Title:

Effectiveness of a genome-wide association study using DNA pooling to make management decisions in feedlot cattle

Author:

Haleigh Prosser

Biographical Sketch:

Growing up in southeast Colorado with a family-owned beef growlot literally in my backyard, I knew from an early age that I would pursue an education and career centered around beef cattle. As a high-risk operation, the family lot has given me extensive experience with cattle that differ from a typical, low-risk herd. My desire to improve the genetics of the beef industry stems from my history, experience, and knowledge of high-risk beef cattle in a feedlot setting. In 2021, I earned my bachelor's degree in Animal Science with a Pre-Veterinary specialization from West Texas A&M University. After graduation, I began pursuing my master's degree, focusing my research efforts on the genetics of beef cattle. My research, evaluating the effectiveness of genetic testing using pooled DNA samples, allows commercial beef producers to obtain the benefits of genetic testing at a feasible cost. This cost-efficient genetic testing method, as well as further studies evaluating the associations between genotypes and economically important phenotypes, will allow me to achieve my lifelong goal improving the genetics in the beef cattle industry.

Advisor Names:

Dr. Thomas Perkins, Ph.D.

Dr. Matthew Scott, DVM, Ph.D.

Advisor Approval:

I certify that I have read and had sufficient time to provide feedback on the attached literature review.

Tommy Perkins — Advisor Name

4/12/2022 — Date

Advisor Name

4/12/2022 — Date

Effectiveness of a genome-wide association study using DNA pooling to make management decisions in feedlot cattle

Haleigh Prosser, West Texas A&M University, Canyon, TX 79016

INTRODUCTION

Large-scale genetic testing of beef cattle to predict performance in a feedlot setting and carcass value at harvest would revolutionize cattle production. Advancements in genomic marker research for enhancing feeding, sorting, and buying strategies would improve management decisions for beef cattle producers, meat processors, and consumers alike. In short, DNA testing and the assembly of the bovine whole-genome sequence has permitted this area of research to progress, capable of allowing the allocation of DNA components, such as specified loci or polymorphisms, to production traits. By applying phenotype associations with genomic loci, patterns, and gene-level markers, insights into the biological mechanics and specific genetic components of a desirable or undesirable phenotype can be ascertained. These advancements may further breeding decisions to propagate or eliminate the presentation of production-critical alleles. However, DNA testing on an individual basis, while ideal, cannot be rationalized in a traditional feedlot setting due to cost and logistical setbacks within large-scale beef production systems. Therefore, we propose that combining individual DNA samples into pools based upon predetermined criteria, such as animal source, arrival date, and arrival weight, producers and feedlot managers can perform genetic testing for 10-30 animals at the cost of an individual test. If these pooled DNA tests can accurately predict feedlot performance, the economic advantage achieved by the genetic-based decisions could outweigh the cost of testing.

REVIEW OF LITERATURE

Introduction to DNA Testing and Applications

The development and promotion of deoxyribonucleic acid (DNA) testing continues to revolutionize the ability to predict animal performance, such as weight gain and feed efficiency. Consequently, new technology tends to drive research in applied genetics and DNA evaluation, and the needs exposed by past and current technologies influence the development of further technologies. DNA molecules, composed of a double helix of polynucleotide chains, contain a series of nucleotide bases that correspond to amino acids produced by an organism. This

