# Genomic analyses with emphasis on single-step

**Condensed notes for short-course taught in 2015 in Poznan, Poland**
Ignacy Misztal et al., 11/20/14 - 5/9/16

**Introduction**
In the last couple of years, the animal breeding research focused on genomic selection. Early studies focused on the design of SNP chips for many species and developed basic methodologies based either on SNP estimation or GBLUP. Successive studies looked at the number of genotyped animals and size of SNP chip for a reasonable accuracy, optimal SNP selection, and SNP weighting. Later studies pondered imputation with low density chips and gains with high density chips, best choice of animals for genotyping, validation methods, etc. Now the most trending topic is the use of sequence data.

The animal-breeding group at the University of Georgia (UGA) has long been active in tools for commercial genetic evaluation across species. When genomic selection appeared, the group focused on development of technology that would be suitable for a commercial application. The method was single-step GBLUP, a BLUP with a different relationship matrix. While the basic theory for ssGBLUP is simple, there are many details. In particular, details that seem to be unimportant for one species become paramount in another species.

The last major steps in the development of ssGBLUP were algorithms for inversion of genomic relationship matrix and the pedigree relationship matrix for a subset of genotyped animals. These developments removed limits to the number of genotyped animals. The success of the first algorithm raises doubts about meaningful accuracy gains with high-density chips (dimensionality of genomic information less than 15,000) and paves the way for genomic selection becoming a routine procedure like BLUP.

Single-step GBLUP has been applied by some of the largest companies/institutions in the US. As a result, UGA has access to perhaps the most comprehensive data sets anywhere. With access to data sets and practical problems, the group solved many puzzles, developed many tools, and discovered many solutions.

The purpose of this material is to present an outline of developments in the genomic selection from a practical point of view, for use in a new short course taught by Misztal, Aguilar and Lourenco in 2015 in Poznan, Poland. Due to scarcity of time, nearly all references in this draft are to papers associated with UGA and collaborators. See those papers for references and names of the “UGA team.” For more additional information please see 2016 course notes as well as BLUPF90 manual available at nce.ads.uga.edu.

**Models, Methods, and Programs**

Good genomic predictions require a few components:

1. Appropriate models for traits of study
2. Appropriate methods for parameter estimation
3. Appropriate methods for genetic evaluation
4. Software that can implement mixed models for both parameter estimation and genetic prediction

Here we assume that all models are analyzed by mixed model methodology.

Good models are a prerequisite for any genetic evaluation including genomic. Bad models = bad GEBVs. All models are approximations. Many effects influence the traits of interest. Giving large data sets, any effect is likely to be statistically significant although many may be unimportant in practice. We can define practically equivalent models as such where correlations of (G)EBV for selection candidates are at least 0.98. Then, our desired model is as simple as possible (parsimonious) that gives essentially same predictions as more complicated models. Please note that very complex models may give bad predictions due to traits not following model assumptions or due to computing instability. Also, commonly used statistical criteria for model selection such as BIC often lead to complex models and not best predictions as determined by cross-validation.

**Comments on Models**

**Single-trait (repeatability with repeated records) animal model.**

Basic model when other traits are not available or are weakly correlated

**Multiple trait model**

A model that accounts for correlations among traits (genetic and environmental). Important when some traits are missing, when the recording is sequential and when genetic correlations are not close to 0. Inclusion of traits under selection in the model helps remove biases form analyses of all traits.

**Random regression model (RRM)**

Useful for analyses of traits that are continuous with time (e.g., milk or weight) or other variables (e.g., temperature-humidity index or herd-level). The key part of RRM is the use of several parameters to describe variability of continuous traits. Functions that use such parameters may be regular polynomials (simple but poor numerical stability), Legendre orthogonal polynomials (better numerical properties), or splines (less variability at extremes). Type and order of functions determine shape of covariance trajectories. In particular, too high polynomial order in RRM shows modeling artifacts where (co)variance curves have large fluctuations and are not supported by biological or management reasons. One-way of simplifying RRM is the use of linear splines as they are equivalent to multiple-trait models with traits corresponding to points at knots (Misztal, 2006) and cross-validation properties are good (Bohmanova et al., 2008).

**Unknown parent groups (UPG)**

If the base population is heterogeneous, under selection the merit of missing parents may be uneven. Then, pedigrees can be augmented by “phantom parents” via unknown parent groups. Groups can model the merit of unknown parents by year of birth, sex, path of selection, genetic origin or breed.

Use of UPG is dangerous due to possibility of confounding or large SE (UPG are fixed effects). Therefore, when UPG are used, the solutions of UPG should be examined for excessive fluctuations, trends, or just common sense. One way of reducing the confounding/fluctuations is treating UPG as random. In the simplest case, this is by treating UPG as animals (var(UPG)=additive variance).

