Bases for Genomic Prediction

Andres Legarra Daniela A.L. Lourenco Zulma G. Vitezica

2024-05-11

Contents

1 Foreword by AL (it only engages him) 6

2 Main notation 7

3 A little bit of history 8

4 Quick look at SNP Data 11
4.1 From crude SNP file to usable genotype file . . . 11
4.2 Basic checking of marker information . . . 14
4.2.1 Call rates . . . 14
4.2.2 Allele frequencies and Minor Allele Frequencies (MAF) . . . 14
4.2.3 Hardy-Weinberg equilibrium . . . 15
4.2.4 Genotypic frequencies in crosses . . . 16
4.2.5 Sex chromosomes and unmapped markers . . . 16
4.2.6 Mendelian conflicts and assigning parents . . . 16
4.2.7 Duplicate genotypes . . . 16

5 A quick tour of Linkage Disequilibrium 17
5.1 Within-family and population linkage disequilibrium . . . 18
5.1.1 Why QTL are easier to trace within family . . . 18
5.2 Quantifying linkage disequilibrium from gametes' genotypes or from individuals' genotypes . . . 19

6 Quantitative genetics of markers, or markers as quantitative traits 21
6.1 Gene content as a quantitative trait . . . 21
6.2 Mean, variance and heritability of gene content . . . 21
6.3 Covariance of gene content across two individuals. . . . 22
6.4 Quality control using heritability of gene content . . . 23
6.5 Gengler's method to estimate missing genotypes and allelic frequencies at the base population . . . 23
6.6 Gene Content Multiple-Trait BLUP . . . 24

7 Imputation in a nutshell 25
7.1 Classical imputation . . . 25
7.1.1 Quick and dirty imputations . . . 25
7.1.2 Linear imputation . . . 25

8 Bayesian inference 27
8.1 Example of Bayesian inference . . . 27
8.2 The Gibbs sampler . . . 28
8.3 Post Gibbs analysis . . . 29

9 Models for genomic prediction 31
9.1 Simple marker model . . . 31
9.1.1 Multiallelic . . . 31
9.1.2 Biallelic . . . 32
9.2 Why markers can't be well chosen: lack of power and the Beavis effect . . . 33
9.2.1 Lack of power . . . 33
9.2.2 The Beavis (or winner's curse) effect . . . 33
9.3 Fit all markers . . . 36
9.3.1 Multiple marker regression as fixed effects . . . 36
9.4 Bayesian Estimation, or Best Prediction, of marker effects . . . 37
9.4.1 Best Predictions as a regularized estimator . . . 38
9.5 The ideal process for genomic prediction . . . 38

10 SNP effect-based methods: SNP-BLUP and Bayesian regressions 39

10.1 Allele coding. 39
10.2 Effect of prior information on marker estimates 41
10.2.1 Marker effect is fixed 42
10.2.2 Marker effect is random 42
10.3 Genetic variance explained by markers 42
10.3.1 Total genetic variance explained by markers 43
10.3.2 Genetic variance explained by markers after fitting the data 44
10.4 Prior distributions for marker effects 44
10.5 RR-BLUP or SNP-BLUP 44
10.5.1 Mixed Model equations for SNP-BLUP 45
10.5.2 How to set variance components in BLUP-SNP 46
10.5.3 Solving for marker effects 46
10.6 Estimating variances from marker models: BayesC with Pi=0 48
10.7 Transforming marker variance into genetic variance 49
10.7.1 Example with mice data 50
10.8 Differential variances for markers 50
10.8.1 REML formula for estimation of single marker variances 50
10.8.2 Bayesian estimation of marker variances 51
10.9 BayesA 51
10.10BayesB 52
10.11BayesC(Pi) 53
10.11.1Markers associated to the trait 54
10.12Bayesian Lasso 56
10.12.1Parameterization of the Bayesian Lasso 56
10.13Stochastic Search Variable Selection 57
10.14Overall recommendations for Bayesian methods 58
10.15Empirical single marker variances from marker estimates 58
10.16VanRaden's NonLinear methods 58
10.17The effect of allele coding on Bayesian Regressions 59
10.18Reliabilities from marker models 59
10.18.1Standard errors from Bayesian methods by MCMC 59
10.18.2Reliabilities 60

11 Genomic relationships **61**
11.1 Reminder about relationships 61
11.2 Identity by state and identity by descent of two individuals 61
11.2.1 Covariance between individuals 62
11.3 Relationships across individuals for a single QTL 63
11.3.1 Negative relationships 63
11.3.2 Centered relationships and IBS relationships 64
11.3.3 Inbreeding at a single QTL 64
11.4 Genomic relationships: Relationships across individuals for many markers 64
11.4.1 VanRaden's first genomic relationship matrix 64
11.4.2 VanRaden's second genomic relationship matrix 65
11.4.3 Allelic frequencies to put in genomic relationships 66
11.4.4 Properties of G 66
11.4.5 Weighted Genomic relationships 68
11.5 Genomic relationships as estimators of realized relationships 68
11.5.1 Other estimators of (genomic) relationships 70
11.5.2 Genomic inbreeding 72
11.6 Compatibility of genomic and pedigree relationships 72
11.6.1 Use of Gengler's method 73
11.6.2 Compatibility of genetic bases 73
11.6.3 Compatibility of genetic variances 73
11.6.4 Compatibility of genetic bases and variances 74
11.7 Singularity of G 75
11.8 Including residual polygenic effects in G 75

11.9 Multiallelic genomic relationships . . . 75

12 GBLUP **77**
12.1 Single trait animal model GBLUP . . . 77
12.2 Multiple trait GBLUP . . . 77
12.3 Reliabilities from GBLUP . . . 78
12.4 All Genomic relationships are equal . . . 78
12.5 GBLUP with singular G . . . 78
12.6 Backsolving from GBLUP to marker estimates . . . 79
12.7 Backsolving when matrix G has been "tuned" . . . 80
12.8 Backsolving with metafounders . . . 81
12.9 Indirect predictions using marker effects . . . 82
12.10Reliabilities of indirect predictions with "tuned G" . . . 83
12.11Reliabilities of indirect predictions with metafounders . . . 85
12.11.1Reliability with assumed allele frequencies for the reference metafounder . 86
12.12Bayesian distribution of marker effects from GBLUP . . . 86
12.13Frequentist distribution of marker effects from GBLUP. . . . 88
12.13.1Example of marker predictions from GBLUP . . . 88
12.14GREML and G-Gibbs . . . 89
12.15Complicated things in GBLUP . . . 89
12.15.1Variances of pseudo-data, DYD's, and de-regressed proofs . . . 89
12.15.2Some problems of pseudo-data . . . 90

13 Non-additive genetic effects in genomic selection **91**
13.1 Dominant genomic relationships . . . 91
13.2 Animal model GDBLUP . . . 93
13.3 Another parameterization . . . 94
13.4 Inbreeding depression . . . 95
13.5 Genomic relationship matrices in absence of HWE . . . 96
13.6 Epistatic genomic relationships . . . 97
13.7 Word of caution . . . 99

14 Single Step GBLUP **100**
14.1 SSGBLUP as improved relationships . . . 100
14.2 SSGBLUP as linear imputations . . . 101
14.3 ssGBLUP mixed model equations . . . 102
14.4 Some properties of matrix H . . . 103
14.4.1 Inbreeding in H . . . 104
14.5 Mixing G and A: blending and compatibility of pedigree and genomic relationships.104
14.5.1 Blending . . . 104
14.6 Compatibility of G and A . . . 105
14.6.1 Fitting G to A . . . 105
14.6.2 Fitting A to G . . . 106
14.6.3 Unknown Parent Groups . . . 106

15 Unknown Parent Groups and Metafounders **109**
15.1 Quick tour of Unknown Parent Groups . . . 109
15.2 Unknown Parent Groups and Single Step GBLUP . . . 110
15.3 Metafounders . . . 110
15.4 Definition of metafounders based on allele frequencies . . . 111
15.4.1 Relationships with other metrics of genomic similarity and inbreeding . . 114
15.5 Pedigree and pedigree-genomic relationships with metafounders and use in genetic evaluation and ssGBLUP . . . 115
15.6 Example with 2 metafounders . . . 115
15.7 Variance components when using metafounders . . . 117
15.8 Estimation of $\mathbf{\Gamma}$ . . . 118

16 Use of method LR to assess potential bias due to design of cross-validation analysis. 120
16.1 Introduction . . . 120
16.2 Bias due to using pre-corrected data or De-Regressed Proofs (DRP) . . . 121
16.2.1 Bias due to ignoring the effect of selection on genetic variance . . . 121
16.2.2 Bias due to pre-correction by fixed effects . . . 121
16.3 Method LR to the rescue . . . 121
16.3.1 LR in a nutshell . . . 121
16.3.2 Examples of interpretation . . . 123
16.3.3 Practicalities . . . 123
16.3.4 The importance of several comparisons . . . 124
16.3.5 Estimation of genomic accuracies vs pedigree ones . . . 124

17 References 125

18 Appendix 135

1 Foreword by AL (it only engages him)

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 http://genoweb.toulouse.inra.fr/~alegarra. I thank Guillermo Martinez-Boggio, 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.

May 13 2016. Corrected error in the Bayesian example (thanks Jesús Piedrafita).

Oct 4 2016. Added posterior variance of marker effects from GBLUP.

August 2017. Added backsolving of GBLUP to SNPBLUP when there is tuning of $\mathbf{G}$

November 2017. Slight correction on same topic

April 2018. Large additions for UGA course.

May 2022. Few typos, plus addition of Method LR, Reliabilities using SNP effects from (ss)GBLUP.

The cover is a drawing of BlackBelly Sheep in Barbados made by José Javier Legarra. Gracias Tebe !

2 Main notation

$\mathbf{X}$,$\mathbf{b}$ Incidence matrix of fixed effects and fixed effects

$\mathbf{a}$ Marker effects

$\mathbf{u}$ Polygenic or additive genetic effects

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$

$\mathbf{G}$ Genomic relationship matrix

p_i Allele frequency at marker i

$\mathbf{A}$ Pedigree-based relationship matrix

3 A little bit of history

Based on Lourenco et al., 2017, BIF Conference.

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

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

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

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

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

Initial allure of MAS led to a massive redirecting of grant funds to this type of research, greatly contributing to the current shortage of qualified quantitative geneticists in animal breeding (Misztal). Despite some of the initial setbacks using MAS, in 2001, some researchers envisioned that genomic information could still help animal breeders to generate more accurate breeding values, if a dense SNP assay that covers the entire genome became available. Extending the idea of incorporating marker information into BLUP (using genotypes, phenotypes and pedigree information), introduced by(Fernando and Grossman 1989), Meuwissen et al. (2001) proposed some methods for what is now termed genome-wide selection or genomic selection (GS). This paper used simulation data to show that accuracy of selection was doubled using genomic selection compared to using only phenotypes and pedigree information. With the promise of large accuracy gains, this paper generated enormous excitement in the scientific community. Some conclusions from this study included: 1) using SNP information can help to increase genetic gain and to reduce the generation interval; 2) the biggest advantage of genomic selection would be for traits with low heritability; 3) animals can be selected early in life prior to performance or progeny testing. With all of this potential, genomic selection was an easy sell.

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

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

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

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

As only a fraction of livestock are genotyped, Misztal et al. (Misztal *et al.* 2009) proposed a method that combines phenotypes, pedigree, and genotypes in a single evaluation. This method is called single-step genomic BLUP (ssGBLUP) and involves altering the relationships between animals based on the similarity of their genotypes. As an example, full-sibs have an average of 50% of their DNA in common, but in practice this may range from 20% to 70% (Lourenco *et al.* 2015a). The ssGBLUP has some advantages over multistep methods. It can be used with multi-trait and maternal effect models, it avoids double counting of phenotypic and pedigree information, it ensures proper weighting of all sources of information, and it can be used with both small and large populations and with any amount of genotyped animals. Overall, greater accuracies and less inflation can be expected when using ssGBLUP compared to multistep methods. Not long after the implementation of GS, single-step was first applied to a dairy population with more than 6,000 genotyped animals (Aguilar *et al.* 2010 ; Christensen and Lund 2010).

An early application of ssGBLUP in beef cattle used simulated data with 1500 genotyped animals in an evaluation for weaning weight with direct and maternal effects (Lourenco *et al.* 2013). Although a small number of genotyped animals was used, gains in accuracy were observed for both direct and maternal weaning weight. Next ssGBLUP was applied to a real breed association data set (Lourenco *et al.* 2015b). This study showed a comprehensive genomic evaluation for nearly 52,000 genotyped Angus cattle, with a considerable gain in accuracy in predicting future performance for young genotyped animals. This gain was on average 4.5 points greater than the traditional evaluations.

4 Quick look at SNP Data

The most abundant polymorphisms at the DNA level are SNPs: Single Nucleotide Polymorphisms. By the art of biochemistry and the joint effort of industry and academia, it is now possible to massively (many of them), accurately (the genotype read is the actual genotype) and economically (the cost is relatively low) read the same set of SNPs across several individuals; the technology is commonly called SNP chips or SNP genotyping. For more information, you may read, for instance, https://www.illumina.com/techniques/popular-applications/genotyping.html .

Triallelic SNPs exist in nature but they are not used for SNP chips. Thus, the possible alleles for SNP loci are all pairwise combinations among (A,C,G,T): A/C, A/G, A/T, C/G, C/T, G/T.

4.1 From crude SNP file to usable genotype file

I (A.L.) do not have much, but I have some, experience in dealing with crude SNP data. These are, for practical purposes such as national genomic evaluations, handled by experienced teams and read, stored in databases; for instance, description of such a process is in (Wiggans *et al.* 2010; Groeneveld and Lichtenberg 2016). However, it is good to be exposed to the crude output of genotyping. This is an excerpt from some real analysis; the name of this file was `..._Custom-FinalReport.txt`:

```
[Header]
GSGT Version    1.9.4
Processing Date 3/16/2012 9:11 AM
Content         OvineSNP50_B.bpm
Num SNPs        54241
Total SNPs      54241
Num Samples     36
Total Samples   36
[Data]
Sample ID       Sample Name     SNP Name        Allele1 - Top   Allele2 - Top   GC Score
ES140000270478  PLACA_CIC_12_96 250506CS3900065000002_1238.1    G       G       0.8932
ES140000270478  PLACA_CIC_12_96 250506CS3900140500001_312.1     A       G       0.7341
ES140000270478  PLACA_CIC_12_96 250506CS3900176800001_906.1     A       G       0.7532
ES140000270478  PLACA_CIC_12_96 250506CS3900211600001_1041.1    A       A       0.9674
ES140000270478  PLACA_CIC_12_96 250506CS3900218700001_1294.1    G       G       0.8178
ES140000270478  PLACA_CIC_12_96 250506CS3900283200001_442.1     C       C       0.6684
ES140000270478  PLACA_CIC_12_96 250506CS3900371000001_1255.1    G       G       0.4565
ES140000270478  PLACA_CIC_12_96 250506CS3900386000001_696.1     A       A       0.4258
ES140000270478  PLACA_CIC_12_96 250506CS3900414400001_1178.1    G       G       0.8690
ES140000270478  PLACA_CIC_12_96 250506CS3900435700001_1658.1    A       A       0.5153
ES140000270478  PLACA_CIC_12_96 250506CS3900464100001_519.1     A       G       0.8116
ES140000270478  PLACA_CIC_12_96 250506CS3900487100001_1521.1    A       G       0.7448
ES140000270478  PLACA_CIC_12_96 250506CS3900539000001_471.1     G       G       0.5248
ES140000270478  PLACA_CIC_12_96 250506CS3901012300001_913.1     A       A       0.7413
ES140000270478  PLACA_CIC_12_96 250506CS3901300500001_1084.1    G       G       0.7990
ES140000270478  PLACA_CIC_12_96 CL635241_413.1  A       A       0.8176
ES140000270478  PLACA_CIC_12_96 CL635750_128.1  A       G       0.7978
ES140000270478  PLACA_CIC_12_96 CL635944_160.1  A       G       0.7283
```

This data contains genotypes for *one* animal: `ES140000270478` for the SNPs that are listed in SNP Name. The columns Allele1 and Allele2 contain the readings in nucleotide form (Adenine, Guanine, Citosine and Thymine – A,G,C,T). For instance in SNP Name `250506CS3900065000002\_1238.1`, this animal is homozygous G/G, but for `CL635750\_128.1` . the animal is heterozygote A/G. The Allele1/ Allele2 notation is, for our purposes, arbitrary: we do not know which one came from the sire and which one came from the dam.

Now, you can see that the same animal ES140000270478 is repeated over and over; there is one line per marker. The file is constituted by a header, and one line per individual and per marker. At some point we arrive to the next animal:

```
ES140000270478 PLACA_CIC_12_96 s76040.1 G G 0.6173
ES140000270478 PLACA_CIC_12_96 s76043.1 A A 0.7965
ES150010016299 PLACA_CIC_10_02 250506CS3900065000002\_1238.1 G G 0.8932
ES150010016299 PLACA_CIC_10_02 250506CS3900140500001\_312.1 A G 0.7341
ES150010016299 PLACA_CIC_10_02 250506CS3900176800001\_906.1 A G 0.7668
```

And so on. There is thus a lot of redundancy here.

Another type of files has the `final_report` format with

```
[Header]
GSGT Version    1.9.4
Processing Date 01/01/2018 10:11 AM
Content     BovineSNP50_v2_C.bpm
Num SNPs    54609
Total SNPs  54609
Num Samples 1
Total Samples   1
[Data]
SNP Name    Sample ID   Allele1 - Forward   Allele2 - Forward   Allele1 - Top   Allele2 - Top   Allele1 - AB
    Allele2 - AB  GC Score  X   Y
ARS-BFGL-BAC-10172  USA201811   G   G   G   G   B   B   0.9506  0.012   1.036
ARS-BFGL-BAC-1020   USA201811   G   G   G   G   B   B   0.9673  0.005   0.652
ARS-BFGL-BAC-10245  USA201811   C   C   G   G   B   B   0.7579  0.092   1.417
ARS-BFGL-BAC-10345  USA201811   A   A   A   A   A   A   0.9276  1.143   0.008
ARS-BFGL-BAC-10365  USA201811   G   G   C   C   B   B   0.5335  0.004   0.862
ARS-BFGL-BAC-10375  USA201811   A   G   A   G   A   B   0.9567  0.478   0.581
ARS-BFGL-BAC-10591  USA201811   A   G   A   G   A   B   0.9003  0.386   0.473
ARS-BFGL-BAC-10867  USA201811   G   G   C   C   A   A   0.9434  0.776   0.004
ARS-BFGL-BAC-10919  USA201811   A   A   A   A   A   A   0.8526  1.232   0.036
ARS-BFGL-BAC-10951  USA201811   T   T   A   A   A   A   0.5140  0.539   0.017
ARS-BFGL-BAC-10952  USA201811   A   A   A   A   A   A   0.9512  0.987   0.030
ARS-BFGL-BAC-10960  USA201811   G   G   G   G   B   B   0.9528  0.018   0.826
ARS-BFGL-BAC-10972  USA201811   G   C   C   G   A   B   0.8759  0.917   0.743
ARS-BFGL-BAC-10975  USA201811   A   G   A   G   A   B   0.8142  0.979   0.739
ARS-BFGL-BAC-10986  USA201811   G   G   C   C   B   B   0.9309  0.055   0.731
ARS-BFGL-BAC-10993  USA201811   C   C   G   G   B   B   0.9014  0.023   1.094
ARS-BFGL-BAC-11000  USA201811   T   T   A   A   A   A   0.9686  0.561   0.013
ARS-BFGL-BAC-11003  USA201811   T   T   A   A   A   A   0.9215  1.171   0.040
ARS-BFGL-BAC-11007  USA201811   T   C   A   G   A   B   0.9454  0.884   0.675
ARS-BFGL-BAC-11025  USA201811   G   G   C   C   B   B   0.9082  0.015   0.740
ARS-BFGL-BAC-11028  USA201811   A   G   A   G   A   B   0.9678  0.182   0.288
ARS-BFGL-BAC-11034  USA201811   T   C   A   G   A   B   0.9509  0.566   0.592
ARS-BFGL-BAC-11039  USA201811   C   C   G   G   B   B   0.9658  0.000   0.889
ARS-BFGL-BAC-11042  USA201811   A   G   A   G   A   B   0.8506  0.947   0.786
ARS-BFGL-BAC-11044  USA201811   T   C   A   G   A   B   0.9654  0.726   0.689
ARS-BFGL-BAC-11047  USA201811   T   T   A   A   A   A   0.9465  0.973   0.015
```

This format is apparently more confusing but it is explained here: https://www.illumina.com/documents/products/technotes/technote_topbot.pdf . In short, what we need to look at is the A/B 's in the columns. For one marker, A and B may "mean" A and T whereas in another locus they may "mean" T and A. However, the A/B notation is less ambiguous (or more accurate) than the A/C/G/T due to the way the chemistry works. Note that the "A" in the A/C/G/T system is *not* the same as the "A" in the A/B system.

Anyway, there is a need to gather this information into a more condensed format. This format is usually comprised of:

- A *map file* with the names of all markers and (if possible) its position (chromosomes, physical location in basepairs). The names of all markers can be found in the file that we just saw, whereas the locations in the genome can be found in the web pages of consortia such as, in sheep, http://www.sheephapmap.org . Or from a close collaborator.
- The *genotype file* with the actual nucleotides typically has, in animal breeding, one line by animal and two columns per SNP marker. It looks like this (6 markers):

```
ES1400NAB40571 G G G G A A A C . . A G
ES1400NAB40573 G G G G G G A C G G A G
ES1400NAB40574 A G G G A G A C G G A A
ES1400NAB40159 G G G G A G A C G G A A
ES1400NAB40528 A G A G A G C C A G A A
ES1500VI492705 G G A G G G A C G G A G
ES1500SSA40533 A G G G A G C C G G A A
```

The . . implies that there is no lecture for this marker and individual: this is a missing genotype. SNP chips have few missing genotypes, but other technologies like GBS (Genotyping By Sequence) have very large amounts of missing genotypes. Imputation "fills in" those gaps; it will be mentioned later. Or, it could be (just making it up)

```
ES1400NAB40571 B B A A B B A B . . A A
...
```

We can compact it even further, noting that (1) SNPs are biallelic (2) the paternal/maternal origin is unknown and (3) you can represent as an integer which represents the number of copies (*gene content*: these wording will be used over and over) of a *reference* allele of the two that are polymorphic at one SNP marker. This is also known as *allele coding.* For most practical purposes, which one is the reference allele is irrelevant. For instance, assume that for ES1400NAB40571 the reference allele is B in *all* markers. So these are the integer codes:

```
ES1400NAB40571 2 0 2 1 5 0
```

Or, in a more compact way (this is what in these notes we call in these notes the "UGA format"),

```
ES1400NAB40571 202150
```

It can be seen that the correspondences are:

code	genotype
0	AA
1	AB or BA
2	BB
5	missing

The reference allele can vary across loci. For instance, consider the same animal

```
ES1400NAB40571 G G G G A A A C . . A G
```

And consider that the reference alleles for each of the 6 markers are (G,G,A,C,G,A). Using these reference alleles would give

```
ES1400NAB4057 222151
```

Which is different from the coding above. Actually, the table of correspondences would be:

Code	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6
0	AA	AA	GG	CC	AA	GG
1	AG or GA	AG or GA	AG or GA	AC or CA	AG or GA	AG or GA
2	GG	GG	AA	AA	GG	AA
5	missing	missing	missing	missing	missing	missing

Note that here we put "Marker 1", "Marker 2", etc, but actually the names are more complex, for instance,

```
250506CS3900539000001_471.
250506CS3901012300001_913.1
250506CS3901300500001_1084.1
CL635241_413.1
CL635750_128.1
CL635944_160.1
```

for this reason, it is essential to keep track of the *names* of the markers that we use.

For most purposes the coding is irrelevant, but it needs to be coherent for every batch of new animals. This is why it is mandatory to, either to stick to one of the alleles (say the B) in Illumina's A/B system, or (better) to store the whole data base with readings in the formats A/C/G/T and A/B. It is also mandatory to keep track of the *names* of the SNP markers in the files. Joining files with integers and no associated file with SNP names is dangerous.

There is software available to convert from the *long* format of Illumina output to the *compact* format. One example is illumina2preGS Alternatively, a self-written script or software can be used.

Note that the Plink format is different from the *UGA format* described here, but there are converters between formats, or you may program your own.

4.2 Basic checking of marker information

A cool feature of the integer format is that somethings are very easy to get and compute.

4.2.1 Call rates

The call rate is the number of observed genotypes:

- **per animal**: of the, say, 54K markers in the chip, 50K have been genotyped for a particular animal, the "call rate animal" is 50K/54K=93%
- **per marker**: of the, say, 900 animals genotyped for marker `CL635944_160.1`, how many have actually been successfully read? Assume that 600 have been read, then the "call rate marker" is $600/900 = 67\%$

Measuring call rate reduces to "count" the number of missing (either ". ." or "5") in the genotype file per row or per column. Animals that have low call rate (i.e. too many markers not genotyped) are eliminated. This is often due to bad conservation of the DNA. Markers that have low call rate (i.e. they have not been read for many animals) are also eliminated. Typically, this is due to poor biochemistry.

Typical thresholds for quality control of call rates are 90% or 95%. Below this level, either the marker or the individual is discarded.

4.2.2 Allele frequencies and Minor Allele Frequencies (MAF)

The allele frequency p is simply the frequency of the reference allele. For instance, consider

```
ES1400NAB40571 G G
ES1400NAB40573 G G
ES1400NAB40574 A G
ES1400NAB40159 G G
ES1400NAB40528 A G
ES1500VI492705 G G
ES1500SSA40533 A A
```

If the reference allele is G, we have 10G against 4A: $p = \frac{10}{14} \approx 0.71$, and the frequence of allele A is $q = 1 - p \approx 0.29$. The funny thing is, that this is very easy to compute looking at the UGA format:

```
ES1400NAB40571 2
ES1400NAB40573 2
ES1400NAB40574 1
ES1400NAB40159 2
ES1400NAB40528 1
ES1500VI492705 2
ES1500SSA40533 0
```

From this format, we have (quite obviously) 2 "G"s for each "2", 1 "G" for each "1" and 0 "G"s for each 0. So, quite obviously, and skipping the columns with missing information:

$$p = \frac{\text{sum of the column}}{2 \times number\ of\ lines}$$

For instance, this *awk* script computes allele frequencies:

```
#!/bin/awk -f
#
# This script computes allele frequencies from marker file with UGA format
# AA=0, aa=2, Aa=1, aA=1, no missing genotypes)
#
BEGIN{ }
```

```
{
    nsnp=length($2)
    split($2,aux,"")
        for (i=1; i<=nsnp; i++){ cnt[i]=cnt[i]+aux[i]    }
}
END {
    for(i=1; i<=nsnp; i++){print(cnt[i]/(2*(NR))) }
}
```

The Minor Allele Frequency (MAF) is the lowest of the two allele frequencies: p and $q = 1 - p$; in Fortran terms, `maf=maxval((/p,q/))` . It is used as a measure of the informativity of the marker. A marker that has $p = 1$ is said to be *monomorphic* and does not give much information, as all individuals are identical for this marker. Therefore, we may ignore it. But accordingly, we may ignore markers for which *almost all* individuals have the same genotype; for instance, if $p = 0.9999$. Where do we put the limit? A rule of thumb for genomic prediction with SNP chips is to remove markers with MAF<5%, or MAF<1%. In practice, this does not change much the results for prediction. But if the objective is to investigate rare variants, then we should not edit markers by MAF.

4.2.3 Hardy-Weinberg equilibrium

In an unselected population, the distribution of genotypes is expected to follow Hardy-Weinberg proportions. In practice, most populations are selected, so Hardy-Weinberg equilibrium (HWE) does not always hold. We can check proportions of observed and expected counts of genotypes. If n is the total number of animals and n_0, n_1, n_2 are the counts of each genotype:

Genotype	0	1	2
Observed	n_0	n_1	n_2
Expected	nq^2	$n2pq$	np^2

From this table, it is possible to make a statistical test to test the hypothesis that the data is under HWE. The statistic is

$$\chi^2 = \sum_{0:2} \frac{(\text{Observed}_i - \text{Expected}i)^2}{\text{Expected}_i}$$

where $\text{Expected}(0:2) = n\left(q^2, \text{2pq}, p^2\right)$. Another way of getting the same statistic directly from the counts (Emigh 1980):

$$\chi^2 = 16n \frac{\left(n_0 n_2 - \frac{n_1^2}{4}\right)^2}{(2n_0 + n_1)^2 (n_1 + 2n_2)^2}$$

This statistic follows a χ^2 distribution with 1 degree of freedom (Emigh 1980). For instance, in the above example with 7 animals:

$$\chi^2 = \frac{(1 - 7 \times 0.29^2)^2}{7 \times 0.29^2} + \frac{(2 - 7 \times 2 \times 0.29 \times 0.71)^2}{7 \times 2 \times 0.29 \times 0.71} + \frac{(4 - 7 \times 0.71^2)^2}{7 \times 0.71^2} = 0.63$$

Which has a non-significant p-value of 0.43 using, in R:

```
pchisq(0.63,1,0,lower.tail=FALSE)
```

You must be very careful if you use HWE statistic to do quality control. In practice, HWE *approximately* holds but it never holds *exactly*. For this reason, with large data sets, the hypothesis is rejected. In practice, a more sensible approach is to reject the marker if the number of observed heterozygotes deviates to much from the expectation. In other words, one marker *may* be rejected if

$$\left|\frac{n_1}{n} - 2pq\right| > t$$

For some threshold t. The value used by default in the BLUPF90 suite of programs is 0.15 following (Wiggans *et al.* 2009).

4.2.4 Genotypic frequencies in crosses

HWE does not hold in crosses, for instance in F1 crosses, so it should *not* be checked. We can however present what should be the genotypic frequencies in F1 crosses. If alleles frequencies in breed A and breed B are p_A and p_B, then the Expected$(0:2) = n\,(q_A q_B, p_A q_B + p_B q_A, p_A p_B)$. This may be useful to check your data.

4.2.5 Sex chromosomes and unmapped markers

The sex chromosomes (X and Y in mammals, Z and W in birds) present some complexities for genomic analysis. Females in mammals carry two alleles at sex chromosomes, but males carry two alleles only in the pseudo-autosomal part (chromosome Y and its counterpart in X). Therefore, these chromosomes are almost systematically eliminated from the analysis. Literature presents methods to deal with sex-linked inheritance, both in the classical pedigree way (Fernando and Grossman 1990) and in the genomic way (Su *et al.* 2014).

Maps (physical situation of markers in chromosomes) are typically constructed by consortia (e.g. http://bovinegenome.org , http://www.sheephapmap.org). It happens that markers may be genotyped but the exact situation is unknown and these are markers are assigned to "chromosome 0". These markers are typically discarded – they are very few and not knowing the position makes some analysis difficult.

4.2.6 Mendelian conflicts and assigning parents

Some genotypes might be incompatible with the declared pedigree. The most typical cases are (1) conflicting genotypes one parent and offspring: a father "AA" can *not* sire an offspring "aa", and (2) conflicting genotypes both parents and offspring: "AA" x "AA" cannot sire "Aa". If such an event is found looking at genotype and pedigree data, it may be a single genotyping error – however, if several of them are found for a couple or trio of individuals, either there is a problem in the pedigree or a misidentification of the sample.

One of the possibilities, if there is a conflict or the sire/dam is not in the pedigree, is to find the sire or dam of one individual based on the observed genotypes Hayes (2011). One such program is seekparentf90.

4.2.7 Duplicate genotypes

Unless there are clones or monozygotic twins, we do not expect identical genotypes – so, a very high concordance of genotypes of two animals is suspicious. In most cases, this is due to mislabeling – two DNA samples from the same animal received two different name tags.

5 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.

Consider first allelic frequencies within loci are: for the first locus,

$$p_1 = \text{freq}(A) = 0.6$$

; for the second locus,

$$p_2 = \text{freq}(B) = 0.5$$

However, to consider the joint frequency at the two loci, we need a frequency table of these diplotypes, as follows:

Table 4: Example of two loci in Linkage disequilibrium

freqs	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. In linkage equilibrium, alleles across loci are distributed at random. For instance, $\text{freq}\,(\text{AB}) = \text{freq}\,(A) \times \text{freq}\,(B) = 0.30$. If these were the case, the table should be as follows:

Table 5: 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" (or segments) 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 F1, F2…Fn generations. At the end, the chromosomes become a fine-grained mosaic of grey and black. However, complete mixture is difficult to attain.

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 locus is a proxy for the other one, and this is why association analysis implicitly uses linkage disequilibrium to map genes.

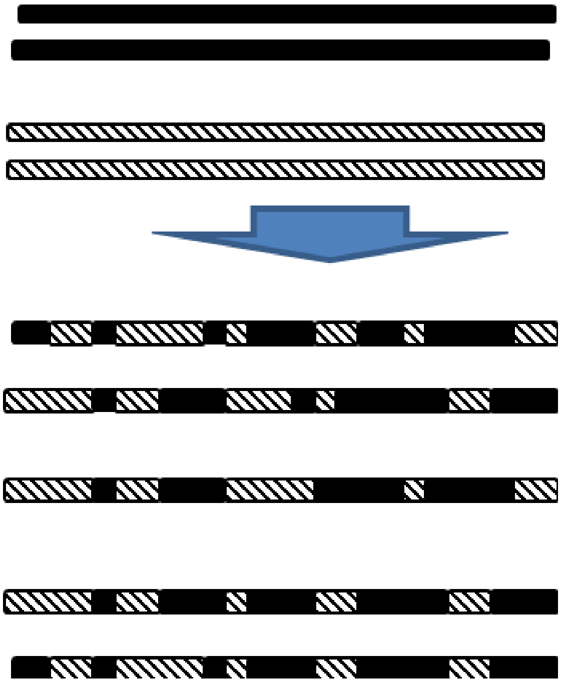

Figure 1: Chunks of ancestral chromosomes after cross of pure lines and several generations

5.1 Within-family and population 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:

Individually considered, the two families have strong within-family linkage disequilibrium. In family 1, pairs "AB" and "ab" come together, but in family 2 pairs "Ab" and "aB" come together. Still, 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 Tenesa *et al.* (2007) .

5.1.1 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 two pieces of information from each family are contradictory when pooled together. 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 $\frac{(3\times 10+1\times(-10))}{4} = 5$, in family one. In family two this is exactly the opposite: $\frac{(1\times 10+3\times(-10))}{4} = -5$, and across all families, Locus B/b would have an effect of $\frac{(3\times 10+1\times(-10))+(1\times 10+3\times(-10))}{4+4} = 0$. Therefore, allele B is a good predictor both within families 1 and 2, but not across families.

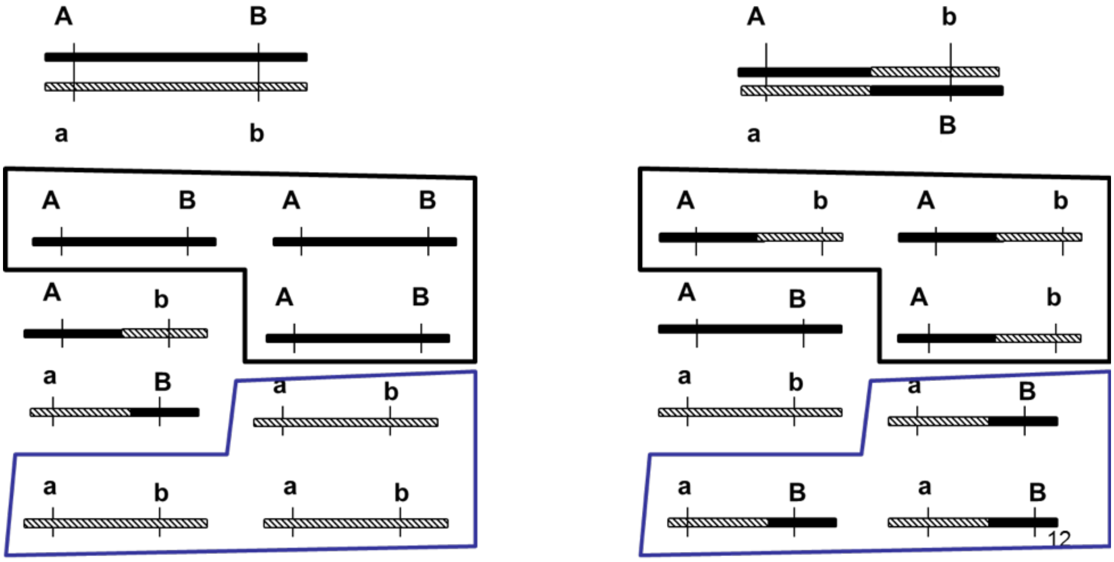

Figure 2: Two sires and eight gametes of the progeny, where each family shows linkage disequilibrium but there is no population linkage disequilibrium

5.2 Quantifying linkage disequilibrium from gametes' genotypes or from individuals' genotypes

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

$$D = freq\,(\mathrm{AB}) - freq\,(A)\,freq(B)$$

Hill and Robertson (1968) proposed, for biallelic loci, to assign numerical values based on gene contents (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 at gametes: {AB, AB, ab, aB, ab, ab, Ab, AB, Ab, AB} can be written as two variables, one X for "A",

$$X = \{1, 1, 0, 0, 0, 0, 1, 1, 1, 1\}$$

and one Y for "B",

$$Y = \{1, 1, 0, 1, 0, 0, 0, 1, 0, 1\}$$

We can get the correlation from R

```
X=c(1,1,0,0,0,0,1,1,1,1)
Y=c(1,1,0,1,0,0,0,1,0,1)
cor(X,Y)
0.4082483
```

and therefore $r = 0.41$. It can be shown that $r = \frac{D}{\sqrt{p_A q_A p_B q_B}}$ where $p_A = 1 - q_A = \mathrm{freq}(A)$. 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 (e.g. it is not the same to use as a reference A or a) but r^2 is invariant to the reference allele.

We just said that we need genotypes at gametes. This implies that we need to know the *phases* of the genotypes. But the phases are not known, although they may be deduced using some phasing software. We may still use Hill 1968 and compute correlations of gene contents. Our example was:

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

But we actually have 5 individuals, with genotypes (note the semicolon separating individuals):

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

If we put this in form of gene content it gives the following table:

X	2	0	0	2	2
Y	2	1	0	1	1

And therefore we get a correlation as

```
X=c(2,0,0,2,2)
Y=c(2,1,0,1,1)
cor(X,Y)
[1] 0.6454972
```

In this example, this value of $r = 0.65$ is not quite the previous estimate of $r = 0.41$, but in practice using genotypes instead of phased gametes results in good estimates (Rogers and Huff 2009).

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

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 also called *individual gene frequencies*, a confusing term).

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

Table 7: Example of gene content for multiallelic blood group

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

Table 8: Example of gene content for biallelic blood group

	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.

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", just as we count milk yield, height, or number of piglets born. The funny thing is that gene content *can also be studied as a quantitative measure* - just like milk yield, height, or number of piglets born-, and it can be therefore studied as a quantitative trait (although it is *not* a *continuous* trait). Therefore, gene content 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:

$$\text{Var}(z) = E\left(z^2\right) - E(z)^2$$

Table 9: Variance of gene content

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

The expectation $E\left(z^2\right)$ can be computed by weighting the column z^2 with the column *Frequency*. Therefore $\sigma_z^2 = \text{Var}\left(z\right) = 4p^2 + 2pq - \left(2p\right)^2 = 2pq$

The *heritability of gene content* is the ratio of genetic to environmental variances. 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 (if you think on it, the substitution effect is exactly $\alpha = 1$). Also, there is no residual error as the gene content is measured (in principle) perfectly. Therefore, the heritability is 1.

6.3 Covariance of gene content across two individuals.

Let's write the gene content of two individuals i and j as z_i, z_j . The covariance is $Cov\left(z_i, z_j\right)$. 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 $\Theta_{\text{ij}} = A_{\text{ij}}/2$, where Θ is known as Malecot "coefficient de parenté", kinship, or coancestry and A_{ij} is the additive relationship. This is just standard theory – two alleles from two individuals are identical if they are IBD. Therefore

$$Cov\left(z_i, z_j\right) = E\left(z_i z_j\right) - E\left(z_i\right) E\left(z_j\right)$$

$E\left(z_i\right) = E\left(z_j\right) = 2p$. $E\left(z_i, z_j\right)$ 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* in two cases. The first one is that the first individual got the allele A (with probability p) and the second one got A as well because it was identical by descent (with probability $A_{\text{ij}}/2$), which yields a probability of $pA_{\text{ij}}/2$. The second case is that the first individual got the allele A (with probability p , the second individual was *not* identical by descent (with probability $1 - A_{\text{ij}}/2$) , *but at the same time* by chance the second individual had the "A" allele with probability p, which yields a probability of $p(1 - A_{\text{ij}}/2)p$. Summing both probabilities we have $pA_{\text{ij}}/2 + p(1 - A_{\text{ij}}/2)p = pqA_{\text{ij}}/2 + p^2$, and multiplying by four possible ways gives $E\left(z_i z_j\right) = A_{\text{ij}} 2pq + 4p^2$. Putting all together gives

$$\text{Cov}\left(z_i, z_j\right) = A_{\text{ij}} 2pq$$

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; Habier *et al.* 2007; Gengler *et al.* 2008).

Extending the result above implies that the gene content in a population can be described like any other trait:

$$E(\mathbf{z}) = \mathbf{2}p$$

$$Var(\mathbf{z}) = \mathbf{A}2pq$$

where $\mathbf{A}$ is the classical numerator relationship matrix.

6.4 Quality control using heritability of gene content

This was explored by (Forneris *et al.* 2015). If gene content is a quantitative trait, we can estimate its heritability. We just need a pedigree file and a data file, although the data is now gene content. The method simply consists in modelling the genotype $\mathbf{z}$ as a quantitative trait:

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

Where $\mathbf{W}$ is a matrix of incidence with 1's for genotyped individuals and 0 otherwise. This is how the data file looks like:

```
1 1 533
0 1 1732
2 1 1207
1 1 952
0 1 678
1 1 2299
0 1 2581
1 1 2845
1 1 3123
```

(gene content, overall mean, animal id). Then we can use REML to estimate the heritability [1]. The result of REML is something like:

```
Final Estimates
Genetic variance(s) for effect 2
0.34311
Residual variance(s)
0.56669E-04
...
h2 - Function: g_2_2_1_1/(g_2_2_1_1+r_1_1)
Mean: 0.99983
```

REML can *not* estimate a heritability of exactly 1, but it should yield almost 1. If not (for instance if $\widehat{h}^2 < 0.98$) we have a problem, either in the genotypes or in the pedigree.

This has been included in the genomic programs of the blupf90 suite (e.g. `preGSf90` or `blupf90+` , with option `OPTION h2_gene_content`), and in the quality control program `qcf90` (with command line flag `--qc h2gc`) and it works very nicely e.g.:

```
snp logL_h2_0.97 logL_h2_0.98 logL_h2_0.99 2(logL_h2_0.97-logL_h2_0.98) h2_max status
         1     -1282.002818         -1282.106147         -1282.472047         0.2066576197         0.9710648424     ok
         2     -1047.334383         -1046.672706         -1046.224648         -1.323353994          1.000000000     ok
         3      1731.884046          1731.881892          1731.671095         0.4306966634E-02     0.9748967801     ok
         4     -964.4946030         -963.7620717         -963.2586319         -1.465062583          1.000000000     ok
         5      1486.562143          1486.718932          1486.659239        -0.3135772773         0.9822424508     ok
         6      2010.251309          2011.104258          2011.780516         -1.705897456          1.000000000     ok
         7      1252.624205          1253.160826          1253.491438         -1.073241969          1.000000000     ok
         8     -1132.914492         -1132.507824         -1132.315053        -0.8133373150         0.9940122911     ok
         9     -618.7259080         -617.2517575         -615.9944968         -2.948301031          1.000000000     ok
        10      200.2745923          190.4188376          180.0119799          19.71150935         0.7961633151     ***
```

6.5 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 major gene (for instance, DGAT1) of interest. It would be useful to have the genotype at the major gene 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 modeling the genotype $\mathbf{z}$ as a quantitative trait (just like in the previous section):

[1] This is an approximation as REML assumes multivariate normality, and gene content has only 3 cases. But gene content can not be modelled as a threshold model, because then the expression $Var(\mathbf{z}) = \mathbf{A}2pq$ is not longer true.

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

where $\mathbf{W}$ 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, $\widehat{\mathbf{u}}$ contains estimates of gene content for all individuals (these are equal to the observed genotype for the genotyped individuals) and $\widehat{\mu}$ actually contains $2\widehat{p}$.

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 major gene than what they actually are. For instance, isolated individuals will have an estimate consisting in $2\widehat{p}$. 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). It also forms the bases of the gene content multiple-trait BLUP that is briefly described next.

6.6 Gene Content Multiple-Trait BLUP

GCMTBLUP can be seen as "Single Step GBLUP for one major gene". The reader is referred to (Legarra and Vitezica 2015) for details and also for simulated examples. Assume that we have a "normal" trait (say growth) and also one major gene. The method of "heritability of gene content" can be expanded to include *both* the "normal" trait y and the gene content z. For instance:

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

$$\mathbf{z} = \mathbf{X}_z\mathbf{b}_z + \mathbf{W}_z\mathbf{u}_z + \mathbf{e}_z$$

With genetic covariance across traits as described above, $\mathbf{G}_0 = \begin{pmatrix} \sigma^2_{u_y} & \sigma_{u_{z,y}} \\ \sigma_{u_{z,y}} & \sigma^2_{u_z} \end{pmatrix}$. The estimated covariance $\sigma_{u_{z,y}} = 2\text{pqa}$ is a function of the effect (a) of the major gene in z on the normal trait in y. This method is more accurate than Gengler's method because it estimates gene content for animals that have not been genotyped based on gene content of relatives but also on "normal" trait information. It actually comes in two flavors: Gene Content Multiple-Trait BLUP (GCMTBLUP) uses estimated $\sigma_{u_{z,y}}$ to estimate EBV's that include the major gene, in an optimal way. But if the gene effect is not known with certitude, $\mathbf{G}_0$ and therefore the effect of the major gene can be estimated by REML (GCMTREML).

7 Imputation in a nutshell

7.1 Classical imputation

Imputation has become part of the regular toolkit of genomic prediction. In essence, the problem is the following. Not all animals have the same kind of genomic information. Omitting the case of sequenced animals, here are the typical cases:

- Animals genotyped with a "medium" chip such as the 50/60K
- Animals genotyped as "low density", for instance 6K
- Animals genotyped as "very low density", e.g. 1000 markers

In addition, there is also the problem that for many animals, some (very few) markers are not genotyped. So that, if there are 50,000 markers in one chip, for a typical animal only 49,800 markers are genotyped. Another more complex cases are Genotyping By Sequence (GBS) and sequencing, but we will not detail such here.

The theory for imputation in animal breeding is well summarized in (VanRaden *et al.* 2011 ; Hickey *et al.* 2011). Output of the programs is usually exact genotypes (the genotype is assumed exactly known), fractional genotypes (probabilities of each genotype) or missing (the genotype of this particular marker and individual is too inaccurate to be imputed). The algorithm for imputation typically proceeds combining two sources of information:

1. If in one individual, a chunk (short enough to assume that there is no recombination) of a chromosome (the paternal or the maternal) can be unambiguously identified as coming from one of the four chromosomes of its parents, then the whole chunk has been transmitted. This is fast and efficient if there are individuals with genotypes and pedigree.

2. If in one individual, at one chunk of a chromosome, a set of markers form a particular pattern, that resembles closely patterns that are already known and that are present in the population, then the "holes" are filled in according to the "known" pattern. This is linkage-disequilibrium based imputation.

A reminder. What imputation *can* do:

1. fill in holes from "lower" to "higher" densities (6K to 50K, 50K to 700K, 700K to sequence)

2. fill in missing markers in the genotypes. For instance, for an animal with a call rate of 99% for a 50K SNP chip, imputation can complete the 500 missing genotypes. This is useful.

Some animals can be imputed without own genotypes using information from genotyped offspring (around 5 offspring gives a decent imputation). In all other cases, it is very hard to impute animals that have not been genotyped for any marker.

7.1.1 Quick and dirty imputations

These forms of imputation are *not* recommended, but it might be useful for quick studies or prototyping, or if the number of missing genotypes is really small (say, individual and animal call rates ≈ 0.99). One form is just assigning the most frequent genotype ("AA", or "Aa", or whatever). Another form is simply assigning genotypes at random based on drawing the genotypes from a distribution with probabilities $(p^2, 2pq, q^2)$; in R this would be

```
z=sample(c(0,1,2),1,prob=c(p^2,2*p*q,q^2))
```

Again, this is *not* recommended. For instance, it can easily give parent-offspring incompatibilities.

7.1.2 Linear imputation

Gengler's (2007) method cites above "imputes" "genotypes" using a linear method (BLUP) for a linear trait (gene content). It ignores all the neighboring markers and it also ignores the Mendelian nature of inheritance of markers, i.e. the offspring of a couple "AA" x "aa" is forcedly "Aa". But the interesting point of Gengler's method is that it can be described analytically, which will eventually lead to the development of Single Step GBLUP (Christensen and Lund 2010).

In particular, the usefulness of the method is because it gives a framework for the error in the imputation. We will see this later.

8 Bayesian inference

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

$$p\left(A,B\right)=p\left(\mathrm{A}\ |B\right)p\left(B\right)=p\left(B|A\right)p\left(A\right)$$

Bayes' theorem says that

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

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 (effects, for instance) knowing A (observed phenotypes). For every value of B we do the following:

1. We compute $p\left(A|B\right)$, 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 $p\left(A|B\right)p\left(B\right)$ to form $p(A)$, which is called the *marginal density of A.*

8.1 Example of Bayesian inference

Assume that we have a collection of quantitative phenotypes $\mathbf{y}$=$\{1,0,-0.8\}$ with $k=3$ records and a very simple model $\mathbf{y}=\mathbf{1}\mu+\mathbf{e}$ with $Var\left(\mathbf{e}\right)=\mathbf{R}=\mathbf{I}\sigma_e^2$ 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

$$p\left(\mu|\mathbf{y}\right)=\frac{p\left(\mathbf{y}|\mu\right)p\left(\mu\right)}{p\left(\mathbf{y}\right)}$$

where

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

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

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

Table 10: example of Bayesian inference with two *a priori* values for μ

	$p(\mu)$	$p(\mathbf{y}\|\mu)$	$p(\mathbf{y}\|\mu)p(\mu)$	$p(\mu\|\mathbf{y}) = \frac{p(\mathbf{y}\|\mu)p(\mu)}{p(\mathbf{y})}$
$\mu = -1$	0.5	0.0051	0.00255	0.40
$\mu = 1$	0.5	0.0076	0.00381	0.60
$p(\mathbf{y})$			0.00636	

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 $E(\mu|\mathbf{y}) = 1 \times 0.60 + -1 \times 0.40 = 0.20$.

If the prior distribution for the mean is continuous, for instance $N(0, \sigma^2_\mu)$ (say $\sigma^2_\mu = 10$, 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

$$p(\mu|\mathbf{y}) = N\left(\widehat{\mu}, lhs^{-1}\right)$$

where

$$lhs = \frac{\mathbf{1}'\mathbf{1}}{\sigma^2_e} + \frac{1}{\sigma^2_\mu}$$

$$\widehat{\mu} = \left(\mathrm{lhs}^{-1}\right)\mathbf{1}'\mathbf{y}\sigma^2_e$$

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

8.2 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^2_e, so we want to evaluate

$$p\left(\mu, \sigma^2_e|\mathbf{y}\right) = \frac{p\left(\mathbf{y}|\mu, \sigma^2_e\right)p(\mu)\,p\left(\sigma^2_e\right)}{p(\mathbf{y})}$$

Writing down in closed form the posterior distributions is impossible. The Gibbs sampler is a numerical Monte Carlo 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^2_e. If we knew σ^2_e, 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:

$$p(\sigma^2_e|\mathbf{y}, \mu)$$

$$p(\mu|\mathbf{y}, \sigma^2_e)$$

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 $\sigma^2_e = 1$. Then we draw a new μ from

$$p\left(\mu|\mathbf{y}, \sigma^2_e\right) = N\left(\widehat{\mu}, lhs^{-1}\right)$$

Then σ_e^2 from

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

which is the conditional distribution assuming flat priors for σ_e^2. Then we plug in this value into $p\left(\mu|\mathbf{y},\sigma_e^2\right)$ 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.

8.3 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.

