Incorporating QTL information in GBLUP and machine learning models for genomic prediction

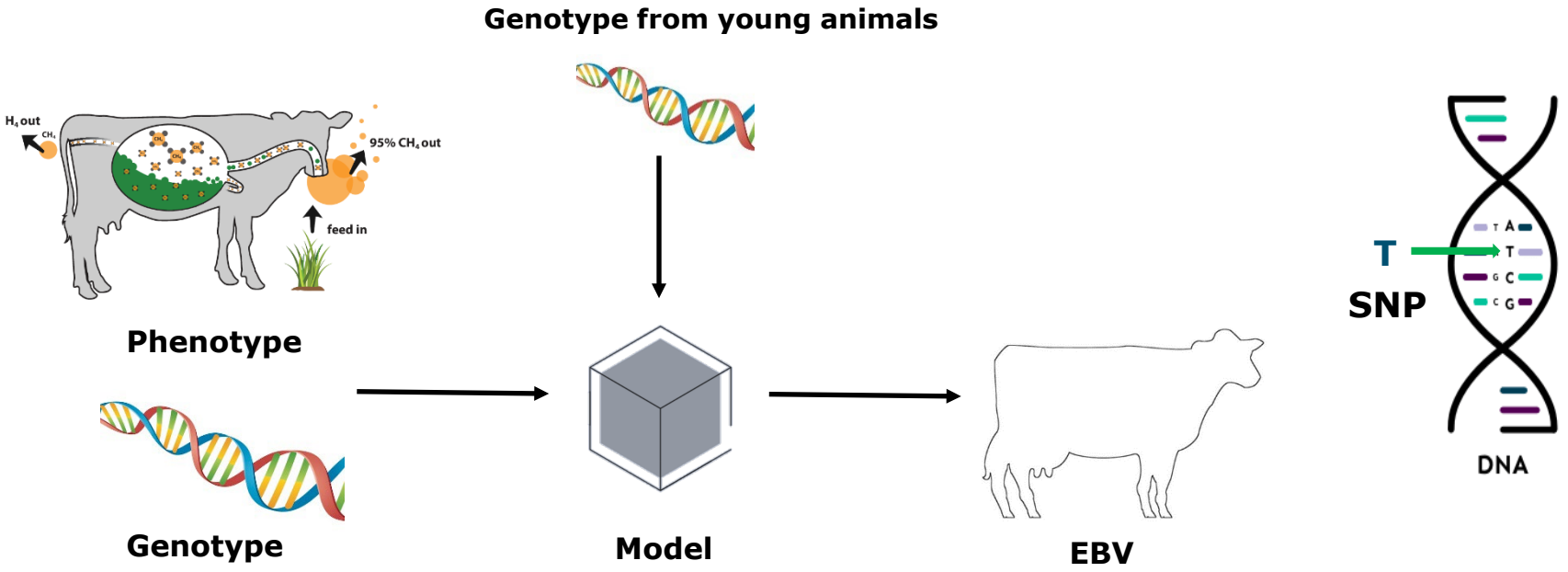
Jifan Yang[1]*, Mario P. L. Calus[1], Theo H. E. Meuwissen[2], Yvonne C. J. Wientjes[1], Pascal Duenk[1]

[1] Animal Breeding and Genomics, Wageningen University & Research, The Netherlands

[2] Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway

WAGENINGEN
UNIVERSITY & RESEARCH

Genomic prediction



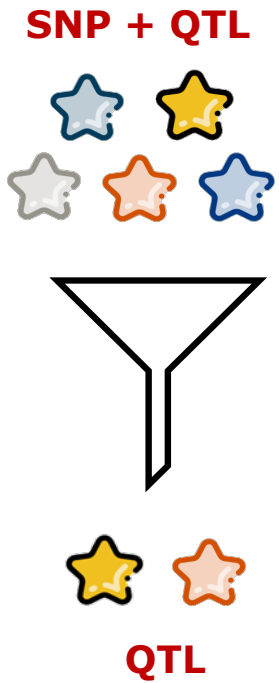
Motivation and Aim

Motivation:

- GBLUP gives SNP and QTL same weight
- Machine learning (ML) models can do feature selection

