GWAS using ssGBLUP

I. Misztal[*], H. Wang[†], I. Aguilar[‡], A. Legarra[§], S. Tsuruta[*], D.A.L. Lourenco[*], B. O. Fragomeni[*], X. Zhang[*], W. M. Muir[#], H. H. Cheng[‖], R. Okimoto[‖], T. Wing[‖], R. R. Hawken[‖], B. Zumbach[‖], R. Fernando[¶].

[*]University of Georgia, Athens, GA, USA; [†]Genus PIC, Hendersonville, TN; [‡]Instituto Nacional de Investigacion Agropecuaria, Las Brujas, Uruguay; [§] INRA, UMR1388, Toulouse, France; [#]Purdue University, West Lafayette, IN, USA; [‖]Cobb-Vantress Inc., Siloam Springs, AR, USA; [¶]Department of Animal Science, Iowa State University, Ames, IA, USA.

ABSTRACT: This study aimed to compare results of genome-wide associations obtained from various methodologies for GWAS when applied to two lines of broiler chicken. Each line contained >250k birds with up to 3 traits and ~5k genotypes with a 60k SNP chip. Methods included single-step GWAS, single marker model and BayesB. Manhattan plots were based on variances of 20-SNP segments, as shorter segments produced noisy plots. Only a few segments explained >1% of the additive variance. One segment explained >20% variance in BayesB but 3% with ssGWAS and <1% with a single marker model. In two lines, no major segment overlapped for any trait. When analyses used slices of generations (1-3,2-4,3-5,1-5), variances for the same segment varied greatly. The plots were more distinct with a new data set that included >16k genotypes, but no segment explained >1.5% of the variance. Strength of associations strongly depends on methodologies and details of implementations.

Keywords: ssGBLUP; GWAS; SNP variance

Introduction

A few years ago NIFA supported a research project on genomic selection in broiler chicken. In this project, two lines of broiler chicken from Cobb were genotyped and phenotyped over a few generations of selection in order to determine gains from genomic selection. As this project coincided with the development of a single-step GBLUP (ssGBLUP) methodology at UGA, many tests of ssGBLUP used the data from this project (e.g., Chen et al. (2011); Simeone et al. (2012)).

Initially, the interest was in genomic selection and comparisons of (G)EBV. Later, the single-step methodology was extended to GWAS (Wang et al. (2012)). This paper documents experiences with GWAS using ssGBLUP and other methodologies mainly in application to the Cobb data set.

Materials and Methods

Single-step genomic association study - ssGBLUP and ssGWAS. The ssGBLUP method is a modification of BLUP with the numerator relationship matrix A^{-1} matrix replaced by H^{-1} (Aguilar et al. (2010)):

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

where $\mathbf{A}_{22}$ is a numerator relationship matrix for genotyped animals and **G** is a genomic relationship matrix. The last matrix can be created following VanRaden (2008) as:

$$\mathbf{G} = \mathbf{ZDZ}'q,$$

where **Z** is a matrix of gene content adjusted for allele frequencies, **D** is a weight matrix for SNP (initially **D** = **I**), and q a normalizing factor. Such a factor can be derived either based on SNP frequencies (VanRaden (2008)), or by ensuring that the average diagonal in **G** is close to that of $\mathbf{A}_{22}$ (Vitezica et al. (2011)). The latter method was used in this study. Briefly, SNP effects and weights for GWAS can be derived as follows (Wang et al. (2012)):

1) Let **D** = **I** in the first step.
2) Calculate $\mathbf{G} = \mathbf{ZDZ}'q$.
3) Calculate GEBV for entire data set using ssGBLUP.
4) Convert GEBV to SNP effects ($\hat{u}$) : $\hat{u} = q\mathbf{DZ}'\left(\mathbf{ZDZ}'q\right)^{-1}\hat{a}$, where $\hat{a}$ is the GEBV of animals which were also genotyped.
5) Calculate SNP weights: $d_i=\hat{u}_i^2 2p_i(1-p_i)$.
6) Normalize SNP weights to remain the total genetic variance constant.
7) Loop to 2. (ssGWAS1) or 4. (ssGWAS2).

