



Genomic Prediction with Golden Helix SNP & Variation Suite

December 16, 2014

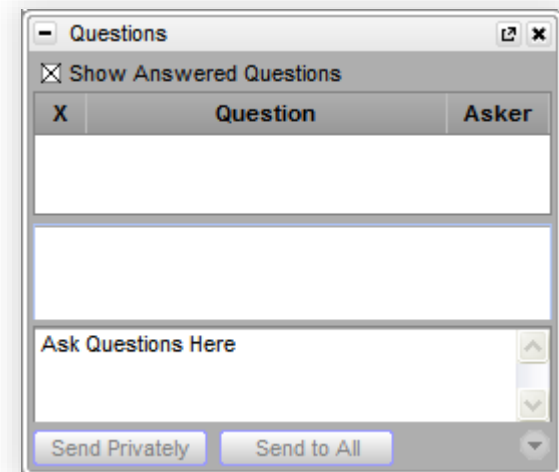
Bryce Christensen
Director of Services



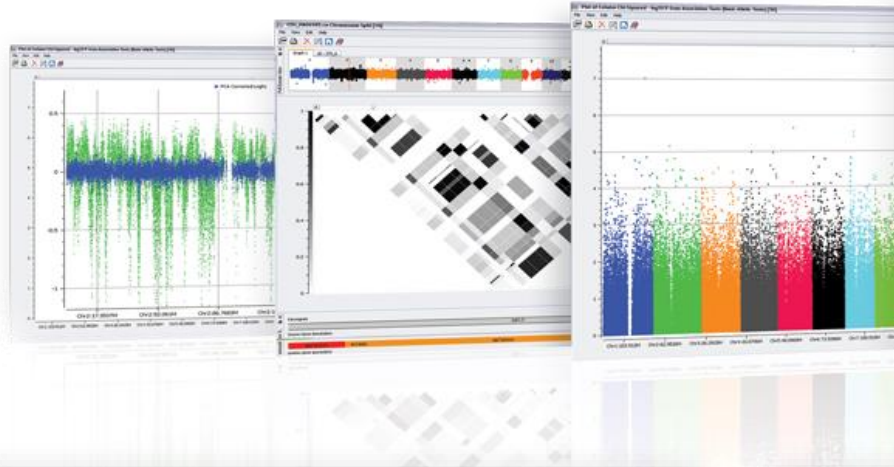


Questions during the presentation

Use the Questions pane in your GoToWebinar window



SNP & Variation Suite (SVS)



GETTING STARTED...

- Create New Project
- Online Tutorials
- Download Example Projects
- Open Existing Project

SVS_Webcast.ghp
SVS_MYH6.ghp
SVS_CephTrio_MYH6_Exons.ghp
SVS_CephTrio_MYH6.ghp
SVS_CephTrio_FakeSignal.ghp
SVS_Norartis_GWAS.ghp
SVS_DNASeq_Training 8.0 Complet...

SUPPORT BULLETINS

SVS 8.0.0 Release Candidate No...
SVS 7.7.8 Release Notes
SVS 7.7.7 Release Notes
Bug: Genotypic Covariates for Mix...
SVS 7.7.6 Release Notes
Firewall Settings for running Gold...
NEW Blazing Fast VCF Importer!

LICENSE INFORMATION

Version 8.0.0 Win64
Released 2013-10-11
License ID 4333
Expires Jul 14 2015

PACKAGE

Power Seat

- ✓ SVS Core
- ✓ GenomeBrowse
- ✓ SNP Analysis
- ✓ CNV Analysis
- ✓ DNA-Seq Analysis
- ✓ RNA-Seq Analysis
- ✓ PBAT Analysis

CONTACT SUPPORT

support@goldenhelix.com
PH: +1.406.585.8137
+1.888.589.4629

ANNOUNCEMENTS

© 2013 Golden Helix

Core Features

- Powerful Data Management
- Rich Visualizations
- Robust Statistics
- Flexibility
- Fully Supported

Applications

- Genotype Analysis
- DNA sequence analysis
- CNV Analysis
- RNA-seq
- Genomic Prediction



1 Genomic Prediction Basics

2 Cattle Case Study

3 Arabidopsis Case Study

4 Interactive SVS Demonstration



- Genomic prediction is a key focus for agrigenomics
- Growing world population requires improved food production
 - 3B in 1960
 - 7.3B today
 - 9.6B projected in 2050

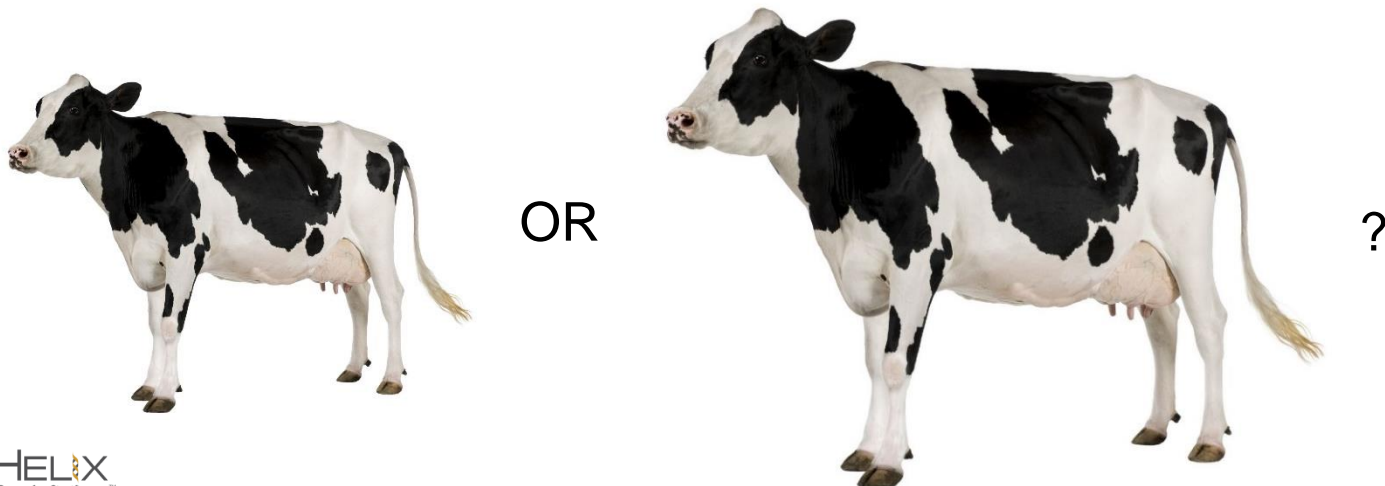
Source: Wikipedia

Country	Population 2010	Population 1990	Growth (%) 1990–2010
<i>World</i>	6,895,889,000	5,306,425,000	30.0%
China	1,341,335,000	1,145,195,000	17.1%
India	1,224,614,000	873,785,000	40.2%
United States	310,384,000	253,339,000	22.5%
Indonesia	239,871,000	184,346,000	30.1%
Brazil	194,946,000	149,650,000	30.3%
Pakistan	173,593,000	111,845,000	55.3%
Nigeria	158,423,000	97,552,000	62.4%
Bangladesh	148,692,000	105,256,000	41.3%
Russia	142,958,000	148,244,000	-3.6%
Japan	128,057,000	122,251,000	4.7%

Why Use Genomic Prediction?



