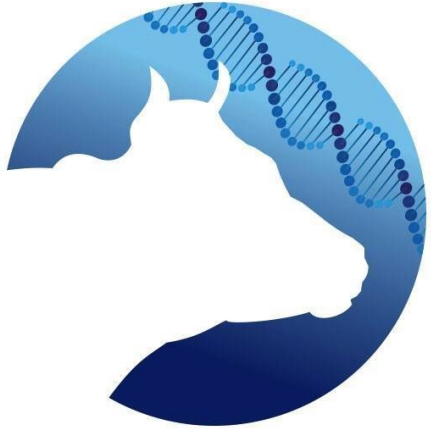
Biology-driven genomic predictions for dry matter intake within and across-breeds using WGS data

Renzo Bonifazi, M Heidaritabar, LD Barlow, AC Bouwman, L Chen, P Stothard, J Basarab, C Li, G Plastow, the BovReg consortium, B Gredler-Grandl



WAGENINGEN UNIVERSITY & RESEARCH

UNIVERSITY OF ALBERTA

BovReg Understanding cattle genomes

EuroFAANG

Biology-driven genomic predictions

- WGS: millions of variants → pinpoint **causal variants** affecting traits of interest
- Key traits: **biological efficiency**, e.g. **feed efficiency**
- BovReg: catalogue of functionally active **genomic features (GF)** in cattle
- **Functional GF** → SNP prioritization → **Biology-driven genomic predictions**

3. Aim of the study

Validation of within- and across-breed

biology-driven genomic predictions using **genomic features**

for **dry matter intake (feed efficiency)**

WAGENINGEN
UNIVERSITY & RESEARCH

UNIVERSITY OF
ALBERTA

Data available for genomic predictions

	NLD	CAN
Breed	Holstein	Beef crosses
Herds	6	14
Dependent variable	DRP on DMI	Pre-corrected DMI
n. animals DMI & geno (training - validation)	~3k (2.2k and 850)	~5.5k (4k and 1.5k)

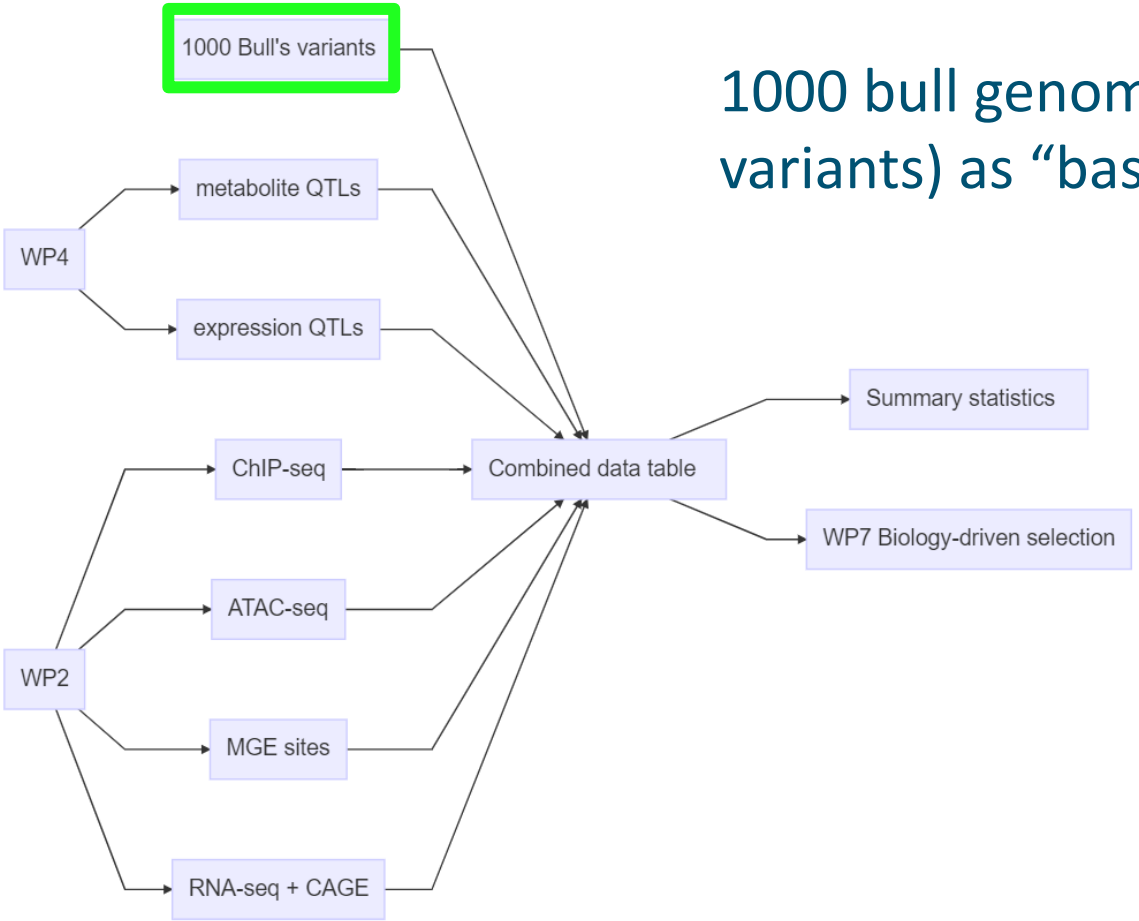
- **forward-in-time validation** (SE via bootstrapping): accuracy and bias
- 50k to imputed WGS (*Beagle*) → prioritize variants based on **GF**

