Genotype imputation accuracy in Holstein Friesian cattle in case of whole-genome sequence data

Rianne van Binsbergen

Marco Bink, Mario Calus, Fred van Eeuwijk,
Ben Hayes, Ina Hulsegge, and Roel Veerkamp

LIVESTOCK RESEARCH
WAGENINGEN UR

Animal Breeding &
Genomics Centre

Background

Whole-genome sequence data might lead to higher accuracy in GWAS and genomic predictions

→ Causal mutation is included (*assumption*)

Large dataset is required = expensive

Solution:

→ Sequence core set of individuals (e.g. founders)

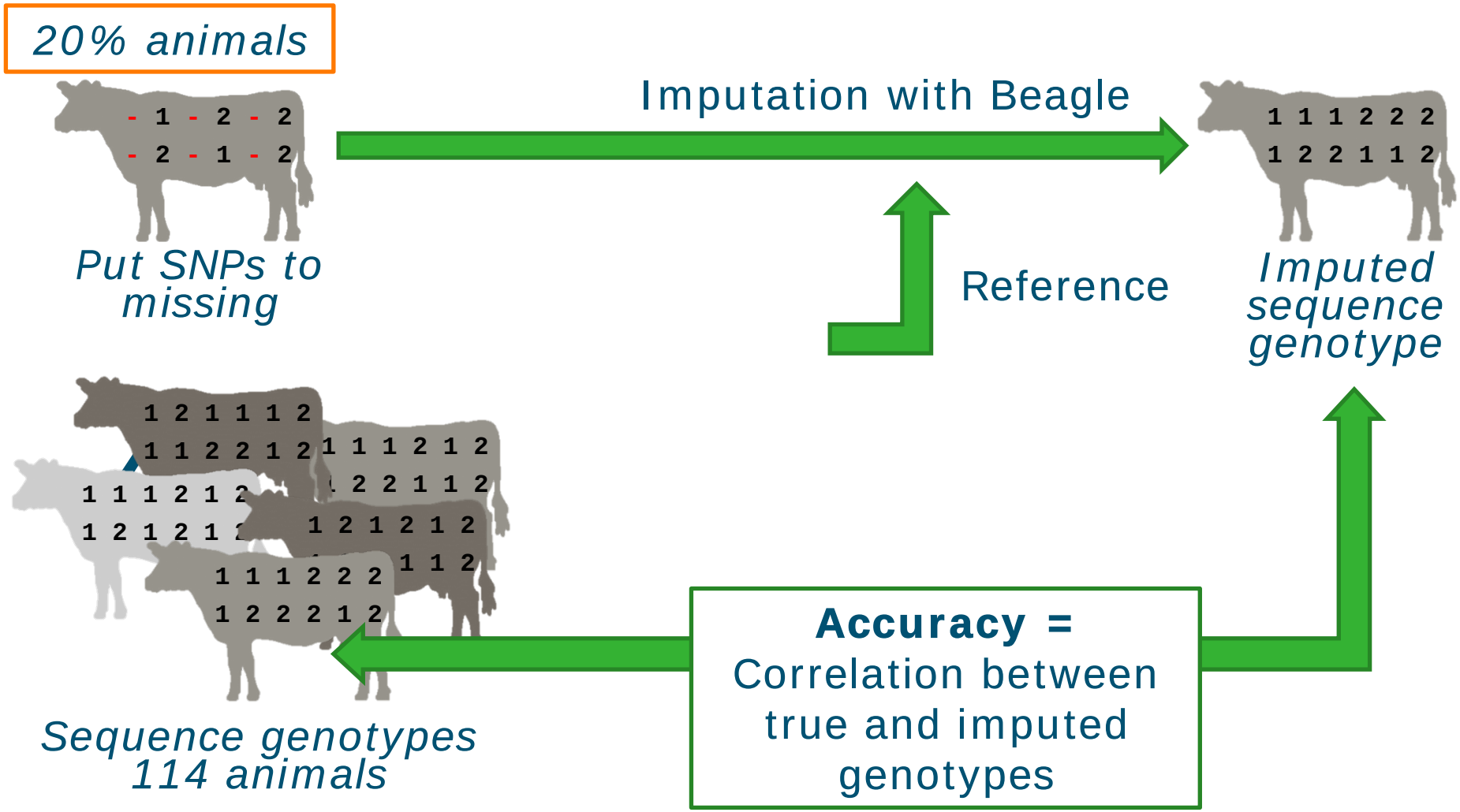
→ Impute whole-genome sequence genotypes of other individuals

