Introduction to Genomic Selection

Notes for a short course taught in 2019 in Piracicaba, Brazil

Daniela Lourenco et al., 7/25/2019

1. A Little bit of history (based on Lourenco et al., 2017 @ BIF Conference)

[http://www.bifconference.com/bif2017/proceedings/01-lourenco.pdf]

Long before genomics found its way into livestock breeding, most of the excitement pertaining to research into livestock improvement via selection involved developments in the BLUP mixed model equations, methods to construct the inverse of the pedigree relationship matrix recursively (Henderson, 1976; Quaas, 1988), parameter estimation and development of new, measureable traits of economic importance. In particular, for several decades (1970's through the early 2000's), lots of resources were invested in finding the most useful evaluation model for various traits. Since the 1970's, the use of pedigree and phenotypic information has been the major contributing factor to the large amount of genetic progress in the livestock industry.

During the late 1970's and early 1980's, geneticists developed techniques that allowed the investigation of DNA, and they discovered several polymorphic markers in the genome. Soller and Beckmann (1983) described the possible uses of new discovered polymorphisms, and surprisingly, their vision of using markers was not much different than how DNA is used today in the genetic improvement of livestock. They hypothesized that markers would be beneficial in constructing more precise genetic relationships, followed by parentage determination, and the identification of quantitative trait loci (QTL). The high cost of genotyping animals for such markers probably prevented the early widespread use of this technology. However, valuable information came along with the first draft of the Human genome project in 2001 (The International SNP Map Working Group, 2001): the majority of the genome sequence variation can be attributed to single nucleotide polymorphisms (SNP).

After all, what are SNPs? The genome is composed of 4 different nucleotides (A, C, T, and G). If you compare the DNA sequence from 2 individuals, there may be some positions were the nucleotides differ. The reality is that SNPs have become the bread-and-butter of DNA sequence variation (Stonecking, 2001) and they are now an important tool to determine the genetic potential of livestock. Even though several other types of DNA markers have been discovered (e.g., microsatellites, RFLP, AFLP) SNPs have become the main marker used to detect variation in the DNA. Why is this so? An important reason is that SNPs are abundant, as they are found throughout the entire genome (Schork et al., 2000). There are about 3 billion nucleotides in the bovine genome, and there are over 30 million SNPs or 1 every 100 nucleotides is a SNP. Another reason is the location in the DNA: they are found in introns, exons, promoters, enhancers, or

intergenic regions. In addition, SNPs are now cheap and easy to genotype in an automated, high-throughput manner because they are binary.

One of the benefits of marker genotyping is the detection of genes that affect traits of importance. The main idea of using SNPs in this task is that a SNP found to be associated with a trait phenotype is a proxy for a nearby gene or causative variant (i.e., a SNP that directly affects the trait). As many SNPs are present in the genome, the likelihood of having at least 1 SNP linked to a causative variant greatly increases, augmenting the chance of finding genes that actually contribute to genetic variation for the trait. This fact contributed to much initial excitement as labs and companies sought to develop genetic tests or profiles of DNA that were associated with genetic differences between animals for important traits. Suddenly, marker assisted selection (MAS) became popular. The promise of MAS was that since the test or the profile appeared to contain genes that directly affect the trait, then potentially great genetic improvement could be realized with the selection of parents that had the desired marker profile. It is not hard to see this would work very well for traits affected by one or a couple of genes. In fact, several genes were identified in cattle, including the myostatin gene located on chromosome 2. When 2 copies of the loss-of-function mutation are present, the excessive muscle hypertrophy is observed in some breeds, including Belgian Blue, Charolais, and Piedmontese (Andersson, 2001). Another example of that has been shown to have a small, but appreciable effect on beef tenderness pertains to the Calpain and Calpastatin (Page et al., 2002) and a genetic test was commercialized by Neogen Genomics (GeneSeek, Lincoln, NE) and Zoetis (Kalamazoo, MI). It is important to notice that all those achievements were based on few SNPs or microsatellites because of still high genotyping costs.

Although there were a few applications in cattle breeding, MAS based on a few markers was not contributing appreciably to livestock improvement simply because most of the traits of interest are quantitative and complex, meaning phenotypes are determined by thousands of genes with small effects and influenced by environmental factors. This goes back to the infinitesimal model assumed by Fisher (1918), where phenotypic variation is backed up by a large number of Mendelian factors with additive effects. Some lessons were certainly learned from the initial stab at MAS: some important genes or gene regions (quantitative trait loci or QTL) were detected; however, the same QTL were not always observed in replicated studies or in other populations, meaning most of them had small effects on the traits (Meuwissen et al., 2016). In addition, the number of QTL associated with a phenotype is rather subjective and depends on the threshold size of the effect used for identifying QTL (Andersson, 2001). Simply put, it appears there are only a few genes that contribute more than 1% of the genetic variation observed between animals for any given polygenic trait.

Initial allure of MAS led to a massive redirecting of grant funds to this type of research, greatly contributing to the current shortage of qualified quantitative geneticists in animal breeding (Misztal and Bertrand, 2008). Despite some of the initial setbacks using MAS, in 2001, some researchers envisioned that genomic information could still help animal breeders to generate more accurate breeding values, if a dense SNP assay that covers the entire genome became available.

Extending the idea of incorporating marker information into BLUP (using genotypes, phenotypes and pedigree information), introduced by Fernando and Grossman (1989), Meuwissen et al. (2001) proposed some methods for what is now termed genome-wide selection or genomic selection (GS). This paper used simulation data to show that accuracy of selection was doubled using genomic selection compared to using only phenotypes and pedigree information. With the promise of large accuracy gains, this paper generated enormous excitement in the scientific community. Some conclusions from this study included: 1) using SNP information can help to increase genetic gain and to reduce the generation interval; 2) the biggest advantage of genomic selection would be for traits with low heritability; 3) animals can be selected early in life prior to performance or progeny testing. With all of this potential, genomic selection was an easy sell.