discovery indicated that the sequence of amino acids makes up an organism's genetic code (Alberts et al., 2002). However, prior to the sequencing of DNA, this genetic code remained a mystery to those attempting to pinpoint the markers and material passed from parents to offspring. Two processes were subsequently developed around 1976, the first being a chain terminator procedure introduced by Sanger and Coulson (1977) and the second a chemical cleavage procedure introduced by Maxam and Gilbert (1977), reduced the common one base per month decoding time to hundreds of bases in an afternoon (Shendure et al., 2017). Fluorescence *in situ* hybridization (FISH), the first technique to bridge cytogenetics and molecular genetics, utilized a fluorescent probe to identify these amino acid sequences to visualize chromosomes and further improved the base generation speed (Durmaz et al., 2015). By 1982, data repositories contained over half a million bases and in just four years later contained nearly 10 million. In 1987, Applied Biosystems, Smith, and Hood marketed a Sanger sequencing machine, which could generate 1,000 bases per day utilizing fluorescence-based technologies (Shendure et al., 2017). Since completion of the Human Genome Project, in which researchers mapped and sequenced the entire human genome, technologies derived during the project allow even small labs to offer some form of genetic testing. This laboratory feasibility, in combination with the successful applications of genetic information to human science and medicine, sparked interest in the genomic information of livestock. Similar to human biomedical scientists, animal scientists insisted genetic analyses, completed by ever-changing technological breakthroughs, could provide insight into potential performance and health capabilities of animals; thus, the desire to map and sequence the genome of agriculturally important livestock species stimulated extensive animal genome research.

Bovine DNA Testing

Through the 1970s and early 1980s, the laboratory mouse was the only mammal with linkage maps and well-defined genetic markers. Morris Soller, in the late 1970s, began to champion an effort to map the genes responsible for traits in livestock, specifically traits with economic importance. Using these genetic maps, Soller suggested that marker-assisted selection could inform breeding decisions and enhance positive alleles in the population (Womack, 2012). At this time, however, the technologies to compile a whole-genome map of a livestock species did not exist. The discovery and application of short sequences of repeated DNA motifs known as short tandem repeats (STRs), at the time termed sequence tagged microsatellite sites (STMS),

led Beckmann and Soller (1990) to propose genetic mapping of livestock; soon after the Beckmann and Soller proposal of STR application began a global search for quantitative trait loci (QTL) in livestock animals (Womack, 2012). The quest for QTL became more straightforward as single nucleotide polymorphisms (SNPs), variations at a single nucleotide position, replaced STRs as more efficient and accurate markers for QTL mapping.

Advancing computational technology attributable to the Human Genome Project, in combination with the introduction of SNPs and a scientific desire to genetically map agriculturally important livestock, inspired an abundance of bovine genome work. By 2012, more than 4,600 mapped cattle QTL represented more than 375 different traits in the Cattle QTL Database (Womack, 2012). The repeated use of FISH technology on the same chromosomes in early experiments ultimately resulted in the sequence of DNA on bovine chromosomes (Womack, 2012), and with the identification of marker genes taken from Fries et al. (1993) and the Texas Standard, whole-genome sequencing of bovines began to occur. A variety of bovine maps exist in the present, each possessing different types of markers of differing resolutions (Womack, 2012). The knowledge and technologies that have been developed create the opportunities to use marker and linkage maps to pinpoint QTL information for economically important traits in livestock, just as Soller suggested over half a century ago. Application of QTL to traits continues to occur and advance as new maps, increased marker density, and further technology become more accessible and available.

Genome Wide Association Studies (GWAS)

Genome Wide Association Studies (GWAS) test many genetic variants to identify phenotype-genotype associations. GWAS can be used to associate genes with phenotypically-observed traits, diseases, and performance. GWAS applies statistical analysis to genotypic and phenotypic data to pinpoint associations and locate common SNPs between organisms sharing a specific phenotype. In animal research, genotyping takes place on DNA extracted from blood, tissue, or hair samples, most often. SNP data must then undergo a quality control step to eliminate SNPs that will not be utilized in the further analysis; using PLINK software, SNPs with high miss rates, low allele frequencies, and low Hardy-Weinberg equilibrium p-values are removed (Purcell et al., 2007). Many disease-based GWAS studies use PERMORY software (Pahl and Schäfer, 2010) to run random permutations and reveal significantly associated SNPs with a particular trait, eliminating potential false positives with a multiple test correction (Lee et

al., 2015). Other GWAS studies, often those with multiple traits or more difficult phenotypes to define, utilize the R package rMVP (Yin et al., 2021) and the fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al., 2016). The FarmCPU software controls for false positives and false negatives using the fixed and random effects in the model. These software programs return results displaying the SNPs that possess significant associations with the phenotypic trait or traits in questions. From those significant SNPs, further analyses to determine the genes closest to these SNPs must occur to select candidate genes. With knowledge of these candidate genes, one can continue to investigate their presence or absence in populations, comparing that information with the presence or absence of a specific phenotype.

