J. Dairy Sci. 96:6716–6729
http://dx.doi.org/10.3168/jds.2012-6237

Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle

C. Yao,*[1] D. M. Spurlock,† L. E. Armentano,* C. D. Page Jr.,‡ M. J. VandeHaar,§ D. M. Bickhart,# and K. A. Weigel*
*Department of Dairy Science, University of Wisconsin, Madison 53706
†Department of Animal Science, Iowa State University, Ames 50011
‡Department of Biostatics and Medical Informatics, University of Wisconsin, Madison 53706
§Department of Animal Science, Michigan State University, East Lansing 48824
#Animal Improvement Program Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705

ABSTRACT

Feed efficiency is an economically important trait in the beef and dairy cattle industries. Residual feed intake (RFI) is a measure of partial efficiency that is independent of production level per unit of body weight. The objective of this study was to identify significant associations between single nucleotide polymorphism (SNP) markers and RFI in dairy cattle using the Random Forests (RF) algorithm. Genomic data included 42,275 SNP genotypes for 395 Holstein cows, whereas phenotypic measurements were daily RFI from 50 to 150 d postpartum. Residual feed intake was defined as the difference between an animal's feed intake and the average intake of its cohort, after adjustment for year and season of calving, year and season of measurement, age at calving nested within parity, days in milk, milk yield, body weight, and body weight change. Random Forests is a widely used machine-learning algorithm that has been applied to classification and regression problems. By analyzing the tree structures produced within RF, the 25 most frequent pairwise SNP interactions were reported as possible epistatic interactions. The importance scores that are generated by RF take into account both main effects of variables and interactions between variables, and the most negative value of all importance scores can be used as the cutoff level for declaring SNP effects as significant. Ranking by importance scores, 188 SNP surpassed the threshold, among which 38 SNP were mapped to RFI quantitative trait loci (QTL) regions reported in a previous study in beef cattle, and 2 SNP were also detected by a genome-wide association study in beef cattle. The ratio of number of SNP located in RFI QTL to the total number of SNP in the top 188 SNP chosen by RF was significantly higher than in all 42,275 whole-genome markers. Pathway analysis indicated that many of the top 188 SNP are in genomic regions that contain annotated genes with biological functions that may influence RFI. Frequently occurring ancestor-descendant SNP pairs can be explored as possible epistatic effects for further study. The importance scores generated by RF can be used effectively to identify large additive or epistatic SNP and informative QTL. The consistency in results of our study and previous studies in beef cattle indicates that the genetic architecture of RFI in dairy cattle might be similar to that of beef cattle.
Key words: Random Forest, single nucleotide polymorphism, residual feed intake, dairy cattle

INTRODUCTION

Feed efficiency is an economically important trait in the beef and dairy cattle industries. Typically, feed costs account for 60 to 65% of the total production costs in a beef cattle operation (Sainz and Paulino, 2004) and up to 50% of the total production costs in a dairy cattle operation (VandeHaar and St-Pierre, 2006). Because of ongoing genetic selection for productivity and improvements in herd management, the efficiency of converting feed to milk in US dairy cattle has doubled over the past 60 yr due to dilution of maintenance (VandeHaar and St-Pierre, 2006). Further improvements in feed efficiency are essential, not only to enhance profitability of the dairy industry, but also to feed the growing global human population sustainably. Because of the biological complexity of feed efficiency, methods to evaluate feed efficiency that are independent of the dilution of maintenance are needed.

Residual feed intake (**RFI**), first proposed by Koch et al. (1963) in beef cattle, is the difference between actual and predicted intake for an animal. Predicted intake can be computed from nutritional models based on dietary energy content, or it can be determined statistically by the deviation of an animal's intake from the average intake of its cohort, after adjustment for

Received October 5, 2012.
Accepted June 20, 2013.
[1]Corresponding author: cyao5@wisc.edu

production and known environmental differences. Thus, RFI is a measure of partial efficiency that is independent of production level per unit of BW, unlike gross efficiency as measured by feed conversion ratio, which is the ratio of feed intake to rate of gain (beef cattle) or milk production (dairy cattle) of an individual animal. Genetically, reported heritability estimates of RFI are moderate, ranging from 0.18 to 0.39 (Arthur et al., 2001; Robinson and Oddy, 2004; Schenkel et al., 2004). Estimated breeding values for RFI in beef cattle in Australia were developed by Exton et al. (1999). On a molecular basis, 19 QTL for RFI in beef cattle were identified as significant at the chromosome-wide level by Sherman et al. (2009), and several SNP markers were identified by genome-wide association studies (**GWAS**) in beef cattle (Bolormaa et al., 2011; Snelling et al., 2011; Rolf et al., 2012). In dairy cattle, however, studies of RFI and reports of QTL for RFI are limited, especially in lactating cows.

Adding genomic information from SNP markers has enhanced traditional genetic evaluation of North American dairy cattle by providing higher reliabilities for young selection candidates (VanRaden et al., 2009). To improve the feed efficiency of dairy cattle, genomic evaluation is essential, because routine measurement of individual animal feed intakes in a traditional progeny-testing program would be cost prohibitive. The majority of published genomic studies in dairy cattle have used EBV of bulls, which reflect the sum of additive genetic effects as the dependent variable. However, direct utilization of RFI phenotypes of individual cows as the dependent variable in a genomic analysis allows the study of both additive effects of individual SNP and epistatic interactions between pairs or sets of SNP. Studying epistatic interactions between SNP using conventional statistical methods such as Bayesian regression models is computationally infeasible due to the massive number of potential 2-way and 3-way interactions between SNP.

The Random Forests (**RF**) algorithm (Breiman, 2001) is a machine-learning method that has been widely applied to classification and regression problems, and is particularly well suited for situations in which the number of potential explanatory variables vastly exceeds the number of observations. According to Liaw and Wiener (2002), the importance scores for potential explanatory variables that are generated by RF algorithms take into account both the main effects of these variables and interactions between variables. In a genomic analysis, these represent additive effects of SNP and epistatic interactions between SNP, respectively. As a screening tool for identifying risk-associated SNP, Lunetta et al. (2004) tested the performance of RF to identify risk-associated SNP using a simulated complex disease model. The importance score significantly outperformed the univariate Fisher exact test *P*-value, if risk SNP interacted. The RF algorithm was also implemented to successfully identify epistatic interactions associated with complex traits in humans, using importance scores in both simulation and real data studies (Chen et al., 2007; Jiang et al., 2009). We also applied the Bayesian least absolute selection and shrinkage operator (**LASSO**) of Park and Casella (2008) to estimate the additive effects of individual SNP. This method is effective for genomic selection of dairy cattle as an additive effects model (de los Campos et al., 2009; Weigel et al., 2009; Vazquez et al., 2010), although it is not yet commonly used and needs to be further studied for GWAS. Due to the advantage of simultaneously analyzing multiple SNP, compared with single SNP-based GWAS, Bayesian LASSO has been recently implemented for GWAS and has successfully detected several significant QTL and genes associated with quantitative traits using estimated genetic effects (Yi and Xu, 2008; Li et al., 2011).

In this study, we estimated SNP effects with the importance scores from the RF algorithm, considering both additive effects of individual SNP and epistatic interactions between SNP, as well as the absolute values of estimated genetic effects from Bayesian LASSO, considering only additive effects of individual SNP. By analyzing the structure of trees produced within the RF, the most frequent pairs of ancestor and descendant SNP (i.e., pairs of SNP occurring within the same branch in many trees) are identified as possible epistatic interactions. To our knowledge, this is the first application of RF for detection of potentially epistatic QTL in food animal species.

MATERIALS AND METHODS

Phenotypic Values

The experimental population and the methods for phenotypic data collection were described in detail by Spurlock et al. (2012). Phenotypic records of 402 Holstein cows at the Iowa State University Dairy from 50 to 150 DIM were available, including daily milk yield, weekly protein percentage and lactose percentage, as analyzed by Dairy Lab Services Inc. (Dubuque, IA), monthly milk fat percentage from DHIA records, weekly BW, and daily DMI.

