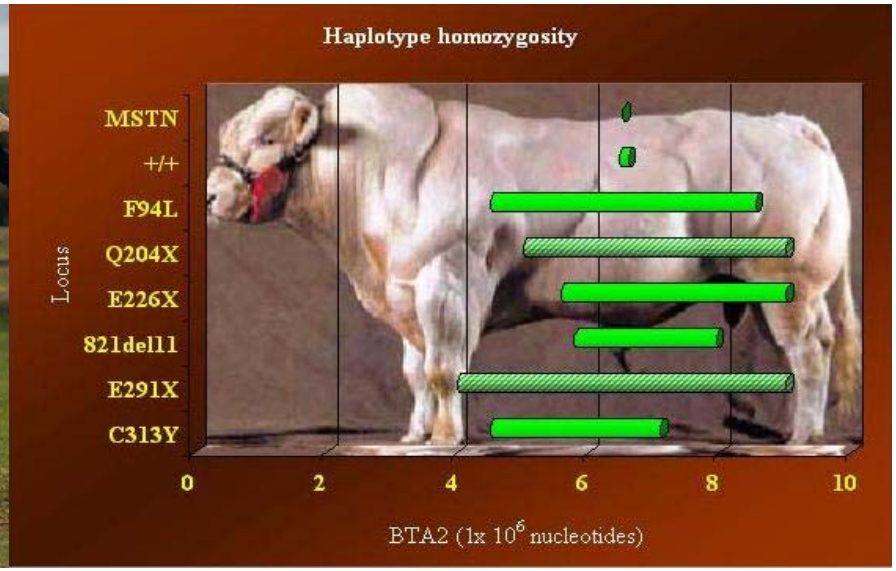
The use of multi-breed reference populations and multi-omic data to maximize accuracy of genomic prediction



M. E. Goddard[1,2], I.M. MacLeod[2], K.E. Kemper[3], R. Xiang[1], I. Van den Berg[1], M. Khansefid[2], H. D. Daetwyler[2] & B.J. Hayes[4]

[1]Faculty of Veterinary & Agricultural Science, University of Melbourne, [2]Agriculture Victoria, Bundoora, [3] Institute for Molecular Bioscience, and [4]*Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia, QLD 4067, Australia.*

THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

This talk

Introduction

Do QTL segregate across breeds?

Why are multi-breed GEBVs hard?

Solutions

THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

Introduction

GEBV accuracy is low if

reference population is small, or

target populations is distantly related to training population

Training populations within breed are too small

numerically small breed

hard to measure traits eg FCE

Therefore, use multi-breed training population

Training on a different breed to target → low accuracy

Aim = Accurate GEBVs for a breed with a small training population based on a multi-breed training population

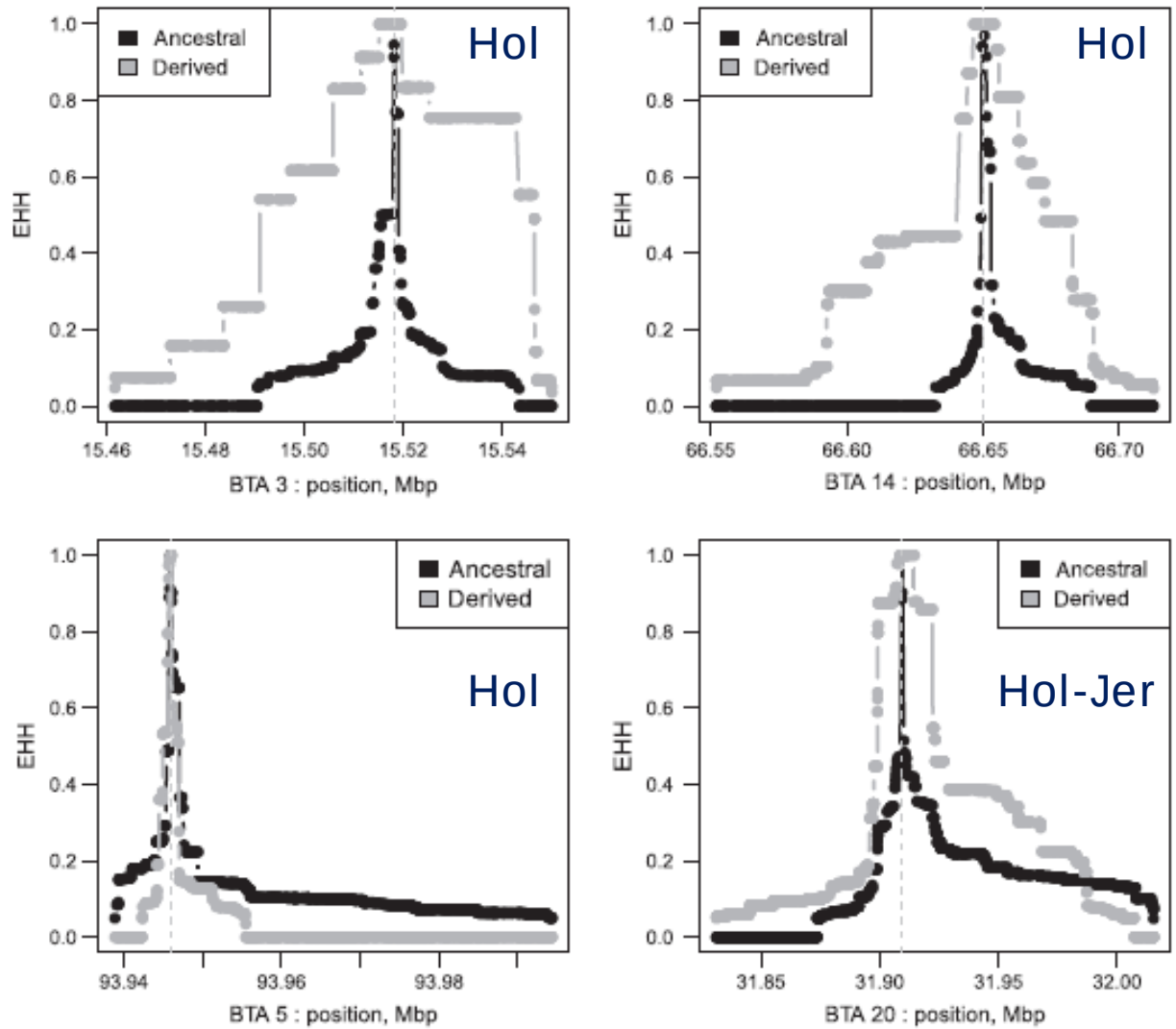
THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

Do QTL segregate across breeds?(Kath Kemper)

Young QTL

Old QTL



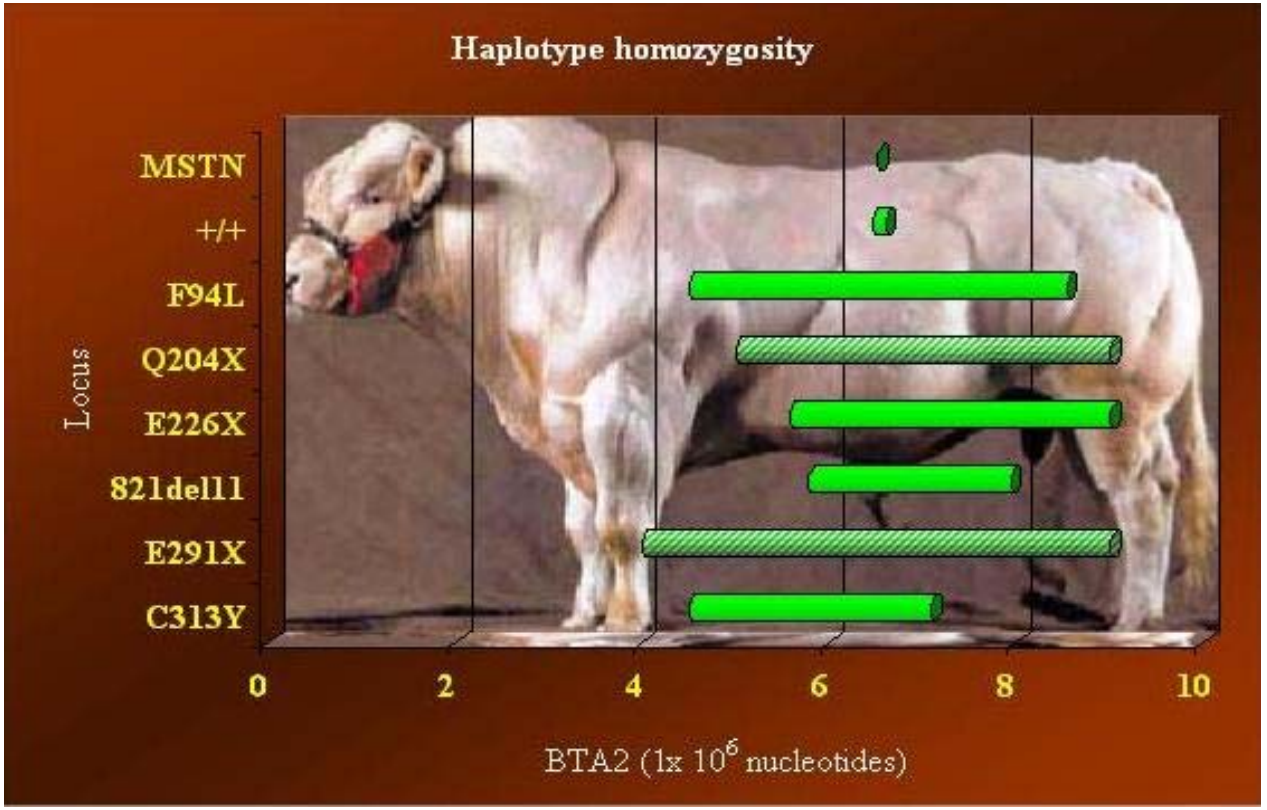
Do QTL segregate across breeds?

Across 11 QTL, length of conserved haplotype (0.4kb-55kb) around mutation suggest age of QTL mutations varies ~ 2,000 to 50,000 generations old

Prior to breed formation

QTL can and do segregate across breeds, although drift and selection can result in fixation

Age of myostatin mutations (50 – 10 gen) (O'Rourke et al)



Why are multi-breed GEBVs hard?