Agricultural geneticists in the animal and plant science sectors notice the practicality of GWAS at an increasing rate; GWAS can associate economically essential performance traits of animals and crops, like meat quality in fed hogs (Gao et al., 2021), milk production in dairy cows (Jiang et al., 2019), and frost tolerance in wheat (Soleimani et al., 2022), for example, with significant genetic markers of interest, as well as identify significant risk markers for animal and crop disease. Understanding the genetic controls of these traits allows for more targeted animal selection, more efficient breeding practices, and faster genetic progress. Lee et al. (2015) used a high-density bovine SNP chip to identify Foot-and-Mouth Disease resistant loci in Holstein cattle; this GWAS research located 3 significant SNPs on a single chromosome. Similar to locating disease-resistant loci in dairy cattle, Xue et al. (2020) used GWAS methodologies to identify 14 significant backfat thickness-associated SNPs and 9 significant loin muscle depth-associated SNPs in a crossbred pig population; utilizing these SNPs to select candidate genes to continue research, genetic selection for those two economically important growth traits can occur. In addition to disease and growth traits, genetic markers of meat quality have also been evaluated using GWAS. Gao et al. (2021) completed GWAS analysis to locate 32 SNPs associated with conductivity, intramuscular fat, marbling score, meat color, moisture, and pH of meat harvested from a crossbred commercial pig population. Using GWAS in agricultural research allows for the application of significant genetic occurrences to commonly observed and economically important phenotypes.

Pooled-Sample DNA Tests

Individual genotyping and individual GWAS studies each possess value in pinpointing the genetic association of a phenotype on an individual level. The practicality, specificity, and

accuracy of individual genotype studies have value in many situations within the beef production industry. For example, detecting disease, predicting traits, and making decisions at an individual level can be economically justified in purebred, breed-association settings where DNA samples and genotyping must be completed for registration, breeding, and sale purposes. In crossbred and commercial populations, however, the cost of individually genotyping animals prohibits the process from occurring. Pooling these animals in the commercial sector (commercial ranches, feedlots, stocker operations, or processing plants) allows a relatively large amount of data analysis for the price of a single genomic test. Especially in groups with extreme phenotypes, like the unrelated animals in a commercial setting, pooling can actually provide a more accurate genetic evaluation (Keele et al., 2021). Additionally, allowing commercial operations to perform genetic evaluations using pooling results in the addition of commercial phenotypes to genetic evaluation (Abrams et al., 2021) and further advances the data availability of bovine genomics.

Much of the hesitation to perform genomic evaluations in the commercial sector occurs because of the sheer mass of data necessary to draw conclusions. Subsequently, commercial operations cannot rationalize the price of mass-sequencing hundreds or thousands of individuals. However, pooling DNA reduces this large number of samples to just a few pools of samples (Huang et al., 2010). Pool designs vary by experimental method and objective, but varying pooling techniques can take account of stratification, inter-loci interactions, and allow further haplotype analysis (Sham et al., 2002). Along with GWAS analysis, pool designs can cost- and time-efficiently associate genetic information possessed in a group with the general phenotype of the group. While some studies incorporate a two-pool design, especially those evaluating a basic allelic association, Sham et al. (2002) propose that large-scale studies should utilize a more-complex design with multiple pools. By reducing the number of individuals in each pool to create further pools and subsets or replicate pools, researchers can reduce the chance of error and provide additional opportunity to find marker-marker associations (Sham et al., 2002).

