Breeding and Genetics: Computational Issues in Genomic Analysis

521 Genomic selection using low-density SNPs. D. Habier, J. C. M. Dekkers*, and R. L. Fernando, *Department of Animal Science and Center for Integrated Animal Genomics*, *Ames, IA*.

Genomic selection (GS) using high-density single nucleotide polymorphisms (HD-SNPs) is promising to improve response to selection. GS is based on estimating effects of HD-SNP alleles using phenotypes in a training data set and using these estimates to obtain GS-EBV for selection candidates based only on HD-SNP genotypes. Genotyping HD-SNPs for all selection candidates may not be cost effective in most livestock species. Thus, we propose to use low-density SNPs (LD-SNPs) to trace chromosomal segments from parents to progeny with sufficient accuracy and estimate GS-EBV of selection candidates. This approach requires inferring HD-SNP haplotypes of the training individuals using parental HD-SNPs. Given estimates of HD-SNP effects, effects of chromosomal segments are estimated as the sum of the HD-SNP effects on that haplotype. To estimate GS-EBV in subsequent generations, only LD-SNPs flanking each segment are utilized to follow the inheritance of segments by estimating the probability of descent of a segment (PDS). To test this approach, a training data set of 1000 individuals with 1000 SNP genotypes on a 10 Morgan genome and phenotypes for an additive trait controlled by 100 QTL with heritability 0.5 was simulated. HD-SNP haplotypes of training individuals were assumed known and divided into 100 segments of 10 cM, flanked by a total of 110 LD-SNPs that were genotyped in subsequent generations. HD-SNP allele effects were estimated by Bayes-B. An MCMC sampler was used to first obtain joint probabilities of segregation indicators for every flanking marker pair. These were then used to calculate the PDS as the probability of maternal origin at the midpoint of flanking LD-SNPs. The correlation between true breeding values and GS-EBVs was used to evaluate the loss in accuracy from using LD-SNPs vs. HD-SNPs in generations following training. The accuracy was 0.66 with HD-SNPs and 0.63 with LD-SNPs in the first generation, dropped to 0.63 and 0.60 in the second, and to 0.61 and 0.56 in the third generation following training. Thus, LD-SNPs can be used for selection candidates with limited loss of accuracy which, depending on the costs of LD-SNPs vs. HD-SNPs, can result in a substantial reduction in costs.

Key Words: Genomic Selection

522 Effects of allele frequency estimation on genomic predictions and inbreeding coefficients. P. M. VanRaden[1], M. E. Tooker*[1], and N. Gengler[2,3], [1]*USDA Animal Improvement Programs Laboratory*, *Beltsville, MD*, [2]*Gembloux Agricultural University*, *Gembloux, Belgium*, [3]*National Fund for Scientific Research*, *Brussels, Belgium*.

Genetic calculations often require estimating allele frequencies, which differ across time due to selection and drift. Data were 50,000 simulated markers and 39,985 actual markers for 2391 genotyped Holstein bulls. Gene content of relatives and gene frequencies in the base (founder) population were estimated using pedigrees and a linear model. Ancestors born since 1950 were included, for a total of 22,088 animals. Because pedigrees were very complete, only one unknown-parent group was used. Convergence to 5 digits of accuracy required about 1000 iterations. Total time was 2 processor days and proportional to number of animals times markers, but actual clock time was reduced by processing loci on separate chromosomes in parallel. Simple allele frequencies were obtained from only the known genotypes. True base frequencies were correlated with estimated base frequencies by 0.98 versus 0.94 with simple frequencies. Genomic predictions and inbreeding coefficients were computed in four ways, using true or estimated base frequencies, simple frequencies, or an “estimate” of .5 for each marker. When allele frequencies estimates were used instead of 0.5 to assign mixed model coefficients, solutions converged more slowly but predictions were more accurate. From simulated data, realized reliabilities for young bulls were 62.8% using either true or estimated base frequencies, 62.6% using simple frequencies, and 62.0% using frequencies set to 0.5. Pedigree and genomic inbreeding coefficients were correlated by 0.73 using true base frequencies, 0.67 using estimated base frequencies, 0.12 using simple frequencies, and 0.72 when frequencies were set to 0.5. Genomic inbreeding coefficients were biased downward by 7% to 9% using either frequency estimate, upward by 31% using 0.5, but were reasonable when true frequencies were used. Frequency estimation had small effects on genomic predictions but large effects on genomic inbreeding coefficients in both simulated and real data.