SNP x breed interactions
- differences in LD phase between breeds
- QTL x breed interactions
  - Due to non-additive gene action
    - typically small variances
  - equivalent to sire x breed interactions
    - typically small
  - Low accuracy even in simulation

Differences in allele frequency
- F_{ST} is low
- QTL segregate across breeds

Why are multi-breed GEBVs hard?

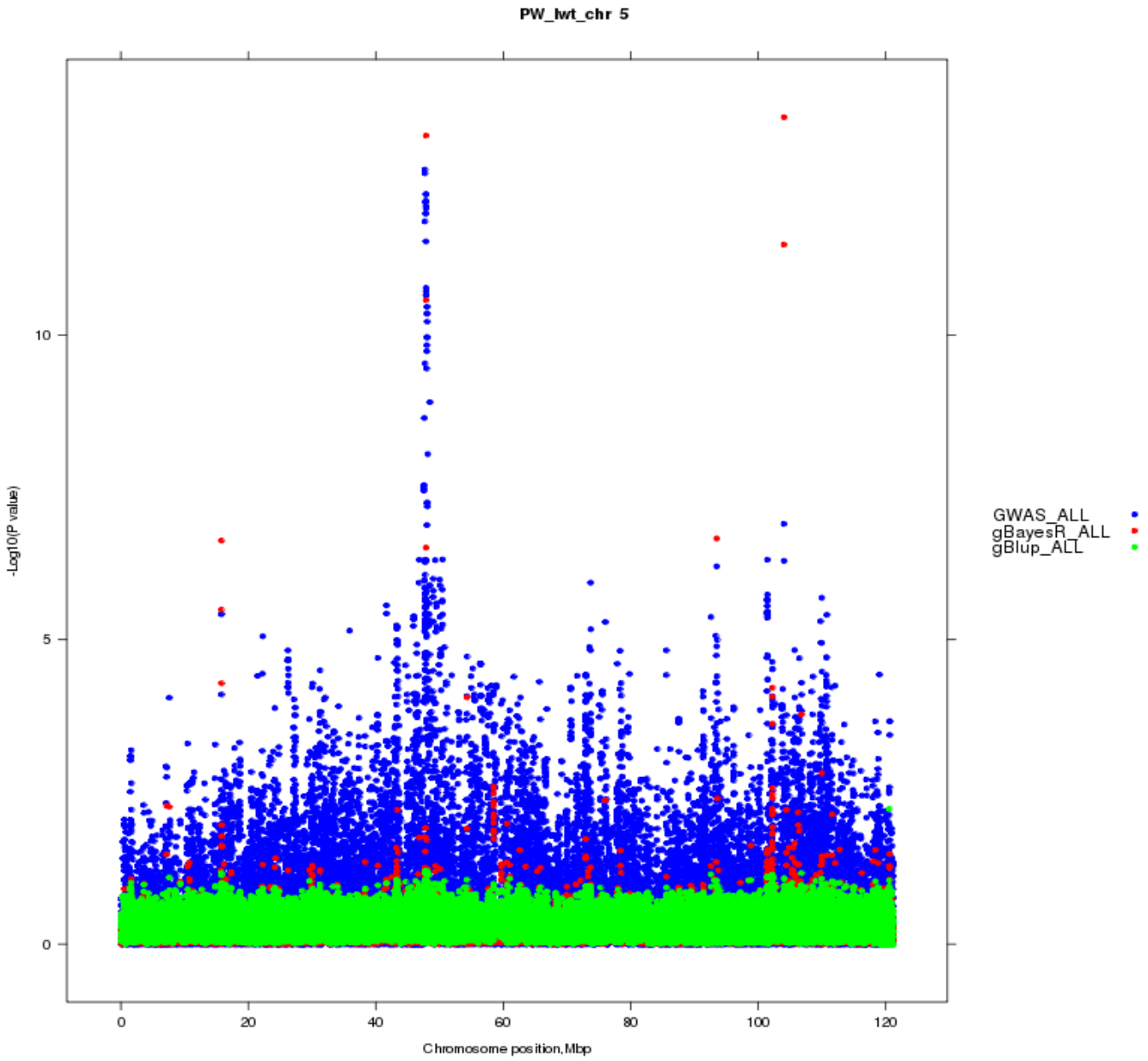
LD phase differs between breeds

Within breed GEBVs estimate the effect of large chromosome segments

This works due to LD within a breed

Effective number of chromosome segments = 5000

That is, segments 600 kb long


PW_lwt_chr 5
-Log10(P value)
10
5
0
0
20
40
60
80
100
120
Chromosome position, Mbp
GWAS_ALL
gBayesR_ALL
gBlup_ALL

Why are multi-breed GEBVs hard?

Within breed GEBVs estimate the effect of large chromosome segments

This works due to LD within a breed

Effective number of chromosome segments = 5000

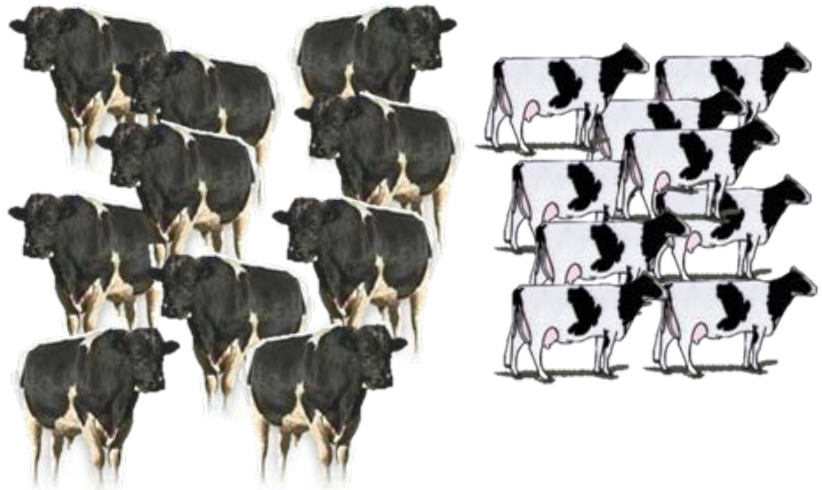
That is, segments 600 kb long

Across breeds conserved segments are much smaller (x10 smaller)

Solutions

Increase size of training population

Include target breed in training population



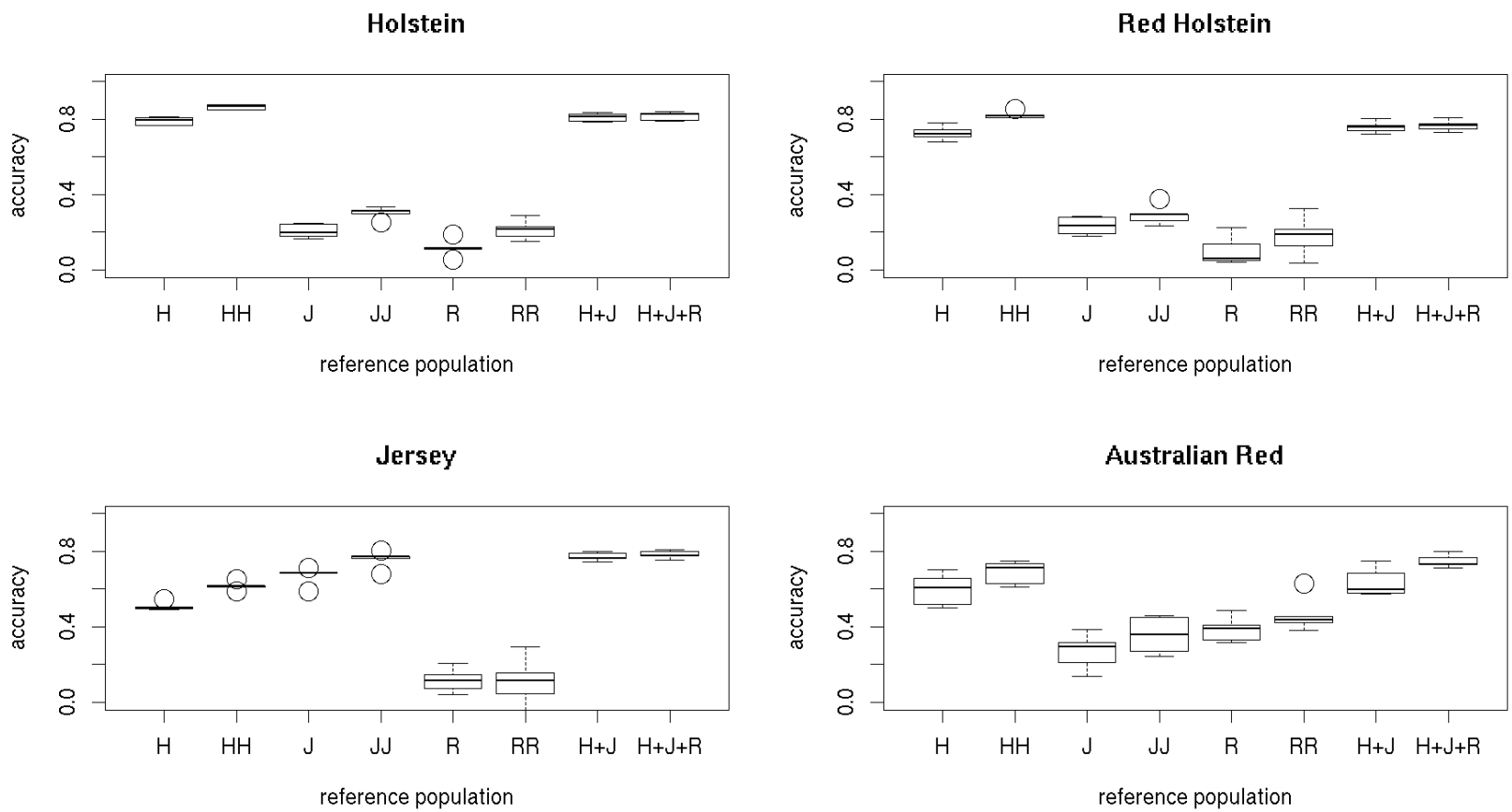
Holstein 4000 bulls, 10023 cows

Jersey 1044 bulls, 4232 cows

Aussie Reds 114 Bulls

Real or imputed 630K SNP for all individuals

Accuracy of Bayes R (Irene van den Berg)



