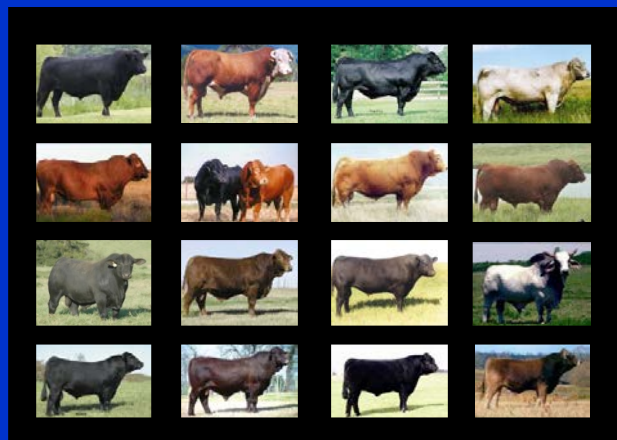
Training, Validation, and Target Populations

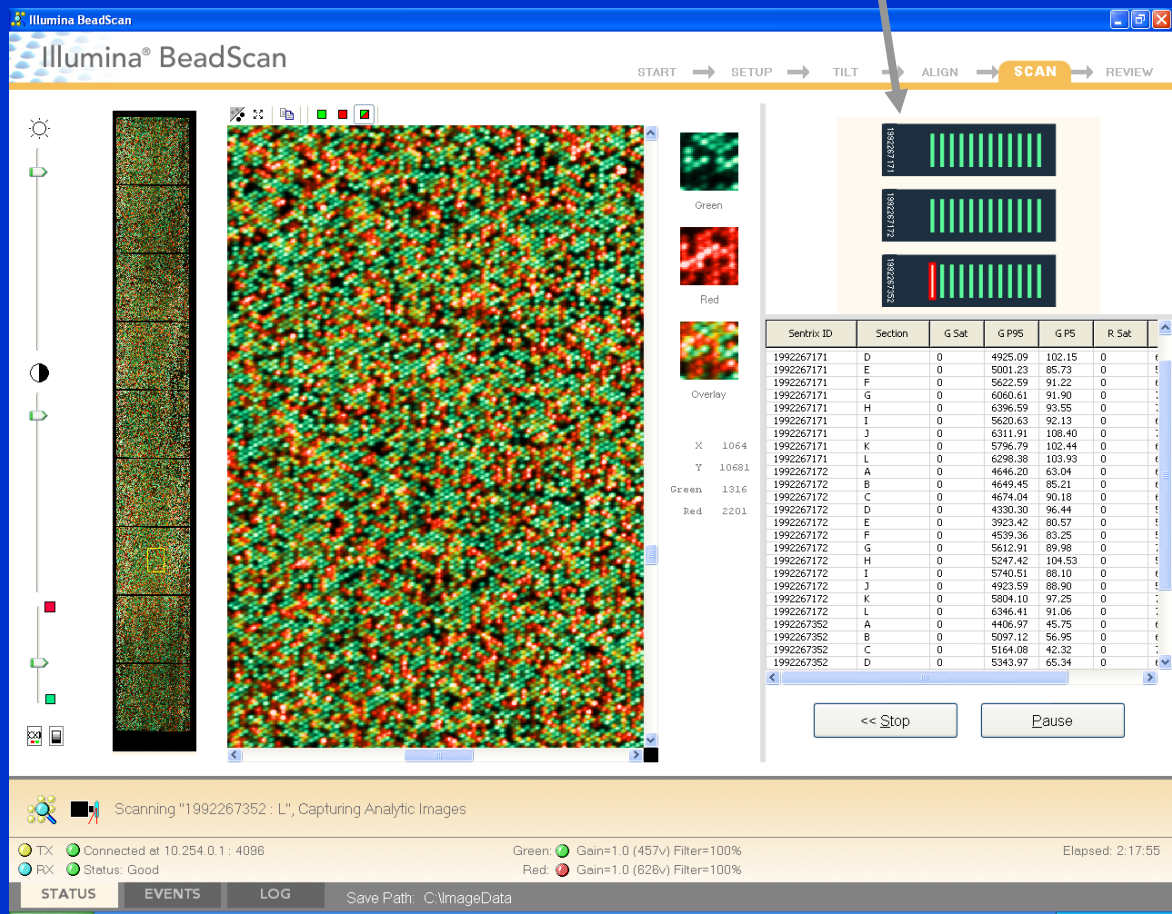
Mark Thallman, Kristina Weber, Larry Kuehn, Warren Snelling, John Keele, Gary Bennett, and John Pollak