However, it took about 8 years from the publication of the Meuwissen et al. (2001) paper until the dense SNP assay required for genomic selection became available for cattle. Researchers from USDA, Illumina, University of Missouri, University of Maryland, and University of Alberta developed a SNP genotyping assay, allowing the genotyping of 54,001 SNP in the bovine genome (Illumina Bovine50k v1; Illumina, San Diego, CA). The initial idea of this research was to use the SNP assay or chip for mapping disease genes and QTLs linked to various traits in cattle (Matukumalli et al., 2009). In 2009, a report about the first bovine genome entirely sequenced (The Bovine Genome Sequencing and Analysis Consortium et al., 2009) was published as an output of a project that cost over $50 million and involved about 300 researchers. With the cattle sequence known, it was possible to estimate the number of genes in the bovine genome: somewhere around 22,000. Armed with the tools to generate genomic information, GS became a reality.

Among all livestock industries in USA, the dairy industry was the first to use genomic selection. More than 30,000 Holstein cattle had been genotyped for more than 40k SNP by the end of 2009 (https://www.uscdcb.com/Genotype/cur_density.html). In January of 2009, researchers from AGIL-USDA released the first official genomic evaluation for Holstein and Jersey. Still in 2009, Angus Genetics Inc. started to run genomic evaluations, but with substantially fewer genotypes, which was also true for other livestock species. After the first validation exercises, the real gains in accuracy were far less than those promised in the paper of Meuwissen et al. (2001). This brought some uncertainties about the usefulness of GS that were later calmed by understanding that more animals should be genotyped to reap the full benefits of GS. VanRaden et al. (2009) showed an increase in accuracy of 20 points when using 3,576 genotyped bulls, opposed to 6 points when using 1,151 bulls. Now, in 2017, Holstein USA has almost 1.9 million and the American Angus Association has more than 400,000 genotyped animals.

When GS was first implemented for dairy breeding purposes, all the excitement was around one specific Holstein bull nicknamed Freddie (Badger-Bluff Fanny Freddie), which had no daughters with milking records in 2009 but was found to be the best young genotyped bull in the world (VanRaden, personal communication). In 2012 when his daughters started producing milk, his superiority was finally confirmed. Freddie's story is an example of what can be achieved with GS, as an animal with high genetic merit was identified earlier in life with greater accuracy. With

the release of genomic estimated breeding values (GEBV), the race to genotype more animals started.

The availability of more genotyped cattle drove the development of new methods to incorporate genomic information into national cattle evaluations. The first method was called multistep, and as the name implied, this method required multiple analyses to have the final GEBVs. Distinct training and validation populations were needed to develop molecular breeding values (MBV) or direct genomic values (DGV), which were blended with traditional EBVs or included as correlated traits (Kachman et al., 2013). This multistep model was the first one to be implemented for genomic selection in the USA. Several studies examining the application of multistep in beef cattle evaluation have been published (Saatchi et al., 2011; Snelling et al., 2011). The main advantage of this approach is that the traditional BLUP evaluation is kept unchanged and genomic selection can be carried out by using additional analyses. However, this method has some disadvantages: a) MBV are only generated for simple models (i.e., single trait, non-maternal models), which is not the reality of genetic evaluations; b) it requires pseudo-phenotypes (EBVs adjusted for parent average and accuracy); c) pseudo-phenotypes rely on accuracy obtained via approximated algorithms, which may generate low quality output; d) only genotyped animals are included in the model; e) MBV may contain part of parent average, which leads to double counting of information.

As only a fraction of livestock is genotyped, Misztal et al. (2009) proposed a method that combines phenotypes, pedigree, and genotypes in a single evaluation. This method is called single-step genomic BLUP (ssGBLUP) and involves altering the relationships between animals based on the similarity of their genotypes. As an example, full-sibs have an average of 50% of their DNA in common, but in practice this may range from 20% to 70% (Lourenco et al., 2015a). The ssGBLUP has some advantages over multistep methods. It can be used with multi-trait and maternal effect models, it avoids double counting of phenotypic and pedigree information, it ensures proper weighting of all sources of information, and it can be used with both small and large populations and with any amount of genotyped animals. Overall, greater accuracies and less inflation can be expected when using ssGBLUP compared to multistep methods. Not long after the implementation of GS, single-step was first applied to a dairy population with more than 6,000 genotyped animals (Aguilar et al., 2010; Christensen and Lund, 2010). An early application of ssGBLUP in beef cattle used simulated data with 1500 genotyped animals in an evaluation for weaning weight with direct and maternal effects (Lourenco et al., 2013). Although a small number of genotyped animals was used, gains in accuracy were observed for both direct and maternal weaning weight. Next ssGBLUP was applied to a real breed association data set (Lourenco et al., 2015b). This study showed a comprehensive genomic evaluation for nearly 52,000 genotyped Angus cattle, with a considerable gain in accuracy in predicting future performance for young genotyped animals. This gain was on average 4.5 points greater than the traditional evaluations.

2. From SNP markers to genomic selection methods

If SNP are just markers located outside genic regions, most of the times, why to use them? Because they may be linked to QTL or genes, fact that can be explained by an event called linkage disequilibrium (LD). The LD is based on expected versus observed allele frequencies and measures the non-random association of alleles across loci. This association gives an idea about physical distance between loci: strong LD means two loci are close. In this way, one loci can be used as a proxy for the others, and gene maps can be created.

Assuming two biallelic loci (A and B), the possible gametes formed will be AB, Ab, aB, ab.

Observed frequency of gametes		Allele frequencies	
Gamete	Frequency	Allele	Frequency
AB	0.35	A	0.7
ab	0.05	a	0.3
aB	0.25	B	0.6
Ab	0.35	b	0.4

A measure of correlation between loci can be calculated as:

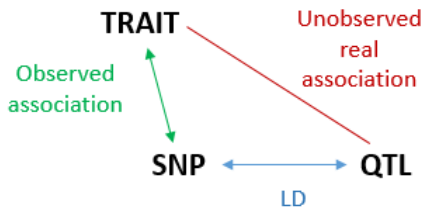
$$r^2 = \frac{D^2}{[p(A)p(a)p(B)p(b)]}; \text{ and } D = p(AB)p(ab) - p(Ab)p(aB) \quad [1]$$

$$r^2 = \frac{[(0.35*0.05)-(0.25*0.35)]^2}{[0.7*0.3*0.6*0.4]} = 0.0972 \quad [2]$$

A value of $r^2 = 0.0972$ is considered low. Moderate values are close to 0.2.

