



Prediction and Meta-Analysis

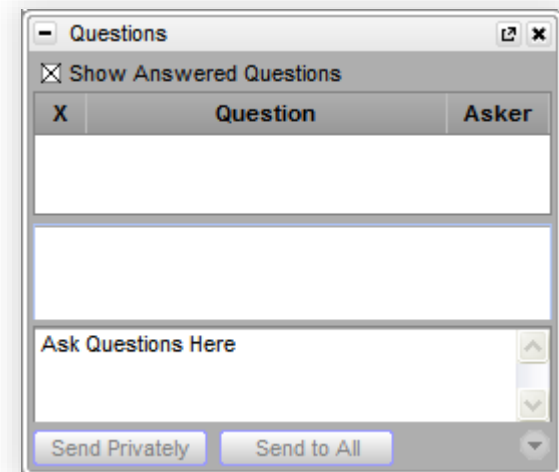
May 13, 2015

Greta Linse Peterson
Director of Product Management
& Quality



Questions during the presentation

Use the Questions pane in your GoToWebinar window



About Golden Helix

Leaders in Genetic Analytics

- Founded in 1998
- Multi-disciplinary: computer science, bioinformatics, statistics, genetics
- Software and analytic services
- Hundreds of literature citations

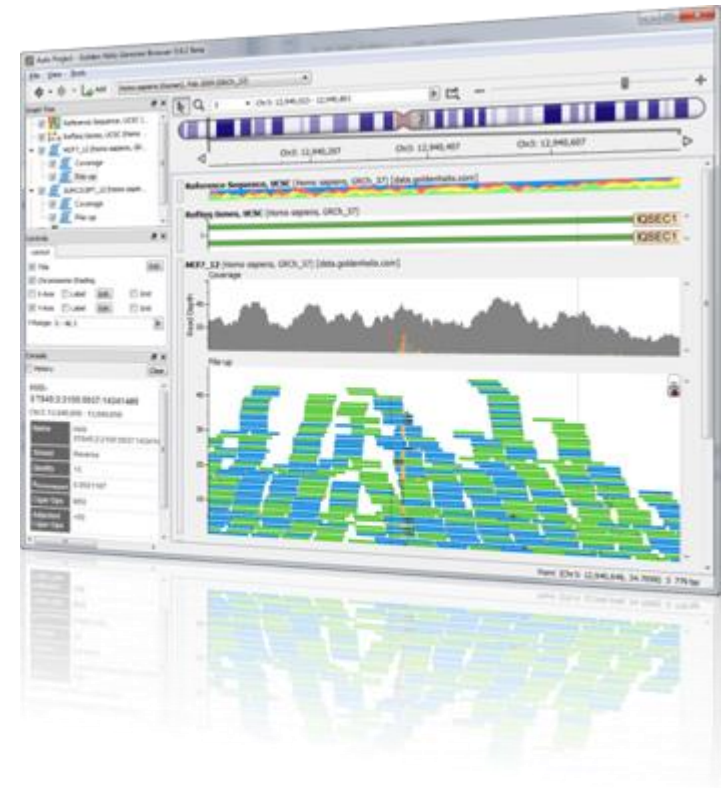
DISCOVERY DR

ENTERPRISE BLVD

GenomeBrowse



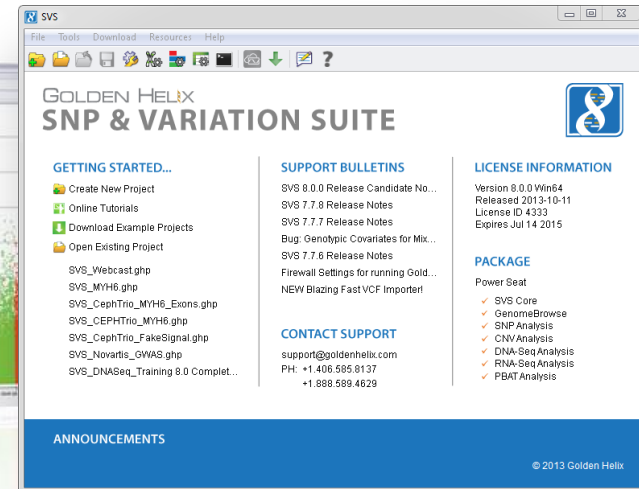
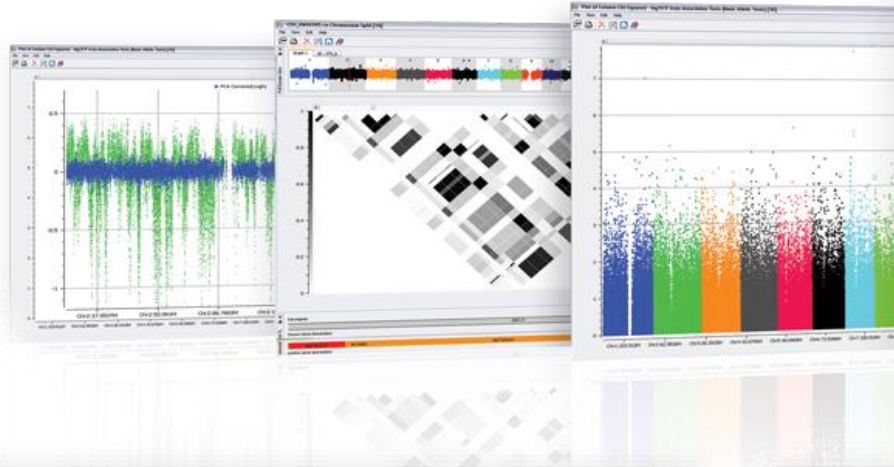
- Powerful visualization software for DNA and RNA sequencing data
- Supports most standard bioinformatics file formats
- Fast and responsive for interactive analysis
- Intuitive controls
- Stream data from the cloud and from your own remote data servers





