Genomic prediction and GWAS with sequence information versus HD or 50k SNP chips

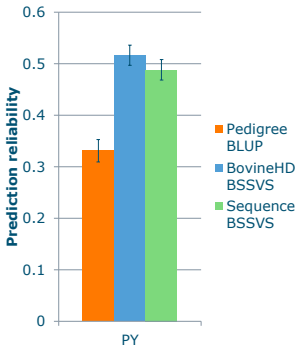
Roel Veerkamp, Aniek Bouwman

Background

- Whole genome sequence data
  - Causal mutation (QTN) is included
  - No dependency on LD between SNP and QTL
- Expected to perform better
  - GWAS
  - WGS: More persistent across generations / breeds

Prediction reliability

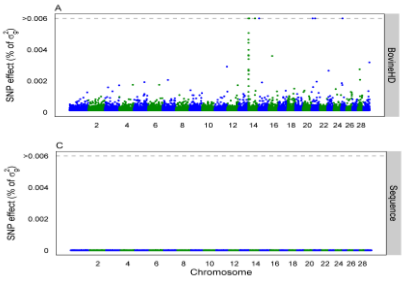
- Not any benefit yet. But it should ….!?
- What are we doing wrong?



GSE

Genomic Prediction with 12.5 Million SNPs for 5503 Holstein Friesian Bulls
R. van Binsbergen[1,2], M.P.L. Calus[1], M.C.A.M. Bink[2], C. Schrooten[2], F.A. van Eeuwijk[2], R.F. Veerkamp[1]

Identifying QTN with GS?



HD SNP: 631,428

SEQ SNP: 12,590,056

Genomic Prediction with 12.5 Million SNPs for 5503 Holstein Friesian Bulls
R. van Binsbergen[1,2], M.P.L. Calus[1], M.C.A.M. Bink[2], C. Schrooten[2], F.A. van Eeuwijk[2], R.F. Veerkamp[1]

Objective of this study

- The potential benefit of sequence data, compared to usual SNP chip, for
  - QTL detection

  → genomic prediction
  - How much genetic variation is explained?
  - Prediction accuracy genomic selection?

Method (1): Imputation to sequence

5556 Holstein Friesian bulls CRV
777K SNP genotypes (Illumina BovineHD BeadChip)

1147 animals (multiple breeds)
28M SNP (whole-genome sequence data)
1000 bull genomes project Run 4

Minimac v2

5556 Holstein Friesian bulls imputed sequence and de-regressed BV for protein yield
3469 bulls used for discovery and training & 2287 bulls used for validation

Aniek Bouwman

Method (2): Statistics GWAS

- Single SNP regression (program GCTA)
  - Include GRM based on HD SNP set
  - MAF >0.01 (13,789,029 SNP)
- Conditional and joint multiple SNP GWAS **(COJO)**
  - Stepwise selection of SNP explaining additional variance

nature genetics

Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits

Method (3): SNP set selection

11 SNP sets selected (based on SNP chip/ significance from GWAS):

	Sequence	HD	50k	COJO
All	1	2	3	
-log(p)>3	4	5	6	7
-log(p)>5	8	9	10	11

How good are these SNP sets for genomic prediction?

Method (4): Two validation methods

Which is the "best" SNP set and how much "better"?

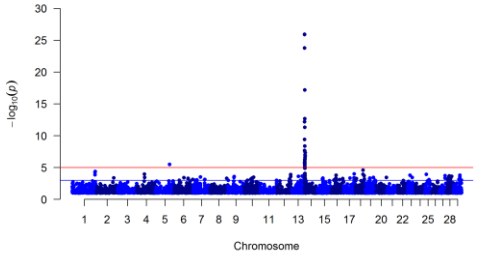
1. Estimate heritability in validation animals using GRM**s** based on selected sets of SNP
2. Train GRM**s** on discovery animals, back solve SNP and predict DGV for 2287 validation animals. Correlate DGV with phenotypes.

