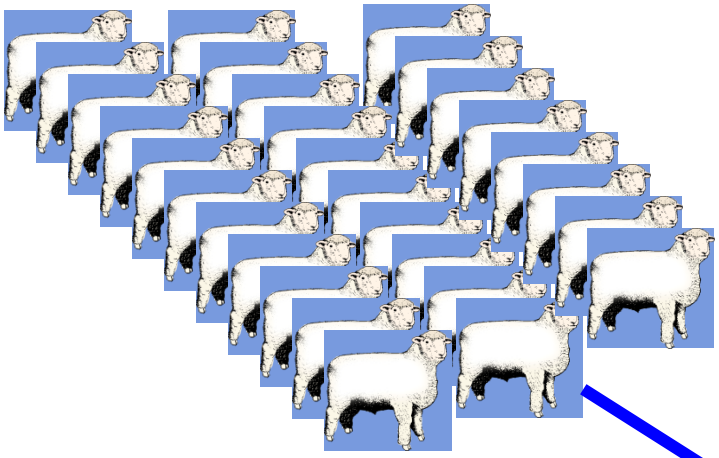
Accuracy of Genomic Prediction

Julius van der Werf

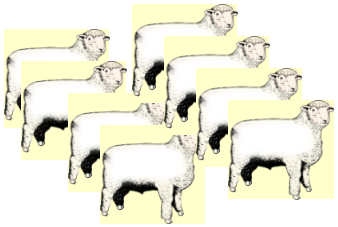
une
University of
New England

SHEEP CRC
An Australian Government Initiative
CRC AUSTRALIA

Genomic Prediction: basic idea



Reference population
measured and DNA tested



Young rams
Only DNA tested

To predict a trait EBV at a young age,

good for for: late traits
hard to measure traits

Genomic prediction accuracy

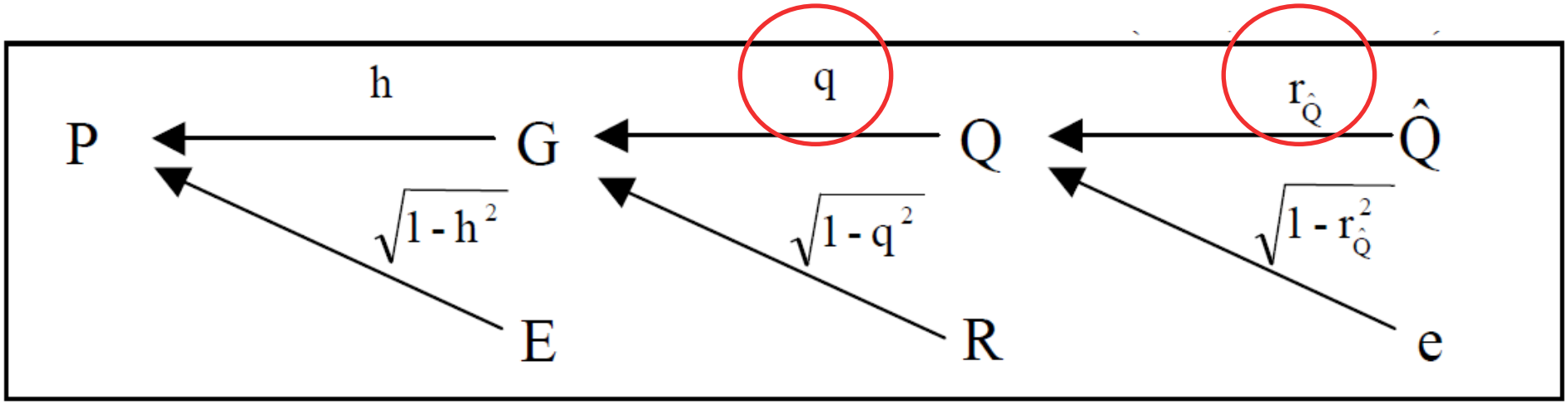
- Derive from the model, e.g. PEV from GBLUP mixed model equations
- Validate with other EBVs or phenotypes
  - Validation population
  - Cross-validation
- Predict in advance based on theory and assumptions about population

