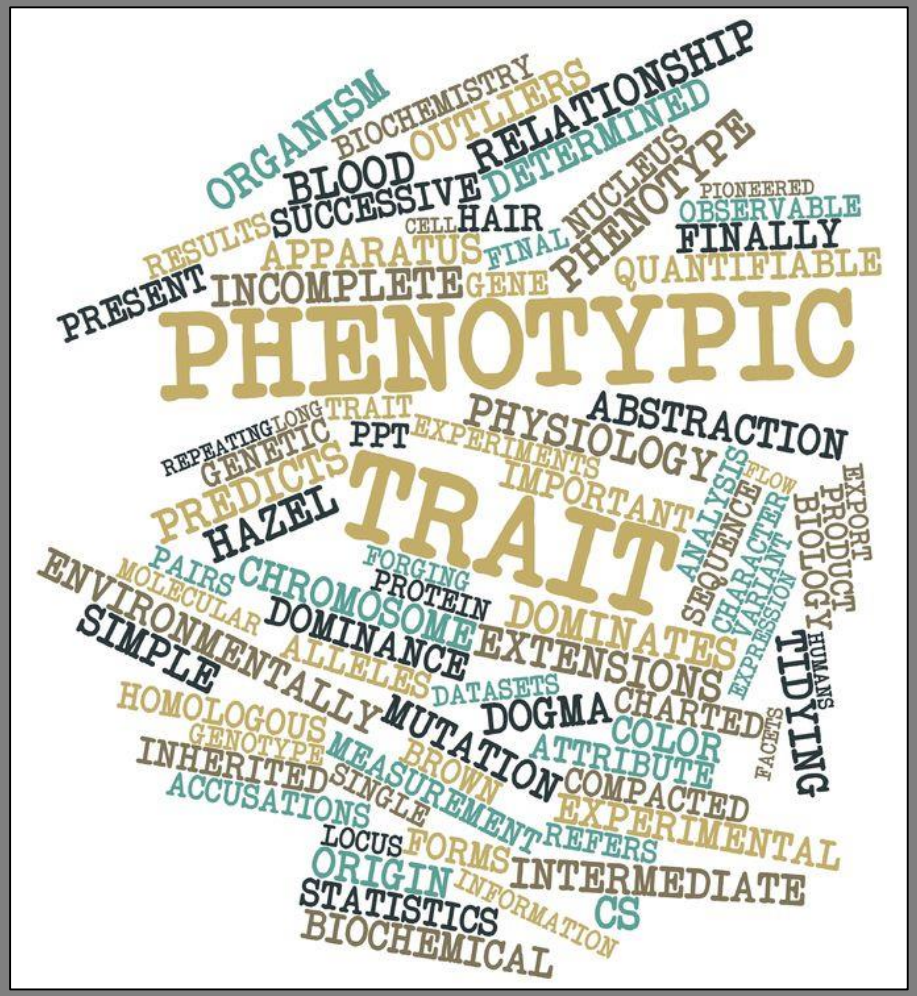


Using Genomic Prediction for Trait Optimization

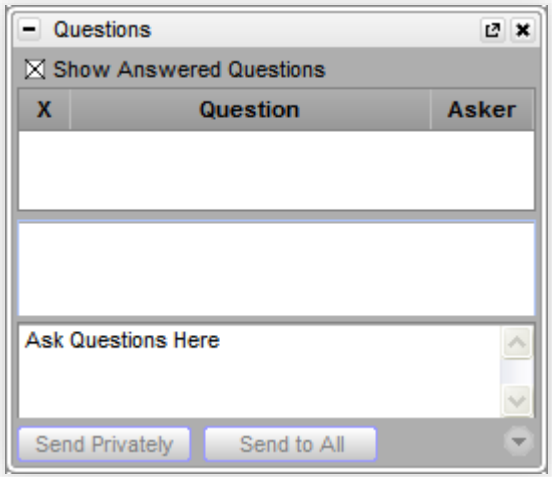
August 26, 2014

Greta Linse Peterson
Director of Product Management and Quality

GOLDEN HELIX
Accelerating the Quest for Significance™

Questions during the presentation

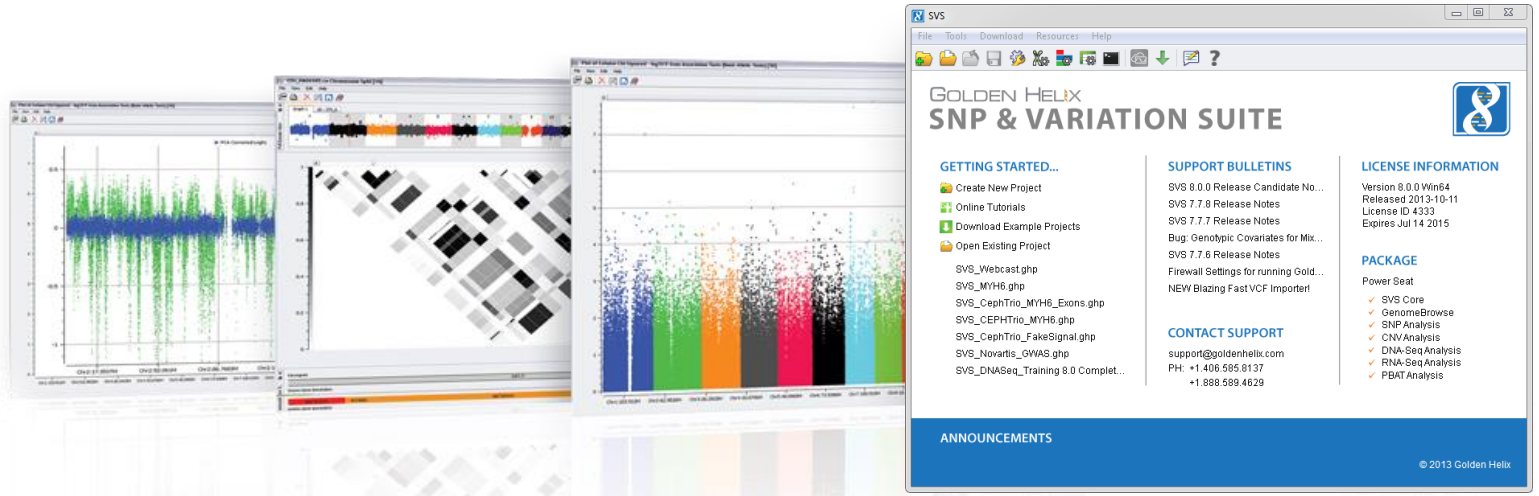
Use the Questions pane in your GoToWebinar window



