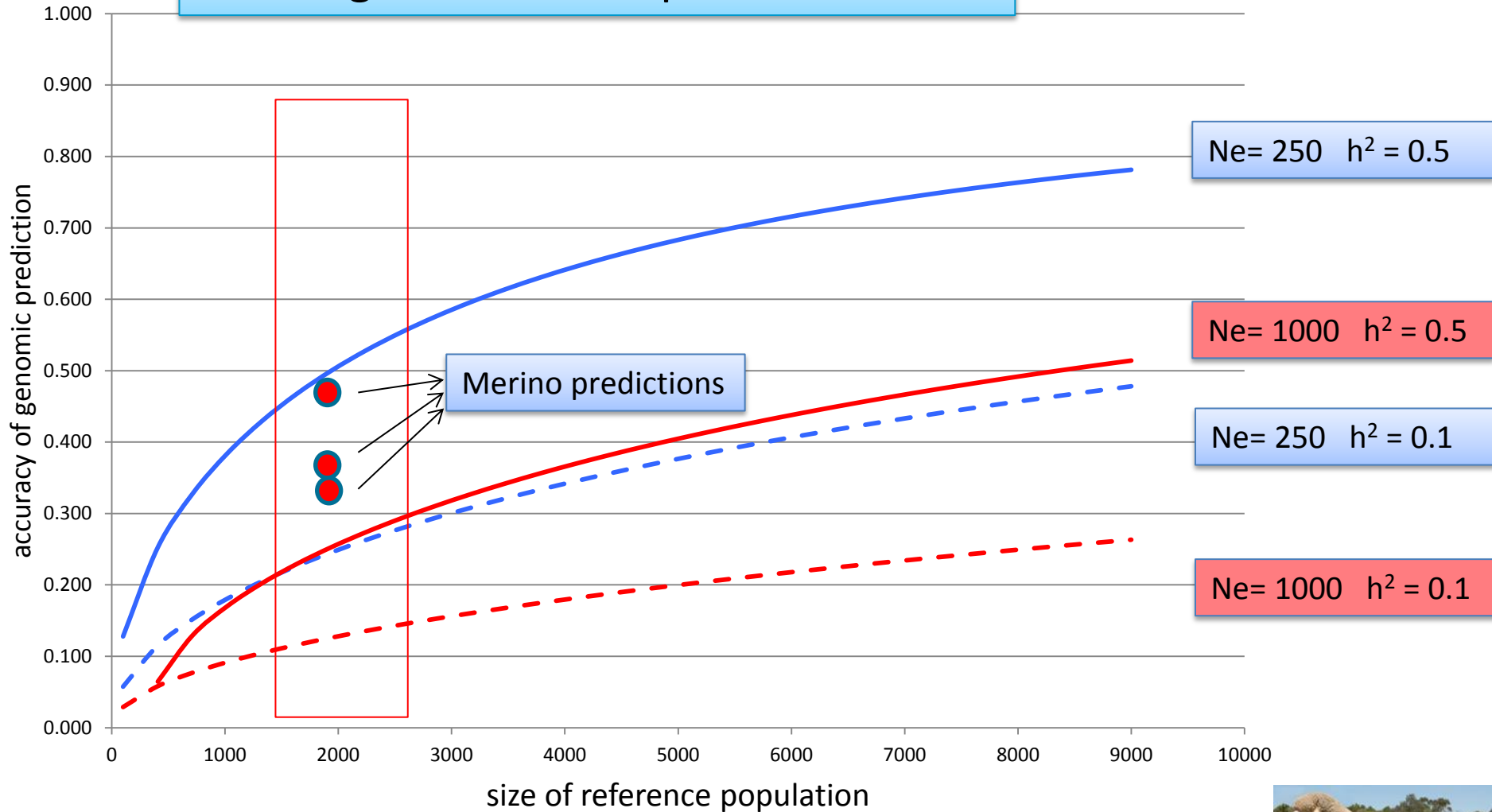
# Genomic prediction accuracy Using Goddard et al, 2011

Does this work?



# Genomic prediction accuracy Using Goddard et al, 2011

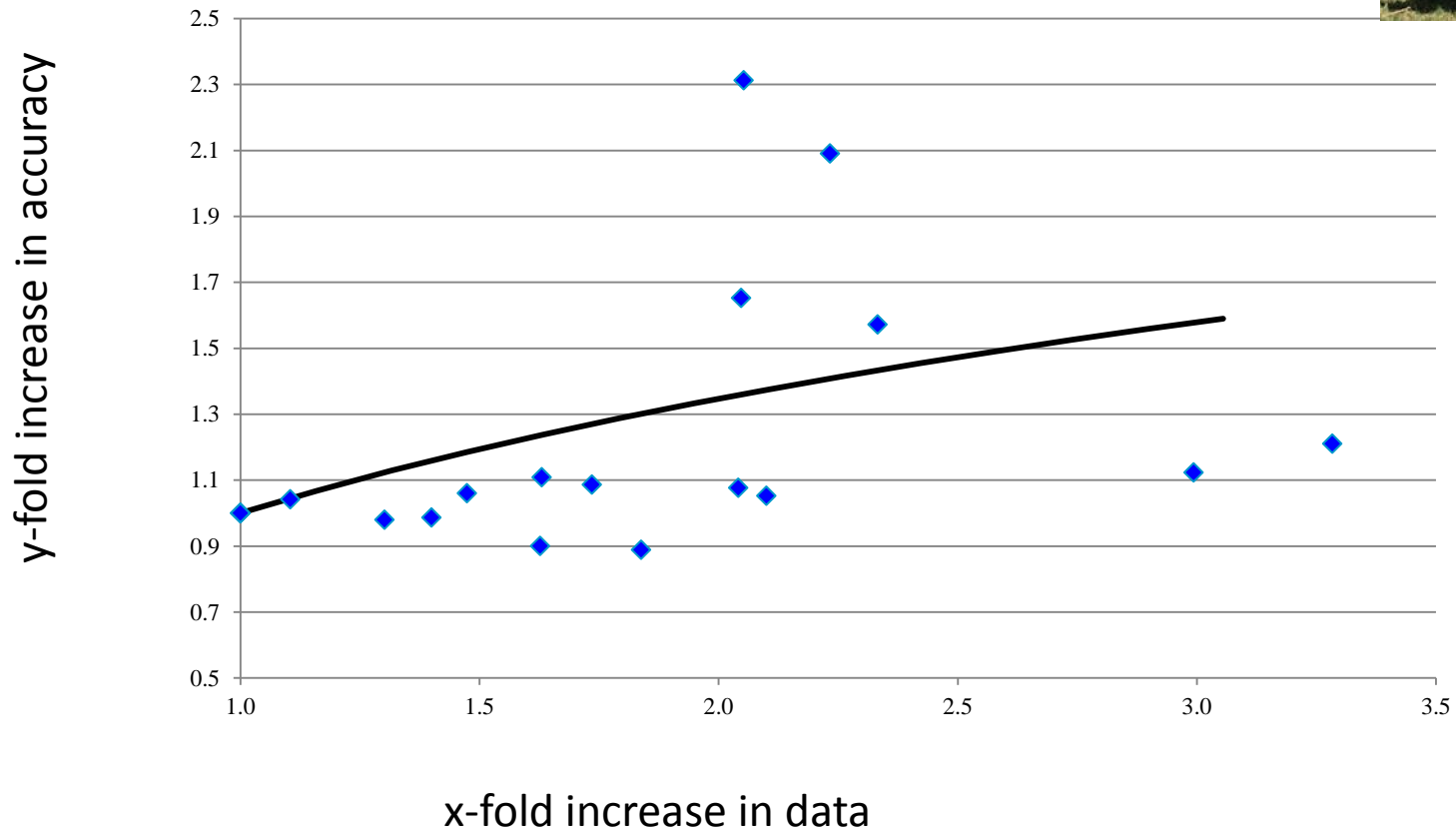
Did we get what we expected?



# Validating 'Genomic Prediction Accuracy'

More data is always good

But does it increase accuracy as expected?



