ABSTRACT: We sequenced the whole-genome of a Danish Jutland bull to identify genetic variants (SNP/indel). Using UnifiedGenotyper from the Genome Analysis Toolkit (GATK), we identified 6,812,198 SNPs and 804,453 indels. There were 2,598,000 (38.1%) novel SNPs and 607,923 (75.6%) novel indels while the remaining was annotated in dbSNP build 133. In-depth annotation of the variants revealed that 45,776 SNPs affected the coding sequences of 11,538 genes, 221 SNPs predicted to cause a premature stop codon, 17 to cause a gain in coding sequence and 20,828 predicted to be non-synonymous. We identified 1,122 indels in coding sequences, 832 predicted to cause frame shift, 89 synonymous. We identified 1,122 indels in coding sequences, 832 predicted to cause frame shift, 89 predicted to be inframe insertion and 115 to be inframe deletion. We detected a higher level of genetic variation in the Jutland bull compared to similar data from Holstein cattle.

Key words: cattle; whole-genome; SNP/indel; annotation

INTRODUCTION

Development of high-throughput sequencing platforms and sequence analysis tools facilitates whole-genome sequencing based variant identification. Over the last few years studies on whole-genome sequencing based identification of genetic variants (Van Tassell et al. (2008); Gibbs et al.(2009); Elsk et al. (2009); Stothard et al. (2011); Zhan et al. (2011); Jansen et al. (2013); Köks et al. (2013); Lee et al. (2013)) described large amount of SNPs across the genomes of modern cattle breeds. Results from these studies provide an insight into the amount of genetic variations segregating between breeds and create basis for the genome-wide association studies (GWAS) to know molecular mechanisms of traits variation and disease mechanisms (Huang et al. (2008); Khatib et al. (2008); Jiang et al. (2010)). Genetic diversity in modern breeds has been reduced due to low effective population sizes and force from selection over long period (Kantanen et al. (2000a)). It is assumed beneficial genetic variants that might have been lost as a result of selection in the modern breeds are still segregating in the purebred population of old breeds. Identification of genetic variants in old breeds could be an advantageous resource to restore favorable alleles underlying economically important traits and to correct inherited genetic defects in modern cattle breeds. Therefore the objective of this study is to identify and in-depth annotate genetic variants (SNP and indel) in a Danish Jutland bull.

MATERIALS AND METHODS

Animal sample. The Old Danish original Jutland cattle breed has been officially documented since 1882. It descended from local cattle herds with black-pied or dun-pied coat colour varieties. Although the first herdbook published in 1881 defined this breed as dual purpose cattle, in the early 1900s this breed was developed into a single purpose dairy breed. In the 1950s, this breed started to decline due to the introduction of imported breeding animals from the Netherlands and Germany. However, a few private breeders kept small herds with the original purebred animals. Since 1955, the Jutland breed officially developed into the black-pied breed called SDM in Danish or Holstein. The breed went through a population bottleneck, as there was a drastic decrease in population size before the conservation programme initiated in 1987. We used one of the seven bulls utilized in the initial conservation programme for this study.

Sequencing, mapping and variant calling. Whole blood was used to extract genomic DNA using commercially available QIAamp DNA Blood Maxi Kit (Qiagen). Paired-end libraries were prepared using genomic DNA according to manufacturer’s protocol (Illumina Inc. San Diego, CA, USA). DNA Sequencing was performed using an Illumina Hiseq 2000 with paired-end libraries to 2×101 bp. We used Burrows-Wheeler Aligner (BWA) (Li and Durbin (2009)) for mapping towards the Bos taurus genome assembly UMD 3.1 (Zimin et al. (2009)). SNP and indel calling were performed using UnifiedGenotyper from the Genome Analysis Toolkit v.2.4.7 (GATK) (DePristo et al. (2011)) with option “--min_base_quality_score 20” and keeping other parameters as default. SNPs and indels from dbSNP build 133 (Sayers et al. (2011)) were used as known sites.

