Using sequence data in genomic selection

Motivation/Opportunity

- Genome wide association study
  - Straight to causative mutation
  - Mapping recessives

- Genomic selection (all hypotheses!)
  - No longer have to rely on LD, causative mutation actually in data set
    - Higher accuracy of prediction?
  - Better prediction across breeds?
    - Assumes same QTL segregating in both breeds
    - No longer have to rely on SNP-QTL associations holding across breeds
  - Better persistence of accuracy across generations

Using sequence data in genomic selection

Motivation/Opportunity

- Genome wide association study
  - Straight to causative mutation
  - Mapping recessives

- Genomic selection (all hypotheses!)
  - No longer have to rely on LD, causative mutation actually in data set
    - Higher accuracy of prediction?
  - Better prediction across breeds?
    - Assumes same QTL segregating in both breeds
    - No longer have to rely on SNP-QTL associations holding across breeds
  - Better persistence of accuracy across generations

Using sequence data in genomic selection

Challenges

- Raw sequence information contains errors
  - Rate varies between technologies
- Reference genomes imperfect
  - Mapping of reads imperfect
- Costly
  - Numbers low, coverage low → power low?

Phantom Variants and Genotypes
Reduced Imputation Accuracy
Impose Upper Bound on Results

Benefits and challenges of using whole-genome sequence in genomic selection

- 36 million SNPs in cattle (1000 Bull Genomes Run4)

- Which method is most appropriate
- Priors
  - BLUP (GBLUP) -> all SNPs in LD with QTL, very small effects
  - BayesA -> some SNPs have moderate to large effects, rest very small
  - BayesB -> many SNPs have zero effect, some have small to moderate effect?

Methods for genomic prediction with full sequence

- Meuwissen and Goddard 2010
  - Simulated population with full sequence data, ~ 900 mutations chosen to be QTL
  - Used GBLUP and BayesB to predict GEBV

The accuracy of the predictions of total genetic value (±SE) in the TEST1 data set when the training data contained $T = 200$ individuals and GWBLUP or BayesB is used to estimate the marker effects

	Causative SNPs			
	GWBLUP		BayesB	
Data	Excluded	Included	Excluded	Included
3 QTL	0.503 ± 0.011	0.508 ± 0.011	0.938 ± 0.013	0.973 ± 0.004
30 QTL	0.491 ± 0.016	0.493 ± 0.010	0.806 ± 0.023	0.826 ± 0.019

Meuwissen, Goddard (2010) Genetics 185:623

Methods for genomic prediction with full sequence

- Meuwissen and Goddard 2010
  - Simulated population with full sequence data, ~ 900 mutations chosen as QTL
  - Used BLUP and BayesB to predict GEBV
  - Large advantage of BayesB over BLUP
    - Prior matches their simulated data -> only 900 QTL amongst millions of SNP
  - 3% advantage of having mutation in data
  - Real data??

Methods for genomic prediction with full sequence

- Meuwissen and Goddard 2010
  - Better persistence of accuracy over generations

Causal SNPs	TEST1: $T = 200$, $L = 1$: 30 QTL	TEST2: $T = 200$, $L = 1$: 30 QTL
Excluded	0.806 ± 0.023	0.806 ± 0.022
Included	0.826 ± 0.019	0.824 ± 0.019

Methods for genomic prediction with full sequence

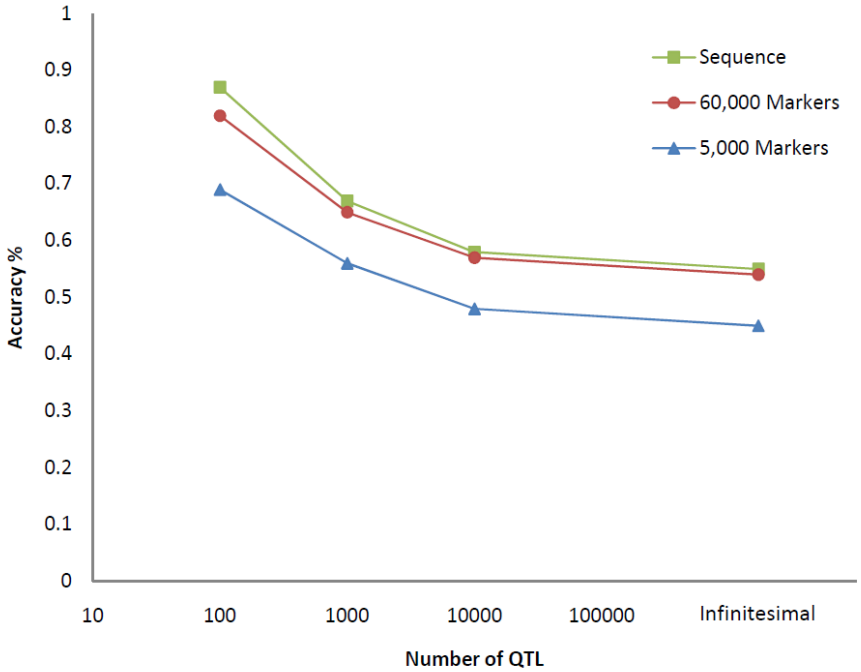


Table 4- Accuracy of the estimated breeding values (±SE) using SNP sequence data using two different methods and two alternative reference populations

