Breeding and Genetics: Whole Genome Selection

612 Utility of genomic relationship matrix to identify genotyping errors.

R. Simeone*[1], I. Misztal[1], and I. Aguilar[1,2], [1]*University of Georgia*, *Athens*, [2]*INIA*, *Las Brujas, Uruguay*.

The purpose of this study was to use the genomic relationship matrix (G) as an indicator of genotyping or other analysis problems in a single-step genomic evaluation procedure. Data was obtained from Cobb-Vantress and consisted of body weights for 183,784 broiler chickens over 3 generations with pedigrees on 186,222 animals. Of these animals 3,284 were genotyped for 57,636 SNP. Loci with no variation or minor allele frequency <0.02 were removed from the data, leaving 48,006 loci for analysis. Construction of G used current allele frequencies. Theoretically, the mean of the diagonal elements in both relationship matrices should be the same. The mean of the diagonal elements of G was 1.03 ± 0.16, however, the distribution of these elements showed 3 peaks: 3,195 in the range from 0.54 to 1.19, 88 in the range from 1.73 to 2.09, and one with a value of 3.12. Animals with a diagonal element >1.2 were assumed to have abnormal genotypes. Genetic predictions were computed by a single-step procedure (SSP) that combined phenotypic, pedigree and genomic information. This procedure was applied with all genotypes or with abnormal genotypes removed and with all phenotypes or only with phenotypes of genotyped animals. Accuracies were computed by dividing the predictive ability by the square root of heritability. Removing genotypes causing abnormal diagonals increased the accuracy from 0.648 to 0.657 when all phenotypes were used and from 0.584 to 0.586 when only phenotypes of genotyped animals were used. The difference between predictions obtained with and without the abnormal genotypes was distributed close to normal but with longer tails. Analysis of diagonals in G may serve as a diagnostic tool to identify erroneous genotypes. Very large diagonals suggest an analysis problem; explanations may be presence of animals of another breed, allele frequency shifts or a genotyping error. Removing suspected genotypes is likely to improve accuracy of genetic evaluation, especially for animals with suspected genotypes or their progenies.

Key Words: genomic relationship, genotyping error, single-step procedure

613 Genetic evaluation including phenotypic, full pedigree, and genomic information: An application in broiler chickens.

C. Y. Chen*[1], I. Misztal[1], I. Aguilar[1,2], S. Tsuruta[1], T. H. E. Meuwissen[3], S. E. Aggrey[4], and W. M. Muir[5], [1]*Department of Animal and Dairy Science, University of Georgia*, *Athens,* [2]*Instituto Nacional de Investigación Agropecuaria*, *Las Brujas 90200, Uruguay*, [3]*Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences*, *NO-1432 As, Norway*, [4]*Department of Poultry Science, University of Georgia*, *Athens,* [5]*Department of Animal Science, Purdue University*, *West Lafayette, IN*.

A complete phenotypic data set (FULL) consisted of 183,784 and 164,246 broilers for 2 lines across 3 generations. Genotyped subset (SUB) consisted of 3,284 and 3,098 broilers in lines 1 and 2 with 57,636 SNP available. Traits were body weight at 6 weeks (BW), ultrasound (US), and binary leg defect score (LEG). Some records were missing for US. Heritability with FULL were 0.17–0.20 for BW, 0.30–0.35 for US, and 0.09–0.11 for LEG. Genetic evaluation was performed by regular BLUP, by a single-step procedure (SSP) that combined relationships based on pedigree and the SNP data, and by Bayes A procedure. While BLUP and SSP could use the complete data set, Bayes A could use only the genotyped subset. Genotyped animals in generation 3 were treated as validation population. The average accuracies of the validation population with BLUP for BW, US, and LEG were 0.46, 0.30, and < 0 with SUB and 0.51, 0.34, and 0.28 with FULL. With SSP, those accuracies were 0.60, 0.34, and 0.06 with SUB and 0.61, 0.40, and 0.37 with FULL, respectively. Accuracies with BayesA were similar to SSP with SUB. Accuracies in lines 1 and 2 were similar for US but different for BW and LEG. For traits with high heritability, the accuracy of the evaluation using the genomic information and only records of genotyped animals may be higher than that using the complete data and BLUP. The opposite is likely for traits with lower heritability, many missing records, or undergoing pre-selection. An optimal genomic evaluation would be multi-trait and would involve all traits and records on which the selection is based.

