J. Dairy Sci. 100:1–7
https://doi.org/10.3168/jds.2016-12199

Detection and assessment of copy number variation using PacBio long read and Illumina sequencing in New Zealand dairy cattle

C. Couldrey,*[1] M. Keehan,* T. Johnson,* K. Tiplady,* A. Winkelman,* M. D. Littlejohn,* A. Scott,* K. E. Kemper,† B. Hayes,‡ S. R. Davis,* and R. J. Spelman*
*Research and Development, Livestock Improvement Corporation, Hamilton, New Zealand 3240
†Institute for Molecular Bioscience, and
‡Centre for Animal Science, University of Queensland, St Lucia 4072, Queensland, Australia

ABSTRACT

Single nucleotide polymorphisms have been the DNA variant of choice for genomic prediction, largely because of the ease of single nucleotide polymorphism genotype collection. In contrast, structural variants (SV), which include copy number variants (CNV), translocations, insertions, and inversions, have eluded easy detection and characterization, particularly in nonhuman species. However, evidence increasingly shows that SV not only contribute a substantial proportion of genetic variation but also have significant influence on phenotypes. Here we present the discovery of CNV in a prominent New Zealand dairy bull using long-read PacBio (Pacific Biosciences, Menlo Park, CA) sequencing technology and the Sniffles SV discovery tool (version 0.0.1; https://github.com/fritzsedlazeck/Sniffles). The CNV identified from long reads were compared with CNV discovered in the same bull from Illumina sequencing using CNVnator (read depth–based tool; Illumina Inc., San Diego, CA) as a means of validation. Subsequently, further validation was undertaken using whole-genome Illumina sequencing of 556 cattle representing the wider New Zealand dairy cattle population. Very limited overlap was observed in CNV discovered from the 2 sequencing platforms, in part because of the differences in size of CNV detected. Only a few CNV were therefore able to be validated using this approach. However, the ability to use CNVnator to genotype the 557 cattle for copy number across all regions identified as putative CNV allowed a genome-wide assessment of transmission level of copy number based on pedigree. The more highly transmissible a putative CNV region was observed to be, the more likely the distribution of copy number was multimodal across the 557 sequenced animals. Furthermore, visual assessment of highly transmissible CNV regions provided evidence supporting the presence of CNV across the sequenced animals. This transmission-based approach was able to confirm a subset of CNV that segregates in the New Zealand dairy cattle population. Genome-wide identification and validation of CNV is an important step toward their inclusion in genomic selection strategies.

Key words: structural variant, copy number variant, cattle, genomic selection, genome

INTRODUCTION

The introduction of genomic selection to dairy cattle breeding has increased the rate of genetic gain. To date, genomic selection has largely focused on the utilization of SNP and very small insertions or deletions (indels). Very little, if any, regard has been given to larger variations such as copy number variations (**CNV**) and other structural variations (**SV**), including deletions, insertions, and duplications. Although CNV and SV account for the greatest amount of total polymorphic content among individual genomes (Weischenfeldt et al., 2013), the focus on SNP and small indels presumably is attributable to the ease with which hundreds of thousands of SNP and indels can be simultaneously genotyped at a minimal cost. However, advances in genomic technologies—initially with microarrays and up to the more recent use of next-generation sequencing—are resulting in an increasing amount of evidence indicating that these larger sequence variations make important contributions to genetic and phenotypic variation (Iafrate et al., 2004; Sebat et al., 2004; Weischenfeldt et al., 2013; MacDonald et al., 2014; Sudmant et al., 2015; Zarrei et al., 2015). No single technology, detection strategy, or algorithm can capture the entire spectrum of SV in the genome. For example, next-generation sequencing technologies typically allow identification of smaller variants of all SV classes, whereas microarray-based platforms (particularly those used primarily

Received October 24, 2016.
Accepted March 12, 2017.
[1]Corresponding author: Christine.couldrey@lic.co.nz.

for SNP genotyping) are limited to the detection of larger deletions and duplications. Similarly, although many different algorithms have been developed for SV identification, there is only limited overlap in the sites discovered, even when multiple algorithms are used to analyze a single data set (Duan et al., 2013). The human 1000 Genomes Project has used a variety of SV detection platforms and detection algorithms for each platform to generate an integrated map of 68,818 SV in unrelated individuals (Sudmant et al., 2015). This map is now considered the gold standard SV list in humans, yet the authors still state that "SV discovery remains a challenge nonetheless, and the full complexity and spectrum of SV is not yet understood" (Sudmant et al., 2015).

