Accuracy of Genomic Prediction

Julius van der Werf

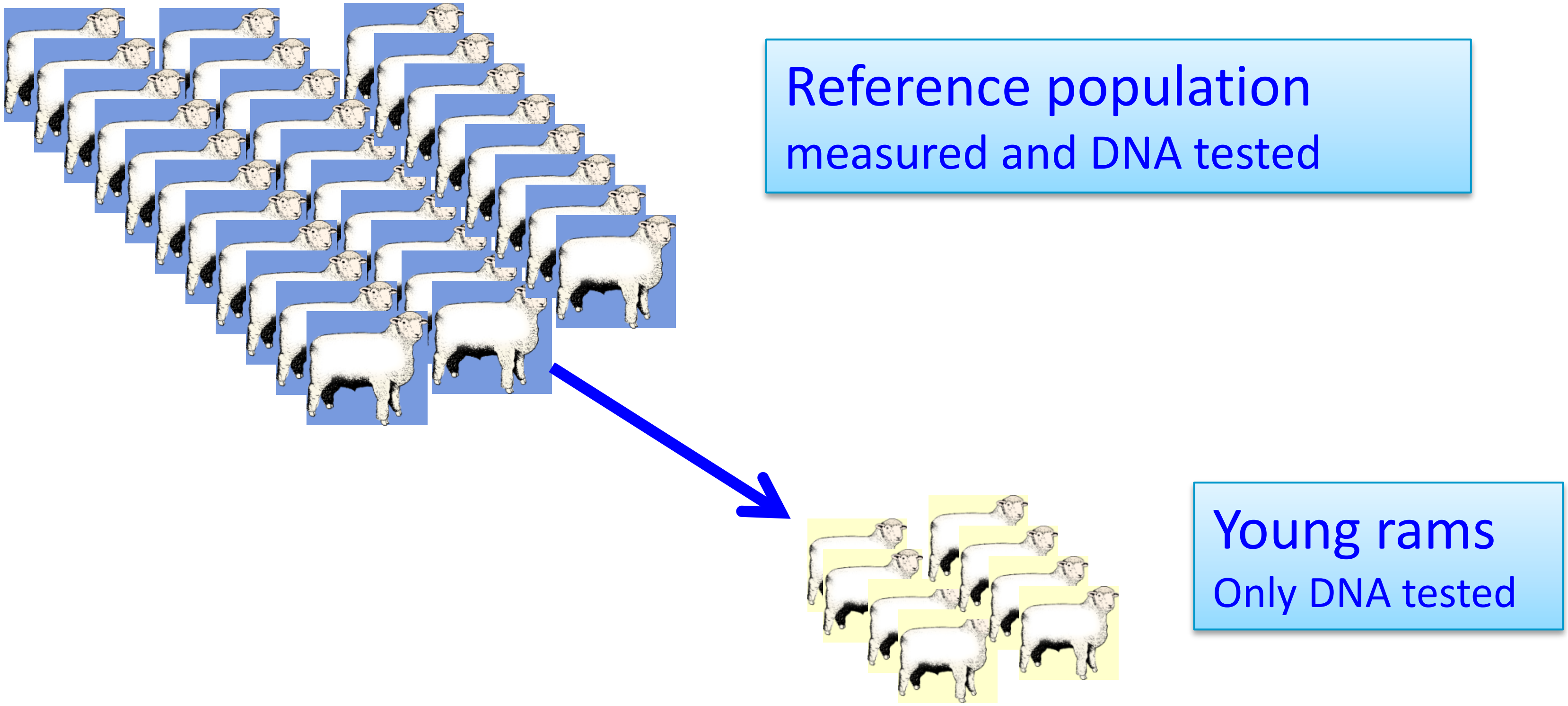
une University of New England

SHEEP CRC

An Australian Government Initiative

CRC AUSTRALIA

Genomic Prediction: basic idea



Reference population
measured and DNA tested

Young rams
Only DNA tested

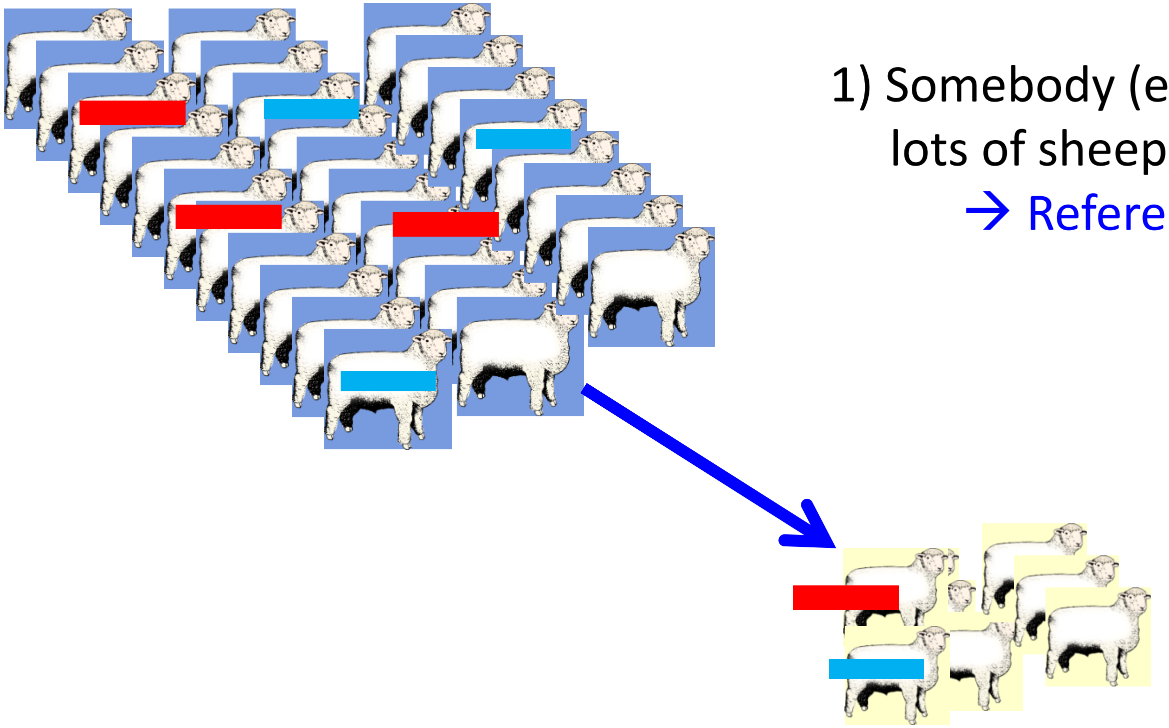
To predict a trait EBV at a young age,

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: basic idea

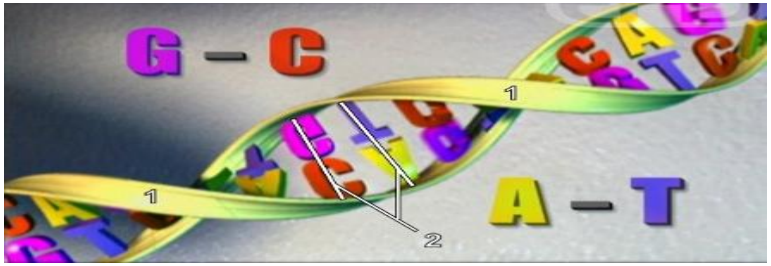


1) Somebody (else) measures
lots of sheep, and their DNA
