Gene Set Enrichment Analysis Using Single Nucleotide Polymorphisms to Identify Genes Associated with Residual Feed Intake in Cattle

Lance Kidder
Spring, 2015

Kristen Johnson and Holly Neibergs, Animal Sciences
College of Agriculture, Human, and Natural Resource Sciences
As thesis advisors for Lance Kidder, we have read this document and find it satisfactory.

Kristen Johnson  Date  3/2/15

Holly Neibergs  Date  3/2/15
The National Program for Genetic Improvement of Feed Efficiency in Beef Cattle is supported by the Agriculture and Food Research Initiative Competitive Grant no. 2011-68004-30214 offered by the United States Department of Agriculture National Institute of Food and Agriculture.

Special thanks to:

Dr. Andrzej Wojtowicz
Dr. Jeremy F. Taylor
Dr. Christopher M. Seabury
Jennifer Michal
Jennifer Kiser
Précis:

Feed comprises 66% and 77% of the total cost of beef cattle calf and yearling finishing systems, respectively. Heritabilities for feed intake and feed efficiency (FE, estimated as residual feed intake, or RFI) have ranged from 0.08 to 0.46 in previous studies, highlighting the potential for genetic selection to bring about significant gains in feed efficiency and profitability within the beef industry. The objective of this study was to identify gene pathways significant for FE (as measured by RFI) through the use of Gene set enrichment analysis-single nucleotide polymorphism (GSEA-SNP) using single nucleotide polymorphisms (SNPs) as proxies for bovine genes. A population of 847 Hereford cattle (181 purebreds and 666 Hereford cross animals) consisting of 23 females and 824 males ranging in age from 210 to 496 d from a single ranch were evaluated for a period ranging from 70 to 140 days on feed (DOF). Only 31 animals were fed over 72 days. Average daily gain (ADG), dry matter intake (DMI), initial weight (IW), mid-test metabolic weight (MMWT), and DOF were recorded across the feeding period for each individual. Covariates for the genome-wide association study (GWAS) were age, % Hereford, and a series of 6 contemporary groups based on harvest date. GWAS was followed by GSEA-SNP of SNP data with Bos taurus gene sets sourced from GO, KEGG, Panther, Reactome, and Metacyc. A total of 20,692 bovine genes were mapped within gene sets, and proxy SNPs were mapped to genes located within 20 kilobase pairs. The null distribution of the GSEA-SNP test statistic was approximated using 10,000 permutations. A majority of genotypes were obtained from the Illumina bovine HD BeadChip, while the remainder were obtained using the Illumina bovine 50K BeadChip and imputed to 778,000 SNPs using Beagle. The Metacyc pathway ‘metabolism of proteins’ with 257 genes tended towards significance for RFI with a normalized enrichment score (NES) of 2.917. There were a total of 108 leading edge genes in the pathway. The top 10 of the 108 genes were: DNAJB11, RPL24, RPS3, PROS1, EIF4A2, ST6GAL1, RPS14, TBCE, EIF3I, and CCL2. Heritability for RFI was estimated to be 0.30. These results suggest that genetic selection for RFI has potential to improve the efficiency and, therefore, profitability of beef cattle production.
Table of Contents

I. Introduction and Literature Review - - - - - 06
II. Thesis Activity - - - - - - - - 31
III. Materials and Methods - - - - - - - - 32
IV. Results and Discussion - - - - - - - - 39
V. Conclusions - - - - - - - - - - 41
VI. References Cited - - - - - - - - - - 43

VII. Appendix - - - - - - - - - - 54
VIII. Figure 1: Gene Set Enrichment Analysis-SNP Flow Diagram - 54
IX. Table 1: Genome-wide Rapid Association using Mixed Model and Regression (GRAMMAR) mixed linear model GWAS Results - - - - 55
X. Figure 2: Identification of loci associated with feed efficiency - Manhattan Plot - - - - - - - - - - - 56
XI. Figure 3: Q-Q plot of RFI GWAS distributions before correction for population structure - - - - - - - - - - - 57
XII. Figure 4: Q-Q plot of RFI GWAS distributions following correction for population structure - - - - - - - - - - - 58
XIII. Table 2: Gene set enrichment analysis-SNP results for the three pathways that tended to be associated with feed efficiency - - - - 59-61
XIV. Figure 5: Running Enrichment Scores Identifying Leading Edge Genes for REACT_17015.1, Metabolism of Proteins - - - - - - 62
XV. Figure 6: Running Enrichment Scores Identifying Leading Edge Genes for P00020, FAS Signaling Pathway - - - - - - - - 63
XVI. Figure 7: Running Enrichment Scores Identifying Leading Edge Genes for GO:0044706, Multi Multicellular Organism Process - - 64
I. Introduction

American agriculture is among the most innovative production systems in the world. The need to feed a growing human population, coupled with alterations in climate that affect ecosystem services critical to food production, presents challenges and opportunities to develop both new efficiencies and environmentally sustainable practices. In 2010, California, Idaho, Montana, Oregon and Washington (NW) produced 32% of the nation’s milk supply, 12% of the beef and 28% of the lambs marketed in the US, and accounted for $11.9 billion in livestock inventory value (NASS, 2011; NASS, 2010) and $16.7 billion in sales of ruminants and their products (NASS, 2011). In many global climate change scenarios, this rich agricultural area is threatened with serious water quality and scarcity issues, increased wildfires and insect outbreaks and accelerated weed invasion. Regional temperatures are projected to increase by 3-10°F within the century (Karl et al., 2009) resulting in declining snow pack and reduced summer stream flows. This reduction in water supply will coincide with growing water demand from agricultural, municipal, and recreational needs.

In the same time scale as the alterations in climate are predicted, the population of the world is expected to approach 9.2 billion. In fact, worldwide food demand in 2050 is expected to be 70% greater than that of 2010 exerting a greater than ever pressure on agricultural production (World Summit, 2009). This increase will affect all of agriculture, but is expected to especially affect beef production as worldwide income increases and demand for beef summarily rises. The challenge is to increase beef production to meet demand and to use less resources (e.g. land and water). To
accomplish this task, metrics must be developed for measurement of efficiency that will promote efficiency gains once widely adopted and implemented.

At the animal level, improvements can be made in several key realms. One critical area is genetic selection. Given the animal-animal variation in most traits, it is possible to use selection strategies to enhance efficient production. For example, feed costs in beef cattle yearling finishing production settings make up 77% of total operating costs (Anderson et al. 2005). Because feed costs make up the majority of total costs, a 1% improvement in feed efficiency (the amount of feed necessary to provide a certain weight gain) has the same economic impact as a 3% increase in rate of gain (pounds/time; National Consortium, 2012). Feed efficiency improvements reduce feed costs without impacting weight gain and represent an important area of research.

The source of genetic improvement in the beef cattle industry is located at the individual farm level; very little vertical marketing strategy exists. Therefore, for genetic improvement to be efficacious it must penetrate to the level of the beef cattle breeder and be relevant to his or her operation. Conventional selection strategies for traits like feed efficiency require a long time and extensive resources to determine an individual animal’s phenotype and then make selection and breeding decisions. Recent advances in genomic sciences have allowed unprecedented research in this field to occur and have resulted in the development of molecular-based selection strategies. These strategies, if deployed across the industry, may have a significant positive impact on the sustainability of the beef cattle industry worldwide.

_Literature Review_
Feed Efficiency in Beef Cattle

Until recently, the typical producer’s method of improving efficiency involved producing cattle with higher rates of gain to reap an assumed reduction in maintenance costs for these faster growing cattle which were marketed at a younger age. Selection for cattle with high rates of gain and fixed harvest weight favors large frame size (Luiting et al., 1994). The relatively fixed harvest weight of today’s production setting causes these large-frame animals to be harvested at a younger and more efficient age, thus explaining their higher feed efficiency. This trend, however, is not efficient when the entire production cycle is considered. If and when slaughter weights change, the perceived efficiency is no longer apparent. In addition, breeding stock of large-frame progeny have themselves higher mature weights and increased maintenance energy needs. These confounding factors cause breeding for larger frame sizes to inherently cause higher feed costs, intake requirements, and environmental impacts for the entire beef production system (Okine et al., 2004).