- **Calculate breeding value (gEBV) for all subjects in a population**
 - May be more accurate than breeding selection based only on pedigree and trait data
- **Predict breeding values for subjects with unknown phenotypes**
 - May avoid costly and lengthy field trials
 - May not always be possible to measure the phenotype
- **Identify genetic markers with best predictive power for a trait**
 - Assist in development of predictive tests and other assays





- **Assumes all loci contribute to phenotype**
- **Incorporates genomic relationship matrix (GRM) in mixed linear model framework to account for relatedness among samples**
- **Calculates allele substitution effect (ASE) for each SNP**
- **Computes estimated breeding values (GEBV) and predicted phenotypes for all samples**
- **Also calculates:**
 - Pseudo-heritability of trait
 - Genetic component of trait variance
 - Error component of trait variance

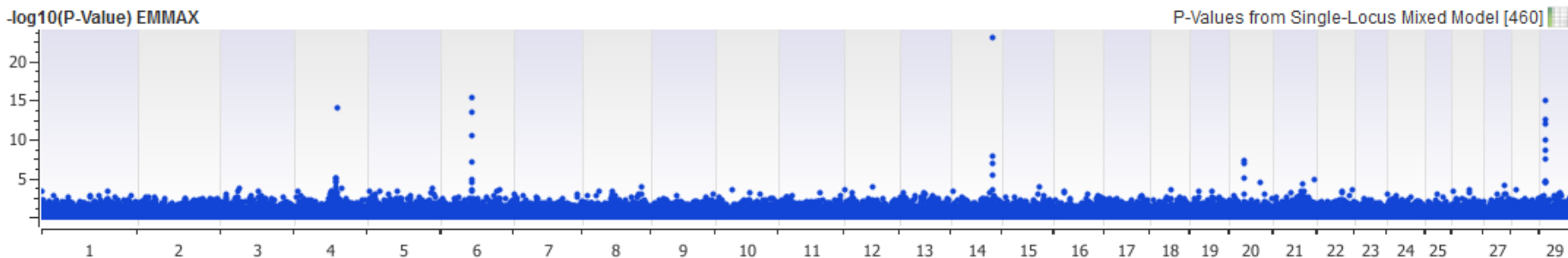
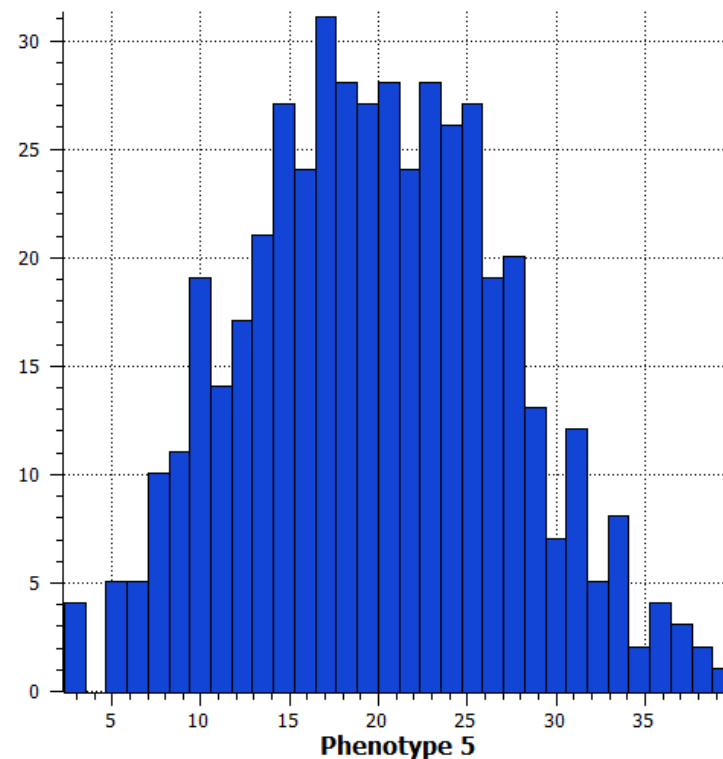


- Included in SVS 8.3
- Bayesian prediction methods estimate effects of gene loci together with parameters required to define probability distribution over effects
- Gibbs sampling (MCMC) used to obtain parameter estimates
- The π (pi) parameter is the prior probability that any SNP will have no effect.
- Pi is fixed in Bayes C, typically at $\pi=0.9$
- Pi is considered unknown and allowed to vary by Bayes C-pi
- Both methods return ASE, gEBV, other parameters of final model
- SVS implementation incorporates GRM