			Method	
	No. QTL	Reference population	Bayes B	gBLUP
QTL	100	1	0.87 (0.009)	0.58 (0.014)
	1000	1	0.67 (0.012)	0.60 (0.017)
	10,000	1	0.58 (0.013)	0.58 (0.015)
	IM	1	0.54 (0.015)	0.55 (0.012)
QTL	100	2	0.81 (0.021)	0.39 (0.020)
	1000	2	0.53 (0.017)	0.35 (0.013)
	10,000	2	0.38 (0.012)	0.34 (0.015)
	IM	2	0.34 (0.012)	0.35 (0.017)

Clark et al, 2011. GSE

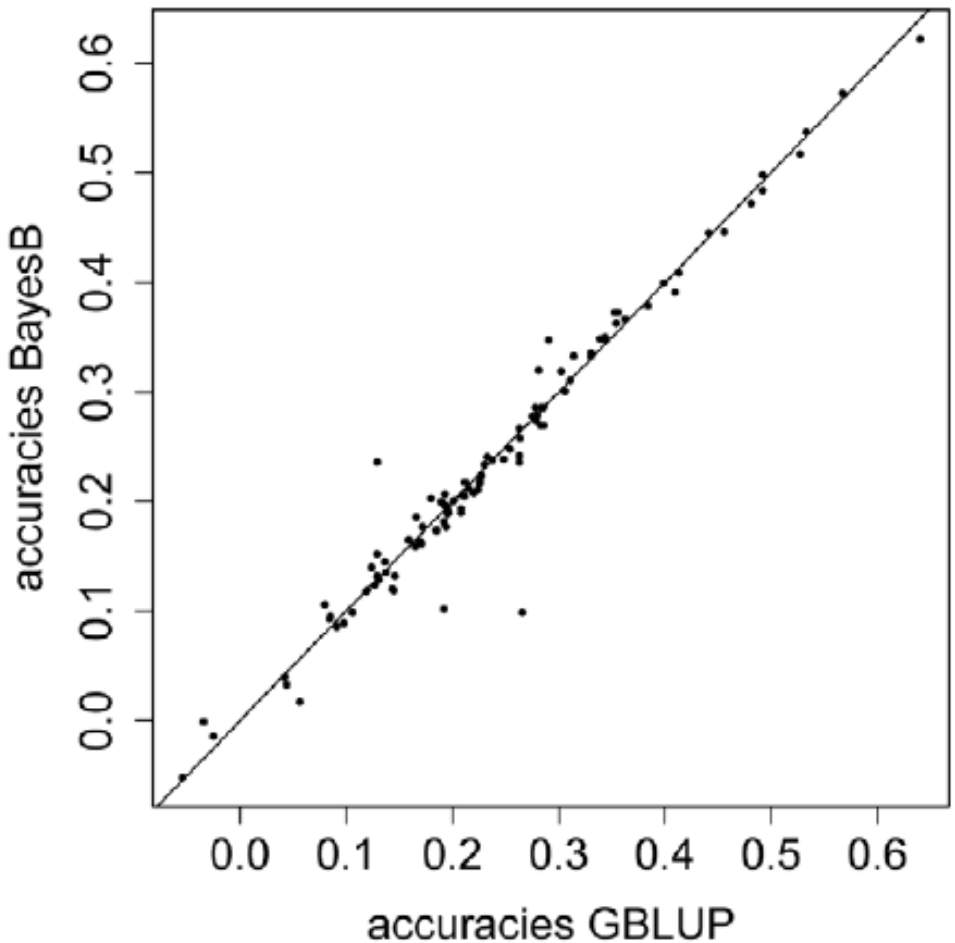
- Sequence slightly higher accuracy if number of QTL low
- Ne simulated at approximately 100, Me about 600