# What effective population size?

*Kijas et al 2012*

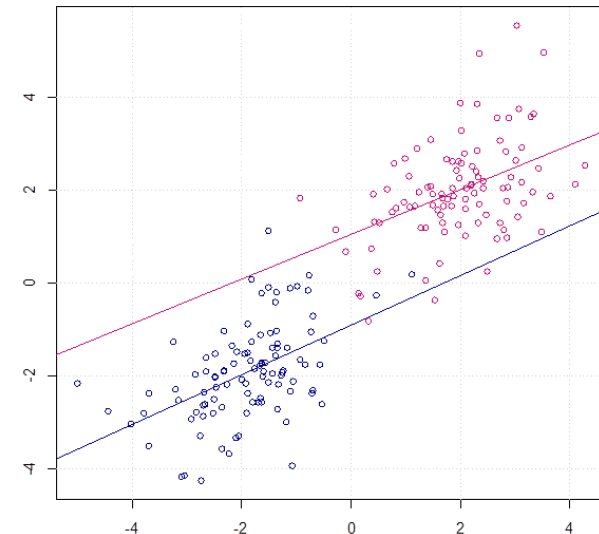
- Sampling?



Populations not homogeneous.

Within and between breed/line accuracies

Some accuracy due to population structure



# Summary so far

- Theory exists to predict genomic prediction accuracy in advance: depends on nr. effective segments, nr records
- Relies on assumptions regarding effective population size
  - Theory assumes everyone in the population is equally unrelated
  - Some (unclear) theory about effective nr of loci
- Ignores heterogeneity of populations and relationships
- *We observe more initial acc and less increase with more data*



# How to derive the effective number of loci?

$$M_e$$

$$y = Wu + e,$$

$u$  = QTL effects,

$W$  = genotypes

$G = WW'$

$G$  = covariance matrix among marker genotypes:  $r^2 = 1 / (1 + 4N_e \times c)$

$$\begin{array}{ccccc}
 1 & r^2_{1,2} & \cdots & r^2_{1,M-1} & r^2_{1,M} \\
 r^2_{2,1} & 1 & \cdots & r^2_{2,M-1} & r^2_{2,M} \\
 \vdots & \vdots & \ddots & \vdots & \vdots \\
 r^2_{M-1,1} & r^2_{M-1,2} & \cdots & 1 & r^2_{M-1,M} \\
 r^2_{M,1} & r^2_{M,2} & \cdots & r^2_{M,M-1} & 1
 \end{array}$$

Need var(G)  
 mean  $r^2$  of all  
 these elements  
 $\rightarrow$  integrate

$$\text{Var}(G) = 1/M_e$$

# How to derive the effective number of loci?

$M_e$  is a (almost linear) function of  $N_e$  and genome size

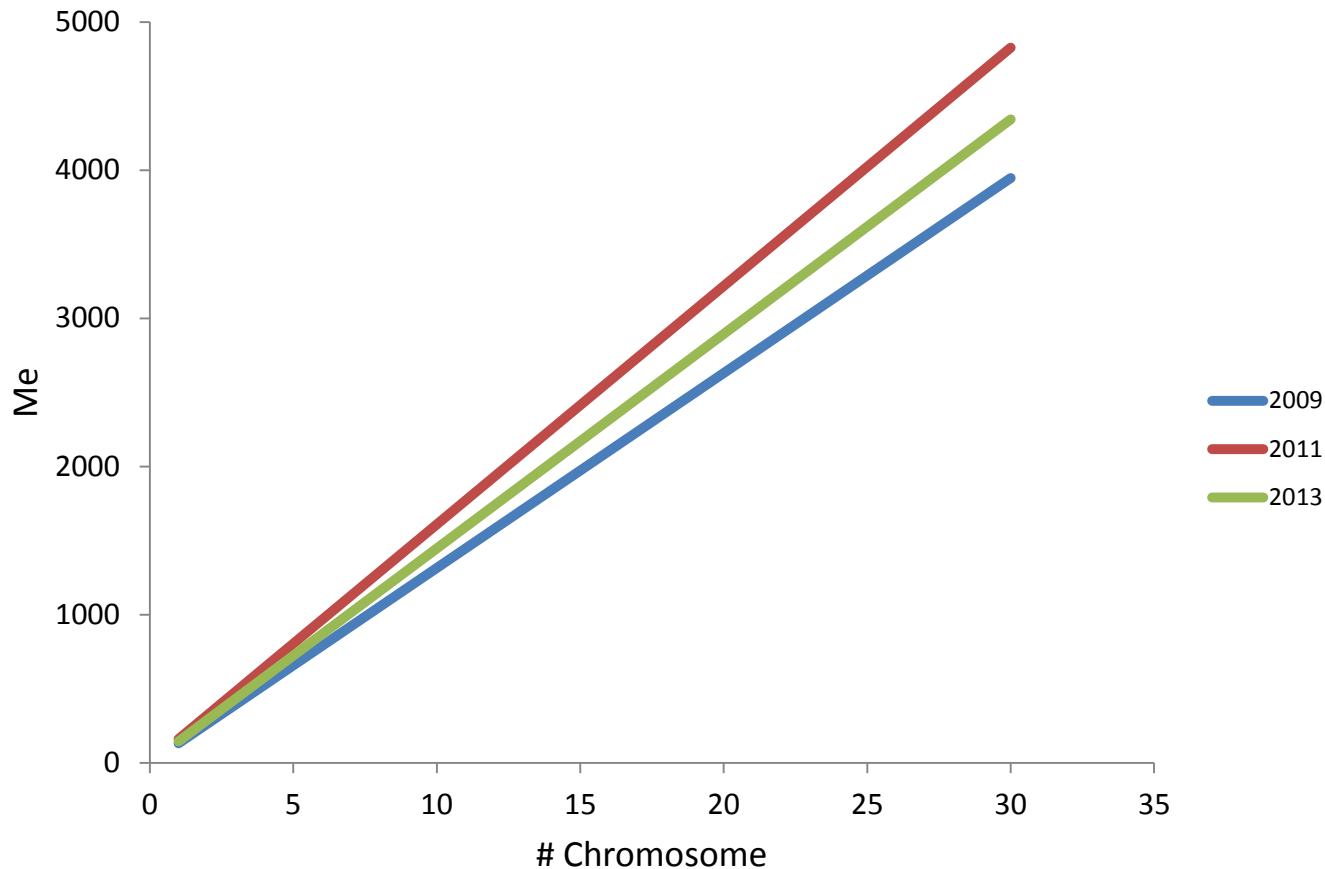
- $M_e = 2N_eLN_{chr} / \ln(4N_eL)$  (Goddard 2009)
- $M_e = 2N_eLN_{chr} / \ln(N_eL)$  (Goddard et al. 2011)
- $M_e = 2N_eLN_{chr} / \ln(2N_e)$  (Meuwissen et al. 2013)

Different assumptions about multiple chromosomes etc.

$N_e$  = Effective pop.size  
L = Av. Length per chrom (1M)  
 $N_{chr}$  = 30

Hard to 'know'  $N_e$

# Difference among the formulas



Example:  $N_e = 500$ ,  $L=1M$      $h^2 = 0.5$  and  $n = 5000$ ,  
→ accuracy = 0.62, 0.58, 0.60



# Validating 'Effective number of segments'

empirically

Can use actual data on A and G to test this

Compare G and A matrices  $G - A = D + E$

Note, this is different from Goddard et al:  $\text{var}(D) = 1/M_e$  assumed  $A = I$