Component percentages were set to missing values if less than 1% or greater than 10% (fat percentage and protein percentage) or less than 0% or greater than 10% (lactose percentage), and less than 1% of total percentage observations were removed. To impute daily records from weekly or monthly data points while also smooth-

ing variation in measurements of BW, fat percentage, protein percentage, and lactose percentage, local linear regression (Wand and Jones, 1995) was used and the bandwidth was selected with the direct plug-in methodology described by Ruppert et al. (1995). If fewer than 10 data points were available for fat percentage, protein percentage, and lactose percentage, the daily values were imputed from the average of available records. Net energy required for lactation (only including energy in milk) was calculated based on the gross energy per kilogram for fat, protein, and lactose according to NRC (2001) using the following formula:

$$\mathrm{NE_L\ (Mcal)} = (0.0929 \times \text{fat \%} + 0.0563 \times \text{protein \%} + 0.0395 \times \text{lactose \%}) \times \text{daily milk yield (kg)}.$$

To calculate RFI, DMI was adjusted using the following mixed linear model:

$$y_{ijklm} = \mu + YS_i + CalvCat_j + YSR_k + Cow_l + \beta_1 DIM_{ijklm} + \beta_2 NE_{Lijklm} + \beta_3 BW_{ijklm} + \beta_4 \Delta BW_{ijklm} + e_{ijklm},$$

where y_{ijklm} is the DMI observation for an individual animal; μ is the overall mean; YS_i ($i = 1, 2, \ldots, 9$) is a categorical fixed effect of year and season of calving; $CalvCat_j$ ($j = 1, 2, \ldots, 14$) is a categorical fixed effect of parity and age at calving; YSR_k ($k = 1, 2, \ldots, 9$) is a categorical fixed effect of year and season of recording; $Cow_l \sim N\left(0, \mathbf{I}\sigma^2_{cow}\right)$ is the random animal effect, where $\mathbf{I}$ is an identity matrix and σ^2_{cow} is the variance of animal effect; DIM_{ijklm} is a continuous fixed effect of DIM, with regression coefficient β_1; NE_{Lijklm} is a continuous fixed effect of NE_L (Mcal), with regression coefficient β_2; BW_{ijklm} is a continuous fixed effect of BW (kg), with regression coefficient β_3; ΔBW_{ijklm} is a continuous fixed effect of BW change (kg/d), with regression coefficient β_4; and $\mathbf{e}_{ijklm} \sim N\left(0, \mathbf{I}\sigma^2_e\right)$ is the random error, where σ^2_e is the error variance. The levels of categorical fixed effects were combined with adjacent levels if the number of records within a level was less than 1,000. The linear mixed model was implemented using lme4 R package version 0.999999-0 (http://lme4.r-forge.r-project.org/). After fitting the model shown above, the estimate of the random animal effect was considered as the RFI phenotype for that cow in the subsequent genomic analysis.

Genetic Markers

The Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA) was used for genotyping. Cows with more than 5% missing genotypes were excluded, and SNP with more than 1% missing values or minor allele frequencies less than 5% were removed. Missing genotypes were imputed without pedigree using BEAGLE 3.3.2 with default options (Browning and Browning, 2009). After quality control editing, 42,275 SNP in 395 cows were available for further analysis. The SNP genotypes at each locus were coded as 0, 1, or 2, according to the number of copies of the minor allele.

Description of the RF Algorithm

General Description. The RF algorithm, which was introduced by Breiman (2001), produces an ensemble of tree predictors (for regression or classification), where each tree is grown from a different bootstrap sample. Furthermore, at each split point (node) in each tree, a different subset of m_{try} features are evaluated to choose the best feature for splitting, where m_{try} is an input parameter to the RF algorithm. In the current study, regression on SNP genotypes was implemented using the randomForest package in R (Liaw, and Wiener, 2011).

The performance of each tree in the RF can be evaluated using the mean squared error (**MSE**) of the "out-of-bag" (**OOB**) data, which are the data points not included in the bootstrap sample. The formula for MSE_{OOB} in the RF is given as follows:

$$\mathrm{MSE_{OOB}} = n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2,$$

where n is the number of animals, $\hat{y}_i$ is the average of the OOB predictions for ith animal, and the OOB predictions are from trees in which the animal was OOB.

To complete the regression analysis, let $n = 395$ be the number of animals that passed genotype quality control, $p = 42{,}275$ be the number of SNP, $\mathbf{y}_{(n \times 1)}$ be the phenotype vector of RFI, and $\mathbf{X} = \{\mathbf{x}_i\}$ be the marker matrix, where $\mathbf{x}_i$ is a $(1 \times p)$ vector containing the p SNP of an individual animal. Draw n_{tree} bootstrap samples from the whole data set to grow the same number of trees. To build a node within a tree, do not choose the best split from all p of the original SNP, but rather search within a random sample of m_{try} SNP to find the best split SNP. The fitted value is the average prediction over n_{tree} trees (Liaw and Wiener, 2002).

Pairwise Epistatic Interactions Between SNP. The tree structures generated by RF are informative for identifying interactions between potential explanatory variables (in this case, epistatic effects of pairs of SNP). An example is illustrated in Figure 1. Assume that SNP C and E jointly have a large epistatic effect on RFI.

The combination of SNP C and E will appear more frequently in the same branch of a tree than in other branches or trees, which will form a parent-descendant (child, grandchild, and so on) pair hereinafter referred to as a descendant pair. In this example, adding SNP E improves prediction accuracy conditionally on the split produced by its ancestor SNP C, which appears at a higher level of the same branch. If 2 SNP have large but independent main effects on the response variable, such as SNP B and C, they will also appear frequently within the same tree, but not necessarily as descendant pairs within the same branch. Thus, the descendant pairs that occur most frequently in the RF can be recognized as possible pairwise epistatic interactions.

SNP Importance Scores. The importance score (ΔMSE%) for the mth SNP is defined as the average percentage increase in MSE when generating a prediction of the OOB data if the value of the mth SNP is randomly permuted, whereas genotypes for all other SNP remain unchanged (Breiman, 2001). Both additive effects and interactions with other SNP will contribute to increasing the importance scores of SNP. In other words, SNP with large positive importance scores are those for which random permutation of the SNP genotype will increase prediction error, and this increase in MSE will reflect both the additive effect of the SNP and its epistatic interactions with all other SNP. For SNP that are not associated with the response variable, the importance score will be approximately centered at 0. However, due to random sampling the actual importance scores will be slightly positive or slightly negative with equal probability.

Therefore, one way to differentiate SNP with real effects on the phenotype from SNP with spurious effects is to use the absolute value of the most negative importance score as the cutoff level for declaring SNP effects as significant—any SNP with importance scores higher than this threshold can be considered as transmitting a real signal.

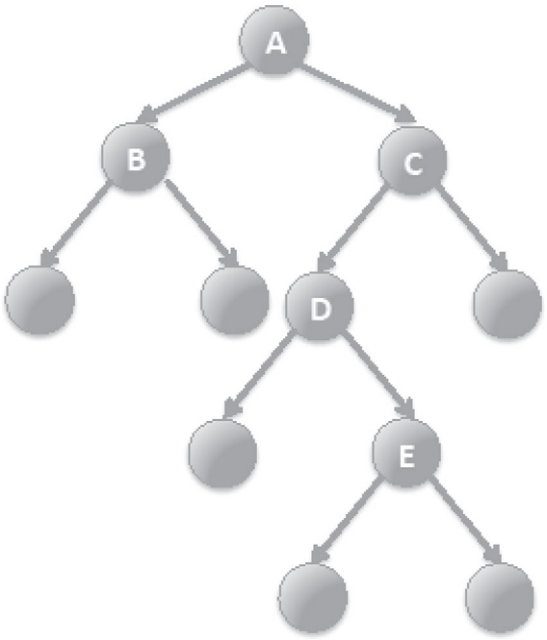

Figure 1. Example tree generated by the Random Forests algorithm. Single nucleotide polymorphism pairs A and B, A and C, A and D, A, and E, C and D, C and E, and D and E represent descendant pairs, which may indicate epistatic genetic effects, whereas SNP pairs B and C, B and D, and B and E represent nondescendant pairs, which may indicate independent additive genetic effects.

Implementation of the RF Analysis

Pairwise Interactions Between SNP. To reduce the computational burden, a 2-step implementation of the RF algorithm was used, in which preselection of SNP was used to remove the majority of SNP that did not have strong additive or epistatic effects. Two parameters, the number of trees grown within each random forest (n_{tree}) and the number of SNP chosen per random bootstrap sample at each node of the trees (m_{try}), were tuned via 5-fold cross-validation with a greedy search algorithm. The parameters chosen in this tuning step were $n_{tree} = 700$ and $m_{try} = 1{,}000$, and $n_{tree} = 700$ and $m_{try} = 100$ for steps 1 and step 2, respectively.

The procedure for SNP selection in step 1 is illustrated in Figure 2. Initially, a total of 1,000 RF were divided into 10 groups with 100 forests apiece. Within each group, and for each forest, the top n_1 SNP, as ranked by importance score, were selected from the total set of 42,275 SNP. Assume that the probability of a single SNP being selected within 1 forest represents a Bernoulli random variable with $p_1 = n_1/42{,}275$. For all 100 forests within a group, the number of times each SNP is selected follows a Binomial distribution, with the number of trials equal to 100 and success probability of p_1 (i.e., probability that a given SNP was included in the top n_1) in each trial. A probability of 0.05 was used as the threshold to obtain a significance level of 0.05, and the n_2 SNP that appeared at least j times with binomial probability (j, p_1) of less than 0.05 were chosen for that group. Repeating this procedure produces 10 groups that, in turn, follow a Binomial distribution, with the number of trails equal to 10, each with probability of success p. The SNP that were represented in more than k groups had binomial probability (k, p) less than 0.05, and these were selected for the analysis of pairwise interactions in step 2. To stabilize the results in step 2, a total of 100 RF were generated for the analysis, and the number of descendant pairs over all trees was counted.