Key Words: chicken, genetic evaluation, genomic prediction

614 Scaling the genomic relationship matrix for single-step evaluation using phenotypic, pedigree and genomic information.

S. Forni*[1,2], I. Aguilar[3,2], I. Misztal[2], and N. Deeb[1], [1]*PIC/Genus Plc*, *Hendersonville, TN,* [2]*University of Georgia*, *Athens,* [3]*INIA*, *Las Brujas, Uruguay*.

Data included litter sizes for 338,346 PIC sows, of which 1,919 had genotypes using the porcine 60k SNP chip. Genotypes were also available for 70 sires. Analyses involved a complete data set or a subset of genotyped animals and their parents (n = 5,090). A genomic relationship matrix was constructed using equal (G05) or observed gene frequencies (GOB). Additional relationship matrices were the pedigree-based relationship matrix (A) and a combined pedigree-genomic matrix (H). For genotyped animals, the mean of diagonal elements in A (G05, GOB) was 1.00 (1.25, 0.94). The mean of off-diagonal elements was 0.03 (0.59, 0.00). A normalized matrix (GN) was obtained by multiplying GOB by a constant to achieve an average diagonal of 1. Using A and the complete data set, the estimate of the additive variance was 1.26(±0.03). With H that included G05, GOB or GN the additive variance estimates were 1.28(±0.03), 1.28(±0.03) and 1.27(±0.03), respectively. Using A and the subset of the data, the estimate of the additive variance was 2.28(±0.52). With H that included G05, GOB or GN the additive variance estimates were 3.43(±0.56), 2.42(±0.39) and 2.25(±0.36), respectively. Accuracies for the complete data set were estimated by inversion. The average accuracy for genotyped animals using A, G05, GOB and GN were 0.23, 0.38, 0.31, and 0.30, respectively. When the genomic relationship matrix has a different scale than the pedigree-based matrix, the estimates of the additive variance may be biased especially for small data sets. Also, estimates of the accuracies of evaluation obtained by inversion may be inflated. One solution to normalize the genomic relationship matrix is by using realized gene frequencies and scaling this matrix to obtain an average diagonal close to 1.

Key Words: genomic selection, swine, single-step evaluation

615 Accuracies of direct genomic breeding values estimated in dairy cattle with a principal component approach.

N. P. P. Macciotta*[1], M. A. Pintus[1], R. Steri[1], C. Pieramati[2], E. L. Nicolazzi[3], E. Santus[4], D. Vicario[5], J. T. van Kaam[6], A. Nardone[7], A. Valentini[7], and P. Ajmone-Marsan[3], [1]*Università di Sassari*, *Sassari, Italia*, [2]*Università di Perugia*, *Perugia, Italia*, [3]*Università di Piacenza*, *Piacenza, Italia*, [4]*ANARB*, *Bussolengo, Italia*, [5]*ANAPRI*, *Udine, Italia*, [6]*ANAFI*, *Cremona, Italia*, [7]*Università della Tuscia*, *Viterbo, Italia*.

A severe risk of overfitting due to the huge asymmetry between number of markers and phenotypes usually represents the main constraint for the implementation of genomic selection in livestock species. In the present work, the number of predictors for calculating direct genomic breeding values (DGV) is reduced by using principal component (PC) analysis. Sires of 3 dairy cattle breeds farmed in Italy were genotyped with the 54K Illumina beadchip: 863 Holstein (H), 749 Brown (B), and 479 Simmental (S). SNPs retained after edits were 40,658, 37,254, and 40,179 and the number of PC extracted 2,564, 2,257, and 2,476 for H, B, and S respectively. Effect of PC on polygenic EBV was estimated in the reference population with a BLUP model. Traits considered were milk yield, protein percentage, udder score and economic index. To create reference and validation population, bulls were tagged either by birth year or randomly. Accuracies were calculated as correlation between DGV and polygenic EBV in validation bulls. High DGV accuracies are obtained with reference animals selected at random (Table 1). When older animals are used to predict younger bulls, DGV accuracy drops dramatically for milk yield, especially for B and H, while it remains almost unchanged for udder score, protein percentage in B and milk yield in S.