**UPG and historical data**

In many analyses, the use of more than 2-3 historical generations does not influence GEBV of the youngest generation or may even improve it (Lourenco et al., 2014). This is due to a number of reasons including 1) decays of genomic predictions, 2) changes in traits over time, 3) imperfect modeling of fixed effects over time. So one of the simplest ways to limit computations is by truncating phenotypes and pedigrees. Truncating data also limits or even eliminates the need for UPG.

**Common features of software**

Software for mixed models should generally support:

- missing traits
- different models per trait
- heterogeneous residual variance

**Approaches for variance component estimation (VCE)**

VCE is usually accomplished by REML or by Bayesian analyses via Gibbs sampling. For general properties, see paper “Reliable computing in estimation of variance components.” Below are some of the more popular features of common algorithms.

General REML programs rely on sparse matrix software. They have approximately a quadratic cost with the number of animals and the cubic cost with the number of traits. They are likely to fail when the number of traits is too high or resulting covariance matrices are close to singularity.

AI-REML is fast with simpler models but often crashes with models close to the parameter space.

EM-REML is more stable but very slow and there are no good estimates of SE. EM-REML is faster with less “missing information.” EM_REML gets “stuck” with RRM if the starting parameters are low.

Special version of REML (canonical transformation) can be very stable and inexpensive with a large number of traits but add model limitations.  See Karin Meyer's page for many options in REML for complicated models.

Bayesian methods via Gibbs sampling (BAGS) may be slow or fast depending on optimizations. Usually the number of required samples is higher with model complexity and with the number of parameters. For instance, a reduced animal model may require 10 times fewer samples than a regular animal model. Analyzing the output of BAGS is a necessity.

An optimized BAGS (e.g., gibbs1f90 and later versions) is resistant to non-positive definite matrices and can be very fast with large number of traits if models per trait are not too different. Complex models (e.g., threshold) are much easier to implement via BAGS than via REML.

**Approaches for national genetic evaluation**

For national evaluations, the number of effects in the model can be so large that the mixed model equations would not fit in memory even for the largest computer. Subsequently solutions are computed by "matrix free" or "iteration on data " approaches, where the data is read every round of iteration and coefficients of mixed model equations are recreated. Instead of being stored, these coefficients are used immediately to create quantities used by iteration methods. A particularly efficient yet simple to implement iteration method is preconditioned conjugate gradient (Tsuruta et al., 2001). This method can use coefficients of mixed model equations in any order.

**Why single step?**

Current methods require de-regressions and creation of an index combining different sources of information (e.g., parent average, direct genomic value, pedigree prediction). As both rely on accuracies, when accuracies are approximated both the deregressions and the index are approximate. Additionally, deregression on a mix of high and low accuracy animals can create double-counting (Legarra et al., 2014; Lourenco et al., 2015).

**Derivation of H**

The derivations follow Legarra et al. (2014) and Aguilar et al. (2010). Let A be a numerator relationship matrix and let the genetic variance be set to 1.0. Let indices 1 refer to ungenotyped and 2 to genotyped animals.

$$var(u) = A\sigma_u^2$$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

Based on conditional distributions:

$u_1|u_2 \sim N(A_{12}A_{22}^{-1}u_2, A_{11} - A_{12}A_{22}^{-1}A_{21})$
$u_2 \sim N(0, G)$
$u_1 = E(u_1|u_2) + \varepsilon$
$= A_{12}A_{22}^{-1}u_2 + \varepsilon$

Calculate variances and covariances:
$var(u_1) = var(A_{12}A_{22}^{-1}u_2 + \varepsilon)$
$= var(A_{12}A_{22}^{-1}u_2) + var(\varepsilon)$
$= A_{12}A_{22}^{-1}GA_{22}^{-1}A_{21} + A_{11} - A_{12}A_{22}^{-1}A_{21}$
$= A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} + A_{11}$

$cov(u_1, u_2) = cov(A_{12}A_{22}^{-1}u_2, u_2) = A_{12}A_{22}^{-1}var(u_2) = A_{12}A_{22}^{-1}G$
$var(u_2) = G$

Finally:

$$H = \begin{pmatrix} var(u_1) & cov(u_1, u_2) \\ cov(u_2, u_1) & var(u_2) \end{pmatrix}$$

$$= \begin{pmatrix} A_{12}A_{22}^{-1}GA_{22}^{-1}A_{21} + A_{11} - A_{12}A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{pmatrix}$$

$$=A + \begin{bmatrix} A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}(G - A_{22}) \\ (G - A_{22})A_{22}^{-1}A_{21} & G - A_{22} \end{bmatrix}$$

