Breeding and Genetics Symposium: Relevance of modeling in the genomics era

38 Is complex modeling important in the age of genomic selection? Guilherme J. M. Rosa*, *University of Wisconsin, Madison, WI.*

Statistical methodology has always played a fundamental role in modern animal breeding and genetics. For example, regression and ANOVA techniques have been developed and applied extensively in the context of estimation of genetic parameters and prediction of genetic merit for complex traits. Later, linear mixed model approaches such as best linear unbiased prediction (BLUP) and residual maximum likelihood (REML) estimation of (co-)variance components became prevailing in the analysis of pedigreed data, given their flexibility to accommodate unbalanced data, complex genetic relationships and overlapping generations. Several extensions of mixed models techniques have also been applied in animal breeding, such as the analysis of binary and count data, growth curves, survival models, and gene mapping in outbred populations. These complex models have been frequently implemented using Bayesian and MCMC techniques, facilitated by recent advances in computing technology. More recently, accessibility to genomic technologies has allowed fine mapping of causative loci, high throughput functional genomics studies, and whole-genome prediction of complex traits in livestock species. However, advancements in genomic technologies have also brought several new challenges from data-storage and data-mining standpoints, given the dimensionality of current data sets. Nowadays, not only efficient computer algorithms are required for data storage and data management, but also carefully tailored data mining tools are essential to deal with issues of multiple testing, potential of over fitting, spurious associations, and nonlinearities and complex interactions inherent to genomic data. In this presentation I will review some of the contemporary statistical and data mining methods currently used in animal breeding and genetics, for both prediction and causal inference, with especial emphasis on mixture regression models and graphical models, and the incorporation of biological knowledge into the analyses. Through some examples, I will illustrate the importance of complex modeling in the age of genomic selection.

Key Words: statistical models, genomic data, animal breeding

39 BLUP, REML, and other tools in the age of genomic selection. Esa A. Mäntysaari* and Martin Lidauer, *Natural Resouces Institute Finland, Green Technology, Jokioinen, Finland.*

Onset of genomic selection changed the focus of animal evaluation experts into estimation of genomic breeding values (GEBV). This was because of enormous potential of genomic information, but also because of similar intellectual challenges in methodologies. Still, also GEBV rely on phenotypes as a source of information. The GEBVs and the ordinary estimated breeding values (EBV) have the same need of well-defined models to attain accurate and unbiased results. Milk and component yield EBVs can illustrate the value of accuracy. Although $EBV_{protein}$ or EBV_{fat} can be used as a indirect estimates of EBV_{milk} (correlations to milk EBV 0.91 and 0.79), the $GEBV_{milk}$ trained on protein (fat) gave validation reliability of 0.27 (0.10), while the training on milk gave $R^2 = 0.45$. In this presentation we discuss particulars of breeding value estimation models and approaches for estimation of variance components (VC). The examples used are the joint Nordic test day model and the multiple trait-across country (MACE) model. The Nordic test day model is used to evaluate bulls and cows in Finland, Sweden and Denmark for 4 breeds: Holstein, Red Dairy Cattle, Jersey and FinnCattle. The challenges are the varying production conditions and admixed populations. The model is a multilactation, multitrait (milk, protein, fat) random regression. In every evaluation run the heterogeneity of variance is estimated for each trait and herd-year. During the implementation, VC for 1,827 parameters were estimated for each country and breed combination. Estimation was done using Monte Carlo REML and EM-algorithm. Most genomic evaluations rely on genotype exchange and MACE results for training. Accuracy of MACE depends on the assumed correlations across countries. Currently Interbull estimates correlations among breeding values of bulls from 31 countries, 6 breeds and 40 traits. The largest single VC estimation is for the Holstein production traits involving all countries. The challenges are computing, and the lack of genetic ties among smaller countries. Current estimation is by subsets of countries, but another alternative is to use MC REML and all countries simultaneously.

Key Words: breeding value estimation, variance components

40 Practical implications for genetic modeling in the genomics era for the dairy industry. Paul M. VanRaden*, *Animal Genomics and Improvement Lab, Agriculture Research Service, USDA, Beltsville, MD.*