9 Models for genomic prediction

If SNP are just markers located outside genic regions, most of the times, why to use them? Because they may be linked to QTL or genes, fact that can be explained by an event called linkage disequilibrium (LD). The LD is based on expected versus observed allele frequencies and measures the non-random association of alleles across loci. We have seen LD before. The strength of the association between two loci is measured by the correlation. We assume that, if neighboring SNPs are tightly correlated, then QTLs that are "in the middle" should be strongly correlated as well (this might not be true – for instance if all QTLs have very low frequency, but that seems unlikely).

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

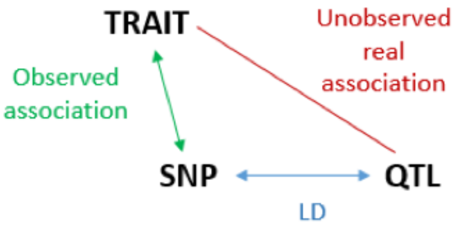

Figure 3: Indirect association QTL - markers - phenotype

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

There are mainly two classes of methods for genomic selection:

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

For most of the livestock populations, the number of SNP is greater than the number of genotyped animals, which results in the famous "small n big p problem". As the number of parameters is greater than the data points used for estimation, a solution is to assume SNP effects are random; in this way, all effects can be jointly estimated. To present the SNP effect-based method we will start with a single gene and we will move towards more of them.

9.1 Simple marker model

9.1.1 Multiallelic

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:

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

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

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

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

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

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 $\mathbf{y} = \{12, 35, 6\}$ this gives

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

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

9.1.2 Biallelic

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

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

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. Then $\mathbf{Z}$ becomes a vector $\mathbf{z}$ and the vector $\begin{pmatrix} a_A \\ a_B \end{pmatrix}$ becomes a scalar a_A . So for individuals $\{BB, AA, BA\}$ we have that

$$\mathbf{z}a = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix} a_A$$

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 polygenic term $\mathbf{u}$, based on pedigree $Var(\mathbf{u}) = \mathbf{A}\sigma_u^2$, like for instance in

$$\mathbf{y} = \mathbf{Xb} + \mathbf{z}a + \mathbf{Wu} + \mathbf{e}$$

You may realize that this follows the chapter "Values and means" in (Falconer and Mackay 1996).

9.2 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 Genome-wide Association Study (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.

9.2.1 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.

9.2.2 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

$$\text{estimated effect } = \text{real effect} + \text{"estimation noise"}$$

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.

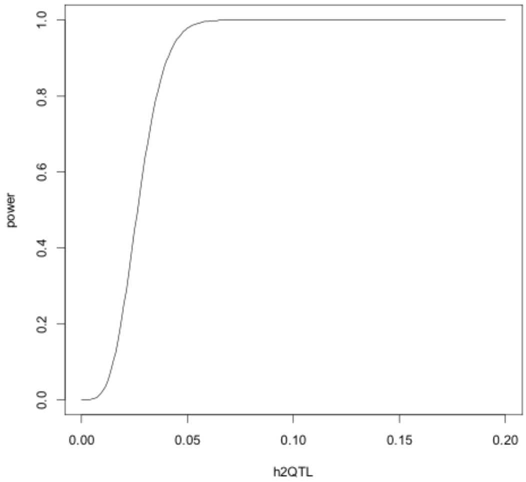

Figure 4: Power of detection of QTL effects perfectly tagged explaining from zero to 100% phenotypic variance

```
#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.

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).

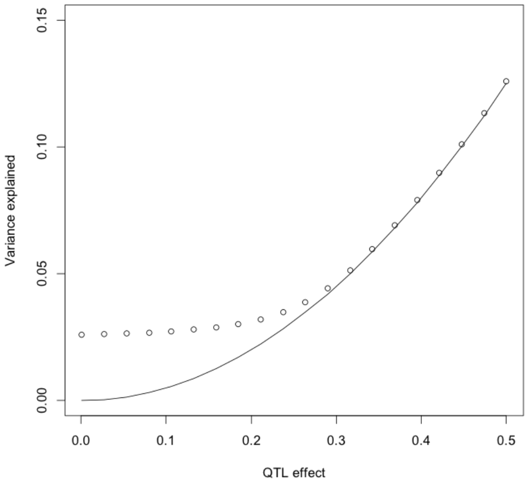

Figure 5: True (straight line) and apparent (dotted line) variance explained by QTL effects going from zero to 0.5 genetic standard deviations

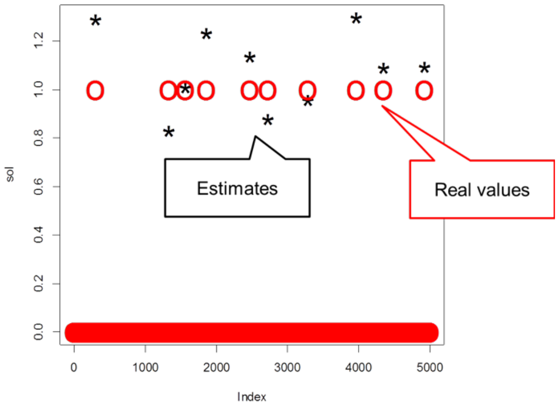

Figure 6: Real (**O**) and estimated (*****) effects after GWAS-like simulations with 10 true QTLs in 5000 markers, 1000 individuals.

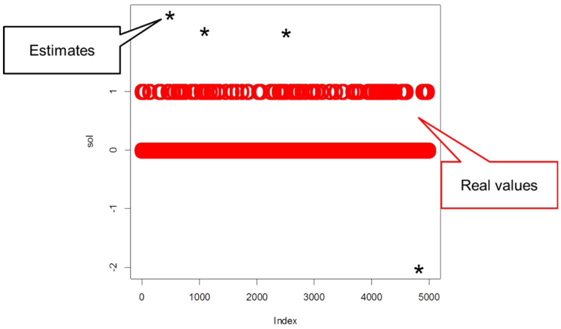

Figure 7: Real (**O**) and estimated (*****) effects after GWAS-like simulations with 100 true QTLs in 50000 markers, 1000 individuals

9.3 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. One example is the bovine myostatin gene (GDF8), i.e. both the Belgian Blue and South Devon breeds carry the same GDF8 mutation, but they have different conformation and double-muscling phenotypes (Smith *et al.* 2000; Dunner *et al.* 2003).

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.

Individual i has a breeding value u_i. According to the previous paragraphs, we will try to predict the breeding value of an individual defined as a sum of marker effects a_k (there are m of them). An individual has genotypes coded in $\mathbf{z}_i$, its breeding value is the sum of marker effects $a_k$weighted by the coefficients in $\mathbf{z}_i$: $u_i = \sum_{k=1,m} z_{ik} a_k = \mathbf{z}_i \mathbf{a}$. For all individuals this becomes $\mathbf{u} = \mathbf{Z}\mathbf{a}$.

9.3.1 Multiple marker regression as fixed effects

9.3.1.1 Multiallelic 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:

$$\mathbf{y}\text{=Xb+}\mathbf{Za}\text{+}\mathbf{e}$$

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 $\{\text{BC/EE}, \text{AA/EF}, \text{BD/FF}\}$. Then

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

9.3.1.2 Biallelic With biallelic markers, we can reduce the number of unknowns to just one effect per marker – the effect of the reference allele. Assume now that we have just three individuals with two biallelic markers: a two-allele $\{A, B\}$ locus, and a two-allele locus with alleles $\{E, F\}$ and three individuals with genotypes $\{\text{BA/EE}, \text{AA/EF}, \text{BB/FF}\}$. If we fit one effect per allele the system of equation is:

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

And if we reduce the effects to one effect per marker, we get

$$\mathbf{Za}=\begin{pmatrix}1 & \vdots & 2\\ 2 & \vdots & 1\\ 0 & \vdots & 0\end{pmatrix}\begin{pmatrix}a_A\\ \cdots\\ a_E\end{pmatrix}$$

Again, estimation of **a** can proceed by least squares.

9.3.1.3 Massive number of markers Imagine that we have 20 markers and 3 individuals, matrix **Z** looks like:

1 1 2 2 1 0 0 1 0 0 2 0 0 2 0 2 2 0 1 1

0 1 2 1 2 1 0 1 2 2 2 2 0 2 1 0 1 0 0 1

2 0 2 0 0 2 1 0 0 0 1 1 0 2 2 1 0 0 0 1

But SNP chips yield *thousands* of markers. 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 **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.

9.4 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 **a** the marker effects, and **y** the data, the *Posterior Mean*, or *Conditional Expectation* of (estimators of) marker effects is given by the expression

$$\widehat{\mathbf{a}}=E\left(\mathbf{a}\,|\mathbf{y}\ \right)=\frac{\int \mathrm{a}\ p\left(\mathbf{y}\mid\mathbf{a}\right)p\left(\mathbf{a}\right)d\mathbf{a}}{\int\ \ p\left(\mathbf{y}\mid\mathbf{a}\right)p\left(\mathbf{a}\right)d\mathbf{a}}$$

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" **a** and its estimator, $\widehat{\mathbf{a}}$ (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 ; Goffinet and Elsen 1984; Fernando and Gianola 1986). The introduction of the prior distribution $p\left(\mathbf{a}\right)$ 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 $p(\mathbf{a})$
2. The likelihood of the data given the marker effects, $p(\mathbf{y} \mid \mathbf{a})$

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

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

where matrix $\mathbf{R}$ 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 $p(\mathbf{a})$ 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.

9.4.1 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 $|\mathbf{a}| < \lambda$. Another example is the Ridge Regression, where there is a penalty function of a_i^2 – the larger the square of the effect, the more penalized. 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.

9.5 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 11: A form of prediction equation.

Locus	Allele	Effects estimates
1	A	+10
2	E	+5

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

10 SNP effect-based methods: SNP-BLUP and Bayesian regressions

For most of the livestock populations, the number of SNP is greater than the number of genotyped animals, which results in the famous "small n big p problem". As the number of parameters is greater than the data points used for estimation, a solution is to assume SNP effects are random (or that they have a prior distribution); in this way, all effects can be jointly estimated. Even if the number of genotyped animals is large, still it makes sense to fit SNPs as random, because the prior information says that small effects are frequent and large effects are unlikely.

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 $p(\mathbf{y} \mid \mathbf{a}) = \text{MVN}(\mathbf{Xb} + \mathbf{Za}, \mathbf{R})$ and a prior distribution for markers, $p(\mathbf{a})$. 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.

10.1 Allele coding.

Allele coding is the assignment of genotypes to numerical values in matrix $\mathbf{Z}$. 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:

aa	Bb
AA	bb
Aa	bb
aa	bb

This can be coded with *one effect by allele*:

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

where a_{2B} is the allele "B" of the 2^{nd} 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 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	$2a_i$	$(2 - 2p_i)\,a_i$

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 $\mathbf{Z}$ matrices:

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

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

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

for the "centered" coding, allelic frequencies where 0.375 and 0.125; it can be verified that each column of centered $\mathbf{Z}$ sums to 0. This will be true if allelic frequencies are computed from observed data. VanRaden (VanRaden 2008) defined matrix $\mathbf{M}$ as $\mathbf{Z}$ with 101 coding and then $\mathbf{Z}$=$\mathbf{M}-\mathbf{P}$, where $\mathbf{P}$ is a matrix with $2(p_i - 0.5)$ or $\mathbf{P} = \mathbf{2p}'$.

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

Table 13: 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, $\mathbf{u} = \mathbf{Za}$ will be the same regardless of the coding.

The following Julia code illustrates this

```julia
# test SNP-BLUP with different allele frequencies
using Random
nsnp=10
nanim=4
vary=10
h2=0.3
vare=(1-h2)*vary
varu=(h2)*vary


function SNP_BLUP(X,Z,y,vare,vara)
    RHS=[X Z]'*y/vare
    LHS=[X Z]'*[X Z]/vare
    #display(LHS)
    LHS[(1+1):(1+nsnp),(1+1):(1+nsnp)] += I/vara
    sol=LHS\RHS
    return sol
end

Random.seed!(1234)
p=rand(Beta(2,2),nsnp)
Z=zeros(nanim,nsnp)
for i in 1:nsnp
```

```
    for j in 1:nanim
        Z[j,i]=Float64.(rand(Binomial(2,p[i]),1))[1]
    end
end

# sample y (not from an actual genetic theory - this is just
# white noise)

y=rand(Normal(0,vary),nanim)

# we use the same variance component for all SNP-BLUPd
# and it is actually neither of the allele frequencies, the sampled or the observed
vara=varu/(2*nsnp*0.5*0.5)

# compute observed frequencies (result of sum is a row vector, we cobvert to
#column vector)
pobs=(sum(Z,dims=1)/(nanim*2))'

# use observed frequencies
unos=ones(nanim)
Zstar= Z - 2*unos*pobs'
# mean
X=unos

sol_obs=SNP_BLUP(X,Zstar,y,vare,vara)
display(sol_obs)

# use drawn frequencies
Zstar= Z - 2*unos*p'
sol_drawn=SNP_BLUP(X,Zstar,y,vare,vara)
display(sol_drawn)

display(isapprox(sol_obs[1+1:1+nsnp] , sol_drawn[1+1:1+nsnp]))

# use .5 frequencies
Zstar= Z .- 1
sol_05=SNP_BLUP(X,Zstar,y,vare,vara)
display(sol_05)

display(isapprox(sol_obs[1+1:1+nsnp] , sol_05[1+1:1+nsnp]))

# use stupid frequencies
pstupid=rand(Normal(0,1),nsnp)
Zstar= Z - 2*unos*pstupid'
sol_stupid=SNP_BLUP(X,Zstar,y,vare,vara)
display(sol_stupid)
display(isapprox(sol_obs[1+1:1+nsnp] , sol_stupid[1+1:1+nsnp]))
```

10.2 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 the prior distribution 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

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

$$\widehat{a} \ = lhs^{-1}\mathbf{1}'\mathbf{y}/\sigma_e^2$$

10.2.1 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.

10.2.2 Marker effect is random

In this case, the marker has different effects across populations because it is on feeble LD with some QTL. Then the “true” effect of the marker may change all the time, because at each generation LD will be different. Thus, we can say that the marker effect is random and 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.

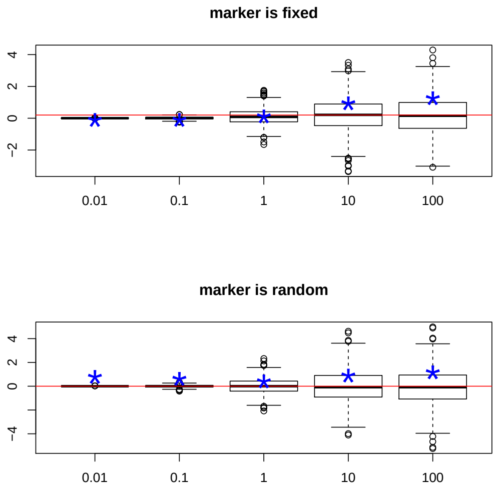

Figure 8: Distribution (boxplots) of errors in the estimate of one marker effect for different levels of shrinkage (X axis). No error is the red line. Blue stars indicate the square root of the mean square error

10.3 Genetic variance explained by markers

A population of n individuals has different breeding values $u_1 \ldots u_n$. These individuals have a certain genetic variance $Var\left(u\right) = \sigma_u^2$. 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 $Var\left(u\right) = E\left(u^2\right) - E\left(u\right)^2$, which is developed in the following Table

Table 14: 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		$4p^2a_i^2 + 2pqa_i^2$	$2pa_i$

So, finally the variance explained by one marker is $4p^2a_i^2 + 2pqa_i^2 - \left(2pa_i\right)^2 = 2pqa_i^2$. Markers with intermediate frequencies will explain most genetic variation. This is one of the reasons to ignore markers with low allele frequency.

10.3.1 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 $\mathbf{z}$ 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 $Var\left(z_1\right) = 2p_1q_1$. The term $Cov\left(z_1, z_2\right)$ turns out to be $\left(z_1, z_2\right) = 2r\sqrt{p_1q_1p_2q_2}$, 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. Although Bulmer effect generates, in selected populations, linkage disequilibrium, we will ignore it here; in this case, on average this term will typically cancel out.

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

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

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

If we assume that all markers have the same variance *a priori* σ_{ao}^2 (say $\sigma_{a1}^2 = \sigma_{a2}^2 = \sigma_{a3}^2 = \ldots = \sigma_{a0}^2$, then $\sigma_u^2 = 2\sum_i^{\text{nsnp}} p_iq_i\sigma_0^2 \ = 2\sigma_0^2\ \sum_i^{\text{nsnp}} p_iq_i$ .We can factor out σ_{ao}^2 and we have the famous identity (Fernando *et al.* 2007; Habier *et al.* 2007 ; VanRaden 2008; Gianola *et al.* 2009):

$$\sigma_{a0}^2 = \frac{\sigma_u^2}{2\sum_i^{\text{nsnp}} p_iq_i}$$

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.

10.3.2 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 $\widehat{a}$ for each marker. We may say that each marker i explains a variance $2p_iq_i\widehat{a}_i^2$. 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 $2\sum p_iq_i\widehat{a}_i^2$ underestimates the total genetic variance, because estimates $\widehat{a}_i$ are shrunken towards 0. Better estimators will be presented later in, among others, GREML and BayesC.

10.4 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 they will have similar incidence matrices in **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)

10.5 RR-BLUP or SNP-BLUP

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

The SNP-BLUP model for the phenotypes is typically something like:

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

with **b** fixed effects (i.e., an overall mean), **a** marker effects, and **e** residual terms, with $Var\left(\mathbf{e}\right) = \mathbf{R}$ and usually $\mathbf{R}$=$\mathbf{I}\sigma_e^2$. Matrix **Z** contains genotypes coded in any of the forms that we have described previously (usually centered, 012 or 101).

The prior for markers can be written as:

$$p\left(\mathbf{a}\right) = \prod_{i=1,nsnp} p(a_i)$$

where

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

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.

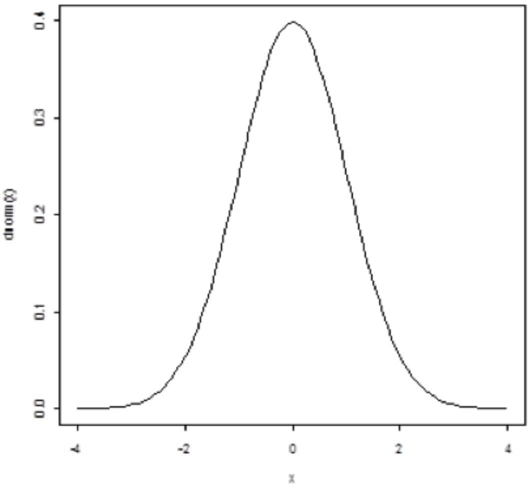

Figure 9: Standard normal distribution

From the Figure, it can be remarked that in a normal distribution 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 few 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:

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

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

10.5.1 Mixed Model equations for SNP-BLUP

10.5.1.1 Single trait 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:

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

Note that this is a linear estimator. If $Var(\mathbf{a}) = \mathbf{D} = \mathbf{I}\sigma_{a0}^2$ and $Var\left(\mathbf{e}\right) = \mathbf{R} = \mathbf{I}\sigma_e^2$, then we can simplify them to

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

with $\lambda = \sigma_e^2/\sigma_{a0}^2$. This expression is also known as *Ridge Regression*, although the Ridge Regression literature presents $\mathbf{I}\lambda$ (or $\mathbf{D}$) merely as a computational device to warrant correct estimates, and genetics literature presents λ as the ratio of residual to genetic variances. (We don't like the name Ridge Regression for this reason). Following traditional notations, we will talk about *lhs* (left hand side of the equations) and *rhs* (right hand side): $lhs \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \end{pmatrix} = rhs$.

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 $\mathbf{Z}^{'}\mathbf{Z}$ is completely dense and full.

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

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

10.5.1.2 Multiple trait

For a multiple trait model, the equations are as above:

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

but $\mathbf{R}$ and $\mathbf{D}$ include multiple trait covariances, e.g. $\mathbf{R}=\mathbf{I}\otimes\mathbf{R}_0$ and $\mathbf{D}=\mathbf{I}\otimes\mathbf{S}_{a0}$.

10.5.2 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

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

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. Also, sometimes the genetic variance that is used in the "large" (national) genetic evaluations does not match well the genetic variance existing in the population with genotypes. But the equation $\sigma_{a0}^2 = \frac{\sigma_u^2}{2\sum_i^{\mathrm{nsnp}} p_i q_i}$ is usually a good guess.

As for the residual variance, it can be taken as well from previous studies.

For the multiple trait case, $\mathbf{S}_{a0} = \mathbf{G}_0/2\sum_i^{\mathrm{nsnp}} p_i q_i$ where $\mathbf{G}_0$ is a matrix with estimates of the genetic covariances across traits.

10.5.3 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^2markers 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 $\mathbf{Z}$ 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:

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

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", $\widehat{\mathbf{e}}^l$:

$$\left(\mathbf{z}_i'\mathbf{z}_i \ + \ \lambda\right)\widehat{a}_i^{l+1} \ = \ \mathbf{z}_i' \ \left(\widehat{\mathbf{e}}^l+\mathbf{z}_i\widehat{a}_i^l\right) = \mathbf{z}_i'\widehat{\mathbf{e}}^l \ + \mathbf{z}_i'\mathbf{z}_i\widehat{a}_i^l$$

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

$$\widehat{\mathbf{e}}^{l+1} \ = \widehat{\mathbf{e}}^l \ - \ \mathbf{z}_i\left(\widehat{a}_i^{l+1}-\widehat{a}_i^l\right)$$

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
```

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

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

Again, only matrix $\mathbf{Z}$ is used but its cross-product $\mathbf{Z}'\mathbf{Z}$ 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 markers in minutes. PCG has a (much) faster convergence than GSRU: see the graphs below. This makes it attractive for large application. However, GSRU can be converted with very few changes into a Gibbs Sampler application.

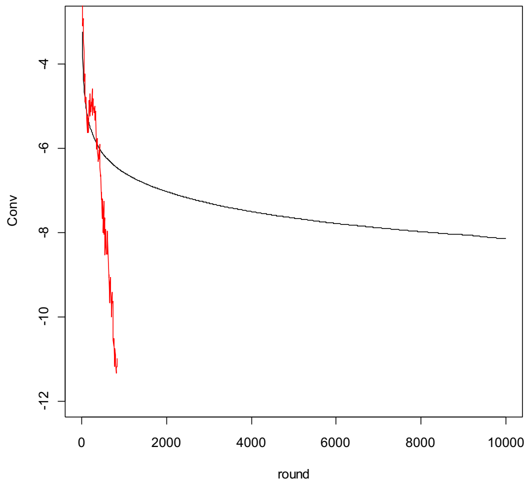

Figure 10: Convergence time for a large Holstein data set (left, GSRU in black, PCG in red)

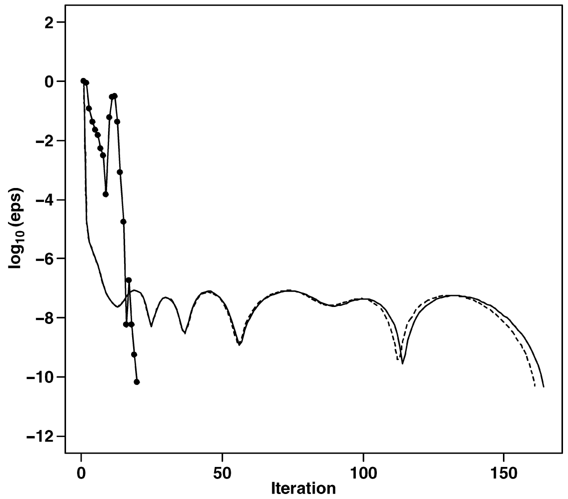

Figure 11: Convergence time for a mice data set (right, PCG line with points)

10.6 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. Although this is typically done using GREML, it can also be done in marker models. This was the case of (Legarra *et al.* 2008) in mice, and it has later been used to estimate genetic variances in wild populations (Sillanpaa 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
```

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 whereas 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 $E\left(\sigma_e^2 | S_e, \nu_e\right) = S_e^2/\nu_e$. One may use previous estimates and put therefore

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

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

NOTE In other parameterizations $E\left(\sigma_e^2 | S_e, \nu_e\right) = S_e^2$ and $E\left(\sigma_a^2 | S_a, \nu_a\right) = S_a^2$ 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 (Aguilar *et al.* 2018).

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

10.7 Transforming marker variance into genetic variance

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

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

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 much missing heritability – estimates of genetic variance with pedigree or markers usually agree up to, say, 10% of difference.

10.7.1 Example with mice data

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 15: 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 (times 10^{-4})	8.37	1.04×10^{-3}	3.93
Body length	0.040	9.09×10^{-6}	0.034
Body Mass Index (times 10^{-4})	2.49	0.80×10^{-3}	3.02

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) . There are essentially two methods to compare estimates from two different bases, presented by (Legarra 2016; Lehermeier *et al.* 2017), although they refer more to a GBLUP framework.

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.

10.8 Differential variances for markers

Real data, shows 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:

$$\widehat{a}_i = \frac{\frac{\mathbf{z}_i'\tilde{\mathbf{y}}}{\sigma_e^2}}{\frac{\mathbf{z}_i'\mathbf{z}_i}{\sigma_e^2} + \frac{1}{\sigma_a^2}}$$

Where $\tilde{\mathbf{y}}$ means “$\mathbf{y}$ corrected by all other effects” and σ_{ai}^2 is the shrinkage of that marker. In BLUP-SNP, we assume $\sigma_{\text{ai}}^2 = \sigma_{a0}^2$ 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 $\widehat{a}_i$, 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. 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.

10.8.1 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:

$$\widehat{\sigma}_{\mathrm{ai}}^{2} = \widehat{a}_{i}^{2} + C^{\mathrm{ii}}$$

where C^{ii} is the element corresponding to the i-th marker on the inverse of the Mixed Model Equations $\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\,\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\,\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\,\mathbf{Z}{+}\mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$.

This expression has two parts, the first, $\widehat{a}_{i}^{2}$, 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). *However*, the equation is *almost certainly wrong* because there is just one marker effect, and even if it was, the estimate is very inaccurate, because there is only one marker effect to estimate one variance component.

10.8.2 Bayesian estimation of marker variances

Marker "variances", can, however, be included 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:

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

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

$$p\left(\sigma_{\mathrm{ai}}^2|\mathrm{something}\right) = p(\ldots)$$

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.

10.9 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:

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

$$p\left(\sigma_{\mathrm{ai}}^2|S_a, \nu_a\right) = S_a\chi_{\nu_a}^{-2}$$

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):

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

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.

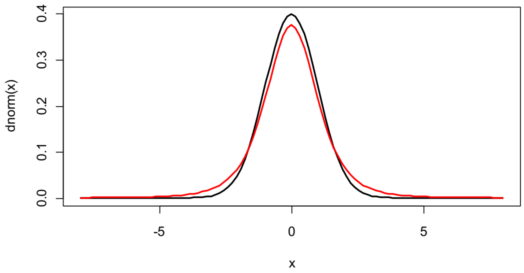

Figure 12: A priori distributions for BLUP-SNP (black) and BayesA (red)

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
```

10.10 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^2_{ai} 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:

$$p\left(a_i|\sigma^2_{\mathrm{ai}}\right)=N\left(0,\sigma^2_{\mathrm{ai}}\right)$$

$$\begin{cases} p\left(\sigma^2_{\text{ai}} \middle| S_a, \nu_a\right) = S_a \chi^{-2}_{\nu_a} \; \textit{with probability } 1-\pi \\ p\left(\sigma^2_{\text{ai}} \middle| S_a, \nu_a\right) = 0 \; \textit{with probability } \pi \end{cases}$$

Then, when $\sigma^2_{\text{ai}} = 0$ 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.

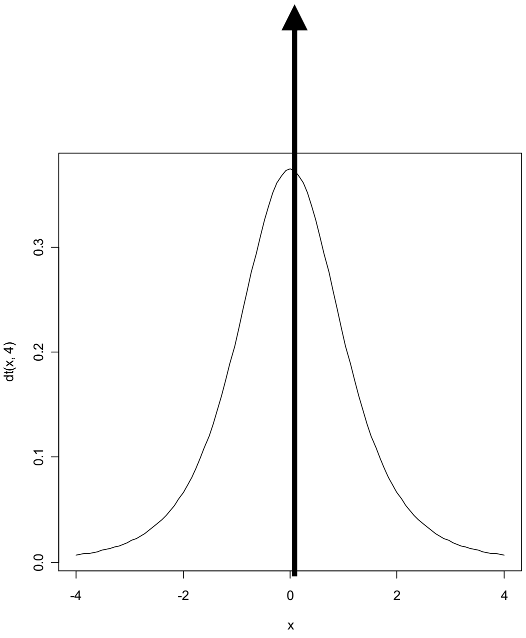

Figure 13: A priori distribution for BayesB

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).

10.11 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^2_a, 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^2_{a0}. 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:

$$p\left(a_i|\delta_i\right) = \begin{cases} N\left(0, \sigma^2_{\text{ai}}\right) \text{ if } \delta_i = 1 \\ 0 \; \textit{otherwise} \end{cases}$$

$$p\left(\sigma^2_{a0} \middle| S_a, \nu_a\right) = S_a \chi^{-2}_{\nu_a}$$

$$p\left(\delta_i = 1\right) = 1 - \pi$$

Where S_a can be set to something like S^2_a=$\sigma^2_{a0}\nu_{a0}$ with

$$\sigma^2_{a0} = \frac{\sigma^2_u}{(1-\pi)\, 2 \sum p_i q_i}$$

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^2_{a0} 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
```

10.11.1 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

$$\sigma^2_u = \sigma^2_{a0}\,(1-\pi)\, 2 \sum p_i q_i$$

Thus, the same total genetic variance can be achieved with large values of σ^2_{a0} 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.* 2013) 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 $\widehat{\delta_i}$, 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:

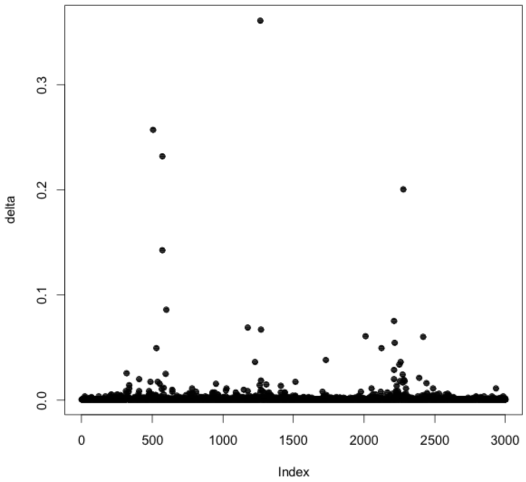

Figure 14: QTL signals from BayesCPi with Pi=0.999

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

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

In our case this is:

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

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) sometimes used in QTL detection (Varona *et al.* 2001 ; 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). We compared several strategies for GWAS including BayesCPi and our conclusion is that all result in similar results (Legarra *et al.* 2015b).

10.12 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.

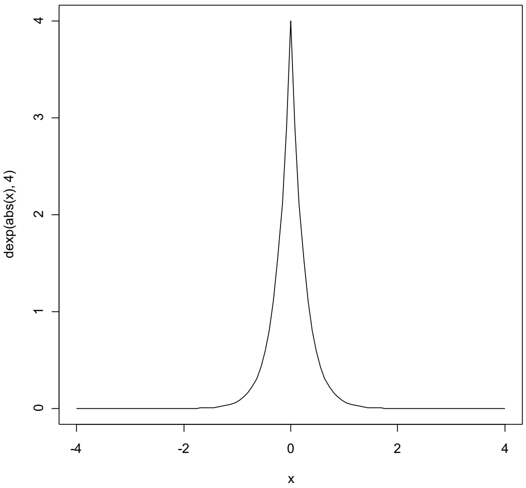

Figure 15: Prior distribution of marker effects for the Bayesian Lasso

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

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

where the density function is on the absolute value of the marker and not on its square like in the normal distribution. Coming back to our notion of variance of markers, (Park and Casella 2008 ; de los Campos *et al.* 2009) showed that the model is equivalent to a model with individual variances by marker, that is:

$$p\left(a_i|\sigma^2_{\text{ai}}\right) = N\left(0, \sigma^2_{\text{ai}}\right)$$

$$p\left(\sigma^2_{\text{ai}}|\lambda\right) = \frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\frac{\sigma^2_{\text{ai}}}{\sigma^2}\right)$$

(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.

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.* 2013)).

10.12.1 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 $\sigma^2 = \sigma^2_e$, the residual variance. (Legarra *et al.* 2011b) objected that it was unnatural to model the distribution of markers on the

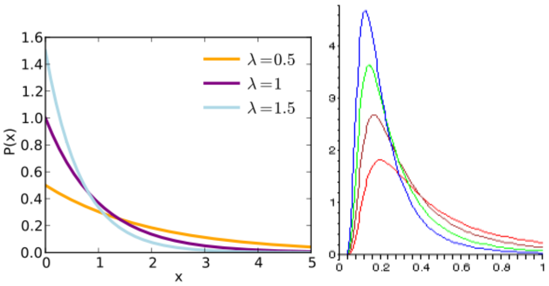

Figure 16: Shapes of the prior distribution of marker variances for the Bayesian Lasso (left) and Bayes A (right)

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 $Var\left(a_i|\lambda\right) = 2/\lambda^2$. 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
```

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()).

10.13 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^2_{al}) and one for the small effects (σ^2_{as}). 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.

10.14 Overall recommendations for Bayesian methods

BayesB seems to be little robust. The other methods are reasonably robust. My (AL) *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.

10.15 Empirical single marker variances from marker estimates

In SNP-BLUP (or equivalently, from GBLUP) it is easy to get marker estimates but running a full Bayesian analysis can be long or impossible. So, people came with ideas to get these weights (VanRaden 2008 ; Wang *et al.* 2012 ; Fragomeni *et al.* 2017)

- Quadratic: $\widehat{\sigma}^2_{\mathrm{ai}} \propto \widehat{a}^2_i$. In (Wang *et al.* 2012), the weight is actually $\widehat{\sigma}^2_{\mathrm{ai}} \propto 2p_iq_i\widehat{a}^2_i$, but this is a mistake as $2p_iq_i\widehat{a}^2_i$ the variance in the population explained by the marker, and not the variance of the marker effect itself. This quadratic weight tends to diverge as markers tend to extreme values.
- FastBayesA (Sun *et al.* 2012): $\widehat{\sigma}^2_{ai} \propto \frac{\widehat{a}^2_i+\nu S^2}{\nu+1}$, where the variance is regressed towards some prior value S^2. This scheme also tends to diverge
- nonlinearA (VanRaden 2008):

$$\widehat{\sigma}^2_{ai} = \sigma^2_{a0} 1.125^{\left|\frac{\widehat{a_i}}{sd(\widehat{a})}-2\right|}$$

10.16 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

$$\sigma^2_{\mathrm{ai}} = \sigma^2_{a0} \left(c^{\left(\frac{\left|\widehat{a_i}\right|}{\mathrm{sd}\left(\widehat{a_1}, \dots, \widehat{a_n}\right)} - 2 \right)} \right)$$

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^2_{ai} 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 $\frac{\left|\widehat{a_i}\right|}{\mathrm{sd}\left(\widehat{a_1}, \dots, \widehat{a_n}\right)}$ (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.

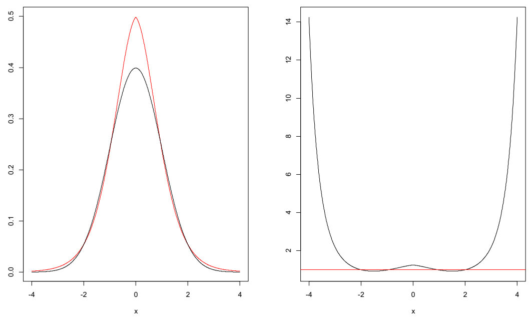

Figure 17: (Left) Shapes of the prior distribution of marker effects for VanRaden nonlinearA (red) and normal BLUP-SNP (black). (Right) Ratio of nonlinearA/normal densities.

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.

10.17 The effect of allele coding on Bayesian Regressions

We have explained how allele coding should (or can) proceed. [(Strandén *et al.* 2017) 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 $\mathbf{Z}$ for genotypes, *not* the frequencies used in $\sigma^2_{a0} = \frac{\sigma^2_u}{2\sum_i^{\text{nsnp}} p_i q_i}$.

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 $\widehat{\mathbf{a}}$ and estimated genetic values $\widehat{\mathbf{u}}$=$\mathbf{Z}\widehat{\mathbf{a}}$ are invariant to parametrization of $\mathbf{Z}$ (centered, 101 or 012 or 210), 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 $\mathbf{1}'\widehat{\mathbf{u}}$, 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 $\mathbf{1}'\widehat{\mathbf{u}}$=$\mathbf{1}'\mathbf{Z}\widehat{\mathbf{a}} = 0$. As for the marker variance σ^2_{a0} estimated by, say, BayesC, they also proved that it is invariant to parameterization of $\mathbf{Z}$.

In other words, we can use any coding (centered, 101 or 012 or 210) in $\mathbf{Z}$ for Bayesian methods. The estimated $\widehat{\mathbf{u}}$ will be the same, the estimated σ^2_{a0} or π will be the same, and the estimated genetic variance computed using, for instance, $\sigma^2_u = \sigma^2_{a0} 2 \sum_i^{\text{nsnp}} p_i q_i$ will be the same too.

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 $\widehat{\mathbf{u}}$ is *not* invariant to parameterization, because there will be a part of the overall mean absorbed, or not, by $\mathbf{Z}\widehat{\mathbf{a}}$. This implies that reports of the posterior variance of $\widehat{\mathbf{u}}$ will depend on the allele coding. The same result applies to GBLUP, as we will see later.

10.18 Reliabilities from marker models

10.18.1 Standard errors from Bayesian methods by MCMC

In these methods, at iteration t, samples of distribution of marker effects is obtained in the form of samples of these effects ($\widetilde{\mathbf{a}}_{(t)}$). At iteration t, samples of the breeding values can be obtained as $\widetilde{\mathbf{u}}_{(t)} = \mathbf{Z}\widetilde{\mathbf{a}}_{(t)}$. At the end of the MCMC process, the final estimate of the breeding value for, say, individual i consist of a *posterior* mean of all $\widetilde{u}$ for that animal,

$$\widehat{u}_i = \overline{\widetilde{u}}_i$$

and a posterior variance $Var(\widehat{u}_i) = Var(\widetilde{u}_i)$. This variance (or rather, its square root: the standard error) can be used in itself as a descriptor of the incertitude of the breeding value. A 95% confidence interval for the breeding value is roughly $\widehat{u}_i \pm 2\text{sd}(\widehat{u}_i)$.

10.18.2 Reliabilities

Reliabilities are only well defined for a multivariate normal model – SNP-BLUP with fixed σ^2_{a0}. The first method uses $Var(\widehat{u}_i)$ as above (i.e. from MCMC). Reliability can be obtained as

$$\text{Rel}_i = 1 - \frac{Var(\widehat{u}_i)}{\mathbf{z}_i\mathbf{z}_i^{'}\sigma^2_{a0}}$$

The second method uses the complete *a posteriori* distribution of marker effects:

$$Var(\mathbf{a}|\mathbf{y}) = \mathbf{C}^{\text{aa}}$$

That can be obtained by MCMC or by inversion of the SNP-BLUP equations. From here we can derive that:

$$Var(\mathbf{u}|\mathbf{y}) = \mathbf{Z}\mathbf{C}^{\text{aa}}\mathbf{Z}^{'}$$

And therefore $Var(\widehat{u}_i) = \mathbf{z}_i\mathbf{C}^{\text{aa}}\mathbf{z}_i^{'}$. The rest proceeds as before.

Imagine for instance that we have 50K markers and 1 million animals in predictions. Imagine that we use SNP-BLUP equations and we can obtain by inversion $\mathbf{C}^{\text{aa}}$, which is a 50K by 50K matrix. Then, for each animal, we compute $\mathbf{z}_i\mathbf{C}^{\text{aa}}\mathbf{z}_i^{'}$ (which has high cost) and $\mathbf{z}_i\mathbf{z}_i^{'}\sigma^2_{a0}$ (which has negligible cost).

These reliabilities have a problem. We know that both $\widehat{u}_i$ and $Var(\widehat{u}_i)$ are invariant to parametrization (coding of $\mathbf{Z}$). But $\mathbf{z}_i\mathbf{z}_i^{'}$ depends on the parametrization, and therefore we can obtain exactly the same breeding values but different reliabilities in function of the chosen coding.

11 Genomic relationships

11.1 Reminder about relationships

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. The pedigree relationship is *not* equal to 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 $F_k = \theta_{\mathrm{ij}}$. Also, $\theta_{\mathrm{kk}} = (1 + F_k)/2$.
- Additive relationship, or relationship in short, is equal to twice the coancestry: $A_{\mathrm{ij}} = 2\theta_{\mathrm{ij}}$. Also, $A_{\mathrm{kk}} = 1 + F_k$.
- The genetic covariance between two individuals is $\mathrm{Cov}\,(u_i, u_j) = 2\theta_{\mathrm{ij}}\sigma_u^2 = A_{\mathrm{ij}}\sigma_u^2$.

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.

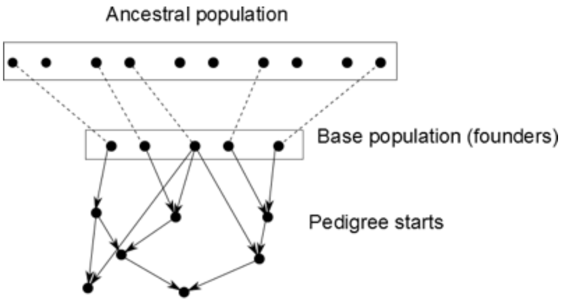

Figure 18: Representation of a pedigree. Continuous lines represent known pedigree links. Dotted lines represent unknown lineages

11.2 Identity by state and identity by descent of two individuals

The probability of Identity by state (IBS), or "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 $r_{\mathrm{Mij}} = 2f_{\mathrm{Mij}}$ (to be on the same scale as A_{ij}). These r_{Mij} are sometimes called "similarity index" but also as "total allelic relationship" (Nejati-Javaremi *et al.* 1997)). For the two-allele case, this is summarised in the following table:

Table 16: 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)

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

Where z_i is coded as $\{0,1,2\}$. 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 ; Li and Horvitz 1953 ; 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 individual i and another allele sampled from individual j. They can be identical because they were identical by descent (with probability $A_{\mathrm{ij}}/2$), or because they were *not* identical by descent (with probability $1 - A_{\mathrm{ij}}/2$) but they were identical just by chance (with probability $p^2 + q^2$). Therefore, $f_{\mathrm{Mij}} = \theta_{\mathrm{ij}} + (1 - \theta_{\mathrm{ij}})\left(p^2 + q^2\right)$ where $\theta_{\mathrm{ij}} = A_{\mathrm{ij}}/2$ is the pedigree coancestry, and

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

also,

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

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

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

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 $\mathbf{G}$ that will be detailed later, up to a constant. Therefore, the results will be identical using IBD or IBS relationships. (We will come back to this later).

11.2.1 Covariance between individuals

What does it mean "covariance between individuals"? The covariance is always computed across several pairs of things:

```
a b
[1,] 1 1
[2,] 2 2
[3,] 3 3
[4,] 4 3
```

Here , $\mathrm{Cov}\,(a, b) = 1.17$

So, how can we define the covariance the genetic value of two individuals $i$and j (for instance bulls ALTACEASAR and BODRUM)? These guys are just two – you can't compute a covariance with just one pair. 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.

11.3 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, $\mu = 2pa$:

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

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

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 2pqσ_a^2. It follows from regular rules of variances and covariances that

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

If we define $z_i^* = z_i - 2p$, in other words, we use the "centered" coding instead of "012", then the covariance between two individuals is equal to $z_i^* z_j^* \sigma_a^2$.

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

Table 17: Relationships r_{Qij} between individuals for a single QTL with $p = 0.5$

	AA	Aa	aa
AA	2	0	-2
Aa	0	0	0
Aa	-2	0	2

Table 18: Relationships r_{Qij}between individuals for a single QTL with $p = 0.25$

	AA	Aa	aa
AA	6	2	2
Aa	2	2/3	-2/3
Aa	-2/3	-2/3	2/3

11.3.1 Negative relationships

Now, this is puzzling because we have negative relationships. 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 either 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.

11.3.2 Centered relationships and IBS relationships

It can be noted that the Table above with $p = 0.5$ is equal to the Table 16 of molecular (or IBS) relationships before, minus a value of 1, times 2:

$$2\left(\begin{bmatrix}2 & 1 & 0\\1 & 1 & 1\\0 & 1 & 2\end{bmatrix}-\begin{bmatrix}1 & 1 & 1\\1 & 1 & 1\\1 & 1 & 1\end{bmatrix}\right)=\begin{pmatrix}2 & 0 & -2\\0 & 0 & 0\\-2 & 0 & 2\end{pmatrix}$$

This shows that relationships at the QTL can be expressed as IBS at the QTL (Nejati-Javaremi *et al.* 1997), and they can be interpreted as twice a probability, as regular relationships in $\mathbf{A}$. The constant value of 1 across all IBS relationships 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.

11.3.3 Inbreeding at a single 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).

11.4 Genomic relationships: Relationships across individuals for many markers

These methods use SNP to infer relationships among individuals, quantifying the number of alleles shared between two individuals. Genomic relationships start from identical by state (IBS) because they consider the fact that two alleles randomly picked from each individual are identical, independently of origin. However, they are later modified to conform to Identity by Descent. Pedigree relationships are identical by descent (IBD) because they consider the shared alleles come from the same ancestor. However, they are also incorrect for two reasons. First, they consider that the genome is infinite, whereas the genome is in fact not infinite, and therefore the pedigree relationships will be correct only on average. Second, the pedigree is not infinite – it is known for a number of generations (the most that I've handled is 40).

11.4.1 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 below:

Table 19: Additive coding for marker effects at locus i with reference allele A.

Genotype	101 Coding	012 Coding	Centered coding
aa	$-a_i$	0	$-2p_ia_i$
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

$$\mathbf{u} = \mathbf{Za}$$

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

$$Var\left(\mathbf{a}\right)=\mathbf{D}$$

With

$$\mathbf{D}=\begin{pmatrix}\sigma_{a1}^2 & 0 & \dots & 0\\ 0 & \sigma_{a2}^2 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & \sigma_{\text{an}}^2\end{pmatrix}$$

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

$$Var\left(\mathbf{u}\right) = \mathbf{Z}Var\left(\mathbf{a}\right)\mathbf{Z}^{'} = \mathbf{ZDZ}^{'} = \mathbf{ZZ}^{'}\sigma_{a0}^2$$

Do <u>not</u> confound $Var\left(\mathbf{u}\right)$ (which is a matrix) with $Var(u)$ (which is a scalar $Var\left(u\right) = E\left(u^2\right) - \overline{E\left(u\right)}^2 = \sigma_u^2$). Elements in $\mathbf{ZZ}^{'}\sigma_{a0}^2$ 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

$$\sigma_u^2 = 2\sum_{i=1}^{\text{nsnp}} p_i q_i \sigma_{a0}^2$$

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

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

When we divide $\mathbf{ZZ}^{'}$ by $\sum p_i(1-p_i)$, $\mathbf{G}$ becomes analogous to the numerator relationship matrix ($\mathbf{A}$). Quoting VanRaden: "The genomic inbreeding coefficient for individual j is simply $G_{\text{jj}} - 1$, and genomic relationships between individuals j and k, which are analogous to the relationship coefficients of Wright (1922), are obtained by dividing elements G_{jk} by square roots of diagonals G_{jj} and G_{kk}." The $\mathbf{G}$ matrix measures the number of homozygous loci for each individual in the diagonals, and it also measures the number of alleles shared among individuals in the off-diagonals. These measures are not IBS – they are IBS modified by the centering due to allelic frequencies, and therefore they become a good approximation of *real* IBD, and this approximation is better than the approximation than we obtain of the pedigree. This is one of the reasons why genomic predictions are better than pedigree prediction.

11.4.2 VanRaden's second genomic relationship matrix

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

$$\mathbf{G} = \mathbf{ZD}_w\mathbf{Z}^{'}$$

with

$$\mathbf{D_w}=\begin{pmatrix} \frac{1}{n\ 2p_1q_1} & 0 & \dots & 0 \\ 0 & \frac{1}{n\ 2p_2q_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{n\ 2p_nq_n} \end{pmatrix}$$

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

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

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

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

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 $\mathbf{G}$”

The second problem is that it assumes that the contribution of each marker to the overall $\mathbf{G}$ 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 $\sigma_{\text{ai}}^2 = \sigma_u^2/n2p_iq_i$. 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 (AL).

11.4.3 Allelic frequencies to put in genomic relationships

There is some confusion on the allelic frequencies to use in the construction of $\mathbf{G}$. (Strandén and Christensen 2011) proved that, if the form is $\mathbf{G}$=$\mathbf{ZZ}^{'}/2\sum p_iq_i$, the allele frequencies used to construct $\mathbf{Z}$ 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 $\sigma_u^2 = 2\sum_{i=1}^{\text{nsnp}} p_iq_i\sigma_{a0}^2$ 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.

11.4.4 Properties of G

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

11.4.4.1 The average value of u is 0 The first property is that the average value of $\mathbf{u}$ is 0, because $\mathbf{Z}$ is centered.

11.4.4.2 The average value of G is 0 The second property is that, the average value of $\mathbf{G}$ is 0. The reason for this is that, by centering the matrix $\mathbf{Z}$, the product $\mathbf{1'Z}$ is equal to a row vector of 0's, as each column of $\mathbf{Z}$ sums to 0 by its centering. And $mean(\mathbf{G}) = \frac{(\mathbf{1'Z})(\mathbf{Z'1})}{m^2 2\sum p_i q_i}$ with m the number of animals.

A related property is that in case of Linkage Equilibrium, terms of $\mathbf{Z}^{'}\mathbf{Z}$ 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:

$$\begin{aligned} E\left(\mathbf{z_i'z_j}\right) = &(1-p_i)(1-p_j)(-p_i)(-p_j) + \\ &(p_i)(1-p_j)(1-p_i)(-p_j) + \\ &(1-p_i)(p_j)(-p_i)(1-p_j) + \\ &(p_i)(p_j)(1-p_i)(1-p_j) = 0 \end{aligned}$$

A verbal explanation is that, if the average value of $\mathbf{u}$ 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.

11.4.4.3 The average value of the diagonal of G is 1 if there is no inbreeding This requires Hardy-Weinberg (but not linkage equilibrium). This can be seen by noting that $\text{tr}\left(\mathbf{ZZ}^{'}\right) = tr\left(\mathbf{Z}^{'}\mathbf{Z}\right)$ where tr is the trace operator. The expression $\text{tr}\left(\mathbf{Z}^{'}\mathbf{Z}\right)$ is the sum of squared covariables corresponding to effects of alleles “a” and “A”, which occur in m animals with respective frequencies $1-p_i$ and p_i in locus i. This is:

$$\mathbf{z}_i^{'}\mathbf{z}_i = 2m\left[(1-p_i)\,p_i^2 + p_i\,(1-p_i)^2\right] = 2mp_i\,(1-p_i) = 2mp_iq_i$$

Therefore, the diagonal of $\mathbf{G}$ has an average of

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

If there is inbreeding there is not Hardy-Weinberg, and there is an inbreeding of F then the genotypes are distributed according to $\{q^2+pqF, 2pq\,(1-F)\,, p^2+pqF\}$ Falconer and Mackay 1996(). Then we multiply each squared value of z by its frequency, and after some algebra (not shown) we arrive to a simple expression:

$$\begin{aligned} \sum \mathbf{z}_i\mathbf{z}_i^{'} =&2m(\\ &(0-2p_i)^2\left(q_i^2+p_iq_iF\right) + \\ &(1-2p_i)^2\left(2p_iq_iF\right) + \\ &(2-2p_i)^2\left(p_i^2+p_iq_iF\right) \\ &) = 2m(1+F)p_iq_i \end{aligned}$$

The diagonal of G has in this case an average of

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

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

11.4.4.4 The average value of the off-diagonal of 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:

$$\text{avoff}\left(\mathbf{G}\right)=\frac{1}{m\left(m-1\right)}\left(\text{sum}\left(\mathbf{G}\right)-diag\left(\mathbf{G}\right)\right)=\frac{m}{m\left(m-1\right)}=\frac{1}{m-1}$$

which is very close to zero.

