Bavarian State Research Center
for Agriculture

How do imputation errors affect genomic breeding values?

Eduardo Pimentel, C. Edel, R. Emmerling, K.-U. Götz

Institute of Animal Breeding

Motivation / Objective

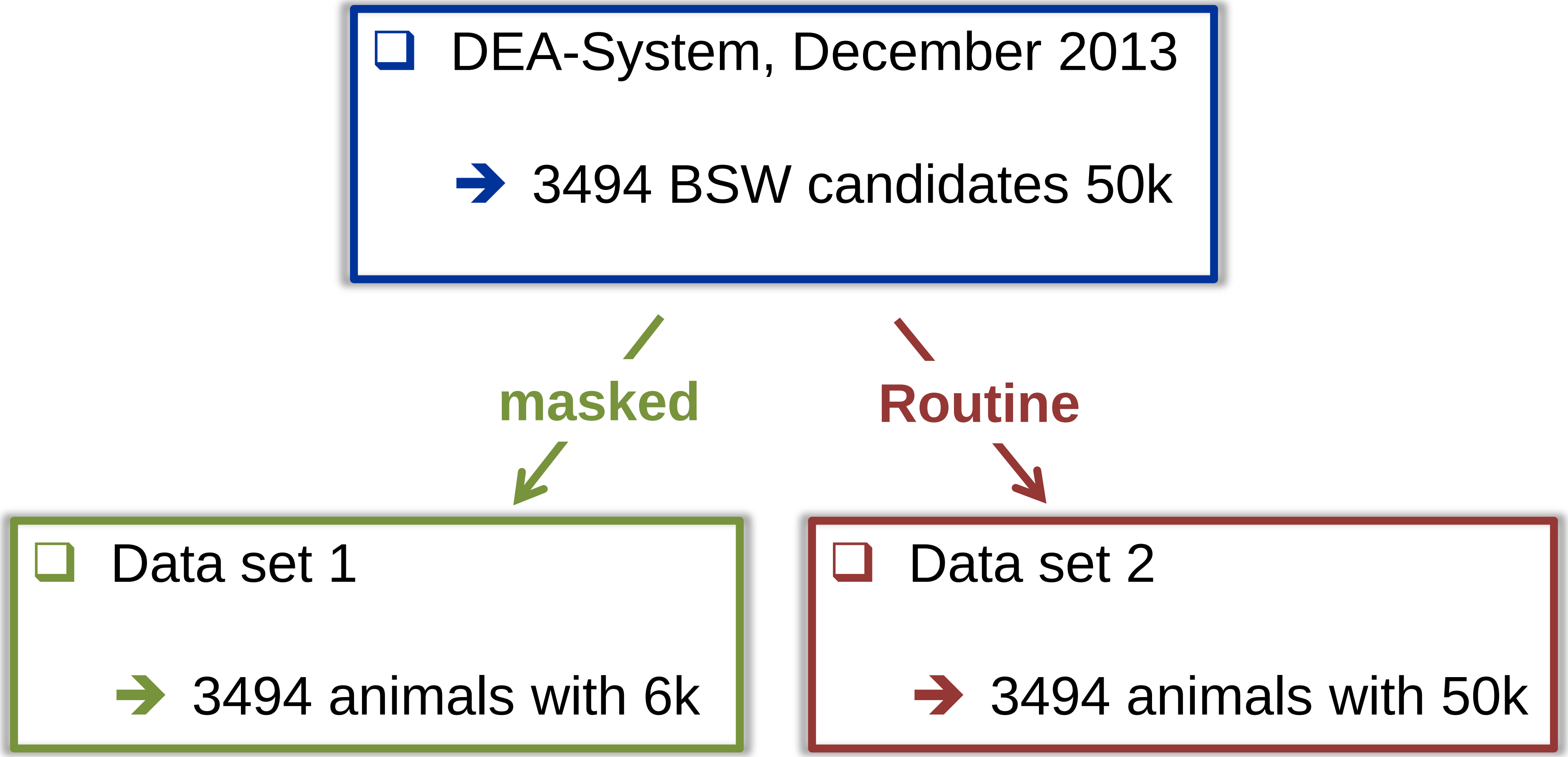
❑ Most studies about the effects of imputation report:

➔ overall correlations between GEBV

➔ comparisons between software

❑ Further investigate the causes and patterns underlying the bias in GEBV due to imputation errors