When the effective population size (Ne) is small, the chromosome segments are shorter, and LD is stronger. If we compare cattle and swine populations, LD would be stronger in swine because of the smaller Ne. The LD also depends on recent and precious recombination events, as it is broken down by recombination. In cattle, moderate LD ($r^2 = 0.2$) is observed in distances smaller than 0.1cM and strong values ($r^2 = 0.8$) are observed in very short distances.

Instead of talking about association between loci, let's assume we can use SNP to deduce the genotype of animals at each unobserved QTL. By having dense SNP panels (e.g., 50,000 SNP), it is more likely that QTL will be in LD with at least one SNP. If QTL A is linked to SNP B, depending on the strength of this linkage, once SNP B is observed it will imply QTL A was inherited together. In this way, genomic selection relies on the LD between SNPs and QTL, and although we do not observe the QTL, an indirect association between SNP and trait phenotype can be observed:



The effectiveness of genomic selection can be predicted based on the proportion of variance on the trait the SNP can explain.

There are mainly two classes of methods for genomic selection:

1) SNP effect-based methods
2) Genomic relationship-based methods

2.1. SNP effect-based methods:

For most of the livestock populations, the number of SNP is greater than the number of genotyped animals, which results in the famous "small *n* big *p* problem". As the number of parameters is greater than the data points used for estimation, a solution is to assume SNP effects are random; in this way, all effects can be jointly estimated. The most common methods used for this purpose are:

1) RR-BLUP or SNP-BLUP

This method assumes normal distribution for SNP effects and constant variance, which means all SNP explain the same proportion of variance on the trait, which is seldom true. If we have the following mixed model:

$$\mathrm{y} = \mathbf{X}b + \mathbf{Z}a + e \qquad [3]$$

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \mathbf{X'}\mathrm{y} \\ \mathbf{Z'}\mathrm{y} \end{bmatrix} \qquad [4]$$

where **X** is the incidence matrix for the fixed effects in *b*, **Z** is a matrix of SNP content, *a* is a vector of SNP effects, *e* is the residual term, λ is the ratio of residual (σ_e^2) to SNP variance (σ_a^2); $\sigma_a^2 = \sigma_u^2/nq$; σ_u^2 is the additive genetic variance; $nq = 2\sum p_i(1-p_i)$; p_i is the the allele frequency of SNP *i*.

Considering the following example with 5 animals genotyped for 7 SNP:

Animal	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	Trait
1	2	0	0	0	2	0	0	9.87
2	1	1	0	0	1	1	0	14.48
3	0	2	0	0	0	0	2	8.91
4	1	0	1	0	1	0	0	14.64
5	1	0	0	1	1	1	0	9.55

For simplicity, λ is assumed to be 1. Solutions for the system of equations in [4] are:

Effect	Solution
b	12.45
SNP1	-0.35
SNP2	0.28
SNP3	1.45
SNP4	-1.37
SNP5	-0.35
SNP6	0.54
SNP7	-1.64

RR-BLUP or SNP-BLUP provides SNP effects, but genomic estimated breeding values (u) can be derived as linear combinations of the SNP effects:

$$u = \mathbf{Z}a \qquad [5]$$

2) Bayesian Methods

Before talking about the Bayesian methods, let's remember the Bayes's theorem, which combines conditional, marginal, and joint probability:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \qquad [6]$$

For statistical inference, *A* is unknown (can be a parameter) and *B* is known (can be a trait phenotype). Therefore, we want to infer values *A* by knowing *B*. the terms in the formula are:
$p(A|B)$: posterior probability of unknown *A* given *B* is known.
$p(B|A)$: likelihood function, determined by both *A* and *B*.
$p(A)$: prior probability of unknown *A*.
$p(B)$: probability to observe *B* without having any knowledge of *A*.

We can drop $p(B)$ from the formula because it is just scaling the posterior distribution. Therefore, the kernel of the Bayes' theorem is:

$$(A|B) \propto p(B|A)p(A) \qquad [7]$$

2.1) BayesA (Meuwissen et al., 2001)

Assumes that all SNP have effect on the trait; more precisely, very few SNP have large effect and many SNP have small effect. Therefore, different variances are assumed for each SNP.

$$\beta_i \sim N\left(\mu, \sigma^2_{\beta_i}\right) \qquad [8]$$

We assume some prior information for the SNP variance. This prior has an inverted chi-squared distribution with ν_β degrees of believe and $\nu_\beta S^2_\beta$ scale:

$$\sigma^2_{\beta_i} \sim \chi^2\left(\nu_\beta, \nu_\beta s^2_\beta\right) \qquad [9]$$

$$p(\sigma^2_{\beta_i}|\mu, \beta, \sigma^2_e, \boldsymbol{D}, y) \sim \chi^2\left(1 + \nu_\beta, \beta_i^2 + \nu_\beta S^2_\beta\right) \qquad [10]$$

Therefore, there are *i* variances to be estimated (**D** is a diagonal matrix of variances for each SNP), whether the SNP *i* is relevant to the model or not. For the degrees of believe in [10], 1 degree of believe comes from the data and ν_β from the prior, which is suggested to be 4 in the literature. In this way, the posterior is heavily controlled by the prior. Another problem of this method is that *n* SNP effects and *n* variances have to be estimated, which would require *n* priors for variance, but in real situation, the same prior is used for all variances.
Bayesian methods are non-linear and likely to be affected by shrinkage, meaning small effects became even smaller and big effects even bigger.

2.2) BayesB (Meuwissen et al., 2001)

Holds the same assumption of BayesA, but for a fraction of SNPs. This is because a common thought, when this method was proposed, was that not many QTL were affecting the traits. This method states that a proportion (π) of the SNP have no effect and 1- π have a non-zero effect. Because of this assumption, many loci have zero variance.

$$p\left(\beta_i \middle| \sigma^2_{\beta_i}, \pi\right) = \begin{cases} N\left(0, \sigma^2_{\beta_i}\right) \, with \, probability \, (1-\pi) \\ 0 \, with \, probability \, \pi \end{cases} \qquad [11]$$

and

$$p(\sigma^2_{\beta_i}|\nu_\beta, S^2_\beta) \sim \chi^2\left(\nu_\beta, S^2_\beta\right) \qquad [12]$$

When π = 0, BayesB becomes BayesA.
One problem of BayesB is a distribution with high frequency of zeros when π is large, resulting in a posterior distribution that is unknown (not in a closed form); this requires the use of Metropolis-Hastings as an extra step in Gibbs Sampling, so SNP effects and variances can be jointly sampled. This extra step makes BayesB very slow.