50,000 Markers on a Chip (50K Chip)

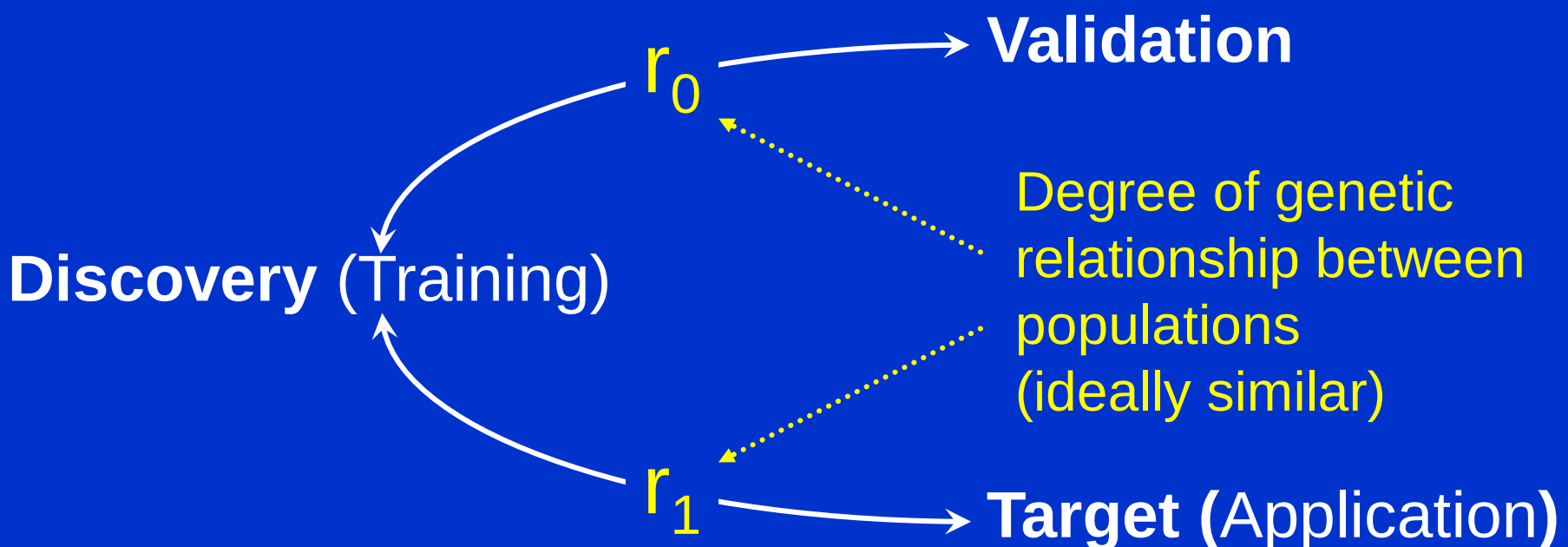
Illumina Infinium Bovine BeadChip

- ~ 50,000 SNP markers across the bovine genome
- -High resolution (1 SNP per 60,000 base pairs)
- - Multiple breeds used for SNP discovery