Objectives

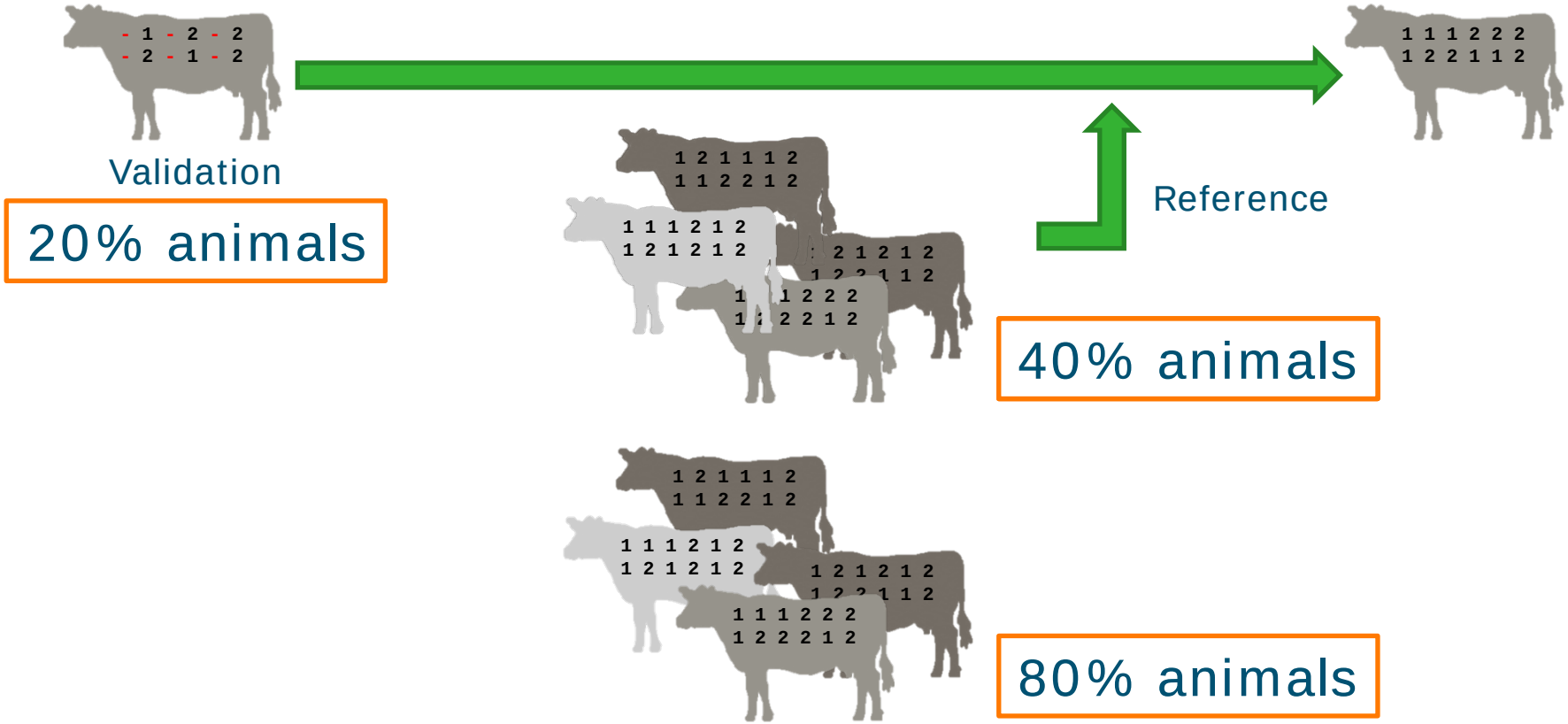
1. Investigate *mean* accuracy of imputation from SNP panel genotypes to whole-genome sequence data in Holstein Friesian dairy cattle

