1 Exact p-values for large-scale single step genome-wide association, with an application for
2 birth weight in American Angus

3

4 Ignacio Aguilar¹, Andres Legarra², Fernando Cardoso^{3,4}, Yutaka Masuda⁵, Daniela Lourenco⁵,
5 Ignacy Misztal⁵

6

7 ¹Instituto Nacional de Investigacion Agropecuaria, Canelones, Uruguay 90200

8 ²UMR GenPhySE, INRA Toulouse, BP52626, 31326 Castanet Tolosan, France

9 ³Department of Animal Science, Federal University of Pelotas, Rio Grande do Sul, Brazil

10 ⁴Embrapa Pecuária Sul, 96400-031 Bagé/RS, Brazil

11

12 ⁵Department of Animal and Dairy Science, University of Georgia, Athens, GA, USA

13

14

15 Corresponding Author: Andres Legarra (andres.legarra@inra.fr)

16

17

18 Email addresses:

19 IA: iaguilar@inia.org.uy

20 AL: andres.legarra@inra.fr

21 FC: fernando.cardoso@embrapa.br

22 YM: Yutaka@uga.edu

23 DL: danilino@uga.edu

24 IM: ignacy@uga.edu

25 ABSTRACT

26 BACKGROUND: Single Step GBLUP (SSGBLUP) is the most comprehensive method for
27 genomic prediction. Point estimates of marker effects from SSGBLUP are often used for
28 Genome Wide Association Studies (GWAS) without a formal framework of hypothesis
29 testing. Our objective was to implement p-values for GWAS studies in the ssGBLUP
30 framework, showing algorithms, computational procedures, and an application to a large
31 beef cattle population.

32 METHODS: P-values were obtained based on the prediction error (co)variance for SNP,
33 which uses the inverse of the coefficient matrix and formulas to compute SNP effects.

34 RESULTS: Computation of p-values took a negligible time for a dataset with almost 2 million
35 animals in the pedigree and 1424 genotyped sires, and no inflation was observed. The SNP
36 passing the Bonferroni threshold of 5.9 in the $-\log_{10}$ scale were the same as those that
37 explained the highest proportion of additive genetic variance, but the latter was penalized
38 (as GWAS signal) by low allele frequency.

39 CONCLUSION: The exact p-value for SSGWAS is a very general and efficient strategy for QTL
40 detection and testing. It can be used in complex data sets such as used in animal breeding,
41 where only a proportion of pedigreed animals are genotyped.

42

43 INTRODUCTION

44

45 Detection and mapping of causal genes and QTLs in livestock genetics is usually done by
46 Genome-Wide Association Studies (GWAS). The most frequent GWAS method is the single-

marker fixed regression, where one marker is fit at a time in the model, and correction for polygenic genetic effects is made using a relationship matrix. This is also known as EMMAX [1–3]. However, use of EMMAX implies that genotyped animals have phenotype and vice versa. In livestock, this requirement is impossible for animals such as dairy bulls that do not have female-related phenotypes (e.g. milk yield). In general, many animals that have own phenotype (e.g. growth) would benefit from phenotypic information from relatives (e.g. growth in daughters, ancestors and collaterals). Typically, phenotypes of related animals are “projected” into genotyped animals [4–6] in a process called de-regression, which has been successfully used to detect and map QTLs [7,8].

De-regression is a cumbersome process, usually not optimal because it is an approximation that lost information and can lead to inaccuracies due to unaccounted for selection, or common effects that are just ignored. In particular using de-regression leads to double counting when the genotyped population includes both sires and their progeny. Legarra and Vitezica [9] showed that an extension of the single-marker EMMAX regression is a two-trait variance component model, where the two traits are actual phenotype and “gene content” at the marker. However, the method would be very slow for a whole genome scan because it requires maximization of a REML likelihood at each marker.