BARC (ARS)
USMARC
University of Missouri
University of Alberta

Populations Involved in Genomic Predictions of Economically Important Traits



3 Fundamental Types of Discovery Populations

- Purebreds of a Single Breed
- Purebreds of Multiple Breeds
- Crossbreds

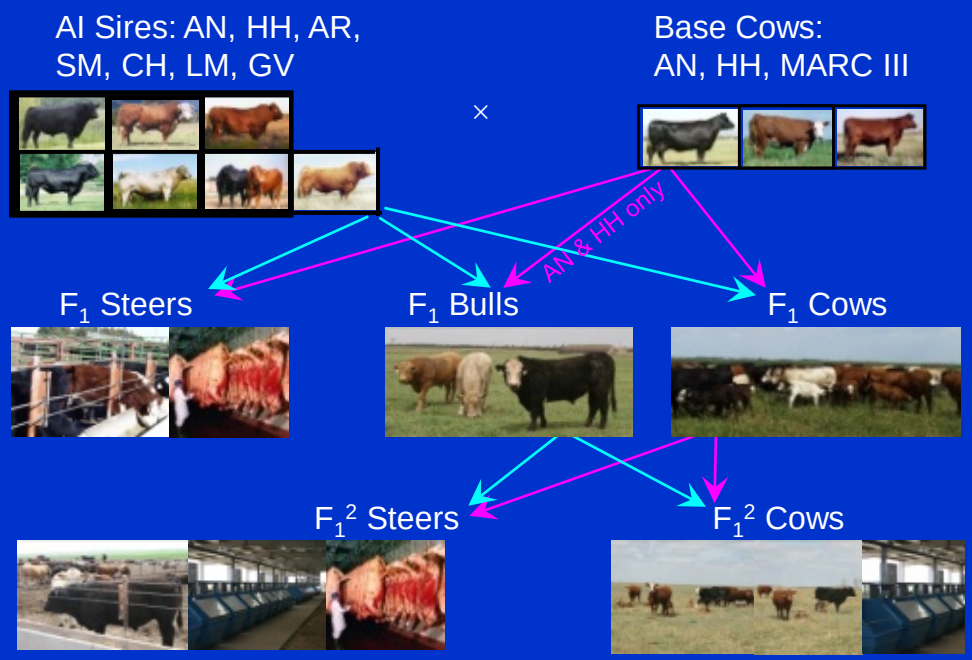
2 Fundamental Types of Discovery Data

- AI Sires with High Accuracy EPDs

- Individuals with Own Phenotypes

Two Resource Populations at USMARC

USMARC Cycle VII
USMARC Ongoing GPE



2,000 Bull Project

2000 Bull Project

- Collaborative Effort
  - Researchers
  - Breed Associations
- Breed associations provided semen for DNA on influential sires
- USMARC ran the 50K SNP chip on those 2,000 sires
- USMARC provides extensively phenotyped animals for use as training data set

2,000 Bull Project: Number of Sires Sampled

- Angus 402
- Hereford 317
- Simmental 253
- Red Angus 173
- Gelbvieh 136
- Limousin 131
- Charolais 125
- Shorthorn 86

- Brangus 68
- Beefmaster 64
- Maine-Anjou 59
- Brahman 53
- Chiangus 47
- Santa Gertrudis 43
- Salers 42
- Braunvieh 27

2026

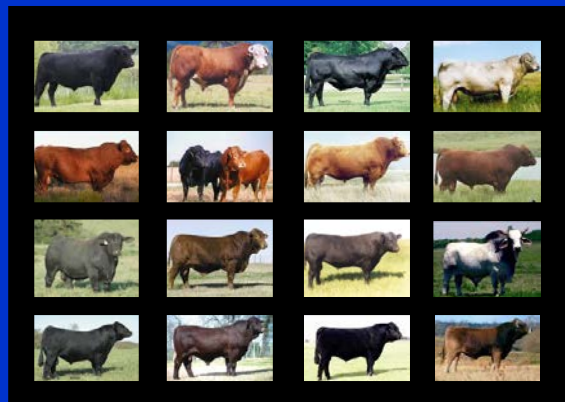
Deregression of EPDs

An approach used to scale EPDs so that high and low accuracy animals can be included in the same analysis