In theory, research using pools of DNA evaluates the genomic makeup of a set of individuals in the cost and time of an individual genomic evaluation. Experiments using pooled DNA can reduce costs by 90% compared to traditional individual genotyping (Keele et al., 2021). However, this break in cost must produce accurate results. In a DNA pooling project evaluating fertility in Holstein cattle, Huang et al. (2010) concluded the significant SNPs in the pooled DNA data also showed significance in individual genotypes, demonstrating the validity

of selective DNA pooling. Similarly, Macgregor et al. (2008) found that DNA pooling, while cost effective, can also capture greater than 80% of the power of individual genotyping in GWAS. If DNA pooling proves significantly accurate as compared to individual genotyping, pools can replace individuals in many studies that do not require individual genotypes.

While a seemingly optimal replacement to individual genome assessment, analyses utilizing pooled DNA must be evaluated and closely observed to assure errors involved in the pooling process do not obscure results. Pool construction error, one of the most common errors in pooling research, includes errors in DNA concentration, sample mixing, pipetting, extraction efficiency, and other often laboratory-based errors (Keele et al., 2021). For this reason, laboratory methodological procedures must be reliable, optimized, and quantitative to prevent technical biases (Sham et al., 2002). Unsurprisingly, reducing experimental error, most commonly in allele frequency estimation, increases the power in selective DNA pooling (Huang et al., 2010). Technical errors, like variation between allele contribution, occur most often because of these pool construction errors; some of this error can be eliminated with use of technical replicates, but replications reduce the cost efficiency of the pooling method (Huang et al., 2010; Keele et al., 2021). Hernandez-Rodriguez et al. (2017) suggest that creating these replicates forms equilibrated pools as well as increases data yield, possibly offsetting the additional cost. Additionally, linear regressions and analysis of variance can allow researchers to evaluate different variables as sources of technical variation (Hernandez-Rodriguez et al., 2017). Sampling error during pool construction arguably creates some of the most difficult decisions for researchers. Estimates must be made to compromise cost and accuracy, to ensure the most accurate data, and to see significant and accurate genetic observations. In theory, and as proved by Keele et al. (2021), based on the Dirichlet distribution, a larger planned animal contribution (thus, a smaller pool) results in larger pool construction error and more variation within pools. Huang et al. (2010), using similar rationality, recommend pooling as many individuals as possible, as allowed by the total mapping population size. While the optimal number for pool size depends largely on both total population and other experimental-specific factors, Barratt et al. (2002) suggest an optimal DNA pool consists of equal concentrations of DNA obtained from 50 individuals. The potential errors associated with pooled DNA samples and research create a need for error prevention and statistical interpretation, but the knowledge of errors common to pooled DNA projects allows researchers to prevent these errors from dismissing the accuracy of

the results.

Alternatively, rather than traditional collection of DNA and pooling methodology, Abrams et al. (2021) demonstrated a process to construct pools prior to DNA extraction. One of the most attractive features of this practice is the ability to further reduce the price of genetic testing by requiring only one DNA extraction per pool. Additionally, this process also ensures equal representation of individuals within pools, creating a more uniform sample and more accurate results. The project utilized white blood cell counts to construct pools of samples; then, DNA and genotyping was completed using the one sample. Because white blood cells contain equal concentrations of DNA (Abrams et al., 2021), adding individual samples to the pool by a certain count of white blood cells ensures an equal individual contribution of DNA. These equal concentrations eliminate the errors addressed above stemming from a variable individual contribution produced by a traditional fluorometric or photometric DNA quantification method. Abrams et al. (2021) found the use of white blood cell count to construct pools predicted the sample representation equally or more accurately than traditional pooling methods. The accuracy of this suggested pooling method only contributes to part of its appeal; the economic benefits of this process are outstanding. In a theoretical scenario containing 100 individuals, the individual genotyping would cost approximately $2,800 and the genotyping of pooled DNA would cost approximately $325 for the pool of 100 samples, but the genotyping of pooled DNA based on white blood cell counts would cost approximately $228 for the pool of 100 samples (Abrams et al., 2021). Utilizing this alternative form of pooled DNA studies, research involving bovine genetics can potentially be both more cost effective and cheaper.

