

JOURNAL OF ANIMAL SCIENCE

The Premier Journal and Leading Source of New Knowledge and Perspective in Animal Science

Really big data: processing and analysis of very large datasets

J. B. Cole, S. Newman, F. Foertter, I. Aguilar and M. Coffey

J ANIM SCI published online November 18, 2011

The online version of this article, along with updated information and services, is located on the World Wide Web at:

<http://jas.fass.org/content/early/2011/11/18/jas.2011-4584>



American Society of Animal Science

www.asas.org

Really big data: processing and analysis of very large datasets¹

J. B. Cole,^{*2} S. Newman,[†] F. Foertter,[†] I. Aguilar,[‡] and M. Coffey[§]

^{*}Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350; [†]Genus plc, Hendersonville, TN, 37075; [‡]Instituto Nacional de Investigación Agropecuaria, INIA Las Brujas, Las Piedras, Canelones, Uruguay, 90200; [§]Scottish Agricultural College, Easter Bush Campus, Midlothian, UK, EH25 9RG

¹ Based on presentations at the Breeding and Genetics Symposium titled “Really Big Data: Processing and Analysis of Very Large Datasets” at the Joint Annual Meeting, July 10-14, 2011, New Orleans, LA. This symposium was sponsored, in part, by the European Association for Animal Production and Genus plc.

² Corresponding author: john.cole@ars.usda.gov

ABSTRACT: Modern animal breeding datasets are large and getting larger, due in part to recent availability of high-density single nucleotide polymorphism arrays and cheap sequencing technology. High-performance computing methods for efficient data warehousing and analysis are under development. Financial and security considerations are important when using shared clusters. Sound software engineering practices are needed, and it is better to use existing solutions when possible. Storage requirements for genotypes are modest, although full-sequence data will require greater storage capacity. Storage requirements for intermediate and results files for genetic evaluations are much greater, particularly when multiple runs must be stored for research and validation studies. The greatest gains in accuracy from genomic selection have been realized for traits of low heritability, and there is increasing interest in new health and management traits. The collection of sufficient phenotypes to produce accurate evaluations may take many years, and high-reliability proofs for older bulls are needed to estimate marker effects. Data mining algorithms applied to large datasets may help identify unexpected relationships in the data, and improved visualization tools will provide insights. Genomic selection using large data requires a lot of computing power, particularly when large fractions of the population are genotyped. Theoretical improvements have made possible the inversion of large numerator relationship matrices, permitted the solving of large systems of equations, and produced fast algorithms for variance components estimation. Recent work shows that single-step approaches combining BLUP with a genomic relationship (G) matrix have similar computational requirements to traditional BLUP, and the limiting factor is the construction and inversion of G for many genotypes. A naive algorithm for creating G for 14,000 individuals required almost 24 h to run, but custom libraries and parallel computing reduced that to 15 m. Large datasets also create challenges for the delivery of genetic evaluations that must be overcome in a way that does not disrupt the transition from conventional to genomic evaluations. Processing time is important, especially as real-time systems for on-farm decisions are developed. The ultimate value of these systems is to decrease time-to-results in research, increase accuracy in genomic evaluations and accelerate rates of genetic improvement.

Key words: computation, genetic evaluation, genomic selection, phenotypes, visualization

INTRODUCTION

From a scientific perspective, 'Big Data' is the point at which the amount of data to be analyzed is larger than the capacity of the available hardware and software to obtain results in a reasonable amount of time. Simply put, one has more data than one can handle in a pre-defined, discrete time step. For financial firms, Big Data means delivering trading-related data analysis at near-instantaneous intervals. At the European Organization for Nuclear Research, the Large Hadron Collider generates one petabyte (i.e., 10^{15} bytes) of data every second during a single experiment (ACM, 2011), most of which is filtered and immediately discarded because analysis of all of those data is infeasible.

Over the last 5 yr, the amount of information created globally has grown by a factor of nine, and the zettabyte (i.e., 10^{21} bytes) mark was surpassed for the first time in 2010. In a recent report (IDC, 2011), it was estimated that in 2011 alone 1.8 zettabytes of information would be created and replicated. At the same time, the cost of creating, capturing, managing, and storing data today is 1/6 of what it was in 2005 (IDC, 2011). This challenge is just as significant in the sciences as in financial services or other industries.

The primary challenge with Big Data in genetic improvement of livestock is that our ability to gather data far outpaces our ability to analyze it. With these thoughts in mind, the objective of this paper is to provide an overview of issues and technologies involved in the management and utilization of very large datasets. Specifically, a look at hardware and software technologies important to high performance computing, manipulation and visualization of very large datasets, the evolution of genetic evaluation programs to handle very large datasets, and the implementation of commercial genetic evaluation schemes.

DATA COLLECTION

Before calculations can be performed, phenotypes must be defined, observations collected, and their quality assessed. Incomplete or low-quality data may lead to inconclusive results or, more importantly, erroneous conclusions. Once a suitable collection of data has been assembled, they can be analyzed to identify patterns and associations among variables. When working with big data, none of these tasks are trivial. In this section, strategies for dealing with each of these problems are addressed.

Current and Novel Phenotypes

Information of interest varies across species, but observations generally are related to animal identification, production environment, and performance. This section will focus specifically on dairy cattle, but points applicable regardless of species will be emphasized. Accurate animal identification is necessary to associate phenotypes with genotypes. Performance data are recorded on the farm, where they are associated with an animal identification code, and records from herds enrolled in the national milk recording program are collected in the National Dairy Database (**NDDB**). Pedigree and conformation data are available for registered (i.e., purebred) cattle. Phenotypes may be placed into several broad groups, including milk volume and composition, reproduction, and health and fitness. The number of phenotypes can become very large (**Table 1**); for example, genetic evaluations are routinely calculated for 31 traits in U.S. Holsteins. Despite the broad range of phenotypes currently evaluated, interest in additional traits continues to grow.

In the U.S., SNP genotypes were now available for 116,980 Holstein cows and bulls as of August, 2011, including 43,525 3K, 72,311 50K, and 1,144 700K genotypes. Due to the cost of genotyping, only bulls and a few influential cows were initially genotyped, but the development of the Illumina Bovine3K BEAD chip (Illumina, Inc., San Diego, CA) has resulted in the genotyping of more than 42,500 cows, as of July, 2011 (Wiggans et al., 2011). In addition, an international project is currently underway to generate full-sequence data for 100 cattle representing many taurine and indicine breeds (T. S. Sonstegard, Bovine Functional Genomics Laboratory, USDA, ARS, Beltsville, MD, personal communication).