→ Reference population

2) A breeder tests
DNA on young rams

Illustrating (dis-)similarity of chromosome segments

Genotype information



Father

10100**1**110111**0**01110**0**1110011

01010**0**111000**1**10001**1**0011010

Mother

00010**0**111100**1**010110**0**110011

10101**1**101011**1**111111**1**111110

Chromosome segments are passed on

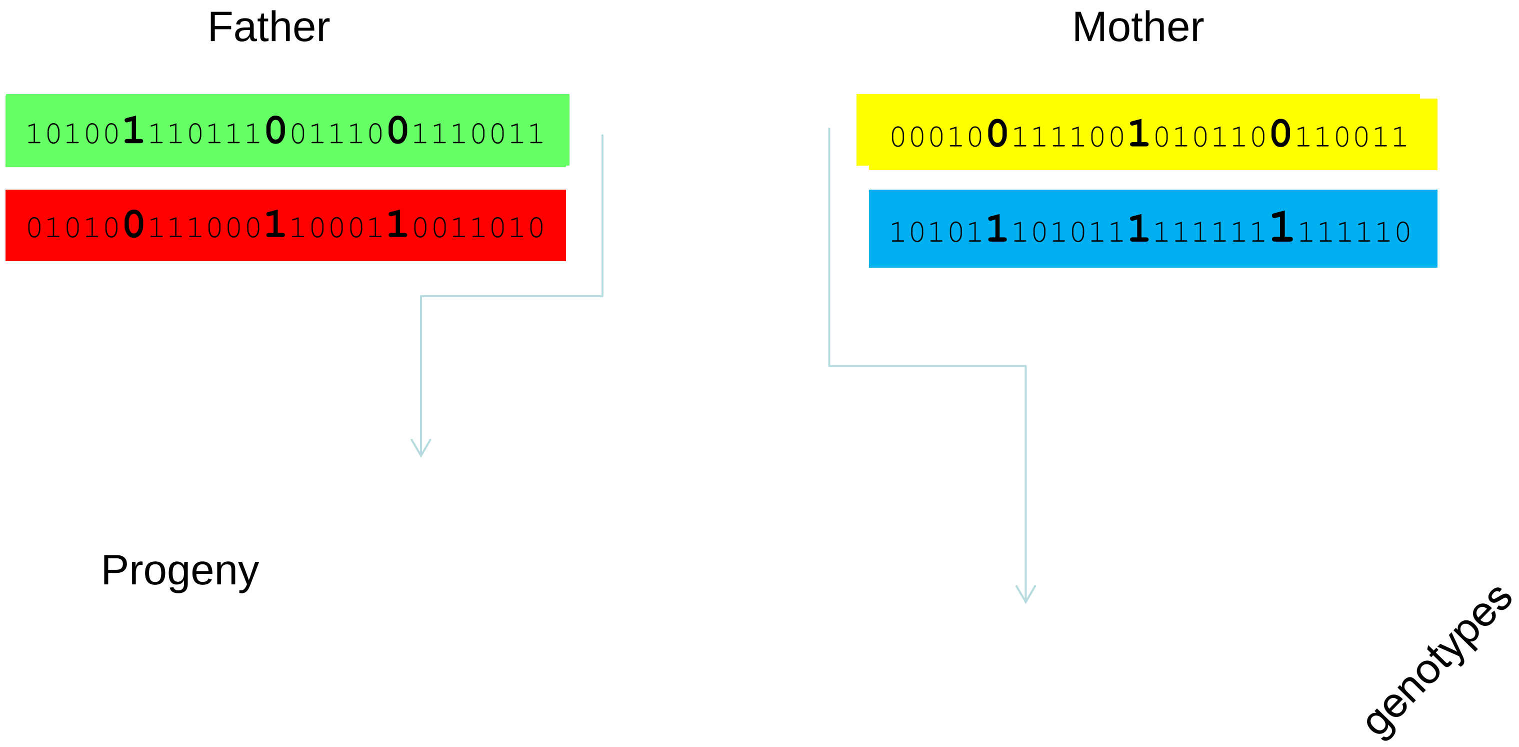
Progeny

10100**1**110111**0**01110**0**1110011

00010**0**111100**1**01011**0**0110011

genotypes

Working out haplotypes (phasing)



