ABSTRACT: The purpose of this study was: (i) to validate and detect copy number variations (CNV) in the cattle genome, and (ii) to find associations between CNV and feed efficiency traits. Average daily gain and dry matter intake were obtained from 3,884 beef cattle. Animals were genotyped with the panels of 50,000 single nucleotide polymorphisms (SNP) or 770,000 SNP. The log R ratio (LRR) was used as an indicator of CNV. Genetic correlations were used to find CNV in the genome (neighboring SNP with highly correlated LRR) and to find associations between CNV and traits. A spectral decomposition of the genomic relationship matrix was obtained to improve computational efficiency. We found 73 CNV regions, and 619 CNV correlated with traits. The results of this study will improve our understanding of structural variants in the cattle genome and help us in selecting more efficient animals.

Keywords: CNV; feed efficiency; GWAS

Introduction

Copy number variations (CNV) are segments of the genome larger than 1kb for which copy number differences have been observed from a reference genome. These can be used to find associations between the genome and phenotypic traits. The purpose of the research was two-fold: (i) to validate and detect CNV in the cattle genome, and (ii) to find associations between CNV and feed efficiency traits in beef cattle. Feed efficiency traits are economically relevant, as feed is the greatest cost to beef operations. Selecting for animals that are more efficient will therefore increase profits by decreasing production costs, and can help mitigate the environmental impact of beef production.

Materials and Methods

Animal care and use. All animals were reared and handled in accordance with the Guide for Care and Use of Agricultural Animals in Research and Teaching and was approved by the Animal Care and Use Committee of the US Meat Animal Research Center.

Data. Average daily gain (ADG) and daily dry matter intake were obtained from 3,884 animals, representing 24 contemporary groups. Contemporary groups consisted of year, season within year (fall and spring), feeding group or phase within season, and sex. The mean number of animals in a contemporary group was 162, with a minimum of 99 and maximum of 261 animals. There were 2,249 steers and 1,635 heifers.

Depending on year and season, steers were fed finishing diets which varied from 2.90 to 3.33 Mcal of ME/kg of dry matter. Heifers were fed growing diets which varied from 2.35 to 2.96 Mcal of ME/kg of dry matter. Individual feed intake was recorded. Cattle were weighed every 2-3 weeks; at the beginning and end of the trial they were weighed on 2 consecutive days. Daily metabolizable energy intake (DMEI) was calculated. ADG and DMEI were standardized to the variance of males.

Animals were genotyped with either the Illumina BovineSNP50 BeadChip v1 or v2, or with the Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA). Only SNP that overlapped between the three platforms were used. These SNP arrays were used to detect CNV. The log R ratio (LRR) was used as an indicator of CNV. It is a normalized measure of total signal intensity for two alleles of a SNP. LRR are positively correlated with copy number.

CNV detection models. The presence of CNV regions is indicated by LRR for consecutive SNP that are highly correlated. To evaluate genetic correlations among SNP, a multivariate model was run where the response variables were the LRR of three adjacent SNP, a “triplet”. Moving along a chromosome, new triplets were formed by dropping the first SNP and adding the adjacent SNP. For each triplet, the model fitted was:

\[ Y = X\beta + u + e \]

where \( Y \) was a matrix of LRR values for each triplet in a chromosome; \( \beta \) was a vector of fixed effects solutions specific to SNP, \( u \) was a vector of random direct genetic effects, \( e \) was a vector of random residuals, and \( X \) was an incidence matrix of fixed effects, which included an overall intercept and platform type (version 1 or 2 of 50k platform, or HD platform). Random effects were distributed as multivariate normal with mean zero and \( r(u) = G = \Sigma \otimes A \), \( var(e) = R = E \otimes I \), \( cov(u,e) = 0 \), and \( var(y) = G + R \). Here \( A \) is the genomic relationship matrix among animals calculated according to VanRaden (2008) estimated from genotypes (coded 0,1 and 2) for all autosomal SNP.

Positive genetic correlations among LRR for consecutive SNP indicate CNV. To detect CNV, we tested the null hypotheses that all genetic covariances among triplets of consecutive SNP were zero against the alternative that covariances were positive within a block of three consecutive SNP. Correlations among consecutive SNP were evaluated with a log-likelihood ratio test in which two models (CNVfull, CNVreduced) with different variance structures were compared with 3 degrees of freedom. The
genetic variance-covariance matrix, $\Sigma$, for the full and reduced models were:

$$\Sigma_{\text{full}} = \begin{bmatrix} \sigma^2_{g1} & \sigma_{g1g2} & \sigma_{g1g3} \\ \sigma_{g2g1} & \sigma^2_{g2} & \sigma_{g2g3} \\ \sigma_{g3g1} & \sigma_{g3g2} & \sigma^2_{g3} \end{bmatrix}, \quad \text{and} \quad \Sigma_{\text{reduced}} = \begin{bmatrix} \sigma^2_{g1} & 0 & 0 \\ 0 & \sigma^2_{g2} & 0 \\ 0 & 0 & \sigma^2_{g3} \end{bmatrix}.$$  

The environmental variance-covariance matrix for both (full and reduced) models included estimates of all variances and covariances.