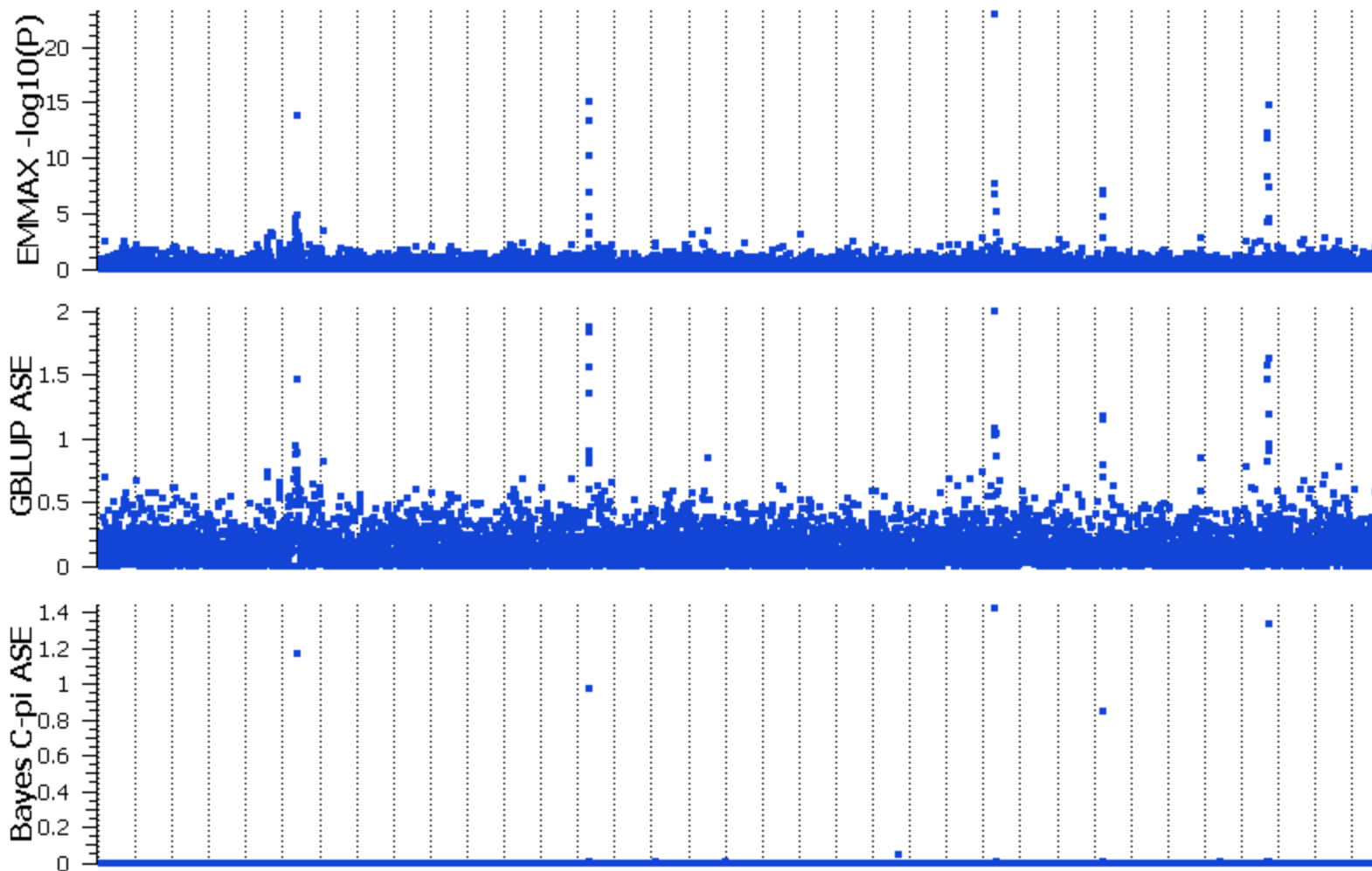
Simulated Cattle Data



- 472 cattle from Bovine HapMap project
- Illumina 50k genotypes
- Simple oligogenic trait simulation
 - 5 SNPs with independent additive effects
 - About 62% of trait explained by simulated genetic effect
- Analyzed three ways:
 - GBLUP in SVS
 - Bayes C-pi in SVS
 - Bayes C-pi in R



Allele Substitution Effects vs. GWAS results



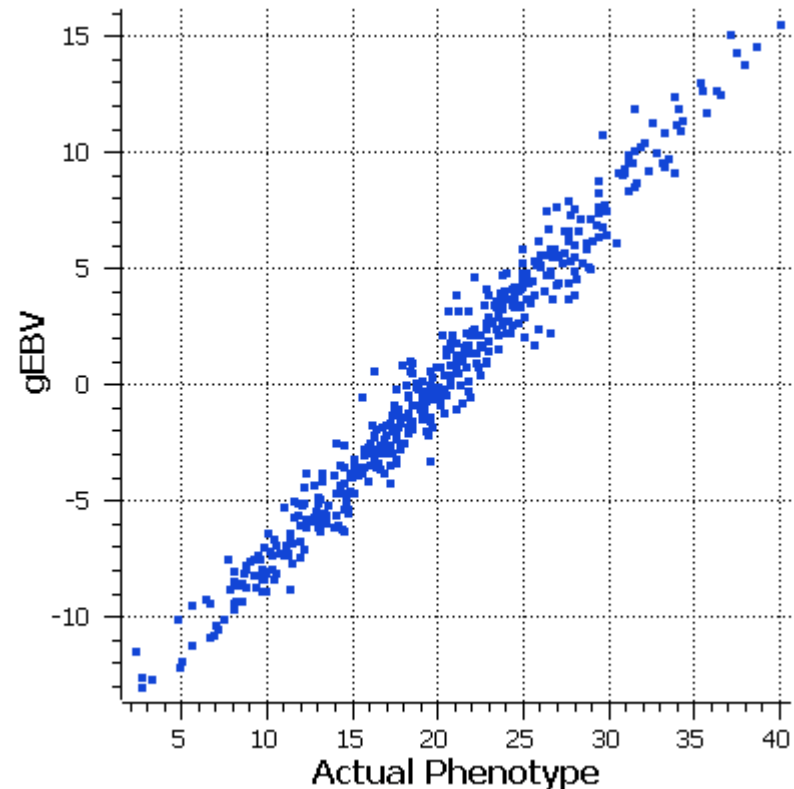


■ SVS Output

- Pseudo-heritability: 0.76
- Genomic Variance: 37.3
- Error Variance: 11.8

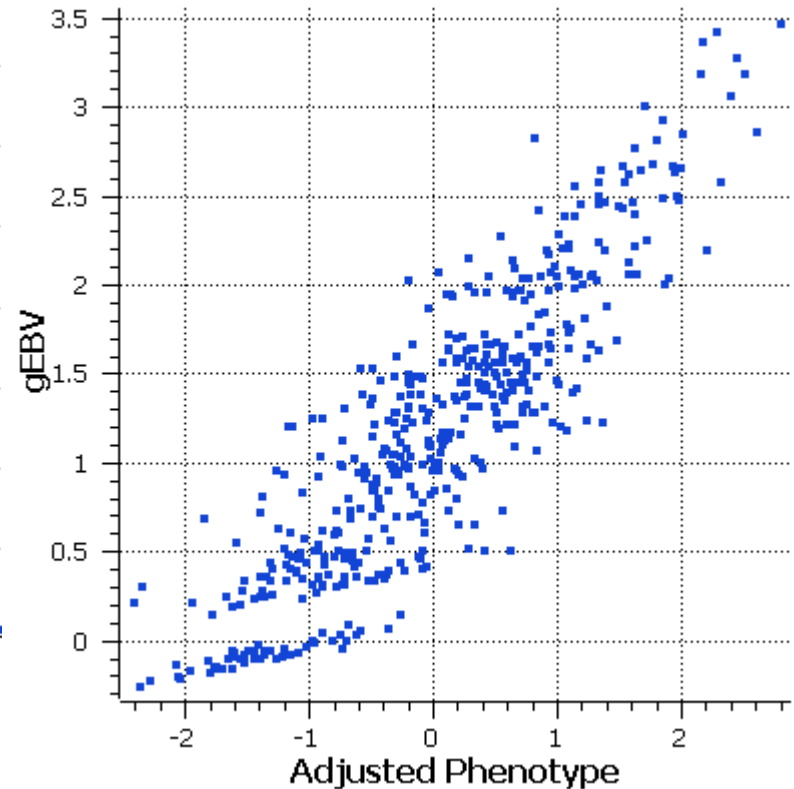
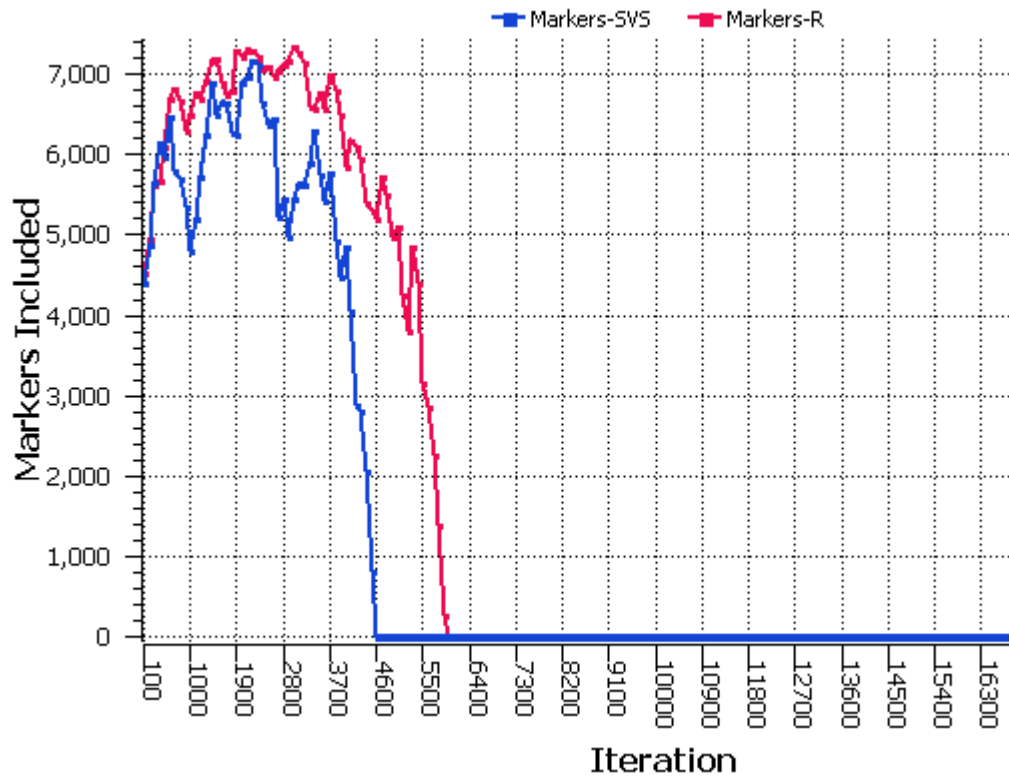
■ Trait simulation