Step 4 is based in the equivalence between GBLUP and SNP-based models (VanRaden (2008)). The SNP weights were calculated iteratively either looping through steps 4-6 (called as ssGWAS1) or through steps 2-6 (called as ssGWAS2). Iterations with both scenarios increase weights of SNP with large effects and decrease those with small effects, essentially regressing them to the mean.

Experiences with simulated data using ssGBLUP (Wang et al. (2012)) indicated that ssGWAS1 was more suitable for identification of SNPs with the largest effects while ssGWAS2 was superior for more accurate GEBV. Also, the highest correlations with QTLs were lower with individual SNP effects but much higher with an average of 8 adjacent SNPs. Similar findings were reported with a method based on GBLUP (Sun et al. (2011)).

Implementation of ssGWAS. All runs of ssGWAS in this study used programs of the BLUPF90 family (Misztal

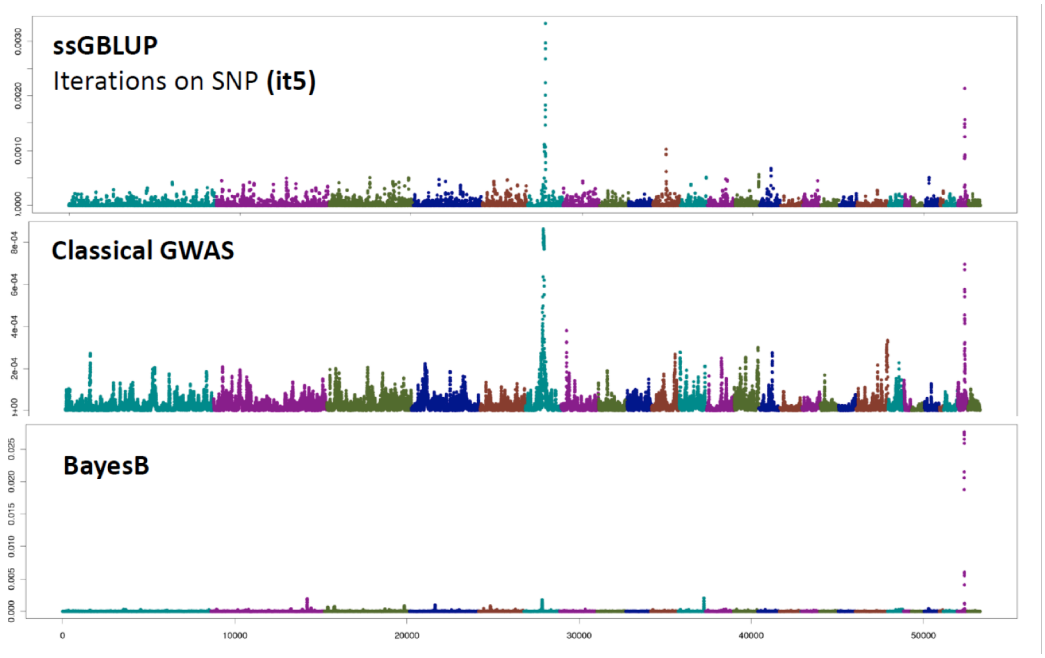

Figure 1. Manhattan plots for Body Weight in chicken obtained by ssGWAS1, Single marker model, and BayesB

et al., (2002)), with modifications by Aguilar et al (2014). The BLUPF90 family includes programs for variance component estimation and for genetic/genomic evaluation. Genomic models are aided by two new components: preGSf90 and postGSf90 (Aguilar et al. (2014)). preGSf90 adds processing of genotypes including extensive quality control for up to 150k genotypes. postGSf90 adds conversion from GEBV to SNP effects, computations of marker weights, and creation of Manhattan plots using moving or overlapping windows. With a new sparse matrix package (Masuda et al., (2014)), REML programs run multitrait models with ≥20k genotypes. Efforts to extend the programs to very large number of genotypes are underway (Misztal et al., (2014); Fragomeni et al., (2014)).

Data. The data for the USDA experiment were provided by Cobb-Vantress Inc. (Siloam Springs, AR). It included two lines of broiler chicken, each with > 250k animals across five generations. Phenotypes were available for body weight, breast meat, and leg score. Approximately 5k animals were genotyped using a 60k SNP chip, with about 1k genotyped animals per generation; quality control with the genotypes included removing Z chromosome SNPs from analyses. More accurate description of the data at an interim stage of selection is available in Chen et al. (2011).