2. Gain insights in factors affecting accuracy of imputation *per SNP*

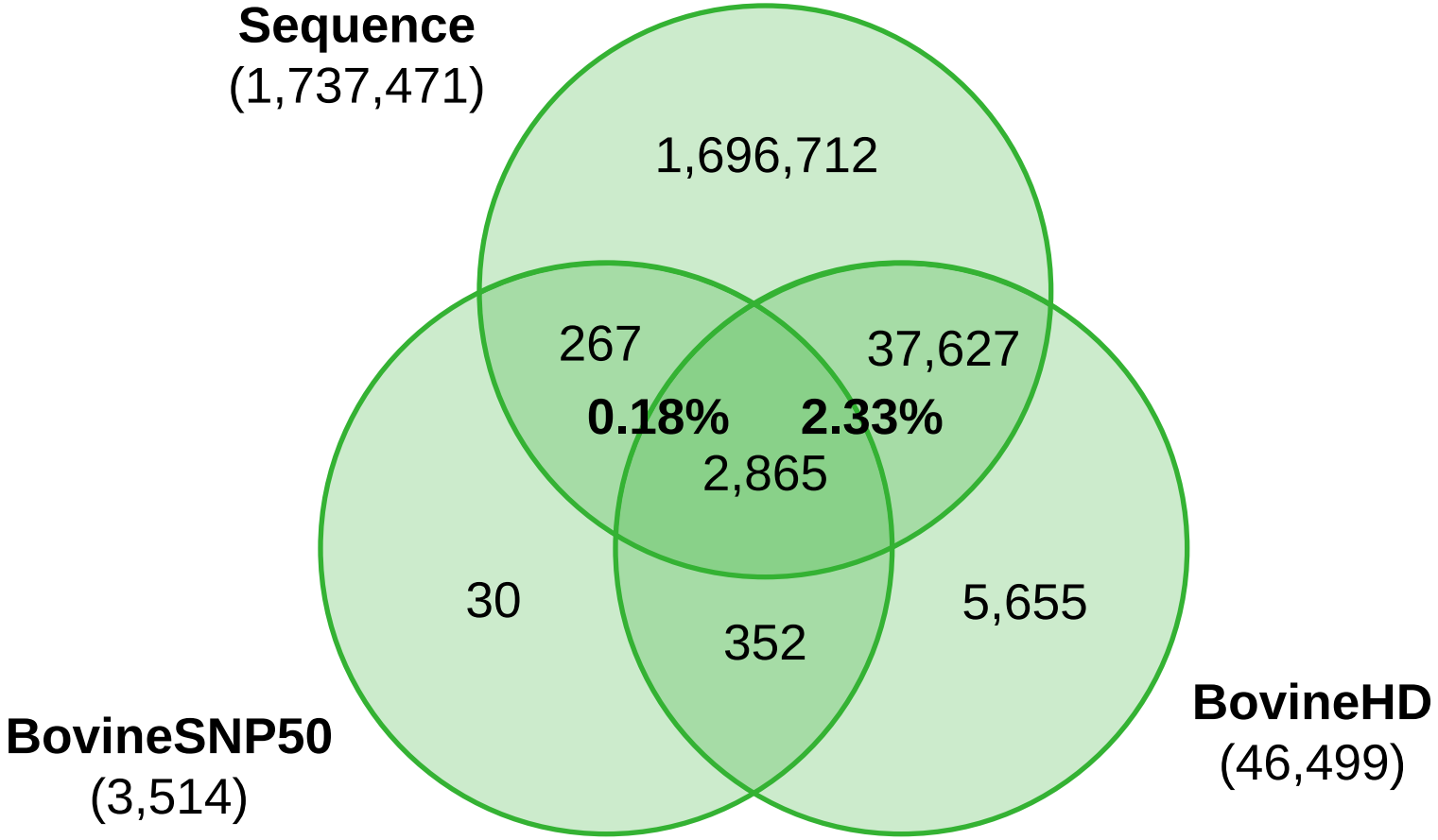
1. General approach



1. Scenarios


- 1 - 2 - 2
- 2 - 1 - 2
Validation
20% animals
1 2 1 1 1 2
1 1 2 2 1 2
1 1 1 2 1 2
1 2 1 2 1 2
2 1 2 1 2
1 1 2 2 2
2 2 1 2
Reference
1 1 1 2 2 2
1 2 2 1 1 2
40% animals
80% animals

