Continuing the transformation

Accuracy of Genomic Prediction

Julius van der Werf

and Sang Hong Lee

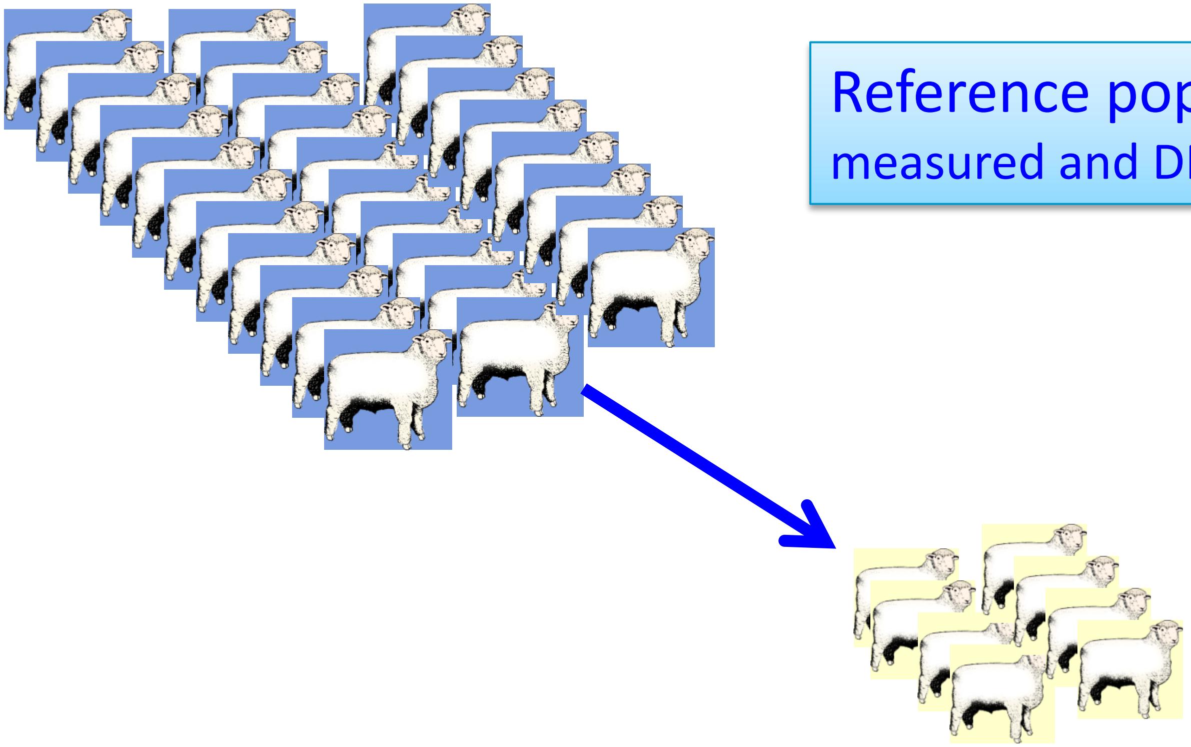
une University of New England

SHEEP CRC

An Australian Government Initiative

CRC AUSTRALIA

Genomic Prediction: basic idea



Reference population
measured and DNA tested

Young sires
Only DNA tested

To predict a trait EBV at a young age,

good for for: late traits
hard to measure traits

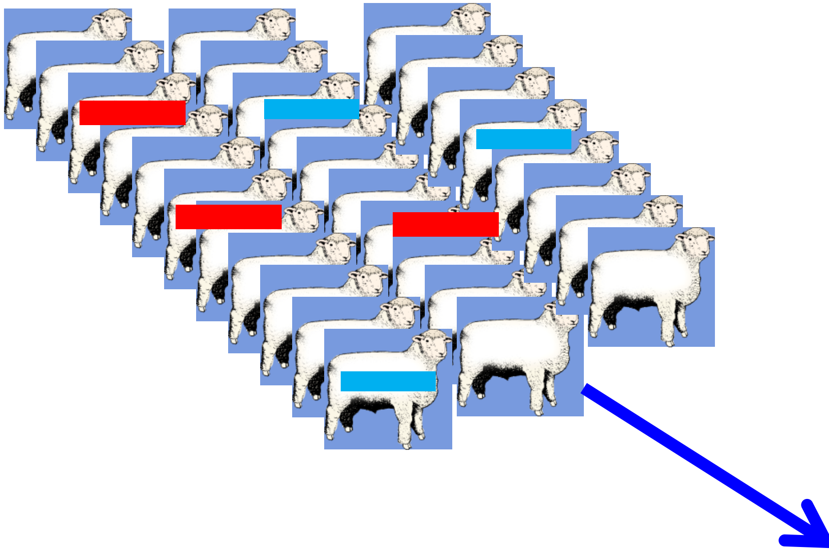
Genomic prediction accuracy

- Derive from the model, e.g. PEV from GBLUP mixed model equations

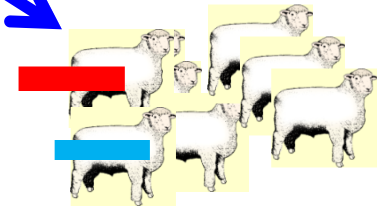
- Validate with other EBVs or phenotypes
  - Validation population
  - Cross-validation

- Predict in advance based on theory and assumptions about population

Genomic Prediction: basic idea



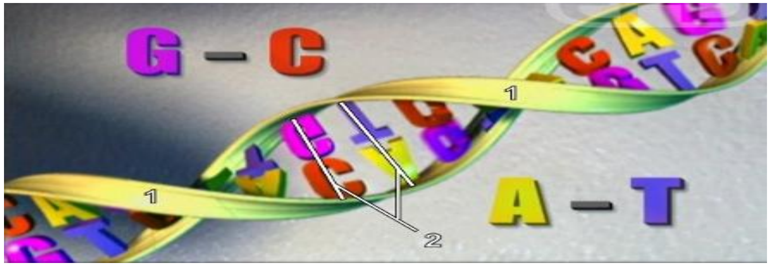
1) Somebody (else) measures
lots of sheep, and their DNA
→ Reference population



