genetic standard deviations was 0.86, 0.64, and 0.58 with realized, equal, and base population allele frequencies, respectively. With scaling by averages, which is currently used in ssGBLUP, bias was 0.07, 0.08, and 0.03, respectively. With automatic scaling, bias was 0.18 regardless of allele frequencies. Accuracies were similar among scaling methods, but about 0.10 lower in the scenario without scaling. The GEBV were more inflated without any scaling, whereas the automatic scaling performed similarly to the scaling by averages. When μ_g was treated as random, with the variance equal to the mean difference between pedigree and genomic relationship matrices, the bias was the same as with the scaling by averages. The automatic scaling is biased, especially when μ_g is treated as a fixed effect and populations underwent strong selection.

Key Words: compatibility between genomic matrices, genomic selection, scaling

203 Partitioning SNP heritability in related individuals. J. Jiang*[1], P. M. VanRaden[2], L. Ma[3], and J. R. O'Connell[4], [1]*Department of Animal Science, North Carolina State University, Raleigh, NC,* [2]*Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD,* [3]*Department of Animal and Avian Sciences, University of Maryland, College Park, MD,* [4]*Department of Medicine, University of Maryland School of Medicine, Baltimore, MD.*

Partitioning SNP heritability by many functional annotations has been a successful tool for understanding the genetic architecture of complex traits in human genetic studies. Similar analyses are being extended to animal research, as (imputed) whole-genome sequence data of many individuals and various functional annotations have become available in livestock animals. Though many approaches have been developed for heritability partition (e.g., linkage disequilibrium score regression [LDSC] and Haseman-Elston regression [HE-reg]), they are mostly based on approximations tailored to human populations and few can produce statistically efficient estimates for animal genomic studies where individuals are often related. To tackle this issue, we present a stochastic MINQUE (Minimum Norm Quadratic Unbiased Estimation) approach for partitioning SNP heritability, which we refer to as MPH. We provide a theoretical analysis comparing LDSC and HE-reg with REML and MPH and demonstrate what LDSC and HE-reg (and similar methods) take advantage of in their approximations: sparse relationships between individuals and relatively weak linkage disequilibrium (LD). We also show that our method is mathematically equivalent to the Monte Carlo REML approach implemented in BOLT. MPH has 3 key features. First, it is comparable to genomic REML in terms of accuracy, while being at least one order of magnitude faster than GCTA and BOLT and using only ~1/4 of memory as much as GCTA, when applied to sequence data and many variance components (or functional annotation categories). Second, it can do weighted analyses if residual variances are unequal (such as DYD). Third, it works for many overlapping functional annotations. Using simulations based on a human pedigree and a dairy cattle pedigree, we illustrate the benefits of our method for partitioning SNP heritability in pedigree-based studies. We also demonstrate that it is feasible to efficiently partition SNP heritability for animal genomes with strong, long-span LD. MPH is freely available at https://jiang18.github.io/mph.

Key Words: SNP heritability, partition, functional annotation

204 Scalable mixed model approach for mapping omnigenic core genes. J. Jiang*[1], P. M. VanRaden[2], L. Ma[3], and J. R. O'Connell[4], [1]*Department of Animal Science, North Carolina State University, Raleigh, NC,* [2]*Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD,* [3]*Department of Animal and Avian Sciences, University of Maryland, College Park, MD,* [4]*Department of Medicine, University of Maryland School of Medicine, Baltimore, MD.*

To make use of exploding genomic data, we present a scalable mixed model approach for genome-wide association studies (GWAS) that can handle millions of genotyped animals, which we refer to as SSGP. Using simulations, we show that our method is as accurate as EMMAX and is a few times faster than BOLT. SSGP can address the genomic inflation issue in large-scale GWAS in domestic animals. Substantial genomic inflation will arise in GWAS in the presence of polygenic inheritance, even when population structure and relatedness have been accounted for. This can be demonstrated by the non-centrality parameter (NCP) for a SNP that is in linkage disequilibrium (LD) with causal variants, $NCP_i \approx N\Sigma_j r_{ij}^2 q_j^2$, where NCP_i is the NCP for SNP i, r_{ij} is the correlation coefficient between SNP i and causal variant j, q^2 is the proportion of phenotypic variance explained by a causal variant, and N is the sample size. If polygenic effects are not well accounted for, the NCP for many tested SNPs may be big in large-scale GWAS, especially those in domestic animals that generally have small effective population sizes and strong, long-span LD on the genome. As a result, we may see significant loci everywhere on the genome, even though any causal variant alone has an undetectable effect. We illustrate this phenomenon by leave-one-chromosome-out (LOCO) GWAS with big cow data and simulations, given the fact that LOCO GWAS does not account for effects of causal variants on the same chromosome as a tested SNP. We also illustrate that only a few loci of significance can be found when whole-genome polygenic effects have been accounted for by SSGP. This finding is in line with the omnigenic core versus peripheral gene model that was recently proposed: the few SSGP significant loci correspond to core genes and those LOCO significant loci everywhere on the genome result from peripheral genes that each have a tiny effect. In summary, our method is useful for finding omnigenic core genes that matter in functional studies and targeted genome editing. SSGP is freely available at https://github.com/jiang18/ssgp.