Father

10100**1**110111**0**01110**0**1110011

01010**0**111000**1**10001**1**0011010

Mother

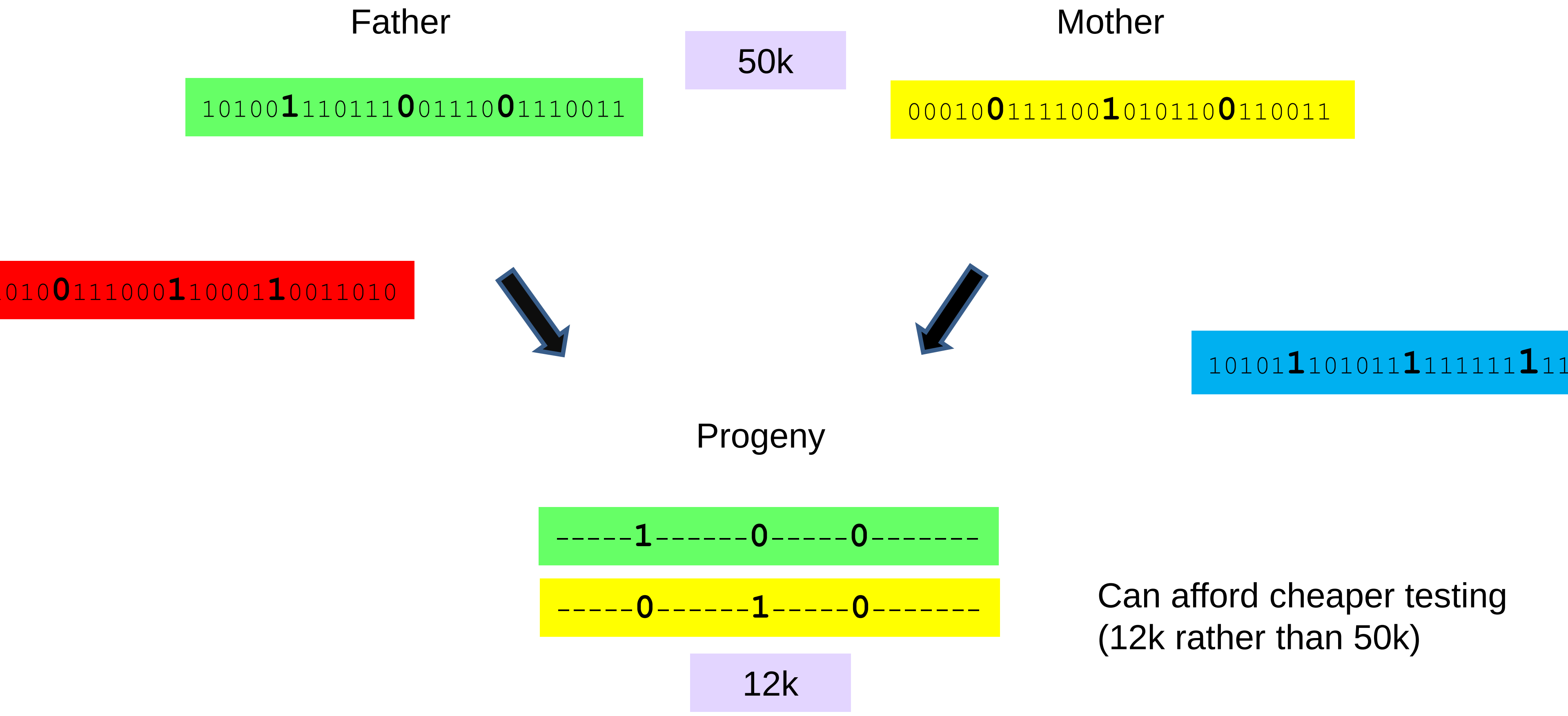
00010**0**111100**1**010110**0**110011

10101**1**101011**1**111111**1**111110

Progeny

genotypes

Filling in the gaps (imputation)



Father

50k

Mother

10100**1**110111**0**01110**0**1110011

00010**0**111100**1**010110**0**110011

010**0**111000**1**10001**1**0011010

10101**1**101011**1**111111**1**11

Progeny

-----**1**------**0**-----**0**-------

-----**0**------**1**-----**0**-------

12k

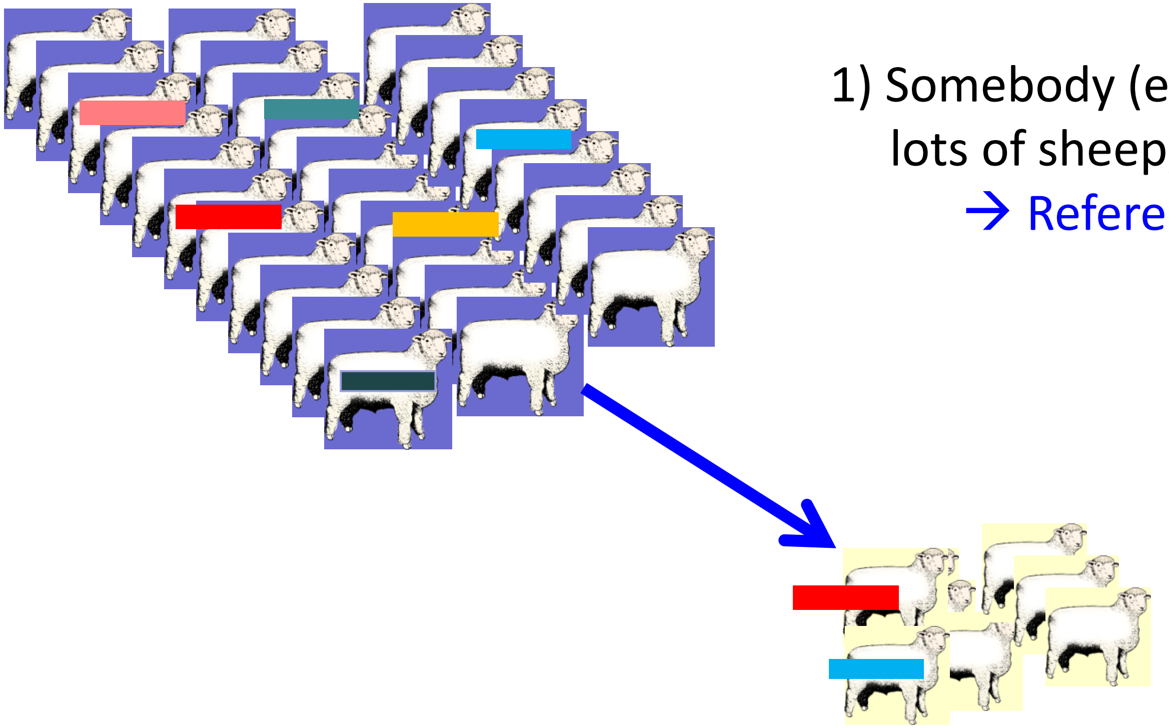
Can afford cheaper testing (12k rather than 50k)