- Genetic variances are the same regardless of accuracy
- Residual variances are heterogeneous depending on the accuracy of the animal

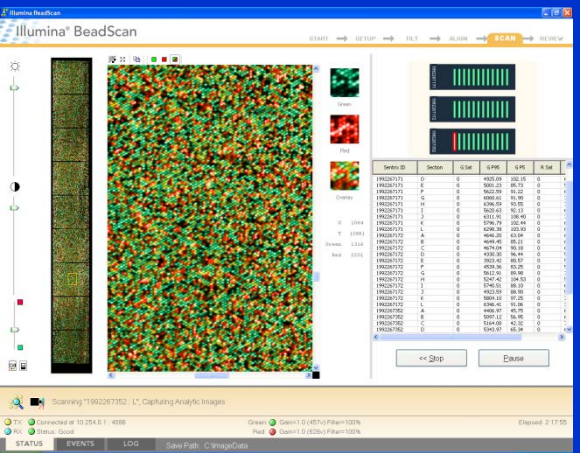
- Allows use of EPDs as phenotypes in genomic analyses

- AI Sires with High Accuracy EPDs



Goals of 2,000 Bull Project

- Demonstrate feasibility and understand challenges of applying whole-genome selection in beef cattle.
- Provide prediction equations for general use
- Provide genomic predictions for the bulls in the project