There are many data of interest that are not routinely recorded, such as information about farm and herd management, but which are currently partially available in the NDDB. There is growing interest in developing cattle that are robust in the face of global climate change, which will require improved understanding of genotype-by-environment interactions. The precise geographic location of animals, as well as climate data (e.g., temperature, relative humidity, and precipitation), is necessary for this research. Methodology for the evaluation and identification of heat tolerant animals using weather data has been reported (Misztal et al. 2010). Calus et al. (2005) have discussed some applications of detailed farm management information, such as housing systems and feeding programs. Feed intake is potentially useful as a proxy for energetic efficiency (Herd and Arthur,

2009), but phenotypes are currently expensive to collect. Therefore, methods to predict body energy status from routinely gathered milk spectral data have been explored (McParland et al., 2011). There also is growing interest in milk composition as it relates to human health, with recent efforts focusing on fatty acid composition (Caroli et al., 2009; Soyeurt et al., 2011).

Other data of interest are either not routinely recorded in on-farm computer systems, or are not transmitted through the milk recording system to the NDDDB. Of particular interest are traits related to animal health and fitness, such as disease incidence, which have been collected in some countries for many years. Cole et al. (2008) have developed a data exchange format for health event data for use in the U.S., and the International Committee for Animal Recording is currently finalizing an international standard (Egger-Danner et al., 2011). Body condition scores (Berry et al., 2003, Banos and Coffey, 2010), birth weights, and mature-weights have applications in herd management, and selection for desirable body condition scores may result in improved fertility, longevity, and lifetime profitability. Female fertility is probably best thought of as a complex of related phenotypes associated with the ability of a cow to cycle normally, display estrus, conceive upon breeding, and carry a pregnancy to term (Interbull, 2011). Some fertility data are included in the NDDDB but more complete recording of phenotypes is desirable, particularly of insemination events. Recording the use of estrous synchronization also may be helpful for avoiding bias in fertility evaluations.

Some technologies, such as radiofrequency ID (**RFID**) tags, are valuable because they improve accuracy of data recording by reducing error rates. New technologies also are enabling collection of many novel phenotypes. For example, many electronic milk meters are capable of recording information on milk yield and milking speed. Some systems are even capable of measuring electrical conductivity, progesterone levels, lactose content, temperature, somatic cell count, and fat and protein concentrations in real-time (e.g., AfiLab, S.A.E. AFIKIM, Kibbutz Afikim, Israel) . Pedometers have proven to be useful tools for monitoring changes in activity associated with estrous behavior and the onset of disease (Edwards and Tozer, 2004; Løvendahl and Chagunda, 2010).

Data Integration and Quality

There are many regular data providers in the U.S. dairy industry, including the national milk recording system, purebred cattle organizations, AI firms, and genotyping laboratories. For example, during the week of July 18 through 22, 2011, dairy records processing centers transmitted to the Animal Improvement Programs Laboratory (ARS, USDA, Beltsville, MD) 1,317,255 test-day records and 630,035 fertility events. Data also are occasionally received from universities under research agreements, and genotypes have been exchanged among the U.S. and several other countries.

Once the data have been received, they are integrated into a single, national dataset after undergoing an extensive series of edits to ensure data quality and consistency (Norman et al., 1994). Current implementation in software is approximately 64,000 source lines of code written in the C and Fortran languages, and the edited data are stored in a relational database. The fundamental principle of this process is that similar data obtained from separate sources must match, and they must exhibit biologically plausible values. For example, when a calving record is received for inclusion in the dystocia and stillbirth dataset, the data are compared to the calving date that initiated the dam's lactation, and the records are flagged if there is an inconsistency. Detailed error reports are sent to data providers so that records may be corrected. In some cases, records can be corrected automatically based on information from other sources, as in the case of erroneous parentage for genotyped individuals. Data also are examined to determine if values are biologically plausible, and abnormal observations may be adjusted (Wiggans et al., 2003). This process results in a national dataset that is reasonably error-free, but data extracted and used for genetic evaluation also undergo another set of edits.

Data quality has generally been assessed in terms of consistency; if data from a new source are consistent with data from other sources, then the new data are of high quality. This is an appealing concept because it is simple to implement, but does not provide metrics that can be used to objectively compare data collection methods. Research by Dechow et al. (2008) indicates that within-herd heritability estimates may be useful as indicators of data quality, and their calculation is relatively straightforward. Widespread use of RFID tags and electronic readers help to improve quality by reducing errors in identification. VanRaden et al. (2011) have shown that many parentage errors can be corrected using SNP genotypes.

Data Mining

The value of Big Data ultimately lies in their ability to inform decisions, whether it is the optimal service sire for a heifer mating or the proper diet for fresh cows. As we accumulate more and more data, the need grows for tools with which we can discover useful, possibly unexpected patterns in those data and report back to farmers in a timely and useful fashion. This can be done using tools for data mining (Tan et al., 2005) and data visualization (Cole and VanRaden, 2010). However, when working with big data, Bonferroni's principle (Shaffer, 1995) must be remembered: if you look hard enough for interesting patterns, you will find them. Many of these relationships will be spurious, and researchers must ensure that they have enough data to support the questions being asked.

Because some datasets are so large, most available software cannot accommodate all of the data at once. It is therefore important that an automated pipeline be developed to extract smaller subsets of the data through data mining. There are four principal tasks in data mining: 1) discovery of interesting relationships among variables in large datasets (i.e., association); 2) division of datasets into a number of discrete groups (i.e., clustering); 3) assignment of observations to groups (i.e., classification), and 4) prediction of real-valued outputs based on attributes of observational units (i.e., regression). Google, for example, developed MapReduce (Lin and Dyer, 2010) to process large amounts of raw web data into more manageable sets of key-value pairs. The MapReduce algorithm is actually a simple word-count algorithm that easily can be parallelized and scaled, and can be repurposed to process many different data types with very little effort.