Genomic features available

GF	Work Package
Meta-GWAS QTL (QTL)	WP4
Expression QTL (eQTL)	WP4
Metabolic QTL (mQTL)	WP4
ATAC-seq	WP2
ChIP-seq	WP2
Mobile Genetic Elements (MGE)	WP2

- Different tissues/samples → Different formats for different GF
- Many files (~1,000) with large size → how to quickly access/query GF information?

Selection of genomic features



1000 bull genome project (~74.9 M variants) as “baseline” to map GF

Selection of genomic features

Variant ID	**CHR**	**POS**	**GF1** (DMI QTL)	**GF2** (liver eQTL)	**GF3** (ATAC tissue1)	**GF .**	**GF*n***
Variant 1	1	659	0	1	0	.	$1.1x10^{-13}$
Variant 2	1	1235	1	0	1	.	NA
Variant 3	1	1578	0	0	1	.	NA
Variant 4	1	1984	0	0	1	.	NA
Variant 5	1	2689	0	0	1	.	$3.9x10^{-5}$
Variant 6	1	3458	0	1	0	.	NA
Variant …	.	.	.	.	.	.	.
Variant 75M	29	49805987	0	0	1	.	NA

For each GF, store information on:

- Position/Intervals as (0/1 coding)
- *p*-value (if available) as (NA/value)

Selection of genomic features

GF mapped to 1000G for all traits/tissues
position/intervals & **p-value**

Select GF group
(across traits/tissues)

Awk/R

List of variants per GF "layer"
(overlapping with 1000G)

