Breeding and Genetics Symposium: Really Big Data: Processing and Analysis of Very Large Datasets

197 High performance computing and really big datasets: Overview and best practices. F. Foertter*, *Genus plc, Hendersonville, TN.*

More often than not, researchers today find themselves overwhelmed with increasingly large data sets. It has become difficult to develop efficient workflows to edit and then analyze large data sets to extract meaningful results. Whereas desktop computers have sufficed in the past, data sets involving high density SNP or sequence data are growing so large that new high performance computing (HPC) methods are under development to allow efficient data housing, searching and analysis. This presentation will provide scientists an insight on how Genus is using HPC to manage large data sets in both research and genomic evaluation. Topics will include a review of hardware choices, including power/temperature and scalability considerations, and data storage and density. Also important are the financial and security considerations between owned clusters, university and national laboratory clusters, and commercial options such as pay-as-you-go clouds. Software issues will also be addressed, including the advantages and disadvantages of commercial versus Open Source and discussion on building in-house codes. Relevant setup scenarios and industry best practices related to Genus' implementation will also be presented. Finally we will demonstrate how Genus is currently leveraging HPC to decrease time-to-results in research, increase accuracy in genomic evaluations, and therefore increasing the rates of genetic improvement.

Key words: computation, genomics, analysis

198 Data structures and visualization. J. B. Cole*, *Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD.*

Genomic tools for genetic improvement have been rapidly adopted in many livestock species over the past few years. This presents new challenges for data collection and management, as well as opportunities for analysis and presentation. The US national dairy database currently includes genotypes for 83,117 bulls and cows and 2,620 imputed dams representing 3 different densities and 4 chip versions. Storage requirements for these genotypes are modest, even when high-density (>500,000K) genotypes are imputed from lower densities. However, storage requirements for intermediate and results files for genetic evaluations are much more substantial, particularly when multiple runs must be stored for research and validation studies. Full-sequence data will be available at reasonable cost in the near future, and will require much more storage. The greatest gains in accuracy from genomic selection have been realized for traits of low heritability, such as fertility and longevity, and there is increasing interest in new health and management traits. In addition to data on novel traits, potentially useful economic and demographic information is being collected by on-farm computer and analytical systems. There is increasing interest in traits such as feed efficiency and resistance to climate change, but the collection of sufficient phenotypes to produce accurate evaluations may take several years, and high-reliability proofs for older bulls are needed to precisely estimate marker effects. As traits proliferate and the number of genotyped animals continues to grow increasingly sophisticated analytical approaches will be tractable. Machine learning algorithms may be useful in identifying previously unrecognized relationships among traits, and the analysis of genetic (co)variances among loci could help identify important gene networks. Improved visualization tools, particularly those capable of processing very large volumes of data in a reasonable amount of time, are needed to help better understand the results of analyses. The challenges and opportunities presented by growing amounts of phenotypic and genomic data are generally similar regardless of the species in question.

Key words: genomics, data structures, visualization

199 Computational challenges in genetic evaluation with really big datasets. I. Aguilar*[1] and I. Misztal[2], *[1]Instituto Nacional de Investigación Agropecuaria, INIA Las Brujas, Canelones, Uruguay, [2]Animal & Dairy Science Department, University of Georgia, Athens.*

Genomic selection poses new computational problems. Genotypes for each individual require a large amount of storage and this amount will increase with larger SNP chips and eventually with individual genome scans. Computations in genomic selection using this data seem to require even more computing power especially when large fractions of population will be genotyped. Looking back in the history of animal breeding, 2 choices exist, brute force or new theoretical developments. For example, storing large A matrices required massive computers and inverting those seemed impossible. Rules to create A^{-1} explicitly by Henderson made these computations trivial. Even with A^{-1}, creating the mixed models explicitly required large resources. The iteration on data algorithm decreased the required resources drastically. Developments in sparse matrix inversion and of the AI algorithm made fast REML a reality. The genomic selection is most likely no different. While the genomic data seems huge, use of larger SNP chips results in limited gains. Sampling for best subset of SNP as in BayesX is time consuming, but methods based on the genomic relationship matrix G seem as efficient especially with larger data sets. In fact, given G the genomic selection may be another BLUP in the form of single-step GBLUP where computations are not much greater than in regular BLUP. The limiting factor in ssGBLUP is constructing and inverting G for many genotypes. Careful programming makes these operations much less expensive. For example, a regular algorithm for creating G for about 14k genotyped individuals required about a day. After using custom libraries and exploiting parallel computing via OpenMP, the computing time was reduced to 15 min. It is possible that G can be made sparse for large number of genotypes and that the number of useful genotypes for prediction will be limited. Hardware improvements have resulted in machines with multiple cores, with much faster speed and bigger cache memory and with more memory. Nevertheless, for successful implementations of large genetic evaluations, improvements in methodology were as important as advances in computer power.

Key words: genetic evaluation, computing methods, genomic selection

200 The implementation of analysis of large data. M. Coffey*, *Scottish Agricultural College, Penicuik, Midlothian, UK.*

Developments in DNA based technologies have led to large amounts of genotype data being available for farmed livestock. This has created great excitement among those engaged in research since data equals papers. However, for those engaged in national genetic evaluations

the logistics of handling so-called large data sets creates unique challenges that generate little scientific interest. These challenges must be overcome to exploit these new technologies and must be overcome in a way that does not create disruption during the transition from conventional evaluations to genomic EBVs (GEBVs). What constitutes large is ill defined but data in the terabytes is now routinely available. It is impractical to throw out existing systems at a whim and business development must take place to generate revenue that stimulates adoption. Thus genetic evaluation centers need to adopt different computing strategies to account for genotype data within systems that run routinely month after month. Data storage cost is not a real issue but processing time is, especially as systems are developed that run in real time for farmers to decide which animals to genotype via web based services and receive GEBVs as a result. This paper will highlight the practical aspects of implementing genomic evaluations.

Key words: genomic breeding value, genetic evaluation