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

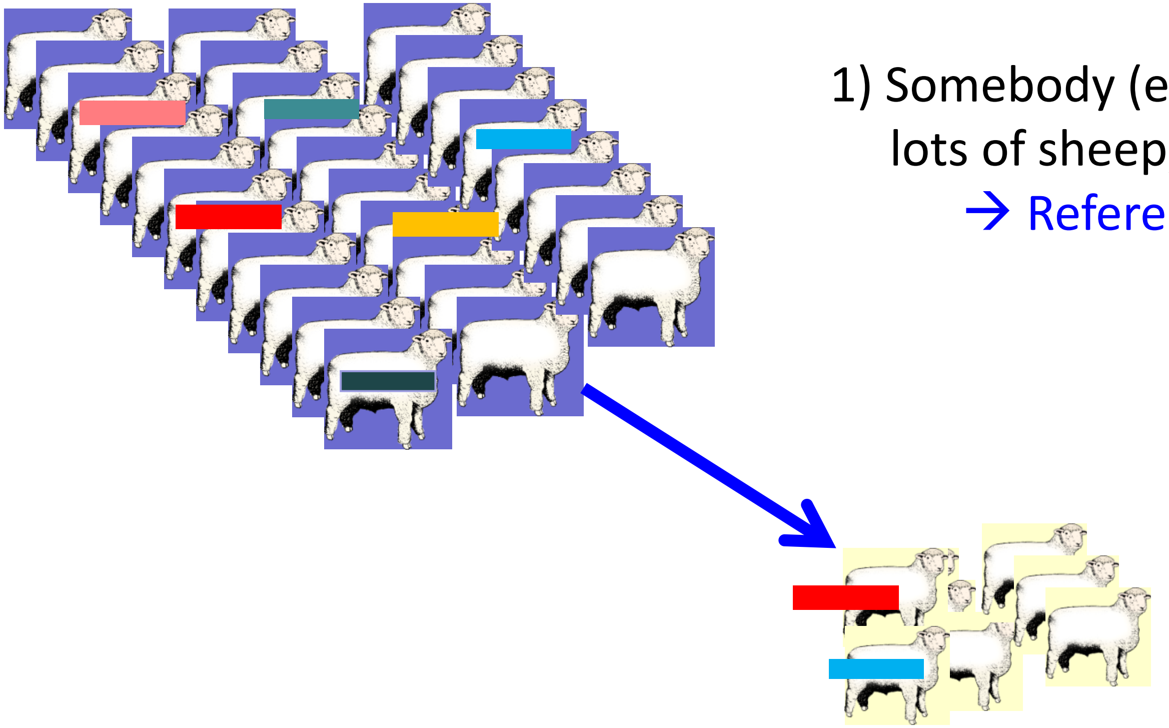
genotypes

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

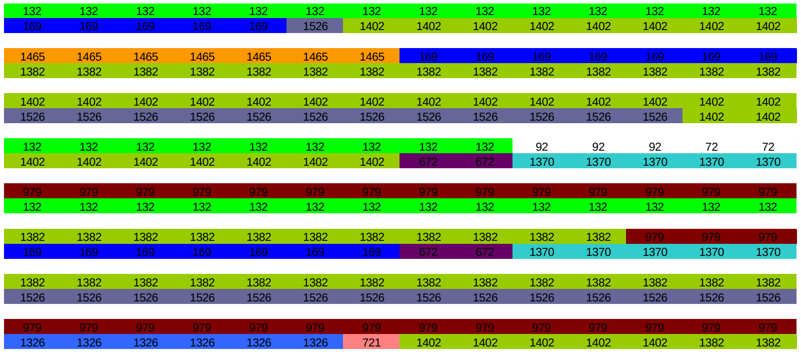


1) Somebody (else) measures lots of sheep, and their DNA
→ Reference population

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

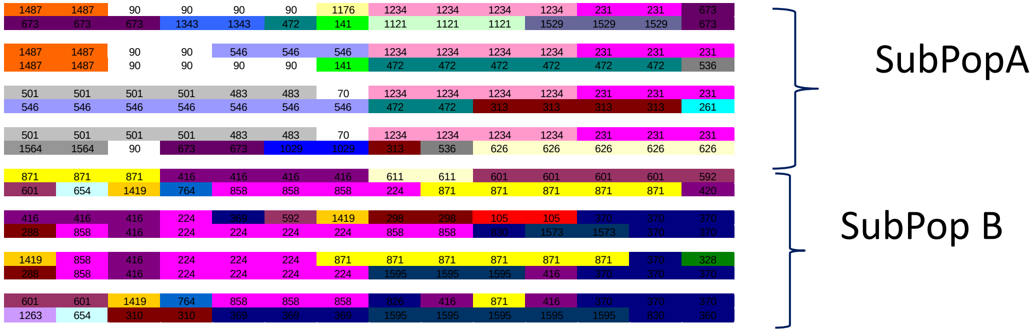
Holstein Friesian, a pig/poultry nucleus

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

Not only recent N_e but also historic N_e is relevant

Genomic prediction accuracy

Design parameters

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

If genome exists of M_e independently segregating ‘effective chromosome segments’

And each segment has variance VA/ $M_{e,}$ then accuracy^2 of estimating each segment

$$\frac{N}{N+V_e / (V_a/M_e)} = \frac{NV_a}{NV_a + V_e M_e} = \frac{h^2}{h^2 + M_e/N}$$

