Proc. Assoc. Advmt. Anim. Breed. Genet. 22:197-200

DETECTION AND ASSESSMENT OF COPY NUMBER VARIATION USING PACBIO LONG READ SEQUENCING IN NEW ZEALAND DAIRY CATTLE

C. Couldrey[1], M. Keehan[1], T. Johnson[1], K. Tiplady[1], A. Winkelman[1], C. Thresher[1], M. D. Littlejohn[1], A. Scott[1], K. E. Kemper[2], B. Hayes[3], S.R. Davis[1] and R.J. Spelman[1]

[1]Research and Development, Livestock Improvement Corporation, Hamilton, New Zealand
[2] Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland, Australia
[3] Centre for Animal Science, University of Queensland, St Lucia, Queensland, Australia

SUMMARY

Structural variants (SVs) have eluded easy detection and characterisation, particularly in non-human species. However, there is increasing evidence that SVs not only contribute a substantial proportion of genetic variation but have significant influence on phenotypes. Here we present discovery of copy number variants (CNVs) (a subset of SVs) in a prominent New Zealand dairy bull using long read PacBio sequencing technology. Validation of CNVs was undertaken utilising whole genome Illumina sequencing of 557 cattle representing the wider New Zealand dairy cattle population. The ability to utilise CNVnator to "genotype" the 557 cattle for copy number across all regions identified as putative CNVs, 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 multi-modal across the 557 sequenced animals. This transmission based approach was able to confirm a subset of CNVs that segregates in the New Zealand dairy cattle population. Genome-wide identification and validation of CNVs is an important step towards their inclusion into genomic selection strategies.

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 utilisation of SNPs and very small insertions or deletions. Very little regard has been given to larger variations such as CNVs. While SVs (including CNVs) account for the greatest amount of total polymorphic content among individual genomes (Weischenfeldt et al. 2013), the focus on SNPs and small indels is presumably due to the ease with which such variation can be genotyped at a minimal cost. However, advances in genomic technologies are resulting in an increasing amount of evidence indicating that these larger sequence variations make important contributions to genetic and phenotypic variation (MacDonald et al. 2014, Zarrei et al. 2015, Sudmant et al. 2015, Weischenfeldt et al. 2013). No single technology, detection strategy, or algorithm can capture the entire spectrum of SVs in the genome. The collective effort of the human 1000 Genomes Project has utilised both a variety of SV detection platforms and algorithms to generate an integrated map of 68,818 SVs in unrelated individuals (Sudmant et al. 2015). This 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 SVs in a population is not unique to human genomics, however, SV detection is critically dependent on the quality of genome assembly, which for species such as cattle, lags behind the quality of the human genome. Furthermore, while SV/CNV detection algorithms invariably report the presence of large numbers of CNVs 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.

In animals such as cattle, a desire to understand 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, Macleod, and Hayes 2014) becomes more common, there is an increasing need to not only capture SNP variation, as CNVs may severely impact imputation (LIC unpublished data), and also 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 (up to 80 kilobases (kb)) provides a new technology for the identifications of CNVs. This technology offers the possibility of single reads that span complex CNVs (Sedlazeck et al. 2015). We have utilised long read single molecule sequencing of a New Zealand Holstein Friesian bull with the vision of improving imputation and ultimately genomic selection and association studies.

MATERIALS AND METHODS

PacBio Sequence and SV Detection: PacBio long read sequences were generated from a Zealand Holstein-Friesian bull by Cold Spring Harbor Laboratories. The PacBio SMRT pipeline was used to generate filtered sub reads in fastq format. Alignment of reads to the UMD 3.1 bovine genome assembly was undertaken using BWA-MEM (v0.7.12; https://arxiv.org/abs/1303.3997) with options "–M –x pacbio".

SVs were called using Sniffles (v0.0.1 https://github.com/fritzsedlazeck/Sniffles). Structural variants displaying > 95% reciprocal overlap with a UMD3.1 contig were removed as these likely represent genome assembly errors. Further filtering retained only SVs present in a single contig.