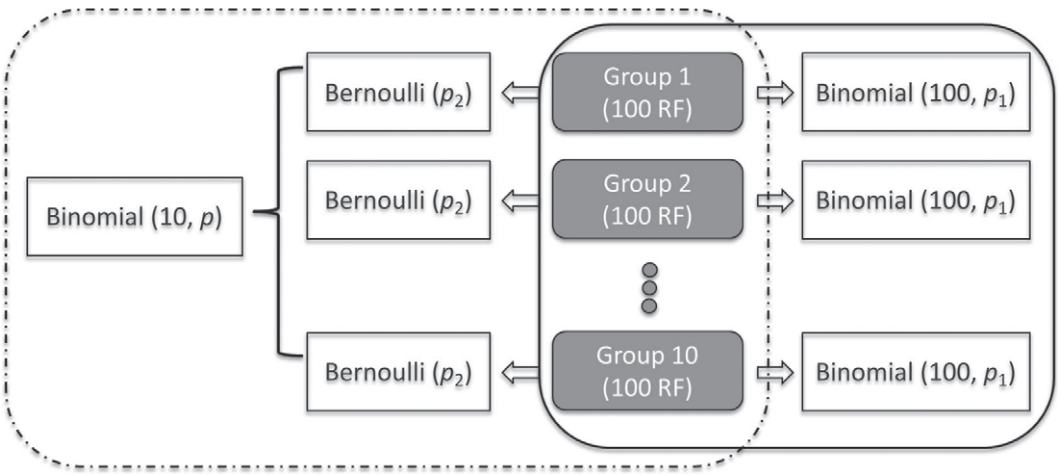

Figure 2. Illustration of the dimension-reduction process (step 1) for selecting SNP to include in the subsequent Random Forests (RF) analysis of pairwise epistatic effects. $p_{(x)}$ = probability that a given SNP was included in the top $n_{(x)}$ SNP.

Important Individual SNP. All 42,275 SNP were ranked according to their importance scores generated by the RF. To improve the stability of the results, the average importance scores from 1,000 RF were used as the final importance scores for individual SNP. The absolute value of the most negative importance score was used as the threshold, and SNP with importance scores greater than this threshold were reported as those with significant effects.

Reference Analysis with Bayesian LASSO

For reference purposes, the additive effects of individual SNP were estimated by regressing RFI phenotypes on SNP covariates using the Bayesian LASSO of Park and Casella (2008). The hierarchical model for Bayesian LASSO is defined as follows:

$$\text{Likelihood: } p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma_\varepsilon^2) = \prod_{i=1}^{n} N(y_i \mid \mathbf{x}_i'\boldsymbol{\beta}, \sigma_\varepsilon^2);$$

$$\text{Prior: } p(\boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{\tau}^2, \lambda^2) = p(\boldsymbol{\beta} \mid \sigma_\varepsilon^2, \boldsymbol{\tau}^2) p(\sigma_\varepsilon^2) p(\boldsymbol{\tau}^2 \mid \lambda) p(\lambda^2)$$
$$= \left[\prod_{j=1}^{p} N(\beta_j \mid 0, \tau_j^2 \sigma_\varepsilon^2)\right] \chi^{-2}(\sigma_\varepsilon^2 \mid \mathrm{df}, S) \times \left[\prod_{j=1}^{p} \mathrm{Exp}(\tau_j^2 \mid \lambda)\right] G(\lambda^2 \mid \alpha_1, \alpha_2),$$

where $N(y_i \mid \mathbf{x}_i'\boldsymbol{\beta}, \sigma_\varepsilon^2)$ and $N(\beta_j \mid 0, \tau_j^2\sigma_\varepsilon^2)$ are normal densities centered at $\mathbf{x}_i'\boldsymbol{\beta}$ and 0, with variances σ_ε^2 and $\tau_j^2\sigma_\varepsilon^2$, respectively, where $\boldsymbol{\beta}$ is the regression of y on SNP covariates and σ_ε^2 is the variance of model residual; $\chi^{-2}(\sigma_\varepsilon^2 \mid \mathrm{df}, S)$ is a scaled-inverted chi-squared density, with degrees of freedom parameter df and scale parameter S; $\mathrm{Exp}(\tau_j^2 \mid \lambda)$ is an exponential distribution assigned to a positive scale parameter τ_j^2, indexed by the prior distribution of the regularization parameter λ; $G(\lambda^2|\alpha_1,\alpha_2)$ is a gamma distribution, with shape and rate parameters α_1 and α_2, respectively. The analysis was performed using the BLR package in R (de los Campos and Perez Rodriguez, 2012).

Pathway Analysis

A pathway analysis was performed to identify biological pathways that may influence RFI. The genomic coordinates of RefSeq genes and Ensembl transcript predictions were downloaded from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/downloads.html#cow). The closest annotated RefSeq genes were identified for each of the top SNP with the highest importance scores using the BEDTools software package (Quinlan and Hall, 2010). Only genes overlapping or within 37,000 bp of each respective SNP marker were chosen for further analysis. The 37,000-bp cutoff was based on the median distance between SNP on the BovineSNP50 BeadChip (Matukumalli et al., 2009). Cattle genome annotation is still ongoing (Childers et al., 2011) and is highly likely to be incomplete. Therefore, SNP with distantly associated genes (i.e., SNP greater than 37,000 bp away from the closest gene) may be linked to cryptic functional genetic elements that have not yet been discovered instead of the closest annotated gene. Out of the current 26,740 annotated Ensembl transcripts, 21,968 transcripts (~82%) were within 37,000 bp of a BovineSNP50 SNP. For RefSeq annotations, 12,137 RefSeq genes (~86%) were within the 37,000 cutoff out of 14,176 total RefSeq genes. Gene enrichment analysis and gene functional analysis were performed using the DAVID web tool (Huang et al., 2009a,b) and pathway analysis was performed using

the PANTHER database (Thomas et al., 2003; Mi et al., 2005). Biochemical pathways were visualized using PathVisio (van Iersel et al., 2008). All annotated *Bos taurus* genes in each respective database were used as background for both the DAVID and PANTHER analyses, as both web tools cannot use gene lists containing over 3,000 entries for background.

RESULTS AND DISCUSSION

The RFI estimates ranged from −4.74 to 3.70 kg of DM/d, with a mean of 0 kg of DM/d and standard deviation of 1.34 kg of DM/d. The correlation between fitted values from the linear mixed model and real observations of DMI was 0.84 as the goodness of fitness for the model. For the RF analysis, the running mean MSE_{OOB} values over different number of RF replications were plotted in Figure 3. As the number of RF replications increased, the variation of mean MSE_{OOB} decreased, which means that the results over 1,000 RF replications in step 1 and 100 RF replications in step 2 were stable.

In the SNP selection step, a total of 5,800 SNP were chosen for the second step, which involved the analysis of pairwise epistatic effects. To select SNP, 4,000 (n_1) SNP were first chosen in each RF, such that $p_1 = 4{,}000/42{,}275 \approx 0.095$. Within each group of 100 RF, the SNP that appeared at least in 15 RF were selected. Based on the frequency distribution SNP that were selected in 1, 2, 3, . . . , 10 groups of 100 RF (Figure 4), it is clear that SNP were not selected randomly. In the absence of real effects of SNP on the phenotype, one would expect a monotonically decreasing pattern in the number of times individual SNP were represented as the number of groups increased. On the contrary, a relatively large number of SNP were selected within all 10 groups, which indicates that they had strong associations with the RFI phenotypes. In total, 5,800 SNP that appeared in at least 5 of the 10 groups were chosen for the second step. For ranking SNP based on individual importance, the threshold value was set to 0.015, which was the most negative value of all importance scores. A total of 188 SNP surpassed this threshold. The distribution of the importance scores for all 42,275 SNP across the 29 autosomes and the X chromosome is given in Figure 5. The closest annotated RefSeq genes identified for each of the 188 SNP are listed in Supplementary Table S1 (available online at http://dx.doi.org/10.3168/jds.2013-6237).

The number of descendant pairs was counted using 100 RF from step 2 with 5,800 SNP, as described above, with 700 trees in each forest. As shown in Table 1, the first 2 pairs included SNP ARS-USMARC-Parent-DQ888312-rs29015945 on *Bos taurus* autosome (**BTA**) 19, which was represented 7 times within the 25 most-frequent descendant pairs and had the high-