Prior to recent developments in data collection, feed efficiency was difficult and costly to estimate. For this reason, these types of data were limited and expensive. The only efficiency statistics commonly measured were feed intake and average daily gain. The relationship between feed intake and gain is not linear (Ferrell and Jenkins, 1984). Optimal efficiencies for these two statistics may occur at different intake levels. Loss of efficiency at high metabolizable energy (ME) intake levels have been attributed to differential internal organ energy requirements, differential muscle mass, immune function, stress response, and higher heat increment at increased ME intake levels (Ferrell and Jenkins 1984; Richardson and Herd, 2004). Estimating feed efficiency using
these metrics is accurate only as long as factors that affect daily gain and efficiency have been controlled for. These limitations caused feed efficiency estimation to be economically infeasible in a production setting, and are one of the reasons that feed efficiency estimation has encountered limited adoption in the commercial beef industry (Wulfhorst et al., 2010).

A relevant and affordable measurement of feed efficiency would be beneficial for more than its intrinsic value as a proxy cost-analyzing metric. Within the beef cattle herd, feed efficiency can provide a measure of herd health and management strategy once a baseline measurement has been obtained. Furthermore, comparison between herds can be facilitated with certain metrics of efficiency thus providing a basis for analysis of herd management as performed by various operators.

**Residual Feed Intake**

Latest developments in data collection have allowed the beef industry to pioneer new tests and metrics for the measurement of feed efficiency. Residual feed intake (RFI) is the most widely-adopted metric for quantifying feed efficiency of today’s beef cattle herds. Although feed efficiency may be confounded by many factors including composition of gain, RFI is phenotypically independent of the production traits used to compute expected intake (Carstens and Tedeschi, 2006). Residual feed intake is defined as the difference between metabolizable energy intake (MEI) and metabolizable energy required (MER: Okine et al., 2004). This allows RFI to have increased sensitivity and precision based on feed quality and energy inputs and outputs. Residual feed intake estimates an individual’s consumed feed utilization by subtracting dry matter
intake (DMI) from predicted DMI using a regression between DMI, average daily gain, and metabolic mean weight across fed cohorts (Basarab et al., 2003). Animals exhibiting positive feed efficiency have MEI greater than MER, while animals whose MEI less than MER exhibit negative RFI and require less energy than is estimated, potentially eating less to gain the same amount of weight (Okine et al., 2004). The magnitude of the RFI value is an approximation of the degree to which their efficiency, or lack of efficiency, exists.

To further explore the relationship between RFI and efficiency, let's consider two animals with divergent feed efficiency. Animal A has been selected for efficiency and therefore possesses an RFI of -3.00. Animal B, on the other hand, is much less efficient and possesses an RFI of +3.00. If, over the course of a day, both animals gain a single pound of weight, animal A will consume 6.00 lbs. less feed than animal B, and 3.00 lbs. less feed than the average of the cohort to which animals A and B belong. Animal B will consume 3.00 lbs. more feed than the average of the same cohort, and 6.00 lbs. more than animal A. Therefore Animal A is more efficient. This illuminates the meaning of the seemingly backwards RFI values, with greater efficiency represented by lower or negative RFI values.

RFI avoids many of the inherent inaccuracies present with the ‘ratio’ traits such as feed conversion ratio (FCR) (intake/gain), partial efficiency of growth (ratio of weight gain to feed intake less maintenance requirements), Kleiber ratio (gain/metabolic mid-weight), and others. Although these ratio traits are simple to calculate, a greater amount of inaccuracy is present in their calculation. ‘Residual’ traits such as RFI avoid a few significant sources of error such as: 1) potential for increase in error variance as a
proportion of the total variance; 2) strong correlation between the ratio and its component traits; and 3) a lack of distinction between the energy used for various purposes (Berry and Crowley, 2013; Gunsett, 1984). Because residual traits are calculated using a least-squares regression, they are independent of their component traits. Although this does not imply genetic independence, it has been found that most estimates of genetic correlation between RFI derived using least squares regression and traits included in the regression model are not significantly different from zero (Herd and Bishop, 2000; Arthur et al., 2001). Selection for ratio traits causes selection pressure in a certain direction that may not be desirable. For example, FCR and ADG are highly correlated. Selection for FCR will therefore, due to correlation, be selection for ADG, which is favorable in larger-framed animals. Because residual traits are nearly genetically uncorrelated, selection for them should not cause appreciable selection in any direction for traits other than RFI (Herd and Bishop, 2000).

Residual (net) feed intake was originally proposed for use in beef cattle production by Koch et al. (1963). From a statistical standpoint, RFI is defined as the residuals, or difference between actual and predicted value, of a linear regression of dry matter intake on such energy sinks as average daily gain (ADG) and estimated metabolic weight: metabolic mean weight (MMWT). This regression is performed using a least-squares approach. Alternatively, RFI can be calculated using standard feed tables such as those published by the National Research Council or other accepted sources to determine the demand for energy posed by each sink and subtract them from the total ME intake (NRC 2001). Residual feed intake calculated in this manner is often termed 'nutritional RFI' (Berry and Crowley, 2013). Because of the inherent
mathematical difference in the calculation methods, a set of RFI values derived using the least-squares method will have an average of zero, while those calculated using tables or other standardized materials will not.

The aforementioned energy sinks are derived from weight data as ADG and MMWT. Metabolic mid-weight is typically defined as live-weight mid-way through the feed test period to the power of 0.75 (Lancaster, et al., 2005; Nkrumah et al., 2007a). This value seeks to estimate the organ mass of the stomach complex, intestines, heart, lung, kidney, and spleen, but can be affected by differences in body composition of gain. Protein and fat require differential energy demands, and for this reason it has been recommended to correct the MMWT value for body composition based on ultrasound measurements (Berry and Crowley, 2013). Fat and protein gain values can be included in the regression model for increased precision. Depositing protein (muscle) requires less energy than depositing the same amount of fat, therefore animals with a greater proportion of lean gain will appear to be more efficient, all else held equal. If not accounted for, this characteristic could potentially contribute to selection for later maturing animals which could again have negative consequences for the overall efficiency of the beef cow herd (Berry and Crowley, 2011). A correlation between RFI and breed has been reported in a number of studies, with early maturing (British) breeds such as Angus and Hereford reporting lower RFI than later maturing (Continental) breeds such as Charolais and Limousin when composition of gain is not a factor in the calculation of RFI (Crowley et al., 2010). To avoid this potential for selection pressure, it is recommended to include some measurement of composition of gain in the regression model.
Feed trial facilities are the origin of much of the currently recorded RFI data, due to the ability for accurate recording of dry matter feed intake and animal weight records at these purpose-built locations. GrowSafe Systems Ltd., and others, have manufactured feeding equipment for the accurate measurement and recording of feed consumption data using scales that continuously monitor feed bunks and individual radio-frequency identification tags that recognize and record each specific animal’s feed consumption. Animals enrolled in these trials are periodically weighed to determine ADG and compile an accurate weight record. These systems account for weight changes due to snow, rain, wind, and other environmental effects in order to maintain accuracy and data integrity. This correction is provided by on-facility weather stations that constantly monitor for confounding effects and use an algorithm to correct for their impact on feed bunk weight (GrowSafe Systems Ltd., 2011).

Individual feeding stations allow for estimation of animal activity level, which can also be included in the regression model to explain further RFI variation. Activity is approximated by frequency of visits to the feed bunk, and has previously explained 3% to 4% of RFI variation (Basarab et al., 2011; Durunna et al., 2011). Although the explained variance in this case is not extreme, it is significant and should be included when feeding frequency data is available.

Further investment in residual feed intake necessitates a review of the maintenance of its accuracy over time. If data gathered on animals at an immature age do not correlate with actual feed efficiency or RFI when mature, there is no benefit to acquiring these data. However, Arthur et al. (2001) discovered that genotypic correlations between residual feed intake in animals of weaning age (274 days) and
yearling age (430 days) were 0.75. Although the age differential between the measurement periods is only 166 days, the genetic correlations indicate that RFI is most likely to be correlated with feed efficiency throughout life. This is favorably compared to genetic correlations of feed conversion ratios at 0.42 for the same animal ages (Arthur et al., 2001). Confidence in the lasting accuracy of RFI measurements as an animal ages can be relatively assured as these data show.

Because accuracy of the linear model depends on linearity of its components, it is necessary to ascertain whether relationships between feed intake (FI) and ADG and MMWT are themselves linear. Although recent studies appear to assume relationships between FI and ADG measurements are linear at least through the duration of the data collection period, this linearity is not inherently ensured. Koch et al. (1963) previously questioned whether linearity of the relationship between FI and ADG is present, and concluded that linearity over a wide range of feed intake, age, or gain values is unlikely. An example might be ADG tapering at advanced age as compared to contemporaries of young age. However, Koch et al. (1963) also stated that linear regression should remain telling over the limited data range present within typical studies. Linearity of the relationship between FI and ADG should not be assumed for studies that include large intake, age, or gain variation (Koch et al., 1963). Feed efficiency can never be directly measured, and this limitation is important to remember when making assumptions regarding its calculation. Additional analysis of these relationships holds great potential for further research within the discipline.