A whole population of haplotypes

Individual																	
1	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	###	1129	1129	1129
2	1088	1088	1088	1088	1088	1088	1088	1088	1192	1192	1192	1192	1192	###	1192	623	623
	178	655	891	891	891	891	891	891	891	1136	1136	1136	1136	735	735	735	735
3	129	129	129	129	129	129	129	655	655	655	655	655	655	###	1038	1038	1038
	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
	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
	1136	1136	1136	1136	1136	1136	1136	1136	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
	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
	655	655	655	1358	1358	1358	1358	1358	1358	342	342	342	342	342	342	1043	1043
8	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	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
	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

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

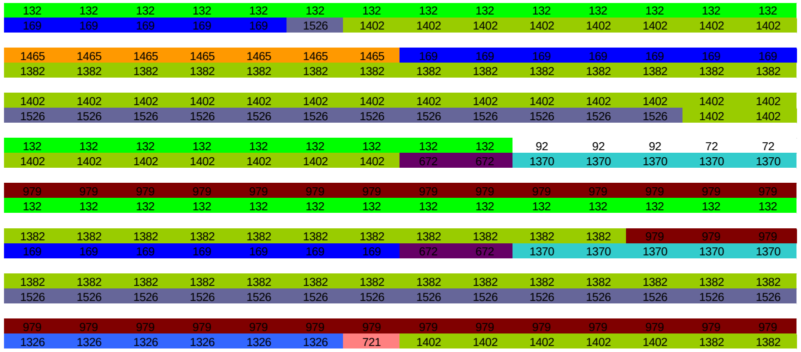


1) Somebody (else) measures lots of sheep, and their DNA
→ Reference population

2) A breeder tests DNA on young rams

Large diversity of segments → less accuracy

populations of haplotypes



132	132	132	132	132	132	132	132	132	132	132	132	132	132
169	169	169	169	169	1526	1402	1402	1402	1402	1402	1402	1402	1402
1465	1465	1465	1465	1465	1465	1465	169	169	169	169	169	169	169
1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	1382
1402	1402	1402	1402	1402	1402	1402	1402	1402	1402	1402	1402	1402	1402
1526	1526	1526	1526	1526	1526	1526	1526	1526	1526	1526	1526	1402	1402
132	132	132	132	132	132	132	132	132	92	92	92	72	72
1402	1402	1402	1402	1402	1402	1402	672	672	1370	1370	1370	1370	1370
979	979	979	979	979	979	979	979	979	979	979	979	979	979
132	132	132	132	132	132	132	132	132	132	132	132	132	132
1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	979	979	979
169	169	169	169	169	169	169	672	672	1370	1370	1370	1370	1370
1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	1382	1382
1526	1526	1526	1526	1526	1526	1526	1526	1526	1526	1526	1526	1526	1526
979	979	979	979	979	979	979	979	979	979	979	979	979	979
1326	1326	1326	1326	1326	1326	721	1402	1402	1402	1402	1402	1382	1382

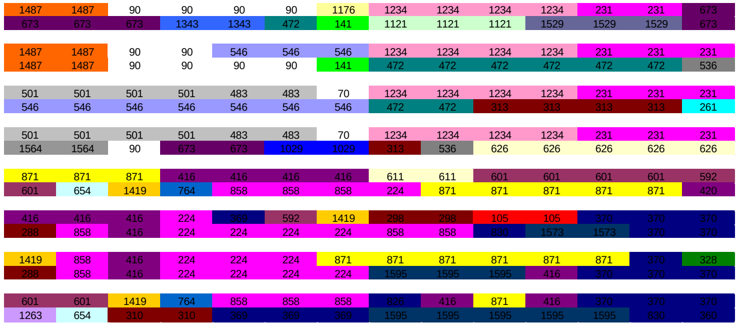
Holstein Friesian, a pig/poultry nucleus

Limited diversity
Long segment sharing

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

Merino sheep, humans

More diversity
Short segment sharing
Sub populations



1487	1487	90	90	90	90	1176	1234	1234	1234	1234	231	231	673
673	673	673	1343	1343	472	141	1121	1121	1121	1529	1529	1529	673
1487	1487	90	90	546	546	546	1234	1234	1234	1234	231	231	231
1487	1487	90	90	90	90	141	472	472	472	472	472	472	536
501	501	501	501	483	483	70	1234	1234	1234	1234	231	231	231
546	546	546	546	546	546	546	472	472	313	313	313	313	261
501	501	501	501	483	483	70	1234	1234	1234	1234	231	231	231
1564	1564	90	673	673	1029	1029	313	536	626	626	626	626	626
871	871	871	416	416	416	416	611	611	601	601	601	601	592
601	654	1419	764	858	858	858	224	871	871	871	871	871	420
416	416	416	224	369	592	1419	298	298	105	105	370	370	370
288	858	416	224	224	224	224	858	858	830	1573	1573	370	370
1419	858	416	224	224	224	871	871	871	871	871	871	370	328
288	858	416	224	224	224	224	1595	1595	1595	416	370	370	370
601	601	1419	764	858	858	858	826	416	871	416	370	370	370
1263	654	310	310	369	369	369	1595	1595	1595	1595	1595	830	360

Fine wool,
small

Coarse wool,
big

Not only recent N_e but also historic N_e is relevant

Genomic prediction accuracy *Using Daetwyler et al, 2008*

Accuracy^2 of estimating a random effect = $n / (n+\lambda)$ $\lambda = V_e / V_a$

