Lecture 10
Multi-Trait Models, Binary and Count Traits, Genome-enhanced prediction

Guilherme J. M. Rosa

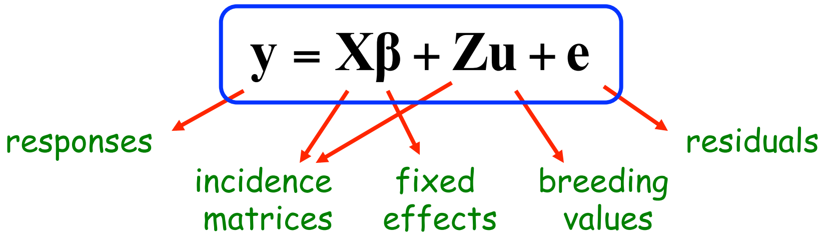
University of Wisconsin-Madison

Introduction to Quantitative Genetics
SISG, Seattle
16 – 18 July 2018

OUTLINE

- Animal Model
- Multiple-trait Model
- Repeatability Model
- Maternal Effects
- Generalized Linear Models
- Genome-enhanced Prediction

Animal Model



$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}}$$

responses

incidence matrices

fixed effects

breeding values

residuals

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 \end{bmatrix} \right)$$

Mixed Model Equations

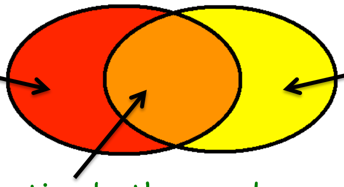
$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}$$

$$\lambda = \frac{\sigma_e^2}{\sigma_a^2} = \frac{1-h^2}{h^2}$$

BLUP: $\hat{\mathbf{u}} = (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{A}^{-1})^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$

Genetic Correlation

Schematic representation of pleiotropy



Genes affecting trait y_1

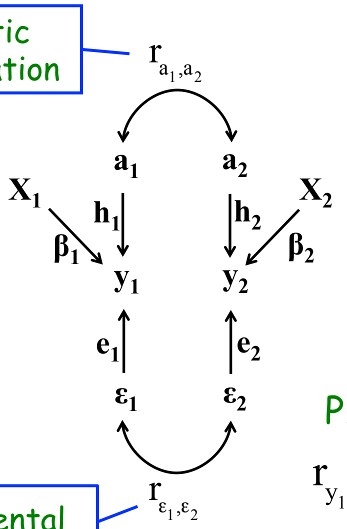
Genes affecting trait y_2

Genes affecting both y_1 and y_2

- Pleiotropic genes affect both y_1 and y_2 resulting in a genetic correlation between the two traits
- In addition to pleiotropy, genetic correlations can be caused also by linkage disequilibrium (LD) between genes affecting the different traits. LD however is a 'temporary' cause of genetic correlation as recombination can breakdown LD over the generations

Multiple (Correlated) Traits

Genetic correlation



Environmental correlation

$$\begin{cases} h_j = \sqrt{h_j^2} \\ e_j = \sqrt{1-h_j^2} \end{cases}$$

Phenotypic correlation:

$$r_{y_1,y_2} = h_1 h_2 r_{a_1,a_2} + e_1 e_2 r_{\varepsilon_1,\varepsilon_2}$$

Multiple (Correlated) Traits

The animal model can be extended for the joint analysis of multiple traits

Let the model for each of k traits be:

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{Z}_j\mathbf{a}_j + \boldsymbol{\varepsilon}_j$$

where j is an index to indicate the trait (j = 1, 2,…,k).

For the joint analysis of the k trait, the model becomes:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon}$$

with design matrices given by:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_k \end{bmatrix} \qquad \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_k \end{bmatrix}$$

Multiple (Correlated) Traits

In this case it is assumed that:

$$\mathrm{Var}\begin{bmatrix} \mathbf{a} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \otimes \mathbf{I} \end{bmatrix}$$

where **G** and **Σ** are the genetic and residual variance-covariance matrices, given by:

$$\mathbf{G} = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_1a_2} & \cdots & \sigma_{a_1a_k} \\ \sigma_{a_1a_2} & \sigma^2_{a_2} & \cdots & \sigma_{a_2a_2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{a_1a_k} & \sigma_{a_2a_k} & \cdots & \sigma^2_{a_k} \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{\varepsilon_1} & \sigma_{\varepsilon_1\varepsilon_2} & \cdots & \sigma_{\varepsilon_1\varepsilon_k} \\ \sigma_{\varepsilon_1\varepsilon_2} & \sigma^2_{\varepsilon_2} & \cdots & \sigma_{\varepsilon_2\varepsilon_2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\varepsilon_1\varepsilon_k} & \sigma_{\varepsilon_2\varepsilon_2} & \cdots & \sigma^2_{\varepsilon_k} \end{bmatrix}$$