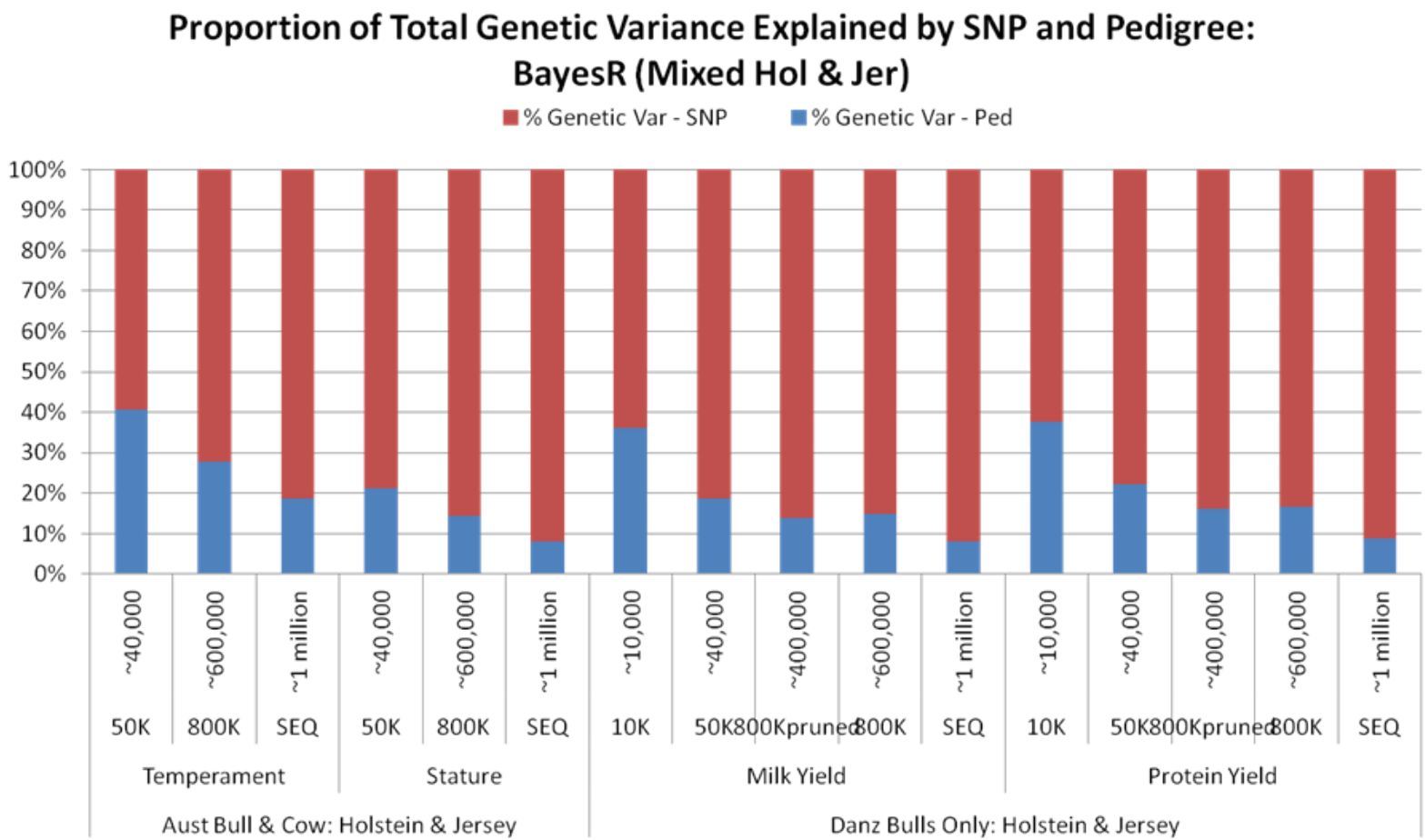
Solutions

Increase size of training population

Include target breed in training population

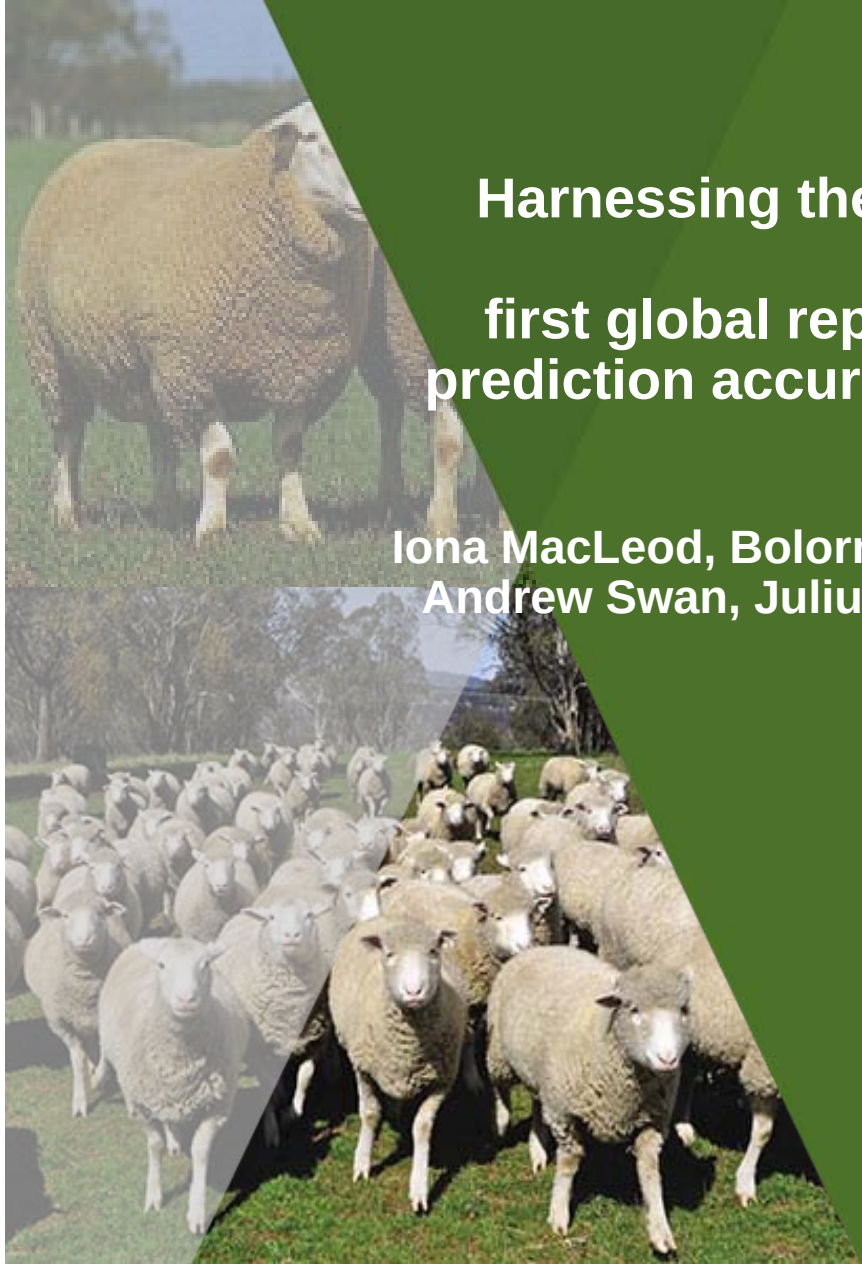
Use denser SNP panels or sequence

Variance explained by SNPs and sequence (Iona Macleod)



THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA



Economic Development, Jobs, Transport and Resources

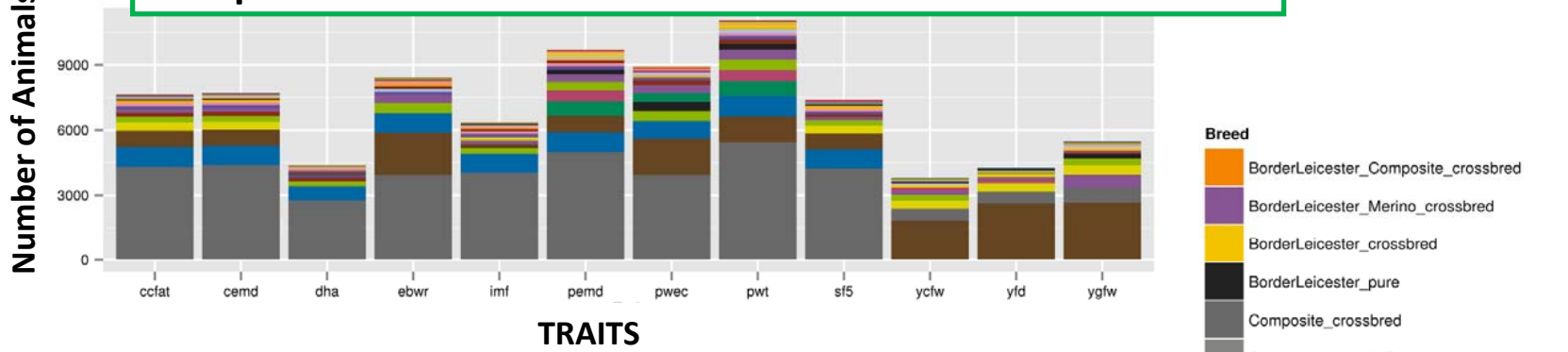
Harnessing the power of whole-genome sequence: first global report of improved genomic prediction accuracy using sequence data in sheep

Iona MacLeod, Bolormaa Sunduimijid, Majid Khansefid, Andrew Swan, Julius van der Werf & Hans Daetwyler