Exclude variants in 50K &
not in within-country WGS

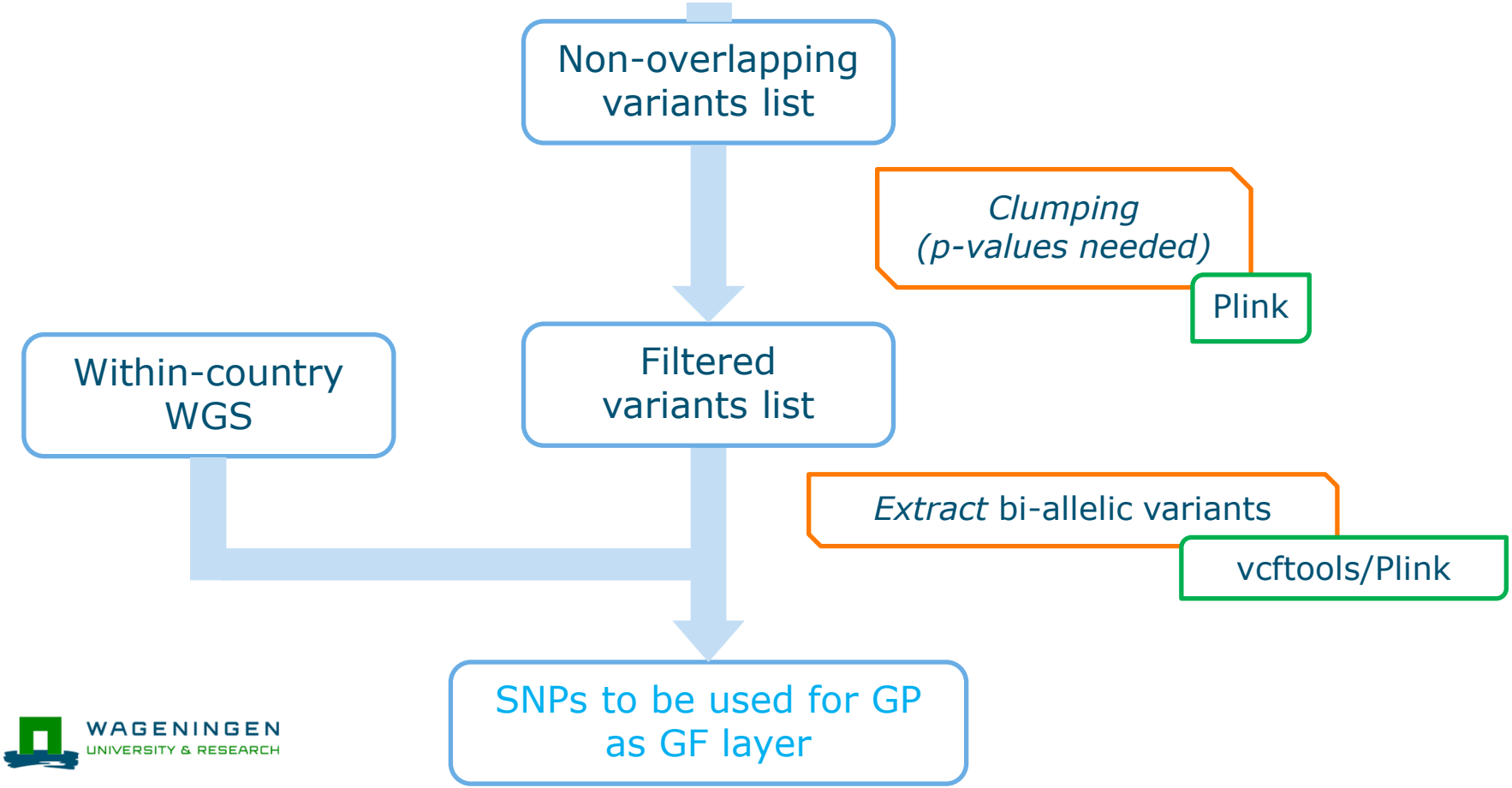
Awk/R

Non-overlapping
variants list

Variant ID	CHR	POS	GF1 (DMI QTL)	GF2 (milk QTL)	GF3 (ATAC tissue1)	GF .	GF*n*
Variant 1	1	659	0	1	0	.	$1.1x10^{-13}$
Variant 2	1	1235	1	0	1	.	NA
Variant 3	1	1578	0	0	1	.	NA
Variant 4	1	1984	0	0	1	.	NA
Variant 5	1	2689	0	0	1	.	$3.9x10^{-5}$
Variant 6	1	3458	0	1	0	.	NA
Variant …	.	.	.	.	.	.	.
Variant 75M	29	49805987	0	0	1	.	NA

