Bases for Genomic Prediction

Andrés Legarra

INRA, Animal Genetics Department,

UMR 1388 GenePhySE,

31326 Castanet Tolosan, France

Version 0.9.1

February 2015

Foreword

This is an incomplete attempt to write a comprehensive review of principles for genomic predictions. The framework is proudly parametric and tries to follow classical quantitative genetics and statistical theory as much as possible. It is incomplete: the wealth of papers being generated makes impossible to follow all the literature. I express my apologies for the resulting self-centered bias.

My own knowledge on the topic owes much to dozens of colleagues with whom I have much worked and discussed. I explicitly thank Ignacy Misztal, Ignacio Aguilar, and all my collaborators for so much joint work and discussion. Financing for these notes was possible by the INRA metaprogram SelGen. They were written in May 2014, during a visit to the University of Georgia (UGA), kindly hosted by Ignacy Misztal; during this visit we taught a course whose material (slides, exercises, and these notes) can be found at http://nce.ads.uga.edu/wiki . Updated versions of these notes can be found at that web page and also at http://genoweb.toulouse.inra.fr/~alegarra. I thank Llibertat Tusell and Paul VanRaden for corrections and comments.

I deeply thank all those people that have produced and made available notes and courses, which have been so useful for me during the years.

Yo no te buscaba y te vi.

September 2014. A large number of mistakes and typos have been corrected.

February 2, 2015. More corrections and few suggestions by Llibertat Tusell and Paul VanRaden.

Table of Contents:

1. Main notation

$\boldsymbol{X}, \boldsymbol{b}$	Incidence matrix of fixed effects and fixed effects
$\boldsymbol{Z}, \boldsymbol{a}$	Incidence matrix of marker effects and marker effects
$\boldsymbol{W}, \boldsymbol{u}$	Incidence matrix of polygenic (or individual) effects and polygenic effects
$\boldsymbol{D}$	Covariance matrix of marker effects, $Var(\boldsymbol{a}) = \boldsymbol{D}$
σ^2_{ai},	Variance of the marker effect a_i
σ^2_{a0}	Variance of marker effects if *all* had the same variance
σ^2_u	Genetic variance
σ^2_e	Residual variance
$\boldsymbol{G}$	Genomic relationship matrix
p_i	Allele frequency at marker i
$\boldsymbol{A}$	Pedigree-based relationship matrix

2. A quick tour of Linkage Disequilibrium

The aim of this section is not really to make a full description, which is beyond the scope of these notes, but to give a few concepts that might be of relevance for practitioners.

In a genome there are many *loci* and loci have *alleles*. In a population, there is a certain distribution of alleles *within* a locus but also *across* loci. This distribution can be described by a regular table. For instance, assume two biallelic loci and that we have 5 individuals, and therefore 10 gametes in our population:

$$\{AB, AB, ab, aB, ab, ab, Ab, AB, Ab, AB\}$$

You may call this haplotypes, diplotypes, or genotypes of the gametes. Allelic frequencies within loci are: $p_1 = freq(A) = 0.6$; $p_2 = freq(B) = 0.5$. A frequency table of these diplotypes is as follows:

Table 1. Example of two loci in Linkage disequilibrium

	A	a
B	0.4	0.2
B	0.1	0.3

The eye sees that allele "A" comes most often associated with "B". But is this any relevant? Does the presence of "A" give any clue on the presence of "B"?

Linkage equilibrium is a common assumption, where alleles across loci are distributed at random. For instance, $freq(AB) = freq(A) \times freq(B) = 0.30$. If these were the case, the table should be as follows:

Table 2. Frequency table if the two loci were in Linkage equilibrium

	A	a
B	0.3	0.2
B	0.3	0.2

Linkage disequilibrium (LD) is the event of non-random association of alleles across loci, and it means that the "observed" table deviates from the "expected" table. The reason why linkage disequilibrium is formed is because some "chunks" of chromosomes are overrepresented in the population and never break down, and this is basically due to finite size of the population (drift, selection) and also to mutation. For instance, consider a cross of two inbred lines and successive F_1, $F_2 \ldots F_n$ generations. At the end, the chromosomes become a fine-grained mosaic of grey and black. However, complete mixture is difficult to attain.

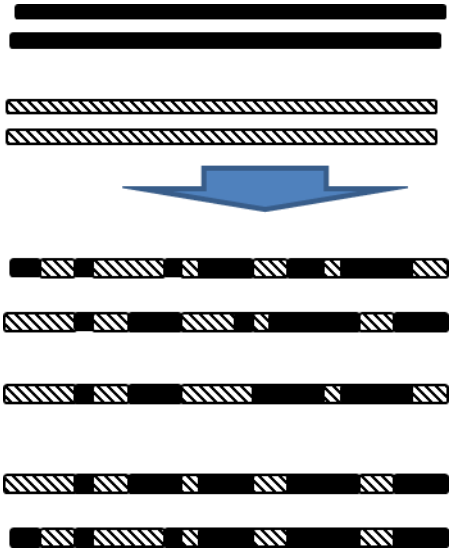


Figure 1. Chunks of ancestral chromosomes after cross of pure lines and several generations.

Linkage disequilibrium describes not-random association of two loci. Nothing more, so, why is it useful? In practice, two loci in LD most often are (very) close. This is because LD breaks down with recombination. Therefore, Linkage disequilibrium of two loci decays on average with the distance, and it serves to map genes. In other words, one loci is a proxy for the other one, and this is why association analysis uses linkage disequilibrium to map genes.

2.1 Within-family and populational linkage disequilibrium

If we study the distribution of alleles within a family (say parents and offspring) we will verify that the linkage disequilibrium is very strong. This is because the chromosomes of the parents are almost completely conserved, because there are very few recombinations in one generation time. Consider for instance the following two sires, and a recombination fraction of 0.25 across the two loci:

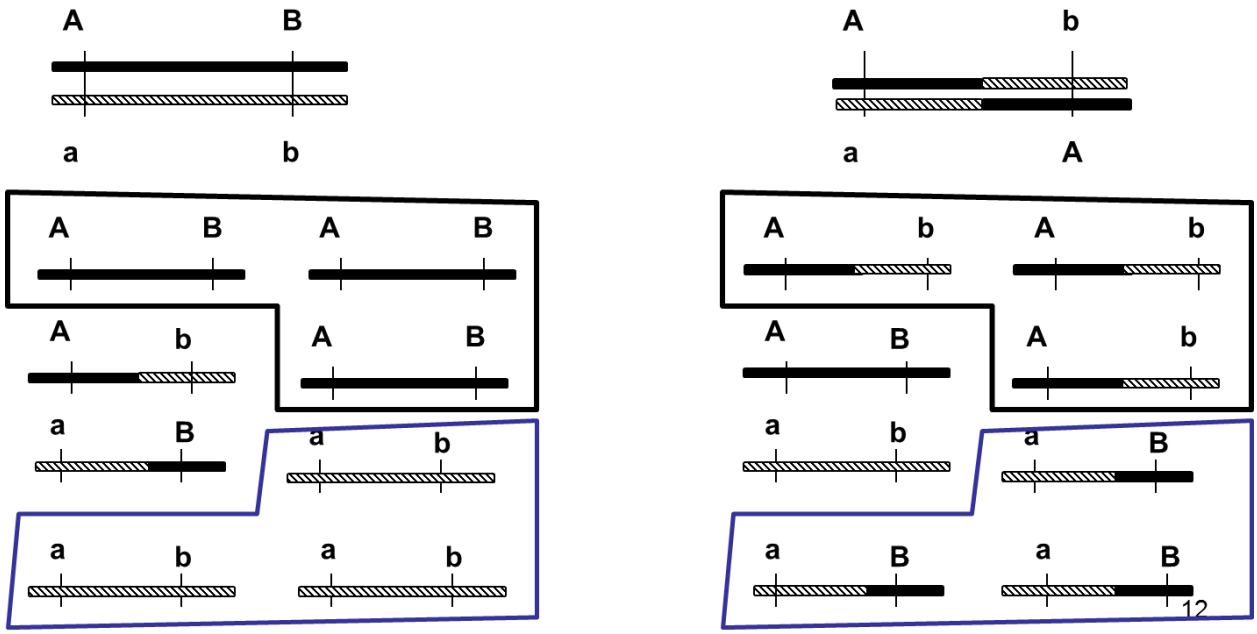


Figure 2. Two sires and eight progeny, where each family shows linkage disequilibrium but there is no populational linkage disequilibrium

Individually considered, the two families have strong within-family linkage disequilibrium, but the population of 16 offspring seen as a whole does not have linkage disequilibrium.

However, populations are large families. Therefore there will be linkage disequilibrium across loci if we look at distances short enough. In general, short-distance linkage disequilibrium reflects old relationships and large-distance linkage disequilibrium reflects recent relationships (SVED 1971; TENESA *et al.* 2007).

2.1.1 Why QTL are easier to trace within family

Now imagine that locus A/a was a QTL with effects of, say, $\{+10, -10\}$ and locus B/b was a genetic marker. It is very easy to trace the QTL within each family, but the information within families are contradictory. Locus B/b would have apparent effects of $\{5, -5\}$ in family one but $\{-5,5\}$ in family two. This can be explained as follows. The four chromosomes carriers of locus B in family one carry three copies of allele A and one copy of allele a. Therefore the apparent effect of allele B is equal to $\frac{(3\times10+1\times(-10))}{4} = 5$, in family one. In family two this is exactly the opposite: $\frac{(1\times10+3\times(-10))}{4} = -5$, and across all families, Locus B/b would have an effect of $\frac{(2\times10+2\times(-10))}{4} = 0$. Therefore allele B is a good predictor both within families 1 and 2, but not across families.

2.2 Quantifying linkage disequilibrium

There are two classical measures. D measures the deviation from *observed* distribution to *expected* distribution:

$$D = freq(AB) - freq(A)freq(B)$$

(HILL and ROBERTSON 1968) proposed, for biallelic loci, to assign numerical values to loci (i.e., $\{A, a\}$ would be $\{0,1\}$ and $\{B, b\}$ would be $\{0,1\}$) and compute Pearson's correlation across loci. In the preceding example, genotypes $\{AB, AB, ab, aB, ab, ab, Ab, AB, Ab, AB\}$ can be written as two variables, one for "A", $X = \{1,1,0,0,0,0,1,1,1,1\}$ and one for "B", $Y = \{1,1,0,1,0,0,0,1,0,1\}$ and therefore $r = 0.41$. It can be shown that $r = \frac{D}{\sqrt{p_A q_A p_B q_B}}$ where $p_A = 1 - q_A = freq(A)$. It has the advantage that r^2 is related to the variance in locus A explained by locus B, and of being easier to understand than D. Both D and r depend on the reference allele but r^2 is invariant to the reference allele.

3. Bayesian inference

Bayesian inference is a form of statistical inference based on Bayes' theorem. This is a statement on conditional probability. We know that

$$p(A,B) = p(A\,|B)p(B) = p(B|A)p(A)$$

Bayes' theorem says that

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

The algebra is valid for either a single-variable A and B or for A and B representing a collection of things (e.g., A can be thousands of phenotypes and B marker effects and variance components).

Its use in statistical inference is as follows. We want to infer values of B knowing A. For every value of B we do the following:

1. We compute $p(A|B)$, which is the probability, or likelihood, of A had we know B.
2. We multiply this probability by the "prior" probability of B, $p(B)$.
3. We cumulate $p(A|B)p(B)$ to form $p(A)$, which is called the *marginal density of A*.

3.1 Example of Bayesian inference

Assume that we have a collection of quantitative phenotypes $\boldsymbol{y} = \{1{,}0, -0.99\}$ with k=3 records and a very simple model $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{e}$ with $Var(\boldsymbol{e}) = \boldsymbol{R} = \boldsymbol{I}\sigma_e^2$ and $\sigma_e^2 = 1$. We will infer μ based on Bayes' theorem; actually, we will infer a whole distribution for μ, what is called the *posterior distribution*, based on

$$p(\mu|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mu)p(\mu)}{p(\boldsymbol{y})}$$

Where

$$p(\boldsymbol{y}|\mu) = MVN(\,\mu, \boldsymbol{I}) = \frac{1}{\sqrt{2\pi^k}|\boldsymbol{R}|}\exp\left(-\frac{1}{2}(\boldsymbol{y} - \mathbf{1}\mu)'\boldsymbol{R}^{-1}(\boldsymbol{y} - \mathbf{1}\mu)\right)$$

is the "likelihood" of the data for a given value of μ.

However, it is unclear what $p(\mu)$ means. This is usually interpreted as a *prior* distribution for μ, which means that we must give probability values to each possible value of μ. These probabilities may come from previous information or just from mathematical or computational convenience, but they must *not* come from the data $\boldsymbol{y}$. Prior distributions require a mental exercise of thinking if μ has been "drawn" from some distribution (e.g., it is a particular farm among a collection of farms), or if there are biological laws that impose prior information – for instance, the infinitesimal model suggests normal distribution for genetic values. If this is the case, such an effect is often called "random" in the jargon.

Finally, $p(\boldsymbol{y})$ is the probability of the data if we average $p(\boldsymbol{y}|\mu)$ across all possible values of μ, weighted by its probability $p(\mu)$.

Consider that there are only two possible values of μ, -1 and 1 with equal *a priori* probabilities of 0.5 and 0.8. Then we can create this table:

	$p(\mu)$	$p(\boldsymbol{y}\|\mu)$	$p(\boldsymbol{y}\|\mu)p(\mu)$	$p(\mu\|\boldsymbol{y}) = \frac{p(\boldsymbol{y}\|\mu)p(\mu)}{p(\boldsymbol{y})}$
$\mu = -1$	0.5	0.051	0.0255	0.40
$\mu = 1$	0.5	0.076	0.0381	0.60
$p(\boldsymbol{y})$			0.05	

So, the final result is that the mean μ has a value of either -1 (with *posterior* probability 0.40) or 1 (with *posterior* probability 0.60). The *posterior expectation* of the mean is $E(\mu|\boldsymbol{y}) = \dot{1} \times\ 0.60 + -1 \times 0.40 = 0.20$.

If the prior distribution for the mean is continuous, for instance $N(0, \sigma_\mu^2)$ (say $\sigma_\mu^2 = 10$) , then the final distribution of μ is continuous as well. Therefore, it is impossible to enumerate all cases as above. In the case that the prior distribution is normal and the likelihood too, the posterior distribution can be derived analytically (e.g. in (SORENSEN and GIANOLA 2002)) and is

$$p(\mu|\boldsymbol{y}) = N(\hat{\mu}, lhs^{-1})$$

Where

$$lhs = \frac{\mathbf{1}'\mathbf{1}}{\sigma_e^2} + \frac{1}{\sigma_\mu^2}$$

$$\hat{\mu} = (lhs^{-1})\mathbf{1}'\boldsymbol{y}/\sigma_e^2$$

So, $\hat{\mu} = 0.064$ on average with a standard deviation of 0.57.

3.2 The Gibbs sampler

Things get more complicated when we have several unknowns in our model. For instance, we might not know the residual variance σ_e^2, so we want to evaluate

$$p(\mu, \sigma_e^2|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mu, \sigma_e^2)p(\mu)p(\sigma_e^2)}{p(\boldsymbol{y})}$$

. Writing down in closed form the posterior distributions is impossible. The Gibbs sampler is a numerical MonteCarlo technique that allows drawing samples from such a distribution. The idea is as follows. If we knew μ, then we could derive the posterior distribution of σ_e^2. If we knew σ_e^2, then we could derive the posterior distribution of μ. These distributions "pretending that we know" are known as *conditional distributions*, and need to be known up to proportionality (this makes algebra less miserable). In our example they are:

$$p(\sigma_e^2|\boldsymbol{y}, \mu)$$
$$p(\mu|\boldsymbol{y}, \sigma_e^2)$$

If these distributions are known, we can draw successive samples from them and then plug these samples into the right hand side of the expressions, "as if" they were true, and iterate the procedure. So we start with, say, $mu = 0$ and $\sigma_e^2 = 1$. Then we draw a new μ from

$$p(\mu|\boldsymbol{y},\sigma_e^2) = N(\hat{\mu}, lhs^{-1})$$

Then σ_e^2 from

$$p(\sigma_e^2|\boldsymbol{y},\mu) = (\boldsymbol{y}-\mathbf{1}\mu)'(\boldsymbol{y}-\mathbf{1}\mu)\chi_k^{-2}$$

Which is the conditional distribution assuming flat priors for σ_e^2. Then we plug in this value into $p(\mu|\boldsymbol{y},\sigma_e^2)$ and we iterate the procedure. After a period, the samples so obtained are from the posterior distribution. Typically thousands of iterates are needed, if not more. The following R code shows a simple simulated example.

```
set.seed(1234)
# simulated n data with mean 100 and residual variance 20
ndata=10
y=100+rnorm(ndata)*sqrt(20)
# Gibbs sampler
#initial values
mu=-1000
vare=10000
varmu=1000
#place to store samples
mus=c()
vares=c()

#sampling per se
for (i in 1:50){
  lhs=ndata/vare+1/varmu
  rhs=sum(y)/vare
  mu=rnorm(1,rhs/lhs,sqrt(1/lhs))
  vare=sum((y-mu)**2)/rchisq(1,ndata)
  cat(mu,vare,"\n")
  mus=c(mus,mu)
  vares=c(vares,vare)
}
```

The beauty of the system of inference is that we decompose a complex problem in smaller ones. For instance, variance component estimation proceeds by sampling breeding values (as in a BLUP "with noise", Robin Thompson *dixit*), and then sampling variance components are estimated as if these EBV's were true.

3.2.1 Post Gibbs analysis

A Gibbs sampler is not converging to any final value, like REML, in which each iterate is better than the precedent. Instead, at the end we have a collection of samples as follows:

```
Mu       vare
38.47288 6832.21
76.12334 323.1892
85.76835 267.1094
91.08181 120.2974
100.1114 19.85989
98.52846 19.85005
98.03879 14.52127
97.54579 20.33205
98.10108 14.76999
99.39184 6.538137
96.90541 13.92563
```

. . .

and these samples define the posterior distribution of our estimator.

The first point is to verify that the chain has converged to the desired posterior distribution. Informal testing plots are very useful. For instance, `plot(vares)` in the above example shows that initial values of σ_e^2 where out of the desired posterior distribution. We can discard some initial values and then keep the rest.

We need to report a final estimate, e.g., of σ_e^2 from this collection of samples. Contrary to REML, the last sample of σ_e^2 is *not* the most exact one, but is all the collection of samples which is of interest, because they approximate the posterior distribution of the estimator. So, a typical choice is the *posterior mean*, which is the average of the samples. In the example above you can for instance discard the first 20 iterations as burn-in and then use the posterior mean across the last 30 samples of the residual variance:

```
> mean(vares[21:50])
[1] 20.28395
```

Which is very close to the simulated value of 20. The post-Gibbs analysis is clumsy but important and packages such as BOA exist in R to simplify things.

4. Models for genomic prediction

4.1 Simple marker model

Assume there is a marker in complete, or even incomplete, LD with a QTL. For example, the polymorphism in the halothane gene (HAL) is a predictor of bad meat quality in swine. The simplest way to fit this into a genetic evaluation is to estimate the effect of the marker by a linear model and least squares:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{marker} + \boldsymbol{e}$$

Where in "marker" we actually introduce a marker with alleles and their effects. More formally, allele effects are embedded in vector $\boldsymbol{a}$ and their incidence matrix is in matrix $\boldsymbol{Z}$:

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

For instance, assume that we have a four-allele $\{A, B, C, D\}$ locus and three individuals with genotypes $\{BC, AA, BD\}$. Then

$$\boldsymbol{Za} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \\ a_C \\ a_D \end{pmatrix}$$

Note that we have put a 2 for the genotype "AA". This means that the effect of a double copy of "A" is twice that of a single copy. This is an *additive model*.

And for $\boldsymbol{y} = \{12,35,6\}$ this gives

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

$$\begin{pmatrix} 12 \\ 35 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 0 & 1 & 1 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \\ a_C \\ a_D \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

Assume now that we do the same with a simple, biallelic marker (say $\{A, B\}$). Consider three individuals with genotypes $\{BB, AA, BA\}$:

$$\boldsymbol{Za} = \begin{pmatrix} 0 & 2 \\ 2 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \end{pmatrix}$$

and

$$\begin{pmatrix} 12 \\ 35 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 0 & 2 \\ 2 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

However, because there is redundancy (if the allele is not A, then it is B) it is mathematically equivalent to prepare a regression of the trait on the number of copies of a single allele, say A. So for individuals $\{BB, AA, BA\}$ we have that

$$\boldsymbol{Za} = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix} a_A$$

and

$$\begin{pmatrix}12\\35\\6\end{pmatrix}=\begin{pmatrix}1\\1\\1\end{pmatrix}\mu+\begin{pmatrix}0\\2\\1\end{pmatrix}a_A+\begin{pmatrix}e_1\\e_2\\e_3\end{pmatrix}$$

The effect of the marker can be estimated by least squares or another regression method. The marker should explain a large part of the variance explained by the gene. The model can be enriched by adding an extra infinitesimal term $\boldsymbol{u}, Var(\boldsymbol{u}) = \boldsymbol{A}\sigma_u^2$, like for instance in

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{Wu} + \boldsymbol{e}$$

4.2 Why markers can't be well chosen: lack of power and the Beavis effect

The method above can be potentially extended to more markers explaining the trait. However, the failure of this method resides in that *we do not know* which markers are associated to the trait. This is a very serious problem, because finding out which markers are linked to a trait generally induces lots of errors – and this because of the nature, and because of the Beavis effect.

Genetic background of complex traits seems to be highly complex and largely infinitesimal: many genes acting, possibly with interactions among them, to give the genetic determinism of one trait. Most of them bearing small effects, some may have large effects. Current alternatives for localization of genes include genomewide association studys (GWAS). This consists in testing, one at a time, markers for its effect on a trait, mostly with a simple linear model as above. The procedure selects those markers with a significant effect after a statistical test, for instance a t-test. This test is usually corrected by Bonferroni to avoid spurious results. However, this way of proceeding leads to lack of power and bias. This will be shown next.

4.2.1 Lack of power

This is because a small effect can rarely be detected. The general formulae for power can be found in, e.g., (LUO 1998) and are implemented in R package `ldDesign`. A very simple version of the formulae for power where the causal variant is truly tagged by a marker is (I owe this expression to Anne Ricard)

$$power = 1 - \Phi\left(Z_{1-\frac{\alpha}{2}} - \beta\sqrt{2pq(n-2)}\right)$$

with $Z_{1-\alpha/2}$ the rejection threshold, that is ≈ 4.81 after Bonferroni correction for 50,000 markers. For instance, in a population of n=1000 individuals, a QTL explaining 1% of the variance and perfectly tagged by a marker will be found 4% of the time. If 100 such QTLs exist in the population, only 4 of them will be found. The following Figure shows the power of detection of a QTL perfectly tagged explaining from 0 to 100% of the phenotypic variance.

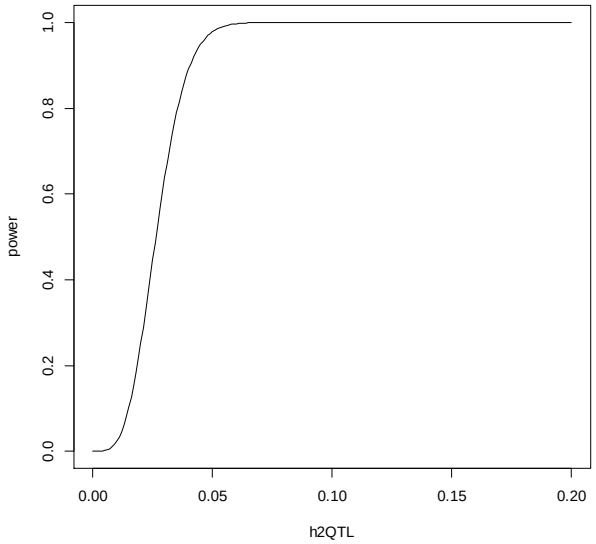


Figure 3. Power of detection of QTL effects perfectly tagged explaining from zero to 100% phenotypic variance.

4.2.2 The Beavis (or winner's curse) effect

This comes as follows. We are mapping QTLs. To declare a QTL in a position, we perform a test (for example a t-test). This test depends on the estimated effect of the QTL, but

$$estimated\ effect\ =\ real\ effect\ +\ «\ estimation\ noise\ ».$$

By keeping selected QTLs, we often keep large and positive noises. This is negligible if there were few QTLs with large effects but this is not the case. Large noises will occur in analysis with many markers, and this biases the estimated QTL effect, making it look much larger than real, in particular if they are small. The problem is exacerbated with GWAS approaches, because of testing many markers.

For instance, assume that a marker with allelic frequency $p = 0.5$ truly explains 5% of the variance. Using formulae in Xu (2003), the variance explained by this marker will be overestimated and show up as 5.1% at regular type-I error. This does not change for more strict Bonferroni-like tests, e.g., $\alpha = 0.05/50000$. However, for markers explaining 0.5% of the variance, the *apparent* variance explained is 0.9% (two times in excess) at $\alpha = 0.05$ and a formidable 2.7% at $\alpha = 0.05/50000$ (a 5-fold overestimation of the explained variance). Therefore, collecting 40 such significant markers may look like capturing all genetic variation whereas in fact they only capture 20% of the variance. The following R script allows these computations.

```
#beavis effect by Xu ,  2003, Genetics 165: 2259-2268
bias.beavis<- function(sigma2=1,n,p=.5,alpha,a){
        # this function computes real and apparent
        #(from QTL detection estimates) variance
        #explained by a biallelic QTL with effect a and
        # allelic frequency p at alpha risk
        #Andres Legarra, 7 March 2014
        gamma=2*p*(1-p)
        sigma2x=gamma
        eps1=-qnorm(1-alpha/2)-sqrt(n*gamma/sigma2)*a
        eps2= qnorm(1-alpha/2)-sqrt(n*gamma/sigma2)*a
        psi1=dnorm(eps1)/(1+pnorm(eps1)-pnorm(eps2))
        psi2=dnorm(eps2)/(1+pnorm(eps1)-pnorm(eps2))
        B=gamma*(sigma2/(n*sigma2x))*(1+eps2*psi1-eps1*psi2)
        var.explained=gamma*a**2
        var.attributed=var.explained+B
        att.over.exp=var.attributed/var.explained
        rel.var.explained=var.explained/sigma2
        rel.var.attributed=var.attributed/sigma2
        list(
                var.explained=var.explained,
                var.attributed=var.attributed,
                rel.var.explained=rel.var.explained,
                rel.var.attributed=rel.var.attributed,
                att.over.exp=att.over.exp
                        )
}
```

The following graph shows the true variance explained by the QTL and the variance *apparently* explained by the QTL, for QTL effects ranging from 0 to 0.5 standard deviations, i.e. explain up to 12% of the variance. It can be seen that small effects are systematically exaggerated.

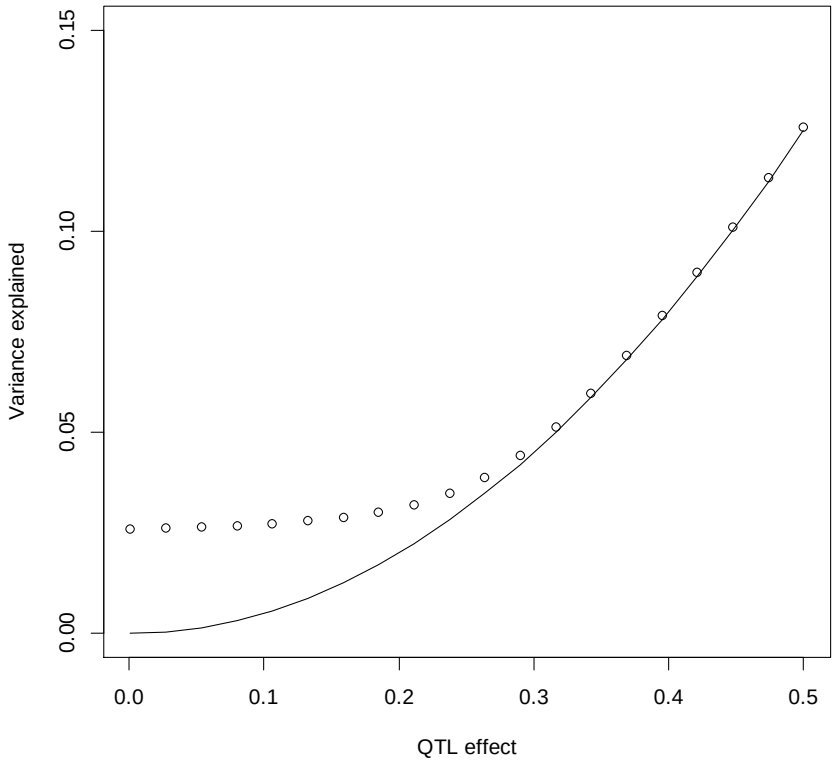


Figure 4. True (straight line) and apparent (dotted line) variance explained by QTL effects going from zero to 0.5 genetic standard deviations.

The two following graphs, from very crude simulations, show both problems. The first one shows no bias, but the second shows, first, that only 3 out of 100 QTL were found (lack of power), and those 3 found are largely overestimated (Beavis effect).

