WCGALP - Method and Tools - Prediction 1

Using additional SNPs selected from whole genome sequence (WGS) data for genomic prediction in Danish Jersey

Aoxing Liu, Mogens Sandø Lund, Didier Boichard,
Sebastien Fritz, Emre Karaman, Yachun Wang, Guosheng Su

Feb 12, 2018

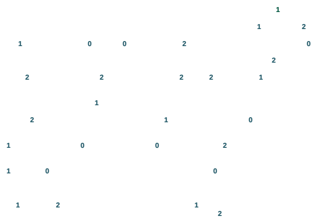
Contents

- Introduction
- Material and methods
- Results and discussion
- Conclusion

Introductio
n

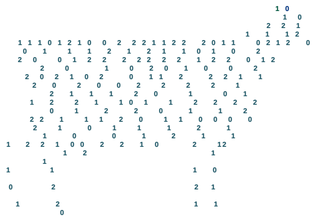
Background

High throughput genotyping



LD (7 K)

MD (54 K)

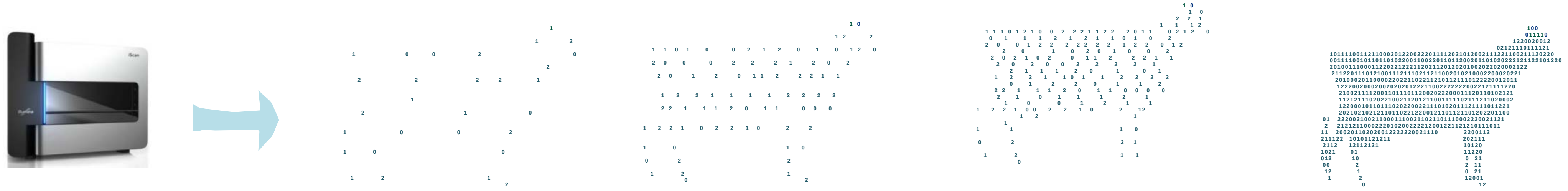


HD (777 K)

WGS (~26,700 K)

Hypothesis: Higher SNP density -> better LD -> higher reliability

Background



High throughput genotyping

LD (7 K)

MD (54 K)

HD (777 K)

WGS (~26,700 K)

Hypothesis: Higher SNP density -> better LD -> higher reliability

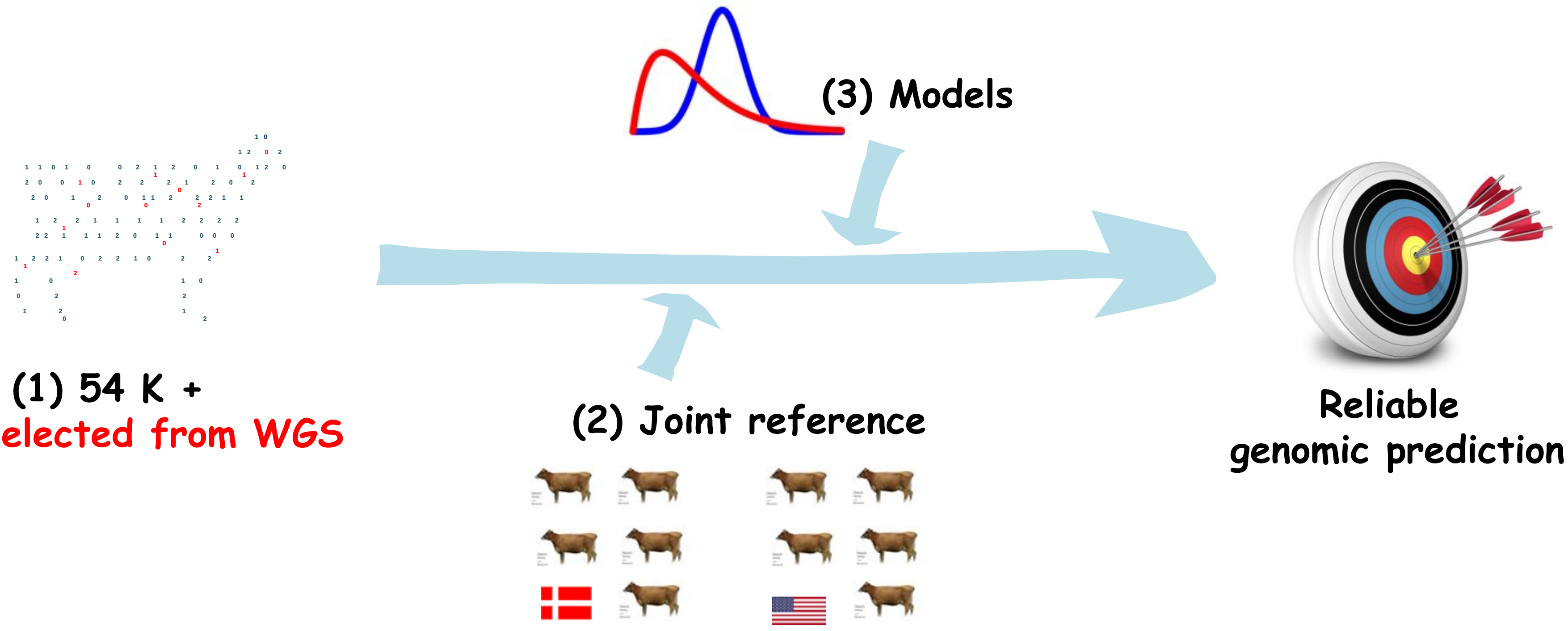
Real data: HD ≈ 54K (Su et al., 2012) & Imputed WGS ≈ HD (Van Binsbergen et al., 2015)