The inverse can be derived from distributions knowing that:

$u_1|u_2 \sim N(A_{12}A_{22}^{-1}u_2, A_{11} - A_{12}A_{22}^{-1}A_{21}) = N(A_{12}A_{22}^{-1}u_2, (A^{11})^{-1})$

$$\begin{aligned} p(\mathbf{u}_1, \mathbf{u}_2) &= p(\mathbf{u}_1 \,|\, \mathbf{u}_2)p(\mathbf{u}_2) \\ &\propto \exp[-0.5(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)'\mathbf{A}^{11}(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)]\exp[-0.5\mathbf{u}_2'\mathbf{G}^{-1}\mathbf{u}_2] \\ &= \exp\left(-0.5\begin{bmatrix} \mathbf{u}_1' & \mathbf{u}_2' \end{bmatrix}\begin{bmatrix} \mathbf{A}^{11} & -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11} & \mathbf{G}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}\right) \\ &= \exp\left(-0.5\begin{bmatrix} \mathbf{u}_1' & \mathbf{u}_2' \end{bmatrix}\begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{G}^{-1} + \mathbf{A}^{22} - \mathbf{A}_{22}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}\right). \end{aligned}$$

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{G}^{-1} + \mathbf{A}^{22} - \mathbf{A}_{22}^{-1} \end{bmatrix} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

## Types of genomic relationship matrix

The most popular matrix was defined by VanRaden (2008). It is based on a SNP model:

$$y_i = \mu + Za + e, \;\; var(a) = I\sigma_a^2, \;\; var(e) = I\sigma_e^2$$

where Z is a matrix of gene content adjusted for allele frequencies and $a$ is a vector SNP effects. When u is a vector of breeding values:

$$u = Za, \qquad var(u) = G\sigma_u^2, \qquad G = ZZ' * \sigma_a^2/\sigma_u^2$$

Alternately,

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

Many other G have been proposed but none seems to provide more accurate solutions. Please note that Z adjusted for allele frequencies is

$$Z = M - 2P$$

where elements of M are 0, 1 or 2 and P is a matrix of allele frequencies. The SNP model and GBLUP have same solutions regardless of gene frequencies (Stranden and Christensen, 2011) but ssGBLUP is affected by P.

The properties of G depend on gene frequencies. With current allele frequencies, the mean of the diagonal elements of G is 1.0 and that of off-diagonals is 0. With equal allele frequencies (0.5), the means of diagonal and off-diagonal elements can be as high as 1.0 and 1.5.

**Decomposition of GEBV**

By analyzing one line of MME we can find the decomposition of GEBV:

$$u_i = w_1 PA + w_2 YD + w_3 PC + w_4 DGV + w_5 PP, \sum w_i = 1$$

where PA is Parent Average, YD is Yield Deviation (phenotypes adjusted for model effects other than additive genetic and error), PC is Progeny Contribution, DGV is the genomic prediction (coming from $\mathbf{G}^{-1}$ or indirectly from SNP effects), PP is the pedigree prediction based on the subset of genotyped animals from **A** (coming from $\mathbf{A}_{22}^{-1}$). When both parents are known and each progeny has a known mate, the weights are:

$$w_1 = \frac{2}{den}\,; w_2 = \frac{n_r/\alpha}{den}\,; w_3 = \frac{n_p/2}{den}$$

$$w_4 = \frac{g^{ii}}{den} \quad ; w_5 = \frac{-a_{ii}^{22}}{den}$$

$$den = 2 + n_r/\alpha + n_p/2 + g^{ii} - a_{22}^{ii}$$

with values for DGV and PP:

$$\mathrm{DGV} = \frac{-\sum\limits_{j,j\neq i} g^{ij}u^j}{g^{ii}}, \;\; \mathrm{PP} = \frac{-\sum\limits_{j,j\neq i} a_{22}^{ij}u^j}{a_{22}^{ii}},$$

The decomposition for young animals (w2=w3=0) is the same as in VanRaden et al. (2009b) except that the effects are estimated jointly. PP accounts for part of PA explained by DGV. In particular, when A=$A_{22}$, PA and PP cancel out. When one or two parents are genotyped, PP will include a fraction of PA. When a genotyped animal is unrelated to a genotyped population, PP=0. Alternately, DGV and PP can be combined into genomic index (GI) as in VanRaden and Wright (2013):

$$w_4 DGV + w_5 PP = w_6 GI$$

$$\mathrm{GI} = \frac{\sum_{j,j\neq i}\left(g^{ij} - a_{22}^{ij}\right)u^j}{g^{ii} - a_{22}^{ii}}, \quad w_6 = \frac{g^{ii} - a_{22}^{ii}}{\mathrm{den}}$$