Material and Methods



□ DEA-System, December 2013

→ 3494 BSW candidates 50k

masked

Routine

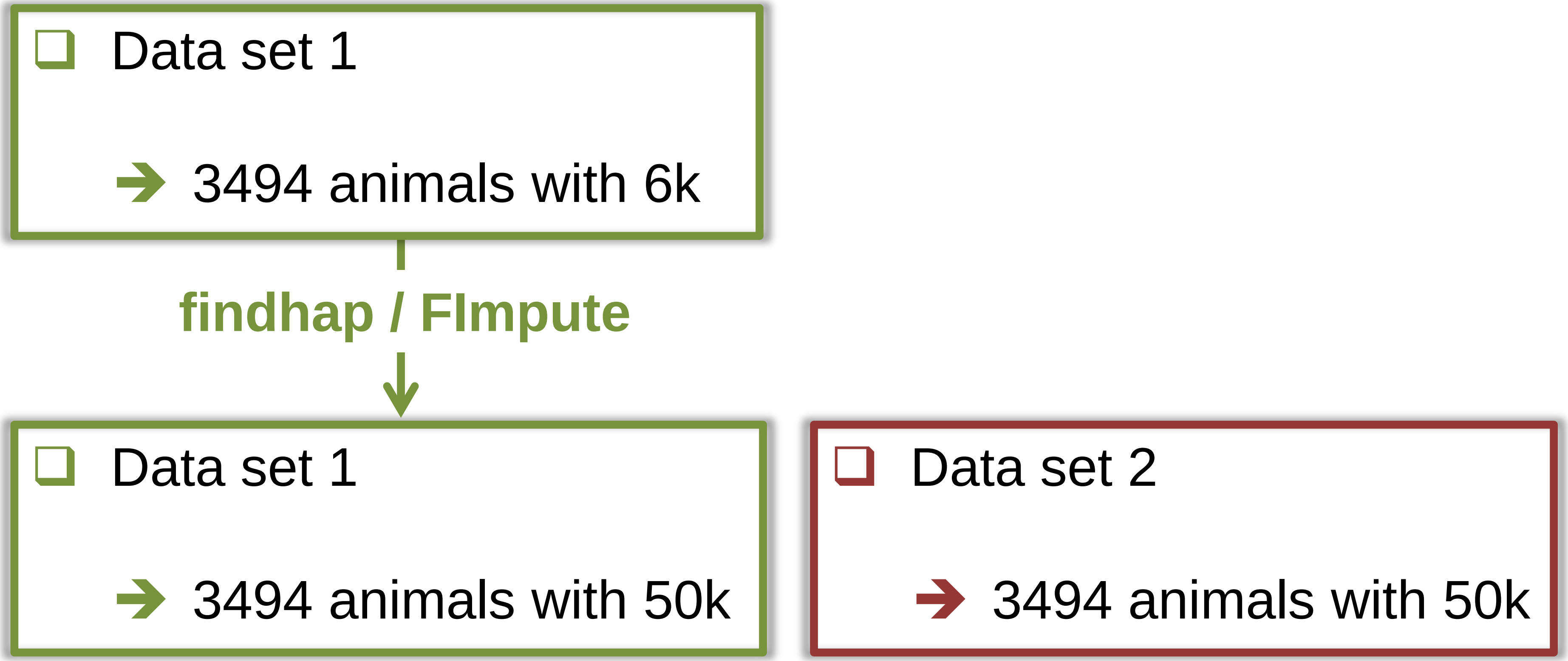
□ Data set 1

→ 3494 animals with 6k

□ Data set 2

→ 3494 animals with 50k

Material and Methods



- □ Data set 1

  → 3494 animals with 6k

findhap / FImpute

- □ Data set 1

  → 3494 animals with 50k

- □ Data set 2

  → 3494 animals with 50k

→ Prediction of GEBV for 37 traits

→ Comparison between GEBV from observed 50k with GEBV from imputed 50k

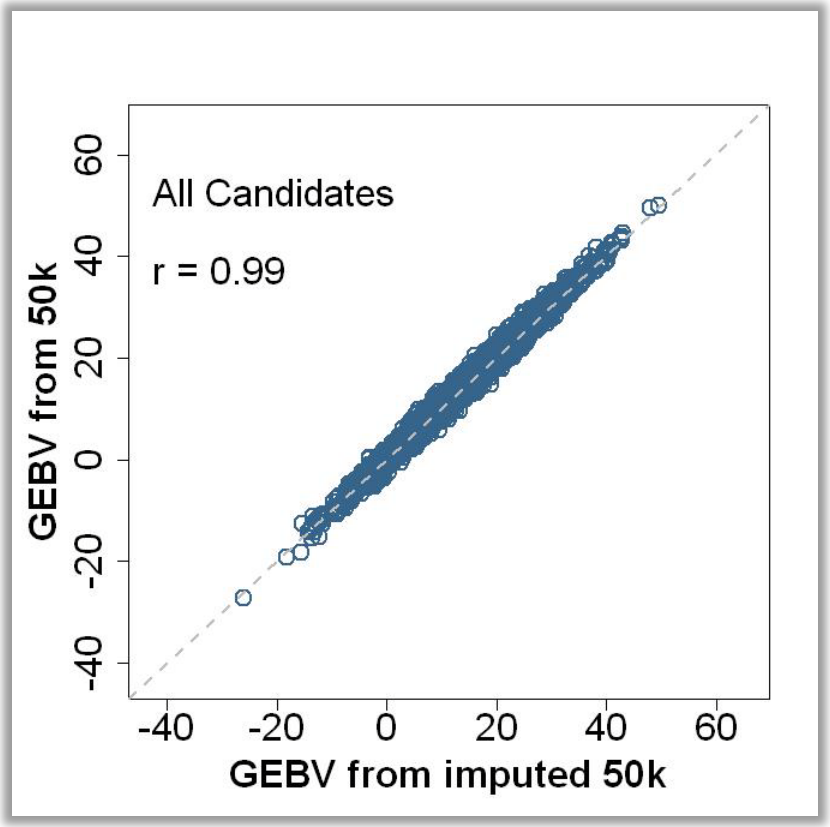
LfL Tierzucht

Changes in ranking within TOP 50 candidates

Trait	Rank correlation		Also top 50 in imputed set	
	findhap	FImpute	findhap	FImpute
Milk (kg)	0.82	0.90	42	44
Fat (kg)	0.90	0.91	42	46
Protein (kg)	0.82	0.91	42	43
SCS	0.79	0.87	43	41
Workability	0.71	0.88	40	44
Udder depth	0.89	0.89	42	40
Feet & legs	0.89	0.93	45	44
Udder	0.80	0.84	44	42
Overall score	0.86	0.89	44	43
Average (n=37)	**0.84**	**0.88**	**44**	**43**

