Conifer Translational Genomics Network
Coordinated Agricultural Project

Genomics in Tree Breeding and Forest Ecosystem Management

Module 14 – Using Markers to Predict Breeding Values

Fikret Isik, North Carolina State University

Module outline

1. Prediction of breeding values based on the genetic relationship matrix derived from pedigrees (**A** matrix)
2. Marker aided selection
3. Imputing missing genotypes
4. Realized genetic covariance (G matrix) for G-BLUP

General combining ability (GCA)

General combining ability (GCA) is the deviation of the parent mean ($\bar{X}$) from the population mean (μ):

$$GCA = \bar{X} - \mu$$

We can use a linear model to define GCA of a parent tree

$$y_i = \mu + GCA_f + e_i$$

The GCA of a tree is 1/2 of the parental additive breeding value (a_f)

$$y_i = \mu + 0.5\, a_f + e_i$$

Where a_f stands for the breeding value of the female parent

Breeding value and genetic value

Breeding value (BV): The value of genes transmitted to progeny for a given parent.

We can define BV of an individual in a linear model as

$$y_i = \mu + 0.5\,(a_m + a_f) + e_i$$

Genetic value (GV) of an individual in a linear model:

$$y_{ij} = \mu + 0.5\,(a_m + a_f) + m_i + e_{ij}$$

Linear models for predicting BVs

• Parental models (GCA model): may be used to estimate parental breeding values: $y_i = \mu + GCA_i + e_i$

• Individual-tree models (BV model): may be used to estimate both parental and progeny BVs in one analysis: $y_i = \mu + a_i + e_i$

- *The additive genetic variance and the numeric relationship matrix are the backbone of the individual-tree model*

The linear mixed model

As shown in previous slide, we use linear models to predict GCA and BV values

These models are usually written in matrix:

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

In order to understand linear mixed models, let's look at the above model in detail

Mixed model to predict BV

Data

Tree (t)	Location (*l*)	Height (y)
1	1	87
2	2	84
3	2	75
4	1	90
5	2	79

The linear model to predict BVs

$$y_{ij} = l_i + t_j + e_{ij} \qquad \text{(Model 1)}$$

Matrix notation

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad \text{(Model 2)}$$

$$\downarrow$$

$$\mathbf{y} = \mathbf{X}\ \mathbf{b} + \mathbf{Z}\ \mathbf{a} + \mathbf{e}$$

$$\begin{bmatrix}87\\84\\75\\90\\79\end{bmatrix} = \begin{bmatrix}1&0\\0&1\\0&1\\1&0\\0&1\end{bmatrix}\begin{bmatrix}l_1\\l_2\end{bmatrix} + \begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}\begin{bmatrix}t_1\\t_2\\t_3\\t_4\\t_5\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\\\varepsilon_5\end{bmatrix} \qquad \text{(Model 3)}$$

Assumptions of mixed models

Random effects have expectations of a mean of zero: $E[\mathbf{a}] = \mathbf{0}$

Residuals have expectations of a mean of zero: $E[\mathbf{e}] = \mathbf{0}$

The variance of random effects is $Var[\mathbf{a}] = \mathbf{G}$

The variance of residuals is $Var[\mathbf{e}] = \mathbf{R}$

Thus, the expectation E[y] and variance (**V**) of the observation vector **y** are given by:

$$E[\mathbf{y}] = \mathbf{Xb}$$

$$Var[\mathbf{y}] = \mathbf{V} = \mathbf{ZGZ^T} + \mathbf{R}$$

Solving mixed model equations

Henderson's mixed model equations

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

The equation can be simplified if the errors (R matrix) are identical for all observations

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

Solutions of mixed model equations

The BLUE of fixed and BLUP of random effects

$$\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \frac{\boldsymbol{\sigma}_e^2}{\boldsymbol{\sigma}_a^2}\mathbf{A}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