- Genomic Variance: 31.6
- Error Variance: 16

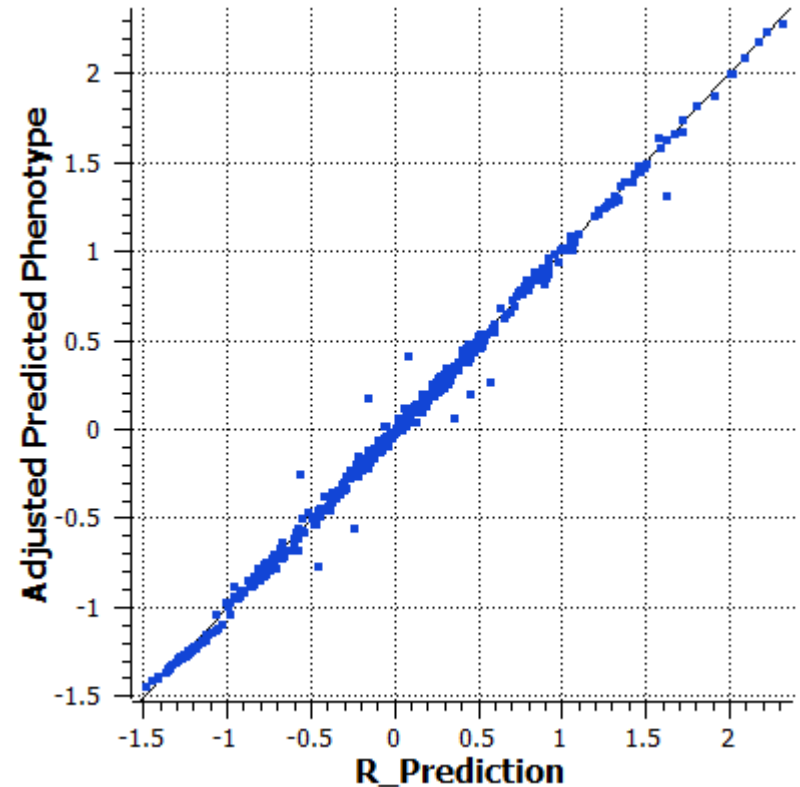
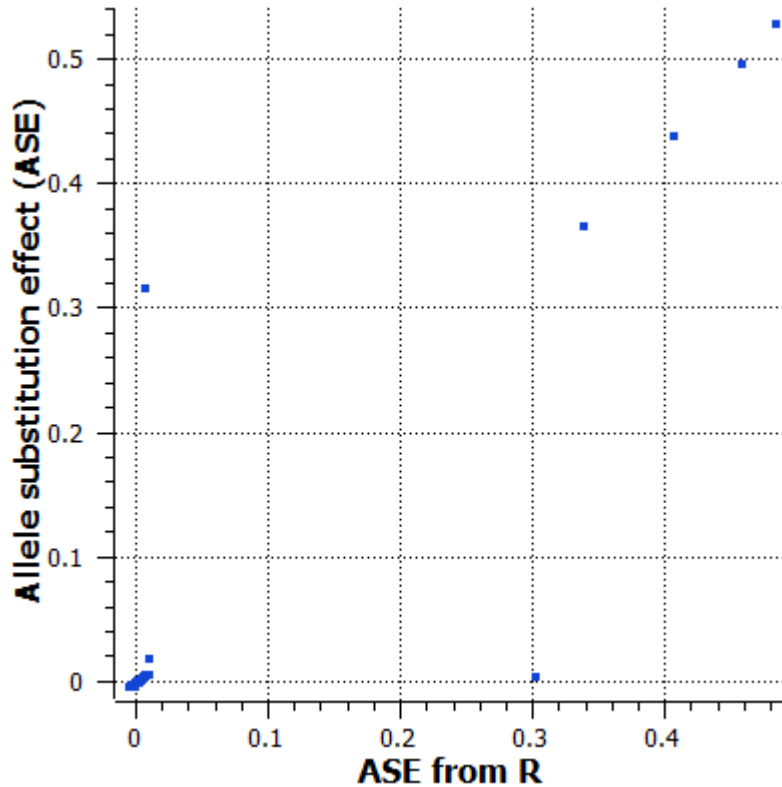




- SVS process ultimately converged on a model with 8 non-negligible SNPs
- Performance very similar to R

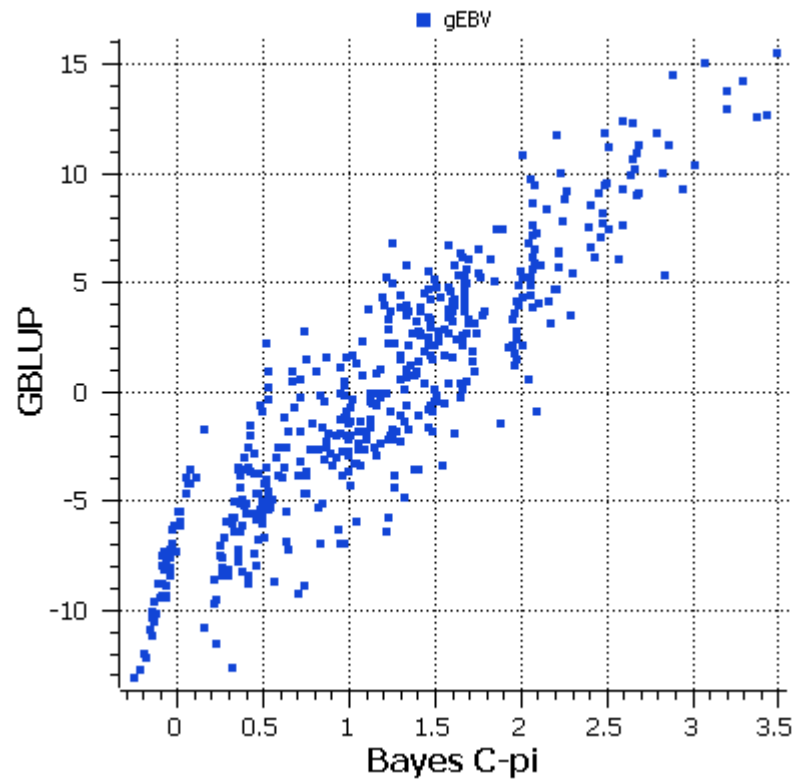
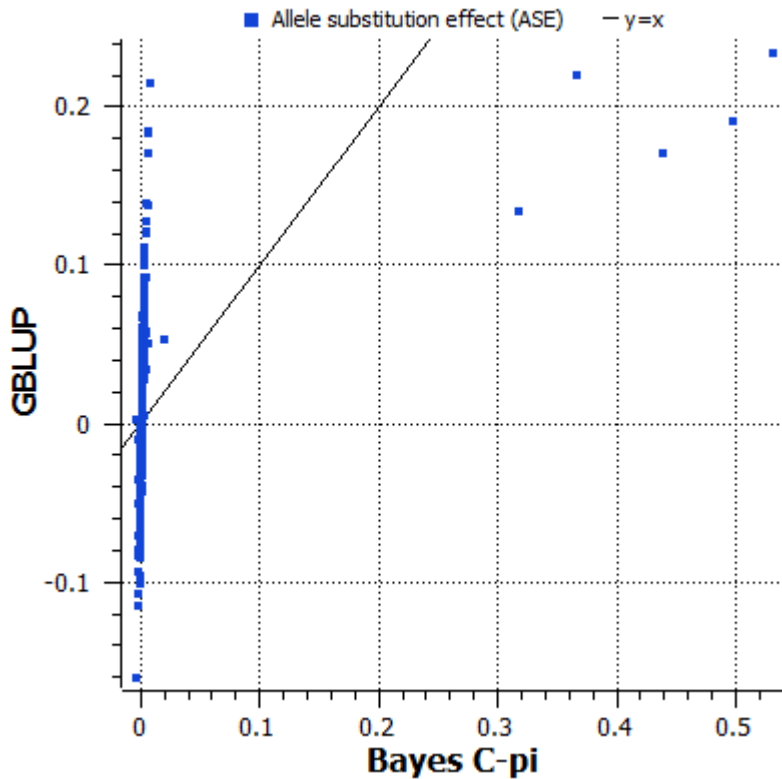