Genomic prediction accuracy *Using Goddard et al, 2011*

Depends on

i) Proportion of genetic variance at QTL captured by markers

i) Accuracy of estimating marker effects



Trait heritability = h^2

G = total BV

Q = genetic effects captured by marker(s)

R = residual polygenic effects

After Goddard et al. (2011, JABG 128); notation after Dekkers (2007, JABG 124)

Model for phenotype: $P = G + E$

Model for BV: $G = Q + R$

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 = # markers

M_e = 'effective number of chromosome segments'

i) Accuracy of estimating marker effects

$$r^2_{Qhat} = V_{qhat}/V_q = N/(N+\lambda)$$

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

$$\text{Accuracy} = \sqrt{(q^2.\, r^2_{Qhat})}$$

$$= q.\, r_{Qhat}$$

Genomic prediction accuracy *Using Goddard et al, 2011*

Depends on

i) Proportion of genetic variance at QTL captured by markers

$b = M/(M_e + M)$

↳ Depends on marker-QTL LD

↳ Depends on

M = # markers

M_e = 'effective number of chromosome segments'

$M_e = 2N_eLk/\ln(2N_e)$

or is it…?

i) Accuracy of estimating marker effects

$V_{qhat}/V_q = N/(N+\lambda)$

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

Accuracy = $\sqrt{\ }$ $b.V_{qhat}/V_q$

Effective number of chromosome segments

Sample size 2000
Heritability 0.05
Number of chromosome 5
Length of the chromosome 1 Morgan
Replicates 100

$M_e = 2N_eLk/\ln(2N_e)$ or is it…?

Ne (=number of generations)	100	1000	5000	Infinity
	number of QTL = 50000			
average	0.556	0.279	0.148	0.045
SD	0.055	0.042	0.032	
Me	223	1184	4465	50000
	Mike's theory			
4NeLk	2000	20000	100000	
2NeLk/log(4NeL)	303	2325	10000	
2NeLk	1000	10000	50000	
2NeLk	1000	10000	50000	
2NeLk/log(NeL)	371	2703	11369	
2NeLk/log($2N_e$)	435	3029	12500	
2NeLk/ln(NeL)	217	1448	5870	
2NeLk/ln(2Ne)	189	1316	5429	

Validating ‘Effective number of segments’