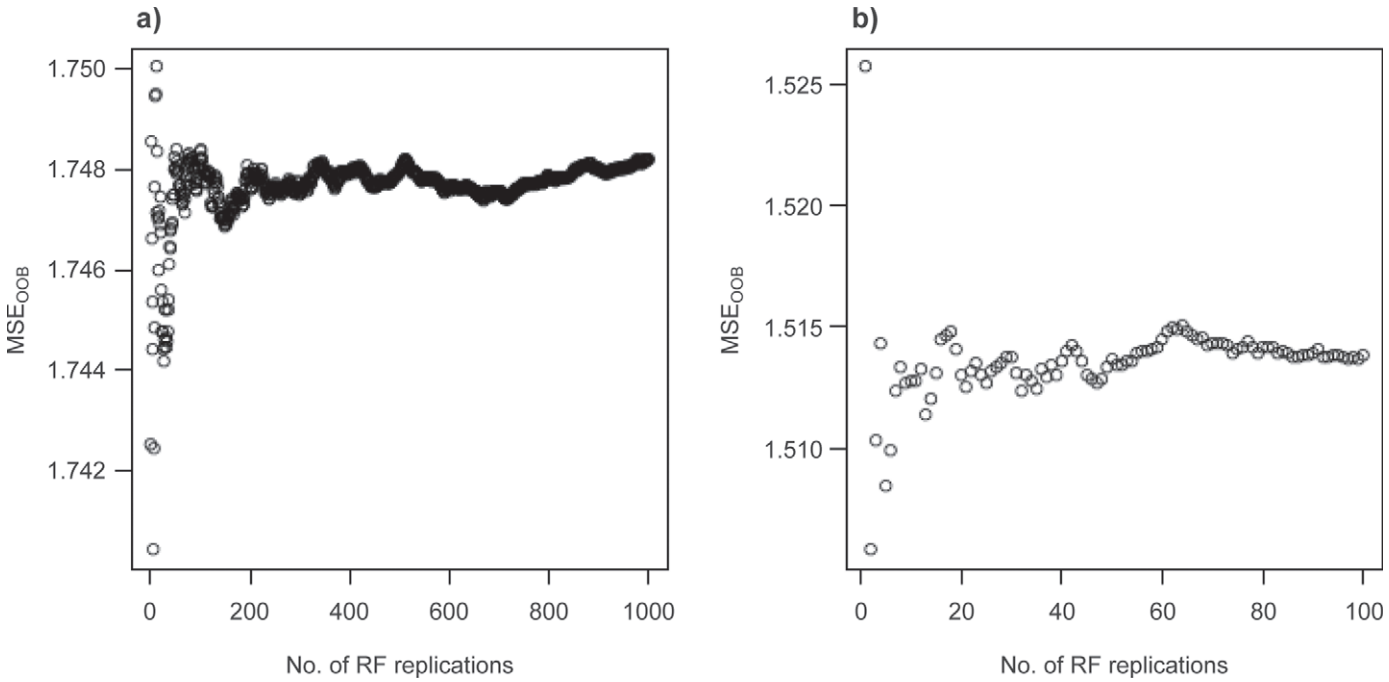

Figure 3. The running means of mean squared error in out-of-bag data (MSE_{OOB}) over different number of Random Forests (RF) replications in (a) step 1 (over 1,000 RF replications) and (b) step 2 (over 100 RF replications). As the number of RF replications increases, the variation of mean MSE_{OOB} decreases.

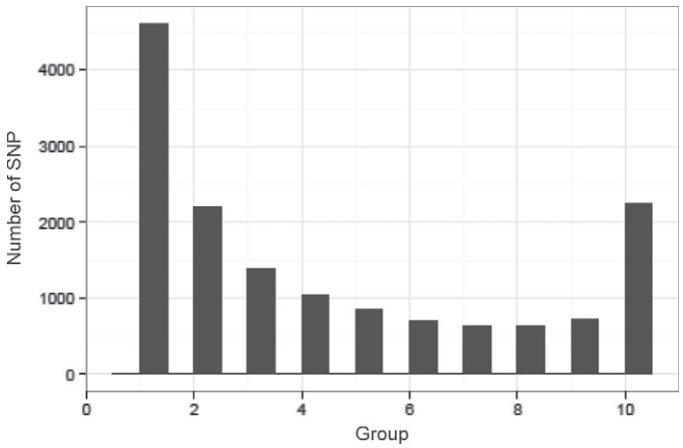

Figure 4. Frequency distribution of representation of individual SNP among the 10 groups created in the dimension-reduction process (step 1) for selecting SNP to include in the subsequent Random Forests analysis of pairwise epistatic effects.

est RF importance score of 0.329 (as shown in Table 2). The SNP ARS-BFGL-NGS-25743 on BTA11 was represented in 16 of the 25 most-frequent descendant pairs and had the second-highest RF importance score (0.286). The SNP ARS-BFGL-NGS-106241 on BTA7 had the third-highest importance score (0.120), and SNP ARS-BFGL-BAC-2599 on BTA22 had the fourth-highest importance score (0.109). These last 2 SNP were represented in the remaining 3 pairs of the top 25 most-frequent descendant pairs, which did not contain SNP with the first- and second-highest importance scores. Thus, SNP that had higher importance scores in the RF analysis were involved in a large number of pairwise interactions.

The significances of the top 25 descendant pairs as possible pairwise epistatic interactions were validated via linear regression. Twenty-one unique SNP were presented in the top 25 descendant pairs. Therefore, the RFI phenotypes of 395 cows were regressed on 21 SNP and 25 pairwise interactions as fixed effects simultaneously (replication 0). Individual SNP were fitted using their genotypes as covariates. Interactions were renumbered as integers 1 to 9, if combinations of genotypes were 11, 00, 01, 02, 10, 12, 20, 21, and 22. Interactions were also fitted as covariates after renumbering due to the limited number of animals. The correlation between fitted RFI and actual RFI was 0.66, and 8 (out of 25) interaction terms had *P*-values below 0.1, which means that 32% of interactions had significance levels below 0.1. To test whether descendant pairs resulted from large main effects of individual SNP instead of interactions, 5 regression models (replications a1 to a5) were fitted using the same 21 SNP as in replication 0, along with 25 interactions that represented randomly pairs of the 21 SNP rather than descendant pairs. Another 5 regression models (replications b1 to b5) were fitted using 25 interactions of pairs of SNP that were randomly sampled from all 5,800 SNPs brought into the second step along with their corresponding main effects. The mean correlations between fitted and actual RFI for replication a1 to a5 and b1 to b5 were 0.62 and 0.61, respectively, and the average numbers of interaction terms with *P*-values below 0.1 were 3 and 2. Interactions in replication a1 to a5 involved the same SNP as in replication 0 but different pairing patterns, which resulted in depressed correlation between fitted and actual RFI and fewer numbers of interactions with *P*-values less than 0.1 than in replication 0. This

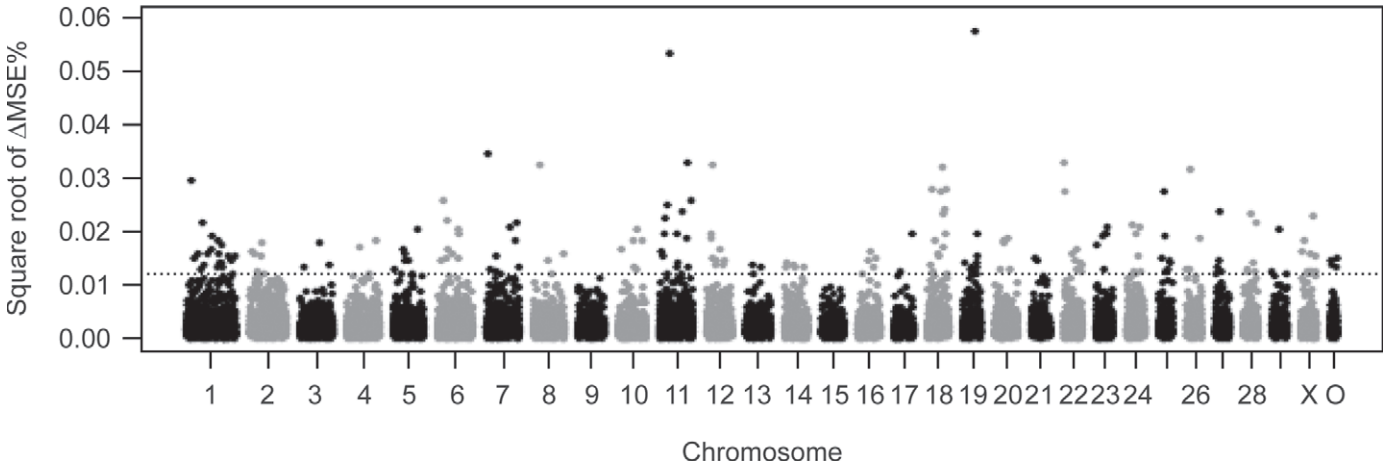

Figure 5. Manhattan plot of the square root of importance scores (ΔMSE%, where MSE = mean squared error) for SNP in the Random Forests analysis. Single nucleotide polymorphisms with negative importance scores are excluded and the dashed line corresponds to the threshold value for SNP selection, which was equal to the absolute value of the most negative importance score.