Genetic models convert data into estimated breeding values and other information useful to breeders. The goal is to provide accurate and timely predictions of the future performance for each animal (or embryo). Modeling involves defining traits, editing raw data, removing environmental effects, including genetic-by-environmental interactions and correlations among traits, and accounting for nonadditive inheritance or nonnormal distributions. Data included phenotypes and pedigrees during the last century and genotypes within the last decade. Genomic data can include markers, haplotypes, and causative effects such as insertions, deletions, or point mutations; most models also include polygenic effects because the markers do not track causative variants perfectly. Total numbers of known variants have increased rapidly from thousands to hundreds of thousands to millions. Nonlinear models add precision for traits influenced by major genes, but linear models work well for traits with more normally distributed genomic effects. Numbers of genotyped animals in US dairy evaluations increased rapidly from a few thousand in 2009 to about 1 million in 2015. Most are young females that will contribute to estimating allele effects in the future, but only about 100,000 have phenotypes so far. Traditional animal models may become biased by genomic preselection because Mendelian sampling of phenotyped progeny and mates is no longer expected to average 0. Single-step models that combine pedigree and genomic relationships can account for such selection, but approximations and new algorithms are needed to avoid excessive computation. Traditional animal models may include all breeds and crossbreds, but most genomic evaluations are still computed within breed. Inclusion of inbreeding, heterosis, dominance, and interactions can improve precision. Multitrait genomic models may be preferred for traits with many missing records or when foreign records are included as pseudo-observations, but most countries use multitrait traditional evaluations followed by single-trait genomic evaluations. A final goal is to explain how the models work so that breeders can more confidently apply the predictions in their selection programs.

Key Words: genetic evaluation, genomic selection, mixed models

41 Experiences in bioinformatics. Luc L. Janss*, *Aarhus University, Tjele, Denmark.*

Knowledge from bioinformatics research can in principle be used to improve genomic predictions. Examples are use of the QTLdb database with collected QTL mapping results, the use of genomic feature annotations, and pathway or Gene Ontology (GO) data. There are, however, several hurdles to appropriately use this information and include it in genomic models for analysis of livestock data. A limitation in the use of QTL mapping results is that results from classical linkage analysis studies have wide confidence intervals such that for any trait a large part of the genome will be tagged as "QTL region". Despite this limitation, SNPs in QTL regions can be found to explain more variance than those outside QTL regions, which is for instance shown for milk and fat yield in dairy cattle, but not for protein yield. A limitation of genomic feature annotations is that this information is not covering the whole genome and is not directly related to traits, or, for pathway or GO data, is mostly categorized in fundamental biological processes. This makes it difficult to attach genuine prior information to this data. The common approach to use this kind of bioinformatics data is to simply try which features or GO groups explain more variance or predict better. This leads to results such as that SNPs in/near genes (but not necessarily from the coding parts of genes) explain more variance in phenotypes. A final significant hurdle to use this information is that our animal populations are highly structured and include relatives. This leads to long-range LD and LD between chromosomes, which effectively spreads QTL effects over the whole genome. This creates a polygenic image of the trait architecture, which matches the assumptions of GBLUP that all SNPs contribute equally, and makes genomic relationships the main driver for genomic prediction within animal populations. Better modeling, notably a separation of effects of linkage and LD in genomic prediction, allows more meaningful inferences from bioinformatics data, and potentially allows to improve genomic predictions where relationships are weak; for example, across breed.

Key Words: bioinformatics, genomic feature, genomic prediction

42 Practical implications for genetic modeling in the genomics era for the beef industry. Andy D. Herring*, *Texas A&M University, College Station, TX.*

The beef cattle industry is based on valuation of phenotypes, and its supply chain components vary across global region. In many areas, producers maintain ownership of animals until sale to an abattoir, yet in many other regions distinct industry segments (cow-calf, grower/stocker, finisher/feedlot, packer) exist where animal ownership changes across segment. Commodity cattle (of unknown genetic and/or management background) have different value potential as compared with cattle with known background. The utility of genomic data and analyses are, and will remain, different regarding these 2 types of cattle. Many genomic approaches have calculated molecular breeding values of animals; however, most beef industry managers other than seedstock producers would much rather have predicted phenotypes (predictions that could be collectively based on breeding value, non-additive genetic value, and environmental value) for improved management and marketing decisions. Current US beef industry trends show increasing carcass (and mature cow) weight and carcass quality grade, but static incidence of respiratory disease in feedlots, and no improvement in percent calf crop weaned in beef herds. Trends for more prevalent and less costly genomic data will also continue. There appears to be large potential to genomically characterize beef cattle for production-related physiological systems (health, growth, body composition, fertility, nutrient utilization) as well as potential interactions for optimal economic management and production system assessment. Better understanding of these systems and their components will require knowledge beyond the DNA sequence including RNA regulatory elements and products and protein function and structure; the roles of fetal programming and epigenetics on economically important traits in beef production remain largely unknown and need investigation. Resource cattle populations with detailed phenotypes and banked biological samples, and that evaluate multiple components of beef cattle production systems remain critically important; partnerships of industry groups and research institutions can assemble large, informative data sets. The incorporation of genomic data into economic assessments is also encouraged.

Key Words: genetic modeling, beef cattle production systems