- Powerful environment for annotation, filtering and visualization of DNaseq data
- Intuitive interface
- Repeatable workflows
- Optimized for clinical applications

SNP & Variation Suite (SVS)



Core Features

- Powerful Data Management
- Rich Visualizations
- Robust Statistics
- Flexible

Applications

- Genotype Analysis
- DNA sequence analysis
- CNV Analysis
- RNA-seq differential expression

Approximate Agenda





1 K-Fold Cross Validation with GBLUP and Bayes C/C-pi

2 Genomic Prediction

3 Meta-Analysis

4 Q&A



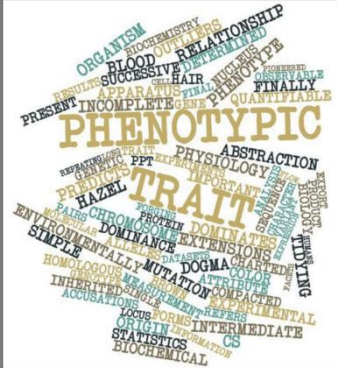
Previous Genomic Prediction Resources

Genomic Prediction with Golden Helix SNP & Variation Suite

December 16, 2014


Bryce Christensen
Director of Services

Using Genomic Prediction for Trait Optimization

August 26, 2014

Greta Linse Peterson
Director of Product Management and Quality




Our 2 SNPs...

Home About Authors Andreas Scherer Gabe Rudy Golden Helix Home

— Precision Medicine – Part III – Tailoring diagnostic and therapeutic strategies — Precision Medicine – Part IV – Adoption by Patients and Health Care Professionals —


Cross-Validation for Genomic Prediction in SVS
Posted on April 28, 2015 by Bryce Christensen

The **SNP and Variation Suite (SVS)** software currently supports three methods for genomic prediction: Genomic Best Linear Unbiased Predictors (GBLUP), Bayes C and Bayes C-pi. We have discussed these methods extensively in previous blogs and webcast events. Although there are extensive applications for these methods, they are primarily used for trait selection in agricultural genetics. Each method can be used to create models that predict phenotypic traits based on genotype data. The model is trained on samples for whom phenotypic data is available, and then used to estimate the same phenotype for samples with unknown phenotypes. But how can you determine if the model is really accurate?

Cross-validation is a powerful method for assessing how well a prediction model may perform in an independent data set. Cross-validation allows you to test the predictive potential of baseline training data internally without biasing the prediction. The basic process is simple: randomly divide the data into several equal subsets, then iteratively create and test predictive models such that each of the subsets is withheld and used for model testing one time while the remaining subsets are used to train the model. This process is known as "K-fold cross-validation," where "K" is the number of iterations used. Figure 1 is a schematic representation of 5-fold cross-validation. In this example, the complete training set is divided into 5 random subsets, and the model training and testing process is repeated five times. In each iteration, one subset is used to test a prediction model that is trained on the other 4 subsets. Upon completion, the known phenotypes for the samples can be compared with the predictions to assess model performance.

Follow...





Our 2 SNPs...

Home About Authors Andreas Scherer Gabe Rudy Golden Helix Home

— Tri-Con 2015 – just 5 days away! — Golden Helix and Fluxion Biosciences Join in a Global, Value-Added Reseller Agreement —



Q&A from our December Genomic Prediction webcast
Posted on February 12, 2015 by Cheryl Rogers

Our **Genomic Prediction webcast** in December discussed using Bayes C-pi and Genomic Best Linear Unbiased Predictors (GBLUP) to predict phenotypic traits from genotypes in order to identify the plants or animals with the best breeding potential for desirable traits.