D = deviation in relationship at QTL

$$\text{Var}(G) = 1/M_e$$

$$M_e = 1/\text{var}(G_{ij})$$

E = error

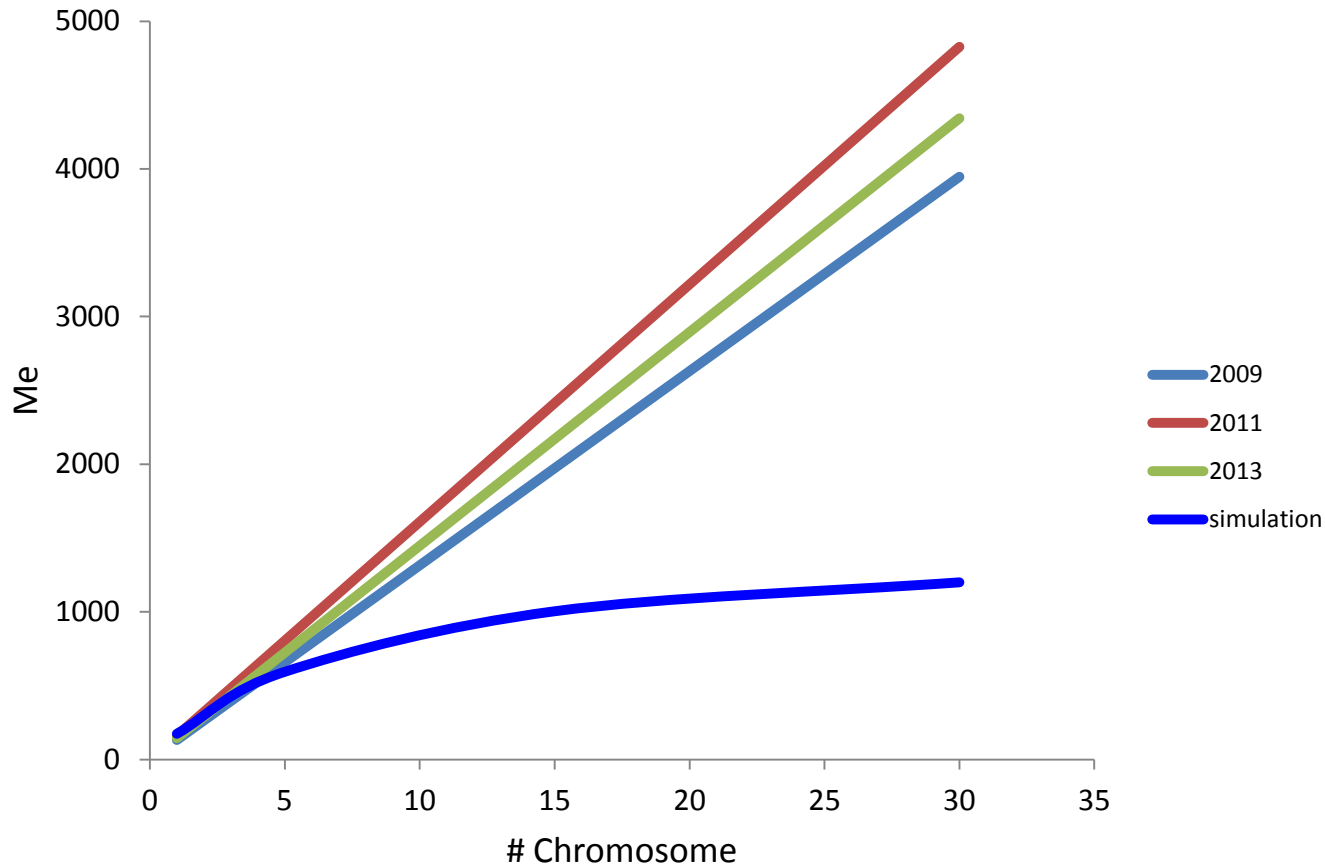
$$\text{Var}(E) = 1/nr \text{ Markers}$$

Given genomic relationships (after collecting data), it is possible to empirically get  $M_e$  from the data

# Simulation

- Coalescence gene dropping
  - $N_e = 500$  for 500 generations
  - $L = 1$  Morgan
  - $N_{chr} = 30$
  - Recombination according to  $L$
  - Mutation rate =  $10E-08$
  - $n = 3000$  in the last generation
- Estimate  $G_{ij}$  and obtain empirical  $M_e$

# Difference from empirical $M_e$



$h^2 = 0.5$  and  $n = 5000$ ,

accuracy = 0.62, 0.58, 0.60 vs. 0.82 (simulation)

# Revisit the theory

$$M_e = \frac{N_{chr}}{[\ln(4N_e L + 1) + 4N_e L(\ln(4N_e L + 1) - 1)] / (8N_e^2 L^2) + (1/3N_e) \times (N_{chr} - 1)}$$

Assuming LD  $r^2 = 1 / (1 + 4N_e \times c)$

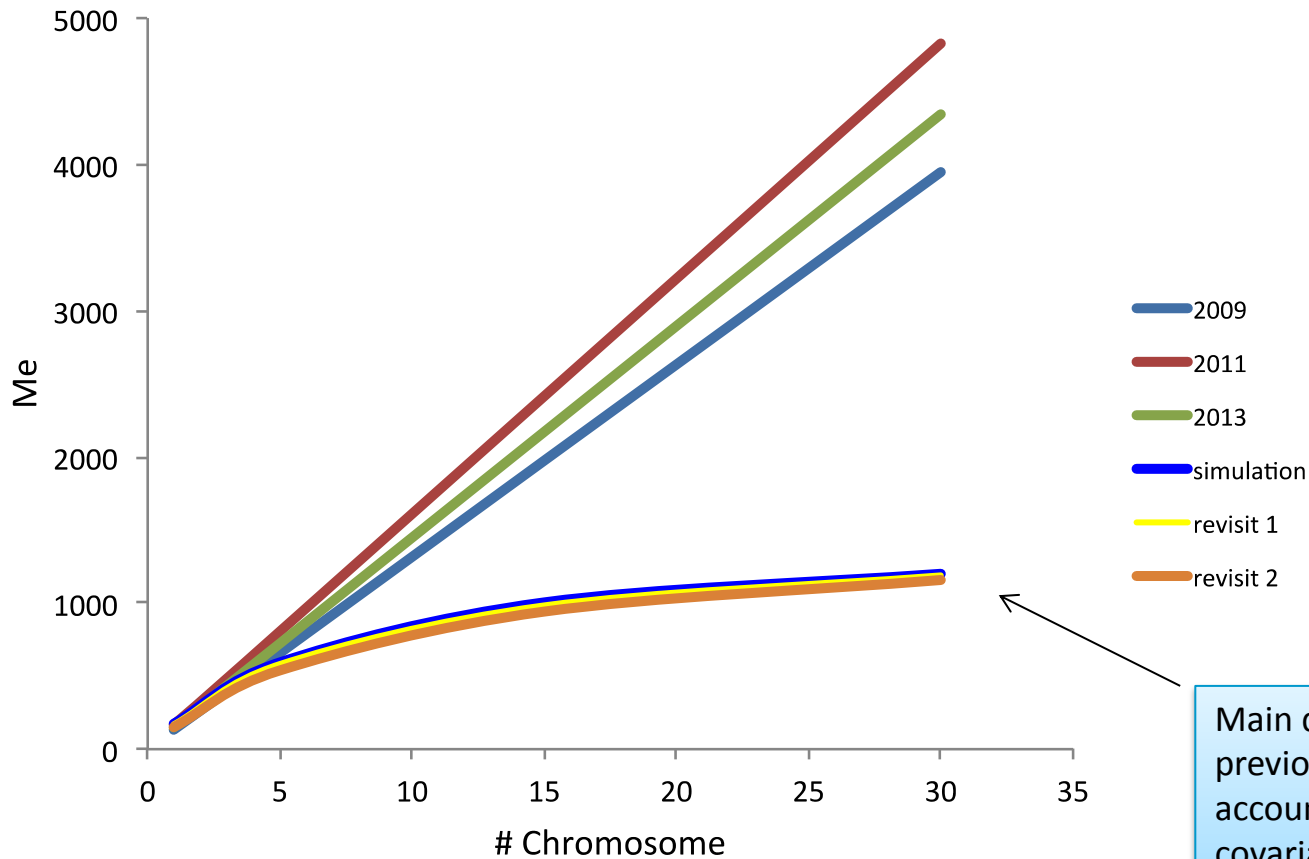
$$M_e = \frac{N_{chr}}{[\ln(2N_e L + 1) + 2N_e L(\ln(2N_e L + 1) - 1)] / (4N_e^2 L^2) + (1/3N_e) \times (N_{chr} - 1)}$$