- Only causative mutations or variants very close to causative mutations can improve reliability

  (van den Berg et al., 2016)

- non-causative mutations bring noise

Objectives



(3) Models

(1) 54 K +
SNPs selected from WGS

(2) Joint reference

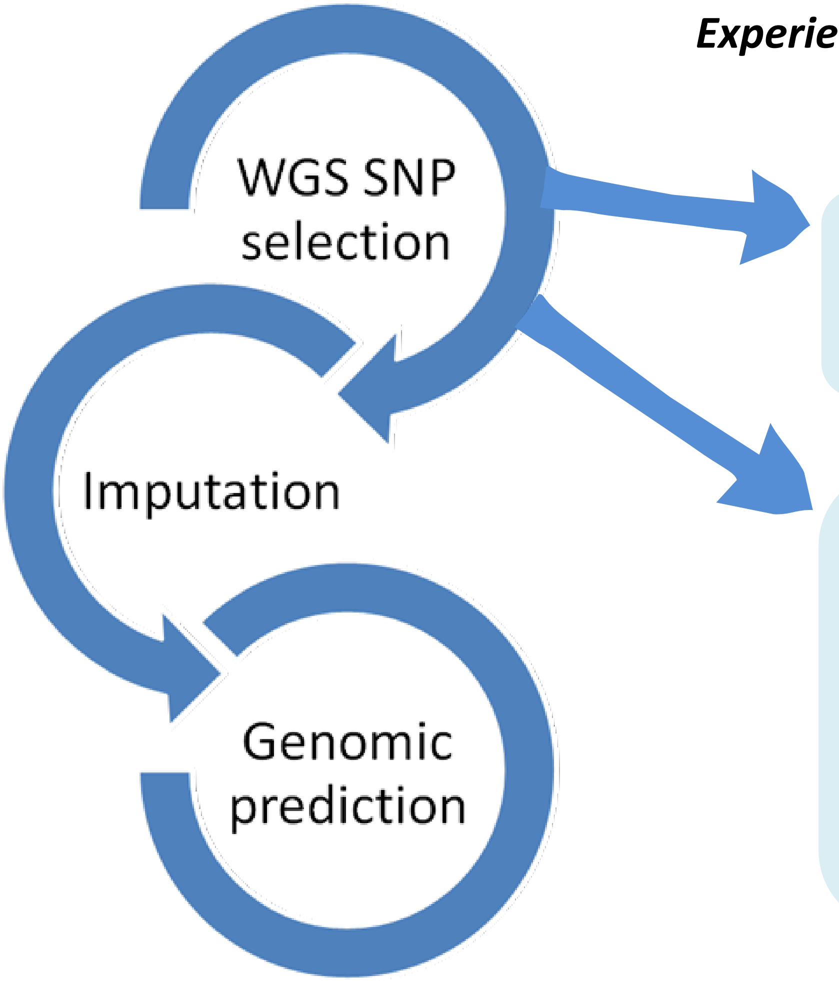
Reliable
genomic prediction

- ➢ Investigate effects of additional WGS SNPs on genomic prediction
- ➢ Effects of using additional WGS SNPS in a joint reference
- ➢ Assessed models on their efficiency to use information of additional WGS SNPs

Material and methods

Workflow

Experience from large scale use of the EuroGenomics custom SNP chip in cattle *(Boichard et al., WCGALP, 2018)*



NOR SNPs (Brondum et al., 2015)

- peaks of QTL from Nordic Holsteins, Nordic Red and Danish Jersey

FR SNPs

- literature
- a strong variant effect predictor annotation (e.g. non-synonymous substitution)
- regulatory regions of genes
- peaks of QTL
- breakpoints of structural SNPs

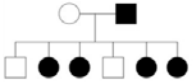
Imputation

Animal

- ➢DK bulls: ~1,300
- ➢US bulls: ~1,200
- ➢DK cows: ~31,000

Genotype

- ➢54K chip
- ➢standard LD chip
- ➢customized LD chip
  - ▪ standard LD chip
  - ▪ NOR SNPs
  - ▪ FRA SNP