GI excludes effect of PA from DGV. See Lourenco et al. (2015)

Assuming that deviations (GEBVs – EBVs) and EBVs are uncorrelated leads to the following approximation [19]:

$$g^{ii} - a_{22}^{ii} \approx \sum_{j,j\neq i}(g_{ij} - a_{22,ij})^2 \approx (n_g - 1)\Delta_i^2$$

where $n_g$ is the number of genotyped animals and $\Delta_i^2$ is the average of the squared difference between genomic and pedigree relationships across all individuals with individual i (Misztal et al., 2012). In practice, $\Delta_i^2$ is about $0.04^2$ (Wang et al., 2014).

**Quality control for G**

Quality control of G is a critical part of genomic selection as poor relationships could lead to many types of biases and losses of accuracy.

**Calling rate for SNP and animals**

Raw genotypes include either the correct SNP or an error code stating that an SNP could not be reliably read (or typed). Calling rate for SNP removes specific SNP markers from all genotyped animals where that SNP could not be reliably typed in a large number of cases. Calling rate for animals removes genotypes of such animals where less than a threshold SNPs are correctly typed. Usually the threshold is 90-95% correct typing.

**Parental exclusions and parent identification**

Parent-progeny genotypes should be compatible. In practice this means that such pairs should not have exclusive genotypes (aa in parent and bb in progeny). For unrelated animals, the number of SNP with opposite genotypes is up to 12.5% (Tsuruta et al., 2009). For parent-progeny pairs, the number is usually less than 3% with raw genotypes but can drop below 0.1% (a few SNP with 50k chip) after imposing the calling rate edits.

If there is exclusion for any parent-progeny pair (> 3% exclusions), possibilities include pedigree or genotyping error. In case of wrong paternity, one solution is to find a compatible parent.

**Distributions of diagonals of G**

Under correct scaling and with current allele frequencies, all diagonals of G should have a quasi normal distribution with a mean of 1.0 and SD < 0.1. Too small or too large diagonals (< 0.6 or > 1.5) indicate either poor genotyping quality or an animal from another line or breed (Simeone et al., 2011).

**Differences between matched G and A22**

For properly matched G and A22, differences are usually < 0.1. Large differences indicate pedigree or genotyping problems, or insufficient pedigree. In populations with missing pedigrees, there is an extra variation of the difference due to unequal pedigree length (Wang et al., 2014).

**Chromosome editing**

SNPs in the sex chromosomes and SNPs not linked to any chromosome (usually called chromosome 0) contribute little if any to GEBV and are routinely eliminated.

**Heritability of gene content**

With correct genotypes, heritability of each SNP as a dependent variable (y) in an animal model is 100% (Forneris et al., 2015). Under special genotyping errors including those from bad imputation, the heritability is lower. In practice, less than 0.99 average heritability for all SNP indicates problems.

**Population stratification using eigenvectors**

Genotypes of a population can be shown graphically as a plot of the two largest eigenvectors. While a single population cluster has an oval shape, separate clusters indicate separate lines or breeds.

**Matching G and A22**

Normal procedure for matching G and A22 include only adjustments to G for same averages of diagonal and off-diagonal elements. When the base population is heterogeneous (pedigrees missing across generations), adjustments are to average pedigree length, and G is too big for animals with short pedigree and too small with long pedigree (Cheng et al., 2011; Vitezica et al., 2011). This results in poor convergence rate when solving ssGBLUP by iteration and in upward/downward biases (Misztal et al., 2012). Several options exist for matching G and A22. One option is to modify the $H^{-1}$ matrix by using a weight for $A22^{-1}$ (omega) smaller than 1.0.

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

Another option is to introduce nonzero relationships between unknown parents like in VanRaden or via metafounders concept (Legarra et al., 2015). The last choice is to cut phenotypes and pedigrees to 2-3 generations. Cutting data to a few generations did not decrease accuracy (Lourenco et al., 2014) and reduces the impact of old pedigrees.

### Correlations between diagonal and off-diagonal elements of G and A22

In typical analyzes, the correlations between G and A22 is about 0.3±0.1 for diagonals and 0.6±0.15 for off-diagonals. The correlations depend on a depth of pedigree. The correlations between the diagonals are treated as those between pedigree and genomic inbreeding. While pedigree-based inbreeding depends on the depth of pedigree, the genomic inbreeding indirectly includes all past inbreeding. In general, new inbreeding is more detrimental than old inbreeding, where effects of deleterious genes are mostly removed by selection. Recent inbreeding with the genomic information can be assessed by runs of homozygosity.