The relationship matrix A:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0.5 & 0 & 0.5 \\ 0 & 1 & 0 & 0.5 & 0 \\ 0.5 & 0 & 1 & 0 & 0.25 \\ 0 & 0.5 & 0 & 1 & 0 \\ 0.5 & 0 & 0.25 & 0 & 1 \end{bmatrix}$$

Tree	Female	Height
3	1	4.5
4	2	2.9
5	1	3.9

Additive genetic relationship matrix (**A**)

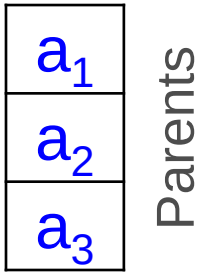
- Accounts for changes in mean and variance of breeding populations with selection
- Allows adjustment for different mating designs, self fertilizations, and other non-random mating
- Uses all the information from relatives to predict breeding values of a tree even if the tree does not have a phenotype. It increases the accuracy of breeding value estimates
- Allows connection of different test series. For example, some trees may not be tested together, but with the **A** matrix, ‘connection’ between tests can be established, allowing them to be analyzed together

GCA (parental) model

GCA model

C matrix

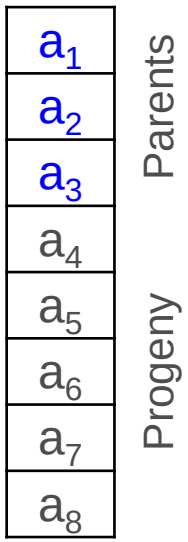
$$\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \frac{\boldsymbol{\sigma}_e^2}{\boldsymbol{\sigma}_a^2}\mathbf{A}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$



The left side of the equation is a vector. These are solutions for fixed effects (**b**) and random effects (**a**). For example, if there are 3 parents in the data, vector **a** includes solutions (a1,a2,a3) for parental GCA values.

BV (individual tree) model

$$\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \frac{\boldsymbol{\sigma}_e^2}{\boldsymbol{\sigma}_a^2}\mathbf{A}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$



- The solutions for part **a** of the vector (left side of the equation) include BV of grandparents, genetic groups, parents, and progeny
- There is no need for extra calculation to obtain breeding values

Accuracy of breeding values

Breeding values are predictions and they are associated with error. Breeders are interested in the reliability of these predictions to make decisions

$$r = \sqrt{1 - \frac{SE}{(1+F)\sigma_A^2}}$$

Accuracy (r) ranges between 0 and 1. Reliable accuracy values are closer to 1.

Where:

SE = Standard error of the predicted breeding values (BV)
F = Inbreeding coefficient, and
σ^2_A = the additive genetic variance (or, the true BV)

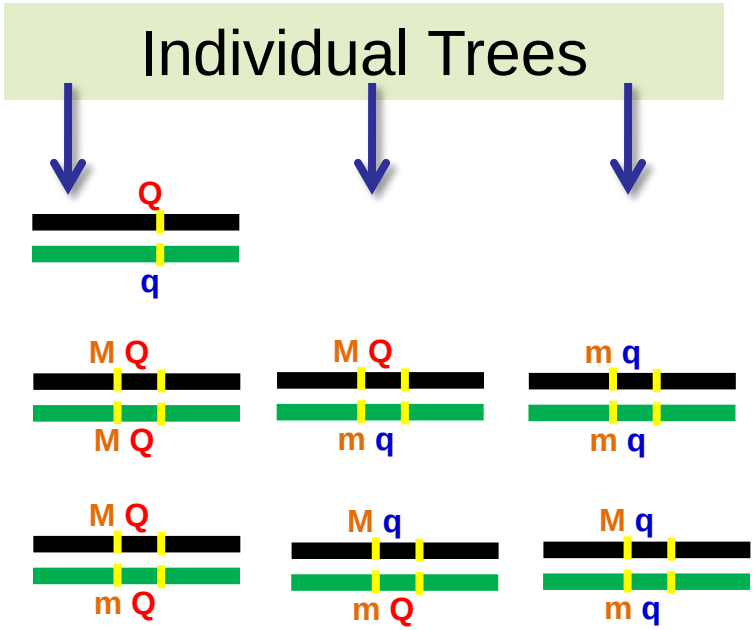
(Gilmour *et al*., 2009)

Genetic basis of MAS

Gene-MAS

LD - MAS

LE - MAS



$y = \mu + M + e$

Figure Credit: Based on Dekkers, 2010 and Dekkers and Hospital,2002.