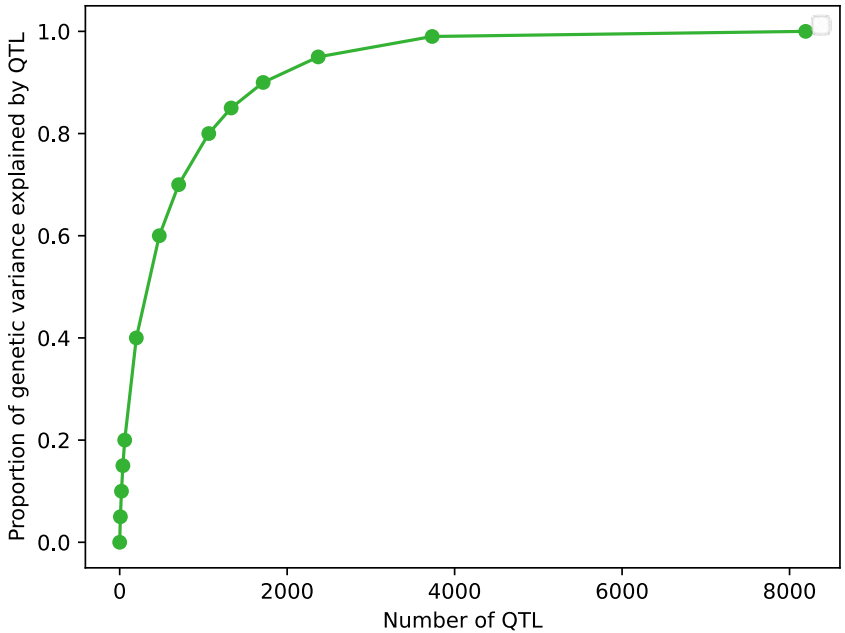
Aim:

Investigate benefit of including QTL genotypes in ML and GBLUP



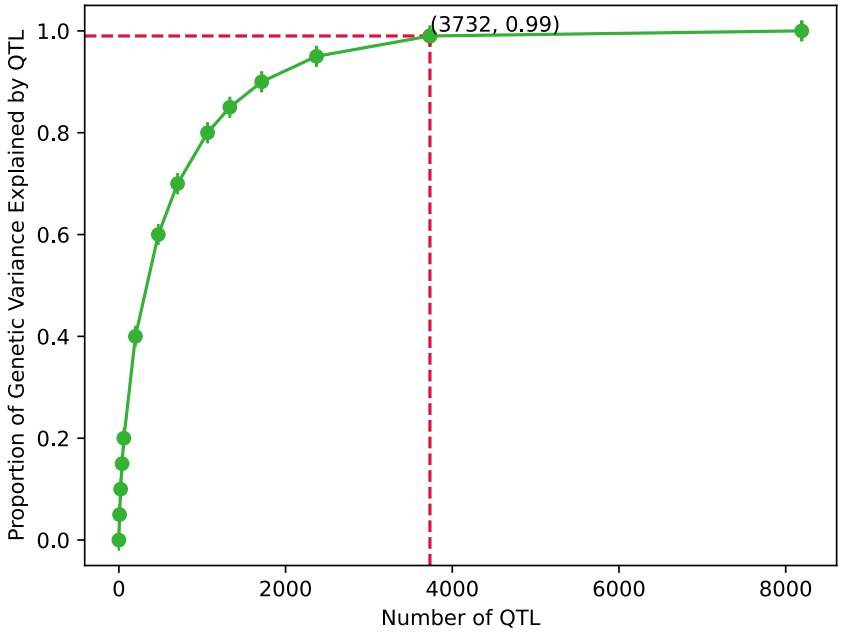
Simulation

- Population under selection
- Additive effect only
- QTL ~ **Γ**, SNP in LD with QTL
- Gen 10-15 for training
- Gen 16 for test



Simulation

- Population under selection
- Additive effect only
- QTL ~ Γ, SNP in LD with QTL
- Gen 10-15 for training
- Gen 16 for test



