PLANT BREEDING & GENOMICS

Genomic Relationships and GBLUP

Presented by

Fikret Isik

Associate Professor

North Carolina State University

Hosted by Shawn Yarnes

Plant Breeding and Genomics

extension

Genomic Relationships and GBLUP

Fikret Isik

Associate Professor

North Carolina State University

Outline

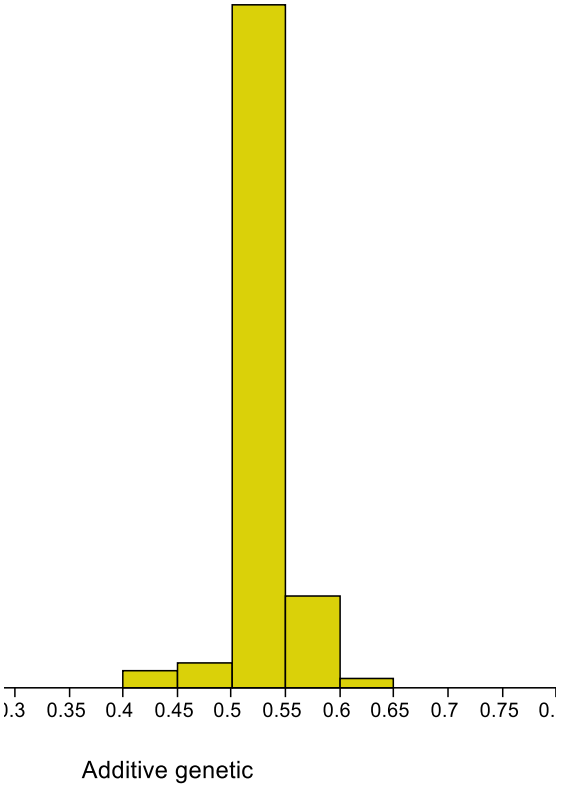
1. Introduction (3)
2. Matrices needed to calculate the G matrix (6)
3. Methods to calculate the G matrix (8), [DEMO]
4. ABLUP (5)
5. GBLUP (5), [DEMO]
6. Conclusions (3)

Total to cover 30 slides out of 44!

Introduction

Average genetic relationships

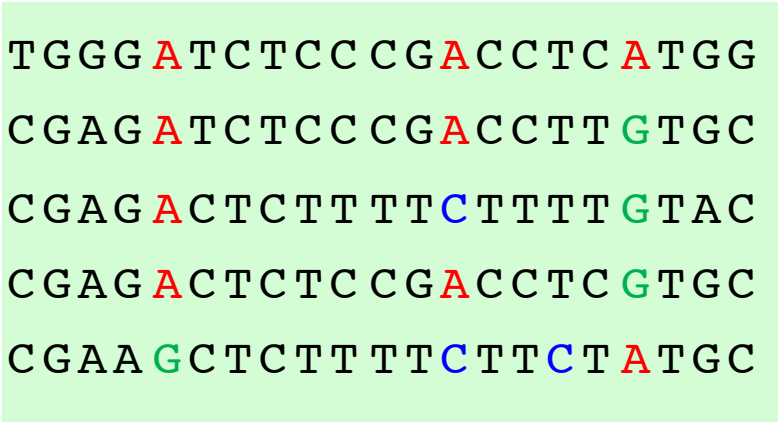
- Probabilities generated from pedigree (**A** matrix) are discrete between close relatives
- For example, we assume that full-sibs share 0.5 of alleles (genome) that are IBD



Markers to estimate similarities

- Genetic markers across the genome can be used to measure genetic similarities and
- may be more precise than pedigree information

(vanRaden 2008)



Shared genome

- Markers estimate *proportion of chromosome segments shared by individuals* including identification of genes identical by state (IBS)

```
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTTCTTCTATGC
```

(vanRaden 2008)

Marker matrices needed to calculate the G matrix

Genotypes and Gene Content

Let's assume that we have 3 diploid individuals and 4 loci. The lower case letters represent the minor (less frequent) alleles at each locus.