CONCLUSION AND IMPLICATIONS TO GENETIC IMPROVEMENT OF BEEF CATTLE

The introduction of genetics into agricultural sciences, and specifically the bovine sector, have driven large-scale genetic testing research exponentially. Cattle are one of the most agriculturally important livestock species, and the bovine genome has received a great deal of attention after the mapping of other genomes, such as mice and humans. The advances in computing technologies that developed during the Human Genome Project, in combination with the outstanding desire to map the genome of livestock and connect genetic markers to economically important traits, placed bovine genome mapping at a high importance. Once the bovine genome was mapped, further technological advances led to the identification of genes and

their corresponding phenotypes; while many of these exist in databases, further research should be completed to continue to locate genes for some of the most economically important traits. By utilizing GWAS methodologies, researchers can detect statistically significant genotype-phenotype relationships by identifying significant SNPs, locating near-distance genes, and further evaluating the presence or absence of the gene of interest in animals showing a certain phenotype. The effectiveness of software like FarmCPU in identifying candidate genes even in complex trait situations, such as meat quality, results in a positive outlook for the identification of genes related to more complex traits; in the beef production industry, disease complexes such as Bovine Respiratory Disease and prediction of carcass merit could reveal a genetic connection through GWAS. Additionally, utilizing DNA pooling processes, research can significantly reduce costs while maintaining much of the statistical power. Research continuously displays the accuracy of pooled SNP associations when compared to individual genotyping, testifying to the effectiveness of pooling. In principle, a 200-fold increase in efficiency is well within reach when using DNA pooling; a situation with 200 cases and 200 controls can be evaluated using two pooled samples rather than 400 individual samples. Further, research suggests that utilizing larger pools would diminish technical error. Additional animals in a pool reduce experimental error, which in turn reduces experimental costs, GWAS with pooling can occur at an even greater level to identify genes associated with some of the most economically important traits. Utilizing these advances in techniques and technology, pooling research can become even more accurate, cheaper, and efficient. Understanding genes associated with production traits of interest can progress beef cattle production, including breeding, feeding, and harvesting programs, and accurate pooled DNA analyses will streamline the process to this genetic understanding.

LITERATURE CITED

Abrams, A.N., T.G. McDaneld, J.W. Keele, C.G. Chitko-McKown, L.A. Kuehn, and M.G. Gonda. 2021. Evaluating accuracy of DNA pool construction based on white blood cell counts. Front Genet. 12:635846. doi:10.3389/fgene.2021.635846

Alberts, B., J. Lewis, M. Raff, K. Roberts. A. Johnson, P. Walter, D. Bray, and J.D. Watson. 2002. Molecular biology of the cell. 4th ed. Garland Science, New York, NY

Barratt, B.J., F. Payne, H.E. Rance, S. Nutland, J.A. Todd, and D.G. Clayton. 2002. Identification of the sources of error in allele frequency estimations from pooled DNA

indicates optimal experimental design. Ann Hum Genet. 66(5-6):393-405. doi: 10.1046/j.1469-1809.2002.00125.x

Beckmann, J.S. and M. Soller. 1990. Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. Biotechnology (N Y). 8(10):930-932. doi:10.1038/nbt1090-930

Durmaz, A.A., E. Karaca, U. Demkow, G. Toruner, J. Schoumans, and O. Cogulu. 2015. Evolution of genetic techniques: Past, present, and beyond. Biomed Res Int. 2015:461524. doi:10.1155/2015/461524

Fries, R., A. Eggen, and J.E. Womack. 1993. The bovine genome map. Mamm Genome. 4:405-428. doi:10.1007/BF00296815

Gao, G., N. Gao, S. Li, W. Kuang, W. Jiang, W. Yu, J. Guo, Z. Li, C. Yang, and Y. Zhao. 2021. Genome-wide association study of meat quality traits in a three-way crossbred commercial pig population. Front Genet. 12:614087. doi:10.3389/fgene.2021.614087

Hernandez-Rodriguez, J., M. Arandjelovic, J. Lester, C. de Filippo, A. Weihmann, M. Meyer, S. Angedakin, F. Casals, A. Navarro, L. Vigilant, H.S. Kuhl, K. Langergraber, C. Boesch, D. Hughes, and T. Marques-Bonet. 2017. The impact of endogenous content, replicates and pooling on genome capture from faecal samples. Mol Ecol Resour. 2018(18):319-333. doi:10.1111/1755-0998.12728

