Evaluation of a functional variant assay for selecting beef cattle

W. M. Snelling,1,2 T. G. McDaneld3, L. A. Kuehn4, A. M. Dickey1 & G. L. Bennett1

1 U.S. Department of Agriculture (USDA). Agricultural Research Service, U.S. Meat Animal Research Center, P.O. Box 166, Clay Center, Nebraska, 68933, United States
2 warren.snelling@ars.usda.gov (Corresponding Author)
3 Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity provider and employer.

Summary

A commercially available genotyping assay for functional variants was chosen to obtain genotypes needed for a selection experiment in populations of pedigreed cattle that have not been extensively genotyped. The assay design included probes for coding sequence variation in 88% of annotated protein coding genes, and variants up- and downstream that may regulate 74% of genes, so function of 97% of genes may be affected by an assayed variant. Slightly more than half (54%) of the 221,115 probed variants were polymorphic in the 3,442 selection population animals genotyped. The low conversion was not unexpected, due to the assay including rare variants that might allow detecting lethal recessives in large populations. The informative variants include coding sequence variation in 63% of genes and variants that may affect regulation of 65%. Overall, 84% of annotated genes may be affected by a polymorphic variant on the functional variant assay. The commercial assay is sufficient to genotype many variants affecting gene function that are segregating in these populations. Strategies to couple this assay with targeted assays for additional functional variation may be needed for some applications.

Keywords: functional variation, genotyping, beef cattle

Introduction

The number of single nucleotide polymorphisms (SNP) and short insertions and deletions described in Bos taurus has grown from just over two million when the bovine genome assembly was first announced (Bovine Genome Sequencing and Analysis Consortium, 2009) to over 100 million today (Sherry et al., 2001; dbSNP Build IDs:130, 150). While the majority of these variants are silent, a fraction affect gene function and ultimately animal phenotypes. Utilizing these functional SNP underlying quantitative trait loci (QTL) may offer an opportunity for genomic predictions that are portable across breeds.

Genotypes are needed for genomic prediction using functional SNP. For well-studied populations having extensive genotypes from moderate (~50,000 SNP; 50K) and high-density (~780,000 SNP; HD) whole genome assays and genomic sequence that covers most haplotypes occurring in the population, imputation from 50K and HD to sequence variant genotypes is a reasonable option. Directly genotyping functional variants is needed when available genotypes and genomic sequence are inadequate to support high accuracy imputation (r > ~0.98). Options for direct genotyping include custom SNP-chip assays, targeted genotyping by sequencing, and the GGP F-250 (F250; Neogen Genomics, Lincoln,
NE, USA), a commercially available functional variant assay, designed to interrogate almost 200,000 predicted functional variants detected in genome and transcriptome sequence of cattle representing several different breeds (Taylor et al., 2016).

The F250 was considered the most viable option for obtaining the functional variant genotypes needed for a “genetics first” approach to selection in four beef cattle populations (Bennett et al., 2018). Developing a custom chip to genotype around 1000 animals each year for this experiment was prohibitively expensive, and too many genotyping by sequencing assays, limited to about 5000 variants in each reaction, would be needed to genotype the 200,000 variants. Objectives of this study were to characterize functional content of the F250, and evaluate performance in the populations under selection.

**Material and methods**

**Populations under selection**

Animals in the selection populations represented three composites (MARC I - 0.25 Limousin, 0.25 Brown Swiss, 0.25 Charolais, 0.125 Angus, 0.125 Hereford; MARC II – 0.25 Hereford, 0.25 Angus, 0.25 Gelbvieh, 0.25 Simmental; MARC III – 0.25 Pinzgauer, 0.25 Red Poll, 0.25 Hereford, 0.25 Angus) and purebred Angus. These populations were originally developed to evaluate retained heterosis in the Germplasm Utilization Project (Gregory et al., 1991) and were subsequently selected for calving ease (Bennett, 2008), and to equalize allele frequencies of specific markers associated with carcass and meat quality characteristics (Bennett et al., 2013; Tait et al., 2014).

**Marker panel for selection**

The F250 was chosen as a readily available assay that could be utilized to genotype functional alleles in the “genetics first” selection experiment. Descriptions of each variant, including position on the UMD3.1 assembly (Zimin et al., 2009), alleles (SNP or indel), probe sequences and sequence flanking each variant were provided by Neogen Genomics. Variant position and alleles were used to assess functional impact relative to Ensembl annotation of UMD3.1 (Aken et al., 2016). Both snpEff (Cingolani et al., 2012) and the Ensembl Variant Effect Predictor (VEP; McLaren et al., 2016) were used to predict functional effects. Variants were assigned to functional categories based on predicted effects. When multiple effects were predicted for a variant, the variant was assigned to the highest-impact category. A BovineHD category was created to include intergenic and intronic variants common to the F250 and BovineHD assays; BovineHD SNP in or near protein-coding sequence were assigned to higher impact categories.