Table 1.

	Random			By Year		
Trait	Holstein	Brown	Simmental	Holstein	Brown	Simmental
Milk yield	0.62	0.82	0.72	0.21	0.18	0.46
Protein percentage	0.52	0.58	0.32	0.37	0.54	0.36
Udder score	0.63	0.64	0.58	0.61	0.52	0.46
Economic index	0.67	0.84	0.64	0.44	0.33	0.28

Key Words: genomic selection, principal components, dairy cattle

616 Choice of parameters for single-step genomic evaluation for type.

I. Misztal*[1], I. Aguilar[1,2], A. Legarra[3], and T. J. Lawlor[4], [1]*University of Georgia, Athens,* [2]*INIA, Las Brujas, Uruguay,* [3]*INRA, Toulouse, France,* [4]*Holstein Association, Brattleboro, VT.*

In a single step procedure, the pedigree-based matrix A is replaced by a matrix H that blends pedigree and genomic relationships. The inverse of matrix H involves an expression $G^{-1} – A_{22}^{-1}$, where G is a genomic relationship matrix and A_{22} is a pedigree relationship matrix for genotyped animals. Two modifications to that expression: $(\alpha\ G + \beta\ A_{22})^{-1}$ - A_{22}^{-1} and $\tau\ (0.95\ G + 0.05\ A_{22}^{-1} – \omega\ A_{22}^{-1}$ were investigated with regard to accuracy and scale of genomic predictions. While the first is equivalent to assuming a genomic and polygenic effect for genotyped animals, the second is equivalent to assuming a double prior for the additive effect. Data included final scores recorded from 1955 to 2009 for 6.2 million Holsteins, pedigrees for 10.5 million animals, and SNP50 genotypes for 6,508 bulls. Analyses used a repeatability animal model. Comparisons involved R^2 and regression coefficients (REG) based on 2004 predictions of young bulls and their 2009 daughter deviations. REG below 1.0 indicate inflation of genomic predictions. The initial expression yielded $R^2 = 0.41$ and REG = 0.75. With the first modification, varying α from 0.6 to 1.2 decreased R^2 less than 0.01 and decreased REG from 0.81 to 0.71. Increasing β from 0 to 0.6 decreased the R^2 and REG by 0.02 or less. With the second modification, varying τ from 0.6 to 1.5 increased R^2 by about 0.02 and increased REG by 0.02 ($\omega = 0$) to 0.15 ($\omega = 1.0$). Decreasing ω from 1.0 to 0 decreased the R^2 by 0.03 and increased REG from 0.2 ($\tau = 1$) to 0.3 ($\tau = 0$). Parameters $\tau = 1.5$ and $\omega = 0.4$ yielded $R^2 = 0.40$ and REG = 1.0. While the scale of G (parameters α and τ) has a small effect on R^2 and REG, matrix G as used here is about 50% too large. The scale of A_{22}^{-1} (parameter ω), which is associated with parental index based on genotyped bulls, has a large impact on inflation of genomic predictions.