Assuming LD  $r^2 = 1 / (2 + 4N_e \times c)$

*For more detail,*

*see Lee et al, 2017 Scientific Reports*

# Empirical $M_e$ and new formula

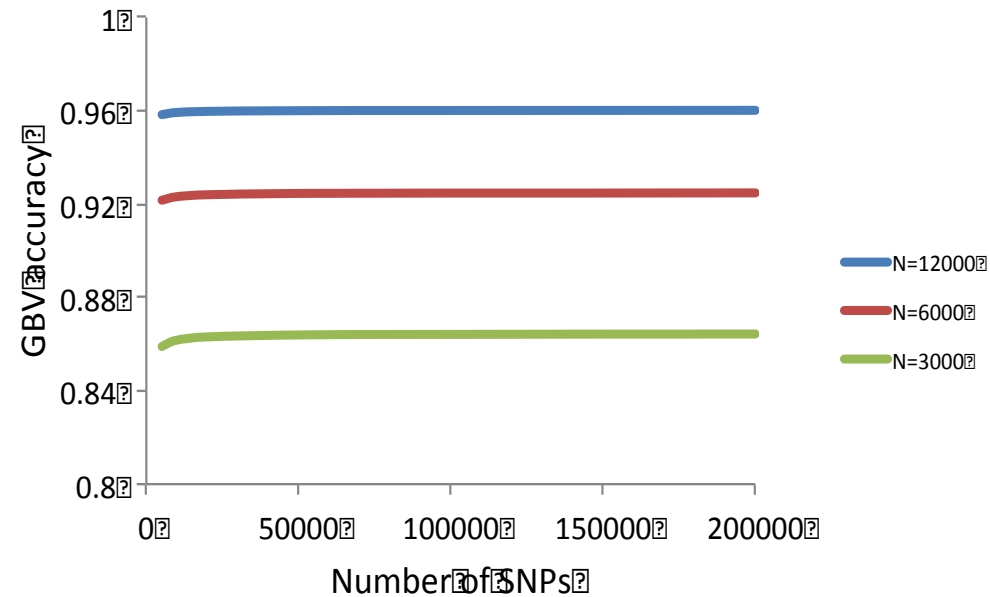
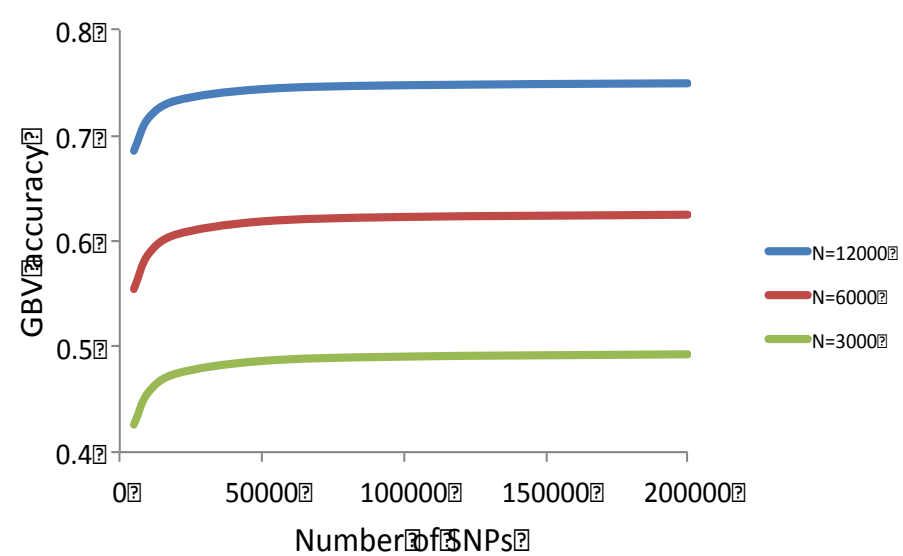


Main difference with previous work is due to accounting for covariance between chromosomes

■ Agreed well

# Genomic prediction accuracy

## Effect of marker density



$N_e = 1,000$

$N_e = 100$

Expect very little improvement with denser markers

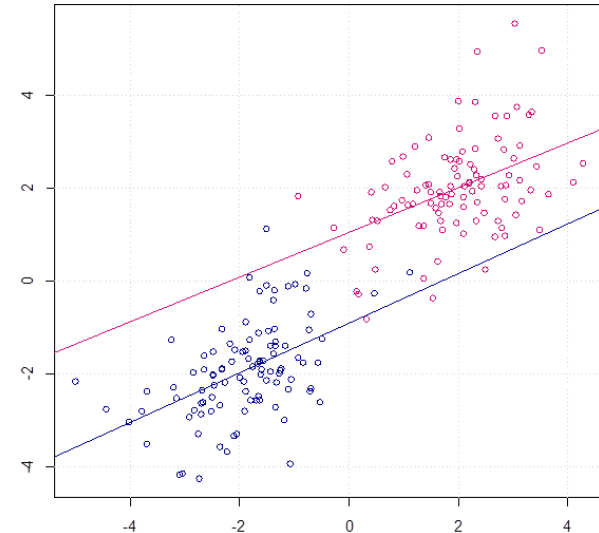
# What effective population size?

<i>Holstein Friesian</i>	$< 100$
<i>Merino Sheep</i>	$\sim 1000 ?$

Populations not homogeneous.

Within and between breed/line accuracies

Some accuracy due to population structure



# How do we validate accuracy?

- Validation population
  - EBV (based on progeny test)
  - Phenotype
  - Is it a homogeneous group?
    - Subpopulations
    - Different cohorts with genetic trend
    - Are there direct relatives between training and validation?
- Cross-validation
  - Across families
  - Random(also within families)
  - Across or within genetic groups (subpopulations)?



# Main question

- How many records are needed in the reference population to achieve a certain accuracy?

But some important sub questions:

- What if you are more related to the reference?
- the value of closer relatives (e.g.own herd) versus the '*general*' reference population