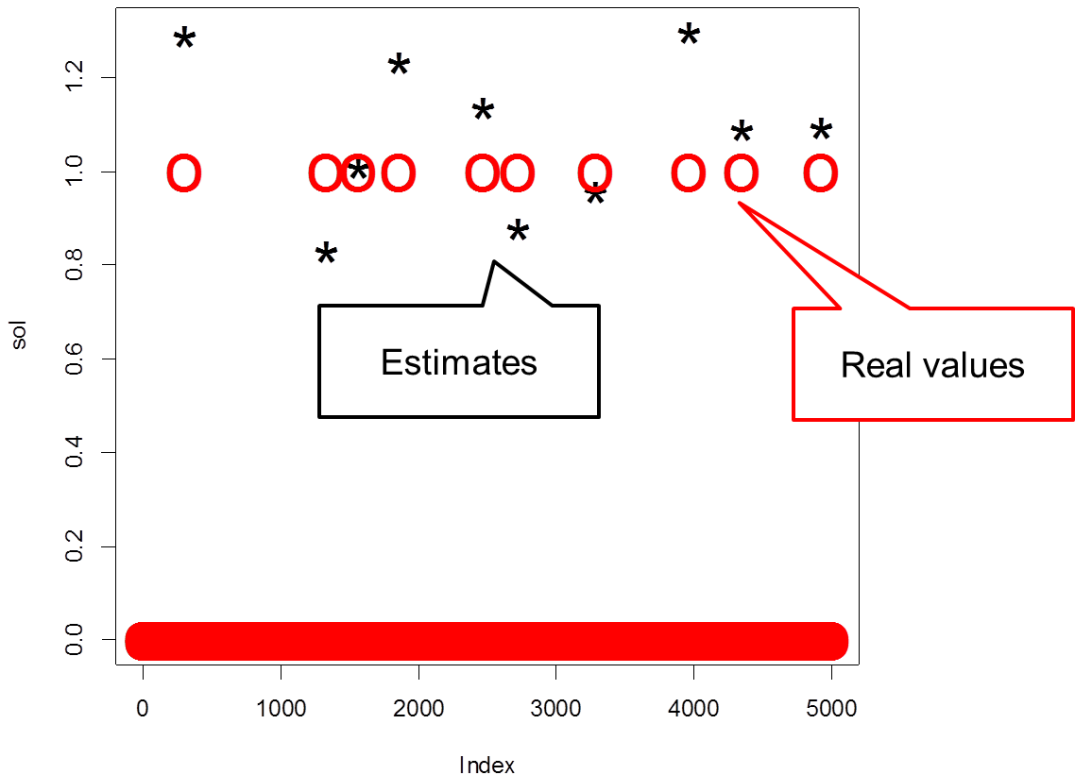


Figure 5. Real (**O**) and estimated (*****) effects after GWAS-like simulations with 10 true QTLs in 5000 markers, 1000 individuals.

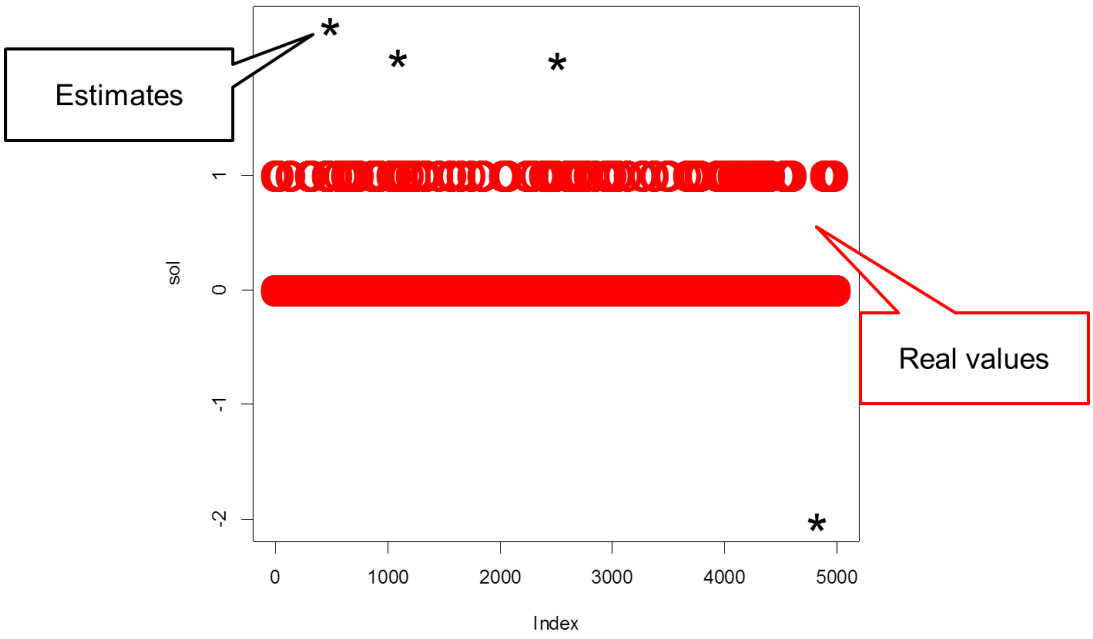


Figure 6. Real (O) and estimated (*) effects after GWAS-like simulations with 100 true QTLs in 50000 markers, 1000 individuals.

4.3 Fit all markers

Lande and Thompson (LANDE and THOMPSON 1990) suggested getting the list of associated markers and their effects from an independent population. Whereas this is typically done –now- in human genetics, it seems impossible to do in agricultural populations. First, the associations are random, and therefore markers associated in one population are not necessarily associated in another one. Second, even the true list of acting genes and QTL will vary across populations due to drift or selection.

These problems plague GWAS and QTL detection analysis. Further, nothing guarantees that markers with no effect at one stage will have no effect at another one, for instance, because of interactions. A simple way to avoid both the lack of power and the Beavis effect is *not to use detection thresholds*. Therefore *all markers are assumed to be QTL*. This simple idea gave (MEUWISSEN *et al.* 2001) the key to attack the estimation of whole genetic value based on markers. First, markers with small effects will be included. Second, no bias will be induced due to the detection process.

Therefore, one should include all markers in genomic prediction. In a way, this makes sense because we use all information without discarding anything. But how is this doable? The simplest is to fit a linear model with the effects of all markers. Note that for this approach to work, you need to cover all the genome; *many* markers are needed.

4.3.1 Multiple marker regression as fixed effects

The multiple marker regression is a simple extension of the single marker regression shown above. First, we construct a model were the phenotype is a function of *all* marker

effects:

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

For instance, assume that we have a four-allele $\{A, B, C, D\}$ locus, another locus with alleles $\{E, F\}$ and three individuals with genotypes $\{BC/EE, AA/EF, BD/FF\}$. Then

$$\boldsymbol{Za} = \begin{pmatrix} 0 & 1 & 1 & 0 & \vdots & 2 & 0 \\ 2 & 0 & 0 & 0 & \vdots & 1 & 1 \\ 0 & 1 & 0 & 1 & \vdots & 0 & 2 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \\ a_C \\ a_D \\ \cdots \\ a_E \\ a_F \end{pmatrix}$$

Again, estimation of $\boldsymbol{a}$ can proceed by least squares. However, this poses two kinds of problems. The first one is practical: we can't (reliably) estimate 50,000 effects from, say, 1,000 data in $\boldsymbol{y}$. The second is conceptual: does it make sense to estimate all these marker effects without imposing any constraints? In fact, one should not expect that a marker has a large effect; rather, we expect them to be restricted to plausible values. For instance, a marker should not have an effect of, say, one phenotypic standard deviation of the trait. In a way, this is an "a priori" information and there must be a way to introduce this information. But this introduces a very old subject of genetic evaluation: prediction. After explaining prediction, we will go back to models.

4.4 Bayesian Estimation, or Best Prediction, of marker effects

Marker effects can be considered as the result of random processes, because they are the result of random buildup of linkage disequilibrium, random generation of alleles at genes, and so on. Therefore they have (or may have) an associated distribution (whether you call this a sampling distribution or a prior distribution is largely a matter of taste). I will generally call this prior information. It is well known (CASELLA and BERGER 1990) that accurate prediction of random effects involves integration of all information, prior information and observed information, that in our case it comes in the form of observed phenotypes.

If we call $\boldsymbol{a}$ the marker effects, and $\boldsymbol{y}$ the data, the *Posterior Mean*, or *Conditional Expectation* of (estimators of) marker effects is given by the expression

$$\widehat{\boldsymbol{a}} = E(\boldsymbol{a}|\boldsymbol{y}) = \frac{\int \boldsymbol{a}\, p(\boldsymbol{y} \mid \boldsymbol{a}) p(\boldsymbol{a}) d\boldsymbol{a}}{\int\ p(\boldsymbol{y} \mid \boldsymbol{a}) p(\boldsymbol{a}) d\boldsymbol{a}}$$

We have already discussed the Posterior Mean in the introduction to Bayesian inference. This is often called as *Best Prediction*, because in a Frequentist context it does minimize, over conceptual repetitions of the procedure, the distance between "true" $\boldsymbol{a}$ and its estimator, $\widehat{\boldsymbol{a}}$ (CASELLA and BERGER 1990). On the other hand this can be seen as a Bayesian estimator as described above. This estimator has an extraordinary advantage over the regular least squares, because it uses all available information (GIANOLA and FERNANDO 1986). Further, it has been proven that Best Predictors are optimal for selection (COCHRAN 1951) (FERNANDO and GIANOLA 1986) (GOFFINET and ELSEN 1984). The introduction of the prior distribution $p(\boldsymbol{a})$ has an effect of "regressing" the estimators towards the *a priori* values, a process that is known as *shrinkage*. Therefore the Best Predictors are "shrunken" or "regressed" estimators.

In the context of genomic predictions, the Best Predictor is composed of two parts:

1. The prior distribution of marker effects $p(\boldsymbol{a})$
2. The likelihood of the data given the marker effects, $p(\boldsymbol{y} \mid \boldsymbol{a})$

Breeders have a fairly decent idea of how to write the latter, $p(\boldsymbol{y} \mid \boldsymbol{a})$. Most often this is written as a normal likelihood, of the form

$$p(\boldsymbol{y} \mid \boldsymbol{a}) = MVN(\boldsymbol{Xb} + \boldsymbol{Za}, \boldsymbol{R})$$

where matrix $\boldsymbol{R}$ contains residual covariances. The model may include further linear terms such as pedigree-based covariances, permanent effects, and so on. However, how to write down the prior distribution $p(\boldsymbol{a})$ is far from being clear, and this has been the subject of frantic research during the last decade. This will be part of the subject of the following sections

4.4.1 Best Predictions as a regularized estimator

Regularized predictors are much used now in Statistics. They are composed of two parts: a likelihood, and a regularization function which prevents the estimators from going "too far away". For instance, the regular Lasso (TIBSHIRANI 1996) can be understood as an estimator that uses a likelihood as above, combined with the restriction $|\boldsymbol{a}| < \lambda$. Another example is the Ridge Regression, where there is a penalty function of $(\boldsymbol{a}^2)$ The explanation of these estimators is largely practical. However, from the point of view of a Bayesian or a Frequentist (or an animal breeder), they are Bayesian (or Best Predictor) estimators with particular sampling or *a priori* distributions. For instance, the Lasso assumes that (marker) effects are *a priori* distributed following a Laplace (double exponential) distribution, and Ridge Regression assumes that effects are *a priori* normally distributed. A by and large advantage of this understanding is that it allows the connection between classical quantitative genetics theory and prior distributions for marker effects.

4.5 The ideal process for genomic prediction

We have prepared the conceptual setup. The process of genomic prediction consists in estimating marker effects using the Conditional Mean of marker effects as above, which is based on phenotypes at the trait(s) of interest and the prior distribution of marker effects. This creates a prediction equation which can be summarized as something like:

Table 3. A form of prediction equation.

Locus	Allele	Effects estimates
1	A	+10
	B	-8
	C	+2.3
	D	-12.5
2	E	+5
	F	-6

For the i-th individual, the product of its genotype (the i-th row, $\boldsymbol{z}_i$ of matrix $\boldsymbol{Z}$) and the alleles' effects (in $\widehat{\boldsymbol{a}}$) gives a genomic estimated breeding value, say $\hat{u}_i = \boldsymbol{z}_i\widehat{\boldsymbol{a}}$. This applies

equally well to animals with or without phenotype. The next section of these notes will describe how this can be accomplished through the so-called *Bayesian regressions.*

Process of genomic prediction

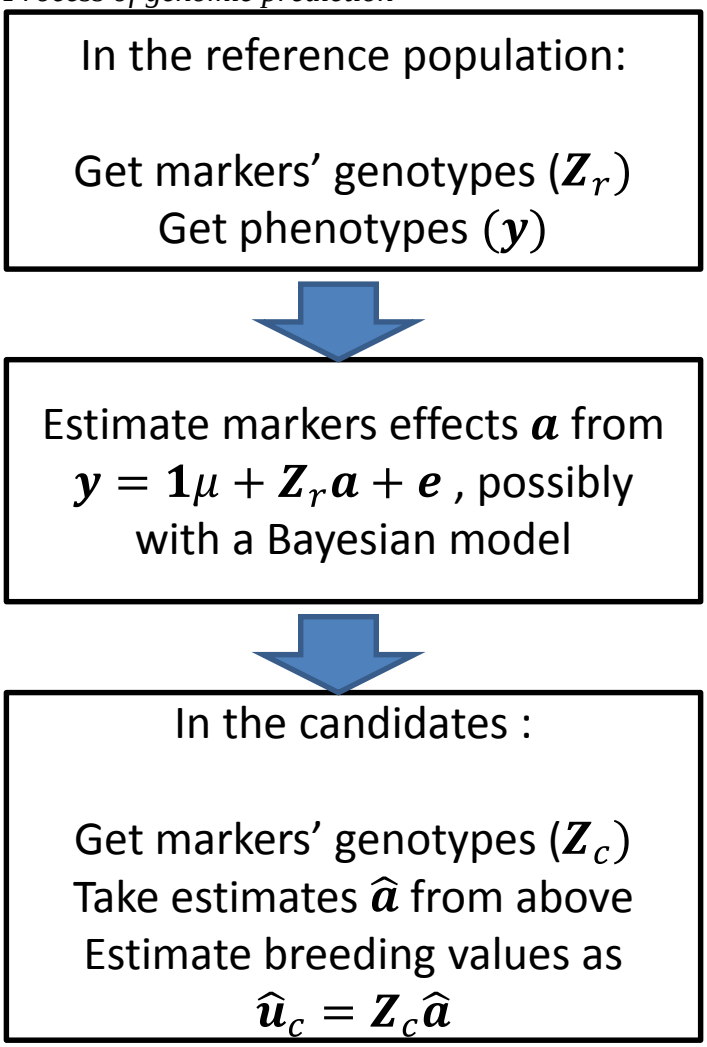


Figure 7. Process of genomic prediction

5. Bayesian regressions

Bayesian regression is another name for the Best Predictor or Conditional Expectation described above, and it describes the fact that we compute Conditional Expectations (another name for regressions (CASELLA and BERGER 1990)) using Bayesian methods. The term was first introduced in the genomic prediction literature by (DE LOS CAMPOS *et al.* 2009) and it is being used since. The Bayesian regression is, as described above, composed of a likelihood $p(\boldsymbol{y} \mid \boldsymbol{a}) = MVN(\boldsymbol{Xb} + \boldsymbol{Za}, \boldsymbol{R})$ and a prior distribution for markers, $p(\boldsymbol{a})$. A full and comprehensive account of Bayesians regressions for genomic prediction is in (DE LOS CAMPOS *et al.* 2013). However, before presenting the different models for Bayesian regressions, we will detail how allele coding should proceed in these methods.

5.1 Allele coding in Bayesian regressions.

Allele coding is the assignment of genotypes to numerical values in matrix $\boldsymbol{Z}$. Strandén and Christensen (STRANDÉN and CHRISTENSEN 2011) studied this in some detail. Markers commonly used for genomic prediction are biallelic markers. Imagine four individuals and two loci, where alleles for the loci are $\{A, a\}$ and $\{B, b\}$. The genotypes of the four individuals are:

$$\begin{matrix} aa & Bb \\ AA & bb \\ Aa & bb \\ aa & bb \end{matrix}$$

This can be coded with one effect by allele:

$$\boldsymbol{Za} = \begin{pmatrix} 0 & 2 & \vdots & 1 & 1 \\ 2 & 0 & \vdots & 0 & 2 \\ 1 & 1 & \vdots & 0 & 2 \\ 0 & 2 & \vdots & 0 & 2 \end{pmatrix} \begin{pmatrix} a_{1A} \\ a_{1a} \\ \cdots \\ a_{2B} \\ a_{2b} \end{pmatrix}$$

where a_{2B} is the allele "B" of the 2nd loci. So, for n markers we have $2n$ effects. Classic theory (e.g. (FALCONER and MACKAY 1996)) shows that this can be reduced to one effect by locus. We code in an additive way, as a regression of genetic value on gene content. The three classical ways of coding are:

Table 4. Additive coding for marker effects at locus *i* with reference allele *A*.

Genotype	101 Coding	012 Coding	Centered coding
aa	$-a_i$	0	$-2p_i a_i$
Aa	0	a_i	$(1-2p_i)a_i$
AA	a_i	$2a_i$	$(2-2p_i)a_i$

where p_i is the frequency of the reference allele ("A" in this case) at the i-th locus. In the example above, we have three possible $\boldsymbol{Z}$ matrices :

101 coding: $\boldsymbol{Za} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & -1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$

012 coding: $\boldsymbol{Za} = \begin{pmatrix} 0 & 1 \\ 2 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$

centered coding: $\boldsymbol{Za} = \begin{pmatrix} -0.75 & 0.75 \\ 1.25 & -0.25 \\ 0.25 & -0.25 \\ -0.75 & -0.25 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$

for the "centered" coding, allelic frequencies where 0.375 and 0.125; it can be verified that each column of centered $\boldsymbol{Z}$ sums to 0. This will be true if allelic frequencies are

computed from observed data. (VANRADEN 2008) defined matrix $\boldsymbol{M}$ as $\boldsymbol{Z}$ with 101 coding and then $\boldsymbol{Z} = \boldsymbol{M} - \boldsymbol{P}$, where $\boldsymbol{P}$ is a matrix with $2(p_i - 0.5)$.

Which allele to pick as a reference is arbitrary. If the other allele is chosen (as in the next Table), then the numbers in $\boldsymbol{Z}$ are reversed.

Table 5. Additive coding for marker effects at locus *i* with reference allele a.

Genotype	101 Coding	012 Coding	Centered coding
aa	a_i	$2a_i$	$(2-2p_i)a_i$
Aa	0	a_i	$(1-2p_i)a_i$
AA	$-a_i$	0	$-2p_i a_i$

As a result, estimates for marker effects a_i will change sign but the absolute value will be the same. Hence, $\boldsymbol{u} = \boldsymbol{Z}\boldsymbol{a}$ will be the same regardless of the coding.

5.2 Effect of prior information on marker estimates

Bayesian regressions are affected by the prior distribution that we assign to marker effects. One of the concerns is to be "fair" about this when making predictions. The problem is that the marker effect can be either too much shrunken (so that its estimate is too small, for instance if there is a major gene) or too little shrunken, in which case the estimate of the marker contains too much error and is completely wrong. Consider one marker. We have a likelihood information for this marker (its effect on the trait) and a prior information from "outside". What happens if this prior information is wrong?

The following two examples illustrate this. In both cases we estimate the marker effect as

$$lhs = \frac{\mathbf{1}'\mathbf{1}}{\sigma_e^2} + \frac{1}{\sigma_a^2}$$
$$\hat{a} = (lhs^{-1})\mathbf{1}'\boldsymbol{y}/\sigma_e^2$$

5.2.1 Marker effect is fixed

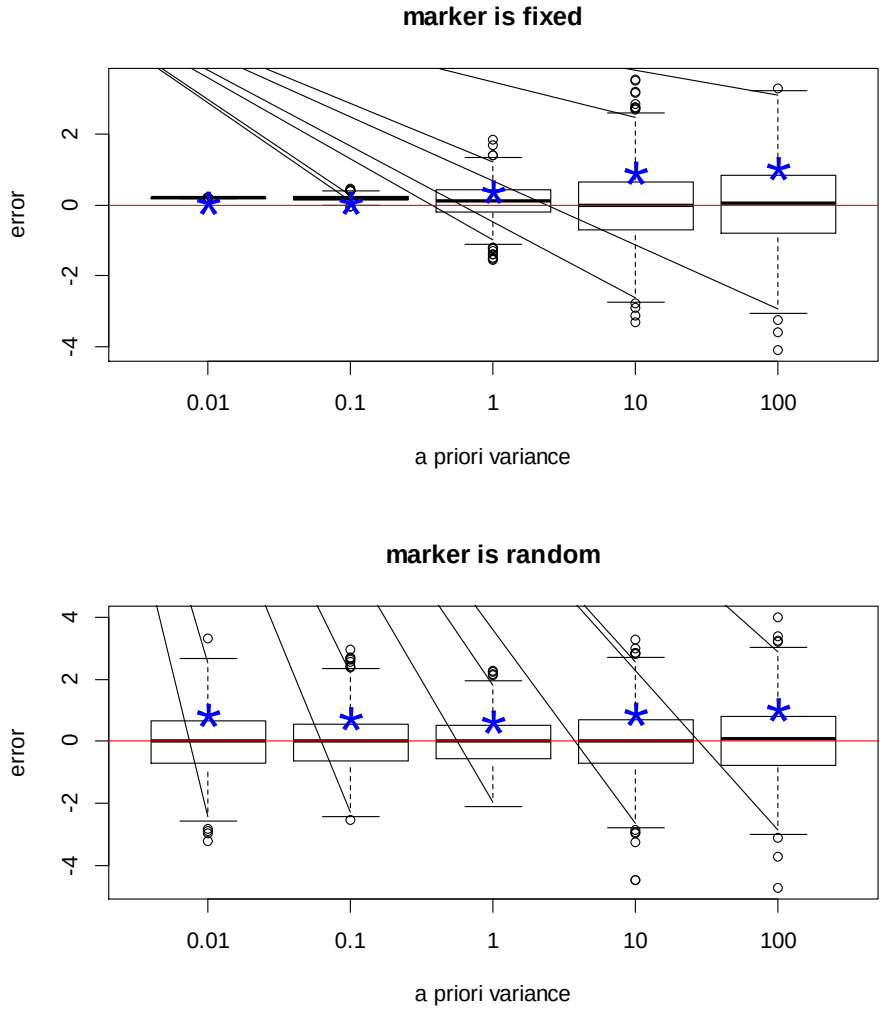
Assume that we have 10 records, and the marker has a "true" effect of 0.2, and this effect is constant across replicates. For instance DGAT1 is a known gene, and it is hard to think that its effect would change across different Holstein populations. We assume different prior variances for the marker, $\sigma_a^2 = \{0.01, 0.1, 1, 10, 100\}$, and $\sigma_e^2 = 1$. We have simulated 1000 data sets, and estimated the marker effect for each replicate; then plotted in the next Figure the error (as a boxplot) against the "no error" (in red), for each assumed marker variance.

It can be seen that when σ_a^2 is "large" the estimator is unbiased (on average there is no error) but each individual estimate has very large error (for instance there are errors of 4). When some shrinkage is used (i.e., for $\sigma_a^2 = 1$) the effect is slightly underestimated but large exaggerations never happen. Thus, across repetitions, the mean square error (blue stars) is minimized for small values of assumed σ_a^2.

5.2.2 Marker effect is random

In this case, the marker has different effects across populations because it is on feeble LD with some QTL. Then its true effect change all the time, so we can say that it comes from some distribution. If the true variance of the marker effect is $\sigma_a^2 = 1$, we obtain the results on the bottom of the Figure. All methods are unbiased (there is no systematic error) but putting the right variance give us the minimum error, as seen by the blue stars.

Figure 8. Distribution (boxplots) of errors in the estimate of one marker effect for different levels of shrinkage (X axis). No error is the red line. Blue stars indicate the square root of the mean square error.



5.3 Genetic variance explained by markers

A population of individuals has a certain genetic variance. If markers are genes: which part of the genetic variance is explained by each marker? This is just basic quantitative genetics. If a marker has an effect of a_i for each copy of the A allele, we have p^2 individuals with a value of $u = +2a_i$, q^2 individuals with a value of $u = 0$, and $2pq$ individuals with a value of $u = a_i$. Then the variance explained by this marker is $(u) = E(u^2) - E(u)^2$, which is developed in the following Table

Table 5. Variance explained by one marker

Genotype	Frequency	u^2	u
AA	p^2	$4a_i^2$	$2a_i$
Aa	$2pq$	a_i^2	a_i
Aa	q^2	0	0
Average		$4pa_i^2 + 2pqa_i^2$	$2pa_i$

So, finally the variance explained by one marker is $4pa_i^2 + 2pqa_i^2 - (2pa_i)^2 = 2pqa_i^2$. Markers with intermediate frequencies will explain most genetic variation.

5.3.1 Total genetic variance explained by markers

These are classic results also. Consider two markers, and consider that we know their effects a_i. The genetic value of an individual with genotype $\boldsymbol{z}$ will be $u = z_1a_1 + z_2a_2$. Variance in the population comes from sampling of genotypes (i.e., some individuals have one genotype while others have another genotype). Then $Var(u) = Var(z_1)a_1^2 + Var(z_2)a_2^2 + 2Cov(z_1, z_2)a_1a_2$. The term $Var(z_1) = 2p_1q_1$. The term $Cov(z_1, z_2)$ turns out to be $(z_1, z_2) = 2r\sqrt{p_1q_1p_2q_2}$, where r is the correlation measuring linkage disequilibrium. The term a_1a_2 implies that marker effects go in the same direction. Therefore, for the covariance between loci to enter into the genetic variance, the two markers need to be on linkage disequilibrium *and* at the same time their effects need to point in the same direction. There is no reason to be so, and on average this term will typically cancel out.

Either assuming linkage equilibrium or assuming that markers are uncorrelated one to each other, then, $Var(u) = Var(z_1)a_1^2 + Var(z_2)a_2^2 = 2p_1q_1a_1^2 + 2p_2q_2a_2^2$, and variances of each marker can simply be added. If we generalize this result to many markers, we have that

$$\sigma_u^2 = Var(u) = 2\sum_i^{nsnp} p_iq_ia_i^2$$

However, in most cases we do not know the marker effects. We may, though, have some prior information on them, like their *a priori* variance (the a priori mean is usually taken as zero). If this is the case, then we can substitute the term a_i^2 by its *a priori* expectation, that is, σ_{ai}^2 and therefore:

$$\sigma_u^2 = Var(u) = 2\sum_i^{nsnp} p_iq_i\sigma_{ai}^2$$

If we assume that all markers have the same variance *a priori* σ_{ao}^2 (say $\sigma_{a1}^2 = \sigma_{a2}^2 = \sigma_{a3}^2 = \cdots = \sigma_{a0}^2$, then $\sigma_u^2 = 2\sum_i^{nsnp} p_iq_i\sigma_0^2 = 2\sigma_0^2\sum_i^{nsnp} p_iq_i$ .We can factor out σ_{ao}^2 and we have the famous identity (GIANOLA *et al.* 2009) (FERNANDO *et al.* 2007) (VANRADEN 2008) (HABIER *et al.* 2007)

$$\sigma^2_{a0} = \frac{\sigma^2_u}{2\sum_i^{nsnp} p_i q_i}$$

This puts the a priori variance of the markers as a function of the genetic variance of the population. This result is used over and over in these notes and in most applications in genomic prediction.

5.3.2 Genetic variance explained by markers after fitting the data

This is actually fairly simple. After fitting the model to the data, there is an estimate $\hat{a}$ for each marker. We may say that each marker i explains a variance $2p_i q_i \hat{a}_i^2$. Therefore, and contrary to common assertions, the genetic variance contributed by each marker is NOT the same across all markers, and this is true for any method. Also, note that $2\sum p_i q_i \hat{a}_i^2$ underestimates the total genetic variance, because estimates $\hat{a}_i$ are shrunken towards 0. Better estimators will be presented later in, among others, GREML and BayesC.

5.4 Prior distributions for marker effects

From previous sections, it is clear that shrinking or, in other words, use of prior distributions for markers is a good idea. Therefore, we need a prior distribution for marker effects, which is notoriously difficult to conceive. Complexity comes, first, because markers are not genes *per se*, rather, they tag genes. But even the distribution of gene effects is unknown. There is a growing consensus in that most complex traits are highly polygenic, with hundreds to thousands of causal genes, most frequently of small effect. So, the prior distribution must include many small and few large effects. Also, for practical reasons, markers are assumed to be uncorrelated – even if they are close. For instance, if two markers are in strong linkage disequilibrium, they will likely show a similar effect *after* fitting the model, because thay will have similar incidence matrices in $\boldsymbol{Z}$. But before fitting the model, we cannot say that their effects will be similar or not. This is even exaggerated because there is arbitrariness in defining the sense of the coding; naming "A" or "a" the reference allele will change the sign of the marker effect.

Many priors for marker effects have been proposed in the last years. These priors come more from practical (ease of computation) than from biological reasons. Each prior originates a method or family of methods, and we will describe them next, as well as their implications.

1. Normal distribution: Random regression BLUP (RR-BLUP), SNP-BLUP, GBLUP
2. Normal distribution with unknown variances: BayesC, GREML, GGibbs
3. Student (t) distribution : BayesA
4. Mixture of Student (t) distribution and spike at 0: BayesB
5. Mixture of Normal distribution and spike at 0: BayesCPi
6. Double exponential: Bayesian Lasso
7. Mixture of a large and small normal distribution: Stochastic Search Variable Selection (SSVS)

5.5 Normal distribution for marker effects: Random regression BLUP (RR-BLUP), also called SNP-BLUP, BLUP-SNP, Ridge Regression, or GBLUP

In this notes, I will keep the name GBLUP for the model using genomic relationship matrices that will appear later, and the name BLUP-SNP for estimating marker effects.

The model for the phenotypes is typically something like:

$$\boldsymbol{u} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}$$

With $\boldsymbol{b}$ fixed effects (i.e., an overall mean), $\boldsymbol{a}$ marker effects, and $\boldsymbol{e}$ residual terms, with $Var(\boldsymbol{e}) = \boldsymbol{R}$ and usually $\boldsymbol{R} = \boldsymbol{I}\sigma_e^2$. Matrix $\boldsymbol{Z}$ contains genotypes coded as we have described previously

The prior for markers can be written as:

$$p(\boldsymbol{a}) = \prod_{i=1,nsnp} p(a_i)$$

Where

$$p(a_i) = N(0, \sigma_{a0}^2)$$

each marker effect follows a priori a normal distribution with a variance σ_{a0}^2 (that we will term hereinafter "variance of marker effects"). Note that the "0" implies that this variance is constant across markers.

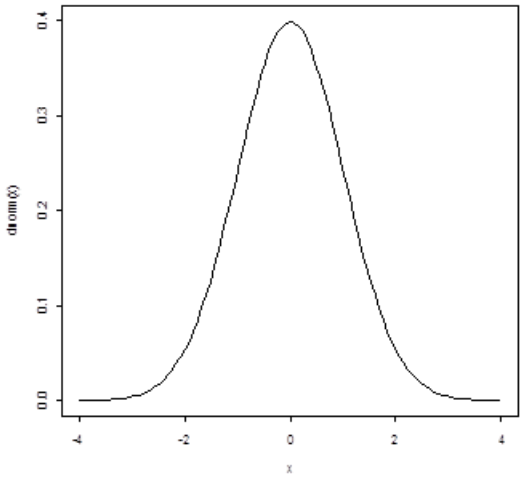


Figure 9. Standard normal distribution

It can be remarked that most effects are concentrated around 0, whereas few effects will be larger than, say, a value of 3. Therefore the prior assumption of normality precludes markers of having very large effects – unless there is a lot of information to compensate for this prior information.