2.3) BayesC

This method has some features of SNP-BLUP and BayesB. It assumes that SNP are drawn from a distribution with constant variance (as in SNP-BLUP) and assumes that some fraction π of SNP have no effect (as in BayesB).

As only one variance has to be estimated, the data points are bigger than one, meaning the data can overwhelm the prior.

$$p(\beta_i|\sigma_\beta^2) = \begin{cases} N(0, \sigma_\beta^2) \; with \; probability \; (1-\pi) \\ 0 \; with \; probability \; \pi \end{cases} \quad [13]$$

BayesC is much faster than BayesB. If π = 0, BayesC becomes SNP-BLUP

2.4) BayesCπ

Is a BayesC with an extra step to estimate π from the data. The prior for π is given by:

$$\pi \sim U(0,1) \quad [14]$$

Although we stopped at BayesCπ, a "whole alphabet" of methods may be available.

2.2 Genomic relationship-based methods:

These methods use SNP to infer relationships among individuals, quantifying the number of alleles shared between two individuals. When SNP are in LD with QTL, it is expected that genomic relationships will provide more accurate information about the covariance between relatives than the pedigree relationships. Genomic relationships are identical by state (IBS) because they account for the probability that two alleles randomly picked from each individual are identical, independently of origin. Pedigree relationships are identical by descent (IBD) because they consider the shared alleles come from the same ancestor.

Let's assume a matrix **Z*** of SNP inherited by each animal, with dimension *n* x *m* where *n* is the number of animals and *m* the number of SNP. Several parametrizations exist, but assuming AA=0, AB=1, and BB=2 (i.e., number of copies of the B allele), **Z*** is centered by allele frequency. Assuming a matrix **P** with elements $2p_i$, with *p* being the minor frequency of allele *i*:

$$\mathbf{Z} = \mathbf{Z}^* - \mathbf{P} \quad [15]$$

Why is **Z** a centered matrix of allele content? Assume we are working with one biallelic marker. If the effect of each copy of the A allele is *a* and the frequency of AA is p^2. Individuals with AA have a value $u = 2a$; individuals BB have $u = 0$ with a frequency of q^2; individuals AB have $u = a$ with a frequency $2pq$. The variance explained by this marker is (Legarra, 2017):

$$Var(u) = \mathrm{E}(u^2) - E(u)^2$$

The average of *u* is $2ap^2 + a2pq$; which becomes $2pa$. The variance explained by one marker is $(2a)^2p^2 + 2pq(a)^2 - (2pa)^2 = 2pqa^2$.

If we genotype two individuals for this marker, what will be the covariance between individuals *i* and *j* based on this marker? Given the average of *u* is $2pa$, as shown above, if we express the breeding values of the animals *i* and *j* as *za* deviated from the population mean (Legarra, 2017):

$$u_i = z_i a - 2pa = (z_i - 2p)a$$

$$u_j = z_j a - 2pa = (z_j - 2p)a$$

According to Legarra (2017), if $Var(a) = \sigma_a^2$, or marker variance, and the genetic variance in Hardy-Weinberg equilibrium is $2pq\sigma_a^2$, the rules of variances and covariances can be applied:

$$Cov(u_i, u_j) = (z_i - 2p)a(z_j - 2p)a = (z_i - 2p)(z_j - 2p)\sigma_a^2$$

If instead of using the allele coding 0,1,2 we use -1,0,1:

$$Cov(u_i, u_j) = z_i z_j \sigma_a^2$$

Note that by using -1,0,1, centering is not needed. However, VanRaden (2008) used -1,0,1 with centering. In this case, $\mathbf{Z} = \mathbf{Z}^* - \mathbf{P}$, but column *i* of **P** was $2(p_i - 0.5)$, which is equivalent to using 0,1,2:

$$\begin{cases} \text{if } z^* = -1, \text{then } z = -1 - 2p + 1 = 0 - 2p \\ \text{if } z^* = 0, \text{then } z = 0 - 2p + 1 = 1 - 2p \\ \text{if } z^* = 1, \text{then } z = 1 - 2p + 1 = 2 - 2p \end{cases}$$

Dividing the covariance by the genetic variance $2pq\sigma_a^2$, we have additive relationships.

Going from one to several markers, we have that the breeding value of an animal can be calculated as the sum of SNP effects weighted by the genotype content ($u = \mathbf{Z}a$). Assuming the same variance per locus, the variance of **u**, $Var(u) = Var(\mathbf{Z}a)$, then:

$$Var(u) = \mathbf{Z}Var(a)\mathbf{Z}' \qquad [16]$$

$$Var(u) = \mathbf{ZDZ}' = \mathbf{ZZ}'\sigma_a^2 \qquad [17]$$

If the genetic variance $\sigma_u^2 = 2\sum_{i=1}^{SNP} p_i(1-p_i)\,\sigma_a^2$, then $\sigma_a^2 = \frac{\sigma_u^2}{2\sum_{i=1}^{SNP} p_i(1-p_i)}$. Replacing σ_a^2 in [17] we have that:

$$Var(u) = \mathbf{ZZ}'\frac{\sigma_u^2}{2\sum_{i=1}^{SNP} p_i(1-p_i)}, \text{ then } Var(u) = \frac{\mathbf{ZZ}'}{2\sum_{i=1}^{SNP} p_i(1-p_i)}\sigma_{\boldsymbol{u}}^2 \qquad [18]$$

According to VanRaden (2008), the genomic relationship (**G**) is given by:

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}, \text{ then } Var(u) = \mathbf{G}\sigma_u^2 \qquad [19]$$

When we divide $\mathbf{ZZ}'$ by $\sum p_i(1-p_i)$, **G** becomes analogous to the numerator relationship matrix (**A**). The **G** matrix contains the number of homozygous loci for each individual in the diagonals, and the number of alleles shared among individuals in the off-diagonals. Other ways to construct the genomic relationship matrix are described in the literature.

If **G** is centered using current allele frequencies, average over all elements is zero, and average diagonal is 1 when there is no inbreeding; although **G** traces inbreeding much further than **A** because it is IBS, **A** is limited by the recent pedigree recording.