# Relationship with reference population

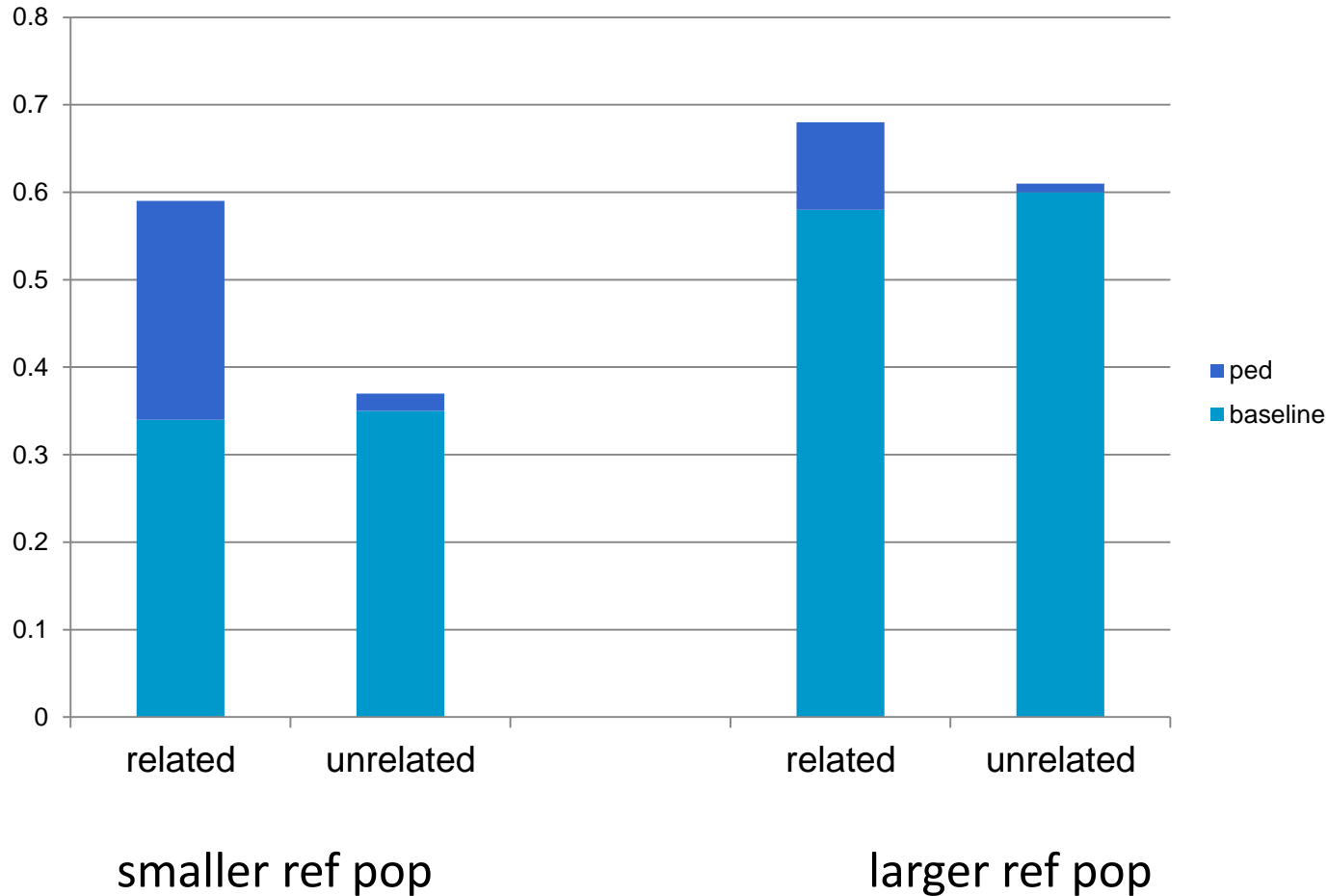
Clark et al 2011

<b>Method</b>	<b>Close</b> Ped 0 - 0.25 Genom 0.08 – 0.35	<b>Distant</b> 0 - 0.125 0.08 – 0.26	<b>Unrelated</b> 0 - 0.05 0.08 – 0.16
BLUP- Shallow pedigree	0.39	0.00	0.00
BLUP- Deep Pedigree	0.42	0.21	0.04
gBLUP	<b>0.57</b>	0.41	<b>0.34</b>

Additional accuracy from family info

'baseline accuracy': graphs predict 0.36  
for  $N_e=100$ ,  $N=1750$ ,  $h^2=0.3$

# Relatedness matters more if the reference population is smaller



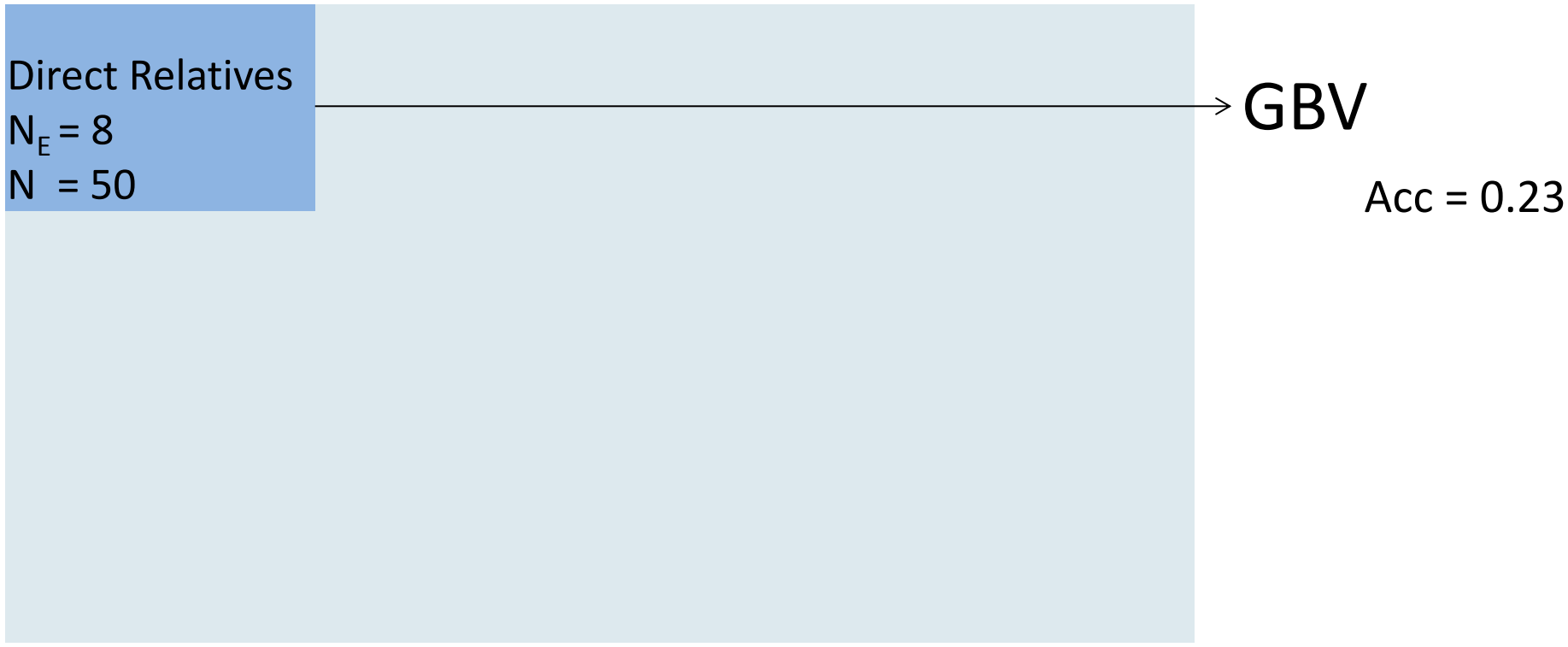
(hypothesis)