Huang, W., B.W. Kirkpatrick, G.J.M. Rosa, and H. Khatib. 2010. A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. Anim Genet. 41:570-578. doi:10.1111/j.1365-2052.2010.02046.x

Jiang, J., L. Ma, D. Prakapenka, P.M. VanRaden, J.B. Cole, and Y. Da. 2019. A large-scale genome-wide association study in U.S. Holstein cattle. Front. Genet. 10:412. doi:10.3389/fgene.2019.00412

Keele, J., T. McDaneld, T. Lawrence, J. Jennings, and L. Kuehn. 2021. Estimation of pool construction and technical error. Agriculture. 11:1091. doi:10.3390/agriculture11111091

Lee, B.-Y., K.-N. Lee, T. Lee, J.-H. Park, S.-M. Kim, H.-S. Lee, D.-S. Chung, H.-S. Shim, H.-K. Lee, and H. Kim. 2015. Bovine genome-wide association study for genetic elements to resist the infection of foot-and-mouth disease in the field. Asian Australas. J. Anim. Sci. 28(2):166-170. doi:10.5713/ajas.14.0383

Liu, X., M. Huang, B. Fan, E.S. Buckler, Z. Zhang. 2016. Iterative usage of fixed and random

effect models for powerful and efficient genome-wide association studies. PLoS Genet. 12(2):e1005767. doi:10.1371/journal.pgen.1005767

Macgregor, S., Z.Z. Zhao, A. Henders, M.G. Nicholas, G.W. Montgomery, and P.M. Visscher. 2008. Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. Nucleic Acids Res. 36(6):e35. doi:10.1093/nar/gkm1060

Maxam, A. and W. Gilbert. 1977. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 74(2):560-564. doi:10.1073/pnas.74.2.560

Pahl, R. and H. Schäfer. 2010. PERMORY: an LD-exploiting permutation test algorithm for genome-wide association testing. Bioinformatics. 26(17):2093-2100. doi:10.1093/bioinformatics/btq399

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, P.C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559-575. doi:10.1086/519795

Sanger, F., S. Nicklen, A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 74(12):5463-5467. doi:10.1073/pnas.74.12.5463

Sham, P., J.S. Bader, I. Craig, M. O'Donovan, and M. Owen. 2002. DNA pooling: A tool for large-scale association studies. Nat Rev Genet. 3:862-871. doi:10.1038/nrg930

Shendure, J., S. Balasubramanian, G.M. Church, W. Gilbert, J. Rogers, and J.A. Schloss. 2017. DNA sequencing at 40: past, present and future. Nature. 550(7676):345-353. doi:10.1038/nature24286

Soleimani, B., H. Lehnert, S. Babben, J. Keilwagen, M. Koch, F.A. Arana-Ceballos, Y. Chesnokov, T. Pshenichnikova, J. Schondelmaier, F. Ordon, A. Börner, and D. Perovic. 2022. Genome wide association study of frost tolerance in wheat. Sci Rep. 12:5275. doi: 10.1038/s41598-022-08706-y

Womack, J.E. 2012. First steps: bovine genomics in historical perspective. Anim Genet. 43(1):2-8. doi:10.1111/j.1365-2052.2012.02382.x

Xue, Y., C. Li, D. Duan, M. Wang, X. Han, K. Wang, R. Qiao, X.-J. Li, and X.-L. Li. 2020. Genome-wide association studies for growth-related traits in a crossbreed pig population. Anim Genet. 52:217-222. doi:10.1111/age.13032

Yin, L., H. Zhang, Z. Tang, J. Xu, D. Yin, Z. Zhang, X. Yuan, M. Zhu, S. Zhao, X. Li, X. Liu.

2021. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genomics Proteomics Bioinformatics. 19(4):619-628. doi:10.1016/j.gpb.2020.10.007