Bayes C-pi: SVS vs. R



- R and SVS selected different SNPs at the chromosome 20 locus
 - The two SNPs are in very strong LD: 96% R^2
- Results are otherwise very similar

Bayes C-pi vs GBLUP: SNP Effects

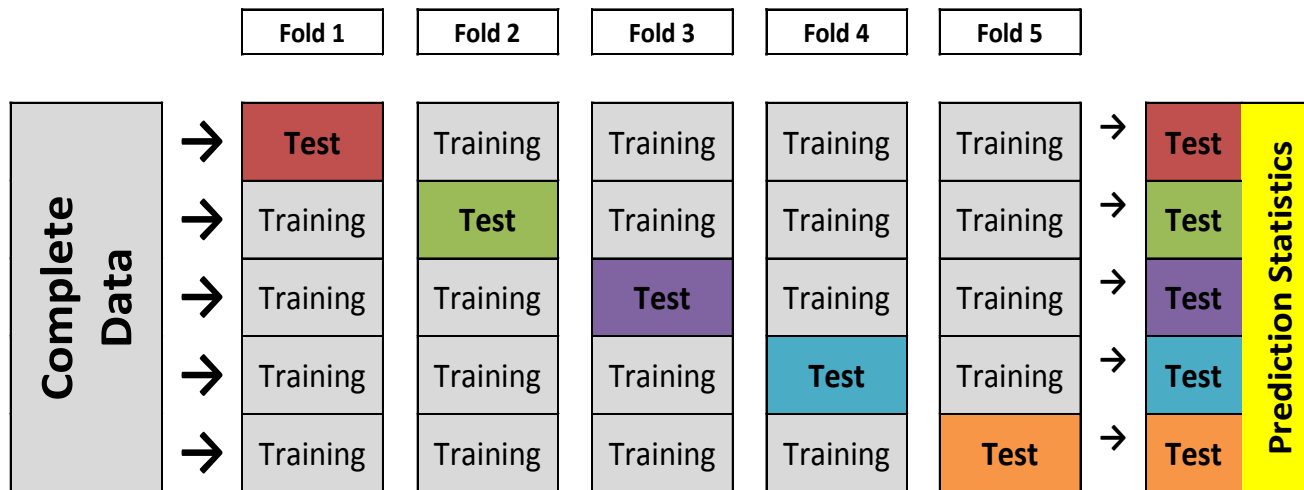


- GBLUP uses all SNPs in model, Bayes C-pi is selective
- Predicted outcomes are ultimately very similar

Cross-Validation Performance



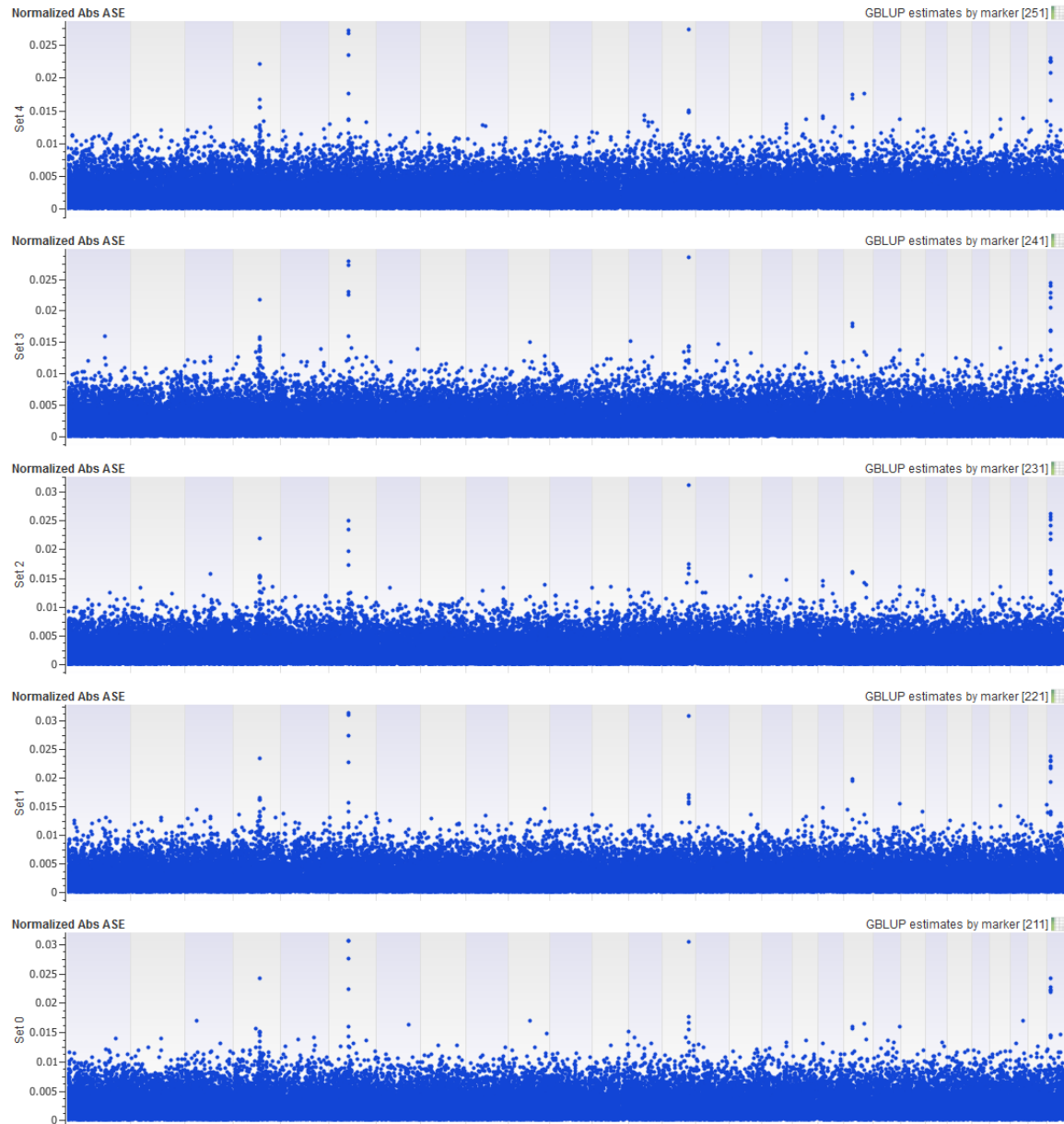
- Cross validation makes it possible to assess the performance of predictions built from a given reference/training dataset
- Five-fold cross validation was used to evaluate the predictive performance of GBLUP and Bayes C-pi for the cattle data