Training Data:
GPE Cycle VII Population

AI Sires: AN, HH, AR,
SM, CH, LM, GV

×

Base Cows:
AN, HH, MARC III

AN & HH only

F_1 Steers

F_1 Bulls

×

F_1 Cows

F_1^2 Steers

F_1^2 Cows

Training Data:
GPE Ongoing Continuous Sampling

AI Sires:
AN, HH, SM, CH, AR, LM, GV, SH,
BN, BM, MA, BR, CI, SG, SA, BV

×

Dams:
AN, HH, CH, SM,
MARC III, Cycle VII F_1

AN, HH, CH, SM

F_1 & BC Steers

F_1 Bulls

×

F_1& BC Heifers

Cross-validation

Training → Validation

USMARC Cycle VII
USMARC Ongoing GPE

AI Sires: AN, HH, AR, SM, CH, LM, GV

×

Base Cows: AN, HH, MARC III

AN & HH only

F_1 Steers

F_1 Bulls

F_1 Cows

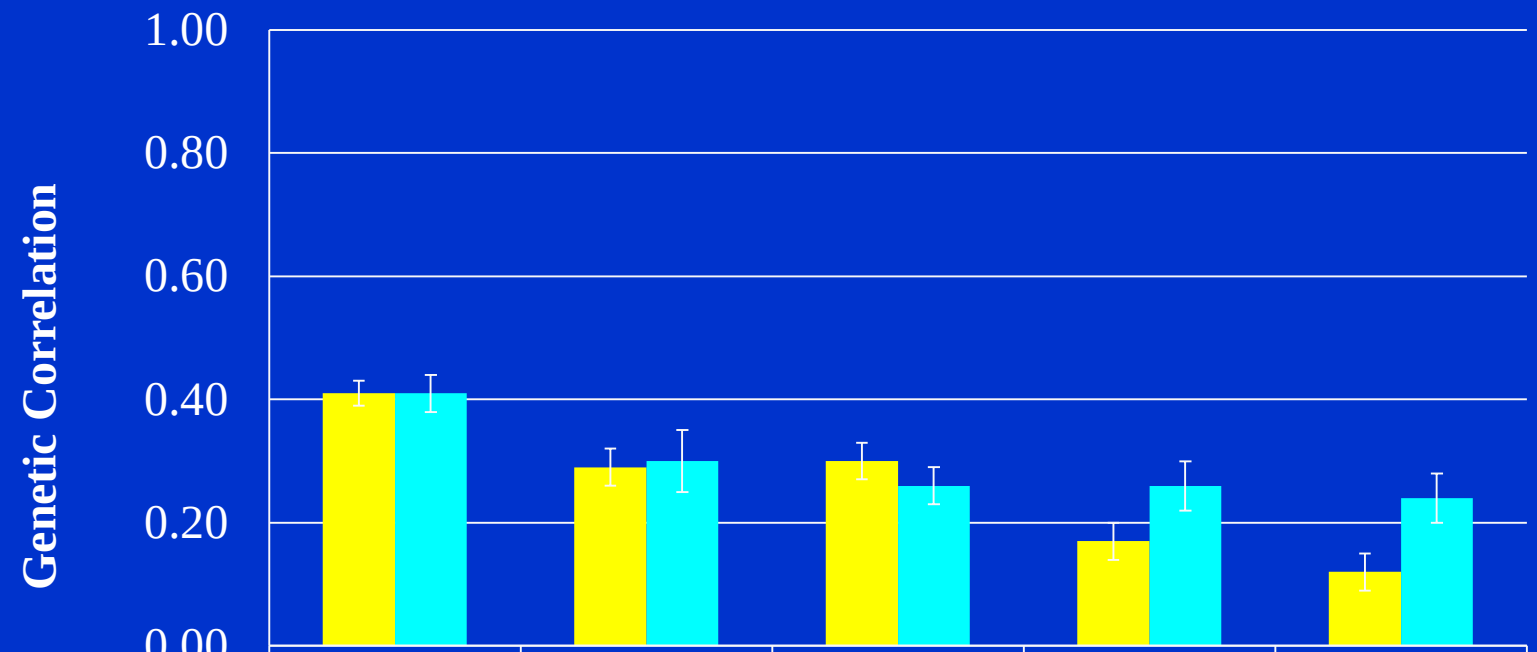
F_1^2 Steers

F_1^2 Cows

2,000 Bull Project

Validation ← Training

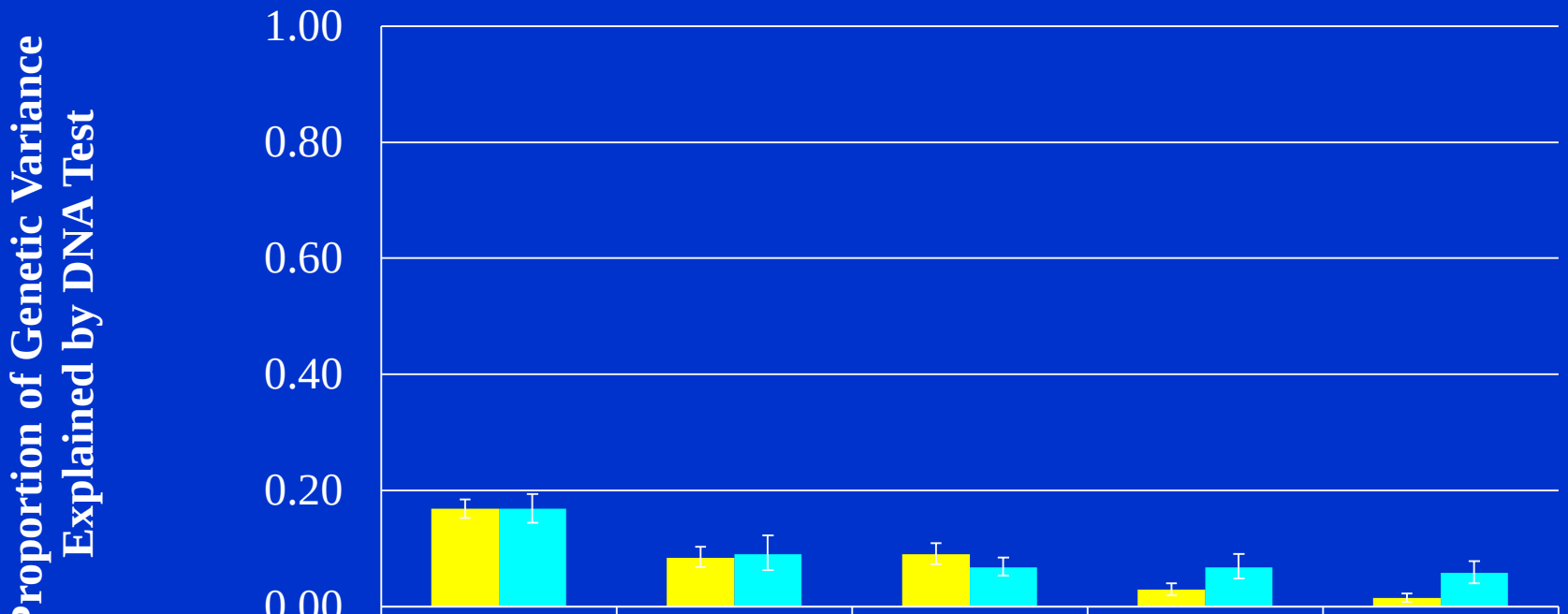
Genetic Correlations in Cross-validation