Agenda

1. Overview
2. Four ways to use genomic prediction
3. Setting up a training and validation dataset
4. Highlights of GBLUP method
5. GBLUP versus Pedigree-based BLUP (ABLUP)
6. Demo
7. Conclusion

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
- Family Based Association

Overview

- **Genomic prediction uses:**
  - genetic information to predict the phenotype or trait for the individuals
  - Phenotypic (trait) data for a subset or all of the individuals.
  - The contribution of each genetic loci to build the model
  - A single mixed model regression equation to solve for:
    - The estimated breeding value (EBV) of individuals
    - The allele substitution effect (ASE) for genetic loci
- **Training and validation can be used to gauge the accuracy of the model**

Agenda

1. Overview
2. Four ways to use genomic prediction
3. Setting up a training and validation dataset
4. Highlights of GBLUP method
5. GBLUP versus Pedigree-based BLUP (ABLUP)
6. Demo
7. Conclusion

Case 1: Predict EBV for all individuals

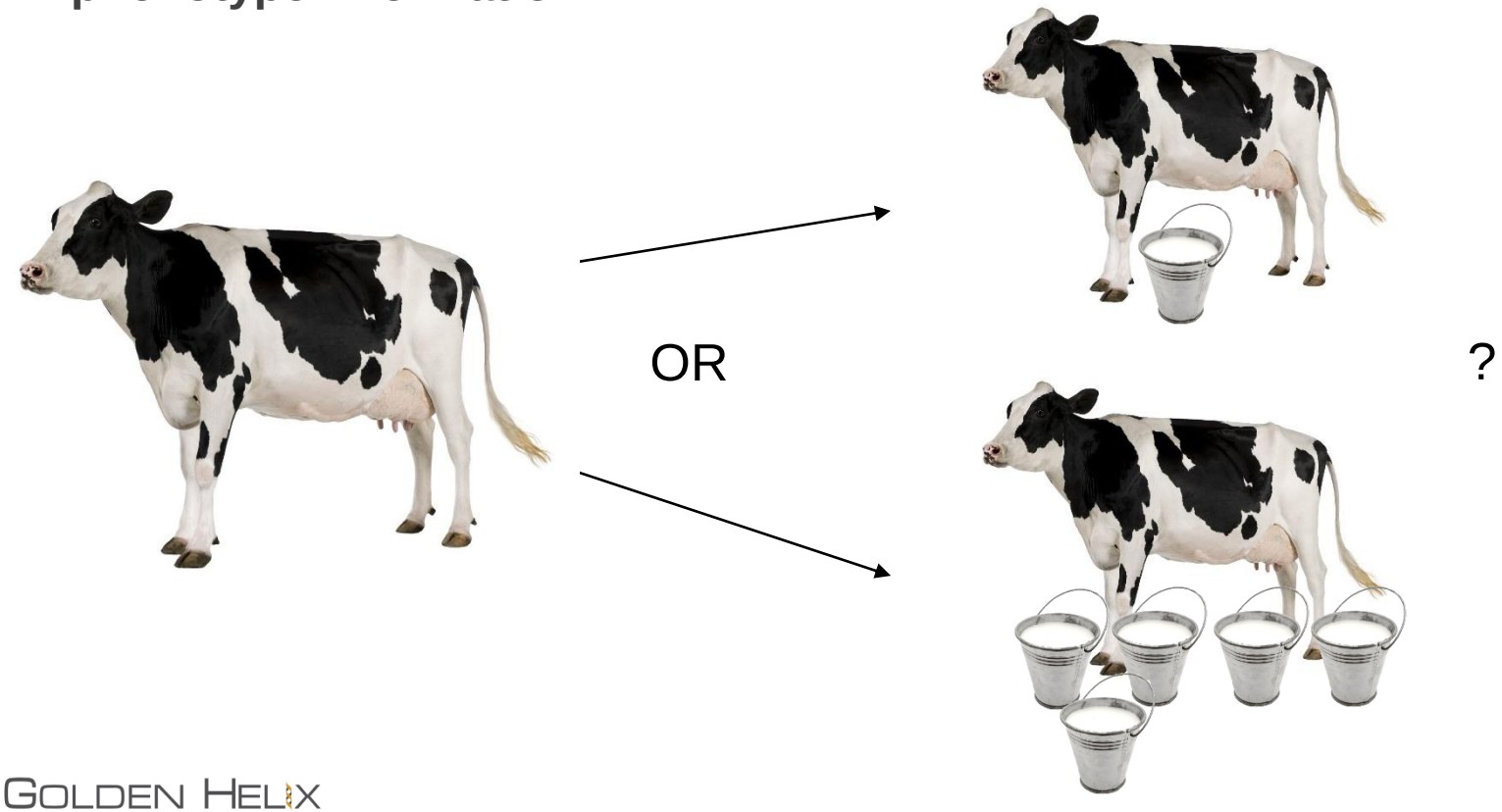
- **Use all individuals as the training set**
- **Identify individuals with the highest EBV to carry forward in breeding programs**

OR

?