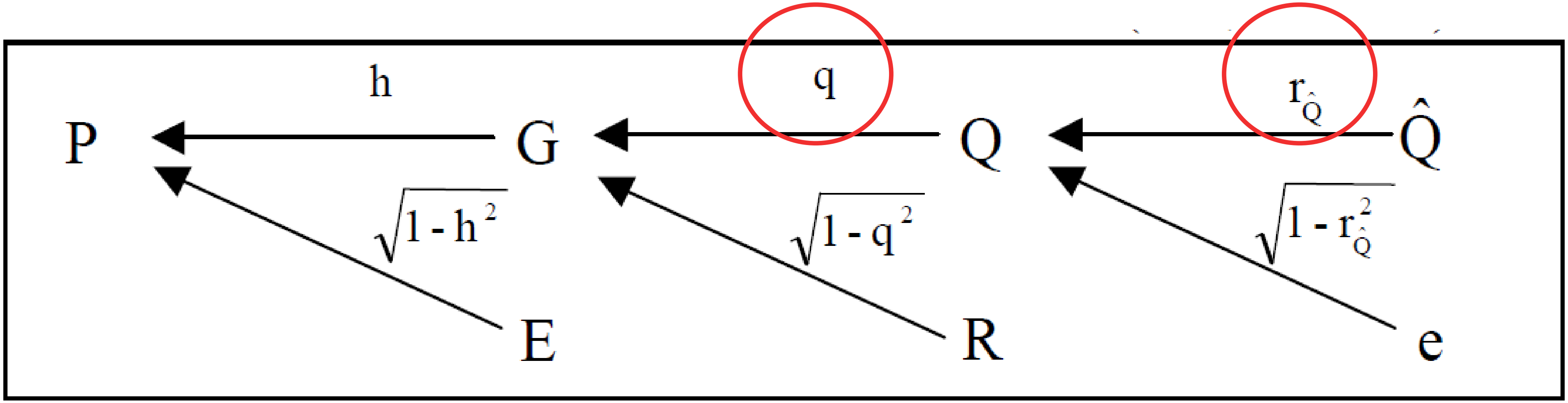
$$r_{g,\hat{g}} = \sqrt{\frac{h^2}{h^2 + M_e / N}}$$

N = nr observations

M_e = effective nr loci

Valid if “all genetic variance is captured by markers”

See also 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: $G = Q + R$

Genomic prediction accuracy *Using Goddard et al, 2011*

Depends on

i) Proportion of genetic variance at QTL captured by markers

q^2

i) Reliability of estimating marker effects

r^2_{Qhat}

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

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

↳ 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

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

$\lambda = M_e/(q^2 . h^2)$

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

$= q . r_{Qhat}$

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

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

$\lambda = M_e/h^2$ same as Daetwyler

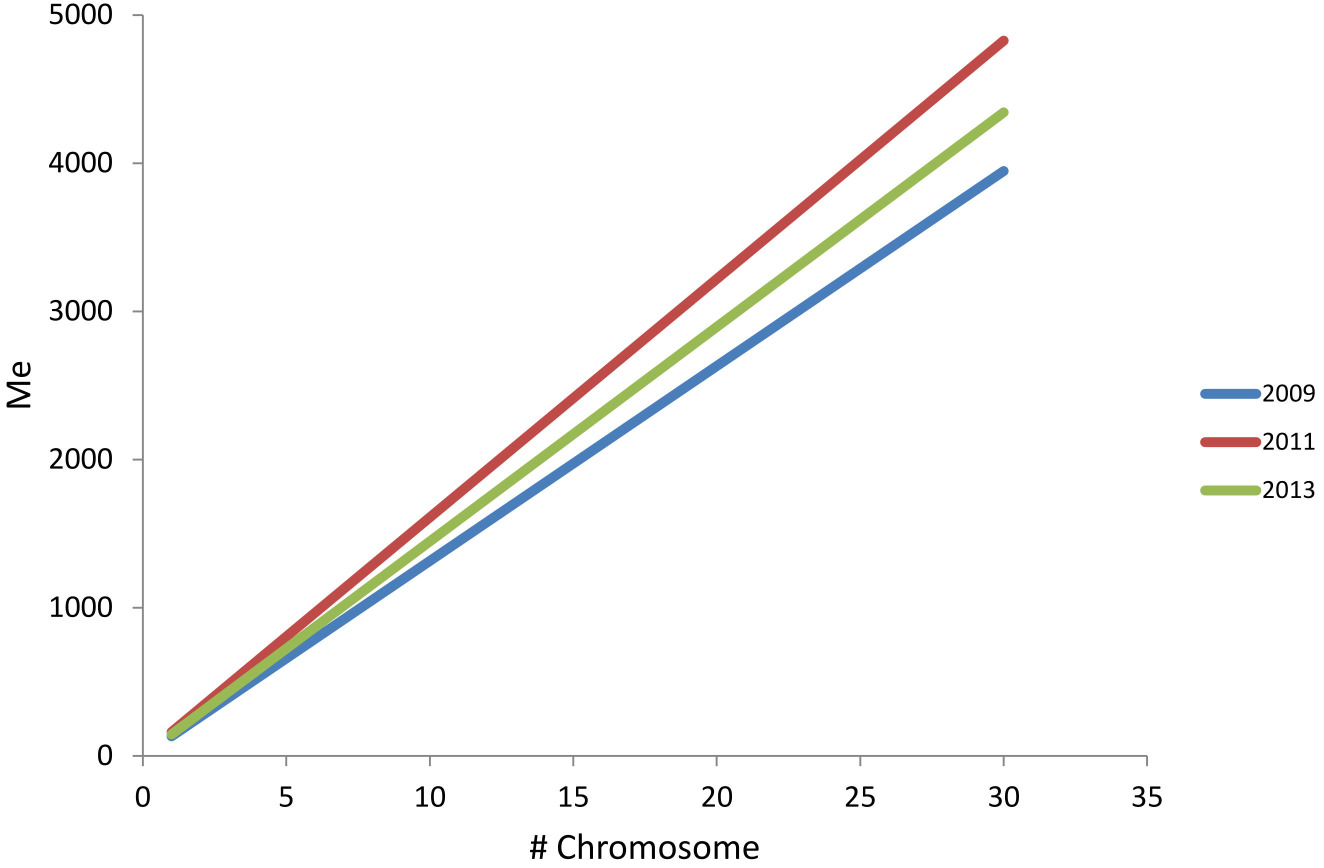
Accuracy = $\sqrt{(\ r^2_{Qhat})}$

$= r_{Qhat}$

M_e is a function of N_e

- $M_e = 2N_eLN_{chr}/\ln(4N_eL)$ (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)