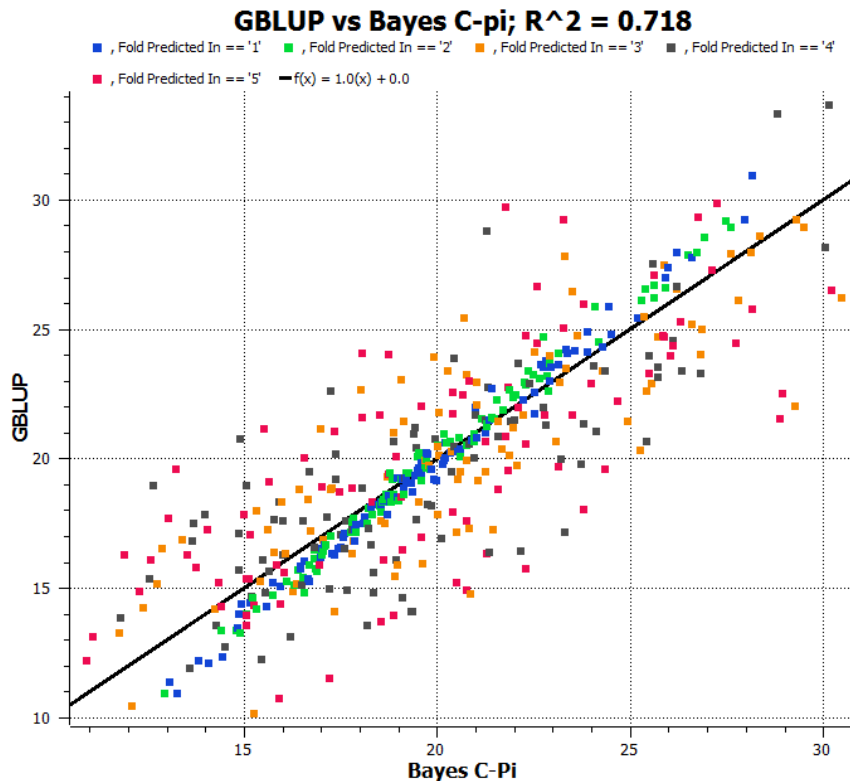
The webcast generated a lot of good questions as our webcasts generally do. I decided to begin to share these Q&A sessions with the community. If the questions below spark new questions or need clarification, feel free to get in touch with us at info@goldenhelix.com.

Question: Does the program (SNP & Variation Suite (SVS)) allow fitting fixed effects in GBLUP?

Answer: The answer to that is, absolutely yes. There is an option to add additional covariates into the model, and any numeric or binary variable or categorical variable, can be accounted for in that manner.

Genomic Prediction Methods Available in SVS



■ GBLUP

- Assumes all loci contribute to the phenotype

■ Bayes C

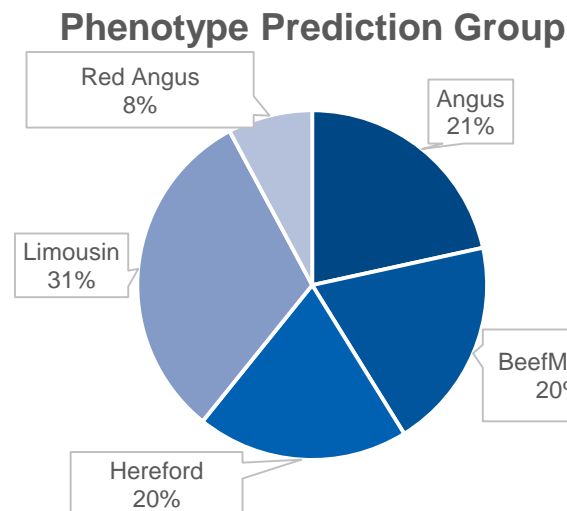
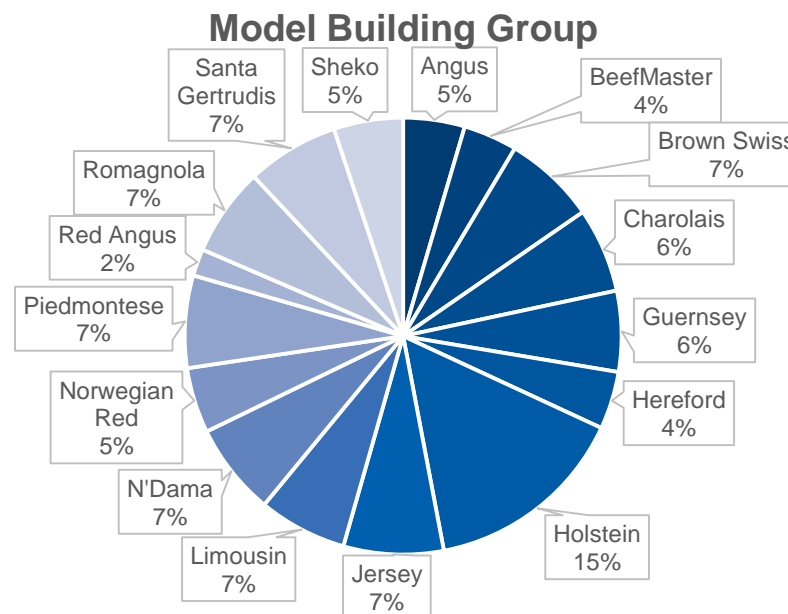
- Estimates effects of gene loci together with parameters required to define probability distribution over events
- Prior probability that any SNP will have no effect fixed