Functional annotation of the variant. SNPs and indels were annotated using NGS-SNP (Grant et al. (2011)). NGS-SNP utilized Ensembl release 72 (Flieck et al. (2011)), dbSNP build 133 (Sayers et al. (2011)), Entrez Gene (Sayers et al. (2011)) and Uniprot (The Uniprot Consortium (2011)) as the source databases during annotation.

RESULTS AND DISCUSSION

Sequencing, mapping and variant identification. Sequencing generated 723,345,316 of raw reads. The numbers of mapped reads were 708,364,997. Mapped reads covered 98.9% of the reference genome with mean 26.4 fold coverage (Figure 1). The genome coverage and mean mapping depth in this study was rational for reliable variant identification (Eck et al.
presented in Table 2. The numbers of intergenic SNPs and indels in each functional class are shown in Figure 1. We identified 804,453 indels (-54 to +44bp); 405,972 were heterozygous and 2,470,687 (36.3%) homozygous SNPs. There were 221 SNPs predicted to cause premature stop loss, 38 of them in splice donor and 47 in splice acceptor sites. The numbers of indels in coding sequences were 1,122 (499 deletions and 623 insertions), 832 were predicted to cause frame shift, 89 were inframe insertions and 115 inframe deletions. Only one indel was detected to affect the first codon of a transcript while 25 were predicted to create amino acid changes in encoded protein without

Table 1. Summary statistics of the genomic variants (SNP/indel) identified in the Jutland bull.

<table>
<thead>
<tr>
<th>SNP/%</th>
<th>Indel/%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>6,812,198</td>
</tr>
<tr>
<td>Homozygous</td>
<td>2,470,687</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>4,341,511</td>
</tr>
<tr>
<td>Novel</td>
<td>2,598,000</td>
</tr>
<tr>
<td>Annotated in dbSNP(known)</td>
<td>4,214,198 (61.9%)</td>
</tr>
</tbody>
</table>

(3.5%) indels were 5kb upstream and 25,366 (3.2%) were downstream of a gene. There were 1,909 indels in 3' UTR and 205 in 5' UTR. We identified 648 splice site indels, 38 of them in splice donor and 47 in splice acceptor sites. The numbers of indels in coding sequences were 1,122 (499 deletions and 623 insertions), 832 were predicted to cause frame shift, 89 were inframe insertions and 115 inframe deletions. Only one indel was detected to affect the first codon of a transcript while 25 were predicted to create amino acid changes in encoded protein without

Table 2. Numbers of SNPs and indels in each functional class.