Pedigree

- ➢6,100 males
- ➢66,000 females

(1)

54K

(2)

54K
+ NOR SNPs + FRA SNPs

Two-step imputation
(Fimpute)

Prediction: SNP

Quality control

- ➢Minor allele frequency > 0.01
- ➢Imputation accuracy
- ▪ correlation > 0.8
- ▪ concordance rate > 0.8

SNPs	No. of SNPs	
	before	after
54K	40,452	33,166
NOR SNPs	1,754	1,270
FRA SNPs	4,325	2,427

Prediction: GBLUP model

CENTER FOR QUANTITATIVE GENETICS AND GENOMICS QGG

- One-component model

$$y = 1\mu + X\underset{\substack{\downarrow \\ \text{54K/ 54K+selected WGS SNPs}}}{g} + e$$

Scenarios	Component_One
54K	54K
54K_NOR	54K+NOR
54K_FRA	54K+FRA
54K_NOR_FRA	54K+NOR+FRA

- Two-component model

$$y = 1\mu + X_{54K}\underset{\substack{\swarrow \\ \text{54K}}}{g_{54K}} + X_{WGS}\underset{\substack{\searrow \\ \text{Selected WGS SNPs}}}{g_{WGS}} + e$$

Scenarios	Component_One	Component_Two
54K_NOR	54K	NOR
54K_FRA	54K	FRA
54K_NOR_FRA	54K	NOR+FRA

Prediction: Reference and validation

Reference

- **DK**: ~1,000 DK bulls born before 2005
- **Joint DK-US**: ~1,000 DK bulls born before 2005
  ~1,200 US bulls

Validation

- ~300 DK bulls born after 2005

Prediction: Model comparision

Compare reliabilities from different models/ scenarios:

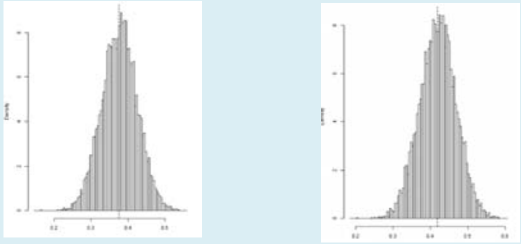
- **SE of reliability:**

  Non-parametric Bootstrap with 10,000 samples

- **Significant test**

  Two-tailed paired t-test with p-value = 0.05

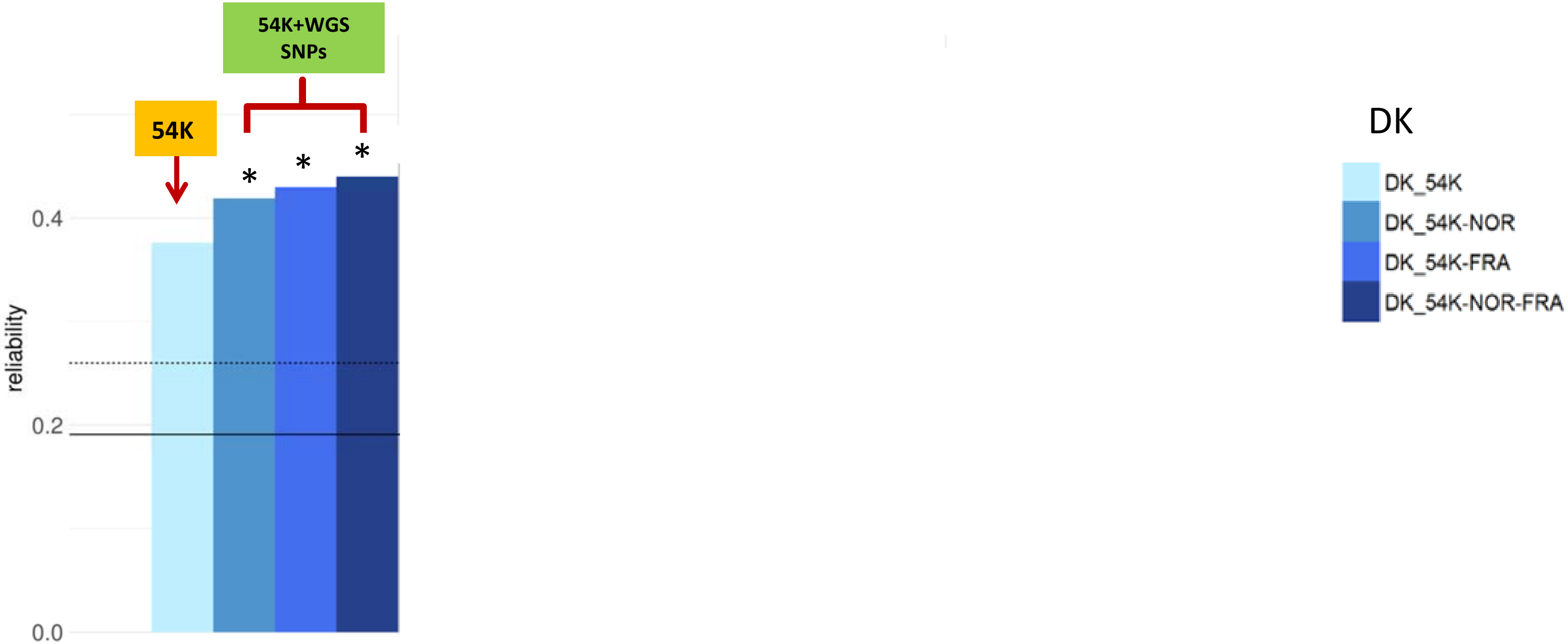
10,000 bootstrap samples of reliabilities