Analyses. Methodologies compared included ssGWAS1 and ssGWAS2, a single marker model (SMM) implemented by WOMBAT (Meyer and Tier (2012)), and BayesB (Meuwissen et al. (2001)) implemented by GenSel (Fernando and Garrick (2009)) with $\pi = 0.9$. Comparison of genomic regions identified between methods was based on plots of genetic variances explained by local SNP regions (20 SNPs).

Results and Discussion

Comparisons of methods based on body weight. Manhattan plots for ssGWAS (3rd iteration), SMM and BayesB are presented in Figure 1. In the first iteration, results from ssGWAS1 were noisy with many small peaks. After 3 iterations the noise was greatly reduced with results similar to that of WOMBAT, but with only 4 out of the top 10 regions in common. In contrast, for BayesB, the noise was also eliminated but to a greater degree, resulting in a plot that was dominated by a single region explaining 23.1% of the genetic variance. This same region was found by ssGWAS1, and with the same rank, but the amount of genetic variation attributed to the region was only 3%. These results highlight that detected associations and strength of association, strongly depends on methodologies and details of implementations. BayesB appears to overly shrink regions to zero, while overestimating the amount of genetic variation attributed to the remaining SNP effects. Lately, several scientists reported large variations in estimates by BayesB and similar methods (Hulzen et al. (2012); J. Taylor, (2013; pers. Comm.)).

Are Manhattan plots similar across lines? Identification of QTLs in one species suggests that Manhattan plots have peaks across lines and even breeds. Figure 2 shows Manhattan plot for 3 traits (Wang et al. (2013)) using ssGWAS1. Several peaks explaining > 1% of the genetic variation were found for body weight, however, peaks for lines 1 and 2 are on different chromosomes. No strong peaks have been observed for breast meat and leg score, and each region for these traits explained < 1% of total genetic variance. Breast meat and leg scores seemed to follow the infinitesimal model. Different peaks for the two lines for body weight suggest different selection goals, or fixation of alleles within lines. Plots for all traits contain few peaks, and no peak clearly overlaps with the other line. Eitan and Soller ((2013)) discussed why the additive variance is maintained despite an intensive selection. While major genes at a time become fixed, other genes become major genes. Subsequently, the additive model is a good approximation for a few generations, but epistasis becomes important over many generations.