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

And each segemnt has variance VA/ $M_{e,}$ then accuracy of estimating each segment

$$\frac{n}{n+V_e / (V_a/M_e)} = \frac{nV_a}{nV_a + V_e\, M_e} = \frac{h^2}{h^2 + M_e/n}$$

n = nr observations

M_e = effective nr loci

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

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'

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

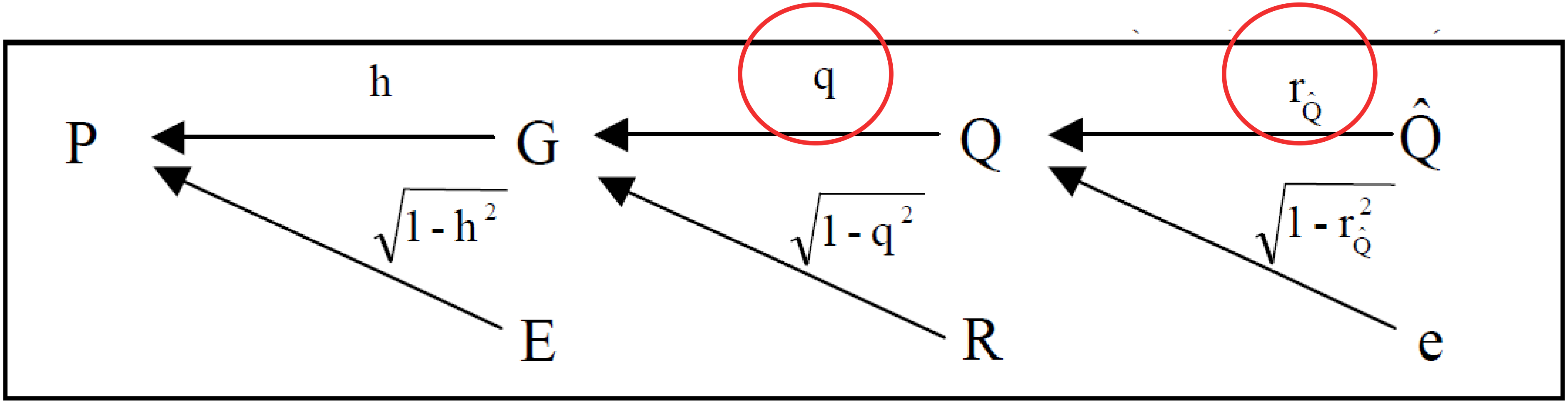
or is it…?

i) Accuracy of estimating marker effects

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

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

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



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

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

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

$\lambda = M_e/h^2$

same as Daetwyler

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

$= r_{Qhat}$

Current question

With very many markers, e.g. sequence, will we be better of?

What if nr markers >>> nr chromosome segments?

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

$\mathrm{Var}(D) = 1/M_e$

E = error

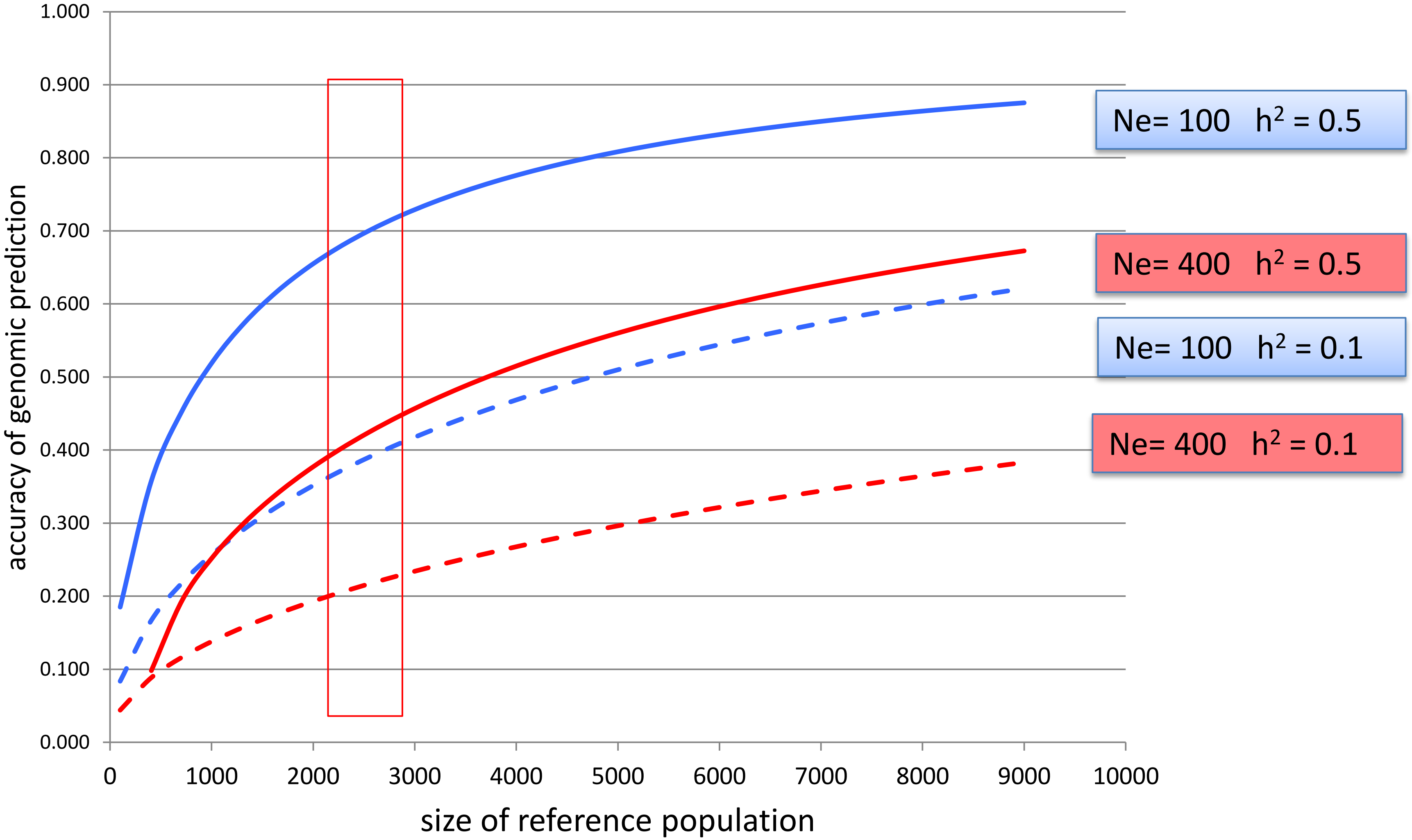
$\mathrm{Var}(E) = 1/\text{nr Markers}$

Empirical validation

Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics 193: 621–631.