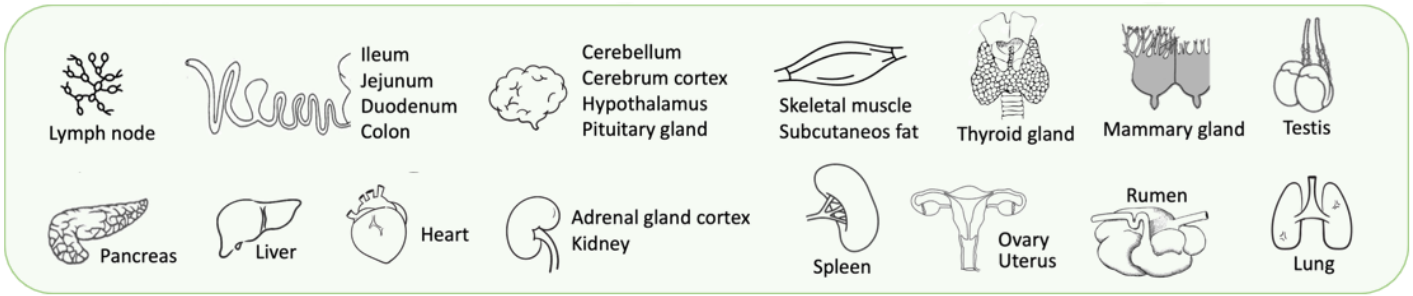
WAGENINGEN
UNIVERSITY & RESEARCH

Selection of genomic features



WAGENINGEN
UNIVERSITY & RESEARCH

Genomic features used

GF	Traits / Tissues
QTL	Meat quality, Growth, Milk production, Morphology, Fertility, Health, Feed efficiency, Methane
eQTL	(Gene, Transcript, Splice) Jejunum, Blood, Liver, Mammary Gland, Adipose, Muscle, Milk, Rumen
ATAC-seq	

From GC Moreira et al., EAAP #939

Scenarios

Scenario	NLD	CAN
50K	48K	46K
50K + **QTL**	48K + 5,416	46K + 4,222
50K + **eQTL**	48K + 12,401	46K + 11,884
50K + **(QTL, eQTL, ATAC)**	48K + 17,796	46K + 16,089

- *NextGP.jl* [1]
- Base 50K: SNPBLUP (BayesC$_0$) common variance across SNPs
- GF as additional layer → SNPBLUP or Bayesian (2 mixture model – no advantage)
- ATAC-seq *consensus peaks* (across all samples and tissues)*
- QTL, eQTL, ATAC → Multi-GF → overlapping GF

