Genomic Analysis Combining Genotyped and Non-Genotyped Individuals

Why a Combined Analysis?

- To exploit all the available phenotypic data in GWAS and genomic prediction
  - Not just the records on genotyped individuals
  - Account for preselection of genotyped individuals
- To ensure that genomic predictions include all available information
- To avoid approximations required in multi-step analyses (that lead to double-counting)

Multi-step Genomic Prediction Analysis

- Mixed model evaluation using all phenotypes and pedigree information to generate EBV and R^2
- Deregression of EBV on genotyped individuals using EBV and R^2 of trios of every genotyped individual, its sire and its dam
- Weighted multiple regression analysis of deregressed EBV to estimate SNP effects
- Genomic prediction DGV of genotyped individuals
- Pedigree prediction of DGV for nongenotyped
- Selection Index blending of DGV & EBV for GE-EBV

Selection Index Blending Assumptions

$$\mathbf{Pb} = \mathbf{g}$$

$$var\begin{bmatrix}\widehat{u}\\ \widehat{m}\\ u\end{bmatrix} = \begin{bmatrix}r_p^2 & r_p^2 r_m^2\\ r_p^2 r_m^2 & r_m^2\\ r_p^2 & r_m^2\end{bmatrix}\begin{bmatrix}r_p^2\\ r_m^2\\ 1\end{bmatrix}\sigma_g^2$$

$$var\begin{bmatrix}u-\widehat{u}\\ m-\widehat{m}\end{bmatrix} = \begin{bmatrix}1-r_p^2 & (1-r_p^2)(1-r_m^2)\\ (1-r_p^2)(1-r_m^2) & 1-r_m^2\end{bmatrix}$$

Kachman (unpublished)

Pedigree Prediction

$$\begin{bmatrix} y_n \\ y_g \end{bmatrix} = \begin{bmatrix} X_n \\ X_g \end{bmatrix} b + \begin{bmatrix} Z_n & 0 \\ 0 & Z_g \end{bmatrix} \begin{bmatrix} u_n \\ u_g \end{bmatrix} + \begin{bmatrix} e_n \\ e_g \end{bmatrix}$$

with

$$var \begin{bmatrix} u_n \\ u_g \end{bmatrix} = \begin{bmatrix} A_{nn} & A_{ng} \\ A_{gn} & A_{gg} \end{bmatrix} \sigma_a^2$$

Where ***A*** is the numerator relationship matrix (from pedigree) with subscripts n=non-genotyped & g=genotyped

Nejati-Javaremi et al (1997)

Replace A with G

$M = k\ columns\ of\ (0,1,2)\ marker\ covariates$

$G = [MM' + (2-M)(2-M)']/k$

Various other authors expanded this with various approaches to center the marker covariates to create a Genomic Relationship Matrix

Fitting G^{-1} in the mixed model equations is known as GBLUP and gives the same estimates of genomic merit as MHG "BLUP"

Genotyped Animals

$$y_g = X_g b + Z_g u_g + e_g$$

Meuwissen, Hayes & Goddard (2001)

$$with\ u_g = M_g \alpha = \sum_{j=1}^{j=\#loci} m_j \alpha_j \delta_j$$

$$\alpha_j = substitution\ effect$$

$$\delta_j = (0,1)\ indicator\ variable$$

Bayesian Alphabet

$$\delta_j = 1,\ \ \sigma^2_{\alpha_j} = (known)\,\sigma^2_\alpha\ was\ "BLUP"$$

$$\delta_j = 1,\ \ \sigma^2_{\alpha_j} = (unknown)\,\sigma^2_{\alpha_j}\ was\ BayesA$$

$$\delta_j = 0\ with\ known\ probability = \pi$$

$$\sigma^2_{\alpha_j} = (unknown)\,\sigma^2_{\alpha_j}\ was\ BayesB$$

Meuwissen, Hayes & Goddard (2001)

$$\delta_j = 0\ with\ (un)known\ probability = \pi$$

$$\sigma^2_{\alpha_j} = (unknown)\,\sigma^2_\alpha\ was\ BayesC\ or\ (BayesC\pi)$$

Kizilkaya et al (2010); Habier et al (2011)

Evolution of "The Model"

Pedigree Relationship Matrix

$$y = Xb + Zu + e$$

$$var[u] = A\sigma_a^2, var[e] = I\sigma_e^2$$

Breeding Value Model

Genomic Relationship Matrix

$$y = Xb + Zu + e$$

$M = k$ columns of $(0,1,2)$ marker covariates

$G = [MM' + (2-M)(2-M)']/k$

$var[u] = G\sigma_a^2, var[e] = I\sigma_e^2$

Nejati-Javaremi et al. (1997)

Breeding Value Model

Equivalent

$$var[u] = var[M\alpha] = MIM'\sigma_\alpha^2$$

Stranden & Garrick (2009)

$u = M\alpha =$ sum of substitution effects

$$y = Xb + ZM\alpha + e$$

$$var[\alpha] = I\sigma_\alpha^2, var[e] = I\sigma_e^2$$

Meuwissen et al. (2001)

Marker Effects Model (MEM)

What to do with the non-genotyped?