11.4.5 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

$$Var\left(\mathbf{u}\right)=\mathbf{Z}Var\left(\mathbf{a}\right)\mathbf{Z}^{'}=\mathbf{ZDZ}^{'}$$

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 $Var\left(\mathbf{u}\right)=\mathbf{ZD}_w\ \mathbf{Z}^{'}\sigma_u^2$ with

$$\mathbf{D}_w\!=\!\begin{pmatrix}\sigma^2_{a1}/\sigma^2_{a0} & 0 & \dots & 0\\ 0 & \sigma^2_{a2}/\sigma^2_{a0} & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & \sigma^2_{an}/\sigma^2_{a0}\end{pmatrix}=\begin{pmatrix}w_1 & 0 & \dots & 0\\ 0 & w_2 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & w_n\end{pmatrix}$$

Note that if $w_1=w_2=\ldots=w_n=1$ 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 (VanRaden 2008) to obtain updates for **D**.

11.5 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.

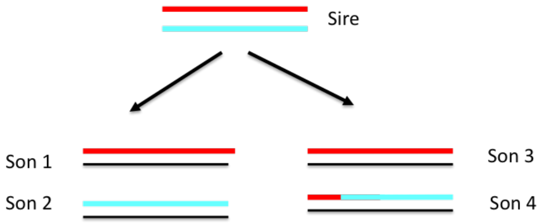

Figure 19: Different transmission of one chromosome from sire to four half-sibs. Different maternal chromosomes are in black.

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. Based on colors, one would say that the relationship of these four sons are something like

$$R = \begin{pmatrix} 1 & 0 & 0.5 & 0.1 \\ 0 & 1 & 0 & 0.4 \\ 0.5 & 0 & 1 & 0.1 \\ 0.1 & 0.4 & 0.1 & 1 \end{pmatrix}$$

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,

$$E\left(R_{\text{ij}}\right) = A_{\text{ij}}$$

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 there will be 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} .

These deviations are skewed and the ratio deviation/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 $\text{Cov}\left(z_i, z_j\right) = R_{\text{ij}}2\text{pq}$ actually involves realized relationships[2]. Then, we can reverse the formulae and estimate those *realized* relationships as $\widehat{R}_{\text{ij}} = \text{Cov}\left(z_i, z_j\right)/2\text{pq}$. For instance: consider three individuals and 20 markers, matrix $\mathbf{Z}$ looks like:

1 1 2 2 1 0 0 1 0 0 2 0 0 2 0 2 2 0 1 1

0 1 2 1 2 1 0 1 2 2 2 2 0 2 1 0 1 0 0 1

2 0 2 0 0 2 1 0 0 0 1 1 0 2 2 1 0 0 0 1

The covariance $\text{Cov}\left(z_i, z_j\right)$ of individuals 1 and 2 is $\text{Cov}\left(\mathbf{z}_1, \mathbf{z}_2\right) = 0.11$. This covariance does not depend on gene coding or allele frequencies. But now, but what 2pq do we divide? We may use the “average” 2pq, in other words, $\frac{2}{m}\sum p_i q_i$ where m is the number of markers. Imagine that frequencies are

0.7 0.31 0.54 0.72 0.83 0.95 0.98 0.84 0.75 0.59 0.37

0.93 0.37 0.79 0.32 0.27 0.14 0.53 0.58 0.78

Then $\frac{2}{m}\sum p_i q_i = 0.354$, and therefore

$$\widehat{R}_{\text{ij}} = \frac{\text{Cov}\left(\mathbf{z}_1, \mathbf{z}_2\right)}{2\sum p_i q_i} = \frac{0.11}{0.354} = 0.31$$

The animals are related (close to “cousins”).

We have just reinvented the wheel. VanRaden’s first $\mathbf{G}$ is:

[2] although we usually assume that “realized” R_{ij} and “expected” A_{ij} are close enough to use $\text{Cov}\left(z_i, z_j\right) \approx A_{\text{ij}}2pq$, for instance to estimate heritability of gene content.

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}^{'}}{2\sum p_k q_k}$$

For two individuals, this is

$$\frac{\mathbf{z}_i\mathbf{z}_j}{2\sum p_k q_k} = \frac{\sum z_{\mathrm{ik}} z_{\mathrm{jk}}}{2\sum p_k q_k}$$

But, when using "centered" coding, then $\sum z_{\mathrm{ik}} z_{\mathrm{jk}}$ is simply $\sum z_{\mathrm{ik}} z_{\mathrm{jk}} = \mathrm{mCov}\left(\mathbf{z}_i, \mathbf{z}_j\right)$. Thus, VanRaden's genomic relationships, 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

$$E\left(\mathbf{G}\right) = \mathbf{R}$$

But also, realized relationships are deviations from expected pedigree relationships, and we have that

$$E\left(\mathbf{R}\right) = \mathbf{A}$$

Therefore

$$E\left(\mathbf{G}\right) = \mathbf{A}$$

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

All these ideas (and many more) were described in depth by (Toro *et al.* 2011 ; Thompson 2013).

11.5.1 Other estimators of (genomic) relationships

We can construct a matrix of r_{Mij}, relationships based on IBS coefficients or "coefficients of similarity" $\mathbf{G}_{\mathrm{IBS}}$. The terms in $\mathbf{G}_{\mathrm{IBS}}$ are usually described in terms of identities or countings:

$\mathbf{G}_{\mathrm{IB}S_{\mathrm{ij}}} = r_{\mathrm{Mij}} = \frac{1}{n}\sum_{m=1}^{n} 2\,\frac{\sum_{k=1}^{2}\sum_{l=1}^{2} I_{\mathrm{kl}}}{4}$,

where I_{kl} measures the identity (with value 1 or 0) of allele k in individual i with allele l in individual j, and single-locus identity measures are averaged across k loci. It has a nice feature: elements in $\mathbf{G}_{\mathrm{IBS}}$ are probabilities (contrary to other $\mathbf{G}$'s)

In the conservation genetics literature, these are usually called molecular relationships (r_{Mij}).

11.5.1.1 Corrected IBS

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:

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

There are many variants of this expression (Lynch 1988 ; Toro *et al.* 2011 ; Ritland 1996) . Extended to several markers (Toro *et al.* 2011):

$$\widehat{R}_{\text{ij}} = \left(\frac{1}{2}\right)\frac{\frac{r_{\text{Mij}}}{2} - \overline{p}^2 - \overline{q}^2 - 2Var(p)}{2\left(\overline{pq} - Var(p)\right)}$$

How do we compute molecular relationships? Consider the following table that compares two genotypes at a time:

Table 20: Molecular coancestry (**bold**) and molecular relationship r_{Mij} (*italic*) comparing two genotypes

	AA		Aa		aa	
AA	**1**	*2*	**0.5**	*1*	**0**	*0*
Aa	**0.5**	*1*	**0.5**	*1*	**0.5**	*1*
aa	**0**	*0*	**0.5**	*1*	**1**	*2*

Values in the left columns table give molecular *coancestries* – the probability that one allele sampled at random from each individual is Identical By State to another allele drawn at sample from the other individual. For instance, two individuals Aa and Aa have a probability of ½ that if we draw one allele from each, the alleles will be identical. You multiply these coancestries by 2 to get molecular relationships r_{Mij} on the right columns.

There is a *much* faster way to get molecular relationships r_{Mij} based on $r_{\text{Mij}} = z_i z_j - z_i - z_j + 2$ where $z_i,\ z_j$ are coded as $\{0, 1, 2\}$. Then it can be shown that the whole array of r_{Mij} can be computed at once as a matrix $\mathbf{G}_{\text{IBS}}$ using (Garcia-Baccino *et al.* 2017)

$$\{r_{\text{Mij}}\} = \mathbf{G}_{\text{IBS}} = \frac{1}{m}\left(\mathbf{Z}_{101}\mathbf{Z}_{101}^{'}\right) + \mathbf{11}'$$

As a crossproduct of $\{0, 1, 2\}$ matrices $\mathbf{Z}_{101}$ (we mean by this using coded as $\{-1, 0, 1\}$ coding), and where $\mathbf{11}'$ is a matrix of 1's. Using the same example as before

1 1 2 2 1 0 0 1 0 0 2 0 0 2 0 2 2 0 1 1

0 1 2 1 2 1 0 1 2 2 2 2 0 2 1 0 1 0 0 1

2 0 2 0 0 2 1 0 0 0 1 1 0 2 2 1 0 0 0 1

with frequencies:

0.7 0.31 0.54 0.72 0.83 0.95 0.98 0.84 0.75 0.59 0.37

0.93 0.37 0.79 0.32 0.27 0.14 0.53 0.58 0.78

Yields $r_{\text{Mij}} = 1.1$ and $\widehat{R}_{\text{ij}} = 2\frac{\frac{r_{\text{Mij}}}{2} - \overline{p}^2 - \overline{q}^2 - 2Var(p)}{2(\overline{pq} - Var(p))} = -0.50$

Quite different from the other estimator (you may check the number). However, if many markers are used, all estimators tend to be very similar.

Values of $\widehat{R}_{\text{ij}}$ 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 ($\widehat{R}_{\text{ij}}$ corrected like that do not form a positive definite covariance matrix).

11.5.1.2 VanRaden with 0.5 allele frequencies One option is to *pretend* that all frequencies are $p_i = 0.5$. Then VanRaden's $\mathbf{G}$ is constructed using $\mathbf{Z}_{101}$ (coded as $\{-1, 0, 1\}$) and dividing by $2\sum p_i q_i = m/2$ with m the number of markers:

$$\mathbf{G}_{05} = \frac{\mathbf{Z}_{101}\mathbf{Z}_{101}^{'}}{m/2} = 2\frac{\mathbf{Z}_{101}\mathbf{Z}_{101}^{'}}{m}$$

In turn, (Garcia-Baccino *et al.* 2017) proved that

$$\mathbf{G}_{\text{IBS}} = \frac{1}{2}\mathbf{G}_{05} + \mathbf{1}\mathbf{1}'$$

So, the $\mathbf{G}_{\text{IBS}}$ is basically a particular case of VanRaden's $\mathbf{G}$.

11.5.2 Genomic inbreeding

From all $\boldsymbol{G}$'s, we have a few estimators of genomic inbreeding. For individual i the genomic inbreeding can be defined as $G_{\text{ii}} - 1$ and it defines its homozygosity with respect to the assumed allelic frequencies. The genomic inbreeding has a few funny properties:

- Genomic inbreeding may be negative. These animals are "more heterozygote than what is expected".
- Genomic inbreeding usually (but not always) correlate well with pedigree inbreeding. Remember, pedigree inbreeding is also an approximation to "true" relationships. In general, if the pedigree is good enough (no missing parents) and the allele frequencies used in $\boldsymbol{G}$ are close to those of the population, the correlation between genomic and pedigree inbreeding is > 0.5.
- Genomic inbreeding of $\mathbf{G}_{\text{IBS}}$ or $\mathbf{G}_{05}$ is directly proportional to *observed* homozygosity of the individual, i.e. the number of markers for which the individual is homozygosity.

11.6 Compatibility of genomic and pedigree relationships

VanRaden's $\mathbf{G}$ 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 $\bar{\mathbf{u}}$=0 across your population. This is contradictory with the pedigree, which imposes $\bar{\mathbf{u}}$=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 $\mathbf{u}$, 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. $\mathbf{u}_{\text{base}}$ are the animals of the genetic base of the pedigree (i.e., the founders). $\mathbf{u}_2$ are genotyped animals, and $\mathbf{u}_1$ are ungenotyped animals.

11.6.1 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).

11.6.2 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 $\mathbf{u}_2$). We know that, by construction of $\mathbf{G}$, the mean of $\mathbf{u}_2$ is set to zero: $p(\mathbf{u}_2) = N(0, \mathbf{G}\sigma_u^2)$. The difference of both means can be modelled as random : $\mu = \overline{\mathbf{u}}_2 - \overline{\mathbf{u}}_{\text{base}} = \overline{\mathbf{u}}_2 = \frac{1}{m}\mathbf{1}'\mathbf{u}_2$ where m is the number of individuals in $\mathbf{u}_2$.

In an infinite population with no selection, there would be no difference between $\overline{\mathbf{u}}_2$ and $\overline{\mathbf{u}}_{\text{base}}$. However, in a finite population there is selection, drift, or both. In this case we can model that $\mathbf{u}_2$ has an a priori mean $p(\mathbf{u}_2|\mu) = N(\mu, \mathbf{G}\sigma_u^2)$. 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 $p(\mathbf{u}_2|\mu)\,p(\mu) = N(\mu, \mathbf{G}\sigma_u^2)N(0, \sigma_\mu^2)$ we have that

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

where $\mathbf{G}^* = (\mathbf{G} + \mathbf{1}\mathbf{1}'a)\sigma_u^2$ 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 $\mathbf{G}$. Informally we may write $\mathbf{G}^* = a + \mathbf{G}$.

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

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

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

Adding constant a as in $\mathbf{G}^*$=$\mathbf{G} + \mathbf{1}\mathbf{1}'a$ 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 $\mathbf{G}$ with current allele frequencies underestimates relationships from the base population. We estimate this underestimation from the average difference between $\mathbf{G}$ and $\mathbf{A}_{22}$. Adding a constant to every element of $\mathbf{G}$ ensures that genomic relationships are, on average, on the same genetic base than pedigree relationships.

11.6.3 Compatibility of genetic variances

In VanRaden's formulation of $\mathbf{G}$=$\mathbf{Z}\mathbf{Z}'/2\sum p_i q_i$, 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

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

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 $\mathbf{u}_2$; I will call this S_{u2}^2 to stress that this is a variance of a particular population, *not* the variance of the genetic base. This is $S_{u2}^2 = \frac{1}{m}\mathbf{u}_2'\mathbf{u}_2 - \overline{\mathbf{u}}_2^2$. This S_{u2}^2 has a certain distribution under either pedigree or genomic modeling. 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:

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

Under genomic relationships we have that:

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

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

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

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

$$b = 1 - \overline{F}_p$$

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, and by an amount (as predicted by the theory) of $1 - \overline{F}_p$. Thus, the multiplication by b corrects for the fixation of alleles due to inbreeding.

11.6.4 Compatibility of genetic bases and variances

With the two pieces above, it is easy to see that a compatible matrix $\mathbf{G}^* = a + b\mathbf{G}$ 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 $a = \overline{\mathbf{A}}_{22}$ 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: $\frac{\mathrm{tr}(\mathbf{G})}{m}b + a = \frac{\mathrm{tr}(\mathbf{A}_{22})}{m}$ and $a + b\overline{\mathbf{G}}=\overline{\mathbf{A}}_{22}$. 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 $a + b\mathbf{G}{=}\mathbf{A}_{22} + \mathbf{E}$. This idea was generalized to several breed origins by (Harris and Johnson 2010). The distribution of $\mathbf{E}$ 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 $\mathbf{E}$ can be very large.

Finally, (Christensen *et al.* 2012) argued that relationships in $\mathbf{G}$ 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 $\mathbf{A}$ to match genomic relationships in $\mathbf{G}$. 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, $\mathbf{A}\left(\gamma\right)$ whose founders have a relationship matrix $\mathbf{A}_{\text{base}} = \gamma + \mathbf{I}(1-\gamma/2)$. Parameter s, used in $\mathbf{G}$=$\mathbf{Z}\mathbf{Z}^{'}/s$ 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.

11.7 Singularity of G

Matrix $\mathbf{G}$ might (and usually is) singular. There are two reasons for this. First, if there are clones or identical twins, two genotypes in $\mathbf{Z}$ will be identical and therefore two animals will show a correlation of exactly 1 in $\mathbf{G}$. Second, if genotypes in $\mathbf{Z}$ 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 $\mathbf{G}$ and then use $\mathbf{G}^{-1}$ in the mixed model equations, there are two ways. The first one is to use a modified $\mathbf{G}_w$= $\left(1-\alpha\right)\mathbf{G} + \alpha\mathbf{I}$,with α a small value (typically 0.05 or 0.01). The second option consists in mixing genomic and pedigree relationships. If $\mathbf{A}_{22}$ is the matrix of genotyped animals, we might use a modified "weighted" $\mathbf{G}_w$= $\left(1-\alpha\right)\mathbf{G} + \alpha\mathbf{A}_{22}$. This is the default in the Blupf90 package, which uses $\alpha = 0.05$. A more detailed explanation is in the next section.

11.8 Including residual polygenic 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.* 2012b)).

Let us decompose the breeding values of genotyped individuals in a part due to markers and a residual part due to pedigree, $\mathbf{u}$=$\mathbf{u}_m$+$\mathbf{u}_p$ with respective variances $\sigma^2_u = \sigma^2_{u,m} + \sigma^2_{u,p}$.

It follows that $Var\left(\mathbf{u}_2\right) = \left(\left(1-\alpha\right)\mathbf{G}+\alpha\mathbf{A}_{22}\right)\sigma^2_u$ where $\alpha = \sigma^2_{u,p}/\sigma^2_u$ is the ratio of pedigree-based variance to total variance.

Therefore, the simplest way to include the residual polygenic effects is to create a modified genomic relationship matrix $\mathbf{G}_w$ ($\mathbf{G}$ in (Aguilar *et al.* 2010); $\mathbf{G}_w$ in (VanRaden 2008 ; Christensen *et al.* 2012) as $\mathbf{G}_w = \left(1-\alpha\right)\mathbf{G}+\alpha\mathbf{A}_{22}$. In practice, the value of α is low and has negligible effects on predictions.

11.9 Multiallelic genomic relationships

In population genetics there are several methods to estimate (pedigree) relationship matrices through markers; these methods were proposed basically in conservation genetics (Ritland 1996; Caballero and Toro 2002). These methods are not very satisfying because they need parameters such as base population allele frequencies that are elusive.

Thus we would be happy if we could extend VanRaden's $\mathbf{G}$ to the multiallelic case. We (Marchal *et al.* 2016) developed it as an extension of the Multiple marker regression model.

Imagine that each allele at each locus produces an effect. We saw such a model in section 9, with an example with 2 loci, 4 alleles in the first and 2 in the second:

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

To center the breeding value we can just center each allele by its allele frequency. So if allele frequencies in this case are, for the first trait, $p_{1,1:4} = (0.1, 0.2, 0.6, 0.1)$ and for the second locus $p_{2,1:2} = (0.3, 0.7)$, each column needs to be substracted by the appropriate allele frequency $2p$ and the matrix becomes

$$\mathbf{Z} = \begin{pmatrix} -0.2 & 0.6 & -0.2 & -0.2 & 1.4 & -1.4 \\ 1.8 & -0.4 & -1.2 & -0.2 & 0.4 & -0.4 \\ -0.2 & 0.6 & -1.2 & 0.8 & -0.6 & 0.6 \end{pmatrix}$$

On the other hand the heterozygosity contributed by each locus is $Het_i = 1 - \sum_j p_j^2$.

As a result, we have two expressions for $\mathbf{G}$, one equivalent to VanRaden 1 :

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{\sum_i Het_i}$$

and another equivalent to VanRaden 2, which assumes that genetic variance explained by each of the n loci is identical:

$$\mathbf{G} = \frac{1}{n} \sum_i \frac{\mathbf{z}_i \mathbf{z}_i'}{Het_i}$$

With our fictious example, the two matrices are as usual not *that* different:

```
G1
  4.4  0.8  -1.2
  0.8  5.2   0.2
 -1.2  0.2   3.2
```

```
 G2
  5.08046  1.05747    -1.58621
  1.05747  4.58785     0.0147783
 -1.58621  0.0147783   2.99507
```

12 GBLUP

12.1 Single trait animal model GBLUP

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

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

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

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

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

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

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

These equations are identical to regular animal model, with the exception that genomic relationships $\mathbf{G}$ 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 $\mathbf{W}$ is set to 0.

3. Regular software (blupf90, asreml, wombat...) works if we include a mechanism to include $\mathbf{G}^{-1}$.

4. Developments including mixed model equations apply to GBLUP as well. Therefore, GREML and G-Gibbs are simple extensions.

12.2 Multiple trait GBLUP

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

$$\begin{pmatrix} \mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}^{'}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}^{'}\mathbf{R}^{-1}\ \mathbf{W}+\mathbf{G}^{-1}\bigotimes\mathbf{G}_0^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}^{'}\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

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

12.3 Reliabilities from GBLUP

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

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

where C^{ii} is the i, i 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 $\mathbf{Z}$ 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. Recently, (Tier *et al.* 0011/2018-02-16) suggested to include the base population as an extra individual, which automatically sets all reliabilities on the same scale.

12.4 All Genomic relationships are equal

We saw before, in Bayesian Regressions, that changing the coding of Z to "centered", 101, 012, 021 gave same results. These results apply, partly, to GBLUP and they can be summarized as follows. Consider the most frequent

$$\frac{\mathbf{ZZ}'}{2\sum p_i q_i}\sigma_u^2 = \mathbf{ZZ}'\frac{\sigma_u^2}{2\sum p_i q_i} = \mathbf{G}\sigma_u^2$$

This matrix $\boldsymbol{G}$ is affected by coding and allele frequencies in two places:

- Centering: how do we define $\mathbf{ZZ}'$ (centering, 101, etc).
- Scaling: what divisor do we put in $2\sum p_i q_i$ *and* what variance do we put in σ_u^2.

All ways of centering give the same $\widehat{\mathbf{u}}$, shifted by a constant, provided that $\frac{\sigma_u^2}{2\sum p_i q_i}$ is constant. For instance:

1. If to construct $\mathbf{G}$ we use *any* coding in $\mathbf{ZZ}'$ but we keep $2\sum p_i q_i$ the same, then $\widehat{\mathbf{u}}$ will be the same
2. If we construct $\mathbf{G}_{05} = \frac{\mathbf{Z}_{101}\mathbf{Z}_{101}'}{m/2}$, but then we use as genetic variance $\frac{\sigma_u^2}{2\sum p_i q_i}\frac{m}{2}$, then $\widehat{\mathbf{u}}$ will be the same
3. If we *estimate* genetic variance by REML, then EBVs will be the same but the estimated genetic variance is not necessarily the same.

These properties of $\mathbf{G}$ do *not* hold for SSGBLUP – we will see that later.

12.5 GBLUP with singular G

If $\mathbf{G}$ is singular, one can use alternative mixed model equations (Harville 1976 ; Henderson 1984):

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

Or a symmetric form that fits better into regular algorithms. First, we predict an auxiliary vector α:

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

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

12.6 Backsolving from GBLUP to marker estimates

Because $\mathbf{G}$ 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 $\mathbf{Z}$, same p's, same variances, etc). This is up to the numerical error produced by forcing $\mathbf{G}$ to be invertible; this numerical error is most often negligible. More formal proofs can be found in (Henderson 1973 ; Strandén and Garrick 2009). We present here how to obtain marker effects.

If breeding values $\mathbf{u}$=$\mathbf{Za}$ and $Var\left(\mathbf{a}\right)=\mathbf{D}$, then the joint distribution of breeding values $\mathbf{u}$ and marker effects $\mathbf{a}$ is (Henderson 1973 ; Strandén and Garrick 2009):

$$Var\begin{pmatrix}\mathbf{u}\\ \mathbf{a}\end{pmatrix}=\begin{pmatrix}\mathbf{ZDZ'} & \mathbf{ZD}\\ \mathbf{DZ}' & \mathbf{D}\end{pmatrix}$$

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

$$p\left(\mathbf{u}|\mathbf{a}\right)=N(\mathbf{Za},\mathbf{0})$$

which means that if marker effects are known, then breeding values are exactly known, and their estimate is simply:

$$\widehat{\mathbf{u}}\ |\widehat{\mathbf{a}}{=}E\left(\mathbf{u}|\mathbf{a}\ {=}\widehat{\mathbf{a}}\right){=}\mathbf{Z}'\widehat{\mathbf{a}}$$

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

However, the opposite is not necessarily true and, conditional on breeding values marker effects can have several possible values:

$$p\left(\mathbf{a}|\mathbf{u}\right)=N\left(\mathbf{DZ}'\left(\mathbf{ZDZ}'\right)^{-1}\mathbf{u},\mathbf{D}-\mathbf{DZ}'\left(\mathbf{ZDZ}'\right)^{\mathbf{-1}}\mathbf{ZD}\right)$$

Thus, the estimate of marker effects conditional on breeding values has a conditional mean. If $\left(\mathbf{ZDZ}'\right){=}\mathbf{G}\sigma^2_u$ then:

$$\widehat{\mathbf{a}}|\widehat{\mathbf{u}}{=}E\left(\mathbf{a}|\mathbf{u}\ {=}\widehat{\mathbf{u}}\right)=\mathbf{DZ}'\left(\mathbf{ZDZ}'\right)^{-1}\ \widehat{\mathbf{u}}{=}\ \mathbf{D}\ \mathbf{Z}'\ \mathbf{G}^{-1}\ \sigma_u^{-2}\ \widehat{\mathbf{u}}$$

or, if, $\mathbf{D}=\mathbf{I}\sigma^2_u/2\Sigma p_i q_i$:

$$\widehat{\mathbf{a}}|\widehat{\mathbf{u}}{=}\frac{1}{2\Sigma p_i q_i}\mathbf{Z}'\ \mathbf{G}^{-1}\ \ \widehat{\mathbf{u}}$$

with associated variance

$$Var\left(\mathbf{a}|\mathbf{u}\ {=}\widehat{\mathbf{u}}\right)=\mathbf{D}-\mathbf{DZ}'\left(\mathbf{ZDZ}'\right)^{\mathbf{-1}}\mathbf{ZD}$$

Note that $\mathbf{D}{-}\mathbf{D}\ \mathbf{Z}'\left(\mathbf{ZDZ}'\right)^{\mathbf{-1}}\ \mathbf{ZD}$ maybe semipositive definite (not invertible) i.e., if two markers are in complete LD. This variance ignores that $\widehat{\mathbf{u}}$ is an estimate.

12.7 Backsolving when matrix G has been "tuned"

In general, the matrix $\mathbf{G}$ has undergone some "tuning" (1) to be invertible, (2) to be on the same scale as pedigree relationships. Usually this is:

$$\mathbf{G}^{\text{tuned}}= (1-\alpha)\left(\mathbf{11}^{'}a + b\mathbf{ZDZ}^{'}\right)+\alpha\mathbf{A}_{22}$$

where the coefficient a adds the extra "average relationship" and *at the same time* models the difference μ from pedigree base to genomic base (Vitezica *et al.* 2011 ; Hsu *et al.* 2017) , the coefficient b considers the reduction in variance due to drift, and α is the part of genetic variance assigned to (pedigree) residual polygenic effects.

In order to extract correctly the marker effects, we need to take that into account. The expression above can be rewritten as:

$$\mathbf{u} = \mathbf{u}_m + \mathbf{u}_p$$

$$\mathbf{u}_m = \mathbf{1}\mu + \mathbf{u}_m^*$$

where $\mathbf{u}_p$ are "pedigree" BVs, $\mathbf{u}_m$ are "marker" breeding values put (shifted) on a pedigree scale (that of $\mathbf{A}$) and $\mathbf{u}_m^*$ are "marker" breeding values put (shifted) on a genomic scale. The respective variances are:

$$Var\left(\mathbf{u}_p\right) = \mathbf{A}_{22}\sigma_u^2\alpha$$

$$Var\left(\mathbf{u}_m^*\right) = b\mathbf{ZDZ}^{'}\left(1-\alpha\right)$$

which implies that, in fact, we have *reduced* the variance of marker effects from an *a priori* variance of $\mathbf{D}$ to another variance of $Var\left(\mathbf{a}\right) = \ b\left(1-\alpha\right)\mathbf{D}$.

Finally,

$$Var\left(\mu\right) = a\left(1-\alpha\right)\sigma_u^2$$

From these elements, we retrieve that $Var\left(\mathbf{u}\right) = Var\left(\mathbf{u}_p + \mathbf{u}_m^* + \mathbf{1}\mu\right) = \mathbf{G}^{\text{tuned}}$

From here we can derive the covariance structure:

$$Var\begin{pmatrix}\mu\\ \mathbf{u}_p\\ \mathbf{u}_m^*\\ \mathbf{u}\\ \mathbf{a}\end{pmatrix} = \begin{pmatrix} a\left(1-\alpha\right)\sigma_u^2 & 0 & 0 & a\left(1-\alpha\right)\mathbf{1}'\sigma_u^2 & 0\\ 0 & \mathbf{A}_{22}\sigma_u^2\alpha & 0 & \mathbf{A}_{22}\sigma_u^2\alpha & 0\\ 0 & 0 & b\left(1-\alpha\right)\mathbf{ZDZ}^{'} & b\left(1-\alpha\right)\mathbf{ZDZ}^{'} & b\left(1-\alpha\right)\mathbf{ZD}\\ a\left(1-\alpha\right)\mathbf{1}\sigma_u^2 & \mathbf{A}_{22}\sigma_u^2\alpha & b\left(1-\alpha\right)\mathbf{ZDZ}^{'} & \left(1-\alpha\right)\left(\mathbf{1}\mathbf{1}^{'}a\sigma_u^2 + b\mathbf{ZDZ}^{'}\right)+\alpha\mathbf{A}_{22}\sigma_u^2 & b\left(1-\alpha\right)\mathbf{ZD}\\ 0 & 0 & b\left(1-\alpha\right)\mathbf{DZ}^{'} & b\left(1-\alpha\right)\mathbf{DZ}^{'} & b\left(1-\alpha\right)\mathbf{D}\end{pmatrix}$$

Under the usual assumption $\mathbf{D} = \mathbf{I}\sigma_u^2/2\Sigma p_i q_i$ we put σ_u^2 as common term:

$$Var\begin{pmatrix}\mu\\ \mathbf{u}_p\\ \mathbf{u}_m^*\\ \mathbf{u}\\ \mathbf{a}\end{pmatrix} = \begin{pmatrix} a\left(1-\alpha\right)\sigma_u^2 & 0 & 0 & a\left(1-\alpha\right)\mathbf{1}^{'}\sigma_u^2 & 0\\ 0 & \mathbf{A}_{22}\sigma_u^2\alpha & 0 & \mathbf{A}_{22}\sigma_u^2\alpha & 0\\ 0 & 0 & b\left(1-\alpha\right)\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i}\sigma_u^2 & b\left(1-\alpha\right)\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i}\sigma_u^2 & b\left(1-\alpha\right)\mathbf{Z}\frac{1}{2\Sigma p_i q_i}\sigma_u^2\\ a\left(1-\alpha\right)\mathbf{1}\sigma_u^2 & \mathbf{A}_{22}\sigma_u^2\alpha & b\left(1-\alpha\right)\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i}\sigma_u^2 & \left(1-\alpha\right)\left(\mathbf{1}\mathbf{1}^{'}a\sigma_u^2 + b\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i}\sigma_u^2\right)+\alpha\mathbf{A}_{22}\sigma_u^2 & b\left(1-\alpha\right)\mathbf{Z}\frac{1}{2\Sigma p_i q_i}\sigma_u^2\\ 0 & 0 & b\left(1-\alpha\right)\mathbf{Z}^{'}\frac{1}{2\Sigma p_i q_i}\sigma_u^2 & b\left(1-\alpha\right)\mathbf{Z}^{'}\frac{1}{2\Sigma p_i q_i}\sigma_u^2 & b\left(1-\alpha\right)\frac{1}{2\Sigma p_i q_i}\mathbf{I}\sigma_u^2\end{pmatrix}$$

$$Var\begin{pmatrix}\mu\\ \mathbf{u}_p\\ \mathbf{u}_m^*\\ \mathbf{u}\\ \mathbf{a}\end{pmatrix}=\begin{pmatrix}a\left(1-\alpha\right) & 0 & 0 & a\left(1-\alpha\right)\mathbf{1}' & 0\\ 0 & \mathbf{A}_{22}\alpha & 0 & \mathbf{A}_{22}\alpha & 0\\ 0 & 0 & b\left(1-\alpha\right)\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i} & b\left(1-\alpha\right)\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i} & b\left(1-\alpha\right)\mathbf{Z}\frac{1}{2\Sigma p_i q_i}\\ a\left(1-\alpha\right)\mathbf{1} & \mathbf{A}_{22}\alpha & b\left(1-\alpha\right)\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i} & \left(1-\alpha\right)\left(\mathbf{11}^{'}a+b\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i}\right)+\alpha\mathbf{A}_{22} & b\left(1-\alpha\right)\mathbf{Z}\frac{1}{2\Sigma p_i q_i}\\ 0 & 0 & b\left(1-\alpha\right)\mathbf{Z}^{'}\frac{1}{2\Sigma p_i q_i} & b\left(1-\alpha\right)\mathbf{Z}^{'}\frac{1}{2\Sigma p_i q_i} & b\left(1-\alpha\right)\mathbf{I}\frac{1}{2\Sigma p_i q_i}\end{pmatrix}\sigma_u^2$$

Let's call $\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i}$=$\mathbf{G}^{\text{untuned}}$. Also, $\mathbf{G}^{\text{tuned}}$$=\left(1-\alpha\right)\left(\mathbf{11}^{'}a+b\mathbf{ZZ}^{'}\frac{1}{2\Sigma p_i q_i}\right)$ $+\alpha\mathbf{A}_{22}$. Thus:

$$Var\begin{pmatrix}\mu\\ \mathbf{u}_p\\ \mathbf{u}_m^*\\ \mathbf{u}\\ \mathbf{a}\end{pmatrix}=\begin{pmatrix}a\left(1-\alpha\right) & 0 & 0 & a\left(1-\alpha\right)\mathbf{1}^{'} & 0\\ 0 & \mathbf{A}_{22}\alpha & 0 & \mathbf{A}_{22}\alpha & 0\\ 0 & 0 & b\left(1-\alpha\right)\mathbf{G}^{\text{untuned}} & b\left(1-\alpha\right)\mathbf{G}^{\text{untuned}} & b\left(1-\alpha\right)\mathbf{Z}\frac{1}{2\Sigma p_i q_i}\\ a\left(1-\alpha\right)\mathbf{1} & \mathbf{A}_{22}\alpha & b\left(1-\alpha\right)\mathbf{G}^{\text{untuned}} & \mathbf{G}^{\text{tuned}} & b\left(1-\alpha\right)\mathbf{Z}\frac{1}{2\Sigma p_i q_i}\\ 0 & 0 & b\left(1-\alpha\right)\mathbf{Z}^{'}\frac{1}{2\Sigma p_i q_i} & b\left(1-\alpha\right)\mathbf{Z}^{'}\frac{1}{2\Sigma p_i q_i} & b\left(1-\alpha\right)\mathbf{I}\frac{1}{2\Sigma p_i q_i}\end{pmatrix}\sigma_u^2$$

And from here, we can get the equations for backsolving (the factor σ_u^2 cancels out):

- Difference between pedigree and genomic bases:

$$\widehat{\mu}|\widehat{\mathbf{u}}=E\left(\mu|\mathbf{u}=\widehat{\mathbf{u}}\right)=a\left(1-\alpha\right)\mathbf{1}'\mathbf{G}^{\text{tuned}^{-1}}\ \widehat{\mathbf{u}}$$

- Pedigree part of estimated breeding values:

$$\widehat{\mathbf{u}}_p|\widehat{\mathbf{u}}=\mathbf{A_{22}}\alpha\mathbf{G}^{\text{tuned}^{-1}}\ \widehat{\mathbf{u}}$$

- "marker" estimated breeding values on the genomic base

$$\widehat{\mathbf{u}}_m^*|\widehat{\mathbf{u}}=b\left(1-\alpha\right)\mathbf{G}^{\text{untuned}}\mathbf{G}^{\text{tuned}^{-1}}\ \widehat{\mathbf{u}}$$

- SNP effects

$$\widehat{\mathbf{a}}|\widehat{\mathbf{u}}=b\left(1-\alpha\right)\mathbf{Z}'\frac{1}{2\Sigma p_i q_i}\mathbf{G}^{\text{tuned}^{-1}}\ \widehat{\mathbf{u}}$$

- indirect predictions from marker effects

$$\widehat{\mathbf{u}}_m^*|\widehat{\mathbf{a}}=\ \mathbf{Z}\widehat{\mathbf{a}}$$

which upon substituting $\widehat{\mathbf{a}}$ is strictly identical to $\widehat{\mathbf{u}}_m^*$ above

- Finally, to get the $\mathbf{u}_m$"marker" estimated breeding values on the pedigree scale:

$$\widehat{\mathbf{u}}_m=\widehat{\mu}+\widehat{\mathbf{u}}_m^*$$

12.8 Backsolving with metafounders

When metafounders are used, there are some differences:

1. matrix $\mathbf{G}$ is not "tuned" so there is no implicit μ
2. allele frequencies to construct $\mathbf{G}$ are assumed to be all $p_i=0.5$
3. Pedigree relationships are $\mathbf{A}_{\Gamma 22}$

Thus the prior covariance of marker effects $\mathbf{D}$ is $\mathbf{D}=\mathbf{I}\frac{1}{2\sum p_i q_i}=\mathbf{I}\frac{2}{m}$ with m number of markers and $\mathbf{G}=\mathbf{ZDZ'}=\frac{2}{m}\mathbf{ZZ'}$ with $\mathbf{Z}$ coded as {-1,0,1} for the three genotypes.

In fact, the role of difference between the genomic base and the pedigree base is played by the metafounder solution, which gives the difference between the genetic level of an ideal genotyped population with $p=0.5$ and the animals at the base population represented by the pedigree.

However, there may still be some “blending” as

$$\mathbf{G}^{\text{blended}} = (1-\alpha)\,\mathbf{G} + \alpha\mathbf{A}_{\Gamma 22} = (1-\alpha)\left(\frac{2}{m}\mathbf{Z}\mathbf{Z}'\right) + \alpha\mathbf{A}_{\Gamma 22}$$

and the equations are like above but with $a = 0$,$b = 1$, and there is *no* difference between pedigree and genomic bases:

- Pedigree part of estimated breeding values:

$$\widehat{\mathbf{u}}_p|\widehat{\mathbf{u}} {=} \alpha\mathbf{A}_{\Gamma 22}\mathbf{G}^{\text{blended}^{-1}}\ \widehat{\mathbf{u}}$$

- “marker” estimated breeding values

$$\widehat{\mathbf{u}}_m|\widehat{\mathbf{u}} {=}\ (1-\alpha)\left(\frac{2}{m}\mathbf{Z}\mathbf{Z}'\right)\mathbf{G}^{\text{blended}^{-1}}\ \widehat{\mathbf{u}}$$

Note that if you sum $\hat{\mathbf{u}}_p$ and $\hat{\mathbf{u}}_m$ you get $\hat{\mathbf{u}}$.

- SNP effects

$$\widehat{\mathbf{a}}|\widehat{\mathbf{u}} {=}\ (1-\alpha)\,\mathbf{Z}'\frac{2}{m}\mathbf{G}^{\text{blended}^{-1}}\ \widehat{\mathbf{u}}$$

- indirect predictions from marker effects

$$\widehat{\mathbf{u}}_m|\widehat{\mathbf{a}} {=}\ \mathbf{Z}\widehat{\mathbf{a}}$$

12.9 Indirect predictions using marker effects

Imagine that we need to make indirect predictions (e.g. once a week) and we want to use estimates of SNP effects without going through the whole process of running GBLUP (or ssGBLUP). We genotype and we obtain a matrix with genotypes $\mathbf{Z}_{new}$. This matrix needs to be centered by the same allelic frequencies as the original GBLUP or ssGBLUP[3].

For indirect predictions of newborn animals, there are two parts: $\mathbf{u}{=}\mathbf{u}_m + \mathbf{u}_p$. The first part is obtained as a sum of marker effects, plus the difference between genomic and pedigree bases μ, (**which is 0** in the case of using metafounders):

$$\widehat{\mathbf{u}}_m = \widehat{\mu} + \widehat{\mathbf{u}}^*_m {=} \widehat{\mu} + \mathbf{Z}_{new}\widehat{\mathbf{a}}$$

whereas the pedigree part of indirect predictions has to be obtained as the parent average of the parents’ pedigree part, *not* of the complete breeding value). Note that this parent average only accounts for a small α part of the genetic variance. This is, for individual i

$$\widehat{u}_{p,i} = 0.5\left(\widehat{u}_{p,sire(i)} + \widehat{u}_{p,dam(i)}\right)$$

see

If the animal has no records indirect predictions are the same as GBLUP predictions. In SSGBLUP (that we have not described yet), they are the same if both parents are genotyped and almost the same if not. The reason is that $\mathbf{H}$ matrix is slightly different for the ungenotyped parents if the genotyped animal is included in the SSGBLUP or not.

In the case of ssGBLUP the expressions for indirect predictions of the pedigree-based part are more complicated and can be found in (Vandenplas *et al.* 2023).

[3] another reason to use metafounders with allelic frequencies at 0.5

12.10 Reliabilities of indirect predictions with “tuned G”

1. Note: this section will need reordering
2. In this section all matrices **G** are what we call “tuned” unless otherwise staten

We may consider the reliability of u_m (breeding value referred to the base of the pedigree) or the reliability of u_m^* (breeding value referred to the genomic population). These two acuracies are for a single individual[4]:

- $Rel_{base-pedigree} = 1 - \frac{PEV(\hat{u}_m)}{Var(u_m)} = \frac{Var(\hat{u}_m)}{Var(u_m)}$
- $Rel_{base-genomic} = 1 - \frac{PEV(\hat{u}_m^*)}{Var(u_m^*)} = \frac{Var(\hat{u}_m^*)}{Var(u_m^*)}$

The difference between the two is the difference between genomic and pedigree bases, μ, that although it is not explicitely computed in ssGBLUP, it is there. And in fact, this μ is estimated with some uncertainty.

For instance, for a single individual we have that

$$Var(u_m^*) = b(1-\alpha)G_{ii}^{untuned}\sigma_u^2$$

$$Var(u_m) = Var(\mu + u_m^*) = a(1-\alpha)\sigma_u^2 + b(1-\alpha)G_{ii}^{untuned}\sigma_u^2 = (1-\alpha)G_{ii}\sigma_u^2$$

where, for the purposes of indirect prediction, it may be easier to build as (for individual j) $G_{jj}^{untuned} = \frac{\mathbf{z}_j^{'}\mathbf{z}_j}{2\sum p_i q_i}$.

Then we need the $Var(\hat{u}_m^*)$. This can be obtained as a function of $\hat{\mathbf{u}}_m^* = \mathbf{Z}_{new}\widehat{\mathbf{a}}$ such that

$$Var(\hat{\mathbf{u}}_m^*) = \mathbf{Z}_{new}Var(\widehat{\mathbf{a}})\mathbf{Z}_{new}^{'}$$

(note that **Z** is the matrix for animals in the (SS)GBLUP evaluation, whereas $\mathbf{Z}_{new}$ is for animals in indirect prediction) with

$$Var(\widehat{\mathbf{a}}) = (1-\alpha)b\frac{1}{2\Sigma p_i q_i}\mathbf{Z}^{'}\mathbf{G}^{-\mathbf{1}}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{\mathrm{uu}})\mathbf{G}^{-\mathbf{1}}\mathbf{Z}\frac{1}{2\Sigma p_i q_i}(1-\alpha)b$$

which gives the alternative (and not necessarily better) expression

$$Var(\hat{\mathbf{u}}_m^*) = (1-\alpha)b\frac{1}{2\Sigma p_i q_i}\mathbf{Z}_{new}\mathbf{Z}^{'}\mathbf{G}^{-\mathbf{1}}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{\mathrm{uu}})\mathbf{G}^{-\mathbf{1}}\mathbf{Z}\mathbf{Z}_{new}^{'}\frac{1}{2\Sigma p_i q_i}(1-\alpha)b =$$

$$= (1-\alpha)b\mathbf{G}_{new,old}^{untuned}\mathbf{G}^{-\mathbf{1}}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{\mathrm{uu}})\mathbf{G}^{-\mathbf{1}}\mathbf{G}_{old,new}^{untuned}(1-\alpha)b$$

for $\mathbf{G}_{new,old}^{untuned} = \frac{\mathbf{Z}_{new}\mathbf{Z}^{'}}{2\Sigma p_i q_i}$ which makes sense as the variance of a selection index.

A bit trickier is to obtain the scalar $Var(\hat{u}_m) = Var(\hat{\mu} + \hat{u}_m^*)$

$$Var(\hat{u}_m) = Var(\hat{\mu} + \hat{u}_m^*) = Var(\hat{\mu}) + Var(\hat{u}_m^*) + 2Cov(\hat{\mu}, \hat{u}_m^*)$$

or, in matrix form,

$$Var(\hat{\mathbf{u}}_\mathbf{m}) = Var(\mathbf{1}\hat{\mu} + \hat{\mathbf{u}}_m^*) =$$

[4] We made use of the Hendersonian identity $Var(\hat{x}) = Var(x) - PEV(x)$. This holds because $PEV(x) = Var(x - \hat{x}) = Var(x) - 2Cov(x, \hat{x}) + Var(\hat{x}) = Var(x) - Var(\hat{x})$ because $Cov(x, \hat{x}) = Var(\hat{x})$ if “Best” properties of BLUP hold

$$= \mathbf{1}\mathbf{1}'Var(\hat{\mu}) + \mathbf{1}Cov(\hat{\mu}, \hat{\mathbf{u}}_m^{'*}) + Cov(\hat{\mathbf{u}}_m^{*}, \hat{\mu})\mathbf{1}^{'} + Var(\hat{\mathbf{u}}_m^{*})$$

With $Var(\hat{\mathbf{u}}_m^{*})$ that we already saw. The $Var(\hat{\mu})$ is as follows.

$$Var(\hat{\mu}) = a(1-\alpha)\mathbf{1}'\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{uu})\mathbf{G}^{-1}\mathbf{1}a(1-\alpha)$$

which is equal to $a^2(1-\alpha)^2$ times the sum of elements of $\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{uu})\mathbf{G}^{-1}$. This is not difficult.

Then we need $Cov(\hat{\mu}, \hat{\mathbf{u}}_m^{*'})$. This is a row vector (which multiplied by $\mathbf{1}$ gives a matrix) whose interpretation is “how much does the uncertainty in μ affects each animal” (i.e. because its genomic information is poor). Anyway, this is

$$\begin{aligned} Cov(\hat{\mu}, \hat{\mathbf{u}}_m^{*'}) &= a(1-\alpha)\mathbf{1}'\mathbf{G}^{-1}Var(\hat{\mathbf{u}})\mathbf{Z}\mathbf{Z}_{new}^{'}\frac{1}{2\Sigma p_i q_i}b(1-\alpha) \\ &= a(1-\alpha)\mathbf{1}'\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{uu})\mathbf{Z}\mathbf{Z}_{new}^{'}\frac{1}{2\Sigma p_i q_i}b(1-\alpha) \end{aligned}$$

which is rather cumbersome and where it appears $\mathbf{G}_{old,new}^{untuned} = \frac{\mathbf{Z}\mathbf{Z}_{new}^{'}}{2\Sigma p_i q_i}$ which describes how close are new to old animals.[5]

It is easier to work with the joint distribution of $\hat{\mu}, \hat{\mathbf{a}}$ as follows:

$$Var\begin{pmatrix}\hat{\mu}\\ \hat{\mathbf{a}}\end{pmatrix} = \begin{bmatrix} a(1-\alpha)\mathbf{1} \\ b(1-\alpha)\frac{1}{2\sum p_i q_i}\mathbf{Z}' \end{bmatrix} \mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{uu})\mathbf{G}^{-1} \begin{bmatrix} a(1-\alpha)\mathbf{1}' & b(1-\alpha)\frac{1}{2\sum p_i q_i}\mathbf{Z} \end{bmatrix}$$

which expanded gives:

$$Var\begin{pmatrix}\hat{\mu}\\ \hat{\mathbf{a}}\end{pmatrix} = \begin{pmatrix} a(1-\alpha)\mathbf{1}\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{uu})\mathbf{G}^{-1}\mathbf{1}'(1-\alpha)a & a(1-\alpha)\mathbf{1}\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{uu})\mathbf{G}^{-1}\mathbf{Z}\frac{1}{2\sum p_i q_i}(1-\alpha)b \\ b(1-\alpha)\frac{1}{2\sum p_i q_i}\mathbf{Z}'\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{uu})\mathbf{G}^{-1}\mathbf{1}'(1-\alpha)a & b(1-\alpha)\frac{1}{2\sum p_i q_i}\mathbf{Z}'\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{uu})\mathbf{G}^{-1}\mathbf{Z}\frac{1}{2\sum p_i q_i}(1-\alpha)b \end{pmatrix}$$

Consider now a single individual with row vector of genotypes $\mathbf{z}_{new}$. Its total breeding value through indirect prediction (including μ) is $\hat{u}_m = \begin{pmatrix}1 & \mathbf{z}_{new}\end{pmatrix}\begin{pmatrix}\hat{\mu}\\ \hat{\mathbf{a}}\end{pmatrix}$ and thus the final expression for PEV of the indirect prediction, using $Var\begin{pmatrix}\hat{\mu}\\ \hat{\mathbf{a}}\end{pmatrix}$ above, is:

$$Var(\hat{u}_m) = \begin{pmatrix}1 & \mathbf{z}_{new}\end{pmatrix} Var\begin{pmatrix}\hat{\mu}\\ \hat{\mathbf{a}}\end{pmatrix}\begin{pmatrix}1\\ \mathbf{z}_{new}'\end{pmatrix}$$

and the final expression for Reliability is

$$Rel_{base-pedigree} = \frac{Var(\hat{u}_m)}{Var(u_m)}$$

with $Var(\hat{u}_m)$ as above and $Var(u_m) = (1-\alpha)G_{new,new}\sigma_u^2 = (1-\alpha)(a + b\frac{1}{\sum 2p_i q_i}\mathbf{z}_{new}\mathbf{z}_{new}')\sigma_u^2$

[5] I (AL) believe this to be a small quantity except in weird cases (a Jersey cow in a Holstein evaluation), the reason is that $\mathbf{G}_{old,new}^{untuned}$ generally has an average close to 0. But this should be checked.

12.11 Reliabilities of indirect predictions with metafounders

Note: here $\mathbf{G}$ is $\mathbf{G}^{blended}$ (but not "tuned" because in metafounders tuning is not needed).