■ Bayes C-pi

- Prior probability that any SNP will have no effect unknown and allowed to vary



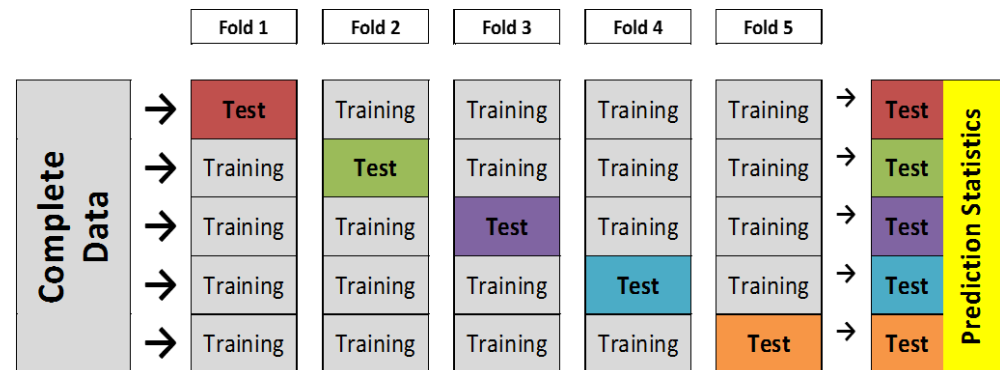
- **402 Bos taurus cattle from Bovine HapMap project**
- **Illumina 50k genotypes**
- **Simple oligogenic trait simulation**
 - 5 SNPs with independent additive effects
 - About 62% of trait explained by simulated genetic effect
- **Split into two groups:**
 - Model Building group – 351 samples from 16 breeds
 - Phenotype prediction group – 51 samples from 5 breeds



K-Fold Cross-Validation



- Use K-Fold Cross-Validation to build a model that can be applied to new genetic data to predict a phenotype
- Can be used with GBLUP, Bayes C, Bayes C-pi
- Requires all samples have a phenotype value
- Can include covariates



Cross-Validation with Multiple Iterations



K-Fold Cross Validation (for Genomic Prediction)

Computations
Perform k-fold cross validation on GBLUP and Bayes C\C-pi

Method(s)

Genomic Best Linear Unbiased Predictors (GBLUP)
 Bayes C-pi
 Bayes C

Bayesian Options

Number of Iterations:
Burn-in:
Thinning:
Initial Pi (for Bayes C this will be the fixed value):

Correct For Gender
Choose Sex Column:
Chromosome that is hemizygous for males:

Use Pre-Computed Genomic Relationship Matrix
Pre-computed genomic relationship matrix spreadsh...

Correct for Additional Covariates
Add Columns
Remove Selected
Clear List

Impute Missing Genotypic Data As:
 Homozygous major allele Numerically as average value

Stratify Folds by
Grouping...

K-Fold Options

Number of Folds
Number of Iterations

Spreadsheet Options

Delete intermediate spreadsheets with results for each fold?

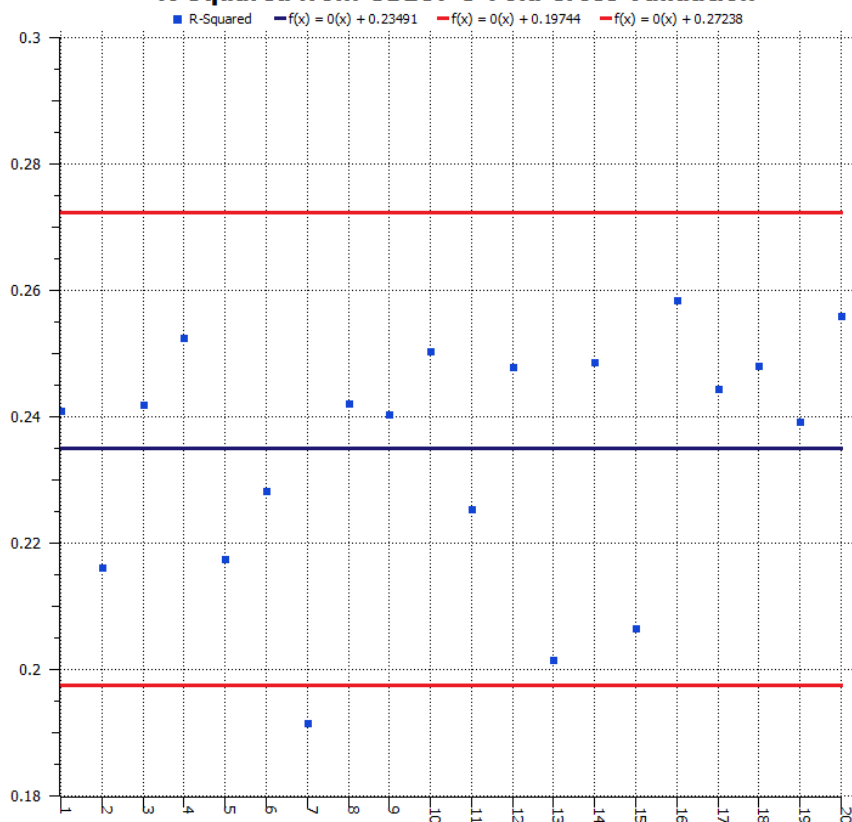
NOTE: If no pre-computed genomic relationship matrix spreadsheet is selected, a genomic relationship matrix will be computed from the genotype data and used for this analysis.