Erbe M, Gredler B, Seefried FR, Bapst B, Simianer H (2013) A Function Accounting for Training Set Size and Marker Density to Model the Average Accuracy of Genomic Prediction. PLoS ONE 8(12): e81046. doi:10.1371/journal.pone.0081046

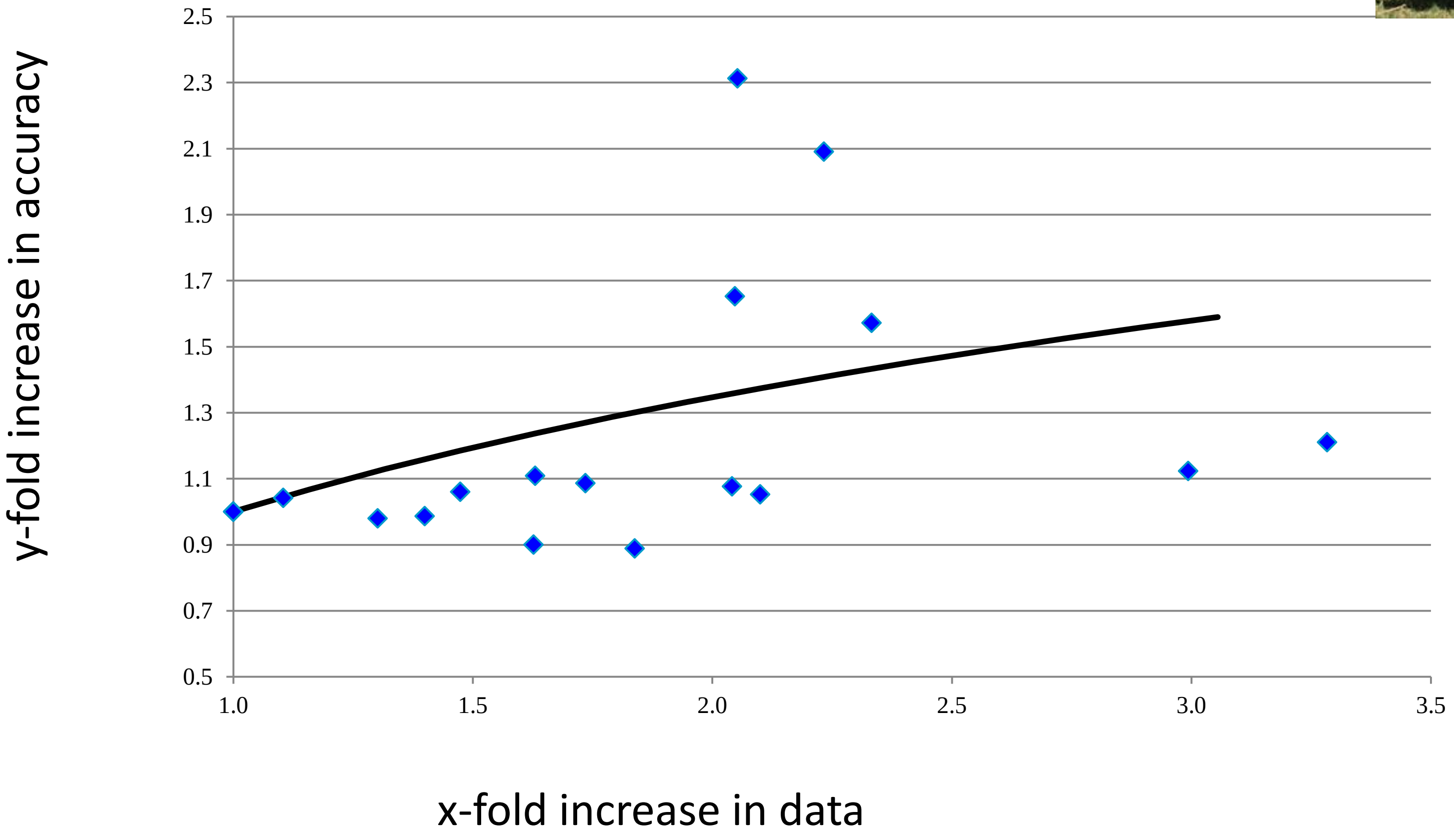
Genomic prediction accuracy *Using Goddard et al, 2011*



Validating ‘Genomic Prediction Accuracy’

More data is always good

But does it increase accuracy as expected?



What effective population size?

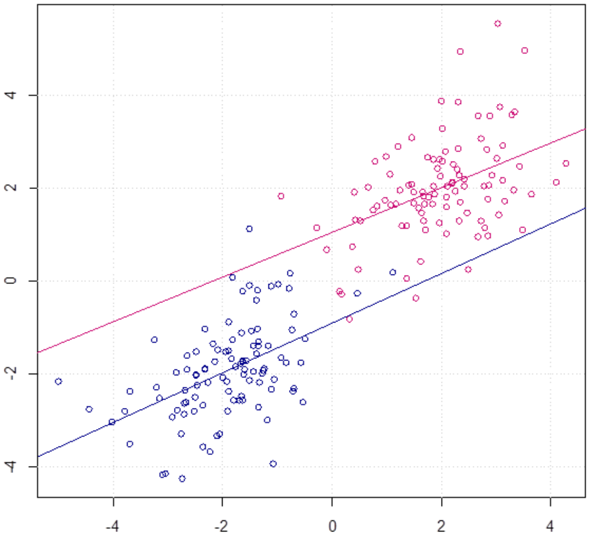
Kijas et al 2012

- Sampling?

