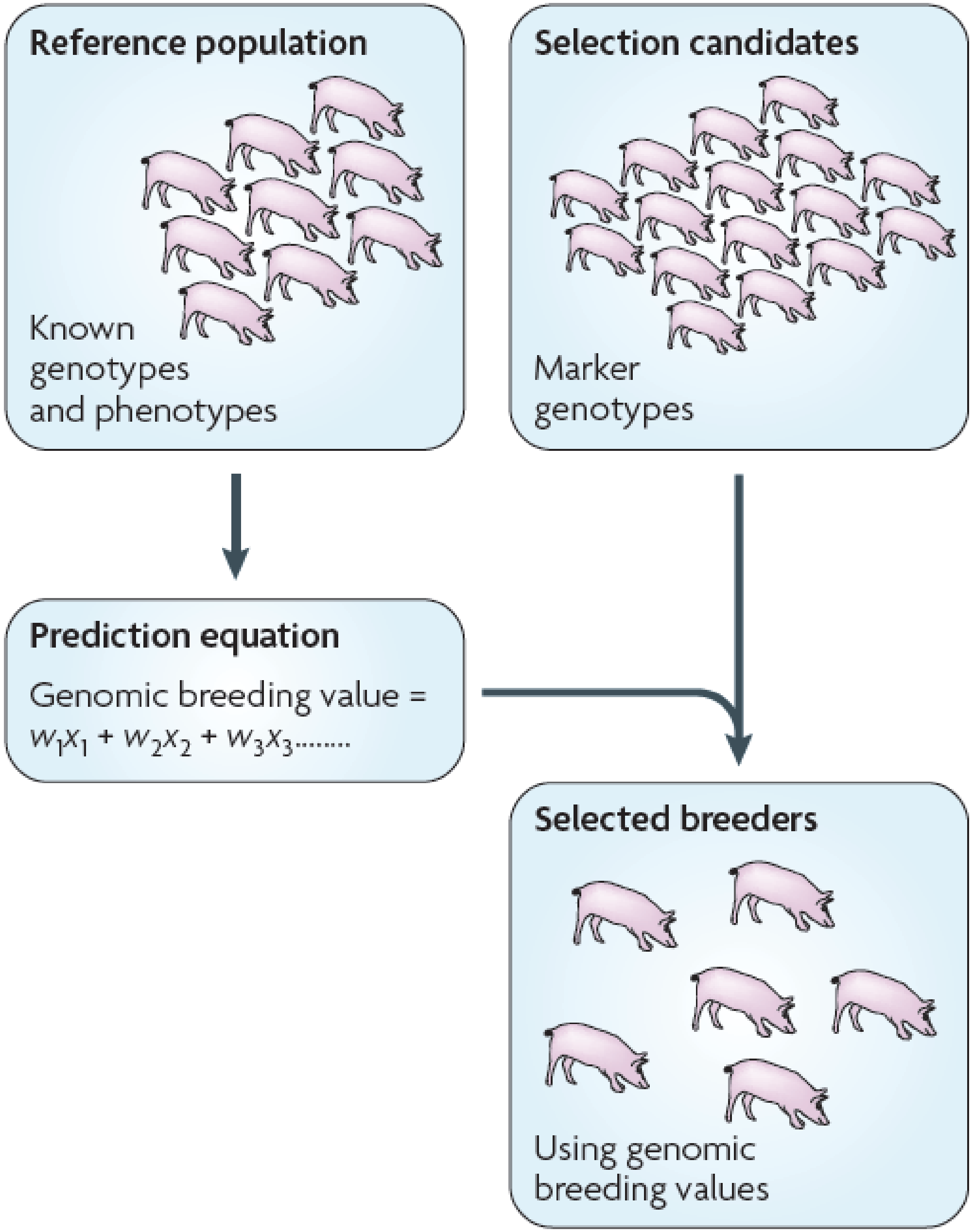
Multi-breed genomic prediction from whole genome sequence data in dairy cattle

Ben Hayes, Kath Kemper, Iona Macleod, Hans Daetwyler,
Mike Goddard and the 1000 Bull Genomes Consortium

Department of Environment and Primary Industries
State Government Victoria

Genomic selection

- > 2 million dairy cattle
- Accelerating genetic gains in livestock & crops for yields



Reference population

Known genotypes and phenotypes

Selection candidates

Marker genotypes

Prediction equation

Genomic breeding value = $w_1x_1 + w_2x_2 + w_3x_3$........

Selected breeders

Using genomic breeding values

Limitations

- Largely based on 50K SNP + BLUP
- Strictly within populations (breed)
  - linkage disequilibrium does not persist across breeds

Reference Set	Validation set	Protein	Fat	Milk	Prot%	Fat%
Holstein only	Holstein	0.49	0.44	0.59	0.61	0.62
	Jersey	-0.06	-0.02	-0.02	-0.06	0.23

Limitations

- Largely based on 50K SNP + BLUP
- Strictly within populations (breed)
  - linkage disequilibrium does not persist across breeds
- Accuracy erodes rapidly with generations

Why?

- QTL don’t segregate across breed?
- Need method other than BLUP, higher density genotypes?
- Sequence data to improve multi-breed predictions?

Do QTL segregate across breeds?

- Kathryn Kemper et al. 2015.
  - *How old are quantitative trait loci and how widely do they segregate?* J. Anim Breed. Genet. 132:12

- Detect milk production QTL in Holstein cattle
  - 632,003 SNP, GWAS, local GEBV analysis for production
  - Identify in 8478 Holsteins cows, 3049 Holstein bulls
  - Impute to sequence (1000 bull genomes Run4), identify putative causal mutations

- Do these QTL segregate in Jersey?
  - 3917 Jersey cows imputed to sequence

Do QTL segregate across breeds?

Detected in Holstein cattle………

QTL	ID	Variant	Gene
1	Chr3:15518228	Downstream	SLC50A1
2	Chr5:31330413	Intergenic	LALBA
3	Chr5:75722422	Upstream	CSF2RB
4	Chr5:93945991	Intron	MGST1
5	Chr6:87296809	Intergenic	CSN1S2
6	Chr10:46585703	Intergenic	USP3
7	Chr14:1802266	Missense	DGAT1
8	Chr14:66649826	Downstream	COX6C
9	Chr20:31909478	Missense	GHR
10	Chr27:36206783	Intergenic	AGPAT6
11	Chr29:41978596	Downstream	SLC22A6

Do QTL segregate across breeds?

Validated in Jersey cattle

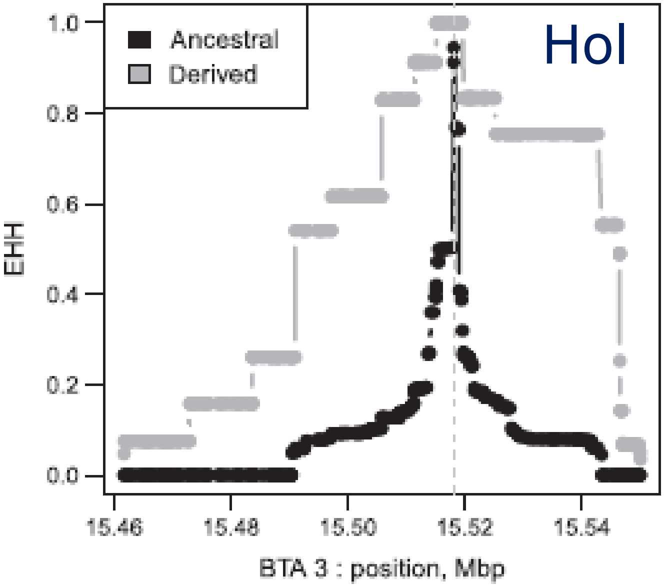
QTL	ID	Variant	Gene
1	Chr3:15518228	Downstream	SLC50A1
2	Chr5:31330413	Intergenic	LALBA
3	**Chr5:75722422**	**Upstream**	**CSF2RB**
4	Chr5:93945991	Intron	MGST1
5	**Chr6:87296809**	**Intergenic**	**CSN1S2**
6	**Chr10:46585703**	**Intergenic**	**USP3**
7	**Chr14:1802266**	**Missense**	**DGAT1**
8	Chr14:66649826	Downstream	COX6C
9	**Chr20:31909478**	**Missense**	**GHR**
10	**Chr27:36206783**	**Intergenic**	**AGPAT6**
11	Chr29:41978596	Downstream	SLC22A6

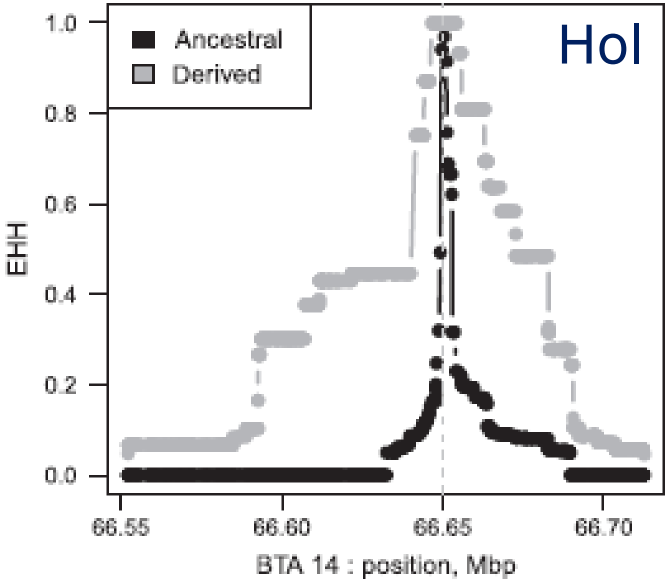
Do QTL segregate across breeds?