Note: $\otimes$ represents the direct (Kronecker) product

Multiple (Correlated) Traits

The MME for multi-trait analyses are of the same form as before, i.e.:

$$\begin{bmatrix} \mathbf{X}'(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I})\mathbf{X} & \mathbf{X}'(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I})\mathbf{Z} \\ \mathbf{Z}'(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I})\mathbf{X} & \mathbf{Z}'(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I})\mathbf{Z}+\mathbf{G}^{-1}\otimes\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I})\mathbf{y} \\ \mathbf{Z}'(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I})\mathbf{y} \end{bmatrix}$$

from which the BLUEs and BLUPs of **β** and **a** can be obtained.

Multiple (Correlated) Traits

The dimensionality of multi-trait MME, however, can become a hurdle for solving it when more than two or three traits are considered

An alternative for the analysis of multiple traits is to use a canonical transformation of the traits, which consists of transforming the vectors of correlated traits into a new vector of uncorrelated variables

In such case, each transformed variable can be analyzed independently using standard single trait models, and subsequently the estimated breeding values are transformed back to the original scale of measurement

Repeatability Model

Repeatability Model

For the analysis of repeated measurements, environmental effects can be partitioned into permanent and temporary effects

In this case, the mixed model, usually called 'repeatability model', can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{p} + \boldsymbol{\varepsilon}$$

where $\mathbf{p} \sim N(\mathbf{0}, \mathbf{I}\sigma_p^2)$ is the vector of permanent environmental effects, with each level pertaining to a common effect to all observations of each animal

Repeatability Model

It is often assumed that **a**, **p**, and **ε**, which are independent from each other

Under these assumptions, the MME becomes:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'W} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\lambda_a\mathbf{A}^{-1} & \mathbf{Z'W} \\ \mathbf{W'X} & \mathbf{W'Z} & \mathbf{W'W}+\lambda_p\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \\ \mathbf{W'y} \end{bmatrix}$$

with $\lambda_a = \sigma^2_\varepsilon / \sigma^2_a$ and $\lambda_p = \sigma^2_\varepsilon / \sigma^2_p$

Repeatability Model

An important definition related to repeated measurements refers to repeatability (r), which is given by the intraclass correlation, i.e., the ratio of the within-individual (or between repeated measurements) to the phenotypic variances:

$$\mathrm{r} = \frac{\sigma^2_{\mathrm{a}} + \sigma^2_{\mathrm{p}}}{\sigma^2_{\mathrm{y}}} = \frac{\sigma^2_{\mathrm{a}} + \sigma^2_{\mathrm{p}}}{\sigma^2_{\mathrm{a}} + \sigma^2_{\mathrm{p}} + \sigma^2_{\varepsilon}}$$

The repeatability coefficient measures the correlation between records on the same animal, and so it is useful for example in the estimation of producing ability of an animal

Maternal Effects

Maternal Effects

There are some traits of interest in livestock, such as weaning weight in beef cattle, in which progeny performance is affected by the dam's ability to affect the calf's environment, such as in the form of nourishment through her milk production, the quantity and quality of which is in part genetically determined

In such cases, dams contribute to the performance of their progeny not only through the genes passed to the progeny (the "direct genetic effects") but also through their ability to provide a suitable environment (the "indirect genetic effects")

Maternal Effects

Maternally influenced traits can be analyzed by using a model as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Za} + \mathbf{Km} + \mathbf{Wp} + \boldsymbol{\varepsilon}$$

where **m** is a vector of random maternal genetic effects, and **p** is a vector of random maternal permanent environmental effects

It is assumed that **m** ~ $N(\mathbf{0}, \mathbf{A}\sigma_m^2)$ and **p** ~ $N(\mathbf{0}, \mathbf{I}\sigma_p^2)$, and quite often a covariance structure between direct and maternal additive genetic effects is considered, assumed equal to $A\sigma_{a,m}$

Computing Strategies

Solving the MME does not necessary require the inversion of the coefficient matrix ***C***

More computationally convenient alternatives for solving high dimensional systems of linear equations include methods based on iteration on the MME, such as the Jacobi or Gauss-Seidel iteration, and the "iteration on the data" strategy, which is commonly used methodology in national genetic evaluations involving millions of records