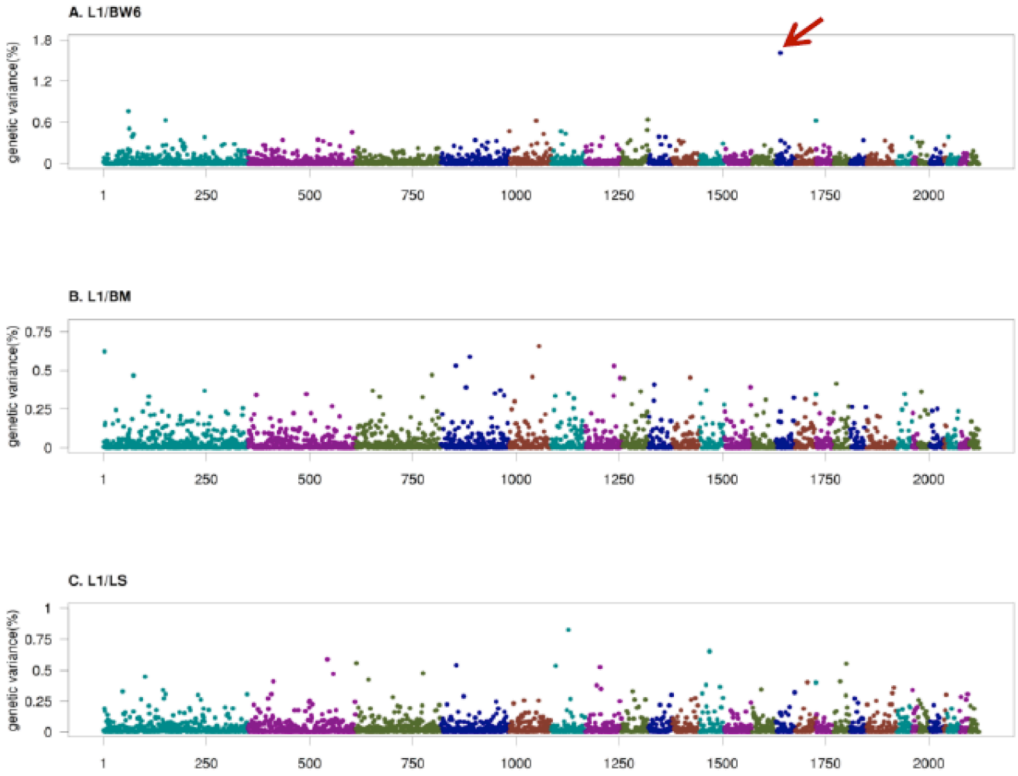

Figure 2. Manhattan plots for body weight, breast meat and leg score in Line 1 obtained by ssGWAS1

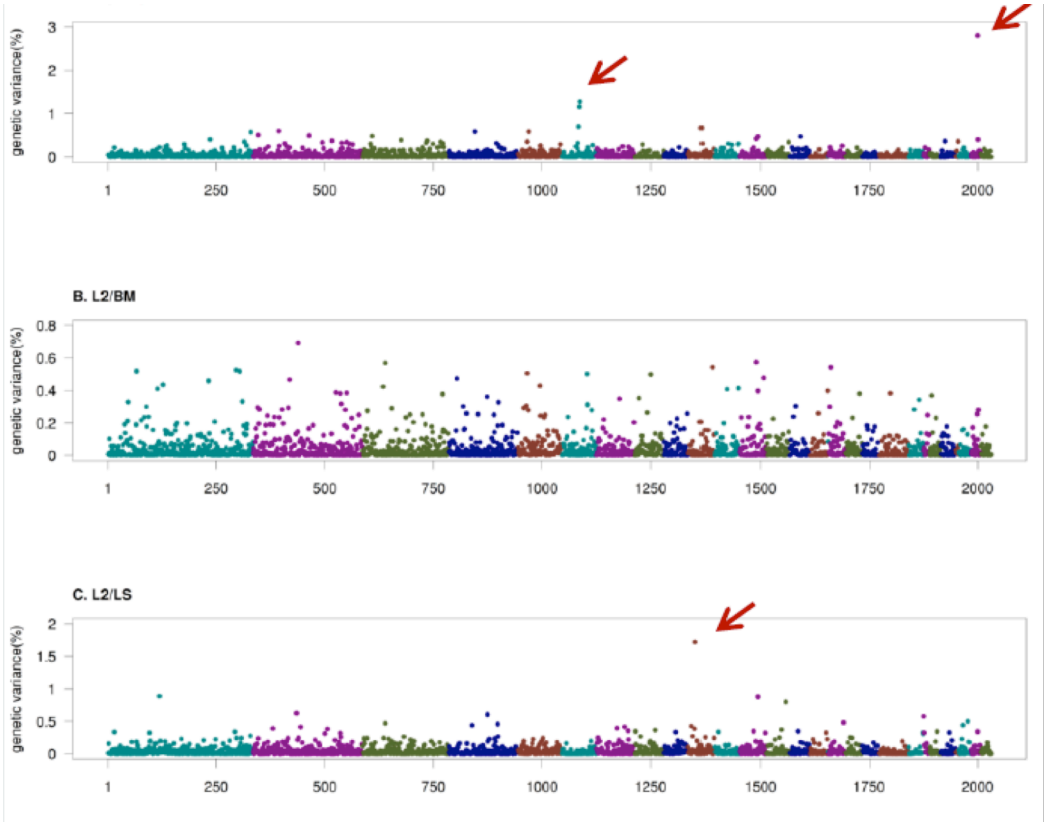

Figure 3. Manhattan plots for body weight, breast meat and leg score in Line 2 obtained by ssGWAS1