Table 1. The 25 most frequently represented descendant pairs of SNP in the Random Forests (RF) analysis, which represent pairs of SNP with potentially important pairwise epistatic interactions associated with residual feed intake[1]

			Individual rank by RF	
Pair rank	SNP 1	SNP 2	SNP 1	SNP 2
1	ARS-BFGL-NGS-25743	ARS-USMARC-Parent-DQ888312-rs29015945	2	1
2	ARS-BFGL-NGS-106241	ARS-BFGL-NGS-25743	3	2
3	ARS-BFGL-NGS-25743	ARS-BFGL-NGS-30459	2	17
4	ARS-BFGL-NGS-65789	ARS-USMARC-Parent-DQ888312-rs29015945	35	1
5	ARS-USMARC-Parent-DQ888312-rs29015945	BTA-43831-no-rs	1	8
6	ARS-BFGL-BAC-2599	ARS-BFGL-NGS-25743	4	2
7	ARS-BFGL-NGS-108391	ARS-BFGL-NGS-25743	10	2
8	ARS-BFGL-BAC-2599	ARS-BFGL-NGS-30448	4	6
9	ARS-BFGL-NGS-25743	BTA-56614-no-rs	2	33
10	ARS-BFGL-NGS-106241	ARS-BFGL-NGS-30448	3	6
11	ARS-BFGL-NGS-20025	ARS-BFGL-NGS-25743	9	2
12	ARS-BFGL-NGS-25743	BTA-19869-no-rs	2	2,509
13	ARS-BFGL-NGS-106241	Hapmap43850-BTA-43155	3	12
14	ARS-BFGL-NGS-25743	ARS-BFGL-NGS-30448	2	6
15	ARS-BFGL-NGS-25743	ARS-BFGL-NGS-84065	2	45
16	ARS-USMARC-Parent-DQ888312-rs29015945	BTA-07453-no-rs	1	128
17	ARS-BFGL-NGS-25743	Hapmap59539-rs29025538	2	17,875
18	ARS-BFGL-NGS-113918	ARS-BFGL-NGS-25743	105	2
19	ARS-BFGL-NGS-25743	BTA-29995-no-rs	2	74
20	ARS-BFGL-NGS-25743	Hapmap43850-BTA-43155	2	12
21	ARS-BFGL-NGS-100251	ARS-BFGL-NGS-25743	157	2
22	ARS-BFGL-NGS-25743	Hapmap43850-BTA-43155	2	12
23	ARS-BFGL-BAC-2599	ARS-USMARC-Parent-DQ888312-rs29015945	4	1
24	ARS-BFGL-NGS-35632	ARS-USMARC-Parent-DQ888312-rs29015945	21	1
25	ARS-USMARC-Parent-DQ888312-rs29015945	BTA-29995-no-rs	1	74

[1]Names and importance score rankings of individual SNP involved in each descendant pair are provided.

indicates that significances of descendant pairs were not caused by main effects (additivity or dominance) of the involved SNP, but rather by their interactions. Results from replication b1 to b5 showed that descendant pairs were also more significant than interactions formed by randomly generated SNP. Hence, descendant pairs from RF helped to choose informative pairwise epistatic interactions.

Table 2 shows the chromosomal locations and importance scores of the highest-ranking SNP from the RF analysis, based on importance scores, along with their ranking based on absolute estimated additive genetic effects from Bayesian LASSO analysis. It is clear that most of the 188 SNP that were associated with RFI in the RF analysis had much lower rankings in the Bayesian LASSO analysis. For example, SNP ARS-USMARC-Parent-DQ888312-rs29015945 ranked first in the RF analysis but only 890th in the Bayesian LASSO analysis, whereas SNP ARS-BFGL-NGS-106241 ranked third in the RF analysis and 99th in the Bayesian LASSO analysis. In other words, these SNP had relatively small additive genetic effects, but they contributed considerably to the RFI phenotype through interactions with other SNP. On the other hand, SNP ARS-BFGL-NGS-25743 ranked second in both the RF analysis and the Bayesian LASSO analysis, which means that it not only had a large additive effect, but also potential critical epistatic effects in conjunction with other SNP.

Mapping to The Bovine Genome Database with Bovine UMD3.1 Assembly Chromosome Genome Browser (http://bovinegenome.org/cgi-bin/gbrowse/bovine_UMD31/), 38 of the 188 SNP that surpassed the RF threshold were located in QTL regions for RFI in beef cattle (had been transformed into the physical position in the database), as reported in a previous study by Sherman et al. (2009). We considered the ratio (R_1) of 38 SNP that were located in RFI QTL to 188 SNP that surpassed the RF threshold as $R_1 = 38/188 = 0.202$, which means that 20.2% of significant SNP in our study were located within QTL regions reported by Sherman et al. (2009). Within all 42,275 SNP located across the whole genome, 5,813 SNP were located in RFI QTL regions when mapped to the UMD3.1 assembly. The second ratio (R_2) was calculated as $R_2 = 5{,}813/42{,}275 = 0.138$, which means that about 13.8% of whole-genome markers were covered by RFI QTL regions reported by Sherman et al. (2009). Besides RFI, previously reported QTL for 166 other traits in dairy and beef cattle were mapped by our top 188 SNP in the UMD3.1 assembly. Similar R_1 and R_2 statistics were then calculated for each of the 166 traits. The density of the difference, $(R_1 - R_2)$, for 167 traits (including

Table 2. Chromosomal locations, nearest known genes mapped in the UMD3.1 assembly (http://bovinegenome.org/cgi-bin/gbrowse/bovine_UMD31/), importance scores (ΔMSE%, where MSE = mean squared error), and rankings in the Random Forests (RF) and Bayesian LASSO (BL) analyses for the 25 SNP with highest importance scores in the RF analysis, which represent individual SNP with potentially important additive and pairwise epistatic associations with residual feed intake

RF rank	BL rank	SNP name	ΔMSE%	Chromosome	Gene[1]
1	890	ARS-USMARC-Parent-DQ888312-rs29015945	0.329	19	
2	2	ARS-BFGL-NGS-25743	0.286	11	
3	99	ARS-BFGL-NGS-106241	0.121	7	*SPOCK1*
4	28	ARS-BFGL-BAC-2599	0.109	22	
5	1,059	Hapmap39438-BTA-81105	0.109	11	*TBC1D8*
6	40	ARS-BFGL-NGS-30448	0.106	8	
7	282	ARS-BFGL-NGS-24482	0.106	12	
8	48	BTA-43831-no-rs	0.103	18	*LOC785907*
9	30	ARS-BFGL-NGS-20025	0.100	26	
10	307	ARS-BFGL-NGS-108391	0.088	1	*RRP1B*
11	169	ARS-BFGL-NGS-12047	0.078	18	*LOC510844*
12	45	Hapmap43850-BTA-43155	0.078	18	
13	32	ARS-BFGL-NGS-2693	0.076	25	*LOC515570*
14	2,272	BTA-12313-rs29024268	0.076	18	
15	3,698	ARS-BFGL-NGS-100820	0.075	22	*ATXN7*
16	25	Hapmap53281-rs29026129	0.068	11	*AFF3*
17	722	ARS-BFGL-NGS-30459	0.067	6	*LOC100299906*
18	1,387	ARS-BFGL-NGS-12715	0.063	11	*LMAN2L*
19	63	Hapmap30625-BTA-43445	0.058	18	*LSM14A*
20	487	BTB-00471723	0.058	11	
21	530	ARS-BFGL-NGS-35632	0.057	27	
22	31	BTB-00718231	0.055	18	*LSM14A*
23	1,839	ARS-BFGL-NGS-91390	0.055	28	*RASSF4*
24	132	Hapmap43378-BTA-85949	0.053	X	
25	2,317	ARS-BFGL-NGS-116293	0.051	11	

[1]*SPOCK1* = sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1; *TBC1D8* = TBC1 domain family, member 8 (with GRAM domain); *RRP1B* = ribosomal RNA processing 1 homolog B (*Saccharomyces cerevisiae*); *ATXN7* = spinocerebellar ataxia type 7 protein; *AFF3* = AF4/FMR2 family, member 3; *LMAN2L* = lectin, mannose-binding 2-like; *LSM14A* = LSM14A, SCD6 homolog A (*Saccharomyces cerevisiae*); *RASSF4* = Ras association (RalGDS/AF-6) domain family member 4.

RFI) was plotted in Figure 6, and it approximated a normal distribution $N(-0.002, 0.028^2)$. For RFI, the difference ($R_1 - R_2$) was 0.064, and it exceeded the 0.99th percentile, which corresponds to a significance level of 0.01 for a one-tailed test. This indicates that the SNP with largest importance score in our RF analysis were overrepresented in QTL regions for RFI in beef cattle as reported by Sherman et al. (2009). Among the top 188 SNP ranked by Bayesian LASSO, 39 SNP overlapped with top 188 SNP identified by RF, and 35 SNP were located in QTL regions for RFI as reported by Sherman et al. (2009), slightly fewer than in the top 188 SNP by RF. According to studies on Angus steers (Herd and Arthur, 2009), heat production from metabolic processes, body composition, and physical activity explained up to 73% of the variation in RFI. It is likely that many of these metabolic processes would be the same for beef and dairy cattle, and the results of our study indicate a genetic basis for the similarities. Furthermore, it is possible that the SNP identified as important in the present study can be used to help to refine the locations of previously reported QTL for RFI in beef cattle.