The GBLUP or SNP-BLUP framework [10–12] allows for the joint estimation of marker effects and the automatic correction of genetic structure in the population. The so-called single step methods (i.e., Single Step GBLUP (SSGBLUP) and Single Step SNP-BLUP (SSSNP-BLUP)) project phenotypes into phenotyped individuals, using pedigree relationships [13–16]. These “single step methods” allow the estimation of breeding values and also marker effects [15,17,18],

71 and the latter have been used for GWAS analysis [19–21], typically by size of estimated
 72 marker effects or a similar measure such as the proportion of genetic variance explained by
 73 a marker or segment. A common pitfall of GWAS methods based on size of effects or
 74 variances explained is the lack of a formal framework for acceptance/rejection of hypothesis
 75 – in particular, there is no clearly defined statistic to be used for hypothesis testing, and
 76 there is no empirical statistical law or closed-form solution under the null hypothesis. The
 77 most common form of declaring significance is to use a threshold on explained genetic
 78 variance (but: how much?) by a set of contiguous markers (but: how many?) [22]. For
 79 instance, [19] studied “the 20 largest explanatory loci”, [20] studied “the 10 windows
 80 explaining the largest amount of genomic variance for gene annotation, gene network and
 81 pathway analyses” whereas [21] considered “1.5 Mb SNP windows that explained more than
 82 0.50 % of the genetic variance”. Thus, there is no formal hypothesis testing. Although using
 83 estimates of markers effects considers correctly the *magnitude* of the estimated marker
 84 effects, it does not always consider correctly the *uncertainty* in the estimation of marker
 85 effects. Even worse, the use of different iterative schemes or the definition of “windows” of
 86 adjacent SNPs can be very arbitrary. This renders interpretation of signals more difficult..

87

88 Recently, equivalences between GBLUP and single marker GWAS results have been shown in
 89 a series of papers [23–25]. Namely, the statistic of single marker GWAS is equivalent to the
 90 statistic obtained from the marker estimate from GBLUP. In both cases the statistic used is
 91 $\hat{a}_i / sd(\hat{a}_i)$ (i.e., marker effect estimate over its standard deviation). Although in EMMAX the
 92 marker is a fixed effect and in GBLUP it is random, a series of papers [23–25] proved that the
 93 statistic is the same in both methods. An application and comparison with Bayesian methods

was shown in a data set of gait in horses [26]. Still, the method could not be applied to data sets consisting of mixtures of genotyped and ungenotyped animals.

Fortunately, Lu et al. [27] showed that the same logic used in [23–25] can be used in a SSGBLUP context. Simply stated, to obtain a statistical test for the effect of a single marker, we only need estimates of the breeding values of genotyped animals and their sampling distribution, and this can be readily obtained from SSGBLUP. Unfortunately, the article of Lu et al. [27] seems largely unnoticed because it focuses on feed efficiency and uses a small data set (i.e., 7,000 phenotyped animals from which 5,000 are genotyped), which does not show computational tools, challenges and bottlenecks.

In the present work, we show the implementation of Single Step GWAS (SSGWAS) with exact p-values as in [27], together with an application in a very large beef cattle data set. We describe algorithms and computational procedures along with their bottlenecks. As result of the data analysis, we find highly significant signals in loci already described by the literature, and good empirical behavior of the method.

Methods

Theory

The classical GWAS method (EMMAX) for marker i uses a linear model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{z}_i a_i + \mathbf{u} + \mathbf{e}$ where vector $\mathbf{b}$ contains fixed effects, $\mathbf{z}_i$ is a vector with “gene contents” (0, 1, or 2 at the

marker), α_i is the substitution effect of the i^{th} marker, and $\mathbf{u}$ is a vector of genetic values, modelled as $Var(\mathbf{u}) = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i q_i} \sigma_u^2 = \mathbf{G}\sigma_u^2$, where $\mathbf{G}$ is a genomic relationship matrix and $\mathbf{Z}$ is a matrix with gene contents for all markers; p_i and q_i are the frequencies of the major and minor alleles of the i^{th} marker, respectively. We assume that variances σ_u^2 and σ_e^2 are known with high certainty, which is usually the case in large data sets. The normal test for the effect of the marker uses the statistic $\frac{\hat{\alpha}_i}{sd(\hat{\alpha}_i)}$, where both values are obtained from inversion of the linear equations of the model. P-values are obtained as $pval_i = 2 \left(1 - \Phi \left(\left| \frac{\hat{\alpha}_i}{sd(\hat{\alpha}_i)} \right| \right) \right)$ where Φ is the cumulative standard normal function.