The work of (Bermann *et al.* 2023) proposed a coherent framework to define reliabilities when there are several base populations (metafounders), as contrasts from a particular reference metafounder (mf). We try to stick to their notation, but to indicate that we build $\mathbf{G}$ with 0.5 allele frequencies we subscript $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{Z}$ with 05, and the genetic variance re-escaled for metafounders is $\sigma^2_{u,mf}$. Here we derive the reliability of the indirect prediction with metafounders, expressed as the reliability of the contrast. Let $\hat{u}_m = \mathbf{z}_{new}\hat{a}$, contrasted with the reference metafounder, e.g. $\mathbf{u}_m - \mathbf{u}_{mf}$. We will call this contrast $u_c = u_m - u_{mf}$ Reliability is then the squared correlation between the *true* $u_c = u_m - u_{mf}$ and the estimated $\hat{u}_c = \hat{u}_m - \hat{u}_{mf}$. Then

$$Rel_c = \frac{Var(\hat{u}_c)}{Var(u_c)}$$

To obtain $Var(\hat{u}_c)$ we express

$$\hat{u}_c = \begin{pmatrix} -1 & \mathbf{z}_{05new} \end{pmatrix} \begin{pmatrix} \hat{u}_{mf} \\ \hat{\mathbf{a}} \end{pmatrix}$$

First we need to consider the block of the reference metafounder plus genotyped individuals $\mathbf{u}_g$:

$$Var \begin{pmatrix} \hat{u}_{mf} \\ \hat{\mathbf{u}}_g \end{pmatrix} = \mathbf{H}_{05block}\sigma^2_{u,mf} - \mathbf{C^{block}}$$

with $\mathbf{H}_{05block} = \begin{pmatrix} h_{05,mf,mf} & \mathbf{h}_{05,mf,g} \\ \mathbf{h}_{05,g,mf} & \mathbf{G}_{05} \end{pmatrix}$ (because $\mathbf{H}_{05,g,g} = \mathbf{G}_{05}$) and $\mathbf{C^{block}} = \begin{pmatrix} c^{22}_{mf,mf} & \mathbf{c}^{22}_{mf,g} \\ \mathbf{c}^{22}_{g,mf} & \mathbf{C}^{22}_{g,g} \end{pmatrix}$

Note that

$$\begin{pmatrix} \hat{u}_{mf} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & (1-\alpha)\frac{2}{m}\mathbf{Z}_{05}\mathbf{G}_{05}^{-1} \end{pmatrix} \begin{pmatrix} \hat{u}_{mf} \\ \hat{\mathbf{u}}_g \end{pmatrix}$$

from which

$$Var \begin{pmatrix} \hat{u}_{mf} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & (1-\alpha)\frac{2}{m}\mathbf{Z}_{05}\mathbf{G}_{05}^{-1} \end{pmatrix} \left[\begin{pmatrix} h_{05,mf,mf}\sigma^2_{u,mf} & \mathbf{h}_{05,mf,g}\sigma^2_{u,mf} \\ \mathbf{h}_{05,g,mf}\sigma^2_{u,mf} & \mathbf{G}_{05}\sigma^2_{u,mf} \end{pmatrix} - \begin{pmatrix} c^{22}_{mf,mf} & \mathbf{c}^{22}_{mf,g} \\ \mathbf{c}^{22}_{g,mf} & \mathbf{C}^{22}_{g,g} \end{pmatrix} \right] \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{G}_{05}^{-1}\mathbf{Z}'_{05}\frac{2}{m}(1-\alpha) \end{pmatrix}$$
$$= \begin{bmatrix} (h_{05,mf,mf}\sigma^2_{u,mf} - c^{22}_{mf,mf}) & (\mathbf{h}_{05,mf,g}\sigma^2_{u,mf} - \mathbf{c}^{22}_{mf,g})\mathbf{G}_{05}^{-1}\mathbf{Z}'_{05}\frac{2}{m}(1-\alpha) \\ (1-\alpha)\frac{2}{m}\mathbf{Z}_{05}\mathbf{G}_{05}^{-1}(\mathbf{h}_{05,g,mf}\sigma^2_{u,mf} - \mathbf{c}^{22}_{g,mf}) & (1-\alpha)\frac{2}{m}\mathbf{Z}_{05}\mathbf{G}_{05}^{-1}(\mathbf{G}_{05}\sigma^2_{u,mf} - \mathbf{C}^{22}_{g,g})\mathbf{G}_{05}^{-1}\mathbf{Z}'_{05}\frac{2}{m}(1-\alpha) \end{bmatrix}$$

Finally, we create the quadratic form inserting the above expression:

$$Var(\hat{u}_c) = \begin{pmatrix} -1 & \mathbf{z}_{05new} \end{pmatrix} Var \begin{pmatrix} \hat{u}_{mf} \\ \hat{\mathbf{a}} \end{pmatrix} \begin{pmatrix} -1 \\ \mathbf{z}'_{05new} \end{pmatrix}$$

The following step is deriving $Var(u_c)$. This would require building the H_Γ matrix including all previous individuals plus the new one, to build

$$Var(u_c) = (H_{\Gamma,mf,mf} - 2H_{\Gamma,mf,new} + H_{\Gamma,new,new})\sigma^2_{u,mf}$$

The element $H_{\Gamma,mf,mf}$ can be stored from the run. The element $H_{\Gamma,new,new} = \frac{2}{m}\mathbf{z}_{new}\mathbf{z}'_{new}$. As for $H_{\Gamma,mf,new}$, this element seems hard to obtain because it involves in principle all genotyped animals plus the new one.

Another practical solution is to work not with H but with A assuming the following:

$$Var(u_c) \approx (A_{\Gamma,mf,mf} - 2A_{\Gamma,mf,new} + A_{\Gamma,new,new})\sigma^2_{u,mf}$$

where $A_{\Gamma,new,new} = 1 + F_{\Gamma,new}$ can be obtained with an inbreeding algorithm, $A_{\Gamma,mf,mf} = \Gamma_{mf,mf}$ and $A_{\Gamma,mf,new} = \mathbf{q}_{new}\Gamma_{:,mf}$ i.e. a vector of metafounders proportions in *new*, times the mf column in Γ.

The next possibility is probably better and simpler.

12.11.1 Reliability with assumed allele frequencies for the reference metafounder

Another possibility is to use $u_{mf} = (2\mathbf{p}_{mf} - \mathbf{1})\mathbf{a}$ and derive both $Var(\hat{u}_c)$ and $Var(u_c)$. This assumes (perfect) knowledge of $\mathbf{p}_{mf}$, row vector of allele frequencies, which may be *estimated* by Gengler's method (equivalently, GLS) or another method. Note that if allele frequencies of the metafounder were perfectly known, then $H_{\Gamma,mf,mf} = \frac{2}{m}(2\mathbf{p}_{mf} - \mathbf{1})(2\mathbf{p}_{mf} - \mathbf{1})'$ and $H_{\Gamma,mf,new} = \frac{2}{m}(2\mathbf{p}_{mf} - \mathbf{1})\mathbf{z}'_{05new}$.

Then:

$$\begin{aligned}\hat{u}_c &= \begin{pmatrix} -1 & \mathbf{z}_{05new} \end{pmatrix} \begin{pmatrix} \hat{u}_{mf} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} -1 & \mathbf{z}_{05new} \end{pmatrix} \begin{pmatrix} (2\mathbf{p}_{mf} - \mathbf{1})\hat{\mathbf{a}} \\ \hat{\mathbf{a}} \end{pmatrix} \\ &= \begin{pmatrix} -1 & \mathbf{z}_{05new} \end{pmatrix} \begin{pmatrix} \hat{u}_{mf} \\ \hat{\mathbf{a}} \end{pmatrix} = (\mathbf{m}_{new} - 2\mathbf{p})\hat{\mathbf{a}}\end{aligned}$$

with $\mathbf{m}_{new}$ coded as $0, 1, 2$. In fact, $(\mathbf{m}_{new} - 2\mathbf{p})$ is simply the original (VanRaden 2008) coding, with associated

$$Var(\hat{u}_c) = (\mathbf{m}_{new} - 2\mathbf{p})Var(\hat{\mathbf{a}})(\mathbf{m}_{new} - 2\mathbf{p})'$$

and in the denominator we have

$$Var(u_c) = \frac{2}{m}(\mathbf{m}_{new} - 2\mathbf{p})(\mathbf{m}_{new} - 2\mathbf{p})'\sigma^2_{u,mf}$$

which is the relationship of the new individual times the genetic variance. It can be probably shown that for a single metafounder and base allele frequencies $\mathbf{p}$, this is identical to the regular expression involving no metafounders.

The final expression is

$$Rel_c = \frac{Var(\hat{u}_c)}{Var(u_c)} = \frac{(\mathbf{m}_{new} - 2\mathbf{p})Var(\hat{\mathbf{a}})(\mathbf{m}_{new} - 2\mathbf{p})'}{\frac{2}{m}(\mathbf{m}_{new} - 2\mathbf{p})(\mathbf{m}_{new} - 2\mathbf{p})'\sigma^2_{u,mf}}$$

where $Var(\hat{\mathbf{a}})$ was shown above.

If instead of the reference metafounder p's, we use "current" allele frequencies p, the reliability obtained will refer to the current population.

12.12 Bayesian distribution of marker effects from GBLUP

Imagine now that, from GBLUP (or GGibbs…), we obtain the posterior distribution of $\mathbf{u}$, i.e. from inversion of the Mixed Model Equations or from MonteCarlo, as:

$$p(\mathbf{u}|\mathbf{y}) = N(\hat{\mathbf{u}}, \mathbf{C}^{\mathrm{uu}})$$

where $Var(\mathbf{u}|\mathbf{y}) = \mathbf{C}^{\mathrm{uu}}$ is the posterior covariance matrix (I am using Hendersonian notation here). To derive the posterior distribution of marker effects, we multiply the conditional distribution

$p\left(\mathbf{a}|\mathbf{u}\right)$ by the posterior distribution $p\left(\mathbf{u}|\mathbf{y}\right)$. This has two parts, first to account for the incertitude in $\widehat{\mathbf{u}}$ contained in $\mathbf{C}^{\mathrm{uu}}$ as:

$$Var\left(\widehat{\mathbf{a}}|\mathbf{y}\right)=Var\left(\widehat{\mathbf{a}}|Var\left(\mathbf{u}|\mathbf{y}\right)\right)$$

$$=Var\left(\mathbf{DZ}^{'}\left(\mathbf{ZDZ}^{'}\right)^{-1}\ \left(\mathbf{u}-\widehat{\mathbf{u}}\right)\right)=\mathbf{DZ}^{'}\left(\mathbf{ZDZ}^{'}\right)^{-1}\mathbf{C}^{\mathrm{uu}}\left(\mathbf{ZDZ}^{'}\right)^{-1}\mathbf{ZD}$$

and second, for the remaining noise

$$Var(\mathbf{a}\text{–}\widehat{\mathbf{a}}|\mathbf{y})=\mathbf{D}-\mathbf{DZ}^{'}\left(\mathbf{ZDZ}^{'}\right)^{\mathbf{-1}}\mathbf{ZD}$$

which gives

$$Var\left(\mathbf{a}|\mathbf{y}\right)=\mathbf{D}-\mathbf{DZ}^{'}\left(\mathbf{ZDZ}^{'}\right)^{\mathbf{-1}}\ \mathbf{ZD}+\mathbf{DZ}^{'}\left(\mathbf{ZDZ}^{'}\right)^{-1}\mathbf{C}^{\mathrm{uu}}\left(\mathbf{ZDZ}^{'}\right)^{-1}\mathbf{ZD}$$

which is a *Bayesian* distribution like the one obtained, e.g. using Gibbs Sampling as in SNP-BLUP.

Putting $\left(\mathbf{ZDZ}^{'}\right)$ $=\mathbf{G}\sigma_u^2$ and reordering yields

$$Var\left(\mathbf{a}|\mathbf{y}\right)=\mathbf{D}+\mathbf{DZ}^{'}\mathbf{G}^{\mathbf{-1}}\sigma_u^{-2}(\mathbf{C}^{\mathrm{uu}}-\mathbf{G}\sigma_u^2)\mathbf{G}^{\mathbf{-1}}\sigma_u^{-2}\mathbf{ZD}$$

Perhaps is more enlightening to consider the alternative expression

$$Var\left(\mathbf{a}|\mathbf{y}\right)=\mathbf{D}-\mathbf{DZ}^{'}\mathbf{G}^{\mathbf{-1}}\sigma_u^{-2}(\mathbf{G}\sigma_u^2-\mathbf{C}^{\mathrm{uu}})\mathbf{G}^{\mathbf{-1}}\sigma_u^{-2}\mathbf{ZD}$$

which is composed of two terms: first, the *a priori* variance of marker effects, $\mathbf{D}$. Second, an *a posteriori* reduction, given by the data, in their incertitude. This reduction comes from a reduction in the incertitude of the genomic breeding values $(\mathbf{G}\sigma_u^2-\mathbf{C}^{\mathrm{uu}})$ which is in turn transferred to the marker effects *via* the linear operator $\mathbf{DZ}^{'}$**.**

If $\mathbf{C}^{\mathrm{uu}}\approx\mathbf{0}$ (well known animals such as progeny tested bulls) this yields $Var\left(\mathbf{a}|\mathbf{y}\right)$ $\mathbf{D}\text{–}\mathbf{DZ}^{'}\mathbf{G}^{\mathbf{-1}}\sigma_u^2\mathbf{ZD}$.

If $\mathbf{D}=\mathbf{I}\sigma_u^2/2\Sigma p_iq_i$ (the usual assumption where, as discussed in previous sections, $\mathbf{ZDZ}^{'}$$=\mathbf{G}\sigma_u^2$) the expressions above become

The estimate of the marker effects is

$$\mathbf{E}\left(\mathbf{a}|\mathbf{y}\right)=\widehat{\mathbf{a}}|\widehat{\mathbf{u}}=\mathbf{DZ}^{'}\left(\mathbf{ZDZ}^{'}\right)^{-1}\ \widehat{\mathbf{u}}=\frac{1}{2\Sigma p_iq_i}\mathbf{Z}^{'}\ \mathbf{G}^{-1}\ \widehat{\mathbf{u}}$$

with covariance matrix

$$Var\left(\mathbf{a}|\mathbf{y}\right)=\frac{\sigma_u^2}{2\Sigma p_iq_i}\mathbf{I}\text{–}\frac{1}{2\Sigma p_iq_i}\mathbf{Z}^{'}\mathbf{G}^{\mathbf{-1}}(\mathbf{G}\sigma_u^2-\mathbf{C}^{\mathrm{uu}})\mathbf{G}^{\mathbf{-1}}\mathbf{Z}\frac{1}{2\Sigma p_iq_i}$$

So that the full distribution of marker effects can be deduced from breeding values by backsolving using the genomic relationship matrix and markers' incidence matrix.

12.13 Frequentist distribution of marker effects from GBLUP.

This is described in (Gualdrón Duarte *et al.* 2014). The distribution of interest is $Var\left(\widehat{\mathbf{a}}\right)$, the *frequentist* variance of the estimators integrating over the conceptual distribution of all possible $\mathbf{y}$'s. Using results in Henderson (Henderson 1975 ; Henderson 1984) we get that, similar (but not identical) as above:

$$Var(\widehat{\mathbf{a}}) = \mathbf{DZ}^{'}\left(\mathbf{ZDZ}^{'}\right)^{-1}\left(\mathbf{ZDZ}^{'}-\mathbf{C}^{\mathrm{uu}}\right)\left(\mathbf{ZDZ}^{'}\right)^{-1}\mathbf{ZD}_{\mathbf{a}}$$

or

$$Var(\widehat{\mathbf{a}}) = \frac{1}{2\Sigma p_i q_i}\mathbf{Z}^{'}\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{\mathrm{uu}})\mathbf{G}^{-1}\mathbf{Z}\frac{1}{2\Sigma p_i q_i}$$

The difference is the term $\mathbf{D}$ because instead of computing $Var\left(\mathbf{a}|\mathbf{y}\right)$ they compute $Var\left(\widehat{\mathbf{a}}\right)$. In fact, $Var\left(\widehat{\mathbf{a}}\right) = Var\left(\mathbf{a}\right) - Var\left(\mathbf{a}|\mathbf{y}\right)$.

When matrix G has been "tuned", the expression is (Aguilar *et al.* 2019):

$$Var(\widehat{\mathbf{a}}) = (1-\alpha)b\frac{1}{2\Sigma p_i q_i}\mathbf{Z}^{'}\mathbf{G}^{-1}(\mathbf{G}\sigma_u^2 - \mathbf{C}^{\mathrm{uu}})\mathbf{G}^{-1}\mathbf{Z}\frac{1}{2\Sigma p_i q_i}(1-\alpha)b$$

12.13.1 Example of marker predictions from GBLUP

Let there be two individuals and three markers and $\sigma_u^2 = 1$:

$\mathbf{Z}=\begin{pmatrix}1 & 1 & 1\\ 0 & 1 & 0\end{pmatrix}$ and $\mathbf{D}=\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}$, both $\mathbf{ZDZ}^{'}$ and $\mathbf{D}$ are positive definite, however $\mathbf{D}-\mathbf{D}\ \mathbf{Z}^{'}\left(\mathbf{ZDZ}^{'}\right)^{-\mathbf{1}}\mathbf{ZD}$ is not full rank. The reason is complete LD between markers 1 and 3. Therefore for a given value of $\mathbf{u}$, there will be infinite possible combinations. Say that $\widehat{\mathbf{u}}=\begin{pmatrix}1\\ -2\end{pmatrix}$. Then there are many possible solutions of $\mathbf{a}$ yielding these $\widehat{\mathbf{u}}$, for instance $\begin{pmatrix}0 & -2 & 3\end{pmatrix}$ or $\begin{pmatrix}-3 & -2 & 6\end{pmatrix}$. However, $\mathbf{a}$ has an *a priori* structure $\mathbf{D}$ that makes that the effects of the first and third SNP have *a priori* the same size, thus the most likely solution will be $\widehat{\mathbf{a}}=\mathbf{DZ}^{'}\left(\mathbf{ZDZ}^{'}\right)^{-1}\widehat{\mathbf{u}}=\begin{pmatrix}1.5 & -2 & 1.5\end{pmatrix}$ so their effect is averaged. The conditional distribution of $\mathbf{a}$ given $\mathbf{u}$ has a variance

$$Var\left(\mathbf{a}|\mathbf{u}\right) = \mathbf{D} - \mathbf{DZ}^{'}\left(\mathbf{ZDZ}^{'}\right)^{-\mathbf{1}}\mathbf{ZD} = \begin{pmatrix}0.5 & 0 & -0.5\\ 0 & 0 & 0\\ -0.5 & 0 & 0.5\end{pmatrix}$$

which shows well that the first and third markers are in LD (and their estimates cannot be disentangled) whereas the second has a unique solution for a given u. Assume that u are estimated with a posteriori covariance error (or prediction error covariance)

$$Var\left(\mathbf{u}|\mathbf{y}\right)=\mathbf{C}^{\mathrm{uu}}=\begin{pmatrix}1.5 & 0\\ 0 & 0.2\end{pmatrix}$$

Then, the incertitude in the estimation of marker effects is

$$Var\left(\mathbf{a}|\mathbf{y}\right) = \begin{pmatrix}0.925 & -0.1 & -0.075\\ -0.1 & 0.2 & -0.1\\ -0.075 & -0.1 & 0.925\end{pmatrix}$$

The difference between $Var\left(\mathbf{a}|\mathbf{y}\right)$ and $Var\left(\mathbf{a}|\mathbf{u}\right)$ is actually $Var\left(\widehat{\mathbf{a}}|\mathbf{y}\right)$, and has value

$$Var(\widehat{\mathbf{a}}|\mathbf{y}) = \begin{pmatrix} 0.425 & -0.1 & 0.425 \\ -0.1 & 0.2 & -0.1 \\ 0.425 & -0.1 & 0.425 \end{pmatrix}$$

It can be seen that this conditional variance does *not* account for the LD across markers 1 and 3, or, in other words, it ignores the fact that their *sum* is the only thing that can be well estimated.

Last, the $Var(\widehat{\mathbf{a}})$ is

$$Var(\widehat{\mathbf{a}}) = \begin{pmatrix} 0.075 & 0.1 & 0.075 \\ 0.1 & 0.8 & 0.1 \\ 0.425 & 0.1 & 0.075 \end{pmatrix}$$

12.14 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 GREML 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. Imagine that, for the same data set, you try three different matrices of relationships. Let's say that you have genomic, pedigree and kernel with respective matrices $\mathbf{A}_{22}$, $\mathbf{G}$ and $\mathbf{K}$ and variance component estimates $\sigma^2_{u_A}, \sigma^2_{u_G}, \sigma^2_{u_K}$. They do not refer to the same conceptual base populations. We proposed a method to compare estimates (Legarra 2016). The method basically says that in order to be comparable, all matrices should have similar statistics (average of the diagonal and of the matrix itself).

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.

12.15 Complicated things in GBLUP

12.15.1 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, $\mathbf{R}$, 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

$$\mathbf{R}= \begin{pmatrix} 1/w_1 & 0 & 0 \\ 0 & 1/w_2 & 0 \\ 0 & 0 & \dots \end{pmatrix} \sigma^2_e$$

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 modeled 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.

12.15.2 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.

13 Non-additive genetic effects in genomic selection

A recent review has been published (Varona *et al.* 2018), and we refer the reader to it for most of this section.

13.1 Dominant genomic relationships

Under quantitative genetics theory, the additive or breeding value for an *i-th* individual (u_i) involves the substitution effects of the genes (α)

$$\alpha = a + d(q - p)$$

which includes the “biological” additive effect a, the “biological” dominant effect d of the genes and the allele frequencies. So, the breeding values of a set of individuals are $\mathbf{u}$=$\mathbf{Z}\alpha$. With no dominant effect of the gene ($d = 0$), $\alpha = a$ and $\mathbf{u} = \mathbf{Za}$ as was defined in the previous sections.

If we consider one locus with two alleles (A_1 and A_2), a biological effect for each genotype can be defined, $A_1A_1 = \ a$, $A_1A_2 = \ d$ and $A_2A_2 = -a$, for instance as deviations from the midpoint of the two homozygous as in (Falconer and Mackay 1996). Naturally, a model that fits additive and dominant genotypic effects of the gene (or marker) can be written as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ta} + \mathbf{Xd} + \mathbf{e}$$

where “biological” additive effects $\mathbf{a}$ and “biological” dominant effects $\mathbf{d}$ for a set of individuals are included for each of the n markers (Toro and Varona 2010). It will be discussed in more details later in these notes.

This “intuitive” and useful model fits “biological” effect of gene or markers, while traditional quantitative genetic talks about “statistical” effects (Hill *et al.* 2008). Breeding values, dominance deviations, epistatic deviations and their variance components are statistical outcomes defined in a population context.

Thus, a genomic dominant model directly comparable to the classical genetic model (e.g. pedigree-based BLUP) has to involve breeding values $\mathbf{u}$ and dominance deviation $\mathbf{v}$ as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \mathbf{v} + \mathbf{e}$$

As in (Falconer and Mackay 1996) (Table 7.3), the breeding value for an individual is $u_{A_1A_1} = 2q\alpha = (2 - 2p)\,\alpha$, $u_{A_1A_2} = (q - p)\,\alpha = (1 - 2p)\,\alpha$ or $u_{A_2A_2} = (-2p)\,\alpha$, depending on its genotype and p is the frequency of A_1. So, the breeding values of a set of individuals are $\mathbf{u} = \mathbf{Z}\boldsymbol{\alpha}$ (with $\mathbf{Z}$ coded as in (VanRaden 2008)). The element of $\mathbf{Z}$ for an individual i at the marker j is

$$Z_{\text{ij}} = \begin{cases} (2 - 2p_j) \\ (1 - 2p_j) \\ \quad -2p_j \end{cases} \text{for genotypes} \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

Also, the dominant deviation of an individual is $v_{A_1A_1} = -2q^2d$, $v_{A_1A_2} = 2\text{pqd}$ and $v_{A_2A_2} = -2p^2d$. Hence, for a set of individuals, the dominance deviations are $\mathbf{v} = \mathbf{Wd}$ with the element of $\mathbf{W}$ for an individual i at the marker j equal to

$$W_{\text{ij}} = \begin{cases} -2q_j^2 \\ 2p_jq_j \\ -2p_j^2 \end{cases} \text{for genotypes} \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

Note that breeding value (u) involves both “biological” additive and dominant effects of the markers (a and d); dominance deviation (v) only includes a portion of the biological dominant effects of the markers (d).

From this information (also in Table 7.3 in Falconer and Mackay, 1996), the variance of breeding values and the variance of dominance deviations are obtained. The **additive genetic variance** is $\sigma_u^2 = 2\text{pq}\left[a + d(q-p)\right]^2 = 2\text{pq}\alpha^2$ with $E\left(u\right) = 0$. Additive variance includes variation due to the additive and dominant effects of the markers.

Also, like breeding values, the mean of dominance deviation is $E\left(v\right) = 0$,

$$E\left(v\right) = p^2\left(-2q^2d\right) + 2pq\left(2pqd\right) + q2\left(-2p^2d\right) = 0$$

and the **dominance genetic variance** is equal to $\sigma_v^2 = E\left(v^2\right) - \left[E(v)\right]^2 = E\left(v^2\right)$, so

$$\sigma_v^2 = p^2\left(-2q^2d\right)^2 + 2pq\left(2pqd\right)^2 + q2\left(-2p^2d\right)^2 = 4p^2q^2d^2(q^2+2pq+p^2)$$

$$\sigma_v^2 = \left[2pqd\right]^2$$

Dominance deviation variance only include a portion of the biological dominant effect of the markers.

Extended to several markers, and considering marker effects as random, this gives

$$\sigma_u^2 = \sum_{j=1}^{\text{nsnp}}\left(2p_jq_j\right)\sigma_{a0}^2 + \sum_{j=1}^{\text{nsnp}}\left(2p_jq_j\left(q_j-p_j\right)^2\right)\sigma_{d0}^2$$

$$\sigma_v^2 = \sum_{j=1}^{\text{nsnp}}\left(2p_jq_j\right)^2\sigma_{d0}^2$$

where σ_{a0}^2 and σ_{d0}^2 are the SNP variances for additive and dominant components, respectively.

The total genetic variance is $\sigma_g^2 = \sigma_u^2 + \sigma_v^2$, the first term is the additive genetic variance and the second term corresponds to the dominance genetic variance or dominance deviation variance. Note that the "statistical" partition of the variance in statistical components due to additivity, dominance and epistasis *does not* reflect the "biological" effects of the genes (Huang and Mackay 2016) though it is useful for prediction and selection decisions. Even when the genes have a biological or functional dominant action, this variation is mostly captured by the additive genetic variance (Hill and Mäki-Tanila 2015).

The "statistical" or classical parameterization implies linkage equilibrium and a population in Hardy-Weinberg equilibrium. Assuming uncorrelated random marker effects $(a,\ d)$, it can be extended to multiple loci (VanRaden 2008 ; Gianola *et al.* 2009) and obtained

$$Var\left(\mathbf{u}\right) = \frac{\mathbf{ZZ}'}{2\sum_{j=1}^{\text{nsnp}}p_jq_j}\sigma_u^2 = \mathbf{G}\sigma_u^2$$

as in (Vitezica *et al.* 2013) which is the classical **additive genomic relationship matrix *G*-**matrix of GBLUP (VanRaden 2008). Note that the variance component is $\sigma_u^2 = \sum\left(2p_jq_j\right)\sigma_{a0}^2 + \sum 2p_jq_j\left(q_j-p_j\right)^2\sigma_{d0}^2$.

For the dominant deviations $\mathbf{v}$, its variance-covariance matrix is:

$$Var\left(\mathbf{v}\right) = \mathbf{WW}'\sigma_{d0}^2$$

After dividing by the variance of the dominance deviations which is

$$\sigma_v^2 = \sum_{j=1}^{\mathrm{nsnp}} \left(2p_j q_j\right)^2 \sigma_{d0}^2$$

the **dominant genomic relationship matrix**, ***D***, is obtained as

$$Var\left(\mathbf{v}\right) = \frac{\mathbf{WW}'}{\sum_{j=1}^{\mathrm{nsnp}} \left(2p_j q_j\right)^2} \sigma_v^2 = \mathbf{D}\sigma_v^2$$

The dominant genomic matrix has some features as it was presented in a previous section for ***G***. Remember that in a base population in Hardy-Weinberg equilibrium, the average of the diagonal of **G** is one, whereas the average off-diagonal is 0. In the same conditions (base population in Hardy-Weinberg conditions), it turns out that the diagonal of **D** sums to

$$\frac{\left[p^2\left(-2q^2\right)^2 + 2\mathrm{pq}\left(2\mathrm{pq}\right)^2 + q^2\left(-2p^2\right)^2\right]}{\left(2\mathrm{pq}\right)^2}$$

for one locus, which is equal to 1. In addition, the sum of off-diagonal elements of **D** which can be written as

$$\frac{\begin{pmatrix} p^2 & 2\mathrm{pq} & q^2 \end{pmatrix} \begin{pmatrix} -2q^2 \\ 2\mathrm{pq} \\ -2p^2 \end{pmatrix} \begin{pmatrix} -2q^2 \\ 2\mathrm{pq} \\ -2p^2 \end{pmatrix}' \begin{pmatrix} p^2 & 2\mathrm{pq} & q^2 \end{pmatrix}'}{\left(2\mathrm{pq}\right)^2}$$

for one locus, sums to 0. Both features correspond to proper definitions of dominant relationships in a base population.

13.2 Animal model GDBLUP

With dominant genomic relationships defined in the previous section as $Var\left(\mathbf{v}\right) = \mathbf{D}\sigma_v^2$, the use of this matrix in a mixed model context for genomic predictions in GBLUP form is straightforward. We have the following linear mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Hu} + \mathbf{Hv} + \mathbf{e}$$

where **H** is an incidence matrix (here it is not the matrix of SSGBLUP) linking individuals to records (phenotypes). With $Var\left(\mathbf{u}\right) = \mathbf{G}\sigma_u^2$, $Var\left(\mathbf{e}\right) = \mathbf{R}$ and assuming multivariate normality, Henderson's mixed model equations are:

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{H} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{H} \\ \mathbf{H}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{H}'\mathbf{R}^{-1}\mathbf{H}+\mathbf{G}^{-1}\sigma_u^{-2} & \mathbf{H}'\mathbf{R}^{-1}\mathbf{H} \\ \mathbf{H}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{H}'\mathbf{R}^{-1}\mathbf{H} & \mathbf{H}'\mathbf{R}^{-1}\mathbf{H}+\mathbf{D}^{-1}\sigma_v^{-2} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{u}} \\ \widehat{\mathbf{v}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{H}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{H}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

These equations are identical to regular animal model, with the exception that genomic relationships in **G** and **D** are used instead of pedigree relationships. The breeding values and the dominance deviations can be predicted from these equations in a population in Hardy Weinberg and linkage equilibrium.

Note that it is not possible to use progeny records (DYD's) to predict dominance, because dominance deviations average to 0 across the progeny.

With the exception of (Aliloo *et al.* 2016) (for fat yield in Holstein), in most studies, the inclusion of dominance in GBLUP model did not improve predictive ability of the model (Su *et al.* 2012a

; Ertl *et al.* 2014 ; Xiang *et al.* 2016 ; Esfandyari *et al.* 2016 ; Moghaddar and Werf 2017) , whereas inclusion of the effect of inbreeding depression (shown later) does (Xiang *et al.* 2016).

13.3 Another parameterization

Now, we come back to the "intuitively" model fitting "biological" effect of markers,

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ta} + \mathbf{Xd} + \mathbf{e}$$

An additive effect a_j and a dominant effect d_j are included for each of the markers. The covariate t_{ij} is equal to 1, 0, -1, for SNP genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. For the dominant component, x_{ij} is equal to 0, 1, 0 for SNP genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. This model is based on "observed" genotypes and in particular in heterozygotes, so it can be called a "genotypic" model.

From this model proposed by (Su *et al.* 2012a) , we can define $\mathbf{u}^*$ and $\mathbf{v}^*$ as the "genotypic" additive and dominant effects. So, we can write for a set of individuals $\mathbf{u}^*=\mathbf{Ta}$ and $\mathbf{v}^*=\mathbf{Xd}$.

Table 21: Genotypic parameterization

Genotype	Frequency	Additive value	Dominant value	u^*	v^*
A_1A_1	p^2	a	0	$(2-2p)a$	$-2pqd$
A_1A_2	$2pq$	0	d	$(1-2p)a$	$(1-2pq)d$
A_2A_2	q^2	$-a$	0	$-2pa$	$-2pqd$
Average		$(p-q)a$	$2pqd$		

Note that $\mathbf{u}^*$ is *not* a breeding value because $\mathbf{a}$ is NOT a substitution effect, is the part attributable to the additive "biological" effect of the marker. The incidence matrix $\mathbf{T}$ corresponds to the incidence matrix $\mathbf{Z}$ (used in the classical model defined in terms of breeding values and dominance deviations). However, the matrix $\mathbf{X} \neq \mathbf{W}$ ($\mathbf{W}$ is used in the classical model for the dominance deviations).

The **variance of the genotypic additive value** can be obtained as $\sigma^2_{u^*} = E\left(u^{*2}\right) - \left[E(u^*)\right]^2$, and idem for the **variance of the genotypic dominant value** $\sigma^2_{v^*}$. Then

$$\sigma^2_{u^*} = \sum 2p_jq_j\sigma^2_a$$

and

$$\sigma^2_{v^*} = \sum 2p_jq_j(1-2p_jq_j)\sigma^2_d$$

Quite different from

$$\sigma^2_u = \sum_{j=1}^{\mathrm{nsnp}} (2p_jq_j)\,\sigma^2_{a0} + \sum_{j=1}^{\mathrm{nsnp}} \left(2p_jq_j\,(q_j-p_j)^2\right)\sigma^2_{d0}$$

$$\sigma^2_v = \sum_{j=1}^{\mathrm{nsnp}} (2p_jq_j)^2\,\sigma^2_{d0}$$

that we obtained before. The variances $\sigma^2_{u^*}$ and $\sigma^2_{v^*}$ estimated under the "genotypic" model as in Su et al. (2012) are NOT genetic variances. In particular, they do not include dominant effects, but by definition of breeding value, the reproductive value of an individual contains substitution effects, which contain dominant effects. Therefore, σ_u^2 and σ_v^2 are more useful for selection.

Vitezica et al. (2013) showed that also the dominant relationship matrices ($\mathbf{D}$) are different between the classical (statistical) and the "genotypic" model. The parameterization is largely a matter of convenience, both models are able to explain the data ($\mathbf{y}$) but their interpretation is different. The classical model in terms of breeding values and substitution effects (statistical) is more adequate for selection (both for ranking animals and for predicting genetic improvement).

The *only* variance components comparable with pedigree-based estimates are σ_u^2 and σ_v^2 obtained from the statistical genomic model. Using variance components estimated from the "genotypic" model is misleading because they underestimate the importance of additive variance and overestimate the importance of dominance variance (Vitezica *et al.* 2013). From the total genetic variance σ_g^2 it can be verified that $\sigma_u^2 + \sigma_v^2 = \sigma_{u^*}^2 + \sigma_{v^*}^2$. Thus, it is simple to switch variance component estimates between "statistical" (σ_u^2 and σ_v^2) and "biological" ($\sigma_{u^*}^2 + \sigma_{v^*}^2$) models if the distribution of the allelic frequencies is available (Vitezica et al., 2013).

The "statistical" model of Vitezica et al. (2013). This means that introducing new genetic effect (e.g. additive vs. additive plus dominance) in the model does *not* change previous estimates. For instance: going from an additive to an additive + dominant model should not change much neither the estimates of variance components, nor the estimates of breeding values and dominant deviations. However, the "genotypic" model of Su et al. (2012a) is not orthogonal. Including dominance may change greatly the estimate of additive values and variances, and in addition, the estimated additive values are not breeding values – they are "genotypic" additive values.

13.4 Inbreeding depression

Phenomena of inbreeding depression and heterosis may be explained by directional dominance (Lynch and Walsh 1998). In other words, higher percentage of positive than negative functional dominant effects d is expected to happen in reality.

With directional dominance, the mean of dominant effect $\mathbf{d}$ is different from zero. However, typically models assume that $\mathbf{a}$ and $\mathbf{d}$ have zero means. Xiang et al. (2016) show that inclusion of genomic inbreeding (based on SNPs and included as a covariate) accounts for directional dominance and inbreeding depression.

Xiang et al. (2016) proposed to write the model including (biological) additive and dominant effects of the markers as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ta} + \mathbf{Xd}^{*} + \mathbf{X1}\mu_d + \mathbf{e}$$

where $\mathbf{d}^{*} = \mathbf{d} - E(\mathbf{d}) = \mathbf{d} - \mu_d$ and the matrix $\mathbf{X}$ has a value of 1 at heterozygous loci for an individual and 0 otherwise.

The term $\mathbf{X1}$ defined as $\mathbf{h} = \mathbf{X1}$ contains the row sums of $\mathbf{X}$, i.e. individual heterozygosities (how many markers are heterozygotes for each individual). The genomic inbreeding coefficient $\mathbf{f}$ can be calculated as: $\mathbf{f} = \mathbf{1} - \mathbf{h}/N$, where N is the number of markers. For instance, $\mathbf{f}$ is a vector that contains the percentage of homozygous loci for each individual. Then,

$$\mathbf{h} = (\mathbf{1} - \mathbf{f})\,N = \mathbf{1}N + \mathbf{f}\,(-N)$$

and with the mean μ_d

$$\mathbf{h}\mu_d = (\mathbf{1} - \mathbf{f})\,N\mu_d = \mathbf{1}N\mu_d + \mathbf{f}\,(-N\mu_d)$$

Thus, the model can be rewritten as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ta} + \mathbf{Xd}^{*} + \mathbf{1}N\mu_d + \mathbf{f}(-N\mu_d) + \mathbf{e}$$

and finally

$$\mathbf{y} = \mathbf{1}\mu^* + \mathbf{Ta} + \mathbf{Xd}^* {+} \mathbf{f}b {+} \mathbf{e}$$

where the term $\mathbf{1}N\mu_d$ is confounded with the overall mean of the model (μ^*), while the $\mathbf{f}\,(-N\mu_d)$ models the inbreeding depression and $b = (-N\mu_d)$ is the inbreeding depression summed over the marker loci, which is to be estimated.

This important result means that genomic inbreeding can be used to model directional dominance. This model allows to obtain estimates of inbreeding depression parameter in different populations (e.g. breeds or lines) and also in crossbred animals (Xiang *et al.* 2016).

Inclusion of genomic inbreeding must *always* be done in order to obtain a correct estimation of genetic dominance variance (σ_v^2). Otherwise, the genetic dominance variance is inflated. This was confirmed in real data by (Xiang *et al.* 2016 ; Aliloo *et al.* 2016). This has long been known for pedigree analysis (e.g. (De Boer and Hoeschele 1993)); even if dominance is not considered, inbreeding may be considered in genomic evaluations.

13.5 Genomic relationship matrices in absence of HWE

In the classical or "statistical" model that we showed previously, the effects (additive or breeding values, dominance deviations and epistatic deviations) are all orthogonal in linkage and **Hardy-Weinberg equilibrium** (HWE). What does the orthogonal property of the model mean? It means that the estimation of one genetic (e.g. additive) effect is not affected by the presence or absence of other genetic effects in the model (e.g. dominance or epistasis).

This property results in orthogonal partition of the variances. Why? because, substitution effect contributes to the additive genetic variance, the dominance deviation contributes to the dominance genetic variance, etc. There is no covariance between the genetic effects. In other words, introducing new genetic effect (e.g. additive vs. additive plus dominance) in the model does *not* change previous estimates. For instance: going from an additive to an additive + dominant model should not change much neither the estimates of variance components, nor the estimates of breeding values and dominant deviations.

Crossbreeding schemes are widely used in animal breeding (e.g. pigs, chickens) for the purpose of exploiting the heterosis and breed complementarity that often occur in crosses. Theses crosses (e.g. F1) or inbred populations are not in Hardy Weinberg equilibrium. Therefore, we need methods for genomic predictions, preferably including dominance, in these populations.

Additive (**G**) and dominance deviation (**D**) relationship matrices can be built removing the requirement of Hardy Weinberg equilibrium and assuming linkage equilibrium (Vitezica et al., 2017). This generalization of classical model is based in the NOIA orthogonal approach (Álvarez-Castro and Carlborg 2007).

So, the breeding values of a set of individuals are $\mathbf{u} = \mathbf{Z}\alpha$ where α are dominant deviations, and the element of $\mathbf{Z}$ for an individual i at the marker j is

$$\mathbf{z}_{ij} = \begin{cases} -(-p_{A_1A_2} - 2p_{A_2A_2}) \\ -(1 - p_{A_1A_2} - 2p_{A_2A_2}) \\ -(2 - p_{A_1A_2} - 2p_{A_2A_2}) \end{cases} \quad \text{for genotypes} \quad \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

and the dominance deviation is $\mathbf{v} = \mathrm{Wd}$ with the element of $\mathbf{W}$ for an individual i at the marker j is

$$\mathbf{w}_{\mathrm{ij}} = \begin{cases} -\frac{2p_{A_1A_2}p_{A_2A_2}}{p_{A_1A_1}+p_{A_2A_2}-\left(p_{A_1A_1}-p_{A_2A_2}\right)^2} \\ \frac{4p_{A_1A_1}p_{A_2A_2}}{p_{A_1A_1}+p_{A_2A_2}-\left(p_{A_1A_1}-p_{A_2A_2}\right)^2} \\ -\frac{2p_{A_1A_1}p_{A_1A_2}}{p_{A_1A_1}+p_{A_2A_2}-\left(p_{A_1A_1}-p_{A_2A_2}\right)^2} \end{cases} \quad \text{for genotypes} \quad \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

where $p_{A_1A_1}$, $p_{A_1A_2}$ and $p_{A_2A_2}$ are the genotypic frequencies for the genotypes A_1A_1, A_1A_2 and A_2A_2. Under the assumption of HWE, the "statistical" model presented before (as in Vitezica et al., 2013) is a particular case of this model where $p_{A_1A_1} = p^2$, $p_{A_1A_2} = 2\text{pq}$ and $p_{A_2A_2} = q^2$, and the denominator $p_{A_1A_1} + p_{A_2A_2} - (p_{A_1A_1} - p_{A_2A_2})^2 = 2\text{pq}$.

The additive relationship matrix is as:

$$Var\left(\mathbf{u}\right) = \frac{\mathbf{ZZ}'}{tr(\mathbf{ZZ}')/n}\sigma_u^2 = \mathbf{G}\sigma_u^2$$

where tr is the trace of the matrix and n is the number of individuals. In a Hardy Weinberg population, tr($\mathbf{ZZ}'$) corresponds to the heterozygosity of the markers $2\sum$pq.

For the dominance deviations, the relationship matrix is

$$Var\left(\mathbf{v}\right) = \frac{\mathbf{WW}'}{tr(\mathbf{WW}')/n}\sigma_v^2 = \mathbf{D}\sigma_v^2$$

The tr($\mathbf{WW}'$) corresponds to the square of the heterozygosity of the markers $4\sum(\text{pq})^2$ in Hardy-Weinberg equilibrium (Vitezica *et al.* 2013).

Now, we know how to build a model that allows the orthogonal decomposition of the variances in any population in Hardy Weinberg equilibrium or not, and thus the correct estimation of genetic variance components (equivalent to pedigree-based estimates).

13.6 Epistatic genomic relationships

The traditional definition of epistasis is the interaction of the genes, pairwise or higher-order interactions. In fact, the number of epistatic effects and accordingly the number of parameters in the model may be extremely large. Thus, we can define epistatic relationship matrices for individuals as we do in GBLUP, which is more efficient from the computational point of view.

Following this idea, Cockerham (1954) suggested to use the Hadamard product of the additive and dominant relationship (pedigree based) matrices to obtain the epistatic relationship matrices. Remember that the Hadamard product ($\odot$) between two matrices ($\mathbf{B} \odot \mathbf{C}$) produces another matrix ($\mathbf{A} = \mathbf{B} \odot \mathbf{C}$, with the same dimension) where each element of $\mathbf{A}$ is the product of elements ($a_{\text{ij}} = b_{\text{ij}} * c_{\text{ij}}$) of the two original matrices ($\mathbf{B}$ and $\mathbf{C}$).

The construction of the epistatic relationship matrices using the Hadamard product depends on the assumption of Hardy Weinberg equilibrium in other words, non-inbreeding and random mating (Cockerham 1954). The Hadamard product relies on the orthogonal property of the model because no covariance exists between main genetic effects (e.g. additive and epistatic effects). Henderson (Henderson 1985) suggested the use of these matrices in BLUP.

Henderson's approach was extended to the genomic framework by Xu (2013) for an F2 design and used for predicting hybrid performance in a rice F2 population (Xu *et al.* 2014). However, their extension is not general. It is a particular case because genotype frequencies in the F2 are the Hardy Weinberg frequencies corresponding to the allele frequency in the F1 (Falconer and Mackay 1996).

If we assume a more general situation with or without Hardy Weinberg equilibrium, we need to check if **Hadamard product of genomic matrices** is equivalent to the direct estimation of loci-based epistatic effects.

We have the following linear model with epistasis:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \mathbf{v} + \sum_{i=A,D}\sum_{j=A,D}\mathbf{g}_{\text{ij}} + \sum_{i=A,D}\sum_{j=A,D}\sum_{k=A,D}\mathbf{g}_{\text{ijk}} + \ldots + \mathbf{e}$$

where $\mathbf{u}$ is the additive or breeding value, $\mathbf{v}$ is the dominance deviations, $\mathbf{g}_{\text{ij}}$ is the second order epistatic effect and $\mathbf{g}_{\text{ijk}}$ the third-order epistatic effect and so on; and $\mathbf{e}$ is a residual vector. The second-order epistatic genetic effects can be partitioned into additive-by-additive ($\mathbf{g}_{\text{AA}}$), additive-by-dominant ($\mathbf{g}_{\text{AD}}$) and dominant-by-dominant ($\mathbf{g}_{\text{DD}}$). The third-order epistatic genetic effects can be included in the model, but they are as either negligible (Hill and Mäki-Tanila 2015) or too difficult to estimate. Note that this genomic model includes "genetic" effects.

For obtaining genetic variance component estimations comparable to pedigree-based variances, a full orthogonal "statistical" model was proposed by (Varona 2014-08-17/2014-08-22 ; Vitezica *et al.* 2017). We have defined in the previous sections, the breeding values of a set of individuals as $\mathbf{u} = \mathbf{Z}\alpha$ and the dominance deviation as $\mathbf{v} = \mathbf{Wd}$. From here in the text, we rename $\mathbf{Z}$ and $\mathbf{W}$ as $\mathbf{H}_a$ and $\mathbf{H}_d$ respectively. Thus, $\mathbf{u} = \mathbf{H}_a\alpha$ and $\mathbf{v} = \mathbf{H}_d\mathbf{d}$. As you see before, the matrix $\mathbf{H}_a$ has n rows (number of individuals) and m columns (number of markers) containing "additive" coefficients, This matrix can be written as

$$\mathbf{H}_a = \begin{pmatrix} \mathbf{h}_{a_k} \\ \vdots \\ \mathbf{h}_{a_n} \end{pmatrix}$$

where $\mathbf{h}_{a_k}$ is a row vector for the k-th individual with m columns. For individual 1, the vector $\mathbf{h}_{a_1}$ is equal to $(h_{a_{11}}, \ldots, h_{a_{1m}})$.

Álvarez-Castro and Carlborg (2007) proved that the coefficients of the incidence matrix for second-order epistatic effects between two loci can be computed as the Kronecker products of the respective incidence matrices for single locus effects. So, for the interactions, such as additive-by-dominant interaction, the matrix $\mathbf{H}_{\text{ad}}$ can be written using Kronecker products of each row of the preceding matrices as

$$\mathbf{H}_{\text{ad}} = \begin{pmatrix} \mathbf{h}_{a_i} \otimes \mathbf{h}_{d_i} \\ \mathbf{h}_{a_{i+1}} \otimes \mathbf{h}_{d_{i+1}} \\ \ldots \\ \mathbf{h}_{a_n} \otimes \mathbf{h}_{d_n} \end{pmatrix}$$

For instance, for individual 1 the incidence matrix of additive-by-dominant epistatic effects is $\mathbf{h}_{ad_1} = \mathbf{h}_{a_1} \otimes \mathbf{h}_{d_1}$. As example, we have **2 individuals** and **3 loci**,

- for individual 1, the vector $\mathbf{h}_{a_1}$ is equal to $(h_{a_{11}}, h_{a_{12}}, h_{a_{13}})$ and $\mathbf{h}_{d_1} = (h_{d_{11}}, h_{d_{12}}, h_{d_{13}})$.
- for individual 2, the vector $\mathbf{h}_{a_2}$ is equal to $(h_{a_{21}}, h_{a_{22}}, h_{a_{23}})$ and $\mathbf{h}_{d_1} = (h_{d_{21}}, h_{d_{22}}, h_{d_{23}})$

and

$$\mathbf{H}_{\text{ad}} = \begin{pmatrix} h_{a_{11}}h_{d_{11}} & h_{a_{11}}h_{d_{12}} & h_{a_{11}}h_{d_{13}} & h_{a_{12}}h_{d_{11}} & h_{a_{12}}h_{d_{12}} & h_{a_{12}}h_{d_{13}} & h_{a_{13}}h_{d_{11}} & h_{a_{13}}h_{d_{12}} & h_{a_{13}}h_{d_{13}} \\ h_{a_{21}}h_{d_{21}} & h_{a_{21}}h_{d_{22}} & h_{a_{21}}h_{d_{23}} & h_{a_{22}}h_{d_{21}} & h_{a_{22}}h_{d_{22}} & h_{a_{22}}h_{d_{23}} & h_{a_{23}}h_{d_{21}} & h_{a_{23}}h_{d_{22}} & h_{a_{23}}h_{d_{23}} \end{pmatrix}$$

The matrix $\mathbf{H}_{\text{ad}}$ has as many columns as marker interactions (here, 9) and as many rows as individuals. This matrix is of very large size (e.g. for a 50K SNP chip and 1000 individuals the matrix contains 1000 x 50000^2 elements). In addition, $\mathbf{H}_{\text{ad}}\mathbf{H}_{\text{ad}}^{'}$ cross-product (that we need to compute for covariance matrices) is computationally expensive. Hopefully, an algebraic shortcut was found (Vitezica et al., 2017) that allows easy computation of $\mathbf{H}_{\text{ad}}\mathbf{H}_{\text{ad}}^{'}$ and the rest of cross-products for epistatic matrices, even for third and higher orders.

The relationship matrices of epistatic genetic effects can be written as

$$Var\left(\mathbf{g}_{\text{AA}}\right) = \frac{\mathbf{H}_{\text{aa}}\mathbf{H}'_{\text{aa}}}{tr(\mathbf{H}_{\text{aa}}\mathbf{H}'_{\text{aa}})/n}\sigma^2_{g_{\text{AA}}} = \mathbf{G}_{\text{AA}}\sigma^2_{g_{\text{AA}}}$$

$$Var\left(\mathbf{g}_{\text{AD}}\right) = \frac{\mathbf{H}_{\text{ad}}\mathbf{H}'_{\text{ad}}}{tr(\mathbf{H}_{\text{ad}}\mathbf{H}'_{\text{ad}})/n}\sigma^2_{g_{\text{AD}}} = \mathbf{G}_{\text{AD}}\sigma^2_{g_{\text{AD}}}$$

$$Var\left(\mathbf{g}_{\mathrm{DD}}\right)=\frac{\mathbf{H}_{\mathrm{dd}}\mathbf{H'}_{\mathrm{dd}}}{tr(\mathbf{H}_{\mathrm{dd}}\mathbf{H'}_{\mathrm{dd}})/n}\sigma^2_{g_{\mathrm{DD}}}=\mathbf{G}_{\mathrm{DD}}\sigma^2_{g_{\mathrm{DD}}}$$

and with the algebraic shortcut as

$$Var\left(\mathbf{g}_{\mathrm{AA}}\right)=\frac{\mathbf{G}_A\odot\mathbf{G}_A}{\mathrm{tr}\left(\mathbf{G}_A\odot\mathbf{G}_A\right)/n}\sigma^2_{g_{\mathrm{AA}}}=\mathbf{G}_{\mathrm{AA}}\sigma^2_{g_{\mathrm{AA}}}$$

$$Var\left(\mathbf{g}_{\mathrm{AD}}\right)=\frac{\mathbf{G}_A\odot\mathbf{G}_D}{\mathrm{tr}\left(\mathbf{G}_A\odot\mathbf{G}_D\right)/n}\sigma^2_{g_{\mathrm{AD}}}=\mathbf{G}_{\mathrm{AD}}\sigma^2_{g_{\mathrm{AD}}}$$

$$Var\left(\mathbf{g}_{\mathrm{DD}}\right)=\frac{\mathbf{G}_D\odot\mathbf{G}_D}{\mathrm{tr}\left(\mathbf{G}_D\odot\mathbf{G}_D\right)/n}\sigma^2_{g_{\mathrm{DD}}}=\mathbf{G}_{\mathrm{DD}}\sigma^2_{g_{\mathrm{DD}}}$$

using Hadamard products of additive and dominance genomic orthogonal relationships. **A standardization based on the trace of the relationship matrices is needed.** The normalization factor based on the traces was already used by Xu (2013) but several authors ignore it (e.g. (Muñoz *et al.* 2014)). Here the reasoning for pairwise interactions is present but it extends to third and higher order interactions, (e.g., $\mathbf{G}_{\mathrm{AAD}}=\frac{\mathbf{G}_A\odot\mathbf{G}_A\odot\mathbf{G}_D}{\mathrm{tr}(\mathbf{G}_A\odot\mathbf{G}_A\odot\mathbf{G}_D)/n}$).

Note that this approach only assumes linkage equilibrium. In outbred populations (as animal populations), substantial LD (linkage disequilibrium) is present only between polymorphisms in tight linkage (Hill and Mäki-Tanila 2015).

Two other approaches are in the literature to model epistatic interactions. First, a "biological" non-orthogonal model has been proposed by Martini et al. (2016) but it can only be used for prediction and not for the estimation of variance components. Second, the RKHS (Reproducing Kernel Hilbert Space) approach (Gianola *et al.* 2006). However, most kernels consider similarities within loci and not consider joint similarity across loci (Varona *et al.* 2018).

13.7 Word of caution

It is quite easy to fit dominance, epistasis… in a GBLUP context when data sets are not too large. However, there is very little information, and most estimates of variance components have high standard errors. Also, the estimates of dominance and epistatic deviations are not of much accuracy. Thus, the researcher should be cautious when interpreting the results and using them in practice.

14 Single Step GBLUP

The idea for ssGBLUP came from the fact that only a small portion of the animals, in a given population, are genotyped. In this way, the best approach to avoid several steps would be to combine pedigree and genomic relationships and use this matrix as the covariance structure in the mixed model equations (MME). There are two derivations and both are very similar.