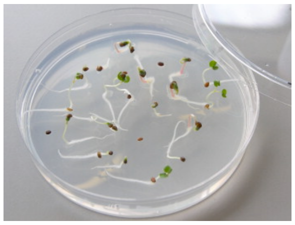
Generalized Linear Mixed Models

The models discussed so far assumed a Gaussian (normal) distribution of the phenotypic traits

Often however phenotypic traits are expressed a a binary (e.g., pregancy in dairy cattle, or germination in seeds) or count variable (e.g., litter size in swine, or fruits in trees)

In such cases the linear (Gaussian) model is not appropriate, and a generalized linear model (GLM) approach is necessary

Generalized Linear Mixed Models



Generalized Linear Mixed Models

GLM can actually model outcomes (response variables) generated from any distribution from the exponential family, which includes the normal, binomial, Poisson and gamma distributions, among others

The GLM consists of three elements:

1. Probability distribution from the exponential family.
2. Linear predictor $\eta = X\beta$
3. Link function g such that $E(Y) = \mu = g^{-1}(\eta)$.

Generalized Linear Mixed Models

Notice that the Gaussian model is a specific case of the GLM, with the normal distribution and an identity link function

In the case of Generalized Linear Mixed Models, including the applications in animal/plant breeding, the model is defined as:

1. Probability distribution from the exponential family.
2. Linear predictor $\eta = X\beta + Zu$
3. Link function g such that $E(Y|u) = \mu = g^{-1}(\eta)$

GLMM in R

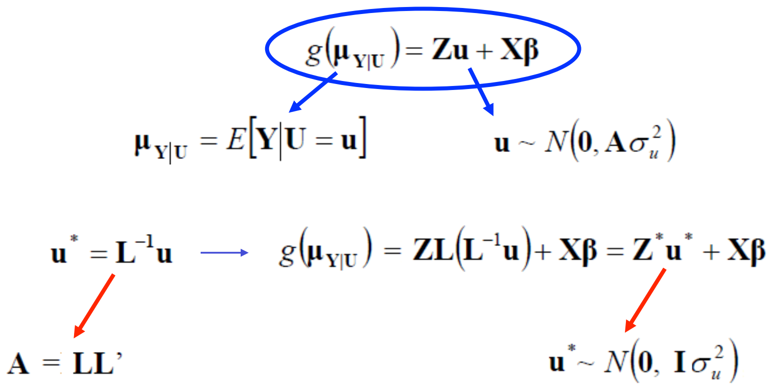
GLMM can be implemented in R using the package lme4

lme4, however, assumes independence between levels of random effects, and as such it is not suitable for many animal/ plant breeding applications

pedigreemm is an R package that uses lme4 with a Cholesky decomposition strategy to overcome this problem

pedigreemm

An R package for fitting generalized linear mixed models in animal breeding



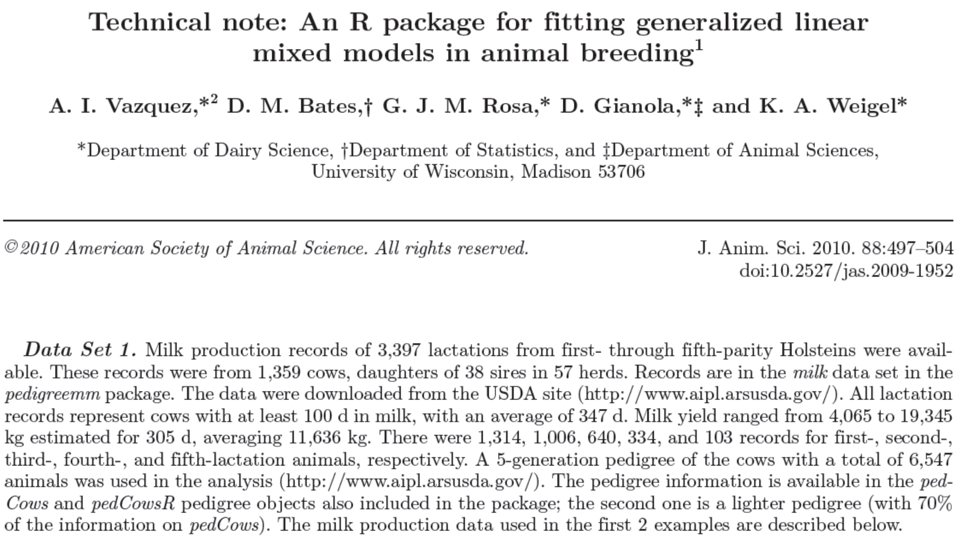
$$g(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{U}}) = \mathbf{Zu} + \mathbf{X\beta}$$

$$\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{U}} = E[\mathbf{Y}|\mathbf{U} = \mathbf{u}] \qquad \mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$$

$$\mathbf{u}^* = \mathbf{L}^{-1}\mathbf{u} \longrightarrow g(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{U}}) = \mathbf{ZL}(\mathbf{L}^{-1}\mathbf{u}) + \mathbf{X\beta} = \mathbf{Z}^*\mathbf{u}^* + \mathbf{X\beta}$$

$$\mathbf{A} = \mathbf{LL}' \qquad \mathbf{u}^* \sim N(\mathbf{0}, \mathbf{I}\sigma_u^2)$$