Illumina Sequence and CNV Genotyping: Illumina HiSeq sequencing of 557 animals representing the population structure of New Zealand dairy cattle and phenotypes of interest has previously been described (Littlejohn et al. 2016). Read-depth-based CNV genotyping analysis was undertaken across the genome of animals sequenced on the Illumina HiSeq platform using CNVnator v0.3 (Abyzov et al. 2011) using a bin size of 150bp. Based on breakpoints identified by Sniffles, copy number was determined for each CNV greater than 100bp in length in each of 557 animals. Mendelian inheritance of copy number was assessed using a mixed linear model. The independent variables were the fixed effect of the mean and the random effect of the animal. The dependent variable was the copy number. The variance of the additive genetic effect of animal was based on a pedigree of each animal and their sire and dam, traced for seven generations. ASREML-r (version 3.0) (Gilmour et al. 2009) was used for estimation of variance components. The variance associated with the animal effect is analogous to the additive genetic variance and heritability is additive genetic variation/phenotypic variation, however, in terms of CNV inheritance, "transmission level" is used instead of the term heritability. A transmission level of 0 indicates either a denovo mutation in Esteem, or a sequencing artefact, or alternatively a transmission level of 1 indicates that the copy number is inherited in a Mendelian fashion.

Effect of CNVs on phasing allelic R^2:Using sequence data from all 557 animals, phasing allelic R^2 (AR^2) was determined for, each SNP within the 936 CNV regions found to have high transmission levels, each SNP outside the CNVs, and each SNP 50, 100, 500, 1000, 3000bp either side of the CNVs.

SNP tagging of CNVs: Correlations between copy numbers for each for the final 2661 CNVs and genotypes from 50K Illumina SNPchip or full sequence were determined.

RESULTS AND DISCUSSION

A total of 32x PacBio coverage of the bovine genome (UMD3.1) was generated. Sniffles software identified a total of 38,709 putative SVs of which 19,797 were CNVs (deletions

n=18,577, duplications n=1220). Of the 3532 CNVs (deletions n=3055, duplications n=477), that remained after filtering, sizes ranged from 1 – 79,450bp with a median size of 321bp (mean size of 818bp). CNVs smaller than 100bp (n=869) were excluded from further analysis as copy number could not be accurately predicted by CNVnator (2661 CNVs remained).

Using CNVnator in genotyping mode we were able to determine copy number at all putative CNV locations identified by PacBio sequence in all 556 animals. These CNV genotypes were used as 'phenotypes' in order to allow the copy number transmission level to be estimated using ASREML in an attempt to make a distinction between real CNVs and the many false positives detected when calling CNVs from short read sequencing. Putative CNVs showed a wide range of transmission levels. Approximately 30% of CNVs called from PacBio sequence showed high transmission level (936 CNVs > 0.70). Sorting CNVs by level of transmission and plotting distribution of copy number in the population indicated a trend of increasing multimodality of copy number with increasing transmission level. Many of the CNVs with a calculated transmission level of greater than 0.6 showed a clear bi- or trimodal distribution of copy number across the 557 animals. The multimodality of copy number, together with visual observation of bam files containing sequencing read-depths, insert size, and the presence of split reads are all consistent with the detection of bona fide CNVs, provided strong evidence that these highly tranmissible CNVs were likely to be present in our population. The observation that many of these trimodal distributions represented deletions (0, 1, vs 2 copies) reflects, at least in part, the relative ease with which deletions are able to be detected relative to duplications, due to the large proportional differences in sequence content for deletions (Abyzov et al. 2011).

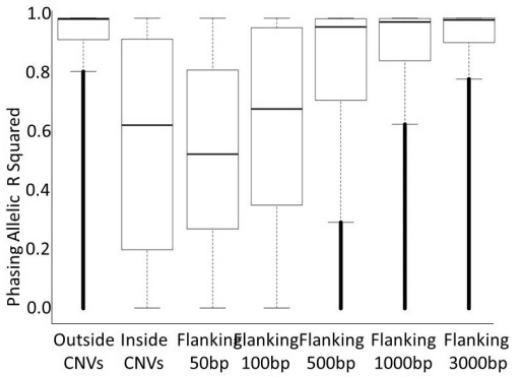

Figure 1 Phasing allelic R2 for SNPs outside, within, and flanking 936 CNVs with high transmission levels

Figure 1 illustrates the detrimental effect of CNVs on the ability to correctly phase SNP genotypes, not only within the CNV itself, but also in the surrounding sequence. Given the vast number of CNVs even in this one individual, it is expected that accuracy of imputation will be negatively affected by the inability to phase the reference sequence accurately.

While the data presented here is not a comprehensive list of CNVs in the New Zealand dairy cattle population, it does illustrate the potential of long read single molecule sequencing as an additional valuable source for identification of CNVs. Furthermore, long read sequence information, combined with independent short read sequencing and pedigree information in 557 animals representative of the population provide compelling evidence of the existence of CNVs in our dairy cattle population, and are not simply false positive results and allows us to begin a catalogue of CNVs. Characterisation of population CNVs has two major benefits to the cattle breeding industry. Firstly, once identified, CNVs may be cheaply identified alongside SNPs by simply adding appropriately designed probes to existing SNP chip genotyping platforms and including CNV genotype information as an additional source of genetic variation in genomic prediction models. Secondly undertaking imputation in a CNV aware manner to bypass poor phasing and increase imputation accuracy.