GBLUP ASE comparison



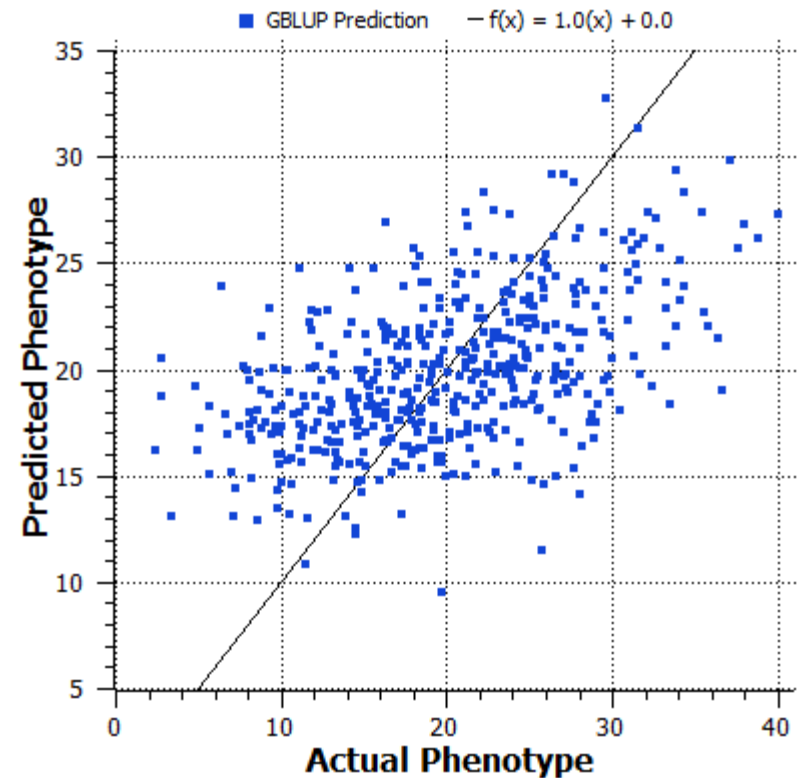
- SNP effects are very similar across subsets with GBLUP



GBLUP Prediction Performance



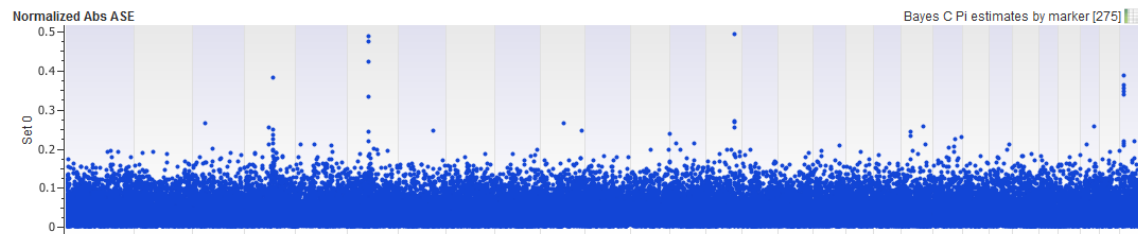
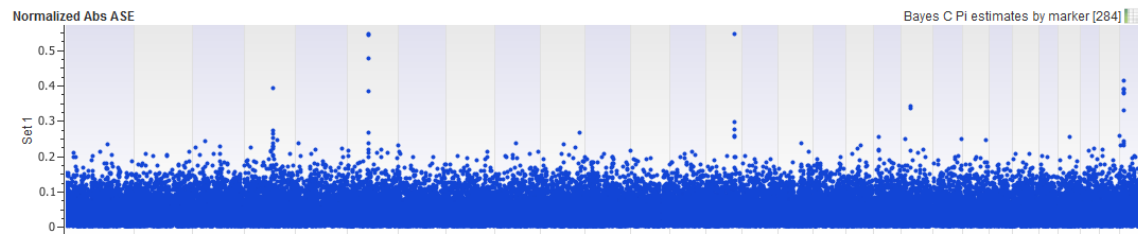
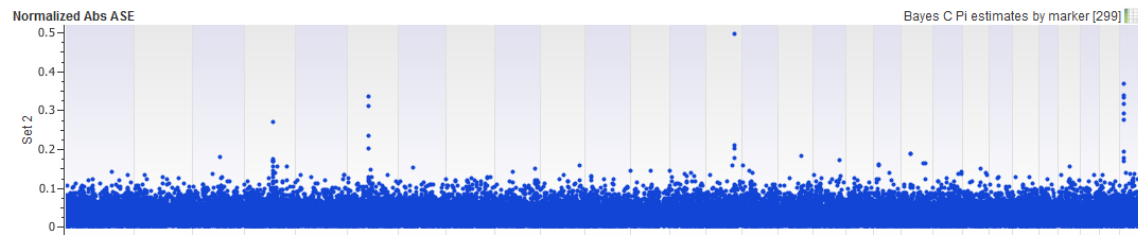
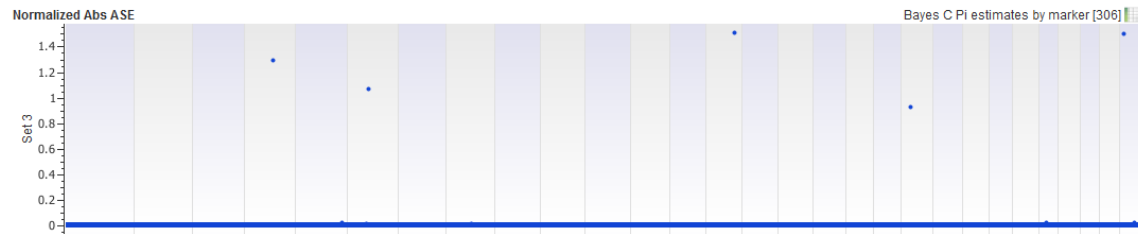
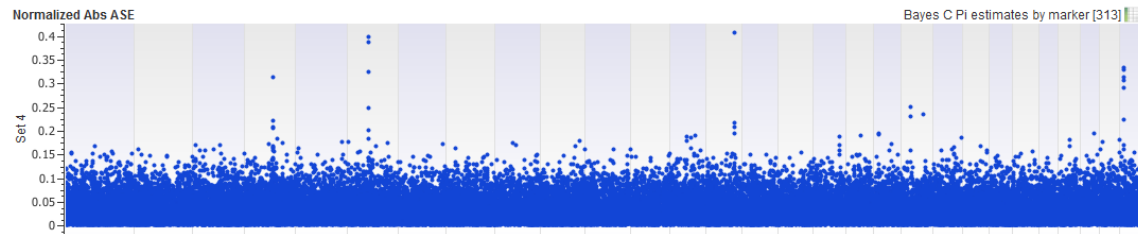
- Linear regression $p=3.5e-33$
- $R^2=0.26$