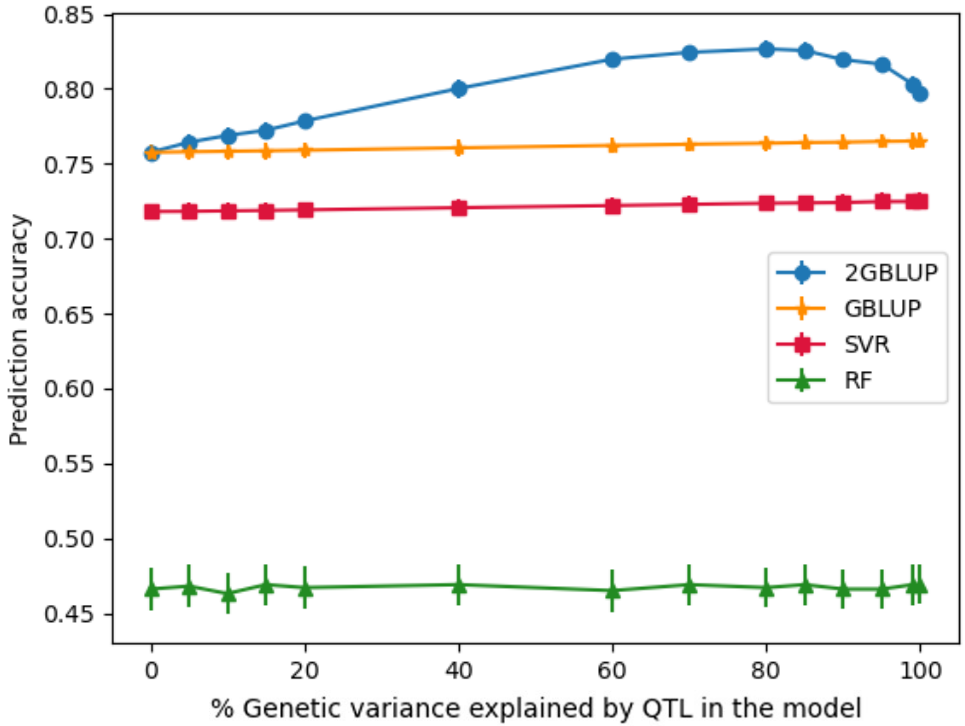
WAGENINGEN
UNIVERSITY & RESEARCH

Models

- GBLUP
  - SNP and QTL are mixed ($\mathbf{u}$) $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e}$
- 2GBLUP
  - SNP ($\mathbf{u_1}$) and QTL ($\mathbf{u_2}$) are separated $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu_1} + \mathbf{Zu_2} + \mathbf{e}$

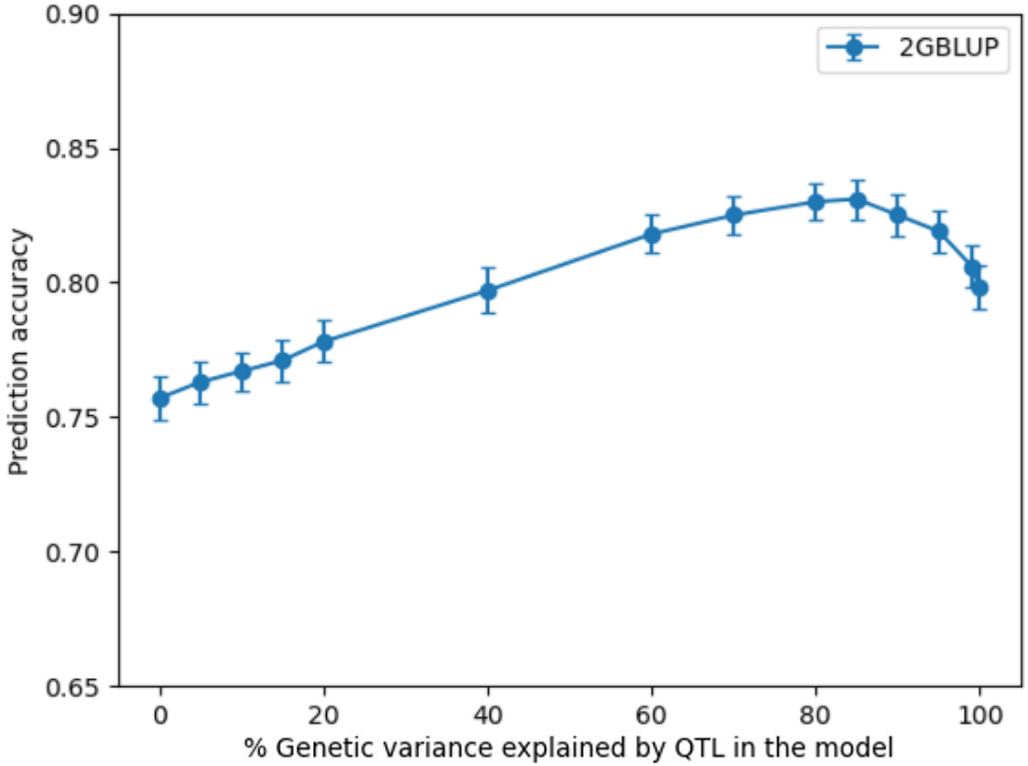
- Machine learning models:
  Random Forest (RF), Support Vector Regression (SVR)
  - SNP and QTL are mixed