Genotypes

	snp1	snp2	snp3	snp4
Ind1	AA	Ct	GG	Ag
Ind2	AA	Ct	Ga	AA
Ind3	tt	CC	GG	AA

The genotypes above are converted to **gene content** (counts of minor allele) as follows. Let's call it the **MAF matrix.**

	snp1	snp2	snp3	snp4
Ind1	0	1	0	1
Ind2	0	1	1	0
Ind3	2	0	0	0

M Matrix

The deviations of 1 from gene content are obtained (generating scores of 1, 0, and -1) for ease of subsequent calculations

	snp1	snp2	snp3	snp4
ind1	-1	0	-1	0
ind2	-1	0	0	-1
ind3	1	-1	-1	-1

With the data formatted, we are ready to compute a matrix of realized genetic similarities among all pairs of individuals (G matrix)

```
# R script
> MAF
     [,1] [,2] [,3] [,4]
[1,]    0    1    0    1
[2,]    0    1    1    0
[3,]    2    0    0    0
> M=MAF-1
> M
     [,1] [,2] [,3] [,4]
[1,]   -1    0   -1    0
[2,]   -1    0    0   -1
[3,]    1   -1   -1   -1
```

(VanRaden, 2008, Forni et al. 2011)

MM' Matrix

The product of **M** matrix with its transpose **M´** is **MM'** matrix

	snp1	snp2	snp3
ind1	2	1	0
ind2	1	2	0
ind3	0	0	4

- Diagonal elements: Counts the # of homozygous loci for each individual. First individual (row 1) has 2 homozygous loci, second individual has 2, third has 4 homozygous loci
- Off-diagonal elements: Measure the # of alleles shared by relatives

(VanRaden, 2008, Forni et al. 2011)

M'M matrix

The product of **M'** matrix with **M** is **M'M** matrix

	snp1	snp2	snp3	snp4
ind1	3	-1	0	0
ind2	-1	1	1	1
ind3	0	1	2	1
ind3	0	1	1	2

- *Diagonals*: Counts the # of homozygous individuals for each locus
  Locus1 has 3 homozygous individuals
  Locus2 has 1 homozygous individual etc..
- *Off-diagonal elements*: Measures the # of times alleles at different loci were inherited by the same individual

(VanRaden, 2008, Forni et al. 2011)

P Matrix

We also need the P matrix

- The columns of **P** are allele frequencies expressed as $P_i = 2(p_i - 0.5)$, where p_i is the MAF of locus i

Example: Let MAF of four loci are

$p_1 = 0.383$, $p_2 = 0.244$, $p_3 = 0.167$, $p_4 = 0.067$

Then the elements of P matrix are $P_i = 2(p_i - 0.5)$,

$$\mathbf{P} = \begin{array}{c|cccc} & \text{snp1} & \text{snp2} & \text{snp3} & \text{snp4} \\ \text{ind1} & -0.234 & -0.512 & -0.666 & -0.866 \\ \text{ind2} & -0.234 & -0.512 & -0.666 & -0.866 \\ \text{ind3} & -0.234 & -0.512 & -0.666 & -0.866 \end{array}$$

The Z matrix

$$\mathbf{Z = M - P =} \begin{vmatrix} -0.766 & 0.512 & -0.334 & 0.866 \\ -0.766 & 0.512 & 0.666 & -0.134 \\ 1.234 & -0.488 & -0.334 & -0.134 \end{vmatrix}$$

- Sets means values of the allele effects to 0
- Subtraction of **P** gives *more credit to rare alleles* than to common alleles when calculating genomic relationships
- Genomic inbreeding coefficient (F) is greater if the individual is homozygous for rare alleles than if homozygous for common alleles

(VanRaden, 2008, Forni et al. 2011)