1. Number of variants on chromosome 1



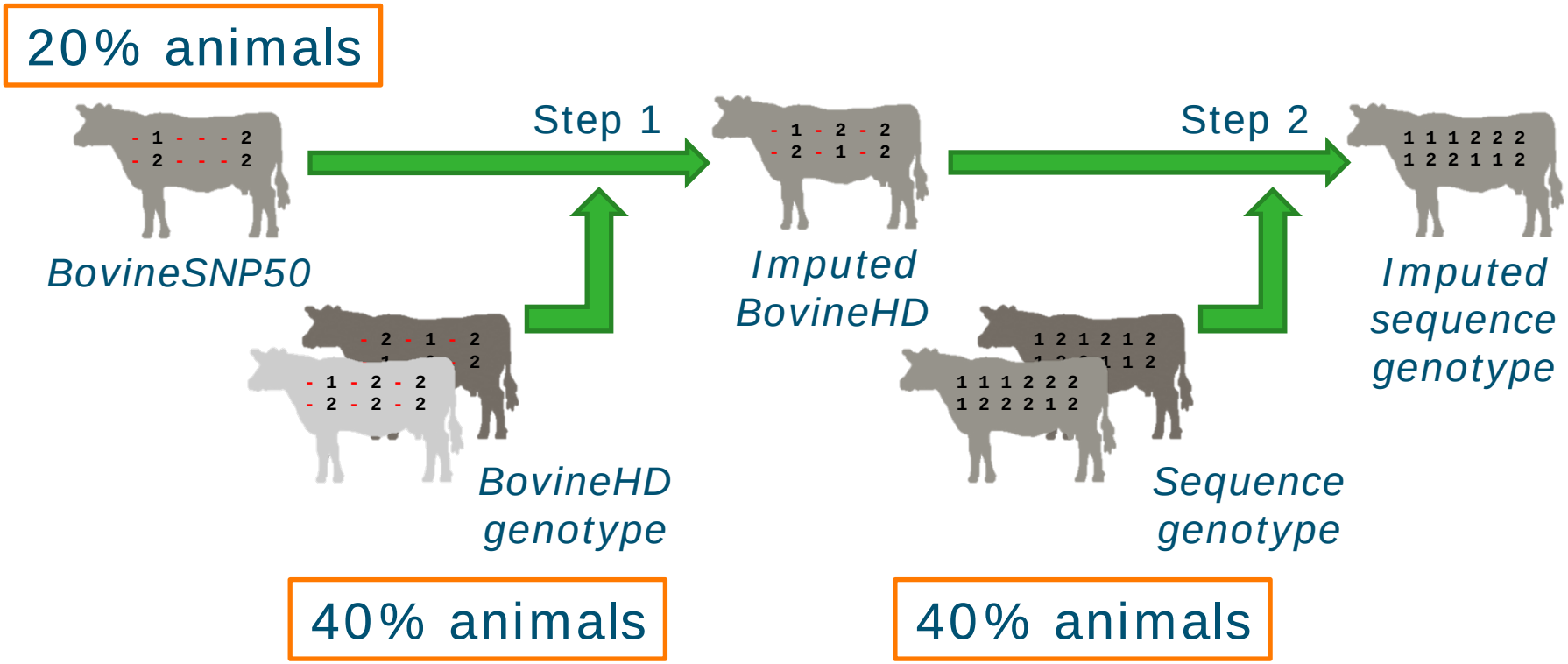
1. Mean accuracy

Scenario	BovineSNP50	BovineHD
80% animals	0.46	0.83
60% animals	0.43	0.81
40% animals	0.37	0.77

Accuracy of imputation was **(too) low**

Accuracy of imputation was **generally high**

1. Two-step approach



1. Mean accuracy

Scenario	BovineSNP50	BovineHD
80% animals	0.46	0.83
60% animals	0.43	0.81
40% animals	0.37	0.77
Two-step approach	0.65	-

Higher accuracy
while
less information
was used!