Key Words: mixed model association, biobank-scale, omnigenic

205 Accounting for X chromosome and allele frequencies in genomic inbreeding estimation. J. P. Nani* and P. M. VanRaden, *ARS-USDA, Beltsville, MD.*

Breeders for many decades used pedigrees to limit increases in inbreeding (pedF), but genomic measures of relationship and inbreeding can provide more precise control. Previous calculations ignored influence of the X chromosome (BTX) when estimating relationships. For mating programs, excluding BTX can cause an increase in inbreeding by mating 2 individuals with the same BTX. Numbers of SNP markers on BTX have increased recently in US national evaluations. The X-specific region has 3.0% of the 79,060 SNPs used and those are coded as 100% homozygous in males, causing homozygosity of females to appear 3% less than males. Allele frequency also has an impact on computing genomic inbreeding (genF). Correlations were high between genF computed using 0.5 ($genF_1$) or base population ($genF_2$) allele frequencies for most but not all breeds. Also, average genF was higher for males than females. The $genF_1$ was further adjusted for BTX ($genF_X$) to obtain better correlations across breeds and sex and to make $genF_X$ more similar to pedF. Haplotype-based inbreeding (hapF) was also estimated for comparison. Future inbreeding was estimated as half an animal's relationship to a recent reference population using pedigree (pEFI), 0.5 ($genFI_1$) or base frequencies ($genFI_2$) or adjusted for BTX markers ($genFI_X$). Definitions were compared for 3,280,753 genotyped animals of 5 breeds using a

pedigree file of 86,924,013 animals. Smaller breeds were more sensitive to the use of different allele frequencies. Correlations with pedF were generally higher using $genF_1$ or $genF_X$ compared with $genF_2$ (average correlations across breeds of 0.67, 0.67 and 0.54 respectively). Correlations with EFI were similar using $genFI_1$, $genFI_2$ or $genFI_X$ (average correlations across breeds of 0.83, 0.84 and 0.83, respectively). Correlations of hapF with pedF were not higher than of $genF_1$ or $genF_x$ with pedF (0.64). Use of $genF_X$ and $genFI_X$ did not affect the correlations within sexes but did improve the mean differences occurring between sexes. The adjustments allow simpler and more accurate comparisons of genomic and pedigree relationships.

Key Words: inbreeding, genomics, relationships

206 Identifying family clusters within the US Holstein population to manage genetic diversity. Y. Steyn*[1], T. Lawlor[2], Y. Masuda[1], D. A. L. Lourenco[1], S. Tsuruta[1], and I. Misztal[1], [1]*University of Georgia, Athens, GA,* [2]*US Holstein Association, Brattleboro, VT.*

Reproductive technology has allowed a few bulls or families to have major genetic contributions on a population. Differences in frequency of alleles that are identical by descent between predominant families may be an indication of multiple paths, i.e., genetic redundancy, leading to a similar polygenic response. The objective of this study was to identify family clusters and their key founders within the Holstein population. Sires with the greatest number of progeny born after 1985 were selected. Final data included 1,145 genotyped sires from 7 countries with birth year ranging from 1962 to 2009. Number of progeny per sire varied from 312 to 49,146. A principal component analysis using the genomic relationship matrix identified founders for 7 potential families. Clustering analysis using k-means was used to separate animals into 5 clusters (C1 to C5) containing 171 (C1), 252 (C2), 200 (C3), 244 (C4), and 278 (C5) animals, respectively. Five clusters allowed for the major progenitors of the Holstein breed to be identified with high genetic relationships within cluster. The 2 most predominant families were C1 and C2. Two families were grouped together in C3, several families primarily used in Canada were in C4, and multiple families used in the US were in C5. While C3 and C4 contained almost no bulls born before 1985, older bulls occurred more in C1 and C5. Indirect genomic predictions (IGP) were obtained for 5 type traits. The family clusters were ranked based on these traits. Rankings were similar for all traits. Differences in trait means were largely due to groups peaking in popularity at different times. Increasing the number of clusters to 10 allowed the predominant sons within a family to be identified. Results from this study suggest that k-means clustering can be used to identify the most influential family groups. Further work will apply this clustering method to the current group of genomic tested selection candidates with the aim of having low within-cluster and high between-cluster variance. This can lead to greater across-family selection and a reduction in loss of genetic diversity while still improving the breed performance.

Key Words: genetic redundancy, optimal selection