Difference among the formulas



- N_e = 500, L=1M h^2 = 0.5 and N = 5000,
- accuracy = 0.62, 0.58, 0.60

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

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

E = error

Var(E) = 1/nr Markers

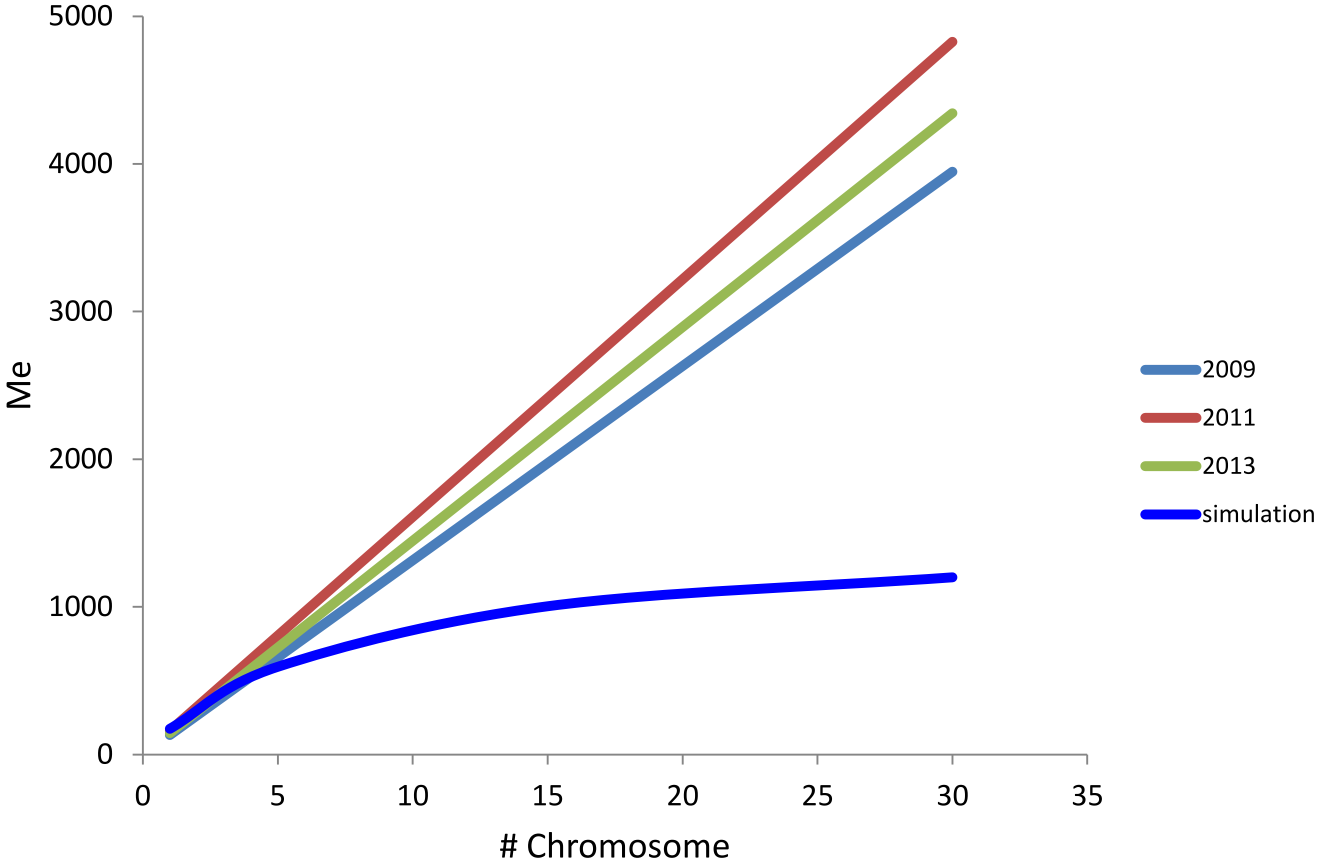
Given genomic relationships (after collecting data), it is possible to empirically get M_e from the data

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 A_{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