### Matching of G and A22 when crossbreeding

Under crossbreeding, ideally G would be matched to A22 for all breed combinations. In particular, blocks of G due to purebred breed combinations would be 0. However, manual manipulation of blocks of matrices often makes the matrix non-positive definite. In a terminal cross model in pigs, the best approach for G was to use average gene frequencies across all animals, without additional steps (Lourenco et al., 2016). In a multibreed evaluation of sheep, where phenotypes were available on crossbreds but not on every purebred, G was generated using an average allele frequency but modeling included explicit unknown parent groups (UPG).

### Importance of matching for different species and genotyping scenarios

The impact of incorrect matching is large in simulated populations, smaller in pigs and chicken, and was almost none in dairy when the reference population included high accuracy bulls only. The impact of the genomic information is negligible on high-accuracy animals and some scaling cancels out for young animals. Also, the impact of scaling is stronger under strong selection. Simulated populations usually include many genotyped generations, each with relatively small low-accuracy animals, and with selection on a single trait. Therefore any problems with scaling will influence the older generations and subsequently the young generation. In real populations, selection is usually multitrait and not very strong on any single trait.

## Validation

### Reliability ($R^2$) of GEBV with older data

In populations with high accuracy males and high number of progeny, validation includes partial and complete data comparisons. A partial data set excludes progeny of the youngest males. Comparisons are made with GEBV or EBV obtained with partial data and EBV or DYD of a male obtained with complete data. When the number of progeny is high, the comparison involves PA+DGV-PP and PC. The decomposition of GEBV indicates when such an approach is appropriate.

### Predictability

When youngest animals have records, the realized accuracies can be computed as Corr(gebv,y-Xb)/h, where gebv is prediction for a given animal obtained without its phenotype, y-Xb is phenotype of that animal adjusted for fixed effects, and h is a square root of heritability

(Legarra, 2008). Predictability seems very objective but is hard to apply for complex models (e.g., maternal effects).

**Mutiple subsets**

When the number of genotyped animals is small, cross-validation is done by creating multiple subsets of genotyped animals, and comparing predictions obtained with all but one subset against phenotypes in that subset. E.g., see papers by Saatchi et al. (2011) or Kachman et al. (2013) The cross-validation can generate very high correlations due to strong relationships across samples. The correlations are a more realistic estimate of accuracy when subsets are chosen to be less related, e.g., by K-means clustering.

**SNP weighting and GWAS in single-step.**

The G matrix used in single-step can be weighted, with weights obtained elsewhere, e.g., by Bayesian or Lasso models. Weights and subsequently Manhattan plots can be obtained from single-step directly by DGV to SNP conversion (Wang et al., 2012; Wang et al., 2014). Assuming that D is a diagonal matrix with weights for SNP, the iterative process is:

1. Set the weight matrix D=I
2. G=ZDZ'/q (q scaling factor; details omitted)
3. Run single-step
4. From GEBV, obtain DGV
5. Convert DGV to SNP effects a= DZ' inv(ZDZ') DGV
6. Calculate SNP variance and use as weight $d_i = 2p_i q_i a_i^2$
7. Normalize D for same sum(D)
8. Go to 2

Usually best weights are obtained after 1-2 rounds.

Step 3 can be done once (if DGV change little across rounds) or every round. Formula in 6 can be different. SNP variances can be plotted as Manhattan plots.

In tests using simulated data sets the estimate of SNP effects were similar to those by BayesB. However, the best estimates were not for SNP effect of QTL but for a cluster of nearby SNPs. This illustrates limited resolution of GWAS in populations with small effective population size.

GWAS by single-step allows for analysis in models where obtaining deregressed proofs is difficult (e.g., maternal, random regression or reaction norm). There are no known methods at this time to assess SNP significance under single-step.

Weighting seems to improve the accuracy for data sets with small number of genotyped animals but little with many (> 10k).

**Indirect prediction in single-step via SNP effects**

When many young animals are genotyped and fast prediction is essential, their prediction can be obtained indirectly:

1. Run single-step using only animals with information
2. Predict SNP effects from DGV as above
3. Use SNP effects for predictions, possible adding parent average.

When the number of genotyped animals in the reference population is high, parent average adds very little to predictions and can be omitted (Lourenco et al., 2015).

