GENOME-WIDE ASSOCIATION MAPPING INCLUDING PHENOTYPES FROM RELATIVES WITHOUT GENOTYPES

by

HUIYU WANG

(Under the Direction of Ignacy Misztal)

ABSTRACT

A common problem for genome-wide association analysis (GWAS) is lack of power for detection of quantitative trait loci (QTLs) and low precision for fine mapping. Here, we present a statistical method, termed "ssGBLUP", which increases both power and precision without increasing genotyping costs by taking advantage of phenotypes from other related and unrelated subjects. The procedure achieves these goals by blending traditional pedigree relationships with those derived from genetic markers, and by conversion of estimated breeding values (EBVs) to marker effects and weights. Efficiency of the method was first examined using simulations with 15,800 subjects, of which 1500 were genotyped. Comparison included two scenarios of ssGBLUP (S1 and S2), classical genome-wide association (CGWAS) and BayesB. For genomic evaluation, the highest accuracy of prediction was obtained by the second iteration of ssGBLUP. Power and precision for GWAS were evaluated by the correlation between true QTL effects and the sum of m adjacent single nucleotide polymorphism (SNP) effects. The best accuracy for QTL mapping occurred for ssGBLUP with m=8, and BayesB with m=16. For simulation data set, ssGBLUP is faster and easier for GWAS without computing pseudo data compared with CGWAS and BayesB.

In the second and third studies, ssGBLUP was extended to GWAS on broiler chickens for single- and multi-trait model. Dataset consisted of 2 pure lines (L1 and L2) across 5 generations for 3 traits: body weight at 6 wk (BW6), ultrasound measurement of breast meat (BM), and leg score (LS) coded 1=no and 2=yes for leg defect. Single-trait model was only based on BW6 of L2. There were 294,632 and 274,776 individuals in pedigree for L1 and L2, of which 4667 and 4553 were genotyped using a SNP60k panel. Results of QTL mapping had express in format of Manhattan plots, which were constructed as proportion of genetic variance explained by each region consisting of 20 consecutive SNPs. Different peaks across traits and lines suggest different selection goals.

The forth study analyzed distribution of differences between pedigree- and genomic-based relationship matrices (G-A). QC reduced differences and was able to identify parent-offspring conflicts. Large discrepancies between G and A imply unidentified errors or limited pedigree depth.

From both simulation and application studies on GWAS, ssGBLUP approach is faster, simpler, and easily applicable to complex models including multi-trait, maternal effects, indirect genetic effects, and random regression.

INDEX WORDS: body weight, breast meat, broiler chickens, genomic-wide association, leg score, SNP

GENOME-WIDE ASSOCIATION MAPPING INCLUDING PHENOTYPES FROM RELATIVES WITHOUT GENOTYPES

by

HUIYU WANG

B.S., China Agricultural University, 2007

M.S., The University of Georgia, 2009

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2013

GENOME-WIDE ASSOCIATION MAPPING INCLUDING PHENOTYPES FROM RELATIVES WITHOUT GENOTYPES

by

HUIYU WANG

Major Professor:	Ignacy Misztal
Committee:	J. Keith Bertrand
	Romdhane Rekaya

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

ACKNOWLEDGEMENTS

I would like to appreciate Dr. Ignacy Misztal for his advice on my research and study. He always inspires me to think horizontally and vertically. Besides academic life, he is also like a friend to encourage me to explore the fun of nature and outdoor adventures.

I would like to thank other committee members: Drs. J. Keith Bertrand and Romdhane Rekaya for all help and assistance; Dr. William M. Muir for his collaboration on publications and inspiring advice; and Dr. Selma Forni Davis who helped me so much academically and personally. Thanks for all the understanding and supervision for all my committee members. It is impossible to accomplish this without their encouragements. Many thanks to all other coauthors of our publications: Drs. Ignacio Aguilar, Andres Legarra, Rohan. L. Fernando, and Rachel Hawken. Thank you for the comments and revision.

As the coworkers and friends, I want to acknowledge Drs. Shogo Tsuruta, Daniela Lino, Ignacio Aguilar, and Ching-Yi Chen. I appreciate all the inspiring conversations on research and happy time of hanging out after school. In addition, I would like to thank all other coworkers in my group, who share the laughter and difficult days with me for so many years. Many thanks to my good friends Kaori Tokuhisa, Ling-Yun Chang, Jing Xu, and Lei Wang who always stand by me. And I want to express appreciation for the consistent help from professors and peers of Iowa State University, who took good care of me during my visit.

Finally, I would like to appreciate my parents and grandma. Their unconditional love encourages me all the time. Thanks for their guidance and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Genomic selection (GS) is a methodology to predict the estimated breeding value (EBV) using a dense set of genetic markers. GS can be implemented by predicting marker effects, usually using Bayesian methods collectively known as BayesX, or by solving BLUP with a genomic relationship matrix (GBLUP). Both methods are equivalent when marker variances are identical although GBLUP is usually run assuming equal variance per marker.

When the population contains both genotyped and ungenotyped animals, GS can be conducted with single- or multiple-step methodologies. In multiple-step methodologies, observations of genotyped animals are augmented by information from ungenotyped animals forming pseudo-observation. Subsequently, pseudo-observations are used with BayesX or GBLUP. An extra step may blend the conventional and genomic EBVs. Multiple step methods are complicated and involve approximations, which reduce accuracy and introduce biases. In single-step genomic BLUP (ssGBLUP), the pedigree and genomic relationships are combined into a comprehensive relationship matrix **H**. The single-step methodology is simpler to use and, due to fewer assumption, can yield more accurate and less biased GEBV. The ssGBLUP has been successfully implemented in dairy cattle, pigs, and chickens.

In BayesX methods, prior distributions of SNP effects are selected to impose stronger shrinkage on small SNP effects. Solutions to SNP effects can be directly used for genome-wide association studies (GWAS), however, the choice of priors can strongly influence these solutions. In particular, hyperparameter π in BayesB or BayesCπ determines how many markers

are allowed to have large effects. When fewer markers are selected, identification of important markers becomes simpler. However, their estimated values may be greatly inflated.

SNP effects can also be obtained via GBLUP or ssGBLUP by conversion of EBV. Regular GBLUP usually follows the "infinitesimal model" as marker variances are assumed equal for all SNPs. This assumption works usually well in practice as (i) most quantitative traits are approaching highly polygenic model of inheritance, and (ii) detailed knowledge of genetic architecture is not necessary to obtain GEBVs. For GWAS by GBLUP or ssGBLUP, a modification is required to adapt the locus-specific variance, forming a trait-specific relationship matrix (TA). Adaptation of GBLUP to TA and GWAS lead to a method called "fastBayesA". An adaptation of ssGBLUP to TA and GWAS is a topic of this dissertation.

Most GWAS have been performed considering one trait at a time although the selection of animals within each line is based on multiple traits. However, the underlying genetic architectures for the same trait across lines might be different, due to different selection pressures. For example, improvement in accuracy of GEBVs varied by lines in broiler chickens, despite similar heritability of each trait. To provide more insight into GWAS across traits and lines, a multiple trait model is desirable to compare results from different lines.

Matrix **H** used by ssGBLUP is based on differences between two relationship matrices for genotyped individuals, one based on the genetic information (**G**) and one on pedigrees ($\mathbf{A_{22}}$). As the coefficients of $\mathbf{A_{22}}$ are proportions that gene pairs are identical by descent (IBD), elements in **G** are identical by state (IBS) and capture parts of the Mendelian sampling. Previous studies indicated small differences (SD around 0.05) between expected and realized relationships. Therefore, large discrepancies between **G** and $\mathbf{A_{22}}$ may indicate errors in SNP chips, pedigree, or limited pedigree depth and could aid in quality control (QC) of genotypes and pedigrees.

The objectives of this dissertation are: 1) to modify the ssGBLUP to adapt to TA and GWAS and test with a simulated dataset; 2) to compare GWAS by ssGBLUP with other GWAS methods widely used, including single-marker model and BayesB; 3) to apply GWAS by ssGBLUP to several traits in two lines of chicken using a multitrait model; and 4) to evaluate differences between $\mathbf{G}$ and $\mathbf{A}_{22}$ derived from a real data set under different levels of QC.

CHAPTER 2

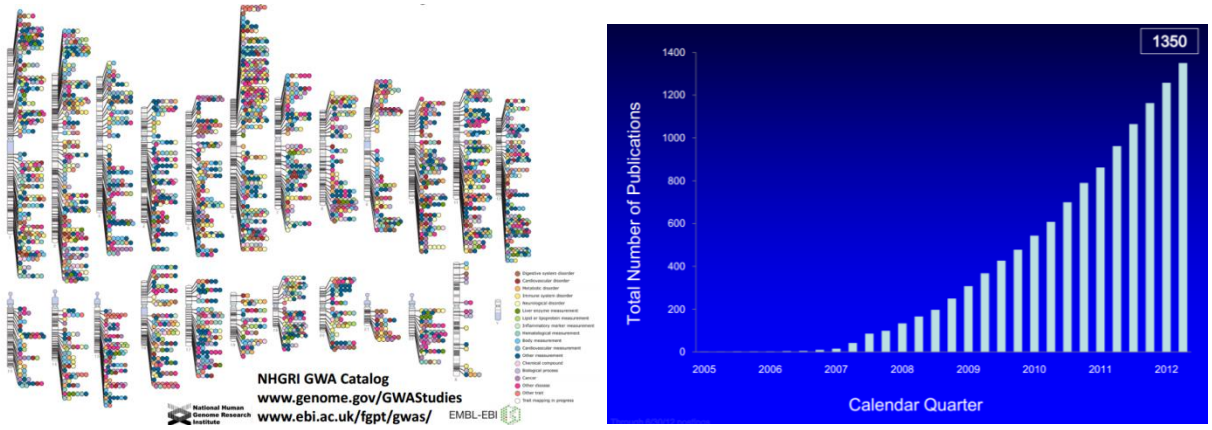
REVIEW OF LITERATURE

GENOME-WIDE ASSOCIATION STUDY (GWAS): CONCEPT AND PROGRESS

According to U.S. Department of Health & Human Services, GWAS is defined as a tool "to identify common factors that influence health and disease" (Cho et al., 2012; Fanous et al., 2012; Stephan et al., 2011; Zhang et al., 2012b). This concept could be extended to any complex quantitative trait in human, animal, plant, and other organisms (Hannum et al., 2009; Hayes et al., 2010; Li et al., 2013b; Yang et al., 2010). Under the general hypothesis of "common disease-common variant", a comprehensive international project in human ("HapMap") has been carried out to seek common variants and link them to specific illnesses. As the most common type of genetic variation in human genome, single nucleotide polymorphisms (SNPs) were used in DNA sequencing. The International HapMap Consortium (2005) reported its completion in phase I and all data were public and freely available to researchers. Through 2005 to Feb 2013 (Figure 2.1), 8621 SNPs on human genome have been identified statistically associated with different complex traits, and the number of publications increased dramatically to 1519 (Hindorff et al., 2013).

Progress of GWAS for domestic animals is rapid in recent decades inspired by research in human. However the implements of GWAS in farm animals rely on different models used in human studies. First, GWAS in human has been carried out on both a global and an individual scale (Norrgard, 2008). In farm animals, population level is more important than individual's

risk. Second, GWAS in human are mostly interested in complex disease for clinical utility, which requires case/control phenotypes. Collection of accurate phenotypes is a big issue, as



(A) (B)

Figure 2.1 A catalog of published Genome-Wide Association Studies (GWAS). (A) Published GWAS through July, 2012 at $p \leq 5x10^{-8}$ for 18 traits categories; (B) published GWAS reports through 2005 to June, 2012.

some symptoms of disease are too subtle to confirm, data requires protection of participants' privacy, and sample size are limited for linkage analyses within families (Im et al., 2012; Stranger et al., 2011; Yoon et al., 2012). In farm animals, the priority is improving production traits, such as milk yield for dairy cattle or sheep, growth and feed efficiency for pigs, beef cattle or sheep and broiler chickens, and egg production for layer chickens (Thornton, 2010). Fertility and disease traits are treated as "secondary traits" (Freeman, 1984; Wesseldijk, 2004). Most phenotypes in animals are continuous. Also, as farm animals are under highly controlled management, phenotype collection is more standardized and complete than in human studies. Third, for GWAS, the density of SNP chips implemented in human and livestock species varies largely. For example, the number of SNP markers identified by HapMap is generally ranging from 250k to 500k, and high density assay consists of nearly 1.2 million loci (e.g. Infinium® HD BeadChips by Illumina). However, assay with 50k to 60k SNPs are more often used in GWAS

for livestock species although BovineHD chips (over 770k) are currently available for limited populations (Rincon et al., 2011). Meanwhile, low density of 3k loci has been widely used for genotyping females (VanRaden et al., 2013). This is not only because of research budget, but also based on characteristics of genome structures. According to demographic history, the extent of linkage disequilibrium (LD) varies largely in human and domestic animal populations, and thus requirements of density of SNP chips are different to achieve similar power of association (Druet and Georges, 2010). The extent of LD also indicates difference of effective population size (N_e). For example, N_e of Utah residents with ancestry from northern and western Europe (CEU) and Yoruba in Ibadan, Nigeria (YRI) are ~ 3100 and ~7500 (Tenesa et al., 2007). For highly inbred species and pure lines of domestic animals, N_e may vary according to selection strategy, but still barely higher than hundred. For example, N_e of current North American and Australian Holstein Cattle is reported as 80 (Goddard et al., 2006; Sargolzaei et al., 2007). For Finnish Landrace and Yorkshire, N_e are respectively 80 and 55 (Uimari and Tapio, 2011). For Spanish Churra sheep, N_e is estimated as 128 (Garcia-Gamez et al., 2012). For commercial lines of chickens, N_e of layers is ~800 (Muir, 1997); and for broilers of Aviagen, N_e ranges from 50 to 200 (Andreescu et al., 2007). The smaller effective population size in animals is due to wide use of artificial insemination using elite sires (Delany, 2003).

METHODOLOGIES OF GWAS BASED ON LINEAR MODEL

Realized relationships based on marker genotypes:

For GWAS in human, unrelated individuals were commonly used to avoid "over-matched" genotypes among relatives (Risch and Teng, 1998; Stranger et al., 2011). However, Visscher (2008) suggested little power will be lost using related individuals, and gains would

include quality control, population stratification, and linkage disequilibrium-linkage analysis (i.e. "LDLA") mapping. Population stratification has been considered as a serious issue during GWAS. Distinct genetic ancestries of case/control groups reflect in allele frequency differences across ethnic groups, which cause spurious associations. This can be avoided if genomic relatedness is used (Cardon and Palmer, 2003; Falush et al., 2003; Lewis and Knight, 2012). For populations of animals (e.g. cattle) which have admixture population structure consisting of multiple strains or lines, breeds/strains/lines should be considered during analyses. They can be included in model based on loadings from principle components in the model, or referred through clustering methodologies for genetic ancestry (Bamshad et al., 2003; Falush et al., 2003; Khatib, 2012; Price et al., 2006; Price et al., 2010).

In animal studies, genetic variances are assumed to be equal within family in GBLUP. And the genomic relationship matrix (**G**) can be calculated and scaled as follows (VanRaden, 2008b):

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{k},$$

where **Z** is an covariate matrix for SNP markers (0/1/2), and k is a scaling parameter $k = 2\sum p_j(1-p_j)$ where p_j is allele frequency for the second allele of j-th SNP, and SNP effects were assumed to be independent (Gianola et al., 2009). G matrix can also be scaled based on other k values (Gianola et al., 2009; Hayes et al., 2009; Legarra et al., 2009). Scaling of G matrix is important for precision of genomic research especially combined with pedigree relationship matrix (**A**) (Forni et al., 2011). Moreover, the G matrix can be used as a diagnostic tool based on large difference of diagonal elements with A for mislabeled animals (Simeone et al., 2011).

The G matrix can also be calculated with additional diagonal matrix with weights (**D**), which makes unequal variances and shrinkage for SNP markers. The elements in **D** are either a

function of allele frequencies where $D_j = \frac{1}{2p_j(1-p_j)}$, or proportional to squared SNP substitution effects with scale (Amin et al., 2007; Leutenegger et al., 2003; Veerkamp et al., 2010; Zhang et al., 2010c).

Statistical framework

There are three essential factors to implement GWAS: large sample size, efficient genotypes that cover whole genome, and powerful and unbiased analytic methods (Cantor et al., 2010). Model and method choice should be based on objective of the research and available resource. On one hand, non-linear models have been considered through Bayesian inference to account for non-normal QTL distribution, or data mining and machine learning methods accounting for interactions of gene by gene, gene by environment and other non-additive effects (Moore et al., 2010; VanRaden, 2008b; VanRaden et al., 2009). On the other hand, for quantitative traits, linear regression (additive model) is more commonly used according to its simplicity, stability, wide available software support, and more other advantages (Cantor et al., 2010; VanRaden, 2008b). To implement GWAS through a linear model, single SNP locus could be fitted as a fixed effect one by one or few tagged SNPs as fixed effect through pre-selection to avoid "small n, large p" problem (p: number of SNPs; n: sample size), improved with including background polygenic effects, population structure, or other environment effects in model (He and Lin, 2011). As the increasing data sets in phenotypes, genotypes and pedigrees, and requirement of fast, powerful and robust computing, software packages have been developed based on various objectives of GWAS research (Kang et al., 2010; Meyer and Tier, 2012; Zhang et al., 2010a). Comparison of several available packages useful for mixed model on GWAS has been shown in Table 2.1(Zhang et al., 2009).

Table 2.1 Software packages useful for mixed model for association mapping.

Category	Program	Web address (http)	Availability	User interface[b]	Flexible modeling	Automatic GWAS[c]	Sample size[d]	Population structure	Build Kinship from pedigree	Build Kinship from marker	Number of Random Effects	Reference
Multi-purpose	TASSEL	www.maizegenetics.net	Free	G/C	No	Yes	S	Yes	Yes	Yes	1	[71]
	SAS	www.sas.com	Licensed	C	Yes	Yes	S	Yes	Yes	Yes	≥1	[73]
	JMP Genomics	www.jmp.com/software/genomics	Licensed	G	Yes	Yes	NA	Yes	NA	Yes	≥1	[74]
Mixed model	ASREML[d]	www.vsni.co.uk/software/asreml	Licensed	C	Yes	Yes	NA	Yes	Yes	No	≥1	[56]
	MTDFREML	aipl.arsusda.gov/curtvt/mtdfreml.html	Free	C	Yes	No	L	Yes	Yes	No	≥1	[75, 76]
	DMU	www.dmu.agrsci.dk	Free	C	Yes	No	L	Yes	Yes	No	≥1	[77]
	QxPak	nce.ads.uga.edu/~ignacy/newprograms.html	Free	C	Yes	Yes	L	Yes	Yes	No	≥1	[78]
	WOMBAT	agbu.une.edu.au/~kmeyer/wombat	Free	C	Yes	NA	L	Yes	Yes	No	≥1	[79]
	EMMA(R)	mouse.cs.ucla.edu/emma	Free	C	No	Yes	M	No	No	Yes	1	[45]
Structure	InStruct	cbsuapps.tc.cornell.edu/InStruct.aspx	Free	C			S	Yes				[38]
	Eigensoft	genepath.med.harvard.edu/~reich/Software.htm	Free	C			M	Yes				[41, 42]
	STRUCTURE	pritch.bsd.uchicago.edu/structure	Free	G			S	Yes				[35]
Kinship	PowerMarker	statgen.ncsu.edu/powermarker	Free	G			S	No	No	Yes		[80]
	SPAGeDi	www.ulb.ac.be/sciences/ecoevol/spagedi.html	Free	C			S	No	No	Yes		[44]

NA: not available.

[a]Software packages are sorted roughly by number of functions and desirable features. The evaluation of functions and features were based solely on authors' judgements, which may be biased. While the software Genstat was included in the text, it is not included in this table because of the authors' lack of familiarity with it.

[b]ASREML is available as standalone and as S language and R (ASREML-S and ASREML-R) add-ons.

[c]A software package was considered to have a graphical (G) user interface when all analyses could be performed by mouse clicking and guided keyboard input; otherwise, it was classified as command (C) line interface.

[d]Automatic GWAS refers to whether a single analysis automatically tests all markers across the genome as opposed to manually testing one marker at a time.

[e]Sample size for which software can perform an association test in minutes per marker or estimate structure or kinship in hours: Small (S): less than 1000; Medium (M): between 1000 and 5000 and Large (L): larger than 5000. Estimates of software capacity are approximate and based on authors' experience rather than exhaustive testing.

However, the univariate model requires a very stringent significance tests, where lots of individuals effects are too small to pass even they do contribute fraction of total genetic variation (Bloom et al., 2013; Yang et al., 2010). Other than that, single marker model ignores the LD information with neighboring markers and thus less powerful compared with models considering multiple loci (Akey et al., 2001; He et al., 2011; Stringer et al., 2011). As quote from Visscher et al. (2012): "… surprisingly large proportion of additive genetic variation is tagged when all SNPs are considered simultaneously". As the level of markers are too large to fit as fixed effects, SNPs effects can be treated as random effects in a mixed model frame (Meuwissen et al., 2001a). Different methodology could be applied based on hypothesis and research objectives. The most commonly used approaches are under two categories: (i) directly fit SNPs as random effects in model, or (ii) indirectly derive SNP solutions from genetic values based on the equivalent model, where both can be implemented by GBLUP or Bayesian regression (de Los Campos et al., 2013; Goddard et al., 2009; Stranden and Garrick, 2009a). For both methods, shrinkage factors would be engaged optionally; the properties of 3 common types of shrinkage have been shown in Figure 2.2 (Chen et al., 2010).

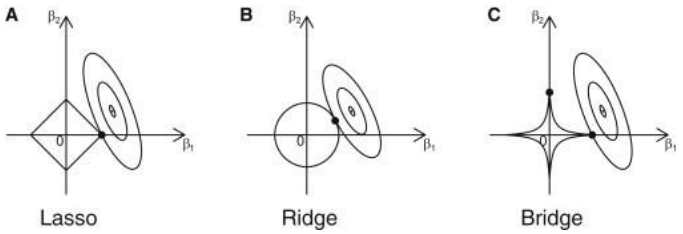


Figure 2.2 A graphic illustration of the properties of different penalty functions. A graphic illustration of the properties of three different penalty functions. The eclipses represent the likelihood contours. (A–C) The square, round, and star shapes represent the lasso, ridge, and bridge constraint, respectively. The dots are the points where likelihood contours are "tangent" to the constraints, i.e., the penalized likelihood estimates. Note that in lasso (A) or bridge (C), the constraint is discontinuous at zero. If the likelihood contour first touches the constraint at point zero, the corresponding parameter estimate is zero, and variable selection is achieved.

Fit SNP effects in model:

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

where **y** is a vector of phenotypes or pseudo-phenotypes (e.g. deregressed EBVs, or daughter yield deviation); μ is overall mean; **u** is vector of polygenic background effect; **X** is genotype covariates matrix (0, 1, 2) with dimension of N×M, where N is the number of individuals and M is the number of SNPs; **g** is a vector of additive marker effects and i-th SNPs g_i~ N(0, σ_{gi}^2); **e** is a vector of residuals.

With Bayesian methods (e.g. BayesA, BayesB and BayesC, etc), SNPs effects, variance and other parameters can be inferred by Markov-Chain Monte Carlo procedure through Gibbs sampling or Metropolis-Hastings steps. Difference of these Bayesian methods is mainly based on prior specifications for **g** (Habier et al., 2010b; Meuwissen et al., 2001a). For BayesA and BayesB, variance of i-th SNP σ_{gi}^2 is locus specific and follows a scaled inversed chi-square distribution with known priors. However, BayesB has an additional parameter π, which is defined as probability that $\sigma_{gi}^2 = 0$, and σ_{gi}^2 is non-zero for all locus in BayesA. That implies stronger shrinkage in BayesB than in BayesA. BayesC is similar to BayesB while the variance is equal to every locus (Verbyla et al., 2009). Therefore, all types of Bayesian methods are

analogous to BayesB: as BayesA is the special case of BayesB where $\pi = 0$, and BayesC is another case of BayesB where σ_{gi}^2 is a constant.