WAGENINGEN UNIVERSITY & RESEARCH

[1] https://github.com/datasciencetoolkit/NextGP.jl

*GC Moreira et al. 2023, EAAP

nf-core/atacseq

Overlapping genomic features

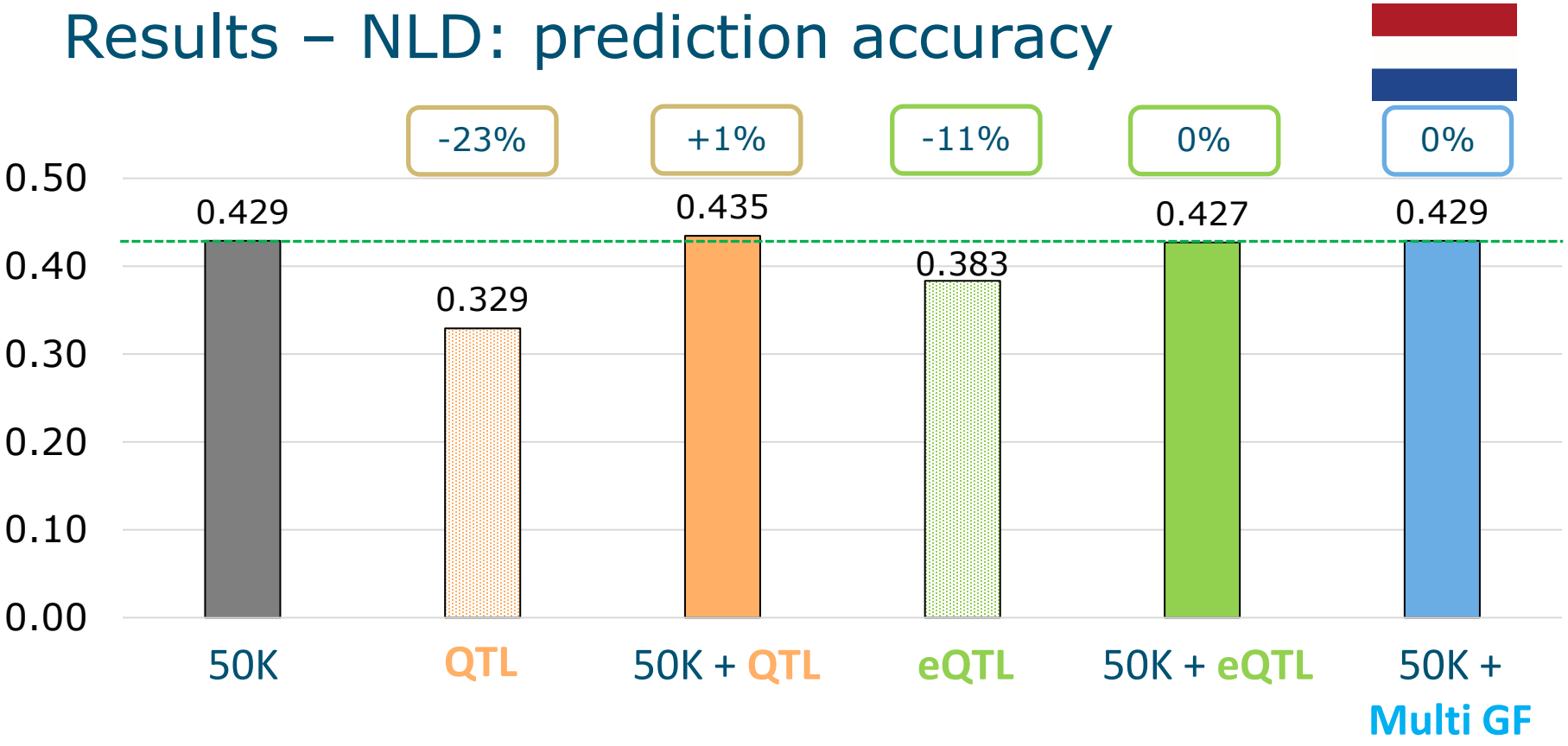
Multi-GF

QTL	eQTL	ATAC	Variants	%
1	1	0	5	0
1	1	1	16	0
1	0	0	2,576	14
1	0	1	2,819	16
0	1	0	5,051	28
0	1	1	7,329	41
			17,796	100

BayesRCπ (2 classes)