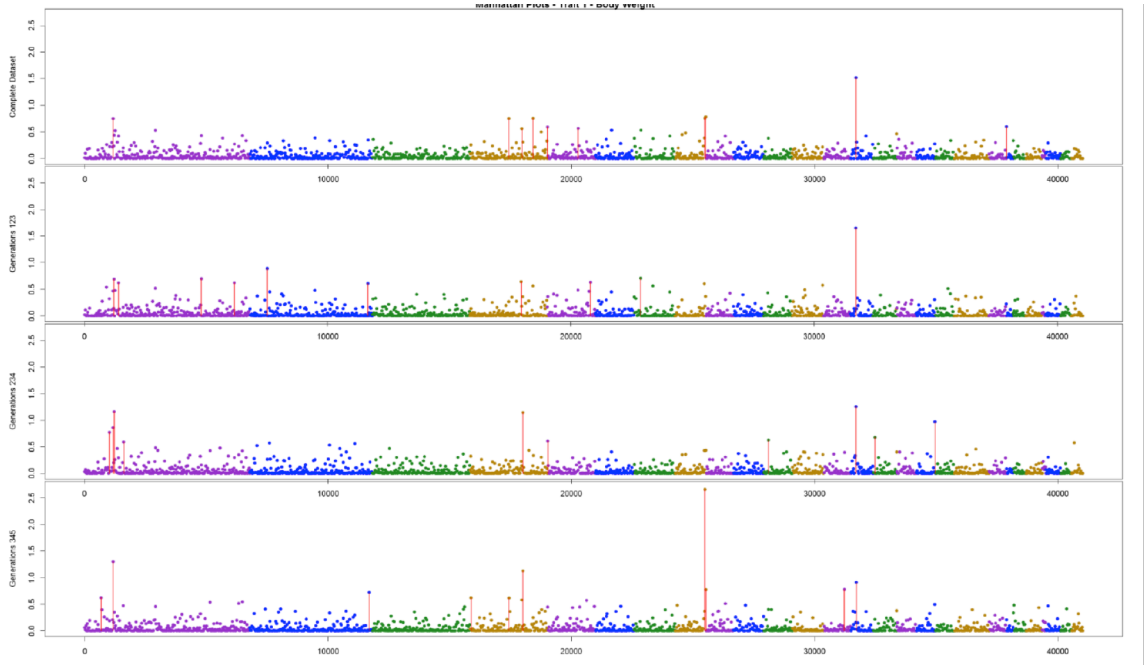

Figure 4. Manhattan plots for Body Weight by GWAS1 calculated with genotypes for generations 1-3, 2-4, 3-5, and 1-5

Are Manhattan plots similar during selection? The purpose of a study done by Fragomeni et al. (2014) was to determine whether the top SNP segments that explain the most variance are stable over multiple generations. The data set was one line of the broiler chicken with phenotypes for body weight, breast meat, and leg score. SNP effects were calculated by ssGWAS1 (3rd iteration) using genotypes from generations 1-3, 2-4, 3-5, and 1-5 (Figure 3). Variances were calculated for segments of 20 SNP. Ten segments for each trait were identified that explained the largest fraction of the variance in any combination of generations. All the segments explained > 0.5% and few explained > 1 % of the total variance. In all the segments the variance explained varied greatly among the combinations of generations. In many cases, a segment identified as top for one combination of generations explained < 0.1% variance for the remaining combinations. Thus, even the top SNP segments identified for a population in broiler chicken may have little predictive power for genetic selection in the following populations.

Why SNP segments and their effects change? Large changes in the variance of SNP windows could be indirectly due to small effective population size and subsequent low number of independent chromosome segments. According to Goddard (2009), the number of such segments is $q=2N_eL/\log(4N_eL)$, where Ne is effective population size and L is the length of chromosome in Morgans. Assuming Ne=50 (lower range showed in Andrescu et al. (2007)) and L=39, q=435. Subsequently there are > 100 SNP per 1 chromosome segment, if we apply the formula in this dataset. Since the boundaries of segments are not fixed but change with populations and additional information, the effects of those segments change as well. Additional issue is high prediction error variance.