14.1 SSGBLUP as improved relationships

Legarra et al. (2009) stated that genomic evaluations would be simpler if genomic relationships were available for all animals in the model. Then, their idea was to look at $\mathbf{A}$ as *a priori* relationship and to $\mathbf{G}$ as an *observed* relationship; however, $\mathbf{G}$ is observed only for some individuals that have $\mathbf{A}_{22}$ as *a priori* relationship. Based on that, they showed the genomic information could be extended to ungenotyped animals based on the joint distribution of breeding values of ungenotyped ($\mathbf{u}_1$) and genotyped ($\mathbf{u}_2$) animals:

$$p(\mathbf{u}_1, \mathbf{u}_2) = p\left(\mathbf{u}_2\right) p(\mathbf{u}_1|\mathbf{u}_2)$$

$$p(\mathbf{u}_2) = \ N(\mathbf{0}, \mathbf{G})$$

If we consider that

$$\text{var}\left(\mathbf{u}\right) = \mathbf{A}\sigma_u^2$$

In the following, we can just omit σ_u^2 in the derivations

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

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

$$p(\mathbf{u}_1|\mathbf{u}_2) = \ N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

This can also be seen as

$$\mathbf{u}_1 {=} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2 + \boldsymbol{\varepsilon}$$

With $Var\left(\varepsilon\right) = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$.

Because the animals with subscript 1 have no genotypes, the variance depends on their pedigree relationships with genotyped animals. The derivation assumes multivariate normality of ε, which holds because these are overall values resulting from a sum of gene effects.

Using rules, variances and covariances are:

$$Var\left(\mathbf{u}_1\right) = var\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2 + \varepsilon\right) = Var(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2) + Var(\varepsilon)$$

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

Rearranging:

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

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

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

Therefore,

$$Var\left(\mathbf{u}_1\right) = \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$

$$Var\left(\mathbf{u}_2\right) = \mathbf{G}$$

$$Cov\left(\mathbf{u}_1, \mathbf{u}_2\right) = Cov\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{u}_2\right) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}Var\left(\mathbf{u}_2\right) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}$$

Finally, the matrix that contains the joint relationships of genotyped and ungenotyped animals is given by (again, assuming for simplicity of presentation $\sigma_u^2 = 1$):

$$\begin{aligned}\mathbf{H} &= \begin{pmatrix} \text{var}\left(\mathbf{u}_1\right) & \text{cov}\left(\mathbf{u}_1, \mathbf{u}_2\right) \\ \text{cov}\left(\mathbf{u}_2, \mathbf{u}_1\right) & \text{var}\left(\mathbf{u}_2\right) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix} \\ &= \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right) \\ \left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}\end{aligned}$$

Which can be simplified to:

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

We usually assume in these notes that $\mathbf{u}_2 = \mathbf{Za}$, which leads to VanRaden's $\mathbf{G}$ (VanRaden 2008). The derivation of (Legarra *et al.* 2009) does not seem to require that $\mathbf{G}$ is actually VanRaden's $\mathbf{G}$ (could potentially be something else), but when they use $\mathbf{A}$ to model relationships, they assume an additive model. So, $\mathbf{G}$ should be "additive", which makes sense for VanRaden's $\mathbf{G}$ but also for similar matrices like $\mathbf{G}_{\text{IBS}}$ or "corrected" $\mathbf{G}_{\text{IBS}}$. One of the key assumptions of the methods in (Legarra *et al.* 2009) is that $E\left(\mathbf{u}_2\right) = 0$, (animals genotyped have 0 expected breeding value) which is not necessarily true if those animals are selected animals. We will see ways to deal with that later. Although $\mathbf{H}$ is very complicated, $\mathbf{H}^{-1}$ is quite simple (Aguilar *et al.* 2010 ; Christensen and Lund 2010).

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

14.2 SSGBLUP as linear imputations

Christensen and Lund (2010) proposed another derivation. They started by inferring the genomic relationship matrix for all animals using inferred (imputed) genotypes for non-genotyped animals; we have seen that this can be obtained using Gengler's (Gengler *et al.* 2007) method, modelling the genotype $\mathbf{z}$ as a quantitative trait: $\mathbf{z}{=}\mathbf{1}\mu + \text{Wu}{+}\mathbf{e}$. If μ ($= 2p$ in the base population) is known, then we can linearly "impute" *centered* gene content for one marker as $\widehat{\mathbf{z}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{z}_2$, which extends to multiple markers as $\widehat{\mathbf{Z}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{Z}_2$.

This provides the "best guess" of genotypes. We may then construct a "poor man" version of G using $\hat{\mathbf{G}} = \widehat{\mathbf{Z}}_1\widehat{\mathbf{Z}}_1'/\sum 2p_jq_j$. This matrix will be incorrect because when we impute, we get a guess – and the guess has an error. However, the missing data theory states that we need 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 $E(\mathbf{Z}_1|\mathbf{Z}_2) = \widehat{\mathbf{Z}}_1$, and the conditional variance expressing the uncertainty about the “guess” is $Var(\ \widehat{\mathbf{Z}}_1 |\mathbf{Z}_2)\ = (\mathbf{A}_{11}{-}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})\mathbf{V}$ where $\mathbf{V}$ contains $2p_kq_k$ in the diagonal. These two results can be combined to obtain the desired augmented genomic relationships. For instance, for the non-genotyped animals,

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

which equals

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

Finally, the augmented covariance matrix is

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

where

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

is the augmented genomic relationship matrix with inverse

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

assuming that $\mathbf{G}$ 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. The key hypothesis here is that the base population allele frequencies are known, which is not necessarily true.

14.3 ssGBLUP mixed model equations

Assuming the following animal model:

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

The MME for ssGBLUP become, for one trait:

$$\begin{pmatrix}\mathbf{X}'\mathbf{X}\sigma_e^{-2} & \mathbf{X}'\mathbf{W}\sigma_e^{-2}\\ \mathbf{W}'\mathbf{X}\sigma_e^{-2} & \mathbf{W}'\mathbf{W}\sigma_e^{-2}{+}\mathbf{H}^{-1}\sigma_u^{-2}\end{pmatrix}\begin{pmatrix}\widehat{\mathbf{b}}\\ \widehat{\mathbf{u}}\end{pmatrix} = \begin{pmatrix}\mathbf{X}'\mathbf{y}\sigma_e^{-2}\\ \mathbf{W}'\mathbf{y}\sigma_e^{-2}\end{pmatrix}$$

And for multiple traits:

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

where $\mathbf{G}_0$ is the matrix of genetic covariance across traits, and usually $\mathbf{R}{=}\mathbf{I}\otimes\mathbf{R}_0$, where $\mathbf{R}_0$ is the matrix of residual covariances. The formulation is as general as pedigree-based BLUP.

Some properties are:

- All models that ran using MME and **A**-matrix run using **H**-matrix. This includes, among others, random regression models, multiple trait models, threshold models, maternal effect models.
- Existing software can be easily recycled to run SSGBLUP including a mechanism to introduce elements $\mathbf{G^{-1}}{-}\mathbf{A}_{22}^{-1}$, that can be computed externally
- REML and Gibbs sampling algorithm work perfectly well without modifications

14.4 Some properties of matrix H

Matrix **H** is full rank (invertible) matrix, because it can be formed as

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

which is a full rank matrix. However, for **H** to be positive definite (which is the requisite for using its inverse in MME), $\mathbf{G} - \mathbf{A}_{22}$ needs to be positive definite. It usually is – maybe after some adjustments for compatibility that will be used later.

The inverse

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

is also full rank, but for it to be positive definite, it needs $\mathbf{G^{-1}}{-}\mathbf{A}_{22}^{-1}$ to be positive definite. Again, if things are done properly, it usually is.

Construction of $\mathbf{H^{-1}}$ is simple, because it follows four steps:

1. Build $\mathbf{A^{-1}}$ using Henderson's rules
2. Build **G** and invert it
3. Build $\mathbf{A}_{22}$ and invert it

The matrix $\mathbf{A}_{22}$ is the relationship matrix of genotyped individuals. This matrix can be constructed using the tabular method, but this is very costly for large data sets. A better option is to use either recursions (Aguilar and Misztal 2008) or Colleau (2002) algorithm. Several strategies were described by Aguilar et al. (2011). We remind also that the $\mathbf{A}_{22}^{-1}$ is *not* the corresponding block of $\mathbf{A^{-1}}$, in other words, it has to be constructed and inverted explicitely.

The diagonal in $\mathbf{G^{-1}}{-}\mathbf{A}_{22}^{-1}$ is usually positive. This implies (roughly) that there is more information in **G** than in $\mathbf{A}_{22}$, because **G** captures realized relationships.

Matrix $\mathbf{H}^1$ is rather sparse. Consider the following two examples:

- Manech Tete Rouse sheep has ~3000 animals (rams) genotyped for a ~500,000 animals pedigree. Thus, $\mathbf{A^{-1}}$ has 4.5 × 10^6 non-null elements, and **G** or $\mathbf{A}_{22}$ have 9 ×10^6non-null elements. Combined, this results in $\mathbf{H^{-1}}$ with 13.5 ×10^6non-null elements. Compare this to what would happen if we could genotype the entire population and do GBLUP (**G** would have 250,000 ×10^6elements !!) or SNP-BLUP ($\mathbf{ZZ}'$ would have 2,500 ×10^6elements).
- Angus cattle has ~11,000,000 animals in pedigree and ~500,000 animals genotyped. Matrix $\mathbf{A^{-1}}$ has 100 ×10^6non-null elements, **G** or $\mathbf{A}_{22}$ have 250,000 ×10^6elements, $\mathbf{H^{-1}}$ has 350 ×10^9elements. If we could genotype the entire population and do GBLUP (**G** would have 121 × 10^{12} elements !!) or SNP-BLUP ($\mathbf{Z'Z}$ would have 2,500 ×10^6elements).

When the number of animals genotyped is very large (larger than the number of markers), matrix **G** gets rather big. For this reason, there are other formulations of ssGBLUP that will be presented later.

Matrix $\mathbf{H}$ 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 $\mathbf{H}$ if their descendants are related in $\mathbf{G}$. Accordingly, two descendants of individuals that are related in $\mathbf{G}$ will be related in $\mathbf{H}$, even if the pedigree disagrees. Indeed, it has been suggested to use $\mathbf{H}$ in mating programs to avoid inbreeding (Sun *et al.* 2013).

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

14.4.1 Inbreeding in H

The diagonal elements of $\mathbf{A}$ contain inbreeding expressed as $F_i = A_{\text{ii}} - 1$. The diagonal elements of $\mathbf{G}$ contain genomic inbreeding expressed as $F_i = G_{\text{ii}} - 1$. Inbreeding is useful to handle genetic variability, and also to compute model based reliabilities as $rel_i = 1 - \frac{\text{PEV}}{(1+F_i)\sigma_u^2}$. Because $\mathbf{H}$ uses all information, the best estimate of inbreeding combining pedigree and genomic information is actually $F_i = H_{\text{ii}} - 1$. Note that we have efficient methods to obtain $\mathbf{H}^{-1}$ but we do not have efficient methods to obtain $\mathbf{H}$ or its diagonal. Xiang et al. (2017) obtained the diagonal of $\mathbf{H}$ by computing the sparse inverse of $\mathbf{H}^{-1}$ as is programmed in YAMS. Anyway, we do not know -yet-how to efficiently obtain “H-inbreeding”. Recently, Colleau et al. (2017) proposed a method that allows to compute overall statistics of $\mathbf{H}$ such as total relationship. The method is quite involved numerically but so far it is the only existing option.

14.5 Mixing G and A: blending and compatibility of pedigree and genomic relationships.

This is a very important chapter because lots of people (including famous researchers) confuse “compatibility”, which tries to put $\mathbf{G}$ and $\mathbf{A}$ in the same scale, and “blending”, which is basically a technique used to assign part of the genetic variance to pedigree – not markers, and at the same time used to have an invertible $\mathbf{G}$.

14.5.1 Blending

14.5.1.1 Blending to include the residual polygenic effect In previous chapter for GBLUP, we saw that we can model the total genetic effect as based, partly, on pedigree, and partly on genomic relationships. Let us decompose the breeding values of all individuals in a part due to markers and a residual part due to pedigree, $\mathbf{u}$=$\mathbf{u}_m$+$\mathbf{u}_p$ with respective variances $\sigma_u^2 = \sigma_{u,m}^2 + \sigma_{u,p}^2$. The “marker-based” part will have a relationship matrix $\mathbf{H}$ in Single Step, whereas the “pedigree-based” part will have a relationship matrix $\mathbf{A}$.

It follows that $Var(\mathbf{u}) = ((1-\alpha)\,\mathbf{H}+\alpha\mathbf{A})\,\sigma_u^2$ where $\alpha = \sigma_{u,m}^2/\sigma_u^2$. In practice, in the SSGBLUP, it is easier to create a modified genomic relationship matrix $\mathbf{G}_w$ ($\mathbf{G}$ in (Aguilar *et al.* 2010); $\mathbf{G}_w$ in (VanRaden 2008 ; Christensen 2012)) as $\mathbf{G}_w = (1-\alpha)\,\mathbf{G}+\alpha\mathbf{A}_{22}$.

This is known as “blending”. In practice, the value of α is low (values oscillate between 0.05 and 0.7) and has mostly negligible effects on predictions. It has been claimed that blending reduces bias of predictions, and this results in different optimal α coefficients for different traits and species. It seems strange to accept that markers would describe one trait up to 95% of its variance, and for the same animals, only 70%. A more coherent approach is to estimate the two components in $\sigma_u^2 = \sigma_{u,m}^2 + \sigma_{u,p}^2$ by REML (Christensen and Lund 2010), fitting explicitly two separate random effects, $\mathbf{u}_m$+$\mathbf{u}_p$ with respective covariance matrices $\mathbf{H}\sigma_{u,m}^2$ and $\mathbf{A}\sigma_{u,p}^2$, and estimate explicitly $\sigma_{u,m}^2$ and $\sigma_{u,p}^2$.

14.5.1.2 Blending to make G invertible Matrix $\mathbf{G}$ is often not invertible, and therefore we can come up with "tricks" to make it invertible. The simplest trick is to add a small constant, say 0.01, to the diagonal of $\mathbf{G}$:

$$\mathbf{G}_w \leftarrow \mathbf{G} + 0.01\mathbf{I}$$

Which gives $\mathbf{G}_w$ nearly identical to $\mathbf{G}$: $\mathbf{G}_w \approx \mathbf{G}$ and therefore $\mathbf{G}_w \approx \frac{\mathbf{ZZ}'}{\sum(2p_j q_j)}$.

Alternatively, we can use the "blending" with the relationship matrix as above

$$\mathbf{G}_w \leftarrow (1-\alpha)\,\mathbf{G}{+}\alpha\mathbf{A}_{22}$$

Here, $\mathbf{G}_w$ is "less" close to the original $\mathbf{G}$ and to $\frac{\mathbf{ZZ}'}{\sum(2p_j q_j)}$. This has repercussions for the backsolving of SNP effects. For this reason, "blending" with the relationship matrix is becoming more cumbersome for some computational strategies such as APY or SNP-based models.

14.6 Compatibility of G and A

14.6.1 Fitting G to A

Based on the way $\mathbf{H}$ is constructed, the central element is $\mathbf{G} - \mathbf{A}_{22}$, which implies both matrices should be compatible (Legarra *et al.* 2014b). VanRaden (2008) stated that $\mathbf{G}$ had to be constructed with base allele frequencies. However, genomic relationships can be biased if $\mathbf{G}$ is constructed based on allele frequencies other than the ones calculated from the base population (VanRaden 2008). Allele frequencies from the base population are not known because of the recent recording of pedigrees (i.e., the base population *per se* is unknown). In some cases, such as dairy cattle, base allele frequencies can be inferred – in other cases such as pigs or sheep, they can not. In the typical case, the allele frequencies are observed a few generations after the start of predigree recording. Allele frequencies p will tend to fixation to the closest extreme (1 or 0) if they are neutral, and towards the favorable allele if they have effects.

Most commonly, allele frequencies used to construct $\mathbf{G}$ are based on the observed population. This brings two problems. The first one is that the machinery of "linear imputation" of Christensen and Lund (2010) fails: the expression $\widehat{\mathbf{Z}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{Z}_2$ as the mean is not 0.

The second problem is that we assumed in the development that the expectation of breeding values for genotyped animals is 0. If the population is under selection, recent animals should have higher genetic values than the base generation. Thus, the assumption $\mathbf{u}_2 \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$. A more sensible approach is to posit a mean for these animals: $\mathbf{u}_2 \sim N(\mathbf{1}\mu, \mathbf{G}\sigma_u^2)$ (Vitezica et al. 2011). In the chapter about genomic relationships, we have seen that if μ is a random effect, this leads to a genomic relationship matrix: $\mathbf{G}^*{=}\mathbf{G} + \mathbf{11}'a$ where $a = \overline{\mathbf{A}}_{22}{-}\overline{\mathbf{G}}$; use of a leads to $\mathbf{u}_2 \sim N(\mathbf{0}, \mathbf{G}^*\sigma_u^2)$ with mean 0. Equivalently, Vitezica et al. (2011) show that a model with explicit estimate of μ leads to the same solution. The idea has been considered also with μ fit as a fixed effect (Hsu *et al.* 2017).

In addition, and as shown in the chapter of genomic relationships, there is a decrease in the genetic variance. This leads to very similar adjustments

$$\mathbf{G}^* = a + b\mathbf{G}$$

with a and b inferred from 2 systems of equations:

$$\frac{\mathrm{tr}\,(\mathbf{G})}{m} b + a = \frac{\mathrm{tr}\,(\mathbf{A}_{22})}{m}$$

$$a + b\overline{\mathbf{G}}{=}\overline{\mathbf{A}}_{22}$$

This adjustments account for genotyped animals being more related through $\mathbf{A}_{22}$ than $\mathbf{G}$ is able to reflect.

14.6.2 Fitting A to G

A second class of method is also detailed in the same chapter, and leads to modify $\mathbf{A}_{22}$ to resemble $\mathbf{G}$ rather than the opposite. Christensen (2012) argued that using *any* allele frequency is subject to uncertainty, and after algebraic integration of allele frequencies he devised a new pedigree relationship matrix, $\mathbf{A}\left(\gamma\right)$ whose founders have a relationship matrix $\mathbf{A}_{\text{base}} = \gamma + \mathbf{I}(1 - \gamma/2)$. Parameter s, used in $\mathbf{G}$=$\mathbf{ZZ}'/s$ can be understood as the counterpart of $2\Sigma p_i q_i$ (heterozygosity of the markers) in the base generation.

Further developments by Garcia-Baccino et al. (2017) showed that the unknown s reduces to $s = 2\sum 0.5^2 = m/2$, simply the number of markers divided by 2, and $\gamma = 8\text{var}\left(p_{\text{base}}\right)$, where p_{base} are the (say, 50K) base allele frequencies. It may be argued that we still need to estimate p_{base} ; true, but using *one* inferred parameter γ to modify $\mathbf{A}$ instead of using 50,000 inferred parameters in p_{base} to construct $\mathbf{G}$ seems a safer strategy. Also, γ needs to be estimated only once and not at each run SSGBLUP when new genotypes are available. Both papers present methods to estimate γ, but the simpler strategy is to estimate p_{base} using Gengler's method and then compute $\widehat{\gamma} = 8\text{var}\left(\widehat{p}_{\text{base}}\right)$. The method has interesting connections with Wright's F_{st} theory (Garcia-Baccino *et al.* 2017) and with genetic distances across populations.

14.6.3 Unknown Parent Groups

Imagine that you are in Europe in 1975 and there is a massive introduction of selected US Holstein bulls into the less-selected European "Friesian" population, as described for instance here. US data are not available, but you want to fit in your model the fact that some groups of "parents" are really different from others. What you do – you assign a "pseudo-parent" at the top of the US pedigree, and a "pseudo-parent" at the top of the European pedigree. These pseudo-parents are not animals per se; they are conceived as infinite pools of animals to draw descendants from; their descendants are not inbred nor related.

This is, presented in a caricature, the origin of Genetic Groups or Unknown Parent Groups (UPGs hereinafter). For more details, go to regular texts on genetic evaluation (e.g.(Mrode and Thompson 2005)) and to Quaas (1988) and Thompson (1979) classic papers. Application of UPG goes through a special matrix $\mathbf{A}^*$ that accomplishes the role of "regular" $\mathbf{A}^{-1}$ in BLUP. This matrix is:

$$\mathbf{A}^* = \begin{pmatrix} \mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q} & -\mathbf{Q}'\mathbf{A}^{-1} \\ -\mathbf{A}^{-1}\mathbf{Q} & \mathbf{A}^{-1} \end{pmatrix}$$

and includes UPG in the upper corner. This matrix *has no inverse* and therefore the "relationship" matrix with groups does *not* exist (this is indeed awkward).

Unknown Parent Groups are used extensively to model:

1. Missing parentship, as in sheep (father is often unknown). Genetic Groups are often defined by year of birth to model genetic progress.
2. Importations, or introduction of foreign material (as in pig companies). Genetic Groups are often defined by country of origin.
3. Crosses (e.g. Angus x Gelbvieh). Genetic Groups are often defined by breed.

The key bit for what we want in these notes is to realize that, in the theory of Unknown Parent Groups,

$$p\left(\mathbf{u}_2\right) = N\left(\mathbf{Qg}, \mathbf{A}\sigma_u^2\right)$$

with $\mathbf{g}$ the "breeding value" of the unknown parent group, $\mathbf{Q}$ containing fractions of origin, for instance an animal could have 10% of its genes from "Lacaune France 2002", 15% of its genes from "Lacaune 2000", 25% from "Lacaune 2004", and 50% of its genes from "Lacaune 1996".

Our problem now is that, when doing genomic predictions, we assumed

$$p\left(\mathbf{u}_2\right) = N\left(\mathbf{0}, \mathbf{G}\sigma_u^2\right)$$

Then, why do we need UPG if we have $\mathbf{G}$ and $\mathbf{G}$ is replacing the pedigree information? How can we conciliate these two definitions for $\mathbf{u}_2$? The developments that lead to SSGBLUP fail because we assumed that

$$p(\mathbf{u}_1|\mathbf{u}_2) = \ N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

But this is no longer true in presence of Unknown Parent Groups (in part because of the fact that $\mathbf{A}^*$ cannot be inverted). Some options were reviewed by Misztal et al. (2013). The idea of metafounders was published by Legarra et al. (2015a). We discuss them quickly here. *This is still an open topic for research.*

14.6.3.1 Truncate pedigree and data The simpler option is to remove old data as in (Lourenco *et al.* 2014). If your UPGs model genetic trend, but you're only interested in recent animals, you can remove "old" data and then trace back your pedigree from your data by 3 generations. And you don't use UPGs. This is simple – yet efficient. In addition, in this case $\mathbf{A}$ and $\mathbf{G}$ match almost automatically because the base generation in the (truncated) pedigree is very close to the genotyped animals.

14.6.3.2 Approximate UPGs The default option in blupf90, when there are UPGs, is to build a matrix H^* as follows:

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

Where $\mathbf{A}_{22}$ is constructed as if UPG don't exist, which is an approximation. This works usually well unless you have <u>many</u> animals (e.g. cows or sheep) that have some unknown parent. In this case, there are three other solutions.

14.6.3.3 Fitting UPG as covariates This is simple yet cumbersome. The model fit for genomic evaluation does *not* use $\mathbf{A}^*$, but it fits UPG as covariates:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Q}\mathbf{g} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

With a matrix with covariates and using "regular" $\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$ for $\mathbf{u}$. The final estimate is $\widehat{\mathbf{u}}^*{=}\widehat{\mathbf{u}}{+}\mathbf{Q}\widehat{\mathbf{g}}$.

This model is not quite right. If $\mathbf{G}$ "contains" all needed information, for genotyped animals the group effect is counted twice: in $\mathbf{Q}$ and in $\mathbf{G}$. Taking it to the extreme, in a GBLUP context, it would make no sense to put covariates.

14.6.3.4 Fitting "exact UPGs" This is equivalent to the previous solution, but the Qg part is embedded within $\mathbf{H}^*$:

$$\mathbf{H}^* = \mathbf{A}^* + \begin{pmatrix} \mathbf{Q}_2' \left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) \mathbf{Q}_2 & \mathbf{Q}_2' \left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) \\ \left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) \mathbf{Q}_2 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$$

14.6.3.5 Metafounders Legarra et al. (2015a) suggested a different point of view. First, to apply Christensen (2012) theory to construct $\mathbf{G}$ as $\mathbf{G}_{05}$, "pretending" that all allele frequencies are 0.5. Second, to substitute UPGs by pseudo-animals called metafounders, that have strange relationship coefficients called $\mathbf{\Gamma}$. For instance, an example of these coefficients is $\mathbf{\Gamma} = \begin{pmatrix} 0.7 & 0.4 \\ 0.4 & 0.5 \end{pmatrix}$. These coefficients need to "match" observed relationships in $\mathbf{G}_{05}$; for instance, if Landrace and Yorkshire have an average of 0.4 relationship in $\mathbf{G}_{05}$, then $\gamma_{\text{Landrace},\text{Yorks}h\text{ire}}$ should be 0.4. If Yorkshire animals that are unrelated based on pedigree have an average of 0.7 relationship in $\mathbf{G}_{05}$, then $\gamma_{\text{Yorks}h\text{ire},\text{Yorks}h\text{ire}} = 0.7$. Based on this, we create $\mathbf{A}^{\Gamma}$ and then we use

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

We achieve two things here: (1) modelling different means and (2) automatic compatibility between $\mathbf{A}$ and $\mathbf{G}$. This strategy was used by Xiang et al. (2017).

15 Unknown Parent Groups and Metafounders

15.1 Quick tour of Unknown Parent Groups

There are no easy readings for this. Mrode and Pocrnic (2023) describe Metafounders in chapter 3 and Unknown Parent Groups (also called Genetic Groups, Phantom Parent Groups, or Westell-Thompson groups) in chapter 4. Masuda *et al.* (2022) discussed both from the point of view of genetic evaluation. The best account about UPGs is still, possibly, Quaas (1988) whereas . Elzo (2008) gives a good historical perspective. Thompson (1979) explains in detail why this was needed, and the concept (now ignored) that bulls were "grouped" when they entered the "stud" and they were evaluated *within* cohort (which now would be considered a heresy).

The UPGs are used to assign animals with missing pedigrees to different genetic levels. In practice, there are very few populations with fully complete pedigrees. This is the case for instance of the Merinos de Rambouillet, with complete pedigree dating back 200 years, or of rabbit lines, with pedigrees spanning 50 or more generations. In pigs and chicken, parentages are usually well recorded, but there are introductions of animals from other populations. In ruminants, sometimes the dam is not recorded (in cattle) and sometimes the sire (in sheep). Many countries import animals from foreign countries. So a method is needed to deal with that.

The theory of UPGs was conceived in the 70's to deal, for instance, with different cohorts of proven bulls or importation of foreign animals with different genetic levels. An initial, sensible assumption was to assume that "nothing" was known on the different UPGs and therefore "everything was possible". This in statistical reasoning means that the UPG effects g could potentially be very large, which in Bayesian arguments can be translated as $p(g) \propto b$ for b a constant, which means that g can go from $-\infty$ to $+\infty$. Or, in other words, there was no variance assumed for g, which means that UPGs were fit as "fixed" effects. In practice it is often assumed, mostly by computational convenience, that UPG effects have a certain *a priori* variance typically assumed to be 1 times the genetic variance, in other words the same as an animal: $p(g \mid \sigma_u^2) = N(0, \sigma_u^2)$. For a number of UPGs stacked in vector $\mathbf{g}$, it is also often assumed that they are uncorrelated to each other: $p(\mathbf{g} \mid \sigma_u^2) = N(\mathbf{0}, \mathbf{I}\sigma_u^2)$.

The value of a UPG in its modern form (Quaas 1988) halves at each generation, so if Holstein1990 means +200 kg of milk and Holstein2000 means +500 kg of milk, an animal that is "pure" Holstein1990 has a prior value of +200, an animal that is "pure" Holstein2000 has a prior value of +500, and an animal that is 1/4 Holstein1990 and 3/4 Holstein2000 has a prior value of +425. This is expressed as (following notation in (Quaas 1988))

$$\mathbf{u} = \mathbf{Q}\mathbf{g} + \mathbf{u}^*$$

where $\mathbf{u}$ are total breeding values, $\mathbf{g}$ are UPG effects, $\mathbf{Q}$ are UPG fractions (e.g. our 3/4/ and 1/4 above) and $\mathbf{u}^*$ are deviations, which are assumed to follow regular pedigree relationship rules, i.e. $Var(\mathbf{u}^*) = \mathbf{A}\sigma_u^2$.

The beauty and the danger of UPGs treated in this manner is that they result in very simple BLUP formulation. In fact, it is enough to use the "normal" MME using a matrix $\mathbf{A}^*$ which is "like" the inverse of a relationship matrix but not really an inverse. For instance, for UPGs treated as fixed, $\mathbf{A}^*$ is not positive definite and therefore it cannot be seen as the inverse of the covariance matrix of $\mathbf{u}$.

The fact that we can still use regular pedigree relationship rules for $\mathbf{u}^*$ follows from the fact that contrary to regular animals, under those assumptions UPG do not "create" inbreeding or have Mendelian sampling. In other words, there is no influence of the origin on the genetic variation of an animal. This is hard to believe for several reasons, first because genetic variance reduces within a breed due to drift (inbreeding) (Kennedy 1991) second, because parental breeds have different variances, and, third, because crossbred animals beyond F1s (i.e. F2, 3/4, etc) have higher variance than the average of the parental breeds (this is known as segregation variance, e.g. Garcia-Cortes and Toro (2006)). These problems were also discussed by VanRaden (1992).

UPGs may result in trouble for several reasons. First, because it is esy to do so, people fit just too many (Quaas 1988; Kudinov *et al.* 2020). Second, lack of accuracy in estimating UPG hampers genetic progress (Kennedy 1981; Phocas and Laloë 2004) and jeopardizes estimate of genetic trend Legarra and VanRaden (2023). Third, assuming no relationship among groups is just wrong. Two Danish Jersey animals with unknown parents even from different years should *a priori* be more similar to each other than to a US Jersey with unknown parents.

The fourth inconvenient I will describe in the next section.

15.2 Unknown Parent Groups and Single Step GBLUP

There has been considerable progress on how to fit UPGs in Single Step GBLUP. A good review is available in Masuda *et al.* (2022) which I recommend to read with Strandén *et al.* (2022). The basic model is an extension of Vitezica *et al.* (2010) , explained in previous chapters, in which a mean μ is used to differentiate genetic and genomic bases, and of Hsu *et al.* (2017) in which several means are considered but UPGs are not considered. In the approach of Strandén *et al.* (2022), there are several such $\boldsymbol{\mu}$ means and there are several UPG effects $\mathbf{g}$. When both are defined in the same manner, i.e. we model n UPGs and n μ's, matrices cancel out and we end up with the equations called "altered QP model" in Masuda *et al.* (2022). In these MME we use an "inverse" of the relationship matrix , which are (UPGs are on top, non genotyped animals in the middle, and genotyped animals in bottom):

$$\mathbf{H}^* = \mathbf{A}^* + \begin{pmatrix} \mathbf{Q}_2' \left(-\mathbf{A}_{22}^{-1}\right) \mathbf{Q}_2 & 0 & \mathbf{Q}_2' \left(\mathbf{A}_{22}^{-1}\right) \\ 0 & 0 & 0 \\ \left(\mathbf{A}_{22}^{-1}\right) \mathbf{Q}_2 & 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$$

where $\mathbf{Q}_2$ contains, *only for genotyped animals*, the fraction of each UPG.

An alternative, intuitive derivation is as follows. Misztal *et al.* (2013) derived MME using the so-called Westell-Thompson transformations that resulted in the following equations that in the UGA jargon we call "exact UPG":

$$\mathbf{H}^{''exact''} = \mathbf{A}^* + \begin{pmatrix} \mathbf{Q}_2' \left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) \mathbf{Q}_2 & 0 & \mathbf{Q}_2' \left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) \\ 0 & 0 & 0 \\ \left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) \mathbf{Q}_2 & 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$$

These expression show that $\mathbf{Q}_2$ multiplies $\mathbf{G}^{-1}$ but this is counter-intuitive. The point of $\mathbf{Q}_2$ is to correct for missing pedigree and different original populations but we have SNPs why we need this? So people at UGA came with the idea of dropping $\mathbf{G}^{-1}$ from the multiplications of $\mathbf{Q}_2$, e.g. Bradford *et al.* (2019) (where it is called "ssGBLUP with UPG for A-1") or Masuda *et al.* (2022) , and we end up again with

$$\mathbf{H}^* = \mathbf{A}^* + \begin{pmatrix} \mathbf{Q}_2' \left(-\mathbf{A}_{22}^{-1}\right) \mathbf{Q}_2 & 0 & \mathbf{Q}_2' \left(\mathbf{A}_{22}^{-1}\right) \\ 0 & 0 & 0 \\ \left(\mathbf{A}_{22}^{-1}\right) \mathbf{Q}_2 & 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$$

15.3 Metafounders

The fourth inconvenient of UPGs came with (Single Step) GBLUP. If you genotype animals from different periods and places, markers don't care about pedigree to tell you that some animals are closer than others. So in our example, markers will tell that (on average) Danish Jersey animals will be more similar to each other than to US Jersey animals and this, regardless of whether pedigree is known or not.

See for instance the Figure showing the PCA of animals from seven Spanish and French dairy breeds (Legarra *et al.* 2014a) - it is clear that animals from Manech (M) and Latxa (L) are closer to each other than animals from Lacaune, and that LCR and MTR are very close.

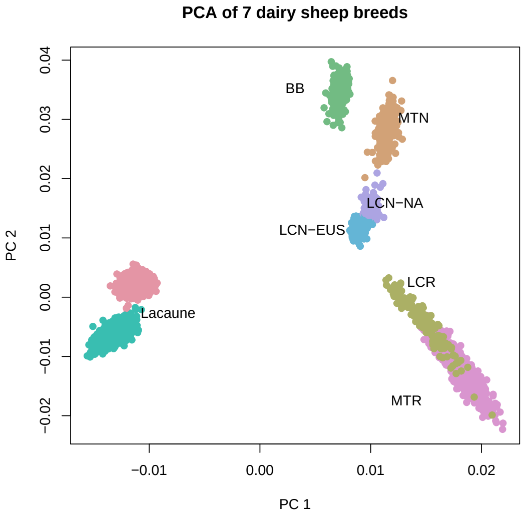

Figure 20: PCA of seven dairy sheep breeds http://dx.doi.org/10.3168/jds.2013-7745

So whereas the "altered QP model" described before corrects correctly for the averages means it does not really make any attempt to consider similarities of populations. It is nevertheless possible to describe how populations are related, and evolutionary and population geneticists have been doing this for years. The full development is a bit twisted, but eventually very usable math was achieved. First, note simply that base populations are "similar", which one may draw using PCA (as in Figure before), using phylogenetic trees (Figure from (Bonhomme *et al.* 2010)), overlapping boxes (see Picture below from (Legarra *et al.* 2015a)) or scatterplots of allele frequencies (this is from work at CDCB). But we need a more mathematical definition to be able to derive what we want: actual numbers to be used in animal breeding.

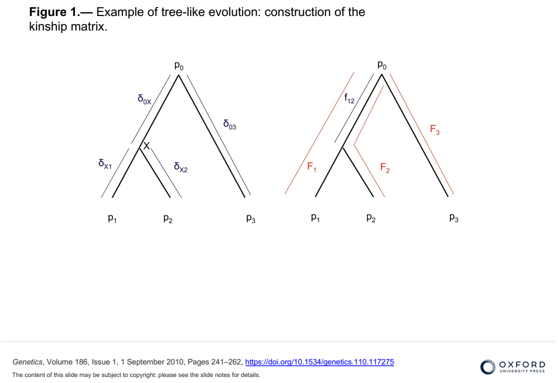
Figure 1.— Example of tree-like evolution: construction of the kinship matrix.

Genetics, Volume 186, Issue 1, 1 September 2010, Pages 241–262, https://doi.org/10.1534/genetics.110.117275

The content of this slide may be subject to copyright: please see the slide notes for details.

Figure 21: Tree-like evolution

15.4 Definition of metafounders based on allele frequencies

The following is a condensation of (Legarra *et al.* 2024a; b).

First we impose that substitution effects α are defined across several populations, so by construction they are "portable" across them. This is an old idea from Stuber and Cockerham (1966) that is very well developed in Christensen *et al.* (2015) whose reading we strongly recommend for more detail. When α is defined across several populations (e.g. Latxa and Lacaune) these differences are picking up genetic differences across breeds - so if a QTL is fixed in one breed for one allele and fixed in the other breed for the alternative allele, when we work within breed we do

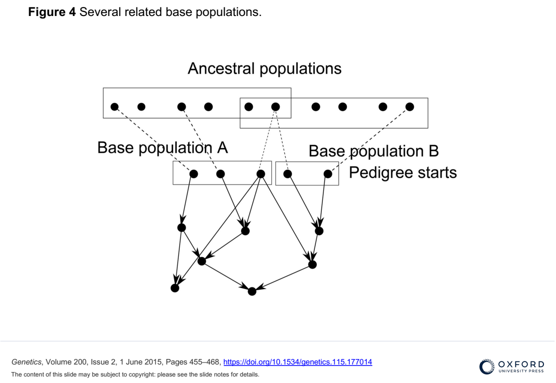

Figure 4 Several related base populations.

Genetics, Volume 200, Issue 2, 1 June 2015, Pages 455–468, https://doi.org/10.1534/genetics.115.177014

The content of this slide may be subject to copyright: please see the slide notes for details.

OXFORD UNIVERSITY PRESS

Figure 22: Overlapping boxes

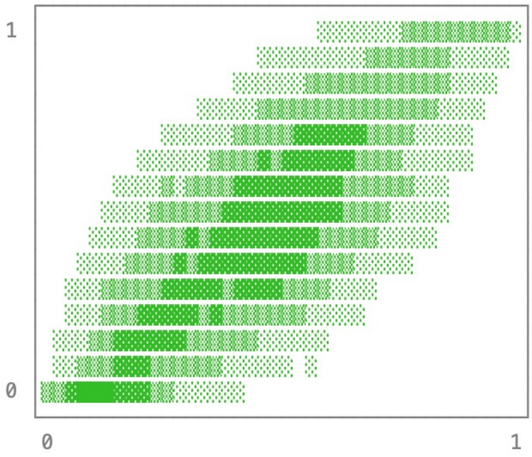

Figure 23: Scatterplot of allele frequencies of two dairy cattle breeds

not “see” this QTL but when we work across breeds we do “see” this QTL. And the substitution effect α trained across breeds will pick up these differences.

In this manner, the average genotypic value of a population i is simply $\mu_i = 2\sum_j p_{(i,j)}\alpha_j$. We have a collection of different populations so the different means are for populations $1, 2...n$ as follows:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ ... \\ \mu_n \end{pmatrix} = \begin{pmatrix} 2\sum_j p_{(1,j)}\alpha_j \\ 2\sum_j p_{(2,j)}\alpha_j \\ ... \\ 2\sum_j p_{(n,j)}\alpha_j \end{pmatrix}$$

So we have the population means and we want to compute some sort of relationship among them. We will call these relationships $\boldsymbol{\Gamma}$. we need to define a covariance across populations, remember that $Cov(X, Y) = E(XY) - E(X)E(Y)$, but what are $E(X)$ and $E(Y)$ here? Thus we need to define a *a priori* “average” or “baseline” of population means - what is this baseline? If populations diverge from an ancestral one (and all do), and we know the starting point of lineages, then we can use the mean at that point. Thus all genetic evaluations refer to a “base population” from which animals, pedigrees, meiosis, stem. However, when dealing with several base populations, it is hard to define one. For instance, in crosses of *Bos Indicus* and *Bos Taurus* we should go back to the ancestral *Bos* probably a few thousand years ago. In crosses of dairy breeds we probably need to go back to 1700 or so. For instance if we deal with Merino sheep populations in Australia, New Zealand, France and South Africa we could use the 1790 Rambouillet flock from which all those breeds diverge. But that’s just too complicated.

So, a convenient point of view is that we just treat allele frequencies as nuisances and we assume (or force) $E(p) = 0.5$ because these are biallelic markers (Christensen 2012). Then we have that $E(\mu) = 2\sum_j 0.5\alpha_j$. This means that **we define all relationships with respect to an ideal population in Hardy-Weinberg with maximum heterozygosity**. This population does not exist, it is merely a convenience.

We can then develop the covariances across populations in a manner similar as we did for GBLUP and genomic relationships:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ ... \\ \mu_n \end{pmatrix} = \begin{pmatrix} 2\sum_j p_{(1,j)}\alpha_j \\ 2\sum_j p_{(2,j)}\alpha_j \\ ... \\ 2\sum_j p_{(n,j)}\alpha_j \end{pmatrix} = \begin{pmatrix} 2\mathbf{p}_{(1)}\boldsymbol{\alpha} \\ 2\mathbf{p}_{(2)}\boldsymbol{\alpha} \\ 2\mathbf{p}_{(n)}\boldsymbol{\alpha} \end{pmatrix}$$

In fact we can conceive the *row vectors* $2\mathbf{p}$, containing two times population frequencies, as “genotypes” of the population. Then we center by the allele frequencies:

$$\boldsymbol{\mu} - E(\mu) = \begin{pmatrix} 2\sum_j p_{(1,j)}\alpha_j \\ 2\sum_j p_{(2,j)}\alpha_j \\ ... \\ 2\sum_j p_{(n,j)}\alpha_j \end{pmatrix} - \begin{pmatrix} 2\sum_j 0.5\alpha_j \\ 2\sum_j 0.5\alpha_j \\ ... \\ 2\sum_j 0.5\alpha_j \end{pmatrix} = \begin{pmatrix} (2\mathbf{p}_{(1)} - \mathbf{1})\boldsymbol{\alpha} \\ (2\mathbf{p}_{(2)} - \mathbf{1})\boldsymbol{\alpha} \\ ... \\ (2\mathbf{p}_{(n)} - \mathbf{1})\boldsymbol{\alpha} \end{pmatrix}$$

So we have defined now that $(2\mathbf{p}_{(i)} - \mathbf{1})$ is the “centered genotype” of population i. We keep following the logic of the genomic relationship matrix in VanRaden (2008) so that genomic relationship between populations i and k is the scaled crossproduct of centered genotypes, like in VanRaden (2008) :

$$\Gamma_{i,k} = \frac{1}{s}(2\mathbf{p}_{(i)} - \mathbf{1})(2\mathbf{p}_{(k)} - \mathbf{1})'$$

but what is s ? s is equal to $2\sum_j p_j q_j$ at the “ideal population” in which allele frequencies are all 0.5, so we get $2\sum_j 0.5 \times 0.5$ resulting in $s = \frac{m}{2}$ for m number of markers. The final expression is then

$$\Gamma_{i,k} = \frac{2}{m}(2\mathbf{p}_{(i)} - \mathbf{1})(2\mathbf{p}_{(k)} - \mathbf{1})'$$

and of a population with itself, it is

$$\Gamma_{i,i} = \frac{2}{m}(2\mathbf{p}_{(i)} - \mathbf{1})(2\mathbf{p}_{(i)} - \mathbf{1})'$$

we may put this in matrix algebra form like we do for **G** matrix:

$$\mathbf{\Gamma} = \frac{2}{m}(2\mathbf{P} - \mathbf{1}\mathbf{1}')(2\mathbf{P} - \mathbf{1}\mathbf{1})'$$

where **P** contains in each row the allele frequencies of each population.

In fact, these expressions using allele frequencies are also "hidden" in the Appendix of Christensen *et al.* (2015) .

15.4.1 Relationships with other metrics of genomic similarity and inbreeding

This is fully developed in Legarra *et al.* (2024b) . There are a few metrics to describe relationships across populations, the better known one is the differentition index F_{ST} . The F_{ST} of two populations is

$$F_{ST(i,k)} = \frac{\sum_i((p_{i,j} - p_{k,j})^2)}{\sum_i((p_{i,j}q_{i,j}) + (p_{k,j}q_{k,j}))}$$

It can be shown that this is equal to

$$F_{ST(i,k)} = \frac{\frac{\Gamma_{i,i}}{8} + \frac{\Gamma_{i,i}}{8} - \frac{\Gamma_{i,k}}{4}}{\frac{1}{2} - \frac{\Gamma_{i,k}}{4}}$$

The F_{ST} explains how much of the variance of the (hypothetical) F1 population is due to mixing populations and not to the variance within populations (this is of course Wright's original interpretation).

As for inbreeding, if $\Gamma_{i,i}$ is a relationship coefficient, then $F_i = \Gamma_{i,i} - 1$ can be seen as an inbreeding coefficient – a measure of homozygosity of the population i, not of any individual. Some math gives:

$$F_i = \Gamma_{i,i} - 1 = \quad 1 - 4\overline{\left(2p_{(i,j)}q_{(i,j)}\right)}$$

If average heterozygosity $\overline{(2p_{i,j}q_{i,j})} = 0$, then $F_i = 1$, meaning that there is complete inbreeding and lack of heterozygosity. If average heterozygosity (under HWE conditions) is maximal: $\overline{(2p_{i,j}q_{i,j})} = 0.5$, then inbreeding $F_i = -1$, meaning complete heterozygosity (under HWE conditions).

Typical values of $\mathbf{\Gamma}$ are a bit surprising because they are high, even across breeds (for instance: 0.7) but this is an artifact from using 0.5 as an absolute reference.

15.5 Pedigree and pedigree-genomic relationships with metafounders and use in genetic evaluation and ssGBLUP

When we know $\mathbf{\Gamma}$, we can derive relationships $\mathbf{A_\Gamma}$ and its inverse $\mathbf{A}_{\mathbf{\Gamma}}^{-1}$ just following the tabular rule and Henderson's algorithm with small modifications (Legarra *et al.* 2015a). This is a convenient matrix to analyse complex populations. Compared to the matrix $\mathbf{A}^*$ used with UPGs, it has a few advantages:

- it is invertible and positive-definite, by construction
- it models more properly if, or how, populations are related to each other
- when you fit UPGs as random you need to assign them a variance, usually $\mathbf{I}\sigma_u^2$, but this value is arbitrary. Here the counterpart of this scale is $\mathbf{\Gamma}\sigma_u^2$, which is based on genetic considerations

There are two inconvenients:

- how to estimate $\mathbf{\Gamma}$ in practice
- the rescaling of the genetic variance

As for the inclusion of metafounders in , it is just an extension of the usual SSGBLUP theory:

$$\mathbf{H}_{\mathbf{\Gamma}}^{-1} = \mathbf{A}_{\mathbf{\Gamma}}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{05}^{-1} - \mathbf{A}_{\mathbf{\Gamma}22}^{-1} \end{pmatrix}$$

The notation $\mathbf{G}_{05}^{-1}$ implies that $\mathbf{G}_{05}$ is computed "pretending" that allele frequencies are 0.5, i.e. as a crossproduct of genotype readings in the format $-1, 0, 1$ and multiplied by $\frac{2}{m}$, then blended with $\mathbf{A}_{\mathbf{\Gamma}22}$.

Contrary to the equations for UPGs before, there is no $\mathbf{Q}_2$ matrix here. Metafounders equations are in the $\mathbf{A}_{\mathbf{\Gamma}}^{-1}$ part.

Once we have $\mathbf{H}_{\mathbf{\Gamma}}^{-1}$ you use it into ssGBLUP equations, e.g. for multiple traits:

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

The backsolving of marker effects has been described previously. The computation of reliabilities has been described previously and in large detail in Bermann *et al.* (2023) .

The *estimate of the metafounder* itself is a bit awkward to interpret but is well shown in Meyer *et al.* (2018) . The value of "0" in a BLUP or a ssGBLUP with metafounders corresponds to the ideal population with allele frequencies of 0.5. The solution of the metafounder measures the deviation of the metafounder with respect to the 0 of the ideal population.

15.6 Example with 2 metafounders

This is the small example in Legarra *et al.* (2024a). The pedigree is

```
ped
14×3 Matrix{Int64}:
  1  0  0
  2  0  0
  3  1  1
  4  1  1
  5  1  2
  6  2  2
  7  2  2
  8  3  4
  9  4  5
 10  6  7
 11   8   9
 12   5  10
 13   8   5
 14  11  12
```

Metafounders are “individuals” 1 and 2 from which all the rest come.

Genotypes are:

```
$ cat exo_genotypes_agamma
       5  11112121112121121102111121110101112021000000202222002222111111112111120220220002202202002
       7  112111121121102111112111121110112012021000112020002220222200022202022022220121001211110111002
       9  20222222202020220202222222222020200202200000020222200222200022202022022220220110111111102002
       11 111111121121102111112111121110112012021000111111112111122200022202022022220121001211110111002
```

The value of $\mathbf{G}_{05}$ (for animals 5,7,9 and 11) is:

```
G
4×4 Matrix{Float64}:
 1.156  0.156  0.956  0.356
 0.156  1.178  0.689  0.956
 0.956  0.689  1.8    0.867
 0.356  0.956  0.867  0.956
```

with inverse:

```
inv(G)
4×4 Matrix{Float64}:
  1.70496    0.794258  -0.925253  -0.589477
  0.794258   5.60418    0.486683  -6.34113
 -0.925253   0.486683   1.64926   -1.63825
 -0.589477  -6.34113   -1.63825    9.09283
```

we have $\mathbf{\Gamma}$:

```
Gamma
2×2 Matrix{Float64}:
 0.408151  0.367189
 0.367189  0.411571
```

we obtain the following values for $\mathbf{A}_{\mathbf{\Gamma}}^{-1}$ and $\mathbf{A}_{\mathbf{\Gamma}22}^{-1}$ (for animals 5,7,9 and 11)

```
A^Gamma^-1
  0.000078 seconds (806 allocations: 38.109 KiB)
14×14 Matrix{Float64}:
  14.92  -10.37  -1.24  -1.24  -0.62   0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
 -10.37   14.73   0.0    0.0   -0.62  -1.25  -1.25   0.0    0.0    0.0    0.0    0.0    0.0    0.0
  -1.24    0.0    1.86   0.62   0.0    0.0    0.0   -1.24   0.0    0.0    0.0    0.0    0.0    0.0
  -1.24    0.0    0.62   2.48   0.61   0.0    0.0   -1.24  -1.23   0.0    0.0    0.0    0.0    0.0
  -0.62   -0.62   0.0    0.61   3.09   0.0    0.0    0.61  -1.23   0.61   0.0   -1.23  -1.23   0.0
   0.0    -1.25   0.0    0.0    0.0    1.87   0.62   0.0    0.0   -1.25   0.0    0.0    0.0    0.0
   0.0    -1.25   0.0    0.0    0.0    0.62   1.87   0.0    0.0   -1.25   0.0    0.0    0.0    0.0
   0.0     0.0   -1.24  -1.24   0.61   0.0    0.0    3.72   0.62   0.0   -1.23   0.0   -1.23   0.0
   0.0     0.0    0.0   -1.23  -1.23   0.0    0.0    0.62   3.07   0.0   -1.23   0.0    0.0    0.0
   0.0     0.0    0.0    0.0    0.61  -1.25  -1.25   0.0    0.0    3.11   0.0   -1.23   0.0    0.0
   0.0     0.0    0.0    0.0    0.0    0.0    0.0   -1.23  -1.23   0.0    3.13   0.66   0.0   -1.31
   0.0     0.0    0.0    0.0   -1.23   0.0    0.0    0.0    0.0   -1.23   0.66   3.12   0.0   -1.31
   0.0     0.0    0.0    0.0   -1.23   0.0    0.0   -1.23   0.0    0.0    0.0    0.0    2.45   0.0
   0.0     0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0   -1.31  -1.31   0.0    2.63
```

```
A^Gamma_22^-1
  0.000059 seconds (112 allocations: 17.320 KiB)
4×4 Matrix{Float64}:
  1.53  -0.2   -0.97   0.04
 -0.2    0.95  -0.07  -0.13
 -0.97  -0.07   2.37  -1.18
  0.04  -0.13  -1.18   1.6
```

and it all sums to $\mathbf{H}_{\mathbf{\Gamma}}^{-1}$:

```
Hi
14×14 Matrix{Float64}:
  14.92  -10.37  -1.24  -1.24  -0.62   0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
 -10.37   14.73   0.0    0.0   -0.62  -1.25  -1.25   0.0    0.0    0.0    0.0    0.0    0.0    0.0
  -1.24    0.0    1.86   0.62   0.0    0.0    0.0   -1.24   0.0    0.0    0.0    0.0    0.0    0.0
  -1.24    0.0    0.62   2.48   0.61   0.0    0.0   -1.24  -1.23   0.0    0.0    0.0    0.0    0.0
  -0.62   -0.62   0.0    0.61   3.22   0.0    0.91   0.61  -1.17   0.61  -0.54  -1.23  -1.23   0.0
   0.0    -1.25   0.0    0.0    0.0    1.87   0.62   0.0    0.0   -1.25   0.0    0.0    0.0    0.0
   0.0    -1.25   0.0    0.0    0.91   0.62   5.89   0.0    0.44  -1.25  -5.38   0.0    0.0    0.0
   0.0     0.0   -1.24  -1.24   0.61   0.0    0.0    3.72   0.62   0.0   -1.23   0.0   -1.23   0.0
   0.0     0.0    0.0   -1.23  -1.17   0.0    0.44   0.62   2.29   0.0   -1.52   0.0    0.0    0.0
   0.0     0.0    0.0    0.0    0.61  -1.25  -1.25   0.0    0.0    3.11   0.0   -1.23   0.0    0.0
   0.0     0.0    0.0    0.0   -0.54   0.0   -5.38  -1.23  -1.52   0.0    9.52   0.66   0.0   -1.31
   0.0     0.0    0.0    0.0   -1.23   0.0    0.0    0.0    0.0   -1.23   0.66   3.12   0.0   -1.31
   0.0     0.0    0.0    0.0   -1.23   0.0    0.0   -1.23   0.0    0.0    0.0    0.0    2.45   0.0
   0.0     0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0   -1.31  -1.31   0.0    2.63
```

The upper 2x2 block of $\mathbf{H}_{\mathbf{\Gamma}}^{-1}$ corresponds to the 2 metafounders of the example.