We assume that markers are independent one from each other. This can be equivalently written as:

$$p(\boldsymbol{a}) = MVN(\boldsymbol{0}, \boldsymbol{D});\ Var(\boldsymbol{a}) = \boldsymbol{D} = \boldsymbol{I}\sigma_{a0}^2$$

where MVN stands for multivariate normal. This formulation including $\boldsymbol{D}$ will be used again throughout these notes.

5.5.1 Mixed Model equations for BLUP-SNP

The great advantage of the normal distribution is its algebraic easiness. Whereas in most cases marker effects are estimated using Gibbs Sampling, as we will see later on, there are closed formulae for estimators of marker effects. We can use Henderson's Mixed Model Equations:

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{D}^{-1} \end{pmatrix}\begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{a}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$$

Note that this is a linear estimator. If $Var(\boldsymbol{a}) = \boldsymbol{D} = \boldsymbol{I}\sigma^2_{a0}$ and $Var(\boldsymbol{e}) = \boldsymbol{I}\sigma^2_e$, then we can simplify them to

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{I}\lambda \end{pmatrix}\begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{a}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{y} \end{pmatrix}$$

with $= \sigma^2_e/\sigma^2_{a0}$. This expression is also known as *Ridge Regression*, although the Ridge Regression literature presents $\boldsymbol{I}\lambda$ (or $\boldsymbol{D}$) merely as a computational device to warrant correct estimates, and genetics literature presents it a s ratio of residual to genetic variances. Following traditional notations, we will talk about lhs (left hand side of the equations) and rhs (right hand side): $hs\begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{a}} \end{pmatrix} = rhs$.

For a multiple trait model, the equations are as above but $\boldsymbol{R}$ and $\boldsymbol{D}$ include multiple trait covariances, e.g. $\boldsymbol{R} = \boldsymbol{I} \otimes \boldsymbol{R}_0$ and $\boldsymbol{D} = \boldsymbol{I} \otimes \boldsymbol{S}_{a0}$.

These equations have unusual features compared to regular ones. First, the dimension is $(number\ of\ fixed\ effects + number\ of\ markers)^2$ but does not depend on the number of animals. Second, they are very little sparse. Matrix is completely dense and full.

For instance, assume $\boldsymbol{Z}\boldsymbol{a} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & -1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ (four individuals and two markers), an overall mean and $\lambda = 0.5$. Then

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{I}\lambda \end{pmatrix} = \begin{pmatrix} 4 & -1 & -3 \\ -1 & 3+0.5 & 0 \\ -3 & 0 & 3+0.5 \end{pmatrix}$$

5.5.2 Solving for marker effects

Mixed model equations as above can be explicitly setup and solved but this is expensive. For instance, setting up the equations would have a cost of n^2 markers times m individuals, and inverting them of n^3. Alternative strategies exist (LEGARRA and MISZTAL 2008) (VANRADEN 2008) (STRANDÉN and GARRICK 2009). They involve working with genotype matrix $\boldsymbol{Z}$ without setting up explicitely the mixed model equations. This can be done using iterative solving, where new solutions are based on old ones, and as iteration proceeds they are better and better until we can stop iterating. Two such procedures are the Gauss Seidel

and the Preconditioned Conjugated Gradients Algorithm or PCG. These were explained in detail by (LEGARRA and MISZTAL 2008).

Gauss Seidel proceeds to solve each unknown pretending than the other ones are known. So, if we deal with the i-th marker at iteration $l+1$, the mixed model equations for that marker reduce to a single equation:

$$(\boldsymbol{z}_i'\boldsymbol{z}_i + \lambda)\hat{a}_i^{l+1} = \boldsymbol{z}_i'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{b}} - \boldsymbol{Z}\widehat{\boldsymbol{a}} + \boldsymbol{z}_i\hat{a}_i^l)$$

This needs n operations for each marker, with a total of n^2 operations for each complete round of the Gibbs Seidel (e.g., 50000^2 for a 50K chip). However, it is easy to realize that the term within the parenthesis is the residual term "so far", $\hat{\boldsymbol{e}}^l$:

$$(\boldsymbol{z}_i'\boldsymbol{z}_i + \lambda)\hat{a}_i^{l+1} = \boldsymbol{z}_i'(\hat{\boldsymbol{e}}^l + \boldsymbol{z}_i\hat{a}_i^l) = \boldsymbol{z}_i'\hat{\boldsymbol{e}}^l + \boldsymbol{z}_i'\boldsymbol{z}_i\hat{a}_i^l$$

So the operation can be changed to a simpler one with a cost of n. The error term needs to be corrected after every new solution of the marker effect, using

$$\hat{\boldsymbol{e}}^{l+1} = \hat{\boldsymbol{e}}^l - \boldsymbol{z}_i(\hat{a}_i^{l+1} - \hat{a}_i^l)$$

With a cost of m (number of records) for each marker, and mn for a complete iteration. This strategy is called *Gauss Seidel with Residual Update* . A pseudo code in Fortran follows; a working code in R is at the Appendix:

```
Double precision:: xpx(neq),y(ndata),e(ndata),X(ndata,neq), &
sol(neq),lambda,lhs,rhs,val
do i=1,neq
      xpx(i)=dot_product(X(:,i),X(:,i)) !form diagonal of X'X
enddo
e=y
do until convergence
      do i=1,neq
            !form lhs X'R-1X + G-1
            lhs=xpx(i)/vare+1/vara
            ! form rhs with y corrected by other effects (formula 1) !X'R-
1y
            rhs=dot_product(X(:,i),e)/vare +xpx(i) *sol(i)/vare
            ! do Gauss Seidel
            val=rhs/lhs
            ! MCMC sample solution from its conditional (commented out
here)
            ! val=normal(rhs/lhs,1d0/lhs)
            ! update e with current estimate (formula 2)
            e=e - X(:,i)*(val-sol(i))
            !update sol
            sol(i)=val
      enddo
enddo
```

Figure 10. GSRU Fortran code

PCG is a strategy that uses a generic solver and proceeds by successive computations of the product $\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{I}\lambda \end{pmatrix}\begin{pmatrix} \widehat{\boldsymbol{b}}^l \\ \widehat{\boldsymbol{a}}^l \end{pmatrix}$. This can be easily done in two steps as

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{I}\lambda \end{pmatrix}\begin{pmatrix} \widehat{\boldsymbol{b}}^l \\ \widehat{\boldsymbol{a}}^l \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}' \end{pmatrix}\left((\boldsymbol{X} \quad \boldsymbol{Z})\begin{pmatrix} \widehat{\boldsymbol{b}}^l \\ \widehat{\boldsymbol{a}}^l \end{pmatrix}\right) + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{I}\lambda\,\widehat{\boldsymbol{a}}^l \end{pmatrix}$$

Again, only matrix $\boldsymbol{Z}$ is used but its cross-product $\boldsymbol{Z}'\boldsymbol{Z}$ is never computed.

Benefits of GSRU and PCG depend on the number of markers, but for large numbers they are extremely fast. For instance a Fortran code with PCG can solve for three thousand records and one million marker in minutes. PCG has a (much) faster convergence than GSRU: see the graph below. This makes it attractive for large application. However, GSRU can be converted with very few changes into a Gibbs Sampler application.

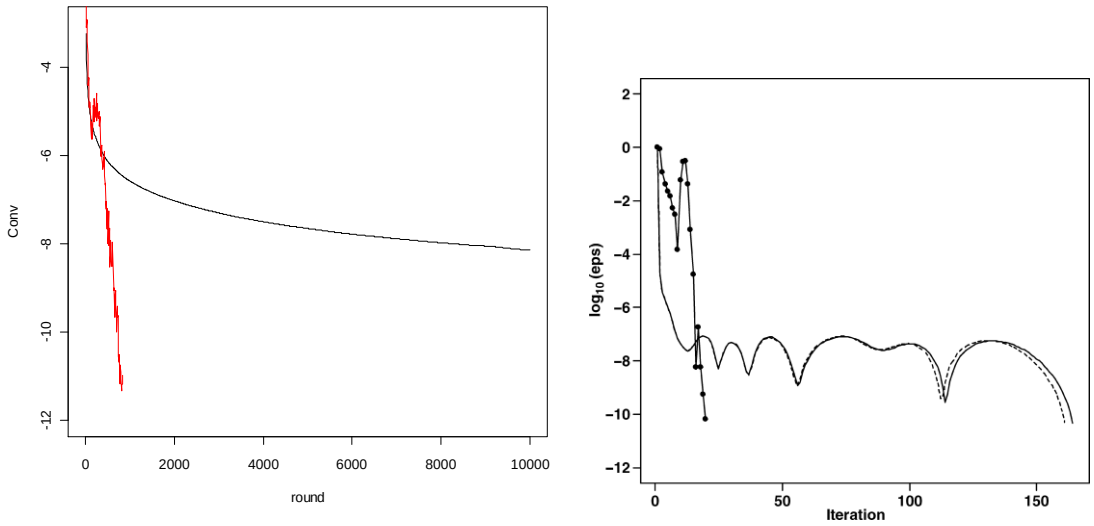


Figure 11. Convergence time for a large Holstein data set (left, GSRU in black, PCG in red) and on a mice data set (right, PCG line with points)

5.5.3 How to set variance components in BLUP-SNP

Henderson's equations assume that you know the values of two variance components, the variance of marker effects (σ_a^2), and the residual variance (σ_e^2). There are two possible strategies. The most common one is to use the relationship between the genetic variance and the *a priori* marker variance and to use

$$\sigma_{a0}^2 = \frac{\sigma_u^2}{2\sum_i^{nsnp} p_i q_i}$$

where σ_u^2 is an estimate of the genetic variance (e.g., obtained from previous pedigree-based studies) and p are marker frequencies ($q = 1 - p$). These allelic frequencies should be the ones in the population where the genetic variance was estimated (e.g., the base population of the pedigree) and *not* the current, observed populations. However, p are usually obtained from the data, so there is some error (although often negligible) and we will come back to this later. As for the residual variance, it can be taken as well from previous studies. For the multiple trait case, $\boldsymbol{S}_{a0} = \boldsymbol{G}_0/2\sum_i^{nsnp} p_i q_i$ where $\boldsymbol{G}_0$ is a matrix with estimates of the genetic covariances across traits.

5.5.4 Variances of pseudo-data, DYD's, and de-regressed proofs

Often, pseudo-phenotypes are used. These can consist in results of field trials, in progeny performances (VANRADEN and WIGGANS 1991), or in own corrected phenotypes. Other type of data are the deregressed proofs (GARRICK *et al.* 2009) (RICARD *et al.* 2013), that consist in

post-processing of pedigree-based genetic evaluations. These pseudo-data do not come from a regular phenotype and have varying variances. However, they do come with a measure of uncertainty (i.e., a bull can have 10 or 10,000 daughters). This can be accounted for in the residual covariance matrix, $\boldsymbol{R}$, which becomes heterogeneous.

In most software (for instance `GS3`, `blupf90` and the `R` function `lm`), this is done using weights. Weight w_i means (informally) the "importance" attached to the i-th record, and (formally) means that the record i behaves like an average of w_i observations, so that

$$\boldsymbol{R} = \begin{pmatrix} 1/w_1 & 0 & 0 \\ 0 & 1/w_2 & 0 \\ 0 & 0 & ... \end{pmatrix} \sigma_e^2$$

More weight means reduced residual variance. There are basically two ways to proceed.

Dairy cattle breeders work with "daughter yield deviations" (DYD). These are the average phenotypes of daughters for every bull, corrected for the EBV of their dam and environmental effects. Also, an "equivalent daughter contribution" (edc) is computed for the DYD, which reflects the number of daughters of that bull. The pseudo-phenotype for each bull is thus modelled as *twice* the DYD. If correction was perfect, a 2DYD for bull i with n_i daughters can be decomposed as:

$$2DYD_i = u_i + 2\frac{1}{n_i}\sum_j \phi_j + 2\frac{1}{n_i}\sum_j e_j = u_i + \frac{1}{n_i}\sum_j \epsilon_j$$

That is, the bull EBV (u_i), (twice) the average of its daughters' Mendelian sampling (ϕ_j), and the average of its daughters' residual deviations (e_j). The two latter terms are confounded into a pseudo-residual ϵ. Then, $Var(\epsilon) = 4Var(\phi) + 4Var(e) = 2\sigma_u^2 + 4\sigma_e^2$, because the variance of the Mendelian sampling is half the genetic variance. Finally,

$$Var(2DYD_i) = \sigma_u^2 + \frac{1}{n_i}\sigma_\epsilon^2$$

Thus, in dairy studies one may use $2DYD$ as a trait, with the typical genetic variance of σ_u^2 and a pseudo-residual variance of $\sigma_\epsilon^2 = 2\sigma_u^2 + 4\sigma_e^2$ with a weight $w_i = n_i$, where n_i is the "equivalent daughter contribution".

For another kind of data, (GARRICK *et al.* 2009) proposed a rather general approach for several kinds of pseudodata. They also provide expressions to put the adequate weights.

5.5.5 Some problems of pseudo-data

Note that the residual covariances of pseudo-data are assumed null. This is wrong. Cows in the same herd will share errors in estimation of the herd effect, and this generates a residual covariance; cows born from the same dam will share errors in estimation of the dam effect, and this also generates a residual covariance; and so on. These errors are ignored. However, Henderson (HENDERSON 1978) showed, in a similar context, that using precorrected data may lead to considerable bias and to loss of accuracy. This is, however, not a problem if pseudorecords used are from progeny testing, in which case the amount of information is so large that covariances among pseudo-data are very small.

5.6 Estimating variances from marker models: BayesC with Pi=0

Often, estimates of variance components from field data are unreliable, too old, or not directly available. In this case, it is simpler to estimate those variances from marker data. This was the case of (LEGARRA *et al.* 2008) in mice, and it has later been used to estimate genetic variances in wild populations (SILLANP\\"A 2011). It is very simple to do using Bayesian inference, and posterior estimates of the variances σ_a^2 and σ_e^2 are obtained. One of such programs is GS3 (LEGARRA *et al.* 2011a). This method has been described by (HABIER *et al.* 2011) as BayesC with Pi=0 and that is how we will cite it.

The algorithm is fairly simple from a GSRU iteration scheme. Instead of iterating the solution, we *sample it*, then we sample the marker variance:

```
do j=1,niter
      do i=1,neq
            !form lhs
            lhs=xpx(i)+1/vara
            ! form rhs with y corrected by other effects (formula 1)
            rhs=dot_product(X(:,i),e)/vare +xpx(i) *sol(i)/vare
            ! MCMC sample solution from its conditional
            val=normal(rhs/lhs,1d0/lhs)
            ! update e with current estimate (formula 2)
            e=e - X(:,i)*(val-sol(i))
            !update sol
            sol(i)=val
      enddo
  ! draw variance components
      ss=sum(sol**2)+ Sa
      vara=ss/chi(nua+nsnp)
      ss=sum(e**2)+ Se
      vare=ss/chi(nue+ndata)
enddo
```

Figure 12. Fortran code for BayesC with Pi=0

The algorithm requires initial values of variances and also prior information for them. Typical prior distributions for variance components are inverted-chi squared (χ^{-2}) scaled by constants (S_a^2 and S_e^2 for marker and residual variances) with some degrees of freedom (ν_a and ν_e). The degrees of freedom represent the amount of information put on those variances and therefore wheras 4 is a small value (and almost "irrelevant") 10,000 is a very strong prior. Typical values used in practice can be 4, for instance. On expectation, if we use *a priori* S_e^2 and ν_e then $E(\sigma_e^2|S_e,\nu_e) = S_e^2/\nu_e$. One may use previous estimates and put therefore

$$S_e^2 = \sigma_e^2 \nu_e$$

$$S_a = \sigma_{a0}^2 \nu_a;\ \sigma_{a0}^2 = \frac{\sigma_u^2}{2\sum_i^{nsnp} p_i q_i}$$

NOTE In other parameterizations $E(\sigma_e^2|S_e,\nu_e) = S_e^2$ and $E(\sigma_a^2|S_a,\nu_a) = S_a^2$ and therefore the Scale factor is in the same scale as the regular variances, and we can use $S_e^2 = \sigma_e^2$ and $S_a^2 = \sigma_{a0}^2$. This is the case for GS3 and the blupf90 family (MISZTAL *et al.*).

This is equivalent to what will be discussed in next chapter about GREML and G-Gibbs.

5.7 Transforming marker variance into genetic variance

We can use the previous result to get the genetic variance from the marker variance:

$$\sigma_u^2 = \sigma_{a0}^2 \, 2 \sum_i^{nsnp} p_i q_i$$

This is ONE estimate of genetic variance. It does not necessarily agree with other estimates for several reasons, mainly, different genetic base, different genetic model, and different data sets. However, published papers in the livestock genetics do NOT show missing heritability – estimates of genetic variance with pedigree or markers usually agree.

An example is interesting here. The mice data set of (LEGARRA *et al.* 2008) produced estimates of genetic variance based on pedigree and of marker variance based on markers, which are summarized in the following table. The column "σ_u^2 – markers" is obtained multiplying σ_{a0}^2 by $2\sum p_i q_i$=3782.05.

Table 6. Variance components in mice data

	σ_u^2 - pedigree	σ_{a0}^2	σ_u^2 - markers
Weight	4.59	$3.52\ 10^{-4}$	1.33
Growth slope*	8.37	$1.04\ 10^{-3}$	3.93
Body length	0.040	$9.09\ 10^{-6}$	0.034
Body Mass Index*	2.49	$0.80\ 10^{-3}$	3.02

*times 10^{-4}

Results are sometimes different, why? One reason is that pedigree estimates in this particular data set are little reliable, because there is a confusion between cage and family. Markers provide more accurate estimates. Another reason is that the genetic variances estimated with pedigrees or with markers refer to two slightly different populations. *Genetic variance estimated with markers* refers to an ideal population in Hardy-Weinberg equilibrium and with certain allele frequencies; these are the hypothesis underlying the expression $\sigma_u^2 = \sigma_{a0}^2 \, 2\sum p_i q_i$. *Genetic variance estimated with pedigree* refers to an ideal population in which founders of the pedigree are unrelated. The fact that we refer to two different ideal populations is referred to as different genetic bases (VANRADEN 2008) (HAYES *et al.* 2009).

It can be shown that if we have a pedigreed population and markers for this population, on expectation both variances are identical in Hardy-Weinberg and absence of inbreeding. We will come back to this notion later on the chapter on GBLUP and genomic relationships, and we will see how to deal with it.

5.8 Differential variances for markers

Simulations, and some trait and species, show the presence of large QTLs (or major genes, if you prefer) in the genome. We have seen before that shrinking markers results in smaller estimates than their "true" value. On the other hand, this avoids too much error in estimation. So how can one proceed? One way is to assign shrinkage differentially. Let's look at the equation for one marker effect:

$$\hat{a}_i = \frac{\frac{\boldsymbol{z}_i'\tilde{\boldsymbol{y}}}{\sigma_e^2}}{\frac{\boldsymbol{z}_i'\boldsymbol{z}_i}{\sigma_e^2} + \frac{1}{\sigma_{ai}^2}}$$

Where $\tilde{\boldsymbol{y}}$ means "**y** corrected by all other effects" and σ_{ai}^2 is the shrinkage of that marker. In BLUP-SNP, we assume $\sigma_{ai}^2 = \sigma_{a0}^2$ to be constant in all markers.

It would be nice to progressively update σ_{ai}^2 in order to get better estimates; intuitively, this means that the larger $\hat{a}_i$, the larger σ_{ai}^2. However this cannot be done easily because we know that giving too much (or too little) value to σ_{ai}^2 results in bad estimates as shown in Figure 7. In turn, this will give bad estimates of σ_{ai}^2 simply because we predict the variance of one marker with the estimate of a single marker.

5.8.1 REML formula for estimation of single marker variances

From old REML literature (e.g., see Ignacy Misztal notes), the EM formula for marker estimation should be:

$$\hat{\sigma}_{ai}^2 = \hat{a}_i^2 + C^{ii}$$

where C^{ii} is the element corresponding to the i-th marker on the inverse of the Mixed Model Equations $\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$. This expression has two parts, the first, $\hat{a}_i^2$, is the marker estimate to the square. However this estimate is way too shrunken (i.e. if the true effect of the marker is 7, the estimate may be 0.3), and the second part, C^{ii}, compensates for this lack of information. It is known as the *missing information*. This estimate can be obtained from a GBLUP context (SHEN *et al.* 2013).

The estimate is very inaccurate, because there is only one marker effect to estimate one variance component.

5.8.2 Bayesian estimation of marker variances

This can be, however, done within a Bayesian framework. The Bayesian framework will postulate a non-normal distribution for marker effects, and this non-normal distribution can be explained as a two-stages (or hierarchical) distribution. In the first stage, we postulate that each marker has *a priori* a different variance from each other:

$$p(a_i|\sigma_{ai}^2) = N(0, \sigma_{ai}^2)$$

In the second stage, we postulate a prior distribution for the variance themselves:

$$p(\sigma_{ai}^2|something) = p(\dots)$$

This prior distribution helps (the estimate of σ_{ai}^2 is more accurate, in the sense of lower mean square error) although it will still be far from reality (e.g.. (GIANOLA *et al.* 2009)). At any rate, this way of working is very convenient because the solving algorithm simplifies greatly. Most Bayesian Regressions are based in this idea.

5.9 BayesA

The simplest idea is to assume that *a priori* we have some information on the marker variance. For instance, this can be σ_{a0}^2 . Thus, we may attach some importance to this value and use it as prior information for σ_{ai}^2 . A natural way of doing this is using an inverted chi-squared distribution with S_a^2=$\sigma_{a0}^2\nu_{a0}$ scale and ν_{a0} degrees of freedom:

$$p(a_i|\sigma_{ai}^2) = N(0, \sigma_{ai}^2)$$
$$p(\sigma_{ai}^2|S_a, \nu_a) = S_a\chi_{\nu_a}^{-2}$$

The value of σ_{a0}^2 should actually be set as

$$\sigma_{a0}^2 = \frac{\nu - 2}{\nu}\frac{\sigma_u^2}{2\sum p_i q_i}$$

Because the variance of a t distribution is $\nu/(\nu - 2)$.

The whole setting is known as BayesA (MEUWISSEN *et al.* 2001). It can be shown that this corresponds to a prior on the marker effects corresponding to a scaled t distribution (GIANOLA *et al.* 2009):

$$p(a_i|\sigma_{a0}^2, \nu_a) = \sigma_{a0}t(0, \nu_a)$$

Which has the property of having "fat tails". This means that large marker effects are not unlikely *a priori*. For instance, having an effect of 4 is 200 times more likely under BayesA with $\nu_a = 4$ than BLUP-SNP. This can be seen in the Figure below.

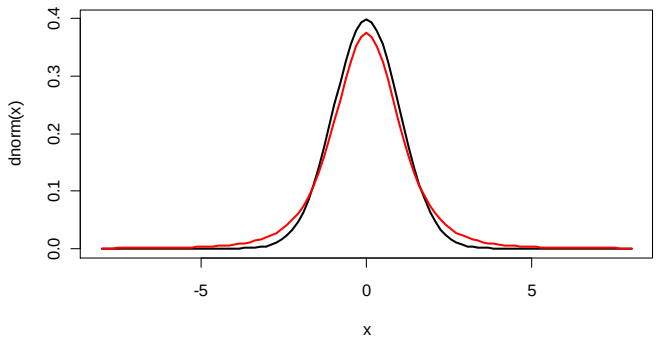


Figure 13. A priori distributions for BLUP-SNP (black) and BayesA (red).

Choosing ν_a is not obvious although small values around 4 are suggested in the literature. High values give the same results as normal distribution and thus BLUP-SNP. The code for BayesA is very simple:

```
do j=1,niter
   do i=1,neq
     !form lhs
     lhs=xpx(i)+1/vara(i)
     ! form rhs with y corrected by other effects
     rhs=dot_product(X(:,i),e)/vare +xpx(i) *sol(i)/vare
     ! MCMC sample solution from its conditional
     val=normal(rhs/lhs,1d0/lhs)
     ! update e with current estimate (formula 2)
     e=e - X(:,i)*(val-sol(i))
     !update sol
     sol(i)=val
     ! draw variance components for markers
     ss=sol(i)**2+nua*Sa
     vara(i)=ss/chi(nua+1)
   enddo
   ! draw variance components for residual
   ss=sum(e**2)+nue*Se
   vare=ss/chi(nue+ndata)
enddo
```

Figure 14. Pseudo code fortran for BayesA.

5.10 BayesB

A very common thought at the beginning of Genomic Evaluation was that there were not many QTLs. So a natural thinking is to consider that many markers do not have effect because they cannot trace QTLs. This originated the method known as BayesB, that simply states that the individual marker variance σ_{ai}^2 is potentially zero, and this can be find out. Note that this cannot happen for BayesA: the a priori chi-squared distribution prevents any marker variance from being zero.

This idea corresponds to a more complex prior as follows:

$$p(a_i|\sigma_{ai}^2) = N(0, \sigma_{ai}^2)$$
$$\begin{cases} p(\sigma_{ai}^2|S_a, \nu_a) = S_a\chi_{\nu_a}^{-2} \; with\; probability\; 1-\pi \\ p(\sigma_{ai}^2|S_a, \nu_a) = 0 \; with\; probability\; \pi \end{cases}$$

Then, when $\sigma_{ai}^2 = 0$ it follows that $a_i = 0$.

Intuitively, this prior corresponds to the following figure. The arrow means that there is a fraction π of markers with zero effect.

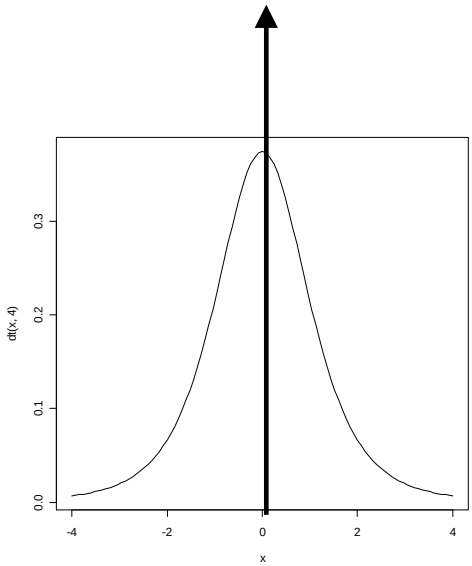


Figure 15. A priori distribution for BayesB

BayesB has a complex algorithm because it does involve the computation of a complex likelihood. Details on its computation can be found on Rohan Fernando's notes (http://www.ans.iastate.edu/stud/courses/short/2009/B-Day2-3.pdf ; slides 20 and 34; http://taurus.ansci.iastate.edu/wiki/projects/winterworkshop2013 , Notes, p. 42). and also in (VILLANUEVA *et al.* 2011).

5.11 BayesC(Pi)

Whereas the premises in BayesB seem interesting the algorithm is not. Further, experience shows that it is sensible to prior values of S_a^2, ν_a and π. As explained in (HABIER *et al.* 2011), this suggests the possibility of a simpler prior scheme where markers *having an effect* would be assigned a "common" variance, say σ_{a0}^2. This is simpler to be explained by introducing additional variables δ_i which explain if the i-th marker has an effect or not. In turn, these variables δ have a prior distribution called Bernouilli with a probability π of being 0. Therefore the hierarchy of priors is:

$$p(a_i|\delta_i) = \begin{cases} N(0, \sigma_{ai}^2)\ if\ \delta_i = 1 \\ 0\ otherwise \end{cases}$$
$$p(\sigma_{a0}^2|S_a, \nu_a) = S_a \chi_{\nu_a}^{-2}$$
$$p(\delta_i = 1) = 1 - \pi$$

Where S_a can be set to something like $S_a^2 = \sigma_{a0}^2 \nu_{a0}$ with

$$\sigma_{a0}^2 = \frac{\sigma_u^2}{(1-\pi) 2 \sum p_i q_i}$$

Experience shows that this prior hierarchy is more robust than BayesB, the reason being that, at the end (*after* fitting the data), the values of σ_{a0}^2 are little dependent on the prior. Thus the model may be correct even if the prior is wrong. Also, the complexity of the algorithm is greatly simplified, and can be summarized as follows:

```
do j=1,niter
    do i=1,neq
       ...
        ! compute loglikelihood for state 1 (i -> in model)
        ! and 0 (not in model)
        ! Notes by RLF (2010, Bayesian Methods in
        ! Genome Association Studies, p 47/67)
        v1=xpx(i)*vare+(xpx(i)**2)*vara
        v0=xpx(i)*vare
        rj=rhs*vare ! because rhs=X'R-1(y corrected)
        ! prob state delta=0
        like2=density_normal((/rj/),v0) !rj = N(0,v0)
        ! prob state delta=1
        like1=density_normal((/rj/),v1) !rj = N(0,v1)
        ! add prior for delta
        like2=like2*pi; like1=like1*(1-pi)
        !standardize
        like2=like2/(like2+like1); like1=like1/(like2+like1)
        delta(i)=sample(states=(/0,1/),prob=(/like2,like1/)
        if(delta(i)==1) then
          val=normal(rhs/lhs,1d0/lhs)
        else
          val=0
        endif
    ...
    enddo
   pi=1- & beta(count(delta==1)+aprioriincluded,
                count(delta==0)+apriori_not_included)
    ss=sum(sol**2)+nua*Sa
    vara=ss/chi(nua+count(delta==1))
    …
enddo
```

Figure 16. Pseudo code fortran for BayesCPi.

5.11.1 Markers associated to the trait

The value of $1-\pi$ (the number of markers having an effect) can be either fixed to a value or estimated from data. This is achieved in the last lines of the code above. How is this possible? Intuitively, we look at the number of markers who have ($\delta = 1$) or not ($\delta = 0$) an effect. Then we add a prior information on π. This comes in the form of a $Beta(a,b)$ distribution, which is a distribution of fractions between 0 and 1, saying that our fraction is *a priori* "like if" we had drawn a black balls and b red balls from an urn to make $\pi = a/(a+b)$.

The genetic variance explained by markers in BayesC(Pi) is equal to

$$\sigma_u^2 = \sigma_{a0}^2(1-\pi)2\sum p_i q_i$$

Thus, the same total genetic variance can be achieved with large values of σ_{a0}^2 and small values of $(1-\pi)$ or the opposite. This implies that there is a confusion between both, and it

is not easy to find out how many markers should be in the model. For instance, (COLOMBANI *et al.* 2012) reported meaningful estimates of π for Holstein but not for Montbeliarde.