Noisy plots: reality or too small data sets? One reason for few peaks in Manhattan plots could be insufficient data although 5k genotypes constitutes a substantial data set. Following the USDA project, Cobb followed with genotyping in a broiler line, which included about 200k birds with genotypes for >15k birds. Their data was analyzed as an effort to test a BayesC-like algorithm for ssGWAS. In this algorithm, only weights for the top 20 SNP were allowed to change, with the standard weight of 1.0 for the remaining SNP (Zhang et al. (2014)). Figure 5 contains Manhattan plot for new ssGWAS and BayesC for one trait. With the larger number of genotypes the peaks are clearer, however, the biggest window explains <1% or <2.5% variance depending on the method.

Can we see large peaks for highly selected populations? High peaks in Manhattan plots can be due to many factors including changes in gene frequencies and pleiotropy. Intensive selection such as in broiler chicken should lead to fixation of major QTL if that QTL has a positive effect on all selected traits. However, a QTL with positive effect on one trait may have a detrimental effect on other traits, with a total economic value close to 0. Therefore the QTL may not reach fixation. For example, in many studies involving milk yield in Holsteins, Manhattan plots show strong peaks close to DGAT1 gene even for small data sets despite strong selection in dairy. This is because both alleles of DGAT1 have been selected, either for large fat content or for large milk yield, and also perhaps for another traits. Tsuruta et al. (2014) performed GWAS for milk yield and mortality rate using a Holstein data set with 35k genotypes and > 6M phenotypes. The biggest region for both milk and mortality was for the region close to DGAT1. If present, pleiotropy can cause serious issues in GWAS (Solovieff et al. (2013)). Another possibility is that a low

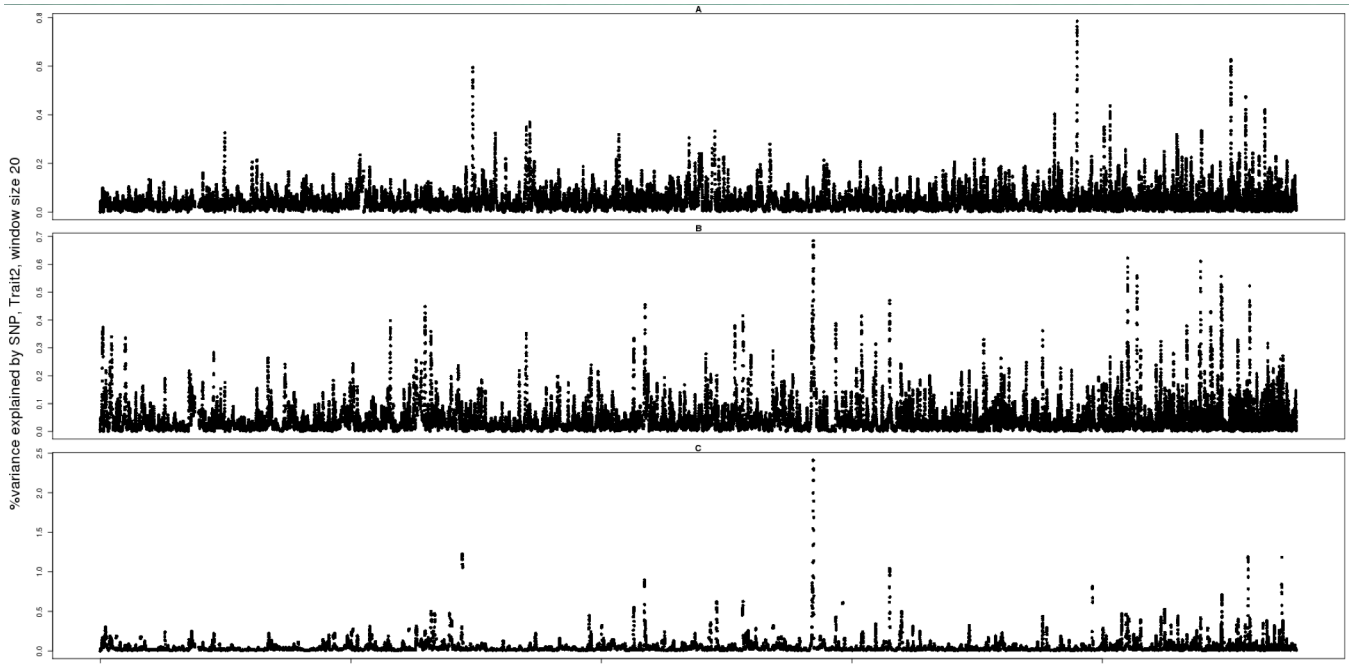

Figure 5. Manhattan plots a new data set containing > 16k genotypes obtained by ssGWAS2 (default and with option to modify weights for the top 20 SNP only) and BayesC with $\pi = 0.99$.