Results and discussion

Results-Reliability of milk

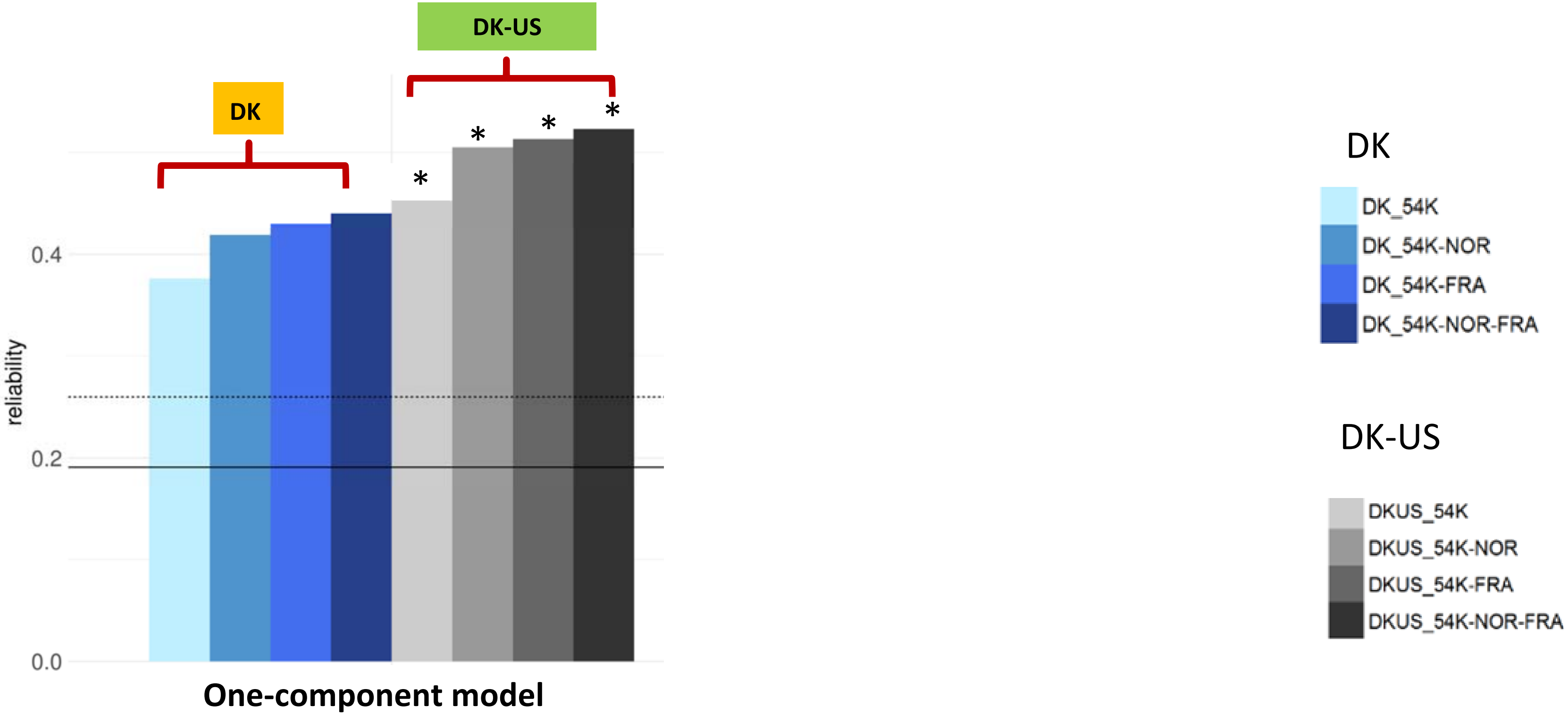
CENTER FOR QUANTITATIVE GENETICS AND GENOMICS QGG



One-component model

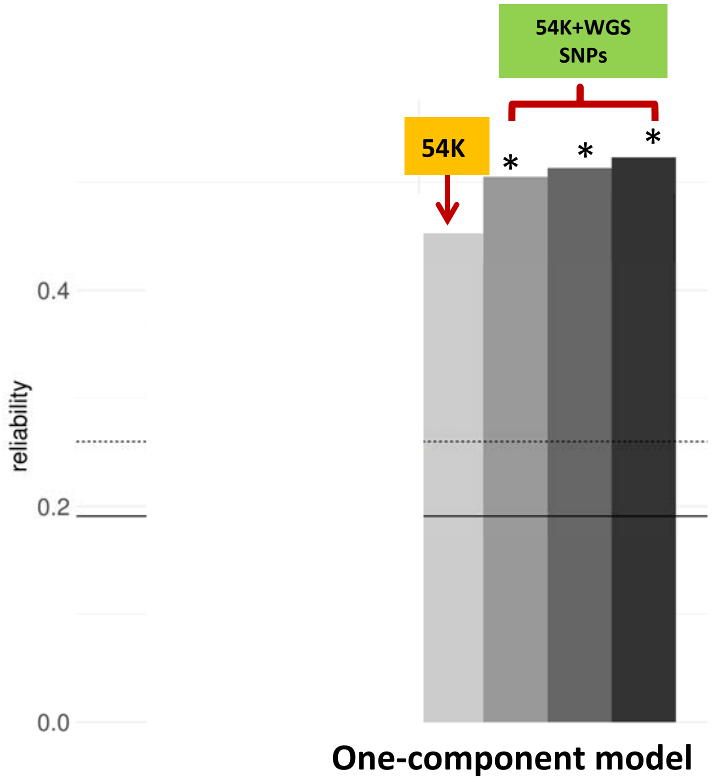
- Inclusion of additional WGS SNPs significantly improved reliability (11.4-17.0%)
- Inclusion of all additional WGS SNPs achieved highest reliabilities

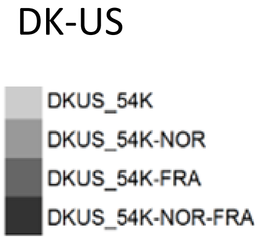
Results-Reliability of milk



- A joint DK-US reference significant better than a DK reference (20%)