# A reference population may have relatives

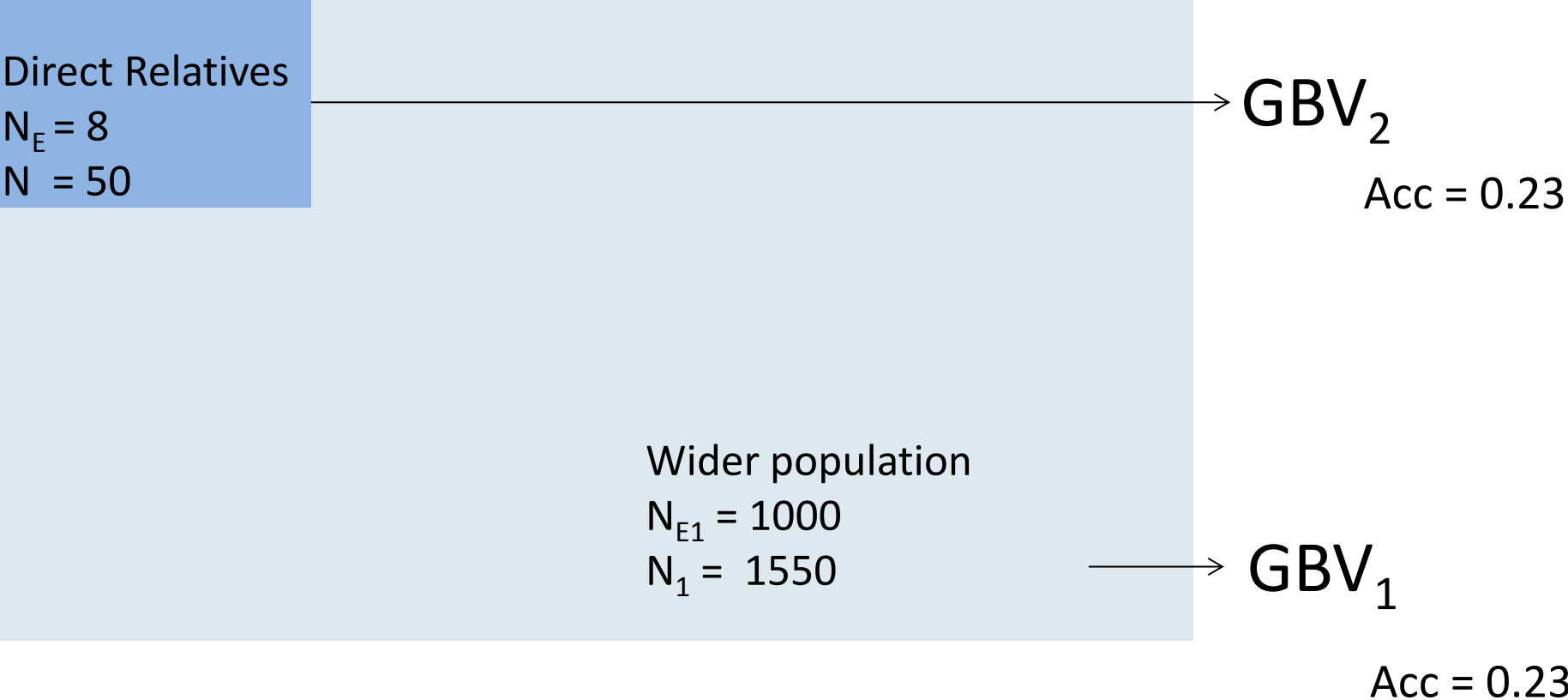


# 'Relatedness' can be represented by effective size

*Hayes et al 2009*



# Information from different subsets can be combined



Calculate overall accuracy using selection index

$$GBV = \sum b_i GBV_i \quad \text{Acc} = 0.31$$

# Using a stratified reference population

-populations are not homogeneous



# Using a stratified reference population

-populations are not homogeneous

Direct Relatives

$$N_{E3} = 8$$

$$N_3 = 50$$

Own Herd

$$N_{E2} = 50$$

$$N_2 = 400$$

Wider population

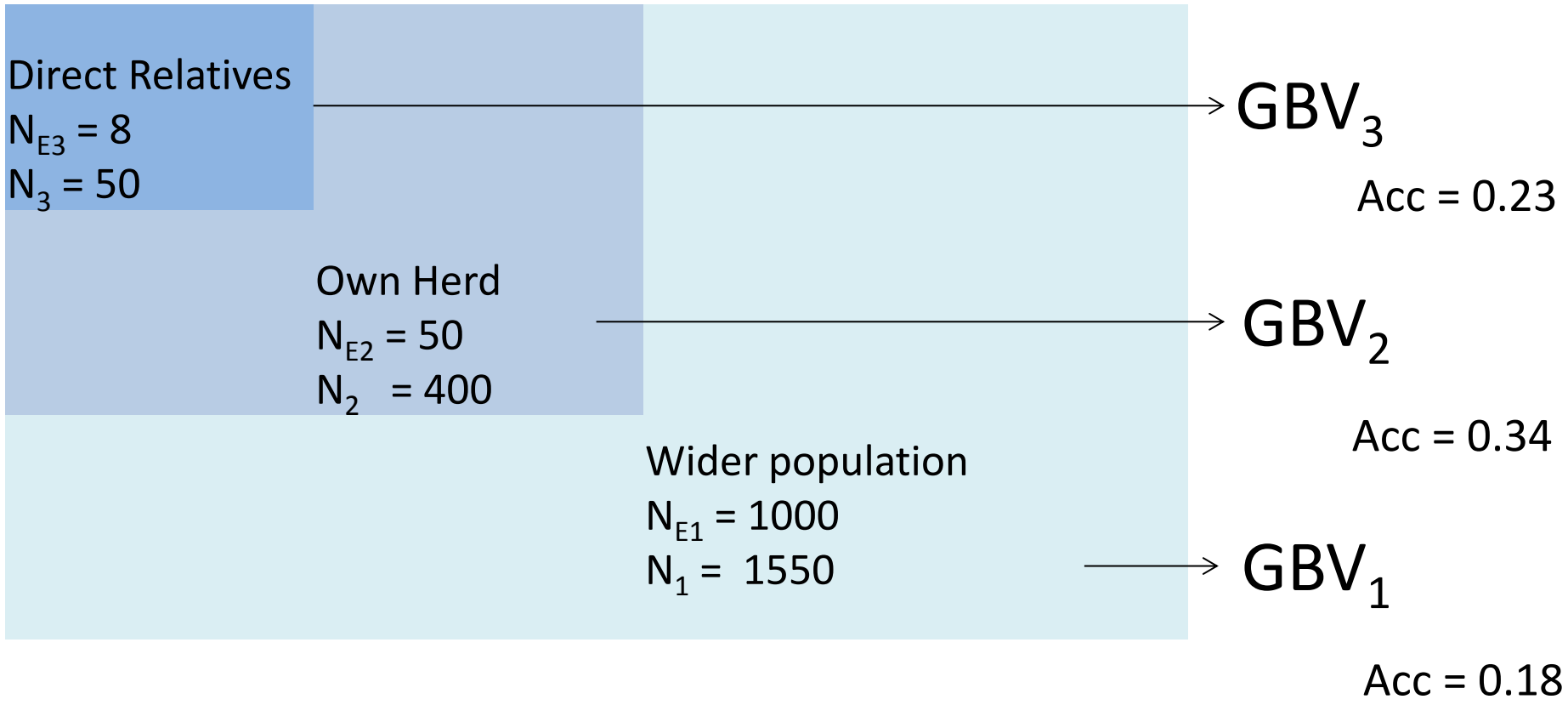
$$N_{E1} = 1000$$

$$N_1 = 1550$$



# Using a stratified reference population

-populations are not homogeneous

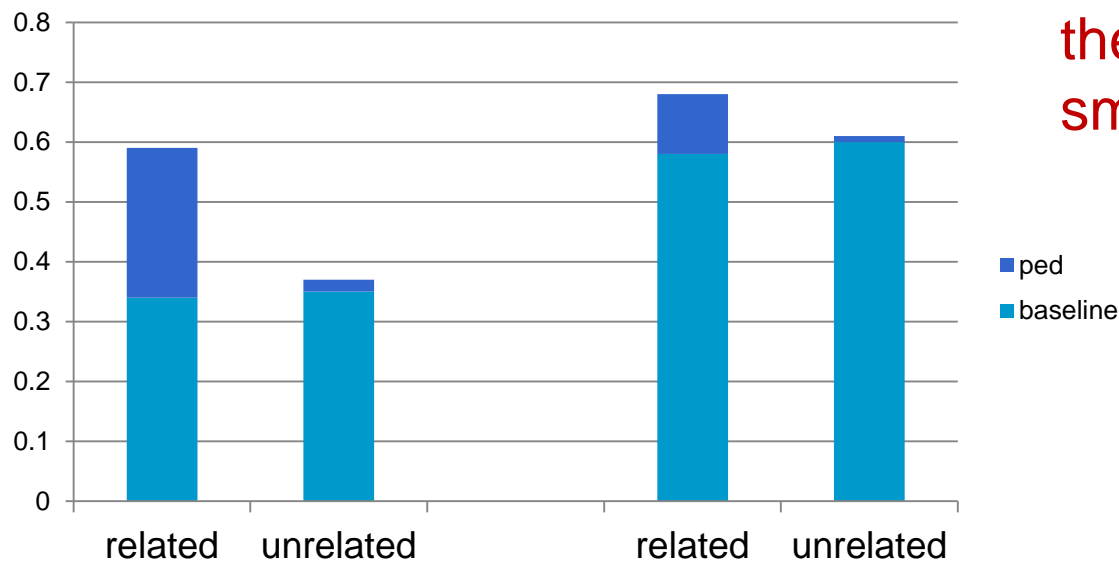


Calculate overall accuracy using selection index

$$GBV = \sum b_i GBV_i \quad Acc = 0.42$$

$$NE_1 = 1000$$

$N_1$	Value of information source			GBV accuracy		diff
	breed (N1)	flock (400)	relatives (50)	all info	breed only	
2,000	16%	52%	21%	0.43	0.22	95%
5,000	31%	39%	15%	0.47	0.32	48%
10,000	45%	26%	10%	0.53	0.42	26%



Relatedness matters more if the reference population is smaller

hypothesis confirmed

$$NE_1 = 1000$$

	Value of information source			GBV accuracy		
$N_1$	breed (N1)	flock	relatives	all info	breed only	diff
		400	50			
2,000	16%	52%	21%	0.43	0.22	95%
5,000	31%	39%	15%	0.47	0.32	48%
10,000	45%	26%	10%	0.53	0.42	26%
$N_1$	breed (N1)	flock	relatives	all info	breed only	diff
		100	10			
2,000	48%	36%	48%	0.28	0.21	36%
5,000	68%	19%	68%	0.36	0.31	15%
10,000	79%	11%	79%	0.45	0.41	7 %