Besides the consistency with Sherman et al. (2009), among the top 188 SNP, Hapmap50758-BTA-69128

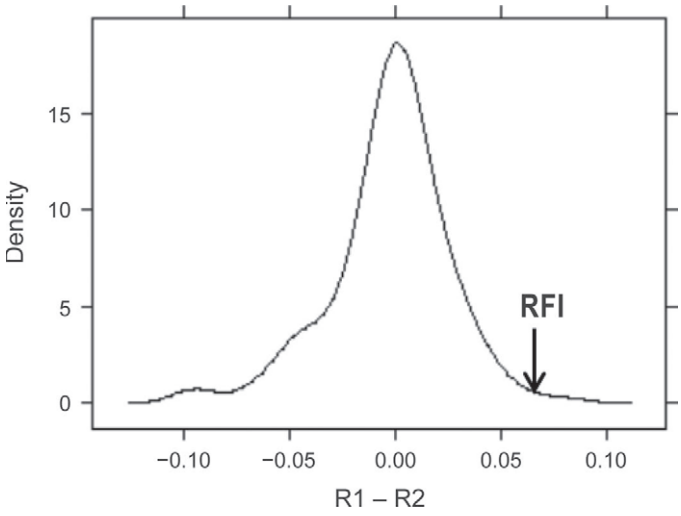

Figure 6. The density plot of ($R_1 - R_2$) for 167 traits including residual feed intake (RFI) that were associated with the 188 SNP with highest importance scores in the Random Forests analysis, where R_1 is the percentage of SNP located in RFI QTL regions for these 188 SNP and R_2 is the same percentage for all 42,275 whole-genome markers.

Table 3. Chromosomal locations, nearest known genes mapped in The Bovine Genome Database (http://bovinegenome.org), importance scores (ΔMSE%, where MSE = mean squared error), and rankings in the Random Forests (RF) and Bayesian LASSO (BL) analyses, for the 188 SNP with highest importance scores in the RF analysis that also mapped to QTL regions for residual feed intake in beef cattle in Sherman et al. (2009), which had been transformed into the physical position in the database

Chromosome	Position (Mb)	SNP name	Gene[1]	ΔMSE%	RF rank	BL rank
3	86.4	ARS-BFGL-NGS-11769	*LOC530929*[2]	0.018	146	19,209
4	26.0	BTA-107055-no-rs		0.030	69	817
	26.2	BTB-01383949		0.015	186	427
	26.2	Hapmap49350-BTA-86626		0.035	56	358
7	13.8	ARS-BFGL-NGS-29980	*KLF1*	0.016	180	170
	14.9	ARS-BFGL-NGS-34694		0.024	91	1,384
	21.3	ARS-BFGL-NGS-76601		0.015	185	426
11	4.6	ARS-BFGL-NGS-111573	*REV1*	0.038	46	69
	5.1	Hapmap53281-rs29026129	*AFF3*	0.068	16	25
	5.4	ARS-BFGL-NGS-25743		0.286	2	2
	5.6	ARS-BFGL-NGS-93607		0.018	141	1,320
	6.1	Hapmap39438-BTA-81105	*TBC1D8*	0.109	5	1,059
	6.2	ARS-BFGL-NGS-116293		0.051	25	2,317
	36.0	BTB-00471723		0.058	20	487
12	88.9	ARS-BFGL-NGS-117411	*COL4A1*[2]	0.039	39	5,948
	90.7	ARS-BFGL-NGS-69018		0.022	108	5,389
	90.9	ARS-BFGL-NGS-26054	*GAS6*	0.023	102	6,553
	91.1	BTB-00509530		0.019	130	3,894
18	34.7	BTB-00708261		0.030	68	13
	38.5	ARS-BFGL-NGS-12047	*LOC510844*	0.078	11	169
	38.5	Hapmap40105-BTA-116218		0.038	44	356
	38.6	Hapmap43850-BTA-43155		0.078	12	45
19	21.9	Hapmap47625-BTA-44726		0.039	40	227
	29.5	Hapmap40907-BTA-121178	*USP43*	0.020	124	1,287
	34.6	Hapmap35517-SCAFFOLD30611_18172	*SLC47A1*	0.018	140	2,306
	35.8	ARS-BFGL-NGS-42120		0.018	144	196
	36.0	Hapmap48343-BTA-45133		0.024	94	90
	36.4	ARS-USMARC-Parent-DQ888312-rs29015945		0.329	1	890
	39.5	ARS-BFGL-NGS-38159		0.017	159	30,122
	58.3	ARS-BFGL-NGS-108439		0.021	120	801
23	45.8	BTA-56614-no-rs		0.043	33	21,294
	45.9	ARS-BFGL-BAC-36395	*LOC784682*	0.031	66	3,324
24	19.5	ARS-BFGL-NGS-38991		0.016	168	8,142
25	8.1	ARS-BFGL-NGS-60044		0.016	177	1,729
	8.2	ARS-BFGL-NGS-113918		0.022	105	484
	12.1	Hapmap24109-BTC-001826	*LOC100139490*	0.021	112	144
	13.6	ARS-BFGL-NGS-4731	*PARN*	0.016	171	3,081
	41.1	Hapmap31341-BTC-065490	*GNA12*	0.020	125	1,691

[1]*KLF1* = Krüppel-like factor 1 (erythroid); *REV1* = REV1, polymerase (DNA directed); *AFF3* = AF4/FMR2 family, member 3; *TBC1D8* = TBC1 domain family, member 8 (with GRAM domain); *COL4A1* = collagen, type IV, α 1; *GAS6* = growth arrest-specific 6; *USP43* = ubiquitin specific peptidase 43; *SLC47A1* = solute carrier family 47, member 1; *PARN* = poly(A)-specific ribonuclease; *GNA12* = guanine nucleotide binding protein (G protein) α 12.

[2]Single nucleotide polymorphisms located within coding region of the gene.

(chromosome 3) and ARS-BFGL-NGS-10829 (chromosome 11) were also identified in a GWAS of RFI by Rolf et al. (2012) in Angus cattle. The SNP ARS-BFGL-NGS-53179 (chromosome 12), ARS-BFGL-NGS-112862 (chromosome 18), BTA-12313-rs29024268 (chromosome 18), Hapmap40907-BTA-121178 (chromosome 19), ARS-BFGL-NGS-42120 (chromosome 19), and ARS-BFGL-NGS-4731 (chromosome 25) were located within a 0.5-Mbp region of significant markers detected by Rolf et al. (2012). No SNP were found to be consistent with studies of Pryce et al. (2012), Snelling et al. (2011), and Bolormaa et al. (2011). Additionally, pathway analysis of the top 188 SNP revealed similar results to those reported by Rolf et al. (2012). One intriguing similarity between Rolf et al. (2012) and this current study was the detection of genes related to the calcium-regulation pathway expressed in smooth and cardiac muscle tissues. The inositol 1,4,5-trisphosphate receptor, type 1 (*ITPR1*) and 5-hydroxytryptamine (serotonin) receptor 4, G protein-coupled (*HTR4*) genes [RefSNP (rs) numbers rs41640891 (UA-IFASA-6532) and rs42809616 (ARS-BFGL-NGS-113598), respectively] code for cell surface receptor proteins that regulate the concentration of calcium ions within the cytoplasm and sarcoplasmic reticulum (Yamada et al., 1994; Khan et al., 1995; Supplementary Figure S1, available

Table 4. Chromosomal locations, nearest known genes mapped in the UMD3.1 assembly (http://bovinegenome.org/cgi-bin/gbrowse/bovine_UMD31/), importance scores (ΔMSE%, where MSE = mean squared error), and rankings in the Random Forests (RF) and Bayesian LASSO (BL) analyses, for the 188 SNP with highest importance scores in the RF analysis that were mapped to annotated genes but not represented into QTL regions for residual feed intake in beef cattle in Sherman et al. (2009)