	BWT	WWT	YWT	REA	MRB
MARC-trained	0.41	0.29	0.30	0.17	0.12
2000 Bull-trained	0.41	0.30	0.26	0.26	0.24

*2,000 bull predictions excluded the sires of the MARC validation populations

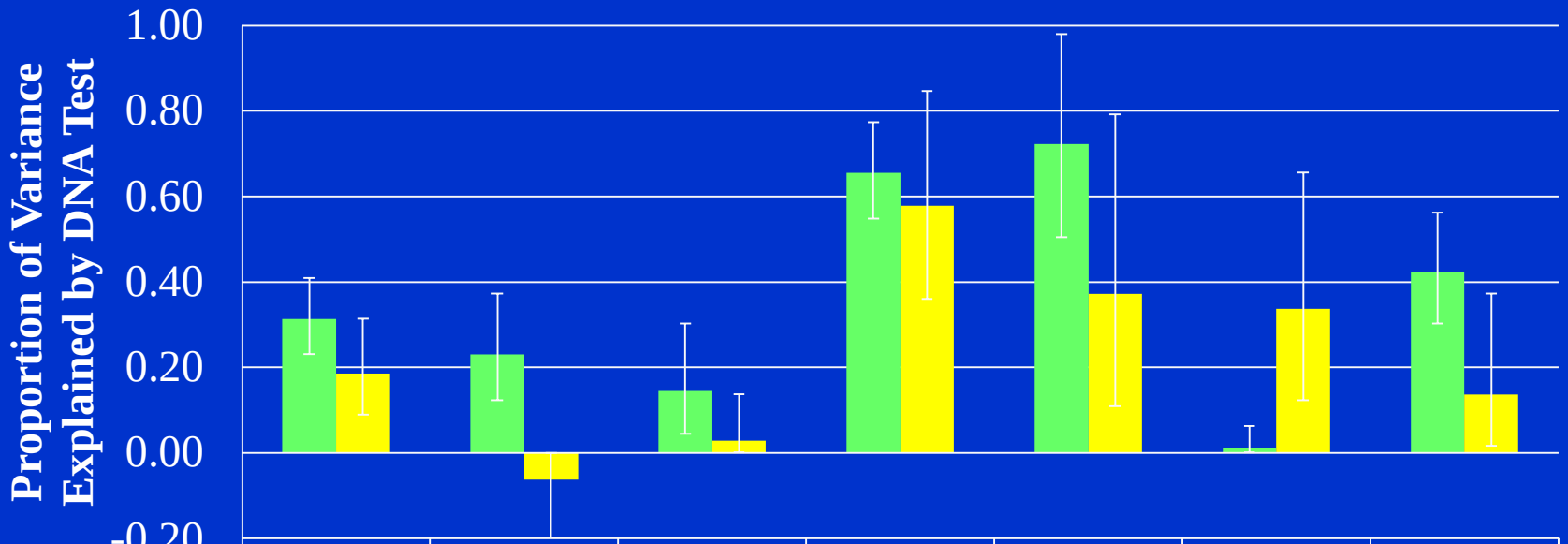
Proportion of Genetic Variance Explained in Cross-validation