Objectives

1. Investigate *mean* accuracy of imputation from SNP panel genotypes to whole-genome sequence data in Holstein Friesian dairy cattle

2. Gain insights in factors affecting accuracy of imputation *per SNP*

2. Factors affecting imputation reliability

LD between imputed SNP and nearest SNP on SNP panel

- Distance (c) (Sved, 1971) $r_{dist}{}^2 = \frac{1}{4*Ne*c+1}$
- MAF difference (Miller, 2013) $r_{dMAF}{}^2 = \frac{1-4dMAF}{2dMAF+1}$

Number of sequenced individuals & MAF of imputed SNP

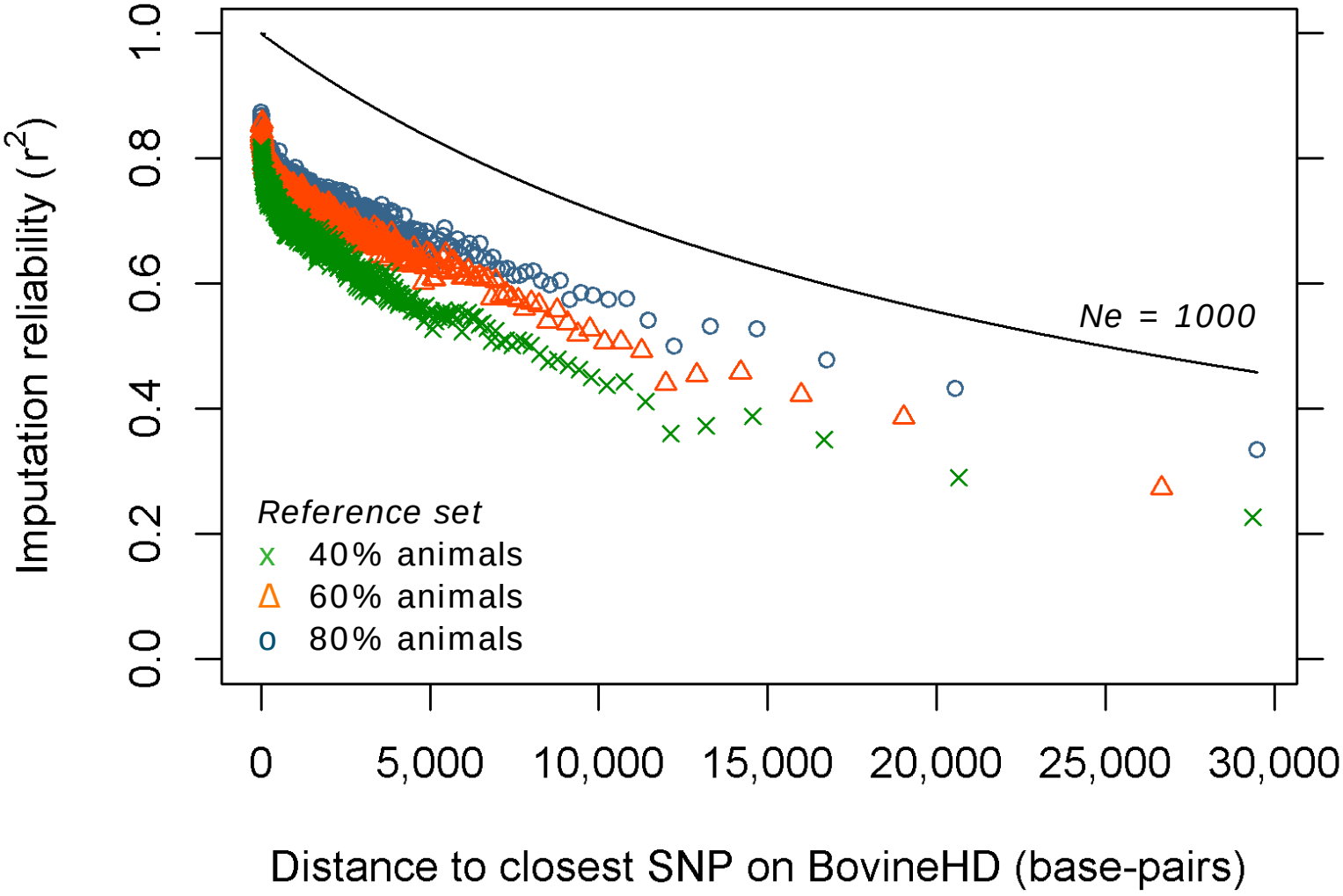
- Empirical Michaelis-Menten function per scenario