For each SNP there were 3 variance estimates, and for adjacent SNP there were 2 co-variance estimates (except for SNP at the beginning and end of chromosomes). Common variance and co-variance estimates among adjacent trios were averaged and a correlation coefficient between adjacent SNP was calculated. Heritabilities of each variance components. On each chromosome we investigated regions where the log-likelihood ratio tests of at least three consecutive triplets were significant ($P < 10^{-10}$), and correlations among SNP were greater than 0.85.

**Trait-CNV association models.** Associations between feed efficiency traits and CNV where investigated with genetic variance-covariance matrix, $\Sigma$, for the full and reduced models included estimates of all variances and covariances.

For each SNP there were 3 variance estimates, and for adjacent SNP there were 2 co-variance estimates (except for SNP at the beginning and end of chromosomes). Common variance and co-variance estimates among adjacent trios were averaged and a correlation coefficient between adjacent SNP was calculated. Heritabilities of each variance components. On each chromosome we investigated regions where the log-likelihood ratio tests of at least three consecutive triplets were significant ($P < 10^{-10}$), and correlations among SNP were greater than 0.85.

The likelihoods for the full and reduced model were obtained, and tested for significance with a log-likelihood ratio test with six degrees of freedom. The estimates of the three variance and two covariance components again were averaged.

**Spectral decomposition of genomic relationship matrix.** We used ASREML v3.0 (Gilmour et al., 2009) and R (R Core Team, 2012) to obtain estimates for the (co)variance components in both models. The genomic relationship matrix ($A$), and more importantly its inverse, is dense, and therefore inverting the left hand side of Henderson’s mixed model equations is computationally intensive. The matrix $A$ is positive definite and therefore can be decomposed into its eigenvectors and eigenvalues, and this decomposition makes the problem more tractable.

The matrix $A$ was decomposed as $A = KLK'$, where $K$ is a matrix of eigenvectors, and $L$ is a diagonal matrix of eigenvalues. The transformed model therefore was:

$$y^* = (I \otimes K')y = (I \otimes K')X\beta + (I \otimes K')Zu + (I \otimes K')e = X'\beta^* + Z'u^* + e^*$$

In this case all covariates were fitted as continuous variables, with no mean fitted. For this model, $E[y^*] = (I \otimes K'X)\beta$, and $\text{var}(y^*) = (I \otimes K')\text{var}(y)(I \otimes K) = (I \otimes K')(G + R)(I \otimes K)$. In our case,

$$\text{var}(u^*) = \Sigma \otimes L. \quad \text{Likewise,}$$

$$\text{var}(e) = R = (I \otimes K)\begin{bmatrix} \sigma^2_{e1} & 1 \sigma_{e1e2} & \ldots \\ 1 \sigma_{e1e2} & \sigma^2_{e2} & \ldots \\ \vdots & \vdots & \ddots \end{bmatrix} (I \otimes K')'$$

and $\text{var}(e^*) = E \otimes I$. because $\text{var}(y^*) = (I \otimes K')\text{var}(y)(I \otimes K) = \Sigma \otimes L + E \otimes I$, and

$$\text{var}(y^*) = \begin{bmatrix} \sigma^2_{g1} & 1 \sigma_{g1g2} & \ldots \\ 1 \sigma_{g1g2} & \sigma^2_{g2} & \ldots \\ \vdots & \vdots & \ddots \end{bmatrix} + \begin{bmatrix} \sigma^2_{e1} & 1 \sigma_{e1e2} & \ldots \\ 1 \sigma_{e1e2} & \sigma^2_{e2} & \ldots \\ \vdots & \vdots & \ddots \end{bmatrix} = \text{var}(u^*) + \text{var}(e^*)$$

The computational improvement from this transformation comes from $\text{var}(u^*) = \Sigma \otimes L$ being trivial to invert because $(\Sigma \otimes L)^{-1} = \Sigma^{-1} \otimes L^{-1}$, and $L$ is diagonal. Because $\text{var}(y^*) = \Sigma \otimes L + E \otimes I$, the system is amendable to sparse matrix techniques. The expected value and variance of $y$ and $y^*$ are identical. Therefore, the transformed and untransformed models are equivalent; variances and covariance estimates are identical.
Results

CNV detection. Likelihood ratio tests of CNV detection models indicated that estimates of covariances between windows of SNP were significantly different from zero in most cases. On average, 80% of windows of SNP achieved nominal significance (\(P < 0.01\)), while 70% and 65% achieved chromosome-wide and genome-wide significance, respectively, with a Bonferroni adjustment for multiple comparisons controlling for type I error at 0.01.