Concerning markers, we have indicators of whether a given marker "is" or "is not" in the model, and these have been used as signals for QTL detection. However this is not often as expected. The output of BayesC(Pi) will be $\hat{\delta}_i$, the *posterior mean* of δ_i. This value will NOT be either 0 or 1 but something in between. So BayesCPi cannot be used to select "the set of SNPs controlling the trait" because such a thing does not exist: there are many possible sets. The following graph shows the kind of result that we obtain:

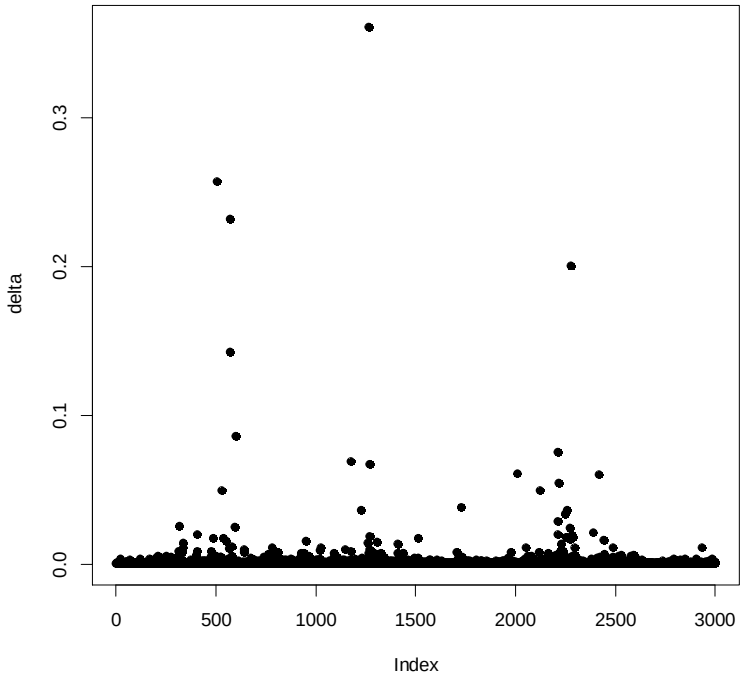


Figure 17. QTL signals from BayesCPi with Pi=0.999

How can we declare significance? There is no such thing as p-values. We may though use the Bayes Factor(WAKEFIELD 2009) (VARONA 2010):

$$BF = \frac{\frac{p(SNP\ in\ the\ model|data)}{p(SNP\ not\ in\ the\ model|data)}}{\frac{p(SNP\ in\ the\ model)}{p(SNP\ not\ in\ the\ model)}}$$

In our case this is:

$$BF_i = \frac{(1-\pi)}{\pi}\frac{p(\delta_i = 1 \mid \boldsymbol{y})}{1 - p(\delta_i = 1 \mid \boldsymbol{y})}$$

What thresholds should we use for BF? Some people suggest using permutations → too long. We can use a scale adapted by (KASS and RAFTERY 1995) used in QTL detection by (VARONA *et al.* 2001) and (VIDAL *et al.* 2005):

- BF= 3-20 "suggestive"
- BF= 20-150 "strong "

- BF>150 "very strong”

Something remarkable is that there is no need for multiple testing (Bonferroni) correction because all SNP were introduced at the same time, and the prior already « penalizes » their estimates (WAKEFIELD 2009).

5.12 Bayesian Lasso

The Bayesian Lasso (PARK and CASELLA 2008) (DE LOS CAMPOS *et al.* 2009) (LEGARRA *et al.* 2011b) suggests a different way to model the effect of markers. Instead of setting *a priori* some of them to 0, it sets them to very small values, as in the following Figure.

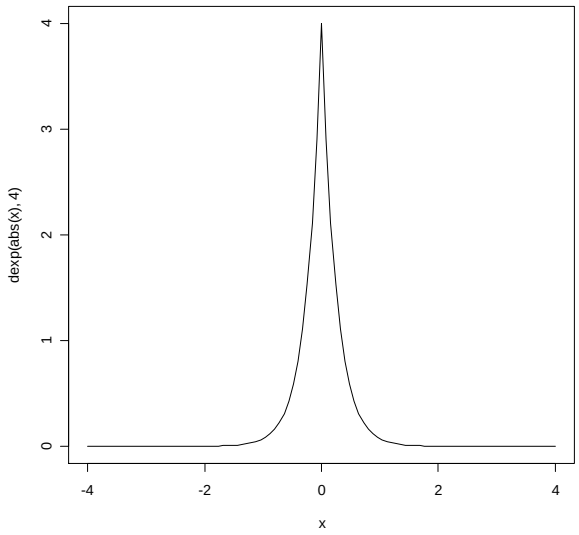


Figure 18. Prior distribution of marker effects for the Bayesian Lasso

This corresponds in fact to the following a priori distribution of markers:

$$p(a_i|\lambda) = \frac{\lambda}{2\sigma}\exp\left(-\frac{\lambda|a_i|}{\sigma}\right)$$

where the density function is on the absolute value of the marker and not on its square like in the normal distribution. Comig back to our notion of variance of markers, (PARK and CASELLA 2008) showed that the model is equivalent to a model with individual variances by marker, that is:

$$p(a_i|\sigma_{ai}^2) = N(0, \sigma_{ai}^2)$$
$$p(\sigma_{ai}^2|\lambda) = \frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\frac{\sigma_{ai}^2}{\sigma^2}\right)$$

(NOTE: the λ here has nothing to do with the λ in BLUP-SNP). The latter density function is a prior distribution on the marker variances that is known as *exponential*. This is very similar to BayesA, in that a prior distribution is postulated for marker variances. The difference is the nature of this prior distribution (exponential in Bayesian Lasso and inverted chi-squared in BayesA), that can be seen in the following Figure. It can be seen that, whereas in Bayesian Lasso very small variances are *a priori* likely, this is not the case in BayesA.

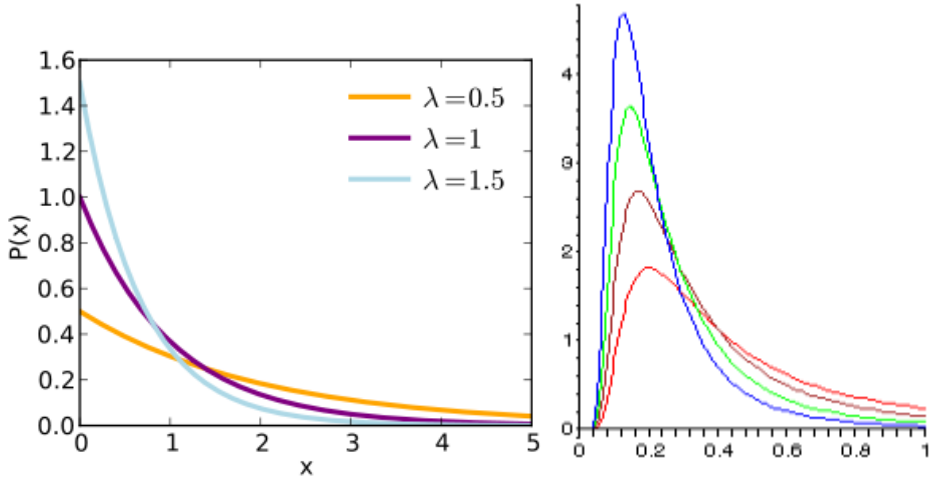


Figure 19. Shapes of the prior distribution of marker variances for the Bayesian Lasso (left) and Bayes A (right)

In practice, we have found that the Bayesian Lasso has a much better convergence than BayesCPi, while being as accurate for predictions (COLOMBANI *et al.* 2012).

5.12.1 Parameterization of the Bayesian Lasso

The term σ in the parameterization above has been subject to small debate. The original implementation of (PARK and CASELLA 2008) considered $\sigma^2 = \sigma_e^2$, the residual variance. (LEGARRA *et al.* 2011b) objected that it was unnatural to model the distribution of markers on the distribution of residuals and suggested setting $\sigma^2 = 1$. In this way, the interpretation of λ is quite straightforward as a reciprocal of the marker variance, because in such case $Var(a_i|\lambda) = 2/\lambda^2$. In this case, a natural way of fitting the prior value of λ is as

$$\frac{2}{\lambda^2} = \frac{\sigma_u^2}{2\sum p_i q_i}$$

This is the default in software GS3. The algorithm with this parameterization is rather simple:

```
do j=1,niter
    do i=1,neq
       !form lhs
       lhs=xpx(i)+1/vara(i)
       rhs=dot_product(X(:,i),e)/vare +xpx(i) *sol(i)/vare
       val=normal(rhs/lhs,1d0/lhs)
       e=e - X(:,i)*(val-sol(i))
       sol(i)=val
       ! draw variance components
       ss=sol(i)**2
       tau2(i)=1d0/rinvGauss(lambda2/ss,lambda2)
    enddo
  ! draw variance components
    ss=sum(e**2)+nue*Se
    vare=ss/chi(nue+ndata)
  ! update lambda
  ...
enddo
```

Figure 20. Fortran pseudocode for the Bayesian Lasso

The alternative implementation takes $\sigma^2 = \sigma_e^2$, and can be found in R package BLR (PÉREZ *et al.* 2010). In this case, a natural way of fitting the prior value of λ is as (PÉREZ *et al.* 2010)

$$\frac{2}{\lambda^2} = \frac{\sigma_u^2}{\sigma_e^2 2\sum p_i q_i}$$

In this case, λ can be thought of as a ratio between marker variance and residual variance (signal-to-noise). Both parameterizations are not strictly equivalent depending on the priors used for λ and the different variances, but they should give very similar results (in spite of (LEGARRA *et al.* 2011b)).

5.13 Stochastic Search Variable Selection

Yet another method, it does postulate two kinds of markers: those with a large effect, and those with a small (but not zero) effect. These are, similarly to BayesC(Pi), reflected in two variances, one for the large effects (σ_{al}^2) and one for the small effects (σ_{as}^2). The idea was from (GEORGE and MCCULLOCH 1993), and details can be found in e.g. (VERBYLA *et al.* 2009). The advantage of this method is that it is rather fast and does not require likelihood computations, although choosing *a priori* the proportions of "large" and "small" effects might be tricky.

5.14 Overall recommendations for Bayesian methods

BayesB seems to be little robust. The other methods are reasonably robust. My personal suggestion is to start from BLUP-SNP, which is very robust, then progress to other methods. Meaningful prior information (for instance how to set up λ from genetic variance) is relevant, if not for anything else, to have correct starting values. Bayesian methods often give similar

precisions than BLUP-SNP, but important exceptions such as fat and protein content in dairy cattle do exist.

5.15 VanRaden's NonLinear methods

Gibbs samplers are notoriously slow and this hampers the implementation of Bayesian methods for genomic predictions. VanRaden (VANRADEN 2008) presented NonLinearA and NonLinearB, iterative methods that do not need samplers and converge in a few iterations. NonLinearA assumes a certain departure from normality, called "curvature" (say c) that oscillates between 1 (regular BLUP-SNP) and 1.25 (COLE *et al.* 2009), such that the distribution would resemble more closely a fat-tailed distribution like Bayesian Lasso or BayesA. In our notation, this means that the marker variance is updated as

$$\sigma_{ai}^2 = \sigma_{a0}^2 \left(c^{\left(\frac{|\hat{a}_i|}{sd(\hat{a}_1,\dots,\hat{a}_n)} - 2 \right)} \right)$$

The role of the curvature is similar (but goes in the opposite direction) to the degrees of freedom in BayesA. The more the curvature, the more large marker effects are allowed. For instance, if $c = 1.25$ and a marker estimate is an outlier in the distribution of marker estimates, and has for instance a standardized value of 2.5, its variance σ_{ai}^2 will be increased by $1.25^{0.5} = 1.12$. To avoid numerical problems, for small data sets, it is recommended to use $c = 1.12$ and to impose a limit of 5 for $\frac{|\hat{a}_i|}{sd(\hat{a}_1,\dots,\hat{a}_n)}$ (VanRaden, personal communication). This algorithm is fast, stable and regularly used for dairy cattle genomic evaluation.

The whole setting is very similar to BayesA or to the Bayesian Lasso, with c playing the role of λ. The prior density for marker effects departs from normality for marker beyond two standard deviations, as shown in the next Figure. It can be seen that large marker effects are much more likely in nonlinearA than in a normal density.

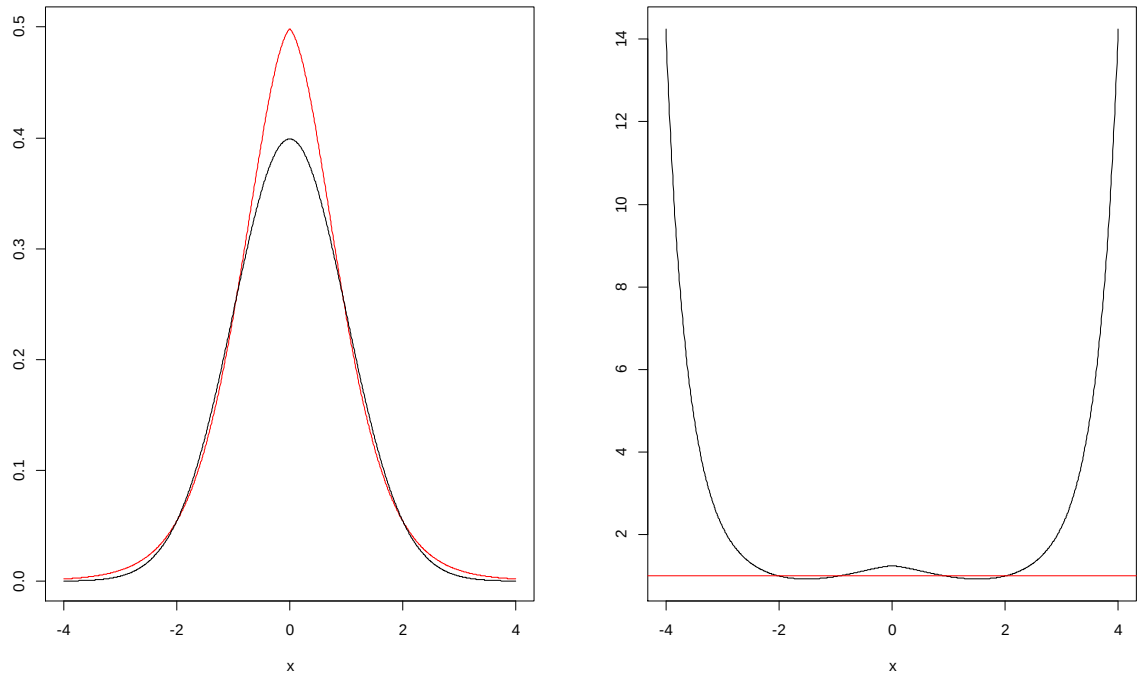


Figure 21. (Left) Shapes of the prior distribution of marker effects for VanRaden nonlinearA (red) and normal BLUP-SNP (black). (Right) Ratio of nonlinearA/normal densities.

The NonLinearB is akin to BayesC(Pi) (some markers are 0 and other share a common variance), whereas NonLinearAB is similar to BayesA (some markers are zero and others have a variance that might change from marker to marker). NonLinearB uses a mixture distribution, in which σ^2_{ai} is obtained from a average of variances weighted by the likelihood that the marker has zero effect or not. However the algorithm will not be further detailed here.

5.16 The effect of allele coding on Bayesian Regressions

We have explained how allele coding should (or can) proceed. (STRANDÉN and CHRISTENSEN 2011) analyzed the result of allele coding in genomic predictions. One need to distinguish carefully two things here. What we mean by allele coding is coding of matrix $\boldsymbol{Z}$ for genotypes, *not* the frequencies used in $\sigma^2_{a0} = \frac{\sigma^2_u}{2\sum_i^{nsnp} p_i q_i}$.

One of their results is that, for any model including a "fixed" effect such as an overall mean μ or a cross-classified effect (e.g., sex) estimates of marker effects $\hat{\boldsymbol{a}}$ and estimated genetic values $\hat{\boldsymbol{u}} = \boldsymbol{Z}\hat{\boldsymbol{a}}$ are invariant to parametrization, up to a constant. This constant will go into the overall mean or fixed effect. Consider for instance the mean. The mean of the genetic values of the population will be $'\hat{\boldsymbol{u}}$ **,** and this mean is not invariant to parameterization, and cannot either be separated from the overall mean of the model, μ. If the centered coding is used, then $\mathbf{1}'\hat{\boldsymbol{u}} = \mathbf{1}'\boldsymbol{Z}\hat{\boldsymbol{a}} = 0$. As for the marker variance σ^2_{a0} estimated by, say, BayesC, they also proofed that it is invariant to parameterization of $\boldsymbol{Z}$.

These results are convenient because they assure us that any allele coding is convenient. However, this result does not apply to the all features. For instance, the standard deviation (and therefore, in animal breeding words, the "model-based" reliability) of estimated genetic values $\widehat{\boldsymbol{u}}$ is *not* invariant to parameterization, because there will be a part of the overall mean absorbed, or not, by $\boldsymbol{Z}\widehat{\boldsymbol{a}}$. This implies that reports of the posterior variance of $\widehat{\boldsymbol{u}}$ will depend on the allele coding. The same result applies to GBLUP, as we will see later.

6. Quantitative genetics of markers, or markers as quantitative traits

6.1 Gene content as a quantitative trait

This small chapter wants to put forward an idea that goes often unnoticed and that was highlighted by (GENGLER *et al.* 2007) (GENGLER *et al.* 2008). A detailed but terse account is in (COCKERHAM 1969). Consider a marker, not necessarily biallelic. An individual is carrier of a certain number of copies, either 0, 1 or 2. This number of copies is usually called *gene content*, sometimes called *individual gene frequencies*, a bit of a confusing term..

For instant consider the blood groups AB0 (multiallelic) or Rh (biallelic +/-) the following table:

Table 7. Example of gene content for blood groups

Individual	Genotype	Gene count for A	Gene count for B	Gene count for 0
John	AB	1	1	0
Peter	A0	1	0	1
Paul	00	0	0	2

	Genotype at Rh	Gene count for +	Gene count for -
John	++	2	0
Peter	+ -	1	1
Paul	- -	0	2

For a biallelic marker, the table is simpler, because the gene content with one reference allele will be 2 minus the gene content of the other allele.

The gene content of one individual is equal to the value in one of the parameterizations (the 012) for the (Bayesian) regression of traits on markers that we have seen before, e.g. in single marker association as in

$$\boldsymbol{Za} = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix} a_A$$

And $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}$ like in

$$\begin{pmatrix} 12 \\ 35 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix} a_A + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

For individuals with genotypes $\{aa, AA, aA\}$. For this reason, in the next, we will denote the gene content of individual i as z_i which will take values $\{0,1,2\}$.

The gene content can thus be "counted", and can be studied as a quantitative measure, and therefore as a quantitative trait (although it is not a continuous trait). Therefore it can be treated by standard quantitative genetics methods. In the following we will deal with gene

content of biallelic markers such as SNPs but many of the results apply to multiallelic markers such as haplotypes or microsatellites.

6.2 Mean, variance and heritability of gene content

If the alleles are $\{A, a\}$ in a population, and A is the reference allele, *the average gene content* $E(z)$ is equal to the number of occurrences of A, which is twice the allelic frequence: $E(z) = 2p$. In Hardy-Weinberg equilibrium, the *variance of gene content* is calculated as:

$$Var(z) = E(z^2) - E(z)^2$$

Table 8. Variance of gene content

Genotype	Frequency	z^2	z
AA	p^2	4	2
Aa	$2pq$	1	1
Aa	q^2	0	0
Average		$4p^2 + 2pq$	$2p$

Therefore $\sigma_z^2 = Var(z) = 4p^2 + 2pq - (2p)^2 = 2pq$

The *heritability of gene content* is the ratio of genetic to environmental variance. Clearly, all variance is genetic because the gene content is fully determined by transmission from fathers to offspring, and all the genetic variance is additive because gene content is additive by construction. Therefore the heritability is 1.

6.3 Gengler's method to estimate missing genotypes and allelic frequencies at the base population

A common case is a long pedigree where some, typically young, animals have been genotyped for a QTL of interest or for markers. If the QTL has an effect, it is important to be able to include its genotype for all individuals (KENNEDY *et al.* 1992). Using expressions above, (GENGLER *et al.* 2007) (GENGLER *et al.* 2008) suggested a way to estimate gene content for all individuals in a pedigree, as well as allele frequencies. The method simply consists in modelling the genotype $\boldsymbol{z}$ as a quantitative trait:

$$\boldsymbol{z} = \mathbf{1}\mu + \boldsymbol{W}\boldsymbol{u} + \boldsymbol{e}$$

Where $\boldsymbol{W}$ is a matrix of incidence with 1's for genotyped individuals and 0 otherwise. A heritability of 0.99 is used to estimate it through mixed model equations; on exit, $\hat{\boldsymbol{u}}$ contains estimates of gene content for all individuals (these are equal to the observed genotype for the genotyped individuals) and $\hat{\mu}$ actually contains $2\hat{p}$. The method has some defaults, mainly, the estimate of gene content is a regressed estimate and therefore individuals tend to be more alike at the QTL than what they actually are. For instance, isolated individuals will have an estimate consisting in $2\hat{p}$. However, Gengler method is very important for two reasons: the first is that it provides an analytical tool to deal with gene content at missing genotypes (and it was completed by (CHRISTENSEN and LUND 2010)) and second, it serves to estimate allelic frequencies at the base population when it is not genotyped (VANRADEN 2008), although the estimate can sometimes go out of parametric space.

6.4 Covariance of gene content across two individuals.

This is $Cov(z_i, z_j)$. Individuals i and j have two copies at the marker. If we draw one copy from i and another from j, the probability of them being identical (by descent) is $\Theta_{ij} = A_{ij}/2$, where Θ is known as Malecot "coefficient de parenté", kinship, or coancestry and A_{ij} is the additive relationship. Therefore

$$Cov(z_i, z_j) = E(z_i z_j) - E(z_i)E(z_j)$$

$E(z_i) = E(z_j) = 2p$. $E(z_i, z_j)$ can be obtained by as follows. There are four ways to sample two alleles. For each way, the product $z_i z_j$ will be 1 *only* if the first individual got the allele A (with probability p) and the second one got A as well, either because it was identical by descent (with probability $A_{ij}/2$) or because it was not identical by descent (with probability $1 - A_{ij}/2$) but by chance had the "A" allele (with probability p), that is $pA_{ij}/2 + p(1 - A_{ij}/2)p = pqA_{ij}/2 + p^2$, and multiplying by four possible ways gives $E(z_i z_j) = A_{ij}2pq + 4p^2$. Putting all together gives

$$Cov(z_i, z_j) = A_{ij}2pq$$

, which means that the covariance between relatives at gene content is a function of their relationship A_{ij} and the genetic variance of gene content $2pq$. In other words, two related individuals will show similar genotypes at the markers. This result was utilized by (GENGLER *et al.* 2007; GENGLER *et al.* 2008) and (HABIER *et al.* 2007).

7. Genomic relationships

7.1 Reminder about relationships

Wright (WRIGHT 1922) introduced the notion of relationships as *correlation* between genetic effects of two individuals. For practical reasons, it is more convenient to use what is often called "numerator relationship" (QUAAS 1976) or simply "relationship" or "additive relationship". This equals the standardized *covariance* (*not* the correlation) between the additive genetic values of two individuals, which does not equal the correlation if there is inbreeding. There are several terms used to talk about relationships, and here we will present the classical definitions according to pedigree:

- Coancestry: θ_{ij}, also called Malecot "*coefficient de parenté*" or *kinship*. This is the probability that two alleles, one picked at random from each one of two individuals i and j, are identical (by descent). If the individual is the same, alleles are sampled with replacement
- Inbreeding F_k: probability that the two alleles in individual k are identical by descent. If k is the offspring of i and j, then $F_k = \theta_{ij}$. Also, $\theta_{kk} = (1 + F_k)/2$.
- Additive relationship, or relationship in short, is equal to twice the coancestry: $A_{ij} = 2\theta_{ij}$. Also, $A_{kk} = 1 + F_k$.
- The genetic covariance between two individuals is $Cov(u_i, u_j) = A_{ij}\sigma_u^2$.

All these measures of relatedness are defined with respect to a base population constituted by *founders*, which are assumed unrelated and carriers of different alleles at causal QTLs. This generates, as a byproduct, that relationships estimated using pedigrees are strictly positive. However, this is not the case when we consider marker or QTL information.

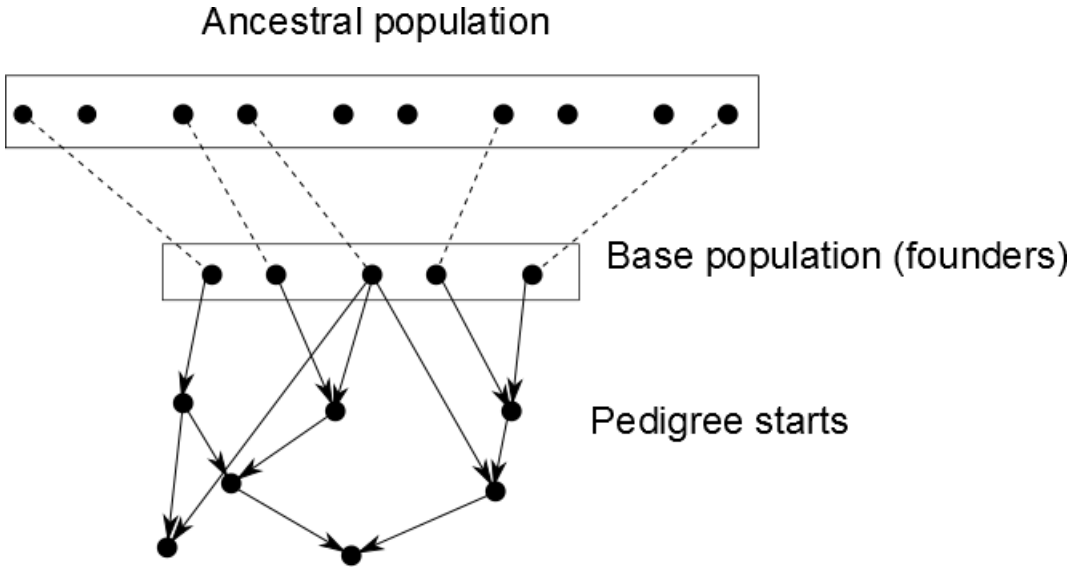


Figure 22. Representation of a pedigree. Continuous lines represent known pedigree links. Dotted lines represent unknown lineages.

7.2 Identity by state and identity by descent of two individuals

Identity by state (IBS) "molecular" coancestries (that we will denote f_{Mij}) refers to the numbers of alleles shared by two individuals, and it is equal to the probability that two alleles picked at random, one by individual, are identical. For the purposes of these notes we will refer to *molecular relationships*, which are $r_{Mij} = 2f_{Mij}$ (to be on the same scale as A_{ij}). These are sometimes called "similarity index" but also as "total allelic relationship" (NEJATI-JAVAREMI *et al.* 1997). For the two-allele case, this is summarized in the following table:

Table 9. Molecular relationships for combinations of different genotypes

	AA	Aa	aa
AA	2	1	0
Aa	1	1	1
aa	0	1	2

In fact, the molecular relationship can be obtained in a mathematical form without counting because (TORO *et al.* 2011):

$$r_{Mij} = z_i z_j - z_i - z_j + 2$$

This expression, connected with genomic relationships, will show its utility later on.

The identity by state reflected in the molecular relationship r_{Mij} and the identity by descent (IBD) reflected in the pedigree relationships A_{ij} have a well-known relationship that is periodically revisited (LI and HORVITZ 1953) (RITLAND 1996) (EDING and MEUWISSEN 2001) (POWELL *et al.* 2010) (TORO *et al.* 2011). A formal derivation can be found in (COCKERHAM 1969) (see also (TORO *et al.* 2011)). A simple one is as follows. Consider one allele sampled from i and another allele sampled from j. They can be identical because they were identical by descent (with probability $A_{ij}/2$), or because they were *not* identical by descent (with probability $1 - A_{ij}/2$) but they were identical just by chance (with probability $p^2 + q^2$). Therefore

$f_{Mij} = \theta_{ij} + (1 - \theta_{ij})(p^2 + q^2)$ where $\theta_{ij} = A_{ij}/2$ is the pedigree coancestry, and

$$r_{Mij} = A_{ij} + (2 - A_{ij})(p^2 + q^2)$$

also,

$$A_{ij} = \frac{r_{Mij} - 2p^2 - 2q^2}{2pq}$$

Thus, IBS is biased upwards with respect to IBD. Reordering we have that:

$$(1 - f_{Mij}) = (1 - \theta_{ij})(1 - p^2 - q^2)$$

Which is in the form of Wright's fixation indexes. This means that molecular heterozygosity (or in other words, "not alikeness" of two individuals equals "not alikeness" by descendance times "not alikeness" of markers.

There is another important point. The expression above to get IBD relationships from IBS relationships is identical to VanRaden's $\boldsymbol{G}$ that will be detailed later, up to a constant. Therefore, the results will be identical using IBD or IBS relationships.

7.2.1 Covariance between individuals

What does it mean "covariance between individuals"? In reality, individuals *i* and *j* (for instance bulls ALTACEASAR and BODRUM) have a defined true genetic value, *that we don't know*. So you cannot calculate a covariance between their true breeding values, because there is only one repetition of the pair. However, the mental construction is as follows. If I repeated events (or I simulate) in my cattle pedigree (transmission of QTL from parents to offspring) many times, individuals ALTACEASAR and BODRUM would have inherited different QTLs and therefore show different genetic values at different repetitions. The covariance of these two hypothetical vectors of genetic values is what we call the covariance between individuals.

7.3 Relationships across individuals for a single QTL

Assume that you are studying one species with a single biallelic quantitative gene. You genotype the individuals and you are asked, what is the covariance between individuals i and j, for which the genotype is known? Let express the breeding values as functions of the genetic value (za) deviated from the population mean, $\mu = 2pa$:

$$u_i = z_i a - 2pa = (z_i - 2p)a$$
$$u_j = z_j a - 2pa = \big(z_j - 2p\big)a$$

Where z_i is expressed as $\{0,1,2\}$ copies of the allele of reference of the QTL having the effect a_i (let's say allele A). If the effect of the QTL has some prior distribution with variance $Var(a) = \sigma_a^2$, and the genetic variance in Hardy-Weinberg equilibrium is $2pq\sigma_a^2$. It follows from regular rules of variances and covariances that

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

Which is equal to $z_i z_j \sigma_a^2$ if we use the "centered" coding instead of "012".