Methods to calculate genomic relationships (G matrix)

GOF (1)

$$\mathbf{G} = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$$

- Derived from observed allele frequencies
- **Z** is incidence matrix for markers
- The denominator scales the **G** to be similar to the **A** matrix
- p_i are the observed MAF of all genotyped individuals regardless of inbreeding and selection

(VanRaden 2008)

GD (2)

$$\mathbf{GD} = \boldsymbol{ZDZ'}$$

- A variation of GOF
- Markers are weighted by reciprocals (D) of their expected variance

- Where **D** is diagonal matrix with elements

$$D_{ii} = \frac{1}{m[2p_i(1-p_i)]}$$

(Amin et al., 2007, Leutenegger et al., 2003)

G05 (3)

- When MAF in the base population is unknown 0.5 is used for all values of p_i

GMF (4)

- MAF set to mean of observed
- When MAF in the base population is unknown average MAF of genotyped population is used to calculate p_i

(VanRaden 2008)

Greg (regression method)

$$\mathbf{MM}' = g_0\mathbf{11}' + g_1\mathbf{A} + \mathbf{E}$$

- *g0* is the intercept, *g1* is the slope
- ***E*** includes differences of true from expected fraction of DNA in common, plus measurement error to account for markers being a subset of the DNA

Solving for **A** results in substituting **G** for **A**

$$G = \frac{\boldsymbol{MM}' - g_0\mathbf{11}'}{g_1}$$

(VanRaden 2008)

GN (normalized method)

$$\mathbf{GN} = \frac{\mathbf{ZZ'}}{\{trace[\mathbf{ZZ'}]\}/_{n}}$$

- ZZ' is weighted by its trace
- This assures compatibility with **A** when the mean inbreeding or the # of generations is low
- Higher levels of inbreeding can be accommodated by substituting *n* (dimensions of Z) with *1+F*
- Diagonals can be less than 1

(Forni et al. 2011)

Problems with the Inverse of G

- The genomic relationship matrix is positive semidefinite but it can be singular if
  - Number of loci is limited
  - Two subjects have identical genotypes across all markers
  - # of markers is smaller than the # of individuals genotyped

