Breeding and Genetics: Genomic Selection Methods II

635 Mating programs including genomic relationships. C. Sun*[1] and P. VanRaden[2], [1]*National Association of Animal Breeders, Columbia, MO,* [2]*Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD.*

Computer mating programs have helped breeders minimize pedigree inbreeding and avoid recessive defects by mating animals with parents that have fewer common ancestors. With genomic selection, breed associations, AI organizations, and on-farm software providers could use new programs to minimize genomic inbreeding by comparing genotypes of potential mates. Relationships could be computed between (1) only requested males and females via a web query, or (2) all genotyped females with only the marketed males (e.g., >200,000 females and >1,500 bulls), because (3) relationships between all >300,000 genotyped animals are difficult to store and transfer. To compare mating strategies, 50 marketed bulls in each of breed (Jersey and Holstein) were selected for top genomic Lifetime Net Merit (LNM), top traditional LNM, or randomly selected. The 500 youngest genotyped females in the largest herd were assigned mates of the same breed with limits of 10 females per bull and 1 bull per cow (for Brown Swiss, only 79 females and 8 bulls were included). Linear programming, a simpler method that assigned least related mates sequentially, and random mating were compared. For each method, calf value was the average of parents' genomic LNM plus the inbreeding loss times average of parents' expected future inbreeding, minus inbreeding loss time parents' genomic or pedigree relationship. A value of $23.11/1% was assumed for inbreeding loss for all mating methods. Compared with random mating, assigning mates using pedigree inbreeding gave only about 60% of the advantage of using genomic inbreeding for Holsteins, and the simpler mating strategy gave about 90% of the linear programming advantage. The economic value of a mating strategy that uses linear programming and genomic instead of pedigree inbreeding is already >$2 million per year for Holsteins and will grow as more females are genotyped. Eventually, dominance effects could also be included in mating programs to estimate inbreeding losses more precisely. Software to estimate dominance variance and to estimate the dominance effect for each SNP could allow mating plans to include both dominance effects and genomic inbreeding.

Key Words: linear program, dominance, inbreeding

636 Random regression and reaction norm extensions of whole genome prediction models accounting for genotype by environment interaction. W. Yang*, C. Chen, and R. J. Tempelman, *Michigan State University, East Lansing.*

Whole genome prediction (WGP) improves accuracy of the breeding values (BV) in livestock. However, these accuracies can be badly compromised when genotype by environment interaction (G × E) presents but is not accounted for. Reaction norm (RN) and random regression (RR) models have been proven to be useful in accounting for G × E by modeling BV as linear functions of environmental covariates. We extended these RR/RN models to infer upon SNP-specific intercepts and linear effects of environmental covariates. We considered several alternative specifications for modeling the distribution of the 2 × 2 variance-covariance matrices (VCV) of the SNP effects in WGP models: (1) independent inverted Wishart (IW) densities and (2) independent conjugate densities on the square root free Cholesky decomposition (CD) of the VCV. Three common extensions being specified were all SNP-specific VCV (BayesA-like), a mixture with a point-mass at zero (BayesB-like) and all SNPs having the same VCV (BayesC-like). Here we considered 5 of the 6 possible RR/RN models: IW-BayesC/IW-BayesA/IW-BayesB/CD-BayesA/CD-BayesB, and compare them to a conventional BayesA model. Based on 20 replicates, each involving around 2200 SNP markers and 2000 individuals in an RN simulation study, 3 scenarios based on an average genetic correlation between SNP-specific intercept and slope effects of 0, 0.5 and 0.8 were studied. In general, IW-BayesA had the highest accuracy under 3 scenarios although all 5 RN/RR based-methods demonstrated better performance in predicting BV than the conventional BayesA ($P < 0.0001$). In an RR application of a Duroc × Pietrain resource population at MSU, 2000 randomly chosen SNP markers and 324 F2 animals were analyzed for back fat thickness at wk 10, 13, 16, 19 and 20. RR-based methods have a 2.4% greater cross-validation accuracy ($P < 0.0001$) for predicting phenotypes compared with the conventional BayesA. We believe that when G × E presents, RR/RN extensions to WGP models are useful for improving accuracy of predicting genetic merit compared with current conventional approaches.