Mollandin *et al.* 2022, BMC bioinformatics

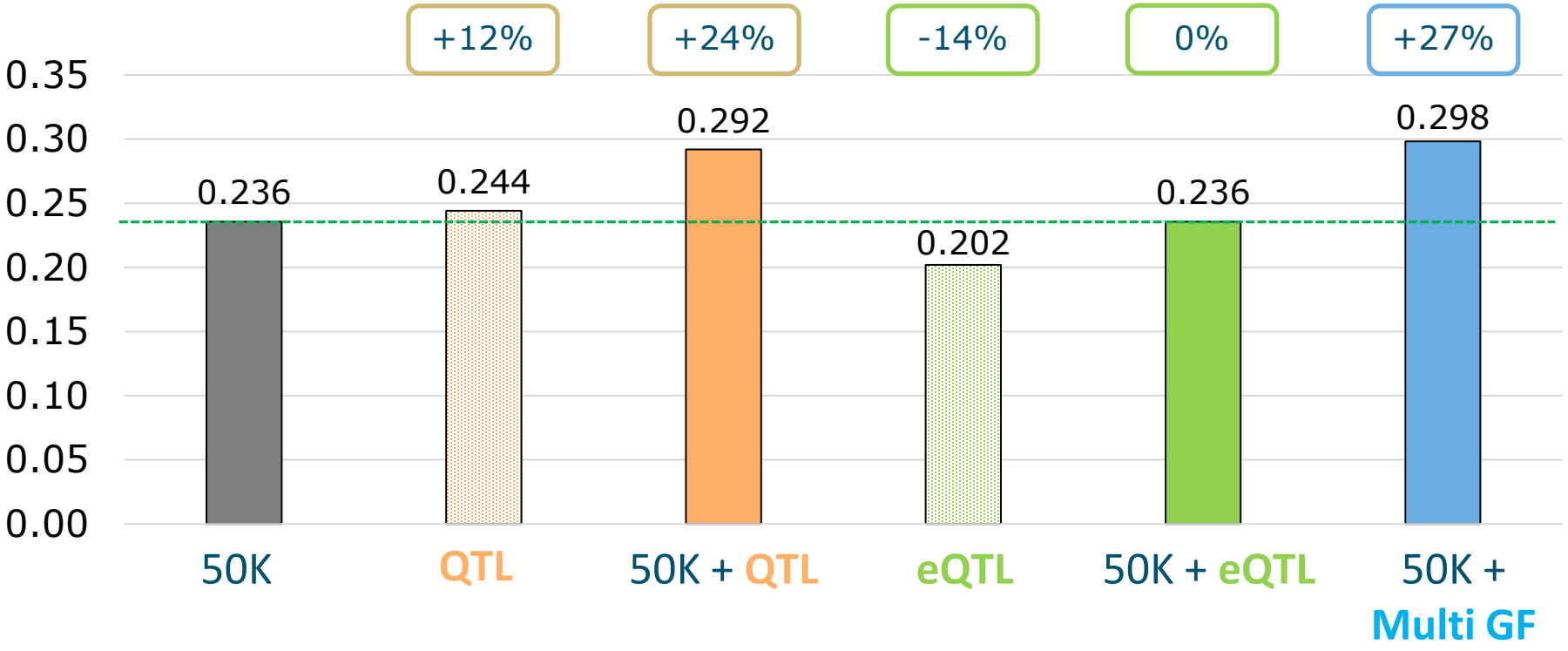
Results – NLD: prediction accuracy



- SE: ±0.03
- No impact on dispersion (slope = 0.68 for 50k)

WAGENINGEN UNIVERSITY & RESEARCH

Results – CAN: prediction accuracy



- SE between ±0.02 and ±0.03
- Similar pattern for dispersion (slope = 0.56 for 50k)

WAGENINGEN UNIVERSITY & RESEARCH

Manhattan plots meta-GWAS QTL for DMI

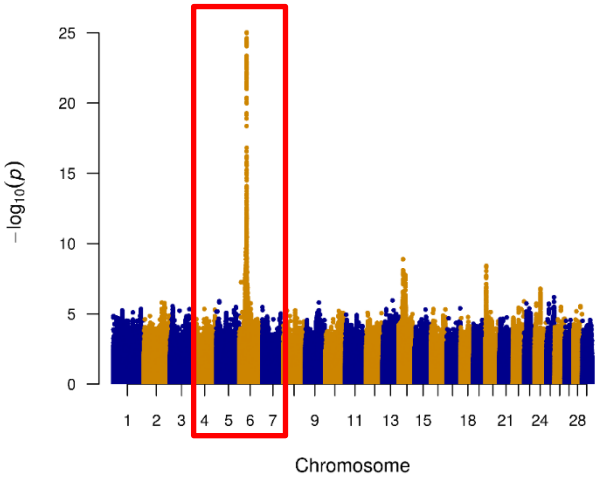
Gredler-Grandl *et al.* 2022, WCGALP

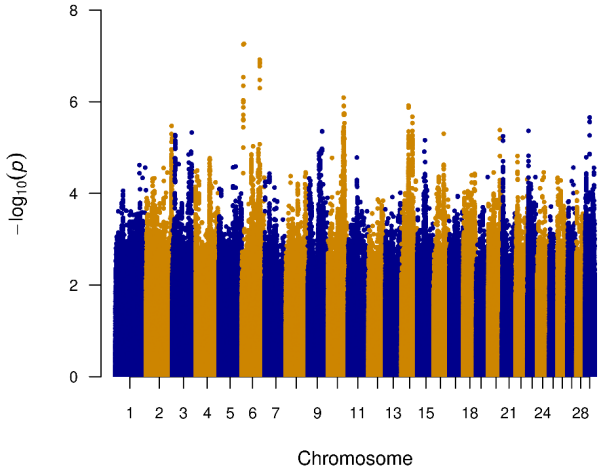
ALL
10,539 animals
Beef, Holstein, Finnish Red
~30M variants