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

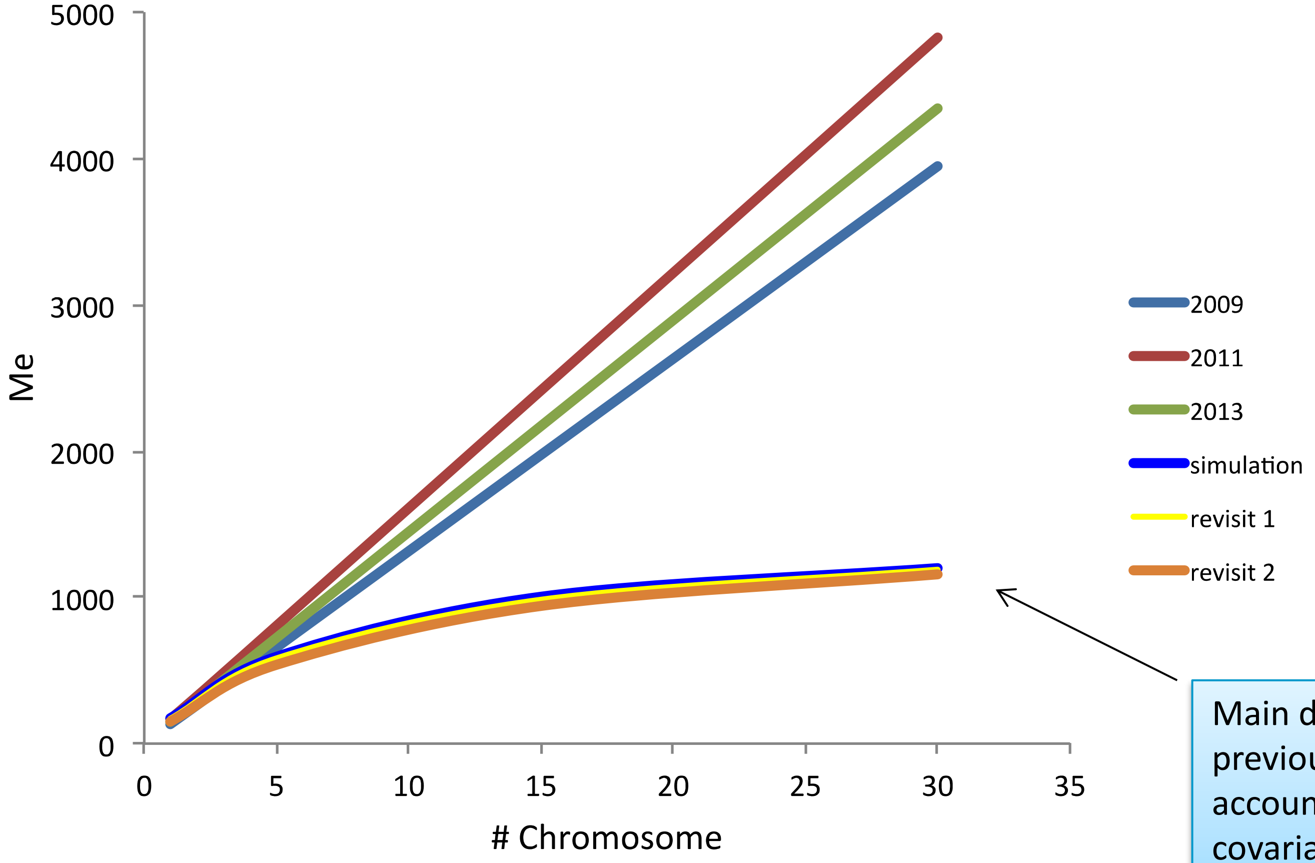
Assuming LD $r^2 = 1 / (1 + 4N_e \times c)$

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

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

For more detail, see a bioRxiv paper Lee *et al*, 2016
doi: http://dx.doi.org/10.1101/054494

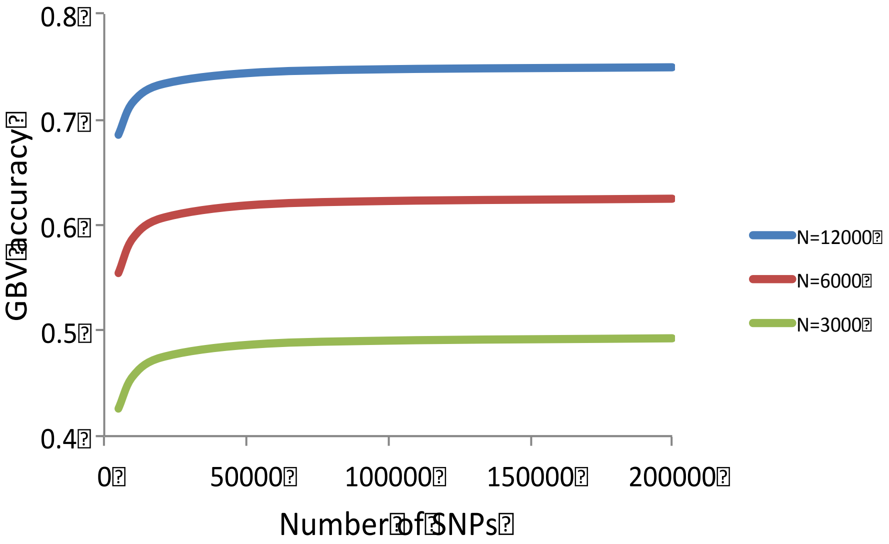
Empirical M_e and new formula

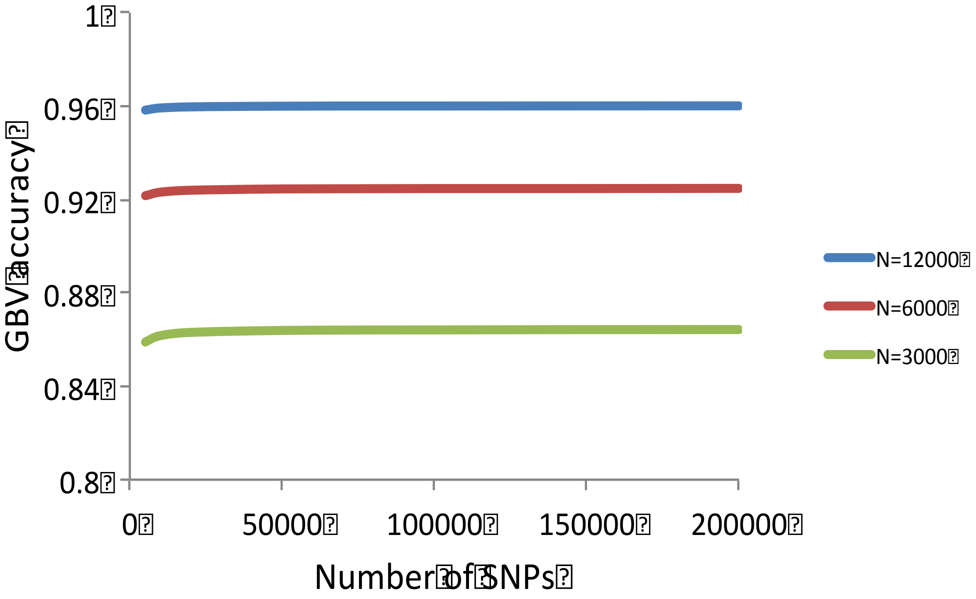


Main difference with previous work is due to accounting for covariance between chromosomes