Can use actual data on A and G to test this

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

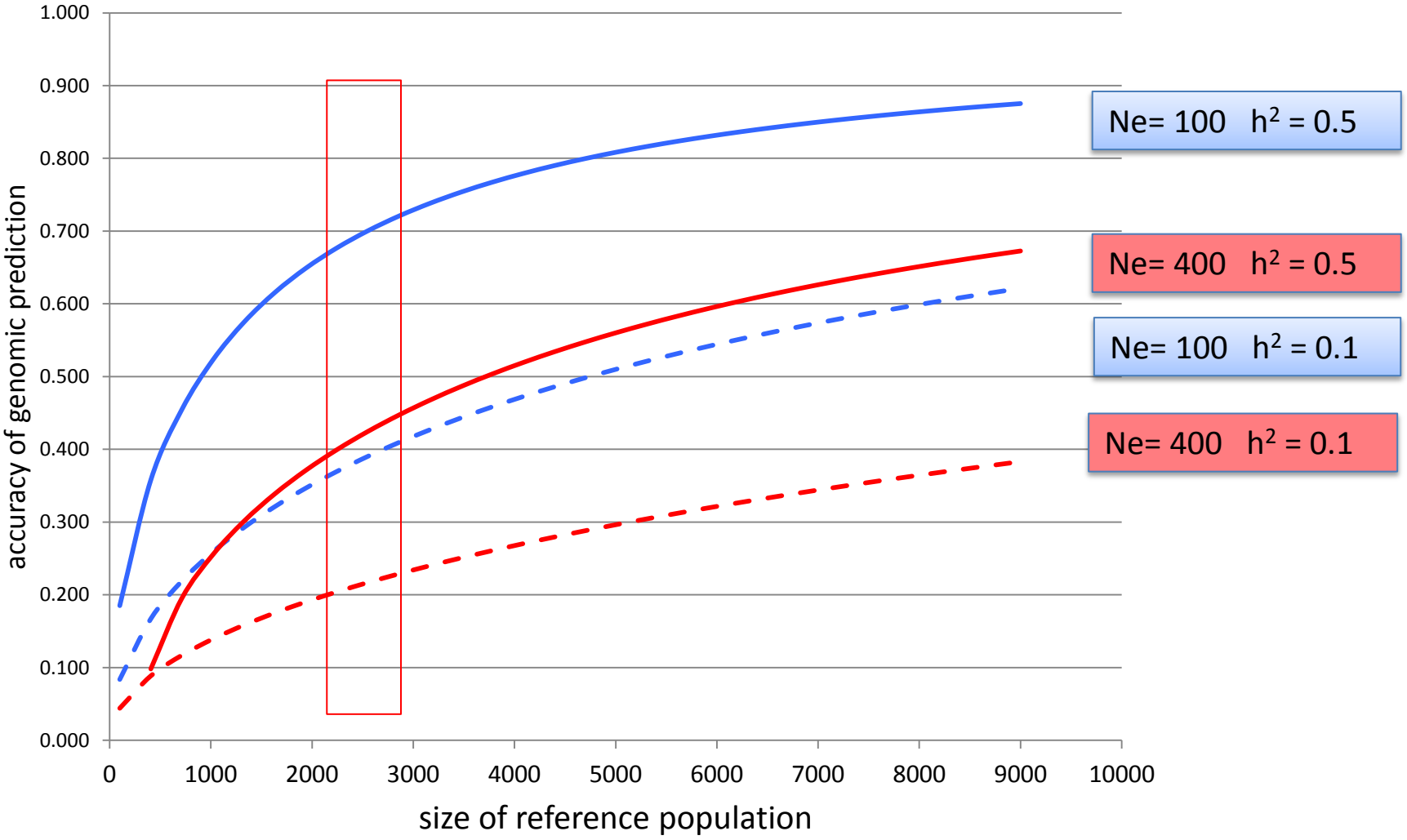
D =deviation in relationship at QTL

Var(D) = $1/M_e$

E = error

Var(E) = 1/nr Markers

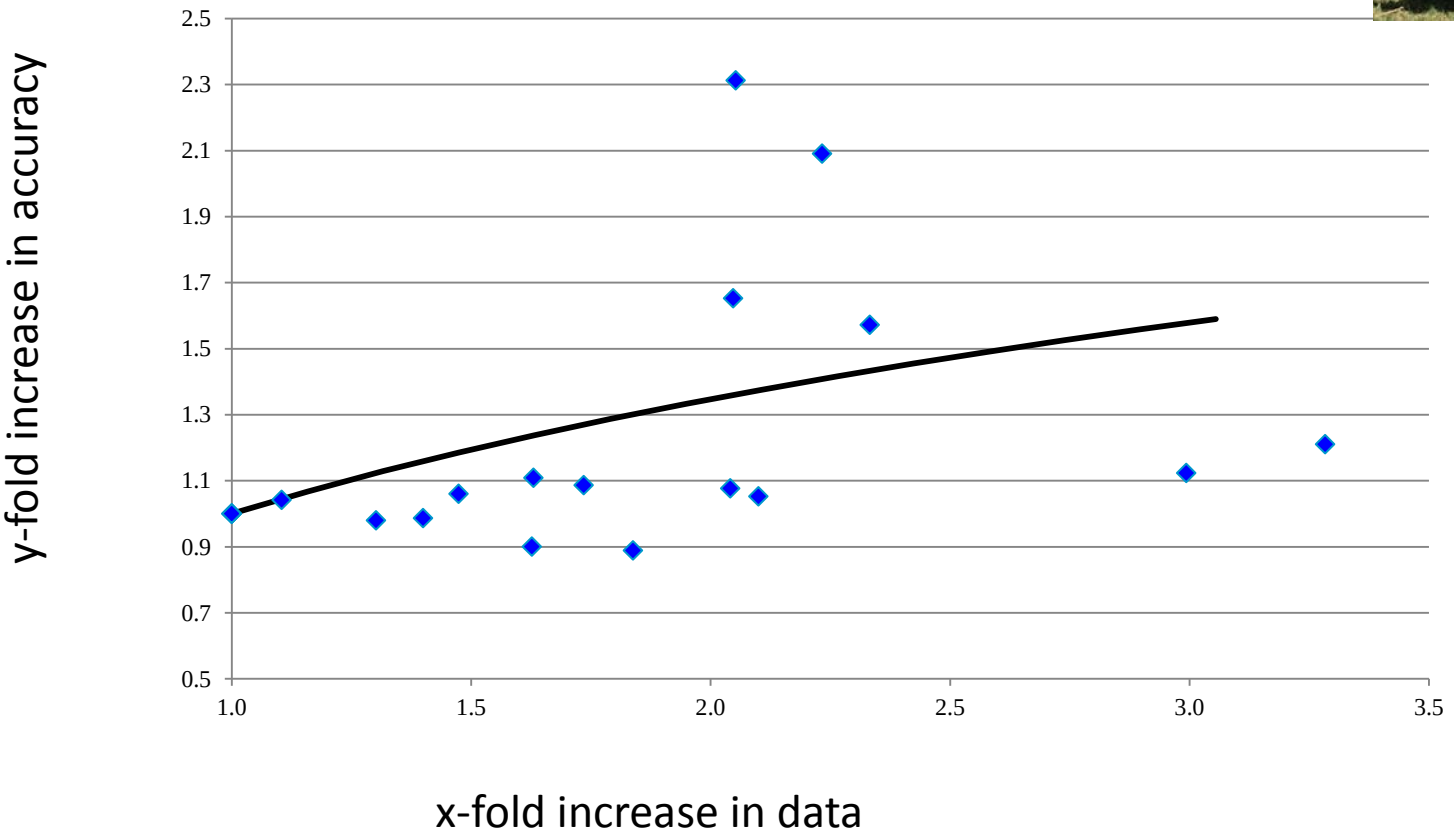