Weighted G matrix

- To avoid potential problems G can be weighted as

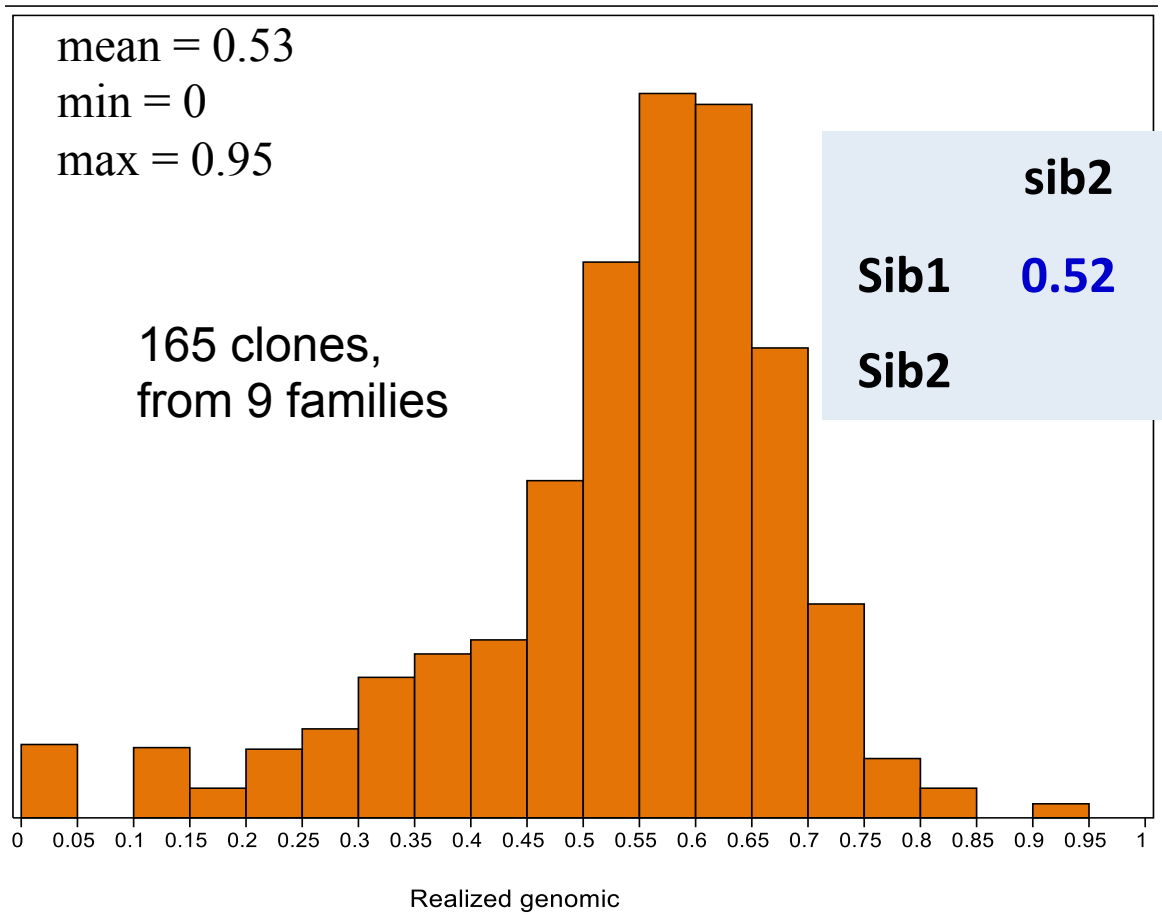
$$\mathbf{G} = w\,\mathbf{Gr} + (1 - w)\mathbf{A}$$

  - **Gr** is unweighted genomic relationship matrix
  - **A** is numerator relationship matrix among only genotyped animals
  - **w** is weight. This value is not critical between values of 0.95 and 0.98 (Aguilar et al. 2010)

DEMO

Calculation of genomic relationships (G matrix)

Realized genomic relationships



	sib2	sib3
Sib1	0.52	0.41
Sib2		0.61

Traditional genetic evaluation

ABLUP

Linear Mixed Model (ABLUP)

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

- **y** vector of observations
- **X** and **Z** are incidence matrices
- **b** vector of fixed factors
- **u** vector of random (genetic) factors ~ N (0, $\mathbf{A}\sigma^2_A$)
- **e** vector of residuals ~ N (0, $\mathbf{I}\sigma^2_e$),

Main assumptions (ABLUP)

$E[\mathbf{u}]=[\mathbf{e}]=0$

$Cov(\mathbf{u},\mathbf{e}) = 0$

$Var(\mathbf{u}) = \mathbf{A}\sigma^2_A = \boldsymbol{G}$

$Var(\mathbf{e}) = \mathbf{I}\sigma^2_e = \mathbf{R}$

$Var(\mathbf{y}) = \mathbf{ZGZ'} + \mathbf{R} = \mathbf{V}$

(Lynch and Walsh 1998)

Mixed Model Equations (ABLUP)

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

$\lambda = \sigma^2_e / \sigma^2_A = (1 - h^2 / h^2)$

λ: shrinkage factor
h^2: narrow-sense heritability

(Henderson 1984, Mrode 2005)

Mendelian Segregation Effect (m)

- When gametes are produced (by meiosis) allele pairs separate, leaving each cell with a single allele
- Sampling of parental alleles is random at each locus during meiosis (Mendel's law of segregation)
- Each progeny receives 50% of parents' DNA

Mendelian Segregation Effect (cont.)