Results-Prediction Accuracy



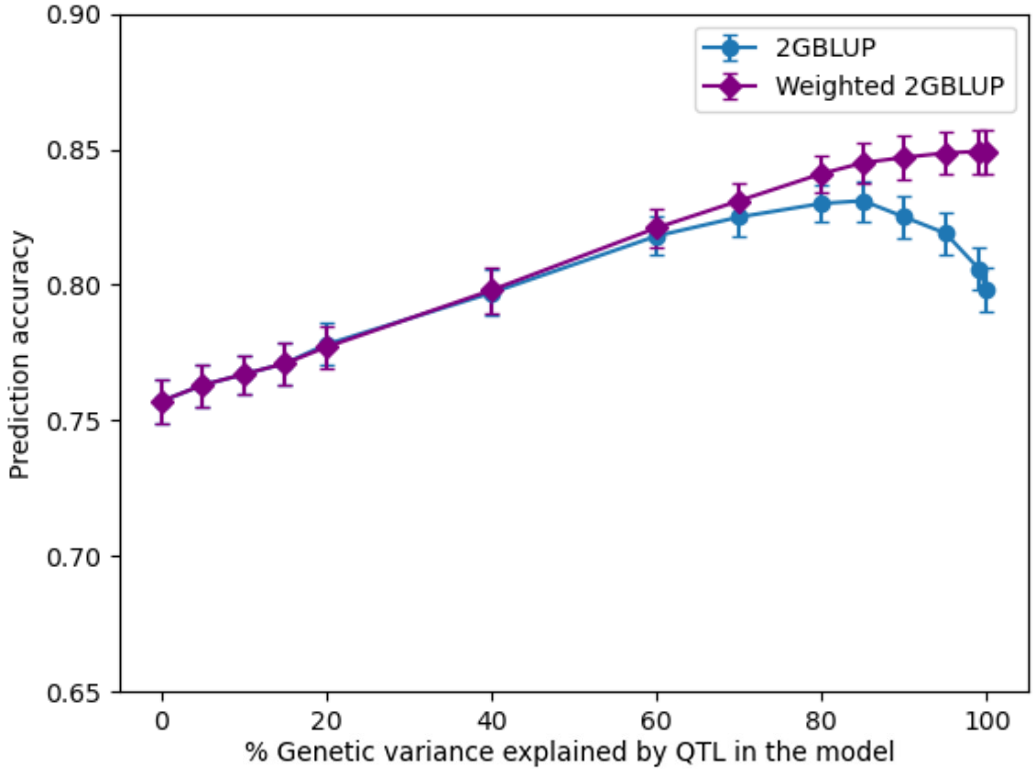
WAGENINGEN
UNIVERSITY & RESEARCH

Results-2GBLUP



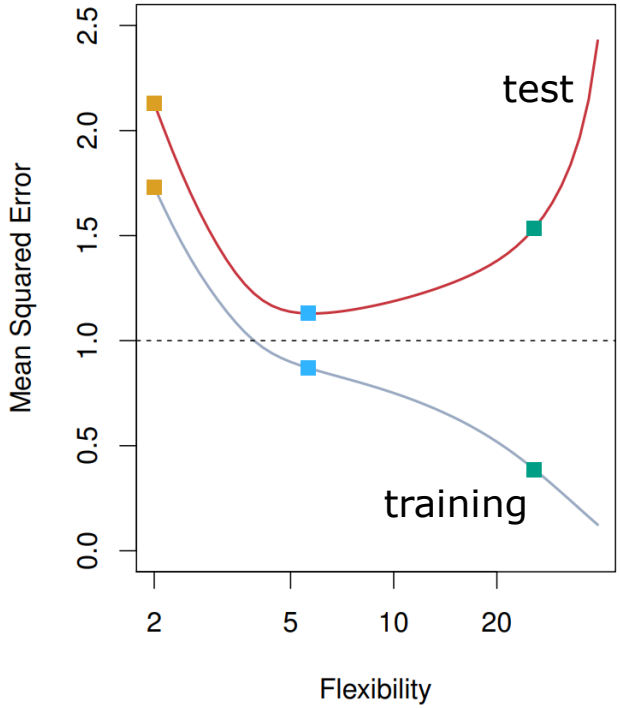
WAGENINGEN
UNIVERSITY & RESEARCH

Results-Weighted 2GBLUP



WAGENINGEN
UNIVERSITY & RESEARCH

Results-overfitting of RF

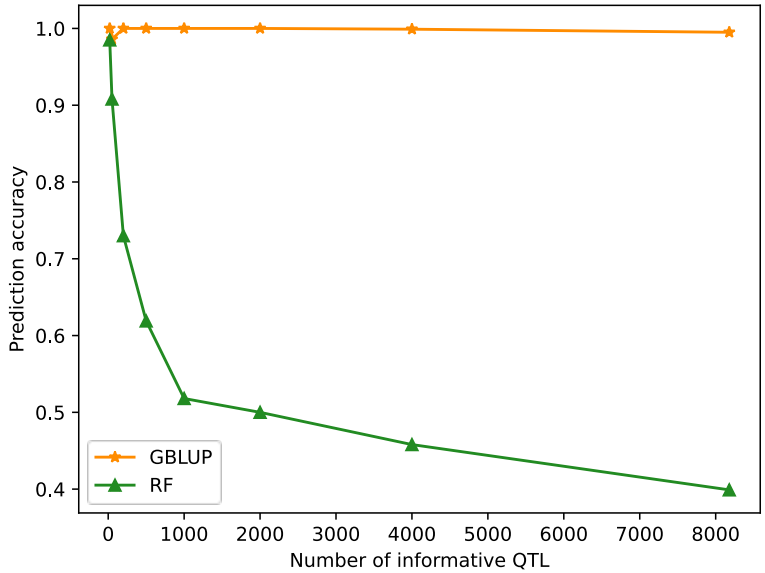


(James et al., 2021)

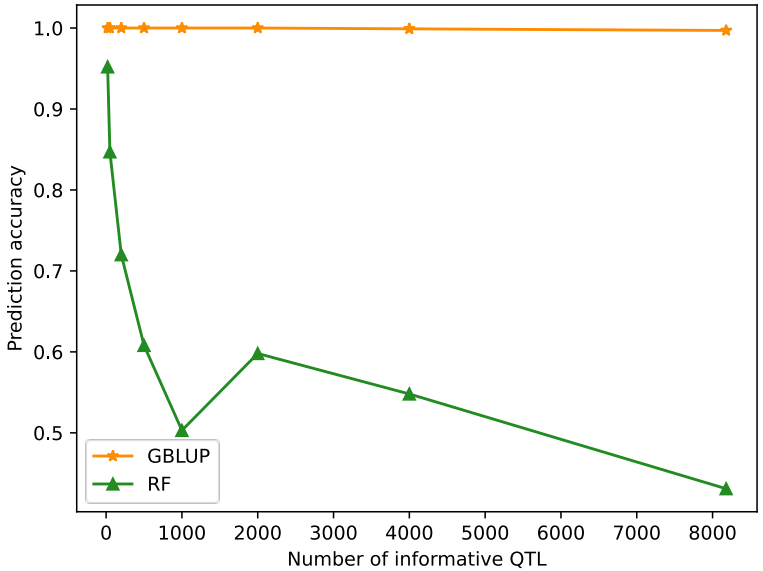
	Correlation (predicted value, training phenotype)
2GBLUP	0.66
RF	0.99
SVR	0.81

Influence of feature numbers

Genetic effect only



QTL only