Case 2: Predict EBV for a subset of individuals

- **Training set includes all individuals with known phenotype information**
- **Phenotype and EBV information is predicted for individuals missing phenotype information**



Case 3: Gauge accuracy of model using Training/Validation

- Randomly choose a subset of individuals to use to train the model
- Set the remaining individuals to have a missing phenotype (validation set)
- Build the model based on the training set and solve for the EBVs (random effects) and phenotypes for all individuals
- Compare the actual phenotypes to the predicted phenotypes or EBVs for the validation set

Case 4: Identify the loci that have the greatest effect on the model

- Use all individuals with phenotype data as the training set
- Examine the allele substitution effect of each loci
- Identify the loci with the greatest normalized ASE (allele substitution effect) and the most influential loci on the model to predict the phenotype or EBVs

Agenda

1. Overview
2. Four ways to use genomic prediction
3. Setting up a training and validation dataset
4. Highlights of GBLUP method
5. GBLUP versus Pedigree-based BLUP (ABLUP)
6. Demo
7. Conclusion

Definitions

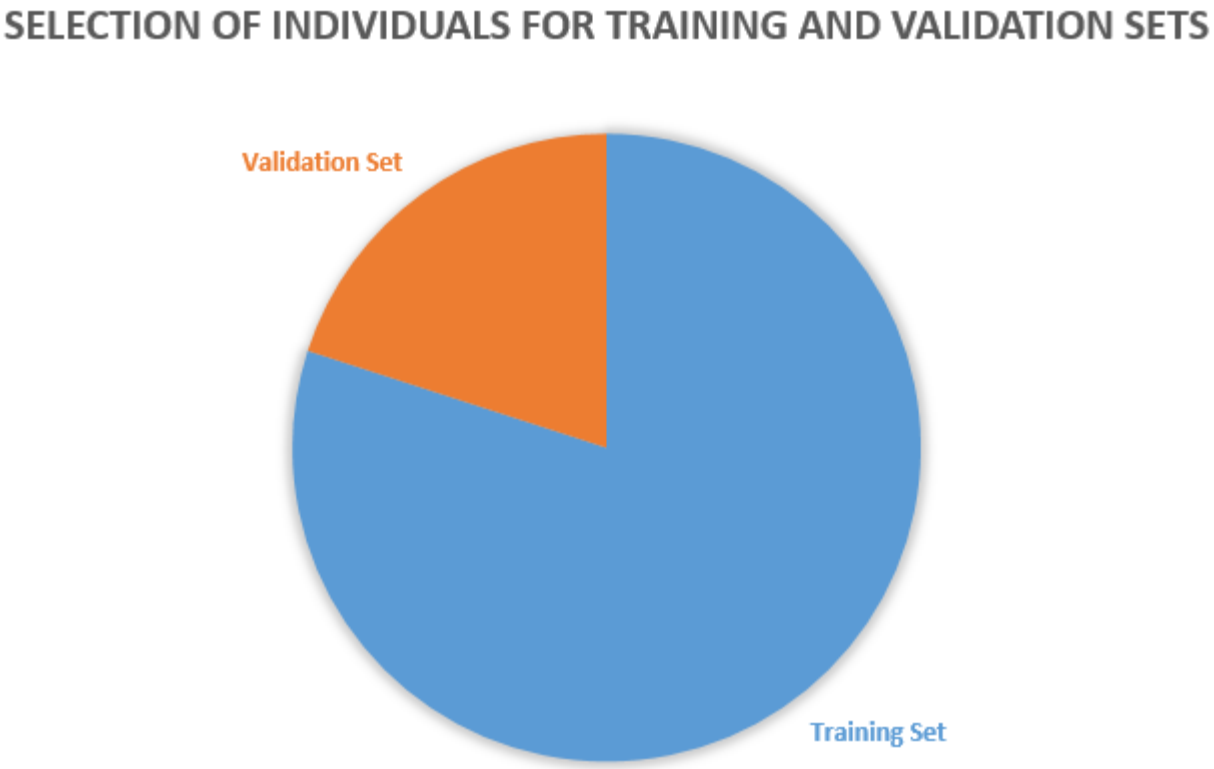
- **Training set:**
  - Subset of individuals used to compute the variance components and parameters of the linear mixed model using known phenotype information
- **Validation set:**
  - Subset of individuals used to predict the y value or phenotype values based on previously defined variance components and parameters of the linear mixed model.
  - Usually in this case the phenotype information is known for these individuals and can be compared against the predicted values.

Selecting individuals for Training/Validation Sets

- **Select the proportion of individuals to use for training:**
  - The larger the proportion of individuals in the training set vs the validation set the more accurate the predictions will be
- **Randomly choose the individuals for training**
- **The remaining individuals will be the validation set**
- **If using categorical covariates, try to select the same proportion from each category**