Key Words: genomic relationships, single-step evaluation, inflation

617 Improved reliability approximation for genomic evaluations in the United States.

G. R. Wiggans* and P. M. VanRaden, *Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD.*

For genomic evaluations, the time required to calculate the inverse of the coefficient matrix for the mixed-model equations increases cubically as the number of genotyped animals increases, and an approximation became necessary for estimating US evaluation reliabilities. The original approximation method used the same contribution to reliability from genomics for all animals. That method was improved by using a weighted sum of the genomic relationships of an animal with predictor animals (ΣG_W), which allowed for individual animal differences. Because calculation time for the genomic relationship matrix only increases quadratically and is routinely available, the sum of relationships of an animal with predictor animals can be obtained. Those relationships were weighted by reliability of the traditional evaluation after removing the contribution to reliability from parent average by first converting both reliabilities to daughter equivalents (DE). Reliabilities from August 2009, the last genomic evaluation for which the coefficient matrix was inverted, were decomposed to extract the genomic contribution in terms of DE calculated with an error-to-sire variance ratio of 14. Of 28,047 genotyped Holsteins, 8,353 bulls and 3,559 cows had genomic evaluations and 16,135 animals did not. Regression of DE on ΣG_W was calculated for those 3 groups. Goodness of fit was assessed by plotting predicted values against mean DE for ΣG_W groups, where groups were by 10. A straight line through the origin provided a good fit except for low ΣG_W. A floor of 30 DE was adopted to improve evaluation accuracy for animals with low ΣG_W. The slope was 0.0584 for evaluated bulls, 0.0557 for evaluated cows, and 0.0506 for animals without evaluations. The higher slope for bulls resulted in a higher reliability for the same ΣG_W. The improved approximation method increased accuracy of genomic reliabilities, particularly when comparing animals with different countries of origin and bulls with only genomic evaluations with progeny-test bulls.

Key Words: reliability, genomic evaluation, genomic relationships

618 Cow adjustments for genomic predictions of Holstein and Jersey bulls.

G. R. Wiggans, T. A. Cooper*, and P. M. VanRaden, *Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD.*

Genomic evaluations are calculated by using values that have been deregressed from traditional PTAs estimating single nucleotide polymorphism (SNP) effects. Previous research indicates that including cow genomic data to calculate SNP effects does not increase reliabilities of genomic evaluations of yield traits. Upward bias in traditional PTA of genotyped cows may be the reason for this. The direct genomic value (DGV) is the sum of an animal's SNP effects. It should be consistent with traditional PTA and is for bulls. For cows, however, the traditional PTA is higher. To make the cow PTA more like those of the bulls for the yield traits (milk, fat and protein), mean and variance adjustments were calculated. Evaluations were stratified by reliability so cow PTA could be adjusted to be similar to bulls with the same reliability. The variance adjustment was the SD of deregressed Mendelian sampling within reliability group for bulls divided by the value for cows. The mean adjustment is the difference between bull and cow evaluations after variance adjustment. Deregressed Mendelian sampling values were adjusted, and then the deregression was reversed to obtain the corrected PTA. To determine gains in reliabilities, predictions were made for bulls with current evaluations that did not have evaluations in August 2006. The predicted values were compared with the bull's actual evaluation from January 2010. For Holstein bulls, predictions using cows' adjusted data were 2.5, 2.8 and 2.1 points higher than those from data without adjustment for milk, fat and protein respectively. Jersey bulls also benefit from cow adjustments with an increase in gains in reliability over parent average of 3.9 points for milk, 5.6 points for fat and 3.5 points for protein. Brown Swiss adjustments could not be evaluated due to low numbers of genotyped cows. Genomic evaluations for Holsteins and Jerseys will be more accurate by better using the information from cows.

Key Words: genomics, prediction, evaluation

619 Investigating bull dam bias in national genetic evaluations. F. Canavesi* and R. Finocchiaro, *Associazione Nazionale Allevatori Frisona Italiana*, *Cremona, Italy*.

Parent averages (PA) are used in combination with direct genomic values to predict genomic breeding values (GEBV). Studies conducted in Germany (2009) showed that classic PA making use of sire and dam EBV deviates from expected contribution to the EBV of sons. This is due to overestimation of bull dams for production traits if compared with a trait like somatic cell score where selection and commercial interests play a minor role. The objective of this study was to investigate the relationship between PAs and realized EBVs in the Italian genetic evaluation system for progeny test bulls. A total of around 800 EBV of bulls born between 2001 and 2003 were used to analyze the relationship between their EBV in January 2010 and their PA in 2004 for production, conformation and somatic cells traits. Regression coefficients of sire and dam, EBV used to predict realized 2010 EBV were examined. Results show that both production and conformation traits deviates from expected values while somatic cell count are close to expected contribution of 0.50 for both EBVs of sire and dam respectively. In agreement with the German study the use of male pedigree information resulted in values close to expected and therefore would be the preferred choice in the prediction of GEBV.

Key Words: parent average, future predictions, bulldam bias

620 Gains in reliability from combining subsets of 500, 5,000, 50,000 or 500,000 genetic markers. P. M. VanRaden and M. E. Tooker*, *Animal Improvement Programs Laboratory, ARS, USDA*, *Beltsville, MD*.