Known as Single-Step "First Attempt"

$$var\begin{bmatrix} u_n \\ u_g \end{bmatrix} = \begin{bmatrix} A_{nn} & A_{ng} \\ A_{gn} & G_{gg} \end{bmatrix} \sigma_a^2$$

Just replace that part of the numerator relationship matrix with genomic relationships

Then need a "brute-force" inversion of the var-cov matrix

Misztal et al (2009)

What to do with the non-genotyped?

Known as Single-Step "Second Attempt" (with brute force inverse)

$$H = var\begin{bmatrix} u_n \\ u_g \end{bmatrix}\sigma_a^{-2} = \begin{bmatrix} A_{nn} + A_{ng}A_{gg}^{-1}G_{gg}A_{gg}^{-1}A_{gn} & A_{ng}A_{gg}^{-1}G_{gg} \\ G_{gg}A_{gg}^{-1}A_{gn} & G_{gg} \end{bmatrix}$$

Legarra et al (2009)

Then with recognition of its simply structured inverse

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G_{gg}^{-1} - A_{gg}^{-1} \end{bmatrix}$$

Aguilar et al (2010)

Offering programming appeal by simply replacing A^{-1} in MME by H^{-1} known as Single-Step GBLUP and variants of which are widely used

What's wrong with Single-Step GBLUP?

- Its predictive ability can be improved by introducing another ad hoc constant κ whose optimal value can be found by trial and error

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \varkappa(G_{gg}^{-1} - A_{gg}^{-1}) \end{bmatrix}$$

What's wrong with Single-Step GBLUP?

- When there are less loci than genotyped individuals, G is singular
- When there are more loci than genotyped individuals, G is singular if locus covariates are centered by allele frequency
  (since G=MM' and M'1=0 then G1=0)
- These problems can be overcome by adhoc regression of **G** towards **A**

What's wrong with Single-Step GBLUP?

- The var-cov matrix involves a blending of **A** and **G** requiring that they represent the same "base"
  - The base in **A** is the pedigree founders but the allele frequencies are not usually known in that population
- It is not clear what to use to center locus covariates in populations of mixed breeds, or populations with variable breed percentages

Issues with single-step GBLUP

- The matrix ***G*** is often singular
  - More animals than markers
  - If ***G*** is centered with observed allele frequency
- The matrix ***G*** must be "on the same base" as ***A***

  Rather than using $G_{gg}^{-1} - A_{gg}^{-1}$

  The model is tuned using

  $\tau[a + b((1-c)G_{gg} + cA_{gg})]^{-1} - \omega A_{gg}^{-1}$

  with some trial and error and

  $\tau \leq 1;\ \omega \leq 1;\ a \leq 0.1;\ b \leq 1;\ 0.05 \leq c \leq 0.2$
- Computing effort increases with numbers genotyped

What's wrong with Single-Step GBLUP?

- It requires brute force inversion of 2 matrices whose order is the number of genotyped individuals (ie **G** and $\mathbf{A}_{gg}$)
  - The inversion effort increase rapidly with number of genotyped individuals
  - Inversion is impractical beyond say 100,000 individuals
- Ignacy now has an "APY" approximation approach for computing these inverses

What's wrong with Single-Step GBLUP?