- Age of mutations?
  - Length of haplotype surrounding mutation
  - Longer haplotype -> more recent mutation
  - Extended haplotype homozygosity (EHH)

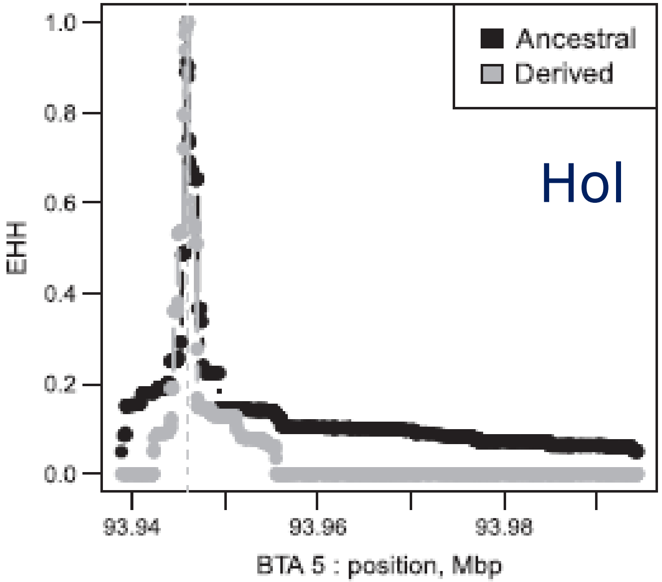
Do QTL segregate across breeds?

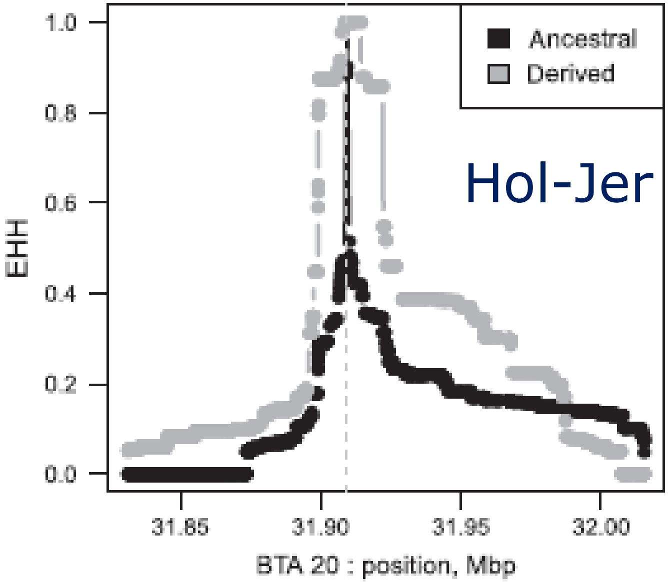
Young QTL





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, selection can result in fixation

Why?

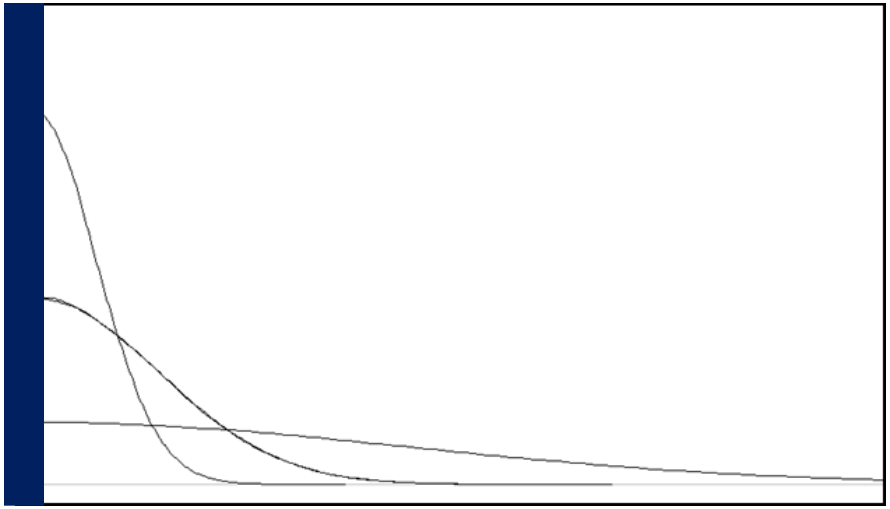
- QTL don’t segregate across breed?

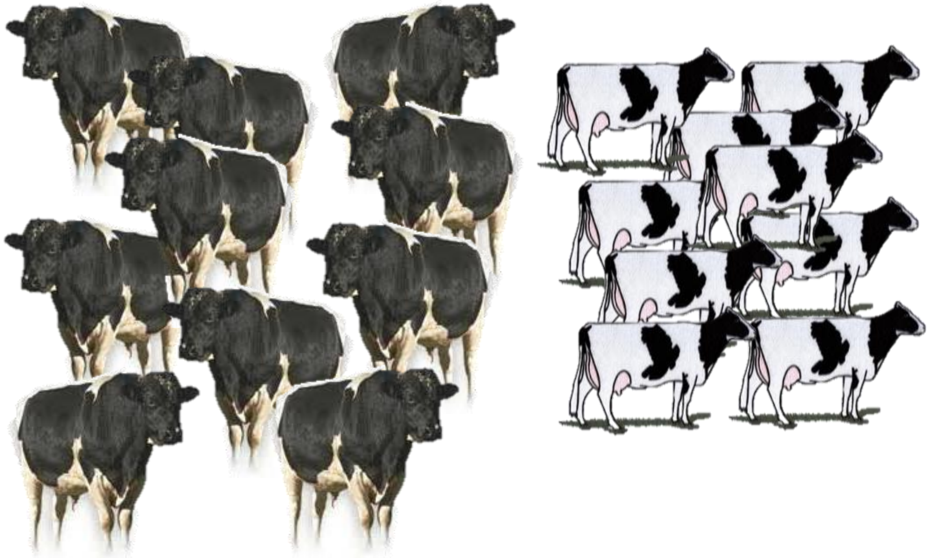
- Need method other than BLUP, higher density genotypes?

- Sequence data to improve multi-breed predictions?

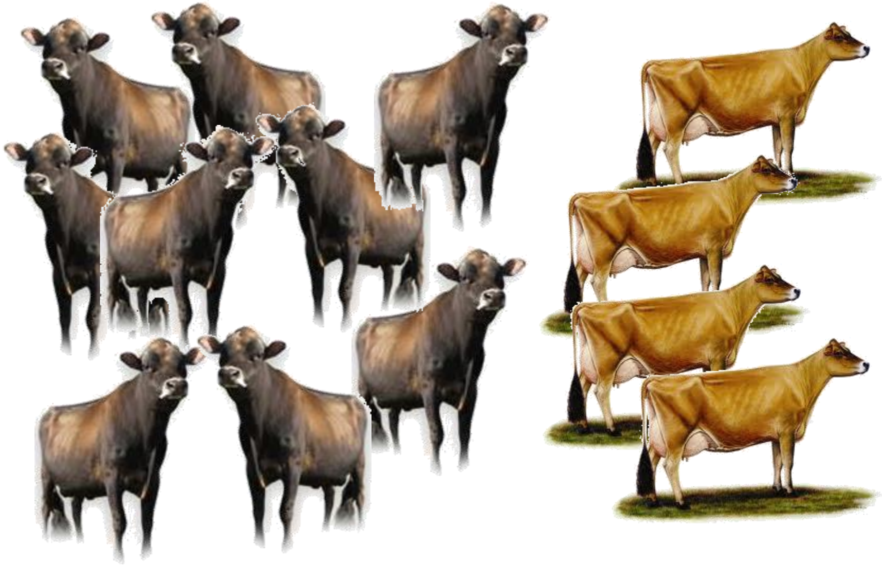
Analysis method?