frequency (but of large effect) allele becomes more frequent and is easily detected. Yet another possibility is that the QTL interacts with the genetic background or the environment so that at some point it becomes of large effect. In general, mutation continuously generates low frequency alleles that are not strongly selected or detected until they reach intermediate frequencies.

Conclusion

For traits of broiler chicken in this study, few windows explain more than 1% of the additive variation, the windows variance may change greatly over time, and no major windows may be common across lines. Detected associations and strength of association strongly depend on methodologies and details of implementations. However, very large peaks do largely agree across methodologies. More clear associations require large number of genotypes, however, the variance explained by one windows is likely to be too small for use in selection. The ssGWAS is a young but potentially useful tool for GWAS when the population contains large number of genotypes and especially if models of analyses are complex.

Literature Cited

Aguilar, I., Misztal, I., Johnson, D.L. et al. (2010). *J. Dairy Sci*. 93:743:752.

Aguilar, I., Misztal, I., Tsuruta, S. et al. (2014). Proc. 10th WCGALP. [Submitted].

Andreescu, C., Avendano, S., Brown, S.R. et al. (2007). *Genetics*. 177, 2161–2169.

Chen, C.Y., I. Misztal, I. Aguilar, A. et al. (2011). *J. Anim. Sci*. 89:2673-2679.

Eitan, Y., Soller, M. (2013). Encyclopedia of Sustainability Science and Technology. 11:8307-8328.

Fernando R.L., Garrick DJ. (2009). Gensel Manual.

Fragomeni, B.O., Misztal, I., Lourenco, D.A.L. et al. (2014). *Fron. Genet*. [Submitted]

Fragomeni, B.O., Misztal, I., Lourenco, D.A.L. et al. (2014). Proc. 10th WCGALP. [Submitted]

Goddard, M. (2009). *Genetica*, 136:2. 245-257.

Hulzen, K.J.E. van, Schopen, G.C.B., Arendonk, J.A.M. van et al. (2012). *J. Dairy Sci*. 95, 2740-2748.

Masuda, Y., Aguilar, I., Tsuruta, S., et al. (2014). Proc. 10th WCGALP. [Submitted].

Meuwissen, T., Hayes, B.J., Goddard, M.E. (2001). *Genetics* 157, 1819–1829.

Meyer, K., Tier, B. (2012). *Genetics*. 190:275-277.

Misztal, I., Tsuruta, S. Strabel, T. et al. (2002). Proc. 7th WCGALP.

Misztal, I., Legarra, A., Aguilar, I. (2014). *J. Dairy Sci*. [In print].

Simeone, R., Misztal, I., Aguilar, I. et al. (2011). *J, Anim. Breed. Genet*. 128:386-393.

Solovieff, S. Cotsapas, C., Lee, P. H. et al. (2013). *Nat. Genet.* 14:483:495.

Sun, X., Fernando, R.L., Garrick, D.J. et al. (2011). *J. Anim. Sci.* 89(E-Suppl2):e11.

Tsuruta, S. Misztal, I, Aguilar, I., et al. (2014). *J. Dairy Sci*. [Submitted]

VanRaden, P.M. (2008). *J. Dairy Sci*.91:4414-4423.

Vitezica, Z.G., Aguilar, I., Misztal, I. et al. (2011). *Genet. Res*. 93:357-366.

Wang, H., Misztal, I., Aguilar, et al. (2012). *Genet. Res*. 94:73-83.

Wang, H., Misztal, I., Aguilar, I. et al. (2013). *J. Anim. Sci.* 91(E-Suppl. 2):191.

Zhang, X., Misztal, I., Lourenco, D.A.L. et al. (2014). Proc. 10th WCGALP. [Submitted].