- Running K-Fold multiple times can provide statistics on the ability of the genotypes to predict the phenotype
- Binary Phenotype:
 - Sensitivity and Specificity
- Quantitative Phenotype:
 - Correlation statistics

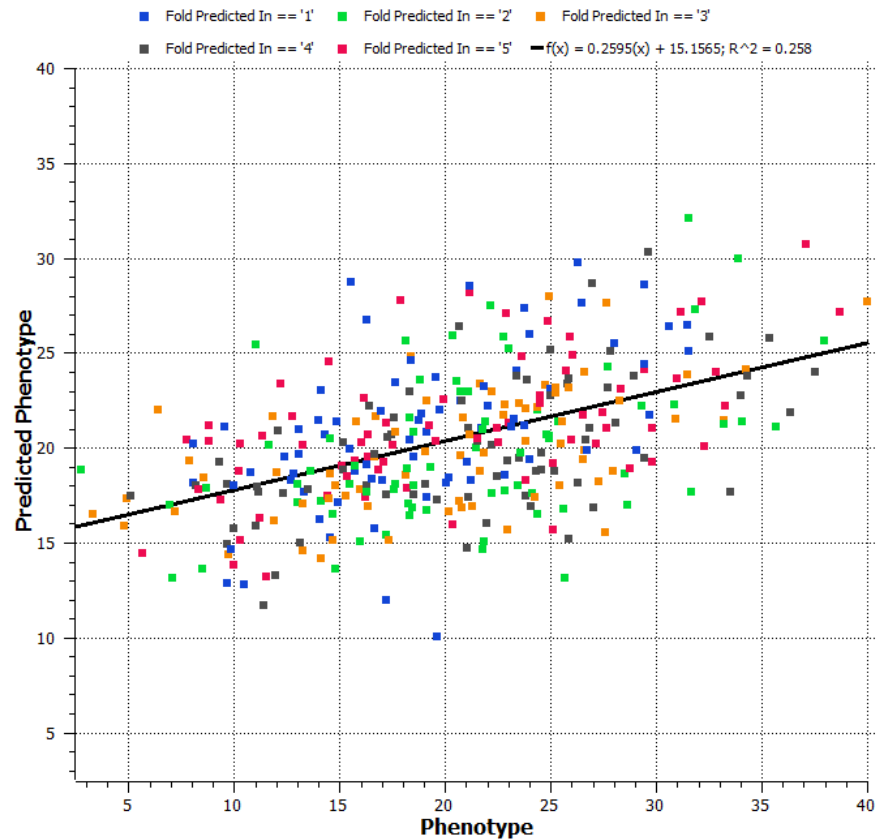
GBLUP 5-Fold Cross-Validation with 20 Iterations