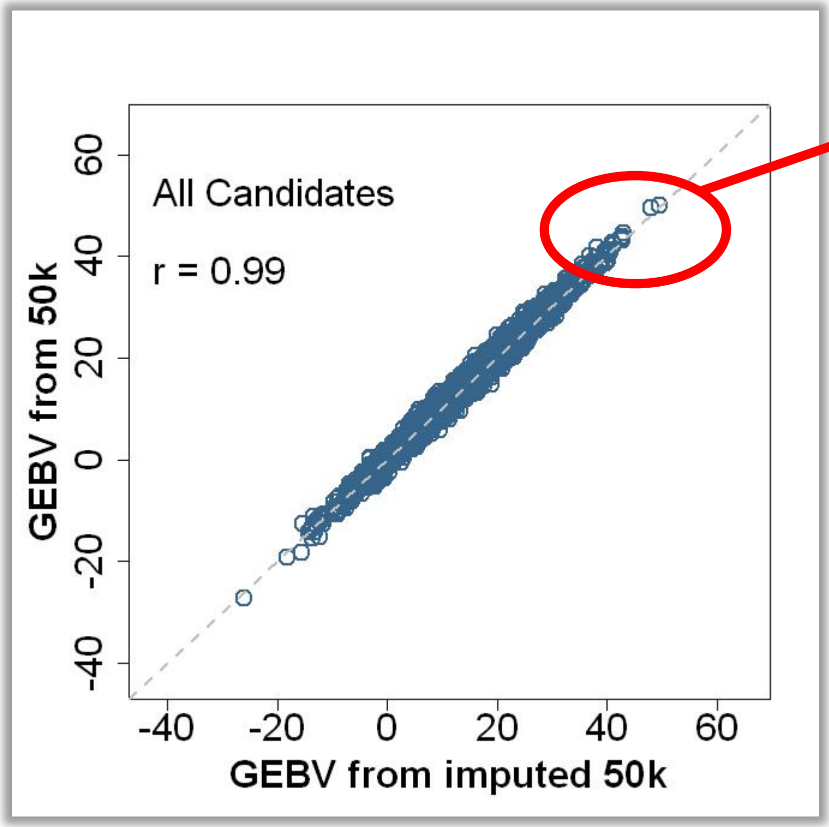
Impact of Imputation errors on GEBV

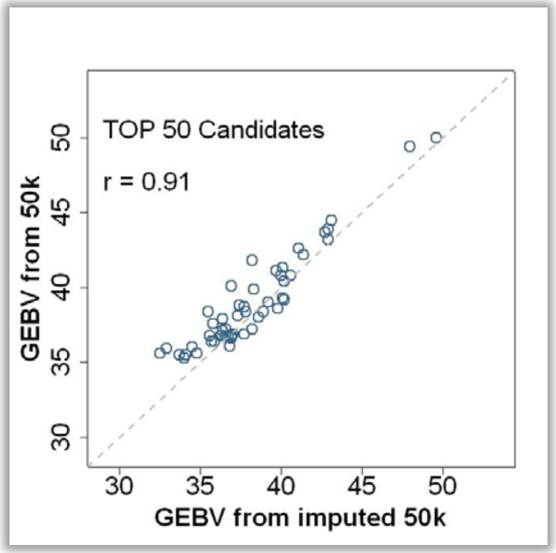
Protein (kg) / FImpute



Impact of Imputation errors on GEBV

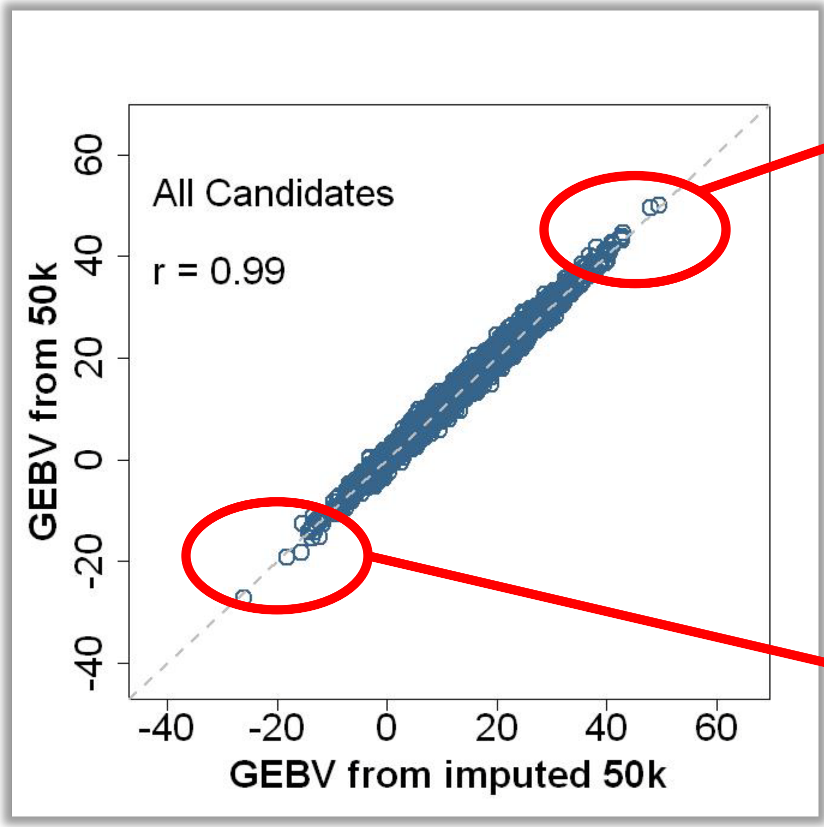
Protein (kg) / FImpute

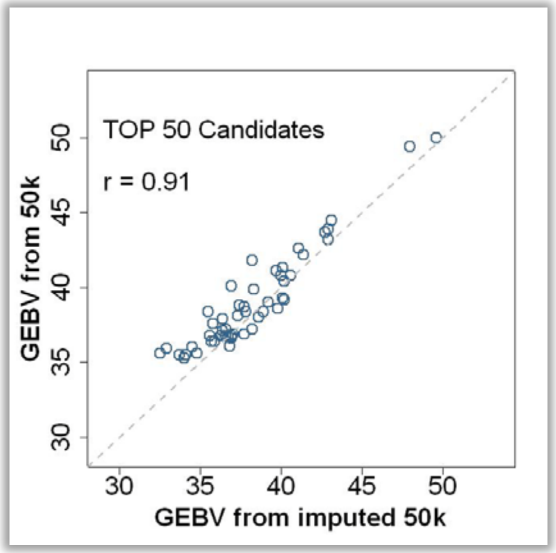


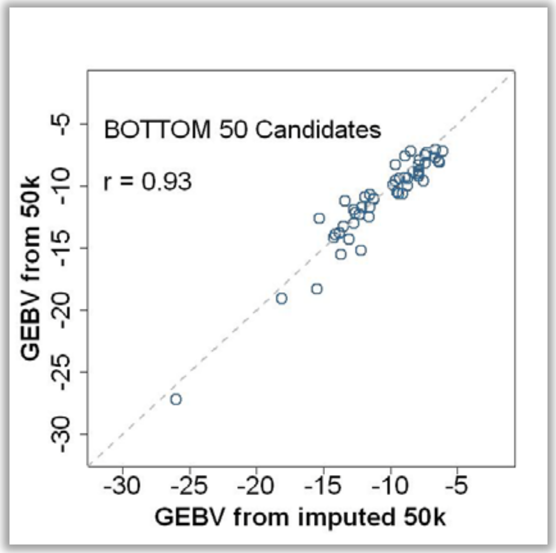


Impact of Imputation errors on GEBV

Protein (kg) / FImpute







A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