$$r_{MAF}{}^2 = \frac{V_{max}*MAF}{K_m+MAF}$$

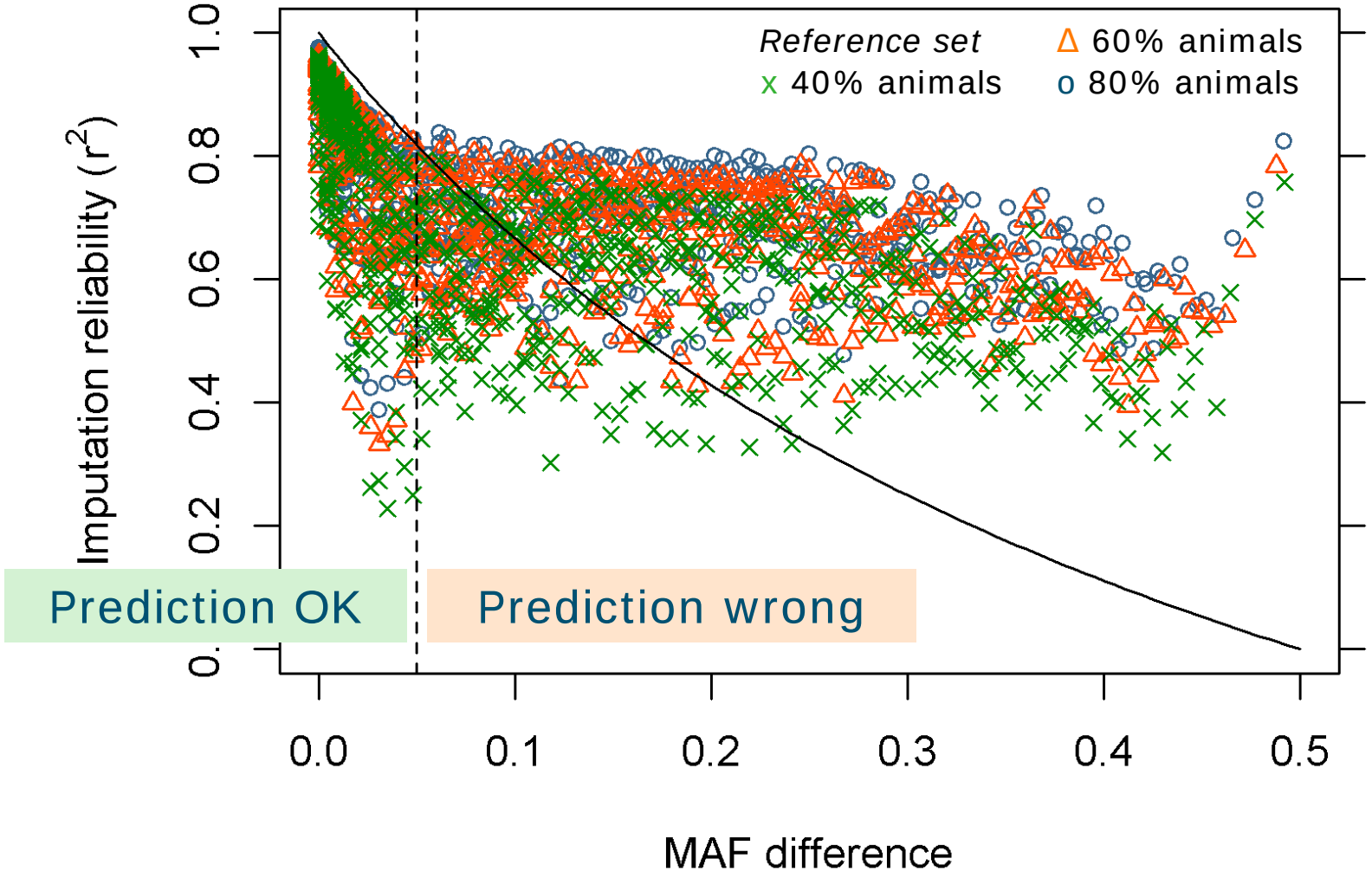
Total predicted imputation reliability = $r_{dist}{}^2 * r_{dMAF}{}^2 * r_{MAF}{}^2$

- Based on SNP in highest LD ($r_{dist}{}^2 * r_{dMAF}{}^2$) of 5 nearest SNPs on SNP chip