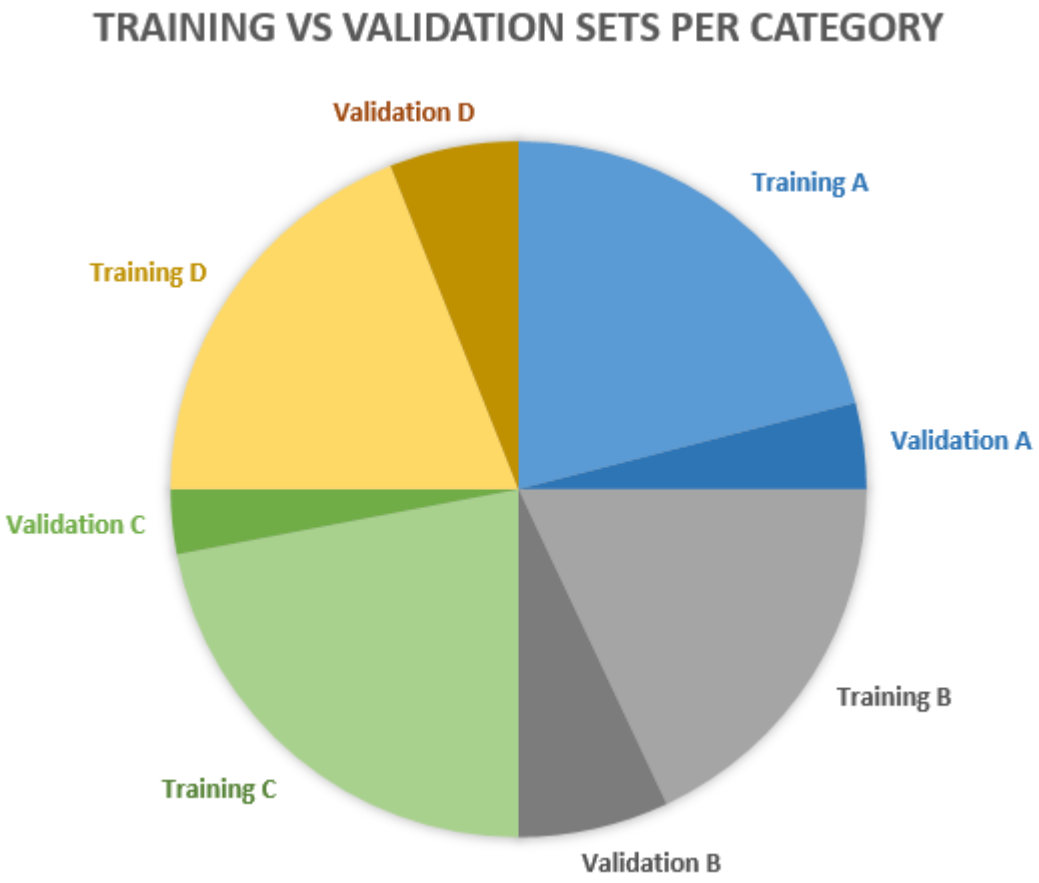
Example 1: No Covariates

- **Choose proportions to be 80% Training / 20% Validation**



Example 2: One Covariate (4 categories)

- **Choose proportions to be 80% Training / 20% Validation for each of the 4 categories**



Agenda

1. Overview
2. Four ways to use genomic prediction
3. Setting up a training and validation dataset
4. Highlights of GBLUP method
5. GBLUP versus Pedigree-based BLUP (ABLUP)
6. Demo
7. Conclusion

Highlights of GBLUP Method

- Formula
- Input Data
- Data Preparation
- Output of GBLUP

J. Dairy Sci. 91:4414–4423
doi:10.3168/jds.2007-0980

Efficient Methods to Compute Genomic Predictions

P. M. VanRaden[1]
Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350

ABSTRACT

Efficient methods for processing genomic data were developed to increase reliability of estimated breeding values and to estimate thousands of marker effects simultaneously. Algorithms were derived and computer programs tested with simulated data for 2,967 bulls and 50,000 markers distributed randomly across 30 chromosomes. Estimation of genomic inbreeding coefficients required accurate estimates of allele frequencies in the base population. Linear model predictions of breeding values were computed by 3 equivalent methods: 1) iteration for individual allele effects followed by summation across loci to obtain estimated breeding values, 2) selection index including a genomic relationship matrix, and 3) mixed model equations including the inverse of genomic relationships. A blend of first- and second-order Jacobi iteration using 2 separate relaxation factors converged well for allele frequencies and effects. Reliability of predicted net merit for young bulls was 63% compared with 32% using the traditional relationship matrix. Nonlinear predictions were also computed using iteration on data and nonlinear regression on marker deviations; an additional (about 3%) gain in reliability for young bulls increased average reliability to 66%. Computing times increased linearly with number of genotypes. Estimation of allele frequencies required 2 processor days, and genomic predictions required <1 d per trait, and traits were processed in parallel. Information from genotyping was equivalent to about 20 daughters with phenotypic records. Actual gains may differ because the simulation did not account for linkage disequilibrium in the base population or selection in subsequent generations.
Key words: genomic selection, mixed model, computer program, relationship matrix

INTRODUCTION

Genomic selection increases the rate of genetic improvement and reduces cost of progeny testing by allowing breeders to preselect animals that inherited chromosome segments of greater merit (Meuwissen et al., 2001; Schaeffer, 2006). Single nucleotide polymorphism (SNP) markers can now cover the genome with high density and are inexpensive to obtain. Evaluations based on SNP genotypes can be computed as soon as DNA can be obtained, which allows selection in both sexes early in life. Application of genomic selection to dairy cattle has just begun (de Roos et al., 2007; van der Beek, 2007; Guillaume et al., 2008). Potential methods and strategies were compared by Meuwissen (2007).

Computer algorithms and programs are needed to incorporate genomic data into genetic evaluations and to process the rapidly expanding numbers of SNP genotypes. Previous algorithms for including markers often fit effects individually rather than simultaneously or fit additional polygenic effects because marker coverage of the genome was not yet complete (de Roos et al., 2007). Iterative algorithms such as Gauss-Seidel and preconditioned conjugate gradient can be used to estimate allele effects (Legarra and Misztal, 2008), but fewer numerical problems may result from direct inversion of variance matrices or mixed model equations (Lee and van der Werf, 2006). Genomic relationships can be included in multitrait derivative-free REML programs (Zhang et al., 2007).

Objectives of this research were 1) to develop computer methods to include genomic data in predictions, 2) to apply the methods to simulated data for actual Holstein and Jersey pedigrees, and 3) to estimate gains in reliability from genomic predictions.

MATERIALS AND METHODS

Predictions were computed by linear and nonlinear systems of equations. The linear predictions assumed that all markers contributed equally to genetic variation (no major genes). The nonlinear (Bayesian) predictions assumed that the prior distribution of marker or QTL effects was not normal. Genetic variance may not be equal across chromosomes or markers because, for example, major genes may exist on some chromosomes. The data vector in both linear and nonlinear predictions was modeled as a linear function of the unknown effects, but solutions for the unknown effects in the nonlinear predictions were nonlinear functions of the data vector. Nonlinear predictions may be better than

Received December 31, 2007.
Accepted June 26, 2008.
[1]Corresponding author: paul.vanraden@ars.usda.gov

GBLUP Formula

- **Mixed Model Equation:**

$$y = X_f \beta_f + u + \epsilon$$

y is a $n \times 1$ vector of observed phenotypes for n individuals

X_f is a $n \times f$ matrix of fixed effects for f fixed effects

β_f is a $f \times 1$ vector of the coefficients of the fixed effects

u is a $n \times 1$ vector of the additive genetic merits (genomic breeding values)