(3) if the most frequent haplotype has a neutral effect on the trait, bottom animals benefit and top animals are penalized.

A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

(3) if the most frequent haplotype has a neutral effect on the trait, bottom animals benefit and top animals are penalized.

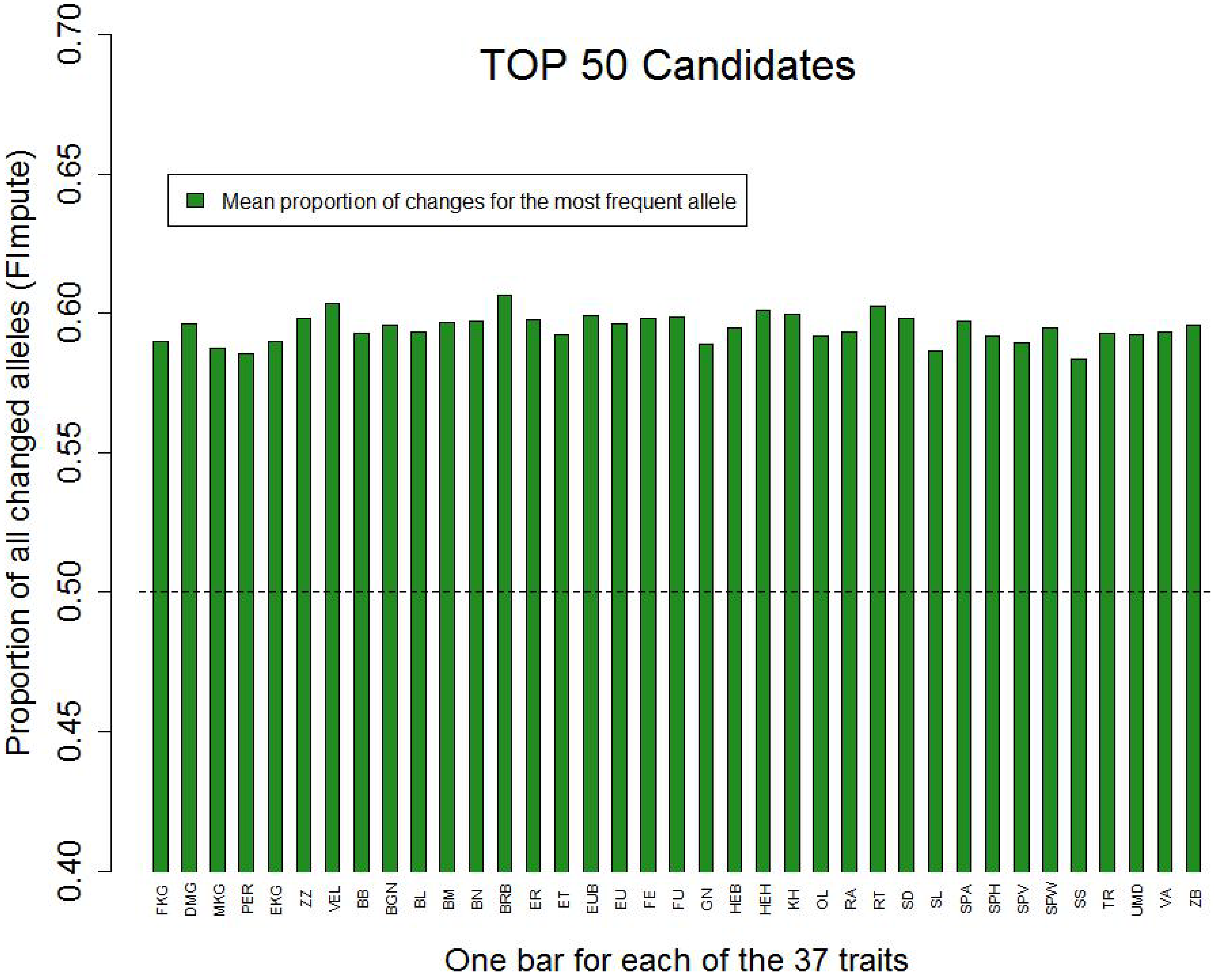
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

