Accuracy of genomic evaluations depends on distance to the reference data

-- Efficiency of sharing genomic information for genomic prediction

Guosheng Su, Gang Guo, and Mogens Sandø Lund

University of Aarhus, Faculty of Agricultural Sciences,
Department of Genetics and Biotechnology, Tjele, Denmark

Background

- Share genomic information to improve genetic evaluation
- How large benefit from sharing genomic data?

Objective

To investigate the efficiency of genomic prediction using:

- Foreign reference data
- Pooled reference data from target and other populations

Data: Simulated data

Simulation scenarios - Genome

Genome: 15 chromosomes, 1200 cM

SNP: 20,000 SNP

Distributed in equal space

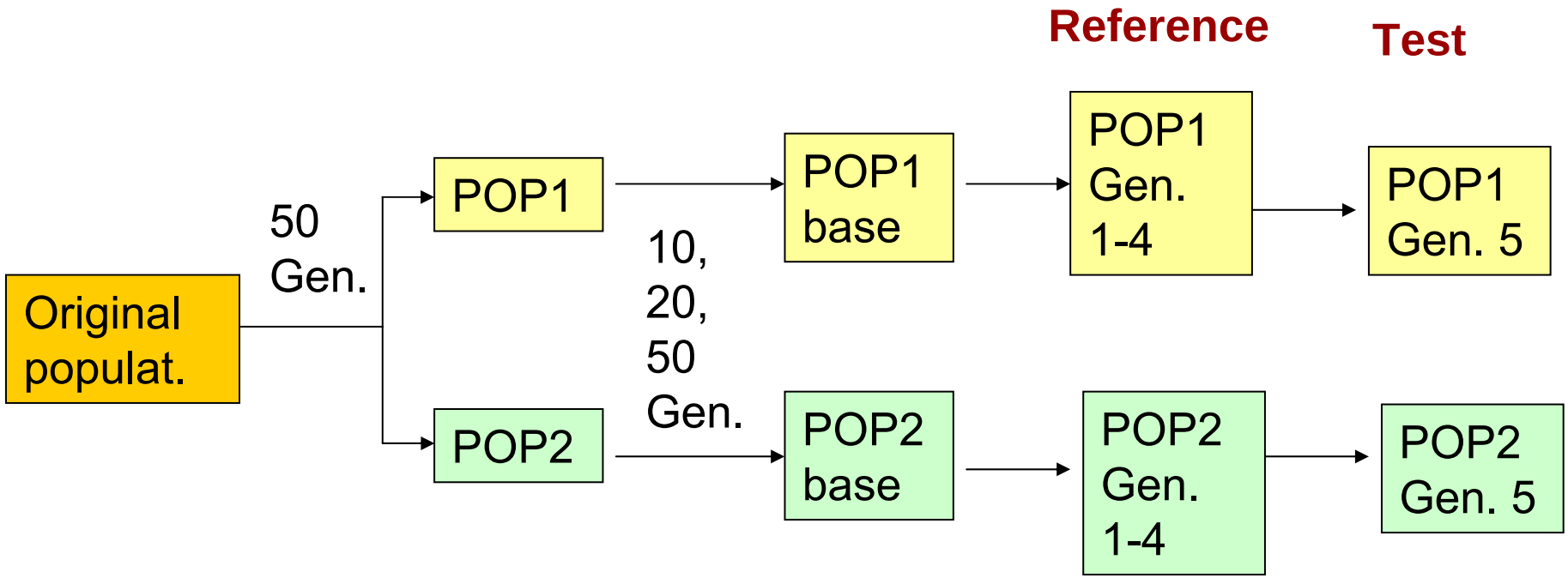
QTL: 500 QTLs

Randomly distributed
Effects of QTLs ~ Gamma(5.4, 0.84)

P = 0.5 for each SNP and QTL in original base generation

No new mutation

Simulation scenario – Population structure



50 sires
50 dams

50 sires, 50 dams per gen.

500 sires, 25000 dams per gen.

Simulation scenarios – BV and phenotypic value

TBV = sum of QTL effects

$Y = TBV + e$

$e \sim N(0, (1-h^2)V_a/h^2)$

$h^2 = 0.3$

5 replicates
for each scenario (10, 20 or 50 diverged generations)

Model to estimate SNP effect for genomic prediction

Bayesian model $\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{m} \mathbf{X}_i \mathbf{q}_i v_i + \mathbf{e}$

y: DYD
μ: intercept
$\mathbf{q}_i$ scaled SNP effects
v_i: scaling factor.

Prior distribution of $\mathbf{q}_i$ and v_i,

$\mathbf{q}_i \sim N(\mathbf{0}, \mathbf{I})$, $v_i \sim N(0, \sigma_v^2)$

(Meuwissen and Goddard, 2004; Villumsen et al., 2009)

Criteria to evaluate efficiency of genomic prediction

1. Reliability of GEBV

$R^2_{GEBV} = Cor^2(GEBV, TBV)$

2. Accuracy of estimates of prediction error variances (PEV)

Let SE = Sqrt(PEV) = Posterior STD of GEBV

95% Confidence interval of GEBV CI = GEBV±1.96SE

If frequency of TBV out of CI significantly differs from 5%, SE is not reliable.

Table 1. Number of markers fixed, mean of minor allele frequency (MAF), degree of LD (r^2), and correpation of LD phases between populations ($Cor_{r1, r2}$, De Roos et al. 2007)

Scenario (No. gen. diverge)	No. markers fixed		MAF		r^2		$Cor_{r1, r2}$
	POP1	POP2	POP1	POP2	POP1	POP2	
G10	544	550	0.30	0.30	0.23	0.23	0.87
G20	1000	993	0.29	0.29	0.26	0.26	0.78
G50	2930	2802	0.27	0.27	0.36	0.35	0.64

Table 2. Comparison on efficiency of genomic prediction based on reference data from target population or from foreign population

Referen. ->	POP1 N=2000			
Test ->	POP1		POP2	
Scenario	R^2_{GEBV}	Out CI %	R^2_{GEBV}	Out CI %
G10	0.70	5.1	0.44	8.0
G20	0.67	4.1	0.32	17.2
G50	0.69	4.7	0.13	23.5

Low reliability **Unreliable SE**

Table 3. Comparison on efficiency of genomic prediction based on pooled reference data (n=4000) or reference data from target population (n=2000)

Referen. ->	POP1 + POP2 N= 2000 + 2000				POP1 N = 2000	POP2 N =2000
Test ->	POP1		POP2		POP1	POP2
Scenario	R^2_{GEBV}	Out CI	R^2_{GEBV}	Out CI	R^2_{GEBV}	R^2_{GEBV}
G10	0.77	5.1%	0.79	4.6%	0.70	0.72
G20	0.74	6.0%	0.75	5.9%	0.67	0.68
G50	0.75	5.4%	0.75	5.2%	0.69	0.69

High reliability

Reliable SE

Conclusions

- Using foreign reference data directly to predict GEBV is not effective (low reliability, unreliable SE), dependent on the genetic distance

- Pooled reference data increases reliability of GEBV because of increasing size of reference data