- It is not computationally straightforward for extension to Single-Step BayesA
- It is not suitable for application of mixture models (BayesB, BayesC, BayesCπ)
  - But these models that provide variable selection are particularly appealing in fine-mapping applications such as with imputed NGS genotypes

Let's revisit the basic idea

$$\begin{bmatrix} y_n \\ y_g \end{bmatrix} = \begin{bmatrix} X_n \\ X_g \end{bmatrix} b + \begin{bmatrix} Z_n & 0 \\ 0 & Z_g \end{bmatrix} \begin{bmatrix} u_n \\ u_g \end{bmatrix} + \begin{bmatrix} e_n \\ e_g \end{bmatrix}$$

$$with\ u_g = M_g \alpha\ for\ genotyped\ individuals$$

$$whereas\ u_n = \widehat{u_n / u_g} + \left(u_n - \widehat{u_n / u_g}\right) = \widehat{u_n / u_g} + \varepsilon_n$$

$$with\ \widehat{u_n / u_g} = A_{ng} A_{gg}^{-1} u_g$$

$$so\ u_n = A_{ng} A_{gg}^{-1} u_g + \left(u_n - A_{ng} A_{gg}^{-1} u_g\right)$$

Substituting these results gives

$$\begin{bmatrix} y_n \\ y_g \end{bmatrix} = \begin{bmatrix} X_n \\ X_g \end{bmatrix} b + \begin{bmatrix} Z_n & 0 \\ 0 & Z_g \end{bmatrix} \begin{bmatrix} u_n \\ u_g \end{bmatrix} + \begin{bmatrix} e_n \\ e_g \end{bmatrix}$$

$$= \begin{bmatrix} X_n \\ X_g \end{bmatrix} b + \begin{bmatrix} Z_n & 0 \\ 0 & Z_g \end{bmatrix} \begin{bmatrix} A_{ng} A_{gg}^{-1} M_g \alpha \\ M_g \alpha \end{bmatrix} + \begin{bmatrix} Z_n & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_n \\ 0 \end{bmatrix} + \begin{bmatrix} e_n \\ e_g \end{bmatrix}$$

$$= \begin{bmatrix} X_n \\ X_g \end{bmatrix} b + \begin{bmatrix} Z_n A_{ng} A_{gg}^{-1} M_g \\ Z_g M_g \end{bmatrix} \alpha + \begin{bmatrix} Z_n \\ 0 \end{bmatrix} \varepsilon_n + \begin{bmatrix} e_n \\ e_g \end{bmatrix}$$

Fernando et al (2014) GSE

With "Hybrid" Mixed Model Equations

$$\begin{bmatrix} X'X & X'ZM & X_n'Z_n \\ M'Z'X & M'Z'ZM + \phi & M_n'Z_n'Z_n \\ Z_n'X_n & Z_n'Z_nM_n & Z_n'Z_n + A^{nn}\lambda \end{bmatrix} \begin{bmatrix} b \\ \alpha \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} X'y \\ M'Z'y \\ Z_n'y_n \end{bmatrix}$$

$$where\ X = \begin{bmatrix} X_n \\ X_g \end{bmatrix}, Z = \begin{bmatrix} Z_n & 0 \\ 0 & Z_g \end{bmatrix}, M = \begin{bmatrix} M_n \\ M_g \end{bmatrix} = \begin{bmatrix} A_{ng} A_{gg}^{-1} M_g \\ M_g \end{bmatrix}, y = \begin{bmatrix} y_n \\ y_g \end{bmatrix}$$

$with\ EBV\ given\ by$

$$\widehat{u_g} = M_g \widehat{\alpha}$$

$$\widehat{u_n} = M_n \widehat{\alpha} + \widehat{\varepsilon_n}$$

NB Single-Step GBLUP
is a special case of the above
(but in this equivalent model no inversion is needed)

$$M_n = A_{ng} A_{gg}^{-1} M_g$$

If everyone is genotyped

$$\begin{bmatrix} X'X & X'ZM & X_n'Z_n \\ M'Z'X & M'Z'ZM+\phi & M_n'Z_n'Z_n \\ Z_n'X_n & Z_n'Z_nM_n & Z_n'Z_n+A^{nn}\lambda \end{bmatrix}\begin{bmatrix} b \\ \alpha \\ \varepsilon_n \end{bmatrix}=\begin{bmatrix} X'y \\ M'Z'y \\ Z_n'y_n \end{bmatrix}$$

These are the MME that form the basis of BayesA, BayesB, BayesC etc

If no one is genotyped

$$\begin{bmatrix} X'X & X'ZM & X_n'Z_n \\ M'Z'X & M'Z'ZM+\phi & M_n'Z_n'Z_n \\ Z_n'X_n & Z_n'Z_nM_n & Z_n'Z_n+A^{nn}\lambda \end{bmatrix}\begin{bmatrix} b \\ \alpha \\ \varepsilon_n \end{bmatrix}=\begin{bmatrix} X'y \\ M'Z'y \\ Z_n'y_n \end{bmatrix}$$