Dividing the covariance by the genetic variance $2pq\sigma_a^2$ we obtain additive relationships produced by the QTL I will call this r_{Qij}. Two exampled for $p = 0.5$ and $p = 0.25$ are shown in the next tables:

Table 10. Relationships r_{Qij} between individuals for a single QTL with $p = 0.5$

	AA	Aa	aa
AA	1	0	-1
Aa	0	0	0
Aa	-1	0	1

Table 11. Relationships r_{Qij} between individuals for a single QTL with $p = 0.25$

	AA	Aa	aa
AA	2.25	0.75	-0.75
Aa	0.75	0.25	-.025
Aa	-0.75	-0.25	0.25

7.3.1 Negative relationships

Now, this is puzzling because we have negative covariances. The reason for this is that we have *imposed* the breeding values to refer to the average of the population. However, there is no error. We need to interpret the values as standardized correlations (VANRADEN 2008) (POWELL *et al.* 2010). This was also frequently done by Wright, who would accept

"negative" inbreedings. The intuitive explanation is that if the average breeding value is to be zero, some animals will be above zero and some below zero. Animals carrying different genotypes will show negative covariances.

These relationships can NOT be interpreted as probabilities. Correcting negative relationships (or genomic relationships) to be 0 is a serious conceptual error and this gives lots of problems, yet it is often done.

7.3.2 Centered relationships and IBS relationships

It can be noted that the Table above with $p = 0.5$ is equal to the Table of molecular (or IBS) relationships before, minus a value of 1. Relationships at the QTL can be obtained as IBS at the QTL (NEJATI-JAVAREMI *et al.* 1997) (VARONA 2010), and they can be interpreted as twice a probability, as regular relationships. However, the value of 1 will be factored out in the mean (STRANDÉN and CHRISTENSEN 2011) and models using either parameterization (and also any assumed p) will give identical estimates of breeding values in the GBLUP context that we will see later on. Therefore *using IBS relationships or genomic relationships gives identical estimates of breeding values* –if associated variance components are comparable.

7.3.3 Inbreeding at a simple QTL

Inbreeding would be the value of the self-relationship r_{Qii} , minus 1. This is puzzling because we have negative values for heterozygotes. What this means is that there is less homozygosity than expected (FALCONER and MACKAY 1996).

7.4 Genomic relationships: Relationships across individuals for many markers

7.4.1 VanRaden's first genomic relationship matrix

We proceed to derive relationships for many markers as we did for one QTL. The derivation is fairly easy and purely statistical. To refer breeding values to an average value of 0, we adopt the "centered" coding for genotypes described before and shown above:

Table 12. Additive coding for marker effects at locus *i* with reference allele *A*.

Genotype	101 Coding	012 Coding	Centered coding
aa	$-a_i$	0	$-2p_i a_i$
Aa	0	a_i	$(1-2p_i)a_i$
AA	a_i	$2a_i$	$(2-2p_i)a_i$

In theory, to refer the breeding values to the pedigree base population, we should use allelic frequencies of the base population but these are rarely available (although Gengler's method can be used). Often current observed frequencies are used. At any rate, we have that

$$\boldsymbol{u = Za}$$

That is, individuals are a sum over genotypes of markers' effects. We have shown that marker effects can be considered to have an a priori distribution, and this a priori distribution has a variance

$$Var(\boldsymbol{a}) = \boldsymbol{D}$$

With

$$\boldsymbol{D} = \begin{pmatrix} \sigma^2_{a1} & 0 & ... & 0 \\ 0 & \sigma^2_{a2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma^2_{an} \end{pmatrix}$$

If we fit different variances by marker, but that is usually assumed as $\boldsymbol{D} = \boldsymbol{I}\sigma^2_{a0}$. Then, the covariance matrix of breeding values is

$$Var(\boldsymbol{u}) = \boldsymbol{Z}Var(\boldsymbol{a})\boldsymbol{Z}' = \boldsymbol{ZDZ}' = \boldsymbol{ZZ}'\sigma^2_{a0}$$

These are however NOT relationships. Relationships are standardized covariances. The variance we need to divide by is the genetic variance or, in other words, the variance of the breeding values of a set of animals. If we assume our population to be in Hardy-Weinberg and Linkage equilibrium, then we have shown that

$$\sigma^2_u = 2\sum_{i=1}^{nsnp} p_i q_i \sigma^2_{a0}$$

Therefore, we can now divide $Var(\boldsymbol{u})$ above by this variance and this gives the genomic relationship matrix (VANRADEN 2008):

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ}'}{2\sum p_i q_i}$$

7.4.2 VanRaden's second (or Yang's) genomic relationship matrix

A second matrix suggested by (VANRADEN 2008) but made popular (and often incorrectly attributed to) (YANG *et al.* 2010) weights each marker differentially, using a matrix of weights $\boldsymbol{D}_w$. $Var(\boldsymbol{u}) = \boldsymbol{ZD}_w\,\boldsymbol{Z}'\sigma^2_u$ where genomic relationships are

$$\boldsymbol{G} = \boldsymbol{ZD}_w\boldsymbol{Z}'$$

with

$$\boldsymbol{D_w} = \begin{pmatrix} \frac{1}{n\,2p_1q_1} & 0 & ... & 0 \\ 0 & \frac{1}{n\,2p_2q_2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \frac{1}{n\,2p_nq_n} \end{pmatrix}$$

Where n is the number of markers. This matrix can be interpreted as a weighted average of genomic relationships, one by marker:

$$\boldsymbol{G} = \frac{1}{nsnp}\sum_{i=1}^{nsnp} \boldsymbol{G}_i = \frac{1}{nsnp}\sum_{i=1}^{nsnp} \frac{\boldsymbol{z}_i\boldsymbol{z}_i'}{2p_iq_i}$$

where $\boldsymbol{z}_i$ is a vector with genotypes for marker i. This corresponds as well to $Var(\boldsymbol{u}) = \boldsymbol{ZDZ}'$ where

$$\boldsymbol{D} = \begin{pmatrix} \frac{\sigma_u^2}{n\ 2p_1q_1} & 0 & ... & 0 \\ 0 & \frac{\sigma_u^2}{n\ 2p_2q_2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \frac{\sigma_u^2}{n\ 2p_nq_n} \end{pmatrix}$$

This "second" genomic relationship, that is quite used, has several problems. The first is that is very sensible to small allelic frequencies, that will give high weight to very rare alleles. For monomorphic alleles ($p = 0$ or 1) the matrix is undefined, which is not the case in the "first $\boldsymbol{G}$"

The second problem is that it assumes that the contribution of each marker to the overall $\boldsymbol{G}$ are identical in terms of variance, which means that markers with small allelic frequencies have large effects. The genetic variance contributed by marker i is equal to σ_u^2/n, irrespectively of its allelic frequence, and $\sigma_{ai}^2 = \sigma_u^2/n2p_iq_i$. Consider two loci with different allelic frequencies $\{0.1,0.5\}$ and $\sigma_u^2 = 1$. The first loci will have $\sigma_{a1}^2 = 5.5$ and the second $\sigma_{a2}^2 = 2$. Therefore, using this matrix imposes different *a priori* variances of markers depending on their frequencies. This has no biological reason, in my opinion.

7.4.3 Allelic frequencies to put in genomic relationships

There is some confusion on the allelic frequencies to use in the construction of $\boldsymbol{G}$. (STRANDÉN and CHRISTENSEN 2011) proved that, if the form is $\boldsymbol{G} = \boldsymbol{ZZ}'/2\sum p_iq_i$, the allele frequencies used to construct $\boldsymbol{Z}$ are irrelevant, and the only change from using different allelic frequencies is that they shift by a constant that is absorbed by the mean. To obtain unbiased values in the same scale as regular relationships, one should use base population allelic frequencies.

However, the allelic frequency in the denominator is more important. The expression $\sigma_u^2 = 2\sum_{i=1}^{nsnp} p_iq_i\sigma_{a0}^2$ puts genetic variance in one population as a function of the allelic frequencies in the same population. Thus, dividing by the current allelic frequencies implies that we refer to the current genetic variance. If there are many generations between current genotypes and pedigree base the genetic variance will reduce. Ways to deal with these will be suggested later.

7.4.4 Properties of G

We will refer here to properties derived for $\boldsymbol{G} = \boldsymbol{ZZ}'/2\sum p_iq_i$ if "observed" genomic relationships are used.

7.4.4.1 The average value of $\boldsymbol{u}$ is 0

The first property is that the average value of $\boldsymbol{u}$ is 0, because $\boldsymbol{Z}$ is centered. This only requires Linkage Equilibrium.

7.4.4.2 The average value of $\boldsymbol{G}$ is 0

The second property is that, the average value of $\boldsymbol{G}$ is 0. This is a corollary of the previous but can be proven as follows. First, we have that $sum(\boldsymbol{ZZ'}) = sum(\boldsymbol{Z'Z})$. In case of Linkage Equilibrium, terms of $\boldsymbol{Z'Z}$ sum to zero, for the following. These are the crossproducts of covariables associated with loci i and j. In LE, these crossproducts occur with frequency $(1-p_i)(1-p_j)$ for the co-occurrence of alleles "a" in i and "a" in j, $(p_i)(1-p_j)$ for "A" and "a", and so on. Then, by summing in order genotypes at respective loci i and j "a" and "a", "a"' and "A", "A" and "a", and "A" and "A", weighted by the respective frequencies:

$$\begin{aligned} E(\boldsymbol{z}_i'\boldsymbol{z}_j) &= (1-p_i)(1-p_j)(-p_i)(-p_j) \\ &+ (p_i)(1-p_j)(1-p_i)(-p_j) \\ &+ (1-p_i)(p_j)(-p_i)(1-p_j) \\ &+ (p_i)(p_j)(1-p_i)(1-p_j) = 0 \end{aligned}$$

A verbal explanation is that, if the average value of $\boldsymbol{u}$ is if 0, then some animals will be more related than the average and others less related than the average – hence the 0 average relationship.

7.4.4.3 The average value of the diagonal of $\boldsymbol{G}$ is 1 if there is no inbreeding

This requires Hardy-Weinberg (but not linkage equilibrium). This can be seen by noting that $\mathrm{tr}(\mathrm{ZZ'}) = \mathrm{tr}(\mathrm{Z'Z})$ where tr is the trace operator. The expression $\mathrm{tr}(\mathrm{Z'Z})$ is the sum of squared covariables corresponding to effects of alleles "a" and "A", which occur in m animals with respective frequencies $1-\mathrm{p_i}$ and $\mathrm{p_i}$ in locus i. This is:

$$\mathrm{z_i'z_i} = 2\mathrm{m}[(1-\mathrm{p_i})\mathrm{p_i^2} + \mathrm{p_i}(1-\mathrm{p_i})^2] = 2\mathrm{mp_i}(1-\mathrm{p_i}) = 2\mathrm{mp_iq_i}$$

Therefore, the diagonal of G has an average of

$$\frac{1}{\mathrm{m}}\mathrm{tr}\left(\frac{\mathrm{ZZ'}}{2\sum \mathrm{p_iq_i}}\right) = \frac{2\mathrm{m}\sum \mathrm{p_iq_i}}{2\mathrm{m}\sum \mathrm{p_iq_i}} = 1$$

If there is inbreeding there is not Hardy-Weinberg, and there is an inbreeding of F then the genotypes are distributed according to $\{\mathrm{q^2 + pqF}, 2\mathrm{pq}(1-\mathrm{F}), \mathrm{p^2 + pqF}\}$ (FALCONER and MACKAY 1996). Then we multiply each value of z by its frequency:

$$\begin{aligned} \mathrm{z_i'z_i} &= 2\mathrm{m}[(1-2\mathrm{p_i})(\mathrm{q_i^2} + \mathrm{p_iq_iF}) + (1-2\mathrm{p_i})(2\mathrm{p_iq_iF}) + (2-2\mathrm{p_i})(\mathrm{p_i^2} + \mathrm{p_iq_iF})] \\ &= 2\mathrm{mp_i}(1-\mathrm{p_i}) = 2\mathrm{m}(1+\mathrm{F})\mathrm{p_iq_i} \end{aligned}$$

The diagonal of G has in this case an average of

$$\frac{1}{\mathrm{m}}\mathrm{tr}\left(\frac{\mathrm{ZZ'}}{2\sum \mathrm{p_iq_i}}\right) = \frac{(1+\mathrm{F})2\mathrm{m}\sum \mathrm{p_iq_i}}{2\mathrm{m}\sum \mathrm{p_iq_i}} = 1 + \mathrm{F}$$

Note that F here is a within-population inbreeding, and can be negative, indicating excess of homozygosity (e.g., in an F1 population).

7.4.4.4 The average value of the off-diagonal of $\boldsymbol{G}$ is almost 0

This is the case if both Hardy-Weinberg and linkage equilibrium hold. If there are m genotyped animals, we have that the value of the off-diagonal is:

$$avoff(\boldsymbol{G}) = \frac{1}{m(m-1)}\big(sum(\boldsymbol{G}) - diag(\boldsymbol{G})\big) = \frac{m}{m(m-1)} = \frac{1}{m-1}$$

which is very close to zero.

7.4.5 Weighted Genomic relationships

We have seen that Bayesian Regressions are an option for genomic selection. Somehow, they consider that different markers may have different variances. This can be implemented using

$$Var(\boldsymbol{u}) = \boldsymbol{Z}Var(\boldsymbol{a})\boldsymbol{Z}' = \boldsymbol{ZDZ}'$$

Alternatively, and mainly for ease of implementation (e.g., in BLUPF90 or AsReml) this can be obtained factorizing out the genetic variance and using a matrix of weights as in $Var(\boldsymbol{u}) = \boldsymbol{ZD}_w\,\boldsymbol{Z}'\sigma_u^2$ with

$$\boldsymbol{D}_w = \begin{pmatrix} \sigma_{a1}^2/\sigma_{a0}^2 & 0 & \dots & 0 \\ 0 & \sigma_{a2}^2/\sigma_{a0}^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_{an}^2/\sigma_{a0}^2 \end{pmatrix} = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_n \end{pmatrix}$$

Note that if $w_1 = w_2 = \cdots = w_n = 1$ this is regular genomic relationships.

Marker variances or weights can be obtained in several ways. (ZHANG *et al.* 2010) and (LEGARRA *et al.* 2011b) suggested to obtain them from Bayesian Regressions, with good results. (SHEN *et al.* 2013) suggested a REML-like strategy that we evoked before, and (SUN *et al.* 2012) proposed a simple (but seriously biased) algorithm to get SNP-specific variances. Another option is to use VanRaden's nonLinearA to obtain updates for $\boldsymbol{D}$.

7.5 Genomic relationships as estimators of realized relationships

The notion of *actual or realized relationship* is of utmost importance for genomic selection. Pedigree relationships assume an infinitesimal model with infinite unlinked genes. At one locus, two full-sibs may share one, two or none alleles. Across all loci, two full sibs share exactly half their genome in the infinitesimal model. This is no longer true with real chromosomes: chromosomes tend to be transmitted together and therefore two half-sibs may inherit vary different dotations, as shown in the Figure below. The paper of VanRaden (VANRADEN 2007) makes a very good review of the subject.

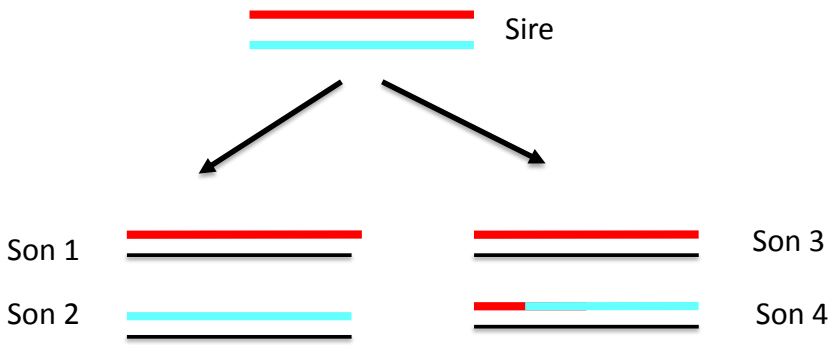


Figure 23. Different transmission of one chromosome from sire to four half-sibs. Different maternal chromosomes are in black.

In this example, sons 1 and 3 are more alike than sons 2 and 4. Therefore, in prediction of son 3, son 1 should be given more weight than sons 2 and 4. These "real" relationships are

called *realized* relationships as opposed to expected relationships. (HILL and WEIR 2011) used the notation R_{ij} to the realized relationship, which we will follow. Expressions for the difference between expected (A_{ij}) and realized (R_{ij}) relationships were given by (VANRADEN 2007) (HILL and WEIR 2011; GARCIA-CORTES *et al.* 2013).

In theory, one can define realized relationships in the same way as regular relationships, assuming an unrelated base population, in which case they are identical by descent relationships. In this case,

$$E(R_{ij}) = A_{ij}$$

This important result means that if we simulate meiosis of chromosomes from the sire to the two half-sibs 1 and 2, at each simulation may have a realized relationship between the two half sibs. This realized relationship will vary between 0 and 0.5, but on average across the simulations it will be 0.25, which is the value of A_{ij} .

. Deviations are skewed and the ratio devation/expectation is high for low related animals. This means that two third-degree cousins may actually not share any allele. Markers can see these differences. (LUAN *et al.* 2012) suggested to obtain realized relationships from a pure identity by descent approach, based on computation of probability transmission from parents to offspring with the help of pedigree and markers (FERNANDO and GROSSMAN 1989) (MEUWISSEN and GODDARD 2010), which assumes that founders of the pedigree are unrelated. This has two drawbacks. The first one is that major genes are ignored (because closely associated markers will be ignored). The second one is that computing becomes rather difficult when genotyped animals do not form a complete pedigree (MEUWISSEN and GODDARD 2010).

However, Cockerham's result $Cov(z_i, z_j) = R_{ij}2pq$ actually involves realized relationships. Then, we can reverse the formulae and estimate those relationships as $R_{ij} = Cov(z_i, z_j)/2pq$. However, summed over many markers, $Cov(z_i, z_j)/2pq$ is VanRaden's genomic relationship, which therefore is an estimator of realized relationship, and an estimator that uses markers to infer relationships. The duality of VanRaden's formulation using genomic relationships is that at the same time it refers to marker effects and to relationships.

If genomic relationships G_{ij} are an unbiased estimator of realized relationships R_{ij}, then

$$E(\boldsymbol{G}) = \boldsymbol{A}$$

This raises another question. If realized relationships R_{ij} can be defined as IBD relationships, then one should not get negative values. Does this means that we should turn negative values in $\boldsymbol{G}$ to zero? The answer is NO. For individuals that are suspected to have 0 relationships, ($A_{ij} = 0$), this means that G_{ij} can oscillate between positive and negative values. However, if we don't use base allelic frequencies, then $\boldsymbol{G}$ is biased with respect to $\boldsymbol{A}$ and underestimates relationships.

7.5.1 Other estimators of (genomic) relationships

In the conservation genetics literature, a common technique is to use molecular relationships (r_{Mij}) corrected by allelic frequencies, using one of the previous results:

$$\hat{R}_{ij} = \frac{r_{Mij} - 2p^2 - 2q^2}{2pq}$$

There are many variants of this expression (LYNCH 1988; TORO *et al.* 2011) (RITLAND 1996) . Values of $\hat{R}_{ij}$ can also be negative, and some set their values to zero. This is a gross mistake, first for the arguments above and second, because it greatly compromises numerical computations ($\hat{R}_{ij}$ corrected like that do not form a positive definite covariance matrix).

7.6 Compatibility of genomic and pedigree relationships

VanRaden's $\boldsymbol{G}$ is dependant on the use of base allelic frequencies. For some populations where old ancestors are genotyped (e.g., some populations of dairy cattle), this is feasible. However this is not the case in many populations. For instance the Lacaune dairy sheep started recording pedigree and data in the 60's, while DNA is stored since the 90's. This causes two problems (that are also problems for Bayesian Regressions):

1. The genetic base is no longer the same for pedigree and marker. We have seen that, by construction, using "centered" coding leads to an imposed average $\overline{\boldsymbol{u}} = 0$ across your population. This is contradictory with the pedigree, which imposes $\overline{\boldsymbol{u}} = 0$ *only* across the founders of the pedigree.

For instance, trying to compare pedigree-based EBV's and genomic-based EBV's, they will be a shift in scale. This shift can be accounted for by selecting a group of animals and referring all EBV's to their average EBV in both cases. Remember that the result of (STRANDÉN and CHRISTENSEN 2011) warrants that there will only be a shift in estimates of $\boldsymbol{u}$, but the differences across breeding values will be identical.

2. The genetic variance changes. The pedigree-based genetic variance σ_u^2 refers to the variance of the breeding values of the founders of the pedigree. The marker-based genetic variance $2\sum p_i q_i \, \sigma_{a0}^2$ refers to the variance of a population with allelic frequencies p_i. These are typically "current" observed allele frequencies. However, in a pedigree markers tend to fix by drift and selection and therefore $2\sum p_i q_i \, \sigma_{a0}^2$ is lower using current frequencies than base allele frequencies.

Equating $\sigma_{a0}^2 = \sigma_u^2 / 2\sum p_i q_i$ will tend to underestimate σ_{a0}^2 . This can be solved if instead of using this expression to obtain σ_{a0}^2, one estimates σ_{a0}^2 or marker variances directly, as in BayesC, Bayesian Lasso, or GREML (see later).

These problems are only relevant if one tries to combine pedigree-based information and genomic-based information. In the following, we will use the following notation. $\boldsymbol{u}_{base}$ are the animals of the genetic base of the pedigree (i.e., the founders). $\boldsymbol{u}_2$ are genotyped animals, and $\boldsymbol{u}_1$ are ungenotyped animals.

7.6.1 Use of Gengler's method

Gengler's method can be used to estimate base allele frequencies(GENGLER *et al.* 2007) (VANRADEN 2008). It has, however, been rarely used; one of the reasons is that estimate may go out of bounds (e.g. allelic frequencies beyond 0 or 1), and (AGUILAR *et al.* 2010) reported poor results.

7.6.2 Compatibility of genetic bases

This is detailed in (VITEZICA *et al.* 2011). If base alleles are not available, one may use current allele frequencies (i.e. frequencies in genotypes of $\boldsymbol{u}_2$). We know that, by construction of $\boldsymbol{G}$, the mean of $\boldsymbol{u}_2$ is set to zero: $p(\boldsymbol{u}_2) = N(0, \boldsymbol{G}\sigma_u^2)$. The difference of both means can be modelled as random : $\mu = \overline{\boldsymbol{u}}_2 - \overline{\boldsymbol{u}}_{base} = \overline{\boldsymbol{u}}_2 = \frac{1}{m}\mathbf{1}'\boldsymbol{u}_2$ where m is the number of individuals in $\boldsymbol{u}_2$.

In an infinite population with no selection, there would be no difference between $\overline{\boldsymbol{u}}_2$ and $\overline{\boldsymbol{u}}_{bas}$. However, in an finite population there is selection, drift, or both. In this case we can model that $\boldsymbol{u}_2$ has an a priori mean $p(\boldsymbol{u}_2|\mu) = N(\mu, \boldsymbol{G}\sigma_u^2)$. This mean is actually the result of random factors (selection and drift) and therefore is a random variable with some variance $\sigma_\mu^2 = a\sigma_u^2$ (a was called α in (VITEZICA *et al.* 2011)). Integrating this mean from the expression $p(\boldsymbol{u}_2|\mu)p(\mu) = N(\mu, \boldsymbol{G}\sigma_u^2)N(0, \sigma_\mu^2)$ we have that

$$p(\boldsymbol{u}_2) = N(0, \boldsymbol{G}^*\sigma_u^2)$$

where $\boldsymbol{G}^* = (\boldsymbol{G} + \mathbf{1}\mathbf{1}'a)\sigma_u^2$ is a "tuned" genomic relationship which takes into account our ignorance as to the difference between pedigree and genomic genetic bases. The $\mathbf{1}\mathbf{1}'$ operator simply adds the constant a to every element of $\boldsymbol{G}$. Informally we may write $\boldsymbol{G}^* = a + \boldsymbol{G}$.

To obtain a value for σ_μ^2, we know based on pedigree that the $Var(\boldsymbol{u}_2) = \boldsymbol{A}_{22}\sigma_u^2$. Therefore $Var\left(\frac{1}{m}\mathbf{1}'\boldsymbol{u}_2\right) = \frac{1}{m^2}(\mathbf{1}'\boldsymbol{A}_{22}\mathbf{1}\sigma_u^2) = \overline{\boldsymbol{A}}_{22}\sigma_u^2$, where $\boldsymbol{A}_{22}$ is the pedigree relationship matrix and the bar means "average over values of $\boldsymbol{A}_{22}$". Based on genomics, this variance would be $Var\left(\frac{1}{m}\mathbf{1}'\boldsymbol{u}_2\right) = \frac{1}{m^2}(\mathbf{1}'\boldsymbol{G}\mathbf{1} + \mathbf{1}'\mathbf{1}\mathbf{1}'\mathbf{1}a)\sigma_u^2 = (\overline{\boldsymbol{G}} + a)\sigma_u^2$. If we equate both variances, we have that

$$a = \overline{\boldsymbol{A}}_{22} - \overline{\boldsymbol{G}}$$

It can be noted that in Hardy-Weinberg equilibrium, $\overline{\boldsymbol{G}} = 0$ and $a = \overline{\boldsymbol{A}}_{22}$.

Adding constant a as in $\boldsymbol{G}^* = \boldsymbol{G} + \mathbf{1}\mathbf{1}'a$ makes, by construction, that both evaluations are in the same scale. This way of getting a value for a is called *method of moments* and guarantees unbiasedness. The genetic interpretation is simple. Constructing $\boldsymbol{G}$ with current allele frequencies underestimates relationships from the base population. We estimate this underestimation from the average difference between $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$. Adding a constant to every element of $\boldsymbol{G}$ ensures that genomic relationships are, on average, on the same genetic base than pedigree relationships.

7.6.3 Compatibility of genetic variances

In VanRaden's formulation of $= \boldsymbol{Z}\boldsymbol{Z}'/2\sum p_i q_i$, the divisor comes because of the assumption that the genetic variance is $\sigma_u^2 = 2\sum p_i q_i \sigma_{a0}^2$. However the product $2\sum p_i q_i$ will be too low if we use current allelic frequencies with respect to base allelic frequencies. Therefore we seek for an adjustment

$$\boldsymbol{G}^* = b\boldsymbol{G}$$

where b accounts for the ratio of "current" $2\sum p_i q_i$ to "base" $2\sum p_i q_i$ and is typically lower than 1 (i.e., the genetic variance has reduced).

The reasoning to solve this issue is as follows. Consider the genetic variance of the genotyped individuals in $\boldsymbol{u}_2$; I will call this S^2_{u2} to stress that this is a variance of a particular population, *not* the variance of the genetic base. This is $S^2_{u2} = \frac{1}{m}\boldsymbol{u}'_2\boldsymbol{u}_2 - \bar{\boldsymbol{u}}^2_2$. This S^2_{u2} has a certain distribution under either pedigree or genomic modelling. As we did with genetic bases, we will equate, on expectation, the two S^2_{u2} .

Under pedigree relationships we have that (SEARLE 1982) p. 355:

$$E(S^2_{u2}) = \left(\frac{1}{m}tr(\boldsymbol{A}_{22}) - \bar{\boldsymbol{A}}_{22}\right)\sigma^2_u = \left(1 + \bar{F}_p - \bar{\boldsymbol{A}}_{22}\right)\sigma^2_u$$

Under genomic relationships we have that:

$$E(S^2_{u2}) = \left(\frac{1}{m}tr(b\boldsymbol{G}) - b\bar{\boldsymbol{G}}\right)\sigma^2_u = b\left(1 + \bar{F}_g - \bar{\boldsymbol{G}}\right)\sigma^2_u$$

where $\bar{F}_p$ is average pedigree inbreeding and $\bar{F}_g$ is average genomic inbreeding. Equating both expectations we have that

$$b = \frac{\left(1 + \bar{F}_p - \bar{\boldsymbol{A}}_{22}\right)}{\left(1 + \bar{F}_g - \bar{\boldsymbol{G}}\right)}$$

A close result was showed by (FORNI *et al.* 2011) who had genomic inbreeding. In Hardy-Weinberg conditions, we have seen that $\bar{\boldsymbol{G}} = 0$ and $\bar{F}_g = 0$ (the average diagonal is 1). On the other hand, if matings are at random, $\bar{F}_p = \bar{\boldsymbol{A}}_{22}/2$. Therefore:

$$b = 1 - \frac{\bar{F}_p}{2}$$

And in that case, $b = 1 - a/2$ above. Which results in $b < 1$. This means that the genetic variance lowered from the pedigree base to the genotyped population. Thus, the multiplication by b corrects for the fixation of alleles due to inbreeding.

7.6.4 Compatibility of genetic bases and variances

With the two pieces above, it is easy to see that a compatible matrix $\boldsymbol{G}^* = a + b\boldsymbol{G}$ can be obtained by the expressions above for a and b. (VITEZICA *et al.* 2011) based on (POWELL *et al.* 2010) observed that relationships in a "recent" population in an "old" population scale can be modelled using Wright's fixation indexes. Translated to our context, this gives $a = \bar{\boldsymbol{A}}_{22}$ and $b = 1 - \frac{a}{2}$, which is the same result as above if Hardy-Weinberg holds.

Christensen et al. (2012) remarked that the hypothesis of random mating population is not likely for the group of genotyped animals, since they would born in different years and some being descendants of others, and suggested to infer a and b from the system of two equations equating average relationships and average inbreeding: $\frac{tr(\boldsymbol{G})}{m}b + a = \frac{tr(\boldsymbol{A}_{22})}{m}$ and $a + b\bar{\boldsymbol{G}} = \bar{\boldsymbol{A}}_{22}$. This is basically a development as above. They further noticed that in practice $b \approx 1 - a/2$ because the deviation from Hardy-Weinberg was small.

(VANRADEN 2008) suggested a regression of observed on expected relationships, minimizing the residuals of $a + b\boldsymbol{G} = \boldsymbol{A}_{22} + \boldsymbol{E}$. This idea was generalized to several breed origins by (HARRIS and JOHNSON 2010). The distribution of $\boldsymbol{E}$ is not homoscedastic and this precluded scholars from trying this approach because it would be sensible to extreme values

(Christensen et al., 2012), e.g., if many far relatives are included, for which the deviations in $\boldsymbol{E}$ can be very large.

