Biology informed genomic predictions for mastitis in Danish Jersey and Nordic Red cattle

Emre Karaman[1], Zexi Cai[1], Arash Chegini[2], Martin Lidauer[2], Luc Janss[1]

[1]Center for Quantitative Genetics and Genomics, Aarhus University (QGG-AU)
[2]Natural Resources Institute Finland (Luke)

Table of Contents

1. Context

2. Data processing and methods

3. Results and conclusions

Table of Contents

1. Context

2. Data processing and methods

3. Results and conclusions

BovReg WP7: Models for biology driven genomic selection

Task 7.2: Validating biology-driven genomic selection within and across small breeds

Sub-task (AU&LUKE):
Incorporating biological information (QTL, mQTL, ATAC-seq etc.) generated in BovReg, into models of genomic prediction for mastitis, in Nordic breeds.

This project received funding from the European Union's H2020 Research and Innovation Program under grant agreement No 815668.: Identification of functionally active genomic features relevant to phenotypic diversity and plasticity in cattle

50K → HD → WGS

- Little or no improvement by using HD instead of 50K SNPs
- Little or no improvement by using WGS instead of HD SNPs
- 50K already captures most of the genetic variation
- WGS data is better utilized if selected SNPs are added to standard 50K chip
    - It maybe relatively easy to include QTL and eQTL SNPs etc., but not those from ATACseq results

Our approach

- Enrich 50K SNPs with results of “-omics” studies
- Can be implemented in routine as weighted GBLUP
- Additional analysis to obtain SNP variances is needed

Table of Contents

1. Context

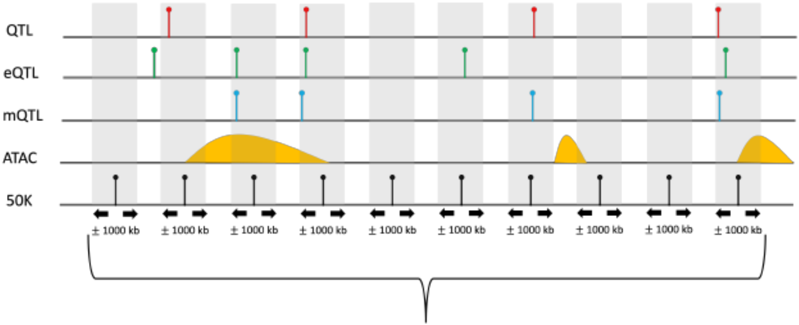
2. Data processing and methods

3. Results and conclusions

Data sets: Phenotypes and genotypes

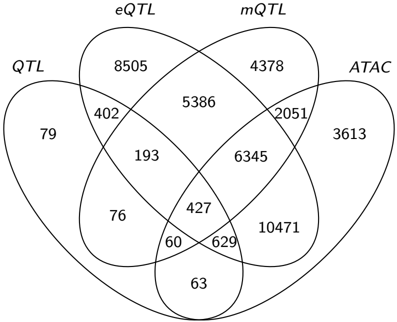
- Danish Jersey (JER) and Nordic Red Cattle (RDC)
- DRPs for mastitis, derived by LUKE were used as phenotypes
- 9,939 JER and 34,394 RDC cows with DRPs
  - Cows born in 2017 or before → reference population (JER: 8,737, RDC: 31,101)
  - Cows born in and after 2018 → validation population (JER: 1,202, RDC: 3,293)
- SNPs on 50K chip were extracted

Data sets: Annotations



$$\mathbf{Q} \rightarrow \begin{array}{l} \text{QTL} \\ \text{eQTL} \\ \text{mQTL} \\ \text{ATAC} \\ \text{50K} \end{array} \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Data sets: Annotations



Method: BayesLV

$$\beta_j \sim N(0, exp(\mathbf{Q}_j\mathbf{q} + \epsilon_j)$$

- $\mathbf{Q}$ is a matrix of known covariates to model the SNP-variances, $\mathbf{Q}_j$ is the j'th row of $\mathbf{Q}$ with covariates for SNP j.
- $\mathbf{q}$ is a vector of effects on the SNP log-variance scale.
- ϵ_i is a residual in the SNP variance model with $\epsilon_j \sim N(0, \sigma^2_\epsilon)$.
- This modelling of SNP variances allows for incorporation of additional biological information, such as genomic annotation groups and QTL information, in genomic predictions. The noise variance σ^2_ϵ is set fixed to a relatively small noise level.
- This forces SNPs that have the same annotations to have close to the same prior Normal distribution ($\mathbf{Q}_j\mathbf{q}$ terms are identical).

Table of Contents

1. Context

2. Data processing and methods

3. Results and conclusions

Reliabilities

	GBLUP	BayesLV				
	50K	+QTL	+eQTL	+mQTL	+ATAC	+All
JER	0.164	0.178	0.173	0.181	0.176	0.187
RDC	0.145	0.139	0.131	0.131	0.135	0.140

The + stands for additional information used on 50K data.

Conclusions

- Enriching 50K SNPs with the prior information from "-omics" analysis, improved reliabilities for JER but not for RDC
- It should be noted that the choice of 1,000 kb for extending SNP region downstream and upstream was arbitrary,
  - Large values may cause a large set of overlapping SNPs across different categories.
  - Small values may cause some important WGS SNPs to be missed, and thereby may lead to loss of information.
- Need further look into the analysis pipelines, to make the approach work also for RDC
- The statistical method (BayesLV) used, and its implemented sampling algorithm has not been thoroughly investigated in such a data integration study for genomic prediction, and they may be improved

AARHUS UNIVERSITY
CRG Centre for Genomic Regulation
diagenode Innovating Epigenetics Solutions
edinethics
EMBL-EBI
ETH zürich
FBN
FACULDADE DE MEDICINA VETERINÁRIA
INSTYTUT GENETYKI I HODOWLI ZWIERZĄT PAN
INRAE
LIÈGE université GIGA Medical Genomics
Université de Limoges
Luke LUONNONVARAKESKUS
MARTIN-LUTHER UNIVERSITÄT HALLE-WITTENBERG
ROSLIN
UNIVERSITY OF ALBERTA
THE UNIVERSITY OF EDINBURGH
Universiteit Utrecht
WAGENINGEN UNIVERSITY & RESEARCH

This project received funding from the European Union's H2020 Research and Innovation Program under grant agreement No 815668.