LD marker effect estimation

AAGCCTTGATAATT
AAGCCTTGCTAATT

Progeny tested parents are grouped by their genotype for a particular SNP, and group means are determined

SNP Genotype		**Phenotype**
AA	--->	20
AC	--->	15
CC	--->	10

SNP effect estimate = +5 for the A allele

Using markers to predict BV

There are at least three ways markers may be used to predict BV:

1. Fitting markers as fixed effects
2. Fitting markers as random effects
3. Genomic-BLUP

In the above methods, markers might be in LD with trait loci or they might capture familial linkages in a population

(Fernando and Grossman, 1989; Dekkers, 2010)

Fixed marker effects (MA-BLUP)

The model may be written in matrix form as

$$\mathbf{y = Xb + Wm + Za + e}$$

b is the non-genetic fixed effects vector
m is the marker fixed effect, and
a is the polygenic effect for all other loci
X, **W**, and **Z** are incidence matrices linking effects **b, m** and **a** to **y**

This is sometimes called marker-assisted BLUP (MA-BLUP). The solutions for **m** are called the ***allelic substitution effect*** at the marker

(Fernando and Grossman, 1989; Dekkers, 2010)

Allelic substitution effect

Fitting markers as fixed effects allows us to calculate allelic substitution effect.

Let's assume we have SNP markers in the data coded 0, 1, 2. The codes correspond with 3 genotypes of a single SNP:

0 = homozygous (AA)

1 = heterozygous (AC)

2 = homozygous (CC)

Allelic substitution effect can be obtained as

$A = (\mu_{AA} - \mu_{CC}) / 2$

Which is the additive effect of alleles

(Falconer and Mackay, 1996)

Fitting markers as random effects

In matrix form the equation becomes

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{m} + \mathbf{Z}_2\mathbf{a} + \mathbf{e}$$

b is the vector of fixed effects with design matrix **X**,
m is the random effect of markers with design matrix $\mathbf{Z}_1$
a is the random effects of loci other than markers with design matrix $\mathbf{Z}_2$
e is the random error term
The expectations of random markers effect (**m**) are a mean of zero and variance of σ^2_m. We summarize this expectation as $\mathbf{m} \sim N(0, \sigma^2_m)$

(Weller *et al*., 2005)

Expanding the model to include multiple markers

The simplest way to model multiple markers is to use multiple regression, fitting all the markers simultaneously

$$y_i = \mu + \sum_{i=1}^{p} X_i g_i + e_i$$

Xi is the design matrix. (Xi = 1 if marker i = 1 and Xi =0 otherwise), p is the number of significant markers

	X1	X2
Tree1	0	1
Tree2	1	1
Tree3	1	0

For example, if we have 2 markers, Xi could take the form

Incomplete genotype data

- Complete data sets (all genotypes for all trees) are seldom, if ever, obtained, even with multiple passes

Tree	locus1	Locus2	locus3	locus4
1	GG	GG	GG	AG
2	.	GG	GG	.
3	AG	GG	.	.
4	GG	.	AG	AG

- Some software programs developed to predict the effect of markers can not handle missing genotypes
- Missing data must be imputed for marker effects to be estimated

Imputing missing genotypes

- For imputation, human geneticists
  - use algorithms that rely on known map positions of SNPs
  - employ a haplotype reference dataset, such as HapMap, for population haplotypes or reference panels and
  - use various software such as IMPUTE, PLINK, MACH, BEAGLE, fastPHASE
- Reference data sets for forest trees generally do not exist since we do not yet have a complete sequenced genome of trees, except for eucalyptus as of 2011
- However, tree breeders, like animal breeders can explore pedigrees or minor allele frequencies (MAF) to impute missing genotypes

Minor allele frequency

To impute missing genotypes, first we need to convert genotypes (e.g., AT) to minor allele frequency.