Results-Reliability of milk

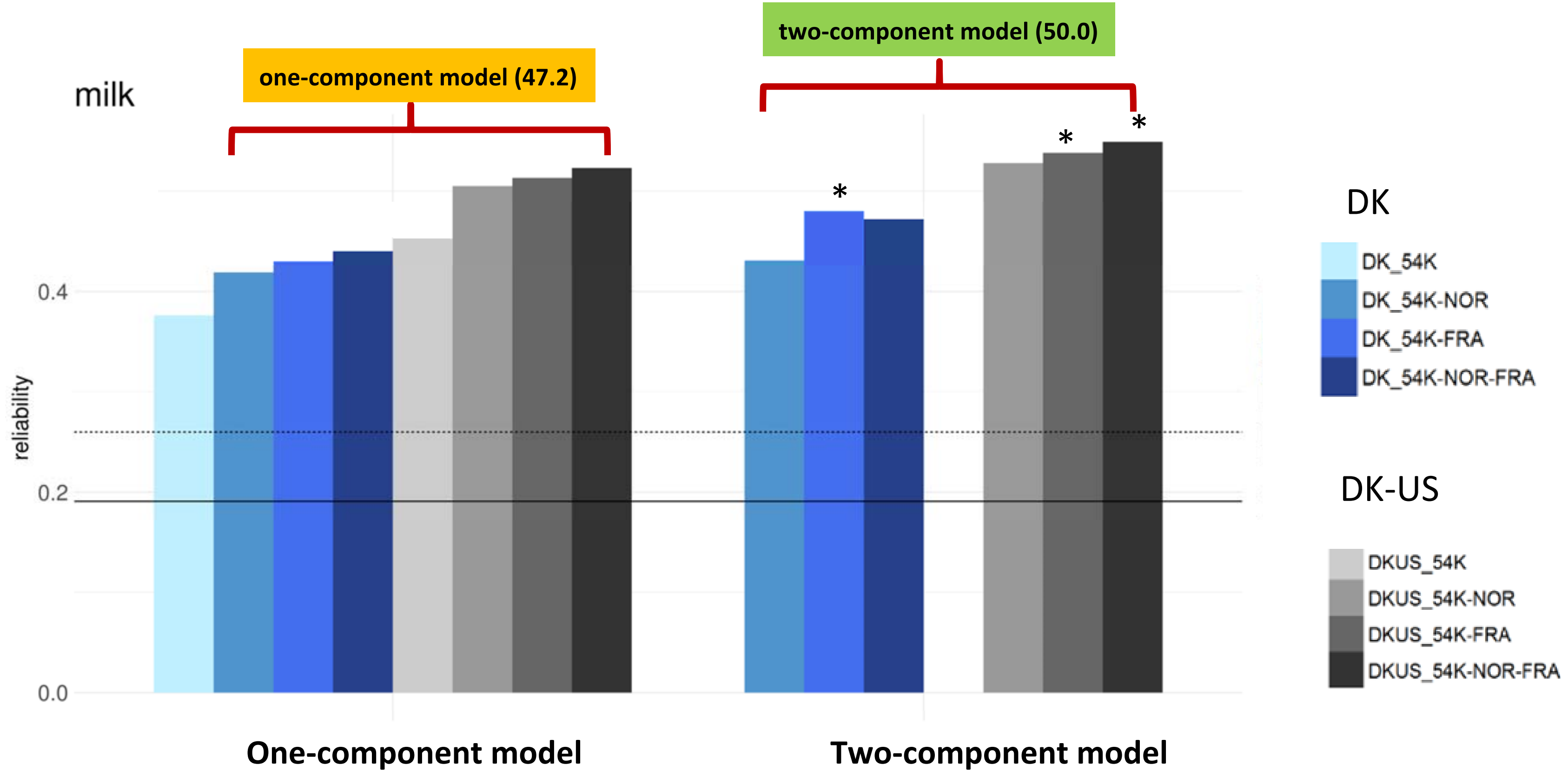




- Additional WGS SNPs improved reliabilities of a joint reference (11.5-13.6%)

Results-Reliability of milk