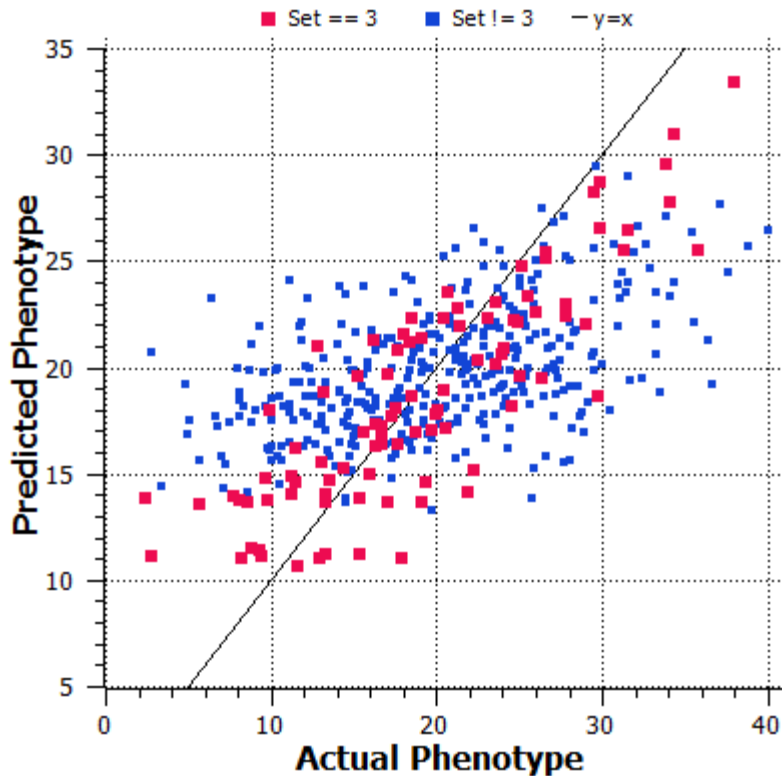
Bayes C-pi SNP effects



- Prediction model did not converge on all subsets
 - Note that some parameters were adjusted between runs for testing purposes.



Bayes C-pi Cross-Validation Results

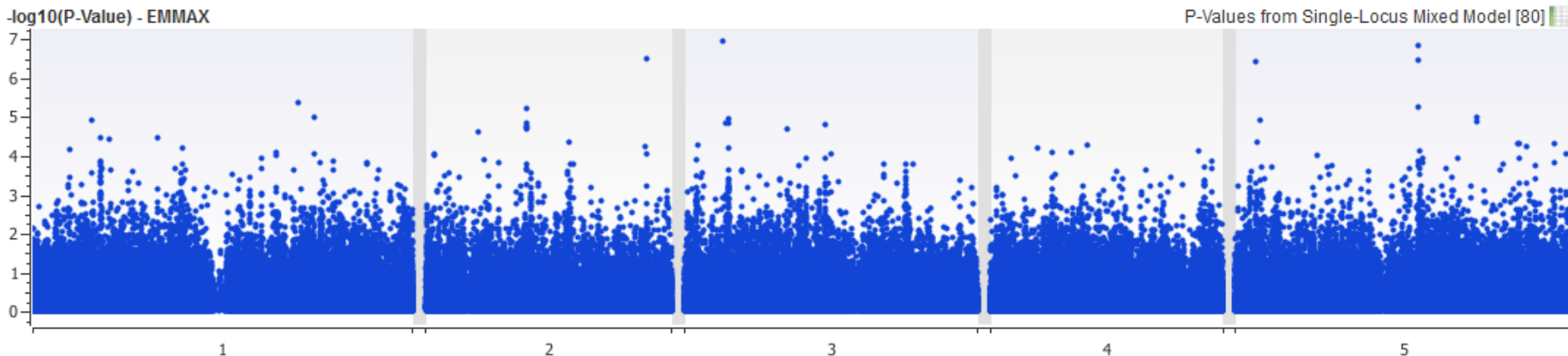
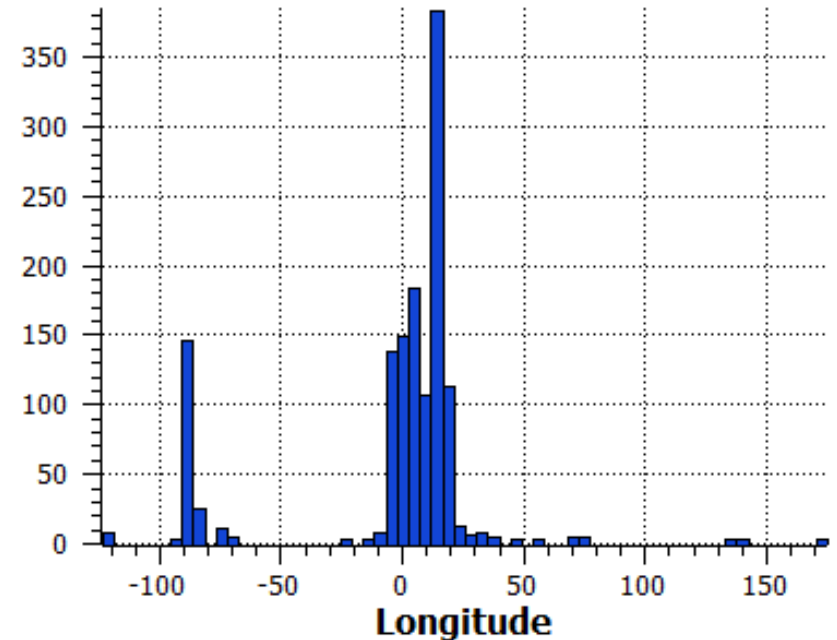


- **Overall:**
 - Linear regression $p=2.2e-48$
 - $R^2=0.37$
- **Set 3 (best convergence):**
 - Linear Regression $p=4.6e-27$
 - $R^2=0.71$
- **Variable selection methods generally perform well with oligogenic traits.**