Let T and C be minor alleles for two loci. Minor allele frequency (MAF) of two loci would be as follows:

Tree	locus1	MAF	locus2	MAF
1	AA	0	CC	2
2	AT/TA	1	CG/GC	1
3	TT	2	GG	0

Now let's look at how we use MAF to impute missing genotypes in the next few slides

Imputing missing genotypes using pedigree

Let's say *q* is the MAF. If a tree is missing a specific locus genotype, the expected genotype would be the average value of gene content number (GCN) of its parents

$$\mathbf{E(q_p) = 0.5(q_s+q_d)}$$

Parental gene content number (*q*) can be replaced by its deviation (*d*) from the population mean (μ) as $d = q - \mu$ for ease of calculation.

Thus, GCN for males is $\mathbf{q_s = d_s + \mu}$ and for females is $\mathbf{q_d = d_d + \mu}$

Expected GCN of progeny: if both parents have known genotype: $\mathbf{E(q_p) = \mu + (d_s + d_d)/2}$

if only maternal grand-sire is known: $\mathbf{E(q_p) = \mu + 0.5d_{mgs}}$

(Gengler *et al*., 2007)

BLUP to impute missing genotypes

We are now going to look at the use of BLUPs to estimate missing genotypes

The model is $\mathbf{q_y} = \mathbf{1'1} + \mathbf{Mu} + \mathbf{e}$

Where $\mathbf{q_y}$ is the vector of allele content number, **1'1** is the vector of 1 to calculate mean, **u** is the individual tree effect. In matrix format the same model is

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{1'M} \\ \mathbf{M'1} & \mathbf{M'M} + \mathbf{A^{-1}}\boldsymbol{\varepsilon} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\mathbf{d}}_{\mathbf{y}} \\ \hat{\mathbf{d}}_{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \mathbf{1'q_y} \\ \mathbf{M'q_y} \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A_{yy}} & \mathbf{A_{yx}} \\ \mathbf{A_{xy}} & \mathbf{A_{xx}} \end{bmatrix} \text{ and } \varepsilon = \sigma_e^2/\sigma_d^2$$

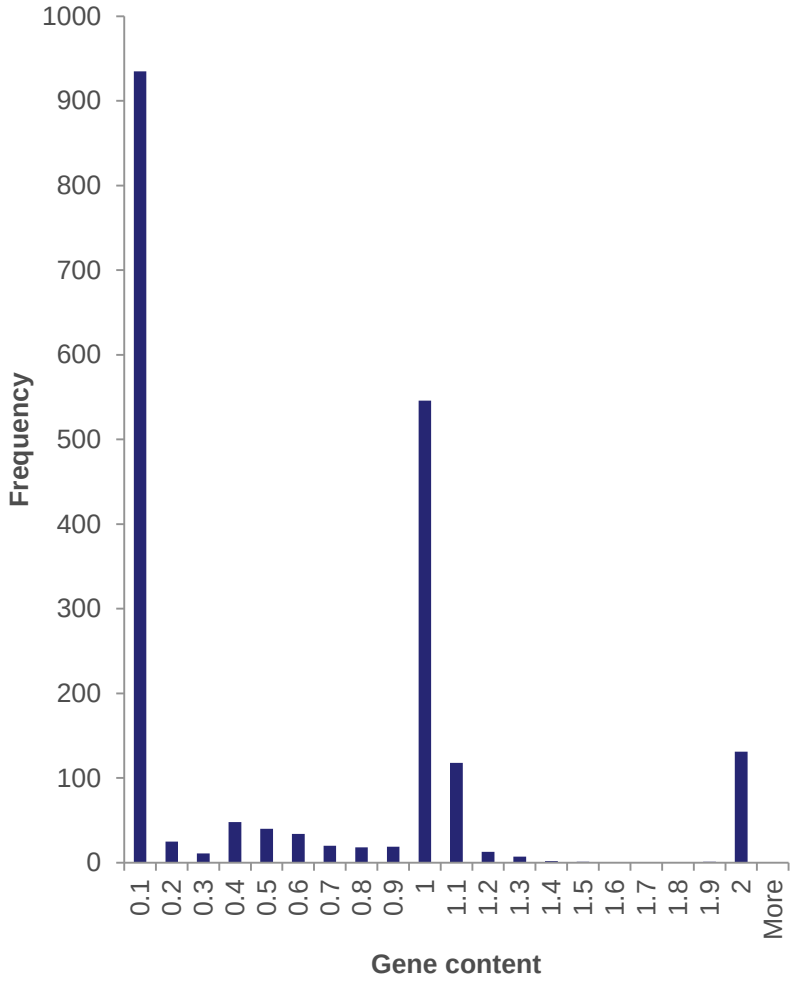
(Gengler *et al.*, 2007)