These MME form the basis of traditional pedigree-based BLUP

Invariant to Covariate Centering

$$Genotyped$$

$$y_g = \mathbf{1}\mu + X_g b + Z_g M_g \alpha + e_g$$

$$= \mathbf{1}\mu + X_g b + Z_g \mathbf{1}c'\alpha + Z_g (M_g - 1c')\alpha + e_g$$

$$define\ \ t = c'\alpha$$

$$y_g = \mathbf{1}(\mu + t) + X_g b + Z_g (M_g - 1c')\alpha + e_g$$

$$= \mathbf{1}\mu^* + X_g b + Z_g M_g^c \alpha + e_g$$

……when all animals genotyped (BayesA, BayesB etc)

But non-genotyped NOT invariant

$$Non-genotyped$$

$$y_n = \mathbf{1}\mu + X_n b + Z_n A_{ng} A_{gg}^{-1} M_g \alpha + Z_n \varepsilon_n + e_n$$

$$= \mathbf{1}\mu + X_n b + Z_n A_{ng} A_{gg}^{-1} \mathbf{1}c'\alpha + Z_n A_{ng} A_{gg}^{-1} (M_g - 1c')\alpha + Z_n \varepsilon_n + e_n$$

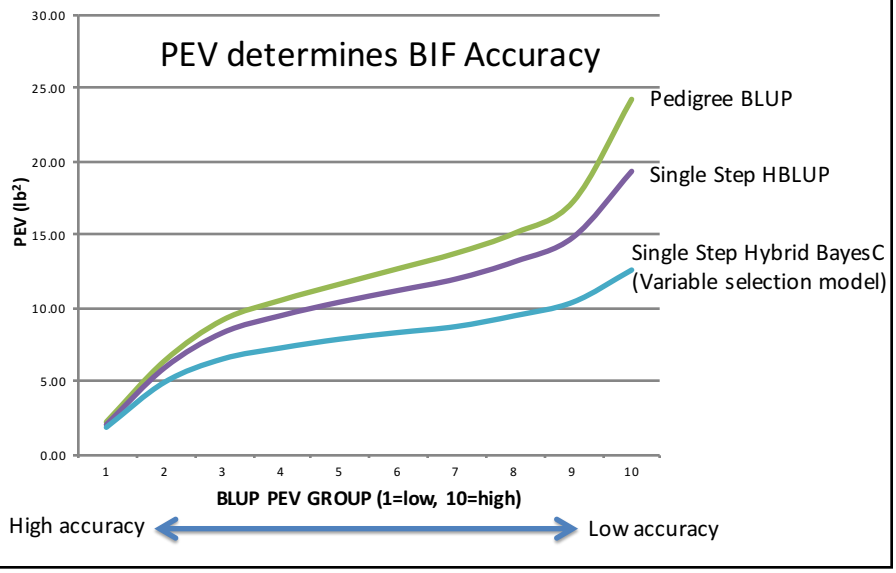
$$= \mathbf{1}\mu + X_n b + Z_n A_{ng} A_{gg}^{-1} \mathbf{1}t + Z_n A_{ng} A_{gg}^{-1} M_g^c \alpha + Z_n \varepsilon_n + e_n$$

So combined analysis of genotyped and non-genotype animals need to include a covariate for *t* if there is arbitrary centering (unless t = 0)

Computational Aspects

- It is easy to compute $A_{ng}A_{gg}^{-1}M_g$
  - And this can be done in parallel
- The computing becomes easier (rather than more difficult or impossible) as more individuals are genotyped
- Readily caters for variable selection or mixture models (eg BayesB, BayesC)
- This formulation is readily extended to multi-breed, maternal effects and multi-trait settings
- In an MCMC framework can provide PEV

Variable Selection Models have more accurate EPD



Summary

- Genomic prediction is an immature technology
- More effort is required to extend algorithms and to develop parallel computing procedures to most efficiently implement the full range of multi-breed, multi-trait, maternal effects and other models that have been routinely applied to large-scale animal prediction in recent decades

Prediction of BVs

$$with\ EBV\ given\ by$$
$$\widehat{u_g} = M_g \widehat{\alpha}$$
$$\widehat{u_n} = M_n \widehat{\alpha} + \widehat{\varepsilon_n}$$
$$or, with\ M_n = A_{ng} A_{gg}^{-1} M_g$$
$$\widehat{u_n} = A_{ng} A_{gg}^{-1} M_g \widehat{\alpha} + \widehat{\varepsilon_n}$$
$$= A_{ng} A_{gg}^{-1} \widehat{u_g} + \widehat{\varepsilon_n}$$