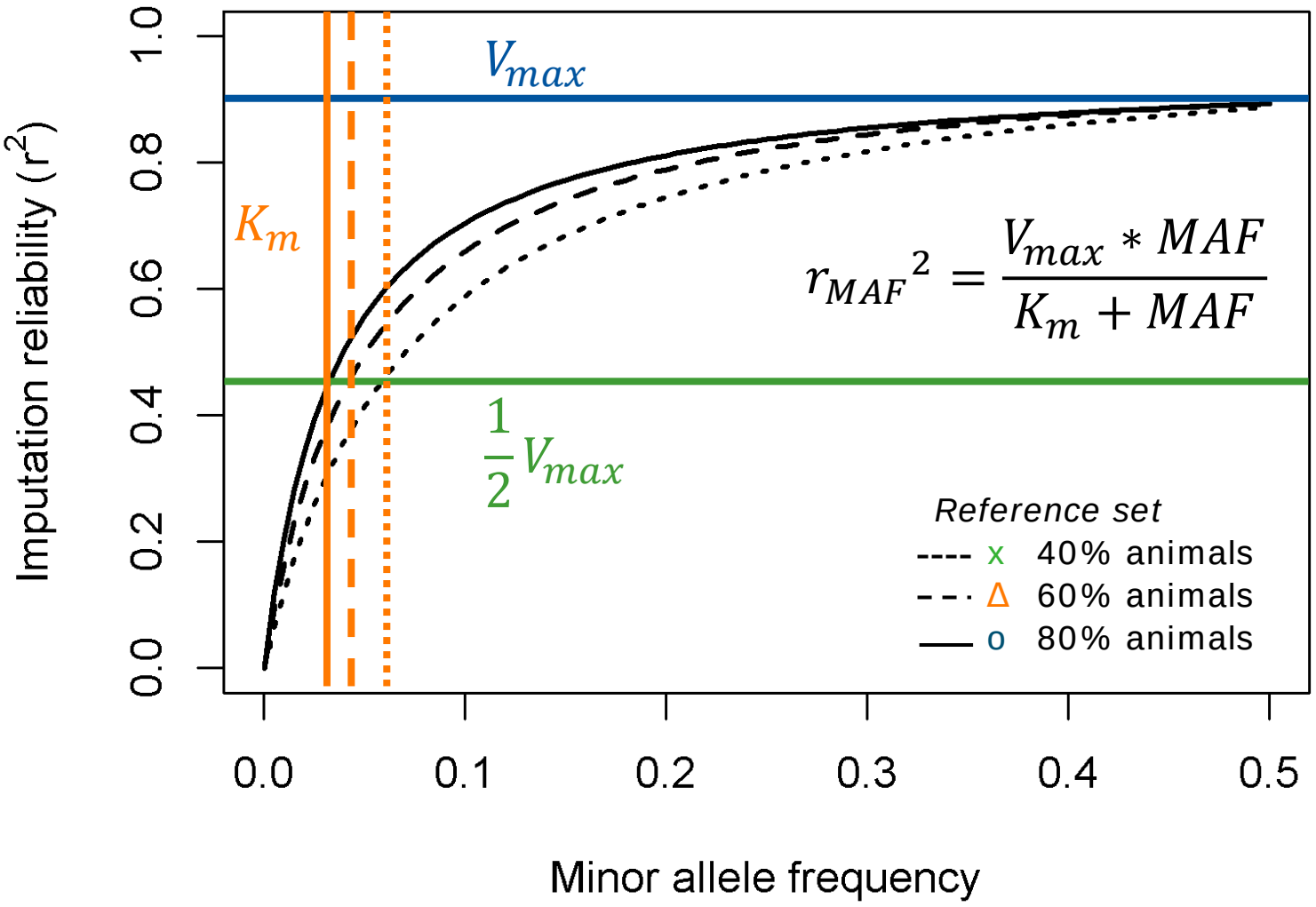
2. Distance


Imputation reliability (r^2)
0.0
0.2
0.4
0.6
0.8
1.0
Ne = 1000
Reference set
x 40% animals
Δ 60% animals
o 80% animals
0
5,000
10,000
15,000
20,000
25,000
30,000
Distance to closest SNP on BovineHD (base-pairs)