Further variation in RFI can be explained using contemporary groups made up of animals with shared characteristics such as origin, age, or weight at beginning of feed
test. Prior studies have seen an increase in the proportion of explained variation in feed intake from 38% with only ADG and MMWT represented in the linear model to 72% when a contemporary group response was also included. When analyzing the variation between these input factors, adjustment for fixed effects should be considered as part of the model (Crowley et al., 2010).

Previous studies have posed inquiries into the nature of RFI heritability and the genomic basis of feed efficiency. Meta-analysis of up to 39 of these publications determined that RFI possesses heritability of $h^2=0.33\pm0.01$ (Berry and Crowley, 2013). This heritability value suggests that RFI can be improved in a similar manner as other moderately heritable traits such as ADG or milk production. The nature and calculation of RFI means that it does away with many of the confounding factors surrounding feed efficiency calculation, which should aid its acceptance with beef producers.

**Biological Basis of Divergent Feed Efficiency**

Variation in residual feed intake between various animals has previously been explained through a variety of means. A study performed by Richardson and Herd (2004) was performed after a generation of selection for divergent RFI identified the following factors in feed efficiency between low-RFI and high-RFI animals: activity (10%), digestibility of diet (10%), body composition (5%), heat increment from ruminal fermentation (9%), feeding behavior (2%), and protein turnover, tissue metabolism, and stress (37%). The cause of a further 27% of variation remains unknown (Richardson and Herd, 2004; Okine et al., 2004; Carstens and Kerley, 2009) These differences are not all-inclusive, as there are many other studies purporting the existence of other
factors affecting feed efficiency of cattle. However, the mentioned study performed an inclusive inquiry into the biological nature of differential feed efficiency. The contribution of such research is important to understanding the biological basis responsible for feed efficiency. Comprehending the influence of these is important to ensure that selection for RFI does not cause a concomitant reduction in fitness (Okine et al., 2004).

*Animal Activity and Feeding Behavior*

Physical activity and feeding behavior play a significant role in efficiency. According to a study of poultry RFI by Luiting et al., (1994), activity level contributed to 80% of the variation in RFI. Similar values have not been reported in cattle studies (as previously mentioned), although many inquiries into correlations between animal performance and feeding activity have been made. A positive correlation exists between total daily feeding duration and RFI (Basarab et al., 2011). Low-RFI animals tend to visit the feed bunk less often than those exhibiting high-RFI (Durunna et al., 2011). The same study found that feeding activity accounted for only 4% to 5% of variation in RFI when calculated using a linear regression model. Although Richardson and Herd (2004) assigned 10% of the variation in RFI to activity, the other estimations may be lower because they take into account solely activity associated with feeding and nothing else. Differences in space allowance and production setting make a difference in activity levels of cattle, and this may be a contributor in the dissimilar findings of activity level’s contribution to RFI (Roca-Fernández et al., 2013).

The energy cost of feeding has been calculated in cattle to be significant, and ingestion rate and duration of feeding are reportedly key factors in determining this cost.
(Susenbeth et al., 1998). Energy spent feeding is closely related to total time spent eating, thus providing further credibility to the results found in the Richardson and Herd (2004) study. Consumption of roughage feedstuffs of varying quality can require 10%-30% of the ME provided by the feedstuff (Susenbeth et al., 1998). These differences were calculated to be responsible for about 2% of the observed difference between the high-RFI and low-RFI animals in the Richardson and Herd (2004) study. The impact of feeding behavior is dependent on the diet fed to the animals (Roca-Fernández et al., 2013).

Robinson and Oddy (2004) analyzed the relationship between RFI and behavioral feeding traits and reported similar positive phenotypic and genetic correlations. However, once the published phenotypic correlation matrix was adjusted for differences in feed intake, there was no phenotypic correlation between RFI and feeding time (-0.01) or eating rate (-0.01) and a weak correlation (0.10) with number of bunk visits. These data suggest that a significant proportion of the observed association between RFI and feeding behavior is an artifact of lower average feed intake in low-RFI animals, at least in this study (Berry and Crowley, 2012).

Although current estimations of the impact of activity and feeding behavior on RFI are moderately low, it is useful to note that heritability estimates for feeding behavior in growing cattle vary from 0.36 to 0.51 (Robinson and Oddy, 2004). This suggests potential efficacy of selection for feeding behavior traits, and possibly activity traits, once economic benefit of such selection is made evident. Cattle activity monitors are presently available, and would have herd management benefits such as detection of estrus and lame and sick animals. Adoption of this technology is mainly limited to
dairies, but could be integrated into selection criteria for beef cattle pending producer demand (Berry and Crowley, 2012).

**Digestibility of Diet**

The digestibility of the diet is another significant factor in feed efficiency. There are many interactions that occur in the digestion process that influence the relative utilization of feedstuffs by cattle. It is widely known that ME intake levels relative to maintenance requirements affect digestion (Richardson and Herd, 2004). There remains, however, a genetic component associated with efficiency of digestion that appears to be a physiological factor in feed efficiency. Channon et al. (2004) selected for steers of low-RFI and high-RFI, and then measured fecal pH to determine starch digestion. The progeny of low-RFI parents generally had lower fecal pH levels than those of high-RFI breeding indicating variation in starch digestibility between the two groups. Fecal pH was highly correlated ($R^2=0.95$) with total starch digestibility (Zinn, 1994). The lower fecal pH in the relatively efficient animals indicates that they received a greater proportion of the starch nutrients from the diet, thus contributing to their efficiency. The cause of these digestive distinctions is unclear, but rate of passage differences may account for a portion of the between-animal variation. Further study examining the microbiome of the ceca and the measurement of volatile fatty acid and lactic acid products of hindgut fermentation may be effective in further understanding the cause of these differential results.

**Body Composition**
Differential composition of gain plays a role in the physiological differences that exist between animals and affect efficiency. As previously mentioned, the addition of ultrasound measurements of body composition remains an important consideration in the calculation of RFI. Differential body composition during gain and maturity is a heritable trait in growing cattle (Nkrumah et al., 2007b) and cows (Pryce et al., 2002). Loss and gain of body condition score is included in the calculation of RFI for lactating animals, because it greatly affects perception of efficiency due to the utilization of body energy reserves instead of energy derived from feed. Preventing apparent but spurious feed efficiency estimation with unstable body composition during a feed test is important for maintaining the validity of the trait. While ultrasound measurements go a long way towards maintaining accuracy, it is important to take into consideration the presence of such confounding phenomena when calculating and comparing animal feed efficiency (Basarab et al., 2011). Because body condition score is a commonly-recorded and observed trait, including it in a selection index would be a simple proposition once standards have been created that maximize producer benefit and value of such selection is made known.

In the Richardson and Herd (2004) study, blood metabolite profiles differed significantly between low-RFI and high-RFI animals. Blood leptin concentration was positively correlated with RFI, and is an indicator of fatness in cattle (Brandt et al., 2007; Minton et al., 1998; Geary et al., 2003). This finding was supported by a carcass analysis of the animals showing the high-RFI group to have carcasses with increased lean composition. Thus perhaps blood leptin concentrations could also be used in refining the determination of RFI.
**Heat Increment**

As ruminal fermentation occurs, heat is produced which constitutes a loss in efficiency and is termed heat increment of feeding (HIF). HIF losses in sheep during the act of eating could account for roughly 40% of total HIF (Webster et al., 1975). The feeding behavior and duration of animals selected for differential RFI has been previously examined, with high-RFI cattle spending more time eating than their more efficient counterparts. Heat increment has a hypothesized impact on RFI of 9%, much of which may be due to aforementioned feeding behavior differences. However, Richardson et al. (2001) found no significant difference in estimated heat production between steers selected for high-RFI and low-RFI. Further research into the role of heat increment on feed efficiency is necessary to make determinations on how best to provoke greater efficiency from cattle in this regard.

**Protein Turnover, Metabolism, Stress, and Mitochondrial Function**

The constant energy requirements of protein turnover and metabolic function are further considerations for RFI differences and may explain a great deal of the between-animal variation. Richardson and Herd (2004) previously estimated the explained variance of these effects to be 37%, representing the largest single factor considered in their study.