Genomic prediction accuracy *Using Goddard et al, 2011*



Validating ‘Genomic Prediction Accuracy’

More data is always good

But does it increase accuracy as expected?



What effective population size?

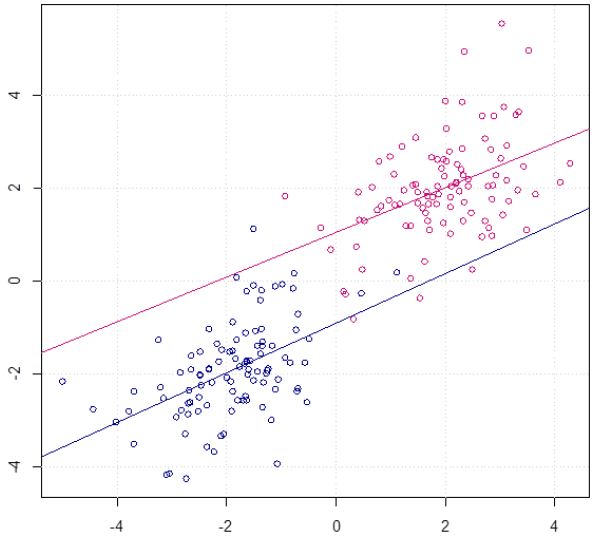
Kijas et al 2012

- Sampling?

Populations not homogeneous.