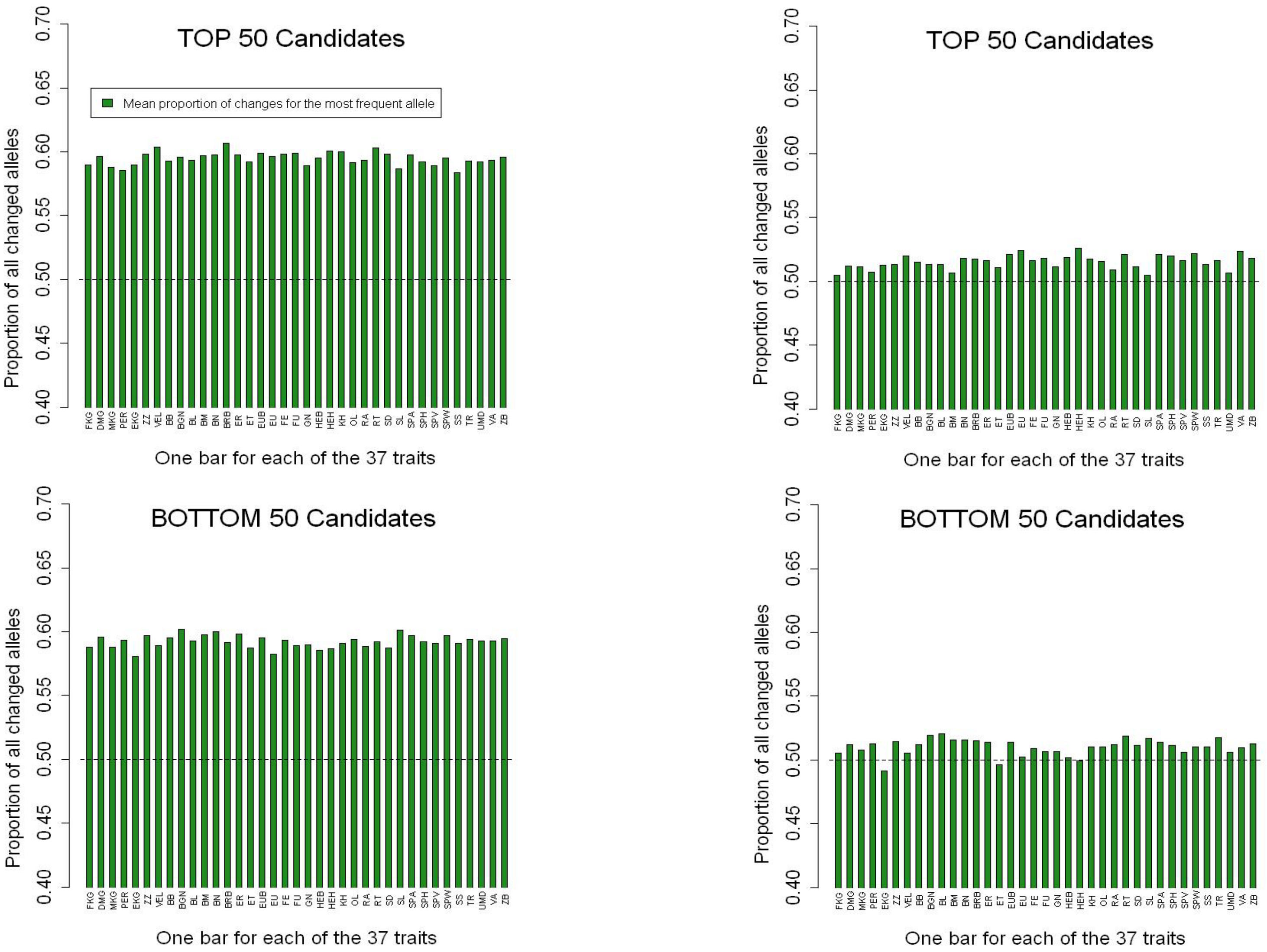
(3) if the most frequent haplotype has a neutral effect on the trait, bottom animals benefit and top animals are penalized.

A possible explanation



A possible explanation

FImpute | findhap



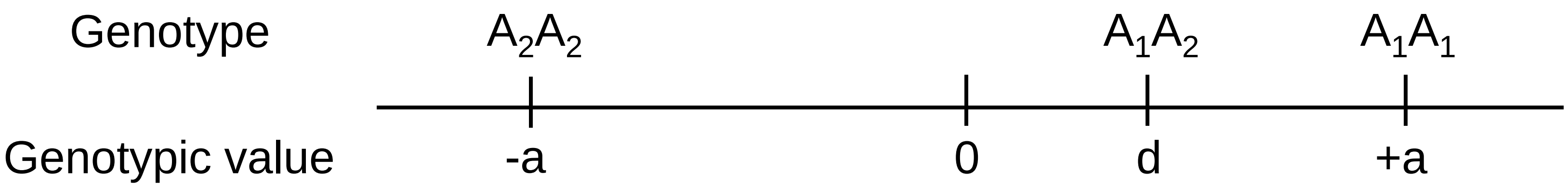
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

(3) if the most frequent haplotype has a neutral effect on the trait, bottom animals benefit and top animals are penalized.

A possible explanation

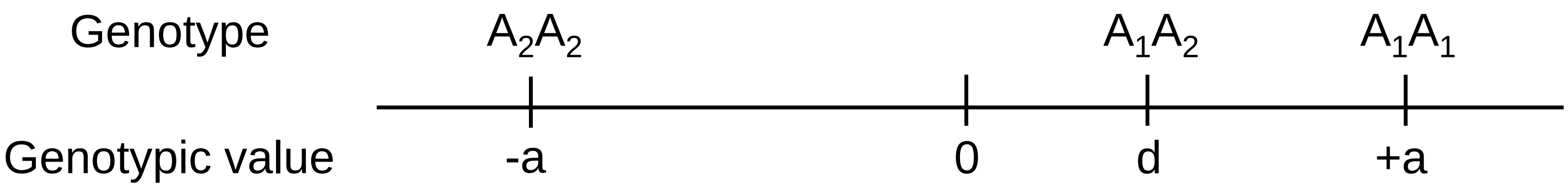


Genotype	A_2A_2		A_1A_2	A_1A_1
Genotypic value	-a	0	d	+a

Assuming Hardy-Weinberg equilibrium:

$$\begin{cases} f(A_1A_1) = p^2 \\ f(A_1A_2) = 2pq \\ f(A_2A_2) = q^2 \end{cases}$$

A possible explanation



Assuming Hardy-Weinberg equilibrium:

$$\begin{cases} f(A_1A_1) = p^2 \\ f(A_1A_2) = 2pq \\ f(A_2A_2) = q^2 \end{cases}$$