SHEEP CRC

AGRICULTURE VICTORIA

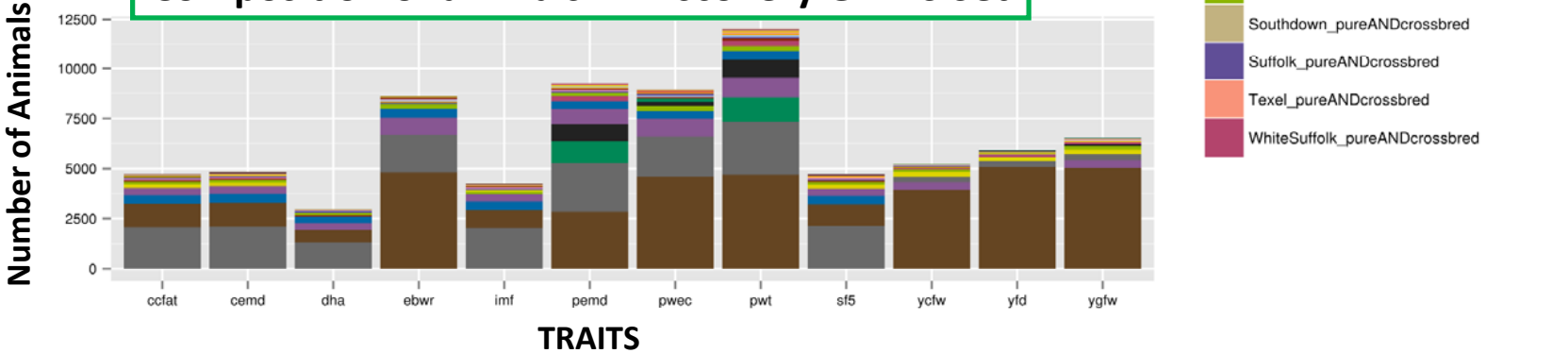
Composition of animals in Genomic Prediction Reference Set



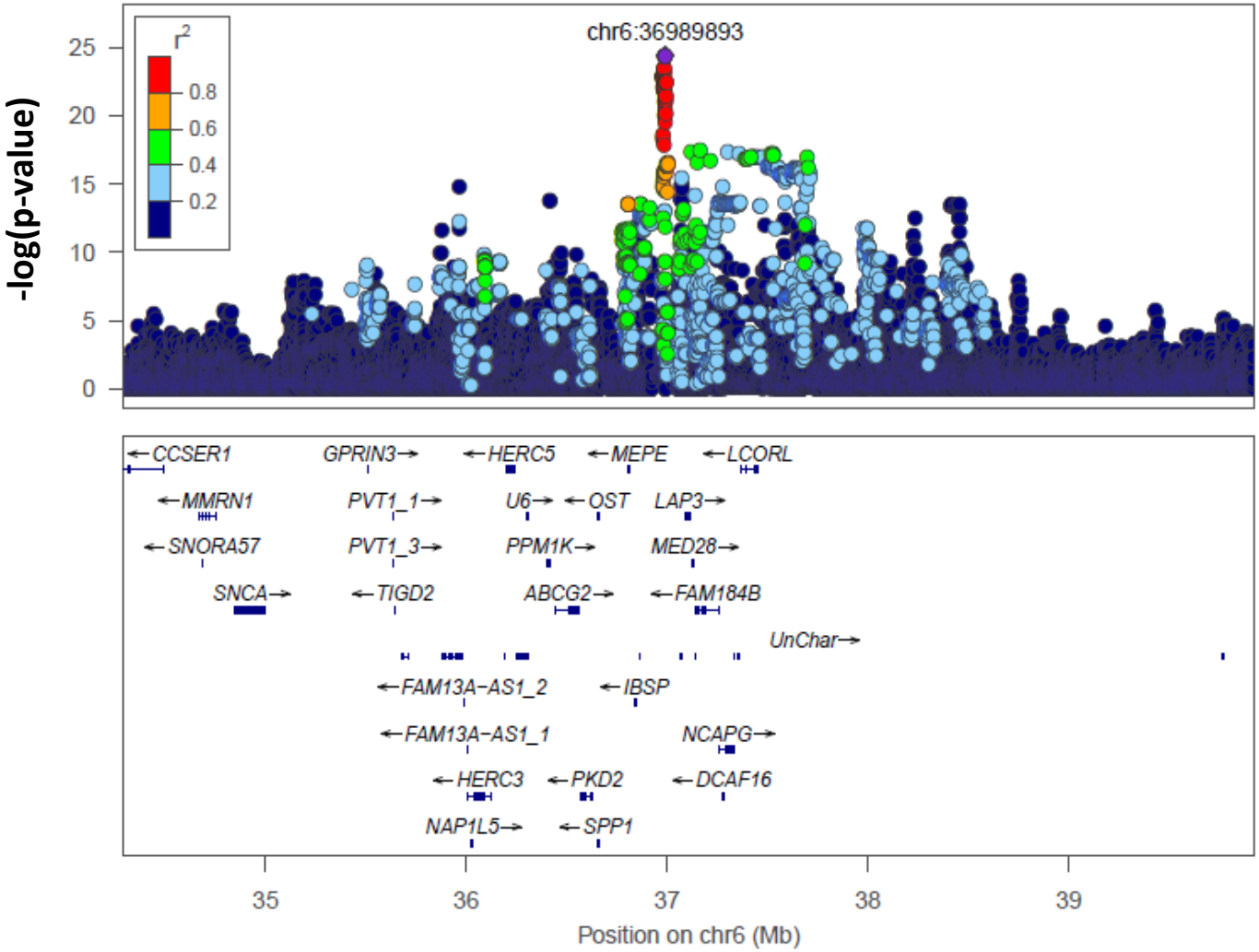
Validation sets - low relationships with Ref.:

1. **Merino**
2. **Merino x Border Leicester F1**

Composition of animals in Discovery GWAS Set



GWAS – Carcass Fat Depth (ccfat)

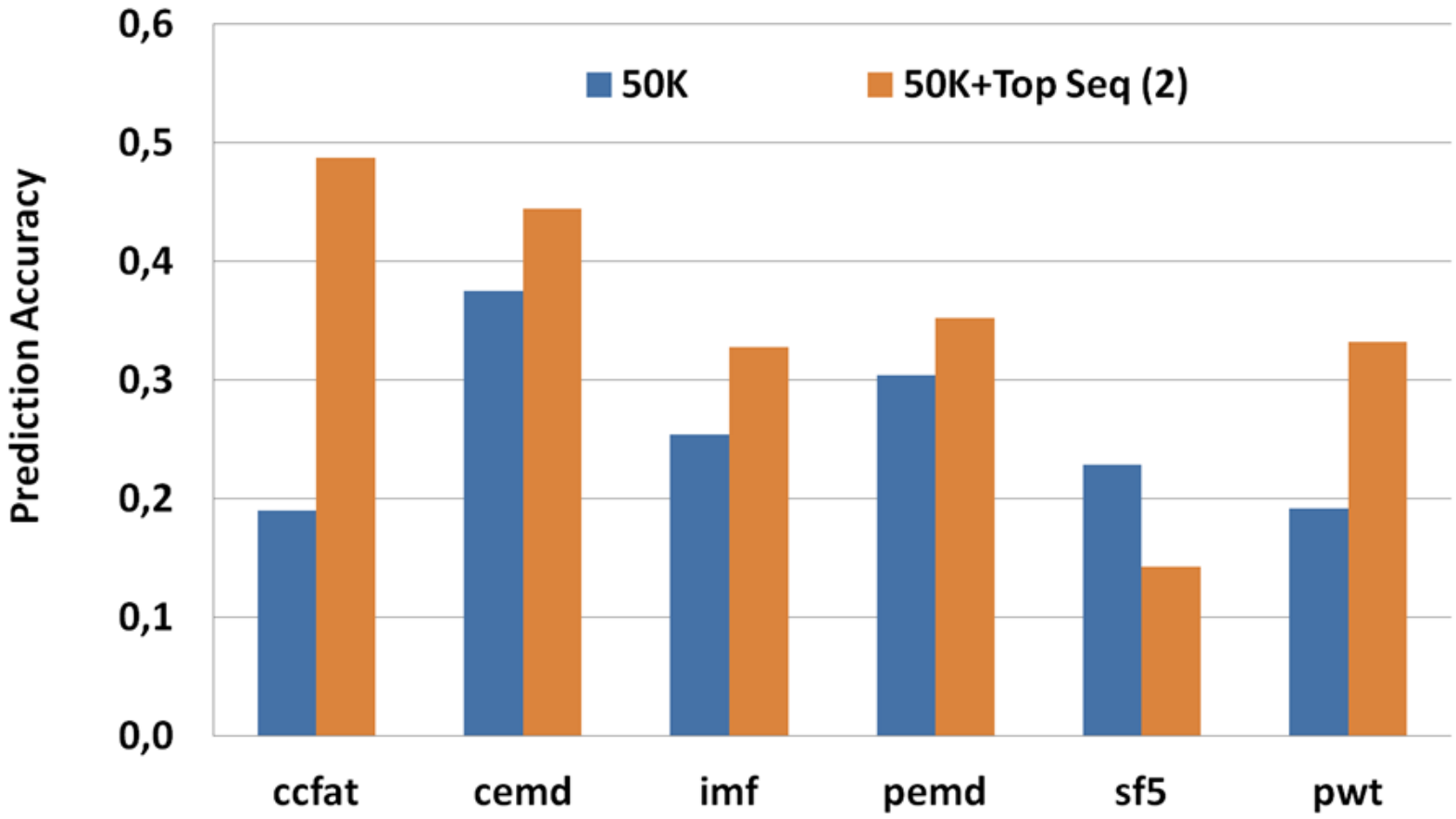


THE UNIVERSITY OF MELBOURNE

ICULTURE VICTORIA

Meat Traits:

GBLUP Accuracy - Merino x Border Leicester



THE UNIVERSITY OF MELBOURNE

JRE VICTORIA

Solutions

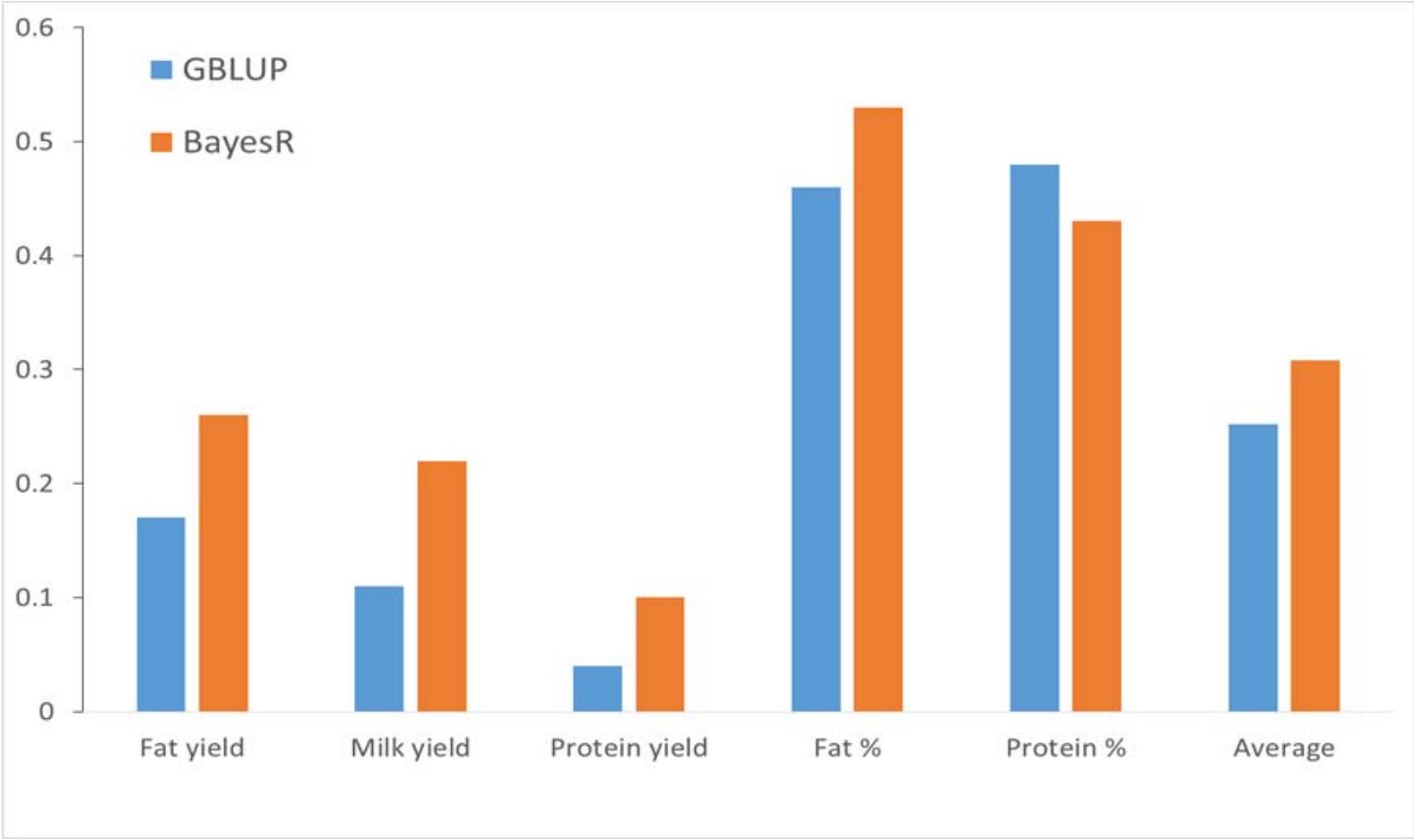
Increase size of training population

Include target breed in training population

Use denser SNP panels or sequence

Use Bayesian statistical method not GBLUP

Accuracy r(DGV,DTD) in Aussie Red Bulls (Iona MacLeod)

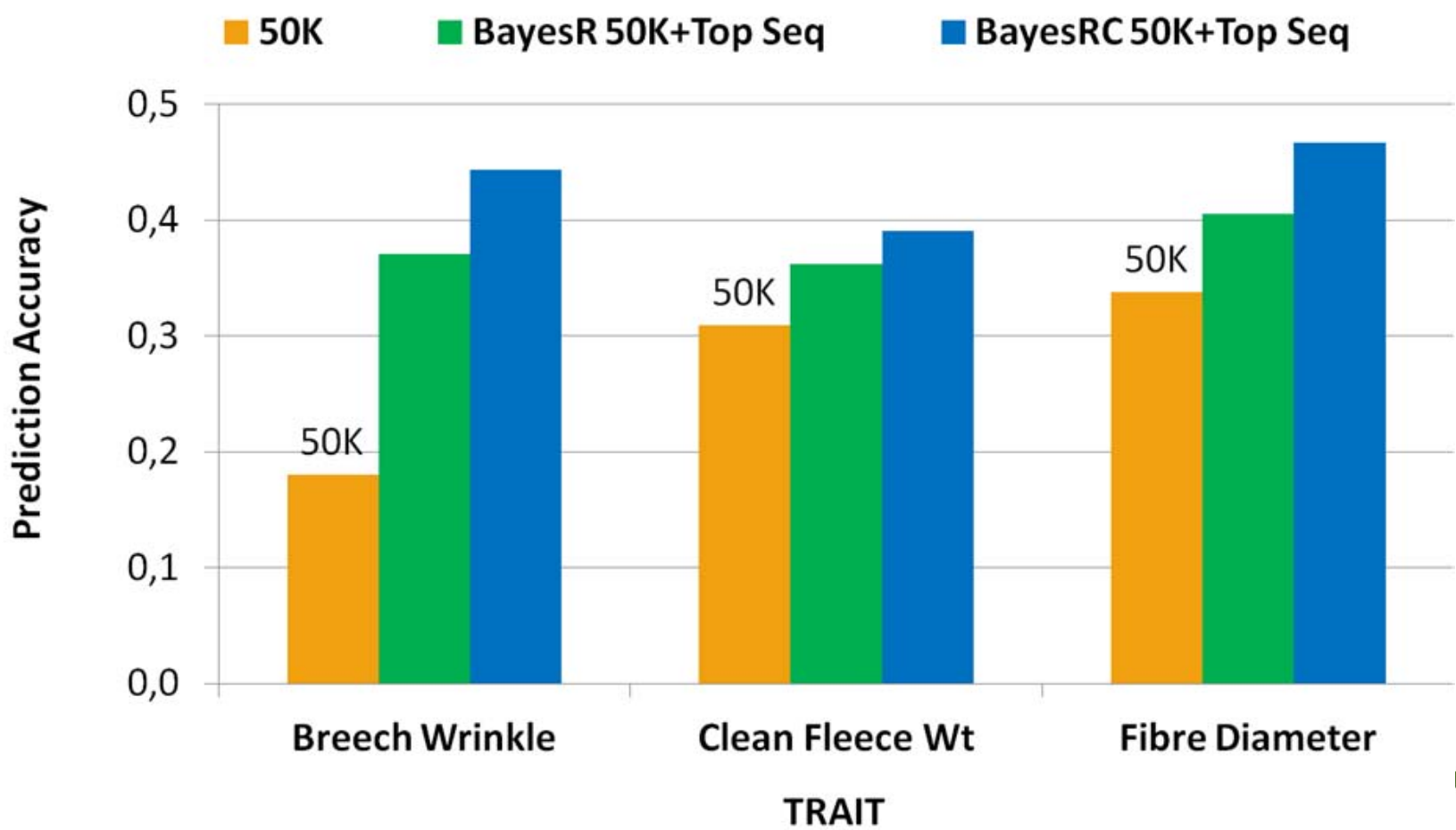


THE UNIVERSITY OF MELBOURNE

TURE VICTORIA

Wool Traits:

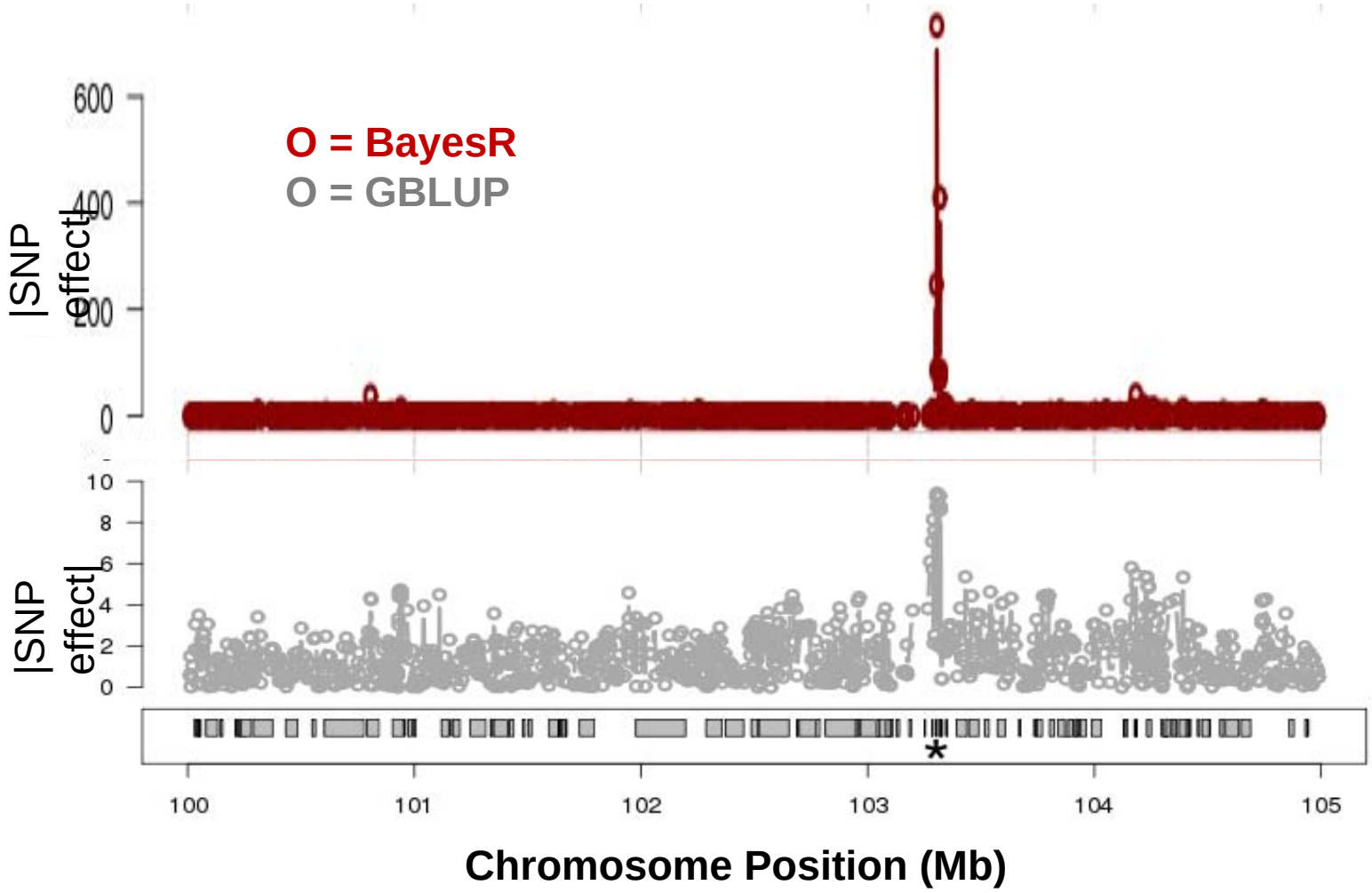
Prediction Accuracy in Merinos



THE UNIVERSITY OF MELBOURNE

URE VICTORIA

BayesR vs BLUP (BTA11)



Solutions

Increase size of training population

Include target breed in training population

Use denser SNP panels or sequence

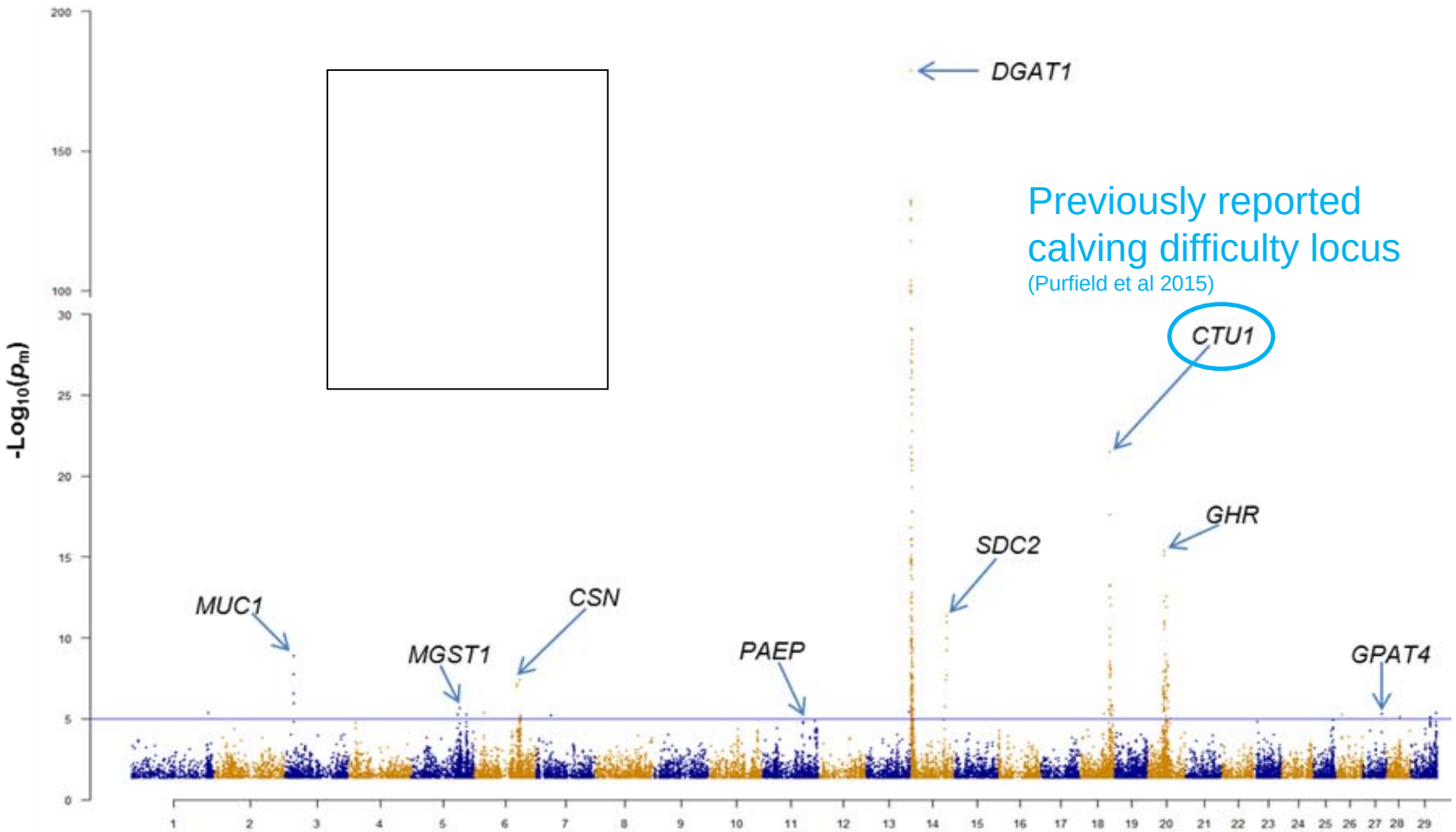
Use Bayesian statistical method not GBLUP

Use multiple traits

THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

Multi-trait GWAS (Ruidong Xiang)



THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

Validation of lead pleiotropic SNPs (Ruidong Xiang)

Select 21 lead pleiotropic SNPs and confirmed by conditional analysis in bulls

Linear index validation of lead pleiotropic SNPs in cows:

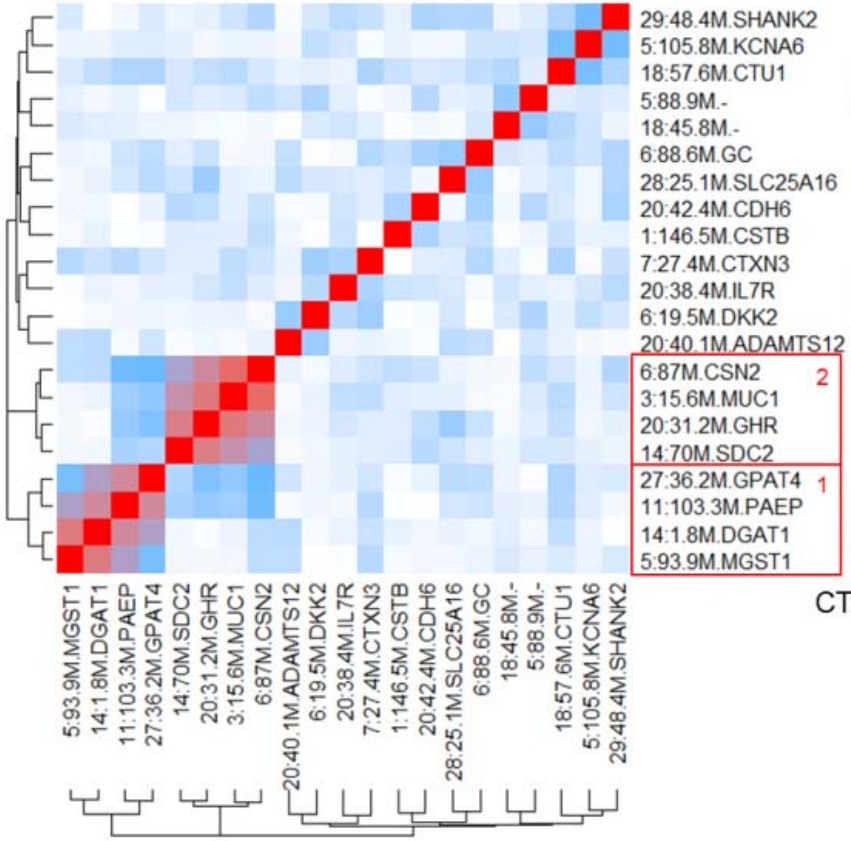
Phenotype	SNPs no.	SNP no. with the same effect directions	Percent	SNPs no. P<0.05 in validation GWAS	Percent
RT	21	21	100%	17	81%
PC		21	100%	18	86%
CT		21	100%	17	81%

THE UNIVERSITY OF MELBOURNE

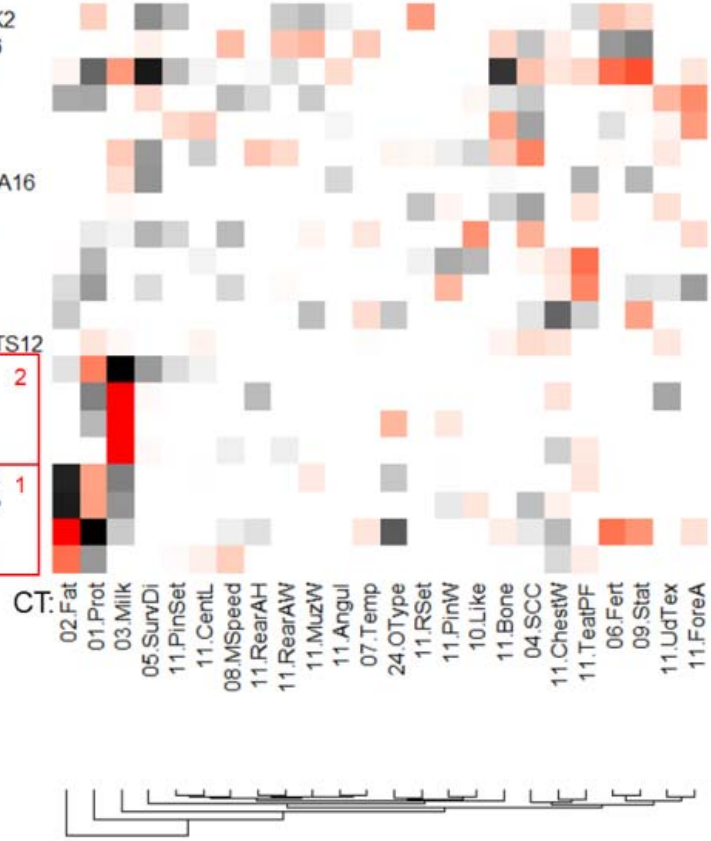
CULTURE VICTORIA

The effects of lead SNPs across independent traits

Cluster of SNPs



Effect of SNPs



THE UNIVERSITY OF MELBOURNE

RE VICTORIA

Solutions

Increase size of training population

Include target breed in training population

Use denser SNP panels or sequence

Use Bayesian statistical method not GBLUP

Use multiple traits

Use gene expression

THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

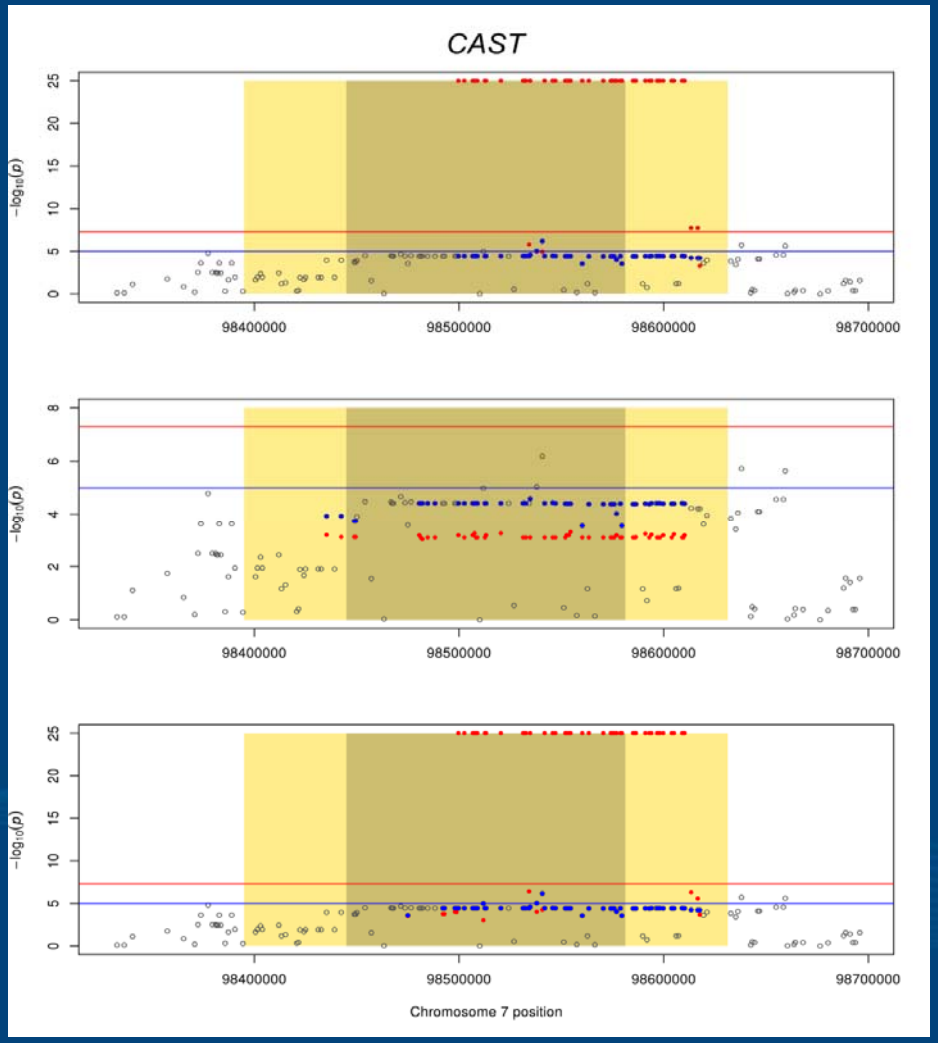
Number of cis eQTL in cattle (Ben Hayes)

	Milk		Blood	
-log10Pvalue	Significant SNP	FDR	Significant SNP	FDR
1	10,019,870	0.958	10,061,484	0.826
2	1,150,197	0.835	1,637,047	0.507
3	173,662	0.553	422,948	0.196
4	40,601	0.237	176,161	0.047
5	15,299	0.063	98,340	0.008
6	6,831	0.014	60,538	0.001
7	3,340	0.003	38,413	0.000
8	2,201	0.000	26,655	0.000

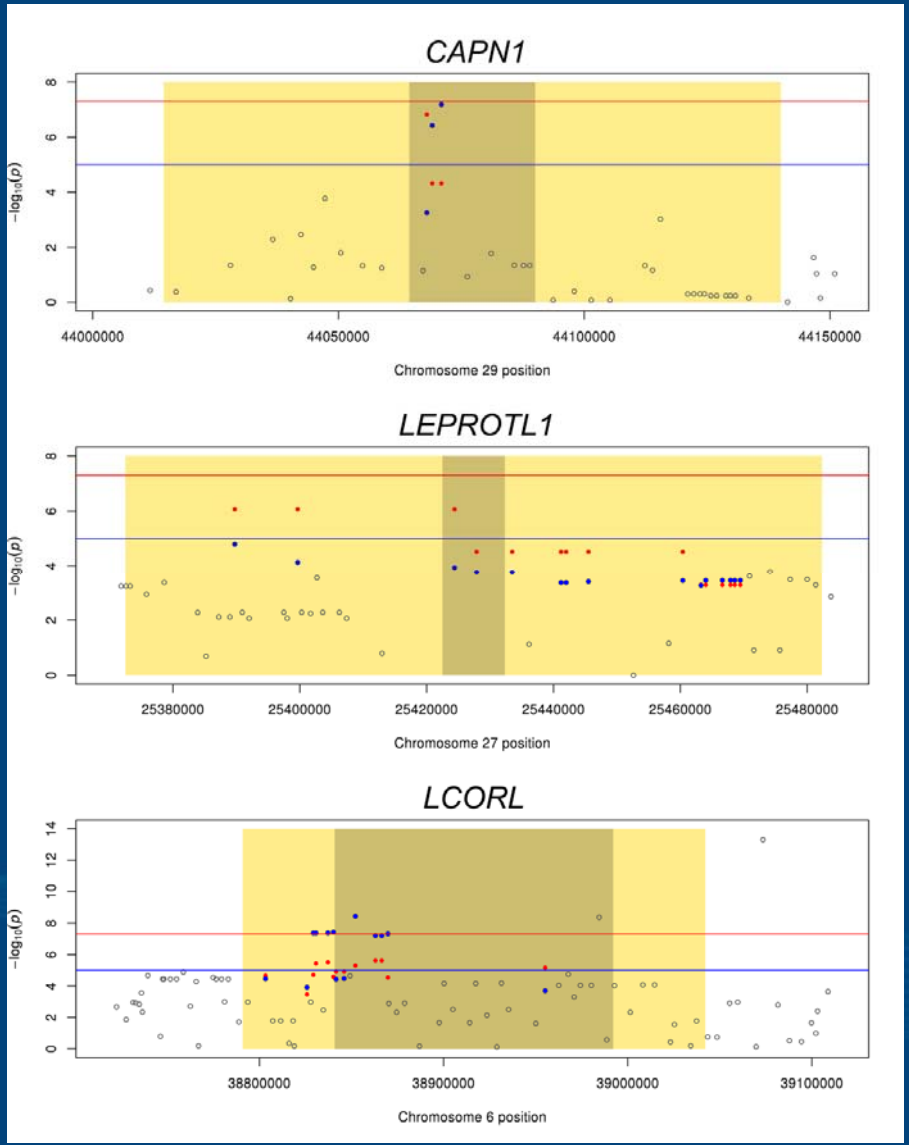
THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

eQTL and QTL (meat quality) comparison within 50kb of calpastatin (Majid Khansefid)



eQTL and QTL (meat quality, PW hip height and multi-trait) overlap



	Effect	P-value	Prop. σ^2_P
Additional traits			
phosphorus conc.	**41.8**	**1.10×10^{-11}**	**0.107**
eSLC37A1	**0.160**	**3.55×10^{-18}**	**0.224**
Key production trait, milk yield			
milk yield – Holstein cows	**-37.6**	**2.19×10^{-3}**	**0.001**
milk yield – Holstein bulls	**-40.3**	**3.17×10^{-3}**	**0.003**
milk yield – Jersey cows	**-45.2**	**3.26×10^{-3}**	**0.002**

That is the allele that *increases* expression of SLC27A1 (an antiporter):

1. *Increases* phosphorus concentration
2. *Decreases* milk yield

(Kemper et al)

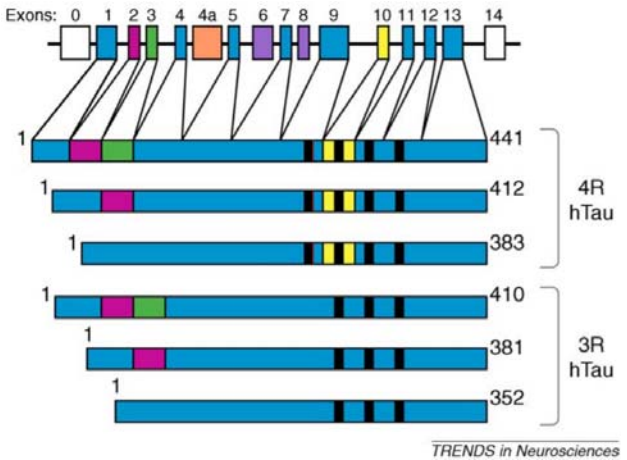
Solutions

Gene expression data

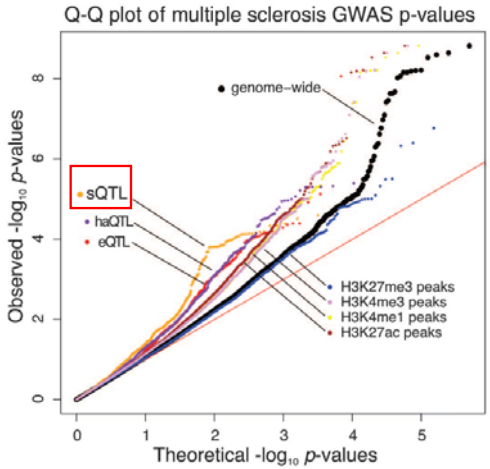
- gene cis eQTL
- splicing cis eQTL
- exon cis eQTL