ϵ is a $n \times 1$ vector of random errors

Where:

$u = M\alpha$ and we assume $E(\alpha) = 0$ and $Var(\alpha) = I\sigma_M^2$

M is a $n \times m$ matrix of minor allele counts per individual per (m) loci and α is a $n \times m$ vector of allele substitution effects per loci

GBLUP Genomic Relationship Matrix

- **Under the above assumptions:**

$$Var(u) = Var(M\alpha) = MVar(\alpha)M' = MM'\sigma_M^2$$

- **Under Hardy-Weinberg equilibrium the sum of the variances would be:**

$$\phi = 2\sum_{k=1}^{m} p_k q_k$$

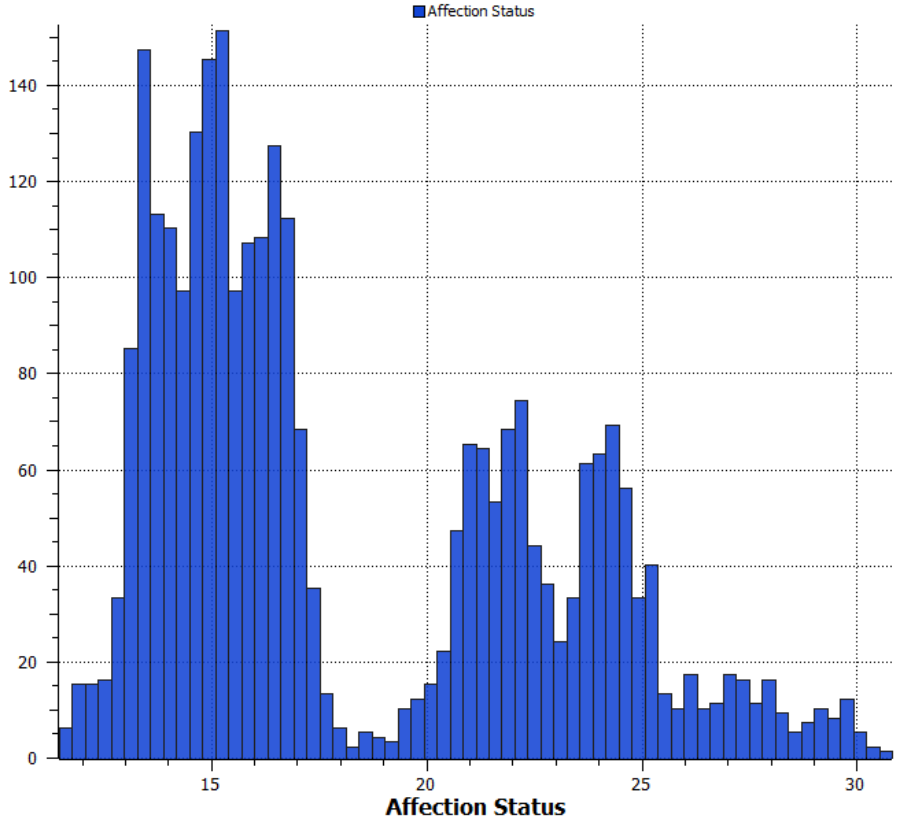
- **Thus giving the normalized variance matrix:**

$$G = \frac{MM'}{\phi}$$

- **We can then show that $Var(u) = \sigma_G^2 G$ where G is the GBLUP Genomic Relationship Matrix (a kinship matrix)**

GBLUP Input Data - Phenotype

- **Phenotype:**
  - At least two non-missing values per categorical covariate group



GBLUP Input Data – Genotype data

- **Genotype data:**
  - Formatted either in minor allele frequency counts (0,1,2) or genotypes (A_A, A_B, B_B)

	G 12	G 13	G 14	G 15	G 16	G 17	G 18	G 19	G
Patients	s16859.1	OARUn.2108_9811.1	s33378.1	s00430.1	s30808.1	s04373.1	OAR21_22786139.1	s69480.1	
AfricanDorper-ADP9	C_G	A_A	G_G	A_G	A_G	G_G	A_A	A_A	
AfricanDorper-ADP6	C_G	A_G	G_G	A_G	G_G	A_G	A_A	A_A	
AfricanDorper-ADP21	G_G	A_G	A_G	A_G	G_G	G_G	A_A	A_G	
AfricanDorper-ADP13	C_G	A_G	G_G	A_G	G_G	A_A	A_C	A_A	
AfricanDorper-ADP5	C_G	A_G	G_G	A_A	A_G	A_G	A_C	A_A	
AfricanDorper-ADP20	C_G	A_G	A_G	G_G	A_G	G_G	A_C	A_G	
AfricanDorper-ADP19	G_G	A_G	G_G	G_G	A_G	G_G	A_A	A_G	
AfricanDorper-ADP11	C_C	A_G	G_G	A_G	A_G	G_G	A_C	A_G	
AfricanDorper-ADP3	C_G	A_A	A_G						
AfricanDorper-ADP25	C_G	A_G	G_G						
AfricanDorper-ADP18	C_G	A_G	A_G						
AfricanDorper-ADP24	G_G	G_G	G_G						
AfricanDorper-ADP17	G_G	A_G	G_G						
AfricanDorper-ADP16	C_C	A_A	A_G						
AfricanDorper-ADP23	G_G	A_G	G_G						
AfricanDorper-ADP15	C_G	A_G	G_G						
AfricanDorper-ADP7	G_G	A_G	G_G						
AfricanDorper-ADP22	C_C	A_G	G_G						
AfricanDorper-ADP8	G_G	A_A	A_G						
AfricanDorper-ADP12	G_G	A_G	G_G						
AfricanDorper-ADP14	C_C	G_G	G_G						
AfricanWhiteDorper-AWD6	C_G	G_G	A_G						
AfricanWhiteDorper-AWD2	G_G	G_G	A_G						
AfricanWhiteDorper-AWD1	G_G	G_G	G_G						
AfricanWhiteDorper-AWD3	G_G	G_G	G_G						
AfricanWhiteDorper-AWD4	C_G	G_G	A_G						
AfricanWhiteDorper-AWD5	G_G	G_G	A_G	A_A	A_A	G_G	C_C	A_A	