	BWT	WWT	YWT	REA	MRB
MARC-trained	0.168	0.084	0.090	0.029	0.014
2000 Bull-trained	0.168	0.090	0.068	0.068	0.058

*2,000 bull predictions excluded the sires of the MARC validation populations

Proportion of Variation in Weight Traits Project from Training on 2,000 Bulls



	AN	AR	CH	GV	HH	LM	SM
Full	0.3136	0.2304	0.1444	0.6561	0.7225	0.0121	0.4225
Reduced	0.1849	-0.0625	0.0289	0.5776	0.3721	0.3364	0.1369

*Full = Prediction of sires including the 2,000 bulls

*Reduced = Predictions excluding the 2,000 bulls

What Can We Do to Improve Prediction Accuracy?

- Add more phenotypes (animals)
- Increase marker density
- Incorporate individual animal DNA sequence on influential sires
- Do a better job of using the information we have (better statistical analyses)

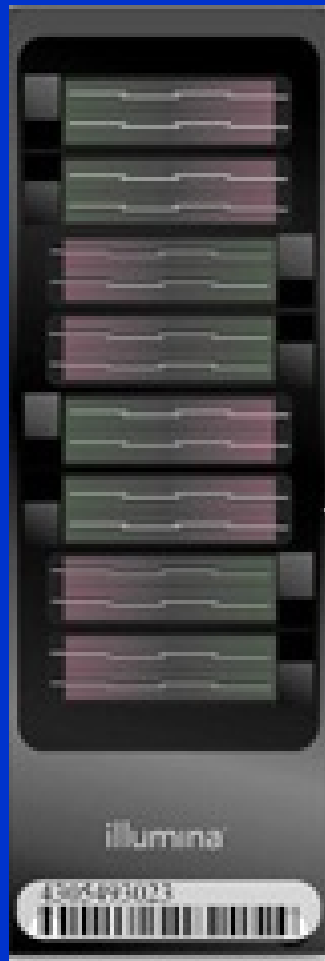
What Can We Do to Improve Prediction Accuracy?

- Add more phenotypes (animals)
- Increase marker density
- Incorporate individual animal DNA sequence on influential sires
- Do a better job of using the information we have (better statistical analyses)