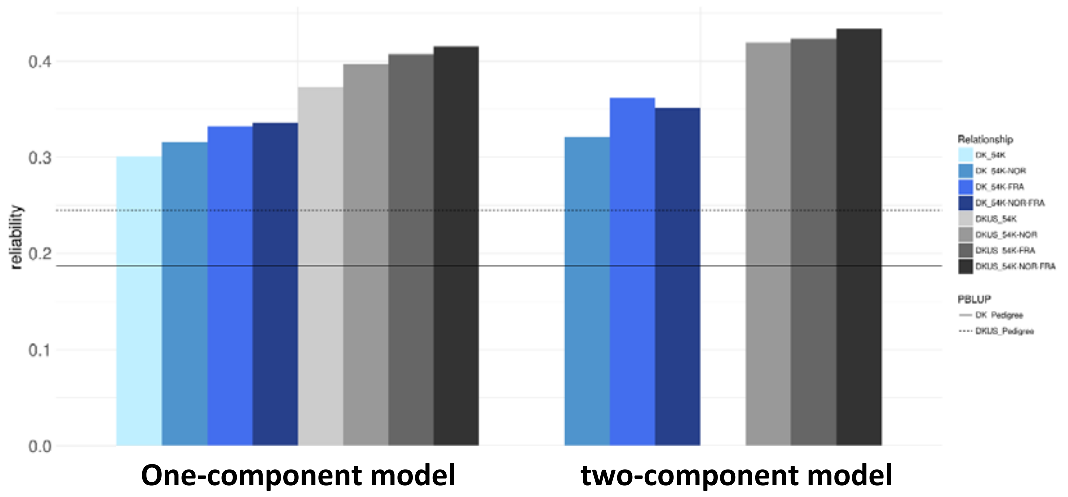
CENTER FOR QUANTITATIVE GENETICS AND GENOMICS QGG



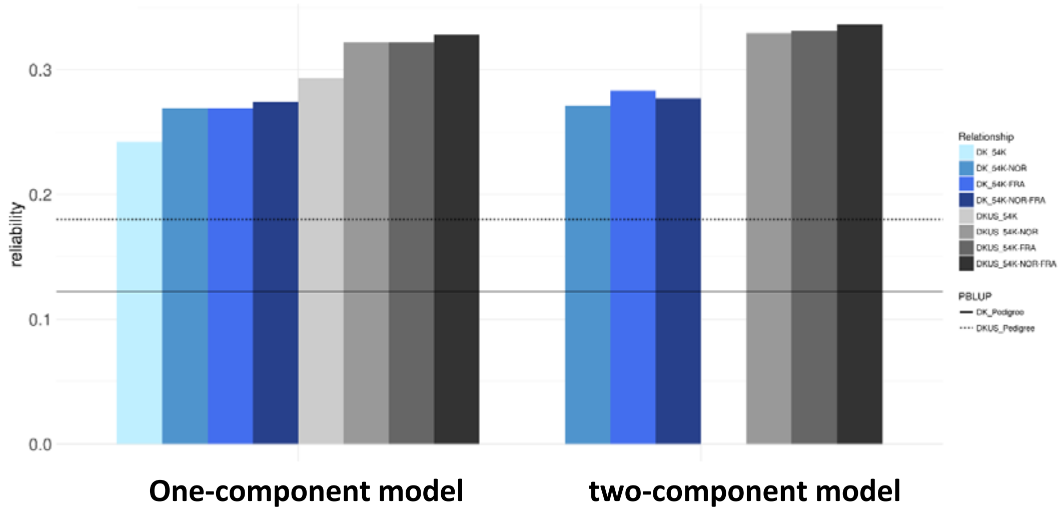
- A two-component model improved reliabilities (4.8%)

Results-Reliability of protein & fat

Protein



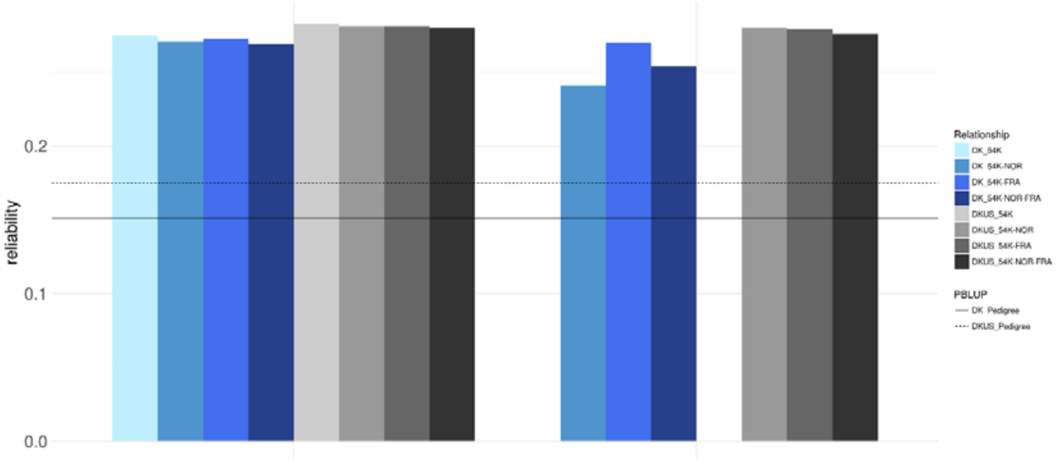
Fat



- Similar to milk

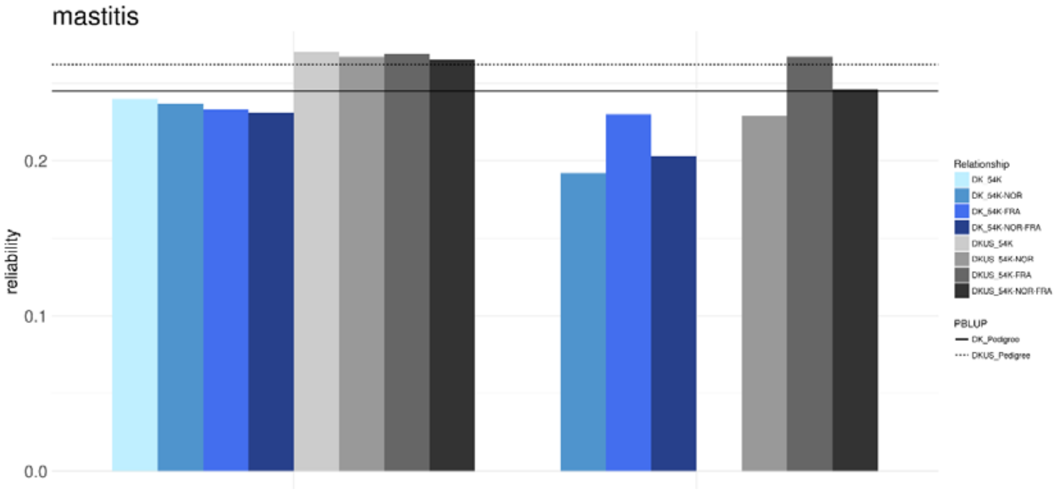
Results-Reliability of fertility & mastitis