It could be argued that much of the CNV variation is already captured by SNP in linkage disequilibrium with CNVs. However, it is unlikely that multi-allelic CNVs would be accurately tagged using bi-allelic SNP, and initial reports indicate that around 20% of large CNVs identified from SNP chip platforms are not well tagged (Xu et al. 2014). Figure 2 illustrates the correlation

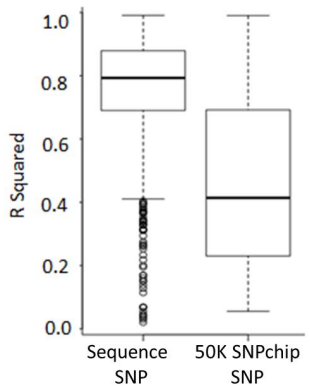

Figure 2 Correlation between copy number and genotype from sequence and 50K SNPchip for 936 CNVs with high transmission levels

between CNV and SNP genotypes on the 50K SNPchip as well as from sequence. Our results indicate that very few of the 936 highly transmissible CNVs are tagged well by SNP on the 50K SNPchip, and unsurprisingly many more CNVs are well tagged by sequence derived SNPs. Our current genomic selection protocols utilise only SNPs present on the 50K SNPchip, and therefore, to date, only a very limited amount of genetic variation from CNVs is being captured and utilised. As a move towards including sequence derived SNPs that tag CNVs could help in improving the accuracy of genomic selection

From a practical perspective, the presence of CNVs may have implications for phasing and imputation of other classes of variants. Given the increasing use of imputation of SNP chip genotypes to whole genome sequence, understanding where CNVs are located in the genome and ideally devising strategies for their correct imputation are of great importance for accurate genome-wide imputation and the generation of accurate genotype information to be utilised in genomic prediction models.

CONCLUSION

We present here the first step towards a gold standard list of CNVs in dairy cattle by utilising both long and short read sequencing technologies together with conservative filtering steps and an easy genome-wide strategy for assessing the Mendelian inheritance. Collectively this provided compelling evidence that these SVs do segregate in the population. Given the increasing use of imputation strategies being used in cattle breeding, identification and characterisation of CNVs (and all classes of SVs) will lead to improved imputation accuracy and will ultimately contribute to improved genomic prediction.

REFERENCES

Abyzov, A., A. E. Urban, M. Snyder, and M. Gerstein. (2011)Genome Res 21 (6):974-84.
Druet, T., I. M. Macleod, and B. J. Hayes. (2014). Heredity (Edinb) 112 (1):39-47.
Duan, J., J. G. Zhang, H. W. Deng, and Y. P. Wang. (2013) PLoS One 8 (3):e59128.
Gilmour, A. R., B. R. Gogel, B.R. Cullis, and R Thompson. (2009). ASREML User Guide Release 3.0.: VSN International Ltd, Hemel Hempstead, UK.
Kadri, N. K., G. Sahana, C. Charlier, T. Iso-Touru, B. Guldbrandtsen, *et al.*. (2014) PLoS Genet 10 (1):e1004049.
Littlejohn, M. D., K Tiplady, T. A. Fink, K. Lehnert, T. Lopdell, *et al.* (2016) Scientific Reports 6.
MacDonald, J. R., R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer. (2014). Nucleic Acids Res 42 (Database issue):D986-92.
Sedlazeck, F.J., P Rescheneder, M Nattestad, and M.C. Schatz. (2015) Genome Informatics, Cold Spring Harbor, New York.
Sudmant, P. H. and 1000 Genomes Project Consortium. (2015) " Nature 526 (7571):75-81.
Weischenfeldt, J., O. Symmons, F. Spitz, and J. O. Korbel. (2013) Nat Rev Genet 14 (2):125-38.
Xu, L., J. B. Cole, D. M. Bickhart, Y. Hou, J. Song, *et al.* (2014) BMC Genomics 15:683.
Zarrei, M., J. R. MacDonald, D. Merico, and S. W. Scherer. (2015Nat Rev Genet 16 (3):172-83.
Zhang, X., R. Du, S. Li, F. Zhang, L. Jin, and H. Wang. (2014) BMC Bioinformatics 15:50.