Moreover, Zhang et al. (2010c) and Sun et al. (2012) proposed similar methods through non-sampling procedure based on different weight elements in matrix of **D** to reconstruct the realized relationship **G** matrix. Both weights are based on SNP marker variance, but with different statistics theories: calculated through posterior/estimated marker effects from BayesB/BLUP $\sigma_{gi}^2 = 2p_i(1-p_i)\hat{g}_i^2$, or derived through EM algorithm $\sigma_{gi}^2 = \frac{\hat{g}_i^2 + \gamma_g S_g^2}{\gamma_g + 1}$; where $\hat{g}_i^2$ is squared estimated marker effects, γ_g and S_g^2 are priors of degree of freedom and scale parameters while $\sigma_{gi}^2 \sim i.i.d. \frac{\gamma_g S_g^2}{\chi_{\gamma_g}^2}$.

Derive SNP effects through the equivalent model:

The method by Sun et al. (2012) could be implemented through iterations and SNP effects (i.e. $\hat{g}_i$) requires solving by an equivalent model (Stranden and Garrick, 2009a):

Assume $$\boldsymbol{a} = \boldsymbol{Xg}$$

where **a** is additive genetic values (genomic breeding values: "GEBVs") with dimension of N×1, then

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{u} + \boldsymbol{Xg} + \boldsymbol{e} \quad \Leftrightarrow \quad \boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{u} + \boldsymbol{Ia} + \boldsymbol{e}$$

According to mixed model equation (Henderson, 1984):

$$\hat{\boldsymbol{a}} = (\boldsymbol{Z'R^{-1}Z} + \lambda_a \boldsymbol{I})^{-1}\boldsymbol{Z'R^{-1}}(\boldsymbol{y} - \mathbf{1}\hat{\mu} + \hat{\boldsymbol{u}})$$

where **Z** is the incidence matrix for **a**; $\mathbf{R}^{-1}$ is the inverse of variance of residuals; and λ_a is the ratio between residual and genetic variances thus $\lambda_a = \frac{\sigma_e^2}{\sigma_a^2}$.

Also, variances of **a** and **u** are:

$$var(\boldsymbol{u}) = \boldsymbol{I}\sigma_u^2 = \frac{\sigma_a^2}{2\sum p_i(1-p_i)}$$

$$var(\boldsymbol{a}) = var(\boldsymbol{Zu}) = \boldsymbol{Z}var(\boldsymbol{u})\boldsymbol{Z}' = \boldsymbol{ZZ}'\frac{\sigma_a^2}{2\sum p_i(1-p_i)} = \boldsymbol{G}\sigma_a^2 \text{ (VanRaden, 2008b).}$$

Full procedure can be referred to Appendix A.

Improved relationship matrix to combine both A and G matrices

As the equivalent model is available to derive marker effects through GEBVs, the methods used for genomic evaluation are also operational for GWAS. Genetic values can be derived based on genotyped individuals where it is only a small portion of the whole population. Although one can include the information from relatives of those genotyped individuals, it requires an extra step with estimation, so called as "multi-step", which may bring bias (Garrick et al., 2009; VanRaden, 2012; Vitezica et al., 2011b). To avoid problems abovementioned, a single-step procedure (SSP) incorporating GBLUP frame has been developed, and it has been applied successfully in various species (Chen et al., 2011b; Christensen and Lund, 2009; Forni et al., 2011; Misztal et al., 2009; Misztal et al., 2002; Vitezica et al., 2010). The main idea of SSP is based on a combined relationship matrix which integrated

G and A matrix together, and its inverse is:

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix} \text{ (Aguilar et al., 2010)}$$

where **H** is the combined relationship pedigree, and $\mathbf{A}_{22}$ is the numerator relationship matrix for genotyped individuals. Additionally, the matrix **G** should be compatible with **A** for more reliable prediction (Chen et al., 2011a; Forni et al., 2011; Simeone et al., 2012).

Currently GWAS progress in chicken

As abovementioned, the SNP chips used in domestic animals are ~50k, which provides decent genome coverage in many species: e.g. pigs, cattle and chickens (Rosa, 2008). In chicken, most traits of interest are still growth relevant traits, including body

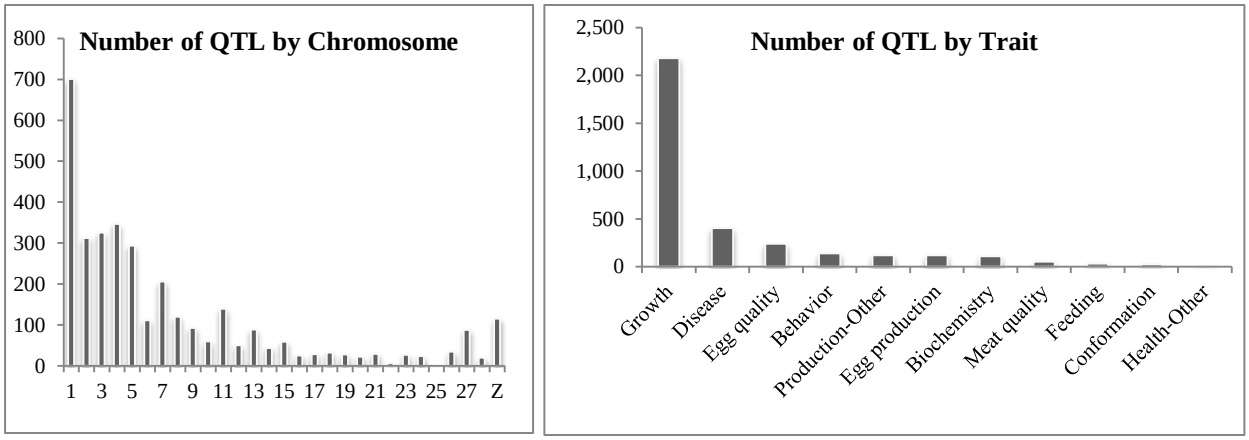


Figure 2.3 Summary of chicken database (QTLdb) by chromosome and by trait.

composition and body size, growth, body weight, fatness, shank, egg weight and uniformity, etc (Ankra-Badu et al., 2010; Gao et al., 2011; Wolc et al., 2012; Xie et al., 2012; Zhang et al., 2012a; Zhou et al., 2006a, b; Zhou et al., 2007). More and more researches are also focus on metabolic traits, resistance to *Salmonella* colonization, avian influenza virus and Marek's disease (Fife et al., 2010; Li et al., 2013a; Sironi et al., 2011). Meanwhile, chicken QTL database has been in progress and updated with newly found QTLs, and summary of reported QTL by trait and chromosome has been listed in Figure 2.3 (Hu et al., 2013).

DISCUSSION ON INTERPRETATION AND UTILIZATION OF RESULTS FROM GWAS

Generally speaking, GWAS achieved great success in finding genetic variants associated with phenotypes of various complex traits in different organisms, rediscovered many genes, and highlighted biological pathways (Cole et al., 2011; Hirschhorn, 2009; Klein and Ziegler, 2011;

Li et al., 2013b; Onteru et al., 2012). There is no doubt that we have found the common variants. However, there are more voices talked about the challenge of GWAS on "missing heritability" issue (Bush and Moore, 2012; Hardy and Singleton, 2009; Maher, 2008; Visscher et al., 2012), as reasons were talked about as follows. First, the underlying hypothesis of "common disease-common variant" might be a flaw. According to this assumption, only small fraction of genetic variation is able to be identified. Moreover, there would be a large heritability gap if the underlying genomic structure of such trait does not match this hypothesis (Zeggini, 2011). For rare variation explaining very large phenotype variance, the loci were discarded during quality control process before GWAS (Figure 2.4). Second, markers used in GWAS are SNPs, which is the single based-pair change in the DNA sequence. However, because of its limitation in variation for a single nucleotide, it has a minimal impact on biological system and causing functional consequences on expression even protein levels. This inspires new projects on other types of markers, for example copy number variation (CNV) studies (Zhang et al., 2011). Third, underlying genetic architecture has been ignored thereby most genetic variation is "missing" in population, such as gene by gene interactions (epistasis), gene by environment interactions, pleiotropic effect, etc (Stranger et al., 2011; Zwarts et al., 2011). Last, it is a big question that how to utilize the results based on GWAS, as those identified genetic variations associated with phenotypes are not necessarily causal variants/genes. This implies the importance of epigenetics and post-GWAS research, which aims to find out "the functional consequences of these loci" from GWAS (Freedman et al., 2011).

In conclusion, GWAS is still a powerful tool that leads us to understand the inherited basis of polygenic traits. It might not be the final step to uncover the veil of truth, and more realistic hypothesis and better methodologies could be involved, but that does not mean

"failure". In extent of "association", GWAS has accomplished its mission. However, there is still a long way to go for detecting functional/causal variants and underlying biology.

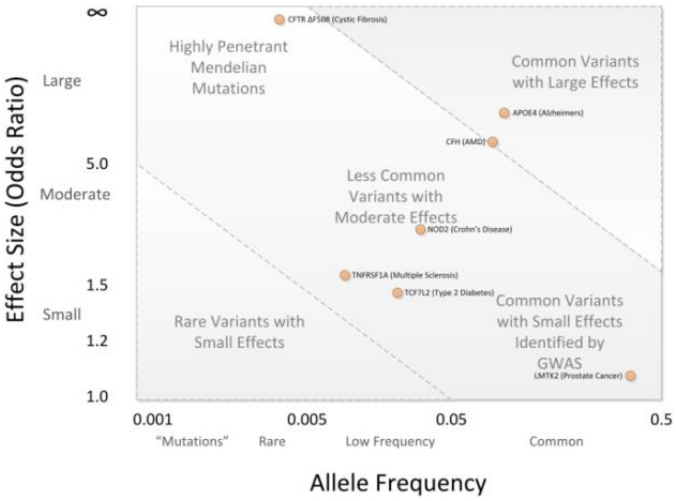


Figure 2.4 Spectrum of disease allele effects. Disease associations are often conceptualized in two dimensions: allele frequency and effect size. Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). The bulk of discovered genetic associations lie on the diagonal denoted by the dashed lines.
doi:10.1371/journal.pcbi.1002822.g001

REFERENCES

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

Akey, J., L. Jin, and M. Xiong. 2001. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9: 291-300.

Amin, N., C. M. van Duijn, and Y. S. Aulchenko. 2007. A genomic background based method for association analysis in related individuals. PLoS One 2: e1274.

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont, and J. C. Dekkers. 2007. Linkage disequilibrium in related breeding lines of chickens. Genetics 177: 2161-2169.

Ankra-Badu, G. A., E. Le Bihan-Duval, S. Mignon-Grasteau, F. Pitel, C. Beaumont, M. J. Duclos, J. Simon, W. Carre, T. E. Porter, A. Vignal, L. A. Cogburn, and S. E. Aggrey. 2010. Mapping QTL for growth and shank traits in chickens divergently selected for high or low body weight. Anim Genet 41: 400-405.

Bamshad, M. J., S. Wooding, W. S. Watkins, C. T. Ostler, M. A. Batzer, and L. B. Jorde. 2003. Human population genetic structure and inference of group membership. Am J Hum Genet 72: 578-589.

Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T. L. Lite, and L. Kruglyak. 2013. Finding the sources of missing heritability in a yeast cross. Nature 494: 234-237.

Bush, W. S., and J. H. Moore. 2012. Chapter 11: Genome-wide association studies. PLoS Comput Biol 8(12): e1002822. doi:10.1371/journal.pcbi.1002822.

Cantor, R. M., K. Lange, and J. S. Sinsheimer. 2010. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet 86: 6-22.

Cardon, L. R., and L. J. Palmer. 2003. Population stratification and spurious allelic association. Lancet 361: 598-604.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011a. Effect of different genomic relationship matrices on accuracy and scale. J Anim Sci.

Chen, C. Y., I. Misztal, I. Aguilar, S. Tsuruta, T. H. Meuwissen, S. E. Aggrey, T. Wing, and W. M. Muir. 2011b. 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, L. S., C. M. Hutter, J. D. Potter, Y. Liu, R. L. Prentice, U. Peters, and L. Hsu. 2010. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. Am J Hum Genet 86: 860-871.

Cho, Y. S., C. H. Chen, C. Hu, J. Long, R. T. Ong, X. Sim, F. Takeuchi, Y. Wu, M. J. Go, T. Yamauchi, Y. C. Chang, S. H. Kwak, R. C. Ma, K. Yamamoto, L. S. Adair, T. Aung, Q. Cai, L. C. Chang, Y. T. Chen, Y. Gao, F. B. Hu, H. L. Kim, S. Kim, Y. J. Kim, J. J. Lee, N. R. Lee, Y. Li, J. J. Liu, W. Lu, J. Nakamura, E. Nakashima, D. P. Ng, W. T. Tay, F. J. Tsai, T. Y. Wong, M. Yokota, W. Zheng, R. Zhang, C. Wang, W. Y. So, K. Ohnaka, H. Ikegami, K. Hara, Y. M. Cho, N. H. Cho, T. J. Chang, Y. Bao, A. K. Hedman, A. P. Morris, M. I. McCarthy, D. Consortium, T. C. Mu, R. Takayanagi, K. S. Park, W. Jia, L. M. Chuang, J. C. Chan, S. Maeda, T. Kadowaki, J. Y. Lee, J. Y. Wu, Y. Y. Teo, E. S. Tai, X. O. Shu, K. L. Mohlke, N. Kato, B. G. Han, and M. Seielstad. 2012. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet 44: 67-72.

Christensen, O. F., and M. S. Lund. 2009. Genomic relationship matrix when some animals are not genotyped. Page 299 in Proc. 60th Annual Meeting EAAP, Barcelona, Spain. Wageningen Press, Wageningen, the Netherlands.

Cole, J. B., G. R. Wiggans, L. Ma, T. S. Sonstegard, T. J. Lawlor, Jr., B. A. Crooker, C. P. Van Tassell, J. Yang, S. Wang, L. K. Matukumalli, and Y. Da. 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics 12: 408.

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.

Delany, M. E. 2003. Genetic diversity and conservation of poultry. In: Muir, W. M. and S. E. Aggrey, editor, Poultry genetics, breeding and biotechnology. 1st ed. CABI, Oxfordshire, UK. p 274.

Druet, T., and M. Georges. 2010. A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184: 789-798.

Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

Fanous, A. H., B. Zhou, S. H. Aggen, S. E. Bergen, R. L. Amdur, J. Duan, A. R. Sanders, J. Shi, B. J. Mowry, A. Olincy, F. Amin, C. R. Cloninger, J. M. Silverman, N. G. Buccola, W. F. Byerley, D. W. Black, R. Freedman, F. Dudbridge, P. A. Holmans, S. Ripke, P. V. Gejman, K. S. Kendler, D. F. Levinson, and C. Schizophrenia Psychiatric Genome-Wide Association Study. 2012. Genome-wide association study of clinical dimensions of schizophrenia: polygenic effect on disorganized symptoms. Am J Psychiatry 169: 1309-1317.

Fife, M. S., J. S. Howell, N. Salmon, P. M. Hocking, P. M. van Diemen, M. A. Jones, M. P. Stevens, and P. Kaiser. 2010. Genome-wide SNP analysis identifies major QTL for Salmonella colonization in the chicken. Anim Genet.

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43: 1.

Freedman, M. L., A. N. Monteiro, S. A. Gayther, G. A. Coetzee, A. Risch, C. Plass, G. Casey, M. De Biasi, C. Carlson, D. Duggan, M. James, P. Liu, J. W. Tichelaar, H. G. Vikis, M. You, and I. G. Mills. 2011. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 43: 513-518.

Freeman, A. E. 1984. Secondary traits: sire evaluation and the reproductive complex. J Dairy Sci 67: 449-458.

Gao, Y., C. G. Feng, C. Song, Z. Q. Du, X. M. Deng, N. Li, and X. X. Hu. 2011. Mapping quantitative trait loci affecting chicken body size traits via genome scanning. Anim Genet 42: 670-674.

Garcia-Gamez, E., G. Sahana, B. Gutierrez-Gil, and J. J. Arranz. 2012. Linkage disequilibrium and inbreeding estimation in Spanish Churra sheep. BMC Genet 13: 43.

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: 55.

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.

Goddard, M. E., B. Hayes, H. McPartlan, and A. J. Chamberlain. 2006. Can the same genetic markers be used in multiple breeds? CD-ROM communication no. 22-16 in Proc. 8th World Congr. Genet. Appl. Livest. Prod., Belo Horizonte, Brazil.

Goddard, M. E., N. R. Wray, K. Verbyla, and P. M. Visscher. 2009. Estimating Effects and Making Predictions from Genome-Wide Marker Data. Statistical Science 24: 517-529.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2010. Extension of the Bayesian Alphabet for genomic selection. Page 468 in Proc. 9th World Congr. Genet. Appl. Livest. Prod, Leipzig, Germany.

Hannum, G., R. Srivas, A. Guenole, H. van Attikum, N. J. Krogan, R. M. Karp, and T. Ideker. 2009. Genome-wide association data reveal a global map of genetic interactions among protein complexes. PLoS Genet 5: e1000782.

Hardy, J., and A. Singleton. 2009. Genomewide association studies and human disease. N Engl J Med 360: 1759-1768.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92: 433-443.

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet 6: e1001139.

He, Q., and D. Y. Lin. 2011. A variable selection method for genome-wide association studies. Bioinformatics 27: 1-8.

He, Y., C. Li, C. I. Amos, M. Xiong, H. Ling, and L. Jin. 2011. Accelerating haplotype-based genome-wide association study using perfect phylogeny and phase-known reference data. PLoS One 6: e22097.

Henderson, C. R. 1984. Applications of Linear Models in Animal Breeding. Guelph Univ. Press, Guelph, Canada.

Hindorff, L. A., J. MacArthur, J. Morales, H. A. Junkins, P. N. Hall, A. K. Klemm, and T. A. Manolio. 2013. A catalog of published genome-wide association studies. Retrieved February 28, 2013, from www.genome.gov/gwastudies.

Hirschhorn, J. N. 2009. Genomewide association studies--illuminating biologic pathways. N Engl J Med 360: 1699-1701.

Hu, Z. L., C. A. Park, X. L. Wu, and J. M. Reecy. 2013. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res 41: D871-879.

Im, H. K., E. R. Gamazon, D. L. Nicolae, and N. J. Cox. 2012. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am J Hum Genet 90: 591-598.

International HapMap, C. 2005. A haplotype map of the human genome. Nature 437: 1299-1320.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42: 348-354.

Khatib, H. 2012. Bovine genomics by james e. Womack. Front Genet 3: 275.

Klein, C., and A. Ziegler. 2011. From GWAS to clinical utility in Parkinson's disease. Lancet 377: 613-614.

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

Leutenegger, A. L., B. Prum, E. Genin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E. A. Thompson. 2003. Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet 73: 516-523.

Lewis, C. M., and J. Knight. 2012. Introduction to genetic association studies. Cold Spring Harb Protoc 2012: 297-306.

Li, D. F., L. Lian, L. J. Qu, Y. M. Chen, W. B. Liu, S. R. Chen, J. X. Zheng, G. Y. Xu, and N. Yang. 2013a. A genome-wide SNP scan reveals two loci associated with the chicken resistance to Marek's disease. Anim Genet 44: 217-222.

Li, H., Z. Peng, X. Yang, W. Wang, J. Fu, J. Wang, Y. Han, Y. Chai, T. Guo, N. Yang, J. Liu, M. L. Warburton, Y. Cheng, X. Hao, P. Zhang, J. Zhao, Y. Liu, G. Wang, J. Li, and J. Yan. 2013b. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45: 43-50.

Maher, B. 2008. Personal genomes: The case of the missing heritability. Nature 456: 18-21.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Meyer, K., and B. Tier. 2012. "SNP Snappy": a strategy for fast genome-wide association studies fitting a full mixed model. Genetics 190: 275-277.

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., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. Blupf90 and related programs (BGF90). Commun. No. 28–07 in Proc. 8th World Congr. Genet. Appl. Livest. Prod., Montpellier, France.

Moore, J. H., F. W. Asselbergs, and S. M. Williams. 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics 26: 445-455.

Muir, W. M. 1997. Genetic selection strategies: computer modeling. Poult Sci 76: 1066-1070.

Norrgard, K. 2008. Genetic variation and disease: GWAS. Nature Education 1 (1).

Onteru, S. K., B. Fan, Z. Q. Du, D. J. Garrick, K. J. Stalder, and M. F. Rothschild. 2012. A whole-genome association study for pig reproductive traits. Anim Genet 43: 18-26.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

Price, A. L., N. A. Zaitlen, D. Reich, and N. Patterson. 2010. New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459-463.

Rincon, G., K. L. Weber, A. L. Eenennaam, B. L. Golden, and J. F. Medrano. 2011. Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys. J Dairy Sci 94: 6116-6121.

Risch, N., and J. Teng. 1998. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res 8: 1273-1288.

Rosa, G. J. M. 2008. Genome-wide Association Analysis (GWAS) in livestock. Page 16 in Proc. 59th Annual Meeting EAAP, Vilnius, Lithuania. Wageningen Press, Wageningen, the Netherlands.

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer. 2007. Estimating effective population size in North American Holstein cattle based on genome-wide linkage disequilibrium Dairy Cattle Breeding and Genetics Committee Meeting, Guelph, Canada.

Simeone, R., I. Misztal, I. Aguilar, and A. Legarra. 2011. Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. J Anim Breed Genet 128: 386-393.

Simeone, R., I. Misztal, I. Aguilar, and Z. G. Vitezica. 2012. Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. J Anim Breed Genet 129: 3-10.

Sironi, L., J. L. Williams, A. Stella, G. Minozzi, A. Moreno, P. Ramelli, J. Han, S. Weigend, J. Wan, G. Lombardi, P. Cordioli, and P. Mariani. 2011. Genomic study of the response of chicken to highly pathogenic avian influenza virus. BMC Proc 5 Suppl 4: S25.

Stephan, R., A. R. Sanders, K. S. Kendler, D. F. Levinson, and P. Sklar, et al. 2011. Genome-wide association study identifies five new schizophrenia loci Nat Genet No. 43. p 969-976.

Stranden, 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.

Stranger, B. E., E. A. Stahl, and T. Raj. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187: 367-383.

Stringer, S., N. R. Wray, R. S. Kahn, and E. M. Derks. 2011. Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. PLoS One 6: e27964.

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.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res 17: 520-526.

Thornton, P. K. 2010. Livestock production: recent trends, future prospects. Philos Trans R Soc Lond B Biol Sci 365: 2853-2867.

Uimari, P., and M. Tapio. 2011. Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds. J Anim Sci 89: 609-614.

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

VanRaden, P. M. 2012. Avoiding bias from genomic pre-selection in converting daughter information across countries. Interbull Bull 45: 5.

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, and G. A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci 96: 668-678.

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

Veerkamp, R. F., H. A. Mulder, and M. P. L. Calus. 2010. Estimation of heritability for dairy traits, combining pedigree with dense SNPs information on some animals. Page 138 in Proc. 9th World Congr. Genet. Appl. Livest. Prod, Leipzig, Germany.

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 (Camb) 91: 307-311.

Visscher, P. M., T. Andrew, and D. R. Nyholt. 2008. Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. Eur J Hum Genet 16: 387-390.

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang. 2012. Five years of GWAS discovery. Am J Hum Genet 90: 7-24.

Vitezica, Z. G., I. Aguilar, and A. Legarra. 2010. One-step vs. multi-step methods for genomic prediction in presence of selection. Page 131 in Proc. 9th World Congr. Genet. Appl. Livest. Prod, Leipzig, Germany.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet Res (Camb) 93: 357-366.

Wesseldijk, B. 2004. Secondary traits make up 26% of breeding goal. Holstein Inter. 11(6):8–11.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, W. G. Hill, and J. C. Dekkers. 2012. Genome-wide association analysis and genetic architecture of egg weight and egg uniformity in layer chickens. Anim Genet 43 Suppl 1: 87-96.

Xie, L., C. Luo, C. Zhang, R. Zhang, J. Tang, Q. Nie, L. Ma, X. Hu, N. Li, Y. Da, and X. Zhang. 2012. Genome-wide association study identified a narrow chromosome 1 region associated with chicken growth traits. PLoS One 7: e30910.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565-569.

Yoon, D., Y. J. Kim, and T. Park. 2012. Phenotype prediction from genome-wide association studies: application to smoking behaviors. BMC Syst Biol 6 Suppl 2: S11.

Zeggini, E. 2011. Next-generation association studies for complex traits. Nat Genet 43: 287-288.

Zhang, D., Y. Qian, N. Akula, N. Alliey-Rodriguez, J. Tang, S. Bipolar Genome, E. S. Gershon, and C. Liu. 2011. Accuracy of CNV Detection from GWAS Data. PLoS One 6: e14511.