With fewer relatives the reliance on the reference population increases

$NE_1 = 1000$

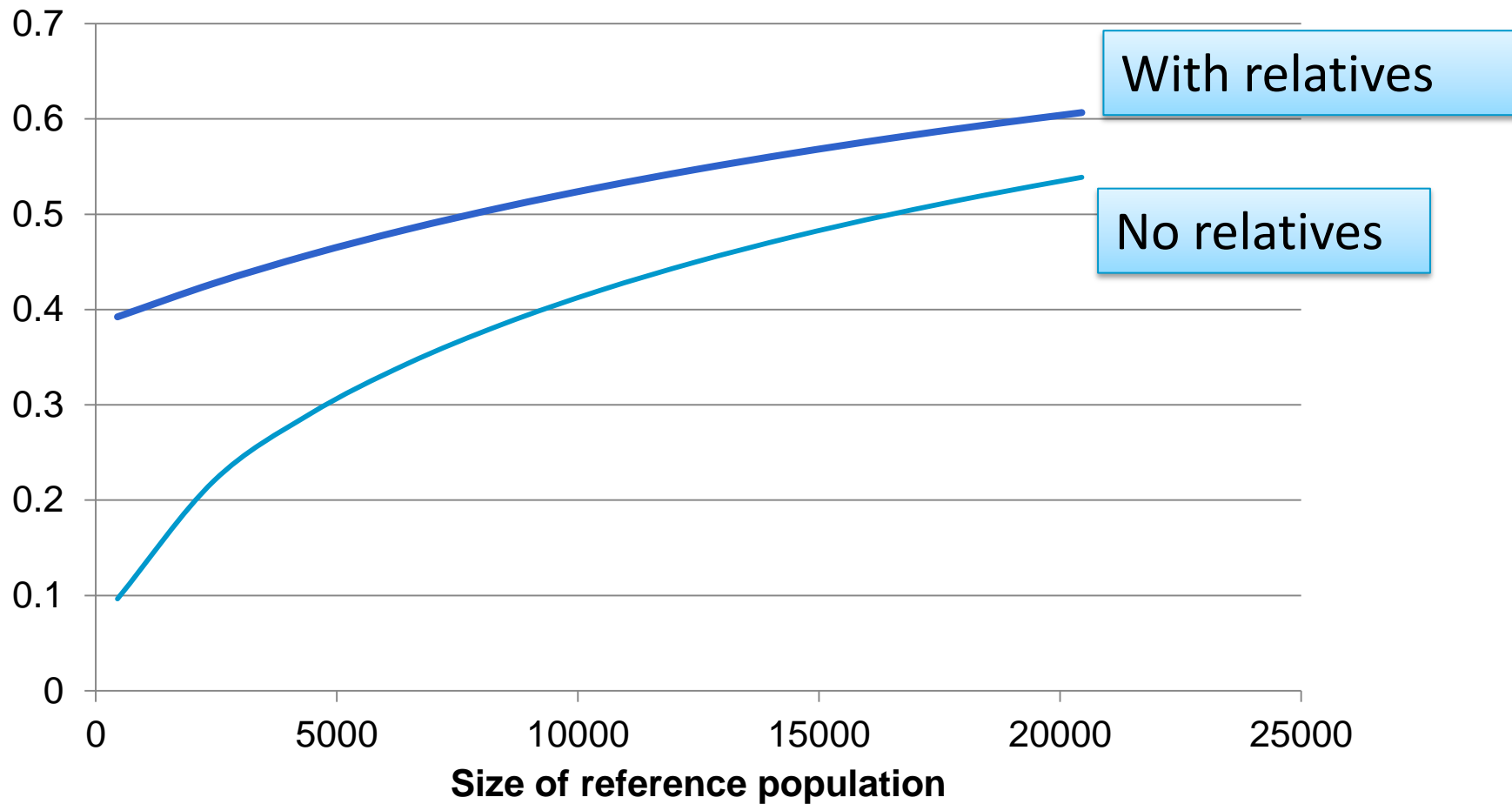
	Value of information source			GBV accuracy		
$N_1$	breed (N1)	flock (400)	relatives (50)	all info	breed only	diff
2,000	16%	52%	21%	0.43	0.22	95%
5,000	31%	39%	15%	0.47	0.32	48%
10,000	45%	26%	10%	0.53	0.42	26%

$NE_1 = 200$

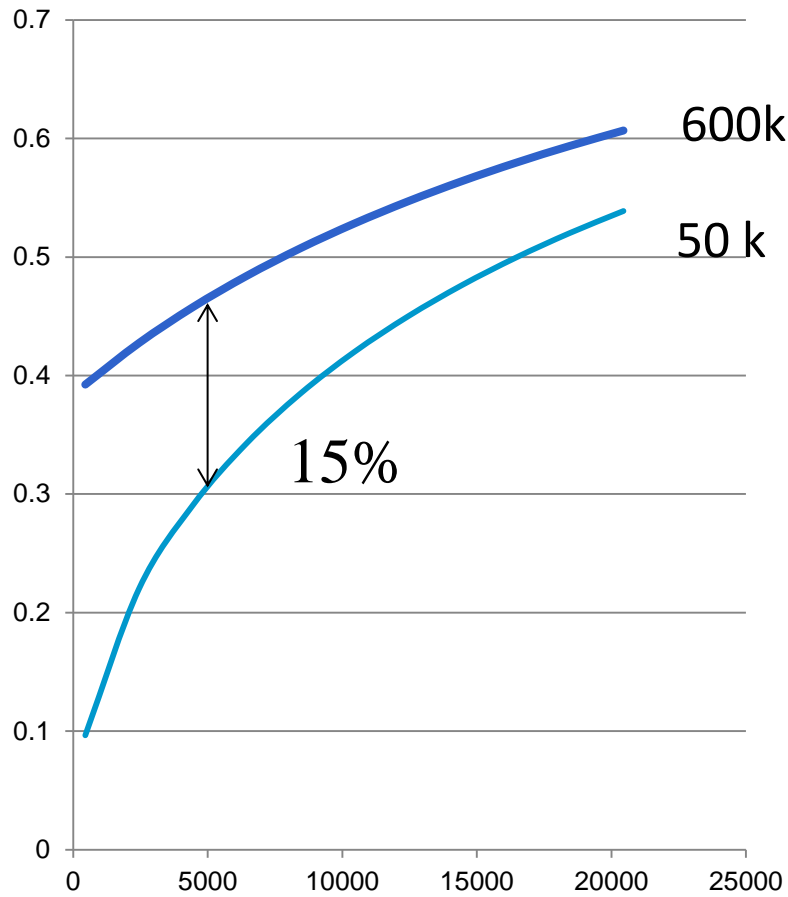
$N_1$	breed (N1)	flock (400)	relatives (50)	all info	breed only	diff
2,000	45%	26%	10%	0.53	0.45	18%
5,000	62%	12%	5%	0.64	0.60	7%
10,000	72%	5%	2%	0.74	0.72	3%

With less diverse populations the relatives matter a lot less

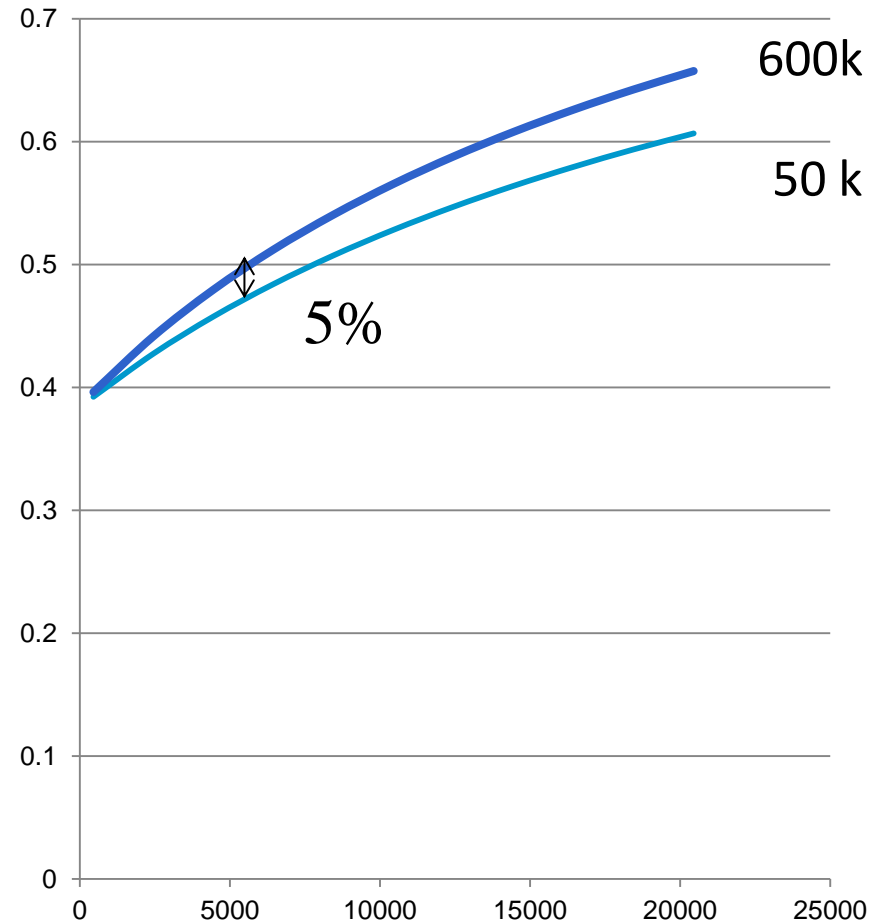
# The effect of a larger reference population