Predicted missing genotypes

treeID	BLUP	SE
1	0.000	1.000
2	0.3504	0.9438
13	0.3504	0.9438
14	0.7007	0.7502
15	0.000	1.000
16	0.7007	0.7502

The predicted missing genotypes are continuous and distributed around 1. Information from all the relatives is used to predict missing genotypes. The predicted genotypes can be used to predict BVs

Distribution of imputed missing genotypes



This histogram displays the predicted gene content number. A high frequency of predicted gene content numbers are around 0, 1 and 2. However, there are some predictions in between these three values, particularly for trees with missing genotypes

Using markers to construct genomic relationships for predictions

Traditional genetic evaluations combine phenotypic data and pedigree information (**A** matrix)

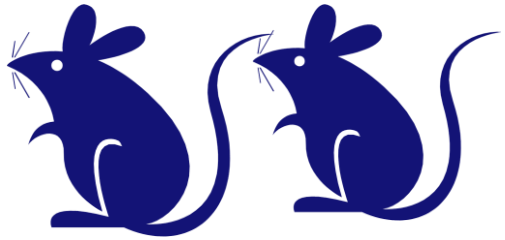
Genetic markers across the entire genome can also be used to measure genetic similarities. Using genetic relationships based on markers may be a more precise way to predict genetic merit of individuals (vanRaden, 2008)

This has important implications in prediction of genetic merit and we shall cover briefly the concept for the rest of this module

Genomic relationships derived from marker genotypes

- Let's first introduce the concept of genetic similarities between individuals using markers
- The principal is based on the idea that if the phenotypes of individuals with the same genotype at a marker (AA) are more similar than the phenotypes of individuals that have a different genotype at the same marker (AC), this suggests that the marker is in LD with a QTL affecting the trait

AA



AC

Calculation of genomic relationships

Let's say **M** is the matrix of 3 loci (columns) and 3 individuals (rows). Elements are set to −1, 0, 1 for a homozygote, heterozygote, and the other homozygote, respectively

$$\mathbf{M} = \begin{vmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & -1 \end{vmatrix}$$

← Tree 1
← Tree 2
← Tree 3

When we multiply the M matrix with its transpose (M`), we get

$$\mathbf{MM`} = \begin{vmatrix} 2 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 3 \end{vmatrix}$$

← Tree 1
← Tree 2
← Tree 3

(Forni *et al*., 2011)

Diagonal – counts the number of homozygous loci for each individual
Off-diagonal – measures the number of alleles shared by relatives

Observed frequencies to obtain G

1) Allele frequencies of three loci

$$\mathbf{P} = \begin{bmatrix} 0.666 & 0.334 & -0.666 \\ 0.666 & 0.334 & -0.666 \\ 0.666 & 0.334 & -0.666 \end{bmatrix}$$