Finally, (CHRISTENSEN 2012) argued that relationships in $\boldsymbol{G}$ do not depend on pedigree depth, and they are exact in some sense. He suggested to take as reference the 101 coding (i.e., set the frequencies to 0.5) and then "tune" pedigree relationships in $\boldsymbol{A}$ to match genomic relationships in $\boldsymbol{G}$. He introduced two extra parameters, γ and s. The γ parameter can be understood as the overall relationship across the base population such that current genotypes are most likely, and integrates the fact that the assumption of unrelatedness at the base population is false in view of genomic results (two animals who share alleles at markers are related even if the pedigree is not informative). More precisely, he devised a new pedigree relationship matrix, $\boldsymbol{A}(\gamma)$ whose founders have a relationship matrix $\boldsymbol{A}_{base} = \gamma + \boldsymbol{I}(1 - \gamma/2)$. Parameter s, used in $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Z}'/s$ can be understood as the counterpart of $2\Sigma p_i q_i$ (heterozygosity of the markers) in the base generation. Both parameters can be deduced from maximum likelihood. This model is the only one which accounts for all the complexities of pedigrees (former ones are based on average relationships) but it has not been tested with real data so far.

7.7 Singularity of G

Matrix $\boldsymbol{G}$ might (and usually is) singular. There are two reasons for this. First, if there are clones or identical twins, two genotypes in $\boldsymbol{Z}$ will be identical and therefore two animals will show a correlation of exactly 1 in $\boldsymbol{G}$. Second, if genotypes in $\boldsymbol{Z}$ use "centered" coding with observed allele frequencies, then the matrix is singular (last row can be predicted from the other ones) (STRANDÉN and CHRISTENSEN 2011).

To obtain an invertible $\boldsymbol{G}$ and then use $\boldsymbol{G}^{-1}$ in the mixed model equations, there are two ways. The first one is to use a modified $\boldsymbol{G}_w = (1 - \alpha)\boldsymbol{G} + \alpha\boldsymbol{I}$,with α a small value (typically 0.05 or 0.01). The second option consists in mixing genomic and pedigree relationships. If $\boldsymbol{A}_{22}$ is the matrix of genotyped animals, we might use a modified "weighted" $\boldsymbol{G}_w = (1 - \alpha)\boldsymbol{G} + \alpha\boldsymbol{A}_{22}$. This is the default in the Blupf90 package, which uses $\alpha = 0.05$. A more detailed explanation is in the next section.

7.8 Including residual polygenics effects in G

One may consider that not all genetic variance is captured by markers. This can be shown by estimating variance assigned to markers and pedigree (LEGARRA *et al.* 2008) (RODRÍGUEZ-RAMILO *et al.* 2014) (JENSEN *et al.* 2012) (CHRISTENSEN and LUND 2010) or because some genomic evaluation procedures give better cross-validation results when an extra polygenic term based exclusively on pedigree relationships is added (e.g. (SU *et al.* 2012)). Let us decompose the breeding values of genotyped individuals in a part due to markers and a residual part due to pedigree, $\boldsymbol{u} = \boldsymbol{u}_m + \boldsymbol{u}_p$ with respective variances $\sigma_u^2 = \sigma_{u,m}^2 + \sigma_{u,p}^2$. It follows that $Var(\boldsymbol{u}_2) = (\alpha\boldsymbol{G} + (1 - \alpha)\boldsymbol{A}_{22})\sigma_u^2$ where $\alpha = \sigma_{u,m}^2/\sigma_u^2$. Therefore, the simplest way to include the residual polygenic effects is to create a modified genomic relationship matrix $\boldsymbol{G}_w$ ($\boldsymbol{G}$ in (AGUILAR *et al.* 2010); $\boldsymbol{G}_w$in (VANRADEN 2008) (CHRISTENSEN 2012) as $\boldsymbol{G}_w = \alpha\boldsymbol{G} + (1 - \alpha)\boldsymbol{A}_{22}$**.** In practice, the value of α is low and has negligible effects on predictions.

8. GBLUP

8.1 Single trait animal model GBLUP

With genomic relationships well defined in the previous section as (rather generally) $Var(\boldsymbol{u}) = \boldsymbol{ZDZ'} = \boldsymbol{ZD}_w\boldsymbol{Z'}\sigma_u^2 = \boldsymbol{G}\sigma_u^2$ (and perhaps after some compatibility "tuning" as before), the construction of genomic predictions in GBLUP form is straightforward. We have the following linear model:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Wu} + \boldsymbol{e}$$

where $\boldsymbol{W}$ is a matrix linking phenotypes to individuals. Then $Var(\boldsymbol{u}) = \boldsymbol{G}\sigma_u^2$, $Var(\boldsymbol{e}) = \boldsymbol{R}$. We may also assume multivariate normality. Under these assumptions, Best Predictios, or Conditional Expectations, of breeding values in $\boldsymbol{u}$ can be obtained by Henderson's mixed model equations as:

$$\begin{pmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}W} \\ \boldsymbol{W'R^{-1}W} & \boldsymbol{W'R^{-1}W + G^{-1}}\sigma_u^{-2} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{W'R^{-1}y} \end{pmatrix}$$

If $\boldsymbol{R} = \boldsymbol{I}\sigma_e^2$, then the variance components can be factored out and the equations become:

$$\begin{pmatrix} \boldsymbol{X'X} & \boldsymbol{X'W} \\ \boldsymbol{W'X} & \boldsymbol{W'W} + \boldsymbol{I}\lambda \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'y} \\ \boldsymbol{W'y} \end{pmatrix}$$

with $= \sigma_e^2/\sigma_u^2$.

These equations are identical to regular animal model, with the exception that genomic relationships $\boldsymbol{G}$ are used instead of pedigree relationships . They have some very nice features:

1. Any model that has been developed in BLUP can be immediately translated into GBLUP. This includes maternal effects model, random regression, competition effect models, multiple trait, etc.
2. All genotyped individuals can be included, either with phenotype or not. The only difference is that the corresponding element in $\boldsymbol{W}$ is set to 0.
3. Regular software works if we include a mechanism to include $\boldsymbol{G^{-1}}$
4. Developments including mixed model equations apply to GBLUP as well. Therefore GREML and G-Gibbs are simple extensions

8.2 Multiple trait GBLUP

This is straightforward as well. The multiple trait mixed model equations are:

$$\begin{pmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}W} \\ \boldsymbol{W'R^{-1}W} & \boldsymbol{W'R^{-1}W + G^{-1}}\otimes\boldsymbol{G_0^{-1}} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{W'R^{-1}y} \end{pmatrix}$$

where $\boldsymbol{G}_0$ is the matrix of genetic covariance across traits, and usually $\boldsymbol{R} = \boldsymbol{I} \otimes \boldsymbol{R}_0$, where $\boldsymbol{R}_0$ is the matrix of residual covariances. Note that these equations work perfectly well with missing traits.

8.2.1 Reliabilities from GBLUP

Nominal reliabilities (NOT cross-validation reliabilities) can be obtained from the Mixed Model equations, as:

$$Rel_i = 1 - \frac{C^{ii}}{G_{ii}\sigma_u^2}$$

where C_{ii} is the element of the inverse of the mixed model equations in its first form (i.e., with explicit σ_u^2). However there is a word of caution. Depending how the coding of $\boldsymbol{Z}$ proceeds, the numerical values of Rel_i change, although EBV's only shift by a constant (STRANDÉN and CHRISTENSEN 2011). This result is problematic because reporting reliabilities becomes tricky.

8.3 GBLUP with singular G

If $\boldsymbol{G}$ is singular, one can use alternative mixed model equations (HARVILLE 1976) (HENDERSON 1984):

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{W} \\ \boldsymbol{G}\sigma_u^2\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} & \boldsymbol{G}\sigma_u^2\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{G}\sigma_u^2\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$$

Or a symmetric form that fits better into regular algorithms:

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{G}\sigma_u^2 \\ \boldsymbol{G}\sigma_u^2\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} & \boldsymbol{G}\sigma_u^2\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{G}\sigma_u^2 + \boldsymbol{G}\sigma_u^2 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{G}\sigma_u^2\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$$

From this, $\hat{\boldsymbol{u}} = \boldsymbol{G}\sigma_u^2\hat{\alpha}$.

8.4 From GBLUP to marker estimates

Because $\boldsymbol{G}$ is formed from marker effects, the algebra warrants that estimates are the same under either GBLUP or BLUP-SNP (VANRADEN 2008), provided that parameterizations are strictly identical (same $\boldsymbol{Z}$, same p's, same variances, etc). This is up to the numerical error produced by forcing $\boldsymbol{G}$ to be invertible; this numerical error is most often negligible. More formal proofs can be found in (HENDERSON 1973) and (STRANDÉN and GARRICK 2009).

If breeding values $\boldsymbol{u} = \boldsymbol{Z}\boldsymbol{a}$ and $Var(\boldsymbol{a}) = \boldsymbol{D}$, then the joint distribution of breeding values $\boldsymbol{u}$ and marker effects $\boldsymbol{a}$ is (HENDERSON 1973) (STRANDÉN and GARRICK 2009):

$$Var\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{a} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' & \boldsymbol{Z}\boldsymbol{D} \\ \boldsymbol{D}_{\boldsymbol{a}}\boldsymbol{Z}_{\boldsymbol{2}}' & \boldsymbol{D}_{\boldsymbol{a}} \end{pmatrix}$$

where, usually, $\boldsymbol{D} = \boldsymbol{I}\sigma_u^2/2\Sigma p_i q_i$. Assuming multivariate normality,

$$\hat{\boldsymbol{u}} \,|\hat{\boldsymbol{a}} = \boldsymbol{Z}'\hat{\boldsymbol{a}}$$

(the breeding value is the sum of marker effects) and

$$\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{u}} = \boldsymbol{DZ}'(\boldsymbol{ZDZ}')^{-1}\,\widehat{\boldsymbol{u}} = \boldsymbol{D}\,\boldsymbol{Z}'\,\boldsymbol{G}^{-1}\,\sigma_u^{-2}\,\widehat{\boldsymbol{u}}$$

If $\boldsymbol{D} = \boldsymbol{I}\sigma_u^2/2\Sigma p_i q_i$ this becomes

$$\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{u}} = \boldsymbol{DZ}'(\boldsymbol{ZDZ}')^{-1}\,\widehat{\boldsymbol{u}} = \frac{\boldsymbol{Z}'\,\boldsymbol{G}^{-1}\,\widehat{\boldsymbol{u}}}{2\Sigma p_i q_i}$$

where (as discussed in previous sections) $\boldsymbol{ZDZ}' = \boldsymbol{G}\sigma_u^2$, so that marker effects can be deduced from breeding values by backsolving using the genomic relationship matrix and markers' incidence matrix.

8.5 GREML and G-Gibbs

Use of genomic relationships to estimate variance components is trivial, and popular methods REML and Gibbs sampler have often been used (CHRISTENSEN and LUND 2010) (RODRÍGUEZ-RAMILO *et al.* 2014) (JENSEN *et al.* 2012). Also, older estimates using relationships based on markers are common in the conservation genetics literature. Often, people call GBLUP something that in fact is GREML. The difference is that in the latter variance components are obtained, whereas in GBLUP these are fixed *a priori*.

As discussed, the estimates obtained by GREML or G-Gibbs refer to a base population with the assumed allelic frequencies (usually the observed ones) and in Hardy-Weinberg equilibrium. Therefore, these estimates are not necessarily comparable to pedigree estimates, that refer to another base population. Further, data sets are often different, making comparison unreliable. In particular, heritability estimates using so-called "unrelated" populations (YANG *et al.* 2010) have large standard errors (making comparisons unreliable) and refer to a very particular population, whereas pedigree-based estimates refer to *another* population.

9. Appendix. Working codes to solve BLUP-SNP (or RR-BLUP) in R

```
# (c) Andres Legarra, INRA
# Legarra & Misztal, Journal of Dairy Science 2008
# X has 1 column of "1"s for the mean and nsnp columns for the snp
genotypes coded numerically (-101, 012, or centered)
# X has "number of individuals" rows
# y has phenotypes; rhs=X'y=crossprod(X,y)


solve_densem_pcg_X <- function(X,rhs,vara,vare){
   # solve [X'X +(I vara)-1] sol = rhs by preconditioned conjugate gradient
for densem A
   # it includes an overall mean, the rest are SNPs
   # all effects are random but the mean

   n=length(rhs)
   m=rep(0,n)
   De=rep(1/vara,n)
   De[1]=0
   for (i in 1:n){
     m[i]=1/( crossprod(X[,i])/vare + De[i])
   }
   sol=rep(0,n)
   p=rep(0,n)
   z=rep(0,n)
   w=rep(0,n)
   r=rhs

   for (k in 1:1000){
     z=m*r
     tau=crossprod(z,r)
      if (k == 1){
         beta=0
              p=z
       }else{
         beta=as.numeric(tau/oldtau)
         p=z+beta*p
      }
      w=crossprod(X,X%*%p)/vare + De*p
      alpha=as.numeric(tau/crossprod(p,w))
      sol=sol+alpha*p
      if ((k%%100) != 0){
           r=r-alpha*w
         }else{
           r=rhs-crossprod(X,X%*%sol)/vare - De*sol
      }
      conv=crossprod(r,r)/crossprod(rhs,rhs)
      #print(c('round ',k,'   convergence=',conv))
      if (conv < 1e-14) break
      oldtau=tau
   }
   #print(c(k,' iterations,   convergence criterion=',conv))
   sol
}


GSRU <- function(y,X,vare,vara,ahat=rep(0,dim(X)[2])){
```

```
  # this function solves a blup for genomic selection
  # given design matrix X, data y, residual variance vare, SNP variance
vara
  # solve by Gauss Seidel
  neq=dim(X)[2]
  e=y-X%*%ahat
  #ahat=rep(0,neq)
  xpx=rep(0,neq)
  epsit=rep(0,1)
  # this is GSRU,
  #set up diag(X'X)
  for (i in 1:neq) {
    xpx[i]=crossprod(X[,i],X[,i])
  }
  mu=0
  # do until convergence
  for (iter in 1:1000) {
    #Gauss Seidel
    eps=0
    for (i in 1:neq){
      lhs=xpx[i]/vare
      if(i>1) lhs=lhs+1/vara
      rhs=crossprod(X[,i],e)/vare + xpx[i]/vare*ahat[i]
      val=rhs/lhs
      eps=eps+((val - ahat[i])**2)
      e = e - X[,i]*(val - ahat[i])
      ahat[i]=val
    }
    eps=eps/sum(sol**2)
    if(iter%%10==0) print(c(iter,eps))
    epsit[iter]=eps
    if(eps<1E-10) break
  }
  print(c(iter,eps,date()))
  ahat
}
```

10. References

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

Casella, G., and R. L. Berger, 1990 *Statistical inference*. Duxbury Press Belmont, CA.

Christensen, O. F., 2012 Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genetics Selection Evolution 44**:** 37.

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

Cochran, W., 1951 Improvement by means of selection, pp. 449-470 in *Second Berkeley Symposium on Mathematical Statistics and Probability*.

Cockerham, C. C., 1969 Variance of gene frequencies. Evolution 23**:** 72-84.

Cole, J., P. VanRaden, J. O'Connell, C. Van Tassell, T. Sonstegard *et al.*, 2009 Distribution and location of genetic effects for dairy traits. Journal of Dairy Science 92**:** 2931-2946.

Colombani, C., A. Legarra, S. Fritz, F. Guillaume, P. Croiseau *et al.*, 2012 Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds. Journal of Dairy science.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler and M. P. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193**:** 327-345.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182**:** 375-385.

Eding, H., and T. Meuwissen, 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118**:** 141-159.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*. Longman New York.

Fernando, R., and D. Gianola, 1986 Optimal properties of the conditional mean as a selection criterion. Theoretical and Applied Genetics 72**:** 822-825.

Fernando, R., and M. Grossman, 1989 Marker assisted selection using best linear unbiased prediction. Genetics, Selection, Evolution: GSE 21**:** 467.

Fernando, R. L., D. Habier, C. Stricker, J. C. M. Dekkers and L. R. Totir, 2007 Genomic selection. Acta Agriculturae Scandinavica, A 57**:** 192-195.

Forni, S., I. Aguilar and I. Misztal, 2011 Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43**:** 1.

Garcia-Cortes, L. A., A. Legarra, C. Chevalet and M. A. Toro, 2013 Variance and Covariance of Actual Relationships between Relatives at One Locus. PLoS ONE 8**:** e57003.

Garrick, D. J., J. F. Taylor and R. L. Fernando, 2009 Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 41**:** 44.

Gengler, N., S. Abras, C. Verkenne, S. Vanderick, M. Szydlowski *et al.*, 2008 Accuracy of prediction of gene content in large animal populations and its use for candidate gene detection and genetic evaluation. Journal of Dairy Science 91**:** 1652-1659.

Gengler, N., P. Mayeres and M. Szydlowski, 2007 A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. animal 1**:** 21-28.

George, E. I., and R. E. McCulloch, 1993 Variable selection via Gibbs sampling. Journal of the American Statistical Association 88**:** 881-889.
Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi and R. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. Genetics 183**:** 347-363.
Gianola, D., and R. L. Fernando, 1986 Bayesian Methods in Animal Breeding Theory. Journal of Animal Science 63**:** 217.
Goffinet, B., and J. Elsen, 1984 Critere optimal de selection: quelques resultats generaux. G{\'e}n{\'e}tique s{\'e}lection {\'e}volution 16**:** 307-318.
Habier, D., R. L. Fernando and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177**:** 2389-2397.
Habier, D., R. L. Fernando, K. Kizilkaya and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12**:** 186.
Harris, B. L., and D. L. Johnson, 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J Dairy Sci 93**:** 1243-1252.
Harville, D., 1976 Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. The Annals of Statistics 4**:** 384-395.
Hayes, B. J., P. M. Visscher and M. E. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91**:** 47-60.
Henderson, C., 1973 Sire evaluation and genetic trends, pp. 10-41.
Henderson, C., 1978 Undesirable properties of regressed least squares prediction of breeding values. Journal of Dairy Science 61**:** 114-120.
Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph.
Hill, W., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38**:** 226-231.
Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res (Camb)**:** 1-18.
Jensen, J., G. Su and P. Madsen, 2012 Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC genetics 13**:** 44.
Kass, R. E., and A. E. Raftery, 1995 Bayes factors. Journal of the American Statistical Association 90**:** 773-795.
Kennedy, B., M. Quinton and J. Van Arendonk, 1992 Estimation of effects of single genes on quantitative traits. Journal of Animal Science 70**:** 2000-2012.
Lande, R., and R. Thompson, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124**:** 743-756.
Legarra, A., and I. Misztal, 2008 Technical note: Computing strategies in genome-wide selection. J Dairy Sci 91**:** 360-366.
Legarra, A., A. Ricardi and O. Filangi, 2011a GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and BayesCp), pp.
Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume and S. Fritz, 2011b Improved Lasso for genomic selection. Genet Res (Camb) 93**:** 77-87.
Legarra, A., C. Robert-Granié, E. Manfredi and J.-M. Elsen, 2008 Performance of genomic selection in mice. Genetics 180**:** 611-618.
Li, C. C., and D. G. Horvitz, 1953 Some methods of estimating the inbreeding coefficient. Am J Hum Genet 5**:** 107-117.
Luan, T., J. Woolliams, J. Odegard, M. Dolezal, S. Roman-Ponce *et al.*, 2012 The importance of identity-by-state information for the accuracy of genomic selection. GENETICS SELECTION EVOLUTION 44**:** 28.

Luo, Z., 1998 Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity 80**:** 198-208.
Lynch, M., 1988 Estimation of relatedness by DNA fingerprinting. Mol Biol Evol 5**:** 584-599.
Meuwissen, T., and M. Goddard, 2010 The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185**:** 1441-1449.
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.
Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet *et al.*, BLUPF90 and related programs (BGF90), pp.
Nejati-Javaremi, A., C. Smith and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. J Anim Sci 75**:** 1738-1745.
Park, T., and G. Casella, 2008 The Bayesian Lasso. Journal of the American Statistical Association 103**:** 681-686.
Pérez, P., G. de Los Campos, J. Crossa and D. Gianola, 2010 Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. The Plant Genome 3**:** 106-116.
Powell, J. E., P. M. Visscher and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet 11**:** 800-805.
Quaas, R. L., 1976 Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32**:** 949-953.
Ricard, A., S. Danvy and A. Legarra, 2013 Computation of deregressed proofs for genomic selection when own phenotypes exist with an application in French show-jumping horses. Journal of Animal Science 91**:** 1076-1085.
Ritland, K., 1996 Estimators for pairwise relatedness and individual inbreeding coefficients. Genetical research 67**:** 175-185.
Rodríguez-Ramilo, S. T., L. A. García-Cortés and Ó. González-Recio, 2014 Combining Genomic and Genealogical Information in a Reproducing Kernel Hilbert Spaces Regression Model for Genome-Enabled Predictions in Dairy Cattle. PLoS ONE 9**:** e93424.
Searle, S. R., 1982 *Matrix algebra useful for statistics*. John Wiley.
Shen, X., M. Alam, F. Fikse and L. Rönnegård, 2013 A novel generalized ridge regression method for quantitative genetics. Genetics 193**:** 1255-1268.
Sillanp\\"a, a., , MJ, 2011 On statistical methods for estimating heritability in wild populations. Molecular Ecology 20**:** 1324-1332.
Sorensen, D., and D. Gianola, 2002 *Likelihood, bayesian and MCMC methods in quantitative genetics*. Springer.
Strandén, I., and O. F. Christensen, 2011 Allele coding in genomic evaluation. Genet Sel Evol 43**:** 25.
Strandén, I., and D. J. Garrick, 2009 Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci 92**:** 2971-2975.
Su, G., P. Madsen, U. S. Nielsen, E. A. Mäntysaari, G. P. Aamand *et al.*, 2012 Genomic prediction for Nordic Red Cattle using one-step and selection index blending. Journal of Dairy science 95**:** 909-917.
Sun, X., L. Qu, D. J. Garrick, J. C. Dekkers and R. L. Fernando, 2012 A Fast EM Algorithm for BayesA-Like Prediction of Genomic Breeding Values. PLoS ONE 7**:** e49157.
Sved, J., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical population biology 2**:** 125-141.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Research 17**:** 520-526.
Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58**:** 267-288.
Toro, M. Á., L. A. García-Cortés and A. Legarra, 2011 A note on the rationale for estimating genealogical coancestry from molecular markers. Genet Sel Evol 43**:** 27.
VanRaden, P., 2007 Genomic measures of relationship and inbreeding. Interbull Bulletin**:** 33.
VanRaden, P., and G. Wiggans, 1991 Derivation, calculation, and use of national animal model information. Journal of Dairy Science 74**:** 2737-2746.
VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91**:** 4414-4423.
Varona, L., 2010 Understanding the use of Bayes factor for testing candidate genes. Journal of Animal Breeding and Genetics 127**:** 16-25.
Varona, L., L. A. García-Cortés and M. Pérez-Enciso, 2001 Bayes factors for detection of quantitative trait loci. Genet Sel Evol 33**:** 133-152.
Verbyla, K. L., B. J. Hayes, P. J. Bowman and M. E. Goddard, 2009 Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet Res 91**:** 307-311.
Vidal, O., J. Noguera, M. Amills, L. Varona, M. Gil *et al.*, 2005 Identification of carcass and meat quality quantitative trait loci in a Landrace pig population selected for growth and leanness. Journal of Animal Science 83**:** 293-300.
Villanueva, B., J. Fernández, L. García-Cortés, L. Varona, H. Daetwyler *et al.*, 2011 Accuracy of genome-wide evaluation for disease resistance in aquaculture breeding programs. Journal of Animal Science 89**:** 3433-3442.
Vitezica, Z., I. Aguilar, I. Misztal and A. Legarra, 2011 Bias in genomic predictions for populations under selection. Genetics Research 93**:** 357-366.
Wakefield, J., 2009 Bayes factors for genome-wide association studies: comparison with P-values. Genetic Epidemiology 33**:** 79-86.
Wright, S., 1922 Coefficients of inbreeding and relationship. Am. Nat. 56**:** 330-338.
Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42**:** 565-569.
Zhang, Z., J. Liu, X. Ding, P. Bijma, D.-J. de Koning *et al.*, 2010 Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS ONE 5**:** e12648.

11. Appendix B: The Single Step

Livestock Science ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Livestock Science

journal homepage: www.elsevier.com/locate/livsci

Single Step, a general approach for genomic selection

Andres Legarra [a,*], Ole F. Christensen [b], Ignacio Aguilar [c], Ignacy Misztal [d]

[a] INRA, UMR1388 GenPhySE, BP52627, 31326 Castanet Tolosan, France
[b] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Blichers Alle 20, P.O. BOX 50, DK-8830 Tjele, Denmark
[c] Instituto Nacional de Investigación Agropecuaria, Las Brujas 90200, Uruguay
[d] Department of Animal and Dairy Science, University of Georgia, Athens 30602-2771, USA

ARTICLE INFO

Keywords:
Genetic evaluation
Genomic evaluation
Marker genotypes
BLUP
Relationship

ABSTRACT

Genomic evaluation methods assume that the reference population is genotyped and phenotyped. This is most often false and the generation of pseudo-phenotypes is uncertain and inaccurate. However, markers obey transmission rules and therefore the covariances of marker genotypes across individuals can be modelled using pedigree relationships. Based on this, an extension of the genomic relationship matrix can be constructed in which genomic relationships are propagated to all individuals, resulting in a combined relationship matrix, which can be used in a BLUP procedure called the Single Step Genomic BLUP. This procedure provides so far the most comprehensive option for genomic evaluation. Several extensions, options and details are described: compatibility of genomic and pedigree relationships, Bayesian regressions, multiple trait models, computational aspects, etc. Many details scattered through a series of papers are put together into this paper.

1. Introduction: brief excursion into methods for genomic evaluation

1.1. Marker information

Genetic progress by selection and mating is based on prediction of the ability of the parents to breed the most efficient descendants. This process of prediction is called genetic evaluation or prediction. Genetic evaluation in plants and livestock has, for the last century, been based on the use of phenotypes at the traits of interest, together with pedigree. In most cases, these evaluations ignore the physical base of heredity, i.e., DNA, and use a simplified conceptual representation of the transmission of genetic information from parents to offspring; namely, each parent passes on average half its genetic constitution, associated with an unknown sampling known as Mendelian sampling. Recent technical developments allow stepping further into biology and peeking at the genome in the form of single nucleotide polymorphisms, known as SNP markers. These markers depict, in an incomplete manner, the differences between DNA inherited by two individuals. They can be used in multiple ways; in this section we will present very briefly how they are typically used in genetic evaluation (or prediction or estimation of breeding values: EBV hereinafter) in a parametric framework. Most genomic evaluations follow the principle of estimating the *conditional expectation* of the breeding value in view of all information, which has optimal properties if the assumptions of the model hold (e.g., Fernando and Gianola, 1986). This (parametric) paradigm has been extremely fruitful over the last decades, allowing for the development of BLUP, REML, Bayesian estimators and giving a coherent framework to solve many applied problems in animal breeding (e.g., Gianola and Fernando, 1986).

The notion of prediction or estimation of random effects is absent in many statistical textbooks (but check, for instance, Casella and Berger (1990)). However, it has been treated as

* Corresponding author. Tel.: +33 561285182; fax: +33 561285353.
E-mail addresses: andres.legarra@toulouse.inra.fr (A. Legarra), OleF.Christensen@agrsci.dk (O.F. Christensen), iaguilar@inia.org.uy (I. Aguilar), ignacy@uga.edu (I. Misztal).

http://dx.doi.org/10.1016/j.livsci.2014.04.029
1871-1413/

early as Smith (1936) with key references e.g. in Cochran (1951), Henderson (1973) or Fernando and Gianola (1986). Based on those authors, the "correct" model of prediction consists in writing down the statistical association between phenotypes and breeding values, then derive the EBVs from the conditional distribution of breeding values given the phenotypes.

1.2. Bayesian regression

Typically, in genomic predictions, the phenotypes of a population are considered as a function of the breeding values, and the breeding value of individuals, $\boldsymbol{u}$ (or part of it) is decomposed into a sum of marker effects $\boldsymbol{a}$ (e.g., Meuwissen et al., 2001; VanRaden, 2008). These marker effects are summed according to the genotype of the individual, coded as (0,1,2) for the (*AA*,*Aa*,*aa*) genotypes. In matrix notation $\boldsymbol{u}=\boldsymbol{Ma}$. It follows that one way of estimating breeding values is to estimate marker effects and then use $\hat{\boldsymbol{u}}=\boldsymbol{M}\hat{\boldsymbol{a}}$. In order to estimate marker effects, one needs to assume a prior distribution for them. The process of estimation of marker effects using the statistical model for phenotypes $p(\boldsymbol{y}|\boldsymbol{a})$ and the prior for markers $p(\boldsymbol{a})$ is often called *Bayesian Regression on markers*. A difficult decision is the choice of the prior for markers. An extensive literature in the subject shows higher accuracy, for some traits and populations, of using "heavy-tailed" a priori distributions (e.g., VanRaden et al., 2009).

1.3. RR-BLUP or GBLUP

If multivariate normality is assumed for the effect of markers, interesting things happen in the algebraic developments. The first one is that the Bayesian Regression becomes what is called RR-BLUP (or SNP-BLUP). The second is the existence of closed forms for the RR-BLUP estimators of marker effects, in the form of Henderson's Mixed Model Equations; these estimators greatly simplify computations and can be easily extended, e.g. for multiple trait situations. The third is the existence of a so-called equivalent model, in which breeding values (and not marker effects) are directly computed by Henderson's Mixed Model Equations using a covariance matrix $Var(\boldsymbol{u})=\boldsymbol{ZD}_a\boldsymbol{Z}'$ (VanRaden, 2008), where $\boldsymbol{Z}=\boldsymbol{M}-2\boldsymbol{P}$ and $\boldsymbol{P}$ contains p_k, the allelic frequencies of markers. This is most often called GBLUP. In the most common case it is assumed that $Var(\boldsymbol{a})=\boldsymbol{D}_a=\boldsymbol{I}\sigma^2_u/2\Sigma p_k q_k$, where σ^2_u is the genetic variance, so that that $Var(\boldsymbol{u})=\sigma^2_u\boldsymbol{G}$, where $\boldsymbol{G}=\boldsymbol{ZZ}'/2\Sigma p_k q_k$. The matrix $\boldsymbol{G}$ is called the *genomic relationship matrix* and will frequently be referred to later. Properties of $\boldsymbol{G}$ for populations in Hardy-Weinberg equilibrium are an average diagonal of 1 and an average off-diagonal of 0. Genomic evaluation using $\boldsymbol{G}$ (GBLUP) gives the same estimated breeding values as a marker-based RR-BLUP and has the additional advantage of fitting very well into ancient developments (e.g., for multiple trait) and current software. An interesting feature of the genomic relationship matrix is that it can be seen as an "improved" estimator of relationships based on markers instead of pedigrees (VanRaden, 2008; Hayes et al., 2009), and is closely related to estimators of relationships based on markers used in conservation genetics (Ritland, 1996; Toro et al., 2011).