(Harville and Callanan 1989)



Technical note: An R package for fitting generalized linear mixed models in animal breeding[1]

A. I. Vazquez,*[2] D. M. Bates,† G. J. M. Rosa,* D. Gianola,*‡ and K. A. Weigel*

*Department of Dairy Science, †Department of Statistics, and ‡Department of Animal Sciences,
University of Wisconsin, Madison 53706

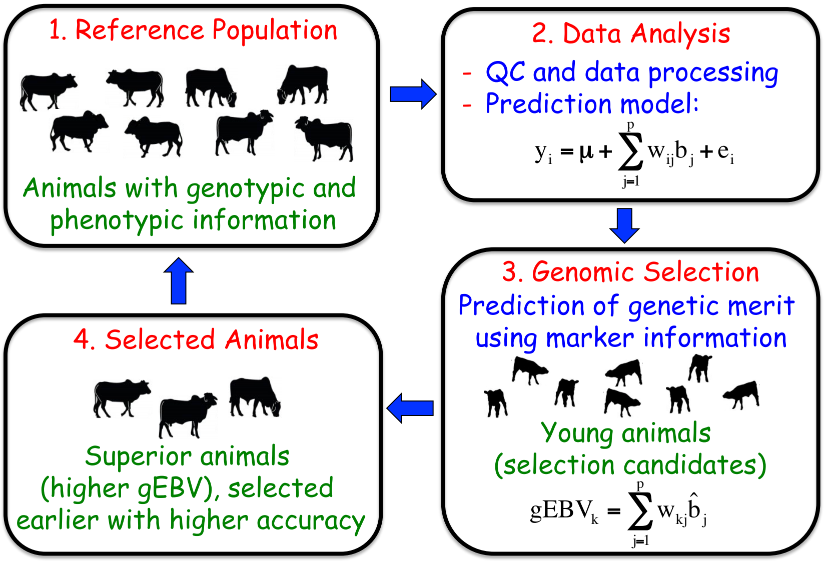
©2010 American Society of Animal Science. All rights reserved.

J. Anim. Sci. 2010. 88:497–504
doi:10.2527/jas.2009-1952

Data Set 1. Milk production records of 3,397 lactations from first- through fifth-parity Holsteins were available. These records were from 1,359 cows, daughters of 38 sires in 57 herds. Records are in the *milk* data set in the *pedigreemm* package. The data were downloaded from the USDA site (http://www.aipl.arsusda.gov/). All lactation records represent cows with at least 100 d in milk, with an average of 347 d. Milk yield ranged from 4,065 to 19,345 kg estimated for 305 d, averaging 11,636 kg. There were 1,314, 1,006, 640, 334, and 103 records for first-, second-, third-, fourth-, and fifth-lactation animals, respectively. A 5-generation pedigree of the cows with a total of 6,547 animals was used in the analysis (http://www.aipl.arsusda.gov/). The pedigree information is available in the *pedCows* and *pedCowsR* pedigree objects also included in the package; the second one is a lighter pedigree (with 70% of the information on *pedCows*). The milk production data used in the first 2 examples are described below.

pedigreemm example

Genome-enhanced Selection



1. Reference Population

Animals with genotypic and phenotypic information

2. Data Analysis

- QC and data processing
- Prediction model:

$$y_i = \mu + \sum_{j=1}^{p} w_{ij} b_j + e_i$$

3. Genomic Selection

Prediction of genetic merit using marker information

Young animals (selection candidates)

$$gEBV_k = \sum_{j=1}^{p} w_{kj} \hat{b}_j$$

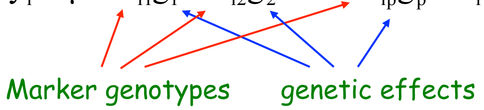
4. Selected Animals

Superior animals (higher gEBV), selected earlier with higher accuracy

Genome-enhanced Selection

(Meuwissen et al., 2001)

$$y_i = \mu + x_{i1}g_1 + x_{i2}g_2 + \ldots + x_{ip}g_p + e_i$$



Marker genotypes

genetic effects

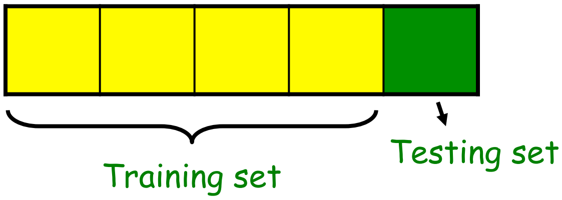
Genomic EBV: $\text{GEBV} = x_{i1}\hat{g}_1 + x_{i2}\hat{g}_2 + \ldots + x_{ip}\hat{g}_p = \sum_{j=1}^{p} x_{ij}\hat{g}_j$

