Accuracy of imputation and prediction in dairy cattle using whole-genome sequence data

Rianne van Binsbergen[*,§], Mario P.L. Calus[§], Marco C.A.M. Bink[*], Fred A. van Eeuwijk[*], and Roel F. Veerkamp[§]

Expectation

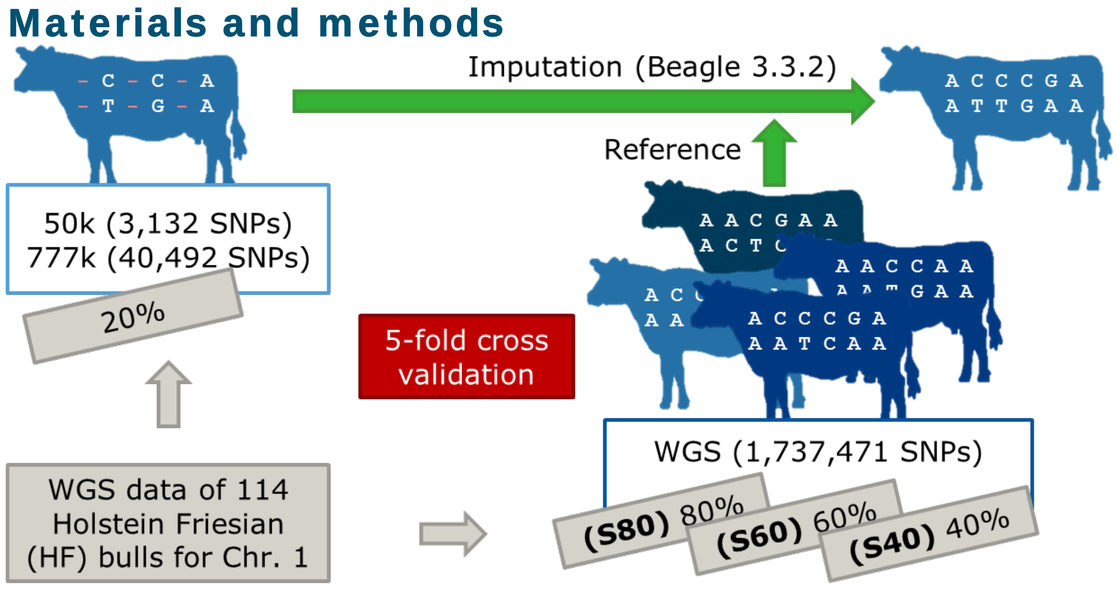
Whole-genome sequence (WGS) data enables estimation of direct effects of actual mutations on a trait, which is expected to increase prediction accuracy compared to currently used commercial SNP genotype panels.

Challenges

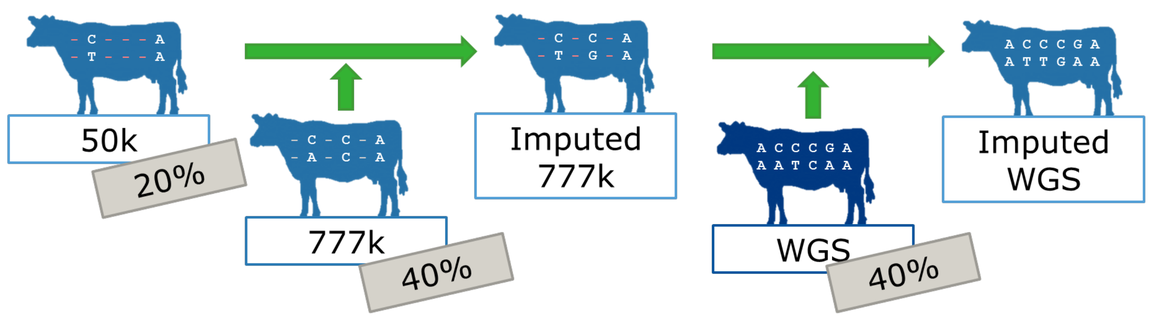
1) A large number of sequenced individuals is needed. However, sequencing is still too expensive.
2) So far, the utilization of WGS data for genomic prediction in cattle was not studied with real data.