Methods for genomic prediction with full sequence

- Ober et al (2012) PLoS Genetics 8(5): e1002685

- Sample size
  - 157 fly lines

- No difference
  - GBLUP vs BayesB



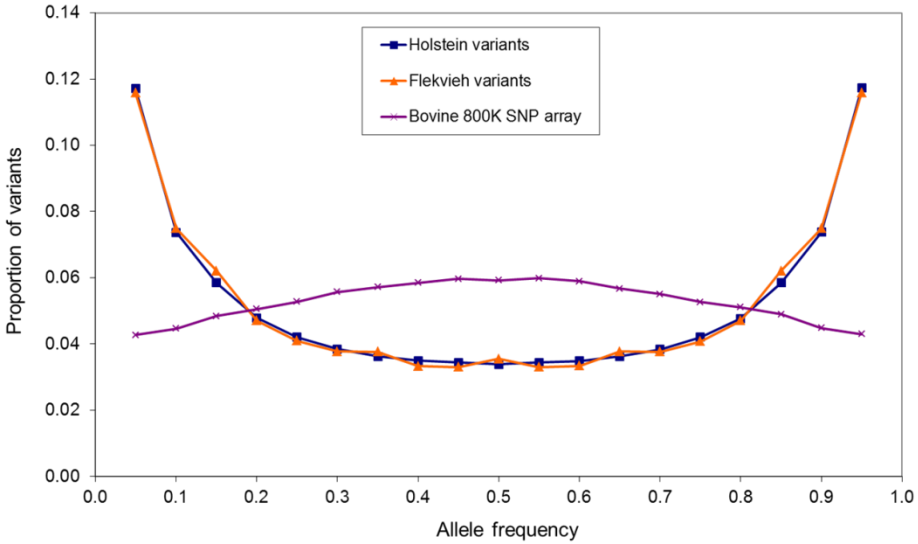
Will sequence data increase genomic prediction accuracy?

- Two different opinions:
  - Yes and No Camps

- Rationale of NO Camp
  - Why would sequence be different to HD chips?
  - Accuracy based overwhelmingly on close relationships
  - Sequence variants adds just noise and more data points in already long chromosome segments being estimated

Will sequence data increase genomic prediction accuracy?

- Rationale of YES Camp
  - Allele frequency spectrum of sequence different to HD chips
    - More low MAF variants



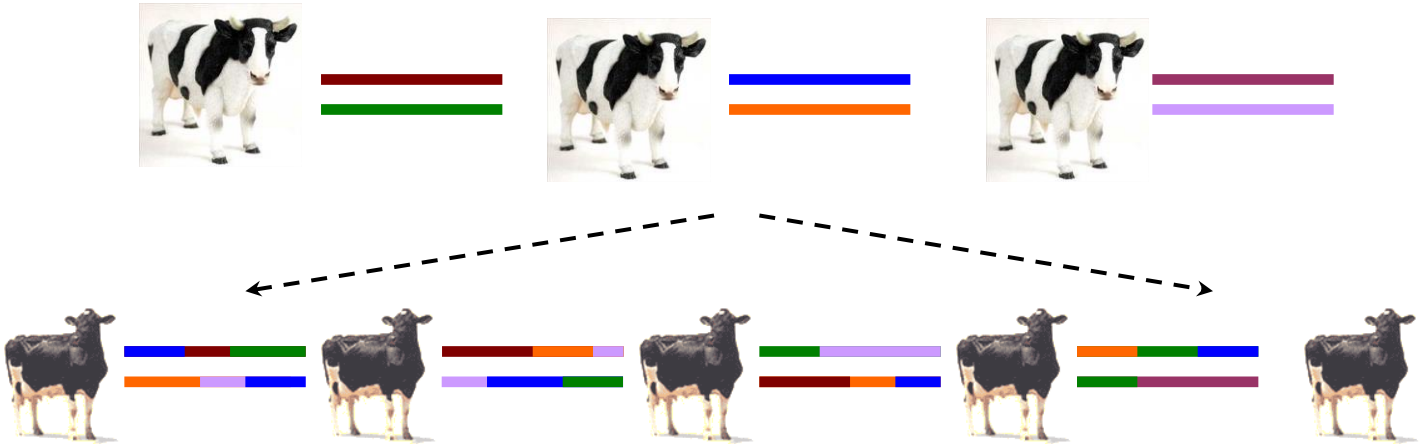
  - Causative mutations in data and higher overall LD
  - Need Bayesian methods and diverse populations to take advantage of much denser SNP

Will sequence data increase genomic prediction accuracy?

- Only a few ways that accuracy can be improved by sequence data
- GBLUP accuracy is roughly independent of number of QTL
- We have ‘approximately’ shown that Bayesian approaches have higher accuracy than GBLUP when number of QTL is lower than number of chromosome segments (M_e)

Number of independent chromosome segments Me

- Measure of population diversity
  - Depends on effective population size and genomes length
- Empirical estimates place it at ~ 1000 in Holstein
- What matters is Me in your breeding/reference population



Will sequence data increase genomic prediction accuracy?