Zhang, H., X. Hu, Z. Wang, Y. Zhang, S. Wang, N. Wang, L. Ma, L. Leng, S. Wang, Q. Wang, Y. Wang, Z. Tang, N. Li, Y. Da, and H. Li. 2012a. Selection signature analysis implicates the PC1/PCSK1 region for chicken abdominal fat content. PLoS One 7: e40736.

Zhang, R., J. D. Yan, R. K. Valenzuela, S. M. Lu, X. Y. Du, B. Zhong, J. Ren, S. H. Zhao, C. G. Gao, L. Wang, T. W. Guo, and J. Ma. 2012b. Further evidence for the association of genetic variants of ZNF804A with schizophrenia and a meta-analysis for genome-wide significance variant rs1344706. Schizophr Res 141: 40-47.

Zhang, Z., E. S. Buckler, T. M. Casstevens, and P. J. Bradbury. 2009. Software engineering the mixed model for genome-wide association studies on large samples. Brief Bioinform 10: 664-675.

Zhang, Z., E. Ersoz, C. Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, and E. S. Buckler. 2010a. Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355-360.

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

Zhou, H., N. Deeb, C. M. Evock-Clover, C. M. Ashwell, and S. J. Lamont. 2006a. Genome-wide linkage analysis to identify chromosomal regions affecting phenotypic traits in the chicken. I. Growth and average daily gain. Poult Sci 85: 1700-1711.

Zhou, H., N. Deeb, C. M. Evock-Clover, C. M. Ashwell, and S. J. Lamont. 2006b. Genome-wide linkage analysis to identify chromosomal regions affecting phenotypic traits in the chicken. II. Body composition. Poult Sci 85: 1712-1721.

Zhou, H., N. Deeb, C. M. Evock-Clover, A. D. Mitchell, C. M. Ashwell, and S. J. Lamont. 2007. Genome-wide linkage analysis to identify chromosomal regions affecting phenotypic traits in the chicken. III. Skeletal integrity. Poult Sci 86: 255-266.

Zwarts, L., M. M. Magwire, M. A. Carbone, M. Versteven, L. Herteleer, R. R. Anholt, P. Callaerts, and T. F. Mackay. 2011. Complex genetic architecture of Drosophila aggressive behavior. Proc Natl Acad Sci U S A 108: 17070-17075.

CHAPTER 3

GENOME-WIDE ASSOCIATION MAPPING INCLUDING PHENOTYPES FROM RELATIVES WITHOUT GENOTYPES[1]

[1] H. Wang, I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. *online Journal of Genetics Research (Cambridge)*. 94: 73-83. Reprinted here with permission of publisher.

ABSTRACT

A common problem for genome-wide association analysis (GWAS) is lack of power for detection of QTLs and precision for fine mapping. Here we present a statistical method, termed Single-step GBLUP (ssGBLUP), which increases both power and precision without increasing genotyping costs by taking advantage of phenotypes from other related and unrelated subjects. The procedure achieves these goals by blending traditional pedigree relationships with those derived from genetic markers, and by conversion of EBVs to marker effects and weights. Additionally, the application of mixed-model approaches allows for both simple and complex analyses that involve multiple traits and confounding factors, such as environmental, epigenetic, or maternal environmental effects. Efficiency of the method was examined using simulations with 15,800 subjects of which 1500 were genotyped. Thirty QTLs were simulated across genome and assumed heritability was 0.5. Comparisons included ssGBLUP applied directly to phenotypes, BayesB and classical GWAS (CGWAS) with deregressed proofs. An average accuracy of prediction 0.89 was obtained by ssGBLUP after one iteration, which was 0.01 higher than by BayesB. Power and precision for GWAS applications were evaluated by the correlation between true QTL effects and the sum of *m* adjacent SNP effects. The highest correlations were 0.82 and 0.74 for ssGBLUP and CGWAS with $m = 8$, and 0.83 for BayesB with $m = 16$. Standard deviations of the correlations across replicates were several times higher in BayesB than in ssGBLUP. The ssGBLUP method with marker weights is faster, more accurate, and easier to implement for GWAS applications without computing pseudo-data.

Key words: genomic evaluation, genome-wide association mapping, QTL, simulation, single step procedure

INTRODUCTION

As a result of commercial availability of highly dense SNP chips in humans, genome-wide association analysis (GWAS) has proven to be a powerful tool to identify genes for common diseases and complex traits (Hirschhorn & Daly, 2005; Visscher *et al.*, 2007). Similarly, GWAS has been applied to animals for the discovery of genes that are associated with disease and production traits (Bennett *et al.* 2010; Bolormaa *et al.* 2010; Karlsson *et al.* 2007; Orr *et al.* 2010; Pryce *et al.* 2010). In animal breeding, a closely related procedure that makes use of the same SNP chips, but for an entirely different purpose, is the estimation of breeding values (GEBVs) for genomic selection (GWMAS), a form of marker assisted selection. GWMAS is often performed with procedures called BayesA or BayesB that consider all genetic associations derived from markers (Meuwissen et al., 2001). Moreover, BayesA and BayesB solutions provide SNP effects; thus, these methods can be applied to GWAS (Goddard & Hayes, 2009; Sun *et al.*, 2011) with the additional advantage of accounting for population stratification and cryptic relatedness (Sillanpaa, 2011). The classical analysis of GWAS (CGWAS) is based on a test of a single marker, which treats each SNP marker as a covariate in the model (Hirschhorn & Daly, 2005). The main advantage of CGWAS is the ease of significance testing; however, it is likely to result in reduced fit to the data compared to methods where all SNP are jointly considered. Additionally, neither Bayesian methods nor single-marker analysis can directly include genetic association found in the pedigree of animals that have not been genotyped. Although such information can be considered indirectly in multiple-step procedures in which phenotypic data from relatives are summarized to create pseudo-data for genotyped individuals (VanRaden *et al.*, 2009), new problems can arise, such as information loss, heterogeneity caused by different amounts of information in the original data set, and bias (Vitezica *et al.*, 2011).

Thus, multiple-step methods for computing genomic predictions are not only complicated but likely suboptimal for GWAS. This is particularly true in livestock species, where pedigrees are complex, and nuclear families are the exception rather than the rule. In contrast Misztal *et al.* (2009) and Christensen and Lund (2010) proposed a single step approach (ssGBLUP) that integrates phenotypes, genotypes and pedigree information. Such information can be combined with genomic data for greater detection power and estimation precision through a properly scaled and augmented relationship matrix (Legarra *et al.*, 2009; Misztal *et al.*, 2009). The ssGBLUP method has been shown to provide more consistent solutions and better accuracy than the multiple-step approach (Aguilar *et al.*, 2010; Chen *et al.*, 2011; Forni *et al.*, 2011).

A limitation of the ssGBLUP methodology is that it is based on an infinitesimal model, which assumes equal variance for all SNP marker-QTL associated effects. An advantage of the infinitesimal model is that the resulting genomic relationship matrix is identical for all traits within a population (Aguilar *et al.*, 2010). In contrast, although BayesA or BayesB is limited in that neither can include phenotypic information from non-genotyped individuals, they remove the assumption of equal variance for all SNP marker-QTL associated effects, which appears to be a more realistic situation. Unfortunately, relaxing this assumption comes at a cost of orders of magnitude more computing time in a Bayesian framework. Combining the strengths of both methods (i.e. allowing for unequal variances in a ssGBLUP context) could improve the accuracy of the estimation of GEBVs for breeding and selection applications, and precision for the estimation of SNP effects for GWAS applications.

Estimation of weights for SNP variances can be achieved without sampling. Zhang *et al.* (2010) derived SNP weights as functions of squares of SNP effects and incorporated those variances as weights in GBLUP. Sun *et al.* (2011) developed an iterative procedure for GBLUP,

in which GEBVs were converted to SNP effects and weights were obtained similar to those in Zhang *et al.* (2010). However, neither studies could directly utilize phenotypes of ungenotyped animals.

The objectives of this research were to investigate the optimal weights on marker variances for improving accuracy and precision in GWAS and GEBVs by ssGBLUP, and to compare results from ssGBLUP, CGWAS and the BayesB methods as described by Meuwissen *et al.* (2001).

MATERIALS AND METHODS

Data simulation

Data were simulated using QMSim (Sargolzaei & Schenkel, 2009) for an additive trait with a mean of 5.0, phenotypic variance 1.0, and heritability 0.5. Two 100 cM chromosomes were simulated, and each chromosome contained 15 uniformly distributed QTLs. For chromosome 1 and chromosome 2, on average 1552 and 1448 SNP markers respectively were evenly distributed. Both SNP markers and QTLs were assumed to be bi-allelic, and no marker loci overlapped with the QTLs. Minor allele frequencies were > 0.05. Effects of QTLs were randomly sampled from a Gamma distribution with a shape factor of 0.4 and a scale factor of 1.36. All additive genetic variance resulted from the QTLs. A simulated population started at generation -1001 (i.e. base population) and consisted of 100 individuals. For generations -1001 to -1, mutation rate of 0.000025 was simulated for each locus of both QTLs and SNPs per generation, and non-overlapping generations were simulated with population size per generation increasing gradually from 100 to 2800. In generations 0 to 4, 80 randomly chosen males and 520 randomly chosen females were genotyped and produced 2600 progenies by random mating. The

phenotypic information was recorded for all animals in generations 0-5. Genotypes were recorded for all parents in generations 3 and 4, and 300 random individuals in generation 5. For recent generations 0-5, the complete data sets contained 15,800 individuals in pedigree with records, of which 1500 individuals were genotyped. The simulation was replicated 10 times. Some statistics of the simulated data set are shown in Table 3.1.

Model and Methodology

The single-trait model for ssGBLUP was:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z_a a} + \mathbf{e} \qquad (1)$$

where **y** is a vector of simulated observations (phenotypes); **1** is a vector of all ones; μ is the overall mean of phenotypic records; $\mathbf{Z_a}$ is an incidence matrix that relates individuals to phenotypes; **a** is a vector of individual animal effects; **e** is a vector of residuals. The variances of **a** and **e** are:

$$\text{var}\begin{bmatrix}\mathbf{a}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{H}\sigma_a^2 & 0\\ 0 & \mathbf{I}\sigma_e^2\end{bmatrix} \qquad (2)$$

where σ_a^2 and σ_e^2 are total genetic additive and residual variances respectively, and **H** is a matrix that combines pedigree and genomic relationships as in Aguilar *et al.* (2010), and its inverse is:

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix}0 & 0\\ 0 & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}}\end{bmatrix} \qquad (3)$$

where **A** is a numerator (pedigree) relationship matrix for all animals; $\mathbf{A_{22}}$ is a numerator relationship matrix for genotyped animals; and **G** is a genomic relationship matrix. Matrix **G** was constructed based on VanRaden *et al.* (2008) that assumed allele frequencies of the current population and adjusted for compatibility with $\mathbf{A_{22}}$, which was applied in "GC" and "BLUP_α" in Chen *et al.* (2011) and Vitezica *et al.* (2011).

(iii) *Derivation of SNP effects from breeding values*

Let the animal effects be decomposed into those for genotyped ($\mathbf{a_g}$) and ungenotyped ($\mathbf{a_n}$) animals. The animal effects of genotyped animals are a function of SNP effects:

$$\mathbf{a_g} = \mathbf{Zu} \qquad (4)$$

where **Z** is a matrix relating genotypes of each locus, and **u** is a vector of SNP marker effects. Thus, the variance of the animal effects is:

$$\text{var}(\mathbf{a_g}) = \text{var}(\mathbf{Zu}) = \mathbf{ZDZ'}\sigma_u^2 = \mathbf{G}^*\sigma_a^2 \qquad (5)$$

where **D** is a diagonal matrix of weights for variances of SNPs (**D** = **I** for GBLUP); σ_u^2 is the genetic additive variance captured by each SNP marker when no weights are present; and $\mathbf{G}^*$ is the weighted genomic relationship matrix.

The joint (co)variance of animal effects ($\mathbf{a_g}$) and SNP effects (**u**) is:

$$\text{var}\begin{bmatrix}\mathbf{a_g}\\ \mathbf{u}\end{bmatrix} = \begin{bmatrix}\mathbf{ZDZ'} & \mathbf{ZD'}\\ \mathbf{DZ'} & \mathbf{D}\end{bmatrix}\sigma_u^2 \quad (6)$$

Subsequently:

$$\mathbf{G}^* = \frac{\text{var}(\mathbf{a_g})}{\sigma_a^2} = \frac{\text{var}(\mathbf{Zu})}{\sigma_a^2} = \mathbf{ZDZ'}\frac{\sigma_u^2}{\sigma_a^2} = \mathbf{ZDZ'}\lambda \qquad (7)$$

where λ is a variance ratio or a normalizing constant. According to VanRaden *et al.* (2008), $\lambda = \frac{\sigma_u^2}{\sigma_a^2} = \frac{1}{\sum_{i=1}^{M} 2p_i(1-p_i)}$, where M is the number of SNPs and p_i is allele frequency of the second allele of the *i*-th marker. Following Stranden and Garrick (2009) one can derive:

$$\hat{\mathbf{u}} = \frac{\sigma_u^2}{\sigma_a^2}\mathbf{DZ'}\mathbf{G}^{*-1}\hat{\mathbf{a}}_\mathbf{g} \qquad (8)$$

Therefore, the equation for predicting SNP effects that uses weighted genomic relationship matrix $\mathbf{G}^*$ becomes:

$$\hat{\mathbf{u}} = \lambda \mathbf{DZ'G}^{*-1}\hat{\mathbf{a}}_\mathbf{g} = \mathbf{DZ'[ZDZ']}^{-1}\hat{\mathbf{a}}_\mathbf{g} \quad (9)$$

This is the best predictor of SNP effects given animal effects (Henderson, 1973). Estimates of SNP effects can be used to estimate individual variance of each SNP effect (Zhang *et al.*, 2010):

$$\hat{\sigma}^2_{u,i} = \hat{u}_i^2 2p_i(1-p_i) \quad (10)$$

Computing algorithm

The above formulas can be used to create an algorithm for estimation of **D** from ssGBLUP. Denote *t* as an iteration number and *i* as the *i*-th SNP. The algorithm proceeds as follows:

1. $t = 0$, $\mathbf{D}_{(t)} = \mathbf{I}$; $\mathbf{G}^*_{(t)} = \mathbf{ZD}_{(t)}\mathbf{Z'}\lambda$
2. Compute $\hat{\mathbf{a}}_\mathbf{g}$ by ssGBLUP
3. Calculate $\hat{\mathbf{u}}_{(t)} = \lambda \mathbf{D}_{(t)}\mathbf{Z'}\mathbf{G}^{*-1}_{(t)}\hat{\mathbf{a}}_\mathbf{g}$
4. Calculate $d^*_{i_{(t+1)}} = \hat{u}^2_{i_{(t)}} 2p_i(1-p_i)$ for all *i* as in Zhang *et al.* (2010)
5. Normalize $\mathbf{D}_{(t+1)} = \frac{tr(\mathbf{D}_{(0)})}{tr(\mathbf{D}^*_{(t+1)})}\mathbf{D}^*_{(t+1)}$
6. Calculate $\mathbf{G}^*_{(t+1)} = \mathbf{ZD}_{(t+1)}\mathbf{Z'}\lambda$
7. $t = t+1$
8. Exit, or loop to step 2 or 3.

In looping to step 3 (scenario S1), one applies the revised $\mathbf{G}^*$ only for the prediction of SNP effects while calculating the animal effects only once, thus $\hat{\mathbf{a}}_\mathbf{g}$ does not change during iterations. In looping to step 2 (scenario S2), both animal and SNP effects are recomputed. Whether

scenario S1 is sufficient as opposed to scenario S2 and how many iterations are necessary is not clear and needs to be determined experimentally. In particular, scenario S1 is applicable to multiple-trait models where the relationship matrix needs to be identical for all traits.

(v) *Computations*

Computations with ssGBLUP involved program BLUPF90 (Misztal *et al.*, 2002) modified for genomic analyses (Aguilar *et al.*, 2010), and used simulated parameters. Comparisons involved BayesB procedure as implemented in the GenSel package (Habier *et al.*, 2010). These procedures used the model:

$$\tilde{\mathbf{y}} = \mathbf{1}\mu + \mathbf{Z_u u} + \mathbf{e} \qquad (11)$$

where $\tilde{\mathbf{y}}$ is a dependent variable (DV) for genotyped animals, with options being non-weighted deregressed proofs (DP) or weighted DP (Stranden & Garrick, 2009). For non-weighted DP, all weights were assumed equal to each other being 1; for weighted DP, the weight for *i*-th individual was calculated as $w_i = \frac{(1-h^2)}{[c+(1-r_i^2)/r_i^2]h^2}$ based on equation (10) in Garrick *et al.* (2009), where *c* is fraction of the genetic variance not accounted for by SNPs, and was assumed to be 0.1 (Ostersen *et al*., 2011); h^2 is the heritability, and r_i^2 is reliability of DP for the *i*-th individual. Moreover, **1** is a vector of all ones; μ is the overall mean; $\mathbf{Z_u}$ is a matrix relating SNP marker effects to phenotypic information; **u** is a vector of SNP marker effects; **e** is a vector of residuals distributed as $N(0, \mathbf{D_b}\sigma_e^2)$, where $\mathbf{D_b}$ is a vector of weights as in Stranden & Garrick (2009). For BayesB, marker effects were assumed to be distributed as $u_j \sim N(0, \sigma_{u_j}^2)$, where $\sigma_{u_j}^2$ is the variance of the *j*-th SNP, and proportion of SNPs with no effects ($\sigma_{u_j}^2 = 0$) was set to 90%. As for ssGBLUP, the total genetic variance of BayesB methods was equal to the simulated value

of 0.5. Priors for variances of SNP effects and residuals followed a scaled inverse Chi-square distribution with degrees of freedom 4 and 10, respectively. The Monte Carlo Markov Chain was run for 100,000 iterations (first 10,000 rounds were discarded as burn-in) with Gibbs sampling, with 100 of Metropolis-Hastings sampling within each Gibbs sampling cycle. Estimates of GEBVs and SNP effects were based on the posterior means according to the remaining 90,000 iterations. Accuracies of genotyped animals were defined as correlations between true breeding values (TBVs) and GEBVs. Accuracy of GWAS was determined by correlations of QTL effects with the sum of *m* SNP solutions adjacent to each QTL, where *m* varied from 1 to 40. We did not attempt to declare detection thresholds, or *p*-values, because they are difficult to define and compare to classical frequentist test of hypothesis, in the context of shrunken or Bayesian estimators, as is the case here (Servin & Stephens, 2007; Wakefield, 2009).

For comparisons, SNP solutions were also estimated by CGWAS using a "Snappy" approach implemented in WOMBAT (Meyer & Tier, 2012). When CGWAS analyses are repeated for a large number of SNPs, the computing time can be large, especially for large SNP panels. In "Snappy", matrices common to all SNPs are precomputed, greatly reducing the computation time for the complete scan.

RESULTS AND DISCUSSION

Accuracy of estimated breeding values

Estimated breeding values had been obtained through regular BLUP, ssGBLUP and Bayesian methods (BayesB using non- or weighted DP), respectively. Accuracies of genotyped animals are shown in Table 3.2, and defined as correlations between TBVs and estimated breeding values: EBVs for regular BLUP, and GEBVs for other approaches. Accuracies of

ssGBLUP ranged from 0.87 (0.02) to 0.89 (0.01) depending on iterations, and they were always higher than EBVs for BLUP. For S1, accuracies of GEBVs remained 0.87 (0.01). This result occurred because GEBVs were not recomputed (only SNP effects). For S2, however, the accuracy increased to 0.89 (0.01) by the second round, then dropped to 0.88 (0.01 or 0.02) until the sixth round, and then dropped to 0.87 (0.02). The slight decrease of accuracy in the later rounds could be due to excessive weights given to SNPs associated with few QTLs with larger effects, and reduced weights for numerous QTLs with smaller effects.

For BayesB methods, the accuracies of non-weighted DP were 0.88 (0.02) and were the same as result of weighted DP (c = 0.1). As using DP as DV yields more reliable breeding value solutions than using EBVs in genetic evaluation (Ostersen *et al.*, 2011), other types of DV (e.g. phenotypic records and EBVs) were not considered in this study. For both scenarios of using non- and weighted DP as DV, accuracies from BayesB methods were similar to ssGBLUP with slightly larger standard deviations (SDs) across replications. While the Bayesian methods lose accuracies when pseudo-data are used (Vitezica *et al.*, 2011), that loss of accuracy seems to be similar to the loss of accuracy in ssGBLUP by assuming variances of all SNPs are equal. In the work of Vitezica *et al.* (2011), genotyped animals do not have observations of their own, whereas here genotyped animals do have associated phenotypes. Therefore information from related animals added little to EBV accuracies.

Accuracy of QTL estimates

Table 3.3 presents accuracies of ssGBLUP for QTLs defined as correlations between QTL effects and the sum of m adjacent SNP marker effects, where m varied from 1 to 40. SNP effects under scenarios of S1 and S2 were updated iteratively resulting in similar results. For both S1 and S2, and all iterations, accuracies of QTLs increased up from m = 1 to m = 8, and

decreased sharply for $m = 40$. Iterations improved the accuracy of the S1 and S2 options but only for $m = 8$ and $m = 16$. With iteration and subsequent recomputation of SNP weights, small SNP effects were reduced every round while the large effects became even larger. Iteration for new GEBVs (S2) allowed corrections to SNP with small effects. The highest correlation at $m = 1, 2,$ and 4 was after the first iteration. The highest correlation was 0.82 with $m = 8$ and the second iteration. In both S1 and S2, iteration for GEBVs maximized the accuracy of GEBVs given weights. The highest accuracy was achieved by having a combination of weights that minimized estimation errors but reflected the reality that SNPs adjacent to a QTL contribute to estimation of that QTL.

The advantage of S2 over S1 is dependent on the number and distribution of QTL effects. With many QTL effects and relatively equal distribution, assigning differential weights to SNPs does not greatly improve the accuracy of GEBVs, and therefore little is gained by iteration on GEBVs. Greater improvements with S2 are expected when differential weights on SNP improve accuracy to a greater degree. In a separate study (results not reported), the realized accuracy of S2 improved up to the third iteration for some traits, while deteriorating for other traits in subsequent round. Further research may establish an optimum number of rounds for each particular situation.

Relatively lower correlations are not unexpected at low *m*. Zondervan and Cardon (2004) have found that the closest SNP marker is not always the best predictor of its neighboring QTL. There should be an "optimal haplotype length" according to the marker density and extent of linkage disequilibrium in the population (Villumsen *et al.*, 2009). Density of SNP markers and QTLs in the simulated genome were, on average, 0.067 cM and 6.06 cM, respectively. With each QTL distributed approximately every 90 SNP markers, the QTL effects could be best

approximated by the sum of the adjacent 90 SNP effects. However, due to recombination and mutation for 1000 generations, the best haplotype length can be much shorter than expected. In this study, approximations with 8 SNPs were the most accurate, while those with close to 40 or more were not. Decreases of accuracies in later iterations can be explained by excessive weights on larger SNPs in later iterations. A different algorithm to calculate weights of SNP effects, e.g. with a lower bound similar to Sun *et al.* (2011), may improve accuracies in later rounds. The form of constructing weights used here is indeed suboptimal, because it considers that the estimate of the *j*-th SNP effect $\hat{u}_j$ is the true value, whereas in fact it is a regressed value. An optimal procedure would consider the uncertainty in the estimation of SNP effects by expectation-maximization (EM) or by Bayesian procedures (Legarra *et al.*, 2011; Xu, 2010).

Table 3.4 shows the correlation between QTL effects and the sum of *m* adjacent SNP solutions for BayesB using non- or weighted DP, and for CGWAS using non-weighted DP. When BayesB was applied, weighting DP had little effect on the correlations, which most likely was due to the simple population structure in our simulated study and subsequently similar weights for most genotyped animals. Compared to ssGBLUP and iteration 1, the correlations resulting from application of BayesB were smaller for $m \leq 4$ and slightly higher for $m \geq 16$. Although the average correlations using BayesB were the same as, or even slightly better, than when ssGBLUP was used, the SDs calculated over 10 replications were much higher for BayesB than ssGBLUP. Even in the best situation, the SDs were 0.07 for BayesB with m = 16, as compared to 0.02 (or 0.03) for ssGBLUP/S1 (or S2) with m = 8. For other m, SDs ranged from 0.08 to 0.27 for BayesB, and from 0.02 to 0.09 for ssGBLUP. Larger SDs with BayesB could be due to its sampling structure (Gianola *et al.*, 2009), which also made BayesB less robust than ssGBLUP. With CGWAS, the correlations were higher than any other methods with m = 1,

matched ssGBLUP in iteration 1 with $m = 2$, and were lower than the other methods with $m \geq 4$. Due to fitting a single SNP as a fixed effect, CGWAS is best for identifying a single causative SNP, but seems less efficient in identifying regions containing the QTLs. In general, SDs with CGWAS were lower than with BayesB, but higher than with ssGBLUP.