Key Words: Genotype, Genomic Selection, Allele Frequency

523 Strategies to incorporate genomic prediction into population-wide genetic evaluations. N. Gengler*[1,2] and P. M. VanRaden[3], [1]*Gembloux Agricultural University*, *Gembloux, Belgium*, [2]*National Fund for Scientific Research*, *Brussels, Belgium*, [3]*USDA Animal Improvement Programs Laboratory*, *Beltsville, MD*.

Most current research on genomic selection is focusing on the accurate prediction of genomic breeding values. However selection solely based on genomic breeding values, despite being theoretically promising, is in practice only suboptimal for several reasons. The two most important are that only few animals are genotyped therefore having genomic prediction directly available and that rankings will change. With genomic breeding values potentially available in the near future, strategies are required to avoid any confusion in the mind of users. The aim of this study is to present three different strategies that could be used to incorporate genomic prediction into population-wide genetic evaluation. The three strategies are: 1) using selection index theory to combine both sources of information into a single set of breeding values; 2) for ungenotyped animals, compute conditional expectation of gene contents for SNP given molecular and pedigree data and use these predicted gene contents; and 3) integrate genomic breeding values as external information into genetic evaluation using a Bayesian framework. If strategy 1) is straight forward, additional steps have to be done to adjust breeding values for changes in those of relatives. A practical implementation is to use reliabilities of the genomic prediction, the population-wide genetic evaluation PA, and PA from the genotyped subset to set up a 3 x 3 matrix for each animal, with off-diagonal elements being functions of the 3 reliabilities. The use of strategy 2) is computationally much more challenging but leads directly to the needed covariance structures combining genomic relationship if known with pedigree relationships. Strategy 3) is potentially a good compromise because the theory is well established and has already been used in beef cattle to incorporate external breeding values. Also current genetic evaluation software can be easily modified to incorporate genomic breeding values.

Key Words: Genomic Prediction, Incorporation, Breeding Value Estimation

524 Selection of single nucleotide polymorphisms and genotype quality for genomic prediction of genetic merit in dairy cattle. G. R. Wiggans*[1], T. S. Sonstegard[1], P. M. VanRaden[1], L. K. Matukumalli[1,2], R. D. Schnabel[3], J. F. Taylor[3], F. S. Schenkel[4], and C. P. Van Tassell[1], [1]*ARS, USDA*, *Beltsville, MD*, [2]*George Mason University*, *Manassas, VA*, [3]*University of Missouri*, *Columbia*, [4]*University of Guelph*, *Guelph, ON, Canada*.

A process to prepare high-density genotypic data for use in genomic prediction was developed. Marker genotypes from >51,000 single nucleotide polymorphisms (SNP) were generated for 3,139 Holstein bulls on the Illumina Bovine SNP50 chip. The SNP were grouped by minor allele frequency (MAF); 10,249 SNP with a MAF of <5% were excluded. Number of SNP for each of 45 MAF categories was uniform (800 to 1,009). Hardy-Weinberg equilibrium was assessed by comparing observed to expected heterozygosity for each locus. For 6 SNP assigned to chromosome 7, no bulls were heterozygous, which confirms the latest assembly that places those SNP on the X chromosome. Observed heterozygosity was within 2% of that expected for 96% of SNP. Linkage between adjacent autosomal SNP was analyzed to determine if the data set could be reduced for downstream analysis. For 1,237 pairs of adjacent SNP, marker genotypes were either both homozygous or both heterozygous (<10 bulls differed for each pair), and the first SNP from each pair was excluded; mean physical distance between those SNP pairs was much smaller (37 kb) than between 39,386 autosomal SNP (64 kb). Sire and son data for 2,566 bulls with 204 genotyped sires were compared to validate sample identification and Mendelian inheritance. For those bulls with >100 inheritance errors, correct sire was determined through comparison with other sires of sons. For sons with the correct sire, 99.99% of SNP with genotypes did not conflict. Comparison of genomic and pedigree relationships detected 3 members of a clonal family, a set of identical twins, and some possible pedigree errors. Genotyping consistency was investigated for 9 bulls genotyped twice and for the twins and clones. Most differences were caused by an inability to determine the genotype for one of the paired SNP; however, one clone had 24 SNP conflicts (99.94% concordance). Although evaluation of the SNP set is ongoing, only minor changes are expected for the final set. This largest set of high-quality SNP data for Holsteins to date should provide the basis for successful genomic prediction.

Key Words: Genomic Prediction, Genotyping, Single Nucleotide Polymorphism

525 Analysis of high dimension marker data in the presence of gene interactions: A machine learning approach. K. R. Robbins, J. K. Bertrand, and R. Rekaya*, *The University of Georgia*, *Athens*.

In recent years there has been much focus on the use of single nucleotide polymorphisms (SNP) for the fine mapping of genomes in an effort to identify causative mutations and important genomic regions for traits of interest; however, many studies focus only on the marginal effects of markers, ignoring potential gene interactions. Simulation studies have show that this approach may not be powerful enough to detect important loci when gene interactions are present. While several studies have examined potential gene interaction, they tend to focus on a small number of SNP markers. Given the prohibitive computational cost of modeling interactions in studies involving a large number SNP, methods need to be developed that can account for potential gene interactions in a computationally efficient manner. This study adopts a machine learning approach by adapting the ant colony optimization algorithm (ACA), coupled with logistic regression on haplotypes and genotypes, for association studies involving large numbers of SNP markers. The proposed method is compared to genotype (GA) and haplotype analysis, implemented using sliding windows (SW). Each algorithm was evaluated using a binary trait simulated using an epistatic model and HapMap ENCODE genotype data. Two simulations scenarios, varying the strength of the epistatic relationship, were replicated five times each. The ACA yielded increases in the power to detect genomic regions associated with the simulated trait of 66.7 % over the next best method in both simulated scenarios. Based on these results it is clear that methods accounting for potential gene interactions are necessary to obtain good power for association studies examining complex traits under the control of interacting genes.

Key Words: Ant Colony Algorithm, Machine Learning, Single Nucleotide Polymorphism

526 Statistical design of validation studies for transcriptional profiling experiments. J. P. Steibel*[1], R. J. Tempelman[1], and G. J. M. Rosa[2], [1]*Michigan State University*, *East Lansing*, [2]*University of Wisconsin*, *Madison*.

Microarrays and quantitative reverse transcription polymerase chain reaction (qRT-PCR) are the most commonly used techniques for transcriptional profiling in animal tissues. Microarrays are commonly used as a first stage screening step, followed by a qRT-PCR experiment intended to validate the results from the first stage. While the design of microarray experiments has been extensively studied, the design and analysis of qRT-PCR validation studies has not. We address this issue as it pertains to the qRT-PCR validation of genes that are concluded to be differentially expressed based on results from a previous microarray experiment. Required sample sizes and expected significance levels for the validation experiment were determined, assuming that biological replicates are independent from those used in the microarray experiments. The level of replication was set to the minimum necessary to control the false discovery rate (FDR) at a certain level while maximizing the power or sensitivity of the overall experiment. We show that the number of replicates depends on the ratios of the FDR and the sensitivity between the two experiments as they depend on the true effect sizes (true mean differences divided by standard deviations). Our results also indicate that the traditional P-value thresholds of 0.05 or 0.01 for statistical significance are potentially too stringent for a second-stage validation experiment. In particular, if the FDR of the microarray experiment was controlled at 30% and the effect sizes are moderately large (i.e., near 1.0), setting the significance level of the validation experiment to 0.1 can control FDR at 5% while attaining greater than 80% power. Additionally, we consider the case were the same samples are used for the validation and microarray experiments (technical validation). We conclude that re-using the same samples in both stages invariably leads to a reduced power and increased FDR compared to the use of independent biological replicates. This increase in FDR is particularly large when the correlation between the two tests is high.

Key Words: Microarrays, qRT-PCR, Experimental Design

527 Model selection in gene-specific mixed linear models for microarray data with application to joint analysis of multiple experiments. L. Qu, N. Bacciu*, D. Nettleton, and J. C. M. Dekkers, *Iowa State University*, *Ames*.

Detecting differential gene expression using microarrays often suffers from low statistical power and accuracy due to small sample sizes and/or high variation. Model selection is hence an important task to obtain accurate significance levels and to improve power, in particular for joint analysis of multiple datasets. We propose four methods that use information across all genes to select one mixed linear model to be fit separately to data from each gene in a microarray experiment. We compare the results of these model selection methods on the joint analysis of two related microarray experiments in pigs. Method 1 is based on the information criteria (e.g., AIC) averaged over all genes to select a model that balances complexity and fit. Method 2 is based on cross-validation which selects the model that minimizes squared best linear unbiased prediction (BLUP) residuals, standardized and averaged over all genes. Method 3 is a multi-gene graphical method that uses principle component analysis on the residuals/BLUPs to compare distributional differences for a candidate factor. Method 4 is similar to method 3, but uses the multiresponse permutation procedure to formally test differences in residuals/BLUPs, ignoring the dependency among subjects. These four methods were applied to joint analysis of two random block design experiments, where the primary question was whether the variances can be pooled across experiments to increase power. Results show that the model that pooled both block and error variances and the one that only pooled block variances performed similarly and were preferred by each method, whereas separate analysis and the model that only pooled error variances were disfavored by each method. Although the methods generally agreed well for this data set, simulations are needed to further investigate which method is preferred and under which situations. In summary, our proposed methods can be useful tools for model selection in microarray analysis, especially when joint analysis of multiple datasets is used to increase statistical power. (Supported by USDA-NRI-2005-3560415618.)

Key Words: Microarrays, Model Selection, Mixed Models

528 Reconstruction of metabolic pathways for the cattle genome. S. Seo* and H. A. Lewin, *Institute for Genomic Biology, University of Illinois*, *Urbana*.

Metabolic reconstruction of microbial, plant and animal genomes is a necessary step toward understanding the evolutionary origins of metabolism and species-specific adaptive traits. The aims of this study were to annotate the recently sequenced cattle genome using a metabolism-centered approach, to reconstruct conserved metabolic pathways, and to identify metabolic pathways with missing genes and proteins. The MetaCyc database and PathwayTools software suite were chosen for this work because they are widely used and easy to implement. An amalgamated cattle genome database was created using the NCBI and Ensembl cattle genome databases (based on build 3.1) as data sources. PathwayTools followed by comprehensive manual curation were implemented for reconstruction of metabolic pathways. The curated database, CattleCyc 1.0, consists of 217 metabolic pathways. A total of 64 mammalian-specific metabolic pathways were modified from the reference pathways in MetaCyc, and two pathways previously identified but missing from MetaCyc were added. Comparative analysis of metabolic pathways revealed the absence of mammalian genes for 22 metabolic enzymes whose activity was reported in the literature. We also identified six human metabolic protein coding genes for which the cattle ortholog is missing from the sequence assembly. CattleCyc is a powerful tool for understanding the biology of ruminants and other cetartiodactyl species. In addition, the approach used to develop CattleCyc provides a framework for the metabolic reconstruction of other newly sequenced mammalian genomes. Having multiple annotated mammalian genomes hosted in BioCyc will facilitate comparative analysis of metabolic pathways among different species and a systems approach to comparative physiology.

Key Words: Genomics, Systems Biology, Metabolic Reconstruction