Table 1 summarizes the CNV regions detected when significant correlations between adjacent SNP were greater than or equal to 0.85. There were on average 3 regions with 3 consecutive highly correlated SNP on each chromosome. On BTA1 there were 7 regions with 3 consecutive SNP correlated with each other. However, on several chromosomes there was only one such region (BTA2, 10, 12, 14, 24, 27), and on others none (BTA15, 16, 25). There were 9 chromosomes where regions consisted of 5 or more consecutive SNP.

Table 1. Summary of regions of CNV where consecutive SNP were significantly correlated

<table>
<thead>
<tr>
<th>Consecutive SNP</th>
<th>No.</th>
<th>Ave. cor</th>
<th>Ave. h²</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>73</td>
<td>0.91</td>
<td>0.42</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>0.92</td>
<td>0.38</td>
</tr>
<tr>
<td>5+</td>
<td>9</td>
<td>0.92</td>
<td>0.41</td>
</tr>
</tbody>
</table>

1All correlations were significant at a nominal \(P<1.0E-10\)
2Correlation level \(\geq 0.85\)
3Number of regions of the respective number of consecutive
4SE of correlations always \(< 0.13\); SE of \(h²\) always \(< 0.05\)

CNV-trait associations. There were a total of 619 SNP where likelihood ratio tests indicated that correlations with traits were different from zero at \(P < 0.01\); at \(P < 0.001\) only 38 SNP were significantly correlated with traits. Of these, 11 SNP had significant correlations with traits at a chromosome-wide level using Bonferroni adjustment for multiple comparisons. Genome-wide significance was not achieved for any SNP.

Table 2 summarizes correlations between SNP and traits at different significance levels. All chromosomes had SNP that were nominally correlated with traits at a \(P < 0.01\) level. On BTA 1, 2, 7, 8, 9, 10, 11, 15, 17, 18, 19, 20, 23, and 28 there were SNP significantly correlated with traits at a \(P < 0.001\) level. On BTA 7, 10, 15, and 23 there were SNP correlated with traits at a chromosome-wide significance level. Correlations between SNP and traits varied, at \(P < 0.01\), maximum correlation with ADG was 0.36 on BTA11, and with MEI was 0.28 on BTA8. At \(P < 0.001\), the highest correlation with ADG was 0.31 on BTA8.

Table 2. Windows of SNP significantly correlated with traits, at different significance levels.

<table>
<thead>
<tr>
<th>Significance level</th>
<th>No.</th>
<th>Ave. h²</th>
<th>Max. cor</th>
<th>ADG</th>
<th>MEI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P &lt; 0.01)</td>
<td>619</td>
<td>0.30</td>
<td>0.37</td>
<td>0.37</td>
<td>0.41</td>
</tr>
<tr>
<td>(P &lt; 0.001)</td>
<td>181</td>
<td>0.37</td>
<td>0.31</td>
<td>0.31</td>
<td>0.28</td>
</tr>
</tbody>
</table>

1ADG = average daily gain; MEI = metabolizable energy intake

Discussion

CNV Detection. We found several regions in the genome where there is evidence of informative CNV. Several of these regions overlap with previously reported ones (e.g., Liu et al., 2010; Hou et al., 2011). Identifying these regions is valuable as it provides a better understanding of the cattle genome. There is evidence in many species that CNV are prevalent, and in humans CNV comprise approximately 12% of the genome (Wong et al., 2007). Our novel way of identifying important CNV regions in the cattle genome will contribute to the growing repository of such genic effects.

CNV-trait associations. Finding associations between feed efficiency traits and CNV proved difficult. Although there were associations between LRR and traits, most of these were not in CNV regions. Most of the associations we found did not pass a multiple comparisons test. We may need more power (more phenotypes) to be able to better map associations. Even with this, the results we found are valuable and provide guidelines of genomic regions that harbor CNV for feed efficiency.

Conclusion

The results of this study imply that there are many CNV regions in the cattle genome. These can be used in genome-wide association studies, and to study structural variants of the genome. Our novel way of identifying variants will add to our knowledge of polymorphisms in the cattle genome. The spectral decomposition of the genomic relationship matrix substantially reduces computation time, and can be applied to solve problems where large matrices would otherwise need to be inverted.

Mention of a trade name, proprietary product, or specific equipment does not constitute a guarantee or warranty by the USDA and does not imply approval to the exclusion of other products that may be suitable. USDA is an equal opportunity provider and employer.

Literature Cited