- So, if the number of QTL is lower than Me, then BayesR accuracy of sequence data will be higher

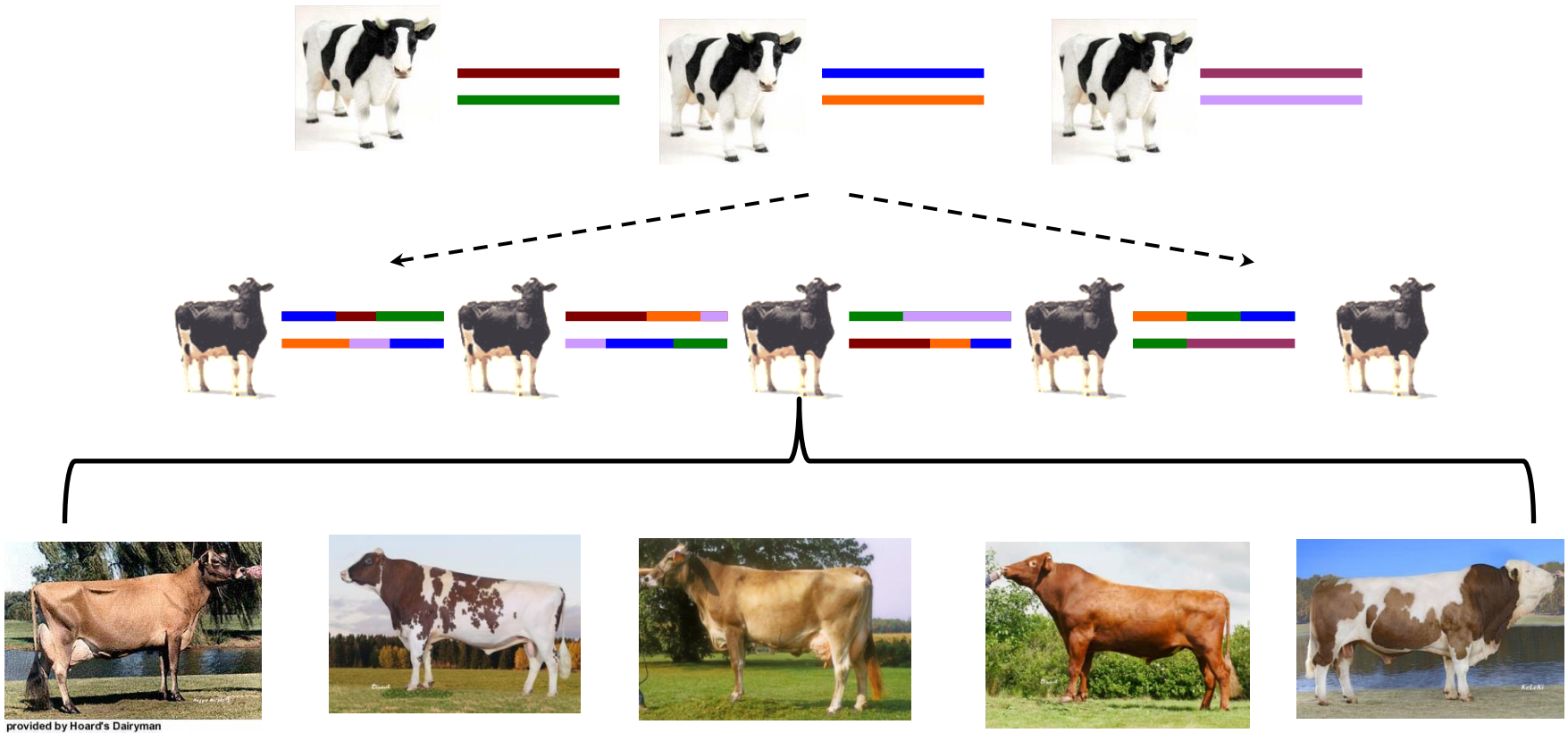
- In Holstein, Me is approximately 1000
  - Unlikely that we have <1000 QTL affecting most Holstein traits

Will sequence data increase genomic prediction accuracy?

- Can we increase Me?
  - Yes, e.g. multi-breed analysis (diverse populations)

Will sequence data increase genomic prediction accuracy?

- Can we increase Me?
  - Yes, e.g. multi-breed analysis (diverse populations)



Will sequence data increase genomic prediction accuracy?

- Accuracy of genomic selection increased if number of QTL *is less* than Me of combined/diverse reference population