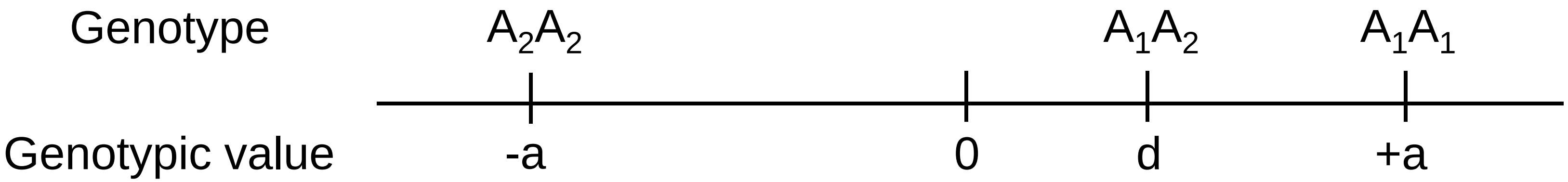
Population mean: $M = a(p - q) + 2pqd$

*Average effect of A_1: $\alpha_1 = q[a + d(q - p)]$

*Average effect of A_2: $\alpha_2 = -p[a + d(q - p)]$

*as deviation from the population mean

A possible explanation



Genotype	A_2A_2		A_1A_2	A_1A_1
Genotypic value	-a	0	d	+a

Assuming Hardy-Weinberg equilibrium:

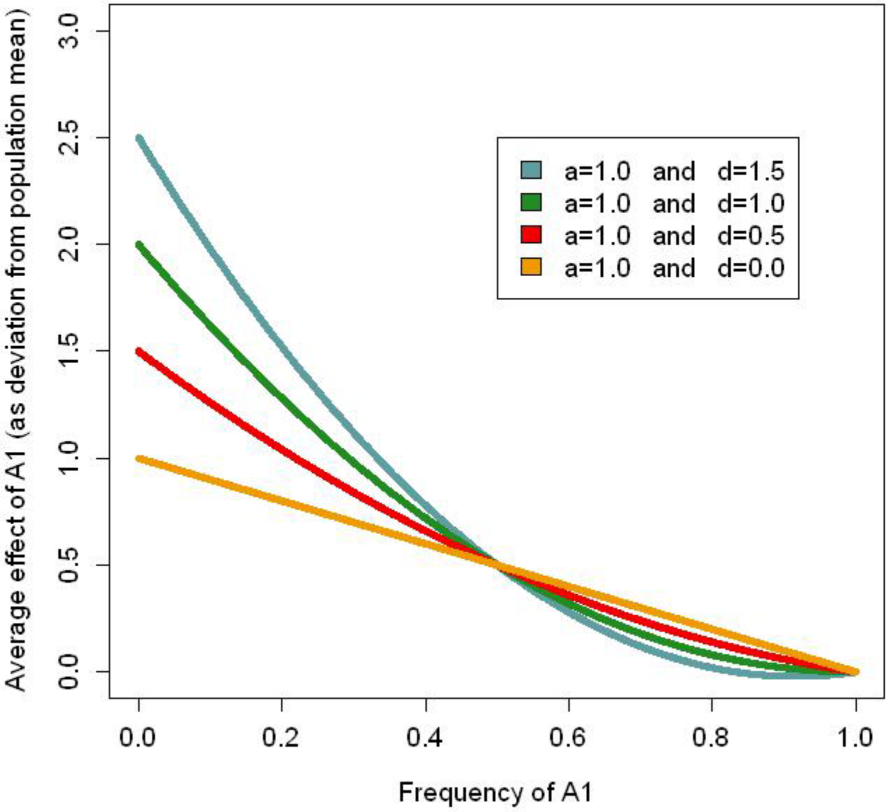
$$\begin{cases} f(A_1A_1) = p^2 \\ f(A_1A_2) = 2pq \\ f(A_2A_2) = q^2 \end{cases}$$