**Approximations of accuracies**

Accuracies in genomic models can be approximated by calculating the amount of information due to genomics (Misztal et al., 2013). In general the information $d_i$ for animal i is composed of information due to phenotypes, pedigrees, and genomic information

$$d_i = d_i^{ph} + d_i^{ped} + d_i^{gen}$$

With reliabilities calculated as

$$rel_i = \frac{d_i}{\alpha + d_i}$$

where alpha is variance ratio for the animal model if the information is in terms of effective number of records, or for the sire model if the information is in terms of effective daughters.

Approximately, the genomic accuracy is:

$$d_i^{gen} \sim \sum_{j,j \neq i} \left(g_{ij} - a_{22,ij}\right)^2 acc_j^2$$

where $acc_j$ is "independent" accuracy of animal j. Independent means not shared with other animals.

One proposed way of obtaining accuracy is by approximated a left hand side of MME by the following and subsequent inversion:

$$LHS^{gen} = D^{ph} \quad + D^{ped} + I + G^{-1} - A_{22}^{-1}$$

Linear adjustments to D's may improve the approximation.

**Computing single-step with a large number of genotyped animals**

**REML and YAMS**

Efficiency with REML computations in animal models is based on MME being sparse and subsequent use of sparse matrix software. With genomics, MME contained large dense blocks due to G, and sparse matrix computations are less efficient with dense matrices than dense matrix techniques. Dense blocks also occur with large multiple-trait and random-regression models. One solution is the use of a computational package that treats dense-blocks in otherwise sparse matrix as dense and applies efficient computations including parallel processing. A package that implements this technique is called YAMS and was developed by

Yutaka Masuda. When applied to (AI)REMLF90 for complex models, the programs run up to 20 times faster with individual computations and up to 100 times faster with sparse inversion. With efficient matrix factorization and inversion, other operations become bottlenecks. There is little improvement for very simple single-trait models with no dense blocks.

**Structure of genotyped populations**

Initial genotyping focuses on high accuracy animals. When all high accuracy animals are genotyped, genotyping moves to lower-accuracy animals that provide much less information. With regard to young animals, many are genotyped but few are selected. For instance, over 1 million Holsteins have been genotyped by 2015, but only 25k are high accuracy bulls, < 50k are cows with phenotype eligible for a national evaluation, and the rest are animals unimportant for genomic selection except for possible reduction in bias due to pre-selection. Subsequently computing choices may either use all genotypes in analyses or select only the important ones for the main analysis and indirect predictions for the rest.

**Efficient inversion of G by APY algorithm**

The rank of G for large populations as shown by the proportion of eigenvalues explaining 99% of variation is limited due to small effective population size (Pocrnic et al., 2016). Such a rank was found to be around 10k in Holstein and Angus, 6k in pigs, and 4k in broiler chicken. This is due to a limited number of independent chromosome segments (ICS), or limited number of independent SNP cluster. Such a fact can be used to derive an inverse of G efficiently (Misztal et al., 2014; Fragomeni et al., 2015; Misztal et al., 2015; Misztal 2016; Masuda et al., 2016). Let **a** be a vector of effects of ICS (or uncorrelated SNP segments). Divide individuals into core individuals denoted as *c* and other (non-core) individuals denoted as *n*. Then $\mathbf{u}_c = \mathbf{Z}_c\mathbf{a} + \boldsymbol{\varepsilon}_c$ and $\mathbf{u}_n = \mathbf{Z}_n\mathbf{a} + \boldsymbol{\varepsilon}_n$. When the number of core animals is the same as the number of ICS: $\hat{\mathbf{a}} \square \mathbf{Z}_c^{-1}\mathbf{u}_c$ and $\mathbf{u}_n = \mathbf{P}\mathbf{u}_n + \boldsymbol{\varepsilon}_n$, where **P** is a matrix that relates BVs of non-core to core individuals. Denote:

$$\mathrm{var}\begin{bmatrix}\mathbf{u}_b \\ \mathbf{u}_c\end{bmatrix} = \mathbf{G} = \begin{bmatrix}\mathbf{G}_{bb} & \mathbf{G}_{bc} \\ \mathbf{G}_{cb} & \mathbf{G}_{cc}\end{bmatrix}\sigma_u^2,$$

Using the recursion equations,

$$\mathbf{G}^{-1} = \begin{bmatrix}\mathbf{G}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} + \begin{bmatrix}-\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathbf{I}\end{bmatrix}\mathbf{M}^{-1}\begin{bmatrix}-\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} & \mathbf{I}\end{bmatrix}.$$

The above inverse has cubic computations for core animals and linear cost for noncore animals. This inverse is also sparse. In tests, APY inversion of a 570k x 570k matrix for Holsteins took < 2 h of computing time and used less than 100 Gbytes of memory (Masuda et al., 2015). The APY inverse is the real inverse, and simulations showed slightly higher accuracy of GEBV compared to using the regular inversion of G (with blending).