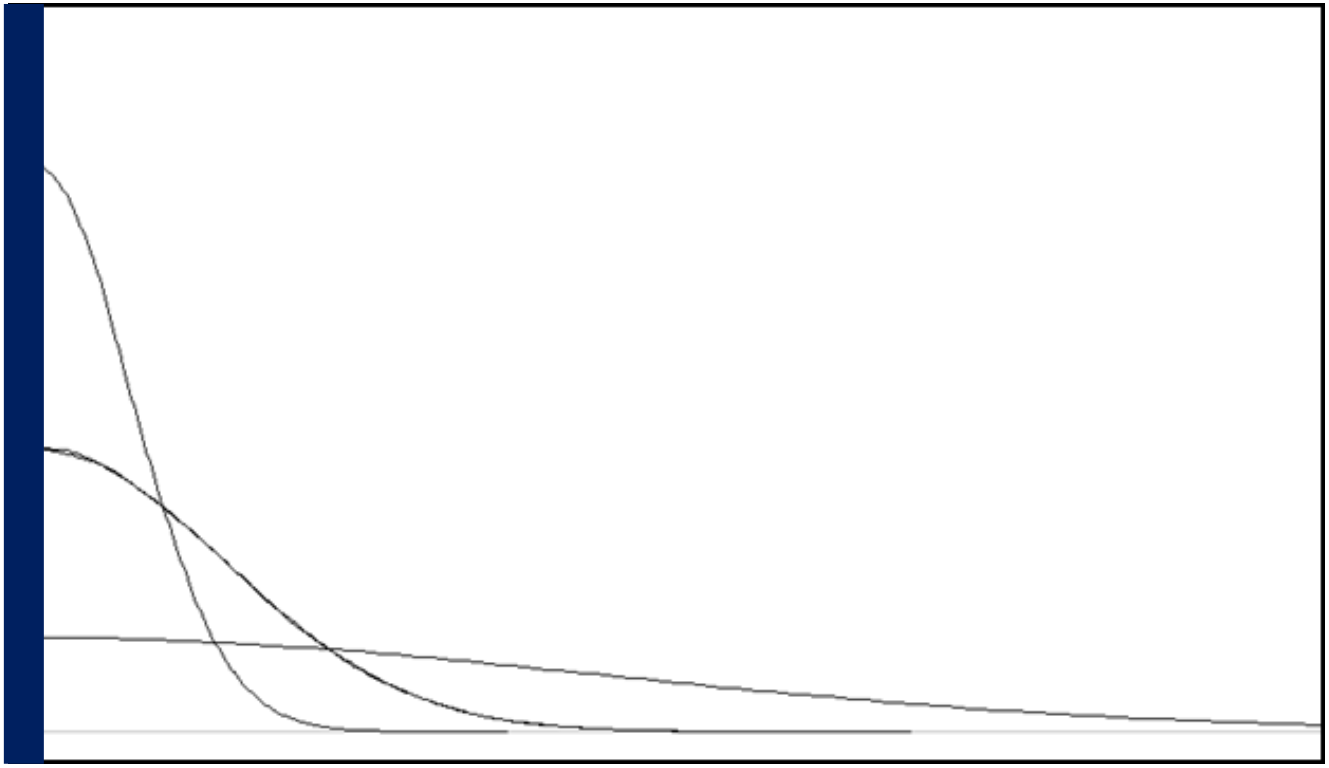
- But
  - need to also increase reference population size
  - There are many unknowns
    - E.g. how many QTL are shared between breeds?

Example: Genomic Prediction With Sequence in Dairy Cattle

Data Set	Breed	Reference	Validation
AusBullsCows	Holstein	11,527 (inc. 8478 cows)	
	Jersey	4687 (inc. 3917 cows)	
	Total	16,214	***114 "Aussie red" bulls (Scandinavian origin)***

Genomic Prediction With Sequence

- BayesR -> variants belong to one of 4 distributions, with zero, very small, small, medium variance
- Posterior proportion of variants in each distribution



Genomic Prediction With Sequence

- BayesR -> variants belong to one of 4 distributions, with zero, very small, small, medium variance
- Posterior proportion of variants in each distribution

- Biological information: *BayesRC* -> different classes of variant
  - allow different proportion of variants, in each distribution, for each class
- Do some classes have more variants of larger effect?

Genomic Prediction With Sequence

- 1 million Run3 variants in genes, +/- 2kb from genes
- Functional versus “regulatory” variants
- ***Seq_BayesRC*** classes:
  - 1. FUNC: Missense mutations
  - 2. REG: Upstream/downstream variants
  - 3. Rest