R-Squared from GBLUP 5-Fold Cross Validation



GBLUP 5-Fold X-Validation Iteration 16





GOLDEN HELIX

SNP & VARIATION SUITE

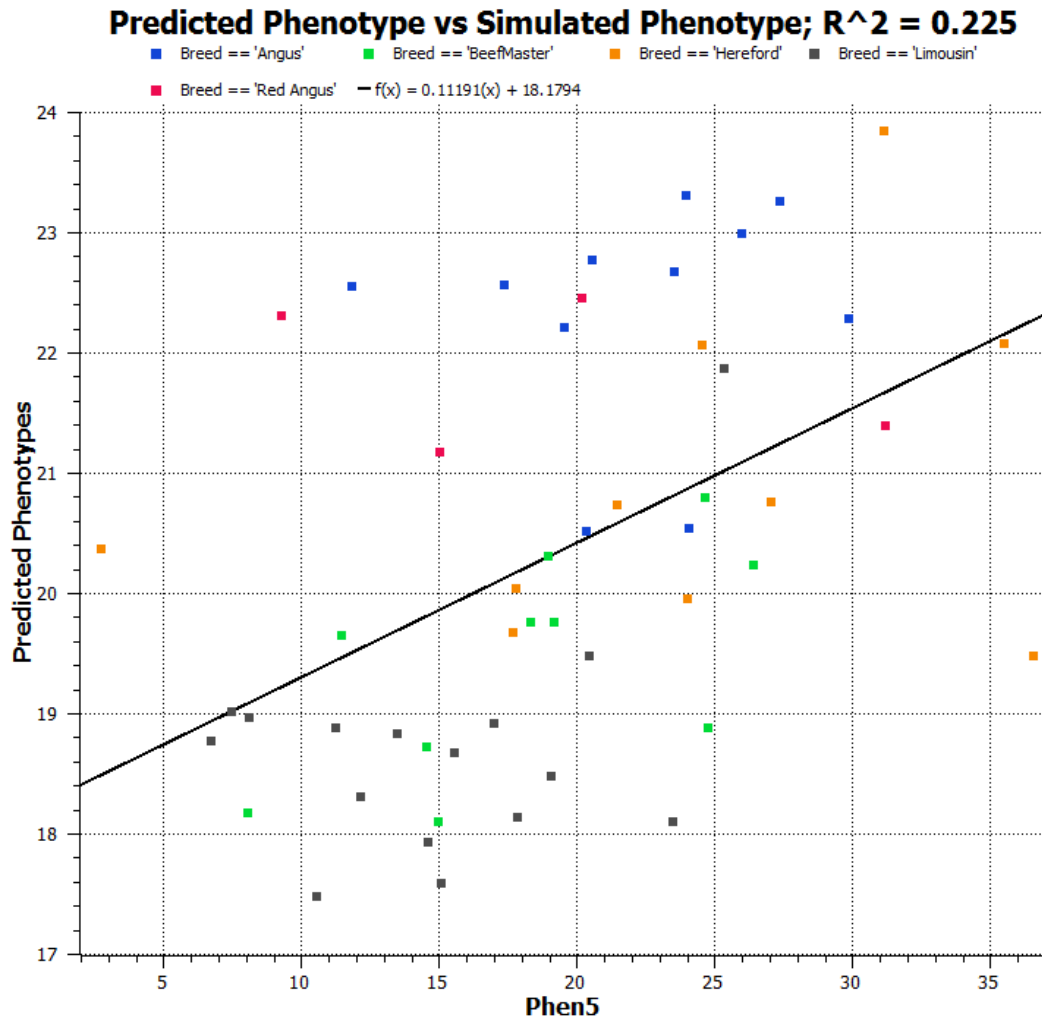
Demonstration

Applying a Prediction Model

A screenshot of a software dialog box titled "Predict Phenotypes From Existing Results". The dialog is divided into several sections: "Computations" with a checkbox for "Use Reference Spreadsheet for strand correction" and a "Reference spreadsheet" field with a "Select Sheet" button; "Computation Method(s)" with two radio buttons: "Centered (genotype values will be coded as 0, 1, or 2, then centered by the mean) (Recommended for GBLUP Results)" and "As is (genotype values will be 0, 1, or 2) (Recommended for Bayesian Results)"; "Impute Missing Genotypic Data As:" with two radio buttons: "Homozygous major allele" and "Numerically as average value"; "Correct For Gender" with a "Choose Sex Column:" field and a "Select Column" button, and a "Chromosome that is hemizygous for males:" field with the value "X"; "Homozygous Markers:" with two radio buttons: "Include (Recommended for GBLUP Results)" and "Remove (Recommended for Bayesian Results)"; "Correct for Additional Covariates" with a large empty list box and buttons for "Add Columns", "Remove Selected", and "Clear List"; "Transformed Data" with input fields for "Mean" and "Standard Deviation", both set to "0"; and "Model Values" with two checkboxes: "Allele Substitution Effects" and "Fixed Effect Coefficients", each with a "Select Sheet" button. At the bottom are "OK", "Cancel", and "Help" buttons.

- Starts from a spreadsheet of genotype data or numeric data
- Recodes to numeric if necessary
- Adjusts the recoding based on strand as needed
- Takes from K-Fold output:
 - Allele Substitution Effects
 - Fixed Effect Coefficients (needs the Intercept at a minimum)
- → Predicted phenotype value