It has been shown [23–25] that joint estimates of marker effects α from SNP-BLUP (or GBLUP models) of the form $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e}$ with prior assumption $Var(\alpha) = \mathbf{I} \frac{\sigma_u^2}{2\sum p_i q_i}$ lead themselves to statistics $\frac{\hat{\alpha}_i}{sd(\hat{\alpha}_i)}$ that are equivalent to the statistic obtained for the same marker in the “fixed regression” framework described above. The application of SSSNP-BLUP instead of SNP-BLUP is straightforward, because $\hat{\alpha}_i$ and $sd(\hat{\alpha}_i)$ are immediately available (e.g. [15]). In GBLUP and SSGBLUP, the same values can be obtained as linear transformations of the solutions for genetic values $\hat{\mathbf{u}}$ and their Prediction Error (Co)Variances [23–26]. These can be obtained from the inverse of the Mixed Model Equations [28,29].

Algorithm

The algorithm for SSGWAS, which includes genotyped and non-genotyped animals, has been implemented in the blupf90 suite of programs [30] as follows. It combines the algorithms for SSGBLUP (e.g.[16]) and back-solving to obtain marker estimates and their associated p-values from estimates of breeding values [25]:

1. Construct the joint pedigree-genomic relationship matrix $\mathbf{H}^{-1} = \mathbf{A}^{-1} +$

$$\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}, \text{ that is the inverse of}$$

$$\mathbf{H} = \begin{pmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix}, \text{ which projects genomic}$$

relationships $\mathbf{G} = \mathbf{Z}\mathbf{Z}'/2\sum p_i q_i$ from genotyped animals (labelled as “2”) to non-

genotyped animals (labelled as “1”). Matrix $\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$ is the pedigree-based

relationship matrix. Matrix $\mathbf{G}$ is (usually) constructed as $\mathbf{G} = (1 - \alpha) \left(a +$

$$b \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i q_i} \right) + \alpha \mathbf{A}_{22}$$

with a and b obtained to equate average inbreeding and average

relationships in $\mathbf{G}$ and $\mathbf{A}_{22}$, and α a small value (typically from 0 to 0.05). In this way,

both relationships are “compatible” [31,32] and $\mathbf{G}$ is invertible [10]. Matrix $\mathbf{Z}$

contains centered gene content as in [10] but using observed allele frequencies.

2. Construct the mixed model equations for SSGBLUP. In a simple case (a model with a single genetic effect) this would be:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X}\sigma_e^{-2} & \mathbf{X}'\mathbf{W}\sigma_e^{-2} \\ \mathbf{W}'\mathbf{X}\sigma_e^{-2} & \mathbf{W}'\mathbf{W}\sigma_e^{-2} + \mathbf{H}\sigma_u^{-2} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y}\sigma_e^{-2} \\ \mathbf{W}'\mathbf{y}\sigma_e^{-2} \end{pmatrix}$$

Where $\boldsymbol{\beta}$ are “fixed” effects and $\mathbf{u}$ are individual animal (not marker) effects.

3. Factorize and obtain the sparse inverse of the coefficient matrix. The whole inverse cannot be obtained directly as it is typically too big [33]. Most exactly, a “sparse inverse” ($\mathbf{C}$) is obtained with selected elements of the inverse corresponding to the

non-zero entries of a (sparse) Cholesky factor ($\mathbf{LL}'$). In our case, this is achieved using supernodal sparse factorization and inversions as programmed in YAMS [34,35]. Factorization is the computing bottleneck of the procedure and is roughly cubic on the number of genotyped animals; YAMS reduces the computing time by, roughly, one order of magnitude.