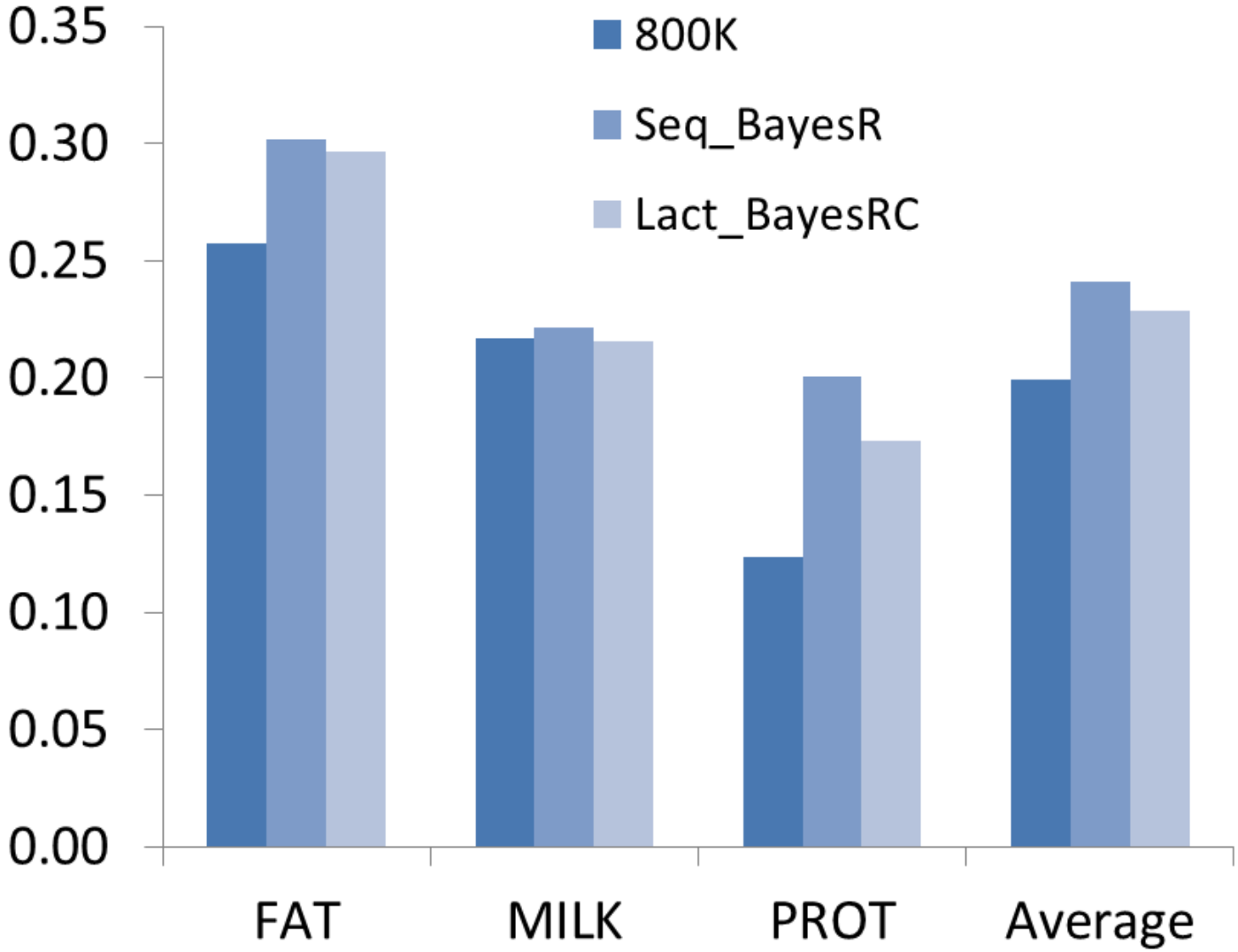
Genomic Prediction With Sequence

- 1 million Run3 variants in genes, +/- 2kb from genes
- Functional versus “regulatory” variants
- ***Seq_BayesRC*** classes:
  - 1. FUNC: Missense mutations
  - 2. REG: Upstream/downstream variants
  - 3. Rest

- Set of 792 genes that were differentially expressed under treatments leading to higher milk production
- ***Lact_FUNC_BayesRC*** classes:
  - 1. FUNC mutations in differentially expressed genes
  - 2. Other mutations in the differentially expressed genes
  - 3. Rest