← Tree 1
← Tree 2
← Tree 3

2) Matrix **P** is subtracted from **M** to obtain **Z**

$$\mathbf{Z} = \mathbf{M} - \mathbf{P} = \begin{bmatrix} 0.334 & -0.334 & -0.334 \\ -0.666 & -0.334 & 0.666 \\ 0.334 & 0.666 & -0.334 \end{bmatrix}$$

3) Matrix **Z** is used to obtain **G**, the genomic relationship matrix

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

(Forni *et al*., 2011)

Genomic-BLUP

- The genomic relationships (**G**) matrix based on markers is then used in linear mixed models to predict breeding values
- This application is another way of using markers for genome-wide selection in a breeding population.
- With these new tools we can incorporate non-additive genetic effects (Mendelian segregation effects) in BLUPs to estimate BVs
- G-BLUP for predicting breeding values will be addressed in module 15.

Summary of the module

- Linear mixed models are powerful tools to predict genetic merit of trees in breeding populations. Traditional genetic evaluation is based on phenotype and pedigree relationships for predictions
- DNA markers can be incorporated into linear models to increase the reliability of genetic merit. However, marker-aided selection has had limited applications in forest tree breeding mainly due to poor correlations between markers and trait at the population level
- With high-throughput genotyping technologies genomic selection approaches are feasible for forest trees

References cited

- Dekkers, JCM, and F. Hospital. 2002. Multifactorial genetics: The use of molecular genetics in the improvement of agricultural populations. Nature Reviews Genetics 3: 22-32. (Available online at: http://dx.doi.org/10.1038/nrg701) (verified 29 Feb 2012).
- Falconer, D. S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics 4th edition. Longman, Essex, England.
- Fernando, R. L., and M. Grossman. 1989. Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution 21: 467–477.
- Forni, S., I. Aguilar, I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43: 1. (Available online at: http://dx.doi.org/10.1186/1297-9686-43-1) (verified 29 Feb 2012).

References cited (cont'd)

- Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: Application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1: 21-28. (Available online at: http://dx.doi.org/ 10.1017/S1751731107392628) (verified 29 Feb 2012).
- Mrode, R. A. 2005. Linear Models for the Prediction of Animal Breeding Values. 2nd edition, CABI, Oxfordshire, United Kingdom.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414-4423. (Available online at: http://dx.doi.org/ 10.3168/jds.2007-0980) (verified 29 Feb 2012).
- Weller, J. I., M. Shlezinger, and M. Ron. 2005. Correcting for bias in estimation of quantitative trait loci effects. Genetics Selection Evolution 37: 501-522. (Available online at: http://dx.doi.org/ 10.3168/jds.2007-0980) (verified 29 Feb 2012).

External Links

- Dekkers, J. Genomic selection in livestock: Introduction and motivation [Online short course notes]. 2010 Animal Breeding & Genetics Short Courses, Iowa State Unversity. Available at: http://www.ans.iastate.edu/stud/courses/short/2010short.html (verified 29 Feb 2012).
- Gilmour, A. R., B. J. Gogel, B. R. Cullis, and T. Thompson. 2009. ASReml user [Online user guide]. VSN International Ltd. Available at: http://www.vsni.co.uk/downloads/asreml/release3/UserGuide.pdf (verified 29 Feb 2012).

Additional resources

- Hayes, B. QTL mapping, MAS, and genomic selection. [Online short course notes] 2007 Animal Breeding & Genetics Short Courses, Iowa State Unversity. Available at: http://http://www.ans.iastate.edu/stud/courses/short/2007/) (verified 29 Feb 2012).
- Meuwissen, T.H.E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome wide dense marker maps. Genetics 157: 1819-1829.

Thank You.

Conifer Translational Genomics Network
Coordinated Agricultural Project

CTGN CAP

UCDAVIS

USDA

United States
Department of
Agriculture

National Institute
of Food and
Agriculture