Protein turnover is defined as catabolism and anabolism of body protein (Richardson and Herd, 2004). Several observations have been made that support the hypothesis of protein turnover being a significant factor in RFI (Richardson and Herd, 2004). Primarily, low-RFI cattle have a higher proportion of lean mass at slaughter, thus
indicating a more efficient system of protein deposition with lower rates of protein degradation as compared to animals in the high-RFI group. Another factor is the positive association between blood urea nitrogen (BUN) concentrations and RFI in previous studies by Richardson et al. (2001). Fluctuations in BUN can occur due to many factors including degradation of dietary protein in the rumen, increased urea recycling at the liver and kidney level or increased protein degradation in the liver. Furthermore, it was found that BUN concentrations are not indicative of divergent feed intake, suggesting other mechanisms are responsible. A final observation supporting this hypothesis is the positive correlation between RFI and concentration of blood aspartate amino-transferase which is a marker of liver function (Richardson et al., 2001). This finding indicates that higher levels of protein catabolism occur in livers of less efficient animals. These findings, when taken together, may represent significant metabolic differences between high-RFI and low-RFI animals.

Accompanying tissue metabolism, stress response has an impact on efficiency that is moderately significant. Just as in other animals, stress response in cattle is determined in part by genetic influences which contribute to greater or lesser glucocorticoid release when stressors are encountered. The glucocorticoid family of steroid hormones, of which adrenaline is a member, moderate the bodily stress response. When stressed, organisms increase their rate of metabolism in order to produce adequate amounts of cellular energy for movement and alert demeanor. This increase in basal metabolic rate is a factor in feed efficiency, as animals more frequently stressed will use, on average, a greater amount of energy (Mader, 2003).
Significant positive relationships have been found between average blood cell levels and RFI. This increase in average leukocyte and erythrocyte concentrations is caused by splenic contractions as a result of stress response following excitation (Gartner et al., 1969). In addition to other effects, the previously described glucocorticoid release elicits changes in the leukocyte profile including an increase in the neutrophil to lymphocyte ratio. This event was observed in cattle studied by Richardson and Herd (2004). The ratio was further elevated in less efficient animals (Richardson and Herd, 2004). Furthermore, blood cortisol levels have a positive correlation with sire estimated breeding values for RFI, while offspring of high-RFI sires have phenotypically higher blood cortisol levels as compared with low-RFI contemporaries (Richardson and Herd, 2004).

Stress response observed in metabolic systems also potentially plays a large role in the divergent efficiency of animals with various stress response. Beta-hydroxy butyrate (βOH) and BUN at weaning were shown to be positively correlated with RFI in a group of steers studied for feed efficiency (Richardson and Herd, 2004). Beta-hydroxy butyrate can be indicative of lipolysis in animals under stress (Warriss, 1984; Whitaker, 1997). BUN, as previously discussed, can be indicative of protein catabolism and ammonia absorption (Cameron, 1992). In previous studies, stress has been reported to promote energy mobilization in the form of lipolysis (Brockman and Laarveld, 1986) and protein catabolism (Cole et al., 1988). Furthermore, βOH presence has also been linked to stress provoked by feeding interruption (Warriss et al., 1995). The presence of these metabolites is indicative of the energy partition and metabolic changes associated with stress and the inefficiency of said changes. Because anabolism of lipids requires
greater energy input than is returned in lipolysis, whenever lipolysis is experienced future efficiency and lifetime carcass quality is compromised. Intramuscular fat (IMF) is highly desirable in beef cattle carcasses, and is the main determinant of quality grade (Wheeler et al., 1994). Intramuscular fat is the last fatty tissue to be deposited and is the first fatty tissue to be catabolized during lipolysis (Jeong et al., 2012). Because of this, lipolysis in beef cattle on feed is undesirable and has great impacts on carcass merit, quality grade, and carcass value.

Stressors inherently exist in production settings, and owing to that fact these effects can have large impact on performance if animals are constantly stressed due to hyperactive stress response followed by glucocorticoid release. Further inquiry on the difference in stress response and feed efficiency between cattle raised in varying production settings is an area in which further research must be conducted. It is important to note that stress response could theoretically be reduced to nothing if all environmental and internal stressors were removed. Of course, this statement ignores basal endocrine release that causes glucocorticoids to be released in the absence of stimuli (Borer et al., 1992). Inquiry into basal glucocorticoid release in animals selected for divergent RFI is also a potential area of further research. At the current research state, the relationship between stress and efficiency is very clear and the results of past experiments extol the numerous benefits of reduced environmental stressors and diminished stress response for greater efficiency.

From a molecular standpoint, mitochondrial energy generation has long been suspected as a source of between-animal feed efficiency variation. Ramos and Kerley (2013) examined mitochondrial respiratory complex 1 protein via microplate assay and
identified differences in complex 1 protein levels between the most and least efficient animals. All four of the respiratory chain complexes showed greater activity in low-RFI broilers as compared to high-RFI animals. The rate of mitochondrial respiration was greater in more efficient cattle, and there was a significant correlation ($P < 0.05$) between the ratio of respiratory complex I/respiratory complex II and RFI and between the ratio of respiratory complex I/respiratory complex III and RFI. In addition, more efficient animals displayed lower rates of both electron leak and reactive oxygen species production (Bottje et al., 2004). These results provide evidence to support the molecular impact of mitochondrial function on energy metabolism and feed efficiency.

*Other Impacts of Selection for Feed Efficiency*

Although one of the most attractive aspects of RFI is its independence from other genetic traits, selection for individual phenotypes always has potential to introduce deleterious effects or traits. Kerley (2010) determined that during times when forage is abundant for livestock consumption efficient cows require less forage, and therefore consume less, and during times when forage quantity is limited efficient cows will outperform inefficient cows. This means that efficient herds allow for greater carrying capacity on pasture, and during periods of modest resources, herds can be sustained on less feed. The benefits from such characteristics are two-fold; with fewer required feed resources and a reduction in the cost of performance.

Berry and Crowley (2013) briefly mention the correct protocol in regards to data collected during animal illness. Although some would argue that such data should be discarded, the possibility exists for RFI to be somehow linked to disease susceptibility. If
this is the case, discarding the data would be a mistake, as it would skew efficiency in favor of increased disease risk. These potential interactions are distinct possibilities, and the animal community must be vigilant to prevent deleterious effects from diminishing the utility of RFI.

Although the energy needs of a single bodily system may seem insignificant, each and every energy sink that is not taken into account is a source of inaccuracy for the calculation of RFI. In a recent study by Carroll and Sanchez (2014), energy usage by the immune function is analyzed. Immune function represents a large energy sink for the body, especially when one considers the implications of raising body temperature during a fever episode. Research by Kluger and Rothenburg (1979) found that raising the bovine body temperature by a single centigrade degree requires an increase of 10-13% metabolizable energy. Carroll and Sanchez noted that body temperature rose roughly 2 degrees centigrade following a lipopolysaccharide immune challenge. Furthermore, Van Eerden et al. (2004) found that divergent RFI layer hens mounted differential immune responses when exposed to an antigen. It is believed that less efficient animals have more energy to devote to immune response, and are therefore able to mount a greater antibody defense when compared to low-RFI animals. It would be prudent to ensure that efforts to select for feed efficiency do not lead to animals with reduced immune function.

**Limitations of RFI**

In developing RFI as a tool for producers and researchers to rely on, there is progress to be made in the standardization of the metric. For instance, because RFI
data calculated using least-square regression cannot be compared between farms or studies without the original data, regression coefficients should be included in every article published on the subject. This will facilitate collaboration and allow valuable comparisons to be made and updated (Berry and Crowley, 2013). It will also allow progress to be measured, as past studies provide baselines for future data and research. Another aspect of RFI that needs to be addressed is its complexity. The findings of Wulfhorst et al. (2010) concluded that nearly 50% of producers would be interested in adopting RFI for their operation and that respondents consider feed efficiency to be an important trait for their operation. This details the value producers see in improving feed efficiencies, and therefore saving input costs spent on feed. However, the same study also concluded that RFI is “. . . complex and not readily understood . . .” even by those specifically trained in animal agriculture (Wulfhorst et al., 2010). This perception must change if residual metrics are to have a place in production agriculture. As the study of genomics advances into new realms, it is expected that some form of marker-assisted selection will allow prediction of feed efficiency without the need to collect feed trial data, but instead a DNA sample to predict phenotype. To discover such markers necessitates effective statistical tests for association between phenotype and genomic data.