Genomic Prediction With Sequence

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



Differences between whole-genome sequence and genotyping-by-sequencing

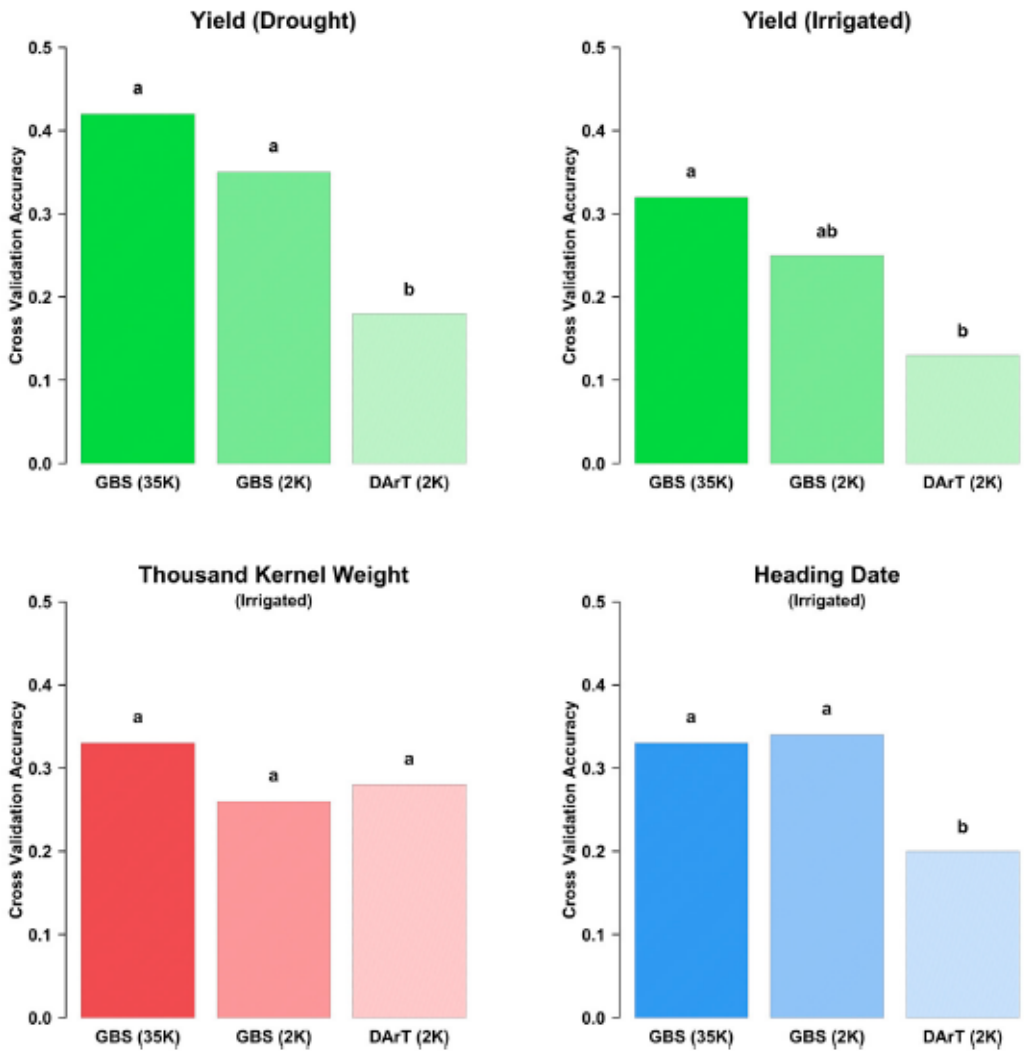
- Whole-genome
  - Usually higher coverage
  - No targeting of regions
  - Aim is a ‘complete’ inventory of variants in individual

- Genotyping-by-sequencing
  - Lower coverage
    - Highly multiplexed to reduce cost
  - Reduce genome space that is sequenced
    - Cut DNA with restriction enzymes
    - Some target specific genome regions
    - Allows for higher coverage in remaining regions
  - Some have high missing data
    - Rely heavily on imputation
  - Aim is to genotype more cheaply than with a SNP chip

Example: Genotyping-By-Sequencing (Genome Complexity Reduction) in Wheat

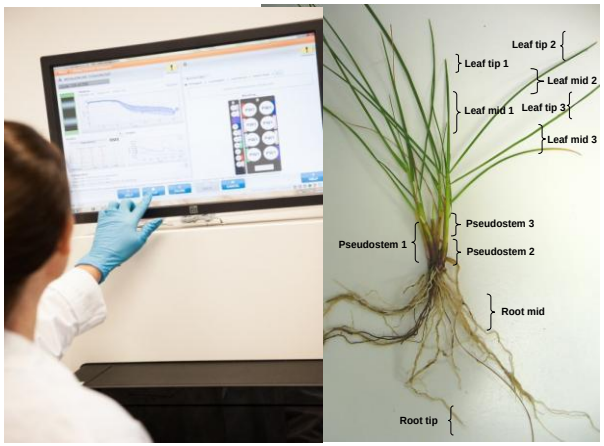
Poland et al, 2012, The Plant Genome

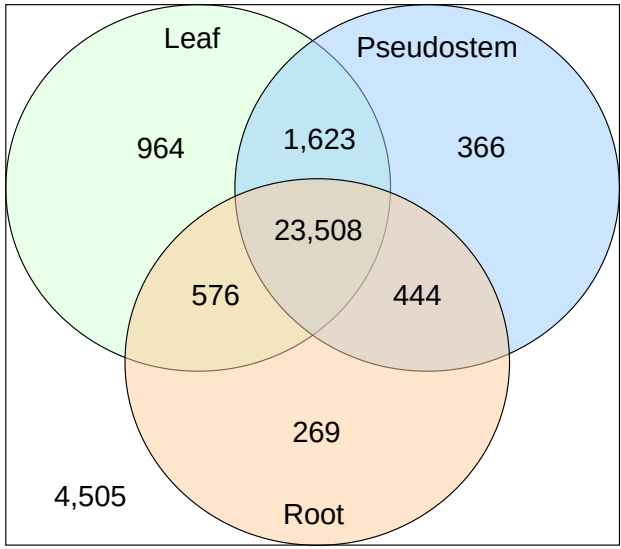
- Extended Elshire protocol (Elshire et al 2011, PloS ONE)
  - Cut up DNA with restriction enzyme
  - Sequence a subset of fragments
  - Impute missing using non-map methods
- 254 wheat breeding lines
  - Cross-validation accuracy
  - Compared accuracy to DArT marker