<table>
<thead>
<tr>
<th>Functional class</th>
<th>SNP (%)</th>
<th>Indel (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>intergenic_variant</td>
<td>4,639,873 (68.1)</td>
<td>539,745 (67.1)</td>
</tr>
<tr>
<td>intron_variant</td>
<td>1,676,710 (24.6)</td>
<td>207,372 (25.8)</td>
</tr>
<tr>
<td>upstream_gene_variant</td>
<td>230,365 (3.4)</td>
<td>27,829 (3.5)</td>
</tr>
<tr>
<td>downstream_gene_variant</td>
<td>197,827 (2.9)</td>
<td>25,366 (3.2)</td>
</tr>
<tr>
<td>3_prime_UTR_variant</td>
<td>12,428 (0.2)</td>
<td>1,909 (0.2)</td>
</tr>
<tr>
<td>5_prime_UTR_variant</td>
<td>2,613 (0.0)</td>
<td>205 (0.0)</td>
</tr>
<tr>
<td>splice_region_variant</td>
<td>4,072 (0.1)</td>
<td>563 (0.1)</td>
</tr>
<tr>
<td>splice_acceptor_variant</td>
<td>142 (0.0)</td>
<td>47 (0.0)</td>
</tr>
<tr>
<td>splice_donor_variant</td>
<td>142 (0.0)</td>
<td>38 (0.0)</td>
</tr>
<tr>
<td>stop_gained</td>
<td>221 (0.0)</td>
<td>-</td>
</tr>
<tr>
<td>stop_lost</td>
<td>17 (0.0)</td>
<td>-</td>
</tr>
<tr>
<td>frameshift_variant</td>
<td>-</td>
<td>852 (0.1)</td>
</tr>
<tr>
<td>initiator_codon_variant</td>
<td>45 (0.0)</td>
<td>1 (0.0)</td>
</tr>
<tr>
<td>frameshift_variant</td>
<td>-</td>
<td>89 (0.0)</td>
</tr>
<tr>
<td>frameshift_variant</td>
<td>-</td>
<td>115 (0.0)</td>
</tr>
<tr>
<td>missense_variant</td>
<td>20,783 (0.3)</td>
<td>25 (0.0)</td>
</tr>
<tr>
<td>stop_retained_variant</td>
<td>11 (0.0)</td>
<td>-</td>
</tr>
<tr>
<td>synonymous_variant</td>
<td>24,595 (0.4)</td>
<td>-</td>
</tr>
<tr>
<td>coding_sequence_variant</td>
<td>104 (0.0)</td>
<td>111 (0.0)</td>
</tr>
<tr>
<td>non_coding_exon_variant</td>
<td>2,209 (0.0)</td>
<td>149 (0.0)</td>
</tr>
<tr>
<td>nc_transcript_variant</td>
<td>13 (0.0)</td>
<td>15 (0.0)</td>
</tr>
<tr>
<td>mature_miRNA_variant</td>
<td>28 (0.0)</td>
<td>22 (0.0)</td>
</tr>
</tbody>
</table>
changing the frame. In total 186 indels were located in non-coding genes, 149 in non-coding exon, 22 in mature miRNAs and 15 in the transcript of non-coding RNA. Variants located in miRNAs might be associated with specific phenotypes of the animal as miRNAs have a regulatory role in post-transcriptional gene expression (Ramesh (2005)).

The higher number of genetic variants identified in this study could explain the anticipated differences between an old breed and a modern breed as the reference genome is based on the Hereford breed (Zimin et al. (2009)). The numbers of non-synonymous SNPs segregating in Jutland breed were higher than those identified in Danish Holstein (Zhan et al. (2011)) and North American Holstein (Stothard et al. (2011)). This finding is not surprising as 93% genes segregating in Danish Holstein have North American origin (Sorensen et al. (2005)). The Jutland breed was also found to be highly polymorphic in microsatellite data compared with Finnish Holstein-Frisian breed (Kantanen et al. (2000b)). The higher level of polymorphisms in the Jutland bull could be explained by this breed being bred more at random without selection and therefore maintaining a substantial part of the genetic variation despite the low population size since the 1950s. The modern day Holstein breed has been maintained with low effective population size (Sorensen et al. (2005)) and selection for specific breeding goals. Both possibilities can reduce genetic diversity in a population and make them genetically more distinct from its original population. Therefore rare alleles segregating in the Jutland breed might be lost from the modern Holstein population.

**Conclusion**

We identified a substantial number of novel genetic variations along with a high rate of non-synonymous exchange in the Jutland cattle genome. A comparison with similar data from Holstein cattle (Das et al. in prep.) showed a higher level of genetic variation in the Jutland bull than in the Holstein breed. This could be an effect of the high selection imposed on Holstein cattle or of the long period with a low effective population size of the Holstein breed causing a reduction in genetic variation. Valuable genetic variations conserved in this ancestral breed could be used for reintroduction back into the modern cattle breeds through genomic selection.

**Acknowledgments**

This research was financially supported by the Danish Research Council for Technology and Production, grant 0602-01729B. We thank the technicians Mette Jeppesen and Mahesha Perera for the work in the sequencing laboratory.

**LITERATURE CITED**