Association analysis in a data-mining context (Aggarwal and Yu, 1998) is based principally on counting methods, which is not equivalent to the genome-wide association analysis (e.g., Maltecca et al., 2011). Clustering (e.g., Everitt et al., 2001) is used to separate items into distinct groups such that items within a group are similar to one another, and items in separate groups are dissimilar. In partitional clustering, items may belong only to a single group, such as when SNP genotypes are called. In hierarchical schemes, items are nested in a tree-like structure and may belong to multiple groups, and are often used to represent relationships among species. Classification models (e.g., Hand, 1997) use rules to assign individuals into classes based on their attributes, and typically involve training and validation steps. There are many classification methods, including

Bayesian belief networks, decision trees, nearest-neighbor classification, neural networks, rule-based classification, and support vector machines. All of these classification methods could be implemented as a MapReduce problem. The results could feed into regression models (e.g., Cook and Weisberg, 1999) that are well known in animal breeding, and focus on the prediction of real-valued outputs, such as breeding values, feed intake, or milk yield.

Visualization

Data mining and visualization are often viewed as complementary topics; data mining uses numerical approaches to discover relationships in data, while visualization provides a way of representing many numbers in a compact form while retaining the information in the data. Cole and VanRaden (2010) recently described some approaches for the graphical display of data from animal breeding programs. There is general agreement on best practices (e.g., Tufte, 1983) for presenting data, and a wide array of available software solutions (e.g., Wickham, 2009). Several examples will be used to demonstrate techniques for sound graphical design that are applicable to a wide variety of situations.

Figure 1 uses color to denote the chromosome on which a marker is located, and marker area is proportional to the magnitude of SNP effects. The three-generation pedigree for the Holstein bull Co-Op O-Style Oman Just-ET (001HO09167) is shown in **Figure 2**, and traces the occurrence of crossovers and the inheritance of individual haplotypes for chromosome 15 from generation to generation. A verbal or written description would occupy more space and provide less information. The line graph in **Figure 3** shows changes in average inbreeding from 1990 to 2010 for U.S. Brown Swiss, Holstein, and Jersey cattle. Breeds are differentiated from one another by both pattern and color, minimizing the risk that readers will confuse one series with another, particularly if the figure is later reproduced in grayscale. Tufte (2006) has proposed that small graphics be embedded in text and used similarly to words. **Figure 4** shows an example from Cole and VanRaden (2009) demonstrating how so-called sparklines might be used in an animal breeding context. Effective visualization will grow in importance as relationships among traits continue to increase in importance.

COMPUTATIONAL CHALLENGES

Big data pose a number of computational challenges that need to be addressed in parallel, much like big data themselves. On the hardware side, limitations are imposed by network bandwidth, component speeds, storage capacity, and the increasing complexity of modern central processing units, particularly when cost is a limiting factor. Some of these problems also can be attacked using high-performance programming languages, improved software engineering practices, and contributions from a broad-based community. Potential solutions to these problems based on recent experience are presented in this section.

Computing Infrastructure for Big Data

The increasing size of datasets exerts pressure on specific parts of the computing infrastructure. Moving a large file across a network is limited by available bandwidth and network connections speeds. Many institutions restrict connection speeds, sometimes to as little as 100 MB/s. At such speeds, it takes 10 min to move a 60 GB file, and almost 3 h to transfer a 1 TB file. A 1 GB/s network reduces the time required to move data, but other users of the network can be affected when the movement of large files saturates available bandwidth. One workaround is high-speed connections between machines engaged in moving big files while isolating them from the rest of the network by the use of switches. A second, decidedly more low-tech, option is to physically transfer external hard disks from one machine to another, avoiding the need to transmit data over the network. Limitations on the ability to move large data also place constraints on the special relationship between data and processing, which may favor cloud computing or put it at a disadvantage relative to local resources when Big Data has to be regularly refreshed. A more efficient but expensive solution is to use high-throughput distributed storage that can serve files at 10 Gigabit Ethernet (1.25 GB/s) or InfiniBand (250 MB·s⁻¹·channel⁻¹ to 3 GB·s⁻¹·channel⁻¹) speeds.

System Speed vs. Component Speed

The operational speed of a component is often used to describe how fast a system operates. However, the speed of the system as a whole is more important than that of individual components because the whole system is used for computation. Software development time also is affected by individual components, making

the choice of programming tools and the availability of skills important. One database engine may be better than another, and flat files and the C programming language may be fast, but all systems have limiting factors. It is increasingly the case that programmer skills and corporate information technology services are limiting factors. A very fast C program running on a very fast server with a slow network connection supported by a programming team that specializes in databases may not turn out to be fast in the long term.

Data Storage

Storage is perhaps the most fundamental problem with respect to Big Data facing the scientific community today. How much and where the data are stored depends on the stage of the research project, that is, retrieval, analysis, or archiving. These stages need not be distinct from one another, and they do not exclude other intermediary stages, such as data creation, pre-processing, cleaning, and mining. However, they do describe key stages in the data lifecycle that must be considered when working with large datasets. A researcher must at a minimum answer the following within the proper temporal scope: “How will I acquire, process, and archive the data?”

When planning for the storage of Big Data, it is not enough to rely solely on density. During analysis, disk speed, network speed, and network bandwidth also must be considered in the context of data-size and time-to-results requirements. Reading data from, and writing to, hard disks can have a high computational cost. On computers where random access memory (**RAM**) is limited, or where the job cannot fit entirely into available RAM, data are often paged in and out of main memory to a temporary storage area on disk. This process reduces computation speed because the central processing unit (**CPU**) spends many cycles on disk I/O. This can be accelerated by increasing memory or installing solid state disk drives (**SSD**), which have no moving parts, but possess very fast access speeds. These disks are currently expensive per GB and are best used tactically where they can be of greatest use. The use of SSD for temporary storage represents an ideal use case and can increase the performance of RAM-limited programs substantially.

The costs of post-analysis archiving of Big Data may be greater than costs of regenerating the data *de novo*. The current cost of standard consumer-grade disk storage is about \$21 per terabyte (S. Gilheany,

ArchiveBuilders.com, Manhattan Beach, CA, personal communication). However, more resilient storage is required for commercial services, and costs are typically 2 to 6 times more expensive for systems with built in redundancy. Unless the data will be frequently reused or shared, it may be more worthwhile to simply discard them, regenerating the data when necessary.