Phenotypic differences due to splicing

- Human Tau gene splicing related to the Alzheimer's disease

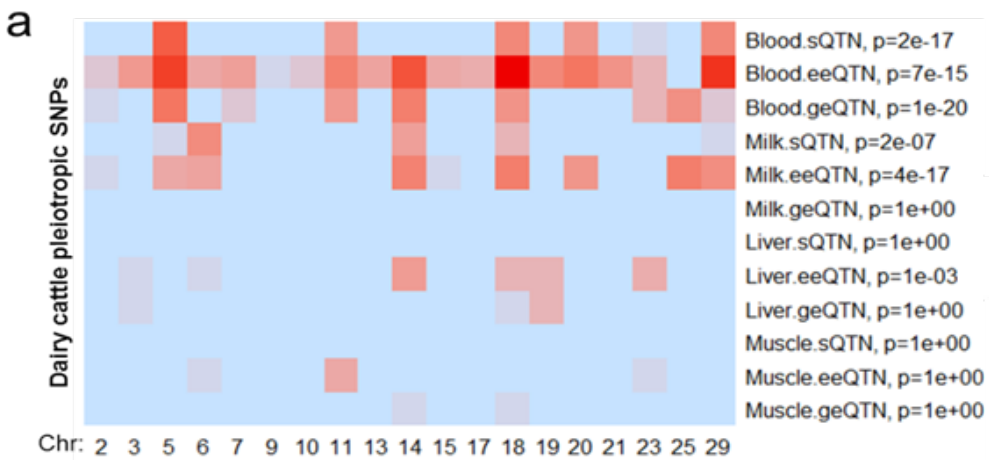


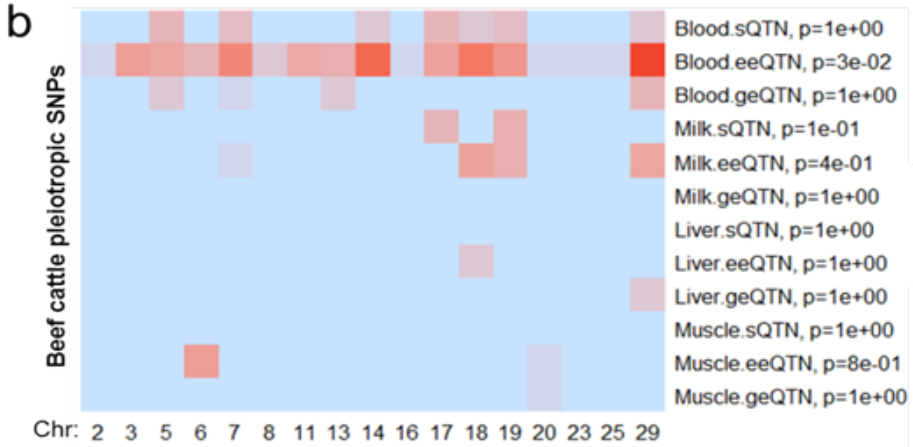
- Many genome variants affecting gene splicing, sQTL contribute to human diseases

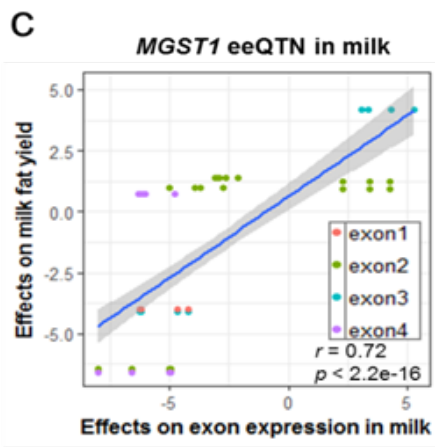


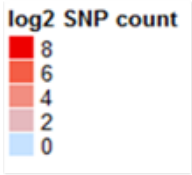
Li et al., 2016

Overlap between eQTL and milk QTL (Ruidong Xiang)

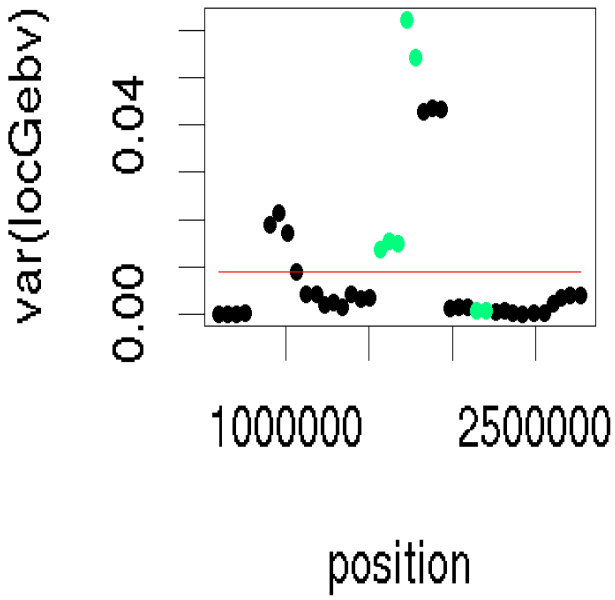


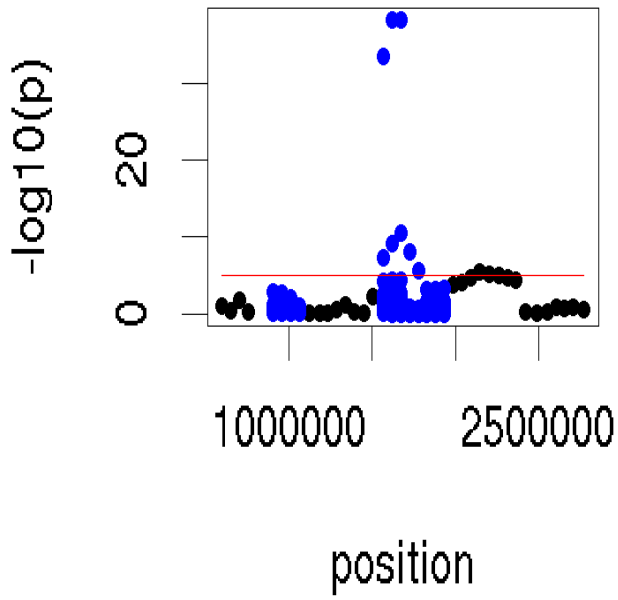






Example: FUK, chr 18, fat yield (Irene van den Berg)





Solutions

Include target breed in training population

Use denser SNP panels or sequence

Use Bayesian statistical method not GBLUP

Multi-trait analysis e.g. gene expression data

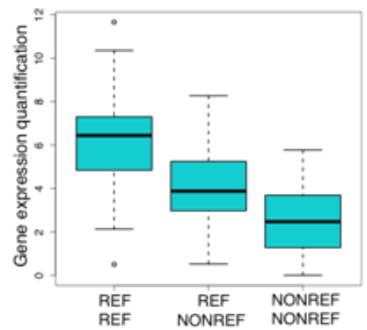
Use functional annotation of genome

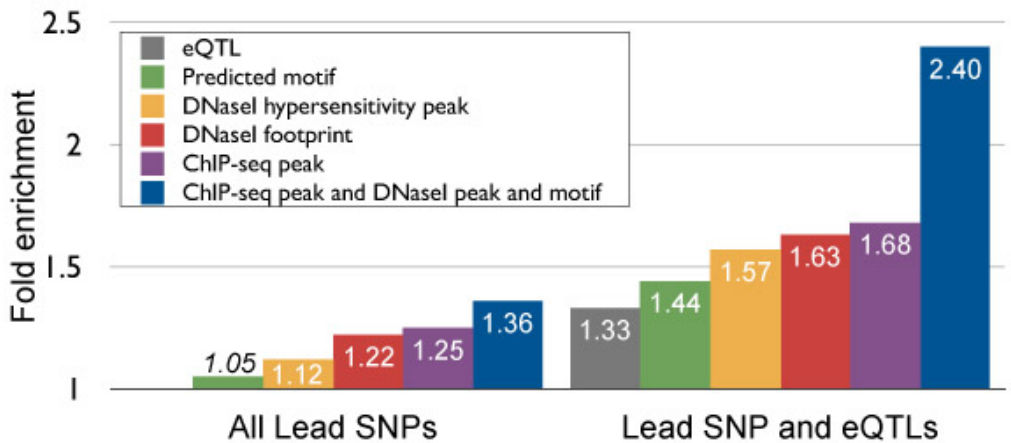
THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

SNP effects at cellular level

- Quantify the impact of a mutation on gene expression levels



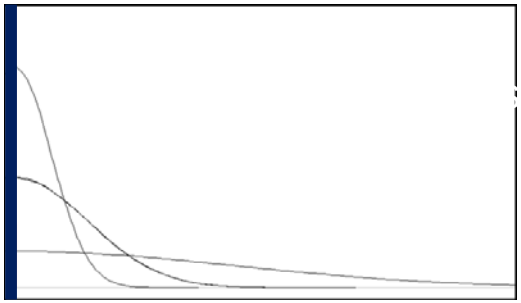


Schaub et al 2012

Genomic prediction – Milk (Iona MacLeod)

- BayesR

	0.0	0.0001	0.001	0.01
	1	2	3	4
Total SNP	Zero	Tiny	Small	Medium
905,813	**99.3%**	**0.69%**	**0.004%**	**0.001%**



- BayesRC

		0.0	0.0001	0.001	0.01	
SNP Class	No. SNP	1	2	3	4	Variance explained
Lact genes + NSC	3768 (**0.4%**)	**95.0%**	**4.3%**	**0.58%**	**0.12%**	**11%**
Lact other	57722 (**6%**)	**99.3%**	**0.7%**	**0.05%**	**0.004%**	**12%**
All others	847905 (**93%**)	**99.5%**	**0.5%**	**0.01%**	**0.000%**	**77%**

THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

Cattle stature (Aniek Bouwman, Ben Hayes et al)

Annotation class	Number
intergenic_variant	83
upstream_gene_variant	11
5_prime_UTR_variant	1
intron_variant	55
missense_variant	5
downstream_gene_variant	8
ChiP-SEQ peaks*	8
WBC eQTL	10

THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA

The bad news

Accuracy only improves a little

You need to capture a high proportion of total variance

Conclusion

Data from the target breed is the most useful

But, training data from other breeds helps

Advantage to use sequence data and Bayesian method

Sequence imputation loses accuracy

Identify near perfect markers and genotype them directly

Expression data and functional annotation helps select best variants

THE UNIVERSITY OF MELBOURNE

AGRICULTURE VICTORIA