**Genotyping and quality control**

DNA from all cattle (calves, cows and bulls) in the selection populations in 2015 was submitted for genotyping on an Illumina platform, with genotypes called by the GenCall algorithm (Steemers & Gunderson, 2007). Heifer calves born in 2016, and all 2017-born calves were also genotyped on the same platform. Because the clusters used by GenCall were updated with a more comprehensive reference before the 2017 calves were called, all previously genotyped samples were re-called with the updated clusters to maintain consistency between the 2017 and earlier calls.
Each batch of genotypes was initially screened for call rate by animal, and animals with unusually low call rates resubmitted for genotyping. Occurrences of each genotype (AA, AB, BB) were counted, and observations of each allele (A, B) were determined as nA = 2nAA + nAB, nB = nAB + 2BB, where nA and nB are counts of the A and B alleles, and nAA, nAB and nBB are counts of homozygous AA, heterozygous AB and homozygous BB genotypes. Variants were checked for excess heterozygosity and allele counts less than the low error rate for Illumina Beadchip assays (0.2%; Jiang et al., 2013). Variants placed on chromosome X but outside the pseudoautosomal region (paX; chromosome X:143,861,798-148,823-899 bp; Mao et al., 2016) were discarded if more than 1% of males were called heterozygous. If fewer males were called heterozygous, the heterozygous male calls discarded but female calls retained. Call rates by marker were also determined, and a minimum marker call rate was established based on the animal call rate distribution.

Variants passing call rate, genotype and allele count filters were further assessed for imputation accuracy, because low accuracy imputation may indicate suspicious genotypes. For this evaluation, complete F250 genotypes of 185 sires, 1081 dams, and 1,281 non-parent animals born prior to 2017 served as the reference. Genotypes of 895 2017-born calves were split into genotypes for 32,872 markers shared with the BovineHD assay and genotypes for markers unique to the F250. Findhap.f90 (VanRaden et al., 2013), with available pedigree including 2017 calves assigned to natural service sires by exclusion, was used to impute the 2107-born calves from their BovineHD genotypes to content unique to the F250. Each autosome, chromosome X and paX were imputed separately. Chromosome X was treated as sex-linked, and paX like an autosome. Accuracy was assessed by correlation (r) between observed and imputed F250 genotypes, and variants not meeting a threshold of r < 0.95 were removed.

Results and Discussion

GGP F-250 content

Information provided about the F250 placed 221,064 variants on autosomes or chromosome X. Annotated protein-coding sequence contained 117,537 of these variants, another 29,837 were in 5’ or 3’ untranslated regions (UTR) or located within 5000 bases up- or downstream of protein coding genes. Non-coding RNA features contained 1,556 variants, 678 were located in pseudogenes, and the remainder were intergenic or intronic. About 97% of the 19,994 annotated protein coding genes may be altered or regulated by at least one F250 variant. Both snpEff and VEP predicted that amino acid sequence of 16,769 genes could be altered by 94,641 non-synonymous SNP, and 8,300 high-impact, loss-of-function (LOF) variants could interrupt the protein coded by 5,751 genes. Minor differences between snpEff and VEP predictions were observed in predicted effects on pseudogenes and variants located between protein coding genes and non-coding RNA.

Genotype quality control

Genotypes of 206,667 F250 variants were called with the updated GenCall clusters, 5,311 fewer variants than were called with the original clustering. The markers not called with the updated clusters included markers that were completely monogenic or excessively heterozygous in the original calls. With the updated clusters, 0.9% of called markers are monogenic, and only one variant was excessively heterozygous, with 99.4% heterozygous
calls in the 3,442 selection population animals. More than half (61.2%) of the X chromosome markers were eliminated, however, by excessive (> 1%) heterozygosity in the 1,209 genotyped males. At least partially due to design of the assay and deliberate inclusion of rare variants that may be lethal recessives (Taylor et al., 2016), 36.4% of the called variants were too rare to be informative in the selection populations. Requiring a minor allele frequency (MAF) of 0.2% was based on discrepancies observed between genotypes called on Illumina and Affymetrix platforms (Jiang et al., 2013). That threshold may eliminate functional alleles contributing to mutation load from consideration in the “genetics first” selection experiment, but genotypes observed at such low frequencies are difficult to distinguish from error. Additionally, the large number of chromosome X variants with several males called heterozygous suggests a higher threshold might be appropriate.