- Agreed well

Genomic prediction accuracy





Ne = 1,000

Ne = 100

Expect very little improvement with denser markers

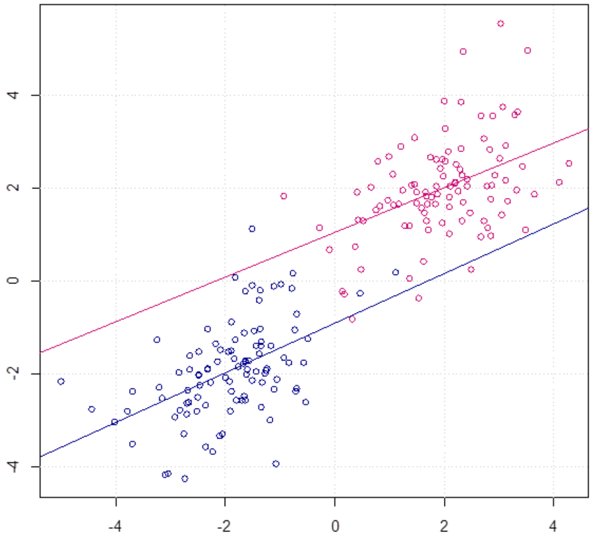
What effective population size?

Hanwoo? ~ 94 (Gondro)

Populations not homogeneous.

Within and between breed/line accuracies

Some accuracy due to population structure



How do we validate accuray?

- Validation population
  - EBV (based on progeny test)
  - Phenotype
  - Is it a homogeneous group?
- Cross-validation
  - Across families
  - Random(also within families)