When the number of genotyped animals is larger than the number of SNP, or if there are similar individuals (e.g., clones), **G** becomes singular, therefore, cannot be inverted. To overcome this problem, we modify **G** to make diagonals bigger. **G** can be blended with 1% or 5% of an identity matrix or the pedigree relationship matrix among genotyped animals:

$$\mathbf{G} = \alpha\mathbf{G}_{raw} + (1-\alpha)\mathbf{A}_{22} \qquad [20]$$

1) Genomic BLUP (GBLUP)

This method is equivalent to SNP-BLUP, but genomic breeding values ($u = \mathbf{Z}a$) are estimated instead of SNP effects (a). It also assumes that SNP explain the same amount of variance; therefore, the majority of SNP have a small effect and very few have moderate to large effect.

Using a simple animal model:

$$\mathrm{y} = \mathbf{X}b + \mathbf{W}u + e \qquad [21]$$

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} \mathbf{X'}\mathrm{y} \\ \mathbf{W'}\mathrm{y} \end{bmatrix} \qquad [22]$$

$$u \sim N(0, \mathbf{G}\sigma_u^2) \qquad [23]$$

where **W** is a matrix that relates observations to animals.

Therefore, GBLUP is a BLUP where the pedigree relationship matrix is replaced by the genomic relationship matrix. The effectiveness of GBLUP will depend on the ability of **G** to approach the realized genetic relationships. In addition, performing a quality control of genomic data before constructing **G** avoids biases and losses of accuracy.

If we assume that not all the genetic variance is explained by markers, an extra polygenic effect can be included to explain the remaining variance. In this case, the model in [21] becomes:

$$\mathrm{y} = \mathbf{X}b + \mathbf{W}u + \mathbf{W_2}g + e \qquad [24]$$

where g is a vector of residual polygenic effect that is not captured by the SNPs. Assuming that δ is the proportion of variance explained by SNPs, the total additive genetic direct effect (u_g) becomes

$$u_g = u + g \qquad [25]$$

$$Var(u_g) = \delta\mathbf{G}\sigma_u^2 + (1-\delta)\mathbf{A}_{22}\sigma_u^2 \qquad [26]$$

$$\text{Therefore, } \mathbf{G}^* = \delta\mathbf{G} + (1-\delta)\mathbf{A}_{22} \qquad [27]$$

In real situations, it is assumed that δ varies from 0.8 to 0.95. Note that this is the same approach used to make **G** invertible ([20]).

GBLUP has been widely used in animal and plant breeding applications. The problem with GBLUP is that **G** contains only information from genotyped animals, so only those animals are included in the model. Because of that, some extra steps need to be taken to combine genomic and pedigree information. When using GBLUP, SNP-BLUP or Bayesian models, the genomic evaluation method is called multistep. The steps involved in multistep are: 1) estimation of EBV using traditional BLUP (all available information); 2) de-regression of EBV, which condenses information from phenotypes (e.g., daughter yield deviation in dairy cattle); 3) estimation of SNP effects using GBLUP or other models; 4) prediction of $\mathbf{Z}a$, which is also known as direct genomic values (DGV); 5) blending DGV with average of parent's EBV, which is known as parent average (PA). Some errors and biases can be introduced during those steps.

2) Single-step Genomic BLUP (ssGBLUP)

The idea for ssGBLUP came from the fact that only a small portion of the animals, in a given population, is genotyped. In this way, the best approach to avoid several steps would be to combine pedigree and genomic relationships and use this matrix as the covariance structure in the mixed model equations (MME). Legarra et al. (2009) stated that genomic evaluations would be simpler if genomic relationships were available for all animals in the model. Then, their idea was to look at **A** as *a priori* relationship and to **G** as an observed relationship; however, **G** is observed only for some individuals that have $\mathbf{A}_{22}$ as *a priori* relationship. Based on that, they showed the genomic information could be extended to ungenotyped animal based on the joint distribution of breeding values of ungenotyped (u_1) and genotyped (u_2) animals:

$$p(u_1, u_2) = p(u_2)p(u_1|u_2) \qquad [28]$$

$$p(u_2) = N(0, \mathbf{G}) \qquad [29]$$

If we consider that

$$var(u_1) = \mathbf{A}\sigma_u^2 \qquad [30]$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \qquad [31]$$

The conditional distribution of breeding values for ungenotyped and genotyped animals is

$$p(u_1|u_2) = N(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}u_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}) \qquad [32]$$

Because the animals with subscript 1 have no genotypes, the variance depends on their pedigree relationships with genotyped animals.

Variances and covariances are:

$$\begin{aligned} var(u_1) &= var(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}u_2 + \varepsilon) \qquad [33] \\ &= var(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}u_2) + var(\varepsilon) \\ &= \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \end{aligned}$$

Rearranging:

$$= \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$
$$= \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{I}\mathbf{A}_{21}$$
$$= \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{22}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$

Therefore,

$$var(u_1) = \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \qquad [34]$$

$$var(u_2) = var(Za) = \mathbf{G} \qquad [35]$$

$$cov(u_1, u_2) = cov(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}u_2, u_2) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}var(u_2) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \qquad [36]$$

Finally, the matrix that contains the joint relationships of genotyped and ungenotyped animals is given by:

$$\mathbf{H} = \begin{pmatrix} var(u_1) & cov(u_1, u_2) \\ cov(u_2, u_1) & var(u_2) \end{pmatrix} \qquad [37]$$

$$= \begin{pmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix}$$

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

Which can be simplified to:

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} [\mathbf{G} - \mathbf{A}_{22}] [\mathbf{I} \quad \mathbf{I}] \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \qquad [38]$$

Although **H** is very complicated, $\mathbf{H}^{-1}$ is quite simple (Aguilar et al., 2010; Christensen and Lund, 2010)