Software Development

Programming for Big Data means coding in such a way to distribute the work over many processing units working in parallel to reduce time-to-results. The amount of programming effort depends on many factors, including hardware architecture, whether or not existing code requires library updates, and the need to port or modify legacy code. Some problems also may be I/O-, memory-, or CPU-bound, and optimal solutions may require code that specifically accounts for those restrictions, as well as the use of different hardware. It is important that code be built to scale as data size increases or hardware capacity increases. Target hardware and programming language choices can greatly reduce the need for ongoing programming efforts as data increases in size.

Each hardware architecture has unique advantages and disadvantages that should be considered relative to programming effort. For example, private clusters allow for rapid deployment of existing open source tools at the cost of managing all of the hardware. Cloud services can be used to expand computational capabilities on an *ad hoc* basis, but there may be few guarantees relative to hardware availability or security. Graphical processing units (**GPU**) are very efficient hardware, but they are memory-limited and hardware-specific coding requirements can substantially increase programming efforts.

Mature open source mathematical libraries optimized for parallel processing are available for most traditional scientific computing languages (e.g., Fortran and C). An important advantage in using such languages is the availability of documentation and technical support. Both languages are commonly used to program supercomputers, in part because they are relatively easy to learn, and because they can be optimized for very fast execution. In contrast, Google's search engine results are provided mostly by software written in high-level languages, such as Java and Python. High-level languages typically support greater abstraction and

provide extensive built-in libraries that reduce overall development time, and recent advances in interpreter technology provide those benefits with small performance penalties.

The Open Compute Language (**OpenCL**) is an example of a programming language that provides high-level tools for parallel programming, but also exposes hardware details as needed. OpenCL uses a data/task-parallel programming model that can be cross-compiled for different hardware architectures, allowing programs to function regardless of hardware upgrades or increases in data size. This avoids unnecessary duplication of effort when developing applications if, for example, the analysis moves from a private cluster to a GPU to a cloud service. More recently, new hybrid languages such as Clojure, PyCUDA, PyOpenCL, Scala, and Go, all feature abstraction, high performance, and hardware portability. In addition, they were created specifically to target Big Data problems running on multiple-core systems, as well as reduce programming effort.

Utilization of New Technology

In animal breeding, many computations are matrix-based and involve the multiplications and inversion of very large arrays (i.e., matrices). These are often a bottleneck in terms of speed, consuming the majority of computational time in any process. Modern graphics cards are becoming very powerful at specific tasks and can run many calculations in parallel on their GPU. This has led to the availability of a programming tool set and hardware platform that can be programmed directly and called from Fortran and C++ programs. The Compute Unified Device Architecture (**CUDA**; NVIDIA Corporation, 2011) enables animal breeders to use legacy Fortran or C/C++ code and offload certain computationally heavy subsections on to the GPU for faster solving. In some instances, increases in speed are at least an order of magnitude and, importantly, GPU are cheap.

Computing with Graphics Hardware

Modern video cards are based on GPU that are highly parallelized, and that processing capacity can be accessed through CUDA. The combination of GPU and CUDA is particularly powerful when applied to vector and matrix operations. Coffey and co-workers (M. Coffey, T. Krzyzelewski, K. Moore, and R. Mrode, unpublished data) have successfully refactored a problem involving multiplication of a large matrix of real numbers into a form that can easily be processed on GPU. Suppose that $\mathbf{C} = \mathbf{A}\mathbf{A}'$, where \mathbf{A} is a 7,072-by-47,280

matrix of floating point numbers. The problem may be broken down into blocks for parallel processing as:

$$\begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} = \begin{bmatrix} A_1 A_2 + B_1 C_2 & A_1 B_2 + B_1 D_2 \\ C_1 A_2 + D_1 C_2 & C_1 B_2 + D_1 D_2 \end{bmatrix}$$

where A_1, \dots, D_1 are submatrices of A , and A_2, \dots, D_2 are submatrices of A' . This implementation results in much faster processing time by using GPU instead of CPU. The inversion of large matrices also is a common problem in genetic and genomic evaluations, and research is underway to develop a system in which a CPU-side process will determine GPU availability and then break-down matrices into suitably sized blocks for piecewise inversion. This should allow for the inversion of any matrix in a way that will utilize all available computing resources, either locally or in a cluster setting.

Creating a Community

Astronomers have pioneered over-wire data collection and dissemination. In the 1980s, two breakthroughs plunged the discipline into the world of Big Data: the internet and charged-coupled devices (Janesick, 2001). Astronomers went from analyzing large photographic plates to downloading high-resolution images from space-based telescopes. Far more data became available than could be analyzed by individuals or even groups of researchers. As internet access became widespread, hobbyists interested in astronomy were recruited to participate in the analysis of those images, a technique that is commonly known as crowdsourcing. This approach is particularly useful for tasks that are easy for humans but difficult for computers, such as identifying meaningful patterns in images.

One example of crowdsourcing is the Zooniverse (Moore, 2011) project of the Citizen Science Alliance. The Zooniverse asks the public to play games with real data using a Web interface. Members of the general public can hunt for supernovae, identify solar flares, or transcribe weather observations recorded aboard ships during World War I. In fact, crowdsourcing does not require input from benevolent individuals. The reCAPTCHA initiative (von Ahn, 2008) adapted a strategy for distinguishing human users from spam bots on websites into a tool for correcting errors in digitized books. Such public engagement provides independent verification of individual data points and error correction. Because of their popularity, social networks and

online games (e.g., Farmville) may provide a conceptual framework for solving scientific problems.

IMPLEMENTATION OF GENOMIC EVALUATIONS

Genomic Evaluation

Historically, theoretical improvements in animal breeding made possible the inversion of the numerator relationship matrix (Henderson, 1975), a problem that commonly exceeded the computing resources available at the time. This allowed the use of mixed model theory for the computation of estimated breeding values and estimation of variance components (Henderson, 1984), although the size of the systems of equations to be solved quickly became impractical for large-scale genetic evaluations. So-called ‘iteration on-data’ methods, which do not require storing mixed model equations in computer memory, were then developed to solve that problem. A series of algorithms were proposed, including Gauss-Seidel or successive over-relaxation (Schaeffer and Kennedy, 1986), second order Jacobi (Misztal and Gianola, 1987), and Jacobi conjugate gradient (Berger et al., 1989) approaches. The current method of choice for solving mixed model equations with iteration-on-data is the preconditioned conjugate gradient (**PCG**), which was introduced in animal breeding by Berger et al. (1989) and implemented in several software packages (e.g., Strandén and Lidauer, 1999; Tsuruta et al., 2001).