Graphs of SNP solutions and their moving averages

Figure 3.1-3.4 present SNP solutions or their 4-point moving averages for several methods. The graphs of SNP solutions are the least noisy for BayesB, and the most noisy for CGWAS, with ssGBLUP in between. While most SNP solutions in BayesB are set to 0, lack of shrinkage in CGWAS results in solutions with more noise. Solutions from the third iteration of ssGBLUP/S1 were more similar to those of BayesB, as each round of ssGBLUP shrinks smaller solutions. With averaging, graphs from all the methods were more similar, with closest similarity between BayesB and ssGBLUP/S1 in iteration 3, and CGWAS and ssGBLUP in iteration 1. The similarities confirm that for this particular data set, most QTLs cannot be located with a single SNP accurately; however, all of the methods are similar in identifying regions containing large QTLs.

Computing considerations

In terms of computing time, one round of ssGBLUP required about 2 min, a run of BayesB required about 5 h, and a run of WOMBAT only required 13 s. Long running time in BayesB is due to long sampling. The extraordinarily fast run in WOMBAT is due to an ingenious algorithm; in testing, WOMBAT was over 100 times faster than previous CGWAS approaches. However, the timing analyses were not fully comparable. Both BayesB and ssGBLUP are useful for creating prediction equations based on computed SNP effects while CGWAS is only useful for GWAS. Comparisons based on computing times are not complete, as

BayesB and CGWAS require a BLUP run to create DP, but no such step is required with ssGBLUP.

When implemented efficiently, the cost of BayesB is linear for the number of SNPs and the number of subjects (Legarra & Misztal, 2008). As currently implemented, the creation of $\mathbf{G}^{-1}$ in ssGBLUP is linear with respect the number of SNPs and cubic with respect to the number of subjects (Aguilar *et al.*, 2011). With efficient implementation, the time to create $\mathbf{G}^{-1}$ is about 1 min for 7k genotypes and 1 h for 30k genotypes (Aguilar *et al.*, 2011). The ssGBLUP method has a potential of smaller than cubic cost with respect to the number of genotypes with nonsymmetric mixed model equations and PCG iteration (Misztal *et al.*, 2009; Legarra and Ducrocq, 2012).

Additional considerations

In practice, GWAS (as practiced in humans') seeks to find loci strongly associated across "unrelated" individuals. Genomic selection works with closely related populations, and this relation generates strong linkage (disequilibrium) within the sample that cannot be ignored. Because results from the three methods are similar, none of the methods do a particularly good job of distinguishing associations from that due to linkage disequilibrium. Additional analyses are required to determine whether markers with large effects are due to associated loci or to linkage disequilibrium.

For the data sets in this study, in the best-case, ssGBLUP delivered more accurate GEBVs than the best-case BayesB. All the methods delivered similar predictions of QTL effects based on sum of 2-SNP effects. The ssGBLUP/S1 method is still relatively new and can benefit from further refinements. In particular, the refinements would involve more accurate sampling of SNP variances as discussed before, and a determination of the optimum number of rounds in

ssGBLUP/S2 for maximum accuracy of GEBVs and GWAS. Another needed refinement for ssGBLUP is methodology for significance testing. Without such testing, the use of ssGBLUP for GWAS is limited to identifying SNPs or regions of SNPs with very large effects.

In general, ssGBLUP/S1 seems to provide more consistent estimates than either BayesB or CGWAS using DP. The ssGBLUP/S1 method is also much simpler and therefore more robust to run as: 1) no pseudo-data are required, and 2) no sampling is used. Mrode *et al.* (2010) found large differences regarding results and computing time among various implementations of BayesB.

Models used in this study were very simple with a relatively balanced population structure. For complicated models, such as a multi-trait, maternal effect, random regression or reaction norm models, DP are hard or near impossible to create. Even if they can be created, approximations of DP (Vitezica *et al.*, 2011) would reduce accuracy. The performance of ssGBLUP is likely to improve with field data and more complex models with additional refinements.

CONCLUSIONS

The ssGBLUP method can be modified to compute SNP effects and estimate variances of SNP effects. Such modifications allow for increased accuracy of GEBVs and enable GWAS. The main advantage of ssGBLUP for GWAS is the ability to incorporate phenotypes of ungenotyped animals directly in a BLUP-like approach, without computing pseudo-data. Modified ssGBLUP may become the method of choice for GWAS in the case where merely a fraction of the population with phenotypes is genotyped. In which case, the model for analysis is too complex for use of other methods, and pseudo-data, such as deregressed proofs, for use with method

BayesB and CGWAS, cannot be obtained with sufficient accuracy. In addition, ssGBLUP has the advantages of fast computing, robust estimates, and simplicity.

ACKNOWLEDGEMENTS

We acknowledge helpful discussions and pointing to the Zhang et al. (2010) study by R. L. Fernando. We used moving averages of SNP solutions following examples by J. Dekkers. This study was partially funded by the Holstein Association and AFRI grants 2009-65205-05665 and 2010-65205-20366 from the USDA NIFA Animal Genome Program.

REFERENCES

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. *Journal of Dairy Science* **93**, 743-752.

Aguilar, I., Misztal, I., Legarra, A. & Tsuruta, S. (2011). Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *Journal of Animal Breeding and Genetics* **128**, 422-428.

Chen, C. Y., Misztal, I., Aguilar, I., Legarra, A. & Muir, W. M. (2011). Effect of different genomic relationship matrices on accuracy and scale. *Journal of Animal Science* **89**, 2673-2679.

Christensen, O. F. & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution* **42**, 2.

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

Garrick, D. J., Taylor, J. F. & Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution* **41**, 55.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. & Fernando, R. L. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* **183**(1), 347-363.

Goddard, M. E. & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* **10**, 381-391.

Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick, D. J. (2010). Extension of the Bayesian Alphabet for Genomic Selection. *In* "The 9th World Congress on Genetics Applied to Livestock Production", pp. 468. German Society for Animal Science, Leipzig, Germany.

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science* **1973**, 10-41.

Hirschhorn, J. N. & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95-108.

Legarra, A., Aguilar, I. & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science* **92**, 4656-4663.

Legarra, A. & Misztal, I. (2008). Technical note: Computing strategies in genome-wide selection. *Journal of Dairy Science* **91**, 360-366.

Legarra, A., Robert-Granie, C., Croiseau, P., Guillaume, F. & Fritz, S. (2011). Improved Lasso for genomic selection. *Genetics Research* **93**, 77-87.

Legarra, A. & Ducrocq, V. (2012). Computational strategies for national integration of phenotypic, genomic and pedigree data in a single-step BLUP. *Journal of Dairy Science* (submitted).

Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-1829.

Meyer, K. & Tier, B. (2012). "SNP Snappy": A strategy for fast genome-wide association studies fitting a full mixed model. *Genetics* **190**, 275-277.

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., Strabel, T., Auvray, B., Druet, T. & Lee, D. H. (2002). BLUPF90 and related programs (BGF90). *In* "The 7th World Congress Genetitcs Application Livestock Production", pp. 28, Montpellier, France.

Mrode, R., Coffey, M. P., I. Stradén, Meuwissen, T. H. E., Kaam, J. B. C. H. M. v., Kearney, J. F., Berr, D. P. (2010). A comparison of various methods for the computation of genomic breeding values of dairy cattle using software at genomicselection.net. *In* "the 9th World Congress on Genetics Applied to Livestock Production", Leipzig, Germany.

Ostersen, T., Christensen, O. F., Henryon, M., Nielsen, B., Su, G. & Madsen, P. (2011). Degressed EBV as the response variable yield more reliablegenomic predictions thtraditional EBV in pure-bre pigs. *Genetics Selection Evolution* **43**(1), 38.

Sargolzaei, M. & Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics* **25,** 680-681.

Servin, B. & Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* **3**, e114.

Sillanpaa, M. J. (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity* **106**, 511-519.

Stranden, I. & Garrick, D. J. (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of Dairy Science* **92**, 2971-2975.

Sun, X., Fernando, R. L., Garrick, D. J. & Dekkers, J. C. M. (2011). An iterative approach for efficient calculation of breeding values and genome-wide association analysis using weighted genomic BLUP. *Journal of Animal Science* **89** (E-Suppl **2**), e11.

VanRaden, P. M., Van Tassell, C. P., 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. *Journal of Dairy Science* **92**, 16-24.

Villumsen, T. M., Janss, L. & Lund, M. S. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *Journal of Animal Breeding and Genetics* **126**, 3-13.

Visscher, P. M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S., Hill, W. G., Hottenga, J. J., Willemsen, G., Boomsma, D. I., Liu, Y. Z., Deng, H. W., Montgomery, G. W. & Martin, N. G. (2007). Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American Journal of Human Genetics* **81**, 1104-1110.

Vitezica, Z. G., Aguilar, I., Misztal, I. & Legarra, A. (2011). Bias in genomic predictions for populations under selection. *Genetics Research* **93**, 357-366.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genetic Epidemiology* **33**, 79-86.

Xu, S. (2010). An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity* **105**, 483-494.

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.

Zondervan, K. T. & Cardon, L. R. (2004). The complex interplay among factors that influence allelic association. *Nature Reviews Genetics* **5**, 89-100.

Table 3.1. Description of genomic data from simulation

Means (SDs[1])		Chr1[2]	Chr2	Total
SNPs	Number	1552 (22)	1448 (22)	3000
	AvgMAF[3]	0.28 (0.004)	0.28 (0.005)	0. 28 (0.005)
QTLs	Number	16 (2)	14 (2)	30
	AgvEffect[4]	0.15 (0.04)	0.16 (0.04)	0.16 (0.04)

[1] SDs: standard deviations.
[2] Chr1 and Chr2: chromosome 1 and chromosome 2.
[3] Average minor allele frequencies of SNPs.
[4] Average effects of QTLs.

Table 3.2. Correlation between true breeding values from simulation with estimated breeding values (EBVs) and deregressed proofs (DP) from regular BLUP, genomic breeding values (GEBVs) from ssGBLUP and from BayesA and BayesB with different types of dependent variables (DV)

BLUP	EBVs	DP						
	0.81	0.77						
	(0.01)	(0.01)						
ssGBLUP	it1[1]	it2	it3	it4	it5	it6	it7	it8
	0.87	0.89	0.88	0.88	0.88	0.87	0.87	0.87
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
BayesB_DP	NW[2]	c=0.1						
	0.88	0.88						
	(0.02)	(0.02)						

[1] GEBV solutions using ssGBLUP from iteration 1 (it1) to iteration 8 (it8).

[2] Non-weighted deregressed proofs, and weighted deregressed proofs with c = 0.1.

Table 3.3. Average correlations (standard deviations) between QTL effects and sum of cluster of SNP effects using ssGBLUP

S1[1]	1[2]	2	4	8	16	40
it1	0.53 (0.07)	0.68 (0.05)	0.79 (0.03)	0.81 (0.02)	0.80 (0.03)	0.62 (0.08)
it2	0.46 (0.07)	0.66 (0.05)	0.78 (0.02)	0.82 (0.02)	0.81 (0.02)	0.63 (0.08)
it3	0.43 (0.07)	0.64 (0.05)	0.77 (0.02)	0.81 (0.02)	0.80 (0.02)	0.62 (0.08)
it4	0.42 (0.07)	0.63 (0.05)	0.77 (0.02)	0.81 (0.02)	0.80 (0.02)	0.62 (0.08)
it5	0.41 (0.07)	0.63 (0.05)	0.76 (0.02)	0.80 (0.02)	0.79 (0.02)	0.61 (0.08)
it6	0.41 (0.07)	0.62 (0.05)	0.75 (0.02)	0.80 (0.02)	0.79 (0.02)	0.61 (0.07)
it7	0.41 (0.07)	0.62 (0.05)	0.75 (0.02)	0.80 (0.02)	0.79 (0.02)	0.61 (0.07)
it8	0.41 (0.07)	0.62 (0.05)	0.75 (0.02)	0.80 (0.02)	0.79 (0.02)	0.60 (0.07)
S2	1	2	4	8	16	40
it1	0.53 (0.07)	0.68 (0.05)	0.79 (0.03)	0.81 (0.02)	0.80 (0.03)	0.62 (0.08)
it2	0.44 (0.09)	0.65 (0.06)	0.77 (0.03)	0.82 (0.03)	0.81 (0.02)	0.63 (0.06)
it3	0.41 (0.08)	0.62 (0.05)	0.75 (0.03)	0.79 (0.03)	0.79 (0.03)	0.65 (0.06)
it4	0.40 (0.07)	0.61 (0.05)	0.73 (0.03)	0.77 (0.03)	0.78 (0.03)	0.64 (0.06)
it5	0.40 (0.07)	0.60 (0.05)	0.72 (0.04)	0.76 (0.04)	0.77 (0.04)	0.64 (0.06)
it6	0.40 (0.07)	0.60 (0.05)	0.72 (0.04)	0.75 (0.04)	0.76 (0.04)	0.63 (0.06)
it7	0.40 (0.07)	0.60 (0.05)	0.72 (0.04)	0.75 (0.04)	0.76 (0.04)	0.63 (0.06)
it8	0.40 (0.07)	0.60 (0.05)	0.71 (0.04)	0.75 (0.04)	0.76 (0.04)	0.63 (0.06)

[1] S1: update weights for SNP effects but not for GEBVs; S2: update weights for both GEBVs and SNP effects in each iteration.
[2] Number of SNPs (i.e. *m* ranges from 1 to 40) in each cluster.

Table 3.4. Average correlations (standard deviations) between QTL effects and sum of cluster of m SNP effects using BayesB and WOMBAT

Item[1]	BayesB		WOMBAT
	NW[2]	c = 0.1	NW
1[3]	0.48 (0.27)	0.47 (0.25)	0.57 (0.14)
2	0.65 (0.16)	0.64 (0.16)	0.68 (0.11)
4	0.78 (0.11)	0.78 (0.10)	0.73 (0.08)
8	0.82 (0.08)	0.82 (0.08)	0.74 (0.07)
16	0.82 (0.07)	0.83 (0.07)	0.73 (0.05)
40	0.66 (0.21)	0.67 (0.21)	0.63 (0.09)

[1] Deregress proofs (DP) used as dependent variables (DV) in BayesB and classical GWAS using WOMBAT.
[2] Non-weighted DP and weighted DP with c = 0.1.
[3] Number of SNPs (i.e. m ranges from 1 to 40) in each cluster.

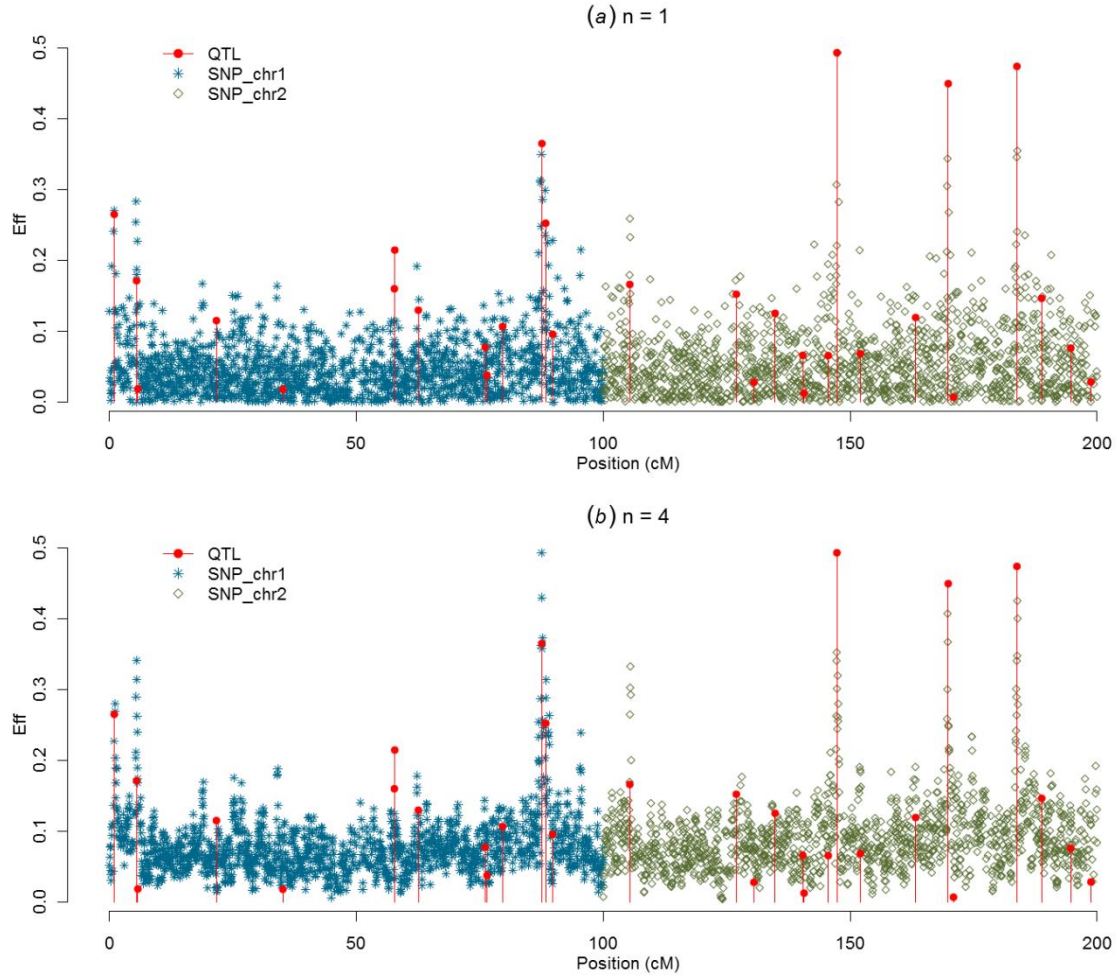


Figure 3.1. SNP solutions and their 4-point moving averages from ssGBLUP/S1 and ssGBLUP/S2 in the first iteration, (a) SNP solutions, and (b) 4-point moving average.

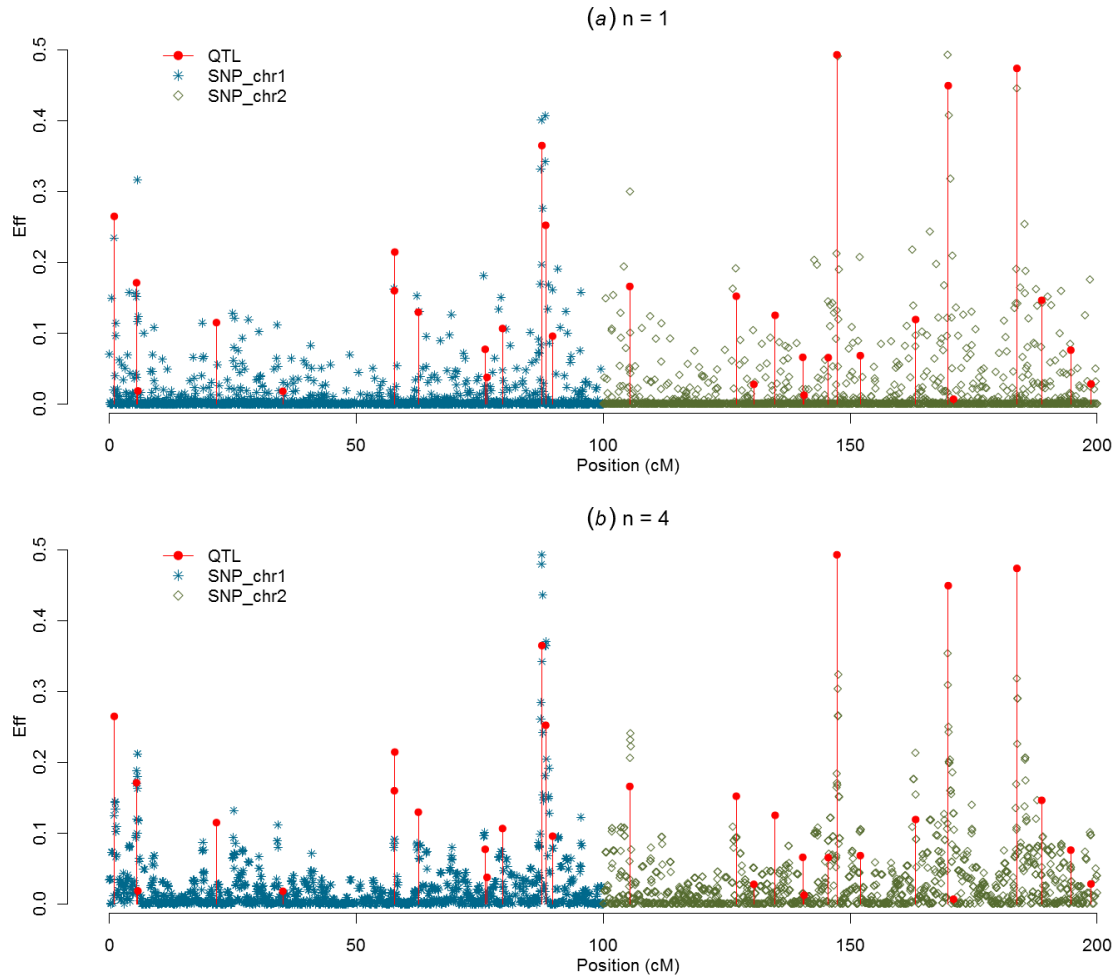


Figure 3.2. SNP solutions and their 4-point moving averages from ssGBLUP/S1 in the third iteration, (*a*) SNP solutions, and (*b*) 4-point moving average.

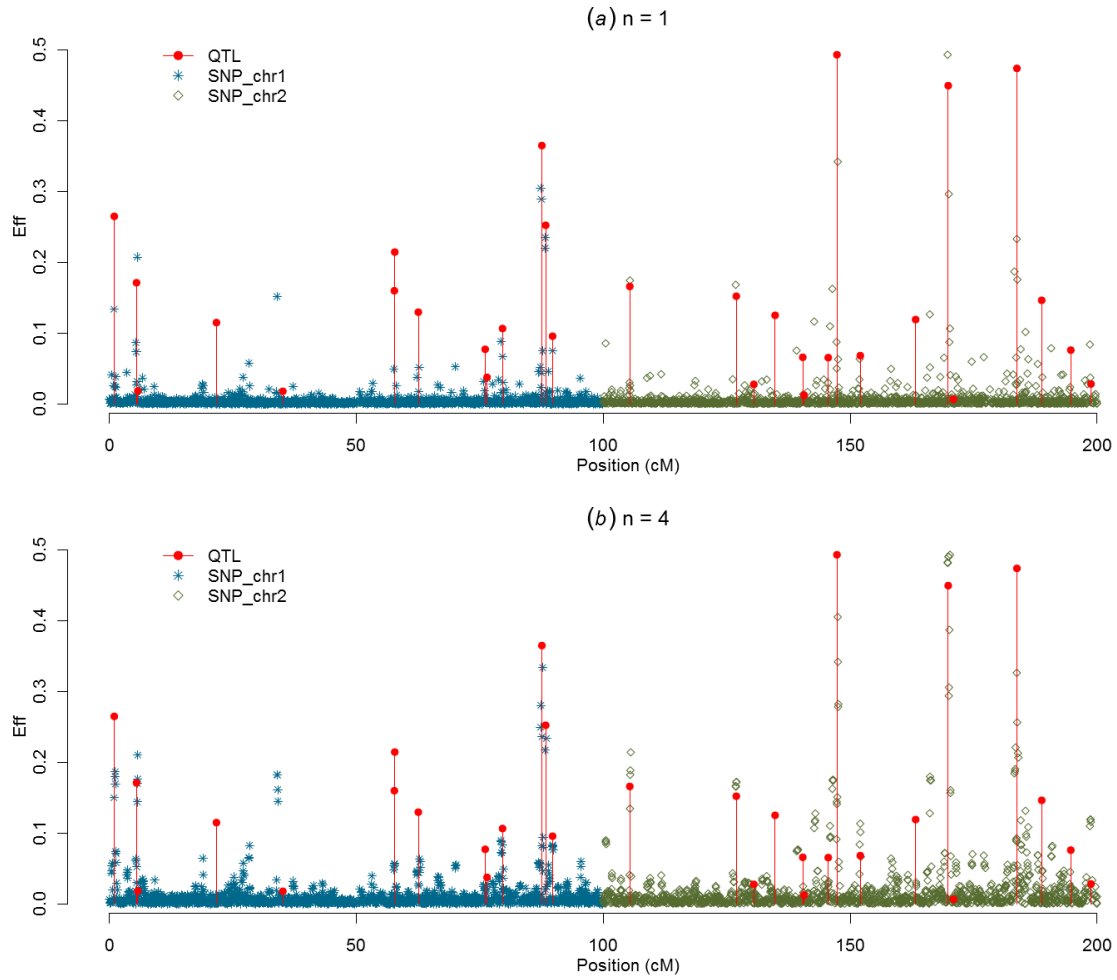


Figure 3.3. SNP solutions and their 4-point moving averages from BayesB with weighted deregressed proofs (c = 0.1) as the dependent variable (DV), (*a*) SNP solutions, and (*b*) 4-point moving average.

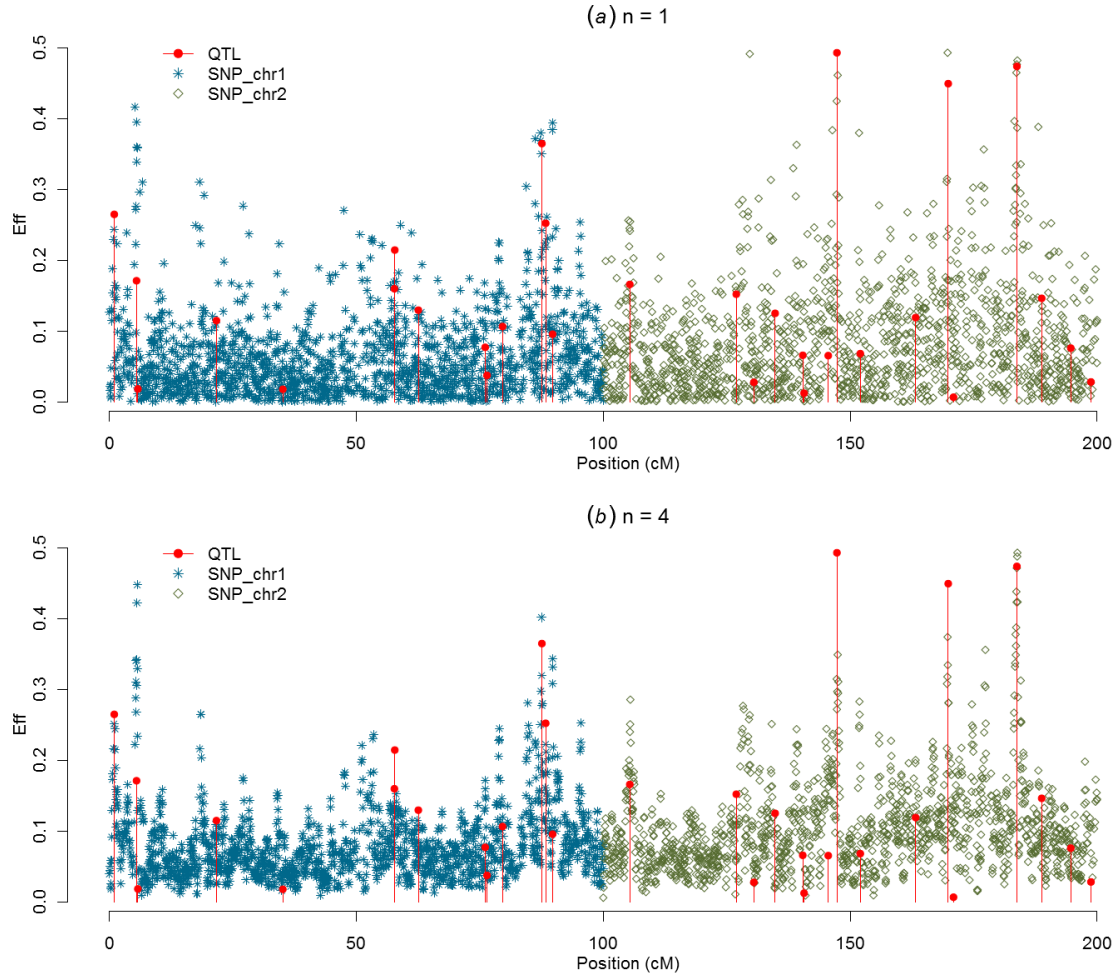


Figure 3.4. SNP solutions and their 4-point moving averages from WOMBAT with non-weighted deregressed proofs as the dependent variable (DV), (*a*) SNP solutions, and (*b*) 4-point moving average.

CHAPTER 4

GENOME-WIDE ASSOCIATION MAPPING INCLUDING PHENOTYPES FROM RELATIVES WITHOUT GENOTYPES FOR 6-WEEK BODY WEIGHT IN BROILER CHICKENS[1]

[1] H. Wang, I. Misztal, W. M. Muir, I. Aguilar, A. Legarra, R. L. Fernando, R. Okimoto, T. Wing, R. Hawken. *To be submitted to Journal of Animal Science.*

ABSTRACT

The purpose of this study was to extend single-step genomic BLUP (ssGBLUP) to genome wide association studies (GWAS) on a production trait in broiler chichens. The ssGBLUP is a procedure that calculates genomic breeding values (GEBVs) based on combined pedigree, genomic and phenotypic information. The procedure achieves these goals by blending traditional pedigree relationships with those derived from genetic markers. In this study, GEBVs were converted to SNP marker effects. Unequal variances for markers were incorporated by deriving weights from SNP solutions, and integrating the calculated weights into a new genomic relationship matrix. Improvements on the SNP weights were obtained iteratively either by only recomputing the SNP effects (S1), or both the SNP effects and GEBVs (S2). Data set included BW at 6 wk (BW6) of 274,776 broiler chickens, of which 4553 were genotyped using a SNP60k chip. After quality control, 40,615 SNP were remained in the analysis. Methods used in the study included S1 and S2 implemented iteratively by ssGBLUP, single marker model implemented by WOMBAT, and BayesB implemented by GenSel. While S1 and S2 of ssGBLUP applied phenotypes directly, the other two methods used deregressed proofs as response variables. The accuracy of prediction for young animals in the latest generation was 0.34 for EBV from BLUP, and 0.44 to 0.52 for GEBVs from S2 according to different iterations. Manhattan plots were carried out by proportion of genetic variances of 20 consecutive SNPs, with similar patterns for all methods. For ssGBLUP, S1 and S2 were thinning with iterations, faster for S2 than for S1. The Manhattan plot for S1 of the third iteration (S1/3) was similar to WOMBAT with 4 out of top 10 regions in common, which explained 3.1% and 2.5% of total genetic variance respectively. For BayesB, the plot was dominated by a single large region explaining 23.1% of the genetic variance, which ranked the same in S1 and ranked sixth in WOMBAT with much

smaller magnitudes. Computing time for S1 and S2 took about 15 min per iteration, while BayesB required about 17 h for 51,000 iterations. The main advantages of ssGBLUP, in addition to computing time, is the ability to include phenotypes from non-genotyped animals, thereby increasing the number of biological samples. More importantly, the method allows GWAS for traits that can only be analyzed using complex models including multi-trait, maternal effects, indirect genetic effects, and random regression.

Key words: body weight, broiler chicken, genome-wide association

INTRODUCTION

To date, GWAS has become a powerful and efficient way to discovery QTL associated with phenotypes and underlying genetic architecture of quantitative traits for broiler chickens (Zhang et al., 2012). Most number of QTL reported in chicken QTLdb were for growth traits (Hu et al., 2013), and pure lines were required for identifying the QTL segregating within breeds (Soller et al., 2006; Fulton 2012). Methods implemented in QTL mapping either fit a single or few SNPs as fixed effect in model (Xie et al., 2012), or analyze all markers simultaneously on whole genome using Bayesian inference (Abasht et al., 2009). An alternative approach explored by Wang et al. (2012) examined GWAS using a method termed ssGBLUP, which increases both power and precision without increasing genotyping costs by taking advantage of phenotypes from both genotyped and ungenotyped subjects. The procedure achieves these goals by blending traditional pedigree relationships with those derived from genetic markers, and by conversion of estimated breeding values (GEBVs) to marker effects and weights.

The application of mixed model approaches allow for both simple and complex analyses that involve multiple traits and partially confounding factors, such as common environment,

epigenetic or maternal environmental effects. Efficiency of the method was examined using simulations and comparisons included ssGBLUP applied directly to phenotypes, BayesB (Habier et al., 2011) and classical GWAS (CGWAS) with deregressed proofs (Meyer and Tier, 2012). The ssGBLUP achieved the highest accuracy for prediction and the highest correlation between simulated QTL and the sum of 8 adjacent SNP effects. It was also faster and simpler to apply than the other methods.

The objectives of this study are to evaluate ssGBLUP for GWAS in BW6 of broiler chickens, and to compare ssGBLUP with other methods regularly used in GWAS.

MATERIAL AND METHOD

Data

Data of broiler chickens were provided by Cobb-Vantress Inc. (Siloam Springs, AR) with BW6 for a dam line across 5 generations (G1, G2 G3, G4, and G5). A brief summary of phenotypic data is presented in Table 4.1. Complete pedigrees were available for all individuals. For generations G1 to G5, 4732 broilers were genotyped with 57,636 SNP markers on a SNP panel across whole genome developed by Groenen et al. (2009), including 8 SNPs from mitochondria. The description of the first 4 generations of the same data set is in Chen et al. (2011).

Quality control (QC) procedures were applied to remove genotyped individuals with pedigree errors and SNP genotypes that were either monomorphic, or displaced segregation distortion (Table 4.2) according to Wiggans et al. (2010) with methodology by Aguilar et al. (2011). After QC, 4553 birds (2205 in G1, 737 in G2, 818 in G3, 793 in G4, and 0 in G5) with 40,615 autosomal SNPs remained, and missing genotypes decreased from 29.8% to 0.32%.

The data set were split into training and validation data sets to estimate accuracy of genetic predictions. The training data set contained 270,661 records for BW6 from G1 to G4, and the validation data set contained 4115 phenotypic records for BW6 in G5, of which none were genotyped. Therefore, the accuracies were not available for CGWAS or BayesB as no genotyped individuals in G5.

Models and computation

The following single-trait model was used for analyses by BLUP and ssGBLUP to determine the increase in accuracy while including SNP information:

$$y = Xb + Wpe + Za + e$$

where **y** is a vector of BW6; **b** is a vector of fixed effects including contemporary group (CG) and sex; **pe** and a are random effects for maternal permanent environmental effect and additive genetic effect; **e** is a vector of residual effects; **W** and **Z** are corresponding incidence matrices for effects of **pe** and **a**, respectively. There were 651 levels for CG and 7929 levels for **pe**.

The distribution of random effects for BLUP was:

$$\text{var}\begin{bmatrix} pe \\ a \\ e \end{bmatrix} = \begin{bmatrix} I_m\sigma^2_{pe} & 0 & 0 \\ 0 & A\sigma^2_a & 0 \\ 0 & 0 & I_n\sigma^2_e \end{bmatrix}$$

where **A** is a numerator relationship matrix based on pedigree for all individuals, $\mathbf{I_m}$ and $\mathbf{I_n}$ are identity matrices with appropriate dimensions. For ssGBLUP, the $\mathbf{A^{-1}}$ matrix was replaced by $\mathbf{H^{-1}}$ according to Aguilar et al. (2010):

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

where $\mathbf{A_{22}}$ is a numerator relationship matrix for genotyped animals and **G** is a genomic relationship matrix constructed as in VanRaden (2008) and scaled for compatibility with $\mathbf{A_{22}}$ as in Chen et al. (2011).

Variance components were estimated using REMLF90 (Misztal et al., 2002) and based on all individuals in the pedigree across 5 generations excluding genomic information. Heritability was calculated as the ratio of estimated genetic variance and total variance, and it was used to estimate the accuracy for both BLUP and S2 of ssGBLUP. Solutions of genetic evaluations for both BLUP and ssGBLUP were obtained using modified BLUP90IOD (Tsuruta et al., 2001; Misztal et al., 2002; Aguilar et al., 2010) based on data through G1 to G4.

For GWAS, SNP effects were obtained from GEBVs of genotyped individuals under two scenarios (S1 and S2) with methodology and algorithm according to Wang et al. (2012). The first scenario S1 involved two steps. In step 1, GEBVs of genotyped animals from ssGBLUP were converted to SNP effects using Equation (9) in Wang et al. (2012), and weight matrix (D) was updated iteratively; in step 2, updated variances of SNP effects were estimated using Equation (10) in Wang et al. (2012) based on squared SNP solutions and allele frequencies. The updated variances were then applied to step 1 and repeated for up to 5 iterations. In the second scenario (S2), an additional step was added preceding step 1 in which the variances were also used to update the genomic relationship matrix (G) in Equation (9) of Wang et al. (2012), and subsequently GEBVs. Note that GEBVs were only updated iteratively in S2, and they were kept the same for all iterations in S1.

Weighted deregressed proofs (DP) were derived from EBV by regular BLUP according to Garrick et al. (2009), where weight for each individual was a function of heritability, accuracy of EBV and fraction of the genetic variance not accounted by SNPs (assumed 0.1) (Ostersen et

al., 2011). The DP were applied to CGWAS using WOMBAT (Meyer and Tier, 2012) and BayesB using GenSel (Habier et al., 2011) based on the following model:

$$y^* = 1\mu + Z_u u + (Z_a a^*) + e$$

where $\mathbf{y}^*$ is a vector of DP for genotyped individuals; **1** is a vector of ones; μ is over all mean; **u** is a fix marker effect for a specific locus (CGWAS), or a vector of all markers (BayesB) where proportion of SNPs with no effect was assumed to be 0.9 (i.e. $\pi = 0.9$); $\mathbf{a}^*$ is background polygenic effect for CGWAS based on pedigree of genotyped individuals, which is not considered in BayesB; **e** is a vector of residual effects; $\mathbf{Z_a}$ is an incidence matrix for effect $\mathbf{a}^*$; and $\mathbf{Z_u}$ is a vector of SNP covariates of each locus for CGWAS, or a matrix of SNP covariates for BayesB. Estimates of genotypic and residual variances from BLUP were used as priors in BayesB, which followed a scaled inverse X^2 with default parameters in GenSel. The use of default parameter $\pi = 0.9$ was due to failure to coverage in BayesCπ (Habier et al., 2011) after 100,000 iterations. Monte Carlo Markov Chain was run for 51,000 rounds with Gibbs samples, of which first 1000 rounds were discarded as burn-in. Within each Gibbs sample cycle, Metropolis-Hastings samples were run for 10 iterations. As WOMBAT and GenSel were not capable to handle missing genotypes, missing codes in SNP file had been replaced by average values.

Manhattan plots based on single-SNP were noisy because of high ratio of the number of SNPs to the number of genotyped individuals. Subsequent Manhattan plots were based on proportion of total variance of GEBVs explained by each non-overlapping window consisting of 20 SNPs. Ranks of top 10 windows that were able to represent most genetic variance were compared among different approaches.

Additionally, prediction accuracies of G5 were compared for EBV from BLUP and GEBVs from ssGBLUP, which were computed by ratio of predictive ability and square root of heritability according to Chen et al. (2010) and Legarra et al. (2008). For CGWAS or BayesB, GEBVs were calculated and defined as the sum of estimated SNP effects for each geneotyped individual through G1 to G4. Accuracies were not available for young animals as no individuals were genotyped in G5.

RESULTS AND DISCUSSION

Genetic evaluation

Variance components of $\hat{\sigma}^2_{pe} = 0.20$, $\hat{\sigma}^2_a = 1.14$ and $\hat{\sigma}^2_e = 3.88$ were calculated from regular BLUP based on all individuals in data set, with heritability estimate of 0.22. Solutions of GEBVs for the first iteration obtained from S1 and S2 of ssGBLUP (S1/1 and S2/1) were equivalent to each other. Table 4.3 presents correlations of (G)EBVs for genotyped individuals. The highest correlations (0.96) between BayesB and ssGBLUP are in S1 (S2/1) and S2/2, where both methods estimated SNP effects on whole genome simultaneously. The correlations between SNP solutions from CGWAS and other methods are lower, which are smaller than 0.80. As SNP effects are calculated in CGWAS individually, estimates for strongly linked SNP are similar, which is likely to cause problem of double counting. Correlations between BayesB and ssGBLUP declined after 2 iterations, implying lower accuracy of GEBVs in later iterations due to over-weighting. This is also confirmed in Table 4.4, which presents accuracy of GEBVs from ssGBLUP as a function of iteration number. The highest accuracy is in S2/2 and S2/3, with declines for later iterations. As genotypes of the latest generation G**5** were not available, accuracies were not available for CGWAS nor BayesB. The results is consistent with Wang et al.

(2012), in which they found the highest accuracy in S2/2 based on a simulated data set. Moreover, Sun et al. (2012) added a constant in the equation to calculate SNP variance, mimicking the structure of such formulas in REML. In our studies involving a constant (results not shown), the accuracy did not decline nor higher than the result of S2/2 from original formula after several iterations. However, adding a constant made identification of top QTL more difficult (Sun et al., 2011).

QTL mapping

Figure 4.1-4.4 show plots of window variances based on different methods. Chromosomes were differentiated by different shades. Windows were neither overlapping nor repetitive, including 20 contiguous SNPs on chromosome. In total, there were 2031 regions, and average length of each region was 0.45 Mbp.

Figure 4.1 shows plots by S1/1, S1/3 and S1/5 from ssGBLUP, which derives weights and solely iterates on SNPs. On one hand, as the iterating progress, plots are less noisy, and peaks indicating largest regions become more distinct. On the other hand, reranking of top few segments arises during such iterationg procedure. In the study by Sun et al. (2011), the number of iteration similar to S1 provided the best identification of top QTL.

Figure 4.2 shows plots by S2/1, S2/3 and S2/5 from ssGBLUP iterating on both SNPs and GEBVs. Please note that the plots for S1/1 and S2/1 are identical. Compared with S1, "thinning" in S2 is more rapid. The plot of S2/3 clearly points to many distinct regions, while the plot in S1/3 seems less so. Since the accuracy of GEBVs peaks at S2/2 to S2/3 and declines afterwards, possibly only the plots of S2/2 and S2/3 are of interest.

Table 4.5 shows the change of ranking of top 10 regions during iterations within ssGBLUP, which is based on S1 (S2/1). Among iterations, the change of rank is minimal within

S1. Even in the fifth iteration, the 36th region ranks originally as forth in S1/1. This could be still considered as small reranking as the total number of regions is 2031. For S2, however, the reranking is minimal at the second iteration (not shown) based on S2/1, but is much stronger in later iterations.

Figure 4.3 shows the plot from CGWAS, where more peaks are found than from ssGBLUP. However, the two largest regions are the same as in S1 through the first to fifth iteration . The presence of many more peaks in CGWAS than other methods is most likely due to lack of removal of strongly linked regions that are false positive (Shen et al., 2013).

Figure 4.4 shows the graph of regions using BayesB. The plot is dominated by a very large region, with all the other regions representing much smaller variances $\leq 2.5\%$. Methods like BayesB are strongly influenced by priors (van Hulzen et al., 2012), and particular by the percentage of SNPs assumed to have null effect (π). Studies on the number of genes influencing a quantitative effect quote the number of < 500 (Otto et al., 2000; Hayes and Goddard, 2001). Here, we assume that 10% of all SNPs (> 4000) have effects. However, each SNP effect partially accounts for the effect of a single QTL. This was mentioned in (Vinkhuyzen et al., 2012) and indirectly confirmed by Daetwyler et al. (2012), where fitting single chromosome in genomic evaluation brought 86% of all gains from using SNPs on all 26 chromosomes.

Table 4.6 shows chromosomal positions and fraction of variances explained by top 10 regions of the 4 methods: CGWAS, BayesB, S1 and S2 of ssGBLUP. For S1/3 and BayesB, regions that represented the largest genetic variance (2.5% and 23. 1%, respectively) were identical on chromosome 27. Extending to top 10 regions in S1/3, there were 2, 4 and 6 common regions with S2/3, CGWAS, and BayesB. However for top 10 regions in BayesB, there were still 6 remained in S1/3, 3 in CGWAS, and only 1 in CGWAS. For CGWAS, the first regions was on

chromosome 6 explaining 3.1% of total additive variance. Of these top 10 regions, there were 4 in common with S1/ 3, 3 with BayesB, and only 2 with S2/3.

In general, the rankings of top 10 regions were similar between ssGBLUP (S1 and S2) and BayesB, with the largest reranking < 107 regions. The reranking in CGWAS was larger compared with the other methods. Additionally, the fraction of explained variance varied greatly among methods. For example, the top region on chromosome 27 that explained 23.1% of the variance in BayesB, 2.5% in S1/3, 5.6% in S2/3, and only 0.8% in CGWAS. There may be several reasons for these differences. First, as mentioned earlier, CGWAS estimates each SNP effect individually, allowing for multiple overlap of several strongly linked regions. Also, CGWAS as implemented by WOMBAT does not take into account relationships among all subjects but only for genotyped individuals, which might lead to detection of spurious associations due to incompleteness (Kang et al., 2009). BayesB is dependent on the choice of parameters and accuracy of dregeression (Garrick et al., 2009; van Hulzen et al., 2012). While S1 or S2 include all available relationships and procedure of deregression is not necessary, the number of rounds is dictated by heuristics at this time. Zeng et al. (2011) and Wang et al. (2012) showed a few methods for GWAS using simulated data sets, and both indicated that all methods were able to identify the same top few regions.

Considerations

Windows were defined with fixed numbers of SNPs (i.e. 20), which might not match every pattern of haplotype blocks. Thus over- or under-estimation of window variances were possible. Moreoer, window variances were calculated based on effect of SNP on each locus, which probably contains error part for estimation, and also brought more variation in results (e.g.

S2 of ssGBLUP). To reduce the noise in estimation process, sliding average values for SNP windows instead of point estimation could be applied.

From simulation with known QTL, very similar results were obtained using all methods (Wang et al., 2012). Due to more complicated data structure in pedigree and genome, similarity decreased for field data, especially for CGWAS with other methods. For BayesB method, different priors affect results greatly (van Hulzen et al., 2012). In this research, $\pi = 0.9$ was applied, as estimation of π did not converge within 100,000 iterations through BayesCπ, which might impact on the accuracy of locating QTL as the appropriate parameter was unobtainable.

It seems that every methodology for GWAS has a weakness. The S1 method of ssGBLUP seems a more useful methodology compared with CGWAS and BayesB when large number of phenotyped subjects are not genotyped, and obtaining deregressed proofs is difficult or impossible. A special weakness of S1 is inability at this time to determine the significance level. Common thoughts are a permutation test (Churchill and Doerge, 1994), or normalizing each SNP solution to a t-like statistic (McClure et al., 2012), which could be difficult to apply to a region including multiple SNPs. However, if the goal of GWAS is identification of most important regions, the significance level for each SNP may not be important, especially given that they are very like to have strongly biases in CGWAS. Future research may determine the level of significance in S1 or S2, e.g., following ideas by Garcia-Cortes and Sorensen (2001).

Computing time

In this study, BayesB and CGWAS required running a regular BLUP, computing accuracies, and creating deregressed proofs. Omitting those procedures, GenSel took 17 h 13 min and WOMBAT took ~6 min. The CGWAS using methods other than WOMBAT would be much slower. The ssGBLUP could be applied directly, and took about 15 min per iteration.

CONCLUSION

This study compares genomic evaluation and association results between different methods: ssGBLUP, CGWAS, and BayesB. There is no evidence for superiority of method choice, but similarity between BayesB and S1 had been shown in various aspects. Advantages of using ssGBLUP includes: 1) no pseudo values are required, 2) complex modeling and multiple-traits are possible, and 3) computing is fast and implementation is simple.

REFERENCES

Abasht, B., E. Sandford, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, A. Hassen, D. Habier, R. L. Fernando, J. C. Dekkers, and S. J. Lamont. 2009. Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations. BMC Genomics 10 (Suppl. 2):S2.

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

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128:422-428.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci. 89:2673-2679.

Churchill, G. A., and R. W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963-971.

Daetwyler, H. D., K. E. Kemper, J. H. van der Werf, and B. J. Hayes. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. J. Anim. Sci. 90:3375-3384.

Garcia-Cortes, L. A., and D. Sorensen. 2001. Alternative implementations of Monte Carlo EM algorithms for likelihood inferences. Genet. Sel. Evol. 33:443-452.

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. 41:55.

Groenen, M. A., P. Wahlberg, M. Foglio, H. H. Cheng, H. J. Megens, R. P. Crooijmans, F. Besnier, M. Lathrop, W. M. Muir, G. K. Wong, I. Gut, and L. Andersson. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 19:510-519.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Hayes, B., and M. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33:209-229.

Hu, Z. L., C. A. Park, X. L. Wu, and J. M. Reecy. 2013. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. 41:D871-879.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42:348-354.

McClure, M. C., H. R. Ramey, M. M. Rolf, S. D. McKay, J. E. Decker, R. H. Chapple, J. W. Kim, T. M. Taxis, R. L. Weaber, R. D. Schnabel, and J. F. Taylor. 2012. Genome-wide

association analysis for quantitative trait loci influencing Warner-Bratzler shear force in five taurine cattle breeds. Anim. Genet. 43:662-673.

Meyer, K., and B. Tier. 2012. "SNP Snappy": a strategy for fast genome-wide association studies fitting a full mixed model. Genetics 190:275-277.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. BLUPF90 and related programs (BGF90). Proc. 7th World Congr. Genet. Appl. Livest. Prod. Montpellier, France. Commun. No. 28-07.

Ostersen, T., O. F. Christensen, M. Henryon, B. Nielsen, G. Su, and P. Madsen. 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genet. Sel. Evol. 43:38.

Otto, S. P., and C. D. Jones. 2000. Detecting the undetected: estimating the total number of loci underlying a quantitative trait. Genetics 156:2093-2107.

Shen, X., M. Alam, F. Fikse, and L. Ronnegard. 2013. A novel generalized ridge regression method for quantitative genetics. Genetics (in press).

Soller, M., S. Weigend, M. N. Romanov, J. C. Dekkers, and S. J. Lamont. 2006. Strategies to assess structural variation in the chicken genome and its associations with biodiversity and biological performance. Poult. Sci. 85:2061-2078.

Sun, X., R. L. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2011. An iterative approach for efficient calculation of breeding values and genome-wide association analysis using weighted genomic BLUP. J. Anim. Sci. 89 (E-Suppl. 2):28. (Abstr.)

Sun, X., L. Qu, D. J. Garrick, J. C. M. Dekkers, and R. L. Fernando. 2012. A Fast EM Algorithm for BayesA-Like Prediction of Genomic Breeding Values. PLoS ONE 7:e49157.

Tsuruta, S., I. Misztal, and I. Stranden. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. J. Anim. Sci. 79:1166-1172.

van Hulzen, K. J., G. C. Schopen, J. A. van Arendonk, M. Nielen, A. P. Koets, C. Schrooten, and H. C. Heuven. 2012. Genome-wide association study to identify chromosomal regions associated with antibody response to Mycobacterium avium subspecies paratuberculosis in milk of Dutch Holstein-Friesians. J. Dairy Sci. 95:2740-2748.

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

Vinkhuyzen, A. A., N. L. Pedersen, J. Yang, S. H. Lee, P. K. Magnusson, W. G. Iacono, M. McGue, P. A. Madden, A. C. Heath, M. Luciano, A. Payton, M. Horan, W. Ollier, N. Pendleton, I. J. Deary, G. W. Montgomery, N. G. Martin, P. M. Visscher, and N. R. Wray. 2012. Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. Transl Psychiatry 2:e102.

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

Wiggans, G. R., P. M. VanRaden, L. R. Bacheller, M. E. Tooker, J. L. Hutchison, T. A. Cooper, and T. S. Sonstegard. 2010. Selection and management of DNA markers for use in genomic evaluation. J. Dairy Sci. 93:2287-2292.

Xie, L., C. Luo, C. Zhang, R. Zhang, J. Tang, Q. Nie, L. Ma, X. Hu, N. Li, Y. Da, and X. Zhang. 2012. Genome-wide association study identified a narrow chromosome 1 region associated with chicken growth traits. PLoS One 7:e30910.

Zeng, J., M. Pszczola, A. Wolc, T. Strabel, R. L. Fernando, D. J. Garrick, and J. C. Dekkers. 2012. Genomic breeding value prediction and QTL mapping of QTLMAS2011 data using Bayesian and GBLUP methods. BMC Proc. 6 (Suppl. 2):S7.

Table 4.1. Descriptive statistics of phenotypic records for BW6[1] in broiler chickens.

Items	BW6, 100g	
	Male	Female
n	132,292	142,484
Mean±SD	25.79 ± 3.22	22.28 ± 2.43
Total *n*	274,776	
Mean±SD	23.97 ± 3.33	

[1]BW6 = body weight at 6 wk, 100g

Table 4.2. Number of genotyped animals and SNPs by reason for removal after quality control.

Category	Number	Reason
Animals	22	Call rate of < 0.9
	157	Parent–progeny conflict rate of > 2%
SNPs	1799	Call rate of < 0.9
	10,894	Minor allele frequency of < 0.05
	6	Parent–progeny conflict rate of > 10%
	4322	genotypes from mitochondrial genome, unknown chromosomes, or sex chromosomes

Table 4.3. Correlations of EBV obtained from regular BLUP and GEBVs[1] obtained from 3 approaches[2] for genotyped individuals.

Correlation	S2/1[3]	S2/2	S2/3	S2/4	S2/5	BayesB[4]	CGWAS
EBV	0.91	0.90	0.88	0.87	0.85	0.90	0.71
S2/1		0.98	0.96	0.93	0.92	0.96	0.78
S2/2			0.99	0.97	0.96	0.96	0.75
S2/3				0.99	0.98	0.93	0.72
S2/4					1.00	0.91	0.71
S2/5						0.90	0.70
BayesB							0.79

[1]GEBVs = genomic breeding values
[2]Single-step genomic BLUP (ssGBLUP), BayesB, and classical genome wide association (CGWAS)
[3]S2/1 = the first iteration of Scenario 2 (S2) in ssGBLUP, which is equivalent to S1
[4]BayesB with $\pi = 0.9$

Table 4.4. Comparison of accuracies of EBV obtained from regular BLUP and GEBVs[1] from ssGBLUP[2] with 5 iterations.

Methods	Accuracy
EBV	0.34
S2/1[3]	0.44
S2/2	0.52
S2/3	0.52
S2/4	0.51
S2/5	0.50

[1]GEBVs = genomic breeding values
[2]ssGBLUP = single-step genomic BLUP
[3]S2/1 = the first iteration of Scenario 2 (S2) in ssGBLUP, which is equivalent to S1

Table 4.5. Rankings of top 10 regions[1] for 5 iterations in ssGBLUP[2].

S1/1(S2/1)[3]	S1/2	S1/3	S1/4	S1/5	S2/2	S2/3	S2/4	S2/5
1	1	1	1	1	1	1	1	1
2	3	3	3	2	9	110	351	479
3	2	2	2	4	6	62	256	472
4	12	21	32	36	2	29	72	100
5	4	4	4	3	16	8	3	2
6	9	11	14	14	20	233	575	766
7	7	7	10	19	19	57	126	179
8	10	15	21	18	8	31	58	86
9	5	8	9	10	7	21	22	25
10	6	6	6	6	5	16	35	50

[1]Each region consists of 20 SNPs, and in totally there are 2031 regions on whole genome
[2]ssGBLUP = single-step genomic BLUP
[3]S1/1 = the first iteration of Scenario 1 (S1) in ssGBLUP, which is equivalent to S2/1

Table 4.6. Rankings top 10 regions among different methods[1].

CGWAS	chr[2]	gVar (%)[3]	S1/3	gVar (%)	S2/3	gVar (%)	BayesB	gVar (%)
1[4]	6	3.07	2	1.29	62	0.38	2	2.35
2	6	2.9	3	0.91	110	0.26	3	1.89
3	6	1.3	4	0.78	8	0.84	40	0.25
4	6	0.98	360	0.09	810	0.01	322	0.06
5	6	0.79	278	0.11	565	0.02	27	0.32
6	27	0.79	1	2.53	1	5.65	1	23.06
7	6	0.6	668	0.04	1216	<0.01	1646	0
8	7	0.48	314	0.1	927	<0.01	99	0.14
9	12	0.48	855	0.03	925	<0.01	387	0.05
10	4	0.45	274	0.11	903	<0.01	173	0.09
total[5]		**11.84**		**5.99**		**7.16**		**28.21**
BayesB	chr	gVar (%)	S1/3	gVar (%)	S2/3	gVar (%)	CGWAS	gVar (%)
1	27	23.06	1	2.53	1	5.65	6	0.79
2	6	2.35	2	1.29	62	0.38	1	3.07
3	6	1.89	3	0.91	110	0.26	2	2.9
4	11	1.39	15	0.43	31	0.55	279	0.08
5	2	1.03	42	0.28	63	0.38	656	0.04
6	3	1	144	0.16	166	0.18	11	0.43
7	4	0.73	9	0.53	105	0.27	450	0.06
8	5	0.68	6	0.59	16	0.72	423	0.06
9	2	0.59	7	0.56	57	0.39	32	0.29
10	2	0.54	264	0.11	119	0.24	53	0.22
total		**33.26**		**7.39**		**9.02**		**794**
S1/3	chr	gVar (%)	S2/3	gVar (%)	CGWAS	gVar (%)	BayesB	gVar (%)
1	27	2.53	1	5.65	6	0.79	1	23.06
2	6	1.29	62	0.38	1	3.07	2	2.35
3	6	0.91	110	0.26	2	2.9	3	1.89
4	6	0.78	8	0.84	3	1.3	40	0.25
5	10	0.72	54	0.41	59	0.22	93	0.15
6	5	0.59	16	0.72	423	0.06	8	0.68
7	2	0.56	57	0.39	32	0.29	9	0.59
8	1	0.54	21	0.67	76	0.19	23	0.35
9	4	0.53	105	0.27	450	0.06	7	0.73
10	12	0.5	13	0.77	357	0.07	31	0.27
total		**8.95**		**10.36**		**8.95**		**30.32**
S2/3	chr	gVar (%)	S1/3	gVar (%)	CGWAS	gVar (%)	BayesB	gVar (%)
1	27	5.65	1	2.53	6	0.79	1	23.06
2	6	2.06	16	0.43	98	0.16	56	0.2
3	2	1.23	20	0.39	125	0.14	29	0.31
4	3	1.02	19	0.4	26	0.32	11	0.54
5	10	0.95	365	0.08	1063	0.02	77	0.17
6	2	0.92	370	0.08	573	0.05	155	0.1
7	14	0.85	82	0.21	606	0.05	41	0.25
8	6	0.84	4	0.78	3	1.3	40	0.25
9	2	0.83	13	0.45	123	0.14	14	0.41
10	12	0.83	152	0.15	555	0.05	118	0.13
total		**15.18**		**5.50**		**3.02**		**25.42**

[1]The third iteration of both scenarios (S1/3 and S2/3) in single-step genomic BLUP (ssGBLUP), BayesB, and classical genome wide association studies (CGWAS)
[2]chr = chromosome number
[3]gVar(%) = proportion of genetic variance each region consisting of 20 SNPs represents
[4]Rankings of each region
[5]Total = sum of gVar(%) of 10 regions of each method

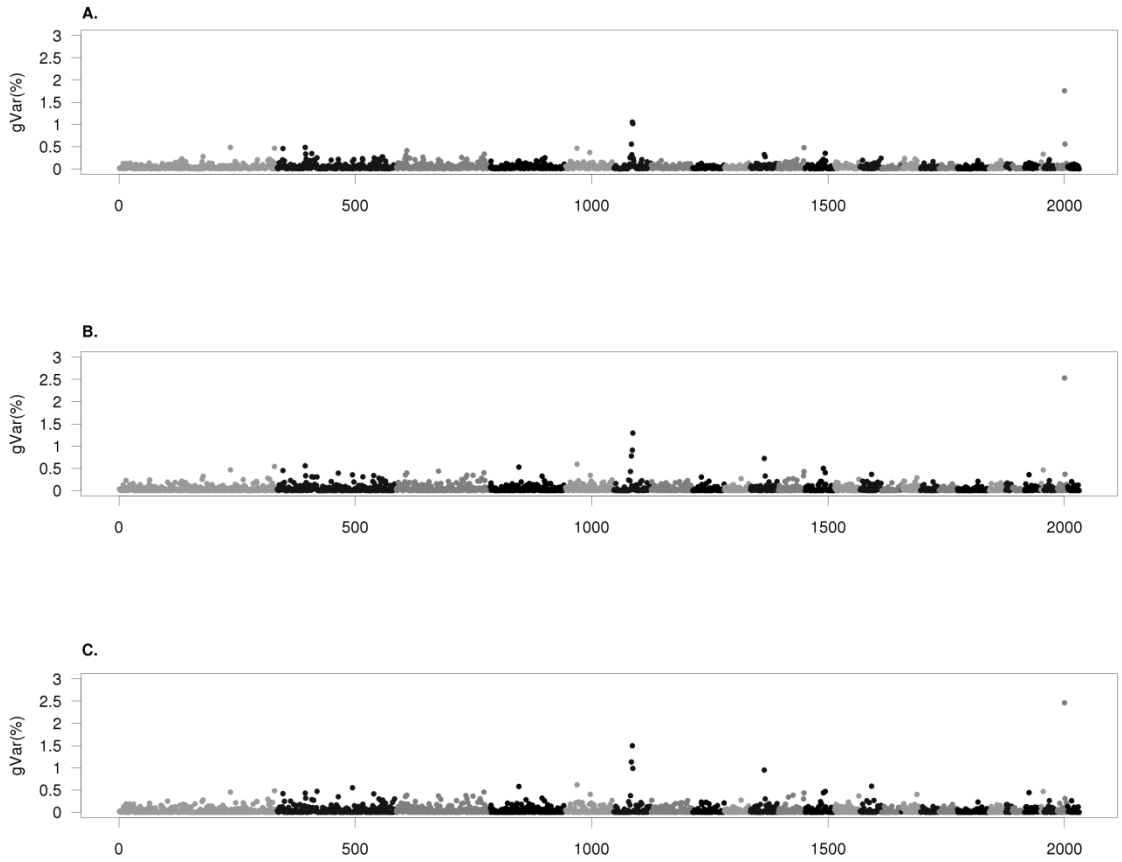


Figure 4.1. Proportion of genetic variance of 20-SNP region under the Senarios 1 (S1) of extend single-step genomic BLUP (ssGBLUP): A. the first iteration (S1/1); B. the third iteration (S1/3); C. the fifth iteration (S1/5). The x-axis represents region location of 20 SNPs. The y-axis represents the proportion of genetic variance of each region.

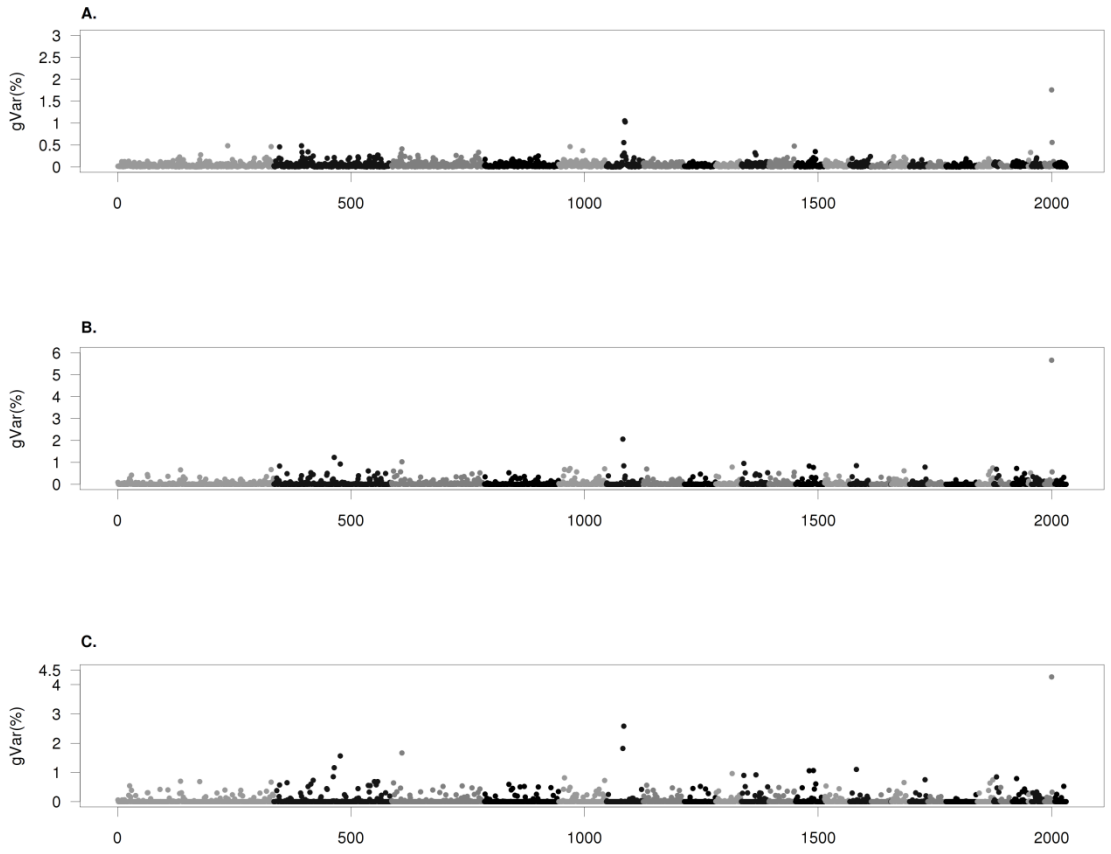


Figure 4.2. Proportion of genetic variance of 20-SNP region under the Senarios 2 (S2) of extend single-step genomic BLUP (ssGBLUP): A. the first iteration (S2/1); B. the third iteration (S2/3); C. the fifth iteration (S2/5). The x-axis represents region location of 20 SNPs. The y-axis represents the proportion of genetic variance of each region.

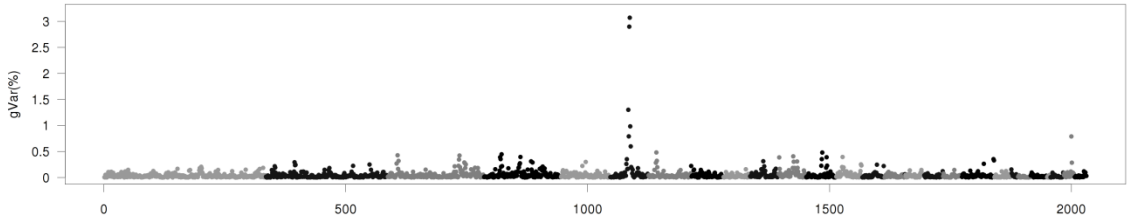


Figure 4.3. Proportion of genetic variance of 20-SNP region using classical genome wide association studies (CGWAS) implemented by WOMBAT: The x-axis represents region location of 20 SNPs. The y-axis represents the proportion of genetic variance of each region.

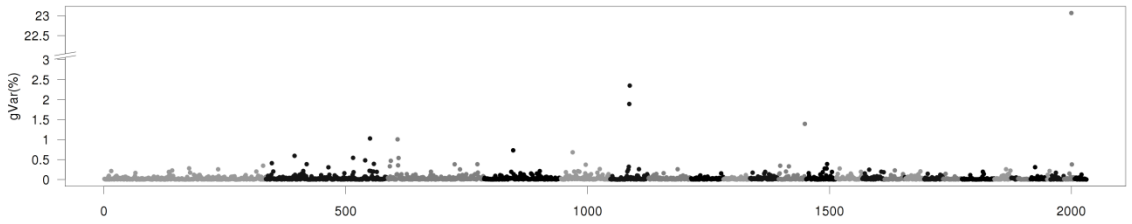


Figure 4.4. Proportion of genetic variance of 20-SNP region using BayesB with $\pi = 0.9$ implemented by GenSel: The x-axis represents region location of 20 SNPs. The y-axis represents the proportion of genetic variance of each region.

CHAPTER 5

GENOME-WIDE ASSOCIATION MAPPING INCLUDING PHENOTYPES FROM RELATIVES WITHOUT GENOTYPES FOR MULTIPLE TRAITS IN BROILER CHICKENS[1]

[1]H. Wang, I. Misztal, W. M. Muir, I. Aguilar, A. Legarra, R. L. Fernando, R. Hawken. *To be submitted to Journal of Animal Science*.

ABSTRACT

The purpose of this study is to extend genome-wide association studies (GWAS) using a single-step method (ssGBLUP) for a multi-trait model in two lines of broiler chickens. Dataset consisted of 2 pure lines (L1 and L2) across 5 generations for 3 traits: body weight at 6 wk (BW6), ultrasound measurement of breast meat (BM), and leg score (LS) coded 1 = no and 2 = yes for leg defect. In total, there were 294,632 and 274,776 individuals in pedigree for L1 and L2, of which 4667 and 4553 were genotyped using a SNP60k panel. After standard quality control, 40,615 SNP markers remained for analyses. For BM, there were ~74% missing phenotypes in both lines. Pedigree, phenotypic and genomic information were combined, and a multi-trait linear model was used through ssGBLUP. Genomic breeding values were calculated for all individuals in pedigree and converted to SNP effects. Variances of markers were calculated from SNP solutions and included as weights in a new genomic relationship matrix. The last step was repeated 5 times. Manhattan plots were constructed as proportion of genetic variance explained by each region consisting of 20 consecutive SNP markers. Several peaks explaining > 1% of the genetic variation were found for BW6, however, peaks for L1 and L2 are on different chromosomes. No strong peaks have been observed for BM and LS, and each region for these traits explained < 1% of total genetic variance. BM and LS seem to follow the infinitesimal model. Different peaks for the two lines for BW6 suggest different selection goals. The ssGBLUP approach allows for simple GWAS with complex models and easy accommodation of information from genotyped animals.

Key words: body weight, breast meat, chicken, GWAS, leg score, ssGBLUP

INTRODUCTION

GWAS enables to map the QTL associated with phenotypes of farm animals and to discover the underlying genetic architecture of quantitative traits of interest (Zhang et al., 2012). For example, over 3400 QTL representing 280 traits have been identified in chicken genome database (Hu et al., 2013). However, most studies for QTL mapping were mostly based on crosses and seldom on pure lines (Abasht et al., 2006). Knowledge of QTLs for traits of interest can be useful for genetic selection for many traits across lines and breeds. QTL effects estimated from crosses may be very different from those on purebreds (Fulton, 2012). Commercial companies usually select a few lines or breeds simultaneously. GWAS could be more useful for selection if QTL for important traits have similar effects across breeds, or at least across lines.

Current methodologies for GWAS include the classical one where each SNP is fit sequentially possibly with pedigree or genomic relationships considered to account for relatedness among subjects (e.g. Meyer and Tier, 2012), and the Bayesian where all SNP are fit simultaneously (e.g. Fernando and Garrick, 2009). Both methodologies analyze one trait at a time. However, breeding objectives for commercial lines are usually multi-trait (Quinton, 2003 and the accuracy of genetic selection increases with a multi-trait model (Jia and Jannink, 2012). As models for GWAS and genetic selection can be treated as equivalent (Goddard et al., 2009; Stranden and Garrick, 2009), a multi-trait GWAS can possibly improve the power and precision of QTL discovery, and help reveal the complexity of the genetic archiecture (Shriner, 2012).

The current methodologies are directly applicable for phenotypes of genotyped subjects. Phenotypes of ungenotyped animals can be utilized indirectly by the use of pseudo-observations for genotyped animals (Garrick et al., 2009). Use of pseudo-information can lead to losses of accuracy and is not applicable to complicated models (Vitezica et al., 2011). Aguilar et al. (2010)

presented a procedure called single-step GBLUP (ssGBLUP) that considers all phenotypic, pedigree and genomic information simultaneously. Wang et al. (2012) extended ssGBLUP to GWAS (ssGWAS). In simulated data, their procedure was as or more accurate than the other procedures while being fast and easy to apply. In analyses of body weight in chicken, estimates of variance for top 20-SNP segments were similar to those by classical GWAS while the same estimates by BayesB were up to 10 times larger. GWAS estimates by the Bayesian methods greatly depend on priors (van Hulzen et al., 2012). ssGWAS allows for GWAS with any model that can be implemented in BLUP.

The first objective of this study was to perform GWAS for 3 traits in 2 lines of broiler chicken by ssGWAS using a multi-trait model. The second objective was to determine whether the most important regions of genome for these traits were overlapping across lines.

MATERIAL AND METHOD

Data

Data of broiler chickens were provided by Cobb-Vantress Inc. (Siloam Springs, AR) with traits of BW at 6 wk (BW6, 100g), breast meat measure based on ultrasound (BM, cm^2), and leg angle (leg score, LS) coded as 1 for acceptable and 2 for not acceptable for a sire line (L1) and dam line (L2) across 5 generations (G1, G2 G3, G4, and G5). A brief summary of phenotypic records has been presented in Table 5.1. Completed pedigrees were available for all individuals. Through G1 to G5, 4940 and 4732 broiler chickens were genotyped with 57,636 SNP markers on a SNP panel across whole genome developed by Groenen et al. (2009) for L1 and L2 respectively, including 8 SNPs from mitochondria. The description of the first 4 generations of the same data set is in Chen et al. (2011b).

Quality control (QC) procedure was applied on both genotyped individuals and SNP genotypes according to Wiggans et al. (2010) with methodology by Aguilar et al. (2011). Genotypes of individuals were removed with call rate < 90% or with parent-progeny conflict for > 2% of SNP. SNP markers would be removed with call rate of < 90%, when caused parent-progeny conflict for > 10% animals, with minor allele frequency of < 5%, or if located on mitochondrial genome, unknown and sex chromosomes. After QC, 4667 and 4553 birds with 42,417 and 40,615 autosomal SNPs remained for 2 Lines.

Model

Multi-trait model was used similar to Chen et al. (2011a) within each line:

$$y_i = X_i b_i + Z_i u_i + (W_i mp_i) + e_i,$$

where i is 1 for BW6, 2 for BM, and 3 for LS; y is the vector of phenotypic records for traits; b is the vector of fixed effects, consisting of contemporary group (house-hatch) and sex; u is the vector of random additive genetic effects, which combines polygenic and genomic values; mp is the vector of random permanent environmental effects, and only available when i = 1 or 2; X, Z, and W are incidence matrices for effects of b, u , and mp; and e is the vector of residuals. For both L1 and L2, about 74% of phenotypic records were missing for BM, and no data were missing for the remaining 2 traits. The (co)variance matrix was assumed as follows within each line:

$$var\begin{bmatrix} u \\ mp \\ e \end{bmatrix} = \begin{bmatrix} A \otimes P & 0 & 0 \\ 0 & I \otimes Q & 0 \\ 0 & 0 & I \otimes R \end{bmatrix},$$

where A is a numerator relationship matrix based pedigree; P, Q, R are covariances matrices for effects of u, mp, and e; and I is identity matrix. In the mixed model, inverse of A (A^{-1}) was

replaced by H^{-1} that incorporated pedigree information with genomic information (Aguilar et al., 2010):

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix},$$

where H is a combined relationships matrix; A_{22} is the fraction of numerator relationship matrix for genotyped individuals; and G is the realized genomic relationship matrix which was scaled to be compatible with A_{22}, and constructed as in Wang et al. (2012):

$$G = \frac{MDM'}{2\sum p_j(1-p_j)},$$

where M is an matrix of covariates for SNP genotypes (0, 1, and 2); p_j is the allele frequency for j-th SNP; and D is a diagonal weight matrix with j-th element calculated as $d_j = p_j(1-p_j)g_j^2$ of which g_j^2 is squared effect of i-th SNP, which is also similar to the way to calculate "trait-specific marker-derived relationship matrix" in Zhang et al. (2010).

RESULTS AND DISCUSSION

Estimates of the variance components and heritability were obtained through multi-trait model based on all available phenotype and pedigree information using average information REML (Table 5.2). Traits of BW6 and BM were genetically correlated in both L1 (0.85) and L2 (0.88), which was similar to the result (0.77) from Le Bihan-Duval et al. (1999). Trait of LS, however, was not genetically correlated with BW6 or BM. The poor correlation between growth and leg disorder traits in poultry had been mentioned in several studies (Kuhlers and McDaniel, 1996; Le Bihan-Duval et al., 1999; Wong-Valle et al., 1993; Zhang et al., 1995).

Manhattan plots involving single SNP were noisy. Therefore, Manhattan plots presented in this study were proportion of genetic variance (%) in windows of 20 consecutive SNP. Figure

5.1-5.2 presents such plots for BW6, BM and LS of L1 and L2. In total, there were 2121 windows for L1 and 2031 windows for L2. As a multi-trait model was used, only "S1" scenario from Wang et al. (2012) was applicable in which results from iteration 3 were shown. Figures 1 presented Manhattan plots for BW6, BM and LS in L1, which have shown very different patterns. For BW6 across iterations, several top peaks were remained while the region explained the largest genetic variance was < 2%. For BM and LS in L1, the top region explained very little (< 1%) proportion of variance of breeding values, and distributions were indistinct and random across iterations. This implies that there might be few medium QTL on genome associated BW6, but the distributions of genes linked with BM and LS in L1 were favoring the infinitesimal model as no large regions were dominating the traits. For genetic architecture, no pleiotropy was found across those 3 traits in L1 according to random peaks in plots. Results of BW6 and BM in L2 were generally similar with L1, except more genetic variances were explained, and regions possible to linked with phenotypes were on different chromosomes. LS in L2 had shown an apparent peak indicating possible QTL. The difference in loci of possible QTL between L1 and L2 indicating divergence of selection goals might affect configurations of the population structure along with linkage disequilibrium and markers segregating within families/lines. This was confirmed before to emphasize limitations in application of QTL across populations (Price et al., 2002; Vikram et al., 2011). Goddard and Hayes (2009) proposed that LD pattern and functional mutations in genes could be population/family specific. Gregersen et al. (2012) found that association between genotypes and phenotypes across 3 Danish pig breeds were rarely overlapping. Similarly, for a research of GWAS on 2 different populations of rice, most significant loci identified were subpopulation-specific (Famoso et al., 2011). This indicated that phenotypes of such trait were controlled by unique alleles within subpopulation.

Moreover, although the magnitudes of genetic variance for all traits and lines were small (< 3%), this was probably due to the extent of shrinkage by using ssGBLUP. For example, when BW6 was analyzed using a single-trait model with BayesB through GenSel, and ssGBLUP, the top region in BayesB explained over 23% of genetic variance, whereas the same region in ssGBLUP also explained the most proportion of variance but only ~2.5% (results not shown).

Table 5.3 showed the proportion of genetic variance and ranks for 3 traits within each line based on top 10 regions in the third iteration, which showed the best prediction and Manhattan plots empirically. For BW6 and LS, top region in L2 explained twice the genetic variance than in L1. All top 10 regions were different across traits and lines, even though BW6 and BM were highly genetically correlated. BM explained very small genetic variation which was < 0.7%, which implied no major QTL associated with breast meat production in these populations, and the inheritance pattern could be highly polygenic.

CONCLUSION

GWAS for large population by a multitrait model is feasible using ssGBLUP. Manhattan plots for windows of 20 SNP indicate several distinct regions for BW6, each explaining < 3% of the total variance. The plots were less distinct for BM and LS where no region explained > 1% of the total variance. The largest regions showed no overlap across traits and lines. for is applicable for GWAS based on multi-trait model and field dataset. Estimates of SNP effects from one line are unlikely to be useful for predicting the performance in the other line.

REFERENCES

Abasht, B., J. C. Dekkers, and S. J. Lamont. 2006. Review of quantitative trait loci identified in the chicken. Poultry science 85: 2079-2096.

Aguilar, I. 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 computation of the genomic relationship matrix and other matrices used in single-step evaluation. J Anim Breed Genet 128: 422-428.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011a. Effect of different genomic relationship matrices on accuracy and scale. Journal of animal science.

Chen, C. Y. et al. 2011b. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. Journal of animal science 89: 23-28.

Famoso, A. N., K. Zhao, R. T. Clark, C. W. Tung, M. H. Wright, C. Bustamante, L. V. Kochian, and S. R. McCouch. 2011. Genetic architecture of aluminum tolerance in rice (Oryza sativa) determined through genome-wide association analysis and QTL mapping. PLoS Genet 7: e1002221.

Fernando, R., and D. Garrick. 2009. GenSel – User manual for a portfolio of genomic selection related analyses, 2nd ed. for version 2.12. Animal Breeding and Genetics, Iowa State University, Ames, IA.

Fulton, J. E. 2012. Genomic selection for poultry breeding. Anim. Front. 2: 30-36.

Garrick D. J., Taylor J. F., Fernando R. L. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55.

Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10: 381-391.

Goddard, M. E., N. R. Wray, K. Verbyla, and P. M. Visscher. 2009. Estimating Effects and Making Predictions from Genome-Wide Marker Data. Statistical Science 24: 517-529.

Gregersen, V. R., L. N. Conley, K. K. Sorensen, B. Guldbrandtsen, I. H. Velander, and C. Bendixen. 2012. Genome-wide association scan and phased haplotype construction for quantitative trait loci affecting boar taint in three pig breeds. BMC Genomics 13: 22.

Groenen, M. A. et al. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome research 19: 510-519.

Hu, Z. L., C. A. Park, X. L. Wu, and J. M. Reecy. 2013. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic acids research 41: D871-879.

Jia, Y., and J. L. Jannink. 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192: 1513-1522.

Kuhlers, D. L., and G. R. McDaniel. 1996. Estimates of heritabilities and genetic correlations between tibial dyschondroplasia expression and body weight at two ages in broilers. Poultry science 75: 959-961.

Le Bihan-Duval, E., N. Millet, and H. Remignon. 1999. Broiler meat quality: effect of selection for increased carcass quality and estimates of genetic parameters. Poultry science 78: 822-826.

Meyer, K., and B. Tier. 2012. "SNP Snappy": a strategy for fast genome-wide association studies fitting a full mixed model. Genetics 190: 275-277.

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

Price, A. H., J. E. Cairns, P. Horton, H. G. Jones, and H. Griffiths. 2002. Linking drought-resistance mechanisms to drought avoidance in upland rice using a QTL approach: progress and new opportunities to integrate stomatal and mesophyll responses. Journal of experimental botany 53: 989-1004.

Quinton, M. 2003. Use of mixed model methodology in poultry breeding: assumptions, limitations and concerns of BLUP-based selection programmes. In: Muir, W. M. and S. E. Aggrey, editor, Poultry genetics, breeding and biotechnology. 1st ed. CABI, Oxfordshire, UK. p 213.

Stranden, I., and D. J. Garrick. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. Journal of dairy science 92: 2971-2975.

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.

van Hulzen, K. J., G. C. Schopen, J. A. van Arendonk, M. Nielen, A. P. Koets, C. Schrooten, and H. C. Heuven. 2012. Genome-wide association study to identify chromosomal regions associated with antibody response to Mycobacterium avium subspecies paratuberculosis in milk of Dutch Holstein-Friesians. J. Dairy Sci. 95:2740-2748.

Vikram, P. et al. 2011. $qDTY_{1.1}$, a major QTL for rice grain yield under reproductive-stage drought stress with a consistent effect in multiple elite genetic backgrounds. BMC genetics 12: 89.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet Res (Camb) 93: 357-366.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics research 94: 73-83.

Wiggans, G. R. et al. 2010. Selection and management of DNA markers for use in genomic evaluation. Journal of dairy science 93: 2287-2292.

Wolc, A. et al. 2012. Genome-wide association analysis and genetic architecture of egg weight and egg uniformity in layer chickens. Animal genetics 43 Suppl 1: 87-96.

Wong-Valle, J., G. R. McDaniel, D. L. Kuhlers, and J. E. Bartels. 1993. Correlated responses to selection for high or low incidence of tibial dyschondroplasia in broilers. Poultry science 72: 1621-1629.

Zhang, H., Z. Wang, S. Wang, and H. Li. 2012. Progress of genome wide association study in domestic animals. Journal of animal science and biotechnology 3: 26.

Zhang, X., G. R. McDaniel, Z. S. Yalcin, and D. L. Kuhlers. 1995. Genetic correlations of tibial dyschondroplasia incidence with carcass traits in broilers. Poultry science 74: 910-915.

Zhang, Z. et al. 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PloS one 5.

Table 5.1. Statistical summary of phenotypic data for 2 lines

Item[1]	L1[2]			L2		
	Male	**Female**	**Total**	**Male**	**Female**	**Total**
BW6, 100g						
No. of records	143,051	151,581	294,632	132,292	142,484	274,776
Mean	99.18	86.51	92.66	90.97	78.62	84.56
SD	17.71	14.20	17.21	11.37	8.57	11.76
BM, cm^2						
No. of records	16,386	58,991	75,377	14,436	56,617	71,053
Mean	48.89	44.79	45.68	46.42	41.18	42.25
SD	7.22	6.97	7.22	5.66	4.82	5.43
LS, 1 and 2						
No. of records	143,051	151,581	294,632	132,292	142,484	274,776
Mean	1.23	1.12	1.17	1.24	1.10	1.17
SD	0.42	0.33	0.37	0.42	0.30	0.37
Animals in pedigree		297,012			277,044	
QC[3]	before	after		before	after	
No. of genotyped animals	4940	4667		4732	4553	
No. of SNPs	57,636	42,417		57,636	40,615	
No. of chromosomes	34	28		34	28	

[1]BW6 = BW at 6 wk; BM = breast meat measurement using ultrasound; LS = leg score
[2]L1 = line 1, and L2 = line 2
[3]QC = quality control

Table 5.2. Genetic (above diagonal) and phenotypic (below diagonal) correlations among traits and estimates of heritability (h^2, bold on diagonal) with each line[1]

L1[2]	BW6	BM	LS
BW6[3]	**0.24**	0.85	0
BM	0.80	**0.25**	0
LS	0.08	0.10	**0.12**
L2	BW6	BM	LS
BW6	**0.22**	0.88	0
BM	0.72	**0.27**	0
LS	0.10	0.12	**0.11**

[1]Estimates of genetic correlations and h^2 were based on multi-trait BLUP for all data through 5 generations
[2]L1 = line 1, and L2 = line 2
[3]BW6 = BW at 6 wk; BM = breast meat measurement using ultrasound; LS = leg score

Table 5.3. Proportion of genetic variances and rankings in the third iteration of 3 traits in 2 lines (L1 and L2)

Trait	L1			L2		
	Region_ID	chr	gVAR	Region_ID	chr	gVAR
BW6	1640	14	1.61%	2000	27	2.80%
	61	1	0.76%	1087	6	1.28%
	1319	8	0.64%	1086	6	1.16%
	151	1	0.63%	1084	6	0.70%
	1726	16	0.63%	1365	10	0.67%
	1049	5	0.62%	394	2	0.60%
	63	1	0.51%	846	4	0.59%
	1318	8	0.49%	969	5	0.59%
	984	5	0.47%	329	1	0.58%
	1109	6	0.47%	347	2	0.51%
BM	1055	5	0.66%	438	2	0.69%
	3	1	0.62%	1490	12	0.57%
	888	4	0.59%	639	3	0.57%
	854	4	0.53%	1389	10	0.54%
	1237	7	0.53%	1662	17	0.54%
	796	3	0.47%	296	1	0.52%
	73	1	0.47%	66	1	0.52%
	1040	5	0.46%	304	1	0.52%
	1422	10	0.45%	966	5	0.51%
	1252	7	0.45%	1104	6	0.50%
LS	1127	6	0.83%	1353	10	1.72%
	1469	11	0.65%	119	1	0.89%
	543	2	0.59%	1494	12	0.88%
	613	3	0.56%	1558	13	0.80%
	1800	18	0.55%	435	2	0.62%
	855	4	0.54%	875	4	0.60%
	1096	6	0.54%	1876	22	0.58%
	1204	7	0.52%	1977	26	0.50%
	775	3	0.48%	1691	17	0.48%
	557	2	0.47%	639	3	0.47%

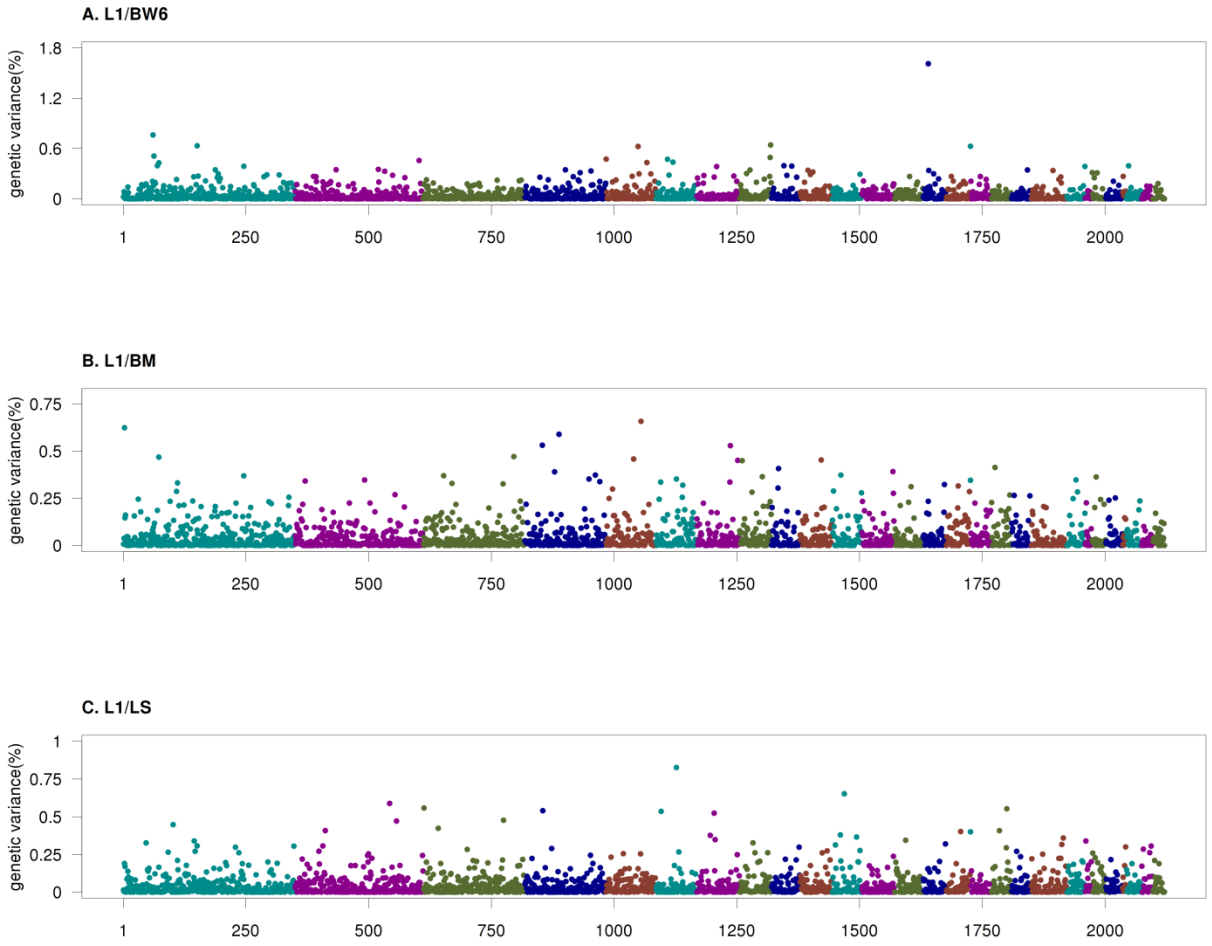


Figure 5.1. Proportion of genetic variance (%) explained by each window of the third iteration for 3 traits in line 1 (L1). A. body weight (BW6); B. breast meat (BM); and C. leg score (LS). X-axis is the window number and chromosomes were differentiated by colors, and y-axis is the proportion of genetic variance (%) each window explains.

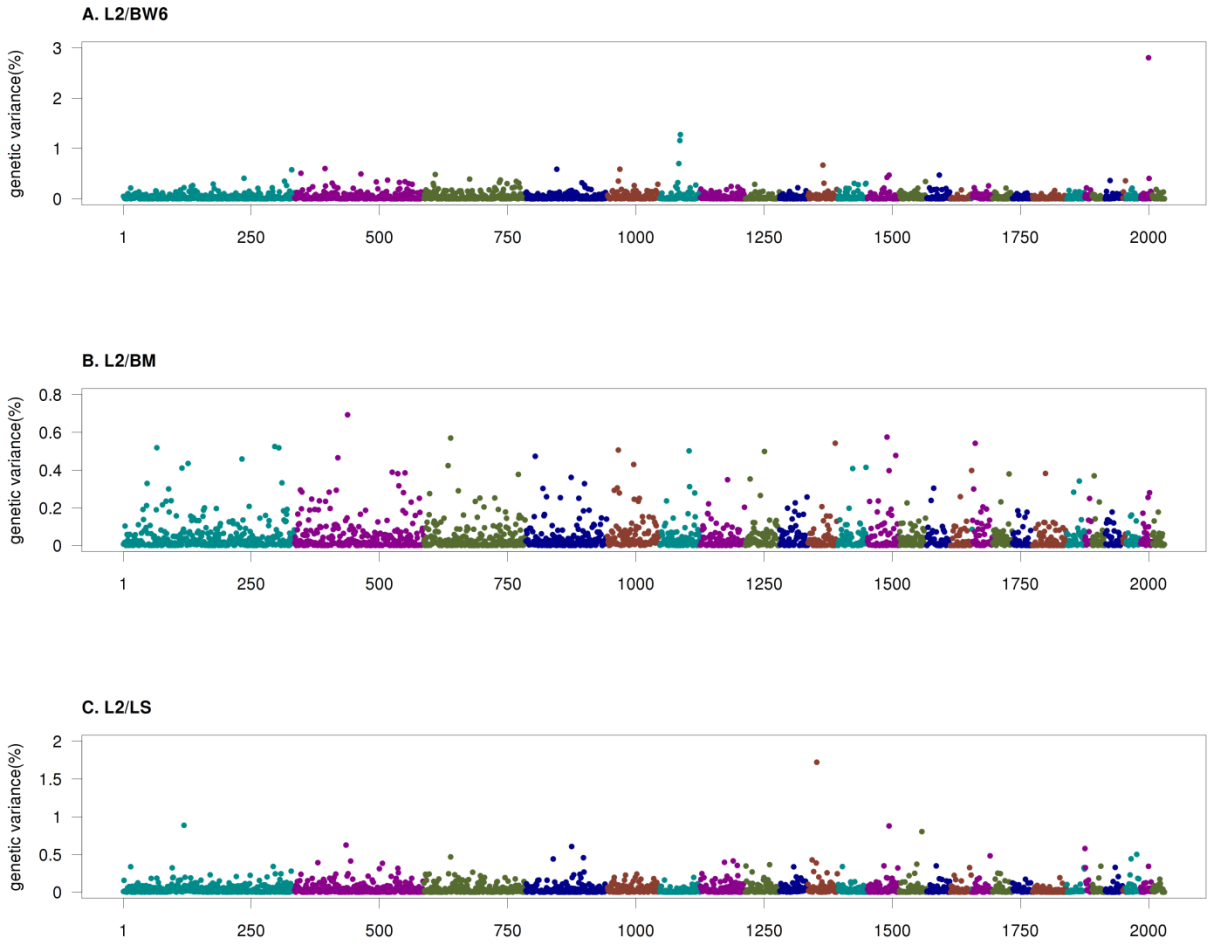


Figure 5.2. Proportion of genetic variance (%) explained by each window of the third iteration for 3 traits in line 2 (L2). A. body weight (BW6); B. breast meat (BM); and C. leg score (LS). X-axis is the window number and chromosomes were differentiated by colors, and y-axis is the proportion of genetic variance (%) each window explains.

CHAPTER 6

DIFFERENCE BETWEEN PEDIGREE- AND GENOMIC-BASED RELATIONSHIPS IN CHICKENS[1]

[1]H. Wang, I. Misztal. *To be submitted to Journal of Animal Science*.

ABSTRACT

As realized relationships would be indicated through SNP genotypes, differences between coefficients of relationships based on genomic and pedigree information should centered at 0 with small variation. This study explored the magnitude and variation of such discrepancy under different levels quality control (QC) and several scenarios of relationship distance for a chicken data. The results indicated large difference existed even if the genomic relationship matrix was scaled, thus QC procedure was required for further tunning. The QC affected mean particularly for diagonal elements, and affect SD to different extents under various situations for off-diagonal and relationships. It also implied that although the large variation could be used as an indicator for pedigree or genotype errors, they were confounding and hardly to distinguish. Furthermore, different ranges of relationship distance were investigated to check the distribution of discrepancy based on 2 matrices. The standard deviation (SD) was larger if two individuals were more closely related, and the degree of correction was directly or indirectly through QC. Therefore, large differences between coefficients of 2 matrices could be used as a diagnostic tool, and QC had been confirmed to be important and necessary before genomic analysese.