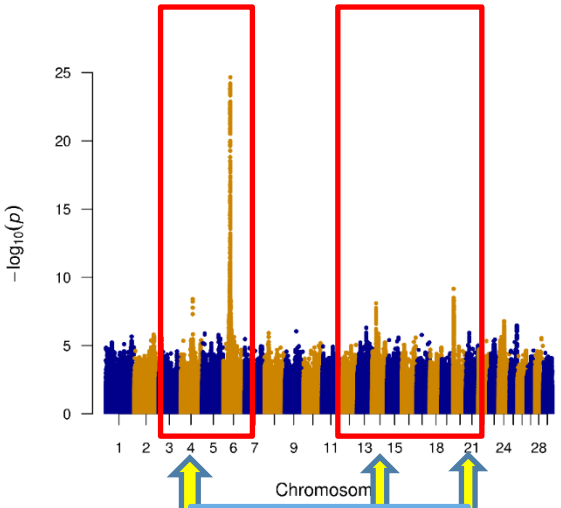
HOL
2,368 animals
Only Holstein
~20M variants

BEEF
7,805 animals
Only Beef
~28M variants

Zhang *et al*. 2020, BMC genomics





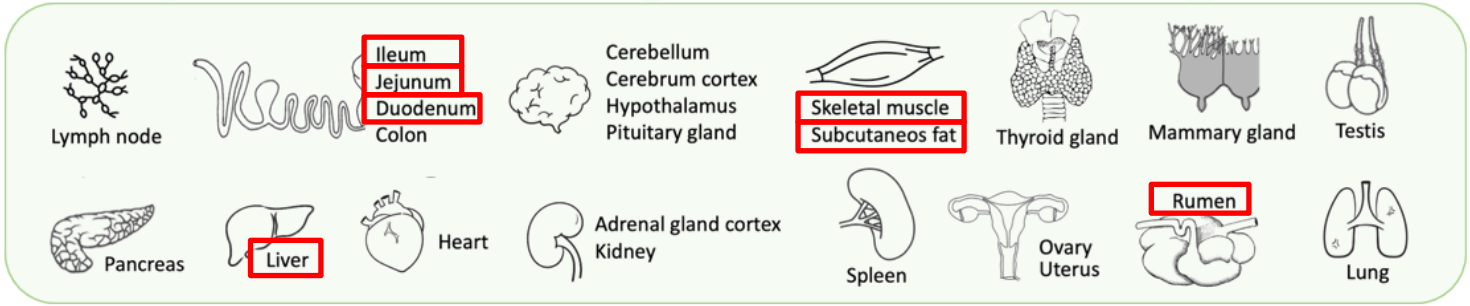


chr 6, 14, 20

data → filter SNPs with $2pq\alpha^2 \geq 0.0001$

chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
count	1	16	6	0	19	21	0	0	1	1	2	0	0	30	6	3	2	1	2	13	0	0	0	2	1	0	1	0	1