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} \qquad [39]$$

Assuming the following animal model:

$$y = \mathbf{X}b + \mathbf{W}u + e \qquad [40]$$

The MME for ssGBLUP becomes:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{H}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} \mathbf{X'}y \\ \mathbf{W'}y \end{bmatrix} \qquad [41]$$

$$u \sim N(0, \mathbf{H}\sigma_u^2) \qquad [42]$$

Based on the way **H** is constructed ([38]), the central element is $\mathbf{G} - \mathbf{A}_{22}$, which implies both matrices should be compatible (Legarra et al., 2014). However, genomic relationships can be biased if **G** is constructed based on allele frequencies other than the ones calculated from the base population (VanRaden, 2008). Allele frequencies from the base population are not known because of the recent recording of pedigrees (i.e., the base population *per se* is unknown). Most commonly, allele frequencies used to construct **G** are based on the recent population. When this is the case, the expectation of breeding values for genotyped animals is 0 (VanRaden, 2008). If the population is under selection, mean breeding values can deviate from 0. Vitezica et al. (2011) proposed an adjustment factor matching averages of **G** to averages of $\mathbf{A}_{22}$. This adjustment accounts for genotyped animals being more related through $\mathbf{A}_{22}$ than **G** is able to reflect, especially when current allele frequencies are used, and can be calculated as:

$$\rho = \frac{1}{n^2}\left(\sum_i \sum_j \mathbf{A}_{22\ i,j} - \sum_i \sum_j \mathbf{G}_{i,j}\right) \qquad [43]$$

where n is the number of elements in $\mathbf{A}_{22}$ and **G**. The new **G** is constructed as

$$\mathbf{G}^* = (1- \rho/2)\mathbf{G} + \mathbf{11}' \rho \qquad [44]$$

$\mathbf{G}^*$ is the adjusted genomic relationship matrix, **1** is a vector of ones, and ρ is Wright's F_{ST}.

In traditional BULP we obtain EBV, which can be described as (VanRaden and Wiggans, 1991):

$$EBV_i = w_1 PA_i + w_2 YD_i + w_3 PC_i \qquad [45]$$

where PA_i is the parent average EBV for animal i, YD_i is the yield deviation (phenotype adjusted for the model effects' solutions other than additive genetic effects and errors) for animal i, and PC_i is the progeny contribution for animal i. When both parents are known, the phenotype is available, and each progeny has a known mate, weights w_1 to w_3 sum to 1.

If ssGBLUP is used, the genomic EBV (GEBV) is composed by (Lourenco et al., 2015a):

$$GEBV_i = w_1 PA + w_2 YD + w_3 PC + w_4 GP + w_5 PP \qquad [46]$$

In this case, weights w_1 to w_5 sum to 1; GP is the direct genomic value and PP is pedigree prediction. The information for GP comes from **G** and the information for PP comes from $\mathbf{A}_{22}$. GP and PP can be retrieved as:

$$\mathrm{GP_i}=\frac{-\sum_{j,j\neq i} g^{ij}\mathrm{u_j}}{g^{ii}};\ \mathrm{PP_i}=\frac{-\sum_{j,j\neq i} a_{22}^{ij}\mathrm{u_j}}{a_{22}^{ii}} \qquad [47]$$

In general, PP accounts for the part of PA that is explained by GP; when all animals are genotyped, A = A22, PA and PP cancel out and GP explains a larger fraction of the GEBV; when a genotyped animal is unrelated to the genotyped population, PP = 0 and GP explains a smaller portion of the GEBV; when both parents are genotyped, PP will include a large part of PA. When a genotyped animal has many progeny, w3 $\approx$ 1 and its GEBV is mainly driven by PC; however, genotyping those animals is useful since they are usually included in the training population. When an animal is not genotyped, w4 = 0 and predictions can be improved due to improved PA and PC if its relatives are genotyped. When an animal is not genotyped and has no phenotypes and no progeny, the GEBV is driven by PA and, in most cases, only a slight improvement in prediction is achieved based on genotyped relatives (Lourenco et al., 2015a).

Weights for all components of GEBV in ssGBLUP can be calculated as:

$$w_1 = \frac{2}{den};\ w_2 = \frac{n_r/\lambda}{den};\ w_3 = \frac{n_p/2}{den};\ w_4 = \frac{g^{ii}}{den};\ w_5 = \frac{-a_{22}^{ii}}{den} \qquad [48]$$

$$den = 2 + n_r/\lambda + n_p/2 + g^{ii} - a_{22}^{ii} \qquad [49]$$

where n_r is the number of records, λ is the variance ratio (residual over genetic variance), and n_p is progeny size.

On the other hand, when GBLUP is the method of choice for genomic predictions, the genomic EBV (GEBV) is composed by:

$$GEBV_i = w_4 DGV_i \qquad [50]$$

SNP effects from GBLUP-based methods:

The assumption for GBLUP and ssGBLUP is that all SNP explain the same proportion of variance on the trait, which is not biologically acceptable. Different weights can be assigned for SNP by backsolving GEBV to SNP effects and calculating variance of each SNP (Wang et al., 2012). Using the idea that GBLUP is equivalent to SNP-BLUP (VanRaden, 2008), the selection index equation for GBLUP is:

$$\hat{u} = \mathbf{G}\left[\mathbf{G} + \mathbf{R}\left(\frac{\sigma_a^2}{\sigma_e^2}\right)\right]^{-1}(\mathrm{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad [52]$$

where **R** is a diagonal matrix accounting for heterogeneous residual variance. If $\hat{u}|\hat{a} = \mathbf{Z}\hat{a}$, replacing the first **G** by $\mathbf{Z}'$, weighted by the ratio of SNP to additive direct variances (k), would allow the calculation of SNP effects (a):

$$\hat{a} = \mathbf{Z}'\mathrm{k}\left[\mathbf{G} + \mathbf{R}\left(\frac{\sigma_a^2}{\sigma_e^2}\right)\right]^{-1}(\mathrm{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad [53]$$

where k is σ_a^2/σ_u^2, and σ_a^2 is the SNP variance. As we saw before, $\sigma_a^2 = \sigma_u^2/2\sum p_i(1-p_i)$. Therefore, k can be reduced to $1/2\sum p_i(1-p_i)$.
Assuming that:

$$w = \left[\mathbf{G} + \mathbf{R}\left(\frac{\sigma_u^2}{\sigma_e^2}\right)\right]^{-1}(\mathrm{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad [54]$$

then,

$$\hat{a} = \mathrm{k}\mathbf{Z}'\mathrm{w} \qquad [55]$$

and therefore,

$$\hat{u} = \mathbf{G}\mathrm{w} \qquad [56]$$

In this way,

$$\mathrm{w} = \mathbf{G}^{-1}\hat{u} \qquad [57]$$

Finally, the SNP effects can be calculated as:

$$\hat{a} = \mathrm{k}\mathbf{Z}'\mathbf{G}^{-1}\hat{u} \qquad [58]$$

as $Var(a) = \mathbf{D}$, the conditional mean of SNP effects given the GEBV is:

$$\hat{a}|\hat{u} = \mathrm{k}\mathbf{D}\mathbf{Z}'\mathbf{G}^{-1}\hat{u} \qquad [59]$$