Key words: chicken, comparison, genomic relationship matrix, numerator relationship matrix, standard deviation

INTRODUCTION

The pedigree-based relationships such as calculated in a numerator relationship matrix (NRM) are measures of expected relationships (Wright, 1934). The value of such relationships depends on the depth and completeness of pedigree (Cassell et al., 2003; Cole and Franke, 2002).

For animals not connected by pedigree, such relationships are zero although methods exist that assign nonzero relationships based on the unknown parent concept (VanRaden, 2008; Lutaaya , 1999). The genomic relationship matrix (GRM) is the measure of actual proportion of identical by descent (IBD) that is created based on SNP genotypes, and does not depend on the pedigree (Hayes et al., 2009; VanRaden, 2008). However, it depends on the algorithm used to create such a matrix, particularly on gene frequencies for each SNP (Aguilar et al., 2011; Habier et al., 2010). Also, the scale of GRM is somewhat arbitrary. Therefore, algorithms to create GRM include mechanisms to adjust GRM for compatibility with NRM, which include scaling GRM for compatibility with NRM (VanRaden, 2008; Chen et al., 2011; Vitezica et al., 2011), or using appropriate allele frequencies (VanRaden, 2008; Yang et al., 2010). After adjustment procedures, animals unrelated in NRM are still likely to be related in GRM. In particular, differences between coefficients of the two matrices can be large for animals from disconnected populations, where the relationships would be 0 in NRM but could be large in GRM. Values of GRM also depend on QC for SNP information (Forni et al., 2011; Nagamine et al., 2012). Standard editing with SNP data includes calling rate for each SNP, calling rate for animals, inconsistencies in Mendelian sampling, and parent-offspring verification (Wiggans et al., 2010).

Hill and Weir (2011) looked at differences between expected and actual relationships as a consequence of Mendelian sampling and linkage. The differences were relatively small with SD generally < 0.05, although these differences were dependent for different types of relationships. For example, the SD of G and A are 0.036 and 0.039, with mean of 0.498 and 0.5 respectively. If GRM are close to actual relationships, large differences between GRM and NRM would point to pedigree or genotyping error (Simeone et al., 2011). One of the best populations to test differences between NRM and GRM is purebred commercial chicken, because lines are closed

so that differences in gene frequencies among lines are not a factor and pedigrees are accurately recorded for many generations so that the base population can be standardized to the same point in time. Therefore, our objectives are to compare coefficients of NRM and GRM under different levels of QC based on a chicken dataset, to investigate the distribution of such difference, and to compare such distributions to those in Hill and Weir (2011).

MATERIAL AND METHOD

Data were obtained from Cobb-Vantress Inc. (Siloam Springs, AR) with a SNP panel used for analysis was described by Groenen et al. (2011) , including 4940 pureline individuals across 5 generations and genotyped for 57,636 SNPs distributing over 34 chromosomes. Three difference levels of QC had been applied including noQC, WeakQC and StrongQC, where WeakQC and StrongQC were similar except for call rates for SNPs and animals. Summary of editing in Table 6.1 is based on Wiggans et al. (2010). The matrices of NRM and GRM for genotyped individuals were obtained according to Aguilar et al. (2011), where GRM was scaled by default to compatible with NRM. Under each level of QC, several scenarios have been investigated to obtain the magnitude and distribution of difference between GRM and NRM (G−A): diagonal and off-diagonal elements, parent-offspring pairs (PO), full-sib pairs (FS), half-sib pairs (HS), and 3 relationship ranges based on pedigree (i.e. 0.1~0.15, 0.2~0.3, and 0.45~0.55).

RESULT AND DISCUSSION

Table 6.2 shows the range, mean and SD of G−A for diagonal, off-diagonal, PO, FS and HS under 3 levels of QC. Range of diagonal, off-diagonal, PO decreased while QC was more

stringent, particularly for diagonal elements. This had been confirmed and implemented by Simeone et al. (2011) where large values of G−A in diagonals could be used to detect pedigree error (Figure 1.). Distribution of G−A approximately followed normal distribution, with mean was 0 which did not changed significantly according to QC. This was as expected, as the realized and average relationship coefficients should be consistent. Moreover, when QC was more stringent, SD decreased for diagonal and PO coefficients, but not for off-diagonal, FS and HS. This implied that although PO conflicts could be controlled through QC based on Mendelian inheritance, other relationships like FS and HS were too complicated to investigate and barely detectable through adjustment based on PO. Large discrepancy of NRM and GRM indicated possible errors in pedigree or SNP chip, or limited pedigree depth. Another sources of differences between A and G is limited pedigree depth. For example, there is no information on which animals are FS in the base population in NRM but GRM contains such information.

Table 6.3 shows statistics of G−A for 3 different ranges of coefficients based on NRM, which indicated 3 levels of distance of relationships among individual pairs in pedigree. When the relationship was closer, the magnitude of range was larger, and number of pairs in pedigree was less, and the SD was bigger. For example, for noQC, there were about 530k relationship pairs in pedigree for 0.1~0.15 (e.g. great grandparent-great grandoffspring, half uncle-nephew, or first cousins), the number decreased to less than 200k for 0.2~0.3 (e.g. grandparent-grandoffspring, HS, or uncle-nephew), and it decreased to about only 32k pairs when range was 0.45~0.55 (e.g. PO or FS). Again, means was around 0, but always negative, which implied some individual pairs indicated no relationships but recorded as PO/FS, HS, first cousins or other relationships in pedigree. The factors were still confounding as difficulty to identify the resources of errors. The QC showed slight adjustment on all elements of G−A, especially for

0.45~0.55 (Figure 2.), as PO was included in this range. Other relationship distances were controlled indirectly through call rates or other QC criteria. However, the SD for range of 0.45~0.55 was much larger than other ranges even PO errors could be corrected to some extent through QC, which might be due to sample size of 0.45~0.55 was much smaller than other ranges.

CONCLUSION

The difference of coefficients in GRM and NRM should follow a normal distribution with mean about 0 and SD < 0.05. Large discrepancy indicated mistakes in pedigree or genotype information with confounding. More completed pedigree and accurate sequencing procedure were required for discover underlying reasons. Conflicts in PO could be detected through Mendelian inheritance but hardly directly implemented on other relationships. Therefore, QC was necessary and powerful before further genomic analyses.

REFERENCES

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J Anim Breed Genet 128: 422-428.

Cassell, B. G., V. Adamec, and R. E. Pearson. 2003. Effect of incomplete pedigrees on estimates of inbreeding and inbreeding depression for days to first service and summit milk yield in Holsteins and Jerseys. Journal of dairy science 86: 2967-2976.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. Journal of animal science.

Cole, J. B., and D. E. Franke. 2002. PyPedal: A package for pedigree analysis using the Python programming language. J. Dairy Sci. 85(Suppl. 1): 323. (Abstr.).

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 : GSE 43: 1.

Groenen, M. A. et al. 2011. The development and characterization of a 60K SNP chip for chicken. BMC genomics 12: 274.

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics, selection, evolution : GSE 42: 5.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics research 91: 47-60.

Hill, W. G., and B. S. Weir. 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetics research 93: 47-64.

Lutaaya, E., I. Misztal, J. K. Bertrand, and J. W. Mabry. 1999. Inbreeding in populations with incomplete pedigree. J. Anim. Breed. Genet. 116:475–480.

Nagamine, Y. et al. 2012. Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping. PloS one 7: e46501.

Simeone, R., I. Misztal, I. Aguilar, and A. Legarra. 2011. Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. J Anim Breed Genet 128: 386-393.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91: 4414-4423.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genetics research 93: 357-366.

Wiggans, G. R. et al. 2010. Selection and management of DNA markers for use in genomic evaluation. Journal of dairy science 93: 2287-2292.

Wright, S. 1934. The method of path coefficients. Ann. Math. Stat 5: 161-215.

Yang, J. et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature genetics 42: 565-569.

Table 6.1. Number of removed genotypes and individuals due to quality control (QC)[1]

StrongQC[2]	No.[2]	Reason
Animals	67	Call rate of < 0.9
	216	Parent–progeny conflict rate of > 2%
SNPs	6095	Call rate of < 0.9
	12,261	Minor allele frequency of < 0.05
	1	Parent–progeny conflict rate of > 10%
	4322	Genotypes from mitochondrial genome, unknown chromosomes, or sex chromosomes
WeakQC	**No.**	**Reason**
Animals	5	Call rate of < 0.7
	219	Parent–progeny conflict rate of > 2%
SNPs	1510	Call rate of < 0.7
	12,261	Minor allele frequency of < 0.05
	1	Parent–progeny conflict rate of > 10%
	4322	Genotypes from mitochondrial genome, unknown chromosomes, or sex chromosomes

[1] noQC, WeakQC and StrongQC; noQC did not remove any animals or SNPs
[2]Numbers of animals or SNPs removed according to QC

Table 6.2. Statistics of difference of coefficients between genomic- and pedigree-based relationship matrices (G−A) for genotyped individuals under different levels of quality control[1] and degree of relationships[2]

G−A	No.[2]	Min[3]	Max[4]	Mean	SD[5]
NoQC[6]					
Diagonal	4940	-0.527	3.113	1.03E-06	0.083
Off-Diagonal	12,199,330	-0.579	1.150	4.40E-06	0.036
PO	6115	-0.579	0.155	-0.040	0.094
FS	9970	-0.189	0.184	-0.016	0.050
HS	69,154	-0.177	0.164	-0.014	0.041
WeakQC					
Diagonal	4727	-0.520	0.876	1.02E-06	0.062
Off-Diagonal	11,169,901	-0.566	1.028	4.43E-06	0.036
PO	5377	-0.323	0.166	-0.016	0.044
FS	9130	-0.185	0.184	-0.017	0.050
HS	59,930	-0.177	0.164	-0.015	0.040
StrongQC					
Diagonal	4667	-0.180	0.840	8.79E-07	0.048
Off-Diagonal	10,888,111	-0.565	1.021	4.44E-06	0.037
PO	5259	-0.158	0.170	-0.011	0.034
FS	9126	-0.185	0.184	-0.017	0.050
HS	59,870	-0.177	0.164	-0.015	0.040

[1]QC: noQC, WeakQC and StrongQC
[2]Parent offspring pairs (PO), full sib pairs (FS), and half sib pairs (HS)
[3]Numbers of animals or SNPs removed according to QC
[4]Minimum value
[5]Maximum value
[6]Standard deviation

Table 6.3. Statistics of difference of coefficients between genomic- and pedigree-based relationship matrices (G−A) for genotyped individuals under different levels of quality control[1] for 3 ranges of relationship coefficients[2]

G−A	No.[3]	Min[4]	Max[5]	Mean	SD[6]
noQC					
0.1~0.15	522,976	-0.238	0.430	-0.020	0.048
0.2~0.3	199,372	-0.330	0.365	-0.024	0.068
0.45~0.55	32,654	-0.557	0.205	-0.045	0.108
WeakQC					
0.1~0.15	472,212	-0.231	0.376	-0.019	0.047
0.2~0.3	182,392	-0.329	0.355	-0.024	0.064
0.45~0.55	30,320	-0.566	0.204	-0.034	0.095
StrongQC					
0.1~0.15	457,836	-0.235	0.372	-0.019	0.048
0.2~0.3	176,298	-0.329	0.357	-0.022	0.063
0.45~0.55	29,528	-0.565	0.208	-0.030	0.092

[1] QC: noQC, WeakQC and StrongQC
[2]Ranges: 0.1~0.15, .0.2~0.3, and 0.45~0.55
[3]Numbers of animals or SNPs removed according to QC
[4]Minimum value
[5]Maximum value
[6]Standard deviation

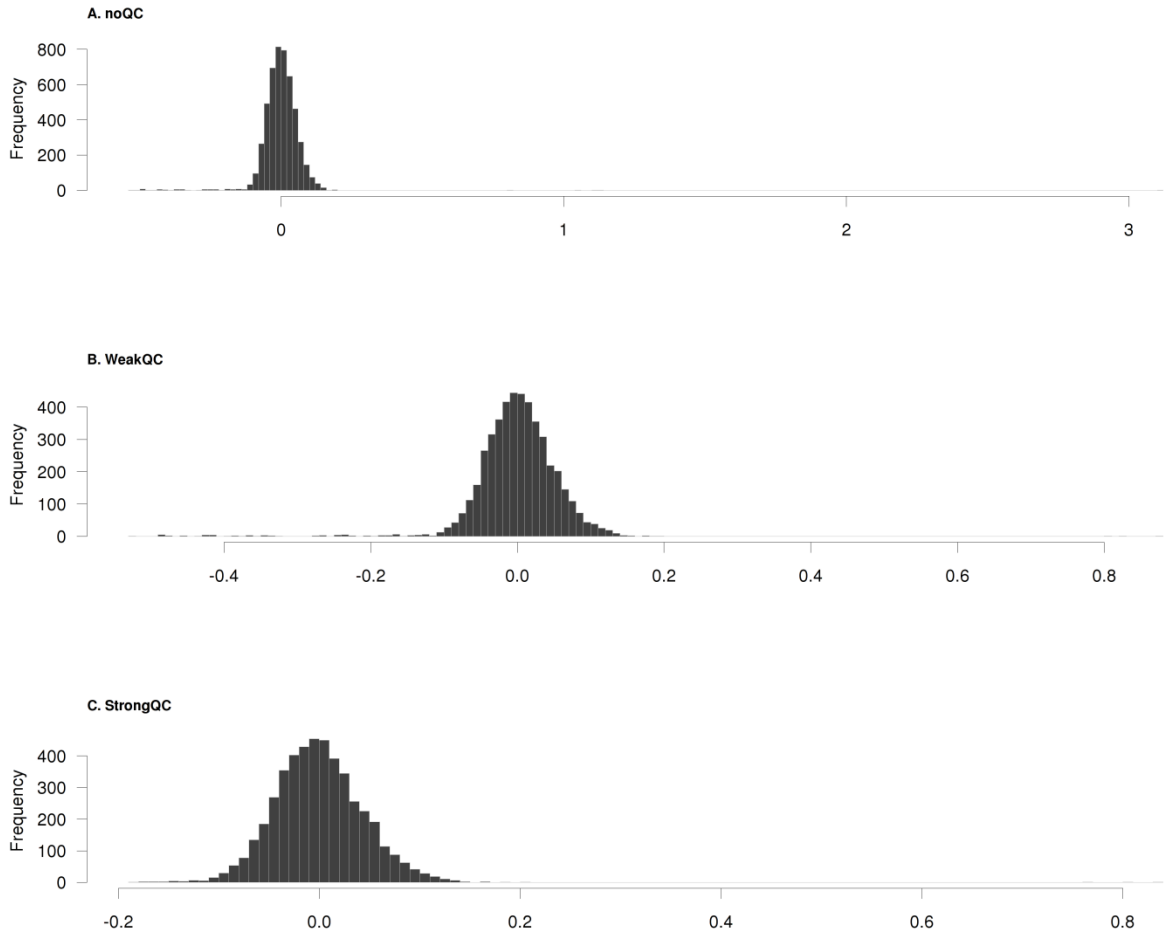


Figure 6.1. Distribution of difference (G−A) in diagonal coefficients between genomic (GRM) and numerator (NRM) relationship matrices under different levels of quality control (QC). A. no any QC, B. weak QC with call rate threshold is 0.7, and C. strong QC with call rate threshold is 0.9. X-axis represents the values of G−A, and y-axis represents frequencies.

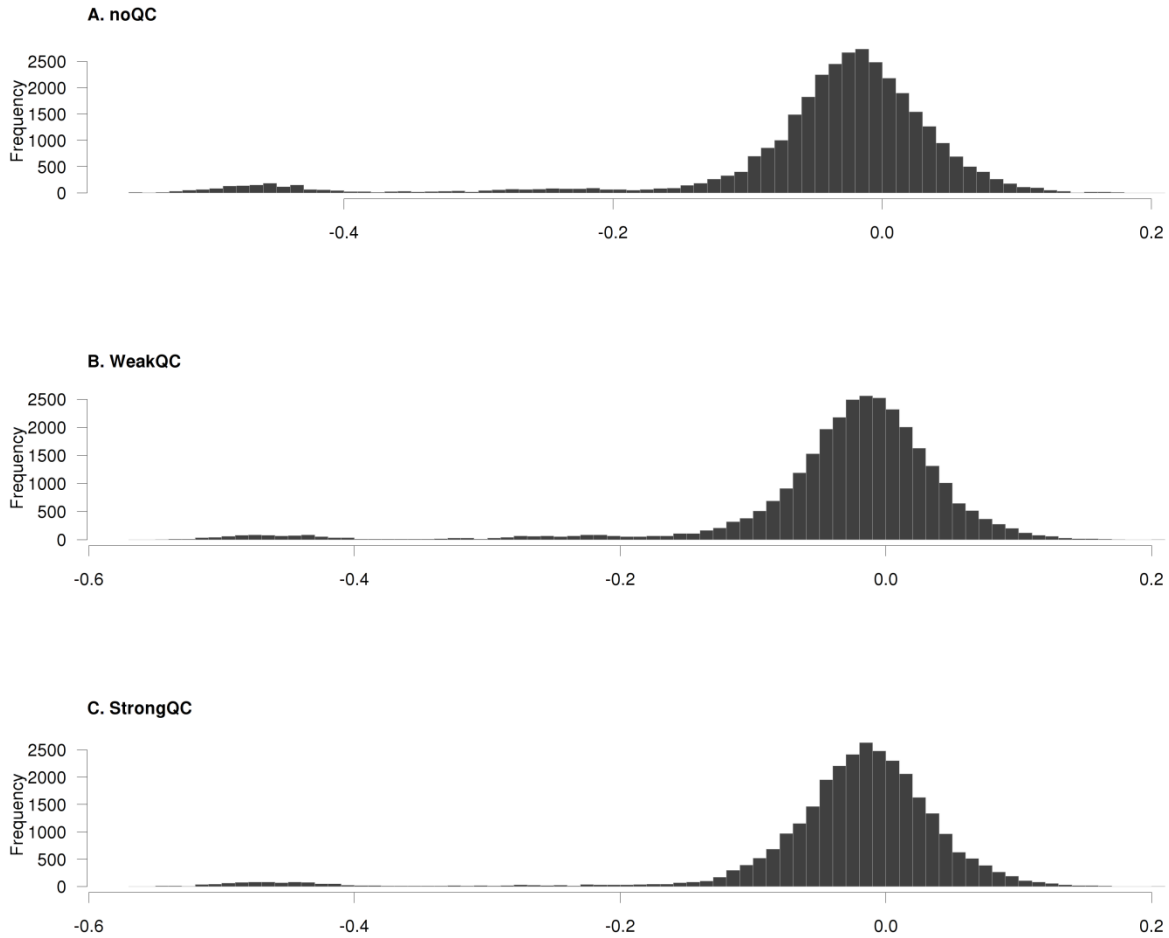


Figure 6.2. Distribution of difference in coefficients (G−A) between genomic (GRM) and numerator (NRM) matrices under different levels of quality control (QC), where all coefficients in NRM are between 0.45 and 0.55. A. no any QC, B. weak QC with call rate threshold is 0.7, and C. strong QC with call rate threshold is 0.9. X-axis represents the values of G−A, and y-axis represents frequencies.

CHAPTER 7

CONCLUSIONS

Genome-wide association study based on ssGBLUP is more powerful and robust than multiple-step approaches. The accuracies of prediction and SNP effects are at least equal to BayesB method, and much better than single-marker analysis for simulation data set. For filed data, as the real QTLs are unkown, there is no evidence for superiority of method choice, but similarity between BayesB and ssGBLUP had been shown in various aspects. BayesB has shown fewer peak(s) referring to regions of large genetic variance, as the use of a shrinkage parameter (i.e. π). The ranks of top 10 regions (consisting of 20 consecutive SNPs) that were able to represent most genetic variance were compared among the methods for body weight at 6 weeks for broiler chickens. Re-ranking occurred during within iterations of ssGBLUP and between different approaches, where several top regions are common between ssGBLUP and BayesB. The ssGBLUP is able to apply for complex models, like multi-trait model with maternal environmental effects. The patterns of the Manhattan plots from ssGBLUP across traits implied QTLs distribution, and across lines implied selection goals. According to our results, trait of body weight of broiler chickens has several medium QTLs associated with phenotypes, but traits of leg problem and breast meat more likely follow infinitesimal model with no apparent genes. Two pure lines may have different selection goals, as the peaks in plots were rarely overlapped. Finally, the advantages of using ssGBLUP includes: 1) no pseudo values are required, 2) complex modeling and multiple-traits are possible, and 3) computing is fast and implementation is simple.

APPENDIX A

GENOME-WIDE ASSOCIATION WITH SINGLE-STEP GBLUP[1]

[1] Developed by I. Aguilar (INIA, Uruguay)

Stranden and Garrick (2009b) presented a derivation of equivalent models to predict genomic effects. Using a linear model based on BLUP methods of Meuwissen et al. (2001b):

$$y = Xb + Zu + e$$

where **Z** is a matrix of marker incidence, **u** the vector of random SNP markers effects; with

$$var(u) = I\sigma_u^2$$

$$var(e) = I\sigma_e^2$$

Solving the mixed model equations, the solutions for the random marker effects are:

$$\hat{u} = (Z'R^{-1}Z + \lambda I)^{-1}Z'R^{-1}(y - X\hat{b})$$

where

$$\lambda = \frac{1}{\sigma_u^2}$$

Estimated breeding values (**EBV**) can be calculated summing SNP markers effects:

$$\hat{a} = Z\hat{u}$$

with

$$var(a) = var(Zu) = Zvar(u)Z' = ZZ'\sigma_u^2$$

This results in an equivalent model to estimate breeding values (Goddard, 2009; VanRaden, 2008a).

From Habier et al. (2007) the variance of SNP effects:

$$var(u) = \frac{\sigma_a^2}{2\sum p_i(1 - p_i)}$$

The model equivalent model:

$$y = Xb + Ia + e$$

where

$$a = Zu$$

and then

$$var(a) = ZZ'\sigma_u^2 = ZZ'\frac{\sigma_a^2}{2\sum p_i(1-p_i)} = G\sigma_a^2$$

Stranden and Garrick (2009b) show in formula 4, that

$$\hat{a} = (R^{-1} + \lambda_a G^{-1})^{-1} R^{-1}(y - X\hat{b})$$

where $\lambda_a = \frac{\sigma_e^2}{\sigma_a^2}$.

Applying matrix inversion lemma to the equiation 1 Stranden and Garrick (2009b) arrives to the following formula:

$$\hat{u} = \lambda^{-1} Z' G^{-1}[(R^{-1} + \lambda_a G^{-1})^{-1} R^{-1}(y - X\hat{b})]$$

where substituting $(R^{-1} + \lambda_a G^{-1})^{-1} R^{-1}(y - X\hat{b})$ by **a** results in :

$$\hat{u} = \lambda^{-1} Z' G^{-1} \hat{a}$$

where

$$\lambda = \frac{1}{\sigma_u^2} = \frac{1}{\frac{\sigma_a^2}{2\sum p_i(1-p_i)}} = \frac{2\sum p_i(1-p_i)}{\sigma_a^2}$$

$$\hat{u} = [\frac{2\sum p_i(1-p_i)}{\sigma_a^2}]^{-1} Z' G^{-1} \hat{a} = \frac{\sigma_a^2}{2\sum p_i(1-p_i)} Z' G^{-1} \hat{a}$$

A matrix G can be constructed using different weights on each marker (Goddard, 2009). In such case

$$G = \frac{ZDZ'}{2\sum p_i(1-p_i)}$$

where D is a diagonal matrix with elements with differential weights for each marker. This will create a relationship matrix that will account for the differential variance explained by each marker. Using a matrix with differential weight for markers results in better predictability (Zhang

et al., 2010b). Legarra et al. (2011b) used a weighted relationship matrix with weights based on the estimated marked effect using a Bayes Lasso methodology.

For weighted G the equation becomes:

$$\hat{u} = \frac{\sigma_a^2}{2\sum p_i(1-p_i)} Z'[\frac{ZDZ'}{2\sum p_i(1-p_i)}]^{-1}\hat{a}$$

This can be implemented in a iterative process, starting with **D**=**I,** and then updating D with

$$D_i = u_i^2 2p_i(1-p_i)$$

REFERENCES

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257.

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.

Legarra, A., C. Robert-Granie, P. Croiseau, F. Guillaume, and S. Fritz. 2011. Improved lasso for genomic selection. Genetics Research 93:77-87.

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.

Stranden, 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.

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

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