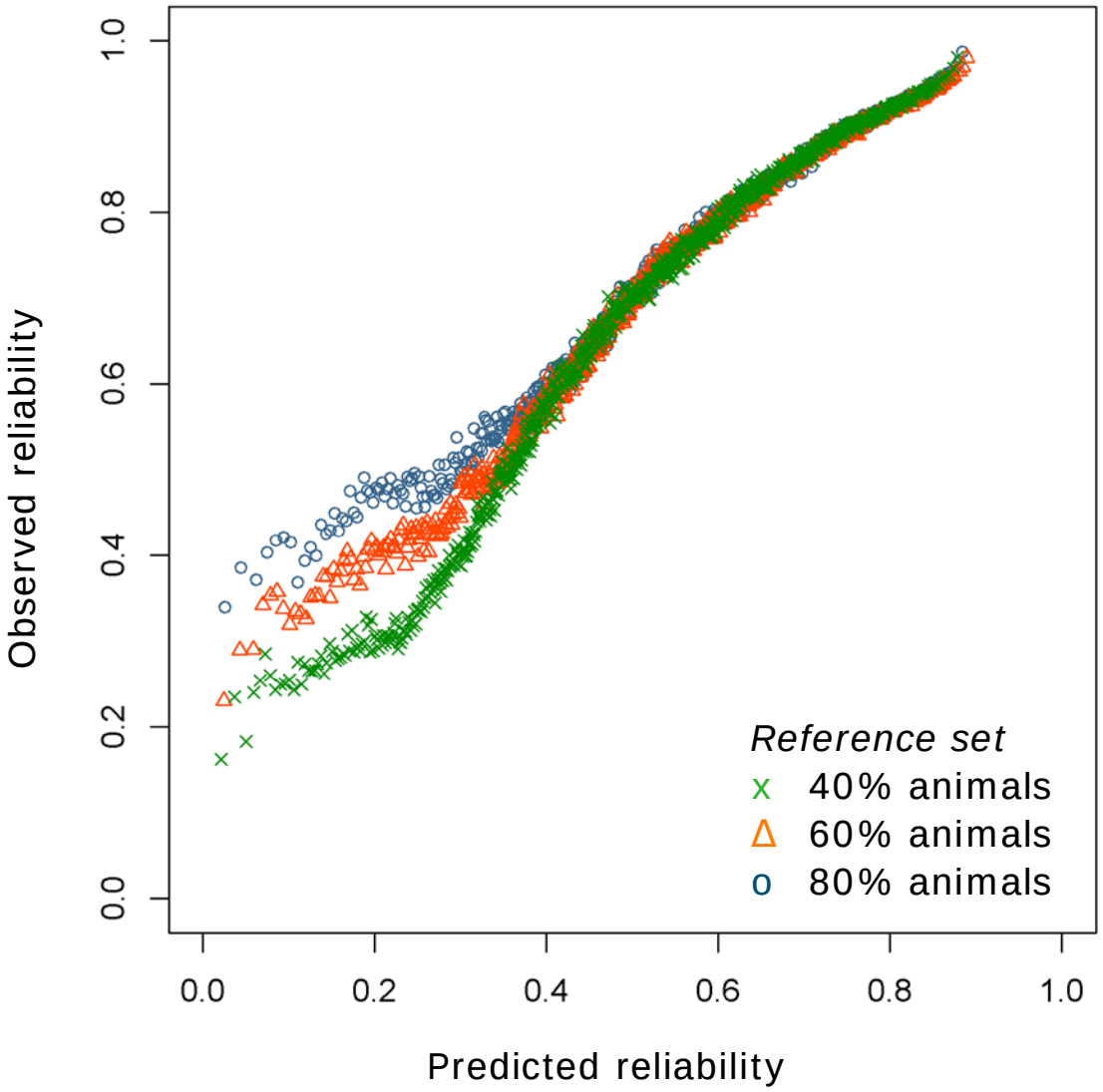
2. MAF difference



2. MAF & Reference set size



2. Total predicted reliability



Conclusions

1. Accuracy of imputation from BovineHD was generally high and for imputation from BovineSNP50 (too) low
   → Stepwise imputation improved accuracy

2. Poor imputation of sequence data variants (including causal mutation?) if
   - poor LD between imputed SNP and SNP chips
   - low MAF of imputed SNP

   → Potentially limits the extra power from using imputed sequence data for GWAS (compared to SNP chips)

Acknowledgements

- 1000 bull genomes project (www.1000bullgenomes.com)
- Breed4Food project (www.breed4food.com)

Thanks for your attention

rianne.vanbinsbergen@wur.nl