Thus, given GEBV from ssGBLUP are available, SNP effects are calculated as (Wang et al., 2012):

$$\hat{a} = \mathrm{k}\mathbf{D}\mathbf{Z}'\mathbf{G}^{-1}\hat{u} \qquad [60]$$

Calculating SNP weights or variances based on GBLUP-based methods is an iterative process. 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**'/k (k scaling factor; details above)
3. Run single-step

4. Convert GEBV into SNP effects $\hat{a} = \text{k}\mathbf{DZ}'\mathbf{G}^{-1}\hat{u}$
5. Calculate SNP weight/variance for SNP i as $d_i = 2p_i q_i a_i^2$
6. Normalize **D** for the same sum(**D**)
7. Go to step 2

Usually, the best weights are obtained after 1-2 rounds. Step 3 can be done once (if changes in GEBV are small across rounds) or every round. Different formulas can be used to calculate SNP weight/variance (e.g., $d_i = a_i^2$). Variances for each SNP can be plotted as a Manhattan Plot; however, no significance test has been developed so far for ssGBLUP or GBLUP. A threshold of 1% of genetic variance can be assumed if the objective is to explore associations between traits and regions in the genome, like in genome-wide association studies (GWAS).

In tests using simulated data sets the estimates of SNP effects were similar to those by BayesB (Wang et al., 2012). 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.

Weighting SNP seems to improve the accuracy of predicting GEBV for data sets with small number of genotyped animals, but marginal or no improvement is observed for large genotyped populations (> 10k; Lourenco et al., 2017), even for less polygenic traits.

3. Validating genomic predictions

Model validation became very popular in animal breeding and genetics with the adoption of genomics. The idea was to compare how superior to EBV the GEBV could be in predicting future performance or future breeding value before the animals have their own/progeny records. For validation, we split the population in two: training and validation. The training population contains informative animals whose phenotypes, pedigree, and genotypes will be used for calculating GEBV and/or SNP effects. The validation population contains animals whose GEBV will be predicted based on their genotypes and pedigree, but no phenotypes, in addition to the information obtained from the training population. As predicting the future is quite vague, we instead, go back and hide phenotypes for a portion of the genotyped animals (validation animals).

If we are using multistep methods, SNP effects are calculated based on the training population and predictions on training animals are obtained as $u = \mathbf{Z}a$.

In single-step, training and validation populations are considered together; however, validation animals have only genotypes and pedigree in the system. Usually, three analyses are needed:

I. Traditional evaluation (BLUP) with complete dataset (pedigree, genotypes, and phenotypes for all animals). This will be the benchmark that we want to reach.

II. Traditional evaluation (BLUP) with reduced dataset (i.e., no phenotypes for validation animals);
III. Genomic evaluation (ssGBLUP) with reduced dataset.

Several ways to perform validation are available and the choice depends on the structure of the dataset and what the benchmark is:

1) *Validation on EBV*: when the benchmark is assumed to be a high accuracy EBV. This type of validation is common when the validation animals have their own records in addition to performance records (e.g., swine populations). The accuracy of predicting the high accuracy EBV (TBV*) is calculated as:

$$\text{Accuracy of EBV} = \text{Cor}(\text{TBV}^*, \text{EBV})$$
$$\text{Accuracy of GEBV} = \text{Cor}(\text{TBV}^*, \text{GEBV})$$

Inflation can be assessed as the regression (b_1) coefficient of TBV* on GEBV, whereas the intercept (b_0) can be considered as a bias estimator; R^2 is reliability or accuracy squared:

$$\text{TBV}^* = b_0 + b_1(\text{G})\text{EBV}$$

2) *Predictive ability (predictivity or predictability)*: when the validation animals have only own records, the benchmark is the phenotype adjusted for fixed effects (Y_c). Note that phenotypes must be adjusted by the fixed effects calculated based on the complete dataset. This is because phenotypes and information on fixed effects for validation animals are only present in the complete dataset.

$$\text{Predictivity EBV} = \text{Cor}(Y_c, \text{EBV})$$
$$\text{Predictivity GEBV} = \text{Cor}(Y_c, \text{GEBV})$$

To put predictivity in the accuracy scale, we divide it by square root of heritability (h). However, this approach can be problematic for traits with very low heritability, given that values can go out of the range for accuracy (0 –1).
This type of validation is common in beef cattle and chicken datasets.

3) *Interbull validation (mostly used in dairy cattle)*: in dairy cattle, the interest is to evaluate bulls but records for most of the traits are available only on females. For a bull to be progeny tested, it may take from 5 to 7 years, given his daughters have to calve for milk production traits started being recorded. In this validation method we remove phenotypes for one generation (~5 years); therefore, validation bulls have no daughters with records in the reduced dataset, but at least *n* (e.g., 10) daughters with records in the complete dataset.

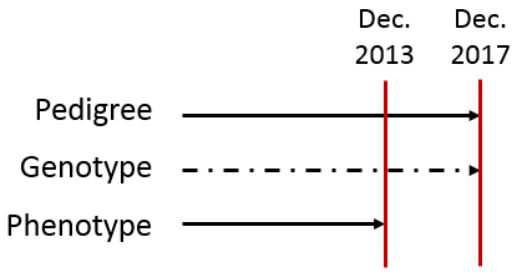


Illustration of a reduced dataset used for the Interbull validation test.

The benchmark in this case is a measure that condenses the progeny information as daughter yield deviation (DYD) or deregressed proof. DYD is the weighted average of yield of all progeny of a bull adjusted for all fixed effects and breeding values of the mates of the bull (Mrode, 2014). The calculation of DYD can be unstable and quite complicated sometimes. VanRaden (2008) suggested a simple way to obtain deregressed proof for bulls (we also call it daughter deviation; DD)

$$DD_{complete} = \frac{EBV_{complete} - PA_{complete}}{R_{complete}} + PA_{complete} \qquad [61]$$

where the fraction is the deregressed Mendelian sampling; Rcomplete is the deregression factor obtained by:

$$R_{complete} = \frac{DE_{progeny}}{DE_{progeny} + DE_{PA} + 1} \qquad [62]$$

$$DE_{PA} = \frac{REL_{PA}}{1 - REL_{PA}} \qquad [63]$$

$$DE_{progeny} = \frac{REL_{EBV}}{1 - REL_{EBV}} - DE_{PA} \qquad [64]$$

where $DE_{progeny}$ is equivalent contributions from progeny (daughters), DE_{PA} is equivalent daughter contributions from parent average (PA), REL is reliability of EBV (as a measure of error associated with EBV prediction). The subscript "complete" indicates that the values were calculated from the complete data set.

