Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Identification and annotation of genetic variants (SNP/Indel) in Danish Jutland cattle

A. Das, F. Panitz and L.-E. Holm
Department of Molecular Biology and Genetics, Aarhus University, Blichers Alle, Tjele, Denmark

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 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); (Elsik et al. (2009); Stothard et al. (2011); Zhan et al. (2011); Jansen et al. (2013); Kõks 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 (Flicek 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. (2009); Kawahara-Miki et al. (2011); Stothard et al. (2011); Zhan et al. (2011)). Using UnifiedGenotyper from GATK we identified 6,812,198 SNPs in the Jutland bull genome; 2,598,000 (38.1%) SNPs were novel and 4,214,198 (61.9%) were annotated in dbSNP build 133 (Sayers et al. (2011)). There were 4,341,511 (63.7%) heterozygous and 2,470,687 (36.3%) homozygous SNPs. We identified 804,453 indels (-54 to +44bp); 405,972 (50.5%) were deletions and 398,481 (49.5%) were insertions. The numbers of novel indels were 607,923 (75.6%) and 196,530 (24.4%) were annotated in dbSNP build 133; 434,506 (54.0%) were heterozygous and 369,947 (36.0%) were homozygous (Table 1).

Functional annotation of the variant. The numbers of SNPs and indels in each functional class are presented in Table 2. The numbers of intergenic SNPs were 4,639,873 (68.1%) and 1,676,710 (24.6%) were intronic. There were 230,365 (3.4%) SNPs located within 5kb upstream and 197,827 (2.9%) in downstream of a transcription start site; 12,428 SNPs were located in the 5' UTR and 2,613 in the 3' UTR. A total of 4,356 SNPs were located in splice sites of 2,966 genes: 142 were in splice-donor sites, 142 were splice-acceptor sites and 4072 were within the region of the splice site. We identified 45,776 SNPs affecting the coding sequences of 11,538 genes. There were 221 SNPs predicted to cause premature stop codon and 17 to cause gain in coding sequence. The numbers of SNPs predicted to be non-synonymous were 20,828. Of the non-synonymous 3,473 SNPs were predicted to have deleterious effects based on their SIFT score (Ng and Henikoff (2001)). The classes of SNPs for non-coding genes encompassed 2,209 non-coding exon, 28 miRNAs and 22 non-coding transcript variants. Intergenic (67.1%) and intronic (25.8%) indels represent the majority of the identified indels. A total of 27,829 (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), 852 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 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 supprising 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.

Acknowledgements

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

DePristo, M. A., Banks, E., Poplin, R. et al. (2011). Nat. Genet. 43(5): 491-498.

Eck, S., Benet-Pages, A., Flisikowski, K. et al. (2009). Genome Biol. 10: R82.

Elsik, C., Tellam, R., Worley, K. et al. (2009). Science. 324:522-528.

Flicek, P., Amode, M., Barrell, D. et al. (2011). Nucleic Acids Res. 39(suppl. 1):D800-D806.

Gibbs, R., Taylor, J., Van Tassell, C. et al. (2009). Science. 324:528-532.

Grant, J., Arantes, A., Liao, X., and Stothard, P. (2011). Bioinformatics. 27:2300-2301.

Huang, W., Maltecca, C., and Khatib, H. (2008). Anim. Genet. 39(5):554-557.
Jiang, L., Liu, J., Sun, D. et al. (2010). PLoS One. 5(10):e13661.
Jansen, S., Aigner, B., Pausch, H. et al.(2011). BMC Genomics. 14:446.
Kantanen, J., Olsaker, I., Holm, L. -E. et al. (2000a). J. Hered. 91(6):446-457.
Kantanen, J., Olsaker, I., Brusgaard, K. et al. (2000b). Genet. Sel. Evol. 32:561-576.
Kawahara-Miki, R., Tsuda, K., Shiwa, Y. et al. (2011). BMC Genomics. 12(1):103.
Khatib, H., Monson, R. L., Schutzkus, V. et al. (2008). J. Dairy Sci. 91(2):784-793.
Kõks, S., Lilleoja, R., Reimann, E. et al. (2013). Mamm. Genome. 24(7-8):309-321.
Lee, K.-T., Chung, W.-H., Lee, S.-Y. et al. (2013). BMC Genomics. 14(1):519.
Li, H., and Durbin, R. (2009). Bioinformatics. 25(14):1754-1760.Ng, P. C., and Henikoff, S. (2001). Genome Res. 1(5):863-874.
Ramesh, S. P. (2005). RNA. 11:1753-1761.
Sayers, E., Barrett, T., Benson, D. et al. (2011). Nucleic Acids Res. 39(suppl. 1):D38-D51.
Sorensen, A. C., Sorensen, M. K., and Berg, P. (2005). J. Dairy Sci. 88(5):1865-1872
Stothard, P., Choi, J., Basu, U. et al. (2011). BMC Genomics. 12:559.
The Uniprot Consortium. (2011). Nucleic Acids Res. 39(suppl. 1):D214-219.
Van Tassell, C., Smith, T., Matukumalli, L. et al. (2008). Nat. Methods. 5:247-252.
Zhan, B., Fadista, J., Thomsen, B. et al. (2011). BMC Genomics. 12(1):557.
Zimin, A., Delcher, A., Florea, L. et al. (2009). Genome Biol. 10(4):R42.