- BayesR -> variants belong to one of 4 distributions, with zero, very small, small, medium variance
- For each SNP, posterior probability in each distribution





Holstein 4000 bulls, 10023 cows

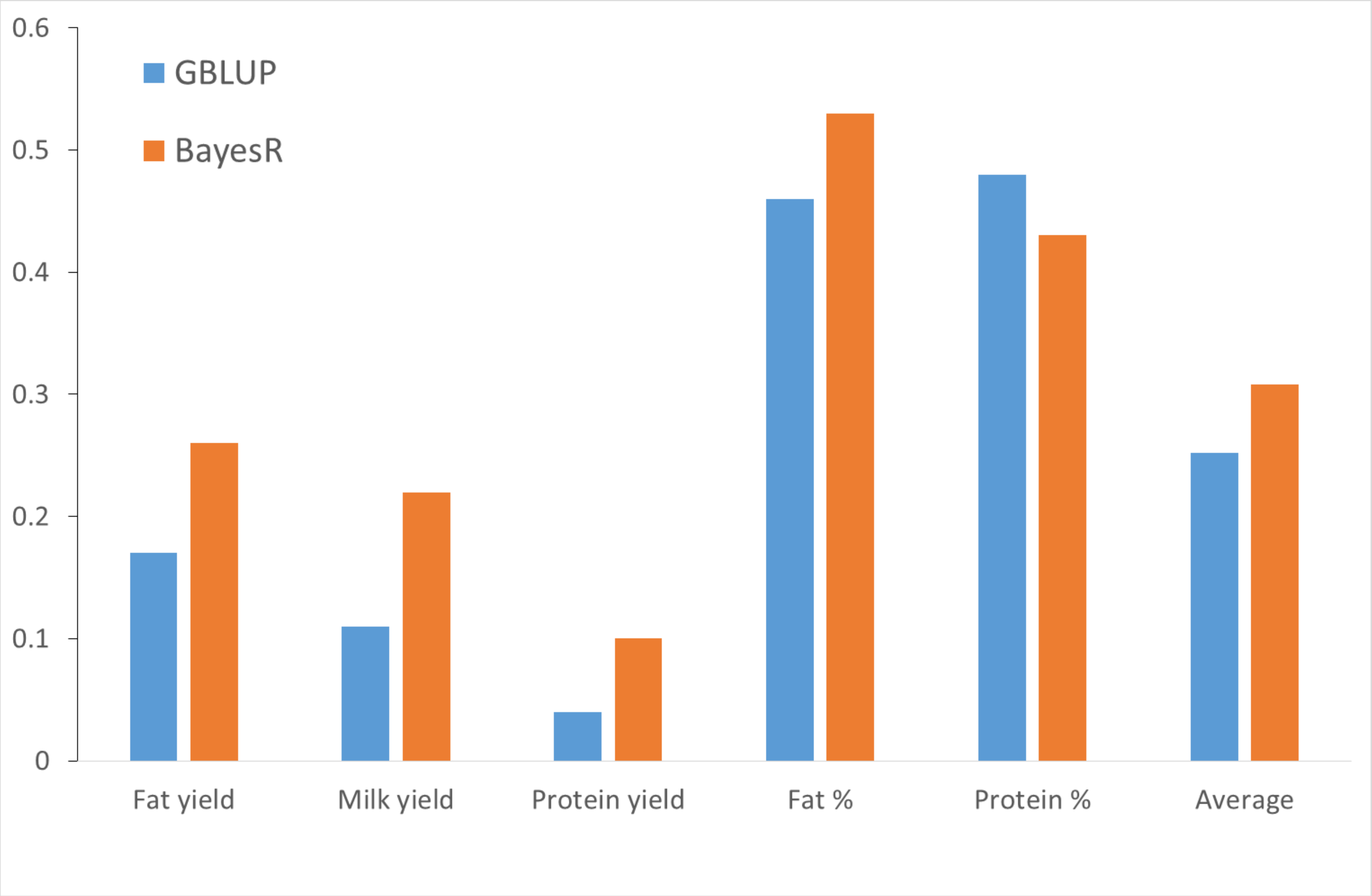


Jersey 1044 bulls, 4232 cows

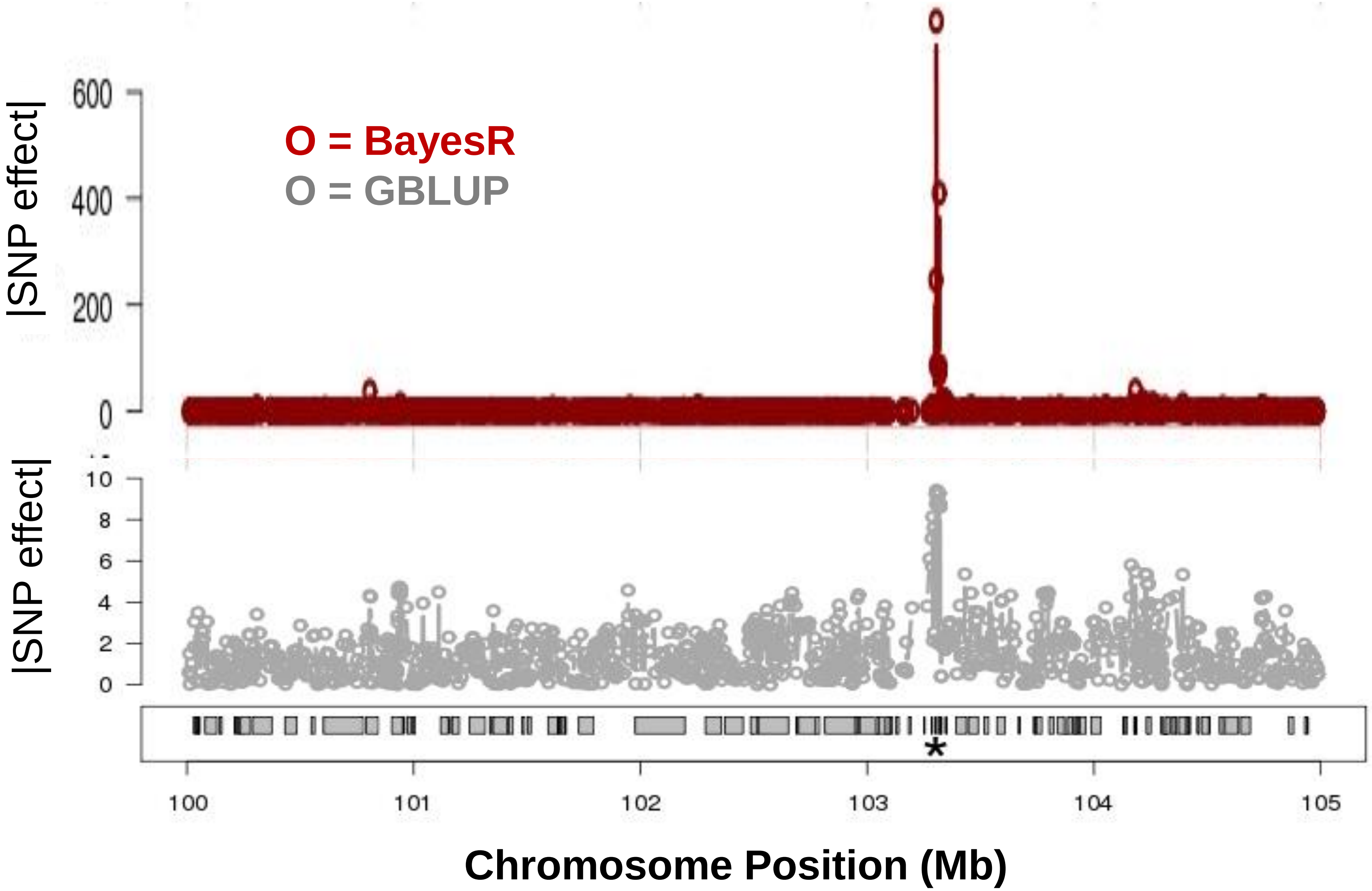
Aussie Reds 114 Bulls

Real or imputed 630K SNP for all individuals

Accuracy r(DGV,DTD) in Aussie Red Bulls



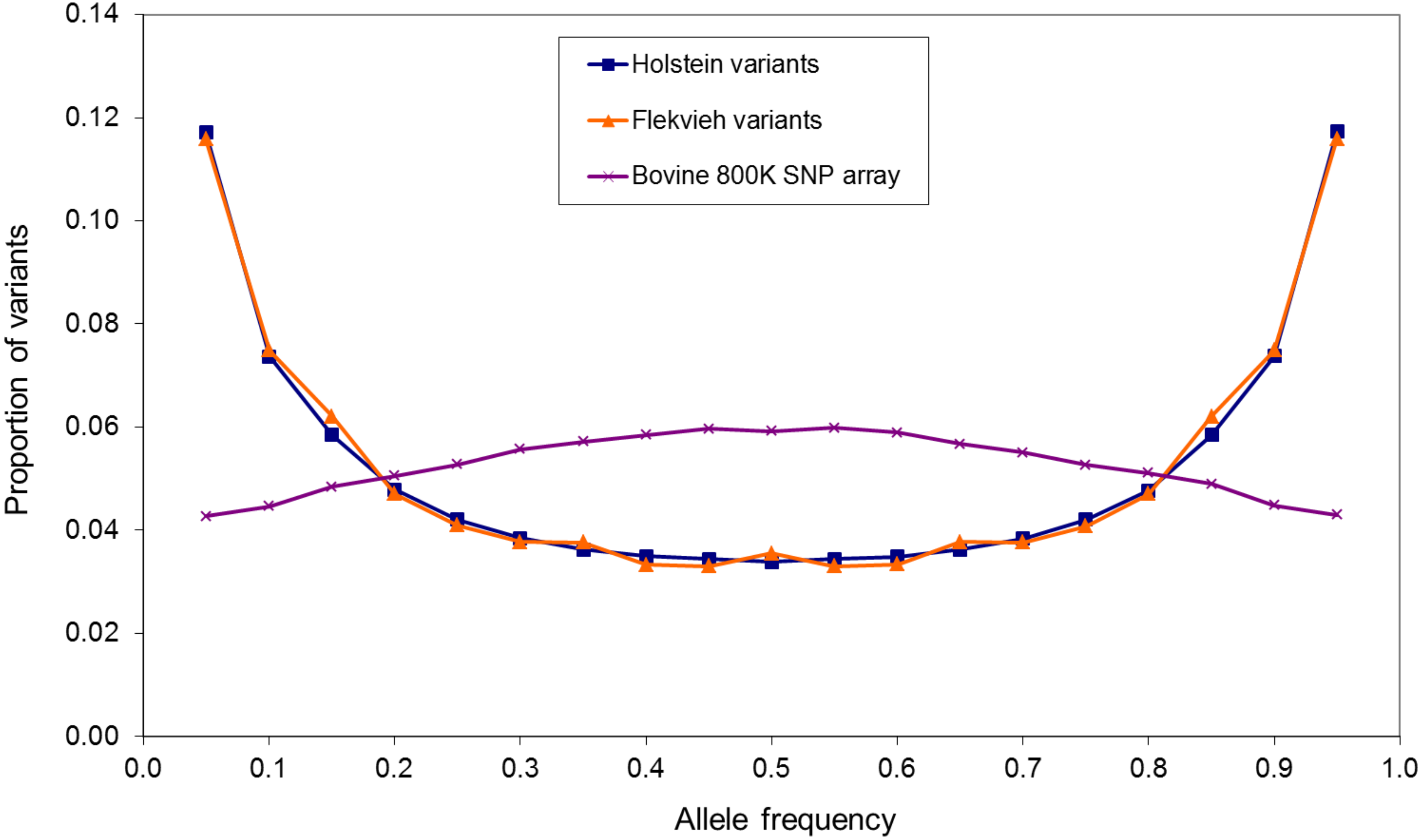
BayesR vs BLUP (BTA11)



HD SNP for multi-breed prediction

- Two issues

Allele frequency spectra Sequence vs HD



HD SNP for multi-breed prediction

- Allele frequency spectra mismatch
  - Lower frequency QTL not detected

- “Ghost QTL”
  - BayesR analysis, using a reference population of Holsteins and Jerseys, predicts some QTL contribute to GEBV in Jersey cattle, even though these QTL do not segregate in Jersey
- Why?
  - much more Holstein than Jersey data, information from Jersey overwhelmed.
  - With HD SNPs ‘correct’ solution not possible – (correct = causal variant segregating in Holstein, but not Jersey)
  - Instead BayesR may utilize number of SNPs to predict effect of QTL, this combination works in Holsteins, misleading in Jerseys

Why these limitations?

- QTL don’t segregate across breed?
- Need method other than BLUP, higher density genotypes?
- Sequence data to improve multi-breed predictions?