* Genome-Wide Association Studies *

Regarding population genetics, the identification of significant alleles within genotype data is of utmost importance. Genotype data is the genetic code within the genome and is in part responsible for all components of bodily function and operation. However, because of the sheer length and complexity of genotype data, identifying
particular loci associated with a trait can prove a great challenge. Genome-wide association studies (GWAS) are a method of identifying regions of the genome associated with a trait. Genotypic data is represented as the identification of alleles, or variant genes, present at each locus. Thousands to millions of DNA variants known as single nucleotide polymorphisms (SNPs) are present within a bovine genome. These SNPs are common variations at particular loci where single nucleotides vary between members of the species. The resolution of detection of SNP markers is dependent on the genotype panel used in the study, with common assays including 7,000, 9,000, 11,000, 50,000, and 778,000 SNP identifications. Sequence data provides even higher resolution and allows for the measurement of allelic state at every locus. For lower resolution genotype data, even if a causal locus is not genotyped in the study, identification of association is still possible due to the manifestation of linkage disequilibrium (LD) via genotyped SNPs nearby (Astle and Balding, 2009).

Linkage disequilibrium is the non-random association of one SNP with another. When in LD, alleles in close proximity to one another on the genome do not independently assort when transferred across generations during the process termed chromosomal crossover. Because crossover is less likely to occur between these loci, they are said to be in LD and are more likely to be inherited in the parental allele order than in nonparent combinations (Astle and Balding, 2009). Genome-wide association studies rely on LD because not all SNPs are identified except in the case of sequence data. Thus it is necessary to use nearby SNPs in LD with causative mutations to identify associations of a genomic region with a trait. This solution is workable for a time until a recombination event occurs between the linked SNP and the true causative mutation. In
this case the LD breaks down and the association is no longer valid (Reich, 2001). LD is both a help and a hindrance to association studies because although stronger LD makes for simpler detection of a causal variation, it also becomes more difficult to precisely map a region due to the many markers highly correlated with the phenotype thus additional computation to determine the causative mutation is required (Astle and Balding, 2009).

These data remain undescriptive but for the ability to determine association of genotype data with phenotypic occurrences. The most important consideration and most frequent confounding effect and source of spurious association in GWAS is population structure. Population structural effects result in a number of consequences that can produce spurious marker-phenotype associations if not corrected for (Astle and Balding, 2009).

When populations have shared ancestry that differs from that which one would expect with panmictic (random-mating) populations, relatedness, or kinship, is introduced into the population structure. For domesticated animals, mating is often controlled and deliberate to leverage any sort of gains that can be made through genetic improvement. Unfortunately for the purposes of association studies, this kinship can generate spurious marker-phenotype associations. In a given population experiencing population stratification, many individuals may possess alleles that represent a certain characteristic widespread within the population, yet has no effect on the phenotype of interest. For example, consider a group of cattle, all with long horns. Many genetic variants are likely to show association with the phenotype “uses horns to scratch.” These alleles will be common within this group of cattle, but do not necessarily “cause”
scratching, which is a learned behavior within the herd (Balding, 2006). Such an association is considered to be spurious and confounds results. Studies with distinct subpopulation groups or studies with populations more closely related with one another than with controls are more likely to exhibit population stratification (Balding, 2006). When such a confounding effect is observed, it is important to correct for its presence and minimize its influence. One method of correcting for population stratification is through the use of mixed models that employ estimates of relatedness known as kinship matrices to correct for one cause of population stratification: relatedness (Astle and Balding, 2009). The kinship matrices are developed using genotype data from the population studied. Mixed model correction is employed in many studies and is especially important for those using data derived from populations believed to suffer from population stratification.

*Gene Set Enrichment Analysis-Single Nucleotide Polymorphism*

While the potential for association studies to determine significance of individual markers cannot be understated, detection of loci with modest effects is often hindered by their insignificant individual effects. As is often the case with biological systems, individual markers usually represent genes that exist as members of biological gene sets that effect change on a larger scale than that of their individual members.

Realizing the limitations of the GWAS approach, gene set enrichment analysis-single nucleotide polymorphism (GSEA-SNP) has been developed to determine the effects of genes when combined into complex biological pathways (Wang et al., 2010). The development of gene pathway-based analysis of GWAS data was motivated by the
use of similar methodology in the analysis of gene expression microarrays (Wang et al., 2010). In past years, GSEA-SNP has become further advanced and currently plays an important role in the analysis of genomic data by utilizing SNPs as proxies for each gene rather than the gene’s expression.

One imperative of performing GSEA-SNP is determining gene sets to be used in the analysis. Gene sets are derived using data from prior research and are separated into categories based on function. Many thousands of gene sets have been created and are freely distributed for use in GSEA-SNP. Furthermore, pathways can be manually curated from recent literature for specific studies based on the needs of a particular area of research, which is especially useful if said discipline is highly specialized or relatively novel. Gene sets are expressed as a list of genes with hierarchical relationships (Wang, 2010). SNP markers are assigned, or ‘mapped,’ to genes within these gene lists. This again works under the principle of linkage disequilibrium. Linkage disequilibrium pruning is then carried out to reduce the instance of multiple SNPs being mapped to the same gene. Multiple mapping in this manner has the potential to introduce bias into the analysis because of varying SNP density and coverage, in addition to biasing towards large gene sets (Wang, 2010). Pathway enrichment scores are calculated with a running sum statistic using GWAS results and the pathways previously mentioned (Wang, 2010). When a gene ranked highly in GWAS (highly associated) is encountered in the pathway being analyzed, the running sum increases. However, when genes not present in the pathway are encountered, the running sum statistic decreases a certain fixed amount such that the running sum statistic reaches zero when the ordered gene list has been completely parsed. The enrichment score
(ES) is the height of the peak reached by the running sum statistic. Enrichment scores are then normalized using the number of genes in the pathway as an adjustment factor. This prevents random chance from allowing larger pathways to have higher significance (Wang, 2010). Correction for multiple testing is applied using a false discovery rate estimate (FDR). False discovery rate estimates the proportion of type 1 (false positive) errors expected to occur when designating a pathway significant.

During GSEA-SNP, significant pathways may be determined along with the genes that make up those pathways. Genes that compose various pathways are ordered based on their contribution to the enrichment of a certain pathway, with primary genes termed 'leading edge genes.'

In this study, the loci associated with RFI were investigated. Genotypes were obtained from DNA isolated from leukocytes and analyzed by GSEA-SNP. Based on similar studies seeking to explain individual variation in RFI, genes expected to play an important role in RFI include those involved in feeding behavior, stress response, immune function, and protein metabolism (Richardson and Herd, 2004; Okine et al., 2004; Van Eerden et al., 2004).

II. Thesis Activity

Residual feed intake phenotypic data were collected along with genomic data for a group of beef cattle. Gene set enrichment analysis-single nucleotide polymorphism was performed on these data to identify loci associated with feed efficiency through the measurement of RFI. The statistical analysis involved in this study represents a method of determining association between genes that work together in a logical set, or
‘pathway’ to effect bodily change, in this case bodily change with significant effect on residual feed intake. The identification of association of these genes with RFI provides a means of using genomic information to predict animals that will be feed efficient. Feed efficiency predictions will then be included in sire summaries so that more feed efficient offspring may be bred. Therefore, the objective of the study was to identify gene sets associated with residual feed intake and, consequently, beef cattle feed efficiency.

III. Materials and Methods

Animals

A group of 851 Hereford cattle (828 steers and 23 heifers) from a single ranch in Nebraska were fed together between 70 to 140 days, with only 31 animals fed for over 72 days. The animals ranged in age from 210 to 496 days when first started on the feed trial. During the feed trial, average daily gain (ADG), dry matter intake (DMI), initial weight (IW), mid-test metabolic weight (MMWT), and days on feed (DOF) were recorded. Further grouping of the animals was based on slaughter date, and 7 contemporary groups were formed referring to the various groups of animals enrolled in the feed study at different times.

DNA Sample Collection

Approximately 20 ml of whole blood was removed from each animal via venipuncture of the coccygeal or jugular vein and collected in EDTA-containing blood vials. Samples were placed on ice or kept at 4°C until the DNA was extracted. DNA was extracted at the University of Missouri using a phenol-chloroform extraction protocol (Sambrook and Russell, 2001).
**Phenol-Chloroform DNA Extraction**

Eight ml of whole blood sample was placed in a 50 ml conical centrifuge tube containing 8 ml 1:1 phenol:chloroform mix. The contents were inverted several times until an emulsion formed and then centrifuged at 2000 x g for 1 minute at room temperature. If the organic and aqueous phases were not well separated, samples were centrifuged again for another 1 minute until separation occurred. The aqueous (upper) phase was removed and saved in a fresh conical centrifuge tube. The organic and interface phases were discarded following this transfer. The complete process was repeated until no protein was visible at the interface of the organic and aqueous phases. DNA was then precipitated with the addition of 1/10 volume of 3M sodium acetate (pH 5.2), and mixed via inversion. Then, 2 to 2.5 volumes (calculated after salt addition) of cold 100% ethanol was added and the tube was inverted. The samples were then placed on ice or at -20°C for 30 min. Following this step, samples were centrifuged at 2000 x g for 15 min. Once centrifuged, the supernatant was carefully decanted and 1 ml of 70% ethanol was added, inverted, and centrifuged for 30 seconds at 2000 x g. The supernatant was carefully decanted and the DNA pellet resuspended in pH 7.8 TE buffer of 100 mM Tris-Cl and 10 mM EDTA. DNA was quantified and diluted to 50ng/ml. Genotyping was conducted at GeneSeek (Lincoln, NE) with the Illumina (San Diego, CA) BovineHD genotyping BeadChip or SNP50 BeadChip. The genotypes were provided to Washington State University to undergo GSEA-SNP.