Prediction Results



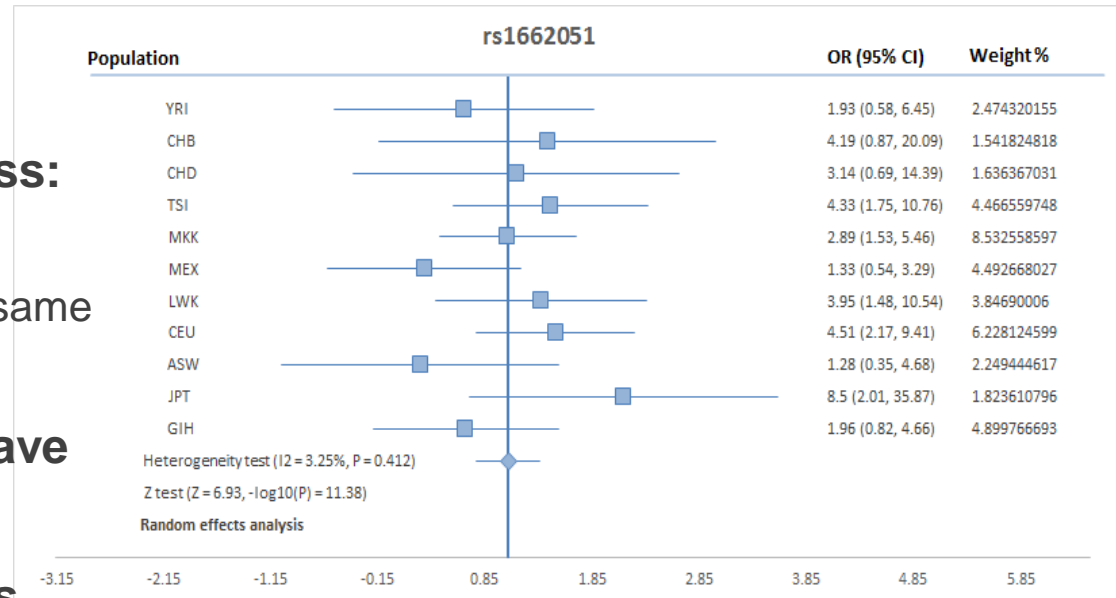


GOLDEN HELIX
SNP & VARIATION SUITE

Demonstration



- **Test effect of marker across:**
 - multiple published studies
 - population groups within the same study
- **Useful when you do not have access to the raw data**
- **Corrects for strand flips as long as the major and minor alleles are provided**
- **Weights studies based on effective sample size**



Meta-Analysis Overview



■ Effect Data Meta-Analysis

- Compare p-values across studies
- Need also:
 - Effect Direction
 - Effective number of samples

■ Inverse-Variance Method

- Compare a combination of Odds ratios and effect sizes
- Need also:
 - Either Odds Ratio CI, or
 - Effect Standard Error

Study has:	Study A	Study B	Study C	Study D	Study E	
P-values	X		X	X		Effect Data Input (P-values and Effective number of samples)
Effect Direction	X	X	X	X		
# Cases & # Controls per Marker			X	X		
# Samples per Marker	X					
Odds Ratio		X		X		Inverse-Variance Based Input
Odds Ratio Confidence Interval		X		X		
Effect Size	X			X	X	
Effect Standard Error	X			X	X	



- Weight is either square root of sample size or inverse variance
- Assumes that all studies are based on the same:
 - Population
 - Phenotype

Table 1. Formulae for meta-analysis

	Analytical strategy	
	Sample size based	Inverse variance based
Inputs	N_i - sample size for study i P_i - P -value for study i Δ_i - direction of effect for study i	β_i - effect size estimate for study i se_i - standard error for study i
Intermediate Statistics	$Z_i = \Phi^{-1}(P_i/2) * \text{sign}(\Delta_i)$ $w_i = \sqrt{N_i}$	$w_i = 1/SE_i^2$ $se = \sqrt{1/\sum_i w_i}$ $\beta = \sum_i \beta_i w_i / \sum_i w_i$
Overall Z-Score	$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$	$Z = \beta/SE$
Overall P -value		$P = 2\Phi(-Z)$

Taken from: Willer, C.J., Li, Y. and Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190--2191. ([link](#))

Random Effects Model Statistics



Assumes:

- Studies included in the meta-analysis are a random sample of all studies
- The effects vary around an overall average effect

Includes:

- Within-study variability aka random error
- Between-study variability aka heterogeneity

	Statistic	Description	
Inputs	w_i	weight for study i	
	β_i	effect size estimate for study i	
	$df = N - 1$	N is the number of studies	
Intermediate Statistics	$\beta = \frac{\sum_i \beta_i w_i}{\sum_i w_i}$	Overall effect size	
	$Q = \sum_i w_i (\beta - \beta_i)^2$	Cochran's Q	
	$I^2 = \begin{cases} \frac{Q - df}{Q} & \text{if } Q > df \\ 0 & \text{if } Q \leq df \end{cases}$	Between-studies variance	
	$\tau^2 = \max\left(0, \frac{Q - df}{\sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}}\right)$	Within-studies variance	
	$w_i^* = \frac{1}{\tau^2 + \frac{1}{w_i}}$	Random effects weight for study i	
	$\beta^* = \frac{\sum_i \beta_i w_i^*}{\sum_i w_i^*}$	RE weighted overall effect estimate	
	$V^* = \frac{1}{\sum_i w_i^*}$	RE variance of the combined effect	
	$se^* = \sqrt{V^*} = \frac{1}{\sqrt{\sum_i w_i^*}}$	RE standard error of the combined effect	
	Overall Z-Score	$Z^* = \frac{\beta^*}{se^*}$	
	Overall χ^2 -Score	$\chi^{2*} = \frac{(\beta^*)^2}{V^*}$	
Overall p-value	$P = 1 - \Phi(\chi^{2*})$		

Borenstein, M., Hedges, L. and Rothstein, H. (2007) Meta-Analysis Fixed effect vs. random effects. www.Meta-Analysis.com ([link](#))

Nordmann, A.J., Kasenda, B. and Briel, M. (2012) Meta-analyses: what the can and cannot do. *Swiss Med Wkly.* **142**:w13518 ([link](#))