2. The problem of missing genotypes and the use of pseudo-data

Genotyping an individual is an expensive process that also requires the availability of a biological sample. Therefore, in most populations either the most recent or the most representative animals (e.g., sires in dairy cattle) have been genotyped. Some individuals are genotyped with low-density chips that genotype only some markers. From these, genotypes at all markers can be efficiently imputed (e.g., VanRaden et al., 2013) and we will consider these individuals as genotyped. A non-genotyped individual is one for which *there is no genotype at any loci*. Therefore, the methods for genomic prediction described above cannot be applied directly, as there is often not phenotype for the individual genotyped and viceversa; this is particularly true for sex-limited traits (milk yield, fertility, prolificacy). Although a sire model could be used, this ignores selection on the female side, and does not yield females' EBVs. Therefore, animal breeders have used pseudo-data or *pseudo-phenotypes*. A pseudo-phenotype is a projection of the phenotypes of individuals close to the genotyped one. In dairy cattle and sheep, pseudo-phenotypes typically used are corrected daughter performances (daughter yield deviations, VanRaden and Wiggans, 1991), whereas in other species de-regressed proofs are often used, with a variety of *ad hoc* adjustments (Garrick et al., 2009; Ricard et al., 2013).

This process is therefore clumsy and we call it *multiple step*. A regular genetic evaluation based on pedigree is run first, and its results are used to create pseudo-performances. Then, a genomic evaluation model is used. This results in losses of information, inaccuracies and biases, whose importance depends on the species and data set. There are several possible problems:

1. The information of a close relative is ignored in the genomic prediction, for instance the dam of a bull if this dam has phenotype but not genotype.
2. The information of a close relative is ignored in the creation of pseudo-phenotypes, for instance a non-genotyped parent. This is serious if the progeny of the genotyped individual is scarce and therefore parental phenotypes are informative (see Ricard et al. (2013) for a discussion in a horse application).
3. Unless estimates of environmental effects are perfect, covariances among pseudo-phenotypes are not correctly modelled. For instance, the yield deviations of two unrelated cows in the same herd will be correlated (e.g., if the herd effect is underestimated both will be biased upwards). This is ignored in the genomic model, which acts as if pseudo-phenotypes were perfectly clean of environmental errors.
4. Many key parameters are difficult to obtain. One of them is precisions of pseudo-phenotypes, which are in most cases rough approximations.
5. There is no feedback. An improved estimation of the breeding value of the genotyped animal should go into the regular pedigree-based genetic evaluation and improve its global accuracy.

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

6. When genomic selection is applied, animals are selected as parents based on their known genotype. The implication is that when phenotypes are obtained from a scheme that has used genomic selection, evaluation based on pedigree becomes biased and is no longer appropriate (Patry and Ducrocq, 2011). Hence, current approaches for constructing pseudo-phenotypes will also become inappropriate due to problems of bias.
7. The process is extremely difficult to generalize. For instance, the multiple-trait generalization of pseudo-phenotypes is basically non-existent, and the pseudo-phenotypes for maternal traits result in much less accurate multiple step predictions (Lourenco et al., 2013).

Some of these defaults can be palliated. VanRaden et al. (2009) used a selection index to *a posteriori* add information from non-genotyped dams to bull genomic evaluations. The procedures of creation of pseudo-phenotypes can be refined over and over, and in dairy cattle they result in very accurate predictions, as accurate as Single Step (Aguilar et al., 2010). In other species the adequacy of multiple step procedures varies more. However, the existence of these problems calls for a unified procedure for prediction of genetic value. This paper will describe such a procedure: the *Single Step*.

3. Development of the Single Step method for genomic evaluation

Legarra et al. (2009) and Christensen and Lund (2010) developed in parallel the basic theory for the Single Step. They started from two somehow different points of view that turned out to result in the same formulation, and we will present both developments, starting with the latter one.

3.1. The Single Step as "imputing" missing genotypes

To some extent, missing genotypes can be deduced from existing genotypes, for instance a dam mated to a sire *AA* producing an offspring *Aa* is necessarily carrier of one allele *a*. In statistical theory, a way to deal with missing information is to augment the model with this missing information (*e.g.*, Tanner and Wong, 1987). This missing information needs to be inferred from the other data, and its joint distribution needs to be considered. This means that a "best guess" of missing information in view of observed data, as suggested by Hickey et al. (2012), who imputed genotypes for the complete nongenotyped population, is not correct enough. Even if one considers the uncertainty of individual "guesses" the across-individual uncertainty is extremely difficult to ascertain or deal with.

An example may clarify this point. Assume a very long complex pedigree and the final generation genotyped for one locus, with allelic frequency $p=frequency(a)$. Due to only having one generation with genotypes and to the long and complex pedigree, best guesses of genotypes in the base animals will be nearly identical and equal to $2p$, for all individuals. Therefore, using "best guess" of genotype without taking uncertainty into account, all base population individuals will be treated by the genomic evaluation as identical, which will force them to have the same estimated breeding value, which is paradoxical. For each individual the uncertainty can be assessed by noting that the distribution of genotypes in this case is approximately *AA* (with probability q^2), *Aa* (with probability $2pq$) and *aa* (with probability p^2), but the joint distribution of genotypes for individuals in the base population is much more difficult to characterize. In principle, incorporation of uncertainty can be done by sampling all possible genotypic configurations of all individuals, e.g. by a Gibbs sampling procedure (e.g. Abraham et al., 2007) but this is computationally infeasible for data of the size used in practical genetic evaluations.

Christensen and Lund (2010), considered the problem as follows. Their objective was to create an extension of the genomic relationship matrix to nongenotyped animals. Following an idea of Gengler et al. (2007), they treated the genotypes as quantitative traits. This makes sense because genotypes are quantitative (0/1/2) and follow Mendelian transmissions. Therefore the covariance of the genotypes z of two individuals i and j is described by their relationship, i.e. $Cov(z_i,z_j)=A_{ij}2pq$ (e.g., Cockerham, 1969). This is less informative than considering the genotype as a union of two discrete entities following Mendelian rules (e.g., sometimes we can exactly deduce a genotype from close relatives) but makes the problem analytically tractable for all cases.

Christensen and Lund (2010) started by inferring the genomic relationship matrix for all animals using inferred (imputed) genotypes for nongenotyped animals; these can simply be obtained as $\hat{\boldsymbol{Z}}_1=\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{Z}_2$, where 1 and 2 stand for nongenotyped and genotyped animals, respectively. This provides the "best guess" of genotypes. However, the missing data theory requires the joint distribution of these "guessed" genotypes. Assuming that multivariate normality holds for genotypes (this is an approximation, but very good when many genotypes are considered), the "best guess" is $E(\boldsymbol{Z}_1|\boldsymbol{Z}_2)=\hat{\boldsymbol{Z}}_1$, and the conditional variance expressing the uncertainty about the "guess" is $Var(\hat{\boldsymbol{Z}}_1|\boldsymbol{Z}_2)=(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})\boldsymbol{V}|$ where $\boldsymbol{V}$ contains $2p_kq_k$ (where $q_k=1-p_k$) in the diagonal. These two results can be combined to obtain the desired augmented genomic relationships. For instance, for the nongenotyped animals,

$$Var(\boldsymbol{u}_1)=\sigma_u^2\left(\frac{\hat{\boldsymbol{Z}}_1\hat{\boldsymbol{Z}}'_1}{2\Sigma p_kq_k}+\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\right),$$

which equals

$$Var(\boldsymbol{u}_1)=\sigma_u^2(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}+\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})$$

Finally, the augmented covariance matrix is

$$Var\begin{pmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{pmatrix}=\sigma_u^2\boldsymbol{H},$$

where

$$\boldsymbol{H}=\begin{pmatrix}\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}+\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\\ \boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G}\end{pmatrix},$$

is the augmented genomic relationship matrix with inverse

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

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

assuming that $\boldsymbol{G}$ is invertible (this will be dealt with later). Therefore, by using an algebraic data augmentation of missing genotypes, Christensen and Lund (2010) derived a simple expression for an augmented genomic relationship matrix and its inverse, without the need to explicitly augment, or "guess", all genotypes for all non-genotyped animals.

3.2. The Single Step as Bayesian updating of the relationship matrix

Legarra et al. (2009) arrived to the same expressions that of Christensen and Lund (2010) in a different manner. They also considered how to construct an extended relationship matrix. However, instead of dealing with individual markers, they dealt with overall breeding values that can be written as $\boldsymbol{u}_2=\boldsymbol{Z}_2\boldsymbol{a}$. They reasoned as follows. Prior to observation of markers, the joint distribution of breeding values is multivariate normal

$$p\begin{pmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{pmatrix}=N(0,\sigma_u^2\boldsymbol{A})$$

with covariance matrix

$$Var\begin{pmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{pmatrix}=\sigma_u^2\boldsymbol{A}=\sigma_u^2\begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{pmatrix}$$

After observing the markers, this covariance matrix will change. The joint distribution above can be split into the product of a marginal and a conditional density; i.e. $p(\boldsymbol{u}_1,\boldsymbol{u}_2)=p(\boldsymbol{u}_1|\boldsymbol{u}_2)p(\boldsymbol{u}_2)$, where

$$p(\boldsymbol{u}_1|\boldsymbol{u}_2)=N(\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{u}_2,\sigma_u^2(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})).$$

In other terms, $\boldsymbol{u}_1=\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{u}_2+\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ and $\boldsymbol{u}_2$ are independent, and $Var(\boldsymbol{\epsilon})=\sigma_u^2(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})$.

As discussed before, in the presence of marker genotypes the genomic relationship matrix can be considered as fully informative about relationships of individuals, without the need to resort to pedigree or knowledge of previous, or future, nongenotyped individuals. Therefore, *after* observing the marker genotypes

$$p(\boldsymbol{u}_2|markers)=N(0,\ \sigma_u^2\boldsymbol{G}).$$

Marker genotypes influence the relationships among nongenotyped individuals and relationships between nongenotyped and genotyped individuals indirectly. Assuming that these relationships are only influenced by marker genotypes through the genomic relationships among genotyped individuals, and assuming that the statistical distribution is determined by these relationships, one can write that

$$p(\boldsymbol{u}_1|\boldsymbol{u}_2,markers)=p(\boldsymbol{u}_1|\boldsymbol{u}_2)$$

Therefore, the joint distribution of breeding values *after* observing the markers is:

$$p(\boldsymbol{u}_1,\boldsymbol{u}_2|markers)=p(\boldsymbol{u}_1|\boldsymbol{u}_2)p(\boldsymbol{u}_2|markers)$$

From these results, expressions for the covariance of breeding values are immediate. For instance, $Var(\boldsymbol{u}_1)=\sigma_u^2(\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}+\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})$ where the part involving $\boldsymbol{G}$ is the variability associated to the conditional mean of breeding values of nongenotyped individuals given the genotyped ones; and the second part is the variability beyond this conditional mean. Finally, the result

$$Var\begin{pmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{pmatrix}=\sigma_u^2\boldsymbol{H}$$
$$=\sigma_u^2\begin{pmatrix}\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}+\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\\ \boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G}\end{pmatrix}$$

is obtained, in full agreement with Christensen and Lund (2010). The reason for this agreement is that in both cases a central assumption is that the influence of marker genotypes on nongenotyped individuals is via relationships determined by the numerator relationship matrix $\boldsymbol{A}$.

3.3. Genetic properties of the extended relationship matrix

Matrix $\boldsymbol{H}$ above can be seen as a modification of regular pedigree relationships to accommodate genomic relationships. For instance, two seemingly unrelated individuals will appear as related in $\boldsymbol{H}$ if their descendants are related in $\boldsymbol{G}$. Accordingly, two descendants of individuals that are related in $\boldsymbol{G}$ will be related in $\boldsymbol{H}$, even if the pedigree disagrees. Indeed, it has been suggested (Sun and Van Raden, 2013) to use $\boldsymbol{H}$ in mating programs to avoid inbreeding.

Contrary to common intuition from BLUP or GBLUP, genotyped animals without phenotype or descendants *cannot* be eliminated from matrix $\boldsymbol{H}$. The reason is that (unless both parents are genotyped) these animals potentially modify pedigree relationship across other animals, notably their parents. For instance imagine two half-sibs, offspring of one sire mated to two nongenotyped, unrelated cows. If these two half sibs are virtually identical, $\boldsymbol{H}$ will include this information and the cows will be made related (even identical) in $\boldsymbol{H}$.

3.4. Single Step genomic BLUP

Because the Single Step relationship matrix provides an explicit and rather sparse inverse of the extended relationship matrix $\boldsymbol{H}$, its application to genomic evaluation is immediate. A full specification of the Single Step Genomic BLUP assumes the following model:

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{b}+\boldsymbol{W}\boldsymbol{u}+\boldsymbol{e}$$

$$Var(\boldsymbol{u})=\boldsymbol{H}\sigma_u^2;\quad Var(\boldsymbol{e})=\boldsymbol{I}\sigma_e^2$$

with $\boldsymbol{H}$ and its inverse as shown above. The logic of BLUP (Henderson, 1973 and many other publications) holds and the only change is to use $\boldsymbol{H}$ instead of the numerator relationship matrix. Genomic predictions estimating simultaneously all breeding values and using all available information are, for the single trait case, the solutions to the mixed model equations (e.g., Aguilar et al., 2010; Christensen and Lund, 2010):

$$\begin{pmatrix}\boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}\\ \boldsymbol{W}'\boldsymbol{X} & \boldsymbol{W}'\boldsymbol{W}+\boldsymbol{H}^{-1}\lambda\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{b}}\\ \hat{\boldsymbol{u}}\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}'\boldsymbol{y}\\ \boldsymbol{W}'\boldsymbol{y}\end{pmatrix}$$

where $\lambda=\sigma_e^2/\sigma_u^2$.

Note that any formulation using relationship matrix $\boldsymbol{A}$ can use $\boldsymbol{H}$ instead, and therefore there is also Single Step

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

REML and Single Step Gibbs, for instance in Legarra et al. (2011a) and Forni et al. (2011).

4. Extensions and refinements of the Single Step

As said above, any model that has been fit as BLUP can be fit as Single Step. We will describe a few of these extensions that are of interest.

4.1. Pseudo-Single Step

Also called "blending" (e.g. Su et al., 2012a), this has been used to include all males of a population with pseudo-phenotypes, where some are genotyped and some are not. This is a compromise between using all information (which might be complex) and ignoring pseudo-phenotypes of non-genotyped males, for instance sires of genotyped males. Accuracy increases, but less than with true Single Step (Baloche et al., 2014).

4.2. Multiple trait

Extension to deal with multiple traits is immediate. The mixed model equations are in the usual notation:

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W}+\boldsymbol{H}^{-1}\otimes\boldsymbol{G}_0 \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{u}} \end{pmatrix}=\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$$

where $\boldsymbol{R}=\boldsymbol{I}\otimes\boldsymbol{R}_0$, $\boldsymbol{R}_0$ is the matrix of residual covariances across traits and $\boldsymbol{G}_0$ is the matrix of genetic covariances across traits. Extension to random regressions or maternal effect models is very similar.

4.3. Marker effect estimates

The GBLUP and other models based on genomic relationship matrices such as the Single Step do not directly provide estimates of marker effects. These are of interest in order to spot locations of major genes (or QTL) and also in order to provide a less computationally demanding evaluation of new born animals that are genotyped but do not have phenotypes. The marker effects can be deduced from estimated breeding values of the genotyped individuals. Consider the joint distribution of breeding values $\boldsymbol{u}$ and marker effects $\boldsymbol{a}$ (Henderson, 1973; Strandén and Garrick, 2009):

$$Var\begin{pmatrix} \boldsymbol{u}_2 \\ \boldsymbol{a} \end{pmatrix}=\begin{pmatrix} \boldsymbol{Z}_2\boldsymbol{D}_a\boldsymbol{Z}_2' & \boldsymbol{Z}_2\boldsymbol{D}_a \\ \boldsymbol{D}_a\boldsymbol{Z}_2' & \boldsymbol{D}_a \end{pmatrix}$$

where, usually, $\boldsymbol{D}_a=\boldsymbol{I}\sigma_u^2/2\Sigma p_i q_i=$ (this assumption will be relaxed later). Assuming multivariate normality, $\hat{\boldsymbol{u}}_2|\hat{\boldsymbol{a}}=\boldsymbol{Z}_2\hat{\boldsymbol{a}}$ (the breeding value is the sum of marker effects) and

$$\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}_2=\boldsymbol{D}_a\boldsymbol{Z}_2'(\boldsymbol{Z}_2\boldsymbol{D}_a\boldsymbol{Z}_2')^{-1}\hat{\boldsymbol{u}}_2=\boldsymbol{D}_a\boldsymbol{Z}_2'\boldsymbol{G}^{-1}\sigma_u^{-2}\hat{\boldsymbol{u}}_2$$

where (as discussed in previous sections) $\boldsymbol{Z}_2\boldsymbol{D}_a\boldsymbol{Z}_2'=\boldsymbol{G}\sigma_u^2$, so that marker effects can be deduced by backsolving using the genomic relationship matrix and markers' incidence matrix. This result has been used, e.g., by Wang et al. (2012), and it will appear later in this paper.

4.4. Extra polygenic effect

It has been often argued that markers do not capture all genetic variation. This can be shown by estimating variance assigned to markers and pedigree (e.g. Legarra et al., 2008) or because some genomic evaluation procedures give better cross-validation results when an extra polygenic term based exclusively on pedigree relationships is added (e.g. Su et al., 2012b). The GBLUP (VanRaden, 2008) and the derivations in the Single Step can accommodate this very easily (Aguilar et al., 2010; Christensen and Lund, 2010). Let us decompose the breeding values of genotyped individuals in a part due to markers and a residual part due to pedigree, $\boldsymbol{u}_2=\boldsymbol{u}_{m,2}+\boldsymbol{u}_{p,2}$ with respective variances $\sigma_u^2=\sigma_{u,m}^2+\sigma_{u,p}^2$. It follows that $Var(\boldsymbol{u}_2)=(\alpha\boldsymbol{G}+(1-\alpha)\boldsymbol{A}_{22})\sigma_u^2$ where $\alpha=\sigma_{u,m}^2/\sigma_u^2$. Therefore, the simplest way is to create a modified genomic relationship matrix $\boldsymbol{G}_w$ ($\boldsymbol{G}$ in Aguilar et al., 2010; $\boldsymbol{G}_w$ in VanRaden, 2008 and Christensen and Lund, 2010) as $\boldsymbol{G}_w=\alpha\boldsymbol{G}+(1-\alpha)\boldsymbol{A}_{22}$ and to plug this relationship matrix in all the expressions before. This has the additional advantage of making $\boldsymbol{G}_w$ invertible, which is not guaranteed for $\boldsymbol{G}$. Equivalently, one can fit *two* random effects, one $\boldsymbol{u}_m$with covariance matrix $\boldsymbol{H}\sigma_{u,m}^2$ and another $\boldsymbol{u}_p$ with covariance matrix $\boldsymbol{A}\sigma_{u,p}^2$.

4.5. Compatibility of genomic and pedigree relationships

This is a key issue in genomic evaluation that has received small attention beyond Single Step developers even though, as shown by Vitezica et al. (2011), it also affects multiple step methods. The derivations above of Single Step mixed model equations include terms such as $\boldsymbol{G}-\boldsymbol{A}_{22}$ and $\boldsymbol{G}^{-1}-\boldsymbol{A}_{22}^{-1}$. This suggests that $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$, the genomic and pedigree relationship matrices, need to be compatible. It has been long known (e.g., Ritland, 1996) that relationships estimated from markers need to use allelic frequencies at the base populations; otherwise a severe bias in the estimated relationships is observed (VanRaden, 2008; Toro et al., 2011). However, typically base population frequencies are unknown because pedigree recording started before biological sampling of individuals. The two derivations of the Single Step assume, either implicitly or explicitly, that the base frequencies are known. In the derivation of Christensen and Lund (2010) the allele frequencies enter explicitly. In the derivation of Legarra et al. (2009) the hypothesis is that the expected breeding value of the genotyped population is 0. This hypothesis will be wrong if either there has been selection or drift, which is commonly the case; the average breeding value will change, and the genetic variance will be reduced. These problems were soon observed by analysis of real life data sets (C.Y. Chen et al., 2011; Forni et al., 2011; Christensen et al., 2012) and verified by simulation (Vitezica et al., 2011).

Several proposals exist so far to make pedigree and genomic relationships compatible. The three first proposals "tune" matrix $\boldsymbol{G}$ to make it compatible with $\boldsymbol{A}_{22}$, in the form $\boldsymbol{G}^*=a+b\boldsymbol{G}$, where a can be understood as an "overall" relationship and b as a change in scale (or genetic variance). VanRaden (2008) suggested a regression of observed on expected relationships, minimizing the residuals of $a+b\boldsymbol{G}=\boldsymbol{A}_{22}+\boldsymbol{E}$. This reflects the fact that over conceptual repetitions

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

of our population (same pedigree but different meiosis and genotypes) $E(\boldsymbol{G})=\boldsymbol{A}_{22}$ if $\boldsymbol{G}$ is the realized relationship and $\boldsymbol{A}_{22}$ is the expected relationship (VanRaden, 2008; Hayes et al., 2009). This idea was generalized to several breed origins by Harris and Johnson (2010). The distribution of $\boldsymbol{E}$ is not homoscedastic (Hill and Weir, 2011; Garcia-Cortes et al., 2013) and this precluded scholars from trying this approach because it would be sensible to extreme values (Christensen et al., 2012), e.g., if many far relatives are included, for which the deviations in $\boldsymbol{E}$ can be very large. A second approach is to model the distribution of the mean of genotyped individuals, i.e., to assume a unknown mean μ for genotyped individuals: $p(\boldsymbol{u}_2)=N(1\mu,\boldsymbol{G})$. This is a random variable: the effect of selection or drift on the trait will vary from one conceptual repetition to another. One can equally write $p(\boldsymbol{u}_2)=N(0, \boldsymbol{G}+11'Var(\mu))$ with μ integrated out. An unbiased method forces the distribution of average values of breeding values ($\overline{\boldsymbol{u}}_2$) to be identical and therefore, the adjustment uses $\boldsymbol{G}^*=a+b\boldsymbol{G}$ with $b=1$ and $a=\overline{\boldsymbol{A}}_{22}-\overline{\boldsymbol{G}}$ where the bar implies average across values of $\boldsymbol{G}$ and $\boldsymbol{A}$. Although this models corrects the change due to genetic trend, it does not consider the fact that there is a reduction in genetic variance from the base population to the genotyped individuals considered in $\boldsymbol{A}_{22}$ but not in $\boldsymbol{G}$. This problem has been tackled twice. The first manner is to consider genotyped individuals as a subpopulation of all individuals in the population and to use Wright's fixation index theory, which allows putting relationships in any scale (Cockerham, 1969, 1973). Translated to our context (Powell et al., 2010) this implies $a=\overline{\boldsymbol{A}}_{22}-\overline{\boldsymbol{G}}$ and $b=1-a/2$ (Vitezica et al., 2011). The value of a can be understood as an overall within-population relationship within the genotyped individuals, with respect to an older population whose genotypes are not observed. This overall relationship cannot be estimated by $\boldsymbol{G}$ for lack of base allele frequencies. The value of $a/2$ can be understood as the "extra" decrease in genetic variance in a random mating population of average relationship $\overline{\boldsymbol{A}}_{22}$. Christensen et al. (2012) remarked that the hypothesis of random mating population is not likely for the group of genotyped animals, since they would be born in different years and some being descendants of others, and suggested to infer a and b jointly based on the drift of the mean of the population (as in Vitezica et al., 2011) and based on the expected genetic variance, which is encapsulated in the average inbreeding observed in $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$. More formally, the empirical variance of breeding values: $S^2_{u_2}=\boldsymbol{u}_2'\boldsymbol{u}_2/n-(\overline{\boldsymbol{u}}_2)^2$ has an expectation $((tr(\boldsymbol{A}_{22}))/(n)-\overline{\boldsymbol{A}}_{22})\sigma^2_u$ or $((tr(\boldsymbol{G}^*))/(n)-\overline{\boldsymbol{G}}^*)\sigma^2_u$ where n is the number of individuals. Forcing unbiasedness implies that a and b should be determined from the system of two equations: $a+b(tr(\boldsymbol{G}))/(n)=(tr(\boldsymbol{A}_{22}))/(n)$ and $a+b\overline{\boldsymbol{G}}=\overline{\boldsymbol{A}}_{22}$. In random mating populations in Hardy-Weinberg equilibrium (for instance in large populations of dairy cattle and sheep, where Hardy-Weinberg equilibrium approximately holds), it turns out that $b=1-a/2$ as in Vitezica et al. (2011). If restricting the group of animals for which compatibility is required to those that are born in a certain generation, the assumption of random mating among those genotyped animals is not unreasonable to assume in many livestock species. All these corrections utilize some estimate of the allelic frequencies to construct $\boldsymbol{G}$, and using observed allele frequencies (either based on all genotyped animals, or based on a subset born in a certain generation) is usually done.

Finally, Christensen (2012) suggested the opposite point of view, to "tune" $\boldsymbol{A}_{22}$ to $\boldsymbol{G}$ instead of the opposite. Pedigrees are arbitrary and depend on the start of pedigree, whereas genotypes at the markers are absolute. Allele frequencies, though, change all the time. He modelled the likelihood of markers given the pedigree as a quantitative trait and then integrated over the uncertain allele frequencies. This amounts to fix allele frequencies at 0.5 and introduce two extra parameters, γ and s. The γ parameter can be understood as the overall relationship across the base population such that current genotypes are more likely, and integrates the fact that the assumption of unrelatedness at the base population is false in view of genomic results (two animals who share alleles at markers are related even if the pedigree is not informative). More precisely, he devised a new pedigree relationship matrix, $\boldsymbol{A}(\gamma)$ whose founders have a relationship matrix $\boldsymbol{A}_{bas}=\gamma+\boldsymbol{I}(1-\gamma/2)$. Parameter s, used in $\boldsymbol{G}=\boldsymbol{ZZ}'/s$ can be understood as the counterpart of $2\Sigma pq$ (heterozygosity of the markers) in the base generation. Both parameters can be deduced from maximum likelihood. This model is the only one which introduces all the complexities of pedigrees (former ones are based on average relationships) but it has not been tested with real data so far (Christensen, 2012).

4.6. Computational algorithms

The use and development of the Single Step has been possible through the use of several state of the art algorithms. Construction and inversion of matrix $\boldsymbol{G}$ are cubic processes, and are much optimized by the use of efficient algorithms and parallel computations (Aguilar et al., 2011). Construction of matrix $\boldsymbol{A}_{22}$ has been possible, for very large pedigrees, by the algorithm of Colleau (2002) which uses Henderson's decomposition of $\boldsymbol{A}=\boldsymbol{TDT}'$ to devise a "solving" that allows easy multiplication of $\boldsymbol{w}=\boldsymbol{Av}$ and computation of $\boldsymbol{A}_{22}$ in quadratic time (Aguilar et al., 2011).

Further, the use of the solver known as preconditioned conjugated gradients (PCG) allows an easy programming to solve the Single Step mixed model equations. PCG proceeds by repeated multiplications ($\boldsymbol{LHS}$)$\boldsymbol{sol}$ where $\boldsymbol{sol}$ is the vector of unknowns. In practice, this product is split into a part

$$\begin{pmatrix} \boldsymbol{X'X} & \boldsymbol{X'W} \\ \boldsymbol{W'X} & \boldsymbol{W'W}+\boldsymbol{A}^{-1}\lambda \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{u}} \end{pmatrix}$$

for which very efficient algorithms already exist (e.g. Strandén and Lidauer, 1999) and a part

$$(\boldsymbol{G}^{-1}-\boldsymbol{A}_{22}^{-1})\lambda\,\hat{\boldsymbol{u}}_2$$

which can be done very efficiently, in particular using parallelization.

In addition, some implementations of the Single Step have used unsymmetric equations to avoid inversion of $\boldsymbol{G}$ (Misztal et al., 2009; Aguilar et al., 2013), with solution by the Bi-Conjugate Gradient Stabilized algorithm. Legarra and Ducrocq (2012) reviewed and suggested implementations of the Single Step with view towards very large data

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

sets such as in dairy cattle. Problems of these data sets are twofold. First, current evaluations use very sophisticated software, first for regular BLUP (e.g., random regressions), and later for genomic evaluations (e.g., Bayesian regressions). Secondly, the large size of the data sets, which may preclude inversion (and even construction) of $\boldsymbol{G}$. They suggested two main alternatives: a non-symmetric system of equations with non-inverted $\boldsymbol{A}_{22}$ and $\boldsymbol{G}$, and an iterative procedure similar to the multiple step but in which results from genomic evaluations would be reintroduced in the regular BLUP evaluation, and results from regular BLUP would be "data" for the genomic evaluations. The non-symmetric system shows slow convergence on large data sets (Aguilar et al., 2013), whereas the iterative method is still untested on large data sets. This is still an active field of research.

4.7. Bayesian regressions in the Single Step

Bayesian or non-linear regressions with non-normal priors for marker effects are certainly more efficient for some traits and species, with the most known example being milk contents in dairy cattle (VanRaden et al., 2009). This has inspired the search for its integration into Single Step.