- Estimation of Mendelian sampling effect requires progeny phenotype
- Or markers to provide such information on which allele at a QTL was transmitted

$$y_i = 0.5\,(u_j + u_k) + \mathbf{m_i} + e$$

Where u_j and u_k are parental contribution to individual i, m_i is the Mendelian term

Genomic BLUP

GBLUP

- GBLUP is relatively easy and does not involve anything that we are not familiar with ABLUP
- All we need to do is substitute the inverse of **A** matrix (Ainv) with the inverse of **G** matrix (Ginv) to predict breeding values

(VanRaden 2008)

GBLUP (cont.)

$$\mathbf{y = Xb + Zu + e}$$

- **Z** is incidence matrix for marker effects
- **u** is vector of additive genetics effects that correspond to allele substitution effects for each marker
- We let the sum **Zu** across all marker loci (m) to be equal to the vector of breeding values **Za = u**

MM Equations (GBLUP)

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

$$\text{EBV}(\hat{\mathbf{u}}) = \mathbf{G}\,[\mathbf{G} + \mathbf{R}\,\lambda]^{-1}\,(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

Lambda is defined as the sum across loci ($2\Sigma p_i 1\text{-}p_i$) times the ratio of error and additive genetic variance.

Accuracy of GEBV

$$\mathbf{G}\left[\mathbf{G}+\mathbf{R}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\mathbf{G}$$

For individuals with observations

$$\mathbf{C}\left[\mathbf{G}+\mathbf{R}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\mathbf{C}'$$

For individuals without observations

C represents the genomic covariance matrix between individuals with and without observations

$$\mathbf{C}=\frac{\mathbf{Z_n Z'}}{2\Sigma p_i(1-p_i)}$$

Fitting GBLUP using ASReml

```
!ARGS 1 2 !rename 1
Title: Asreml code for GBLUP
 tree   !P
 female !P       male   !P
 series !I  site !I     rep !I    row !I   col !I
 height volume !/10

C165pedmatrix.csv !SKIP 1 !ALPHA !SORT #pedigree
Ginv.giv   #IT MUST FOLLOW THIS ORDER
data.csv  !SKIP 1 !DOPART $1    #data file

!PART 2  # GBLUP
volume  ~ mu site  !r tree  # model
1 1 1
0 0 IDEN !S2==14.7
tree 1
tree 0 GIV 7.9 !GF
```