Within and between breed/line accuracies

Some accuracy due to population structure



Relationship with reference population

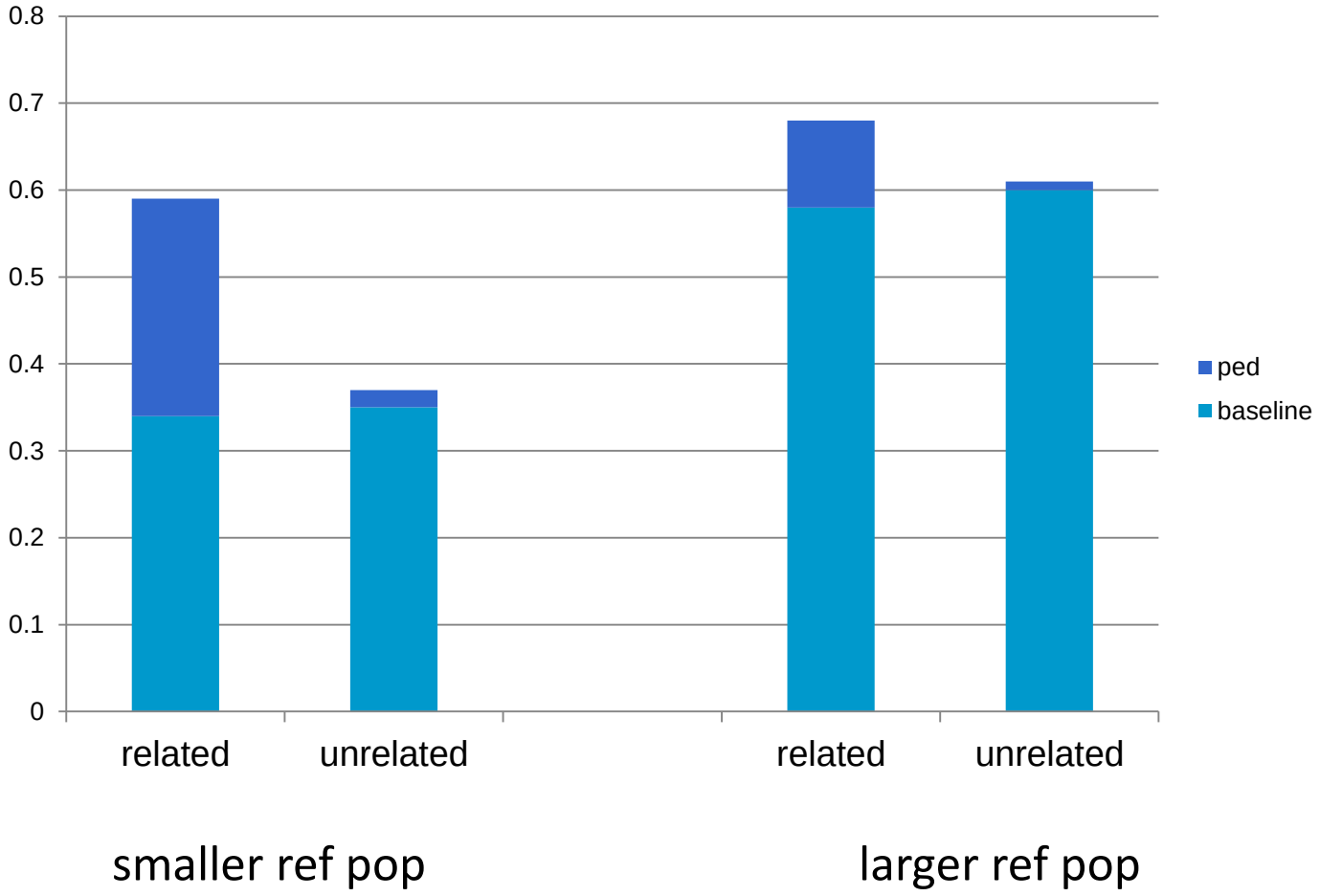
Clark et al 2011

Method	**Close** Ped 0 - 0.25 Genom 0.08 – 0.35	**Distant** 0 - 0.125 0.08 – 0.26	**Unrelated** 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	**0.57**	0.41	0.34