**Gene Set Enrichment Analysis**

**Calculation of the Residual Feed Intake Phenotype**
Four of the original 851 animals were removed from the statistical analysis due to missing weight data necessary for the calculation of RFI. With the remaining 847 animals, a least squares regression of DMI was performed on ADG and MMWT. The statistical model to represent this calculation was:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \]

Where \( Y \) is expected DMI, \( \beta_0 \) is the equation intercept, \( \beta_1 \) and \( \beta_2 \) are the coefficients of the equation, \( X_1 \) is the mid-test metabolic body weight, \( X_2 \) is the average daily gain, and \( \epsilon \) is the residual (Berry and Crowley, 2013).

**Mapping of Single Nucleotide Polymorphisms to Genes**

In GSEA-SNP, DNA variants serve as proxies for genes that comprise the gene sets. Therefore, a critical consideration in GSEA-SNP is the assignment of the DNA variants to the genes that they will represent. In this study, all SNPs within each gene and the surrounding 20kb on each side of the gene were evaluated as to their suitability to serve as the gene’s proxy. The SNP with the greatest evidence for association with RFI was chosen to represent that gene. This process was conducted on 20,692 genes with the location of each gene based on the UMD 3.1 assembly (Zimin et al., 2009).

**Gene Sets**

Gene sets utilized in this study were specifically chosen for GSEA-SNP in cattle that were compiled into GO2MSIG (http://www.go2msig.org/cgi-bin/prebuilt.cgi?taxid=9913). This is a gene set generator that consists of 3493 pathways that were obtained from the gene2go annotation data at the National Center

Kyoto Encyclopedia of Genes and Genomes (KEGG) holds a network of pathways classified in seven groups including metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug development. Each of these groups is further subdivided into multiple pathways identifying biological processes of various types. Gene Ontology (GO) is another database of gene sets with cellular, molecular, and biological function grouped into three primary categories from which a hierarchy of terms is derived. The large, general terms at the upper levels of the hierarchy branch to many lower level pathways, allowing the sets to be linked in this manner. Panther is another curated source that maintains a group of regulatory and metabolic pathways. Pathways within this database are ordered based on evolutionary relationships, with a series of families and subfamilies of proteins that result from expression of the pathways. Metacyc is a network that specializes in primary and secondary metabolic function gene sets, including associated compounds, enzymes, and genes. Reactome maintains a group of molecular and cellular biological pathways that are grouped into molecular events. Pathway entities include intermediary metabolism, signaling, innate and acquired immune function, transcriptional regulation, apoptosis, and disease.

*Quality Control and Covariate Determination*
Single nucleotide polymorphisms with a genotyping call rate of less than 90% or minor allele frequencies of less than 1% were removed from the analysis. Similarly, animals that had a SNP genotyping call rate of less than 90% (n=4) were removed.

Sum of squares and variance inflation factor testing also took place for possible confounding effects including population structure. Linear models were fitted to the data using one dependent variable at a time for age, sex, DOF, weight, percent Hereford, and contemporary groups 1-6 based on the various groups of animals enrolled in the feed study at different times. The results of this testing suggested that the contemporary groups showed high correlation with sex, weight, and days on feed. Correlations between: age and weight, sex and weight, percent Hereford and ADG, weight and ADG, weight and DMI, and weight and mean metabolic weight were all significant (P < 1x10^{-15}). Variation inflation factor was calculated as an indicator of multicollinearity, or correlation between two or more predictor variables, and was determined to be below 2 for % Angus breed, % Red Angus breed, age, DOF, and sex. This indicates low multicollinearity for all tested model components. With these results, it was determined that a statistically accurate covariate group could be derived from these variables. Based on correlations between these groups and the RFI phenotype, they were included as covariates in the enrichment analysis.

Scatterplots were made that compared age, sex, DOF, weight, percent Hereford, and contemporary groups 1-6 to the RFI phenotype independently. If significant grouping based on these characteristics was evident, then they were included in the model as a covariate. The contemporary group effects were redundant with sex, DOF,
and weight, so these variables were left out of the model and the final covariates included age, percent Hereford, and contemporary groups 1-6.

The analytic process is detailed in **Figure 1**. Raw genotype data were taken through quality control as previously described. Following quality control, SNPs were assigned to genes within 20 kb of their mapped location using the ENSEMBL genome browser database of *Bos taurus* genes. Once the 3493 gene sets were identified, 10,000 permutations of the GWAS results were created to form a null distribution and enrichment scores were calculated.

Gene set enrichment analysis-single nucleotide polymorphism was begun by performing a GWAS on the SNP genotypic data using GenABEL open-source software based on the R programming language (http://www.genabel.org/packages/GenABEL). **Table 1** lists the most significant individual genes and a Manhattan plot of these results can be observed in **Figure 2**. Genome-wide Rapid Association using Mixed Model and Regression (GRAMMAR) mixed linear models within the GenABEL statistical package were used to correct for the effects of population structure (http://www.genabel.org/packages/GenABEL). Final covariates included age, percent Hereford, and contemporary groups 1-6 based on various groups of animals enrolled in the feed study at different times.

Quantile-quantile (Q-Q) plots were drawn to compare the expected and observed $-\log_{10}$ P-values that were generated from the genotypes. With no association between the genotype and phenotype, the plotted line will fall on an idealized 45 degree slope. If association is present with a genotype, the line plotted will deviate with a greater slope.
than the idealized 45 degree line. Q-Q plots were created before and after applying
GRAMMAR correction. In Figure 3, the plot of the expected and observed –log10 P-
values deviated greatly from the idealized 45 degree line indicating that there was
significant uncorrected population stratification. In Figure 4, the expected and observed
–log10 P-values equaled roughly the same way as the idealized 45 degree line, with the
exception of loci that were associated with RFI. This indicates that the population
stratification shown in Figure 3 was corrected using mixed linear models as shown in
Figure 4.

The association analysis that preceded the enrichment analysis ranked the
genes by the level of association that they had with RFI. The enrichment analysis relied
on this association analysis to derive the enrichment score. Following the association
analysis, enrichment scores were determined based on the running sum statistic. A
running sum statistic was calculated by comparing the ordered gene list from the GWAS
to the presence or absence of genes in the gene sets. If a highly-associated gene was
present in a pathway, the running sum statistic increased. If, on the other hand, a
highly-associated gene was not present in a pathway, the running sum statistic
decreased (Subramanian, 2005). The increase in the running sum statistic for genes
present in the pathway was based on evidence of an association between the proxy
SNP and RFI, while a decrease in the running sum statistic resulted when a ranked
gene was encountered in the GWAS list that was not in the pathway. The maximum
enrichment score was derived from the height of the peak of the running sum statistic,
which reflected the degree to which gene sets were enriched (overrepresented) and
corresponded to a weighted Kolmogorov-Smirnov-like statistic. This procedure was repeated for each gene set used for the analysis.

The enrichment scores were normalized across pathways through the use of a permutation procedure that calculated an empirical null distribution of enrichment scores. The normalized enrichment score of each pathway or gene set was then compared to the null distribution to determine if the pathway was associated with RFI. A total of 10,000 permutations were used to determine the empirical null distribution. An FDR was calculated using the FDR q-value procedure developed by Story (2002). False discovery rate accounted for multiple testing, and was defined as the probability of a type-1 error associated with a given pathway using P-value as a metric. Gene sets were defined as enriched if the normalized enrichment score (NES) was greater than 3.0 or if the FDR q value was less than 0.05. These values were based on typical enrichment score and FDR thresholds found in the literature (Subramanian, 2005).