Results: number of SNP

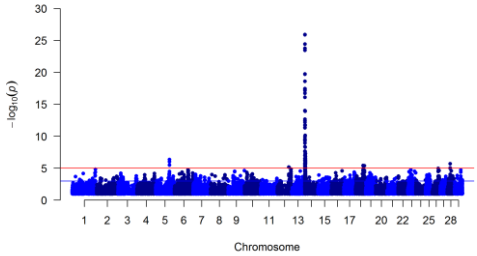
GRMs	Sequence	HD	50k	COJO
All	13,789,029	656,044	49,580	
-log(p)>3	24,387	1,238	120	119
-log(p)>5	2,194	159	27	49

Many more (significant) SNP with sequence info
Reduced with COJO to 49 SNP explaining genetic variance

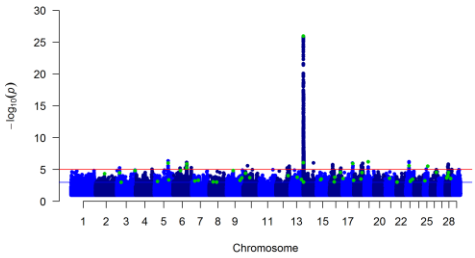
Results: 50K



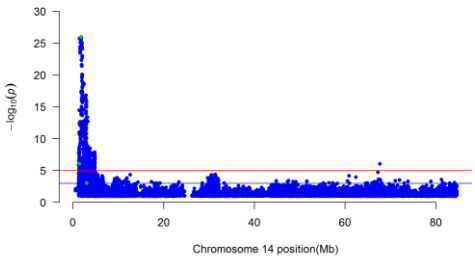
Results: HD



Results: Sequence + cojo5



Results: Cojo5 on Chr14 (DGAT)



Results: Heritability GRMs

h^2 is %variance explained by GRMs

GRMs	Sequence	HD	50k	COJO
All	0.83	0.82	0.81	
-log(p)>3	0.53	0.40	0.22	0.24
-log(p)>5	~~0.60*~~	~~0.43*~~	~~0.22*~~	0.16

*Scale problems with GRM when estimating variances

Considerable reduction when selecting SNP

GRMs	Sequence	HD	50k	COJO
All	13,789,029	656,044	49,580	
-log(p)>3	24,387	1,238	120	119
-log(p)>5	2,194	159	27	49

Results: Heritability GRMs + GRMc

variance explained by selected SNP GRM, whilst accounting for GRMc

All	Sequence	HD	50k
GRMs	0.83	0.78	0.70
GRMc	-	0.04	0.12

Similar LogL when fitting GRMs or GRMc separate

-log(p)>3	Sequence	HD	50k	COJO5
GRMs	0.19	0.15	0.09	0.11
GRMc	0.61	0.65	0.73	0.71

LogL better compared with other models even full sequence

Results: Genomic prediction

Correlation between genomic breeding value and phenotype

GRMs	Sequence	HD	50k	COJO
All	0.68	0.68	0.68	
-log(p)>3	0.58	0.56	0.42	0.38
-log(p)>5	0.39	0.30	0.28	0.31

Separating GS+, SIRE+, SMGS+ to random to conclude

GRMs	Sequence	HD	50k	COJO
All	13,789,029	656,044	49,580	
-log(p)>3	24,387	1,238	120	119
-log(p)>5	2,194	159	27	49

Conclusions

- Simple using sequence within Holstein population, unlikely to improve GS, but helps QTL detection.

→ Another approach?

- Subsets of selected SNP always poorer h^2 and GS
  - Full seq. accuracy GS of 0.68 and h^2 =0.83
  - 51 SNPs accuracy GS of 0.31 and h^2 ≈0.16
- Good way to get realistic expectations from GWAS+QTL.

Acknowledgements

1000 bull genomes consortium
www.1000bullgenomes.com