The 129,817 variants passing heterozygosity and MAF criteria had call rates ranging from 41.8% to 100%. After re-genotyping low call rate samples, animal call rates ranged from 83.0% to 99.9%. There was a break in the continuum of animal call rates at 89%; five animals and 395 variants fell below this threshold.

**Imputation.**

Genotypes of the 13,451 variants called in all 895 2017-born calves and 68 potential natural service sires were compared in order to complete the pedigree needed for imputation. All calves were unambiguously matched to a single bull, with 93.3% of calves matched to one bull with zero opposite homozygote exclusions. Most of the remaining calves were matched with one exclusion. These matches were added to previously established pedigree to evaluate imputation accuracy of the filtered variants unique to the F250 assay.

Mean (median) imputation accuracies by calf for 2017-born calves were 0.996 (0.995), with a minimum of 0.828 for one of the calves with call rate less than 89%. By variant, mean (median) accuracies were 0.951 (0.996). Accuracy for 2.6% of the variants could not be computed because the 2017 calves were monogenic, or imputed to be monogenic. Of the variants with accuracies, 2.3% had correlations less than zero between observed and imputed genotypes, and 9.6% of accuracies were less than 0.95. Minor allele frequency was, at best, weakly related to imputation accuracy ($r^2=0.05$), variants imputed with high or low accuracy covered the complete spectrum of allele frequencies, although the lowest accuracy variants were predominantly low MAF. The majority of rare variants (MAF < 0.01) were imputed with high accuracy. The generally high accuracies observed for imputing the 2017 calves from 32,878 SNP common between the F250 and BovineHD to 96,549 variants unique to the F250 was unexpected. Mean accuracies around 0.85 for imputing from commercial SNP arrays to variants detected in sequenced sires have been reported (Daetwyler et al., 2014; Hayes et al., 2014; Li et al., 2014) and observed in other USMARC data. The difference with these data may be that all parents, not just influential sires, contribute to the reference.

**Informative variant impact**

Between variants never called and called variants not meeting quality control criteria, genotypes for 54.5% of the variants interrogated by the F250 assay appear informative in the selection populations. Most of the variants located outside annotated protein coding regions are polymorphic in these populations, but less than 10% of variants predicted to disrupt coded proteins and one-third of those altering amino acid sequence are useful (Table 1). Probed
variants segregating in these populations still account for variation in coding sequence of 63% of genes, 65% may be affected by variation in 5’ and 3’ regions flanking the protein coding sequence. The F250 appears capable of accounting for much functional variation that is segregating in these populations. Additional variation that is not assayed by the F250 may exist. Further research may explore strategies to couple the F250 with targeted genotyping approaches for a more comprehensive assessment of functional variation.

Table 1. Predicted impact of variants interrogated by GGP F-250 functional variant assay.

<table>
<thead>
<tr>
<th>Functional classification</th>
<th>Probed variants</th>
<th>Genes affected</th>
<th>Usable variants</th>
<th>Genes affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein coding</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>loss of function</td>
<td>8,300</td>
<td>5,751</td>
<td>703</td>
<td>651</td>
</tr>
<tr>
<td>non-synonymous</td>
<td>94,641</td>
<td>16,770</td>
<td>32,057</td>
<td>10,985</td>
</tr>
<tr>
<td>synonymous</td>
<td>14,607</td>
<td>7,175</td>
<td>8,860</td>
<td>5,154</td>
</tr>
<tr>
<td>Potentially regulatory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-coding RNA</td>
<td>1,557</td>
<td>386</td>
<td>915</td>
<td>275</td>
</tr>
<tr>
<td>5’ &amp; 3’ UTR</td>
<td>13,380</td>
<td>7,811</td>
<td>9,892</td>
<td>6,409</td>
</tr>
<tr>
<td>5 kb up- &amp; downstream</td>
<td>16,458</td>
<td>10,007</td>
<td>12,467</td>
<td>8,685</td>
</tr>
<tr>
<td>No annotated effect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BovineHD(^1)</td>
<td>29,035</td>
<td></td>
<td>25,527</td>
<td></td>
</tr>
<tr>
<td>pseudogenes</td>
<td>678</td>
<td>205</td>
<td>252</td>
<td>125</td>
</tr>
<tr>
<td>intergenic</td>
<td>31,041</td>
<td></td>
<td>21,308</td>
<td></td>
</tr>
<tr>
<td>intronic</td>
<td>11,418</td>
<td></td>
<td>8,427</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>221,115</td>
<td>19,359</td>
<td>120,408</td>
<td>16,860</td>
</tr>
</tbody>
</table>

\(^1\) Common variants on both BovineHD and GGP F-250 assays that are not in a protein coding or potentially regulatory category.

List of References