15.7 Variance components when using metafounders

The theory of metafounders assumes that all animals are "related". This implies that we may need to scale the genetic variance used. I will show how this can be visualized with a simple simulation in R.

Assume that we have a very large population of 1000 animals, they tell us that they are unrelated and they have a genetic variance $\sigma_u^2 = 30$. So we're going to see if this is true simulating the BV from $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$.

```
set.seed(1234)
require(MASS)
sigma2u=30
myMu=rep(0,1000)
A=diag(1000)
BV=mvrnorm(1,mu=myMu,Sigma=A*sigma2u)
var(BV)
#[1] 29.84048
```

Well, this seems correct. How does it work with metafounders? Say that $\gamma = 0.7$. The matrix $A_{\mathbf{\Gamma}}$ has in this case values of $1 + \frac{\gamma}{2}$ in the diagonal and γ in the off-diagonals, so the simulation can be

```
set.seed(1234)
require(MASS)
gamma=0.7
sigma2u=30
myMu=rep(0,1000)
Agamma=matrix(gamma,1000,1000)
diag(Agamma)=(1+gamma/2)
BV=mvrnorm(1,mu=myMu,Sigma=Agamma*sigma2u)
var(BV)
#[1] 19.38168
```

so there's something smelly here. In fact the values in $A_{\mathbf{\Gamma}}$ makes animals *very* related to each other, so to compensate for that we need to be more "tolerant" to deviations, in other words to increase the variance. The correct factor derived in Legarra *et al.* (2015a) is in fact $\sigma^2_{u,related} = \sigma^2_{u,unrelated}/(1 - \frac{\gamma}{2})$

```
set.seed(1234)
require(MASS)
gamma=0.7
sigma2u=30
myMu=rep(0,1000)
Agamma=matrix(gamma,1000,1000)
```

```
diag(Agamma)=(1+gamma/2)
sigma2uRelated=sigma2u/(1-gamma/2)
cat(sigma2uRelated,"\n")
# 46.15385
BV=mvrnorm(1,mu=myMu,Sigma=Agamma*sigma2uRelated)
var(BV)
#[1] 29.81797
```

So the scaling seems to work correctly.

The factor $1/(1-\frac{\gamma}{2})$ can be interpreted as follows. $\gamma = 0$ means that animals are "unrelated" so the scale factor is 1. As γ tends to 1 animals are more and more "related" so the variance $\sigma^2_{u,related}$ needs to increase to "allow" for a larger genetic variance.

When there are several metafounders there is no equivalence between variances - for instance, animals from "crosses" of two metafounders have more variance than animals from a single one. An approximate equivalence derived at Legarra *et al.* (2015a) assumes that most animals are a complex cross (which is a very crude approximation) resulting in the scaling factor

$$\sigma^2_{u,related} = \sigma^2_{u,unrelated}/(1+\frac{\overline{diag(\Gamma)}}{2}-\bar{\Gamma})$$

whether this scaling is useful or something else is needed, is still subject to study. A better work would require estimation of the genetic parameters, but this is often complex for large data sets.

15.8 Estimation of $\mathbf{\Gamma}$

There have been a few methods proposed (Legarra *et al.* 2015a; Garcia-Baccino *et al.* 2017; Kudinov *et al.* 2020; Kudinov *et al.* 2022; Legarra *et al.* 2024a). If allele frequencies at each base population are well estimated we just use the equation:

$$\mathbf{\Gamma} = \frac{2}{m}(2\mathbf{P}-\mathbf{1}\mathbf{1}')(2\mathbf{P}-\mathbf{1}\mathbf{1})'$$

Typically we have animals from different populations, with complex pedigrees. Also typically, estimating base allele frequencies is difficult or inaccurate. There are many works on this that I will not mention - they can be found in Legarra *et al.* (2024a) .

I will describe briefly the *exact* ML approach, two approaches that are more general and the extensions to allow for large number of metafounders. Details are in (Kudinov *et al.* 2022; Legarra *et al.* 2024a)

The *exact* ML approach (Christensen and Legarra 2022; Legarra *et al.* 2024a) is useful when there is a single metafounder. Essentially consists in computing $\mathbf{G}_{05}$, $\mathbf{A_2 2}$, and then obtain a few statistics from which a cubic equation is derived, whose solution is the estimate.

The covariance function method from Kudinov *et al.* (2022) can be seen in two manners. It can be seen as based on an implicit model of evolution of allele frequencies with time or, aternatively, of evolution of genomic relationships with time. Essentially, it estimates $\mathbf{\Gamma}$ (named $\mathbf{K}$ in theo=ir work) in a reduced subset of metafounders and then it "expands" it to intermediate points using sort of interpolation matrices called $\mathbf{\Phi}$.

The method of Legarra *et al.* (2024a) observes that the combined (inverse) relationship matrix $\mathbf{H_\Gamma} = \left(\mathbf{A}_{\mathbf{\Gamma}}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{05}^{-1} - \mathbf{A}_{\mathbf{\Gamma}22}^{-1} \end{pmatrix}\right)^{-1}$ "projects" genomic relationships in $\mathbf{G}_{05}$ to the metafounders, and this projection is the new estimate of $\mathbf{\Gamma}$, so starting with an initial value of $\mathbf{\Gamma}$, better estimates can be obtained re-computing $\mathbf{H_\Gamma}$, extracting updated $\mathbf{\Gamma}$ and re-doing the blocks $\mathbf{A}^{-1}$ and $\mathbf{A}_{\mathbf{\Gamma}22}^{-1}$.

This can be done for a group of metafounders where there is enough information to estimate things well. Extension (interpolation) to many groups can be obtained assuming a regular increase of relationships, which can be inferred from the increase of pedigree inbreeding.

The two methods in Kudinov *et al.* (2022) and Legarra *et al.* (2024a) are actually quite similar in philosophy, they use the expression $\mathbf{\Gamma} = \frac{2}{m}(2\mathbf{P} - \mathbf{1}\mathbf{1}')(2\mathbf{P} - \mathbf{1}\mathbf{1})'$ and they model more or less explicitly changes of $\mathbf{\Gamma}$ across time. Thay have not been formally compared so far (May 2024).

16 Use of method LR to assess potential bias due to design of cross-validation analysis.

Andres Legarra, INRAE, 28 Oct 2021.

This project has received funding from the European Unions' Horizon 2020 Research & Innovation programme under grant agreement N°772787 -SMARTER.

16.1 Introduction

In genomic selection, use of early genomic proofs can lead to suboptimal selection decisions if there is bias (see below for description of bias). In sheep and goats, and in particular for traits expressed in females, there is a lack of good tools to evaluate the presence or absence of bias, and the methods to evaluate accuracy of genomic selection are suboptimal.

We do we concern about bias? Selection theory establishes that selection is optimal if each candidate to selection is compared fairly to each other. This means that across individuals, Estimated Breeding Values (EBV, $\hat{u}$) of the selected candidates is equal to the expectation of the (true) Breeding Values (BV, u). When animals are selected, this is true under two conditions: $\bar{u} = \bar{\hat{u}}$ and $cov(u, \hat{u}) = var(\hat{u})$, where the means and the covariances apply across the animals selected in an operation (i.e. at the time of selecting young male lambs). The property $cov(u, \hat{u}) = var(\hat{u})$ is needed because if the distribution of $\hat{u}$ of e.g. young animals is too (or not enough) spread, we will select too many (or too little) young animals. Note that at this point, these properties are not statistical and therefore are neither "frequentist" nor "Bayesian". [6]

These properties can be formalized as

(1) *equality of estimated and true means* :

$$\mathbf{1}'\hat{\mathbf{u}} = \mathbf{1}'\mathbf{u}$$

or equivalently $\frac{1}{n}\sum \hat{u}_i = \frac{1}{n}\sum u_i$ or still $\bar{\hat{u}} = \bar{u}$, and

(2) *slope of true on estimated equal to 1*

$$\frac{1}{n}\sum \left(\hat{u} - \overline{\hat{u}_i}\right)^2 = \frac{1}{n}\sum \left[\left(\hat{u}_i - \bar{\hat{u}}\right)(u_i - \bar{u})\right]$$

or equivalently $cov(u, \hat{u}) = var(\hat{u})$.

Henderson (Henderson (1975), Henderson (1982)) established that the two properties above hold, even if there is selection, *on expectation* for *one* animal across repeated conceptual sampling of its $(u, \hat{u})$. Then Legarra and Reverter (2018) proved that the proof applies to *sets* of EBVs from groups of animals, so we have that the two properties hold *on expectation* for *many* animals across repeated conceptual sampling. By the Law of Large Numbers, when the number of animals is large, a number converges to its expectation. This means that, for a *large* number of animals, $\bar{\hat{u}} = \bar{u}$ *must* hold empirically.

So the theory says that, without invoking some esoteric statistical framework, genetic evaluations should be unbiased. But how can we check this? We don't have u, only $\hat{u}$. In dairy cattle, they compare predictions vs. progeny proofs (or Daughter Yield Deviations) but in other species the number of offspring of each animal is small.

In addition, we're interested in finding out the accuracy of genomic prediction, i.e. $r(u, \hat{u})$. Again, it is difficult to obtain this number in small ruminant cases.

[6] The compensation that NZ farmers got in 2010 for using genomic bulls with biased genomic proofs did not know about priors or sampling distributions :-)

16.2 Bias due to using pre-corrected data or De-Regressed Proofs (DRP)

The following is extracted from Legarra and Reverter (2018). Often we have used precorrected phenotypes y^* or deregressed proofs, and compare predictions $\hat{y}$ with (precorrected) observations y^* (this method is sometimes called "predictability"). The estimator of accuracy is e.g. $r \approx cor(y^*, \hat{y})/h$ for h^2 the heritability (Legarra et al. 2008). But this ignores that precorrection generates a covariance structure in y^*, is *very* sensitive to low values of h^2, and it also ignores that animals used in these studies can be preselected (case for instance of elite males). This leads to paradoxes:

- $r > 1$ (observed in chicken)
- $r_{pedigree} > r_{genomic}$ (observed in dairy cattle for fertility)

16.2.1 Bias due to ignoring the effect of selection on genetic variance

It also ignores that candidates to selection have reduced genetic variance (Bijma (2012)). For instance, for prospective AI rams in dairy sheep, because they're highly selected, their genetic variance is less than the "normal" genetic variance [7].

Consider for instance that we made a study on growth in meat sheep in selected rams in a performance recording station. These rams are selected based on parent average and therefore their genetic variance, is, say, $k = 80\%$ of the populational one, and $h^2 = 0.3$. Through cross-validation we obtain $cor(y^*, \hat{y}) = 0.4$ and we conclude that $r \approx cor(y^*, \hat{y})/h = 0.73$. However this is incorrect because these animals were selected, so that *in these rams*, the heritability is actually $h^{2*} = \frac{kh^2}{1-(1-k)h^2} \approx 0.26$. Coupling in our equation $r \approx cor(y^*, \hat{y})/h^* = 0.78$, quite higher.

16.2.2 Bias due to pre-correction by fixed effects

There is a second, non-negligible source of bias. We use y^* (precorrected data) as it was "exact". This leads to overestimation of accuracies. In Legarra and Reverter (2018)) we worked out that for a balanced design with n_i records per contemporary group, the bias is such that the *relative* overestimation of accuracy is of order $\frac{1}{n_i}$. For instance:

- Dairy sheep: assume 25 animals / contemporary group. This leads to overestimation of accuracy by $1/25 = 4\%$. If $r \approx cor(y^*, \hat{y})/h = 0.73$, this r was overestimated by 4% so that the actual accuracy should be $r = 0.73(1 - 0.04) = 0.70$.
- Beef cattle: 5 animals / contemporary group. This leads to overestimation of accuracy of 20%

16.3 Method LR to the rescue

For all these reasons we want better methods to assess biases and accuracies.

Legarra and Reverter (2018), with further proofs by Bermann *et al.* (2021), extended the machinery developed by Henderson (1975) and Reverter *et al.* (1994/Jan/01) to infer biases and accuracies by splitting the data set. They defined *partial* (p) and whole (w) data sets, so the *partial* data set contains all information until a given date and the *whole* data set contains all information available for the analyst until a later date (not necessarily now). The procedure, called LR from Linear Regression [8] is described next.

16.3.1 LR in a nutshell

You have complete (*whole*) records, pedigree and (perhaps) markers. Consider a cut-off date. Records before these date make the *partial* data set: $\mathbf{y}_p$ whereas all records make the *whole* data set: $\mathbf{y}_w$. Then you run two genetic evaluation with either the *partial* data or the *whole* data, and you keep the entire pedigree and markers in both. In these manner, you have EBVs for all animals in both cases, $\hat{u}_p$ and $\hat{u}_w$ respectively.

[7] Note that the genetic variance is recovered when this animals mate to females in the next generation.

[8] The fact that the initials of the authors are LR is, of course, coincidental :-)

Then you compare EBVs of animals in *partial* and *whole* prediction. You don't include *all* animals; you consider contemporary animals with similar information, in which you have an interest (for instance, males candidates to selection). We call this *focal* animals or *focal* groups. See below for examples.

The comparison is very simple and it just consist in a series of statistics that can be easily computed. We propose several criteria. This can be found in Macedo *et al.* (2020b) which is the most up-to-date source. Note that in the following, whenever we put something like $cov(\hat{\mathbf{u}}_p, \hat{\mathbf{u}}_w)$ we mean a *scalar* (the "observed" covariance) and not a *matrix* (which is the sampling or prior distribution of the vector).

16.3.1.1 Bias This is measured using $\hat{\Delta}_p = \bar{\hat{\mathbf{u}}}_p - \bar{\hat{\mathbf{u}}}_w$. The expectation is 0 (no bias). A positive value means that animals with *partial* information are overevaluated.

16.3.1.2 Slope Also calledr over/underdispersion. This is measured using $\hat{b}_p = \frac{cov(\hat{\mathbf{u}}_p,\hat{\mathbf{u}}_w)}{var(\hat{\mathbf{u}}_p)}$ or, equivalently, computing the slope b_1 of the linear regression "whole on partial" $\hat{u}_w \sim b_0 + b_1\hat{u}_p + \epsilon$. The expectation is 1 (no over- neither under-dispersion), values lower than 1 mean that selected candidates are overestimated. This is the kind of bias commonly reported in dairy cattle studies.

A very small example with 5 individuals follows:

```
# these are actually 5 "proven" bulls
EBV2018=c(999,849,831,953,764)
EBV2019=c(973,833,904,963,807)
Delta_p=mean(EBV2018)-mean(EBV2019) # -16.8
b_p=cov(EBV2019,EBV2018)/var(EBV2018) #0.71
aa=lm(EBV2019~EBV2018)
b_p=aa$coefficients[2] # 0.71
```

16.3.1.3 Accuracies There are two estimators of *relative* accuracies and two estimators of *absolute* accuracies.

- The first statistic is the correlation between *partial* and *whole* EBVs: $\hat{\rho}_{wp} = \frac{cov(\hat{\mathbf{u}}_p,\hat{\mathbf{u}}_w)}{\sqrt{var(\hat{\mathbf{u}}_p)var(\hat{\mathbf{u}}_p)}}$ (or simply `cor(u_p,u_w)`). This has expected value $\frac{acc_p}{acc_w}$ where acc means accuracy.

So, this estimates a *ratio* of accuracies and *not* the absolute accuracy. For instance, Values close to 1 indicate that "partial evaluation" was "as accurate" as "whole" evaluation, but both evaluations could be "little accurate".

A byproduct of $\hat{\rho}_{wp}$ is an estimator of the *relative increase in accuracy.* In effect, $\frac{1}{\hat{\rho}_{wp}} - 1$ has expected value $\frac{acc_w - acc_p}{acc_p}$, which is the relative increase in accuracy from *whole* to *partial* . For instance, boars can be evaluated for carcass traits *before* or *after* some full-sibs have been slaughtered, and $\frac{1}{\hat{\rho}_{wp}} - 1$ gives the relative increase in accuracy.

- The second statistic is $\hat{\rho}^2_{wp} = \frac{cov(\hat{\mathbf{u}}_p,\hat{\mathbf{u}}_w)}{var(\hat{\mathbf{u}}_w)}$, with expected value $\frac{acc^2_p}{acc^2_w}$, i.e. the ratio of *reliabilities* (squared accuracies).

Note that in fact this statistic $\hat{\rho}^2_{wp}$ is the slope b_1 of the regression "partial on whole": $\hat{u}_p \sim b_0 + b_1\hat{u}_w + \epsilon$. A note of caution of this statistic is that the expected value requires that the evaluation is unbiased ($\hat{b}_p = 1$) something that is *not* required for $\hat{\rho}_{wp}$. In principle, the value obtained for $\hat{\rho}^2_{wp}$ should be the square of the value obtained for $\hat{\rho}_{wp}$, but this is not true in practice as it holds only in expectation.

Both statistics are easy to compute:

```
rho_pw=cor(EBV2018,EBV2019) # 0.9101622
rho2_pw=cov(EBV2019,EBV2018)/var(EBV2019)# 1.15944
```

note that in this example $\hat{\rho}^2_{wp}$ is not admissible ($\hat{\rho}^2_{wp} > 1$ would mean that $acc_p > acc_w$) and this is because in the example $\hat{b}_p$ is not even close to 1.

- The first estimator of *absolute* reliability is an estimator of "selected" reliability: $\widehat{acc}_p^2 = \frac{cov(\hat{\mathbf{u}}_p,\hat{\mathbf{u}}_w)}{\sigma^2_{u*}}$. The denominator σ^2_{u*} is the variance of animals in the focal group (and not the variance of the base generation σ^2_u).

When animals are pre-selected (for instance, prospective AI rams selected based on parent average) their genetic variance σ^2_{u*} is less than the "normal" genetic variance σ^2_u. As an example, in Manech Tete Rousse, $\sigma^2_u \approx 500$ but $\sigma^2_{u*} \approx 350$ for young selected rams (for milk yield) Macedo *et al.* (2021). The variance σ^2_{u*} can be estimated using Gibbs Sampling (Sorensen *et al.* (2001),Macedo *et al.* (2020b)).

So, this equation gives the "selected" reliability (Bijma (2012),Dekkers (1992)), which is the "ability" to rank *within* those animals (more difficult when they are selected). However, we can't (easily) use this reliability to predict genetic progress, and we can't compare it with results in less selected animals, say, females. Also, the numbers do not match with those model-based, i.e. by Selection Index theory or from the inverse of the MME. The solution to this was given by Dekkers (1992) and Bijma (2012), and it leads to the last statistic:

- Unselected reliability, $\widehat{rel}_p = 1 - \frac{\sigma^2_{u*}}{\sigma^2_u}(1 - \widehat{acc}_p^2)$. The mathematical explanation of all this is quite boring and convoluted, but some detailed exapmles can be found in Macedo *et al.* (2020a) and Macedo *et al.* (2020b).

16.3.2 Examples of interpretation

Just to give a feeling of what these numbers look like and mean. When we did the first cross-validation approaches in dairy sheep, we used AI rams that after selection based on parent average, were used in progeny testing. In order to compute if genomic selection is good, we can evaluate these rams with ssGBLUP at birth, and then after progeny. The first result that we get is $\hat{\rho}_{wp}$, but it can't be used to predict genetic progress of genomic selection. Then we do better and we compute $\widehat{acc}_p^2$, but we obtain a number that is very small because the animals are highly selected. What we want is the accuracy of the genomic young rams *if they were not selected*, because a genomic selection scheme genotypes a wide basis of animals. To do so, we use the equation above to transform $\widehat{acc}_p^2$ into $\widehat{rel}_p$, which is the number that we want.

For instance, we obtained the following Table (Macedo *et al.* (2020b)):

Method	$\widehat{acc}_p^2$	$\widehat{rel}_p$	$\hat{\rho}_p^2 w$
BLUP-MF	0.22	0.53	0.32
SSGBLUP-MF	0.32	0.59	0.45

In the Table, the numbers of $\widehat{acc}_p^2$ seems "obviously wrong" because, for instance, for BLUP the reliability of the Parent Average from progeny-tested sire and phenotyped mother is usually close to 0.5, much higher than the observed numbers of ~ 0.25. However, the $\widehat{acc}_p^2 = 0.22$ in BLUP is the reliability *within the selected rams*, whereas the reliability *across all possible* rams is in fact $\widehat{rel}_p$, which has a value of 0.53 much closer to what we expect. The value of $\rho_p^2 w$ is more complicated to interpret. However, in the three columns it is obvious that SSGBLUP is more accurate than BLUP.

16.3.3 Practicalities

1. You evaluate the bias and accuracies for a category of animals. We call this *focal* animals or *focal* groups. These are contemporary animals for which the properties above hold, which are "exchangeable" (in other words, we're interested in the group, not in each individual animal) and in which we are interested. For instance young born rams can be a focal group. 1st-lambing females can be a focal group, and rams with first crop of daughters could be a focal group as well. But it is not a good idea to define a focal group composed of 50% progeny-tested rams 4 year old and 50% young animals that are 1 year old, because the

first will be more accurate and the second more shrunken towards the mean. To define the focal animals, the best way is to do it by analyzing the data: for instance, take all m rams born in year say 2010, and from them select those n that had offspring with record in 2014, but not before. Then the number of animals in the focal group is n.

2. Define dates in a way such that the focal individual will have more information in the *whole* than in the *partial* data set. For instance, young rams could have only parents' (and genomic) information in the *partial* data set and offspring information in the *whole* data set. First lambing females could have one record for milk yield in the *partial* and two records in the *whole* ; and similar cases. In the example above, the year of *partial* can be 2010 and the year of *whole* can be 2014.

3. The way we do this is using the data set and "looking forward" from each year. For instance, we take all rams born in 2014 that were used in AI , and few years later (say 2017) we find out which of these rams have daughters with milk yield. This defines a focal group for "partial"=2014 and "whole"=2017 We can do the same for 2014 vs. 2018, 2019, etc.

4. In these manner we have many "pairs" of *whole* and *partial.* For instance you can do "partial" at 2010, 2011,. . . and compare each of them vs. "whole" at 2014, 2015. . . . It is important to do several comparisons because the statistics vary a lot across years. Using several pairs of *whole* and *partial* requires automatic handling of files and data editing, that we do using automated scripts in R, Unix tools, and R scripting. The genetic evaluations, themselves, can be run in any software that you like.

5. In practice we delete "records" (milk yield, etc etc) based on the year, and we keep ALL pedigree and ALL markers. A more refined approach is to keep pedigree and markers only up to the same date, for instance if "partial"= March 2014 we should keep records, pedigree and markers up to March 2014 (because pedigree and markers were used to predict the young rams).

6. In genetic evaluations with Unknown Parent Groups, the EBVs are not estimable functions So you need to refer all EBVs to a common genetic base in order to infer "bias" or not. Typically the genetic base is something like "average EBV of all females born in 2010" or something like that.

All this requires good knowledge of the data sets, the breeding scheme (or the breed), and a good command of scripting and genetic

16.3.4 The importance of several comparisons

The Figure 1 below shows all the estimates of b_{pw} in Macedo *et al.* (2020b). For instance, in the X-axis we see the year of cut-off of *partial*, and the repeated points correspond to several *whole* years: 2010, 2011. . . It is clear that there is a large variation of b_{pw} due to chance, so to assess the unbiasedness of genetic evaluation one should do several pairs of *whole* and *partial* and not rely on a single study. For instance, year 2008 evaluation was clearly biased ($b_{pw} < 1$) whereas the other years were not.

16.3.5 Estimation of genomic accuracies vs pedigree ones

How do I infer if a genomic evaluation is more accurate than a pedigree based one? There are two manners.

The first approach is to use *whole* and *partial* as we have explained so far, and evaluate each run both BLUP and genomic prediction (e.g. SSGBLUP), which yields a Table like above. This gives quite complete information as we can compare accuracies across methods and at different times.

The second approach is to consider that the genomic evaluation has "more data" so the pedigree-based evaluation is *partial* and the genomic evaluation is *whole.* The records $\mathbf{y}$ are the same. in both. Then the statistics above describe the ratio, increase, or absolute accuracies. For instance if we observe $\hat{\rho}_{pw} = 1$, it means that adding genotypes did not change anything. However, if we obtain $\hat{\rho}_{pw} = 0.9$, it means that accuracy increased (relatively) by $\frac{1}{\hat{\rho}_{pw}} - 1 = 0.11$.

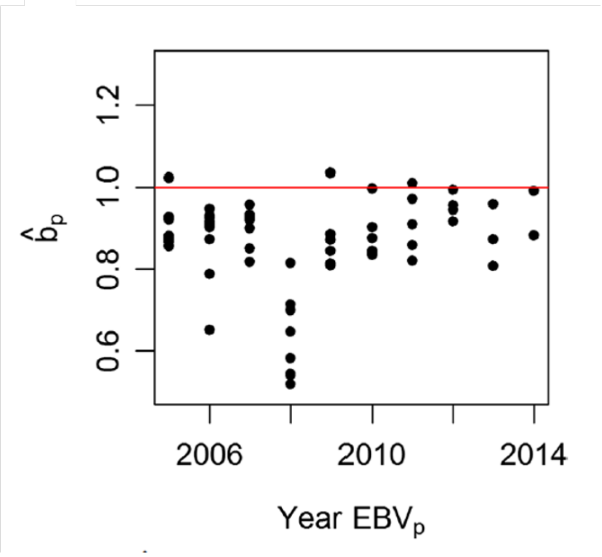

Figure 24: Different estimates of b_{pw}

17 References

Aguilar I., and I. Misztal, 2008 < i> Technical Note: Recursive Algorithm for Inbreeding Coefficients Assuming Nonzero Inbreeding of Unknown Parents. Journal of Dairy Science 91: 1669–1672.

Aguilar I., I. Misztal, D. 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. Journal of Dairy Science 93: 743–752.

Aguilar I., I. Misztal, A. Legarra, and S. Tsuruta, 2011 Efficient computations of genomic relationship matrix and other matrices used in the single-step evaluation. Journal of Animal Breeding and Genetics 128: 422–428.

Aguilar I., S. Tsuruta, Y. Masuda, D. Lourenco, A. Legarra, *et al.*, 2018 BLUPF90 suite of programs for animal breeding with focus on genomics, p. 751 in *Proceedings of the World Congress on Genetics Applied to Livestock Production*, Auckland, New Zealand.

Aguilar I., A. Legarra, F. Cardoso, Y. Masuda, D. Lourenco, *et al.*, 2019 Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. Genetics Selection Evolution 51: 28. https://doi.org/10.1186/s12711-019-0469-3

Aliloo H., J. E. Pryce, O. González-Recio, B. G. Cocks, and B. J. Hayes, 2016 Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genetics Selection Evolution 48: 8. https://doi.org/10.1186/s12711-016-0186-0

Álvarez-Castro J. M., and Ö. Carlborg, 2007 A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. Genetics 176: 1151–1167. https://doi.org/10.1534/genetics.106.067348

Andersson L., 2001 Genetic dissection of phenotypic diversity in farm animals. Nature Reviews Genetics 2: 130–138. https://doi.org/10.1038/35052563

Bermann M., A. Legarra, M. K. Hollifield, Y. Masuda, D. Lourenco, *et al.*, 2021 Validation of single-step GBLUP genomic predictions from threshold models using the linear regression method: An application in chicken mortality. Journal of Animal Breeding and Genetics 138: 4–13. https://doi.org/10.1111/jbg.12507

Bermann M., I. Aguilar, D. Lourenco, I. Misztal, and A. Legarra, 2023 Reliabilities of estimated breeding values in models with metafounders. Genetics Selection Evolution 55: 6. https://doi.org/10.1186/s12711-023-00778-2

Bijma P., 2012 Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. Journal of Animal Breeding and Genetics 129: 345–358.

Bonhomme M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, *et al.*, 2010 Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. Genetics 186: 241–262. https://doi.org/10.1534/genetics.110.117275

Bradford H., Y. Masuda, P. VanRaden, A. Legarra, and I. Misztal, 2019 Modeling missing pedigree in single-step genomic BLUP. Journal of dairy science 102: 2336–2346.

Caballero A., and M. A. Toro, 2002 Analysis of genetic diversity for the management of conserved subdivided populations. Conservation genetics 3: 289. https://doi.org/10.1023/A:1019956205473

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

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

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., P. Madsen, B. Nielsen, T. Ostersen, and G. Su, 2012 Single-step methods for genomic evaluation in pigs. Animal 6: 1565–1571.

Christensen O. F., A. Legarra, M. S. Lund, and G. Su, 2015 Genetic evaluation for three-way crossbreeding. Genetics Selection Evolution 47: 98.

Christensen O. F., and A. Legarra, 2022 Maximum likelihood estimation of metafounder parameters for single and multiple breeds, pp. 1376–1379 in *Proceedings of 12th World Congress on Genetics Applied to Livestock Production*, Wageningen Academic Publishers.

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

Cockerham C. C., 1954 An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39: 859.

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.

Colleau J. J., 2002 An indirect approach to the extensive calculation of relationship coefficients. Genetics Selection Evolution 34: 409–422.

Colleau J.-J., I. Palhière, S. T. Rodríguez-Ramilo, and A. Legarra, 2017 A fast indirect method to compute functions of genomic relationships concerning genotyped and ungenotyped individuals, for diversity management. Genetics Selection Evolution 49: 87. https://doi.org/10.1186/s12711-017-0363-9

Colombani C., A. Legarra, S. Fritz, F. Guillaume, P. Croiseau, *et al.*, 2013 Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCPi methods for genomic selection in French Holstein and Montbéliarde breeds. Journal of Dairy Science 96: 575–591.

Consortium T. B. G. S. and A., C. G. Elsik, R. L. Tellam, and K. C. Worley, 2009 The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science 324: 522–528. https://doi.org/10.1126/science.1169588

De Boer I., and I. Hoeschele, 1993 Genetic evaluation methods for populations with dominance and inbreeding. Theoretical and Applied Genetics 86: 245–258.

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.

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.

Dekkers J. C. M., 1992 Asymptotic response to selection on best linear unbiased predictors of breeding values. Animal Science 54: 351–360. https://doi.org/10.1017/S0003356100020808

Dunner S., M. E. Miranda, Y. Amigues, J. Cañón, M. Georges, *et al.*, 2003 Haplotype diversity of the myostatin gene among beef cattlebreeds. Genetics Selection Evolution 35: 103. https://doi.org/10.1186/1297-9686-35-1-103

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.

Elzo M., 2008 *Animal breeding notes*. University of Florida., http://www.animal.ufl.edu/elzo/.

Emigh T. H., 1980 A comparison of tests for Hardy-Weinberg equilibrium. Biometrics 36: 627–642.

Ertl J., A. Legarra, Z. G. Vitezica, L. Varona, C. Edel, *et al.*, 2014 Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet Sel Evol 46: 40.

Esfandyari H., P. Bijma, M. Henryon, O. F. Christensen, and A. C. Sørensen, 2016 Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. Genetics Selection Evolution 48: 40. https://doi.org/10.1186/s12711-016-0220-2

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. L., and M. Grossman, 1989 Marker assisted prediction using best linear unbiased prediction. Genetics Selection Evolution 21: 467–477.

Fernando R. L., and M. Grossman, 1990 Genetic evaluation with autosomal and X-chromosomal inheritance. Theoretical and Applied Genetics 80: 75–80. https://doi.org/10.1007/BF00224018

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.

Forneris N. S., A. Legarra, Z. G. Vitezica, S. Tsuruta, I. Aguilar, *et al.*, 2015 Quality control of genotypes using heritability estimates of gene content at the marker. Genetics 199: 675–681.

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.

Fragomeni B. O., D. A. L. Lourenco, Y. Masuda, A. Legarra, and I. Misztal, 2017 Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genetics Selection Evolution 49: 59. https://doi.org/10.1186/s12711-017-0335-0

Garcia-Baccino C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, *et al.*, 2017 Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. Genetics Selection Evolution 49: 34. https://doi.org/10.1186/s12711-017-0309-2

Garcia-Cortes L. A., and M. Toro, 2006 Multibreed analysis by splitting the breeding values. Genetics Selection Evolution 38: 601–615.

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., 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.

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.

George E. I., and R. E. McCulloch, 1993 Variable selection via Gibbs sampling. Journal of the American Statistical Association 88: 881–889.

Gianola D., and R. L. Fernando, 1986 Bayesian Methods in Animal Breeding Theory. Journal of Animal Science 63: 217.

Gianola D., R. L. Fernando, and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173: 1761–1776.

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.

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.

Group T. I. S. M. W., 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928–933. https://doi.org/10.1038/35057149

Gualdrón Duarte J. L., R. J. Cantet, R. O. Bates, C. W. Ernst, N. E. Raney, *et al.*, 2014 Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics 15: 246. https://doi.org/10.1186/1471-2105-15-246

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.

Hayes B., 2011 < i> Technical note: Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. Journal of Dairy Science 94: 2114–2117.

Henderson C. R., 1973 Sire evaluations and genetic trends. J Anim Sci Symposium.

Henderson C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics 423–447.

Henderson C. R., 1976 A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32: 69–83.

Henderson C., 1978 Undesirable properties of regressed least squares prediction of breeding values. Journal of Dairy Science 61: 114–120.

Henderson C. R., 1982 Best linear unbiased prediction in populations that have undergone selection, pp. 191–201 in *Proceedings of the world congress on sheep and beef cattle breeding*,.

Henderson C. R., 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph.

Henderson C. R., 1985 Best linear unbiased prediction of nonadditive genetic merits. J. Anim. Sci 60: 111–117.

Hickey J. M., B. P. Kinghorn, B. Tier, J. F. Wilson, N. Dunstan, *et al.*, 2011 A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol 43. https://doi.org/10.1186/1297-9686-43-12

Hill W., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38: 226–231.

Hill W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLOS Genetics 4: e1000008. https://doi.org/10.1371/journal.pgen.1000008

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.

Hill W. g., and A. Mäki-Tanila, 2015 Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. Journal of Animal Breeding and Genetics 132: 176–186. https://doi.org/10.1111/jbg.12140

Hsu W.-L., D. J. Garrick, and R. L. Fernando, 2017 The accuracy and bias of single-step genomic prediction for populations under selection. G3: Genes, Genomes, Genetics 7: 2685–2694.

Huang W., and T. F. C. Mackay, 2016 The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. PLOS Genetics 12: e1006421. https://doi.org/10.1371/journal.pgen.1006421

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.

Kachman S. D., M. L. Spangler, G. L. Bennett, K. J. Hanford, L. A. Kuehn, *et al.*, 2013 Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genetics Selection Evolution 45: 30. https://doi.org/10.1186/1297-9686-45-30

Kass R. E., and A. E. Raftery, 1995 Bayes factors. Journal of the American Statistical Association 90: 773–795.

Kennedy B. W., 1981 Bias and Mean Square Error from Ignoring Genetic Groups in Mixed Model Sire Evaluation. Journal of Dairy Science 64: 689–697. https://doi.org/10.3168/jds.S0022-0302(81)82631-8

Kennedy B., 1991 CR Henderson: The unfinished legacy. Journal of Dairy Science 74: 4067–4081.

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.

Kudinov A. A., E. A. Mäntysaari, G. P. Aamand, P. Uimari, and I. Strandén, 2020 Metafounder approach for single-step genomic evaluations of Red Dairy cattle. Journal of Dairy Science 103: 6299–6310. https://doi.org/10.3168/jds.2019-17483

Kudinov A. A., M. Koivula, G. P. Aamand, I. Strandén, and E. A. Mäntysaari, 2022 Single-step genomic BLUP with many metafounders. Frontiers in Genetics 13: 1012205.

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. Journal of Dairy Science 91: 360–366.

Legarra A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen, 2008 Performance of genomic selection in mice. Genetics 180: 611–618.

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

Legarra A., A. Ricardi, and O. Filangi, 2011a GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and BayesCp)

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., G. Baloche, F. Barillet, J. M. Astruc, C. Soulas, *et al.*, 2014a Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. Journal of dairy science 97: 3200–3212.

Legarra A., O. F. Christensen, I. Aguilar, and I. Misztal, 2014b Single Step, a general approach for genomic selection. Livestock Science 166: 54–65.

Legarra A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal, 2015a Ancestral relationships using metafounders: Finite ancestral populations and across population relationships. Genetics 200: 455–468. https://doi.org/10.1534/genetics.115.177014

Legarra A., P. Croiseau, M. P. Sanchez, S. Teyssèdre, G. Sallé, *et al.*, 2015b A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. Genetics Selection Evolution 47: 6. https://doi.org/10.1186/s12711-015-0087-7

Legarra A., and Z. G. Vitezica, 2015 Genetic evaluation with major genes and polygenic inheritance when some animals are not genotyped using gene content multiple-trait BLUP. Genetics Selection Evolution 47: 89. https://doi.org/10.1186/s12711-015-0165-x

Legarra A., 2016 Comparing estimates of genetic variance across different relationship models. Theoretical population biology 107: 26–30.

Legarra A., and A. Reverter, 2018 Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method [Erratum: Dec. 2019, v. 51 (1), p. 69]

Legarra A., and P. M. VanRaden, 2023 Effect of modelling unknown parent groups and metafounders on the historical genetic trend of fertility traits. Interbull Bulletin 11–14.

Legarra A., M. Bermann, Q. Mei, and O. F. Christensen, 2024a Estimating genomic relationships of metafounders across and within breeds using maximum likelihood, pseudo-expectation–maximization maximum likelihood and increase of relationships. Genetics Selection Evolution 56: 35. https://doi.org/10.1186/s12711-024-00892-9

Legarra A., M. Bermann, Q. Mei, and O. F. Christensen, 2024b Redefining and interpreting genomic relationships of metafounders. Genetics Selection Evolution 56: 34. https://doi.org/10.1186/s12711-024-00891-w

Lehermeier C., G. de los Campos, V. Wimmer, and C.-C. Schön, 2017 Genomic variance estimates: With or without disequilibrium covariances? Journal of Animal Breeding and Genetics 134: 232–241. https://doi.org/10.1111/jbg.12268

Li C. C., and D. G. Horvitz, 1953 Some methods of estimating the inbreeding coefficient. Am J Hum Genet 5: 107–117.

Lourenco D., I. Misztal, H. Wang, I. Aguilar, S. Tsuruta, *et al.*, 2013 Prediction accuracy for a simulated maternally affected trait of beef cattle using different genomic evaluation models. Journal of Animal Science 91: 4090–4098.

Lourenco D., I. Misztal, S. Tsuruta, I. Aguilar, T. Lawlor, *et al.*, 2014 Are evaluations on young genotyped animals benefiting from the past generations? Journal of Dairy Science 97: 3930–3942.

Lourenco D. A., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, *et al.*, 2015a Accuracy of estimated breeding values with genomic information on males, females, or both: An example on broiler chicken. Genetics Selection Evolution 47: 56.

Lourenco D. a. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, *et al.*, 2015b Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. Journal of Animal Science 93: 2653–2662. https://doi.org/10.2527/jas.2014-8836

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.

Lynch M., and B. Walsh, 1998 *Genetics and analysis of quantitative traits*. Sinauer associates.

Macedo F., A. Reverter, and A. Legarra, 2020a Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. Journal of Dairy Science 103: 529–544.

Macedo F. L., O. F. Christensen, J.-M. Astruc, I. Aguilar, Y. Masuda, *et al.*, 2020b Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. Genetics, Selection, Evolution : GSE 52. https://doi.org/10.1186/s12711-020-00567-1

Macedo F. L., O. F. Christensen, and A. Legarra, 2021 Selection and drift reduce genetic variation for milk yield in Manech Tête Rousse dairy sheep. JDS Communications 2: 31–34. https://doi.org/10.3168/jdsc.2020-0010

Marchal A., A. Legarra, S. Tisné, C. Carasco-Lacombe, A. Manez, *et al.*, 2016 Multivariate genomic model improves analysis of oil palm (Elaeis guineensis Jacq.) progeny tests. Molecular Breeding 36: 2.

Martini J. W. R., V. Wimmer, M. Erbe, and H. Simianer, 2016 Epistasis and covariance: How gene interaction translates into genomic relationship. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 129: 963–976. https://doi.org/10.1007/s00122-016-2675-5

Masuda Y., P. M. VanRaden, S. Tsuruta, D. A. L. Lourenco, and I. Misztal, 2022 Invited review: Unknown-parent groups and metafounders in single-step genomic BLUP. Journal of Dairy Science 105: 923–939. https://doi.org/10.3168/jds.2021-20293

Matukumalli L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, *et al.*, 2009 Development and characterization of a high density SNP genotyping assay for cattle. PloS One 4: e5350. https://doi.org/10.1371/journal.pone.0005350

Meuwissen T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Meuwissen T., 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., B. Hayes, and M. Goddard, 2016 Genomic selection: A paradigm shift in animal breeding. Animal Frontiers 6: 6–14. https://doi.org/10.2527/af.2016-0002

Meyer K., B. Tier, and A. Swan, 2018 Estimates of genetic trend for single-step genomic evaluations. Genetics Selection Evolution 50: 39. https://doi.org/10.1186/s12711-018-0410-1

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

Misztal I., Z.-G. Vitezica, A. Legarra, I. Aguilar, and A. Swan, 2013 Unknown-parent groups in single-step genomic evaluation. Journal of Animal Breeding and Genetics 130: 252–258.

Misztal I., Shortage of quantitative geneticists in animal breeding. Journal of Animal Breeding and Genetics 124: 255–256. https://doi.org/10.1111/j.1439-0388.2007.00679.x

Moghaddar N., and J. H. J. van der Werf, 2017 Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. Journal of Animal Breeding and Genetics 134: 453–462. https://doi.org/10.1111/jbg.12287

Mrode R., and R. Thompson, 2005 *Linear models for the prediction of animal breeding values*. Cabi.

Mrode R., and I. Pocrnic, 2023 *Linear Models for the Prediction of the Genetic Merit of Animals, 4th Edition*. CABI.

Muñoz P. R., M. F. R. Resende, S. A. Gezan, M. D. V. Resende, G. de los Campos, *et al.*, 2014 Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. Genetics 198: 1759–1768. https://doi.org/10.1534/genetics.114.171322

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.

Page B. T., E. Casas, M. P. Heaton, N. G. Cullen, D. L. Hyndman, *et al.*, 2002 Evaluation of single-nucleotide polymorphisms in CAPN1 for association with meat tenderness in cattle,. Journal of Animal Science 80: 3077–3085. https://doi.org/10.2527/2002.80123077x

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.

Phocas F., and D. Laloë, 2004 Should genetic groups be fitted in BLUP evaluation? Practical answer for the French AI beef sire evaluation. Genetics Selection Evolution 36: 325. https://doi.org/10.1186/1297-9686-36-3-325

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.

Quaas R. L., 1988 Additive genetic model with groups and relationships. Journal of Dairy Science 71: 1338–1345.

Reverter A., B. L. Golden, R. M. Bourdon, and J. S. Brinks, 1994/Jan/01 Technical note: Detection of bias in genetic predictions. Journal of animal science 72: 34–37. https://doi.org/10.2527/1994.72134x

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.

Rogers A. R., and C. Huff, 2009 Linkage Disequilibrium Between Loci With Unknown Phase. Genetics 182: 839–844. https://doi.org/10.1534/genetics.108.093153

Saatchi M., M. C. McClure, S. D. McKay, M. M. Rolf, J. Kim, *et al.*, 2011 Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genetics Selection Evolution 43: 40. https://doi.org/10.1186/1297-9686-43-40

Schork N. J., D. Fallin, and J. S. Lanchbury, 2000 Single nucleotide polymorphisms and the future of genetic epidemiology. Clinical Genetics 58: 250–264. https://doi.org/10.1034/j.1399-0004.2000.580402.x

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.

Sillanpaa M., 2011 On statistical methods for estimating heritability in wild populations. Molecular Ecology 20: 1324–1332.

Smith J. A., A. M. Lewis, P. Wiener, and J. L. Williams, 2000 Genetic variation in the bovine myostatin gene in UK beef cattle: Allele frequencies and haplotype analysis in the South Devon. Animal Genetics 31: 306–309. https://doi.org/10.1046/j.1365-2052.2000.00521.x

Snelling W. M., M. F. Allan, J. W. Keele, L. A. Kuehn, R. M. Thallman, *et al.*, 2011 Partial-genome evaluation of postweaning feed intake and efficiency of crossbred beef cattle,. Journal of Animal Science 89: 1731–1741. https://doi.org/10.2527/jas.2010-3526

Soller M., and J. S. Beckmann, 1983 Genetic polymorphism in varietal identification and genetic improvement. Theoretical and Applied Genetics 67: 25–33. https://doi.org/10.1007/BF00303917

Sorensen D., R. Fernando, and D. Gianola, 2001 Inferring the trajectory of genetic variance in the course of artificial selection. Genetical research 77: 83–94.

Sorensen D., and D. Gianola, 2002 *Likelihood, bayesian and MCMC methods in quantitative genetics*. Springer.

Stoneking M., 2001 Single nucleotide polymorphisms: From the evolutionary past. . . Nature 409: 821–822. https://doi.org/10.1038/35057279

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.

Strandén I., and O. F. Christensen, 2011 Allele coding in genomic evaluation. Genet Sel Evol 43: 25.

Strandén I., K. Matilainen, G. p. Aamand, and E. a. Mäntysaari, 2017 Solving efficiently large single-step genomic best linear unbiased prediction models. Journal of Animal Breeding and Genetics 134: 264–274. https://doi.org/10.1111/jbg.12257

Strandén I., G. P. Aamand, and E. A. Mäntysaari, 2022 Single-step genomic BLUP with genetic groups and automatic adjustment for allele coding. Genetics Selection Evolution 54: 38. https://doi.org/10.1186/s12711-022-00721-x

Stuber C. W., and C. C. Cockerham, 1966 Gene Effects and Variances in Hybrid Populations. Genetics 54: 1279–1286.

Su G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund, 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., P. Madsen, U. S. Nielsen, E. A. Mäntysaari, G. P. Aamand, *et al.*, 2012b Genomic prediction for Nordic Red Cattle using one-step and selection index blending. Journal of Dairy science 95: 909–917.

Su G., B. Guldbrandtsen, G. P. Aamand, I. Strandén, and M. S. Lund, 2014 Genomic relationships based on X chromosome markers and accuracy of genomic predictions with and without X chromosome markers. Genetics, Selection, Evolution : GSE 46: 47. https://doi.org/10.1186/1297-9686-46-47

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.

Sun C., P. M. VanRaden, J. R. O'Connell, K. A. Weigel, and D. Gianola, 2013 Mating programs including genomic relationships and dominance effects. Journal of Dairy Science 96: 8014–8023. https://doi.org/10.3168/jds.2013-6969

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.

Thompson R., 1979 Sire evaluation. Biometrics 35: 339–353.

Thompson E. A., 2013 Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. Genetics 194: 301–326. https://doi.org/10.1534/genetics.112.148825

Tibshirani R., 1996 Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58: 267–288.

Tier B., K. Meyer, and A. Swan, 0011/2018-02-16 On implied genetic effects, relationships and alternate allele coding, in *Proceedings of the 11th World Congress on Genetics Applied to Livestock Production*, Auckland.

Toro M. A., and L. Varona, 2010 A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol 42: 33.

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.

Vandenplas J., J. ten Napel, S. N. Darbaghshahi, R. Evans, M. P. L. Calus, *et al.*, 2023 Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. Genetics Selection Evolution 55: 37. https://doi.org/10.1186/s12711-023-00808-z

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., 1992 Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. Journal of Dairy Science 75: 3136–3144.

VanRaden P. M., 2007 Genomic measures of relationship and inbreeding. Interbull bull 37: 33–36.

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

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

VanRaden P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel, 2011 Genomic evaluations with many more genotypes. Genetics, Selection, Evolution : GSE 43: 10. https://doi.org/10.1186/1297-9686-43-10

Varona L., 2014-08-17/2014-08-22 A general approach for calculation of genomic relationship matrices for epistatic effects., pp. 11–22 in *10th World Congress on Genetics Applied to Livestock Production.*, Vancouver, Canada.

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.

Varona L., 2010 Understanding the use of Bayes factor for testing candidate genes. Journal of Animal Breeding and Genetics 127: 16–25.

Varona L., A. Legarra, M. A. Toro, and Z. G. Vitezica, 2018 Non-additive Effects in Genomic Selection. Frontiers in Genetics 9. https://doi.org/10.3389/fgene.2018.00078

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, and A. Legarra, 2010 One-step vs. Multi-step methods for genomic prediction in presence of selection. Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany.

Vitezica Z., I. Aguilar, I. Misztal, and A. Legarra, 2011 Bias in genomic predictions for populations under selection. Genetics Research 93: 357–366.

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

Vitezica Z. G., A. Legarra, M. A. Toro, and L. Varona, 2017 Orthogonal Estimates of Variances for Additive, Dominance and Epistatic Effects in Populations. Genetics genetics.116.199406. https://doi.org/10.1534/genetics.116.199406

Wakefield J., 2009 Bayes factors for genome-wide association studies: Comparison with P-values. Genetic Epidemiology 33: 79–86.

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

Wiggans G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, *et al.*, 2009 Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J Dairy Sci 92: 3431–3436.

Wiggans G. R., P. M. VanRaden, L. R. Bacheller, M. E. Tooker, J. L. Hutchison, *et al.*, 2010 Selection and management of DNA markers for use in genomic evaluation. Journal of Dairy Science 93: 2287–2292. https://doi.org/10.3168/jds.2009-2773

Wright S., 1922 Coefficients of inbreeding and relationship. Am. Nat. 56: 330–338.

Xiang T., O. F. Christensen, Z. G. Vitezica, and A. Legarra, 2016 Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genetics Selection Evolution 48: 92. https://doi.org/10.1186/s12711-016-0271-4

Xiang T., O. F. Christensen, and A. Legarra, 2017 Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. Journal of Animal Science 95: 1472–1480.

Xu S., 2013 Mapping Quantitative Trait Loci by Controlling Polygenic Background Effects. Genetics 195: 1209–1222. https://doi.org/10.1534/genetics.113.157032

Xu S., D. Zhu, and Q. Zhang, 2014 Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proceedings of the National Academy of Sciences 111: 12456–12461. https://doi.org/10.1073/pnas.1413750111

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.

18 Appendix

Legarra A., Christensen O. F., Aguilar I., Misztal I., 2014 Single Step, a general approach for genomic selection. Livestock Science 166: 54–65.

Legarra A., Reverter A., 2017 Can we frame and understand cross-validation results in animal breeding? In: AAABG conference, Townsville, Australia. http://agbu.une.edu.au/AAABG_2017.html

Varona L., Legarra A., Toro M. A., Vitezica Z. G., 2018 Non-additive Effects in Genomic Selection. Front. Genet. 9.

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/© 2014 Elsevier B.V. All rights reserved.

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 $Var(\boldsymbol{a})=\boldsymbol{D}_a=\boldsymbol{I}\sigma_u^2/2\Sigma p_k q_k$, where σ_u^2 is the genetic variance, so that that $Var(\boldsymbol{u})=\sigma_u^2\boldsymbol{G}$, 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.