The growing volume of data utilized in genomic selection programs poses computational problems that are shared by functional genetics analyses. Genotypes for each individual require a large amount of storage, increasing as SNP chips increase in density, and will continue to grow as individual sequences become routinely affordable. Full sequence data are of great potential value because the accuracy of genomic evaluations will increase once causal mutations can be tracked, rather than markers in LD with those mutations. There also is growing interest in interactions among genes and regulatory networks that control gene expression (e.g., Fortes et al., 2011), which will benefit from higher-density marker panels. In order to realize that potential, however, the annotation of the bovine genome must continue to improve (Reese et al., 2010).

Efficient methods to compute genomic predictions were described by VanRaden (2008), and current genomic evaluation systems involve several steps (VanRaden et al., 2009; Hayes et al., 2009; Harris and Johnson, 2010). State-of-the-art systems for dairy and beef cattle evaluations were discussed by Wiggans et al.

(2011) and Garrick (2011), respectively. A single-step approach that is conceptually simpler in many respects has been proposed by Misztal et al. (2009). While genomic evaluation systems have been rapidly adopted, there are still a number of technical problems that need to be solved, such as correct adjustment for genomic pre-selection of dairy sires (Patry and Ducrocq, 2011; Vitezica et al., 2011).

In the single-step approach for the incorporation of genomic information in genetic evaluations, the inverse of the numerator relationship matrix (A) based only on pedigree information is replaced by the inverse of a combined relationship matrix (H) based on pedigree and genomic information (Legarra et al., 2009; Aguilar et al., 2010). This resulted in a genetic evaluation with EBV predicted for genotyped and non-genotyped animals using all available information (Aguilar et al., 2010; Christensen and Lund, 2010).

Implementation of the single-step method requires additional matrices, namely the inverse of the genomic relationship matrix and the inverse of the numerator relationship matrix for the subset of genotyped animals. Efficient methods for creation and inversion of such matrices were recently developed (Aguilar et al., 2011a). For example, using a naïve algorithm for creating the genomic relationship matrix for 14,000 genotyped animals required almost one day, but using custom libraries and parallel computing, the time required was reduced to 15 min.

Large-scale genetic evaluations with the single-step approach were run for either single trait (Aguilar et al., 2010) or multiple-trait analyses (Aguilar et al., 2011b; Tsuruta et al., 2011). Computing time for genetic evaluations using genomic information took twice as long as evaluations without genomic data (Tsuruta et al., 2011), but could be reduced by 4 to 11% compared to regular genetic evaluations if custom libraries are used (I. Aguilar, I. Misztal, A. Legarra, and S. Tsuruta, unpublished data).

Big Data in a Commercial Setting

There are several challenges related to Big Data in a commercial setting, such as the Edinburgh Genetic Evaluation Services (**EGENES**; Midlothian, UK), which provides genetic evaluations for 1.85 million dairy cattle in the U.K., as well as web tools for breeding and farm management. In addition to traditional phenotypic and pedigree data, the EGENES database includes 11,500 50K genotypes, 600 800K genotypes, and will

include full-sequence data on 10 bulls. Imputation of genotypes on cows will further increase the size of the genomic dataset. The most notable challenges related to the implementation of genomic evaluations are storage costs and computing time. Storage is of great importance because data audits are necessary and processed data have to be retained, which requires not only disk capacity but a backup facility as well. Estimated breeding values have historically been computed and distributed according to fixed schedules, but there is growing demand for genomic estimated breeding values to be calculated on demand.

While data files used as input to individual runs of a genetic evaluation system may easily be managed, that is not necessarily the case for the intermediate files generated by the evaluation system. However, those data also need to be retained so that audit trails can be constructed if there are questions about results. Files from multiple runs also are needed so that new results may be compared to older results to detect unexpected changes in evaluations. As storage requirements grow, backup times, restore times, media costs, and off-site storage fees increase.

Computing requirements are increasing as customers demand more rapid service, such as more frequent conventional evaluations and on-demand calculation of genomic EBV from genotypes. In addition, marker effects estimates should be frequently updated to reflect new data. Such services require new software, faster hardware, high-capacity, and high-reliability network connections, and they typically involve increased costs for customer support. Consumers of the data often view them as a commodity rather than as premium items, and it can be difficult to recover the cost of development and implementation through increased fees.

SUMMARY AND CONCLUSION

Datasets are getting larger but we have many tools for working with them. To ensure that this remains the case, prototype software must be designed to consider scalability at the outset, which is often not done. Some computational resources, such as memory, disk space, and processing cycles, are relatively inexpensive. Programmer time is much more valuable, and speed should be measured as person-hours to solution rather than simply as data processing time. Programmer training programs need to look back 15 or 20 yr and rediscover strategies that focus on finesse rather than raw computational speed. Good code is good code irrespective of

computing power, and the last decade has seen the growth of profligate programming. Students often have never dealt with size constraints, which is important when working with large data. Software engineering has evolved into a mature discipline, and we need to re-learn and apply good developmental practices that consider scalability at the outset. Animal breeders should seek out more formal training in programming, rather than depending primarily on self-learning.

As a community, consideration should be paid to the idea of best practices. Public and private organizations both can benefit from sharing experiences and discussing ideas for solving difficult problems. Reference datasets against which software may be tested could be developed and shared freely, an approach that has proven to be very useful in the U.S. dairy industry, and does not require the disclosure of any intellectual property or other sensitive information. This also may lead to the development of common formats for data exchange, such as those used by the participants in the International Bull Evaluation Service's (<http://www.interbull.org/>) genetic evaluation programs. Ideally, the community should work towards more open sharing of materials, such as a requirement that genotypes generated by publicly-funded research be deposited in a repository that is freely accessible, a model successfully used by the U.S. National Institutes of Health (NIH, 2007). The need for such sharing may decrease somewhat when it becomes cheaper to resequence than to backup or transmit genotypic data. For old animals, there are limited quantities of DNA available, and it is unreasonable to demand that every scientific community expend often-scarce resources to recreate data that already exist.

History indicates that we never have enough information; the more data we have, the more data we want. The ultimate value in Big Data lies in their ability to help answer questions, which may range from routine calculation of genomic estimated breeding values to predictions of regulatory networks, and there are always new questions. Relationships among traits are of increasing interest as we try to understand the biology that underlies complex phenotypes in livestock species, which will require ever-increasing computational resources. Customers demand rapid turnaround, often in real time, and smart engineering will be required to deliver such services cost-effectively. Ultimately, we hope that more really will be better.