Key Words: whole-genome prediction, genotype by environment interaction

637 Exploring alternative specifications for whole genome prediction bivariate trait models. W. Yang*, C. Chen, and R. J. Tempelman, *Michigan State University, East Lansing.*

Multiple trait (MT) whole genome prediction (WGP) using high density single nucleotide polymorphism (SNP) marker panels may reap benefits for improving accuracy of selection and generation intervals. One current approach is based on specifying independent inverted Wishart prior densities (IW-BayesA) on SNP-specific variance-covariance matrices (VCV). We propose an alternative bivariate WGP model based on a modified Cholesky decomposition (CD) of the VCV as it potentially allows greater flexibility for modeling bivariate heterogeneity of genetic effects across SNP. We consider such a specification across all SNP (BayesA-like) and a specification that allows some SNP have null effects for either or both traits (BayesB-like). We refer to these 2 specifications as CD-BayesA and CD-BayesB. Univariate BayesA/B on Trait 1, univariate BayesA/B on Trait 2, and bivariate IW-BayesA, CD-BayesA and CD-BayesB on both traits, were compared using 20 replicates of simulated data derived from 2000 SNP markers and 500 animals and an average genetic correlation between the 2 traits of 0.5. Furthermore, QTL effects were generated from heterogeneous bivariate gamma densities based on 10 QTL for Trait 1 only, 10 QTL for Trait 2 only and 10 QTL for both traits. Heritabilities for Traits 1 and 2 were specified to be 0.5 and 0.1, respectively. For Trait 2, we found that bivariate WGP models had generally the highest average accuracy of breeding value compared with the univariate models. In particular, CD-BayesA had 8% greater accuracy than univariate BayesA for Trait 2 ($P < 0.0001$), and 5% higher accuracy than univariate BayesB ($P = 0.0145$). For Trait 1, IW-BayesA had a 2% lower accuracy ($P < 0.0001$) compared with univariate BayesA/BayesB, CD-BayesA and CD-BayesB, reflecting somewhat the inflexibility of the IW-BayesA approach. We applied these models to various phenotypes from the Wellcome Trust mice database, using 1787 animals and 1900 randomly selected SNP markers. Based on a cross-validation study, differences in predictive abilities between the competing models were rather heterogeneous depending on the magnitude of the genetic correlations.

Key Words: pleitropy, multiple trait, genetic correlation

638 Prediction of direct genomic values by using a restricted pool of SNP selected by maximum difference analysis. M. Cellesi[1], N. P. P. Macciotta[1], G. Gaspa[1], D. Vicario[2], P. Ajmone-Marsan[3], A. Stella[4], and C. Dimauro*[1], [1]*Dipartimento di Agraria, Sezione Scienze Zootecniche Universita' di Sassari, Sassari, Italy,* [2]*Associazione Nazionale Allevatori Razza Pezzata Rossa Italiana (ANAPRI), Udine, Italy,* [3]*Istituto di Zootecnica, Universita' Cattolica del Sacro Cuore, Piacenza, Italy,* [4]*CNR IBBA, Lodi, Italy.*

In the present research, a new technique able to select SNP-markers significantly associated with a particular trait (T) is proposed. Genotypes of 2,093 Italian Holstein and of 1,310 Simmental bulls were generated with the Illumina's 50K BeadChip. Phenotypes used were deregressed proofs for milk, fat and protein yield. Animals were ranked according to T. Then, the best 100/80 (B) and the worst 100/80 (W) were selected for Holstein and Simmental, respectively. For each SNP, frequency of the 3 genotypes (2 homozygote and one heterozygote) were calculated. Finally the genotype with the highest frequency in B animals (f(B)) was found and compared with the frequency of the same genotype in W (f(W)). The difference f(B)-f(W) represented the measure that gives the name of the method, maximum difference analysis (MDA). A bootstrap procedure was implemented to derive a posterior probability distribution used to declare a SNP positively associated with T. Markers negatively associated with T (i.e., with the maximum genotypic frequency in W) were also detected. Direct genomic values (DGV) were predicted with a BLUP procedure using both MDA-selected or all SNP available (40,780 and 49,870 for Simmental and Holstein, respectively). DGV accuracies were higher with the MDA selected SNP than with all original markers (in parentheses), particularly for Simmental (around 15% on average) (Table 1). These results suggest that a customized assay containing only the MDA selected SNPs could be developed to genotype animals thus reducing costs and computational resources.

Table 1. Number of MDA selected SNP and DGV accuracies obtained with the selected SNP and with all original markers (in parentheses)

	Holstein			Simmental		
	Milk	Fat	Protein	Milk	Fat	Protein
MDA selected SNP	763	557	823	155	177	217
DGV accuracies	0.45 (0.43)	0.51 (0.41)	0.38 (0.39)	0.35 (0.20)	0.39 (0.27)	0.41 (0.24)

Key Words: genomic selection, SNP selection

639 Using identifiability of genetic causal effects as a criterion for covariate choice in genome-enabled selection models. B. D. Valente*, G. J. M. Rosa, D. Gianola, and K. A. Weigel, *University of Wisconsin-Madison, Madison.*

In applications of genome-enabled selection models to study relationships between genome-wide markers genotypes and a trait, it is common to use other phenotypic traits as model covariates. In this study, we demonstrate that in the context of animal breeding, the choice of model covariates is not a purely statistical problem, and that poor decisions in this regard may result in misleading interpretation of inferences. As an example, consider a scenario where trait A is affected by variable G representing genome-wide genotypes. Suppose that A is also affected by a trait B that is not heritable (i.e., G does not affect B). This scenario may be represented by a causal model structured as G→A←B, which suggests that although B is independent of G, conditioning on A renders them associated to each other. Therefore, if one proposes a genome-enabled selection model to study the trait B as a function of genome-wide markers genotypes but decides to use A as a covariate, inferences would indicate a relationship between G and B, even though B is not heritable. Although still allowing genome-enabled predictions of phenotypes, the application of these for selection purposes would be misleading, as B could not be modified by selection. In selection, phenotypes are expected to change as a result of interventions on genotypes, so that the relevant information in this context is not just the statistical association between genotypes and phenotypes, but the causal relationship between them. However, inferences provided by the analysis described above are not relevant for breeding purposes because the expressed association between G and B does not reflect a causal relationship. We review requirements for identifying causal information from data. Considering different scenarios, we demonstrate that ignoring causal assumptions for the identifiability of genetic causal effects may lead to proposing models that may be useful for phenotypic predictions but not for selection. The use of graph criteria for identifying causal effects is suggested for the construction of genome-enabled selection models applied for breeding.

Key Words: genetic effect, genome-enabled prediction, selection

640 Assessing statistical properties of cSNP discovery and genotyping using RNAseq and genotyping chip data. P. D. Reeb*, C. W. Ernst, N. Raney, L. Preeyanon, T. Brown, R. O. Bates, and J. P. Steibel, *Michigan State University, East Lansing.*

A first step in allelic specific expression (ASE) testing using RNaseq data consists of discovering coding SNP (cSNP) and calling genotypes. In this work, we used genotypes from Illumina PorcineSNP60 Beadchip (SNP60) to estimate properties of cSNP calling and genotyping from RNaseq of skeletal muscle samples. Total RNA was extracted from longissimus muscle of 24 pigs genotyped with the SNP60. Individual RNA was reverse transcribed into cDNA, fragmented and labeled into barcoded libraries, sequenced on Illumina HiSeq 2000 (100 bp, paired-end reads). Reads were aligned to reference genome using TopHat. Total and allele-specific read counts at each SNP in the SNP60 set was obtained with mpileup of SAMTools. SNP60 positions to which a minimum number of reads (Rmin) were aligned were used to call cSNP with the VarScan program. Each genomic position from the SNP60 was classified as polymorphic or monomorphic in the sample. Monomorphic sites were used to estimate cSNP false discovery rate (FDR) and polymorphic sites allowed estimating sensitivity (proportion of segregating SNP called as cSNP). For cSNP discovery (Table), sensitivity increased (67% to 96%) with Rmin (20 to 500). In contrast, FDR was constant (2%). Accuracy of estimated MAF also increased with Rmin, but the increase was explained by the genotype call rate (not shown in table). Contrastingly, calling heterozygous genotypes with low FDR required much larger coverage (Rmin > 200). Error in calling heterozygotes will have the greatest impact on estimation of ASE. These results emphasize the importance of studying properties of cSNP calling and genotyping for future eQTL applications.

Table 1.

						Heterozygous genotype rates	
Rmin	#SNP	Sensitivity	cSNP FDR	Genotype call rate	MAF	Sensitivity	FDR
20	3485	0.67	0.020	0.53	0.39	0.37	0.56
50	2312	0.79	0.019	0.68	0.53	0.48	0.31
100	1745	0.86	0.016	0.80	0.70	0.60	0.16
200	1318	0.93	0.016	0.90	0.82	0.72	0.05
500	951	0.96	0.018	0.97	0.94	0.87	0.01

Key Words: RNAseq, cSNP, pig

641 A robust Bayesian regression model for whole-genome analyses. K. Kizilkaya*[1,2], R. L. Fernando[1], and D. Garrick[1], [1]*Iowa State University, Ames,* [2]*Adnan Menderes University, Aydin, Turkey.*

Following the groundbreaking paper of Meuwissen et al. (2001) establishing BayesA and BayesB methods, Habier et al. (2011) extended these to BayesCpi and BayesDpi. The relationships among these models are established using 3 parameters; pi, scale and degrees of freedom. We propose an overarching model for whole-genome analysis, where pi, scale and degrees of freedom are treated as unknown. The models were compared using a simulation study carried out to examine the estimability of these parameters, and by applying them to de-regressed milk, fat, and protein yields, and somatic cell scores. A trait with heritability of 50% was simulated for 5,000 animals based on 50, 500 or 5,000 QTL randomly sampled from real SNPs from the 50k panel. The QTL substitution effects were sampled from t-distributions with 4 or 100 degrees of freedom. Phenotypic values of animals were generated for 5 reps by adding residuals from normal distribution to the sum of the QTL effects. Two sets of SNP genotypes were used for genome-wide analyses: only QTL genotypes (Set1) or all 50k marker except QTL (Set2). Estimates of degrees of freedom from Set1 and 2 converged to the true values within QTL scenarios. Estimates of pi from Set1 approached zero, indicating a BayesA model with low degrees of freedom, or BLUP model with high degrees of freedom. Estimates of pi from Set2 approached the true values, indicating a BayesB model with low degrees of freedom. Accuracies of genomic estimated breeding values from the robust model showed good agreement with those from BayesA, BayesB, BayesCpi or BayesDpi models.

Key Words: robust model, degrees of freedom, pi

642 Genome-wide analysis of case-control data using logit, probit and robit models. K. Kizilkaya*[1,2], R. L. Fernando[1], S. Kachman[3], and D. Garrick[1], [1]*Iowa State University, Ames,* [2]*Adnan Menderes University, Aydin, Turkey,* [3]*University of Nebraska, Lincoln.*

The threshold model using the probit link, is the most commonly used model for genetic evaluation of categorical traits. It has been recently extended to genome-wide analysis. However, the alternative logit and t (robit) links are preferred for many statistical applications. The logit or robit model can be computed by augmenting the joint posterior density with Logistic or t-distributed rather than normally distributed underlying variables. A simulation study was conducted to quantify accuracy of genomic prediction assuming probit, logit and robit models. A binary trait (full data) determined by 50 QTL with heritability 10, 25 or 50% was generated based on incidence rates of 0.01, 0.02, 0.05 or 0.10 using 2,250 purebred training animals. QTL were simulated by randomly selecting loci from 50k SNPs and assigning effects from a normal distribution. Underlying variables of 2,250 animals were generated by summing 50 QTL effects and by adding simulated Normal, Logistic or t distributed residuals. Case-control data were generated by selecting matching controls for every case. Simulations were replicated 10 times. SNP effects were estimated by BayesC with pi = 0.995 assuming Normal, Logistic or t distributed residuals. Accuracies of genomic estimated breeding values were calculated by correlating true and estimated genotypic values of animals. There was no substantial difference among accuracies from the logit, probit or robit models in analyses of either full or case-control data sets. However, case-control data resulted in about half the accuracies of full data. Accuracies increased as incidence rate and heritability increased.

Key Words: case-control, probit, logit

643 A structural model for genetic similarity in genomic selection of admixed populations. E. Hay*, S. Smith, and R. Rekaya, *University of Georgia, Athens.*

Current approaches for dealing with admixed and crossbred populations in genomic selection rely on using different groups of animals in training and validation sets. These approaches gain from increased power as results of increasing the size of the training set. However, they fail at different degrees depending on the genetic similarity between the sub-populations of the admixed population. Our proposed multi-compartment model where the effect of an SNP could be different between breeds and parameterized as a function of its effect on one of the breeds in admixed population through a one to one mapping function, was able to remediate some problems of the pooled data approaches but still suffer from the high dimensionality of the unknown parameters to estimate. To overcome this problem, we propose not to estimate a mapping parameter α for each SNPi rather to build a model for α as a function of information already available in the genotype data via a hierarchical structural model. In this study, α was modeled as a function of the change in minor allele frequencies across lines and potential change in linkage phase. An admixed population consisting of 2 breeds was simulated. Each breed consisted of 2000 individuals genotyped for 50K SNPs and measured for a quantitative trait with 0.40 heritability. Genetic dissimilarity was simulated mainly by changing SNP minor allele frequencies between the 2 breeds. Three analyses were conducted: 1) classical pooled data (M1); 2) pooled data using the multi-compartment model and α for each SNP (M2); and 3) pooled data using multi-compartment model and our structural model for α (M3). For M1, accuracy (correlation between EBVs and GEBVs) was 0.54. The accuracy increased to 0.66 using M2 very likely due to a better accounting for the genetic dissimilarity between the 2 breeds. When a structural model was assumed (M3) the accuracy dropped to 0.63. This small decrease compared with M2 indicates that it is possible to model α as a function of the information already available in the genotype data with little impact in accuracy but with a substantial reduction in the number of parameters to estimate.

Key Words: genomic selection, admixed population