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,

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

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

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

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

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 **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:

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

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

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):

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

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:

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

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) $E(\mathbf{G})=\mathbf{A}_{22}$ if $\mathbf{G}$ is the realized relationship and $\mathbf{A}_{22}$ 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 $\mathbf{E}$ 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 $\mathbf{E}$ 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: $p(\mathbf{u}_2)=N(1\mu,\mathbf{G})$. 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 $p(\mathbf{u}_2)=N(0, \mathbf{G}+11'Var(\mu))$ with μ integrated out. An unbiased method forces the distribution of average values of breeding values $(\overline{\mathbf{u}}_2)$ to be identical and therefore, the adjustment uses $\mathbf{G}^{*}=a+b\mathbf{G}$ with $b=1$ and $a=\overline{\mathbf{A}}_{22}-\overline{\mathbf{G}}$ where the bar implies average across values of $\mathbf{G}$ and $\mathbf{A}$. 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 $\mathbf{A}_{22}$ but not in $\mathbf{G}$. 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 $a=\overline{\mathbf{A}}_{22}-\overline{\mathbf{G}}$ 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 $\mathbf{G}$ 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 $\overline{\mathbf{A}}_{22}$. 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 $\mathbf{G}$ and $\mathbf{A}_{22}$. More formally, the empirical variance of breeding values:$S^2_{u_2}=\mathbf{u}_2'\mathbf{u}_2/n-(\overline{\mathbf{u}}_2)^2$ has an expectation $((tr(\mathbf{A}_{22}))/(n)-\overline{\mathbf{A}}_{22})\sigma^2_u$ or $((tr(\mathbf{G}^{*}))/(n)-\overline{\mathbf{G}}^{*})\sigma^2_u$ where n is the number of individuals. Forcing unbiasedness implies that a and b should be determined from the system of two equations: $a+b(tr(\mathbf{G}))/(n)=(tr(\mathbf{A}_{22}))/(n)$ and $a+b\overline{\mathbf{G}}=\overline{\mathbf{A}}_{22}$. 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 $\mathbf{G}$, 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" $\mathbf{A}_{22}$ to $\mathbf{G}$ 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, $\mathbf{A}$ (γ) whose founders have a relationship matrix $\mathbf{A}_{bas}=\gamma+\mathbf{I}$ $(1-\gamma/2)$. Parameter s, used in $\mathbf{G}=\mathbf{ZZ}'/s$ 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 $\mathbf{G}$ are cubic processes, and are much optimized by the use of efficient algorithms and parallel computations (Aguilar et al., 2011). Construction of matrix $\mathbf{A}_{22}$ has been possible, for very large pedigrees, by the algorithm of Colleau (2002) which uses Henderson's decomposition of $\mathbf{A}=\mathbf{TDT}'$ to devise a "solving" that allows easy multiplication of $\mathbf{w}=\mathbf{Av}$ and computation of $\mathbf{A}_{22}$ 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 $(\mathbf{LHS})\mathbf{sol}$ where $\mathbf{sol}$ is the vector of unknowns. In practice, this product is split into a part

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

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

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

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

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

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:

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

Their final Single Step mixed model equations are

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

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).

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 TopPigs 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 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.

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

CAN WE FRAME AND UNDERSTAND CROSS-VALIDATION RESULTS IN ANIMAL BREEDING?

A. Legarra[1], A. Reverter[2]

[1] UMR 1388 GenPhySE, INRA, Castanet Tolosan, France
[2] CSIRO Agriculture and Food, 306 Carmody Rd., St. Lucia, QLD 4067, Australia

SUMMARY

Performance of genomic selection is typically evaluated by cross-validation. In this work we review and point out some problems and features of the cross-validation metrics. Then we propose a semiparametric alternative using statistics derived from the “Method R”.

INTRODUCTION

Genomic prediction of breeding values via genomic BLUP (GBLUP) is expensive and requires initial and continuous investments in genotyping. State of the art theory so far does not yield convincing *a priori* estimates of the increased accuracy of genomic prediction vs. pedigree-based predictions. Thus, cross-validation has been extensively used (e.g. Legarra *et al.* 2008; VanRaden *et al.* 2009; Mantysaari *et al.* 2010; Christensen *et al.* 2012). The theory of cross-validation is poorly understood in the context of heavily related and selected data (but see (Gianola and Schön, 2016)). For instance, how to evaluate accuracy for maternal traits is very unclear. Here we provide a brief review of this topic and suggest some options.

CROSS-VALIDATION BIAS AND ACCURACY

What cross-validation? Forecasters such as pedigree-BLUP and GBLUP may behave differently according to what the “forecasted” target is. Breeders have a difficult task, namely, to forecast the best reproducers in order to select them. In this, they are different from *machine learners*, whose objective is (from our perspective) to forecast present phenomena. Thus, it is rather obvious that for breeders the best method is such that allows taking the best selection decisions, that it is, the method that best predicts future performance of an individual knowing its genetic background.

We will call this *forward cross-validation*. Its features are three-fold: (1) It needs the definition of a cut-off date; (2) It needs the construction of “Full” and “Reduced” data sets (Mantysaari *et al.* 2010; Olson *et al.* 2011); and (3) In its crudest form, it does not provide any form of randomisation and therefore a point estimate of goodness of prediction is obtained, without any associated measure of uncertainty.

In contrast, the classical *random folding k*-fold cross-validation in its most classic form splits randomly the data into *k* distinct sets and predicts one set from the remaining *k-1* sets. Its key features include: (1) Extremely simple to implement; (2) Provides estimates of standard error of metrics of cross-validation; (3) Not realistic in an animal breeding setting and the ranking of methods is not suitable for practical purposes; and (4) Tends to overfit (case of leave-one-out)

Some more esoteric forms of cross-validation exist. Legarra *et al.* (2008) *split folds “across” or “within” families*, obtaining very different results. But this is undoable (and little useful) for regular animal breeding data. The *k-means for cross-validation* (Saatchi *et al.* 2011) separates individuals into “most distinct” folds, and the i-th fold is predicted from the remaining k-1 folds. This does not answer the breeder’s question, which most often wants to predict from *close*, not from *far* animals.

Which metrics? To assess the *predictive ability* of the different forecasters, animal breeders are highly formatted by Henderson’s BLUP, which in turn was highly dependent upon dairy cattle genetic improvement. Metrics commonly used come from linear regression, named in this paper *predictive abilities*, are:

$$\text{Bias: } b_0 = E(u - \hat{u}); \qquad \text{Slope: } b_1 = \frac{Cov(u,\hat{u})}{Var(\hat{u})}; \quad \text{Accuracy: } r = \frac{Cov(u,\hat{u})}{\sqrt{Var(u)Var(\hat{u})}}$$

Sometimes mean squared error is used ($MSE = b_0^2 + \sigma_u^2(1 + r^2/b_1^2 - 2r^2/b_1)$). Properties of BLUP in absence of selection are no bias, slope of 1, and maximum accuracy. Henderson defined this at the individual level on a frequentist basis (over conceptual repetitions). Bias=0 and slope=1 ensure fair comparisons across old and young animals. This is important if the scheme mixes proven and young animals, like dairy cattle. It seems less relevant in schemes were reproducers are culled quickly (pigs, chicken) with beef species falling someone in the middle, we believe. Deviations may exist if there is selection, because bias and slope are related to genetic gain and dispersion (see Figure 1).

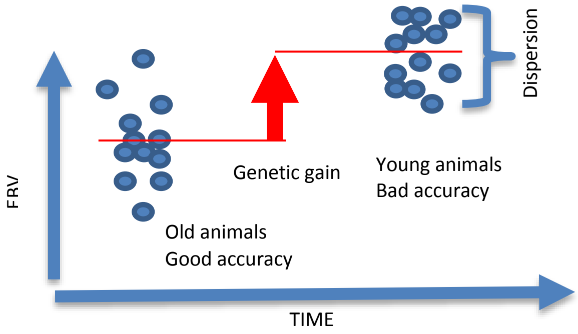

Figure 1 Typical scenario for retrospective analysis.

What is it meant by classical bias? Animal breeders probably agree to Henderson's (1973) sentence "*most users would, I think, be reluctant deliberately to bias comparisons between different groups, for example to underevaluate young sires as compared to older ones*". Here we have an operational definition of bias. In formal terms this implies that at a given point in time:

$$\begin{aligned} b_0^{[Henderson]} &= \left(\mathbf{1}'\hat{\boldsymbol{u}}_{group1} - \mathbf{1}'\hat{\boldsymbol{u}}_{group2}\right) - \left(\mathbf{1}'\boldsymbol{u}_{group1} - \mathbf{1}'\boldsymbol{u}_{group2}\right) \\ &= \left(\mathbf{1}'\hat{\boldsymbol{u}}_{group1} - \mathbf{1}'\boldsymbol{u}_{group1}\right) - \left(\mathbf{1}'\hat{\boldsymbol{u}}_{group2} - \mathbf{1}'\boldsymbol{u}_{group2}\right) \end{aligned}$$

This definition has practical implications: if the candidates are chosen *across* groups, selection decisions are optimal if there is no bias. Thus, it is expected that $b_0^{[Henderson]} = 0$. There may be several definitions of groups: (1) Different conditions (grazing vs. indoor fed cattle). This case should be addressed by the model used for evaluation; (2) Within country, different amounts of information that cumulate in time (progeny-tested vs. genomic bulls). This case is strongly affected by within-country genetic trend (see below); (3) Same amount of information, but different origins (US vs. FR). This case is most affected by wrong estimates of the difference in genetic level across countries (Bonaiti *et al.* 1993; Powell and Wiggans 1994).

The Interbull definition. Interbull uses retrospective tests (Boichard *et al.* 1995; Mantysaari *et al.* 2010) that compare EBV's *before* and *after* progeny testing.

$$b_0^{[Interbull]} = \mathbf{1}'\hat{\boldsymbol{u}}_t - \mathbf{1}'\hat{\boldsymbol{u}}_{t-1}$$

If progeny testing gives exact EBVs, then $\hat{\boldsymbol{u}}_t = \boldsymbol{u}_t$ and $b_0^{[Interbull]} = \mathbf{1}'\boldsymbol{u} - \mathbf{1}'\hat{\boldsymbol{u}}_{t-1}$.Note that $b_0^{[Henderson]} \neq b_0^{[Interbull]}$, but if group1 is "very old" proven bulls and $\hat{\boldsymbol{u}}_t = \boldsymbol{u}_t$ and group2 is genomic bulls (then becoming proven bulls) then $b_0^{[Henderson]} = b_0^{[Interbull]}$. This may be rather obvious, but it only holds for progeny testing data.

What happens under selection? Assume that we want to compare selection candidates with "proven" animals. If there is *no* selection, then $\mathbf{1}'\boldsymbol{u}_{group1} = \mathbf{1}'\boldsymbol{u}_{group2}$ and there is actually no need to make the test. Alas, if *there is* selection, then

$$b_0^{[Henderson]} = \left(\mathbf{1}'\hat{\boldsymbol{u}}_{group1} - \mathbf{1}'\hat{\boldsymbol{u}}_{group2}\right) - \left(\mathbf{1}'\boldsymbol{u}_{group1} - \mathbf{1}'\boldsymbol{u}_{group2}\right) = n\left(\hat{\Delta} - \Delta\right)$$

in other words, unbiasedness requires a correct (unbiased!) estimate of the realized genetic trend.

What is overdispersion, a.k.a {Interbull, genomic} bias? Is it affected by selection?

Dairy cattle breeders are much concerned by overdispersion of genomic proofs. If there is too much dispersion of $\hat{\boldsymbol{u}}_{genomic}$, the retained candidates will have unfairly high $\hat{\boldsymbol{u}}_{genomic}$. This could be staten more formally as "the mean of the EBVs of the selected candidates should be equal to the mean of the TBVs". If selection is by truncation and under multivariate normality, the true mean *after* selection is $\mu_T = (\mathbf{1}'\boldsymbol{u})/n + ir\sigma_u$, but this mean is (implicitly) predicted before selection as $\mu_E = (\mathbf{1}'\hat{\boldsymbol{u}})/n + i\sigma_{\hat{u}}$.

For $\mu_T = \mu_E$ to hold, we need the first unbiasedness condition (b_0 above), plus a second condition, $\sigma_{\hat{u}} = r\sigma_u$. But this condition *only* holds if $Cov(u,\hat{u}) = Var(\hat{u})$, which amounts to the regression coefficient to be 1:

$$b_1 = \frac{Cov(u,\hat{u})}{Var(\hat{u})}$$

This is the Interbull official, and most put forward, test of unbiasedness and nowadays more often called as "bias". It is easy to see why $b_1 = 1$ may not hold, namely, because selection modifies variances in rather unpredictable manners. The expected $Cov(u,\hat{u}) = Var(\hat{u})$ holds under quite restrictive conditions (Henderson 1982).

Evaluations can easily be biased. Unbiasedness of current genetic evaluations is more wishful thinking than an established fact. Unbiasedness exist only if several conditions hold:

- The model is correct (linear model, effects, heritabilities…)
- The selection process is described by the data
- Multivariate normality

Thus, there are many reasons why there is wrong estimate of the genetic trend and thus there will be bias:

- Collinearity of contemporary groups and genetic trend (this is the usual case)
- Genetic groups in the model
- Heritability is wrong (or changes with time)
- Analysis are single trait whereas selection is multiple trait
- Selection decisions not based on data.

In addition, genetic gain can be estimated one generation forward (but no more) unless an explicit selection model is included. In other words, retrospective analysis cannot be done deleting two generations of records. This would need explicit introduction of the selection process.

Why some species/traits seem biased where others do not? Basically, if there is *no* selection then *automatically* $b_0 = 0$ holds (i.e., all possible sets of candidates have 0 average value), and most likely $b_1 = 1$ holds, because selection does not change variances, and if a decent estimator of genetic variance is used, then genetic parameters are such that $b_1 = \frac{Cov(u,\hat{u})}{Var(\hat{u})} = 1$ by construction, in particular in a BLUP context. So, bias is expected to increase more with higher genetic gains.

An example is *pigs*. Christensen *et al.* (Christensen *et al.* 2012) found slopes below 1 (~0.9) for a heritable, selected trait (daily gain), whereas Xiang *et al.* (Xiang *et al.* 2016) found regressions nearly one for hard-to-select trait litter size.

In Lacaune dairy *sheep* (Baloche *et al.* 2014), we can put together the following. Figure 1 shows the regression slopes vs. the expected genetic gain or the expected loss of genetic variance based on Robertson (1977) . In theory, the reduction in variance is accounted for by genetic evaluation (Bijma 2012). In practice, this does not seem to be the case. A possible solution may be to reestimate this variance in each cycle of selection.

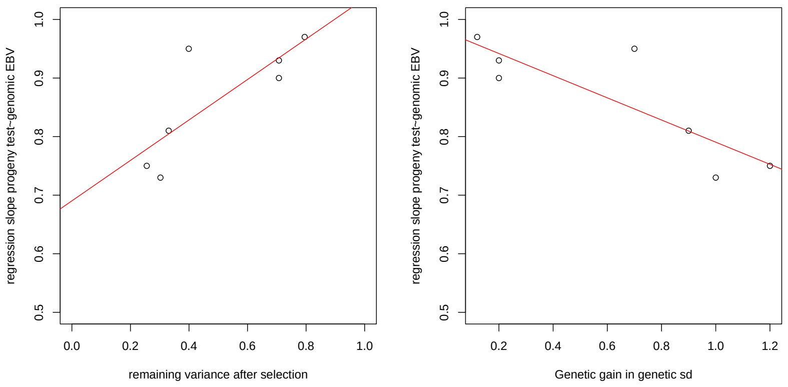

Figure 2 Slope b_1 vs. expected reduction in genetic variance (left) or genetic gain (right) by trait in Lacaune dairy sheep.

Vitezica *et al.* (2011) compared by *simulation* several predictors in selected populations in a SSGBLUP context. Statistic b_1 generally indicated bias, that was higher with less heritability. High heritability increases the selection differential and reduces variances, but it also gives more information. Interestingly, the only method which provided unbiased $b_1 = 0.99$ resulted in strong bias $b_0 = 1.38\sigma_u$. Thus, *both* bias should be checked.

What do we mean by accuracy? In animal breeding textbooks, accuracy (r, with reliability r^2) is presented twice: first, as a component of $\Delta_G = ir\sigma_u$ (so, a populational parameter) and, second, as a measure of uncertainty of $\hat{u}$ (an individual parameter). However, when selecting from real populations, EBVs are correlated across individuals, so the individual accuracies may be meaningless. In other words: it is pointless to obtain $r_i = 0.70$ and $r_j = 0.70$ if $r\left(\hat{u}_i, \hat{u}_j\right) = 0.69$.

Cross-validation accuracies are computed as correlations $r^2 = \frac{Cov(u,\hat{u})}{Var(u)Var(\hat{u})}$. They indicate our ability to rank individuals *within* a cohort. The fact that these accuracies are computed regardless of the correlated structure of both u and $\hat{u}$ has unclear implications. In fact, it can be shown that, if Hendersonian conditions hold, $E(r)^2 = 1 - \frac{\left(\overline{diag(\boldsymbol{C}^{22})} - \overline{\boldsymbol{C}^{22}}\right)}{\left(\overline{diag(\boldsymbol{G})} - \overline{\boldsymbol{G}}\right)}$ is the expectation of the observed reliability. This reliability takes into account the "classical" reliability contained in the diagonal terms but also the relationships a priori (in $\boldsymbol{G}$) and a posteriori (in $\boldsymbol{C}^{22}$) across individuals. If the evaluation method cannot rank correctly *within* the validation sample, then diagonal and off-diagonal values of $\boldsymbol{C}^{22}$ are similar and reliability drops down. This is a desirable behaviour.

Selection also affects observed cross-validation accuracy (Edel et al., 2012; Bijma 2012). If the cross-validation test uses elite animals, accuracies are underestimated. In other words, it is easy to rank all animals, but more difficult to rank elite animals. The reduction is such that

$$r^2_{selected} = 1 - (1 - r^2_{unselected})\frac{\sigma^2_{u_{unselected}}}{\sigma^2_{u_{selected}}} .$$

ISSUES OF CROSS-VALIDATION METRICS

The accuracy of cross-validation metrics. After an experiment has been carried out, the breeder wants to know if the genomic accuracy is really different from the parents average accuracy. A simple method is to use the theoretical standard error of the estimates; for b_0 and b_1 these are from classical regression theory. For the correlation, this is a bit more convoluted, but an option is to use Fisher's z-transform: $z = \frac{1}{2} ln \frac{1+r}{1-r}$ has approximate s.e. $1/\sqrt{n-3}$ where n is the number of data points used. From this a confidence interval can be worked out. For instance, in the Basco-Bearnaise breed genomic predictions of 87 rams were 0.06 more accurate than parent averages (Legarra *et al.* 2014); this implies a rather symmetric 95% confidence interval of $[-0.15,0.27]$.

There is a source of bias and two sources of randomness in cross-validation metrics. The source of bias is that individuals are related both at the stage of prediction (parent average and genomic) and later, at the stage of validation (moment at which they have data; except for the case of progeny-tested animals for which proofs can be assumed uncorrelated). This has been discussed above. The two sources of randomness are: (1) Sampling of the reference population, (2) Sampling of the validation population. Fisher's z-transform and Hotelling-Williams test include both. However, they do not consider that individuals are related, and therefore the accuracy is likely to be overestimated. Again, a theoretical equation can be worked out to estimate $Var(r)$.

(Re)Sampling of the validation population. A more practical approach involves using (re)sampling techniques. In k-fold cross-validation this is immediate but, as discussed before, the setting is not realistic. In (Mäntysaari and Koivula 2012; Legarra *et al.* 2014; Cuyabano *et al.* 2015), sampling of the validation population was addressed by *bootstrapping*, i.e. sampling n individuals with replacement from the original n individuals in the validation data set. This method main virtue is that it avoids strong influence of outliers in the validation data set. It also allows formal comparisons of accuracies. Its main drawback is that it does not addresses the sampling of the reference population.

(Re)sampling of the reference population. Recently, (Mikshowsky *et al.* 2016) bootstrapped, not the validation, but the *reference* population. This also provides distribution of metrics. However, it may be argued that, in a dairy cattle reference population, including a sire twice (what the bootstrapping actually does) is like including it once, because the accuracy of the sire pseudo-phenotype is close to 1 in dairy cattle. Thus, including it twice will not change much the solution for the sire – or the contribution of the sire to SNPs solutions. Therefore, randomness comes from *removing* sires more than by *overrepresenting* sires. In that sense, Mikshowsky *et al.* (2016) bootstrap corresponds to Tukey's jackknife with more than one data point removed.

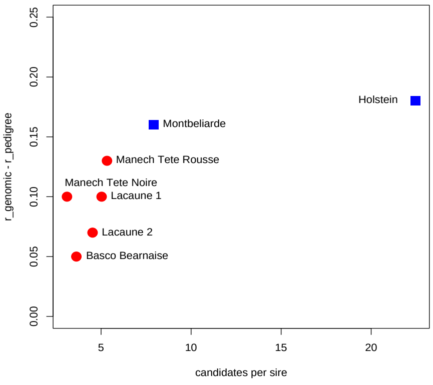

Figure 3 Genomic accuracy and family size.

Superiority of genomic on pedigree predictions is a function of family structure of the validation data set. Consider a set of two generations, a generation of parents and one of descendants: n full-sib families with k offspring each. Parents have information (say, own weight) but there is not information for the offspring. We can ask: is it worth doing genomic prediction?

Families can be easily ranked based on parent average, but there is not possibility to rank within families with pedigree information. However, genomic information *can* rank *within* family as well as *across* families. Thus, the observed benefit of GBLUP by retrospective analysis will be larger in a set composed of *few* families with a large number of candidates *within* families. In the limit, if there is one big family, pedigree prediction has 0 accuracy, whereas if there are n families with 1 offspring each, pedigree and genomic predictions should behave similarly.

This is supported by Figure 3 in which we plot the genomic vs pedigree accuracy for milk yield for five dairy sheep and two dairy cattle breeds in France, as a function of family size. Clearly, the larger the family size, the larger the benefit because genomic selection allows distinguishing sibs. This raises several questions: (1) Do comparisons reflect “genetic architecture” or merely data structure in the validation? (2) Do selection schemes that select across families get less benefit from genomic selection? (3) Is Holstein gaining a lot from genomic selection because it has higher LD than other breeds or just as an artefact of its family structure?

Which variables to use on the metrics? In the dairy industry, sires do not have phenotypes, so that comparisons are between (G)EBV’s and the “true” progeny proofs or deregressed proofs. In other species, it is more common to compare (G)EBV’s to “true” phenotypes, say $\boldsymbol{y}$, using an approximation $r = Corr(GEBV, y)/h$ where h^2 is the heritability (Legarra *et al.* 2008). This is unsatisfactory, for conceptual and practical reasons:

- The equation above for *r* assumes uncorrelated individuals and GEBV’s
- Records $\boldsymbol{y}$ are typically pre-corrected to $\boldsymbol{y}^* = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}$, and the results are sensitive to precorrection. It is unclear what happens if there are contemporary groups in $\boldsymbol{b}$ that are not present in the training data.
- If the whole data set is used for precorrection, then a relationship structure is fit (e.g. pedigree relationships) as $\boldsymbol{y}^* = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'(\boldsymbol{ZAZ}\sigma_u^2 + \boldsymbol{I}\sigma_e^2)^{-1}\boldsymbol{X})^{-})\boldsymbol{y}$ where $\boldsymbol{A}\sigma_u^2$ is assumed to be “correct”. If the assumed relationship is biased or incorrect, so will be $\hat{\boldsymbol{b}}$ and $\boldsymbol{y}^*$, and the bias will be toward the assumed relationship. This may explain some puzzling results, e.g. poor performance of genomic prediction in low heritable traits such as fertility (Hayes *et al.* 2009).
- Even after precorrection, there will be a remaining covariance structure across pre-corrected $\boldsymbol{y}^*$. This structure is notoriously hard to model (and rarely modelled). This may explain phenomena such as $\frac{Corr(GEBV,\boldsymbol{y}^*)}{h} > 1$.
- Some precorrected $\boldsymbol{y}^*$ are too clumsy (Ricard *et al.* 2013) to be believed or computed in practice, for instance maternal effects.

CROSS-VALIDATION ACCURACIES FROM METHOD R

Description of the method. We propose to use the properties of method R to construct metrics of cross-validation. Reverter *et al.* (1994) observed that the regression of EBVs obtained with “whole” (w) data on EBVs estimated with “partial” (p) data, $b_{w,p} = \frac{Cov(\hat{u}_w,\hat{u}_p)}{Var(\hat{u}_p)}$ is 1, and this checks bias (in the sense b_1 before). The correlation of partial on whole (eq. 7-9 in their paper) $\rho_{p,w} = \frac{Cov(\hat{u}_p,\hat{u}_w)}{\sqrt{Var(\hat{u}_w)Var(\hat{u}_p)}}$ is a function of respective accuracies. Invoking exchangeability, both equations can be extended to multivariate forms, and expectations can be taken in both the numerator and the denominator, resulting in:

$$b_{w,p} = \hat{\boldsymbol{u}}'_w\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_p/\hat{\boldsymbol{u}}'_p\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_p$$

where $\boldsymbol{K}$ is a matrix of relationships, $b_{p,w}$ with an expected value of 1, and

$$\rho_{w,p} = \hat{\boldsymbol{u}}'_p\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_w/\sqrt{\hat{\boldsymbol{u}}'_p\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_p\hat{\boldsymbol{u}}'_w\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_w}$$

with an expected value $E\left(\rho_{w,p}\right) = \sqrt{\frac{\mu_{acc_p^2}}{\mu_{acc_w^2}}}$ that is, proportional to the relative increase in average reliabilities. As more data cumulates, $\hat{\boldsymbol{u}}$ tends towards the true breeding values, thus $\hat{\boldsymbol{u}}_w$ is more accurate than $\hat{\boldsymbol{u}}_p$. The empirical covariance $\hat{\boldsymbol{u}}_w'\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_p$ measures the strength of the association between the two, whereas $\hat{\boldsymbol{u}}_p'\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_p$ measures the extent of shrinkage due to lack of information. In other words, the theoretical prediction error covariances are replaced by empirical ones (Thompson 2001). By combining cross-validation and theory from mixed models, we hope to retain the best of both worlds: a measure of accuracy that corresponds to reality and that is little affected by the existence of related, unbalanced data. Therefore, an algorithm to estimate accuracy of (say) PBLUP and GBLUP is:

1. Compute EBV's with all data ("whole") using, say, GBLUP (which method should not be critical if all animals have data or progeny)
2. Choose cutoff date
3. Create "partial" data: Set values after cutoff date to missing
4. Compute EBVs based on "partial" and GBLUP
5. Compute statistic $b_{w,p}^{GBLUP} = \frac{\hat{\mathbf{u}}_p'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_w}{\hat{\mathbf{u}}_p'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_p}$
6. Compute statistic $\rho_{p,w}^{GBLUP} = \frac{\hat{\mathbf{u}}_p'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_w}{\sqrt{\hat{\mathbf{u}}_w'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_w\hat{\mathbf{u}}_p'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_p}}$
7. Compute EBVs based on "partial" and PBLUP
8. Compute statistic $b_{w,p}^{PBLUP} = \frac{\hat{\mathbf{u}}_p'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_w}{\hat{\mathbf{u}}_p'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_p}$
9. Compute statistic $\rho_{p,w}^{PBLUP} = \frac{\hat{\mathbf{u}}_p'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_w}{\sqrt{\hat{\mathbf{u}}_w'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_w\hat{\mathbf{u}}_p'\boldsymbol{K}^{-1}\hat{\mathbf{u}}_p}}$

For forward cross-validation, the statistics should be computed for the focal individuals (i.e., candidates to selection). On exit, $b_{w,p}^{GBLUP}$ should be 1 (unbiased method) and is equivalent to b_1 and $\rho_{p,w}^{GBLUP}$ and $\rho_{p,w}^{PBLUP}$ describes the respective accuracies of GBLUP and PBLUP. An extra statistic is bias $\mu_{wp} = b_0 = (\boldsymbol{1}'\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_w - \boldsymbol{1}'\boldsymbol{K}^{-1}\hat{\boldsymbol{u}}_p)/n$. Matrix **K** should be the "true" relationship matrix across individuals but there should be no great difference in using either genomic or pedigree relationships as far as they are correct. The procedure has several advantages: is completely general (it can be used e.g. for maternal traits or random regression), it is semi-automatic, and can, at least potentially, provide estimates of the accuracy of the cross-validation metric. There are though many points that need to be addressed: robustness to misspecification, the role of selection (and how to avoid biases in the estimates of the different $b's$), how to sample efficiently, etc.

TEST WITH REAL LIFE DATA SETS

In beef cattle, we used genetic and phenotypic resources from Brahman cows (N = 995) and bulls (N = 1,116) outlined in (Porto-Neto *et al.* 2015). The phenotype was yearling body weight. A procedure "method R" as above was introduced to assess accuracy of GBLUP, and random (1000 replicates) splits of the data set in training and validation was used, as animals are quite unrelated and belong to a single generation. We only present very briefly the results. The statistic $b_{w,p} = 0.96 \pm 0.08$ (in the whole population) showed that evaluation was nearly unbiased, whereas $\rho_{p,w} = 0.67 \pm 0.02$ has a correlation of 0.81 with conventional cross-validation accuracy estimated as $\frac{Corr(GEBV, \boldsymbol{y}^{*})}{h}$.

In dairy sheep, we used a large data set (Manech Tete Rousse) of 1,700,000 milk yield performances, 500,000 animals in pedigree and 2,111 sires with 50K genotypes. Data was split at 2011 in training and validation. For *all* individuals, unbiasedness of (SSG)BLUP was checked with results $\mu_{w,p} = b_0 = 0.2\sigma_g = 5$ (liters), $b_{w,p} = b_1 = 0.996$, so genetic evaluation is virtually unbiased for b_1 (slope) but not for b_0 (genetic trend), which is unsurprising because the model includes Unknown Parent Groups. Later, candidates to selection were compared, with $\rho_{w,p}^{SSGBLUP} = 0.55$ vs. $\rho_{w,p}^{BLUP} = 0.39$, and both evaluations where notoriously biased ($b_1^{SSGBLUP} = 0.77$, $b_1^{BLUP} = 0.70$), possibly due to selection not well accounted for. All these results agree well with previous analysis (Legarra *et al.* 2014).

ACKNOWLEDGEMENTS

Toni Reverter and Andrés Legarra benefit funding from the INRA-CSIRO linkage proposals 2016/2017. AL also financed by grants GENOMIA (Poctefa, Feder) and GenoPyr (Feder).

REFERENCES

Baloche G., Legarra A., Sallé G., Larroque H., Astruc J. M., Robert-Granié C., Barillet F. (2014) *J. Dairy Sci.* **97**: 1107–1116.

Bijma P. (2012) *J. Anim. Breed. Genet.* **129**: 345–358.

Boichard D., Bonaiti B., Barbat A., Mattalia S. (1995) *J. Dairy Sci.* **78**: 431–437.

Bonaiti B., Boichard D., Barbat A., Mattalia S. (1993) *Interbull Bull.* **8**

Christensen O., Madsen P., Nielsen B., Ostersen T., Su G. (2012) *Animal* **6**: 1565–1571.

Cuyabano B. C. D., Su G., Rosa G. J. M., Lund M. S., Gianola D. (2015) *J. Dairy Sci.* **98**: 7351–7363.

Edel, C., Neuner, S., Emmerling, R. and Goetz, K.U., 2012. *Interbull Bull.* **46**

Gianola D., Schön C.-C. (2016) *G3 GenesGenomesGenetics* **6**: 3107–3128.

Hayes B. J., Bowman P. J., Chamberlain A. J., Goddard M. E. (2009) *J Dairy Sci* **92**: 433–443.

Henderson C. R. (1973) *J Anim Sci* (**Symposium**) 10-41

Legarra A., Robert-Granié C., Manfredi E., Elsen J.-M. (2008) *Genetics* **180**: 611–618.

Legarra A., Baloche G., Barillet F., Astruc J., Soulas C., Aguerre X., Arrese F., Mintegi L., Lasarte M., Maeztu F. (2014*) J. Dairy Sci.* **97**: 3200–3212.

Mantysaari E., Liu Z., VanRaden P. (2010) *Interbull Bull* **41**.

Mäntysaari E. A., Koivula M. (2012) *Interbull Bull.* **46**

Mikshowsky A. A., Gianola D., Weigel K. A. (2016) *J. Dairy Sci.* **99**: 3632–3645.

Olson K., VanRaden P., Tooker M., Cooper T. (2011) *J. Dairy Sci.* **94**: 2613–2620.

Porto-Neto L. R., Barendse W., Henshall J. M., McWilliam S. M., Lehnert S. A., Reverter A. (2015) *Genet. Sel. Evol.* **47**: 84.

Powell R. L., Wiggans G. R. (1994) *Interbull Bull.* **10**.

Reverter A., Golden B. L., Bourdon R. M., Brinks J. S. (1994*) J. Anim. Sci.* **72**: 34–37.

Ricard A., Danvy S., Legarra A. (2013) *J. Anim. Sci.* **91**: 1076–1085.

Robertson A. (1977) Z. Für Tierz. Zücht. **94**: 131–135.

Saatchi M., McClure M. C., McKay S. D., Rolf M. M., Kim J., et al. (2011) *Genet. Sel. Evol.* **43**: 40.

Thompson R. (2001) *Livest. Prod. Sci.* **72**: 129–134.

VanRaden P. M., Tassell C. P. V., Wiggans G. R., Sonstegard T. S., Schnabel R. D., Taylor J. F., Schenkel F. S. (2009) *J Dairy Sci* **92**: 16–24.

Vitezica Z., Aguilar I., Misztal I., Legarra A. (2011) *Genet. Res.* **93**: 357–366.

Xiang T., Nielsen B., Su G., Legarra A., Christensen O. F. (2016) *J. Anim. Sci.* **94**: 936–948.

frontiers in Genetics

REVIEW
published: 06 March 2018
doi: 10.3389/fgene.2018.00078

Non-additive Effects in Genomic Selection

Luis Varona[1,2]*, Andres Legarra[3], Miguel A. Toro[4] and Zulma G. Vitezica[5]

[1] Departamento de Anatomía, Embriología y Genética Animal, Universidad de Zaragoza, Zaragoza, Spain, [2] Instituto Agroalimentario de Aragón (IA2), Zaragoza, Spain, [3] Génétique Physiologie et Systèmes d'Elevage (GenPhySE), Institut National de la Recherche Agronomique de Toulouse, Castanet-Tolosan, France, [4] Departamento Producción Agraria, ETS Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid, Madrid, Spain, [5] Génétique Physiologie et Systèmes d'Elevage (GenPhySE), Université de Toulouse, Castanet-Tolosan, France

In the last decade, genomic selection has become a standard in the genetic evaluation of livestock populations. However, most procedures for the implementation of genomic selection only consider the additive effects associated with SNP (Single Nucleotide Polymorphism) markers used to calculate the prediction of the breeding values of candidates for selection. Nevertheless, the availability of estimates of non-additive effects is of interest because: (i) they contribute to an increase in the accuracy of the prediction of breeding values and the genetic response; (ii) they allow the definition of mate allocation procedures between candidates for selection; and (iii) they can be used to enhance non-additive genetic variation through the definition of appropriate crossbreeding or purebred breeding schemes. This study presents a review of methods for the incorporation of non-additive genetic effects into genomic selection procedures and their potential applications in the prediction of future performance, mate allocation, crossbreeding, and purebred selection. The work concludes with a brief outline of some ideas for future lines of that may help the standard inclusion of non-additive effects in genomic selection.

Keywords: genomic selection, dominance, epistasis, crossbreeding, genetic evaluation

OPEN ACCESS

Edited by:
Rohan Luigi Fernando,
Iowa State University, United States

Reviewed by:
Romdhane Rekaya,
University of Georgia, United States
Piter Bijma,
Wageningen University & Research,
Netherlands

****Correspondence:***
Luis Varona
lvarona@unizar.es

Specialty section:
This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: *15 December 2017*
Accepted: *19 February 2018*
Published: *06 March 2018*

Citation:
Varona L, Legarra A, Toro MA and Vitezica ZG (2018) Non-additive Effects in Genomic Selection. Front. Genet. 9:78. doi: 10.3389/fgene.2018.00078

INTRODUCTION

Through his experiments on pea plants, Gregor Mendel (1866) realized that some traits are dominant over others (for example "round peas" were dominant over "wrinkled peas"). In Mendel's own words: "As a rule, hybrids do not represent the form exactly intermediate between the parental strains… Those traits that pass into hybrid association entirely or almost entirely unchanged, thus themselves representing the traits of the hybrid, are termed "dominating," and those that become latent in the association, "recessive"". Shortly after the rediscovery of Mendel's rules, it was observed that, in some cases, the addition of the individual action of genes could not explain the mode of inheritance, and Bateson (1909) coined the term "epistasis" to describe the cases in which the actions of two or more genes interact. A distinction must be drawn between biological (functional) genetic effects that correspond to the Mendelian definition (i.e., dominance means that the heterozygote value is higher or lower than the mean of homozygous genotypes) and statistical (population or weighted) effects which depend on allelic frequencies. In the latter, the relevant issue is the contribution of non-additive effects to genetic variance. Some authors argue that non-additive genetic effects may be a general phenomenon whose understanding is important for gaining more knowledge on the nature of quantitative traits, but whose contribution to variance is negligible (Crow, 2010).

From the perspective of quantitative genetics, Fisher (1918) conceived the infinitesimal model which postulates that a very large number of unlinked genes control the genetic variation of quantitative traits. He described the resemblance between relatives in a pure additive model which was quickly extended to incorporate dominance (Fisher, 1918; Wright, 1921). Resemblance between relatives including epistatic effects of second and higher order was also described (Cockerham, 1954; Kempthorne, 1954). However, whilst the formulation of the infinitesimal model in the additive context is evident, its interpretation is not clear when non-additive effects are included (Barton et al., 2017).

The main goal of animal or plant breeding is to identify, select and mate the best individuals of a breeding stock in order to maximize performance in future generations (Falconer and McKay, 1996; Bernardo, 2010). The procedure for computing the breeding values (genetic evaluation) of candidates for selection plays a crucial role. Traditionally, these methods use phenotypic and genealogical information, such as the selection index (Hazel, 1943) or the Best Linear Unbiased Predictor (Henderson, 1973) and rely on the foundations of the infinitesimal model (Fisher, 1918).

Nevertheless, non-additive genetic effects have been ignored in the genetic evaluation of livestock for several reasons: (i) the lack of informative pedigrees, such as large full-sib families; (ii) the calculations involved are more complex; (iii) the fact that statistical additive variance captures biological dominance or higher order interaction effects (Hill, 2010); and, (iv) the difficulty in using dominant values in practice (mate allocation). As a consequence, estimates of non-additive genetic variances are scarce in livestock populations (Misztal et al., 1998; Nguyen and Nagyné-Kiszlinger, 2016).

GENOMIC SELECTION

Since the late 80s and 90s, developments in molecular genetics resulted in a set of neutral molecular markers, such as microsatellites, that were commonly used to detect QTL (Quantitative Trait Loci) in almost all livestock populations. The objective of those studies was to identify polymorphic markers or genes associated with phenotypic variation of traits of interest (www.animalgenome.org/QTL), with the ultimate goal of using them in Marker or Gene Assisted Selection (Dekkers, 2004). However, these strategies became obsolete with the advent of dense genotyping devices (Gunderson et al., 2005) that provided a very large amount of SNP (Single Nucleotide Polymorphism) and that allowed the development of genomic selection (GS) models (Meuwissen et al., 2001).

Genomic selection has become a very successful strategy for the prediction of the breeding values of candidates for selection and has revolutionized the field of animal breeding over the past decade. The basic idea of GS is to develop the following linear model:

$$y_i = \mu + \sum_{j=1}^{n} t_{ij} a_j + e_i$$

The model explains the phenotypic data of m individuals (y_i) with $i = 1 \ldots m$ (or transformations of data, such as daughter yield deviations) by the effects associated with a very large number (n) of SNP (a_j) with $j = 1 \ldots n$. Moreover, t_{ij} is the genotypic configuration (coded additively, e.g., Falconer and McKay, 1996) of the *ith* individual and for the *jth* SNP (0, 1, and 2 for A_1A_1, A_1A_2, and A_2A_2 genotypes, respectively), and e_i is the residual. Furthermore, the prediction of individual breeding values ($\hat{u}_i$) of the candidates for selection can be calculated *a posteriori* from marker effect estimates as $\hat{u}_i = \sum_{j=1}^{n} t_{ij}\hat{a}_j$.

A significant limitation for implementation is that most genomic evaluation models suffer the statistical problem of a larger number of parameters (n) that must be estimated from a smaller number of data (m). The most common method employed for resolving this problem is the use of some type of regularization of SNP marker effects (Gianola, 2013). Several approaches have been suggested, ranging from a simple Gaussian regularization (Meuwissen et al., 2001) to more complex models that involve t shaped (Meuwissen et al., 2001), double exponential (De los Campos et al., 2009b), mixtures of distributions (Meuwissen et al., 2001; Habier et al., 2011; Erbe et al., 2012), or non-parametric or semi-parametric approaches (Gonzalez-Recio et al., 2014). The predictive ability of all these approaches depends on the genetic architecture of the traits being analyzed (Daetwyler et al., 2010), although for polygenic traits, all approaches offer similar results (Wang et al., 2015).

An interesting property of the assumption of a Gaussian prior distribution for marker effects (Random Regression BLUP—RR-BLUP) is that the GS model can be reformulated in terms of individual (animal) effects, using the equations of the Henderson's classic Mixed Model that provide breeding values for all individuals, including candidates for selection (Genomic BLUP or GBLUP). The only difference with standard mixed model equations is that the numerator relationship matrix (**A**) is replaced by the genomic relationship matrix (**G**), as defined by VanRaden (2008). In addition, this approach can be extended for the genetic evaluation of non-genotyped individuals in the Single-Step approach (Aguilar et al., 2010), facilitating the integration of GS procedures in the genetic evaluation of candidates for selection in most livestock breeding programmes. More recently, Fernando et al. (2014) described a Bayesian procedure that can also simultaneously evaluate genotyped and non-genotyped individuals and allows the use of alternative regularization procedures. Nevertheless, computational costs are markedly higher with the Bayesian model than with the Single-Step approach.

Despite the regularization procedure, the genomic evaluation methods are based on the evaluation of marker substitution effects through the construction of the covariates (t_{ij}) or the **G** matrix (above). The additive (or breeding) values capture a large part of dominant and higher-order interaction effects (Hill et al., 2008; Crow, 2010; Hill, 2010). Substitution effects that capture dominance and epistatic functional effects are not necessarily stable across generations or populations due to changes in allelic frequencies. In any case, only additive values (substitution effects) contribute to breeding values and are therefore expressed in the next generation. However, estimates of non-additive genetic effects may be of relevance because: (i) they may contribute to increasing the accuracy of prediction of breeding values and the response to selection (Toro and Varona, 2010; Aliloo et al., 2016; Duenk et al., 2017); (ii) they allow the definition of mate allocation procedures between candidates for selection (Maki-Tanila, 2007; Toro and Varona, 2010; Aliloo et al., 2017); and (iii) they can be used to benefit from non-additive genetic variation through the definition of appropriate crossbreeding or purebred breeding schemes (Maki-Tanila, 2007; Zeng et al., 2013).

GENOMIC SELECTION MODELS WITH DOMINANCE

The simplest approach for the inclusion of dominance in genomic selection models is to extend the basic model with the inclusion of a dominance effect (Toro and Varona, 2010; Su et al., 2012) associated to each SNP marker:

$$y_i = \mu + \sum_{j=1}^{n} t_{ij} a_j + \sum_{j=1}^{n} c_{ij} d_j + e_i$$

where y_i is the phenotypic value of the *ith* individual and μ is the population mean. For each of the *n* SNP markers, a_j and d_j are the additive and dominance effects for the *jth* marker, respectively. The covariates t_{ij} and c_{ij} are 2, 1, and 0 (coded additively) and 0, 1, and 0 (coded in a "biological dominant" manner) for the genotypes A_1A_1, A_1A_2, A_2A_2 of each marker, respectively. In some ways, pedigree-based models for dominance were based on "expected" dominant relationships. Thus, genomic models are based on "observed" heterozygotes. However, when using this model it should be noted that that a_j is no longer the marker substitution effect, but the "biological" additive genotypic effect and individual breeding values are not predicted. In fact, the partition of variance in statistical components due to additivity, dominance, and epistasis does not reflect the "biological" (or "functional") effect of the genes although it is useful for prediction and selection (Huang and Mackay, 2016). The model was reformulated in terms of breeding values and dominance deviations (Falconer and Mackay, 1996) by Vitezica et al. (2013) after the assumption of a Hardy-Weinberg equilibrium within each:

$$y_i = \mu + \sum_{j=1}^{n} w_{ij} \alpha_j + \sum_{j=1}^{n} g_{ij} d_j + e_i$$

where

$$w_{ij} = \begin{cases} (2 - 2p_j) & A_1A_1 \\ (1 - 2p_j) & A_1A_2 \\ -2p_j & A_2A_2 \end{cases} \qquad g_{ij} = \begin{cases} -2q_j^2 & A_1A_1 \\ 2p_jq_j & A_1A_2 \\ -2p_j^2 & A_2A_2 \end{cases}$$

and $\alpha_j = a_j + d_j \left(q_j - p_j\right)$ is now the allelic substitution effect and p_j and q_j are the allelic frequencies for A_1 and A_2 for the *jth* SNP marker. The genetic variance due to a single locus is:

$$\sigma_{Gj}^2 = 2p_jq_j \left[a_j + d_j \left(q_j - p_j\right)\right]^2 + \left(2p_jq_jd_j\right)^2$$

where the additive variance is $\sigma_{Aj}^2 = 2p_jq_j \left[a_j + d_j \left(q_j - p_j\right)\right]^2 = 2p_jq_j\alpha_j^2$ and the dominance variance is $\sigma_{Dj}^2 = \left(2p_jq_jd_j\right)^2$ and the multilocus variances, under linkage equilibrium (LE), are $\sigma_G^2 = \sum_{j=1}^{n} \sigma_{Gj}^2$, $\sigma_A^2 = \sum_{j=1}^{n} \sigma_{Aj}^2$ and $\sigma_D^2 = \sum_{j=1}^{n} \sigma_{Dj}^2$. In fact, "biological" (in terms of genotypic additive and dominant values) and "statistical" (in terms of breeding values and dominance deviations) models are equivalent parameterisations of the same model (Vitezica et al., 2013), and the following expressions:

$$\begin{aligned} \sigma_A^2 &= \sum_{j=1}^{n} \left(2p_jq_j\right) \sigma_a^2 + \sum_{j=1}^{n} \left(2p_jq_j \left(q_j - p_j\right)^2\right) \sigma_d^2 \\ \sigma_{A^*}^2 &= \sum_{j=1}^{n} \left(2p_jq_j\right) \sigma_a^2 \\ \sigma_D^2 &= \sum_{j=1}^{n} \left(4p_j^2q_j^2\right) \sigma_d^2 \\ \sigma_{D^*}^2 &= \sum_{j=1}^{n} \left(2p_jq_j \left(1 - 2p_jq_j\right)\right) \sigma_d^2 \end{aligned}$$

that can be used to switch variance components estimates between "biological" (σ_{A*}^2 and σ_{D*}^2) and "statistical" (σ_A^2 and σ_D^2) models. It can be verified that $\sigma_A^2 + \sigma_D^2 = \sigma_{A*}^2 + \sigma_{D*}^2$. In addition, if $p = q = 0.5$, all variances are identical and if $d = 0$, $\sigma_A^2 = \sigma_{A*}^2$. A further generalization can be also achieved to avoid the requirements of the Hardy-Weinberg equilibrium (Vitezica et al., 2017), by following the NOIA model (Alvarez-Castro and Carlborg, 2007) by replacing w_{ij} and g_{ij} with:

$$w_{ij} = \begin{cases} -\left(-p_{12j} - 2p_{22j}\right) & A_1A_1 \\ -\left(1 - p_{12j} - 2p_{22j}\right) & A_1A_2 \\ -\left(2 - p_{12j} - 2p_{22j}\right) & A_2A_2 \end{cases}$$

$$g_{ij} = \begin{cases} -\dfrac{2p_{12j}p_{22j}}{p_{11j} + p_{22j} - \left(p_{11j} - p_{22j}\right)^2} & A_1A_1 \\ \dfrac{4p_{11j}p_{22j}}{p_{11j} + p_{22j} - \left(p_{11j} - p_{22j}\right)^2} & A_1A_2 \\ -\dfrac{2p_{11j}p_{12j}}{p_{11j} + p_{22j} - \left(p_{11j} - p_{22j}\right)^2} & A_2A_2 \end{cases}$$

where, p_{11j}, p_{12j}, and p_{22j} are the genotypic frequencies for A_1A_1, A_1A_2, and A_2A_2 at the *jth* SNP marker, respectively.

Note that all these models require a regularization process for additive and dominance effects. The simplest approach is to expand the RR-BLUP by the assumption of a prior Gaussian distribution for the additive and dominance effects. It is feasible to assume any other kind of prior distribution for the dominance (as described above) and the additive effects (Acevedo et al., 2015). However, a major advantage of using a Gaussian prior distribution is that the model can be easily transformed into Henderson's Mixed Model equations by using the definition of additive (**G**) and dominance covariance matrices (**D**), as suggested by Vitezica et al. (2013).

Genomic selection models with dominance have been tested in several populations, including dairy cattle (Ertl et al., 2014;

Aliloo et al., 2016; Jiang et al., 2017), pigs (Esfandyari et al., 2016; Xiang et al., 2016), sheep (Moghaddar and van der Werf, 2017), and layers (Heidaritabar et al., 2016) with ambiguous results. Jiang et al. (2017) found a negligible percentage of variation explained by dominance effects for productive life in a Holstein cattle population, although Ertl et al. (2014) suggested that dominance may suppose up to 39% of the total genetic variation for Somatic Cell Score in a population of Fleckvieh cattle. In general, the increase in the accuracy of additive breeding values by including dominance was scarce, with the exception of Aliloo et al. (2016).

DOMINANCE AND INBREEDING DEPRESSION (OR HETEROSIS)

The classical theory of quantitative genetics (Falconer and Mackay, 1996) postulates that inbreeding depression (or heterosis) occurs due to directional dominance. However, the presence of directional dominance (i.e., a higher percentage of positive than negative dominant effects) is in sharp contrast to the assumptions of the procedures described above that use symmetric prior distributions. This drawback can be overcome by the assumption of a mean of dominant effects that is different from zero, e.g., $E\left(d\right) = \mu_d$, as proposed by Xiang et al. (2016). The standard model can be reformulated as:

$$\begin{aligned} y_i &= \mu + \sum_{j=1}^{n} t_{ij}a_j + \sum_{j=1}^{n} c_{ij}\left[d_j^* + \mu_d\right] + e_i \\ &= \mu + \sum_{j=1}^{n} t_{ij}a_j + \sum_{j=1}^{n} c_{ij}d_j^* + \sum_{j=1}^{n} c_{ij}\mu_d + e_i \end{aligned}$$

where $d_j^* = d_j - \mu_d$, then $E\left(d^*\right) = 0$. It should be noted the term $\sum_{j=1}^{n} c_{ij}\mu_d$ is an average of dominance effects for the *ith* individual, because c_{ij} has a value of 1 for heterozygous loci and 0 for homozygous. Inbreeding (or full homozygosity) coefficients f_i can be calculated as:

$$f_i = 1 - \frac{\sum_{j=1}^{n} c_{ij}}{n}$$

So, $\sum_{j=1}^{n} c_{ij}\mu_d = \left(1 - f_i\right) n\mu_d = n\mu_d - f_i n\mu_d$. The first term $n\mu_d$ is absorbed in the overall mean of the model (μ), and the second ($-f_i n\mu_d$) corresponds to a covariate $b = -n\mu_d$ associated with inbreeding (f_i). This covariate can be seen as inbreeding depression (if it has a detrimental effect) caused by genomic inbreeding. In addition, it can be also implemented in the GBLUP models described above with the introduction of a covariate within the mixed model equations.

Nonetheless, it assumes that the expected mean of the dominance effects is the same for all markers. In the literature, there are signs that the decrease in performance is associated heterogeneously within the genomic regions (Pryce et al., 2014; Howard et al., 2015; Saura et al., 2015). Models that consider alternative means of dominance effects within genomic regions may be useful to model inbreeding depression in a more appropriate way.