Populations not homogeneous.

Within and between breed/line accuracies

Some accuracy due to population structure



Relationship with reference population

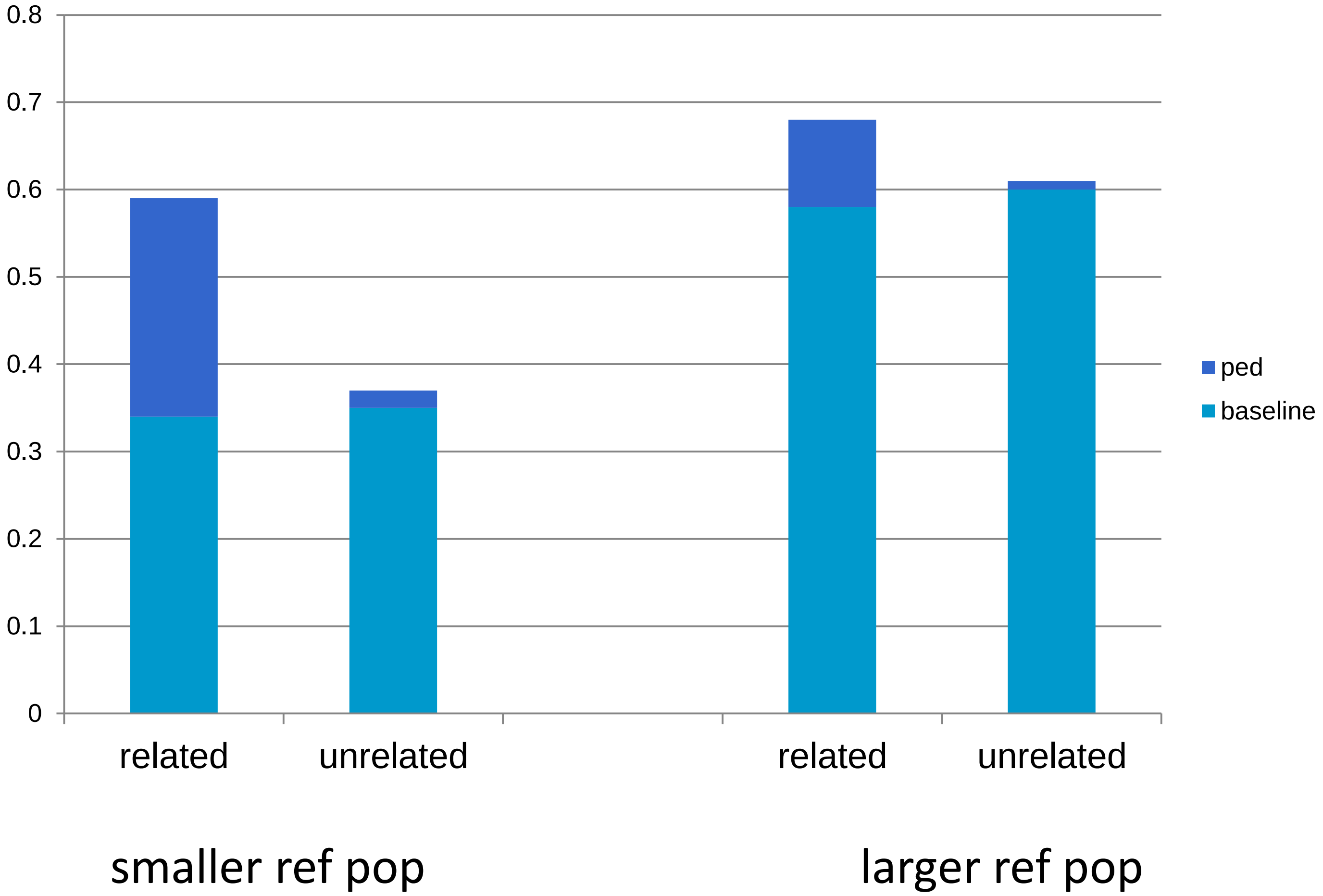
Clark et al 2011

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

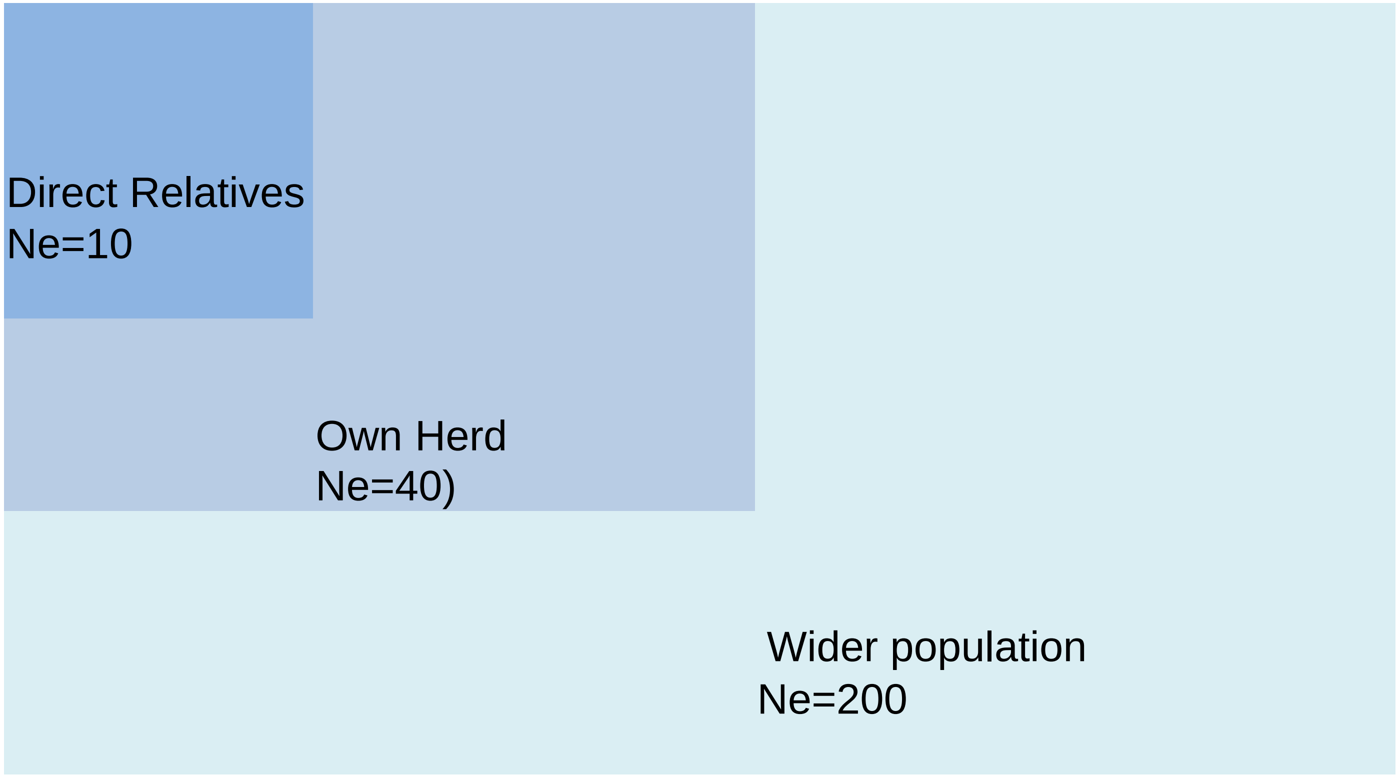
Additional accuracy from family info

'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



Direct Relatives
Ne=10