IV. Results and Discussion

None of the evaluated KEGG, GO, Panther, Metacyc, or Reactome gene sets were associated with RFI. Three gene sets, one from Reactome (REACT_17015.1 metabolism of proteins, NES=2.917, FDR=1), one from Panther (P00020 first apoptosis signaling (FAS) pathway, NES=2.836, FDR=1), and one from Gene Ontology (GO:0044706 multi-multicellular organism pathway, NES=2.815, FDR=1) had tendencies toward an association with RFI (Table 2). Within the pathways identified were genes that were associated with growth, development, and metabolism that could well be important in a trait such as feed efficiency (Table 2) particularly considering
literature data indicated differences in body composition and metabolism as RFI determinants (Basarab et al., 2011; Richardson and Herd, 2004).

Reactome pathway REACT_17015.1, metabolism of proteins ($P = 0.002$), contained 257 genes total and 108 leading edge genes (genes that contributed to the increase in the normalized enrichment score). A plot of the running enrichment score for REACT_17015.1 can be seen in Figure 5. The peak of the figure divided leading-edge genes (on the left of the vertical dashed black line) from non-leading-edge genes (on the right of the dashed black line). The metabolism of proteins gene set contained genes responsible for protein production including translation, post-translational modification, and protein folding (Borer et al., 1992; Chu et al., 2009; Tochigi et al., 2008). Protein turnover is a process that has been shown to be important for feed efficiency (Richardson and Herd 2004). The leading edge genes included a great number of genes controlling proteins involved in ribosomal components, which would be consistent with what is known about the processes involved in feed efficiency.

The second gene set that trended towards significance was Panther P00020, first apoptosis signaling pathway ($P = 0.0003$) which contained 13 genes of which 5 were leading edge genes. A plot of the running enrichment score for P00020 is shown in Figure 6. The FAS signaling pathway contained genes responsible for the mediation of cellular apoptotic signaling through interaction with surface-expressed Fas ligand (FasL) on cells targeted for destruction (Wajant, 2002). Cellular apoptosis, or programmed death, is related to protein turnover and the replenishment of bodily protein resources. This gene set’s trend towards enrichment would be consistent with literature relating the effects of protein turnover rates on feed efficiency (Cameron, 1992).
The third pathway identified as trending towards significance was Gene Ontology GO:0044706, multi-multicellular organism pathway (P = 0.0023), that contained 32 genes of which 15 were leading edge genes. A plot of the running enrichment score for GO:0044706 can be seen in Figure 7. The multi-multicellular organism pathway was composed of genes involved with signaling between cells of the same or different organisms. This encompassed hormones involved in reproduction and parturition (http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0044706). Although the female animals in this study were open, the trend towards enrichment experienced by this pathway is similar to results observed by Perkins et al., (2014) who found that low-RFI cattle express less gonadotropic releasing hormone, and therefore less follicle-stimulating and luteinizing hormones than their less-efficient counterparts.

V. Conclusions

As popular demand for beef products continues to swell in the face of increased consumption due to a rapidly growing protein-seeking population, it is becoming more important than ever to improve the efficiency of bovine production through the utilization of molecular biology. Research into RFI has been increasing in breadth, application, and volume over the past decade, and this type of research is expected to continue at an increasing rate into the future.

The beef industry stands to gain a great deal from integrating molecular biology and genomic information when making breeding decisions. However, despite the promise of greatly increased efficiency of production, it is important to remember the limitations of this technology that has only recently become available to producers. As the technology matures that allows researchers to sequence genomes, the ability for
association studies to discover causative mutations will increase. Genotyping with SNP BeadChips is not designed to identify the causative mutation, but rather to identify a SNP in linkage disequilibrium with the causative mutation. Eventually, linkage disequilibrium between the SNP and the causative mutation will deteriorate and the marker will no longer remain predictive of a certain phenotype. This is an inherent limitation of genotypic data that relies on SNPs in linkage disequilibrium rather than using the causative mutation as the means for predicting the phenotypes. When genotyping is supplanted by whole-genome sequencing, accuracy of the prediction of genotypes will increase and these findings will remain valid over time, because the causative mutation will be included in the sequence, unlike the current situation where marker accuracies are dependent on linkage disequilibrium. The National Program for Genetic Improvement of Feed Efficiency in Beef Cattle hopes to provide the beef cattle industry with solutions for identifying highly efficient animals in a reliable and cost-effective manner. Research to identify loci associated with feed efficiency is the first step toward achieving this goal. Genomics provides an opportunity to brighten the future of the beef industry by facilitating efficient production of wholesome, nutritious protein sources.
VI. References Cited


National Agricultural Statistics Service.

National Agricultural Statistics Service.

National Program for Genetic Improvement of Feed Efficiency in Beef Cattle. (n.d.).


doi:10.2527/jas.2013-7019

relationship between
calving interval and body condition score conditional on milk yield. *J. Dairy Sci.
85*: 1590–1595.

Ramos, M. H., & Kerley, M. S. (2013). Mitochondrial complex I protein differs among
residual feed intake phenotype in beef cattle. *Journal of Animal Science, 91*(7),
3299–3304. doi:10.2527/jas.2012-5589

199.

intake in beef cattle. 2. Synthesis of results following divergent selection.

Richardson, E. C., Herd, R. M., Oddy, V. H., Thompson, J. M., Archer, J. A., & Arthur,
steer progeny of parents selected for and against residual feed intake. *Australian
Journal of Experimental Agriculture, 41*(7), 1065–1072.

fatness, muscle area and feeding behaviour of feedlot finished beef cattle. *Livest.
Prod. Sci.* 90:255–270

communication. Behavioural activities of two dairy cow genotypes (Holstein-
Friesian vs. Jersey x Holstein-Friesian) in two milk production systems (grazing vs. confinement). *Spanish Journal of Agricultural Research, 11*(1), 120. doi:10.5424/sjar/2013111-2682


doi:10.1016/j.neures.2007.10.010


doi:10.1038/nrg2884


doi:10.1002/jsfa.2740350808


Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls (2007). *Nature, 447*:661-678.


Figure 1: Gene Set Enrichment Analysis-SNP Flow Diagram showing the processes involved in identifying gene pathways associated with feed efficiency as measured by RFI on 847 Hereford cattle fed in a GrowSafe system for 70-140 days.

1 Genome-wide association analysis (GWAS) was performed on genotype data using the mixed model regression analyses called Genome-wide Rapid Association using Mixed Model and Regression (GRAMMAR) mixed linear models; 2 Gene set enrichment analysis-SNP was performed with 10,000 permutations to derive the null distribution; 3 Linkage disequilibrium pruning was carried out to reduce the instance of multiple SNPs being mapped to the same gene; 4 Animal quality control included removal of single nucleotide polymorphisms with genotyping call rate of less than 90% or minor allele frequencies of less than 1%; 5 Animals with genotype SNP call rate of less than 90% (n=4) were removed; 6 SNPs were assigned to genes within 20kb of their mapped location using the ENSEMBL genome browser database of Bos taurus genes; 7 Gene sets utilized in this study were specifically chosen for gene set enrichment analysis-SNP in cattle that were compiled into the GO2MSIG database.
Table 1: Genome-wide Rapid Association using Mixed Model and Regression (GRAMMAR) mixed linear models GWAS Results. Genes were determined to be associated with feed efficiency based on Wellcome Trust Consortium Guidelines (http://www.wellcome.ac.uk/).

<table>
<thead>
<tr>
<th>^1BTA Chromosome</th>
<th>^2Position (bp)</th>
<th>^3Mapped Gene</th>
<th>^4P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>84,111,083</td>
<td>SSPN</td>
<td>p = 5.07x10^-6</td>
</tr>
<tr>
<td>5</td>
<td>84,099,507</td>
<td>SSPN</td>
<td>p = 7.99x10^-6</td>
</tr>
<tr>
<td>20</td>
<td>69,808,892</td>
<td>LOC100847985</td>
<td>p = 2.69x10^-5</td>
</tr>
<tr>
<td>3</td>
<td>70,258,600</td>
<td>TYW3, CRYZ</td>
<td>p = 3.03x10^-5</td>
</tr>
<tr>
<td>4</td>
<td>5,407,154</td>
<td>IKZF1</td>
<td>p = 4.18x10^-5</td>
</tr>
</tbody>
</table>