1000 Bull genomes project

- Sequencing more expensive than SNP chips
- 100,000s of animals genotyped with SNP chips

- Alternative strategy
  - *Sequence key ancestors and impute genotypes into all animals genotyped with SNP chips*

Source : CRV

1000 Bull Genomes Run 5

- 1682 Sequenced Animals (1577 Taurus, 115 Indicus), 11x
- ~55 Breeds: Dairy, Beef, Dual Purpose, Crosses, Composites

1000 Bull Genomes Run 5

- 1682 Sequenced Animals (1577 Taurus, 115 Indicus), 11x
- ~55 Breeds: Dairy, Beef, Dual Purpose, Crosses, Composites

Breed	**Taurus**		**Taurus**	**Breed**	**Indicus**
Holstein(BW,RW)	450	Beef Crosses	5	UgandaAdmixed	26
Simmental Fleckvieh Pezzata Rossa	274	Piedmontese	5	IranAdmixed	9
Angus (B,R)	157	Angler	5	Korean	9
Brown Swiss (incl. Braunvieh)	131	Hinterwalder	3	JapanNative	8
Jersey	66	Vorderwalder	3	Brahman	7
Beef Composites	64	Eringer	2	Gir	6
Danish Red	44	Galloway	2	Nelore	5
Gelbvieh	41	Unknown	2	Dehong	2
Hereford	41	Scottish Highland	2	Dengchuan	2
Charolais	39	Romagnola	2	Fujian	2
Limousin	31	Stabilizer	2	Guanling	2
Beef Booster	29	Tyrolean Grey	2	Hasake	2
Montbeliarde	28	Salers	1	Liping	2
Dairy-Beef Crosses	27			Luxi	2
AyrshireFinnish	25			Menggu	2
Normande	24			Nanyang	2
NorwegianRed	24			Qinchuan	2
Guernsey	20			Wenling	2
Swedish Red	16			Yanbian	2
Belgian Blue	10			Dabieshan	2
Marchigiana	8			Xizang	1

1000 Bull Genomes Run 5

Run 5 – Taurus only

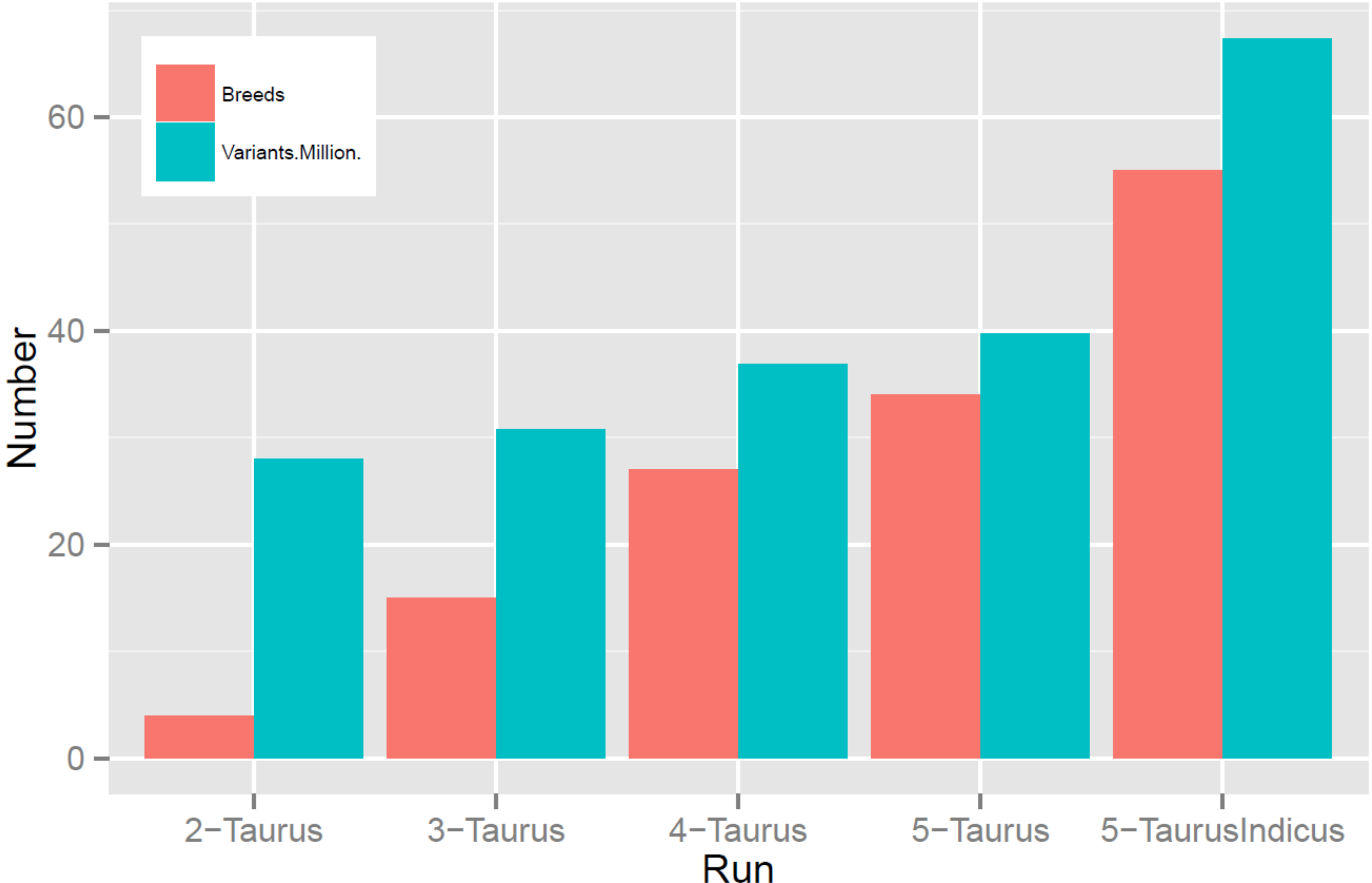
39.7 million filtered variants

38.1 million SNP, 1.7 million Indel

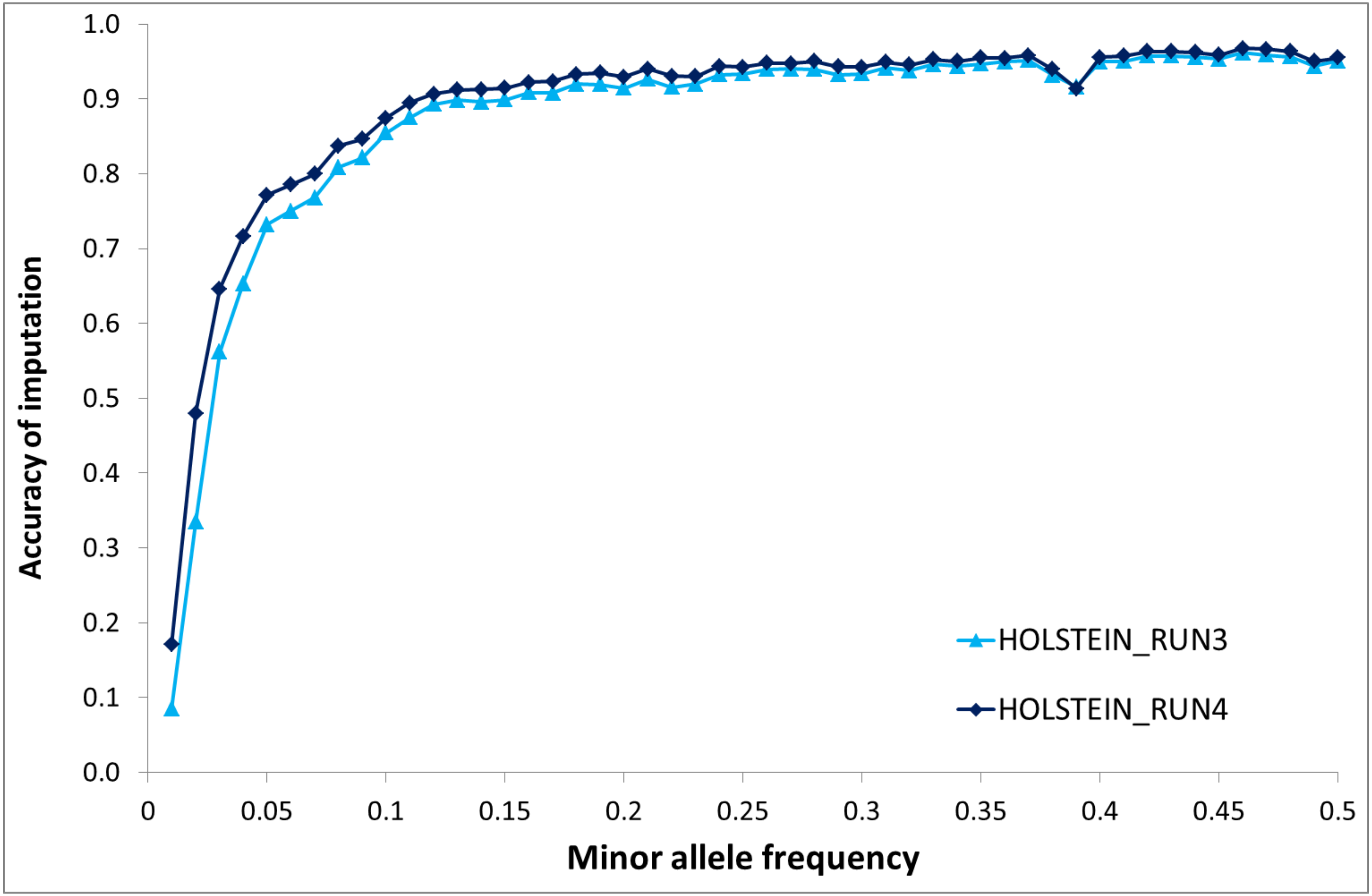
Run 5 – Taurus Indicus

67.3 million filtered variants

64.8 million SNP, 2.5 million Indel



Accuracy of imputation



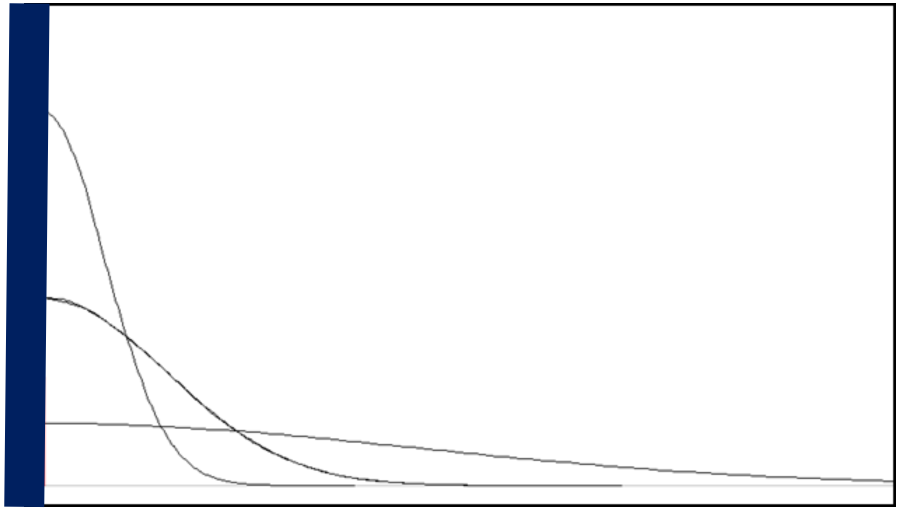
Genomic prediction with sequence

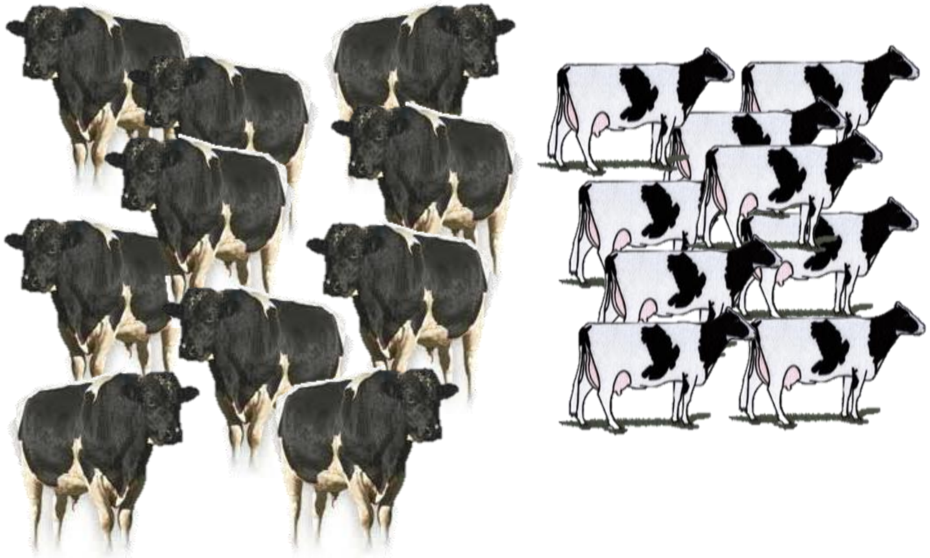
- Causative mutation is in the data set
- But 39.7 million effects to estimate, from same number of records
- Use biological information -> *are some classes of variant more likely to affect trait?*

Analysis method?

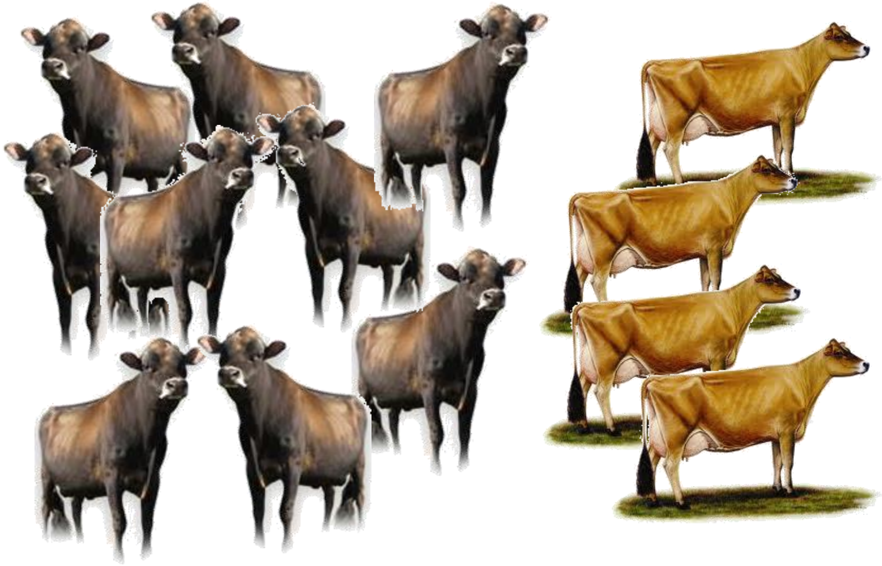
- Incorporate biological information?

- BayesRC allows different classes of variant
  - Allow different proportion of variants, in each of the four distributions, for each class
  - Do some classes have more variants of larger effect?
  - Does not give any class higher probability than any other *a priori*, just groups variants within a class





Holstein 4000 bulls, 10023 cows



Jersey 1044 bulls, 4232 cows

Aussie Reds 114 Bulls

Sequence genotypes all individuals (1.2m variants)

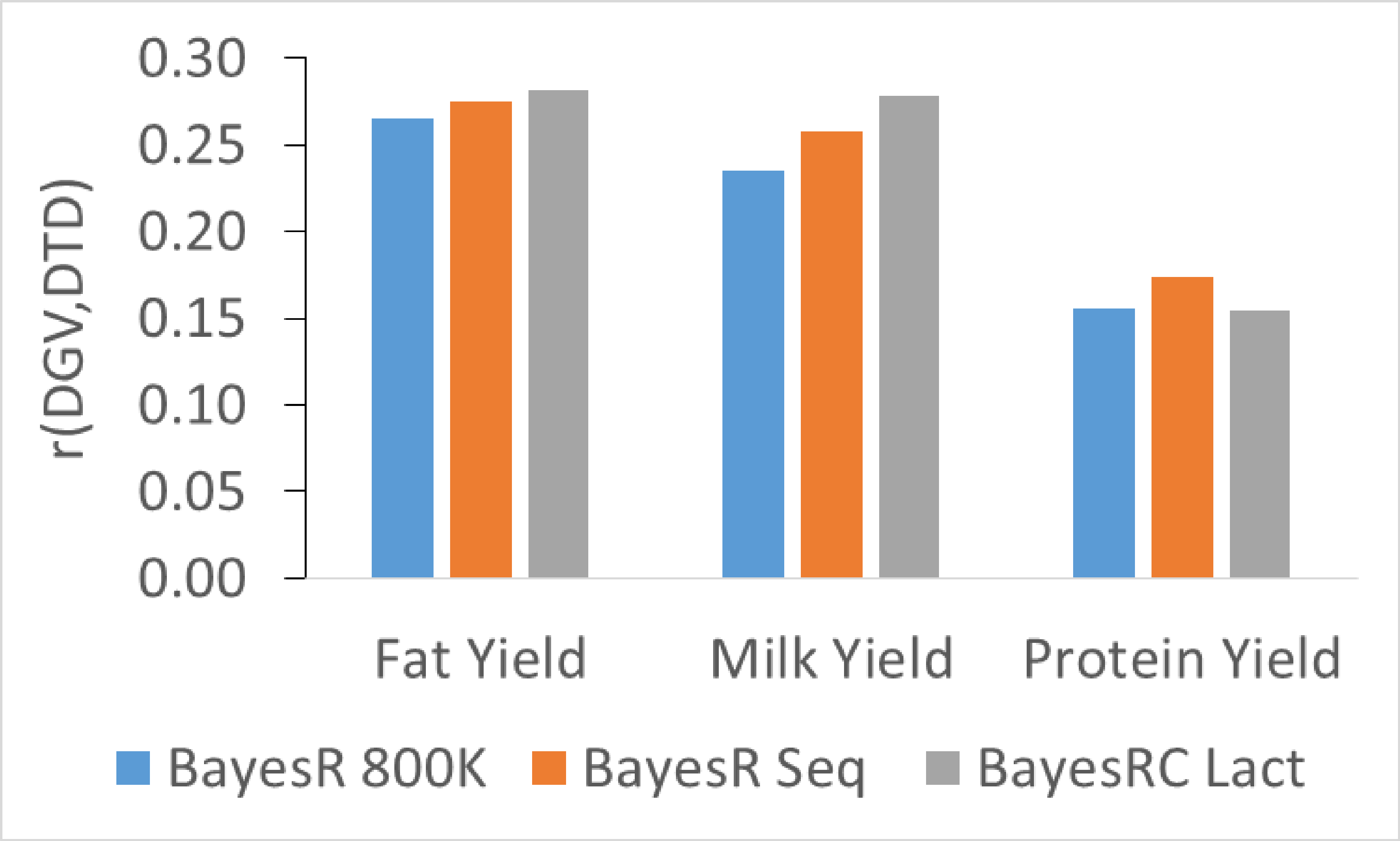
Analysis method?

- Milk production – BayesRC classes
  1. 792 lactation genes, non syn variants
  2. 792 lactation genes, other variants
  3. All other variants

- Compare r(DGV,DTD) in Aussie Red bulls
  - BayesR 800K
  - BayesR Seq (sequence variants)
  - BayesRC Lact (sequence variants, classes above)

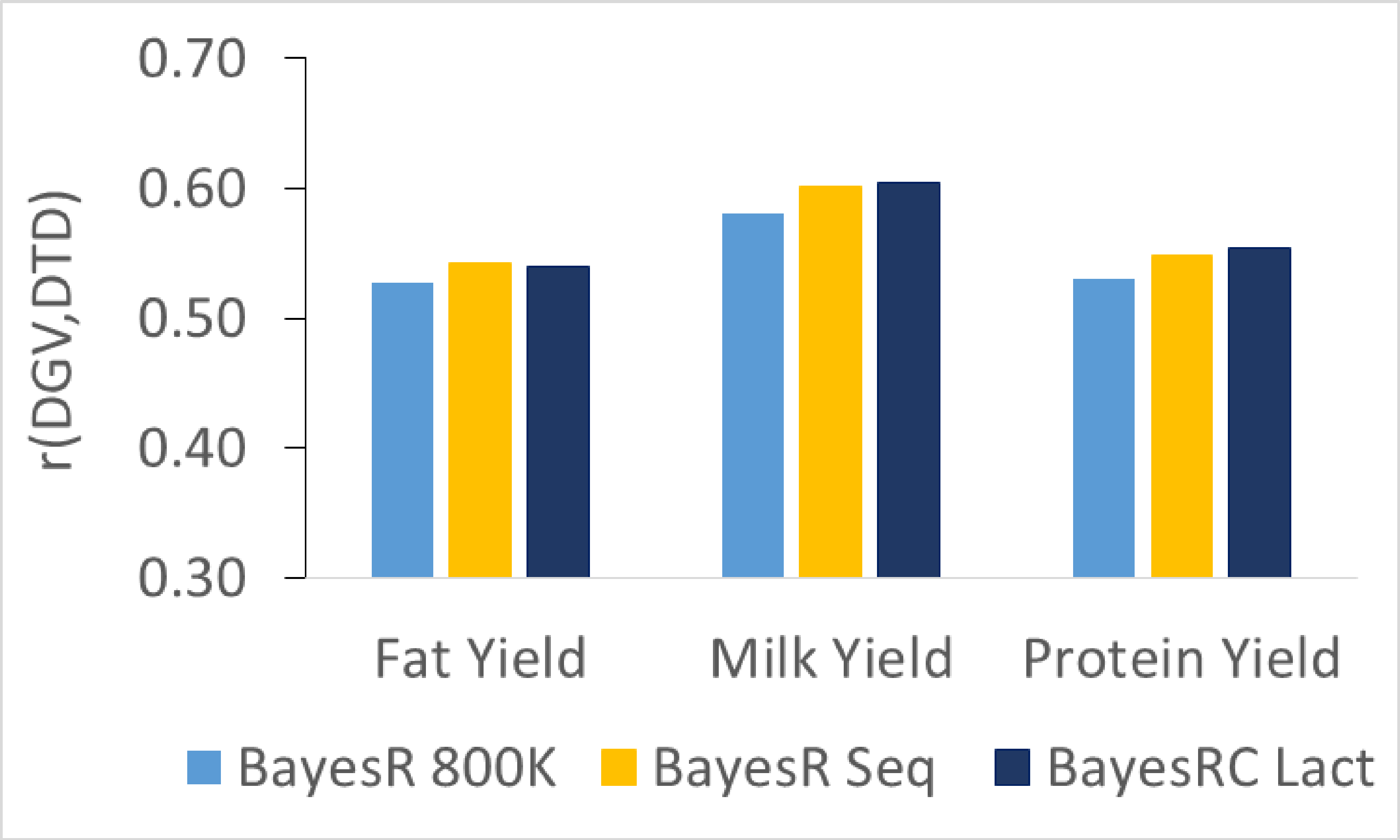
Genomic prediction – results

r(DGV,DTD) (AusBullCows -> **Aussie Reds**)