OBJECTIVE (1)
Study accuracy of genotype imputation using WGS

Materials and methods



Sequential imputation



Results

Mean (and SD) accuracy of imputation (correlation between true and imputed genotype) per SNP

Scenario	50k → WGS	777k → WGS
S40	0.37 (0.36)	0.77 (0.29)
S60	0.43 (0.36)	0.81 (0.27)
S80	0.46 (0.37)	0.83 (0.27)
Sequential	0.65 (0.30)	

CONCLUSIONS (1)
777k → WGS: generally high imputation accuracy
50k → WGS: (too) low imputation accuracy
- Sequential imputation improved accuracy

Van Binsbergen, *et al*. 2014. Genet Sel Evol 46:41

OBJECTIVE (2)
Study reliability of genomic prediction using WGS

Materials and methods

777k and imputed WGS data for 5503 HF bulls
- 2087 validation (youngest bulls) and 3416 training

BLUP (*best linear unbiased prediction*)

BLUP_P: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Za} + \boldsymbol{e}$

BLUP_G: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zg} + \boldsymbol{e}$

$\boldsymbol{y}$ vector of phenotypes, μ overall mean, $\boldsymbol{Z}$ design matrix,

$\boldsymbol{a}$ additive genetic effects; $\mathbf{a}|\mathbf{A}, \sigma_g^2 \sim N(\mathbf{0}, \mathbf{A}\sigma_g^2)$
- $\boldsymbol{A}$ numerator relationship matrix (pedigree)

$\boldsymbol{g}$ genomic values; $\boldsymbol{g}|\boldsymbol{GRM}, \sigma_g^2 \sim N(\mathbf{0}, \boldsymbol{GRM}\sigma_g^2)$
- $\boldsymbol{GRM}$ genomic relationship matrix [1]

$\boldsymbol{e}$ residual effects; $\boldsymbol{e}|\boldsymbol{D}, \sigma_e^2 \sim N(\mathbf{0}, \boldsymbol{D}\sigma_e^2)$
- $\boldsymbol{D}$ diagonal matrix containing weights of phenotypes

BSSVS[2] (*Bayes stochastic search variable selection*)

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zu} + \boldsymbol{X\alpha} + \boldsymbol{e}$$

$\boldsymbol{u}$ polygenic effects of all bulls; $\boldsymbol{u}|\mathbf{A}, \sigma_u^2 \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$

$\boldsymbol{X}$ matrix with genotypes (0,1,2) of all bulls

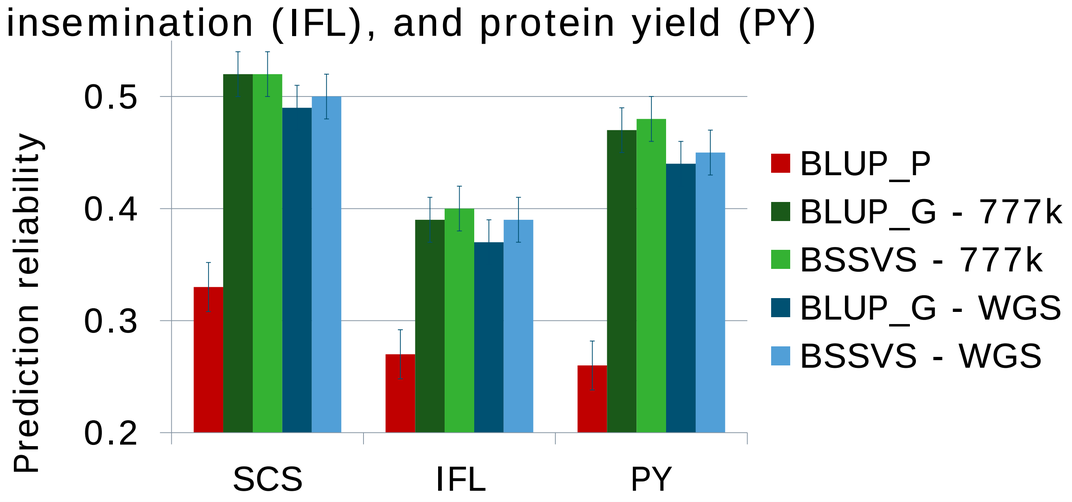
$\boldsymbol{\alpha}$ (random) allele substitution effects with prior distribution:

$$\alpha_j|I_j, \sigma_a^2 = \begin{cases} \sim N\left(0, \frac{\sigma_a^2}{100}\right) when\ I_j = 0 \text{ (SNP small effect)} \\ \sim N(0, \sigma_a^2) when\ I_j = 1 \text{ (SNP large effect)} \end{cases}$$

Prior distribution $I_j = \text{Bernoulli}(1-\pi)$

Results

Mean (and SE) prediction reliability (squared correlation original and predicted phenotype) for somatic cell score (SCS), interval between first an last insemination (IFL), and protein yield (PY)



CONCLUSIONS (2)
WGS compared to 777k did not improve genomic prediction reliability in this study

References

[1] Yang, *et al.* 2010. Nat Genet 42:565-569
[2] Calus. 2014. Genet Sel Evol 46:24

* Wageningen UR Biometris
PO Box 16, 6700 AA Wageningen, NL
www.wageningenur.nl/biometris

Contact: rianne.vanbinsbergen@wur.nl

§ Wageningen UR Livestock Research
Animal Breeding and Genomics Centre
PO Box 338, 6700 AH Wageningen, NL
www.wageningenur.nl/abgc

Acknowledgements

The 1000 bull genomes consortium
www.1000bullgenomes.com