Implementation in SVS (Preview)



Choose Options for Study # 1

Study # 1: Association Tests (Dominant Model) (Reference: Reference) [10]

Select marker name column [or other result label column]

Input Fields for Effect Data

Use Sample-Size-Based Inputs *

Select p-value column

Select effect direction column

Actual number of samples by marker

Select number of samples column

Actual number of samples overall

Number of total samples

Effective number of samples by marker, computed from:

Select number of cases column

Select number of controls column

Effective number of samples overall, computed from:

Number of cases

Number of controls

Use Inverse-Variance-Based (Effect Size) Inputs *

Select effect size column

Select standard error column

Use Inverse-Variance-Based (Odds Ratio) Inputs *

Select odds ratio column

Select column for the CI lower bound

Select column for the CI upper bound

* NOTE: Either sample-size-based inputs or inverse-variance-based inputs must be used exclusively for all studies in the meta-analysis.

Use Genomic Control for this study (Chi-Squared dist. with 1 degree of freedom assumed)

- Options chosen for the first study inform the options for subsequent studies
- For first study can choose
 - Effect Data Method, or
 - Inverse-Variance Based Method
- For subsequent studies only the group of options chosen for the first study will be available

Meta-Analysis Output



■ Fixed Effects Model

- P-value
- Effect Size
- Standard Error
- Z
- Chi-Squared

■ Random Effects Model

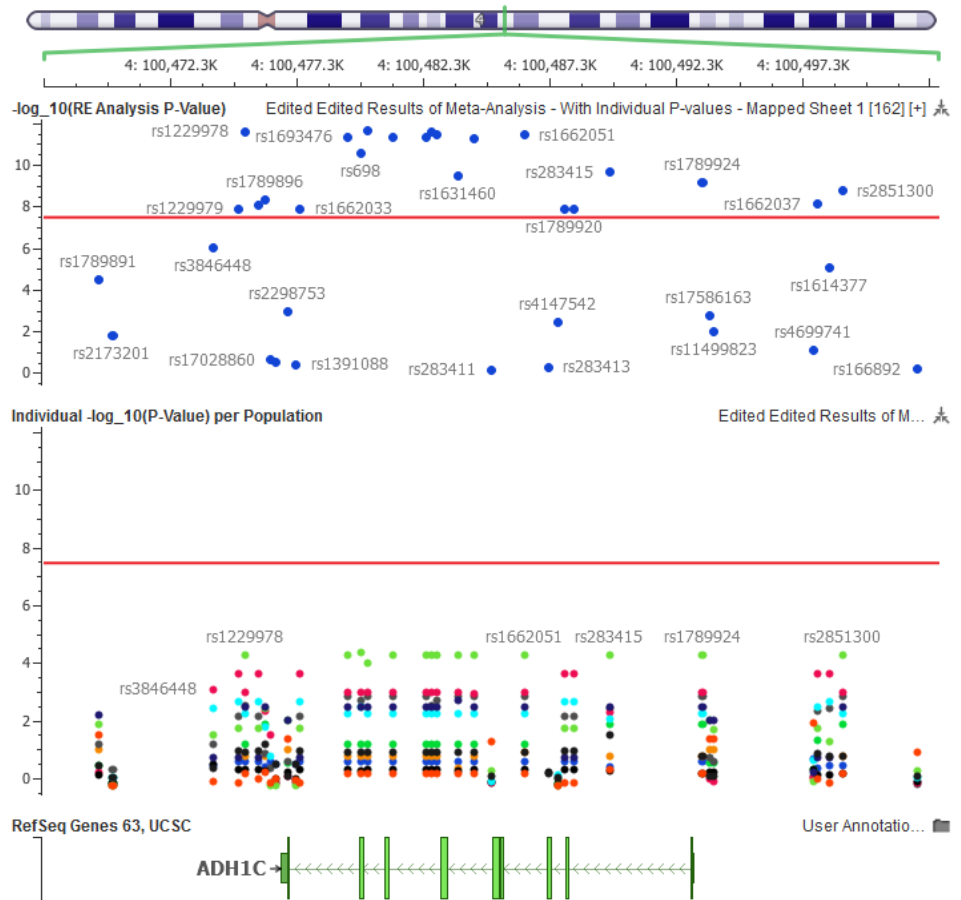
- Same output as Fixed Effects Model

■ Cochran's Q

■ I-Squared

■ Tau-Squared

■ (Optional) Genomic Control





GOLDEN HELIX
SNP & VARIATION SUITE

Demonstration



- **K-Fold Cross Validation can be used to build a genomic prediction model**
- **Prediction models can now be applied to new data without having to worry about merging data**
- **Coming soon SVS will have Meta-Analysis methods available**
- **The power of SVS data manipulation, visualization and user friendly GUIs make these methods easier to learn and use.**





Questions or more info:

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





Questions?

Use the Questions pane in your GoToWebinar window