4. Solve the mixed model equations for $\begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix}$ using the sparse Cholesky factor.
5. Extract from $\mathbf{C}$ the submatrix corresponding to the genotyped animals, $\mathbf{C}^{u_2 u_2}$. This matrix contains the prediction error (co)variances of their estimated breeding values, $\hat{u}_2$, i.e. $Var(\mathbf{u} - \hat{u}_2) = \mathbf{C}^{u_2 u_2}$.
6. Backsolve for SNP effect estimates:

$$\hat{a}|\hat{u} = (1 - \alpha)b\mathbf{Z}' \frac{1}{2\Sigma p_i q_i} \mathbf{G}^{-1} \hat{u}_2$$

- If matrix $\mathbf{G}$ is full rank and compatible with pedigree relationships (for instance, if $\mathbf{Z}$ is built with base allele frequencies) then $\hat{a}|\hat{u} = \mathbf{Z}' \frac{1}{2\Sigma p_i q_i} \mathbf{G}^{-1} \hat{u}_2$ [10,36,37].
7. Obtain individual prediction error variances of SNP effects as [25]

$$Var(\hat{a}_i) = \frac{1}{2\Sigma p_i q_i} (1 - \alpha)b\mathbf{z}_i' \mathbf{G}^{-1} (\mathbf{G}\sigma_u^2 - \mathbf{C}^{u_2 u_2}) \mathbf{G}^{-1} \mathbf{z}_i (1 - \alpha)b \frac{1}{2\Sigma p_i q_i}$$

where $\mathbf{z}_i$ is the i -th column of matrix $\mathbf{Z}$, corresponding to genotypes of marker i across individuals. The values of 0.95 in steps 7 and 8 refer to the “blending” of $\mathbf{G}$ with $\mathbf{A}$ in step 1 and it will change with different choices of blending parameters.

8. Finally, the p-value for marker i is obtained as $pval_i = 2 \left(1 - \Phi \left(\left| \frac{\hat{a}_i}{sd(\hat{a}_i)} \right| \right) \right)$.

Note that this analysis has to be run only once, as opposed to the n individual runs for n markers in the classical “fixed effect” regression or EMMAX (e.g. [38]).

MATERIAL

We re-analyzed a data set from the American Angus Association [39] including birth weight. As the complete data set is very large, with about 7 million individual weights, 52000 genotyped animals and 8 million animals in pedigree, we used only partial data. These included phenotypes recorded in the last four years, specifically from 2009 to 2012, which comprised 1,046,623 birth weights. Three generations of ancestors were traced back, totaling 1,849,865 individuals. All available genotypes of sires with phenotyped offspring were used (i.e., 1,424 genotyped sires). Genotypes were obtained with the BovineSNP50k v2 BeadChip; 38,122 polymorphic markers were used after quality control [39].

The linear model for birth weight included the effects of contemporary group, animal additive genetic effect, and a permanent environmental effect of the mother, which considers the maternal ability during pregnancy. This differs from the model actually used in national evaluations, that also considers a maternal genetic effect. Variance components were fixed at values from the national evaluation, with heritability of 0.48 and a maternal component of 0.10.

Computations were done using the blupf90 software suite. Then, GWAS results were plotted with qqman [40]. Rejection thresholds used a Bonferroni correction for multiple testing of 0.05/38122, which equals to 5.9 in the $-\log_{10}$ scale. Significant regions were explored in AnimalQTLdb and Jbrowse [41] using the bovine map assembly UMD 3.1 [42]. Although a

new genome map assembly has already been published (i.e., ARS-UCD 1.2), the
aforementioned genome browsers are still using the UMD 3.1.

In addition to exact SSGWAS using p-values, we also ran “approximated” SSGWAS using
estimates of marker effects, in the form of percentage of variance explained by marker
effect [18]. This estimates the population genetic variance explained by the marker effect,
and is approximately computed as $2p_i q_i \hat{a}_i^2$. There are no theoretical thresholds in this
approach, and we used an arbitrary threshold of 0.10% of total genetic variance explained
by 1 marker. Opposite to [18], we do not present results of iterating the SSGBLUP using
“weights” for each marker, as this procedure did not result in increased marker effect and
variance (results not shown).