The desire to have a comprehensive list of SV in a population is not unique to human genomics. Attempts have been made to catalog CNV, but not SV, in a wide variety of species, including cattle (Xu et al., 2014), pigs (Jiang et al., 2014), and sheep (Xu et al., 2014; Jenkins et al., 2016). However, CNV and SV detection is critically dependent on the quality of genome assembly, which for species such as cattle lags behind the quality of that for the human genome. Furthermore, although CNV and SV detection algorithms invariably report the presence of large numbers of CNV and SV in each individual, these detection algorithms are plagued with a high rate of false discovery. Without a gold standard with which to compare detected variants, case-by-case validation is a lengthy process and not suited for genome-wide analysis.

The underlying driving force in investigating the human genome has often been the desire to understand disease phenotypes. Conversely, in commercial production animals such as cattle, investigation of the genome is driven by production traits and the desire to predict animal performance at an early age through genomic selection. As widespread genotyping and imputation of genotypes to sequence level (Druet et al., 2014) become more common, there exists an increasing need to capture not only SNP variation but also CNV and SV, which may severely affect imputation (K. Tiplady, unpublished data) and be associated with or contribute to important production trait phenotypes (Kadri et al., 2014; Xu et al., 2014).

The recent availability of long-read single-molecule sequencing has provided another technology for the identification of SV and CNV. Given the long (up to 80 kb) sequence reads that can be achieved, this technology offers the possibility of single reads that span complex SV and actively assess SV in a repetitive region (Sedlazeck et al., 2015). In this study we used long-read single-molecule sequencing of a New Zealand Holstein-Friesian bull together with Illumina short-read sequencing of animals representative of the genes present in New Zealand dairy genetics to begin identifying and characterizing CNV with the vision of improving imputation and, ultimately, genomic selection and association studies.

MATERIALS AND METHODS

PacBio

Sequence. PacBio (Pacific Biosciences, Menlo Park, CA) long-read sequences were generated from a Holstein-Friesian bull (41% New Zealand Holstein-Friesian genetics) by Cold Spring Harbor Laboratories (Cold Spring Harbor, NY). Genomic DNA was fragmented to an average of 10 kb with a g-tube (Covaris, Woburn, MA). The fragmented DNA was then repaired and ligated to SMRTbell adapters (Pacific Biosciences) following the manufacturer's instructions. The ligated library was size selected for fragments >10 kb using the BluePippin (Sage Science, Beverly, MA). The size-selected library was annealed to SMRTbell primers (Pacific Biosciences), bound to the P6 polymerase, and sequenced on the Pacific Biosciences RS II instrument following the manufacturer's instructions. The PacBio SMRT pipeline was used to generate filtered subreads in fastq format. Alignment of subreads to the UMD 3.1 bovine genome assembly (http://bovinegenome.org/?q=node/61) was undertaken using BWA-MEM (version 0.7.12; https://arxiv.org/abs/1303.3997) with options "–M –x pacbio."

SV Detection. The BAM files from BWA-MEM mapping of PacBio sequences were sorted, and SV were called using Sniffles (version 0.0.1; https://github.com/fritzsedlazeck/Sniffles). Structural variants that displayed >95% reciprocal overlap with a UMD 3.1 contig were removed because these likely represent genome assembly errors. In an attempt to reduce the number of false-positive SV identified, further filtering was undertaken to retain only those SV present in a single contig.

Illumina

Sequence. Illumina HiSeq sequencing of 556 animals, including 25 trios and 395 duos, has previously been described (Littlejohn et al., 2016). Briefly, 100-bp paired-end sequencing was performed (Illumina HiSeq2000; Illumina Inc., San Diego, CA) on 137 Holstein-Friesians, 100 Jerseys, 318 Holstein-Friesian × Jersey crossbreeds, and 1 Ayrshire representing the population structure of New Zealand dairy cattle and phenotypes of interest. Additionally, DNA from a Holstein-Friesian bull