Genomic prediction – results

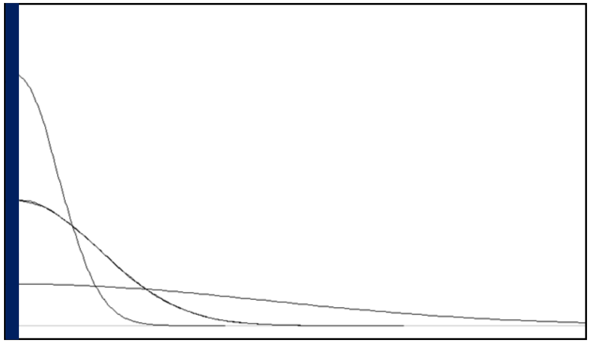
r(DGV,DTD) (AusBullCows -> **Red Holsteins**)



Genomic prediction (Protein)

- BayesR

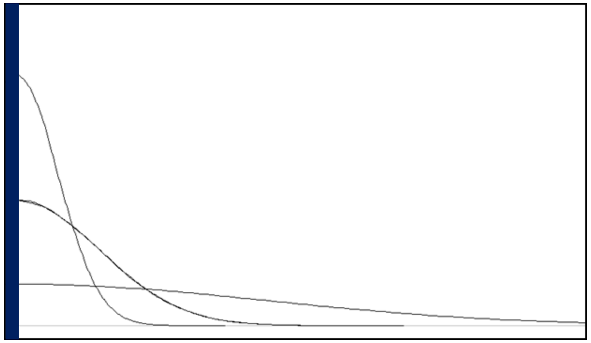
	0.00	0.0001	0.001	0.01
Total SNP	**1**	**2**	**3**	**4**
905,813	**99.3%**	**0.69%**	**0.004%**	**0.001%**



Genomic prediction (Protein)

- BayesR

	0.00	0.0001	0.001	0.01
Total SNP	**1**	**2**	**3**	**4**
905,813	**99.3%**	**0.69%**	**0.004%**	**0.001%**



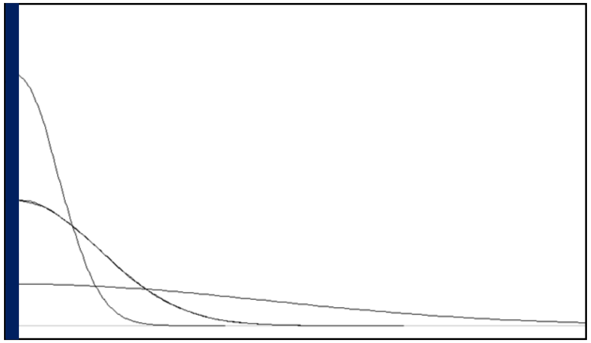
- BayesRC Lact

		0.00	0.0001	0.001	0.01
SNP Class	**No. SNP**	**1**	**2**	**3**	**4**
Diff Ex FUNC	3768 (0.4%)	**95.0%**	**4.3%**	**0.58%**	**0.12%**
Diff Ex other	57722 (6%)	**99.3%**	**0.7%**	**0.05%**	**0.004%**
All others	847905 (99%)	**99.5%**	**0.5%**	**0.01%**	**0.000%**

Genomic prediction (Protein)

- BayesR

	0.00	0.0001	0.001	0.01
Total SNP	**1**	**2**	**3**	**4**
905,813	**99.3%**	**0.69%**	**0.004%**	**0.001%**



- BayesRC Lact

		0.00	0.0001	0.001	0.01	
SNP Class	**No. SNP**	**1**	**2**	**3**	**4**	**Variance explained**
Diff Ex FUNC	3768 (0.4%)	**95.0%**	**4.3%**	**0.58%**	**0.12%**	**10.5%**
Diff Ex other	57722 (6%)	**99.3%**	**0.7%**	**0.05%**	**0.004%**	**12.4%**
All others	847905 (99%)	**99.5%**	**0.5%**	**0.01%**	**0.000%**	**77%**

Conclusions

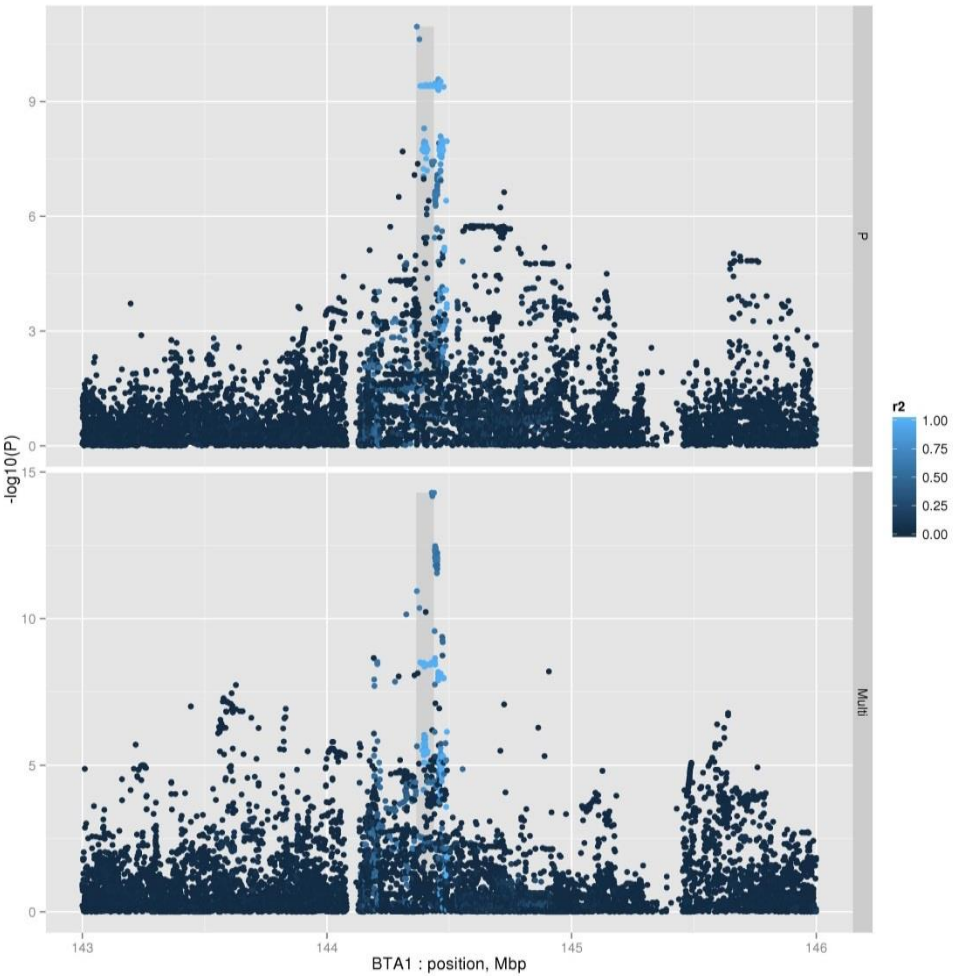
- QTL can and do segregate across breeds
  - Causal mutations for milk production between 2000 and 50,000 generations old
  - ~ 50% of QTL in common Holsteins and Jerseys
  - More breeds in reference -> more QTL in common
- 800K SNP improves across breed predictions, greatest advantage if Bayesian method used
  - Allele frequency mismatch
  - Ghost QTL
- Sequence data + biological information can improve accuracy of multi-breed genomic predictions

Conclusions

- Classes enriched for QTL in BayesRC
- Non-target phenotypes?

QTL for non-target phenotypes as classes

GWAS with imputed sequence



Explains:

10% of variance in phosphorous concentration

0.1% of variance in milk yield

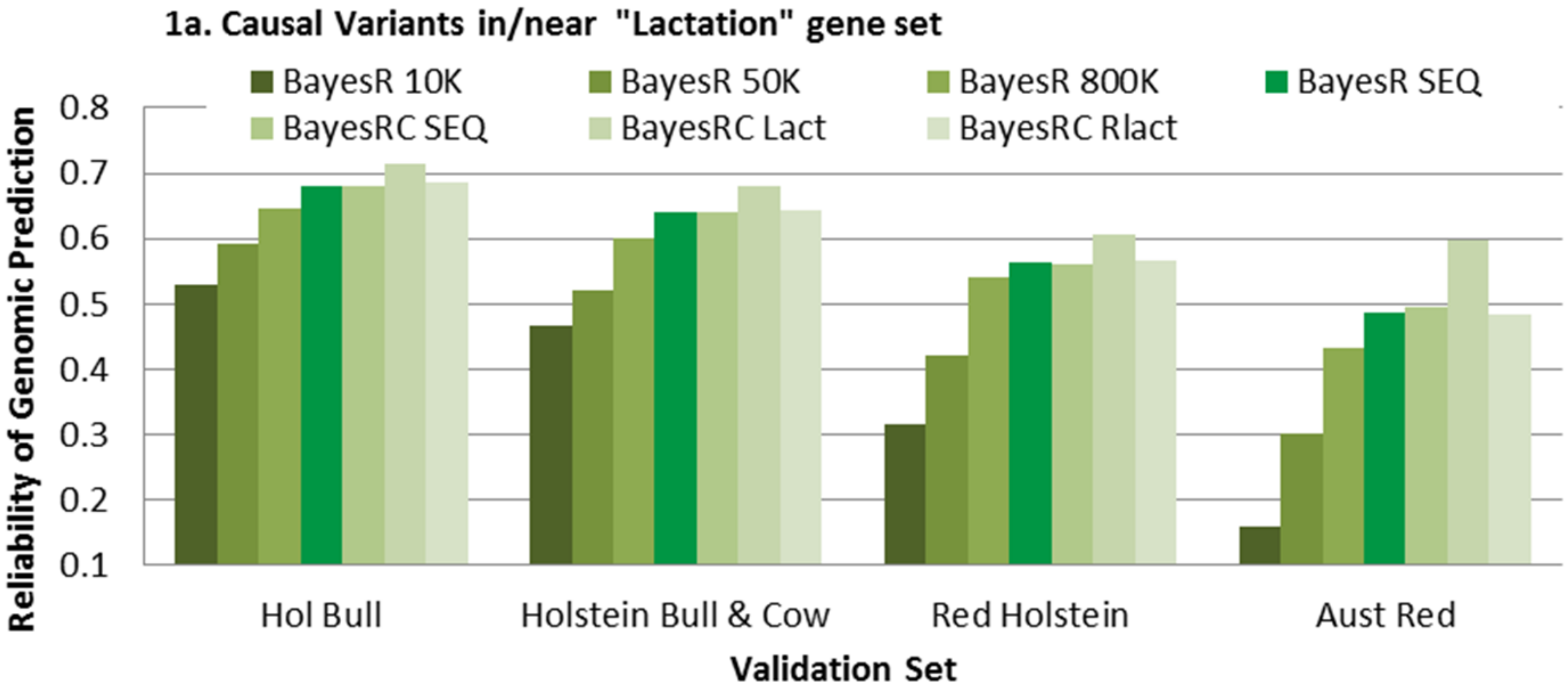
Conclusions

- Classes enriched for QTL in BayesRC

- Non-target phenotypes?

- Annotation of regulatory regions -> FAANG

Improving multi-breed predictions



With thanks

Kathryn Kemper, Iona Macleod, Chris Schrooten (CRV), Hans Daetwyler, Amanda Chamberlain, Mike Goddard (DEDJTR)

1000 bull genomes consortium

Ruedi Fries (Technische Universität München, Germany), Mogens Lund/Bernt Guldbrandtsen (Aarhus University, Denmark), Didier Boichard (INRA, France), Paul Stothard (University of Alberta, Genome Canada), Roel Veerkamp (Wageningen UR, Netherlands), Curt Van Tassell (United States Department of Agriculture), Tom Druet (University of Liege), Birgit Gredler (QualitasAg), Johanna Vilkki (MTT), Erik Mullaart (CRV), Alessandro Bagnato (UNIVERSITA' DEGLI STUDI DI MILANO), Donagh Berry (TEAGASC), DJ deKoning (SLU), Enrico Santus (ANARB), James Reecy (ISU), Jerry Taylor (UofM), Flavio Schenkel (UofGuelph), Cord Droegemueller (UofBern), Steve Miller (AgResearch), Dirk Hinrichs (University of Kiel), Beatriz (INIA), Eileen Wall (SRUC), Lorenzo Bomba (Università Cattolica del Sacro Cuore), Ezequiel (Fondazione Parco Tecnologico Padano), Luis Varona (Universidad de Zaragoza), Joanna Szyda (Wroclaw University of Environmental), Theo Meuwissen (Norwegian University of Life Sciences), Jesús Piedrafita (Universitat Autònoma de Barcelona), Christa Kuhn (Leibniz Institute for Farm Animal Biology)

Dairy Futures CRC

Limitations of 50K + BLUP

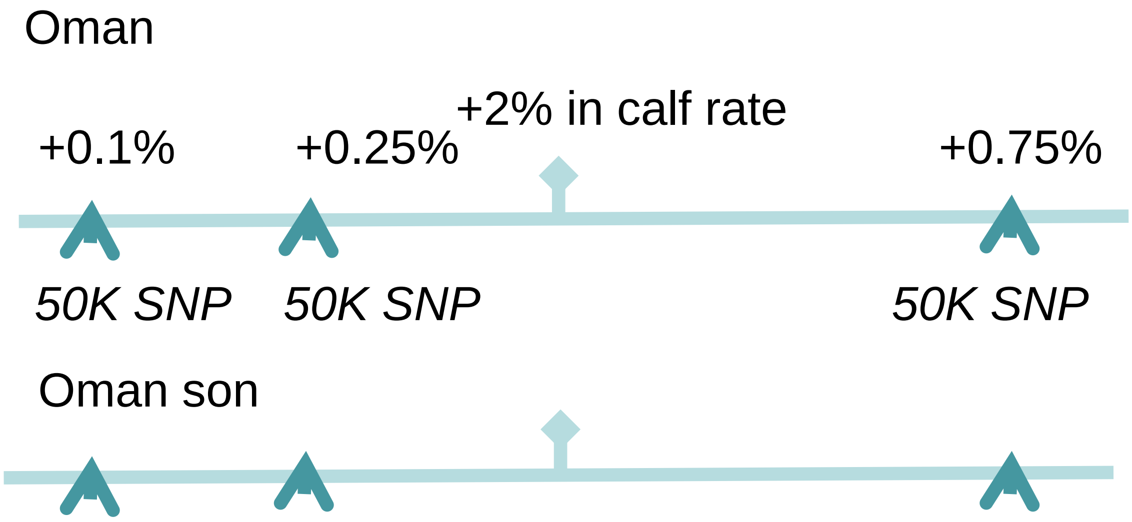
Oman

+2% in calf rate

+0.1% +0.25% +0.75%

50K SNP *50K SNP* *50K SNP*

Limitations of 50K + BLUP



Oman

+2% in calf rate

+0.1%

+0.25%

+0.75%

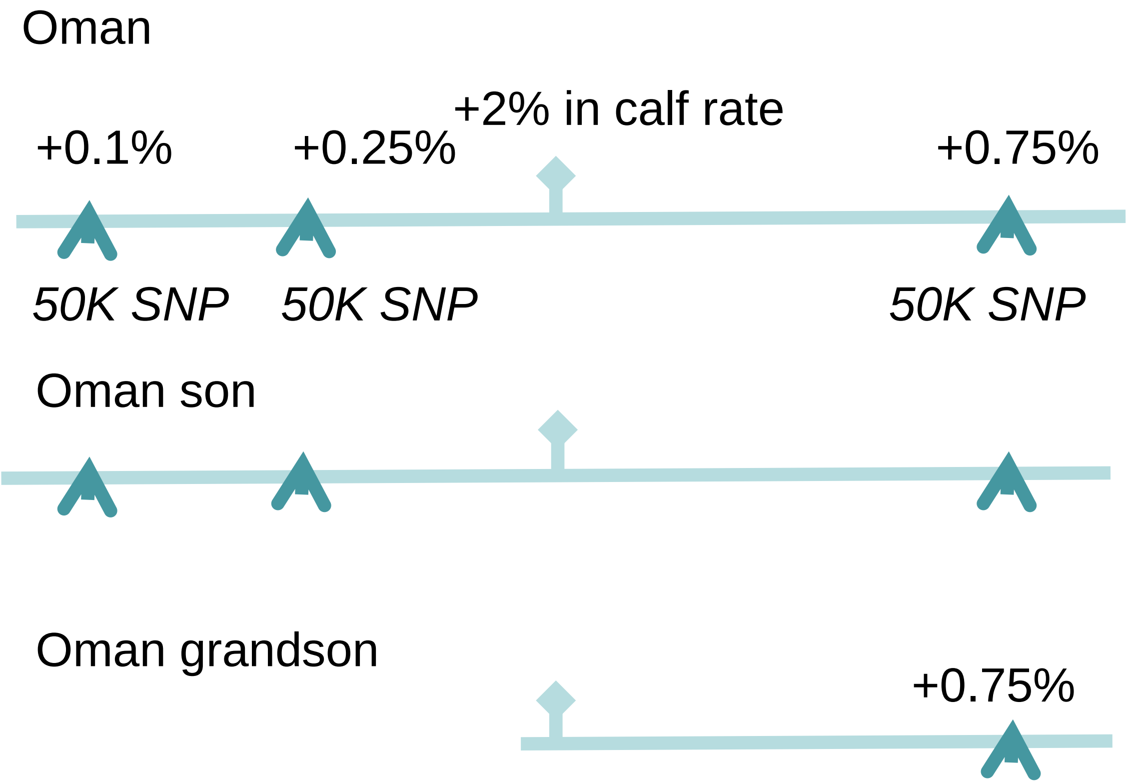
50K SNP

50K SNP

50K SNP

Oman son

Limitations of 50K + BLUP



Oman

+2% in calf rate

+0.1% +0.25% +0.75%

50K SNP *50K SNP* *50K SNP*

Oman son

Oman grandson

+0.75%

Limitations of 50K + BLUP