	1	2	3	4	5	6	7	
gid	wPt.0538	wPt.8463	wPt.6348	wPt.9992	wPt.2838	wPt.8266	wPt.1100	
32	2	2	0	2	2	0	2	
28	2	2	2	2	2	0	2	
231	0	2	2	2	2	0	2	
127	2	2	2	2	2	0	2	
17	2	2	0	2	2	0	2	
6	2	2	2	2	2	0	2	
278	0	2	0	0	0	0	2	
367	0	2	0	2	2	0	2	
461	2	2	2	2	2	0	2	
101	0	2	2	2	2	0	2	
20	2	2	2	2	2	0	0	
239	2	2	2	2	2	0	2	
18	0	2	2	2	2	0	2	

GBLUP Input Data – Genetic Position Information

- **Chromosome & position information needed to identify non-autosomal loci**

Name	Chromosome	StartPos	StopPos	Array	Type	B1	B2	B3	Note
s42208	1	706835	706835	SNP50	SNP	0.996677741	.	.	OARv3.1:OAI
s64747	1	748143	748143	SNP50	SNP	1	.	.	OARv3.1:OAI
s68493	1	785434	785434	SNP50	SNP	0.993355482	.	.	OARv3.1:OAI
OAR1_420114	1	792698	792698	SNP50	SNP	0.996677741	.	.	OARv3.1:OAI
OAR1_537224_X	1	912507	912507	SNP50	SNP	1	.	.	OARv3.1:OAI
s43636	1	954073	954073	SNP50	SNP	0.995726496	.	.	OARv3.1:OAI
s35460	1	999877	999877	SNP50	SNP	0.996677741	.	.	OARv3.1:OAI
s48804	1	1155400	1155400	SNP50	SNP	1	.	.	OARv3.1:OAI
s41127	1	1180263	1180263	SNP50	SNP	0.991735537	.	.	OARv3.1:OAI
s18466	1	1193562	1193562	SNP50	SNP	0.991735537	.	.	OARv3.1:OAI
s40172	1	1245222	1245222	SNP50	SNP	0.991735537	.	.	OARv3.1:OAI
s46291	1	1360820	1360820	SNP50	SNP	0.986440678	.	.	OARv3.1:OAI
s26718	1	1390945	1390945	SNP50	SNP	1	.	.	OARv3.1:OAI
s31488	1	1406764	1406764	SNP50	SNP	0.987341772	.	.	OARv3.1:OAI
s46222	1	1513820	1513820	SNP50	SNP	0.991735537	.	.	OARv3.1:OAI
s00523	1	1563484	1563484	SNP50	SNP	0.996688742	.	.	OARv3.1:OAI
s38369	1	1618342	1618342	SNP50	SNP	0.993355482	.	.	OARv3.1:OAI
s09524	1	1675087	1675087	SNP50	SNP	1	.	.	OARv3.1:OAI
s43961	1	1738285	1738285	SNP50	SNP	0.973421927	.	.	OARv3.1:OAI
s22577	1	1815740	1815740	SNP50	SNP	0.981818182	.	.	OARv3.1:OAI

Compute GRM

- **Filter genetic data to remove:**
  - Non-autosomal loci
  - Loci with minor allele frequency < 0.05
  - Loci in Linkage-Disequilibrium
  - Loci with a poor call rate (e.g. < 0.85)

$$\begin{bmatrix} A\_A & \cdots & A\_B \\ \vdots & \ddots & \vdots \\ B\_B & \cdots & B\_B \end{bmatrix}_{n\times m} \rightarrow \begin{bmatrix} 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 2 & \cdots & 2 \end{bmatrix}_{n\times m} \rightarrow \begin{bmatrix} 1.01 & \cdots & 0.027 \\ \vdots & \ddots & \vdots \\ 0.027 & \cdots & 0.998 \end{bmatrix}_{n\times n} = GRM$$

Output of GBLUP

- **Per individual Genomic Estimated Breeding Values (Sample-wise random effects)**
- **Per marker allele substitution effects**
- **Pseudo-heritability** $ph = \hat{\sigma}_G^2 / Var(y)$
- **P-value of the model** $P\left(X > \left(-2(l_0 - l_1)\right)\right), X \sim \chi_1^2$
- **Genetic component of variance** $V_g(\hat{\sigma}_G^2)$
- **Error component of variance** $V_e(\hat{\sigma}_e^2)$

GBLUP versus Pedigree-based BLUP (ABLUP)