RESULTS

Factorization of the Mixed Model Equations and extraction of $C^{u_2 u_2}$ required 30Gb of RAM
memory and the computing time was 14h (wall-clock time). Computation of $Var(\hat{a}_i)$ and p-
values took a negligible time of a few minutes. Quantile-quantile plot (QQplot) and
Manhattan plots are in Figures 1 and 2. The QQplot does not show large deviations from the
null hypothesis, which means that SSGBLUP captures correctly the structure of the
population *via* relationship matrices. In presence of structure unaccounted for, inflation of p-
values is expected [3,43,44].

The GWAS results pointed two regions significant at genome-wide level, with values of $-\log_{10}(pvalue)$ near 8. The first region at the end of chromosome 7 includes 3 markers: ARS-BFGL-NGS-107035, ARS-BFGL-NGS-101886 and ARS-BFGL-NGS-18900, which are all in very high linkage disequilibrium (correlation > 0.95 in all cases). Very close signals in the same region can be found in *animalQTLdb* for the trait “body conformation trait” in Holstein cattle [8] and “average daily body weight gain” found in Brangus (composite of Brahman and Angus) [45]. The fact that the signal appears in several breeds would imply that this is not a false positive. The second region at the end of chromosome 20 only includes marker Hapmap42635-BTA-68718. This marker is included in a QTL region detected by [46] in Hereford.

Figure 3 shows the percentage of genetic variance attributed to each marker. The top markers are the same as in Figure 2 (ARS-BFGL-NGS-18900 and Hapmap42635-BTA-68718), but the region in chromosome 7 has a shorter peak than the region in chromosome 20. This is at first sight perplexing, because estimates of marker effects and standard errors are similar for peaks at chromosomes 7 and 20 (effects $\pm$ standard deviations of -0.041 ± 0.007 and 0.043 ± 0.008 , respectively). The different height of the peaks in chromosome 7 and chromosome 20 in the variance plot (Figure 3) and in the $-\log_{10}(pvalue)$ plot (Figure 2) is entirely due to the different minor allelic frequencies: 0.28 and 0.46 respectively.

FIGURE 1 – Quantile-quantile plot (QQPLOT) for the birth weight trait. The grey region represents a 95% confidence interval.

The grey region

FIGURE 2 – Manhattan plot (p-values of individual SNP effects) for birth weight. The red line corresponds to Bonferroni rejection threshold for nominal $\alpha=0.05$.

FIGURE 3 – Percentage of genetic variance explained by markers for birth weight in American Angus. The red line corresponds to an arbitrary rejection threshold of 0.10%.

DISCUSSION

Estimates of SNP effects from single-step methods became available in 2012 [18,47]; however, a measure of significance for SNP estimates has not been available until now in current implementations of single-step GBLUP. Expanding the ideas from the equivalence between GBLUP and GWAS [23–25], Lu et al. [27] derived exact GWAS for a single-step GBLUP framework in a relatively small data set. In our study we showed exact p-values calculated for the first time using a large data set including genotypes, phenotypes, and pedigree.

266

267 Known associations from beef cattle were used to validate the proposed methodology and
 268 the results show good agreements with the literature. The qqplot showed no inflation of p-
 269 values, as expected, because the relationship structure is well accounted for [2,3,43].
 270 Additional advantages compared to “approximate” SSGWAS include avoiding the need for
 271 arbitrary choices as the length of segments, percentage of variance explained, and iterative
 272 scheme. The use of percentage of variance explained, as advocated by several authors
 273 [18,22], is misleading because it depends too much on the heterozygosity at the marker.
 274 Variance explained may be useful for breeding purposes, but may not be useful for QTL
 275 detection, where the final objective is a detailed understanding of gene action.

