# Genomic prediction using high-density SNP markers in Nordic Holstein and Red (RDC)

Guosheng Su[1], Rasmus F. Brøndum[1], Peipei Ma[1]
Bernt Guldbrandtsen[1], Gert P. Aamand[2], Mogens S. Lund[1]

[1]Department of Molecular Biology and Genetics,
Aarhus University, Denmark
[2]Nordic Cattle Genetic Evaluation, DK-8200 Aarhus N,
Denmark

## Background

- Expect stronger LD between genes and markers with HD markers
- Expect to get more accurate genomic predictions using HD markers
- However, simulations with large number of QTL show no clear benefit from using HD markers

## Objective

Investigate accuracy of genomic predictions in Nordic Holstein and RDC using imputed HD markers (770k)

# Populations

Nordic bulls with genotype (50k and imputed 770k) and de-regressed proofs (DRP)

| | |
|---|---|
| Holstein: | 4539 |
| RDC: | 4403 |

Training data: animals born before 2001-10-01

Test data : animals born after 2001-10-01

Traits: protein, fertility, udder health

# Marker data

**50 k data**: original 50k data with some markers missing

**imputed 50k data**: missing markers in 50k were imputed using Beagle

**Imputed HD data**: 50k data were imputed to HD data using Beagle

- Holstein: based on 557 HD genotyped bulls (EuroGenomics)
- RDC: based on 706 HD genotyped bulls
- After imputation, delete the markers in complete LD with preceding locus.
- About 500k markers used for genomic prediction

# Statistical Model

1. GBLUP

2. Bayesian mixture model with two normal distributions

   Prior: π = 20% for 50k data

   π = 2% for HD data

   The priors were chosen according to the analysis with various π

# Results

## Allele error rate of imputation

| Breed | N_ref | N_test | Error rate % |
|---|---|---|---|
| Holstein | 457 | 100 | 0.77 |
| RDC | 556 | 150 | 0.96 |

DNK: 1.75%, SWE: 0.59%, FIN: 0.54%

**Validation procedure:**

- Test data were creared by deleting the markers not in 50k chip
- Imput those markers
- Allele error rate = number of wrong alleles / total number of imputed alleles

# Reliability of DGV in Holstein

| Trait | N | GBLUP | | | Bayesian mixture | | |
|---|---|---|---|---|---|---|---|
| | | 50k | $50k_{imp}$ | HD | 50k | $50k_{imp}$ | HD |
| Protein | 1395 | 0.425 | 0.426 | 0.429 | 0.435 | 0.434 | 0.440 |
| Fertility | 1378 | 0.404 | 0.403 | 0.413 | 0.406 | 0.406 | 0.416 |
| Udder health | 1461 | 0.370 | 0.372 | 0.370 | 0.375 | 0.376 | 0.376 |
| **Average** | **1411** | **0.400** | **0.400** | **0.404** | **0.405** | **0.405** | **0.410** |

- **Reliability of DGV using HD is 0.5% higher than 50k**
- **Imputation in 50k has no effect in this population**
- **Bayesian mixture reliability is 0.5% higher than GBLUP**

# Regression of DRP on DGV in Holstein

| Trait | N | GBLUP | | | Mixture | | |
|---|---|---|---|---|---|---|---|
| | | 50k | $50k_{imp}$ | HD | 50k | $50k_{imp}$ | HD |
| Protein | 1395 | 0.853 | 0.847 | 0.863 | 0.855 | 0.845 | 0.862 |
| Fertility | 1378 | 0.972 | 0.963 | 0.994 | 0.968 | 0.958 | 0.996 |
| Udder health | 1461 | 0.952 | 0.933 | 0.946 | 0.948 | 0.927 | 0.946 |
| **Average** | **1411** | **0.926** | **0.914** | **0.934** | **0.924** | **0.910** | **0.935** |

- **HD reduces prediction bias slightly**

# Reliability of DGV in RDC

| Trait | N | GBLUP | | | Bayesian mixture | | |
|---|---|---|---|---|---|---|---|
| | | 50k | $50k_{imp}$ | HD | 50k | $50k_{imp}$ | HD |
| Protein | 923 | 0.346 | 0.358 | 0.358 | 0.346 | 0.357 | 0.359 |
| Fertility | 940 | 0.297 | 0.293 | 0.304 | 0.299 | 0.296 | 0.307 |
| Udder health | 978 | 0.244 | 0.246 | 0.257 | 0.243 | 0.248 | 0.259 |
| **Average** | **947** | **0.296** | **0.299** | **0.306** | **0.296** | **0.300** | **0.308** |

- **$50k_{imp}$ is better than 50k for protein**
- **Bayesian mixture not superior over GBLUP in this population**
- **Reliability from HD is 1.2% higher than 50k, 0.8% higher than $50k_{imp}$**

# Regression of DRP on DGV in RDC

| Trait | N | GBLUP | | | Mixture | | |
|---|---|---|---|---|---|---|---|
| | | 50k | $50k_{imp}$ | HD | 50k | $50k_{imp}$ | HD |
| Protein | 923 | 0.849 | 0.875 | 0.877 | 0.835 | 0.864 | 0.877 |
| Fertility | 940 | 0.934 | 0.939 | 0.980 | 0.933 | 0.940 | 0.980 |
| Udder health | 978 | 0.851 | 0.854 | 0.872 | 0.839 | 0.846 | 0.870 |
| **Average** | **947** | **0.878** | **0.889** | **0.910** | **0.869** | **0.883** | **0.909** |

- **$50k_{imp}$ reduces prediction bias for protein**
- **HD reduces prediction bias for all traits**

# Why only small extra gain by using HD SNP?

-- Is the advantage of increasing LD offset by increasing number of unknowns?

-- Is imputation error rate higher than that in validation?

# Conclusions

- Small extra gain by increasing markers from 50k to 500k (770k chip).
- More sophisticated methods and models are needed to get full benefit from HD markers for genomic prediction