- ⇨ 'big p small n paradigm'
- ⇨ Dimension reduction techniques (e.g. SVD and PLS), and stepwise strategies
- ⇨ Alternatively: penalized regression, shrinkage estimation

Cross-validation

(Predictive Ability)

➡ K-fold



Training set

Testing set

$$\begin{cases} y = X\beta + e \\ \hat{\beta}\text{: estimate of } \beta \end{cases} \Rightarrow \begin{cases} \text{PMSE} = \frac{1}{m}\sum_i (y_i - \hat{y}_i)^2 \\ \hat{y} = X\hat{\beta} \end{cases}$$

➡ Leave-one-out ("n-fold")

GBLUP

Regression with genetic effects with
normal distribution with common variance

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j + \mathbf{e} \quad \text{, with: } g_j \mid \sigma_g^2 \sim N(0, \sigma_g^2)$$

Equivalent Model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{b} + \mathbf{e} \quad \text{, with: } \mathbf{b} \mid \sigma_b^2 \sim N(\mathbf{0}, \mathbf{G}\sigma_b^2)$$

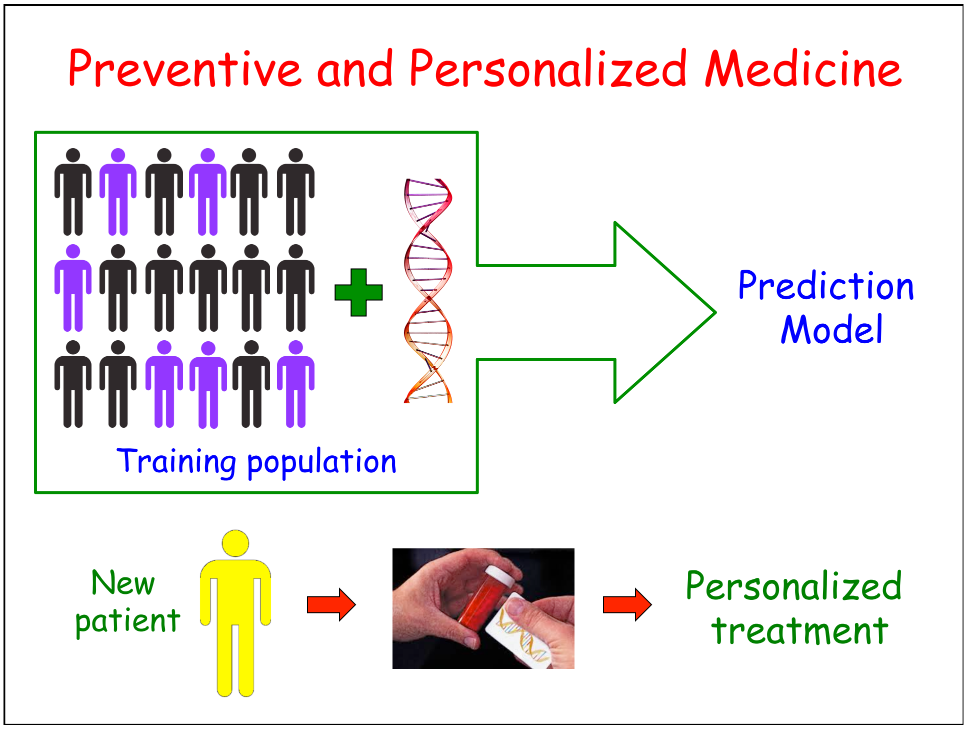
⇨ G is the genomic relationship matrix (VanRaden 2008):

$$\mathbf{G} = \left(2 \sum_{j=1}^{p} p_j (1 - p_j) \right)^{-1} (\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})'$$

ssGBLUP

- Single-step GBLUP (Misztal et al. 2009)
- Single mixed model with all animals (genotyped and non-genotyped) included, with matrix A replaced by H:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$



Preventive and Personalized Medicine

Training population

Prediction Model

New patient

Personalized treatment