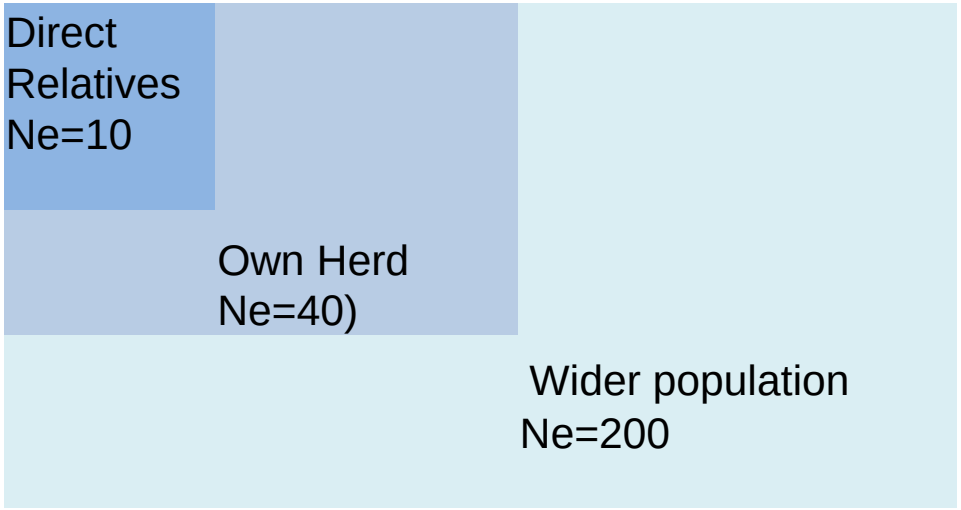
Additional accuracy from family info

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

Relatedness matters more if the reference population is smaller



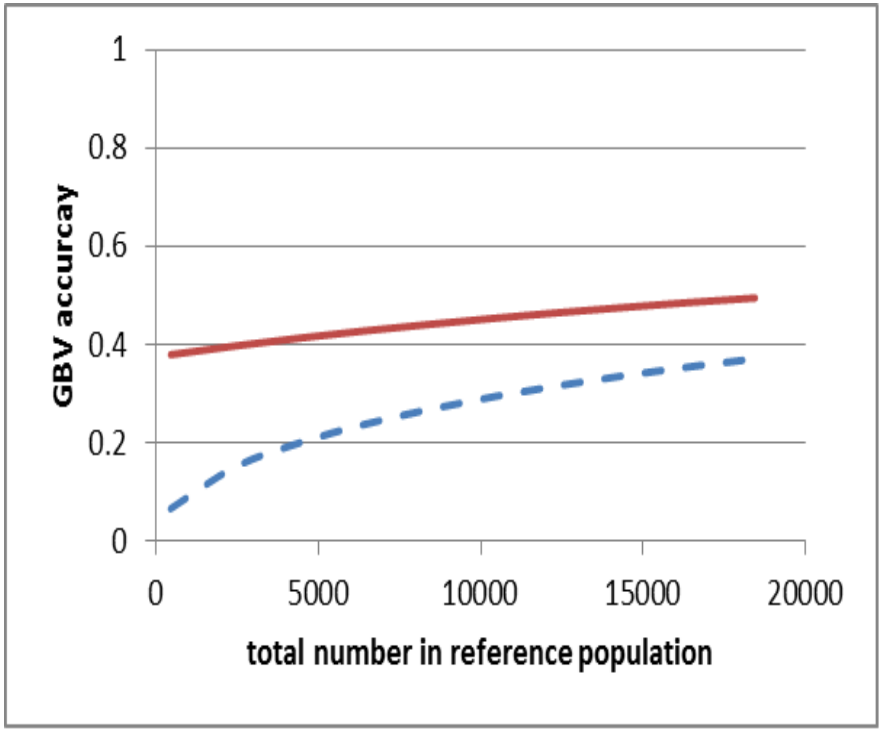
Using a stratified Reference population -populations are not homogeneous