More genetic markers can increase both reliability and cost of genomic selection. Fewer markers can be used to trace chromosome segments within a population once identified by high-density haplotyping. Combinations of marker densities can improve reliability at lower cost. As of January 2010, 33,414 North American Holsteins had been genotyped for 50,000 genetic markers. Genotypes for 500,000 markers were simulated using pedigree data for this same population. Linkage was introduced among base alleles to make correlations among simulated genotypes similar to actual. Reduced subsets were examined using every 10th, 100th, or 1000th marker. In marker regression models, polygenic variance was 70, 30, 10, and 0% of genetic variance with 500, 5,000, 50,000 and 500,000 markers, respectively. Respective reliabilities obtained as squared correlations of estimated and true breeding values averaged across 5 replicates were 39.4, 70.2, 82.6, and 84.0% for 14,061 young bull predictions. At highest density, one processor required 2.5 d to complete 150 iterations for the 5 replicates. A mixed-density data set had 500,000 markers genotyped for 3,515 young bulls and 3,883 bulls with > 90% reliability and 50,000 markers genotyped for the remaining 26,016 animals. This data set had 70% missing genotypes; however, after imputing from haplotypes, only 4% of genotypes were missing, and average reliability was 83.1%. Two other mixed-density data sets had 50,000 markers for cows and progeny-tested bulls but only 5,000 or 500 markers for young animals. Reliabilities averaged 79.6% for young animals if 5,000 markers were genotyped and the other 45,000 imputed. At 500-marker density, inheritance probability was computed for each marker instead of simply assigning either parental haplotype; reliabilities averaged 70.3% when young animals were genotyped for 500 markers and both parents were genotyped for 50,000. Very high marker density can increase reliability slightly (1.4%), whereas low marker density allows breeders to apply cost-effective genomic selection to many more animals.

Key Words: reliability, marker density, genomic evaluation

621 Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. K. A. Weigel*[1], G. de los Campos[1], A. I. Vazquez[1], G. J. M. Rosa[1], D. Gianola[1], and C. P. Van Tassell[2], [1]*University of Wisconsin*, *Madison,* [2]*USDA-ARS*, *Beltsville, MD*.

The objective of the present study was to evaluate the predictive ability of direct genomic values for economically important dairy traits when genotypes at some single nucleotide polymorphism (SNP) loci were imputed, rather than measured directly. Genotypic data consisted of 42,552 SNP genotypes for each of 1,762 Jersey sires. Phenotypic data consisted of predicted transmitting abilities (PTA) for milk yield, protein percentage, and daughter pregnancy rate from May 2006 for 1,446 sires in the training set and from April 2009 for 316 sires in the testing set. The SNP effects were estimated using the Bayesian least absolute selection and shrinkage operator (LASSO) with data of sires in the training set, and direct genomic values (DGV) for sires in the testing set were computed by multiplying these estimates by corresponding genotype dosages for sires in the testing set. The average correlation across traits between DGV (before progeny testing) and PTA (after progeny testing) for sires in the testing set was 70.6% when all 42,552 SNP genotypes were used. When genotypes for 93.1, 96.6, 98.3, or 99.1% of loci were masked and subsequently imputed, mean correlations between DGV and PTA were 68.5, 64.8, 54.8, or 43.5%, respectively. When genotypes were also masked and imputed for a random 50% of sires in the training set, mean correlations between DGV and PTA were 65.7, 63.2, 53.9, or 49.5%, respectively. Results of this study indicate that a low density chip comprised of 3,000 equally spaced SNPs can provide approximately 95% of the predictive ability observed with the BovineSNP50 Beadchip (Illumina, Inc., San Diego, CA), but if fewer than 1,500 SNP are genotyped the accuracy of DGV may be limited by errors in the imputed genotypes of selection candidates.

Key Words: genomics, imputation, Jersey

622 Filling in missing genotypes using haplotypes. P. M. VanRaden*[1], J. R. O'Connell[2], G. R. Wiggans[1], and K. A. Weigel[3], [1]*Animal Improvement Programs Laboratory, ARS, USDA*, *Beltsville, MD,* [2]*University of Maryland School of Medicine*, *Baltimore,* [3]*University of Wisconsin*, *Madison*.