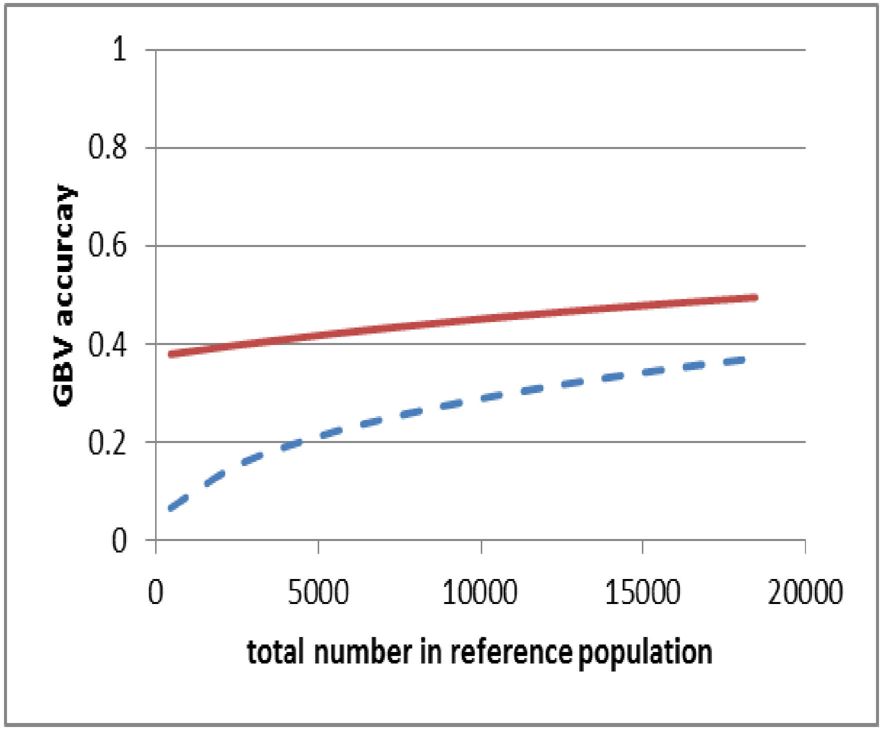
Own Herd
Ne=40)

Wider population
Ne=200

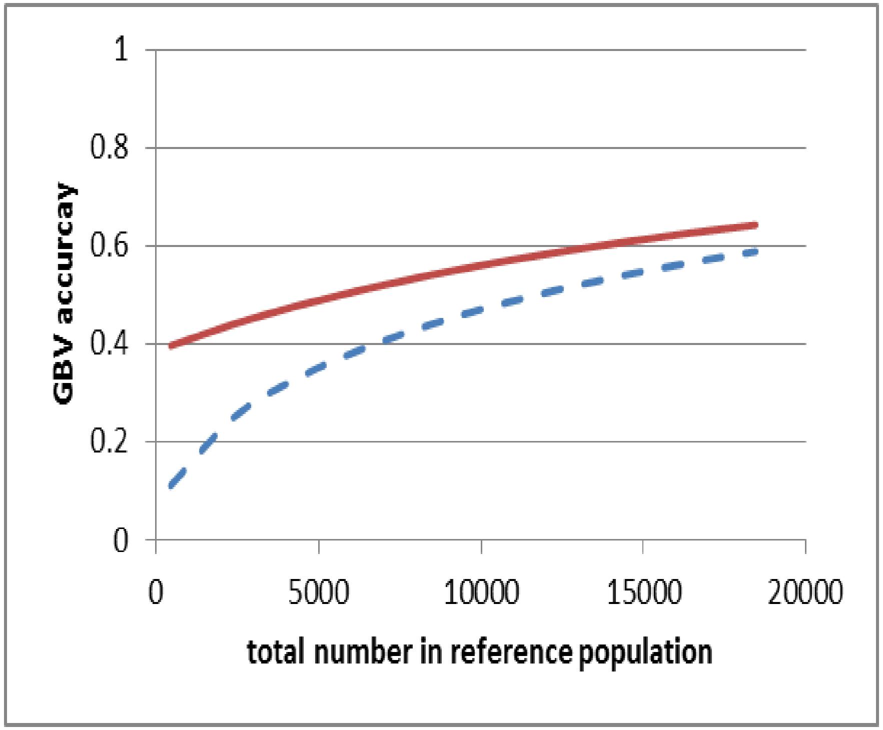
Accuracy of GBV

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).

	Value of information source[1]					
N1	*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: depends on nr. effective segments, nr records
- Relies on assumptions regarding effective population size
- Ignores heterogeneity of populations and relationships