# The effect of denser marker panels



No relatives



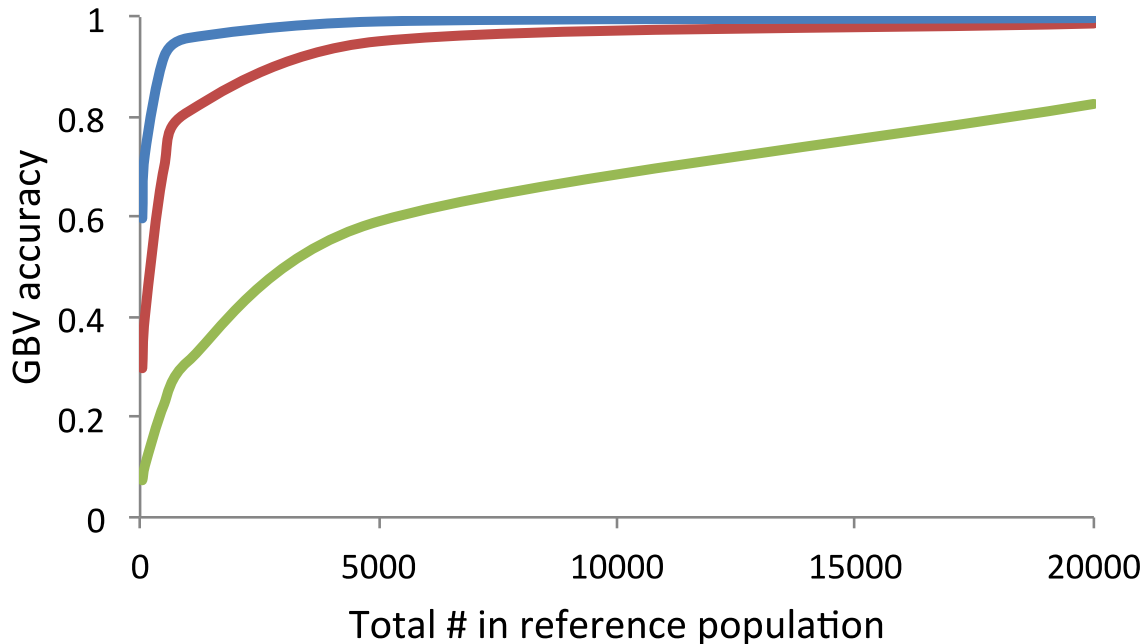
With relatives

# Conclusion

- Theory exists to predict genomic prediction accuracy in advance: depends on population diversity, nr records
- Reference populations are heterogeneous, with closer as well as distant relatives
- Relatives and flock/herd mates will increase accuracy and decrease reliance on wider reference population (and denser marker panels)



# Sample availability

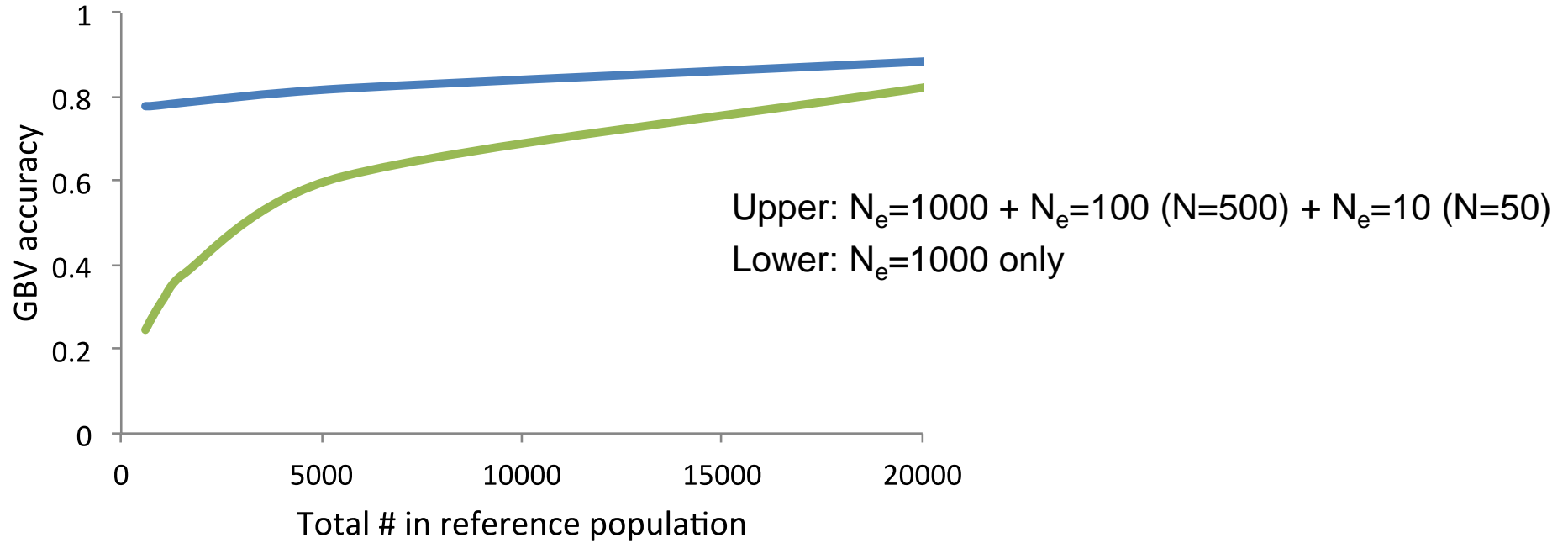


Upper:  $N_e=10$  only  
Middle:  $N_e=100$  only  
Lower:  $N_e=1000$  only

- $h^2=0.25$
- $N_e=10$  would have  $< N = 100$  (maximum acc. = 0.73)
- $N_e=100$  would have  $< N = 1,000$  (maximum acc. = 0.81)
- $N_e=1,000$  can have  $N = 20,000$  (acc. = 0.83)



# Composite design



# Implication

- Marker density
  - For beef cattle or sheep, very dense markers (e.g. 600K) may not be cost-effective, compared to 50K
  - For  $N_e = 1000$ , accuracy is similar between 50K and 600K
- Marker density is not a critical design parameter
  - when  $> 50K$  with  $N_e = 1000$  (livestock)
  - when  $> 200K$  with  $N_e = 10,000$  (human)
- But, it may matter with very large  $N_e$ 
  - Multi-breeds or multi-ethnicities

This theory also assumes an ‘infinitesimal model’

# Implication

- To maximise prediction accuracy
  - give a priority to genotype reference sample of smaller  $N_e$ ,
  - e.g. close relatives > flocks (local, village) > states > country > ...
  - When  $h^2$  is lower, reference sample of smaller  $N_e$  is more important

Note that  $N_e$  can be changed, depending on the target sample

# Implication

## ■ MTG2

<https://sites.google.com/site/honglee0707/mtg2>

Given design parameters, MTG2 can provide the expected accuracy and power

See section 7 and 9 in the manual