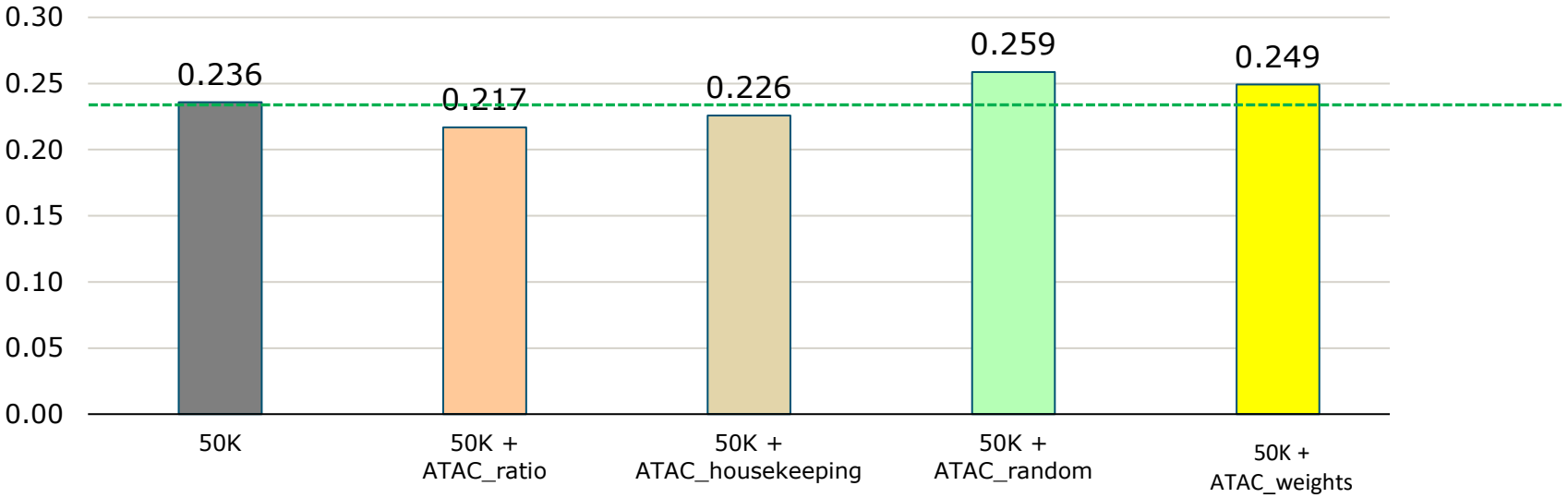
Use of narrow-peaks ATAC-seq



Different scenarios implemented using *ATAC-seq* narrow peaks for selected tissues

Scenario	Description	SNPs
ATAC_ratio	Top 10K variants from narrow peaks ratio (overlaps selected/ total tissues) **+ adjacent SNPs (200Kb)**	210,919 → 19,523 (LD pruning)
ATAC_housekeep	Top 10K variants that overlaps across all narrow peaks ("**housekeeping" set**)	10,000
ATAC_random	**Random** selection Segment size: 150K / 200K / **250K** Percentile cutoff threshold: **top 10%** / top 20% / top 30%	3,318
ATAC_weights	**Higher weights on less frequent variants** (e.g., QTLs) Segment size: 150K / 200K / **250K** Percentile cutoff threshold: **top 10%** / top 20% / top 30%	390

Results (SNPBLUP models)



- Randomly selected variants gave same or higher accuracy (# SNPs)
- ATAC-seq modelled as additional SNP layer → Next: model narrow-peaks as different layers into NextGP
- Narrow-peaks not as useful as consensus-peaks information?
- How did you model/consider ATAC-seq?

Next steps

- Model as detailed annotation layers trait/tissues-specific into NextGP
- BayesLV using *p*-values (test on implemented scenarios)
- Across-breed (using SNP effects) and multi-breed (combined reference population) GP for NLD-CAN

Conclusions

- Inclusion of **Genomic Features** could increase **genomic prediction accuracies** for **Dry Matter Intake**
- Results may vary across breeds/datasets
- Complex traits: find causal variants
- No advantage using Bayesian (2 mixture model) over SNPBLUP for GF
- ATAC-seq narrow-peaks showed low added value

Acknowledgments

Dr Emre Karaman and Dr Luc Janss

Dr Gabriel Costa M Moreira

The BovReg consortium

This work is carried out within the BovReg project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668.

BovReg
Understanding cattle genomes

EuroFAANG
www.eurofaang.eu

EuroFAANG
Research Infrastructure

FAANG
Functional Annotation of Animal Genomes
10 Years

WAGENINGEN
UNIVERSITY & RESEARCH

UNIVERSITY OF ALBERTA

Thanks for your attention

renzo.bonifazi@wur.nl