Figure 1: Plot for genome coverage by chromosome- horizontal axis shows chromosomes of the reference genome. Blue bars represent the size of the reference chromosome while the read bars indicate region of coverage. Left vertical axis indicates scale of the chromosome length while the right vertical axis shows the percentage scale of the coverage. Upper green line shows percentages of sequence coverage.

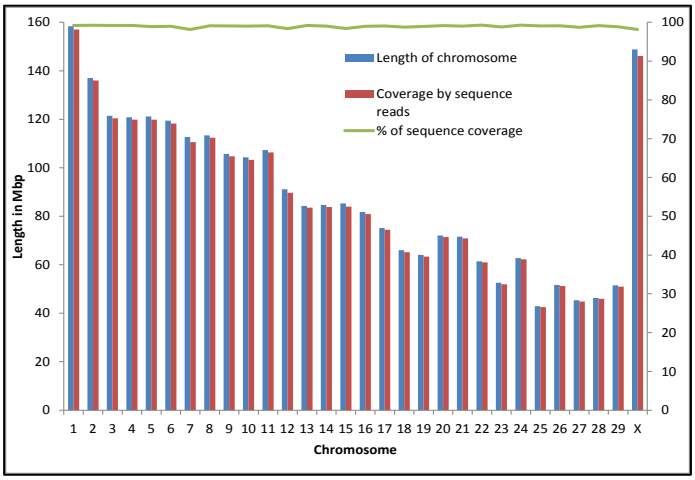

Table 1: Summary statistics of the genomic variants (SNP/indel) identified in the Jutland bull.

	SNP	Indel
Total	6,812,198	804,453
Homozygous	2,470,687 (36.3%)	369,947 (36.0%)
Heterozygous	4,341,511 (63.7%)	434,506 (54.0%)
Novel	2,598,000 (38.1%)	607,923 (75.6%)
Annotated in dbSNP(known)	4,214,198 (61.9%)	196,530 (24.4%)

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

Functional class	SNP (%)	Indel (%)
intergenic_variant	4,639,873 (68.1)	539,745 (67.1)
intron_variant	1,676,710 (24.6)	207,372 (25.8)
upstream_gene_variant	230,365 (3.4)	27,829 (3.5)
downstream_gene_variant	197,827 (2.9)	25,366 (3.2)
3_prime_UTR_variant	12,428 (0.2)	1,909 (0.2)
5_prime_UTR_variant	2,613 (0.0)	205 (0.0)
splice_region_variant	4,072 (0.1)	563 (0.1)
splice_acceptor_variant	142 (0.0)	47 (0.0)
splice_donor_variant	142 (0.0)	38 (0.0)
stop_gained	221 (0.0)	-
stop_lost	17 (0.0)	-
frameshift_variant	-	852 (0.1)
initiator_codon_variant	45 (0.0)	1 (0.0)
inframe_insertion	-	89 (0.0)
inframe_deletion	-	115 (0.0)
missense_variant	20,783 (0.3)	25 (0.0)
stop_retained_variant	11 (0.0)	-
synonymous_variant	24,595 (0.4)	-
coding_sequence_variant	104 (0.0)	111 (0.0)
non_coding_exon_variant	2,209 (0.0)	149 (0.0)
nc_transcript_variant	13 (0.0)	15 (0.0)
mature_miRNA_variant	28 (0.0)	22 (0.0)