Chromosome	Position (Mb)	SNP name	Gene[1]	ΔMSE%	RF rank	BL rank
1	41.3	BTA-95584-no-rs	*LOC100336601*	0.037	49	147
	43.6	BTB-00161977	*COL8A1*	0.034	59	105
	44.0	Hapmap43396-BTA-89742	*C1H3orf26*	0.024	95	726
	59.4	UA-IFASA-2169	*ZBTB20*	0.024	92	97
	71.3	ARS-BFGL-NGS-5124	*TFRC*	0.017	163	9,239
	119.3	ARS-BFGL-NGS-59470	*LOC785299*	0.047	28	162
	131.3	ARS-BFGL-NGS-117553	*LOC782895*	0.022	106	2,699
	144.6	BTA-55340-no-rs	*PDE9A*	0.028	72	7,643
	146.5	ARS-BFGL-NGS-108391	*RRP1B*	0.088	10	307
5	105.7	ARS-BFGL-NGS-98808	*KCNA1*[2]	0.021	109	160
	108.4	Hapmap34169-BES10_Contig488_621	*ERC1*	0.041	36	39,208
6	26.4	ARS-BFGL-NGS-59728	*MTTP*	0.023	101	2,908
7	0.9	BTB-01632886	*MAPK9*	0.043	32	478
	41.1	UA-IFASA-9367	*CLK4*	0.018	147	505
	50.5	ARS-BFGL-NGS-106241	*SPOCK1*	0.120	3	99
	62.1	ARS-BFGL-NGS-113598	*HTR4*	0.017	154	13,704
10	18.4	ARS-BFGL-NGS-107048	*THSD4*	0.028	70	4,154
	40.1	BTB-00418910	*MDGA2*	0.017	155	2,438
11	2.7	ARS-BFGL-NGS-12715	*LMAN2L*	0.063	18	1,387
	48.5	ARS-BFGL-NGS-101636	*REEP1*	0.026	79	20,696
	48.9	Hapmap43414-BTA-96067	*ST3GAL5*	0.018	143	4,751
	67.9	BTA-07453-no-rs	*AAK1*	0.020	128	298
	68.7	ARS-BFGL-NGS-10829	*CAPN14*	0.024	98	279
12	77.6	BTA-29995-no-rs	*HS6ST3*	0.028	74	18,844
13	76.0	ARS-BFGL-NGS-23787	*SLC13A3*	0.015	188	790
	77.6	ARS-BFGL-NGS-97619	*PREX1*	0.018	148	1,482
14	21.0	ARS-BFGL-NGS-102399	*LOC512910*	0.020	123	3,651
16	30.8	BTA-96954-no-rs	*CDC42BPA*	0.026	78	707
	35.8	BTA-06703-rs29021060	*WDR64*	0.021	116	12,282
	69.9	BTB-01188142	*KCNK2*	0.018	145	475
18	52.4	ARS-BFGL-NGS-17369	*ZNF404*	0.015	183	1,715
	52.5	ARS-BFGL-NGS-112862	*LOC100141003*	0.019	135	1,076
	55.6	ARS-BFGL-NGS-31529	*LMTK3*	0.033	62	4,218
	56.2	BTA-43831-no-rs	*LOC785907*	0.103	8	48
20	5.5	BTB-01104181	*CPEB4*	0.017	161	9,889
21	69.9	ARS-BFGL-NGS-1345	*KLC1*	0.023	99	324
22	21.8	UA-IFASA-6532	*ITPR1*	0.019	132	399
	25.1	BTA-53914-no-rs	*CNTN6*	0.022	107	1,193
22	32.6	Hapmap52235-ss46526582	*UBA3*[2]	0.015	184	437
	37.6	ARS-BFGL-NGS-100820	*ATXN7*	0.075	15	3,698
	39.2	BTB-00846141	*PTPRG*	0.016	165	10,112
26	28.2	Hapmap46811-BTA-40771	*SORCS1*	0.035	53	242
28	4.4	Hapmap47519-BTA-117732	*LOC614741*	0.047	27	14,142
	5.3	ARS-BFGL-NGS-282	*SIPA1L2*	0.016	173	1,458
	45.0	ARS-BFGL-NGS-91390	*RASSF4*	0.055	23	1,839
	45.0	ARS-BFGL-NGS-1761	*TMEM72*	0.016	166	49
X	1.4	BTB-01316213	*LOC526880*	0.016	167	249
	3.8	ARS-BFGL-NGS-82123	*SEPT6*	0.016	175	686

[1]*COL8A1* = collagen, type VIII, α 1; *ZBTB20* = zinc finger and BTB domain containing 20; *TFRC* = transferrin receptor (P90, CD71); *PDE9A* = phosphodiesterase 9A; *RRP1B* = ribosomal RNA processing 1 homolog B (*Saccharomyces cerevisiae*); *KCNA1* = potassium voltage-gated channel, shaker-related subfamily, member 1; *ERC1* = ELKS/RAB6-interacting/CAST family member 1; *MTTP* = microsomal triglyceride transfer protein; *MAPK9* = mitogen-activated protein kinase 9; *CLK4* = CDC-like kinase 4; *SPOCK1* = sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1; *HTR4* = 5-hydroxytryptamine (serotonin) receptor 4, G protein-coupled; *THSD4* = thrombospondin, type I, domain containing 4; *MDGA2* = MAM domain containing glycosylphosphatidylinositol anchor 2; *LMAN2L* = lectin, mannose-binding 2-like; *REEP1* = receptor accessory protein 1; *ST3GAL5* = ST3 β-galactoside α-2,3-sialyltransferase 5; *AAK1* = AP2 associated kinase 1; *CAPN14* = calcium-activated neutral proteinase 14; *HS6ST3* = heparan sulfate 6-O-sulfotransferase; *SLC13A3* = solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 3; *PREX1* = phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 1; *CDC42BPA* = CDC42 binding protein kinase α (DMPK-like); *WDR64* = WD repeat domain 64; *KCNK2* = potassium channel, subfamily K, member 2; *ZNF404* = zinc finger protein 404; *LMTK3* = lemur tyrosine kinase 3; *CPEB4* = cytoplasmic polyadenylation element binding protein 4; *KLC1* = kinesin light chain 1; *ITPR1* = inositol 1,4,5-trisphosphate receptor, type 1; *CNTN6* = contactin 6; *UBA3* = ubiquitin-like modifier activating enzyme 3; *ATXN7* = spinocerebellar ataxia type 7 protein; *PTPRG* = protein tyrosine phosphatase, receptor type, G; *SORCS1* = sortilin-related VPS10 domain containing receptor 1; *SIPA1L2* = signal-induced proliferation-associated 1 like 2; *RASSF4* = Ras association (RalGDS/AF-6) domain family member 4; *TMEM72* = transmembrane protein 72; *SEPT6* = septin 6.

[2]Single nucleotide polymorphism located within coding region of the gene.

Table 5. Chromosomal locations of 6 genes mapped in the assembly UMD3.1 (http://bovinegenome.org/cgi-bin/gbrowse/bovine_UMD31/) by 12 adjacent SNP that ranked among the 188 SNP with highest importance scores in the Random Forests analysis, as well as linkage disequilibrium (as measured by R^2) between these SNP

Gene[1]	Adjacent SNP	Chromosome	Distance between SNP (kb)	R^2	Coding region between?	No. of SNP between
LOC100299906	ARS-BFGL-NGS-30459 BTA-94473-no-rs	6	74.9	0.37	Yes	1
RORA	BTA-70155-no-rs BTB-00426034	10	24.5	0.66	No	0
VPS13B	BTA-35285-no-rs ARS-BFGL-BAC-24806	14	357.4	0.20	Yes	6
LSM14A	BTB-00718231 Hapmap30625-BTA-43445	18	38.8	1.00	Yes	0
PPARD	ARS-BFGL-NGS-40073 ARS-BFGL-NGS-61728	23	28.6	0.87	Yes	0
LOC515570	ARS-BFGL-NGS-2693 ARS-USMARC-Parent-AY941204-rs17872131	25	6.3	0.86	Yes	0

[1]*RORA* = RAR-related orphan receptor A; *VPS13B* = vacuolar protein sorting 13 homolog B; *LSM14A* = LSM14A, SCD6 homolog A (*Saccharomyces cerevisiae*); *PPARD* = peroxisome proliferator-activated receptor δ.

online at http://dx.doi.org/10.3168/jds.2013-6237). Identification of this pathway in both studies suggests that myometrial activity (i.e., muscle contraction and relaxation) may influence RFI in both dairy and beef cattle, which would be related to the findings of Herd and Arthur (2009) regarding physical activity and RFI variance.

As shown in Tables 3 and 4, mapping to the UMD3.1 assembly, 68 different annotated genes were associated with 74 of the 188 SNP, with the highest importance scores in the RF analysis. Four SNP [ARS-BFGL-NGS-11769 (BTA3), ARS-BFGL-NGS-98808 (BTA5), ARS-BFGL-NGS-117411 (BTA12), and Hapmap52235-ss46526582 (BTA22)] are in the coding regions of 4 different genes [*LOC530929*; potassium voltage-gated channel, shaker-related subfamily, member 1 (*KCNA1*); collagen, type IV, α1 (*COL4A1*); and ubiquitin-like modifier activating enzyme 3 (*UBA3*)]. The SNP in *COL4A1* also mapped to a QTL region for RFI in beef cattle (Sherman et al., 2009). Additionally, 6 genes [*LOC100299906*, RAR-related orphan receptor A (*RORA*), vacuolar protein sorting 13 homolog B (*VPS13B*), SCD6 homolog A (*Saccharomyces cerevisiae*) (*LSM14A*), peroxisome proliferator-activated receptor δ (*PPARD*), and (*loc515570*)] were mapped by pairs of adjacent SNP (Table 5). Some of these are in high linkage disequilibrium, as indicated by the coefficient of determination values in Table 5. With the exception of *RORA*, coding regions are located between the 2 SNP. Because the current BovineSNP50 BeadChip cannot provide greater resolution, additional studies are needed to investigate genetic variation near or within these genes. Among closest annotated RefSeq genes, the *PPARD* gene encodes a receptor protein that regulates the size and number of peroxisomes in the cell (Xu et al., 1999). Polymorphisms in this gene have been linked to obesity in humans (Shin et al., 2004), so it would follow that the alleles of *PPARD* identified in this study may affect feed efficiency in cattle. Gene functional analysis revealed an enrichment of integral membrane protein and ion transport genes in this gene data set (Supplementary Table S2, available online at http://dx.doi.org/10.3168/jds.2013-6237). The PANTHER biological process analysis also revealed a high percentage of genes involved in cellular communication (33.3%) and cellular processes (43%) (Supplementary Table S3). Cell surface receptors such as those identified in this study often affect critical cellular secondary messaging systems and are excellent starting points for future studies investigating RFI in cattle. Pathway enrichment scores from both PANTHER and DAVID web tools were calculated against the entire gene list for each organism. Although the use of the actual lists of genes within 37,000 bp of each BovineSNP50 SNP position would result in more-accurate enrichment scores, limitations of the existing web tools prevent the use of such large, specialized lists. Regardless, only approximately 14% of the RefSeq genes were not within our designated cutoff for association with BovineSNP50 SNP, so the entire gene list is a close approximation to our actual background gene list for this calculation.