DEMO

Genomic BLUP

Conclusions

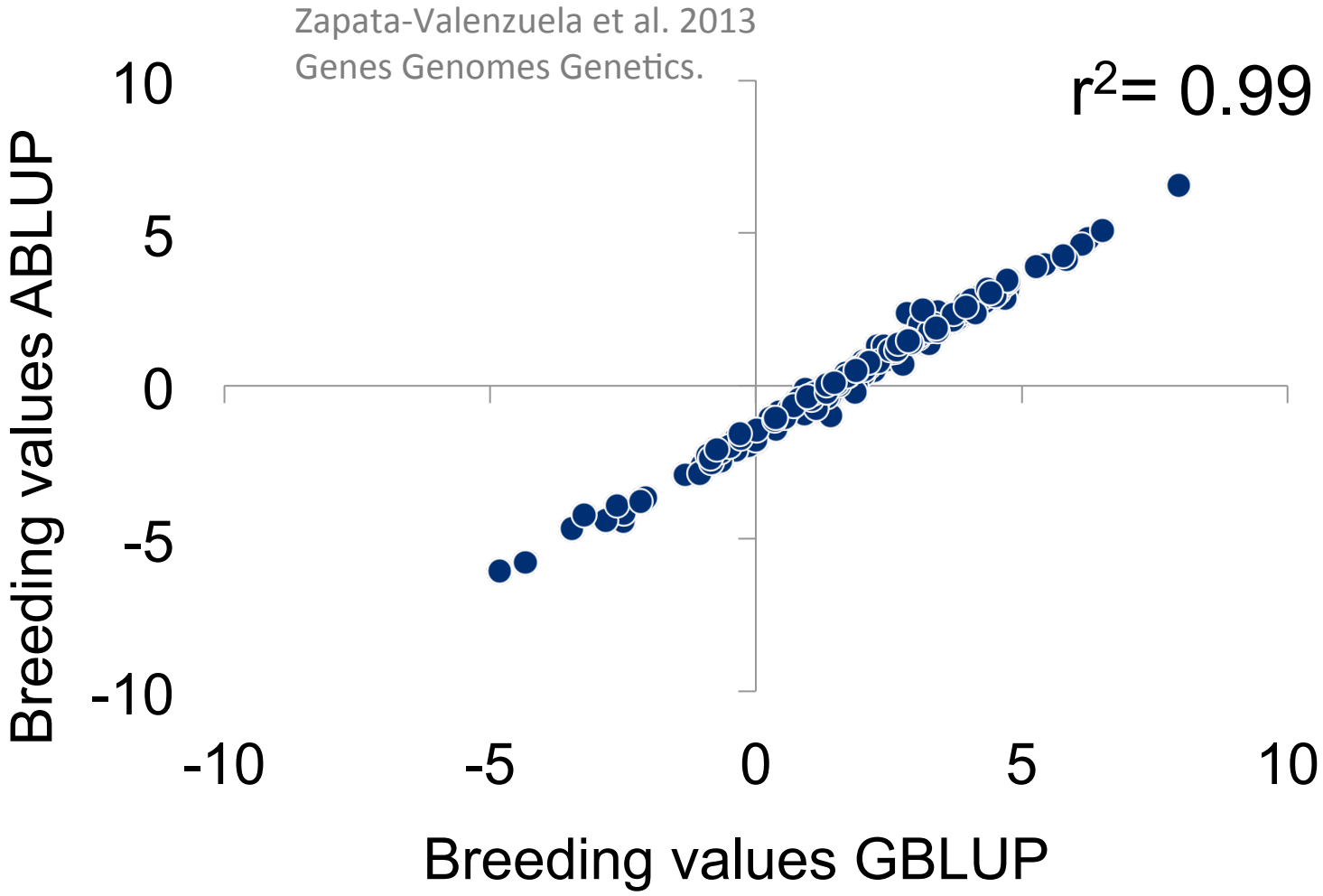
Accuracies of the predictions

Accuracies of predictions from markers (GBLUP) are higher than accuracies of predictions from pedigree based models (ABLUP)

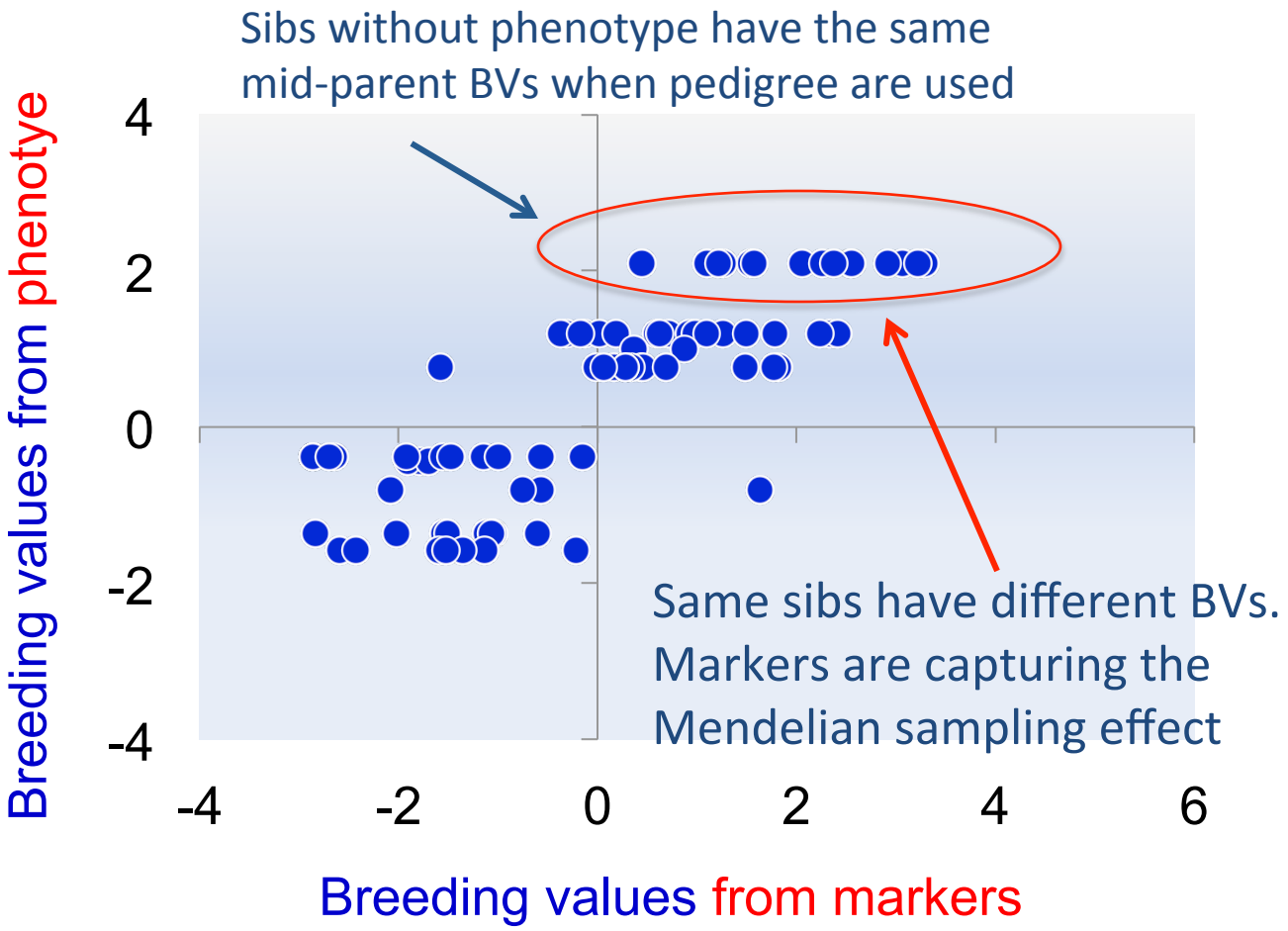
Training / validation	r(ABLUP)	r(GBLUP)
84 / 81	**0.60**	**0.71**
148 / 17	**0.61**	**0.76**

Zapata-Valenzuela et al. 2013 Genes Genomes Genetics.

Correlation between predictions



Predictions without phenotype



Zapata-Valenzuela et al. 2013 Genes Genomes Genetics.

Acknowledgement

- Christian Maltecca (NC State U)
- Jim Holland (NC State U)
- Ross Whetten (NC State U)
- Jaime Zapat Valenzuela (BioForest SA, Chile)
- Funda Ogut (NC State U)
- NC State University Tree Improvement Program

References

- Forni, S., Aguilar, I., & Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution, 43(1).
- Henderson, C. R. (1984). Linear models in animal breeding.
- Lynch, M., & Walsh, B. (1998). Genetics and analysis of quantitative traits.
- Mrode, R. A., & Thompson, R. (2005). Linear models for the prediction of animal breeding values. Cabi, UK.
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. Journal of Dairy Science, 91(11), 4414-4423.

PLANT BREEDING & GENOMICS

Please fill out the survey evaluation.
You will be contacted via email.

Today's Presentation Available
http://www.extension.org/pages/68019

Sign up for PBG News
http://pbgworks.org

Sign up for Future Webinars and View Archive
http://www.extension.org/pages/60426

extension