^1Bos taurus chromosome, ^2Location of genes is given by base pair number from the telomere based on the UMD 3.1 assembly (Zimin et al., 2009), ^3Mapped gene refers to individual genes mapped within 20kb of the SNP position serving as the gene proxy, ^4P-values for the association of the gene’s proxy SNP. Values of P < 5.5 x 10^-5 provide evidence for a moderate association and values P < 5 x 10^-7 provide evidence for a strong association of the genes with feed efficiency (Wellcome Trust Case Control Consortium, 2007).
Figure 2: Identification of loci associated with feed efficiency on BTA3, 5, 8, 10, 11, 12, 13, 15, 19, 20, 21, 22, and 23. Results are presented with $-\log_{10}(p\text{-Value})$ on the Y-axis and Bovine Chromosome Number on the X-axis. Blue and red significance lines are based on the Wellcome Trust Consortium Guidelines for Moderate ($P < 5.5 \times 10^{-5}$) and high ($P < 5 \times 10^{-7}$) association, respectively (Wellcome Trust Case Control Consortium, 2007)
Figure 3: Q-Q plot of RFI GWAS distributions before correction for population structure using Genome-wide Rapid Association using Mixed Model and Regression (GRAMMAR) mixed linear models. The observed -log_{10} P-values are plotted on the vertical axis, while the expected -log_{10} P-values are plotted on the horizontal axis. The large deviation of the entire line above the idealized 45 degree line indicates uncorrected population stratification.
Figure 4: Q-Q plot of RFI GWAS distributions following correction for population structure using Genome-wide Rapid Association using Mixed Model and Regression (GRAMMAR) mixed linear models. The observed -$\log_{10}$ P-values are plotted on the vertical axis, while the expected -$\log_{10}$ P-values are plotted on the horizontal axis. After correction, the distribution of observed and expected -$\log_{10}$ P-values follow the idealized 45 degree line except where associations are detected with RFI.
Table 2: Gene set enrichment analysis-SNP results on the three pathways that tended to be associated with feed efficiency.

<table>
<thead>
<tr>
<th>Gene Set</th>
<th>Leading Edge Gene ID</th>
<th>Gene Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism of Proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REACT_17015.1</td>
<td>DNAJB11</td>
<td>Encodes a soluble glycoprotein of the endoplasmic reticulum (ER) lumen that functions as a co-chaperone of binding immunoglobulin protein</td>
</tr>
<tr>
<td></td>
<td>RPL24</td>
<td>Encodes a ribosomal protein that is a component of the 60S subunit</td>
</tr>
<tr>
<td></td>
<td>RPS3</td>
<td>Encodes a ribosomal protein that is a component of the 40S subunit</td>
</tr>
<tr>
<td></td>
<td>PROS1</td>
<td>Encodes a vitamin K-dependent plasma protein that functions as a cofactor for the anticoagulant protease to inhibit blood coagulation</td>
</tr>
<tr>
<td></td>
<td>EIF4A2</td>
<td>Encodes eukaryotic translation initiation factor 4A2</td>
</tr>
<tr>
<td></td>
<td>ST6GAL1</td>
<td>Encodes a member of glycosyltransferase family 29 that catalyzes the transfer of sialic acid</td>
</tr>
<tr>
<td></td>
<td>RPS14</td>
<td>Encodes a ribosomal protein that is a component of the 40S subunit</td>
</tr>
<tr>
<td></td>
<td>TBCE</td>
<td>Encodes cofactor E, one of four proteins involved in the pathway leading to correctly folded beta-tubulin from folding intermediates</td>
</tr>
</tbody>
</table>

1Gene sets from NCBI Gene2Go (http://www.ncbi.nlm.nih.gov/) that were significant by GSEA-SNP, 2Gene set names and descriptions adapted from 3Reactome (http://www.reactome.org/)
Table 2 (continued): Gene set enrichment analysis-SNP results on the three pathways that tended to be associated with feed efficiency.

<table>
<thead>
<tr>
<th>Gene set ID</th>
<th>Pathway</th>
<th>Leading Edge Gene ID</th>
<th>Gene Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00020</td>
<td>FAS Signaling Pathway</td>
<td>FASLG</td>
<td>Encodes a protein that is the ligand for FAS. Interaction of FAS with this ligand is critical in triggering apoptosis of some types of cells such as lymphocytes.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FAF1</td>
<td>Encodes a protein that binds to FAS antigen and can initiate apoptosis or enhance apoptosis initiated through FAS antigen.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAPK10</td>
<td>Encodes a protein that is a member of the MAP kinase family. MAP kinases are involved in a wide variety of cellular processes.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSN</td>
<td>Encodes a protein that binds to the &quot;plus&quot; ends of actin monomers and filaments to prevent monomer exchange.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FADD</td>
<td>Encodes a protein that is an adaptor molecule that interacts with various cell surface receptors and mediates cell apoptotic signals.</td>
</tr>
</tbody>
</table>

\(^{2}\)Gene set names and descriptions adapted from \(^{4}\)Panther (http://www.pantherdb.org/).
Table 2 (continued): Gene set enrichment analysis-SNP results on the three pathways that tended to be associated with feed efficiency.

<table>
<thead>
<tr>
<th><strong>GO:0044706 Multi Multicellular Organism Process</strong></th>
<th><strong>Leading Edge Gene ID</strong></th>
<th><strong>Gene Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STAT5B</strong></td>
<td><strong>PPP1R1B</strong></td>
<td>Encodes a protein that is a member of the STAT family of transcription factors. This protein mediates the signal transduction triggered by various cell ligands</td>
</tr>
<tr>
<td><strong>PPP1R1B</strong></td>
<td><strong>TIMP1</strong></td>
<td>Encodes a bifunctional signal transduction molecule. Receptor stimulation regulates its phosphorylation and function as a kinase or phosphatase inhibitor.</td>
</tr>
<tr>
<td><strong>TIMP1</strong></td>
<td><strong>BCL2</strong></td>
<td>Encodes a protein that belongs to the TIMP gene family. The protein is able to promote cell proliferation and may also have an anti-apoptotic function.</td>
</tr>
<tr>
<td><strong>BCL2</strong></td>
<td><strong>B4GALT1</strong></td>
<td>Encodes an integral outer mitochondrial membrane protein that blocks the apoptotic death of some cells such as lymphocytes.</td>
</tr>
<tr>
<td><strong>B4GALT1</strong></td>
<td><strong>C1QBP</strong></td>
<td>Encodes a type II membrane-bound glycoprotein that participates both in glycoconjugate and lactose biosynthesis.</td>
</tr>
<tr>
<td><strong>C1QBP</strong></td>
<td><strong>ACVR2A</strong></td>
<td>Encodes a protein known to bind the globular heads of C1q molecules and inhibit C1 activation to yield the first component of the serum complement system.</td>
</tr>
<tr>
<td><strong>ACVR2A</strong></td>
<td><strong>VMP1</strong></td>
<td>Encodes a receptor that mediates the functions of activins, which are members of the TGF-beta superfamily involved in diverse biological processes.</td>
</tr>
<tr>
<td><strong>VMP1</strong></td>
<td></td>
<td>Encodes a pancreatitis-associated transmembrane protein whose expression triggers autophagy in several human diseases</td>
</tr>
</tbody>
</table>

*2 Gene set names and descriptions adapted from *5 Gene Ontology (http://www.geneontology.org/).*
Running enrichment score appears on the vertical axis, and rank of each gene within total GWAS results appears in the horizontal axis. The red line charts the enrichment score of the pathway. When genes appear in the pathway, enrichment score increases, and when a gene is encountered that does not appear in the pathway, enrichment score decreases. The dotted line separates leading edge genes from non-leading edge genes, with leading edge genes appearing to the left of the line. Black lines above the horizontal axis depict genes that appear in the pathway for which running enrichment scores are shown above.
Figure 6: Running enrichment scores identifying leading edge genes for P00020, FAS Signaling Pathway

Running enrichment score appears on the vertical axis, and rank of each gene within total GWAS results appears in the horizontal axis. The red line charts the enrichment score of the pathway. When genes appear in the pathway, enrichment score increases, and when a gene is encountered that does not appear in the pathway, enrichment score decreases. The dotted line separates leading edge genes from non-leading edge genes, with leading edge genes appearing to the left of the line. Black lines above the horizontal axis depict genes that appear in the pathway for which running enrichment scores are shown above.
Figure 7: Running enrichment scores identifying leading edge genes for GO:0044706, Multi Multicellular Organism Process.

Running enrichment score appears on the vertical axis, and rank of each gene within total GWAS results appears in the horizontal axis. The red line charts the enrichment score of the pathway. When genes appear in the pathway, enrichment score increases, and when a gene is encountered that does not appear in the pathway, enrichment score decreases. The dotted line separates leading edge genes from non-leading edge genes, with leading edge genes appearing to the left of the line. Black lines above the horizontal axis depict genes that appear in the pathway for which running enrichment scores are shown above.