CONCLUSIONS

The importance scores generated by RF algorithms can be useful for identifying individual SNP that have large additive and epistatic effects on RFI or other economically important traits. Frequently occurring descendant pairs can be discovered by examining the structure of individual trees within the RF, and in this

manner, possible pairs of epistatic SNP can be identified for further study. Future studies, perhaps with larger data sets, can add to our knowledge regarding the genetic mechanisms underlying RFI in dairy cattle and the possible roles of annotated genes that were mapped by important SNP in our RF analysis. Lastly, the consistency in results between this study and studies of RFI in beef cattle should be explored further, as ongoing projects to describe the underlying genetic basis of RFI are currently underway in both beef and dairy cattle.

ACKNOWLEDGMENTS

This project was supported by Agriculture and Food Research Initiative Competitive Grants no. 2008-35205-18711 and 2011-68004-30340 from the US Department of Agriculture (USDA) National Institute of Food and Agriculture (Washington, DC). Support from Hatch grant no. MSN139239 from the Wisconsin Agricultural Experiment Station (Madison) is acknowledged, and K. A. Weigel acknowledges partial financial support from the National Association of Animal Breeders (Columbia, MO).

REFERENCES

Arthur, P. F., J. A. Archer, D. J. Johnston, R. M. Herd, E. C. Richardson, and P. F. Parnell. 2001. Genetic and phenotypic variance and covariance components for feed intake, feed efficiency, and other postweaning traits in Angus cattle. J. Anim. Sci. 79:2805–2811.

Bolormaa, S., B. J. Hayes, K. Savin, R. Hawken, W. Barendse, P. F. Arthur, R. M. Herd, and M. E. Goddard. 2011. Genome-wide association studies for feedlot and growth traits in cattle. J. Anim. Sci. 89:1684–1697.

Breiman, L. 2001. Random Forests. Mach. Learn. 45:5–32.

Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84:210–223.

Chen, X., C.-T. Liu, M. Zhang, and H. Zhang. 2007. A forest-based approach to identifying gene and gene-gene interactions. Proc. Natl. Acad. Sci. USA 104:19199–19203.

Childers, C. P., J. T. Reese, J. P. Sundaram, D. C. Vile, C. M. Dickens, K. L. Childs, H. Salih, A. K. Bennett, D. E. Hagen, D. L. Adelson, and C. G. Elsik. 2011. Bovine Genome Database: Integrated tools for genome annotation and discovery. Nucleic Acids Res. 39:D830–D834.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385.

de Los Campos, G., and P. Perez Rodriguez. 2012. BLR: Bayesian Linear Regression, R package version 1.3.

Exton, S. C., P. F. Arthur, J. A. Archer, and R. M. Herd. 1999. Strategies for industry adoption of genetic improvement of net feed efficiency in beef cattle. Proc. Assoc. Advmt. Anim. Breed. Genet. 13:424–427.

Herd, R. M., and P. F. Arthur. 2009. Physiological basis for residual feed intake. J. Anim. Sci. 87:E64–E71.

Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4:44–57.

Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009b. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37:1–13.

Jiang, R., W. Tang, X. Wu, and W. Fu. 2009. A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics 10(Suppl. 1):S65.

Khan, N. A., F. Ferriere, and P. Deschaux. 1995. Serotonin-induced calcium signaling via 5-HT1A receptors in human leukemia (K 562) cells. Cell. Immunol. 165:148–152.

Koch, R. M., L. A. Swiger, D. Chambers, and K. E. Gregory. 1963. Efficiency of feed use in beef cattle. J. Anim. Sci. 22:486–494.

Li, J., K. Das, G. Fu, R. Li, and R. Wu. 2011. The Bayesian lasso for genome-wide association studies. Bioinformatics 27:516–523.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.

Liaw, A., and M. Wiener. 2011. randomForest: Breiman and Cutler's Random Forests for classification and regression, R package version 4.6-3.

Lunetta, K. L., L. B. Hayward, J. Segal, and P. Van Eerdewegh. 2004. Screening large-scale association study data: Exploiting interactions using Random Forests. BMC Genet. 5:32.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4:e5350.

Mi, H., B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano, and P. D. Thomas. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 33:D284–D288.

NRC. 2001. Nutrient Requirements of Dairy Cattle. 7th rev. ed. Natl. Acad. Sci., Washington, DC.

Park, T., and G. Casella. 2008. The Bayesian lasso. J. Am. Stat. Assoc. 103:681–686.

Pryce, J. E., J. Arias, P. J. Bowman, S. R. Davis, K. A. Macdonald, G. C. Waghorn, W. J. Wales, Y. J. Williams, R. J. Spelman, and B. J. Hayes. 2012. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. J. Dairy Sci. 95:2108–2119.

Quinlan, A. R., and I. M. Hall. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Robinson, D. L., and V. H. Oddy. 2004. Genetic parameters for feed efficiency, fatness, muscle area and feeding behaviour of feedlot finished beef cattle. Livest. Prod. Sci. 90:255–270.

Rolf, M. M., J. F. Taylor, R. D. Schnabel, S. D. McKay, M. C. McClure, S. L. Northcutt, M. S. Kerley, and R. L. Weaber. 2012. Genome-wide association analysis for feed efficiency in Angus cattle. Anim. Genet. 43:367–374.

Ruppert, D., S. J. Sheather, and M. P. Wand. 1995. An effective bandwidth selector for local least squares regression. J. Am. Stat. Assoc. 90:1257–1270.

Sainz, R. D., and P. V. Paulino. 2004. Residual feed intake. Sierra Foothill Research and Extension Center, University of California, Davis.

Schenkel, F. S., S. P. Miller, and J. W. Wilton. 2004. Genetic parameters and breed differences for feed efficiency, growth and body composition traits of young beef bulls. Can. J. Anim. Sci. 84:177–185.

Sherman, E. L., J. D. Nkrumah, C. Li, R. Bartusiak, B. Murdoch, and S. S. Moore. 2009. Fine mapping quantitative trait loci for feed intake and feed efficiency in beef cattle. J. Anim. Sci. 87:37–45.

Shin, H. D., B. L. Park, L. H. Kim, H. S. Jung, Y. M. Cho, M. K. Moon, Y. J. Park, H. K. Lee, and K. S. Park. 2004. Genetic polymorphisms in peroxisome proliferator-activated receptor δ associated with obesity. Diabetes 53:847–851.

Snelling, W. M., M. F. Allan, J. W. Keele, L. A. Kuehn, R. M. Thallman, G. L. Bennett, C. L. Ferrell, T. G. Jenkins, H. C. Freetly, M. K. Nielsen, and K. M. Rolfe. 2011. Partial-genome evaluation of

postweaning feed intake and efficiency of crossbred beef cattle. J. Anim. Sci. 89:1731–1741.
Spurlock, D. M., J. C. M. Dekkers, R. Fernando, D. A. Koltes, and A. Wolc. 2012. Genetic parameters for energy balance, feed efficiency, and related traits in Holstein cattle. J. Dairy Sci. 95:5393–5402.
Thomas, P. D., M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. 2003. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res. 13:2129–2141.
van Iersel, M. P., T. Kelder, A. R. Pico, K. Hanspers, S. Coort, B. R. Conklin, and C. Evelo. 2008. Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics 9:399.
VandeHaar, M. J., and N. St-Pierre. 2006. Major advances in nutrition: Relevance to the sustainability of the dairy industry. J. Dairy Sci. 89:1280–1291.
VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, and F. Schenkel. 2009. Reliability of genomic predictions for North American dairy bulls. J. Dairy Sci. 92:16–24.
Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, and D. B. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93:5942–5949.
Wand, M. P., and M. C. Jones. 1995. Kernel Smoothing. Vol. 60. Chapman and Hall/CRC, Boca Raton, FL; London, UK; New York, NY; and Washington, DC.
Weigel, K. A., G. de los Campos, O. González-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248–5257.
Xu, H. E., M. H. Lambert, V. G. Montana, D. J. Parks, S. G. Blanchard, P. J. Brown, D. D. Sternbach, J. M. Lehmann, G. B. Wisely, T. M. Willson, S. A. Kliewer, and M. V. Milburn. 1999. Molecular recognition of fatty acids by peroxisome proliferator-activated receptors. Mol. Cell 3:397–403.
Yamada, N., Y. Makino, R. A. Clark, D. W. Pearson, M. G. Mattei, J. L. Guénet, E. Ohama, I. Fujino, A. Miyawaki, and T. Furuichi. 1994. Human inositol 1,4,5-trisphosphate type-1 receptor, InsP3R1: Structure, function, regulation of expression and chromosomal localization. Biochem. J. 302:781–790.
Yi, N., and S. Xu. 2008. Bayesian LASSO for quantitative trait loci mapping. Genetics 179:1045–1055.