LITERATURE CITED

- ACM. 2011. CERN experiments generating one petabyte of data every second. Accessed Aug. 10, 2011.
<http://cacm.acm.org/news/110048-cern-experiments-generating-one-petabyte-of-data-every-second/fulltext>.
- Aggarwal, C. C., and P. S. Yu. 1998. Mining large datasets for association rules. *Data Engin. Bull.* 21:23-31.
- Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010a. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93:743-752.
- Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011a. Efficient computations of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet.* 128:422-428.
- Aguilar, I., I. Misztal, S. Tsuruta, G. R. Wiggans, and T. J. Lawlor. 2011b. Multiple trait genomic evaluation of conception rate in Holsteins. *J. Dairy Sci.* 94:2621-2624.
- Berger, P., G. Luecke, and J. Hoekstra. 1989. Iterative algorithms for solving mixed model equations. *J. Dairy Sci.* 72:514-522.
- Berry, D. P., F. Buckley, P. Dillo, R. D. Evans, M. Rath, and R. F. Veerkamp. 2003. Genetic relationships among body condition score, body weight, milk yield, and fertility in dairy cows. *J. Dairy Sci.* 86:2193-2204.
- Banos, G., and M. P. Coffey. 2010. Genetic association between body energy measured throughout lactation and fertility in dairy cattle. *Animal* 4:189-199.
- Calus, M. P. L., J. J. Windig, and R. F. Veerkamp. 2005. Associations among descriptors of herd management and phenotypic and genetic levels of health and fertility. *J. Dairy Sci.* 88:2178-2189.
- Caroli, A. M., S. Chessa, and G. J. Erhardt. 2009. Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. *J. Dairy Sci.* 92:5335-5352.
- Christensen, O., and M. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42:2 (doi:10.1186/1297-9686-42-2).
- Cole, J. B., D. J. Null, and L. R. Bacheller. 2008. A data exchange format and national database for producer-

- recorded health event data from on-farm management software. *J. Dairy Sci.* 91(E-Suppl. 1):2-3 (Abstr.).
- Cole, J. B., and P. M. VanRaden. 2010. Visualization of results from genomic evaluations. *J. Dairy Sci.* 93:2727-2740.
- Cole, J. B. and P. M. VanRaden. 2011. Use of haplotypes to estimate Mendelian sampling effects and selection limits. *J. Anim. Breed. Genet.* 128:446-455.
- Cook, R. D., and S. Weisberg. 1999. *Applied Regression Including Computing and Graphics.* John Wiley & Sons, New York, NY.
- Dechow, C. D., H. D. Norman, N. R. Zwald, C. M. Cowan, and O. M. Meland. 2008. Relationship between individual herd-heritability estimates and sire misidentification rate. *J. Dairy Sci.* 91:1640-1647.
- Edwards, J. L., and P. R. Tozer. 2004. Using activity and milk yields as predictors of fresh cow disorders. *J. Dairy Sci.* 87:524-531.
- Egger-Danner, C., K. Stock, J. Cole, A. Bradley, J. Pryce, N. Gengler, L. Andrews, and E. Strandberg. 2011. Registration of health traits - strategies of phenotyping, aspects of data quality and possible benefits. 37th ICAR Session, Bourg-en-Bresse, France.
- Everitt, B. S., S. Landau, and M. Leese. 2001. *Cluster Analysis*, 4th ed. Arnold Publishers, London, UK.
- Fortes, M. R. S., A. Reverter, S. H. Nagaraj, Y. Zhang, N. N. Jonsson, W. Barris, S. Lehnert, G. B. Boe-Hansen, and R. J. Hawken. 2011. A single nucleotide polymorphism-derived regulatory gene network underlying puberty in 2 tropical breeds of beef cattle. *J. Animal Sci.* 89:1669-1683.
- Garrick, D. J. 2011. The nature, scope and impact of genomic prediction in beef cattle in the United States. *Genet. Sel. Evol.* 43:17. (doi:10.1186/1297-9686-43-17).
- Hand, D. J. 1997. *Construction and Assessment of Classification Rules.* John Wiley & Sons, New York, NY.
- Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J. Dairy Sci.* 93:1243-1252.
- Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92:433-443.
- Henderson, C. R. 1975. Rapid method for computing the inverse of a relationship matrix. *J. Dairy Sci.* 58:1727-

- Henderson, C. R. 1984. Application of Linear Models in Animal Breeding. University of Guelph, Ontario.
- Herd, R. M., and P. F. Arthur. 2009. Physiological basis for residual feed intake. *J. Anim. Sci.* 87:E64-E71.
- IDC. 2011. The 2011 Digital Universe Study: Extracting Value from Chaos. Accessed Aug. 3, 2011.
<http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm>.
- Interbull. 2011. Interbull Routine Genetic Evaluation for Female Fertility Traits. Accessed July 28, 2011.
http://www-interbull.slu.se/Female_fert/framesida-fert.htm.
- Janesick, J. R. 2001. Scientific Charge-Coupled Devices. SPIE Press, Bellingham, WA.
- Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92:4656-4663.
- Lin, J., and C. Dyer. 2010. Data-Intensive Text Processing with MapReduce (Synthesis Lectures on Human Language Technologies). Morgan & Claypool, San Rafael, CA.
- Løvendahl, P., and M. G. G. Chagunda. 2010. On the use of physical activity monitoring for estrus detection in dairy cows. *J. Dairy Sci.* 93:249-259.
- Maltecca, C., K. A. Gray, K. A. Weigel, J. P. Cassady, and M. Ashwell. 2011. A genome-wide association study of direct gestation length in US Holstein and Italian Brown populations. *Animal Genet.* 42:585-591.
- McParland, S, Banos, G, Wall, E, Coffey, MP, Soyeurt, H, Veerkamp, RF, Berry, DP, 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. *Journal of Dairy Science.* 94: 3651-3661
- Misztal, I. and D. Gianola. 1987. Indirect solution of mixed model equations. *J. Dairy Sci.* 70:716-723.
- Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.* 92:4648-4655.
- Misztal, I., I. Aguilar, S. Tsuruta, J. P. Sanchez, and B. Zumbach. 2010. Studies on heat stress in dairy cattle and pigs. Commun. No. 625 in Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany.
- Moore, J., P. L. Gay, K. Hogan, C. Lintott, C. Impey, and C. Watson. 2011. Facebooking citizen science with the Zooniverse. *Bull. Am. Astronom. Soc.* 43:158.13 (Abstr.).