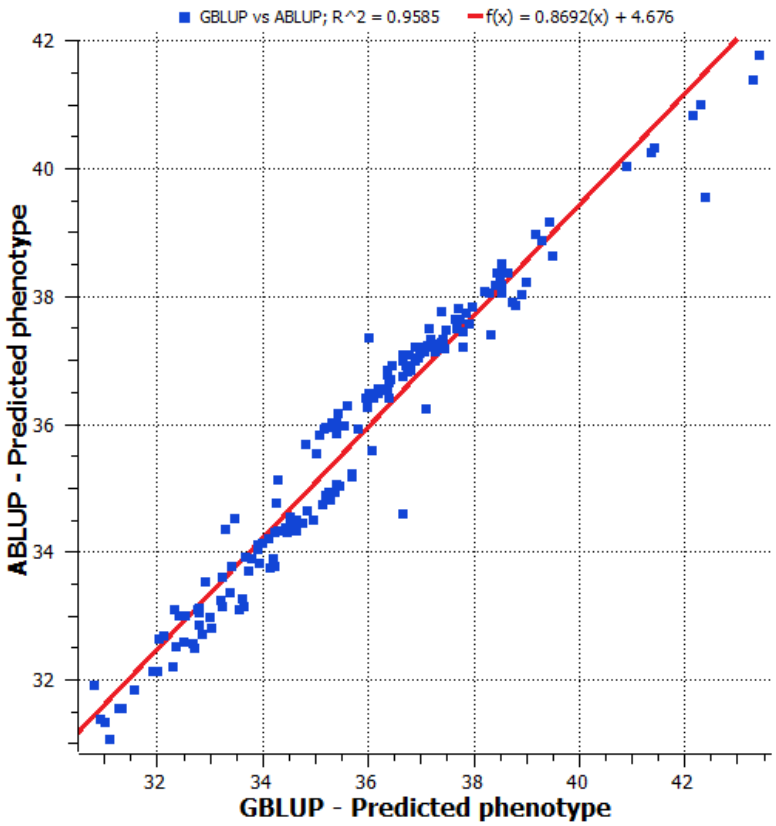
GBLUP

- Uses *genomic* information to infer the relationships between individuals
- Can make predictions without knowing pedigree structure
- Can deal with population sub-groups without needing to perform meta-analysis

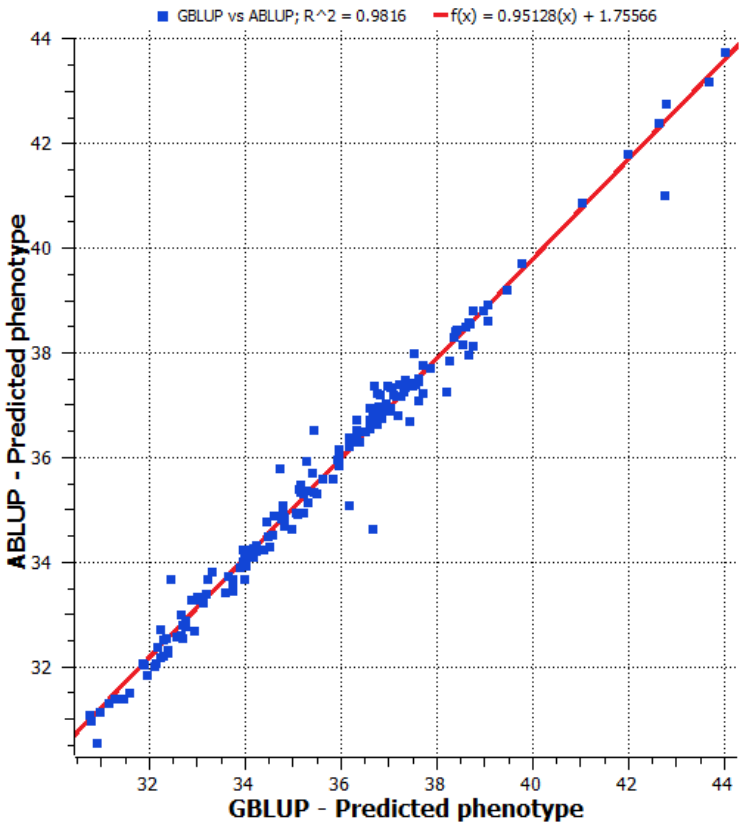
ABLUP

- Uses *pedigree* structure to explicitly define the relationships between individuals
- Can be more accurate if the pedigree information is known for all individuals
- Can be more accurate if within a family the degrees of relatedness are fairly high

GBLUP vs ABLUP Phenotype Predictions for small Pedigrees



All phenotypes known



Training & Validation (80 / 20)