276

277 Although SSGWAS was possible using the last 4 years of phenotypes and all genotyped bulls
 278 present in the American Angus data, the computational burden for the factorization of the
 279 left-hand side of the mixed model equations was not negligible. In our case, matrix $\mathbf{H}^{-1}$
 280 contains (roughly) 16 million non-null elements from pedigree and 2 million non-null
 281 elements from genotypes. However, the matrix $\mathbf{H}^{-1}$ for the whole data set in Lourenco et al
 282 (2015) contained 63 million non-null elements from pedigree and 2704 million non-null
 283 elements from genotypes. The first simple method to reduce the size of the problem is to
 284 include only genotyped ancestors in exact SSGWAS and exclude candidates to selection. This
 285 leads to a number of genotyped animals in the thousands or tens of thousands. The second
 286 strategy is to use the sparse version of $\mathbf{G}^{-1}$ known as APY [39,48,49] that enormously
 287 increases the number of null elements in $\mathbf{G}$. If animals in the core portion of $\mathbf{G}$ are a
 288 representative sample of the population, this also improves the estimates of the SNP effects
 289 [50].

290

291 We emphasize that, contrary to regular GWAS, in SSGWAS computations do not need to be
292 repeated at each marker, therefore, all p-values were obtained in a single run of SSGWAS.

293

294

295 CONCLUSIONS

296

297 The exact Single Step GWAS is a very general and efficient strategy for QTL detection,
298 localization and testing, providing accurate p-values. It can be used in complex data sets
299 such as the ones in animal breeding, with a host of unbalanced effects, very complex models
300 and the presence of genotyped and ungenotyped animals. Our proposed strategy is
301 computationally viable for very large populations and solves the main issues in the current
302 Single Step GWAS, that were precluding the use of the method.

303

304 DECLARATIONS

305 ETHICS APPROVAL AND CONSENT TO PARTICIPATE

306 Not applicable.

307

308 CONSENT FOR PUBLICATION

309 Not applicable.

310

311 AVAILABILITY OF DATA AND MATERIALS

The data that support the findings of this study were provided from the American Angus Association but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The methods described here are included using “OPTION snp_p_value” in the parameter file in software blupf90 (factorization of the mixed model equations and solving of the SSGBLUP equations) and postGSf90 (backsolving of snp effects and computation of p-values), available at <http://nce.ads.uga.edu/software/>.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This study was partially funded by the American Angus Association (St. Joseph, MO) and by Agriculture and Food Research Initiative Competitive Grants no. 2015-67015-22936 from the US Department of Agriculture’s National Institute of Food and Agriculture.

AUTHORS CONTRIBUTIONS

The theory was refined by IA, AL and FC. IA and YM programmed the algorithms. The American Angus data set was analysed by AL and DL. DL and IM contributed critically to interpretation of results and planning of the analyses. The manuscript was initially written by AL and then completed by the rest of authors. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We acknowledge Dan Moser and Stephen Miller for discussions, corrections to the manuscript, and for sharing the American Angus Association data through the Angus Genetics Inc. Discussions with Anne Ricard (IFCE, Toulouse, France) and Toni Reverter (CSIRO, Brisbane, Australia) are highly appreciated. AL did part of this work while visiting the University of Georgia in a four-month sabbatical.

REFERENCES

1. Kang HM, Sul JH, Zaitlen NA, Kong S, Freimer NB, Sabatti C, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*. 2010;42:348–54.
2. Kennedy B, Quinton M, Van Arendonk J. Estimation of effects of single genes on quantitative traits. *Journal of Animal Science*. 1992;70:2000–12.
3. Teyssèdre S, Elsen J-M, Ricard A. Statistical distributions of test statistics used for quantitative trait association mapping in structured populations. *Genet Sel Evol*. 2012;44:32.
4. VanRaden P, Wiggans G. Derivation, calculation, and use of national animal model information. *Journal of Dairy Science*. 1991;74:2737–46.
5. Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol*. 2009;41:44.
6. Ricard A, Danvy S, Legarra A. Computation of deregressed proofs for genomic selection when own phenotypes exist with an application in French show-jumping horses. *Journal of Animal Science*. 2013;91:1076–85.
7. Rupp R, Senin P, Sarry J, Allain C, Tasca C, Ligat L, et al. A Point Mutation in Suppressor of Cytokine Signalling 2 (Socs2) Increases the Susceptibility to Inflammation of the Mammary Gland while Associated with Higher Body Weight and Size and Higher Milk Production in a Sheep Model. *PLOS Genetics*. 2015;11:e1005629.
8. Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Crooker BA, et al. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. *BMC Genomics*. 2011;12:408.
9. Legarra A, Vitezica ZG. Genetic evaluation with major genes and polygenic inheritance when some animals are not genotyped using gene content multiple-trait BLUP. *Genetics Selection Evolution*. 2015;47:89.

10. VanRaden PM. Efficient Methods to Compute Genomic Predictions. *J Dairy Sci.* 2008;91:4414–23.
11. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010;42:565–9.
12. Fernando RL, Habier D, Stricker C, Dekkers JCM, Totir LR. Genomic selection. *Acta Agriculturae Scandinavica, A.* 2007;57:192–5.
13. Aguilar I, Misztal I, Johnson D, Legarra A, Tsuruta S, Lawlor T. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science.* 2010;93:743–52.
14. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. *Genet Sel Evol.* 2010;42:2.
15. Fernando RL, Dekkers JC, Garrick DJ. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genet Sel Evol.* 2014;46:50.
16. Legarra A, Christensen OF, Aguilar I, Misztal I. Single Step, a general approach for genomic selection. *Livestock Science.* 2014;166:54–65.
17. Legarra A, Ducrocq V. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. *Journal of Dairy Science.* 2012;95:4629–45.
18. Wang H, Misztal I, Aguilar I, Legarra A, Muir W. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genetics Research.* 2012;94:73–83.
19. Dikmen S, Cole JB, Null DJ, Hansen PJ. Genome-wide association mapping for identification of quantitative trait loci for rectal temperature during heat stress in Holstein cattle. *PLoS ONE.* 2013;8:e69202.
20. Tiezzi F, Parker-Gaddis KL, Cole JB, Clay JS, Maltecca C. A Genome-Wide Association Study for Clinical Mastitis in First Parity US Holstein Cows Using Single-Step Approach and Genomic Matrix Re-Weighting Procedure. *PLOS ONE.* 2015;10:e0114919.
21. Han Y, Peñagaricano F. Unravelling the genomic architecture of bull fertility in Holstein cattle. *BMC Genetics.* 2016;17:143.
22. Fernando RL, Garrick D. Bayesian methods applied to GWAS. Genome-wide association studies and genomic prediction [Internet]. 2013 [cited 2017 Feb 15]. p. 237–274. Available from: http://link.springer.com/protocol/10.1007/978-1-62703-447-0_10
23. Bernal Rubio YL, Gualdrón Duarte JL, Bates RO, Ernst CW, Nonneman D, Rohrer GA, et al. Meta-analysis of genome-wide association from genomic prediction models. *Anim Genet.* 2016;47:36–48.

24. Chen C, Steibel JP, Tempelman RJ. Genome-Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods. *Genetics*. 2017;206:1791–806.
25. Gualdrón Duarte JL, Cantet RJ, Bates RO, Ernst CW, Raney NE, Steibel JP. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC Bioinformatics*. 2014;15:246.
26. Legarra A, Ricard A, Varona L. GWAS by GBLUP: Single and Multimarker EMMAX and Bayes Factors, with an Example in Detection of a Major Gene for Horse Gait. *G3: Genes, Genomes, Genetics*. 2018;8:2301–8.
27. Lu Y, Vandehaar MJ, Spurlock DM, Weigel KA, Armentano LE, Connor EE, et al. Genome-wide association analyses based on a multiple-trait approach for modeling feed efficiency. *J Dairy Sci*. 2018;101:3140–54.
28. Henderson CR. Best linear unbiased estimation and prediction under a selection model. *Biometrics*. 1975;423–447.
29. Henderson CR. *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph; 1984.
30. Aguilar I, Tsuruta S, Masuda Y, Lourenco D, Legarra A, Misztal I. BLUPF90 suite of programs for animal breeding with focus on genomics. *Proceedings of the World Congress on Genetics Applied to Livestock Production [Internet]*. Auckland, New Zealand; 2018 [cited 2018 Jul 30]. p. 751. Available from: <http://www.wcgalp.org/proceedings/2018/blupf90-suite-programs-animal-breeding-focus-genomics>
31. Vitezica Z, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. *Genetics Research*. 2011;93:357–66.
32. Christensen O, Madsen P, Nielsen B, Ostensen T, Su G. Single-step methods for genomic evaluation in pigs. *Animal*. 2012;6:1565–71.
33. Misztal I, Perez-Enciso M. Sparse Matrix Inversion for Restricted Maximum Likelihood Estimation of Variance Components by Expectation-Maximization. *Journal of Dairy Science*. 1993;76:1479–83.
34. Masuda Y, Baba T, Suzuki M. Application of supernodal sparse factorization and inversion to the estimation of (co) variance components by residual maximum likelihood. *Journal of Animal Breeding and Genetics*. 2014;131:227–36.
35. Masuda Y, Aguilar I, Tsuruta S, Misztal I. Technical note: Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements. *J Anim Sci*. 2015;93:4670–4.
36. Strandén I, Garrick DJ. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J Dairy Sci*. 2009;92:2971–5.

37. Misztal I, Wang H, Aguilar I, Legarra A, Muir B. Genome-wide association mapping using single-step GBLUP. 63 Annual Meeting of the European Association of Animal Production (EAAP) 2012-08-27/2012-08-31, Bratislava, Slovakia. Wageningen Academic Publisher; 2012.
38. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*. 2012;44:821–4.
39. Lourenco D a. L, Tsuruta S, Fragomeni BO, Masuda Y, Aguilar I, Legarra A, et al. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *J Anim Sci*. 2015;93:2653–62.
40. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *The Journal of Open Source Software*. 2018;3:25.
41. Hu Z-L, Fritz ER, Reecy JM. AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucleic Acids Res*. 2007;35:D604-609.
42. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol*. 2009;10:R42.
43. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed model association methods. *Nat Genet*. 2014;46:100–6.
44. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999;55:997–1004.
45. Peters SO, Kizilkaya K, Garrick DJ, Fernando RL, Reecy JM, Weaber RL, et al. Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcass traits in Brangus heifers. *J Anim Sci*. 2012;90:3398–409.
46. Saatchi M, Beever JE, Decker JE, Faulkner DB, Freetly HC, Hansen SL, et al. QTLs associated with dry matter intake, metabolic mid-test weight, growth and feed efficiency have little overlap across 4 beef cattle studies. *BMC Genomics*. 2014;15:1004.
47. Wang H, Misztal I, Aguilar I, Legarra A, Fernando RL, Vitezica Z, et al. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Frontiers in Genetics* [Internet]. 2014;5. Available from: http://www.frontiersin.org/Journal/Abstract.aspx?s=620&name=livestock_genomics&ART_DOI=10.3389/fgene.2014.00134
48. Misztal I. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics*. 2016;202:401–409.
49. Masuda Y, Misztal I, Tsuruta S, Lourenco DA, Fragomeni BO, Legarra A, et al. Single-step genomic evaluations with 570K genotyped animals in US Holsteins. *Interbull Bulletin*. 2015;
50. Lourenco DAL, Legarra A, Moser D, Miller S, Misztal I. Tuning indirect predictions based on SNP effects from single-step GBLUP. *Interbull Bulletin* [Internet]. 2018 [cited 2018 Nov 13];0. Available from: <https://journal.interbull.org/index.php/ib/article/view/1448>

Manhattan plot of Variances