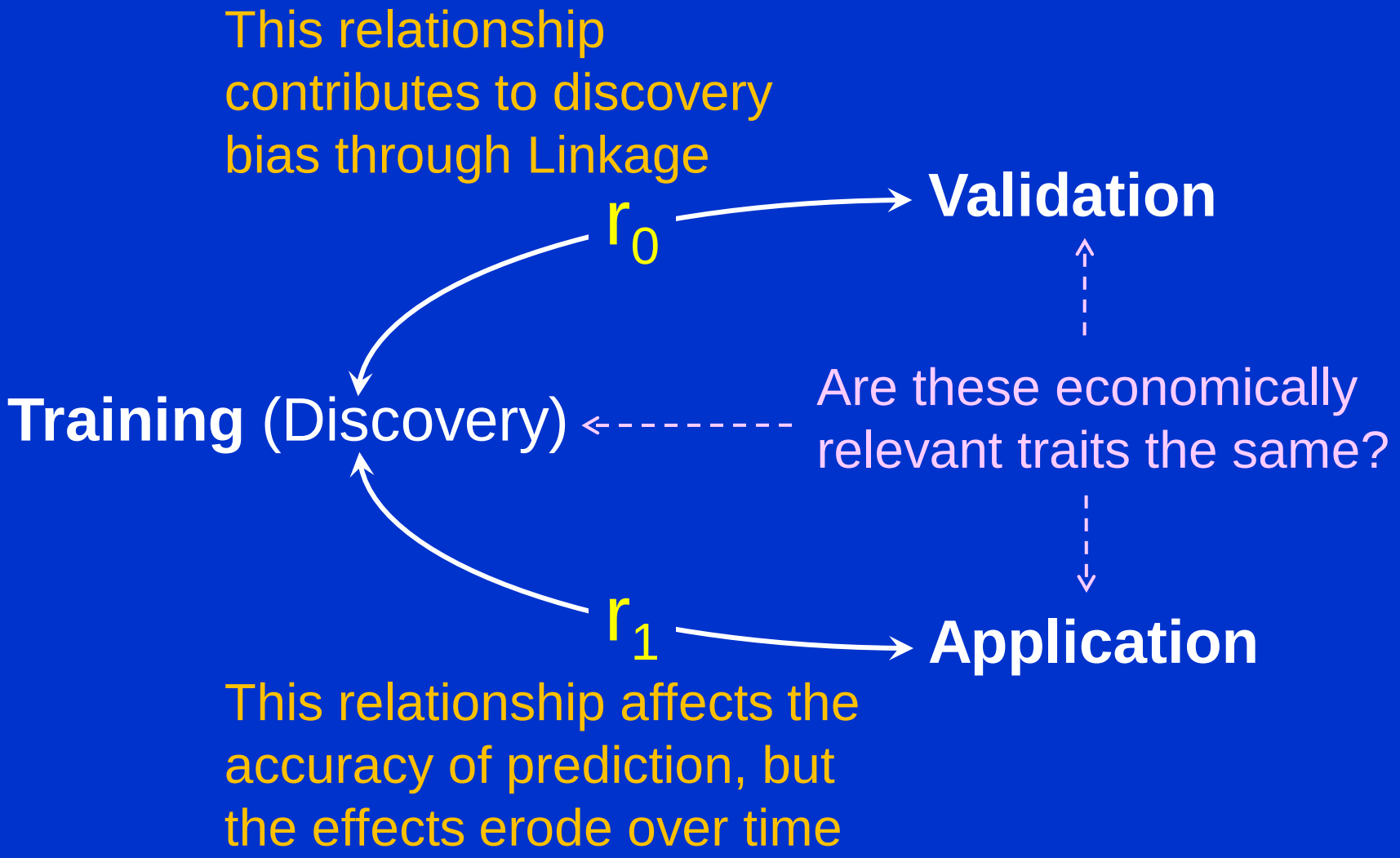
We need to find ways to use information from all available discovery populations, regardless of the target breed(s).

770K Chip

- Illumina BovineHD
- 770,000 Single Nucleotide Polymorpisms (SNP)
- Much higher SNP density than 50K chip
- Should allow predictions to be less breed-specific
- We are in the process of having this new chip run on > 300 sires that have substantial progeny at USMARC



Populations Involved in Whole Genome Selection



Prospects for Moving Beyond Validation Populations

- Conceptually, it should be possible to combine discovery and validation into a single step with a single population.
- There is still considerable work to be done to make this practical.
- This concept assumes that the accuracy of the genomic part of each individual's genetic evaluation should vary depending on its genetic relationship to the discovery population.

Prospects for Moving Beyond Validation Populations

- This concept requires that raw genotypes be available on the training/validation population as well as the target population.
- For traits that are routinely recorded in the target population, phenotypes should be continuously integrated into the training/validation population.

Conclusions

- Within-breed predictions based on the 50K work well.
- Training on multiple purebred populations is more effective than training on only a single, small purebred population.
- With increasing marker density, crossbred populations will likely become increasingly important components of training.