Example: Transcriptome-Based Genotyping-By-Sequencing in Ryegrass

- Newly developed protocol – low level transcript sequencing method
- Aligned to the Impact04 transcript atlas data
  - Based on RNA seq of 11 tissues
  - 85% of genes expressed in all three tissues
- All SNPs genic
  - No target enrichment required → reduced cost
  - May miss variation outside genes
- Generate c. 2 Million sequence reads per genotype
  - c. 1% output from 1 lane HiSeq2500
- Cost is 50$ per sample (HiSeq 2000)

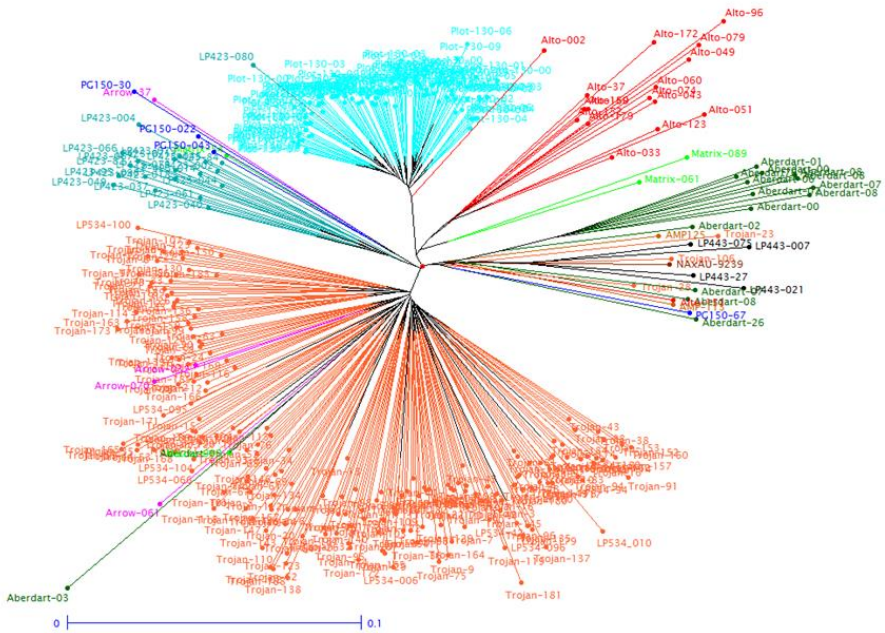


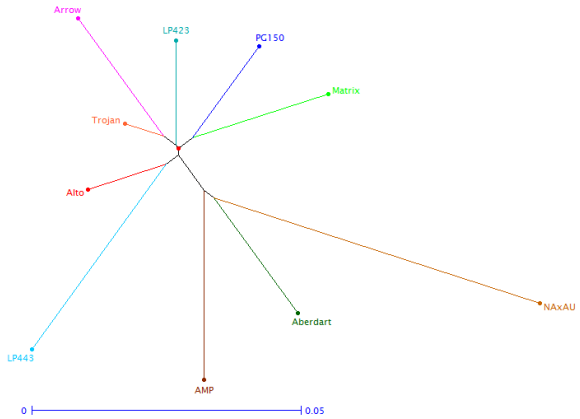


Transcript Analysis: Genotyping-By-Sequencing

- **449,713** SNPs from 85 individuals
- Validated in **288** samples
  - Including an F_1 mapping population
- Assessed segregation ratios in the mapping population
  - **139,772** high quality SNPs

- Population analysis confirmed known relationships and population structure





Within Cultivar Genomic Prediction of Agronomic Traits in Ryegrass

- Within cultivar Alto flowering time (H^2= 0.85) and biomass yield (H^2= 0.43)
- c. 140 individuals
- 9000 GBS SNP
- 5x fold cross-validation GS
- gBLUP and Bayesian methods (some with dominance fitted)

Flowering time

Accuracy	gBLUP			BayesianRR		Bayesian LASSO	
		Imputed	FT loci fitted as fixed effect		Dominance		Dominance
Mean	**0.641**	**0.649**	**0.690**	**0.659**	**0.658**	**0.644**	**0.660**
SE	0.031	0.033	0.036	0.031	0.029	0.034	0.028

Biomass Yield

Accuracy	gBLUP		BayesianRR		Bayesian LASSO	
		Imputed		Dominance		Dominance
Mean	**0.383**	**0.411**	**0.403**	**0.459**	**0.424**	**0.482**
SE	0.060	0.064	0.068	0.064	0.061	0.071