In dairy cattle, the measure of predictivity is reliability (R^2). In this way,

$$DD_{complete} = b_0 + b_1(G)EBV$$

Inflation can be assessed as the regression (b_1) coefficient of TBV* on GEBV, whereas the intercept (b_0) can be considered as a bias estimator; R^2 is reliability or accuracy squared.

Note that there are 2 purposes/notions of accuracy or reliability (Bijma, 2012):

a) Correlation between the "true EBV" and the EBV for selection candidates, which is a population parameter and reflects the response to selection. This value is reduced when selection takes place.
b) Stability of EBV, which is calculated for each individual based on standard errors. This value does not account for selection.
c) It is not possible to obtain a single measure that would reflect correlation between the true and estimated breeding value, and the standard error when the population is under selection.

4) K-fold cross validation: it can be used when the validation animals have only own phenotypes or when all individuals are from a single generation. In this method, the genotyped population is randomly divided into *k* subsets, and phenotypes are removed from one subset a time. Every time phenotypes are removed from one subset, we need to run reduced BLUP and ssGBLUP (or other genomic method).

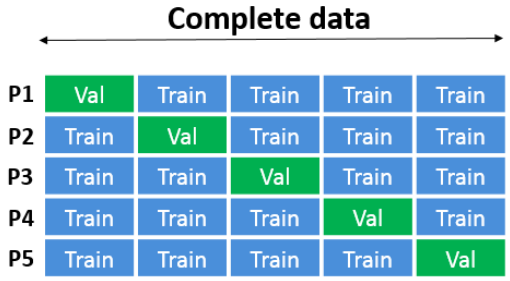


Illustration of a 5-fold cross validation scheme (Fragomeni et al., 2016)

Predictivity for each fold is calculated as below. The final predictivity is the average of all folds.

$$\text{Predictivity EBV} = \frac{\sum_{i=1}^{k} Cor(Y_c, EBV)}{k}$$

$$\text{Predictivity GEBV} = \frac{\sum_{i=1}^{k} Cor(Y_c, GEBV)}{k}$$

This type of validation can give extremely high predictivity, especially in small populations where training and validation populations are highly related. When several generations of data are available, the training population may contain progeny and the validation population may contain parents, which does not make sense.

References

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

Andersson, L. 2001. Genetic dissection of phenotypic diversity in farm animals. Nat. Rev. Genet. 2: 130-138.

Bijma, P. 2012. 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. 129:345-358.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42: 2.

Fernando, R. L., and M. Grossman. 1989. Marker-assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21: 467-477.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. 52: 399-433.

Henderson, C.R. 1976. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics 32:69-83.

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

Legarra, A., O. F. Chistensen, I. Aguilar, and I. Misztal. 2014. Single step, a general approach for genomic selection. Livest. Prod. Sci. 166:54-65.

Legarra, A. 2017. Bases for genomic prediction. http://genoweb.toulouse.inra.fr/~alegarra/GSIP_git.pdf

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. 91: 4090-4098.

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

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. Amen, L. Wang, D. W. Moser, and I. Misztal. 2015b. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. Journal of Animal Science 93: 2653-2662.

Lourenco D.A.L, Fragomeni, B.O., Bradford, H.L., et al. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. J. Anim. Breed. Genet. 134:46-471.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. VanTassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One 4: e5350.

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.

Meuwissen, T. H. E., B. Hayes, and M. Goddard. 2016. Genomic selection: A paradigm shift in animal breeding. Anim. Front. 6: 6-14.

Misztal, I., and J. K. Bertrand. 2008. Challenges of training quantitative graduate students. J. Anim. Sci. 86 (E-Suppl. 1): 165.

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

Mrode, R. 2014. Linear models for the prediction of animal breeding values. CABI - 3rd ed. 343p.

Page, B. T., E. Casas, M. P. Heaton, N. G. Cullen, D. L. Hyndman, C. A. Morris, A. M. Crawford, T. L. Wheeler, M. Koohmaraie, J. W. Keele, and T. P. Smith. 2002. Evaluation of single-nucleotide polymorphisms in CAPN1 for association with meat tenderness in cattle. J. Anim. Sci. 80: 3077-3085.

Quaas, R.L. 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71:1338-1345.

Saatchi, M., M. C. McClure, S. D. Mckay, M. M. Rolf, J. Kim, J. E. Decker, T. M. Taxis, R. H. Chapple, H. R. Ramey, S. L. Northcutt, S. Bauck, B. Woodward, J. C. M. Dekkers, R. L. Fernando, R. D. Schnabel, D. J. Garrick, and J. F. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Gen. Sel. Evol. 43: 40.

Schork, N. J., D.Fallin, and S. Lanchbury. 2000. Single nucleotide polymorphisms and the future of genetic epidemiology. Clin. Genet. 58: 250-264.

Snelling, W. M., M. F. Allan, J. W. Keele, L. A. Kuehn, R. M. Thallman, G. L. Bennett, C. L. Ferrell, T. G. Jenkins, H. C. Freetly, M. K. Nielsen, and K. M. Rolfe. 2011. Partial-genome evaluation of postweaning feed intake and efficiency of crossbred beef cattle. J. Anim. Sci. 89: 1731-1741.

Soller, M., and J. S. Beckmann. 1983. Genetic polymorphism in varietal identification and genetic improvement. Theor. Appl. Genet. 67: 25-33.

Stonecking, M. 2001. From the evolutionary past... . Nature 409: 821-822.

The Bovine Genome Sequencing and Analysis Consortium, C. G. Elsik, R. L. Tellam, and K. C. Worley. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324: 522-528.

The International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.

VanRaden, P. M. and G. R. Wiggans. 1991. Deviation, calculation, and use of national animal model information. J. Dairy Sci. 74:2737-2746.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

VanRaden, P. M., C. P. VanTassel, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16-24.

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.

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.