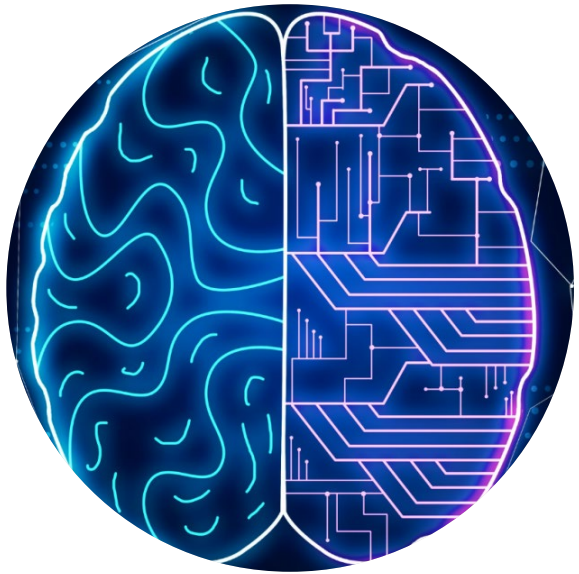


QTL + 10,000 SNP

Take-home message

- (weighted) 2GBLUP can benefit from including QTL
- RF and SVR did not benefit from including QTL
- Random forest performed poorly:
    - overfitting
    - may work when few QTL explain most of genetic variance

Thank you for your attention.



Take-home message

- (weighted) 2GBLUP can benefit from including QTL
- RF and SVR did not benefit from including QTL
- Random forest performed poorly:
    - overfitting
    - may work when few QTL explain most of genetic variance