Unknown genotypes can be made known (imputed) from observed genotypes at the same or nearby loci of relatives using pedigree haplotyping, or from matching allele patterns (regardless of pedigree) using population haplotyping. Fortran program findhap.f90 was designed to

combine population and pedigree haplotyping. Each chromosome was divided into segments of about 100 markers each. Each genotype was matched to the list of currently known haplotypes sorted from most to least frequent for efficiency. If a match was found (no conflicting homozygote), any remaining unknown alleles in the found haplotype were imputed from homozygous genotypes. The individual's second haplotype was obtained by subtracting its first from its genotype, and the second was checked against remaining haplotypes. If no match was found, the new genotype (or haplotype) was added to the list. After completing population haplotyping, pedigrees were examined to resolve conflicts between parent and progeny haplotypes, locate crossovers that created new haplotypes, and impute haplotypes of nongenotyped ancestors from their genotyped descendants. One processor took 2 h to find haplotypes for 43,385 actual markers of 33,414 Holsteins. For the same population, time increased only to 2.5 h with 500,000 simulated markers but with 500 markers per segment. Computing time increased much less than linearly because most haplotypes were excluded as not matching after just the first few markers. Genotype storage required 13 GB for 500,000 markers, but haplotype storage required only 2.5 GB. Shared haplotypes were stored just once, and only index numbers were stored for individuals instead of full haplotypes. Paternal alleles were determined correctly for 95% of heterozygous markers, and linkage was determined correctly for 98% of adjacent pairs of heterozygous markers in simulated data. Population haplotyping correctly filled 95% of missing high density marker genotypes. Pedigree haplotyping can fill missing genotypes efficiently for nongenotyped ancestors or progeny with lower marker density.

Key Words: haplotyping, marker density, imputation

623 Use of haplotypes to predict selection limits and Mendelian sampling. J. B. Cole*, *Animal Improvement Programs Laboratory, ARS, USDA*, *Beltsville, MD*.

Limits to selection and Mendelian sampling terms can be calculated using haplotypes, which are sums of individual additive effects on a chromosome. Haplotypes were imputed for 43,385 actual markers of 3,765 Jerseys using the Fortran program findhap.f90, which combines population and pedigree haplotyping methods. Longer chromosomes had more distinct haplotypes, ranging from 7,287 for *Bos taurus* autosome 1 (BTA) to 2,460 for the X chromosome. This is expected because longer chromosomes undergo recombination more often than shorter ones. Mendelian sampling (MS) variances were calculated for genotyped animals as the sum of squared haplotype differences for each chromosome in the genome. The distribution of MS variances had a heavy right tail (skewness = 0.276), with a mean of 49,290 ± 13,981. Genotypes for each chromosome were constructed from pairwise combinations among the top 5% of haplotypes based on the sum of marker effects for lifetime net merit (NM) for each chromosome. Correlations among raw and adjusted values in the top group ranged from 0.897 on BTA12 to 0.998 on the X chromosome. Selection of the best unadjusted haplotypes for each chromosome results in an animal with an EBV of +$5,243 for NM. Adjusting for inbreeding resulted in a slightly lower EBV of +$4,496. Haplotype values were adjusted to account for changes in homozygosity by adding or subtracting 6% of an additive genetic standard deviation per 1% decrease or increase in homozygosity. The top Jersey bull, ALL LYNNS RESTORE VERNON-ET (29JE03647), had an EBV NM of +$1,180 in the January 2010 evaluation. For 11 chromosomes (BTA 4, 9, 13, 15, 20, 21, 22, 25, 26, 28, and X) the best genotype after adjusting for inbreeding consisted of 2 copies of the same haplotype. Differences between the best and poorest haplotypes ranged from a maximum of $65 for BTA1 to a minimum of $12 for BTAX. Selecting animals rather than chromosomes may result in slower progress, but limits may be the same because most chromosomes will become homozygous with either strategy. Selection on functions of MS could be used to change variances in later generations.

Key Words: genetic gain, haplotyping, mendelian sampling