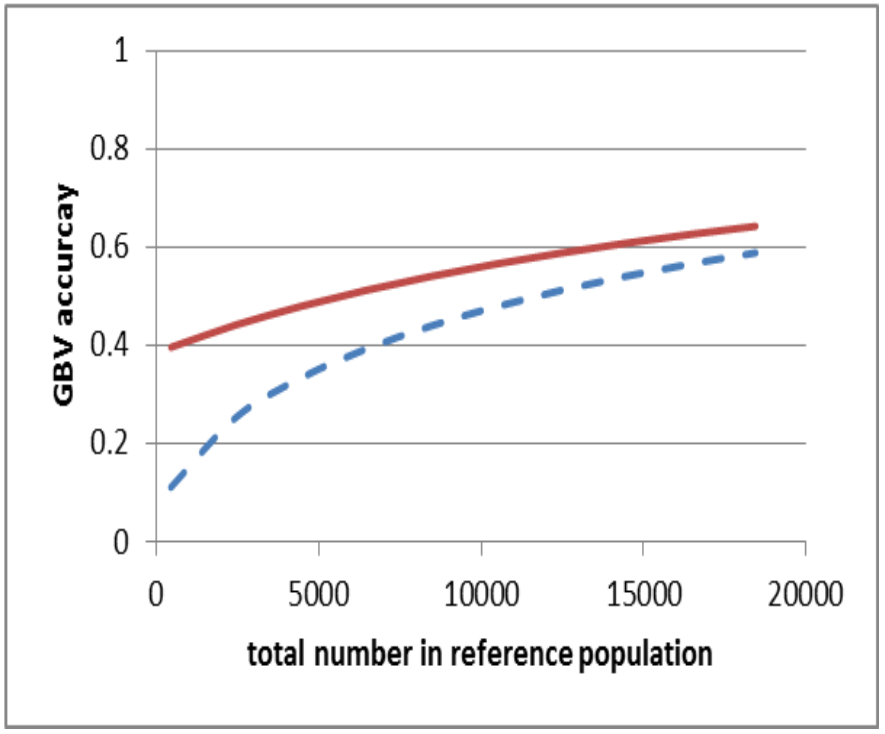
Accuracy of GBV

vary total reference population size

comparing ‘with’ (continuous line) and ‘without’ (dashed line) information on own herd and relatives.



Nmarkers=12k



Nmarkers = 500k

Contribution of different sources

Van der Werf et al, AAABG 2015

Table 1 Value of the various information sources, accuracy of GBV with and without the *flock* and *relatives* information sources[2] and the relative accuracy difference (diff).

N1	Value of information source[1] *breed*	*flock*	*relatives*	GBV_acc_with	GBV_acc_wo	diff[3]
NE1=1000, N2=400, N3=50						
2000	16%	52%	21%	0.428	0.220	95%
5000	31%	39%	15%	0.471	0.318	48%
10,000	45%	26%	10%	0.528	0.420	26%
NE1=1000, **N2=100, N3=10**						
2000	48%	36%	12%	0.279	0.205	36%
5000	68%	19%	6%	0.357	0.309	15%
10,000	79%	11%	4%	0.445	0.414	7%
NE1=200, N2=400, N3=50						
2000	45%	26%	10%	0.528	0.448	18%
5000	62%	12%	5%	0.640	0.599	7%
10,000	72%	5%	2%	0.739	0.718	3%

[1] Percent decrease in accuracy if this information source was removed.

[2] N_{E2} = 50, N_{E3} = 8, Marker density = 50k.

[3] Difference between prediction accuracy with and without information from flock and relatives

Conclusions

- Theory exists to predict genomic prediction accuracy in advance
- Relies on assumptions regarding effective population size
- Ignores heterogeneity of populations and relationships