- NIH. 2007. Policy for Sharing of Data Obtained in NIH-Supported or Conducted Genome-Wide Association Studies (GWAS). Accessed Aug. 3, 2011. <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html>.
- Norman, H. D., L. G. Waite, G. R. Wiggans, and L. M. Walton. 1994. Improving accuracy of the United States genetics database with a new editing system for dairy records. *J. Dairy Sci.* 77:3198-3208.
- NVIDIA Corporation. 2011. CUDA: Parallel Programming Made Easy. Accessed Aug. 9, 2011. http://www.nvidia.com/object/cuda_home_new.html.
- Patry, C., and V. Ducrocq. 2011. Accounting for genomic pre-selection in national BLUP evaluations in dairy cattle. *Genet. Sel. Evol.* doi:10.1186/1297-9686-43-30.
- Reese, J. T., C. P. Childers, J. P. Sundaram, C. M. Dickens, K. L. Childs, D. C. Vile, and C. G. Elsik. 2010. Bovine Genome Database: supporting community annotation and analysis of the *Bos taurus* genome. *BMC Genomics.* 11:645 (doi:10.1186/1471-2164-11-645).
- Schadt, E. E., M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan. 2010. Computational solutions to large-scale data management and analysis. *Nature Rev. Genet.* 11:647-657.
- Schaeffer, L. R., and B. W. Kennedy. 1986. Computing strategies for solving mixed model equations. *J. Dairy Sci.* 69:575-579.
- Shaffer, J. P. 1995. Multiple hypothesis testing. *Ann. Rev. Psych.* 46:561-584.
- Soyeurt H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-Infrared prediction of bovine milk fatty acids across multiple breeds, production systems and countries. *J. Dairy Sci.* 93:1657-1667.
- Strandén, I. and M. Lidauer. 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. *J. Dairy Sci.* 82:2779-2787.
- Tan, P.-N., M. Steinbach, and V. Kumar. 2005. *Introduction to Data Mining*. Addison-Wesley, New York, NY.
- Tsuruta, S., I. Misztal, I. Aguilar, and T. J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. *J. Dairy Sci.* 94:4198-4204.
- Tsuruta, S., I. Misztal, and I. Strandén. 2001. Use of the preconditioned conjugate gradient algorithm as a

- generic solver for mixed-model equations in animal breeding applications. *J. Anim. Sci.* 79:1166-1172.
- Tufte, E. R. 2006. *Beautiful Evidence*. Graphics Press, Cheshire, CT.
- Tufte, E. R. 1983. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414-4423.
- VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92:16-24.
- VanRaden, P.M., O'Connell, J.R., Wiggans, G.R., and Weigel, K.A. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* 43:10.
- VanRaden, P.M., M.E. Tooker, and J. B. Cole. 2009. Can you believe those genomic evaluations for young bulls? *J. Dairy Sci.* 92(E-Suppl. 1):314. (Abstr.)
- Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. *Genet Res.* 93:357-366.
- von Ahn, L., B. Maurer, C. McMillen, D. Abraham, and M. Blum. 2008. reCAPTCHA: Human-based character recognition via Web security measures. *Science* 321:1465-1468.
- Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, NY.
- Wiggans, G. R., T. A. Cooper, P. M. VanRaden, K. M. Olson, and M. E. Tooker. 2011. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. *J. Dairy Sci.* (Accepted.)
- Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2011. The genomic evaluation system in the United States: Past, present, future. *J. Dairy Sci.* 94:3202-3211.
- Wiggans, G. R., P. M. VanRaden, J. C. and Philpot. 2003. Technical Note: Detection and adjustment of abnormal test-day yields. *J. Dairy Sci.* 86:2721-2724.

Table 1. The number of records stored in the U.S. national dairy database¹.

Type of Record	Number of Records ¹
Cow with lactation data	28,394,976
Lactations	68,373,863
Individual test days	508,574,732
Dystocia records	20,770,758
Animals in pedigree file	58,893,009
Bull genotypes	50,393
Cow genotypes	70,687

¹Totals include animals of all breeds.

FIGURE CAPTIONS

Figure 1. This Manhattan plot showing the distribution of marker effects for lifetime net merit in U.S. Holstein (HO) cattle uses color to differentiate among markers on different chromosomes and marker size to emphasize the magnitude of marker effects. Points indicating the magnitude of marker effects are proportional to the absolute value of the effect size.

Figure 2. This three-generation pedigree for the Holstein bull Co-Op O-Style Oman Just-ET (001HO09167) traces the inheritance of individual haplotypes for chromosome 15 from the great-grandparents. Segments of only 7 of 16 great-grandparental chromosomes are present in O-Style's genome.

Figure 3. This line graph shows changes in average inbreeding (%) between 1990 and 2010 for U.S. Brown Swiss (solid, red line), Holstein (short-dashed, green line) and Jersey (long-dashed, blue line) cattle. Lines are differentiated from one another by two separate factors, pattern and color, minimizing the risk that readers will confuse one series with another. Adapted from Cole and VanRaden (2011).

Figure 4. Sparklines are small graphics that may be used in text in a wordlike manner. This example uses sparklines to predict the outcome of a hypothetical mating of the Holstein bull Co-Op O-Style Oman Just-ET (001HO09167) and genotyped animal "Cow C". Used with permission from Cole and VanRaden (2010).