## References and selected publications

### Non UGA papers

Wiggans GR, Cooper TA, VanRaden PM, Cole JB: Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. J Dairy Sci 2011, 94:6188–6193.

Legarra A, Robert-Granie C, Manfredi E, Elsen JM: Performance of genomic selection in mice. Genetics 2008, 180:611–618.

VanRaden PM: Efficient methods to compute genomic predictions. J Dairy Sci 2008, 91:4414–4423.

Bijma P: Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. J Anim Breed Genet 2012, 129:345–358.

VanRaden PM, Wright JR. Measuring genomic pre-selection in theory and in practice. Interbull Bull. 2013;47:147-50.

Stranden I, Christensen OF: Allele coding in genomic evaluation. Gen Sel Evol 2011, 43:25.

Christensen OF: Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet Sel Evol 2012, 44:37.

Garrick DJ, Taylor JF, Fernando RL: Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 2009, 41:55.

### Early papers from University of Georgia and Associates

Tsuruta. S., I. Misztal and I. Strandén. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. J. Anim. Sci. 1166:1172.

Misztal, I. 2006. Properties of random regression model using linear splines. J. Anim. Breed. Genet. 123(2):73-144.

Misztal, I. 2008. Reliable Computing in Estimation of Variance Components. J. Anim. Breed. Genet. 125:363-370.

Bohmanova,J., F. Miglior, J. Jamrozik, I. Misztal and P.G. Sullivan. 2008. Comparison of random regression models with Legendre polynomials and linear splines for production traits and somatic cell score of Canadian Holstein cows. J. Dairy Sci. 91:3627-3638.

## Papers on genomics from University of Georgia and Associates

Tsuruta, S., I. Misztal, and T. J. Lawlor. 2009. Use of low density SNP chip for parental verification in US Holsteins. J. Dairy Sci. 92(Suppl. 1):W52.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. J. Dairy Sci. 92:4648-4655.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656-4663

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743:752.

Chen, C. Y., I. Misztal, I. Aguilar, S. Tsuruta, T. H. E. Meuwissen, S. E. Aggrey, T. Wing, and W. M. Muir. 2011. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: an example using broiler chickens. J. Animal Sci. 89:23-28.

Aguilar, I., I. Misztal , A. Legarra , S.Tsuruta. 2011. Efficient computation of genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128(6):422-428.
150

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43:1.

Aguilar, I., I. Misztal, S. Tsuruta, G. R. Wiggans and T. J. Lawlor. 2011. Multiple trait genomic evaluation of conception rate in Holsteins. J. Dairy Sci. 94:2621-2624.

Simeone, R., I. Misztal, I. Aguilar, and A. Legarra. 2011. Evaluation of the utility of genomic relationship matrix as a diagnostic tool to detect mislabeled genotyped animals in a broiler chicken population. J, Anim. Breed. Genet. 128(5):386-393.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and B. Muir. 2011. Effect of different genomic relationship matrix on accuracy and scale. J. Anim. Sci. 89:2673-2679.

Tsuruta, S., I. Aguilar, I. Misztal, and T. J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J. Dairy Sci. 94:4198-4204.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in Genomic Predictions for Populations Under Selection. Genet. Res. Camb. 93:357–366.

Simeone, R., I. Misztal, I. Aguilar, and Z. Vitezica. 2012. Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. J, Anim. Breed. Genet. 129( 1):3–10.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94(2):73-83.

Faux, P., N. Gengler, and I. Misztal. 2012. A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix. J. Dairy Sci. 95:6093-6102.

Misztal, I., S. Tsuruta, I. Aguilar, A. Legarra, P. M. VanRaden, and T. J. Lawlor. 2013. Methods to Approximate Reliabilities in Single-Step Genomic Evaluation. J. Dairy Sci. 96:647–654.

Misztal, I., Z.G. Vitezica, A. Legarra, I. Aguilar, and A.A. Swan. 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130:252–258.

Tsuruta, S., I. Misztal, and T. J. Lawlor. 2013. Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. J. Dairy Sci. 96:3332-3335.

Elzo, M.A., C.A. Martinez, G.C. Lamb, D.D. Johnson, M.G. Thomas, I. Misztal, D.O. Rae, J.G. Wasdin, J.D. Driver. 2013. Genomic-polygenic evaluation for ultrasound and weight traits in Angus–Brahman multibreed cattle with the Illumina3k chip, Livest. Sci. 153:39-49.

Lourenco, D. A. L., I. Misztal, H. Wang, I. Aguilar, S. Tsuruta, and J. K. Bertrand. 2013. Prediction accuracy for a simulated maternally affected trait of beef cattle using different genomic evaluation models. J. Anim. Sci. 4090-4098.

Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, E. Ezra, M. Ron, A. Shirak, and J. I. Weller. 2014. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. J. Dairy Sci. 97:1742-1752.

Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97:3943-3952

Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, T. J. Lawlor, S. Forni, and J. I. Weller. 2014. Are evaluations on young genotyped animals benefiting from the past generations? J. Dairy Sci. 97:3930-3942.

Tsuruta, S., I. Misztal, D. A. L. Lourenco, and T. J. Lawlor. 2014. Assigning unknown parent groups to reduce bias in genomic evaluations of final score in US Holsteins. J. Dairy Sci.97: 5814-5821

Wang, H., I. Misztal, I. Aguilar, A. Legarra, R. L. Fernando, Z. Vitezica, R. Okimoto, T. Wing, R. Hawken, and W. M. Muir. 2014. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. Frontiers Genet. DOI=10.3389/fgene.2014.00134.

Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, A General Approach For Genomic Selection. Livest. Sci. 166:54-65.

Dufrasne, M., I. Misztal, S. Tsuruta, N. Gengler, and K. A. Gray. 2014. Genetic analysis of pig survival up to commercial weight in a crossbred population. Livest. Sci. 167:19-24.

Wang, H., I. Misztal and A. Legarra. 2014. Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals. J. Animal Breeding Genet. DOI: 10.1111/jbg.12109

Fragomeni, B., I. Misztal, D. Lourenco, I. Aguilar, R. Okimoto and W. Muir. 2014. Changes in variance of top SNP windows over generations for three traits in broiler chicken. Frontiers Genet. doi: 10.3389/fgene.2014.00332

Zhang, X., I. Misztal, M. Heidaritabar, J. W. M. Bastiaansen, R. Borg, R. L. Sapp, T. Wing, R. R. Hawken, D. A. L. Lourenco, Z. G. Vitezica. 2015. Prior genetic architecture impacting genomic regions under selection: an example using genomic selection in two poultry breeds. Livest. Sci. 171:1-11.

Forneris, N. S., A. Legarra, Z. G. Vitezica, S. Tsuruta, I. Aguilar, I. Misztal, and R.J.C. Cantet. 2015. Quality control of genotypes using heritability estimates of gene content at the marker. Genetics. doi: 10.1534/genetics.114.173559.

Fragomeni, B. O., D.A.L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. 2015. Use of genomic recursions in single-step genomic BLUP with a large number of genotypes. J. Dairy Sci. 98:4090–4094.

Fragomeni, B. O., D.A.L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, and I. Misztal. 2015. Use of Genomic Recursions and Algorithm for Proven and Young Animals for Single-Step Genomic BLUP Analyses - A Simulation Study. J. Anim. Breed. Genet. (Accepted).

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics. doi: 10.1534/genetics.115.177014.

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and I. Misztal. 2015. Genetic evaluation using single-step genomic BLUP in American Angus. J. Anim. Sci. doi:10.2527/jas2014-8836.

Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015. Accuracy of estimated breeding values for males and females with genomic information on males, females, or both: a broiler chicken example. Genet. Sel. Evol. 47:56.

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, C. Y. Chen, W. O. Herring, I. Misztal. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. J Anim. Sci. 94:909-919.

Masuda, Y., S. Tsuruta, I. Aguilar, and I. Misztal. 2015. Technical note: Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements. J. Dairy Sci. 93:4670-4674.

Misztal, I., B. O. Fragomeni, D. A. L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, and T. J. Lawlor. 2015. Efficient inversion of genomic relationship matrix by the algorithm for proven and young (APY). Interbull Bull. 49.

Masuda, Y., I. Misztal, S. Tsuruta, D. A. L. Lourenco, B. O. Fragomeni, A. Legarra, I. Aguilar, T. J. Lawlor. 2015. Single-step genomic evaluations with 570K genotyped animals in US Holsteins. Interbull Bull. 49:85-89.

Misztal I. 2016. Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. Genetics. 202:401-409.

Masuda, Y., I. Misztal, S. Tsuruta, A. Legarra, I. Aguilar, D. A. L. Lourenco, B. O. Fragomeni, and T. J. Lawlor. 2015. Implementation of genomic recursions in single-step genomic BLUP for US Holsteins with a large number of genotyped animals. J. Dairy Sci. 98: 4090-4094.

Pocrnic I., D. A. L. Lourenco, Y. Masuda, A. Legarra, I. Misztal. 2016. The Dimensionality of Genomic Information and its Effect on Genomic Prediction. Genetics. 203:573-581.