Fertility



One-component model two-component model

Mastitis

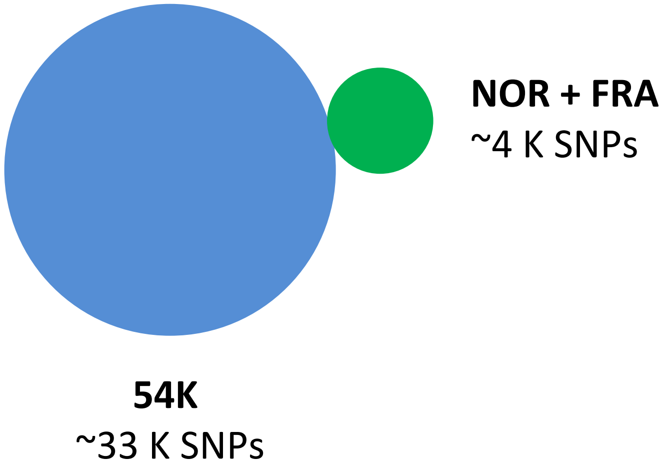


One-component model two-component model

- No significant difference between 54K and 54K + selected WGS SNPs

Results

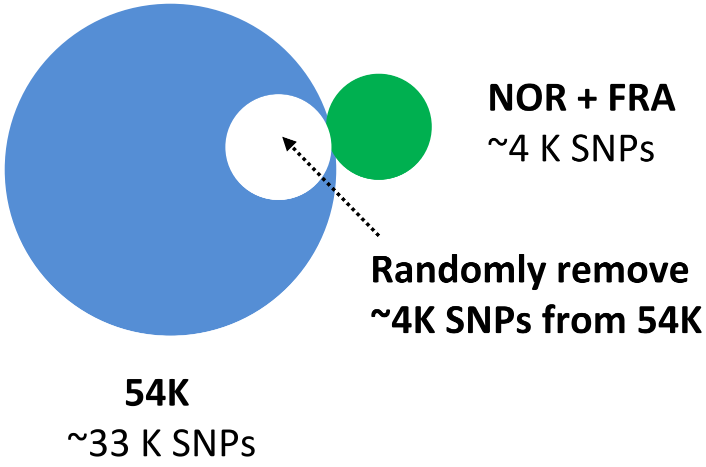
Dose the improvement of reliabilities come from increase of SNP density ?



NOR + FRA
~4 K SNPs

54K
~33 K SNPs

VS.



NOR + FRA
~4 K SNPs

Randomly remove ~4K SNPs from 54K

54K
~33 K SNPs

- ✓ **54K + NOR + FRA**

- ✓ **54Kminus + NOR + FRA**

No. of SNPs is equal to 54K chip!

Results

Reliability (54Kminus + NOR + FRA) – Reliability (54K + NOR + FRA)

Trait	Reference	One-component	Two-Component
Milk	DK	0.003	0.002
	DKUS	0	-0.001
Protein	DK	0.001	0.001
	DKUS	-0.002	-0.003
Fat	DK	0.002	0
	DKUS	-0.001	-0.003

- **No difference between 54K + NOR + FRA and 54Kminus + NOR + FRA**
- **Improvement of reliabilities using additional WGS SNPs not from increase of SNP density**

Conclusion

Conclusion

- Additional WGS SNPs improved reliabilities for milk production, not for fertility and mastitis
- The inclusion of all additional WGS SNPs achieved highest reliabilities
- A joint DK-US reference better than a DK reference for all traits
- Additional WGS SNPs further improved reliabilities of a joint DK-US reference
- A two-component model improved reliabilities for milk production

Acknowledgement

I've Hired This Cow To Deliver The Following Message:
SAUVE
Moo
(That's Cow For "THANK YOU")

GenSAP

VIKINGGENETICS

NAV Nordisk Avlsværdi Vurdering
Nordic Cattle Genetic Evaluation

CHINA SCHOLARSHIP COUNCIL

AARHUS UNIVERSITY

INRA SCIENCE & IMPACT

- Gert Pedersen Aamand, NAV
- Esa Mantysaari, Luke
- Per Madsen, Aarhus University
- Goutam Sahana, Aarhus University
- Ulrik Sander Nielsen, SEGES
- Han Mulder, Wageningen University & Research
- Xiaowei Mao, Cornell University
- Peipei Ma, Shanghai Jiao Tong University

Methods

Imputation accuracy

- Correlation = COR (TRUE, IMPUTED)

- Concordance rate = $\frac{\text{No. of animals with corectly imputed genotypes}}{\text{No. of animals with imputed genotypes}}$

Methods

Non-parametric Bootstrap

A general method for determining the SE of any estimator

1) Read data of 269 bulls in validation population

2) Randomly sample 269 rows with replacement

3) Calculate R2 for SCE1 and SCE2 for each bootstrap sample

4) Repeat this process 10,000 times

5) Differences between reliabilities among scenarios : CI and paired t-test

1)

ID	DRP	R2_DRP	SCE1	SCE2
1	104.1	99	1.92	-7.13
2	88.9	93	-1.38	-11.89
…	…	…	…	..
269	113.0	99	22.40	16.66

2)

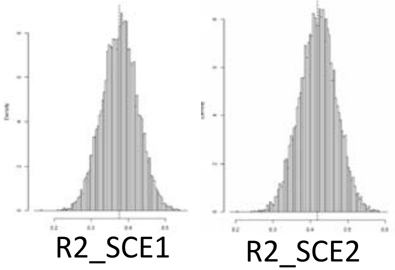
ID	DRP	R2_DRP	SCE1	SCE2
1	104.1	99	1.92	-7.13
1	104.1	99	1.92	-7.13
…	…	…	…	..
269	113.0	99	22.40	16.66

3)

Round	R2_SCE1	R2_DRP
1	0.38	0.42

4)

Round	R2_SCE1	R2_DRP
1	0.38	0.42
…	…	…
10,000	0.39	0.41



5) Two-tailed paired t-test with df=10,000-1

$$= \frac{mean(R2\_SCE1)-mean(R2\_SCE2)}{se}$$