Figure 1

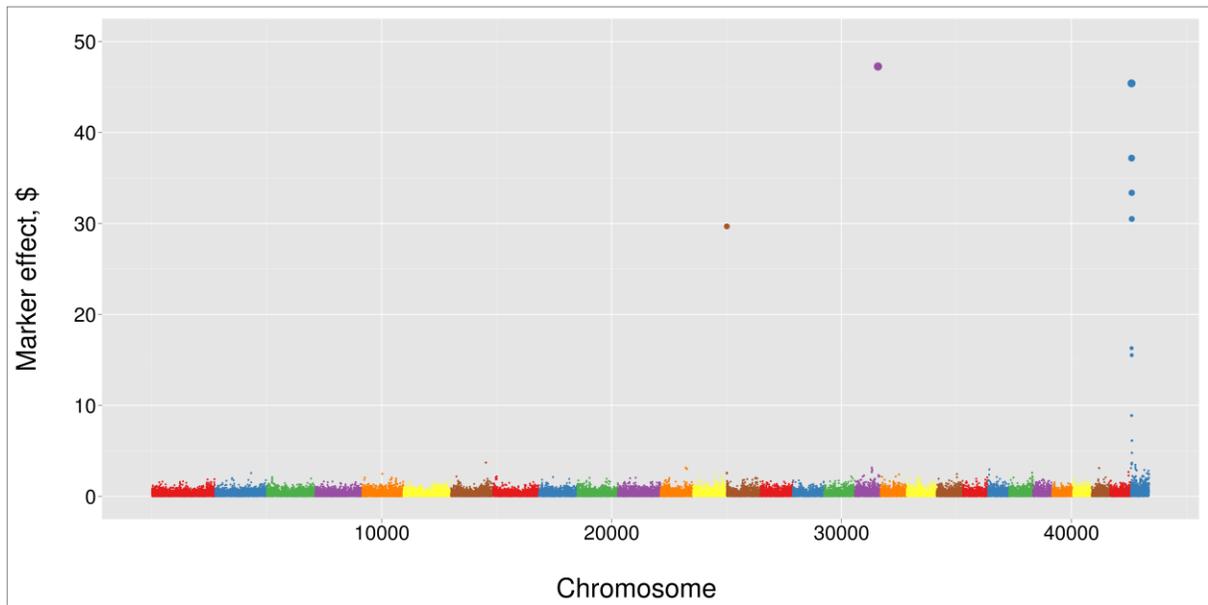


Figure 2

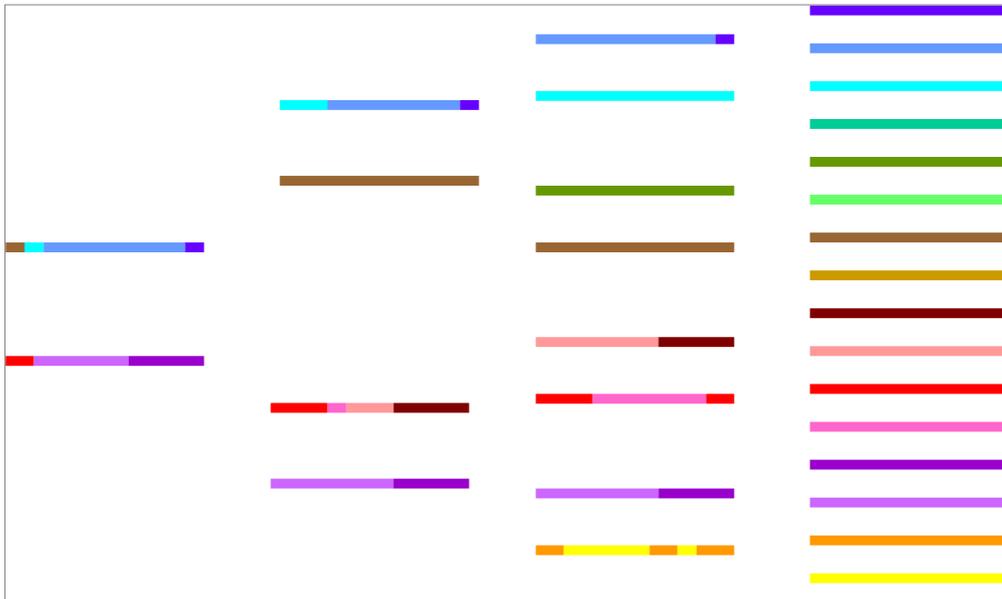


Figure 3

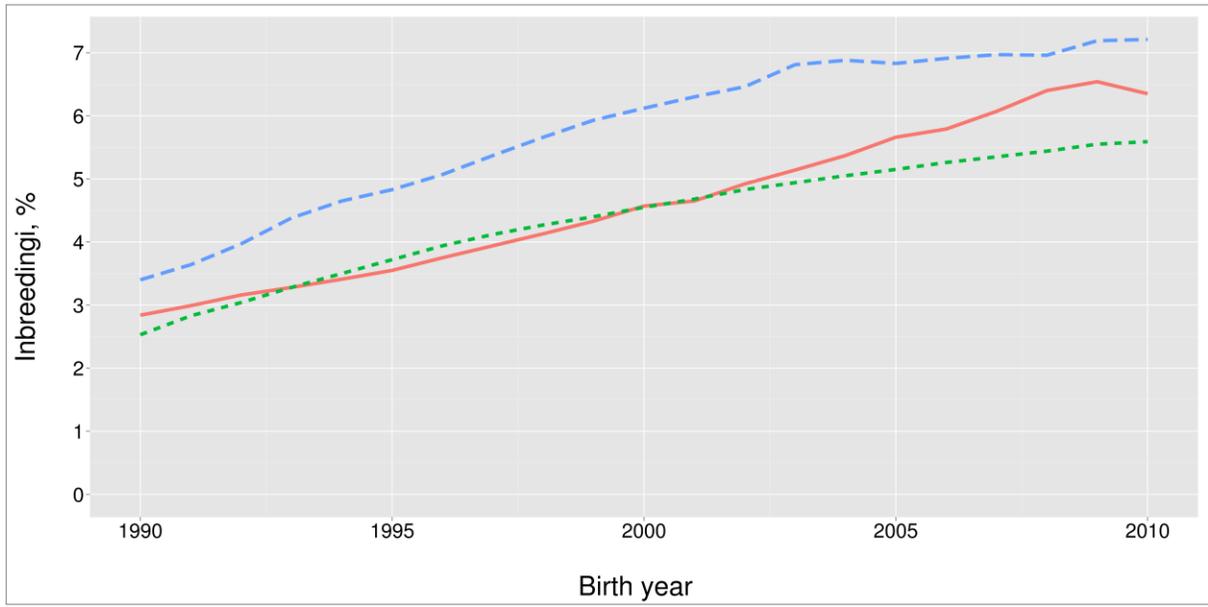


Figure 4

Finally, calf C has a PA NM\$ of +803, placing it in the 99th percentile for all Holsteins. Cow C has very desirable PTA for most chromosomes and also ranks in the 99th percentile in the Holstein breed. O-Style appears to be an excellent mate for cow C, but there may be bulls that better complement her weaknesses, particularly BTA 24 through 28:

