Breeding and Genetics: Genomic Selection Methods I

546 Mixed model methods for genomic prediction and estimation of variance components of additive and dominance effects using SNP markers. Y. Da* and S. Wang, *Department of Animal Science, University of Minnesota, St. Paul.*

Mixed model methods for joint genomic prediction and estimation of variance components for additive and dominance effects using SNP markers were developed based on the quantitative genetics model that partitions a genotypic value into breeding value and dominance deviation. Two sets of formulations were developed for genomic BLUP (GBLUP) and genomic REML (GREML) estimation of variance components of additive and dominance effects using SNP markers. Set 1 of GBLUP and GREML formulations is based on the conditional expectation of breeding values and dominance deviations given phenotypic observations with fixed effects estimated by the best linear unbiased estimator (CE). The CE set of formulations applies to both cases of ‘$q< m$’ and ‘$q>m$’ and applies to cases with singular genomic additive and dominance correlation matrices, where q = number of individuals and m = number of SNP markers. Set 2 of GBLUP and GREML formulations is based on mixed model equations (MME). GREML formulations based on MME are computationally more efficient for the case of ‘$q>m$’ and are less efficient for the case of ‘$q<m$’ than the CE formulations. Reliability formulations were derived for GBLUP of breeding values, dominance deviations and genetic values as summation of breeding values and dominance deviations. GREML is an effective tool to assess the exact type of genetic effects and assess the genetic contribution of the whole genome or targeted chromosome regions and genes to the phenotypic variance. GBLUP of total genetic value that includes additive and dominance effects provide prediction of an individual’s total genetic potential.

Key Words: genomic prediction, variance component, dominance

547 GVCBLUP 2.1: A computing package for genomic prediction and estimation of variance components for additive and dominance effects using SNP markers. C. Wang*[1], D. Prakapenka[2], S. Wang[1], H. B. Runesha[2], and Y. Da[1], [1]*Department of Animal Science, University of Minnesota, St. Paul,* [2]*Research Computing, The University of Chicago, Chicago, IL.*

GVCBLUP is designed for variance component estimation and genomic prediction for additive and dominance effects using SNP markers. Computing speed of GVCBLUP 2.1 increased by about 10 times running on single-core Windows desktops and increased by about 50 times running on 2~4 core Windows desktops using OPENMP. This new version has 3 programs: GREML_CE, GREML_QM, and GCORRMX. The GREML_CE and GREML_QM programs combined the 24 GREML and GBLUP programs in the previous version. GREML_CE is based on the conditional expectation of breeding values or dominance deviations given the phenotypic observations and applies to full-rank and singular genomic additive and dominance relationship matrices, and GREML_QM is based on mixed model equations and is designed for $q > m$, where q = number of individuals and m = number of markers. These 2 programs calculate GREML estimates of variance components of additive effects, dominance effects and random residuals, calculate additive and dominance heritabilities as well as heritability in the broad sense, calculate GBLUP of breeding values, dominance deviations and genetic values as summation of breeding values and dominance deviations for individuals in training and validation data sets, and calculate reliability of each GBLUP. Option is available to calculate GREML and GBLUP for additive effects only or dominance effects only, and for using any of the 3 definitions of genomic additive and dominance relationship matrices. GCORRMX is for calculating genomic additive and dominance relationships for 3 definitions.

Key Words: genomic prediction, variance component, dominance

548 Estimating dominance SNP effects using alternative single-step type genomic prediction equations. N. Gengler*, *ULg-GxABT, Gembloux, Belgium.*

Recently a new and alternative derivation of single-step type genomic prediction equations allowing joint estimation of GEBV and SNP effects based on the partitioning of genetic (co)variances was developed. The method was derived from a random mixed inheritance model where SNP and residual polygenic effects were jointly modeled. The derived equations were modified to allow non-genotyped animals and to estimate directly and jointly GEBV and SNP effects. Several other advantages of the new equations were that weighting of SNP and polygenic effects becomes explicitly and that SNP effects were also estimated simultaneously. This method makes better use of High-Density SNP panels and can be modified to accommodate other type of genetic effects. In the present study modifications of the equations were developed to allow the estimation of dominance SNP effects even if not all animals are genotyped and parental sub-class effects are used. Previous research done to estimate dominance effects were not very successful, the rationale being that they were hindered by the weakness of dominance information. However, by estimating dominance SNP effects and subsequently dominance GEBV, the estimation and the exploitation of specific combining abilities would become finally feasible. Recently, by genotyping animals heterozygosity of a given SNP locus is now precisely known. Therefore, through the direct use of this information with these alternative equations dominance SNP effects can be estimated and used to exploit specific combining abilities of given combinations of animal genomes. Finally, though this development the flexibility of these alternative equations combining advantages of single-step and of explicit SNP effect estimation methods to accommodate other types of genetic effects is shown.

Key Words: dominance effect, single step, genomic prediction

549 A comparison of hidden Markov-based imputation algorithms when applied to livestock data. K. Dhakal*[1], J. M. Hickey[4], A. Kranis[3], M. A. Cleveland[2], and C. Maltecca[1], [1]*North Carolina State University, Raleigh,* [2]*School of Environmental and Rural Science, Armidale, NSW, Australia,* [3]*Aviagen Ltd., Midlothian, United Kingdom,* [4]*Genus plc., Hendersonville, TN.*

Several of the pedigree-free imputation algorithms are hidden markov model (HMM) based approaches that approximate the coalescent and capture linkage disequilibrium information. Some pedigree-free algorithms have found use in livestock, mostly because their performance is reasonable, they are easy to use, and because in some instances pedigree information is not present or suitable in livestock data sets being imputed. The objective of this study was to compare the performance of HMM imputation algorithms in several typical livestock genotyping scenarios with different structures where reference and test panels were different sizes. The data set included genotypes for pigs, and dairy cattle. Three in-silico low-density panels were constructed with densities equivalent to 6,065 (L6k), 3,022 (L3k), and 384 (L384) SNP across the entire genome. Four popular software packages fastPHASE, MaCH (minimac), Impute2, and Beagle were used for imputation, and imputation accuracies were evaluated. Differences in accuracies were found among imputation algorithms and across scenarios, with MaCH (minimac) giving higher accuracy (R-squared >0.85) when L6k and L3k panels were used for both pig and cattle data sets. Accuracy was higher when larger reference sets and test animals in L6k and L3k panels were used. Computational time also varied across scenarios with the MaCH (minimac) algorithm overall being the fastest. In general, the results obtained are helpful in guiding the selection of imputation algorithms for different imputation scenarios and livestock data.

Key Words: genotype imputation, pedigree-free, accuracy

550 Effect of genotype imputation on genome-based prediction of complex traits: An empirical study with mice data. V. P. S. Felipe*[1,2], G. J. M. Rosa[1], H. Okut[3], D. Gianola[1], and M. A. Silva[2], [1]*University of Wisconsin-Madison, Madison,* [2]*Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil,* [3]*University of Yuzuncy Yil, Van, Turkey.*

High-density molecular marker panels have been used in animal and plant breeding for prediction of genetic merit of selection candidates. The prediction models often contain thousands of SNPs, which are fitted simultaneously using shrinkage-based estimation approaches. Quite often, because of cost constrains only a subset of the SNPs are genotyped in the selection candidate population, and genotype imputation methods are applied. The goal of this study was to evaluate the effect of genotype imputation on prediction accuracy of phenotypes. The hypothesis underlying this work was that all genetic signal and information available in a data set is contained entirely on the observed genotypes. A publicly available data set on mice was used, with information of 1,809 SNPs equally spaced along the genome of 1,881 animals. The traits considered were body weight and body mass index. And, from the full set of SNPs, only 201, 453 or 905 were selected as the genotyped SNPs, with the remaining marker imputed using the Beagle software. Then, Bayesian Lasso (BL), reproducing kernel Hilbert spaces (RKHS) and Bayesian regularized artificial neural networks (BRANN) were fitted using the subsets and the full panel of SNPs before and after genotype imputation. RKHS method showed the best predictive accuracy. Genotype imputation seemed to have the same effect on efficiency of BL and RKHS, whereas BRANN resulted in more sensible predictions due to imputation error. In scenarios which genotype imputation accuracy was good and masking rates of 75% and 50%, the genotype imputation did not bring great benefit. However, when genotype information was sparse (90% masking), genotype imputation brought information about important markers and improved predictive ability. The obtained results show that not always the imputation of unknown genotypes is advantageous for phenotypic prediction. The gain of imputing genotypes will depend on the connectedness between reference population and selection candidates, heritability of the trait, number markers available in the original panel, and the method used to predict marker effects.

Key Words: genomic selection, imputation, predictive ability

551 A fast expectation maximization antedependence model for whole genome prediction. C. Chen*, H. Wang, W. Yang, and R. J. Tempelman, *Michigan State University, East Lansing.*

As whole genome prediction (WGP) becomes based on even higher density single nucleotide polymorphism (SNP) marker panels, computational efficiency becomes a greater consideration such that inference strategies other than Markov chain Monte Carlo (MCMC) might be important. Two such popular alternatives are genomic best linear unbiased prediction (GBLUP) and BayesA,B/LASSO like methods based on the use of the expectation maximization (EM) algorithm. A primary limitation of these models is based on the specification of SNP effects being independently distributed, even though one might anticipate sizeable correlation between effects of SNP in close proximity to a major causal variant. Our group has previously developed such a model based on a first order antedependence covariance structure between adjacent SNP, while basing our inference strategy on MCMC. We have demonstrated that modeling this type of non-stationary correlation improves accuracy of breeding value (BV) prediction compared with models assuming independent SNP effects. We have developed a computationally tractable EM analog of this antedependence model that we dub EM-anteBayesA. In a simulation study involving 30 replicates, each involving just over 1000 SNP markers in linkage disequilibrium (LD) with average pairwise LD $r^2 = 0.30$, we compared EM-anteBayesA with a more conventional EM-based BayesA as well as more conventional implementations of anteBayesA and BayesA based on the use of MCMC. Although the EM-based methods tended to lead to slightly less accuracy in BV prediction than their MCMC counterparts, they were extremely competitive computationally thus rendering them to be tractable alternatives. Specifically, EM-anteBayesA demonstrated significantly higher accuracies than conventional EM-BayesA ($P = 0.02$). We also demonstrate the 4 models/inference strategies on the publicly available Wellcome Trust heterogeneous stock mice data. We conclude that EM-anteBayesA is a promising alternative for improving accuracy of WGP compared with other computationally efficient WGP implementations.

Key Words: computational efficiency, genomic selection, expectation maximization (EM)

552 Unknown-parent groups and incomplete pedigrees in single-step genomic evaluation. I. Misztal*[1], Z. Vitezica[2], A. Legarra[3], I. Aguilar[4], and A. Swan[5], [1]*University of Georgia, Athens,* [2]*Université de Toulouse, Castanet-Tolosan, France,* [3]*INRA, Castanet-Tolosan, France,* [4]*INIA, Las Brujas, Uruguay,* [5]*University of New England, Armidale, Australia.*

In single-step genomic evaluation using best linear unbiased prediction (ssGBLUP), genomic predictions are calculated with a relationship matrix that combines pedigree and genomic information. For missing pedigrees, unknown selection processes, or inclusion of several populations, a BLUP model can include unknown-parent groups (UPG) in the animal effect. For ssGBLUP, UPG equations also involve contributions from genomic relationships. When those contributions are ignored, UPG solutions and genetic predictions can be biased. Several options exist to eliminate or reduce such biases. First, mixed model equations can be modified to include contributions to UPG elements from genomic relationships (greater software complexity). Second, UPG can be implemented as separate effects (higher cost of computing and data processing). Third, contributions can be ignored when they are relatively small but they may be small only after refinements to UPG definitions. Fourth, contributions may approximately cancel out when genomic and pedigree relationships are constructed for compatibility; however, different construction steps are required for unknown parents from the

same or different populations. Finally, an additional polygenic effect that also includes UPG can be added to the model (slower convergence rate). Chosen options need to reflect different origins of UPGs: missing pedigrees in a closely selected population, multiple breeds, external lines or combinations of origins. Incomplete pedigrees may also cause biases and convergence problems even when UPGs are not in the model. In such cases, choices include restoration or truncation of pedigrees. Severity of problems with UPG and incomplete pedigrees greatly depends on the population structure. The problems are small in large purebred populations that include many high-accuracy sires (e.g., in dairy). The problems are larger in multi-line/multi-breed populations especially with few high-accuracy animals (e.g., in sheep).

Key Words: bias, genomic evaluation, unknown-parent group

553 Efficient inversion of a large genomic relationship matrix stored on a disk using a multi-core processor and graphic processing units.

Y. Masuda* and M. Suzuki, *Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Japan.*

The objective of this study was to develop new software for quick computation of the inverse of a large genomic relationship matrix stored on a disk by a hybrid system with graphic processing units (GPUs) and multi-core central processing unit (CPU). The matrix was split into submatrices called “panels,” whose elements were stored on a solid state drive (SSD). The process of inversion was described as a set of multiplications and additions between panels. The panels were loaded into main memory and updated if required. The updated elements were immediately written back to the disk. The optimized BLAS libraries, OpenBLAS and CUDA BLAS, were employed for matrix operations. Some computations on GPUs and accesses with the file were parallelized by OpenMP. Our software was written in Fortran 2003 and compiled with GFortran 4.7.2. The program were benchmarked on a computer with Intel Core i7–3770 (quad-core 3.4GHz), 32GB main memory, and NVIDIA GeForce GTX 580 with 1.5GB RAM, running Linux (x86_64). When the matrix had 50,000 of the order and it stored on the disk, the computing time for the inversion was 5.6 min in single precision and 15.0 min in double precision arithmetic. When enough memory was available on GPUs, the computing time was reduced by approximately 30% in single precision and 10% in double precision arithmetic. Although, the matrix was stored on the disk, our implementation completed the inversion 1.8 times (single precision) or 1.3 times (double precision) faster than a system where all data were loaded into main memory and processed by the optimized LAPACK subroutine (DPOTRF/DPOTRI) with a multi-core CPU only. The inversion for a matrix of 110,000 order in single precision (or 80,000 in double precision) was completed within 1 h. This technique is especially useful when the number of genotypes is up to 200,000 because the inverse of genomic relationship matrix can be directly obtained and used for the calculation of genomic predictions and their reliabilities without modifications to existing software.

Key Words: computing, software, GPU