Relationship with reference population

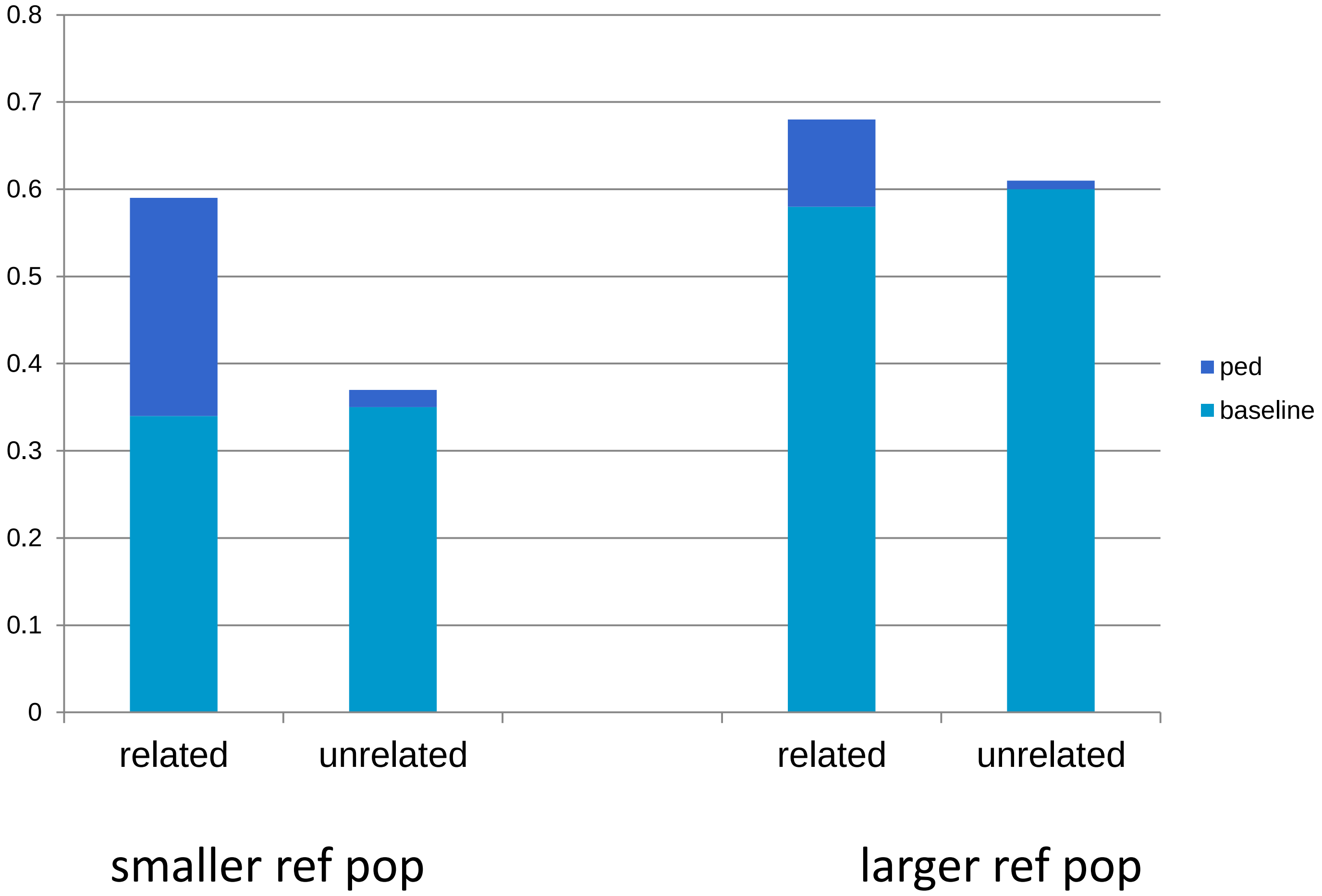
Clark et al 2011

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

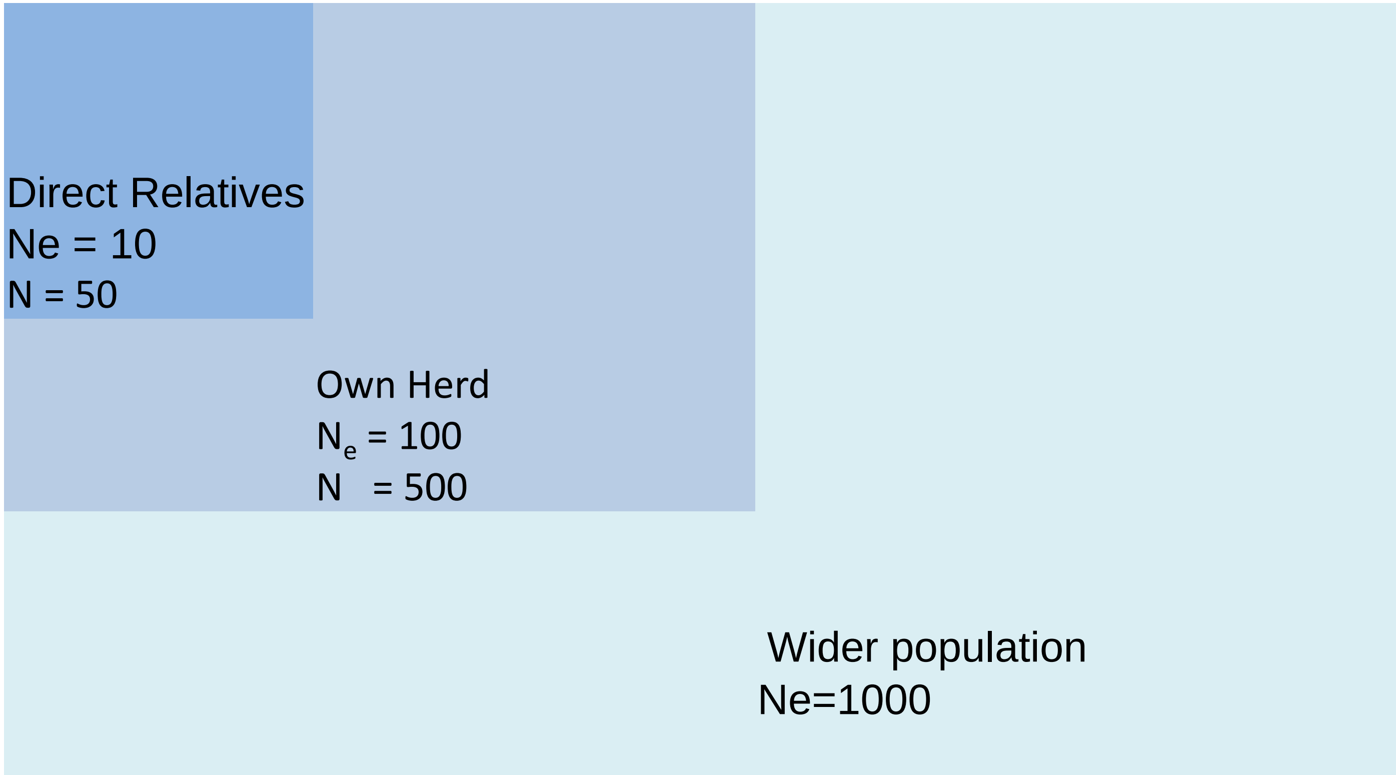
Additional accuracy from family info

‘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

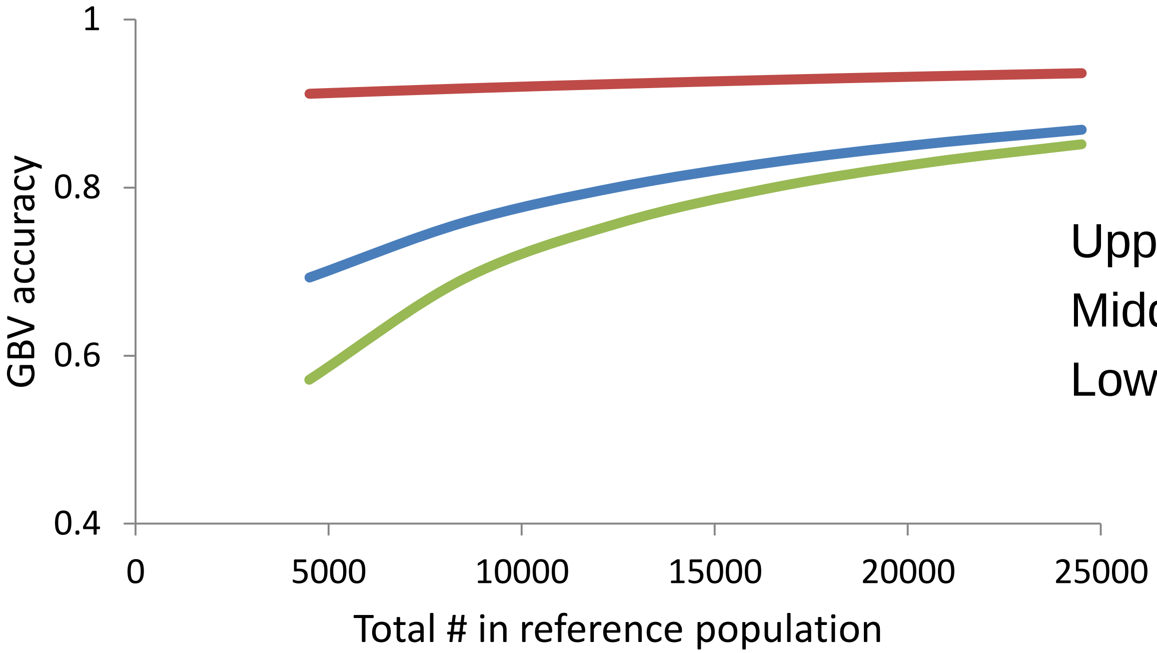


Direct Relatives
Ne = 10
N = 50

Own Herd
N_e = 100
N = 500

Wider population
Ne=1000

Relative importance



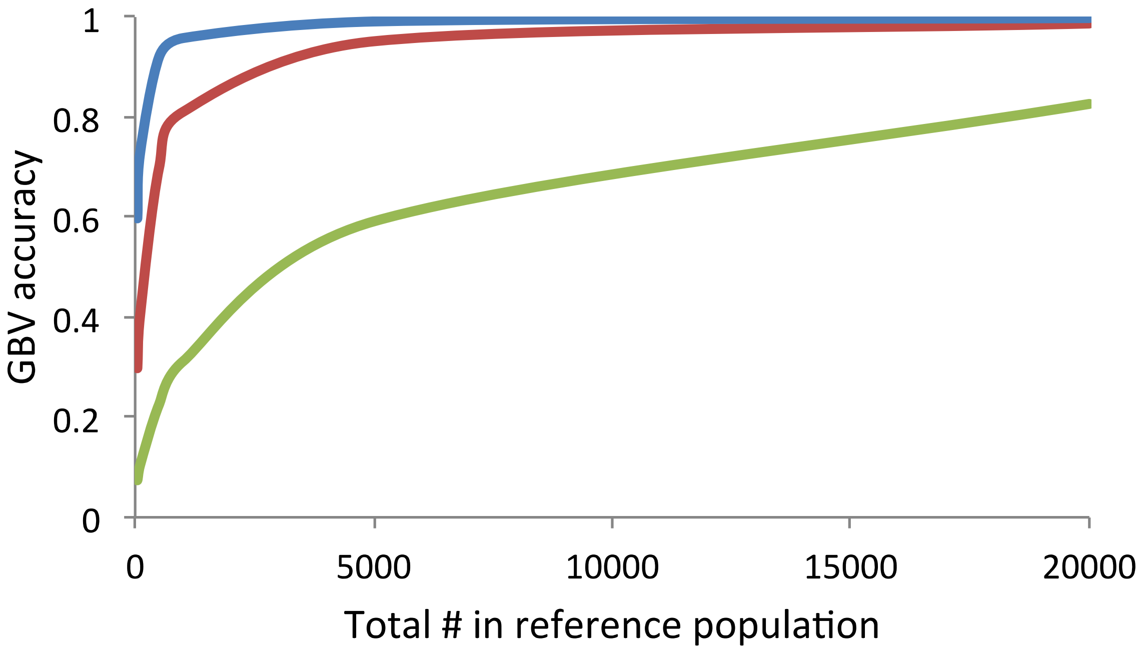
Upper: N_e=1000 + N_e=10 (N=500)

Middle: N_e=1000 + N_e=100 (N=500)

Lower: N_e=1000 only

- h^2=0.25
- Data from smaller N_e is more important

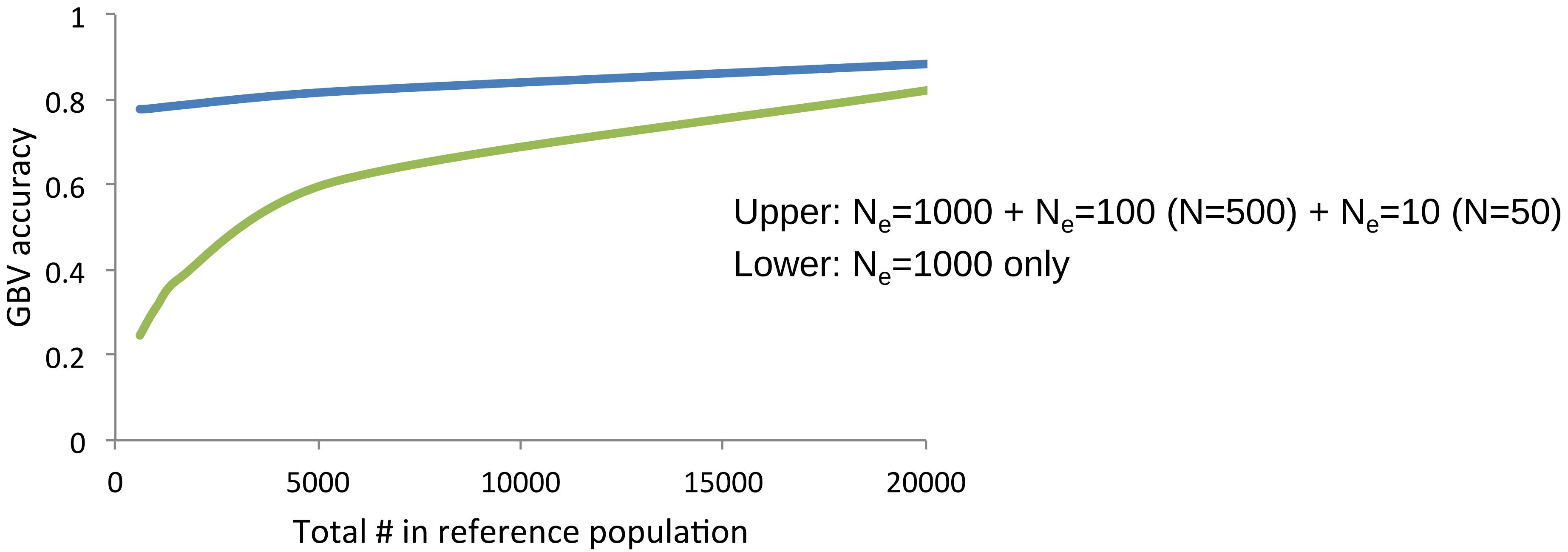
Sample availability



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



Upper: N_e=1000 + N_e=100 (N=500) + N_e=10 (N=50)

Lower: N_e=1000 only

- h^2=0.25
- Smaller N_e is important with smaller total N
- Benefit from large N_e too (0.78 to 0.89)

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
  - > 50K with N_e = 1000 (livestock)
  - > 200K with N_e = 10,000 (human)
- But, it may matter with very large N_e
  - Multi-breeds or multi-ethnicities

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

- ■ To maximise prediction accuracy
  - – Sample availability is much higher for larger N_e (in terms of sample size)
  - – e.g. close relatives < flocks (local, village) < states < country < …
- ■ Heterogeneous stocks are important as well
  - – Unlimited source
    - • Common SNP chips across breeds or ethnicities
    - • Getting cheaper

Implication

- To maximise prediction accuracy
  - Composite design would be desirable
    - N_e=1000 (N=10,000) + N_e=100 (N=500) + N_e=10 (N=50)
- It may useful if one can get the expected prediction accuracy before conducting experiment. For example,
  - When adding a bunch of heterogeneous stocks to your data, how much the accuracy can be increased?
  - When adding a number of newly genotyped individuals, what accuracy can you expect?
  - And, what is the power?

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