Arabidopsis Data



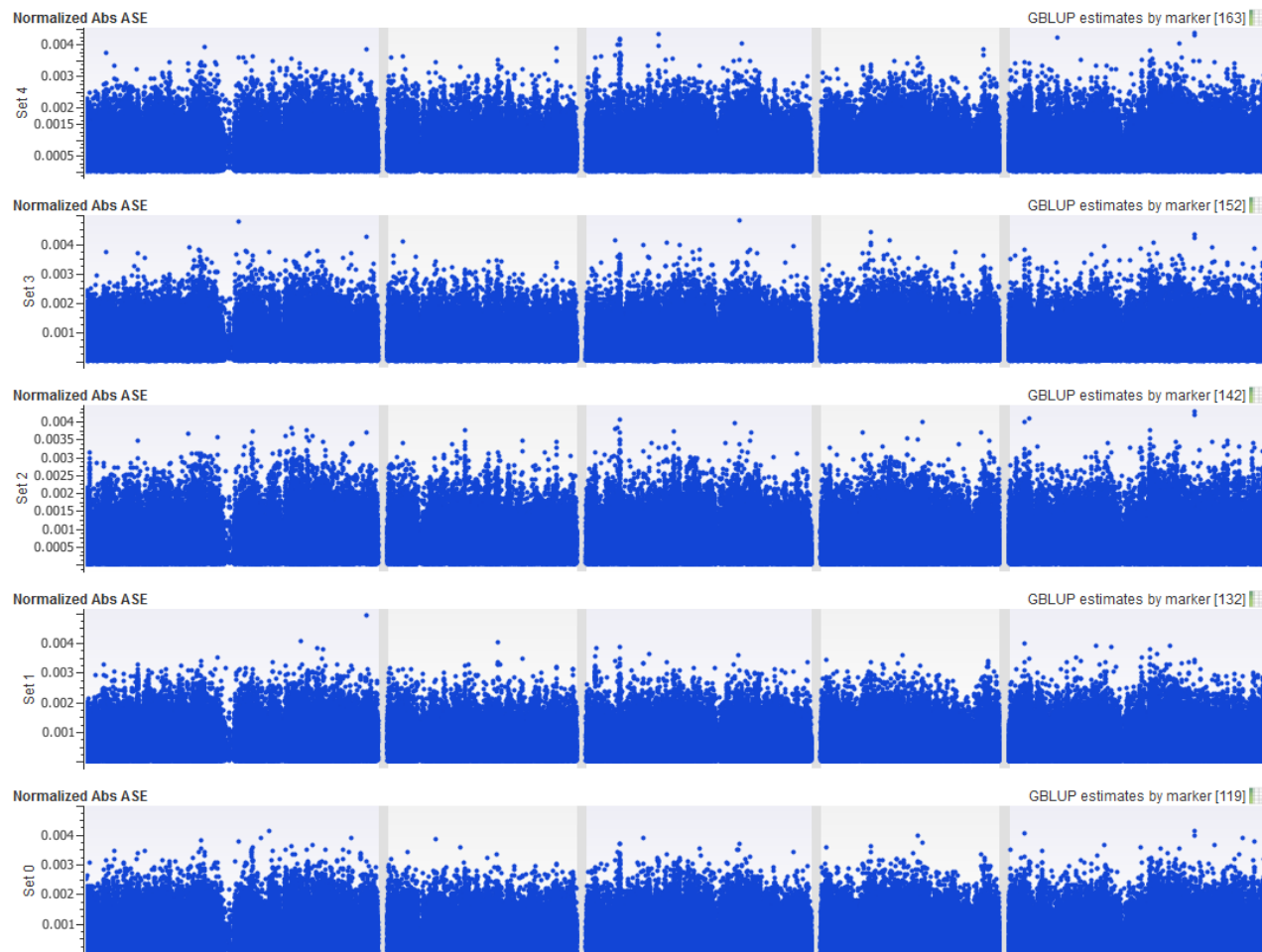
- 1307 Arabidopsis specimens publicly available from AtPolyDB (Gregor Mendel Institute)
- Inbred lines from many countries
- About 214k SNPs available from custom genotyping chip
- Longitude selected as a complex polygenic trait model.



GBLUP SNP Effects



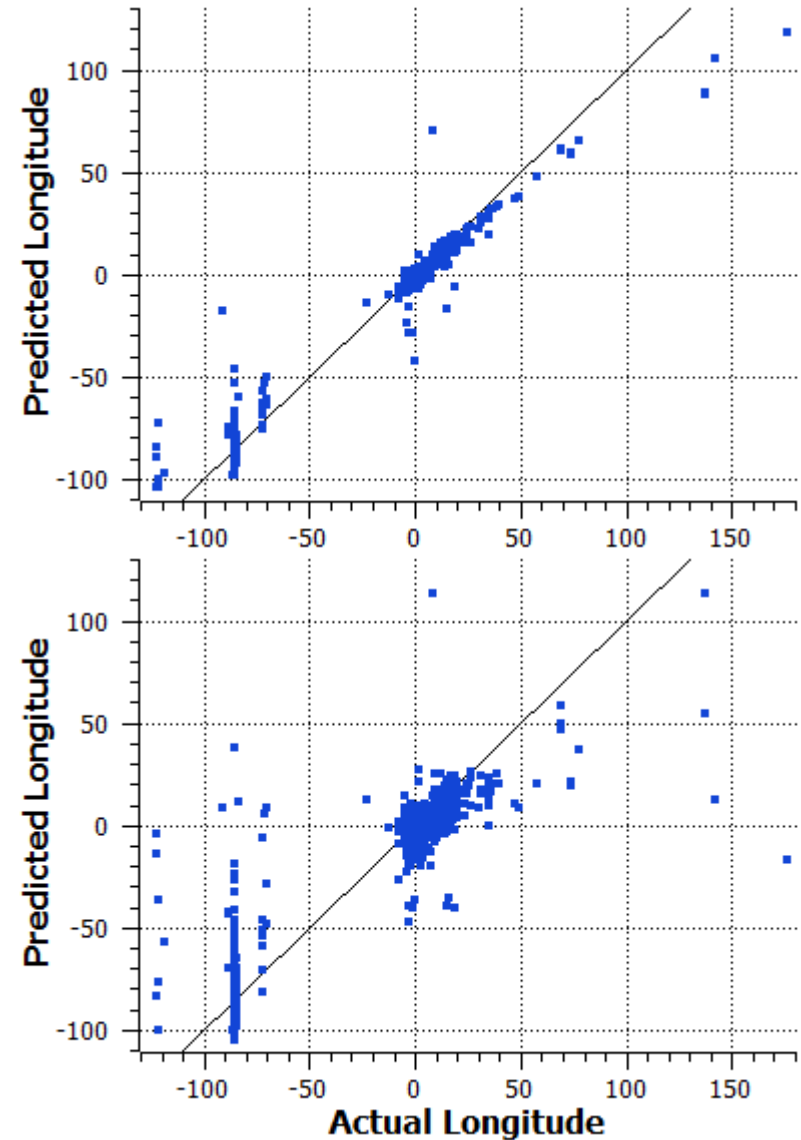
- Fairly uniform effects are observed throughout the genome
- The predictions are not dominated by any particular loci



GBLUP Prediction Performance



- **Within-sample prediction**
 - Top figure
 - Very accurate prediction
 - Overfitted!
- **Out-of sample prediction**
 - Bottom figure
 - 5-fold cross-validation
 - Good general performance: $R^2=0.83$
 - Outliers difficult to predict
- **GBLUP is generally expected to perform well for complex or polygenic traits**





SNP & VARIATION SUITE



[DEMONSTRATION]



- **Genomic prediction techniques can provide**
 - Predicted phenotypes and estimated breeding values
 - Influential loci for the phenotype
- **Genomic prediction can help breeders and researchers make decisions**
 - Which animals are likely to pass on desirable traits
 - Which loci could be used for a targeted assay for diagnostic purposes
- **SVS provides a powerful integrated solution for**
 - Data management
 - Genomic prediction
 - GWAS
 - Visualization



- Additional resources available at www.goldenhelix.com:
 - **Genomic Prediction in Agriculture**
 - eBook by Dr. Andreas Scherer
 - **Using Genomic Prediction for Trait Optimization**
 - Recorded webcast by Greta Linse Peterson
- Visit our exhibit at PAG!



Using Genomic Prediction for Trait Optimization

August 26, 2014

Greta Linse Peterson
Director of Product Management and Quality

GOLDEN HELIX
Accelerating the Quest for Applications

Genomic Prediction in Agriculture

Dr. Andreas Scherer
Golden Helix, Inc.



Questions or more info:

- Email info@goldenhelix.com
- Request an evaluation of the software at www.goldenhelix.com





Special Webcast Offer for SVS

- \$1995 for a single-named user license
- Extended 15 month license (3 months free!)
- Orders must be received by December 31st, 2014
- Register for offer at:

<http://goldenhelix.com/Events/webcasts/forms/offer.html>





Questions?

Use the Questions pane in your GoToWebinar window