Bayesian regressions can be understood as inferring the variances associated to each marker in the expression $Var(\boldsymbol{a})=\boldsymbol{D_a}$, i.e. the elements $\sigma^2_{a,k}$ in the diagonal of $\boldsymbol{D_a}$ being k-SNP specific. Zhang et al. (2010) and Legarra et al. (2011b) checked that running a full Bayesian regression to estimate breeding values, or using it to infer variances in $\boldsymbol{D_a}$ to use $\boldsymbol{G}=\boldsymbol{Z}_2\boldsymbol{D_a}\boldsymbol{Z}_2'$ in a GBLUP gave essentially the same solution. Legarra et al. (2009) suggested to use such $\boldsymbol{G}$ with precomputed variances in the Single Step procedures. Makgahlela et al. (2013) picked, using BayesB, either 750 or 1500 preselected markers to form $=\boldsymbol{Z}_2\boldsymbol{D_a}\boldsymbol{Z}_2'$, which resulted in better accuracies for milk but not for protein, and they concluded that picking the right number of markers was not obvious. No other attempt has been done so far. In a similar spirit, Wang et al. (2012) suggested to compute variances in $\boldsymbol{D_a}$ in an iterative manner within the Single Step. They obtained the marker effects from the expression $\hat{\boldsymbol{a}}|\hat{\boldsymbol{u}}_2=\boldsymbol{D_a}\boldsymbol{Z}_2'(\boldsymbol{Z}_2\boldsymbol{D_a}\boldsymbol{Z}_2')^{-1}\hat{\boldsymbol{u}}_2$, to later infer the k-th marker variance as (proportional to) $\hat{a}_k^2$ (Sun et al., 2012). Note that this estimate is severely biased (it ignores the uncertainty in the estimation of $\hat{a}_k$) and therefore an empirical correction needs to be applied, which is not the case in true Bayesian or maximum likelihood procedures (De los Campos et al., 2009; Shen et al., 2013). After computation of a new $\boldsymbol{G}$, Single Step GBLUP is rerun and markers are re-estimated, and the procedure is iterated a few times. Their simulation showed an increased accuracy of this method for traits with large QTLs.

Legarra and Ducrocq (2012) suggested two ways of dealing with Bayesian regressions. The first one was to use an equivalent set of mixed model equations including marker effects:

$$\begin{pmatrix} \boldsymbol{X}'X & \boldsymbol{X}_1'\boldsymbol{W}_1 & \boldsymbol{Z}_2'\,\boldsymbol{W}_2\boldsymbol{Z}_2 \\ \boldsymbol{W}_1'\boldsymbol{X} & \boldsymbol{W}_1'\boldsymbol{W}_1'+\boldsymbol{A}^{11}\lambda & \boldsymbol{A}^{12}\boldsymbol{Z}_2\lambda \\ \boldsymbol{Z}_2'\,\boldsymbol{W}_2\boldsymbol{X}_2 & \boldsymbol{Z}_2'\,\boldsymbol{A}^{12}\lambda & \boldsymbol{Z}_2'\,\boldsymbol{W}_2\boldsymbol{W}_2\boldsymbol{Z}_2+\boldsymbol{Z}_2'\,(\boldsymbol{A}^{22}-\boldsymbol{A}_{22}^{-1})\boldsymbol{Z}_2\lambda+\boldsymbol{D}_a^{-1}\sigma_e^2 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{u}}_1 \\ \boldsymbol{a} \end{pmatrix} = \begin{pmatrix} X'y \\ \boldsymbol{W}_1'\boldsymbol{y}_1 \\ \boldsymbol{Z}_2'\,\boldsymbol{W}_2'\boldsymbol{y}_2 \end{pmatrix}$$

In this system of equations, Bayesian Regressions are accommodated by using different *a priori* distributions for $Var(\boldsymbol{a})=\boldsymbol{D}_a$ (e.g., in Bayesian Lasso the prior distribution of elements in $\boldsymbol{D}_a$ is double exponential). This system of equations (A1) could then be solved by a Bayesian procedure such as the Gibbs sampler, which solves for $\boldsymbol{D}_a$. In the second option, an equivalent iterative procedure can iterate between solutions to regular BLUP and (Bayesian) genomic predictions; the results of one would be introduced into the other. Because this system does not infer marker variances *per se*, it does not suffer from the bias in variance estimation of Wang et al (2012). Tuning markers to be in the same scale as pedigree in the previous set of equations or in the iterative system would include an extra unknown for the parameter μ in Vitezica et al. (2011).

In addition, Fernando et al. (2013) recently presented another system of equations explicit on marker solutions. Equations include marker effects for *all* individuals, imputed following Gengler's method, and residual pedigree-based EBV for nongenotyped animals, $\boldsymbol{\epsilon}$. This $\boldsymbol{\epsilon}$ is what remains of the breeding value after we fit (imputed) SNP effects to nongenotyped individuals. Therefore total genetic value:

$$\boldsymbol{u}=(\hat{\boldsymbol{Z}}_1\,\boldsymbol{Z}_2)\boldsymbol{a}+\begin{pmatrix}\boldsymbol{\epsilon}\\0\end{pmatrix}=\hat{\boldsymbol{Z}}\boldsymbol{a}+\begin{pmatrix}\boldsymbol{\epsilon}\\0\end{pmatrix}.$$

Their final Single Step mixed model equations are

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{Z}} & \boldsymbol{X}'_1\boldsymbol{W}_1 \\ \hat{\boldsymbol{Z}}'\boldsymbol{W}'\boldsymbol{X} & \hat{\boldsymbol{Z}}'\boldsymbol{W}'\boldsymbol{W}\hat{\boldsymbol{Z}}+\mathbf{I}\frac{\sigma_e^2}{\sigma_a^2} & \hat{\boldsymbol{Z}}'_1\boldsymbol{W}'_1\boldsymbol{W}_1 \\ \boldsymbol{W}'_1\boldsymbol{X}_1 & \boldsymbol{W}'_1\boldsymbol{W}_1\hat{\boldsymbol{Z}}_1 & \boldsymbol{W}'_1\boldsymbol{W}_1+\boldsymbol{A}^{11}\frac{\sigma_e^2}{\sigma_u^2} \end{pmatrix} \times \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{\epsilon}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{y} \\ \hat{\boldsymbol{Z}}'\boldsymbol{W}'\boldsymbol{y} \\ \boldsymbol{W}'_1\boldsymbol{y} \end{pmatrix}$$

in which a Gibbs sampler can iterate to obtain Bayesian estimates. These equations are simpler than previous ones but at the cost of a very dense and large system of equations.

All these methods for Bayesian regressions in Single Step are largely untested, and only Wang et al. (2012) method is efficiently implemented and has been used in real data sets (Dikmen et al., 2013), for which no alternative currently exists.

4.8. Unknown parent groups

Missing genealogy and/or crosses are ubiquitous in animal breeding. A typical solution consists in fitting unknown parent groups, which model different means across groups of founders well identified, i.e. belonging to different generations or breeds. BLUP equations including unknown parent groups are created using an expanded inverse of the relationship matrix $\boldsymbol{A}^{-1}$ (Quaas, 1988).

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

Unfortunately, the Single Step Mixed Model equations do not accommodate this well, because of the additional matrices ($\boldsymbol{G}^{-1}-\boldsymbol{A}_{22}^{-1}$). The problem was explained in detail by Misztal et al. (2013b) who showed that proper equations would imply complex terms of the form $\boldsymbol{Q}_2'(\boldsymbol{G}^{-1}-\boldsymbol{A}_{22}^{-1})\boldsymbol{Q}_2$, implying matrix $\boldsymbol{Q}_2$ with fractions of each unknown parent group for each genotyped animal. These modifications are difficult to compute and program. Current alternatives involve ignoring the term (often with negligible results) or using the original Westell-Robinson model, which is in the form

$$\boldsymbol{y}=\boldsymbol{Xb}+\boldsymbol{Qg}+\boldsymbol{Wu}+\boldsymbol{e}$$

(Quaas, 1988) and fitting unknown parent groups $\boldsymbol{g}$ as covariates. This is satisfactory and involves no approximations, but cumbersome to implement and of slow convergence.

4.9. Accuracies

Individual accuracies can be obtained in principle from the inverse of the Single Step mixed model equations. This is impossible in practice for medium to large data sets. Therefore, Misztal et al. (2013a) suggested extending known approximations in the estimation of accuracy to the Single Step case. Modifications involve use of known approximations for the pedigree-based BLUP and add extra information from ($\boldsymbol{G}^{-1}-\boldsymbol{A}_{22}^{-1}$) to each animal; then to iterate the procedure. This procedure is accurate in dairy species, as attested by Misztal et al. (2013a) and in Manech dairy sheep (Baloche et al., unpublished) where correlations between approximate accuracies and exact accuracies from inverse of the Mixed Model Equations were found to equal 0.95 in both cases.

5. Future developments

Among important possible extensions, we will mention two: crosses and fit of dominance effects.

5.1. Crosses

Development of the Single Step has been done for purebred populations, in which heterosis is absent, genetic variance is assumed constant throughout the generations, and matings are (close to being) at random. In classical theory (e.g., Lo et al., 1997) populations involved in crossing are assumed completely unrelated; this is subject to discussion depending on the genetic architecture of the trait. For instance, Ibáñez-Escriche et al. (2009) obtained the same accuracy fitting markers with the same or different effects across breeds. Recently, Christensen et al. (2014) presented a Single Step in these lines, where the value of a crossbred animal is a sum of gametic effects, each with a different within-pure breed extended relationship matrix. On the other hand, Harris and Johnson (2010, 2013) presented an evaluation system for pure breeds and their complex crosses which considers different breed origins but roughly the same effect of markers across breeds. These aspects need to be further derived. Also, testing in real data sets is most necessary because simulations are unreliable for such complex cases. However, crossbred data sets with genomic information are scarce so far.

5.2. Dominance

Genomic predictions including dominance (e.g., Toro and Varona, 2010; Wellmann and Bennewitz, 2012) are much easier than their pedigree counterparts, which are notoriously difficult, in particular if inbreeding is involved (De Boer and Hoeschele, 1993). Dominance versions of GBLUP have been proposed (Su et al., 2012b; Vitezica et al., 2013) and real data analysis, done (Su et al., 2012b; Ertl et al., 2013; Vitezica et al., 2013). However, these methods need that genotyped animals have a phenotype, which may be precorrected. For animals that have no phenotype (i.e., dairy bulls) there are no methods to generate pseudo-phenotypes including dominance, because all methods to generate pseudo-data involve additive relationships only. For instance, computation of DYD's in dairy cattle will average to zero dominance deviations of the offspring. Therefore Single Step methods for dominance are highly relevant, yet a simple combination of pedigree-based and marker-based methods is difficult because the pedigree-based method is already difficult.

6. Obscure points and limits

6.1. Treatment of linkage

Markers are physically linked and their co-ocurrence is correlated. However, most genomic prediction models, including Bayesian Regressions and the Single Step, assume markers to be unlinked. In addition, the pedigree-based matrix $\boldsymbol{A}$ assumes loci as unlinked as well. Meuwissen et al. (2011) suggested a modified $\boldsymbol{H}$ matrix in which pedigree relationships would not be included using pedigree relationships $\boldsymbol{A}$, but using $\boldsymbol{G}_{FG}$, the Fernando and Grossman (1989) covariance matrix using pedigree and markers. The latter would be computed by means of iterative peeling, producing relationships for all individuals, genotyped or not. This procedure provides in principle a more accurate relationship matrix, and therefore should result in more accurate Single Step evaluations. However, the extent of this extra accuracy has not been evaluated in realistic simulations (e.g., with large genomes and large number of animals) or in real life data sets and it is unknown how this method scales to large pedigrees.

6.2. Convergence of solvers

The convergence rate with regular Single Step when solved by PCG iteration depends on species. The rate is similar to BLUP and poses no problem with complete pedigree and a uniform base population (e.g., chicken). The rate is also good with high-accuracy genotyped animals (dairy bulls). The rate can be poor with complex models when the pedigree contains many generations of animals without phenotypes. In such a case, restricting the pedigree to fewer old animals improves the rate. Poor convergence rate in some models is due to incompatibility

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

between $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$ when the pedigree has missing animals across generations (Misztal et al., 2013). When $\boldsymbol{G}$ is scaled for an average $\boldsymbol{A}_{22}$, elements of $\boldsymbol{A}_{22}^{-1}{}_{22}^{-1}$ due to animals with very long pedigree are larger. Solutions to this problem include modifications to $\boldsymbol{A}$ (e.g., as in Christensen, 2012), or pedigree or even phenotype truncations. Lourenco et al. (in press) investigated the effect of cutting pedigrees and phenotypes on accuracy for the youngest generation. Use of data beyond 2 generations of phenotypes and 4 generations of pedigree did not improve the accuracy while increasing computing costs.

In large data sets with many genotyped individuals (e.g., with genotyped cows) there are reports of lack of, or very slow, convergence (Harris et al., 2013; VanRaden, personal communication). This raises the question if the typical form of the mixed model equations for single-Step, including $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$ is the most appropriate, or alternative forms based on marker effects such as those presented by Legarra and Ducrocq (2012) or Fernando et al. (2013) are better numerically conditioned. No real data testing of these approaches has been shown so far. A limit to testing these approaches is the availability of very general software for BLUP. General software (multiple trait, multiple effects, etc.) does not exist for marker-based methods.

6.3. Computational limits

Computing and inverting $\boldsymbol{G}$ and and $\boldsymbol{A}_{22}$ is challenging and of cubic cost, which will eventually preclude its use for, say, $>100{,}000$ animals, and alternatives have been suggested (Legarra and Ducrocq, 2012; Fernando et al., 2013) but not thoroughly tested. These alternatives would be either highly parallelizable or use indirect representations avoiding explicit computations. However, so far, problems of convergence seem more limiting than size.

7. Current state and practical experiences

7.1. Dairy sheep

In France, the Lacaune, Manech and Basco-Bearnaise genomic evaluations use Single Step in its typical form, with corrections of $\boldsymbol{G}$ to match $\boldsymbol{A}_{22}$ and with the fit of unknown parent groups as covariates. Preliminary research did not show an added accuracy of Bayesian Regressions (Duchemin et al., 2012). Single step results in higher accuracy than GBLUP with pseudo-phenotypes (Baloche et al., 2014) and in a much simpler implementation. Single Step will be the method for genomic prediction in the future Lacaune dairy sheep genomic selection scheme.

7.2. Dairy goat

In France, the dairy goat population is testing genomic selection procedures with the Single Step as the evaluation tool (Carillier et al., 2013) although it is very soon to establish its impact.

7.3. Pigs

In Denmark, routine genetic evaluation of the three DanBred breeds Duroc, Landrace and Yorkshire has since October 2011 been made by Single-Step in its typical form, with corrections of $\boldsymbol{G}$ to match $\boldsymbol{A}_{22}$. The implementation of genomic evaluation via Single-Step was straight-forward and it has resulted in increased accuracy compared to the traditional genetic evaluation. Breeding companies PIC and ToPigs also use Single Step for genomic predictions.

7.4. Dairy cattle

National evaluations are based on multiple step procedures, but most countries are willing to change to Single Step, and many are experimenting (e.g., VanRaden, unpublished; Koivula et al., 2012; Harris et al., 2013). The reason for this change is the conceptual and practical simplicity of the Single Step, and its ability to account for genomic preselection (Patry and Ducrocq, 2011). Due to abundance of data and completeness of genotyping, tests show equivalent accuracies of Single Step and multiple step procedures (e.g., Aguilar et al., 2010). ssGBLUP was always more accurate than GBLUP for several milkability traits (Gray et al., 2012), and slightly more accurate for test-day models (Koivula et al., 2012). Also, Přibyl et al. (2013) showed higher accuracy of the Single Step for Check Republic data.

7.5. Beef cattle

There are no studies on the application of Single Step to real data sets. These data sets are more complex for genomic evaluation than other species because of missing relationships, smaller sibships, and the presence of maternal effects. Real data studies are therefore much needed. However, in a simulation study by Lourenco et al. (2013), accuracies of genomic predictions with ssGBLUP were always higher than with BLUP, which was not the case with BayesC. This was particularly true for maternal traits.

7.6. Chicken

In studies on decay of genomic prediction over generations (Wolc et al., 2011), BayesB was more accurate than single-trait GBLUP but less accurate than 2-trait GBLUP; in that study, GBLUP was applied to a reduced animal model and was equivalent to ssGBLUP. C. Chen et al. (2011) and C.Y. Chen et al. (2011b) also showed higher accuracies of Single Step than with Bayesian regressions.

8. Software

To our knowledge, the only publicly available software packages which can directly run Single Step evaluations are the BLUPF90 family of programs (Misztal et al., 2002; http://nce.ads.uga.edu/wiki) and software DMU (Madsen and Jensen, 2000, http://www.dmu.agrsci.dk/) in which it is fully implemented including regular BLUP, REML, Gibbs samplers (only BLUPF90), threshold models, generalized linear mixed models (only DMU) and iteration on data for very large data sets, and several options (most of them mentioned above).

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

Table 1
Accuracy of Single Step versus other methods in some species.

Authors	Single step	Multiple step	Pedigree BLUP	Species, trait
Aguilar et al. (2010)	0.70	0.70	0.60	Dairy cattle, final score
Baloche et al. (2014)	0.47	0.43	0.32	Milk yield, dairy sheep
C.Y. Chen et al. (2011)[a]	0.36		0.20	Breast meat, chicken
C. Chen et al. (2011)	0.37	0.09	0.28	Leg Score, chicken
Christensen et al. (2012)[a]	0.35	0.35	0.18	Daily gain, pigs
Aguilar et al. (2011)	0.39		0.26	Conception rate at first parity

[a] Predictive abilities: $r(y, \hat{u})$.

Software Mix99 (Vuori et al., 2006) has been modified to include Single Step, although these modifications are not publicly available. Public packages such as Wombat (Meyer, 2007; http://didgeridoo.une.edu.au/km/wombat.php) or ASREML (http://www.vsni.co.uk/software/asreml) can include covariance matrices computed externally, and therefore matrix $\boldsymbol{H}^{-1}$ needs to be computed with an external tool and then fit into the model.

9. Conclusion: overall benefits and drawbacks of the single Step

The Single Step provides a simple method to combine all information in a simple manner, with the additional advantage of requiring little changes to existing software. Accuracy is usually as high as, if not greater than, any other method. Some studies concerning accuracy of the Single Step have been gathered in Table 1. Beyond its extra accuracy, it has the following interesting properties:

1. Automatic accounting of all relatives of genotyped individuals and their performances.
2. Simultaneous fit of genomic information and estimates of other effects (e.g., contemporary groups). Therefore not loss of information.
3. Feedback: the extra accuracy in genotyped individuals is transmitted to all their relatives (*e.g.* Christensen et al., 2012).
4. Simple extensions. Because this is a linear BLUP-like estimator, the extension to more complicated models (multiple trait, threshold traits, test day records) is immediate. Any model fit using relationship matrices can be fit using combined relationship matrices.
5. Analytical framework. The Single Step provides an analytical framework for further developments. This is notoriously difficult with pseudo-data.

As drawbacks, one can cite the following:

1. Programming complexity to fit complicated models for marker effects (Bayesian Regressions, machine learning algorithms, etc.).
2. Lack of experience on very large data sets.
3. Long computing times with current Single Step algorithms methods, for very large data sets.
4. Lack of an easy and elegant way of considering major genes in a multiple trait setting, this is a drawback of multiple step methods as well.

Conflict of interest

Authors declare that they have no conflict of interest.

References

Abraham, K.J., Totir, L.R., Fernando, R.L., 2007. Improved techniques for sampling complex pedigrees with the Gibbs sampler. Gen. Sel. Evol. 39, 27–38.
Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93, 743–752.
Aguilar, I., Misztal, I., Legarra, A., Tsuruta, S., 2011. Efficient computations of genomic relationship matrix and other matrices used in the single-step evaluation. J. Anim. Breed Genet. 128, 422–428.
Aguilar, I., Legarra, A., Tsuruta, S., Misztal, I., 2013. Genetic evaluation using unsymmetric single step genomic methodology with large number of genotypes. Interbull Bull., 47.
Baloche, G., Legarra, A., Sallé, G., Larroque, H., Astruc, J.M., Robert-Granié, C., Barillet, F., 2014. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. J. Dairy Sci. 97, 1107–1116.
Carillier, C., Larroque, H., Palhière, I., Clément, V., Rupp, R., Robert-Granié, C., 2013. A first step toward genomic selection in the multi-breed French dairy goat population. J. Dairy Sci. 96, 7294–7305.
Casella, G., Berger, R.L., 1990. Statistical Inference. Duxbury Press Belmont, CA.
Chen, C., Misztal, I., Aguilar, I., Tsuruta, S., Aggrey, S., Wing, T., Muir, W., 2011. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: an example using broiler chickens. J. Anim Sci. 89, 23–28.
Chen, C.Y., Misztal, I., Aguilar, I., Legarra, A., Muir, W.M., 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim Sci. 89, 2673–2679.
Christensen, O.F., 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Gen. Sel. Evol. 44, 37.
Christensen, O.F., Lund, M.S., 2010. Genomic prediction when some animals are not genotyped. Gen. Sel. Evol. 42, 2.
Christensen, O., Madsen, P., Nielsen, B., Ostersen, T., Su, G., 2012. Single-step methods for genomic evaluation in pigs. Animal 6, 1565–1571.
Christensen, O.F., Madsen, P., Nielsen, B., Su, G., 2014. Genomic evaluation of both purebred and crossbred performances. Gen. Sel. Evol. 46, 23.
Cochran, W., 1951. Improvement by means of selection. In: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pp. 449–470.
Cockerham, C.C., 1969. Variance of gene frequencies. Evolution 23, 72–84.
Cockerham, C.C., 1973. Analyses of gene frequencies. Genetics 74, 679.
Colleau, J.J., 2002. An indirect approach to the extensive calculation of relationship coefficients. Gen. Sel. Evol. 34, 409–422.
De Boer, I., Hoeschele, I., 1993. Genetic evaluation methods for populations with dominance and inbreeding. Theor. Appl. Gen. 86, 245–258.
de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182, 375–385.
Dikmen, S., Cole, J.B., Null, D.J., Hansen, P.J., 2013. Genome-wide association mapping for identification of quantitative trait loci for rectal temperature during heat stress in Holstein cattle. PLoS ONE 8, e69202.
Duchemin, S., Colombani, C., Legarra, A., Baloche, G., Larroque, H., Astruc, J.-M., Barillet, F., Robert-Granié, C., Manfredi, E., 2012. Genomic

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

selection in the French Lacaune dairy sheep breed. J. Dairy Sci. 95, 2723–2733.
Ertl, J., Legarra, A., Vitezica, Z.G., Varona, L., Edel, C., Reiner, E., Gotz, K.-U., 2013. Genomic analysis of dominance effects in milk production and conformation traits of Fleckvieh cattle. Interbull Bull., 47.
Fernando, R., Gianola, D., 1986. Optimal properties of the conditional mean as a selection criterion. Theor. Appl. Gen. 72, 822–825.
Fernando, R.L., Grossman, M., 1989. Marker assisted prediction using best linear unbiased prediction. Gen. Sel. Evol. 21, 467–477.
Fernando, R.L., Garrick, D.J., Dekkers, J.C.M., 2013. Bayesian regression method for genomic analyses with incomplete genotype data. European Federation of Animal Science. Wageningen Press, Nantes, France225.
Forni, S., Aguilar, I., Misztal, I., 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Gen. Sel. Evol. 43, 1.
Garcia-Cortes, L.A., Legarra, A., Chevalet, C., Toro, M.A., 2013. Variance and covariance of actual relationships between relatives at one locus. PLoS ONE 8, e57003.
Garrick, D.J., Taylor, J.F., Fernando, R.L., 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Gen. Sel. Evol. 41, 44.
Gengler, N., Mayeres, P., Szydlowski, M., 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1, 21–28.
Gianola, D., Fernando, R.L., 1986. Bayesian methods in animal breeding theory. J. Anim. Sci. 63, 217.
Gray, K.A., Cassady, J.P., Huang, Y., Maltecca, C., 2012. Effectiveness of genomic prediction on milk flow traits in dairy cattle. Gen. Sel. Evol. 44, 24.
Harris, B.L., Johnson, D.L., 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93, 1243–1252.
Harris, B.L., Winkelman, A.M., Johnson, D.L., 2013. Impact of including a large number of female genotypes on genomic selection. Interbull Bull., 47.
Hayes, B.J., Visscher, P.M., Goddard, M.E., 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91, 47–60.
Henderson, C.R., 1973. Sire evaluation and genetic trends. In: Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush. pp. 10–41.
Hickey, J.M., Kinghorn, B.P., Tier, B., van der Werf, J.H., Cleveland, M.A., 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Gen. Sel. Evol. 44.
Hill, W.G., Weir, B.S., 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93, 47–64.
Ibáñez-Escriche, N., Fernando, R.L., Toosi, A., Dekkers, J.C.M., 2009. Genomic selection of purebreds for crossbred performance. Gen. Sel. Evol. 41, 12.
Koivula, M., Strandén, I., Poso, J., Aamand, G.P., Mäntysaari, E.A., 2012. Single step genomic evaluations for the Nordic Red Dairy cattle test day data. Interbull Bull., 46.
Legarra, A., Robert-Granié, C., Manfredi, E., Elsen, J.-M., 2008. Performance of genomic selection in mice. Genetics 180, 611–618.
Legarra, A., Aguilar, I., Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92, 4656–4663.
Legarra, A., Calenge, F., Mariani, P., Velge, P., Beaumont, C., 2011a. Use of a reduced set of single nucleotide polymorphisms for genetic evaluation of resistance to Salmonella carrier state in laying hens. Poultry Sci. 90, 731–736.
Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., Fritz, S., 2011b. Improved Lasso for genomic selection. Genet. Res. 93, 77–87.
Legarra, A., Ducrocq, V., 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. J. Dairy Sci. 95, 4629–4645.
Lo, L., Fernando, R., Grossman, M., 1997. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. J. Anim Sci. 75, 2877–2884.
Lourenco, D., Misztal, I., Wang, H., Aguilar, I., Tsuruta, S., Bertrand, J., 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., Misztal, I., Tsuruta, S., Aguilar, I., Lawlor, T.J., Forni, S., Weller, J.I., 2014. Are evaluations on young genotyped animals benefiting from the past generations? J. Dairy Sci. 10.3168/jds.2013-776, (in press).
Madsen, P., Jensen, J., 2000. A user's guide to DMU. A Package for Analysing Multivariate Mixed Models. Version 61–33.
Makgahlela, M.L., Knürr, T., Aamand, G., Stranden, I., Mäntyasaari, E., 2013. Single step evaluations using haplotype segments. Interbull Bull. 47.
Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.
Meuwissen, T., Luan, T., Woolliams, J., 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J. Anim. Breed Genet. 128, 429–439.
Meyer, K., 2007. WOMBAT—a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). J. Zhejiang Univ. Sci. B 8 (11), 815–821.
Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., and Lee, D.H. 2002. BLUPF90 and related programs (BGF90). In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August, 2002. Session 28. Institut National de la Recherche Agronomique (INRA). pp. 1–2.
Misztal, I., Legarra, A., Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92, 4648–4655.
Misztal, I., Tsuruta, S., Aguilar, I., Legarra, A., VanRaden, P., Lawlor, T., 2013a. Methods to approximate reliabilities in single-step genomic evaluation. J. Dairy Sci. 96, 647–654.
Misztal, I., Vitezica, Z., Legarra, A., Aguilar, I., Swan, A., 2013b. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed Genet. 130, 252–258.
Patry, C., Ducrocq, V., 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J. Dairy Sci. 94, 1011–1020.
Powell, J.E., Visscher, P.M., Goddard, M.E., 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11, 800–805.
Přibyl, J., Madsen, P., Bauer, J., Přibylová, J., Šimečková, M., Vostrý, L., Zavadilová, L., 2013. Contribution of domestic production records, interbull estimated breeding values, and single nucleotide polymorphism genetic markers to the single-step genomic evaluation of milk production. J. Dairy Sci. 96, 1865–1873.
Quaas, R.L., 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71, 1338–1345.
Ricard, A., Danvy, S., Legarra, A., 2013. Computation of deregressed proofs for genomic selection when own phenotypes exist with an application in French show-jumping horses. J. Anim Sci. 91, 1076–1085.
Ritland, K., 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. 67, 175–185.
Shen, X., Alam, M., Fikse, F., Ronnegård, L., 2013. A novel generalized ridge regression method for quantitative genetics. Genetics 193, 1255–1268.
Smith, H.F., 1936. A discriminant function for plant selection. Ann. Eugen. 7, 240–250.
Strandén, I., Lidauer, M., 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. J. Dairy Sci. 82, 2779–2787.
Strandén, I., Garrick, D.J., 2009. Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92, 2971–2975.
Su, G., Christensen, O.F., Ostersen, T., Henryon, M., Lund, M.S., 2012a. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS ONE 7, e45293.
Su, G., Madsen, P., Nielsen, U.S., Mäntysaari, E.A., Aamand, G.P., Christensen, O.F., Lund, M.S., 2012b. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. J. Dairy Sci. 95, 909–917.
Sun, X., Qu, L., Garrick, D.J., Dekkers, J.C., Fernando, R.L., 2012. A fast EM Algorithm for BayesA-like prediction of genomic breeding values. PLoS ONE 7, e49157.
Sun, C., Van Raden, P., 2013. Mating programs including genomic relationships. J. Dairy Sci. 96, 653.
Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. 82, 528–540.
Toro, M.Á., García-Cortés, L.A., Legarra, A., 2011. A note on the rationale for estimating genealogical coancestry from molecular markers. Gen. Sel. Evol. 43, 27.
Toro, M.A., Varona, L., 2010. A note on mate allocation for dominance handling in genomic selection. Gen. Sel. Evol. 42, 33.
VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423.
VanRaden, P., Wiggans, G., 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74, 2737–2746.

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029

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

VanRaden, P., Null, D., Sargolzaei, M., Wiggans, G., Tooker, M., Cole, J., Sonstegard, T., Connor, E., Winters, M., van Kaam, J., 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 96, 668–678.

Vitezica, Z., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for populations under selection. Genet. Res. 93, 357–366.

Vitezica, Z.G., Varona, L., Legarra, A., 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195, 1223–1230.

Vuori, K., Strandén, I., Lidauer, M., Mäntysaari, E., 2006. MiX99-effective solver for large and complex linear mixed models. In: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Minas Gerais, Brazil. 13–18 August 2006. Instituto Prociência. pp. 27–33.

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

Wellmann, R., Bennewitz, J., 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet. Res. 94, 21.

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J.E., O'Sullivan, N.P., Preisinger, R., Habier, D., Fernando, R., Garrick, D.J., Lamont, S.J., Dekkers, J.C.M., 2011. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Gen. Sel. Evol. 43, 5.

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.-J., Zhang, Q., 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS ONE 5, e12648.

Please cite this article as: Legarra, A., et al., Single Step, a general approach for genomic selection. Livestock Science (2014), http://dx.doi.org/10.1016/j.livsci.2014.04.029