Population mean: $M = a(p - q) + 2pqd$

*Average effect of A_1: $\alpha_1 = q[a + d(q - p)]$

*Average effect of A_2: $\alpha_2 = -p[a + d(q - p)]$

*as deviation from the population mean



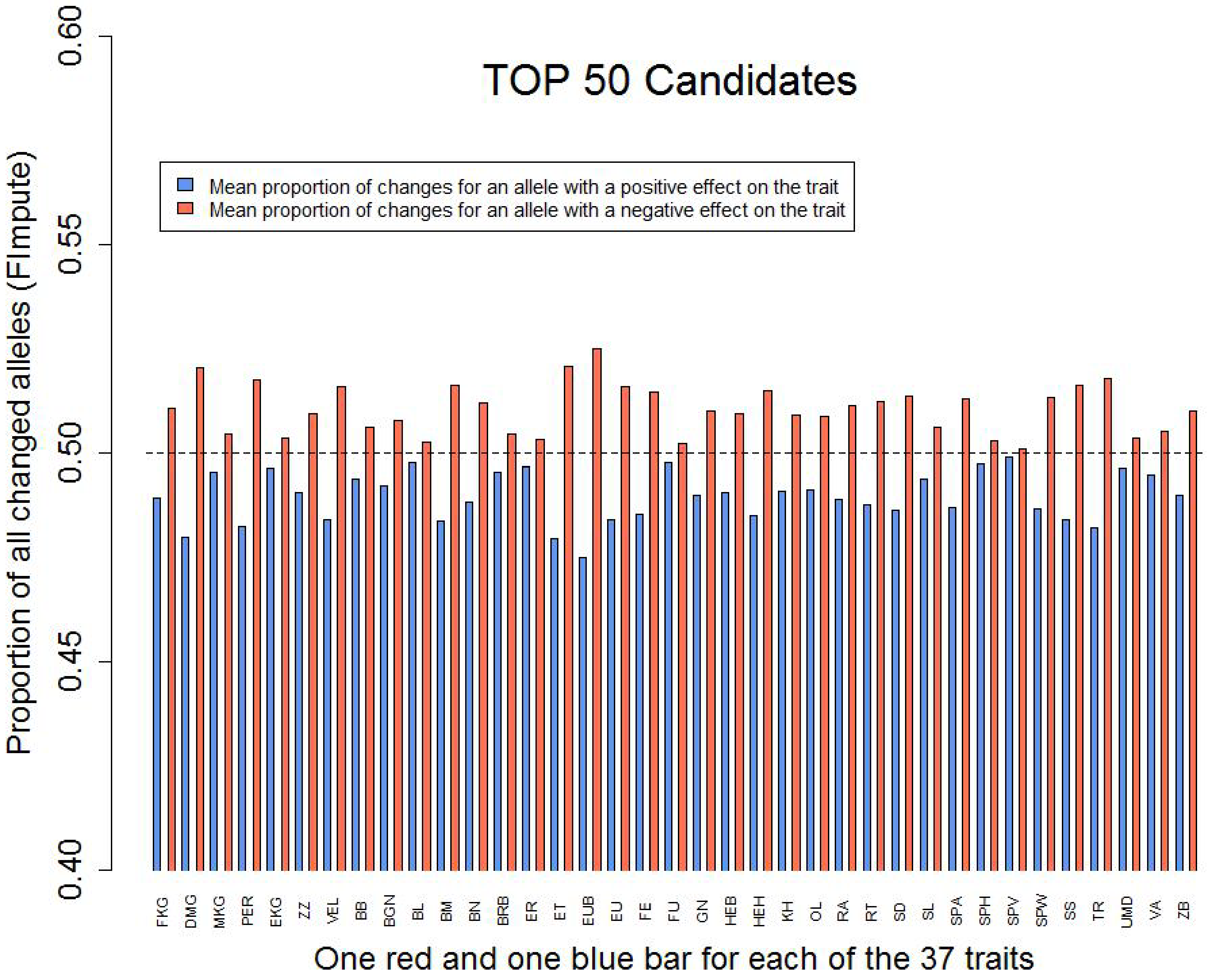
A possible explanation

(1) On average, top animals should have the best haplotypes and bottom animals should have the worst haplotypes;

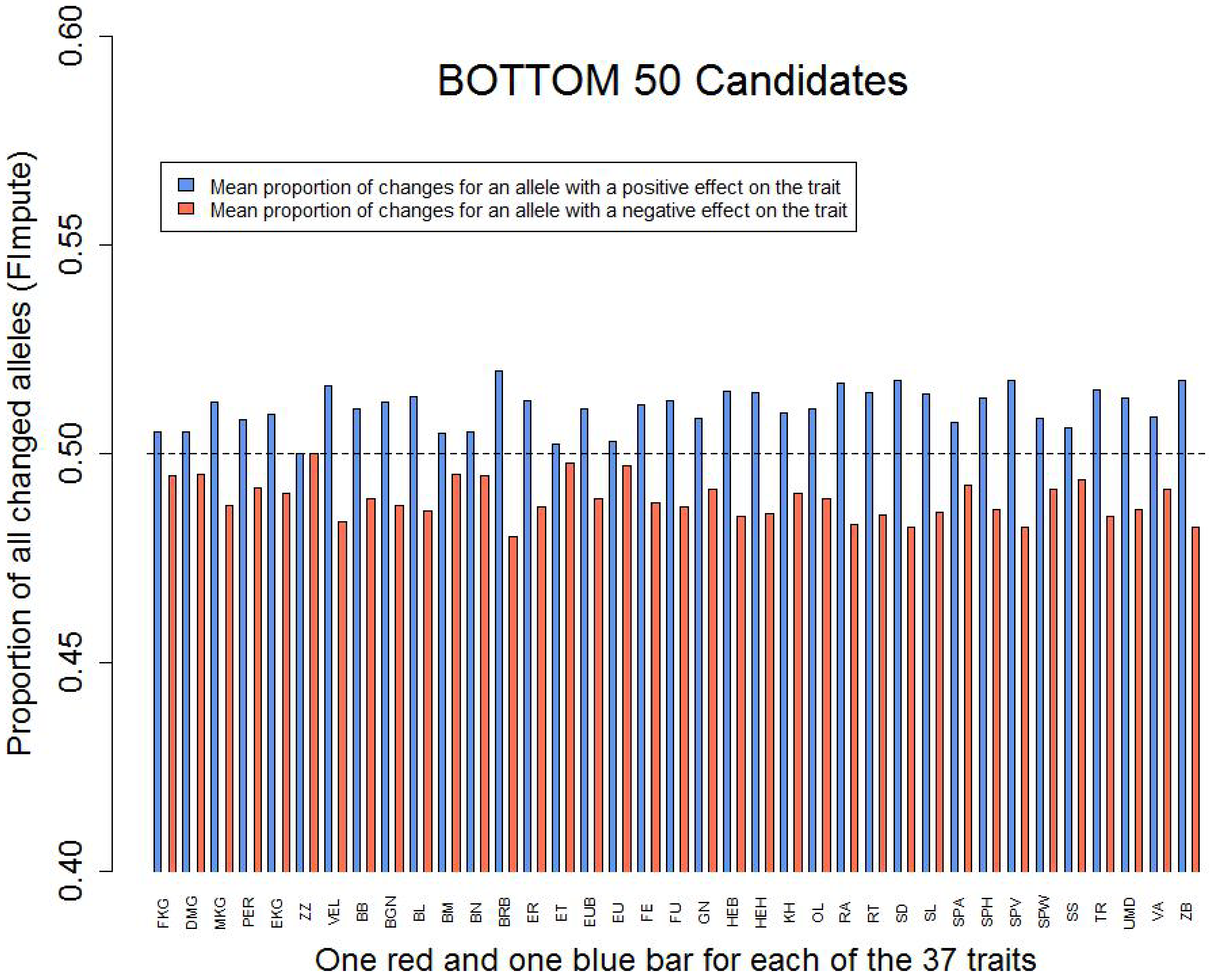
(2) in ambiguous cases, imputation algorithms will suggest the most frequent haplotype as replacement;

(3) if the most frequent haplotype has a neutral effect on the trait, bottom animals benefit and top animals are penalized.