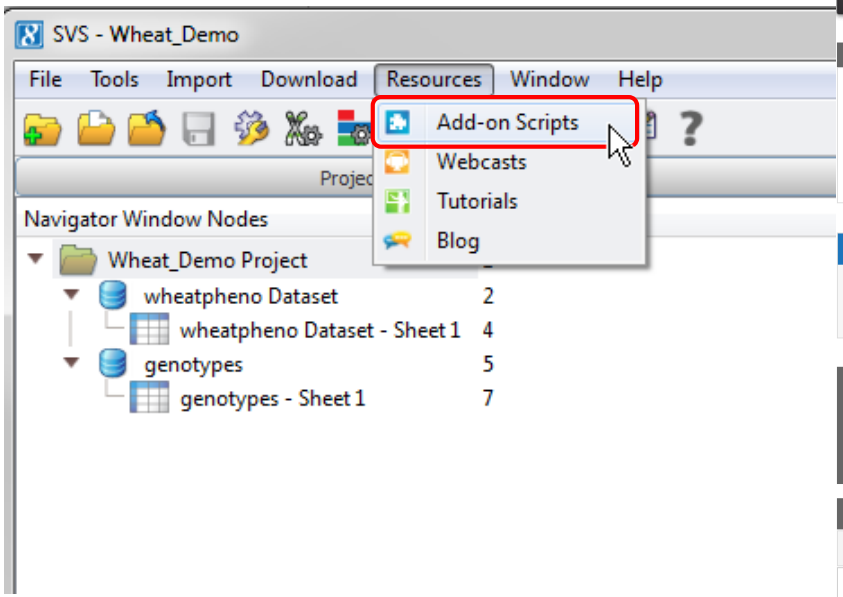
Demonstration

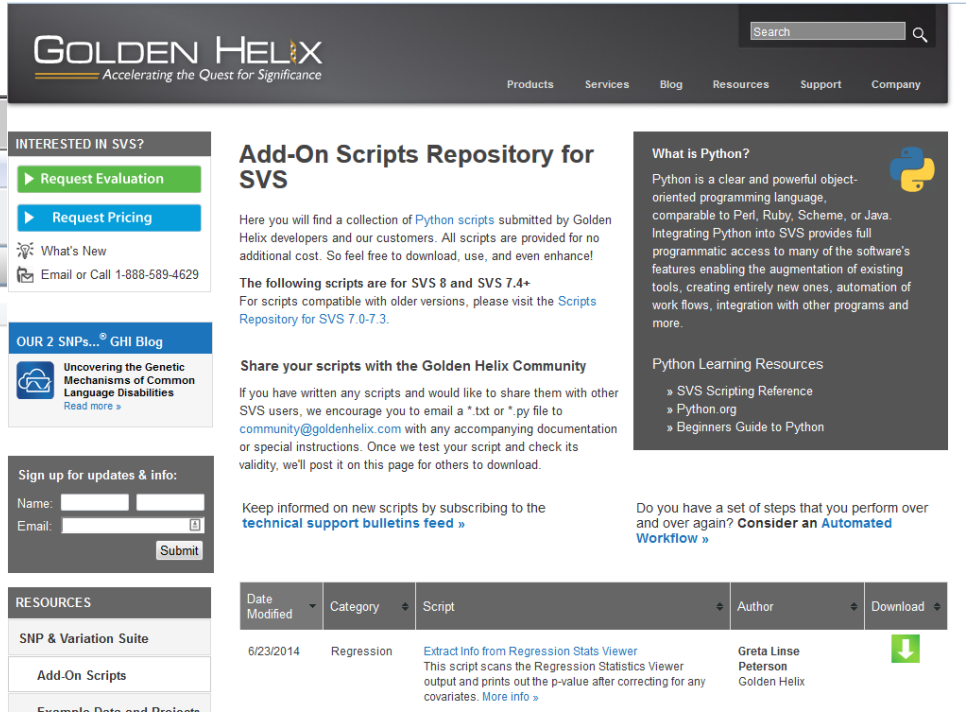
SNP & VARIATION SUITE

[DEMONSTRATION]

Add-On Scripts Used in the Demo

- **Select Random Subset by Category**
- **Create Pseudo Marker Mapped Spreadsheet**





www.goldenhelix.com/SNP_Variation/scripts/index.html

Conclusion

- **Genomic prediction using GBLUP can provide**
  - The Estimated Breeding Value
  - Influential Loci for the phenotype
- **Genomic prediction can help breeders and researchers make decisions**
  - Which animals are likely to pass on their desirable traits
  - Which loci could be used for a targeted assay for diagnostic purposes
- **While other tools are available for Genomic Prediction, SVS combines**
  - Data management,
  - Genomic prediction, and
  - Visualization

  in one powerful package

Future Improvements

- New genomic prediction methods including Bayes C & Bayes Cπ
- Easier expansion/application of trained models on new datasets
- Ability to revise models with new information

- Have a request? Let us know!

Data Obtained From:

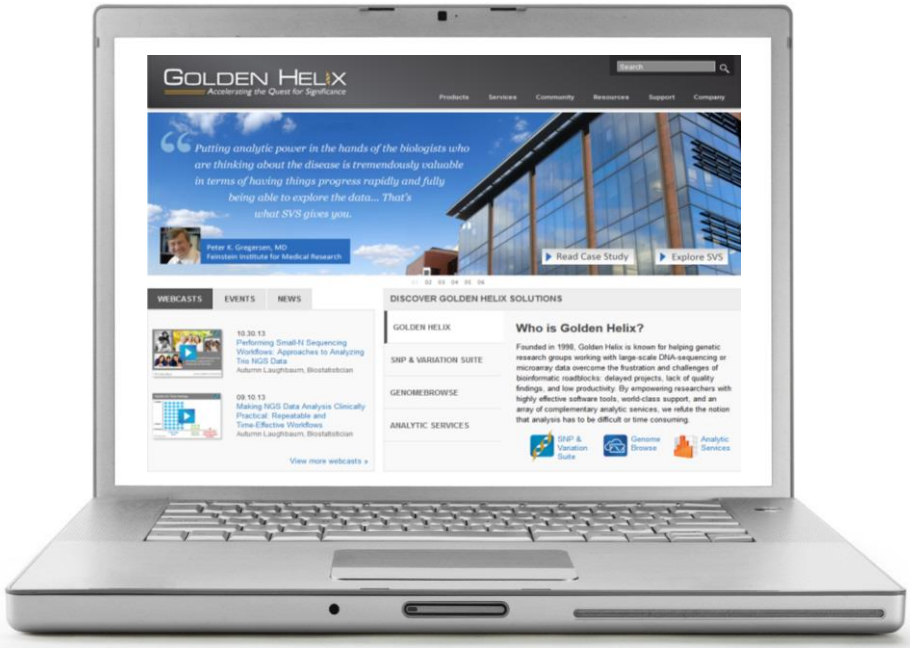
- **International Sheep Genomics Consortium (www.sheephapmap.org)**
  - Provided access to the Sheep HapMap SNP 50k data on request
- **data(wheat) from library(BLR) in R [Pérez, 2010]**

References

- Isik, F. (2013, Aug 1). 'Genomic Relationships and GBLUP' [Webinar]. In *Extension America's Research-based Learning Network*. Retrieved from https://www.extension.org/pages/68019/genomic-relationships-and-gblup#.U_ZRmPldX4a
- Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K, Paiva S, Barendse W, Ciani E, Raadsma H, McEwan J, Dalrymple B; International Sheep Genomics Consortium Members. 'Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection.' *PLoS Biol.* 2012 Feb;10(2):e1001258. doi: 10.1371/journal.pbio.1001258. Epub 2012 Feb 7
- Pérez, P., de los Campos, G., Crossa, J., Gianola, D. (2010) 'Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R' [BLR]. *Plant Genome*, 3(2): 106-116. doi:10.3835/plantgenome2010.04.0005.
- Taylor, J.F. (2013) 'Implementation and accuracy of genomic selection', *Aquaculture*, http://dx.doi.org/10.1016/j.aquaculture.2013.02.017
- VanRaden, P.M. (2008) 'Efficient Methods to Compute Genomic Predictions', *J. Dairy Sci*, 91, pp. 4414–4423.
- Zhang, L., Liu J., Zhao F., Ren H., Xu L., et al. (2013) 'Genome-Wide Association Studies for Growth and Meat Production Traits in Sheep', *PLoS ONE* 8(6):e66569. doi:10.1371/journal.pone.0066569.

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