An alternative approach to explain the phenomenon of inbreeding depression (or heterosis) is the consideration of a possible relationship between additive and dominance biological effects (Wellmann and Bennewitz, 2011). There is theoretical proofs (Caballero and Keightley, 1994) and empirical evidence (Bennewitz and Meuwissen, 2010) that supports this argument. Wellmann and Bennewitz (2012) expanded the "biological" model described above with regularization procedures that allows for this dependence. They defined up to four models (Bayes D0 to D3) based on the Bayes C approach (Verbyla et al., 2009). The last two models (Bayes D2 and D3) included dependencies between genotypic additive and dominance effects. In the first (D2), the dependence was modeled through the prior variance of the dominance effects ($Var\left(d\middle|\left|a\right|\right)$) and in the second (D3), they further expanded it to the prior mean ($E\left(d\middle|\left|a\right|\right)$), where $\left|a\right|$ is the absolute value of the additive effect. Implementation of these models is extremely complex and they have not been thoroughly tested (Bennewitz et al., 2017).

IMPRINTING

Another source of non-additive genetic variation is genomic imprinting (Reik and Walter, 2001). This involves total or partial inactivation of paternal and maternal alleles. Following the quantitative model established by Spencer (2002), Nishio and Satoh (2015) put forward two alternative genomic selection models to include imprinting effects. The first extends the "statistical" model with dominance (in terms of breeding values and dominance deviations) as:

$$y_i = \mu + \sum_{j=1}^{n} w_{ij}\alpha_j + \sum_{j=1}^{n} g_{ij}d_j + \sum_{j=1}^{n} r_{ij}i_j + e_i$$

where

$$w_{ij} = \begin{cases} 2 - 2p_j & A_1A_1 \\ 1 - 2p_j & A_1A_2 \\ 1 - 2p_j & A_2A_1 \\ -2p_j & A_2A_2 \end{cases}$$

$$g_{ij} = \begin{cases} -2q_j^2 & A_1A_1 \\ 2p_jq_j & A_1A_2 \\ 2p_jq_j & A_2A_1 \\ -2p_j^2 & A_2A_2 \end{cases} \text{and } r_{ij} = \begin{cases} 0 & A_1A_1 \\ 1 & A_1A_2 \\ -1 & A_2A_1 \\ 0 & A_2A_2 \end{cases}$$

and i_j is the imprinting effects associated with *jth* marker. The second alternative proposed the distribution of the genetic effects into paternal (p_j) and maternal (m_j) gametic effects and a dominance deviation.

$$y_i = \mu + \sum_{j=1}^{n} l_{ij}p_j + \sum_{j=1}^{n} j_{ij}m_j + \sum_{j=1}^{n} g_{ij}d_j + e_i$$

where

$$l_{ij} = j_{ij} = \begin{cases} q_j & A_1 \\ -\left(1 - q_j\right) & A_2 \end{cases}$$

These models have been implemented in some studies with livestock data: (Hu et al., 2016) did not find an increase in predictive ability when imprinting effects were included in the model. In addition, estimates of the percentage of phenotypic variation caused by imprinting were small and ranged between 1.3 and 1.4% in pigs (Guo et al., 2016) and from 0.2 to 2.1% in dairy cattle (Jiang et al., 2017). However, this latter study reported that imprinting effects supposed more than 20% of the total genetic variance in some reproductive traits, like pregnancy or conception rate.

EPISTASIS

The last and most complex source of non-additive genetic variation is the epistatic interactions between two or more genes. An immediate approach for genomic evaluation including epistatic interactions is to define an explicit model by including pairwise or higher order epistatic effects:

$$\begin{aligned} y_i = \mu &+ \sum_{j=1}^{n} t_{ij} a_j + \sum_{j=1}^{n} c_{ij} d_j + \sum_{j=1}^{n}\sum_{k=1}^{n} t_{ij} t_{ik} aa_{jk} \\ &+ \sum_{j=1}^{n}\sum_{k=1}^{n} t_{ij} g_{ik} ad_{jk} + \sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij} g_{ik} dd_{jk} \\ &+ \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} t_{ij} t_{ik} t_{il} aaa_{jkl} + \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} t_{ij} t_{ik} g_{il} aad_{jkl} \\ &+ \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} t_{ij} g_{ik} g_{il} add_{jkl} + \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} g_{ij} g_{ik} g_{il} ddd_{jkl} \\ &+ \ldots + e_i \end{aligned}$$

where aa_{jk}, ad_{jk}, and dd_{jk} are second order additive x additive, additive x dominant and dominant x dominant epistatic effects between the *jth* and *kth* SNP effects and aaa_{jkl}, aad_{jkl}, add_{jkl} and ddd_{jk} are third order additive x additive x additive, additive x additive x dominant, additive x dominant x dominant and dominant x dominant x dominant epistatic effects. Despite the method of regularization used, the number of parameters to estimate is extremely large. Consequently, the computational requirements are enormous and the amount of information available, in the statistical sense, for the estimation of each epistatic effect is very small. Therefore, the most efficient (at least from a computational point of view) method for including epistatic interactions in genomic selection models is to define appropriate covariance matrices between individual effects, in the same way that the standard GBLUP model uses the genomic relationship matrix, but, in this case, taking into account the interactive nature of the genetic effects. There are two main approaches in the published literature: (1) the definition of genomic relationship matrices that consider epistatic interactions (Varona et al., 2014; Martini et al., 2016; Vitezica et al., 2017), and (2) the application of Kernel-based statistical methods (Gianola et al., 2006; de los Campos et al., 2009a; Morota and Gianola, 2014).

This simplest method for defining genomic relationship matrices is the *extended GBLUP model (EGBLUP)*, described by Jiang and Reif (2015) and Martini et al. (2016). These authors start from a reduced version of the "biological" model:

$$y_i = \mu + \sum_{j=1}^{n} t_{ij} a_j + \sum_{j=1}^{n}\sum_{k=1}^{n} t_{ij} t_{ik} aa_{jk} + e_i$$

and they define an equivalent model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{g_1} + \mathbf{g_2} + \mathbf{e}$$

where μ is the general mean, $\mathbf{y}$ is the vector of phenotypic data and $\mathbf{e}$ is the vector of the residuals. In addition, the model includes one "biologically" additive ($\mathbf{g_1}$) and one epistatic ($\mathbf{g_2}$) multivariate Gaussian term with the following distributions:

$$\mathbf{g_1} \sim N\left(0, \mathbf{G_1}\sigma^2_{g1}\right) \qquad \mathbf{g_2} \sim N\left(0, \mathbf{G_2}\sigma^2_{g2}\right)$$

Where $\mathbf{G_1} = \mathbf{TT}'$ and $\mathbf{G_2} = \mathbf{G_1} \circ \mathbf{G_1}$ being:

$$\mathbf{T} = \begin{bmatrix} t_{11} \cdots t_{1n} \\ \vdots \quad \ddots \quad \vdots \\ t_{k1} \cdots t_{kn} \end{bmatrix}$$

and the Hadamard product. Moreover, n is the number of SNP markers and k the number of individuals. However, with this model the additive and epistatic effects are not orthogonal and dominant effects are not included. Therefore, it can only be used for prediction of the phenotypes and not for the estimation of variance components (Martini et al., 2016). To avoid this inconvenience, Varona et al. (2014) and Vitezica et al. (2017) developed a full orthogonal model. They start with the expansion of the individual genotypic effect into additive, dominance and epistatic effects:

$$\begin{aligned} \mathbf{y} = \mathbf{1}\mu + \mathbf{g} + \mathbf{e} = \mathbf{1}\mu + \mathbf{g_A} + \mathbf{g_D} + \sum_{i=A,D}\sum_{j=A,D} \mathbf{g_{ij}} \\ + \sum_{i=A,D}\sum_{j=A,D}\sum_{k=A,D} \mathbf{g_{ijk}} + \ldots + \mathbf{e} \end{aligned}$$

Where $\mathbf{g}$ is the vector of the individual genotypic effects, $\mathbf{g_A}$ is the vector of additive effects, $\mathbf{g_D}$ the vector of individual dominance effects, $\mathbf{g_{ij}}$ is the second order epistatic effects, $\mathbf{g_{ijk}}$ the third order epistatic effects and so on. For simplicity, each individual effect is defined by the sum of SNP (or combination of SNP) effects $\mathbf{h}$ with equal prior Gaussian variability and weighted by an incidence matrix ($\mathbf{H}$). So, for the additive and dominant effects, $\mathbf{g_A} = \mathbf{H_A a}$ and $\mathbf{g_D} = \mathbf{H_D d}$: :

$$\mathbf{H_A} = \begin{pmatrix} \mathbf{h}_{\mathrm{A1}} \\ \cdots \\ \mathbf{h}_{\mathrm{Ak}} \end{pmatrix} \qquad \text{and } \mathbf{H_D} = \begin{pmatrix} \mathbf{h}_{\mathrm{D1}} \\ \cdots \\ \mathbf{h}_{\mathrm{Dk}} \end{pmatrix}$$

Where each $\mathbf{h}$ vector is composed by n (number of SNP markers) elements ($\mathbf{h}_{\mathrm{Ai}} = \left\{h_{Ai1}, h_{Ai2}, \ldots, h_{Ain}\right\}$ and

$\mathbf{h}_{\mathrm{Di}} = \left\{h_{Di1}, h_{Di2}, \ldots, h_{Din}\right\}$) and **a** and **d** are the vectors of the SNP additive and dominant effects. These $\mathbf{h}_{\mathrm{Ai}}$ and $\mathbf{h}_{\mathrm{Di}}$vectors can be defined in several ways, depending of the reference point or the assumption of the Hardy-Weinberg equilibrium, among others. However, orthogonal partitioning of variances must follow the NOIA approach (Alvarez-Castro and Carlborg, 2007):

$$h_{Aij} = \begin{cases} -\left(-p_{12j} - 2p_{22j}\right) \; A_1A_1 \\ -\left(1 - p_{12j} - 2p_{22j}\right) \; A_1A_2 \\ -\left(2 - p_{12j} - 2p_{22j}\right) \; A_2A_2 \end{cases}$$

$$h_{Dij} = \begin{cases} -\dfrac{2p_{12j}p_{22j}}{p_{11j} + p_{22j} - \left(p_{11j} - p_{22j}\right)^2} \; A_1A_1 \\ \dfrac{4p_{11j}p_{22j}}{p_{11j} + p_{22j} - \left(p_{11j} - p_{22j}\right)^2} \; A_1A_2 \\ -\dfrac{2p_{11j}p_{12j}}{p_{11j} + p_{22j} - \left(p_{11j} - p_{22j}\right)^2} \; A_2A_2 \end{cases}$$

Therefore, and under the assumption that SNP additive or dominant effects follow a Gaussian distribution, the additive and dominant "genomic" (co) variance relationship matrices can be computed as:

$$Cov\left(\boldsymbol{g}_A\right) = \frac{\mathbf{H_A H_A^{'}}}{tr\left(\mathbf{H_A H_A^{'}}\right)/n}\sigma_A^2 \qquad Cov\left(\boldsymbol{g}_D\right) = \frac{\mathbf{H_D H_D^{'}}}{tr\left(\mathbf{H_D H_D^{'}}\right)/n}\sigma_D^2$$

where the division by traces standardizes the variance components to an ideal infinite "unrelated" population. For second order epistatic effects ($\mathbf{g}_{\mathrm{AA}}$, $\mathbf{g}_{\mathrm{AD},}$ and $\mathbf{g}_{\mathrm{DD}}$), Alvarez-Castro and Carlborg (2007) proved that:

$$\mathbf{h}_{\mathrm{AAij}} = \mathbf{h}_{\mathrm{Ai}} \otimes \mathbf{h}_{\mathrm{Aj}} \qquad \mathbf{h}_{\mathrm{ADij}} = \mathbf{h}_{\mathrm{Ai}} \otimes \mathbf{h}_{\mathrm{Dj}} \qquad \mathbf{h}_{\mathrm{DDij}} = \mathbf{h}_{\mathrm{Di}} \otimes \mathbf{h}_{\mathrm{Dj}}$$

and, as a consequence, the matrices $\mathbf{H}_{\mathrm{AA}}$, $\mathbf{H}_{\mathrm{AD}}$ and $\mathbf{H}_{\mathrm{DD}}$ can be written as:

$$\mathbf{H_{AA}} = \begin{pmatrix} \mathbf{h_{A1}} \otimes \mathbf{h_{A1}} \\ \mathbf{h_{A2}} \otimes \mathbf{h_{A2}} \\ . \\ \mathbf{h_{An}} \otimes \mathbf{h_{An}} \end{pmatrix} \mathbf{H_{AD}} = \begin{pmatrix} \mathbf{h_{A1}} \otimes \mathbf{h_{D1}} \\ \mathbf{h_{A2}} \otimes \mathbf{h_{D2}} \\ . \\ \mathbf{h_{An}} \otimes \mathbf{h_{Dn}} \end{pmatrix}$$

$$\mathbf{H_{DD}} = \begin{pmatrix} \mathbf{h_{D1}} \otimes \mathbf{h_{D1}} \\ \mathbf{h_{D2}} \otimes \mathbf{h_{D2}} \\ . \\ \mathbf{h_{Dn}} \otimes \mathbf{h_{Dn}} \end{pmatrix}$$

and, as before, under the assumption of Gaussian distribution of second-order epistatic effects, the covariance between them can be calculated as:

$$Cov\left(\boldsymbol{g}_{AA}\right) = \frac{\mathbf{H_{AA} H_{AA}^{'}}}{tr\left(\mathbf{H_{AA} H_{AA}^{'}}\right)/n}\sigma_{AA}^2 = \boldsymbol{G_{AA}}\sigma_{AA}^2$$

$$Cov\left(\boldsymbol{g}_{AD}\right) = \frac{\mathbf{H_{AD} H_{AD}^{'}}}{tr\left(\mathbf{H_{AD} H_{AD}^{'}}\right)/n}\sigma_{AD}^2 = \boldsymbol{G_{AD}}\sigma_{AD}^2$$

$$Cov\left(\boldsymbol{g}_{DD}\right) = \frac{\mathbf{H_{DD} H_{DD}^{'}}}{tr\left(\mathbf{H_{DD} H_{DD}^{'}}\right)/n}\sigma_{DD}^2 = \boldsymbol{G_{DD}}\sigma_{DD}^2$$

and the covariance between any higher order epistatic effects must be:

$$Cov\left(\boldsymbol{g}_{ijk}\right) = \frac{\mathbf{H_{ijk} H_{ijk}^{'}}}{tr\left(\mathbf{H_{ijk} H_{ijk}^{'}}\right)/n}\sigma_{ijk}^2 = \boldsymbol{G_{ijk}}\sigma_{ijk}^2$$

However, **H** matrices are extremely large and calculation of $\mathbf{HH'}$ cross-products is computationally expensive; each **H** matrix has as many columns as marker interactions and as many rows as individuals. Nevertheless, Vitezica et al. (2017) provided an algebraic shortcut that allows calculation from the additive and dominance matrices, described above, as:

$$Cov\left(\boldsymbol{g}_{AA}\right) = \frac{\mathbf{G_A} \circ \mathbf{G_A}}{tr\left(\mathbf{G_A} \circ \mathbf{G_A}\right)/n}\sigma_{AA}^2 = G_{AA}\sigma_{AA}^2$$

$$Cov\left(\boldsymbol{g}_{AD}\right) = \frac{\mathbf{G_A} \circ \mathbf{G_D}}{tr\left(\mathbf{G_A} \circ \mathbf{G_D}\right)/n}\sigma_{AD}^2 = \boldsymbol{G_{AD}}\sigma_{AD}^2$$

$$Cov\left(\boldsymbol{g}_{DD}\right) = \frac{\mathbf{G_D} \circ \mathbf{G_D}}{tr\left(\mathbf{G_D} \circ \mathbf{G_D}\right)/n}\sigma_{DD}^2 = \boldsymbol{G_{DD}}\sigma_{DD}^2$$

For higher order interactions the results are equivalent. As an example, the covariance matrix for the AAD epistatic interaction can be calculated as:

$$Cov\left(\boldsymbol{g}_{AAD}\right) = \frac{\mathbf{G_A} \circ \mathbf{G\ _D} \circ \mathbf{G_D}}{tr\left(\mathbf{G_A} \circ \mathbf{G_D} \circ \mathbf{G_D}\right)/n}\sigma_{ADD}^2 = \boldsymbol{G_{ADD}}\sigma_{ADD}^2$$

It should be noted that $\mathbf{G} \circ \mathbf{G}\ldots$ products tend to **I** and higher order epistatic effects tend to be confused with residuals. Nevertheless, this orthogonal approach assumes linkage equilibrium between SNP molecular markers. Linkage disequilibrium (LD) modifies the distribution of the variance into additive, dominance and epistatic components, and orthogonal partition is not possible (Hill and Maki-Tanila, 2015). In outbred populations, substantial LD is present only between polymorphisms in tight linkage (Hill and Maki-Tanila, 2015). However, whilst the distribution of epistatic effects is still unclear (Wei et al., 2015, there is evidence of epistatic interactions between linked loci (Lynch, 1991). Alternative approaches, such as those of Akdemir and Jannick (2015) and Akdemir et al. (2017) have been developed to define locally epistatic relationship matrices. These studies used a RKHS (Reproducing Kernel Hilbert Space) to define these matrices and average them.

The RKHS approach to model epistatic interactions relies on the idea that the relationship between phenotypes and genotypes may not be linear (Gianola et al., 2006; de los Campos et al., 2009a). The main objective is to predict the performance of each individual given its marker genotype through a function that maps the genotypes into phenotypic responses. One of the simplest methods is to consider that this function is linear and, consequently, the results are equivalent to the GBLUP approach. Nevertheless, the power of the Kernel concept relies on the possibility of using alternative functions of marker genotypes. In short, RKHS procedures result in some non-parametric functions g() of a SNP markers set (**X**):

$$\mathbf{y} = \mu + \mathrm{g}\left(\mathbf{X}\right) + \mathbf{e}$$

and define a cost function to minimize

$$J = \left(\mathbf{y} - g\left(\mathbf{X}\right)\right)' \left(\mathbf{y} - g\left(\mathbf{X}\right)\right) + \lambda \left\| g\left(\mathbf{X}\right)\right\|_H^2$$

where the term $\left\| g\left(\mathbf{X}\right)\right\|_H^2$ is a norm under a Hilbert space. Kimeldorf and Wahba (1971) found that $\mathbf{g}(\mathbf{X})$ can be reformulated as:

$$g\left(\mathbf{X}\right) = \alpha_0 + \sum_{i=1}^{n} \alpha_i \mathbf{K}\left(\mathbf{x}{-}\mathbf{x_i}\right)$$

where $\mathbf{K}$ is a positive semi-definite matrix that meets the requisites of a Kernel Matrix. It defines the similarity between individuals and meets the distance requirements in a Hilbert space (Wootters, 1981). The performance of the method depends on an adequate choice of $\mathbf{K}$ that can be chosen from among a very large number of options. The easiest RKHS option is to use the genealogical ($\mathbf{A}$) or genomic ($\mathbf{G}$) relationship matrices as kernel matrices (Rodríguez-Ramilo et al., 2014), this leads to the standard BLUP or the GBLUP as particular cases of RKHS. However, they only are able to capture the additive genetic variation and if the model tries to accommodate dominance or epistatic interactions, an alternative Kernel matrix has to be implemented for a pair of SNP vectors of two individuals ($\mathbf{x}$ and $\mathbf{x}'$). Most kernels proposed so far (Gianola et al., 2006; Piepho, 2009; Morota et al., 2013; Tusell et al., 2014) consider the similarity across individuals within loci (i.e., similarities within loci are summed). Using Taylor series expansions, it can be shown that kernels of this type are a weighted sum of the additive ($\mathbf{G}$) and dominance covariance matrices ($\mathbf{D}$), and therefore implicitly account for dominance (Piepho, 2009). However, these kernels do not consider joint similarity across loci. A kernel that includes epistasis should measure similarities simultaneously between pairs, triplets etc., of loci across individuals, as described in Jiang and Reif (2015) and Martini et al. (2016).

APPLICATIONS OF GENOMIC SELECTION WITH NON-ADDITIVE GENETIC EFFECTS

Predictive Performance

The most direct application of the genomic prediction models is to predict the performance of an individual for continuous or categorical phenotypes. Here the introduction of non-additive genetic effects in the procedures of prediction becomes relevant, as the main objective is to predict performance conditioned on the genotype of the individual, despite the additive, dominant or epistatic gene action. In fact, simulation studies show up to 17% more accurate predictions based on the sum of additive and dominance effects compared to prediction based on only additive effects (Wellmann and Bennewitz, 2012; Da et al., 2014). However, the performance of semi-parametric or non-parametric approaches such as RKHS methods seems to be appropriate because they are designed to maximize predicting ability over a given individual and not to predict the future performance of the progeny; they are also designed to capture complex and non-explicit interactions. Moreover, some new research fields have merged with genomic evaluation for predicting future performance, examples include: microbiomics (Ramayo-Caldas et al., 2016; Yang et al., 2017), metabolomics (Fontanesi, 2016) and precision farming (Banhazi et al., 2012). Over time they will provide a global picture of the genetic and environmental circumstances that affect the future performance of individuals and they will contribute to the development of more accurate prediction models.

Mate Allocation

In the past, there was a strong belief in "nicking": pairs of individuals that, wisely selected, would give rise to very efficient offspring (Lush, 1943). In terms of quantitative genetics, the existence of "nicking" would imply that there is large variance of dominant deviations (or epistasis) compared to the variance of breeding values, something that finally turned out to be generally false. Even so, there is room for mate allocation within a population (Toro and Varona, 2010). Under models that include dominance effects, the output of the genomic selection procedure can be used to calculate the prediction of performance of future mating (G_{ij}) between the *ith* and *jth* individual as:

$$E\left(G_{ij}\right) = \sum_{k=1}^{n} \Big[pr_{ijk}\left(A_1A_1\right)\hat{a}_J + pr_{ijk}\left(A_1A_2\right)\hat{d}_J \\ -pr_{ijk}\left(A_2A_2\right)\hat{a}_J\Big]$$

where $pr_{ijk}(A_1A_1)$, $pr_{ijk}(A_1A_2)$, and $pr_{ijk}(A_2A_2)$ are the probabilities of the genotypes A_1A_1, A_1A_2, and $\mathrm{A_2A_2}$ for the combination of the *ith* and *jth* individual and the *kth* marker, $\hat{a}_k$ and $\hat{d}_k$ are the estimates of the additive and dominance effects for the same marker and n is the number of markers. Later, optimisation procedures like linear programming (Jansen and Wilton, 1985) or heuristic approximations (simulated annealing, Kirkpatrick et al., 1983) can be used to define a set of mates that maximize performance in the future generation. In a simulated example, Toro and Varona (2010) compared random mating vs. mate selection with a model including dominance and found advantages that ranged between 6 and 22% of the expected response. Sun et al. (2013), Ertl et al. (2014), and Aliloo et al. (2017) have confirmed these improvements with dairy cattle data. However, its implementation in livestock populations is limited because it must be taken into account that the accuracy of the prediction of a potential mate will be low and the advantage will be only relevant when traits have a large amount of non-additive genetic variance. In addition, it requires the genotyping of male and females in the population that is not always available. Moreover, the use of models that include more complex interactions, such as models with epistatic effects or non-parametric approaches, is not so immediate. In fact, the predicted performance of a mate should be calculated after integrating the predictive performance over all possible future genotypic configurations of the expected progeny. For epistasis (but not for dominance) these genotypic configurations also depend on recombination fractions across the genome.

Selection for Crossbreeding

There is consensus that profit from non-additive genetic effects in a selection program can be obtained when commercial animals are the product of mating with those that do not participate in the maintenance of a breeding population. The typical way to proceed is to produce two-way or three-way crosses between populations maintained and selected separately (i.e., in pigs). Selection is carried out within lines to benefit from additivity and, in addition, the value of the cross may increase due to the heterosis. Some of the most popular livestock production systems, including pig, poultry, and rabbit production, involve regular crossbreeding schemes, with the aim of capturing the complementarity between the performance of the purebred populations and heterosis. The breeding goal within pure lines is to select individuals to maximize the response in the crossbred population. The traditional approach for this objective was Reciprocal Recurrent Selection—RRS—(Comstock et al., 1949). RRS postulates the selection of individuals in purebred populations based on the performance of their crossbred progeny. If the source of information is the performance of these crossbred progeny, the main drawback of the practical application of RRS is the increase of generation intervals that reduce overall genetic response. In practical terms, the performance of the pure lines is used, and a high genetic purebred/crossbred correlation is sought in order to warrant correct genetic progress (Wei and van der Werf, 1994), however, this may not be the case because of non-additive effects or genotype x environment (G x E) interactions.

The use of genomic information can provide a very useful tool to improve the ability of prediction of breeding values in purebred populations based on crossbred performance without the need to wait for recording crossbred progeny. Ibánez-Escriche et al. (2009) designed a first approach of the use of GS for crossbred performance under a purely additive model. This study defined a breed specific genomic selection model as:

$$y_i = \mu + \sum_{j=1}^{n} \left(t^S_{ijk}\alpha^S_{jk} + t^D_{ijl}\alpha^D_{jl} \right) + e_i$$

where t^S_{ijk} is the SNP allele at the *jth* locus from breed k and received from the sire of the *ith* individual that can take values 0 or 1, and α^S_{jk} is the breed-specific substitution effect for the *jth* locus and the *kth* breed. Similarly, t^D_{ijl} and α^D_{jl} were defined for the alleles received from the dam of the *lth* breed. The objective of this approach was to estimate allele substitution effects within breed. Even under the assumption of absence of G x E interactions, SNP allele substitution effects may differ between populations due to: (1) Specific population patterns of linkage disequilibrium with the QTL, or (2) The presence of genotypic dominance effects. The allelic substitution effects of the A (or B) population (α_A or α_B) on performance of A x B depends on the biological additive (a) and dominance (d) effects, and the allelic frequencies of B–p_B- (or A–p_A -) as $\alpha_A = a + (1 - 2p_B)\, d$ or $\alpha_B = a + (1 - 2p_A)\, d$). Under dominance, Kinghorn et al. (2010) demonstrated a clear advantage of this approach, assuming the estimation of SNP effects was perfect. This model has been expanded by Sevillano et al. (2017) to a three-way crossbreeding scheme, after the evaluation of a procedure to trace the breed-of-origin of alleles in three-way crossbred animals (Sevillano et al., 2016). This is an example of the "partial genetic" approach (substitution effects defined within populations). Stuber and Cockerham (1966) showed that gene substitution effects can be defined within populations or across populations, and, if all the (non-additive) effects are accounted for, both approaches are equivalent. Christensen et al. (2015) proposed an alternative model called the "common genetic" approach. Both models were compared by Xiang et al. (2016, 2017) in the same data set with very similar results, but more research is still needed.

Crossbreeding implies mating between individuals of parental populations and a formal description of the additive and dominance variance in the crossbred population is required to evaluate the relevance of mate allocation when the crossbreds are generated. Toosi et al. (2010) and Zeng et al. (2013) extended the aforementioned model to include additive and dominance effects and proved (in both cases with simulated data) its superiority over the strictly additive model if dominance variance is present. These results were confirmed by Esfandyari et al. (2015), who proved that the response to selection for crossbreeding performance is increased by training on crossbred genotypes and phenotypes, and by tracking the allele line origin when pure lines are not closely related. Later, Vitezica et al. (2016) described the substitution effects and dominance deviations within the scope of an F1 population and showed that the additive and dominant variance in a crossbred population is:

$$\begin{aligned}
\sigma^2_{A(A)} &= 2p_Aq_A\alpha^2_A = 2\left[p_Aq_Aa^2 + 2p_Aq_A\left(q_B - p_B\right)ad \right. \\
&\quad \left. +\ p_Aq_A\left(q_B - p_B\right)^2 d^2\right] \\
\sigma^2_{A(A)} &=\ 2p_Aq_A\left[a + \left(q_B - p_B\right)d\right]^2 \\
\sigma^2_{A(B)} &= 2p_Bq_B\alpha^2_B = 2\left[p_Bq_Ba^2 + 2p_Bq_B\left(q_A - p_A\right)ad \right. \\
&\quad \left. +\ p_Bq_B\left(q_A - p_A\right)^2 d^2\right] \\
\sigma^2_{A(B)} &= 2p_Bq_B\left[a + \left(q_A - p_A\right)d\right]^2 \\
\sigma^2_D &= 4p_Aq_Ap_Bq_Bd^2
\end{aligned}$$

where $\sigma^2_{A(A)}$ and $\sigma^2_{A(B)}$ are the additive variance generated by the purebred populations A and B, respectively, σ^2_D is the dominance variance, p_A, q_A, p_B and q_B are the allelic frequencies in purebred populations, and *a* and *d* are the additive and dominance effects.

However, all these approaches assume that the additive and dominance effects have the same magnitude in pure and crossbred populations and this implies an absence of G x E interaction. To avoid this restriction, Vitezica et al. (2016) and Xiang et al. (2016) proposed a multivariate genomic BLUP that is capable of considering different additive and dominance effects and their correlations between pure and crossbred populations.

Selection in Purebred Populations

The response to selection in purebred populations depends on the magnitude of the additive variance and on the prediction of the additive breeding values for the candidates for reproduction. It is usually assumed that it is not worth selecting individuals with the highest dominance values because they will go back to zero as a result of random mating. However, Toro (1993, 1998) proposed two mating strategies that can be used to take advantage of dominance in a closed population. The first (Toro, 1993), was a method that basically consists of performing two types of mating: (a) minimum coancestry mating in order to obtain the progenies that will constitute the commercial population and will also be utilized for testing, and (b) maximum coancestry mating from which the breeding population will be maintained. Toro's second strategy (Toro, 1998) advocates the use of the selection of grandparental combinations. Both strategies are analogous with reciprocal-recurrent selection (Comstock et al., 1949) in that they rely on the crucial distinction between commercial and breeding populations. Nevertheless, they have been exclusively tested by simulation and with a reduced set of genes with known additive and dominance effect. Their efficiency has yet to be verified using a large number of SNP markers.

FINAL REMARKS

Despite huge efforts in the development of statistical models for the implementation of genomic selection with non-additive effects, there are still some issues that have to be dealt with before the use of these models in genomic evaluation becomes standard. A major obstacle is the lack of serious testing as this requires extensive data sets with genotypes and phenotypes, and these data sets are rare. In fact, non-additive genetic variance is expected to be low for most traits (Crow, 2010; Hill et al., 2010), with the exception of fitness related traits. Therefore, the inclusion of non-additive effects in genomic selection models will provide very low (or negligible) improvement in the genetic response or the ability of prediction.

Non-additive effects are easily incorporated into GBLUP procedures (Vitezica et al., 2013, 2017) but efforts must be made to define a single-step approach (Aguilar et al., 2010) that is able to use phenotypic data from non-genotyped individuals and the complete genealogical information of breeding schemes. The major limitation of the GBLUP or single-step approaches is the calculation of the inverse of the genomic relationship matrices (**G**), the introduction of non-additive effects will involve the calculation of the inverse of additional matrices related with dominance or epistatic effects. Nevertheless, this is really a constraint in populations with a large number of genotyped individual (i.e., Holstein), while most of the livestock populations do not suffer for any limitations. In fact, the computational cost for inverting additive and non-additive genomic relationship matrices is equivalent. On the other hand, using current pedigree-based BLUP models based on dominance (de Boer and Hoeschele, 1993) seems futile because the models are computationally complicated.

Recent studies (Xiang et al., 2016) have shown that inbreeding depression can be modeled and included in GS approaches through a covariate with the average individual heterozygosity. Nevertheless, this approach only considers the effects of the dominance in inbreeding depression and the role of epistatic interactions in inbreeding depression (Minvielle, 1987) has not yet been studied. However, directional dominance is not necessary requisite for having a substantial dominance variance. In fact it would be interesting to know if there are traits with substantial dominance variance and without inbreeding depression, because they would be good candidates for successful strategies of using dominance. In addition, it should be mentioned that the genetic architecture of non-additive genetic effects and its relationship with inbreeding depression and heterosis is a relevant subject of future research.

The presence of dominance with inbreeding implies the existence of up to five variance components in pedigree-based analysis (Smith and Maki-Tanila, 1990; de Boer and Hoeschele, 1993): additive; dominance between non-inbred; dominance between inbred; covariance between additive; and, inbred dominance values and inbreeding depression. As far as we know, this model has only been used twice with real data in animal breeding (Shaw and Woolliams, 1999; Fernández et al., 2017); their equivalence with the variance components captured by SNP marker effects has to be clarified.

Finally, the parametric approach for the estimation of epistatic effects (Vitezica et al., 2017) fails when linkage disequilibrium is present. A full description of the effect of the genes and their interactions in populations under linkage disequilibrium and the definition of predictive effects has not been reformulated within the scope of genomic selection. It is unclear what we mean by genetic variances when there is linkage disequilibrium, particularly because linkage disequilibrium is population specific and unstable across generations or subpopulations. Nevertheless, Mäki-Tanila and Hill (2014) showed that when the number of loci increases, epistatic variance disappears. At the same time, the proportion of dominance variance stays the same. Thus, dominance variance is the main non-additive component even with linkage disequilibrium (Hill and Maki-Tanila, 2015).

AUTHOR CONTRIBUTIONS

LV prepare the initial draft of the review and it was corrected and improved by AL, MT, and ZV. The final manuscript was read and approved by all the authors.

ACKNOWLEDGMENTS

This work was financed by the INRA SELGEN metaprogram—project EpiSel (ZV, AL), CGL2016-80155 (LV) and CGL2016-75904-C2-2-P (MT) of Ministerio de Economía y Competitividad of Spain.

REFERENCES

Acevedo, C. F., de Resende, M. D. V., Silva, F. F., Viana, J. M. S., Valente, M. S. F., Resende, M. F. R., et al. (2015). Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC Genetics* 16:105. doi: 10.1186/s12863-015-0264-2

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and 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. doi: 10.3168/jds.2009-2730

Akdemir, D., and Jannick, J. (2015). Locally epistatic genomic relationships matrices for genomic association and prediction. *Genetics* 199, 857–871. doi: 10.1534/genetics.114.173658

Akdemir, D., Jannick, J., and Isidro-Sanchez, J. (2017). Locally epistatic models for genome-wide prediction and association by importance sampling. *Genet. Sel. Evol.* 49:74. doi: 10.1186/s12711-017-0348-8

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., and Hayes, B. J. (2016). Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. *Genet. Sel. Evol.* 48:186. doi: 10.1186/s12711-016-0186-0

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., Goddard, M. E., and Hayes, B. J. (2017). Including non-additive genetic effects in mating programs to maximize dairy farm profitability. *J. Dairy Sci.* 100, 1203–1222. doi: 10.3168/jds.2016-11261

Alvarez-Castro, J. M., and Carlborg, O. (2007). A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* 176, 1151–1167. doi: 10.1534/genetics.106.067348

Banhazi, T. M., Lehr, H., Black, J. L., Crabtree, H., Schofield, P., Tscharke, M., et al. (2012). Precision livestock farming: an international review of scientific and commercial aspects. *Int. J. Agric. Biol. Eng.* 5, 1–9. doi: 10.3965/j.ijabe.20120503.001

Barton, N. H., Etheridge, A. M., and Véber, A. (2017). The infinitesimal model: definition, derivation and implications. *Theor. Pop. Biol.* 118, 50–73. doi: 10.1016/j.tpb.2017.06.001

Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge, UK: Cambridge University Press Warehouse. doi: 10.5962/bhl.title.44575

Bennewitz, J., and Meuwissen, T. H. E. (2010). The distribution of QTL additive and dominance effects in porcine F2 crosses. *J. Anim. Breed. Genet.* 127, 171–179. doi: 10.1111/j.1439-0388.2009.00847.x

Bennewitz, J., Edel, C., Fries, R., Meuwissen, T. H. E., and Wellmann, R. (2017). Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. *Genet. Sel. Evol.* 49:7. doi: 10.1186/s12711-017-0284-7

Bernardo, R. (2010). *Breeding for Quantitative Traits in Plants, 2nd Edn.* Woodsbury, MN: Stemma Press.

Caballero, A., and Keightley, P. D. (1994). A pleiotropic nonadditive model of variation in quantitative traits. *Genetics* 138, 883–900.

Christensen, O. F., Legarra, A., Lund, M. S., and Su, G. (2015). Genetic evaluation for three-way crossbreeding. *Genet. Sel. Evol*, 47:177. doi: 10.1186/s12711-015-0177-6

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39, 859–882.

Comstock, R. E., Robinson, H. F., and Harvey, P. H. (1949). A breeding procedure designed to make maximum use of both general and specific combining ability. *Agron. J.* 41, 360–367. doi: 10.2134/agronj1949.00021962004100080006x

Crow, J. F. (2010). On epistasis: why it is unimportant in polygenic directional selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 1241–1244. doi: 10.1098/rstb.2009.0275

Da, Y., Wang, C., Wang, S., and Hu, G. (2014). Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS ONE* 9:e87666. doi: 10.1371/journal.pone.0087666

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185, 1021–1031. doi: 10.1534/genetics.110.116855

de Boer, I., and Hoeschele, I. (1993). Genetic evaluation methods for populations with dominance and inbreeding. *Theor. Appl. Genet.* 86, 245–258. doi: 10.1007/BF00222086

de los Campos, G., Gianola, D., and Rosa, G. J. (2009a). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J. Anim. Sci.* 87, 1883–1887. doi: 10.2527/jas.2008-1259

De los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., et al. (2009b). Prediction quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385. doi: 10.1534/genetics.109.101501

Dekkers, J. C. (2004). Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J. Anim. Sci.* 82, E313–E328. doi: 10.2527/2004.8213_supplE313x

Duenk, P., Calus, M. P. L., Wientjes, Y. C. J., and Bijma, P. (2017). Benefits of dominance over additive models for the estimation of average effects in the presence of dominance. *G3 Genes Genomes Genetics* 7, 3405–3414. doi: 10.1534/g3.117.300113

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswani, S., Bowman, P. J., Reich, C. M., et al. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95, 4114–4129. doi: 10.3168/jds.2011-5019

Ertl, J., Legarra, A., Vitezica, Z. G., Varona, L., Edel, C., Emmerling, R., et al. (2014). Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. *Genet. Sel. Evol.* 46:40. doi: 10.1186/1297-9686-46-40

Esfandyari, H., Bijma, P., Henryon, M., Christensen, O. F., and Sorensen, A. C. (2016). Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. *Genet. Sel. Evol.* 48:40. doi: 10.1186/s12711-016-0220-2

Esfandyari, H., Sorensen, A. C., and Bijma, P. (2015). A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genet. Sel. Evol.* 47:76. doi: 10.1186/s12711-015-0155-z

Falconer, D. S., and McKay, T. (1996). *Introduction to Quantitative Genetics*. Harlow: Pearson Education Limited.

Fernández, E. N., Legarra, A., Martínez, R., Sánchez, J. P., and Baselga, M. (2017). Pedigree-based estimation of covariance between dominance deviations and additive genetic effects in closed rabbit lines considering inbreeding and using a computationally simpler equivalent model. *J. Anim. Breed. Genet.* 134, 184–195. doi: 10.1111/jbg.12267

Fernando, R. L., Dekkers, J. C., and Garrick, D. J. (2014). A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analysis. *Genet. Sel. Evol.* 46:50. doi: 10.1186/1297-9686-46-50

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian Inheritance. *Trans. R. Soc. Edinburgh* 52, 399–433. doi: 10.1017/S0080456800012163

Fontanesi, L. (2016). Metabolomics and livestock genomics: insights into a phenotyping frontier and its application in animal breeding. *Anim. Front.* 6, 73–79. doi: 10.2527/af.2016-0011

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776. doi: 10.1534/genetics.105.049510

Gonzalez-Recio, O., Rosa, G. J. M., and Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide predictin of complex traits. *Livest. Sci.* 166, 217–231. doi: 10.1016j.livsci.2014.05.036

Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G., and Chee, M. S. (2005). A genome-wide scalable SNP genotyping asay using microarray technology. *Nat. Genet.* 37, 549–554. doi: 10.1038/ng1547

Guo, X., Christensen, O. F., Ostersen, T., Wang, Y., Lund, M. S., and Su, G. (2016). Genomic prediction using models with dominance and imprinting effects for backfat thickness and average daily gain in Danish Duroc pigs. *Genet. Sel. Evol.* 48:67. doi: 10.1186/s12711-016-0245-6

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186. doi: 10.1186/1471-2105-12-186

Hazel, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics* 28, 476–490.

Heidaritabar, M., Wolc, A., Arango, J., Zeng, J., Settar, P., Fulton, J. E., et al. (2016). Impact of fitting dominance and additive effects on accuracy of genomic prediction of breeding values in layers. *J. Anim. Breed. Genet.* 133, 334–346. doi: 10.1111/jbg.12225

Henderson, C. R. (1973). "Sire evaluation and genetic trends," in *Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. Jay L. Lush 10-41* (Champaing, IL: ASAS and ADSA).

Hill, W. G. (2010). Understanding and using quantitative genetic variation. *Philos. Trans. R. Soc. Lond. B Sci.* 365, 73–85. doi: 10.1098/rstb.2009.0203

Hill, W. G., and Maki-Tanila, A. (2015). Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. *J. Anim. Breed. Genet*. 132, 176–186. doi: 10.1111/jbg.12140

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*. 4:e1000008. doi: 10.1371/journal.pgen.1000008

Howard, J. T., Haile-Mariam, M., Pryce, J. E., and Maltecca, C. (2015). Investigation of regions impacting inbreeding depression and their association with the additive genetic effect for United States and Australia Jersey dairy cattle. *BMC Genomics* 16:813. doi: 10.1186/s12864-015-2001-7

Hu, Y., Rosa, G. J. M., and Gianola, D. (2016). Incorporating parent-of-origin effects in whole-genome prediction of complex traits. *Genet. Sel. Evol*. 48:34. doi: 10.1186/s12711-016-0213-1

Huang, W., and Mackay, T. F. C. (2016). The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS Genet*. 10:e1006421. doi: 10.1371/journal.pgen.1006421

Ibánez-Escriche, N., Fernando, R. L., Toosi, A., and Dekkers, J. C. (2009). Genomic selection of purebreds for crossbred performance. *Genet. Sel. Evol*. 41:12. doi: 10.1186/1297-9686-41-12

Jansen, G. B., and Wilton, J. W. (1985). Selecting mating pairs with linear programming techniques. *J. Dairy Sci*. 68, 1302–1305. doi: 10.3168/jds.S0022-0302(85)80961-9

Jiang, J., Shen, B., O' Connell, J. R., VanRaden, P. M., Cole, J. B., and Ma, L. (2017). Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. *BMC Genomics* 18:425. doi: 10.1186/s12864-017-3821-4

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768. doi: 10.1534/genetics.115.177907

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proc. R. Soc. Lond. B Biol. Sci*. 143, 102–113. doi: 10.1098/rspb.1954.0056

Kimeldorf, G., and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl*. 33, 82–95. doi: 10.1016/0022-247X(71)90184-3

Kinghorn, B. P., Hickey, J. M., and van der Werf, J. H. J. (2010). "Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals," in *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production* (Leipzig), 36.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680. doi: 10.1126/science.220.4598.671

Lush, J. L. (1943). *Animal Breeding Plans, 2nd Edn*. Ames, IA: The Collegiate Press Inc., Iowa.

Lynch, M. (1991). The genetic interpretation of inbreeding depression and outbreeding depression. *Evolution* 45, 622–629. doi: 10.1111/j.1558-5646.1991.tb04333.x

Maki-Tanila, A. (2007). An overview on quantitative and genomic tools for utilising dominance genetic variation in improving animal production. *Agric. Food Sci*. 16, 188–198. doi: 10.2137/145960607782219337

Mäki-Tanila, A., and Hill, W. G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics* 198, 355–367. doi: 10.1534/genetics.114.165282

Martini, J. W., Wimmer, V., Erbe, M., and Simianer, H. (2016). Epistasis and covariance: how gene interaction translates into genomic relationship. *Theor. Appl. Genet*. 129, 963–976. doi: 10.1007/s00122-016-2675-5

Mendel, G. (1866). Versuche über Pflanzen-Hybriden. – Verhandlungen des Naturforschenden Vereines, Abhandlungern, Brünn, 4, 3–47. Editions in different languages published by Matlová (1973). doi: 10.5962/bhl.title.61004

Meuwissen, T. H., Hayes, B., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Minvielle, F. (1987). Dominance is not necessary for heterosis: a two-locus model. *Genet. Res*. 49, 245–247. doi: 10.1017/S0016672300027142

Misztal, I., Varona, L., Culbertson, M., Gengler, N., Bertrand, J. K., Mabry, J., et al. (1998). Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. *Biotechnol. Agron. Soc. Environ*. 2, 227–233.

Moghaddar, N., and van der Werf, J. H. J. (2017). Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. *J. Anim. Breed. Genet*. 134, 453–462. doi: 10.1111/jbg.12287

Morota, G., and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Front. Genet*. 5:363. doi: 10.3389/fgene.2014.00363

Morota, G., Koyama, M., Rosa, G. J. M., Weigel, K. A., and Gianola, D. (2013). Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. *Genet. Sel. Evol*. 45:17. doi: 10.1186/1297-9686-45-17

Nguyen, T. N., and Nagyné-Kiszlinger, H. (2016). Dominance effects in domestic populations. *Acta Agraria Kaposvariensis* 20, 1–20.

Nishio, M., and Satoh, M. (2015). Genomic best linear unbiased prediction method including imprinting effects for genomic evaluation. *Genet. Sel. Evol*. 47:32. doi: 10.1186/s12711-015-0091-y

Piepho, H. P. (2009). Ridge regression and extensions for genomewide selection in maize. *Crop Sci*. 49, 1165–1176. doi: 10.2135/cropsci2008.10.0595

Pryce, J. E., Haile-Mariam, M., Goddard, M. E., and Hayes, B. J. (2014). Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. *Genet. Sel. Evol*. 46:71. doi: 10.1186/s12711-014-0071-7

Ramayo-Caldas, Y., Mach, N., Lepage, P., Levenez, F., Denis, C., Lemonnier, G., et al. (2016). Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits. *ISME J*. 10, 2973–2977. doi: 10.1038/ismej.2016.77

Reik, W., and Walter, J. (2001). Genomic imprinting, parental influence on the genome. *Nat. Rev. Genet*. 2, 21–32. doi: 10.1038/35047554

Rodríguez-Ramilo, S. T., García-Cortés, L. A., and González-Recio, O. (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. doi: 10.1371/journal.pone.0093424

Saura, M., Fernández, A., Varona, L., Fernández, A. I., De Cara, M. A. R., Barragán, C., et al. (2015). Detecting inbreeding depression for reproductive traits in Iberian pigs using genome wide data. *Genet. Sel. Evol*. 47:12. doi: 10.1186/s12711-014-0081-5

Sevillano, C. A., Vandenplas, J., Bastiaansen, J. W. M., and Calus, M. P. L. (2016). Empirical determination of breed-of-origin of alleles in three-way crossbred pigs. *Genet. Sel. Evol*. 48:55. doi: 10.1186/s12711-016-0234-9

Sevillano, C. A., Vandenplas, J., Bastiaansen, J. W. M., Bergsma, R., and Calus, M. P. L. (2017). Genomic evaluation for a three-way crossbreeding system considering breed-of-origin of alleles. *Genet. Sel. Evol*. 49:75. doi: 10.1186/s12711-017-0350-1

Shaw, F. H., and Woolliams, J. A. (1999). Variance component analysis of skin and weight data for sheep subjected to rapid inbreeding. *Genet. Sel. Evol*. 31, 43–59. doi: 10.1186/1297-9686-31-1-43

Smith, S. P., and Maki-Tanila, A. (1990). Genotypic covariance matrices and their inverses for models allowing dominance and inbreeding. *Genet. Sel. Evol*. 22, 65–91. doi: 10.1186/1297-9686-22-1-65

Spencer, H. G. (2002). The correlation between relatives on the supposition of genomic imprinting. *Genetics* 161, 411–417.

Stuber, C. W., and Cockerham, C. C. (1966). Gene effects and variances in hybrid populations. *Genetics* 64, 1279–1286

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., and Lund, M. S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS ONE* 7:e45293. doi: 10.1371/journal.pone.0045293

Sun, C., VanRaden, P. M., O'Connell, J. R., Weigel, K. A., and Gianola, D. (2013). Mating programs including genomic relationships and dominance effects. *J. Dairy Sci*. 96, 8014–8023. doi: 10.3168/jds.2013-6969

Toosi, A., Fernando, R. L., and Dekkers, J. C. (2010). Genomic selection in admixed and crossbred populations. *J. Anim. Sci*. 88, 32–46. doi: 10.2527/jas.2009-1975

Toro, M. A. (1993). A new method aimed at using the dominance variance in closed breeding populations. *Genet. Sel. Evol*. 25, 63–74. doi: 10.1186/1297-9686-25-1-63

Toro, M. A. (1998). Selection of grandparental combinations as a procedure designed to make use of dominance genetic effects. *Genet. Sel. Evol*. 30, 339–349. doi: 10.1186/1297-9686-30-4-339

Toro, M. A., and Varona, L. (2010). A note on mate allocation for dominance handling in genomic selection. *Genet. Sel. Evol*. 42:33. doi: 10.1186/1297-9686-42-33

Tusell, L., Pérez-Rodríguez, P., Forni, S., and Gianola, D. (2014). Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. *J. Anim. Breed Genet*. 131, 105–115. doi: 10.1111/jbg.12070

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci*. 91, 4414–4123. doi: 10.3168/jds.2007-0980

Varona, L., Vitezica, Z. G., Munilla, S., and Legarra, A.- (2014). "A general approach for calculation of genomic relationship matrices for Epistatic effects," in *Proceedings from the 10th World Congress on Genetics Applied to Livestock Production* (Vancouver, BC), 11–22.

Verbyla, K. L., Hayes, B. J., Bowman, P. J., and Goddard, M. E. (2009). Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genet. Res*. 91, 307–311. doi: 10.1017/S0016672309990243

Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal estimates of variances for additive, dominance and epistatic effects in populations. *Genetics* 206, 1297–1307. doi: 10.1534/genetics.116.199406

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195, 1223–1230. doi: 10.1534/genetics.113.155176

Vitezica, Z. G., Varona, L., Elsen, J. M., Misztal, I., Herring, W., and Legarra, A. (2016). Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. *Genet. Sel. Evol*. 48:6. doi: 10.1186/s12711-016-0185-1

Wang, X., Yang, Z., and Xu, C. (2015). A comparison of genomic selection methods for breeding value prediction. *Sci. Bull*. 60, 925–935. doi: 10.1007/s11434-015-0791-2

Wei, M., and van der Werf, J. H. J. (1994). Maximizing genetic response in crossbreds using both purebred and crossbred information. *Anim. Prod*. 59, 401–413. doi: 10.1017/S0003356100007923

Wei, W. H., Hemani, G., and Haley, C. S. (2015). Detecting epistasis in human complex traits. *Nat. Rev. Genet*. 15, 722–733. doi: 10.1038/nrg3747

Wellmann, R., and Bennewitz, J. (2011). The contribution of dominance to the understanding of quantitative genetic variation. *Genet. Res*. 92, 139–154. doi: 10.1017/S0016672310000649

Wellmann, R., and Bennewitz, J. (2012). Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genet. Res*. 94, 21–37. doi: 10.1017/S0016672312000018

Wootters, W. K. (1981). Statistical distance and Hilbert space. *Phys. Rev. D* 23, 357–363. doi: 10.1103/PhysRevD.23.357

Wright, S. (1921). Systems of mating. I. The biometric relations between parent and offspring. *Genetics* 6, 111–123.

Xiang, T., Christensen, O. F., and Legarra, A. (2017). Technical note: genomic evaluation for crossbred performance in a single-step approach with metafounders. *J. Anim. Sci*. 95, 1472–1480. doi: 10.2527/jas2016.1155

Xiang, T., Christensen, O. F., Vitezica, Z. G., and Legarra, A. (2016). Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. *Genet. Sel. Evol*. 48:92. doi: 10.1186/s12711-016-0271-4

Yang, H., Huang, X., Fang, S., He, M., Zhao, Y., Wu, Z., et al. (2017). Unraveling the fecal microbiota and metagenomic functional capacity associated with feed efficiency in pigs. *Front. Microbiol.* 8:1555. doi: 10.3389/fmicb.2017.01555

Zeng, J., Toosi, A., Fernando, R. L., Dekkers, J. C. M., and Garrick, D. J. (2013). Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genet. Sel. Evol*. 45:11. doi: 10.1186/1297-9686-45-11

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.