A possible explanation

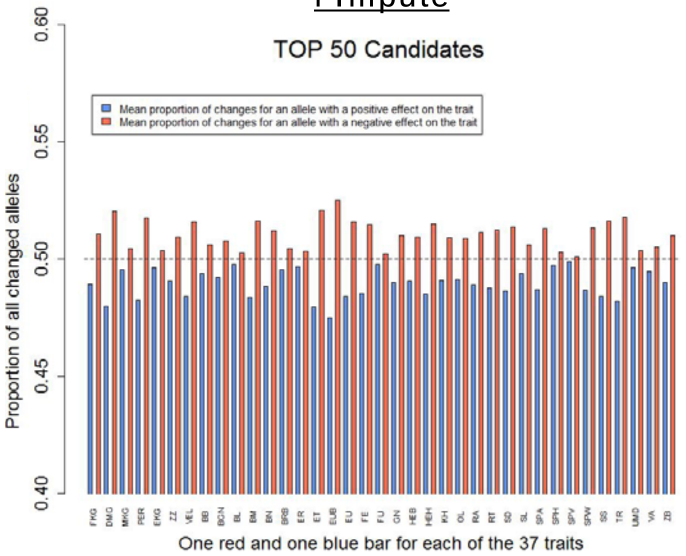


A possible explanation

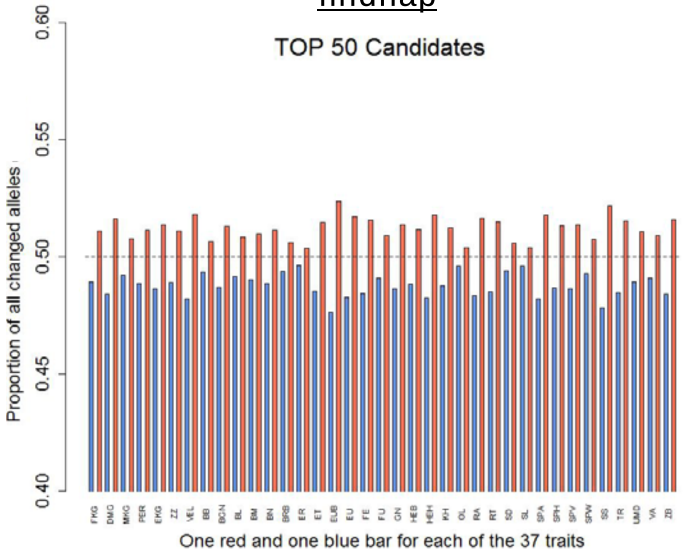


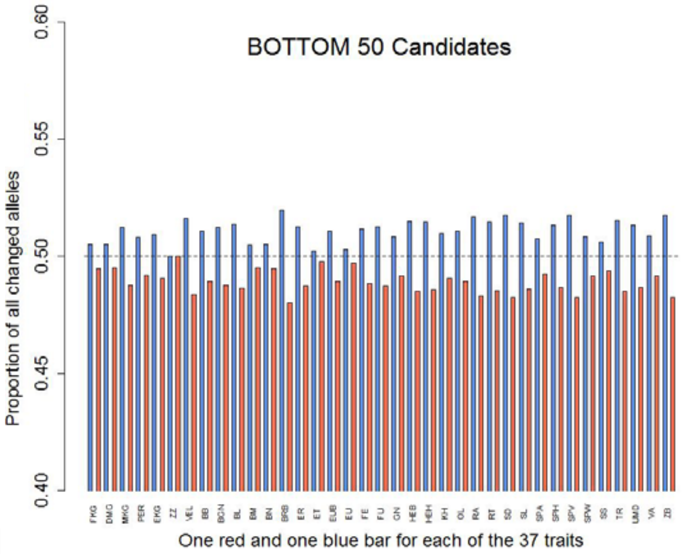
A possible explanation

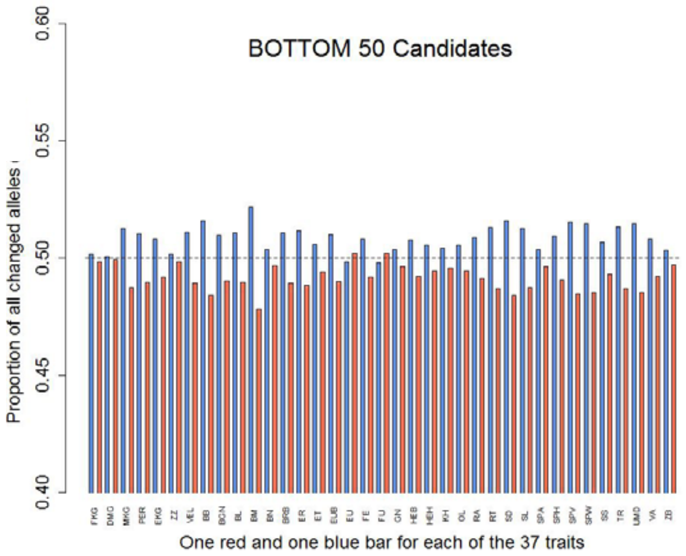
FImpute



findhap







Conclusions

- Bias in GEBV due to imputation errors

  ➔ downwards in top and upwards in bottom segment

Conclusions

- ❑ Bias in GEBV due to imputation errors

  ➔ downwards in top and upwards in bottom segment

- ❑ Imputation algorithms usually suggest haplotypes with higher frequency and more neutral effects

  ➔ disadvantage for top and advantage for bottom animals

Conclusions

❑ Bias in GEBV due to imputation errors

➔ downwards in top and upwards in bottom segment

❑ Imputation algorithms usually suggest haplotypes with higher frequency and more neutral effects

➔ disadvantage for top and advantage for bottom animals

❑ Might have implications, especially for mixed pools of candidates genotyped at different densities

Aknowledgements

- Paul VanRaden (findhap)
- Mehdi Sargolzaei (FImpute)
- intergenomics

Thanks!

LfL
Tierzucht