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.

¹Corresponding 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

(“Esteem”) chosen for PacBio long-read sequencing was sequenced using Illumina chemistry.

SV Detection. Read depth–based CNV analysis was undertaken across the genomes of the 557 animals sequenced on the Illumina HiSeq platform using CNVnator version 0.3 (Abyzov et al., 2011). A bin size of 150 bp was selected based on the average read depth and read depth distribution of the sequenced samples. 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 CNV identified, further filtering was undertaken to retain only those CNV present in a single contig.

CNV Genotyping. Based on breakpoints identified by Sniffles (for PacBio sequencing) and CNVnator (for Illumina sequencing) in Esteem’s DNA, copy number was determined for each CNV greater than 100 bp in length (PacBio, n = 2,661; Illumina, n = 31,510) in each of 556 animals. Frequency of assigned copy number for regions overlapping genes and exons was assessed visually for multimodality. Mendelian inheritance of copy number was assessed by estimating the heritability of a copy number 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 7 generations. The estimation of variance components was done using ASREML-r (version 3.0; Gilmour et al., 2009). Use of the term *heritability* for transmission of CNV from parents to offspring deviates from its common use; however, the variance associated with the animal effect is analogous to the additive genetic variance, and heritability is additive genetic variation and phenotypic variation. Rather than *heritability*, the term *transmission level* has been used throughout this study. A transmission level of 0 indicates either a de novo mutation in Esteem or a sequencing artifact, and a transmission level of 1 indicates that the copy number is inherited in a Mendelian fashion.

RESULTS

PacBio

A total of 10,694,517 PacBio reads (85,464,862,655 bp) were generated from 204 SMRT cells representing 32× coverage of the bovine genome. The distribution of sequence length (496–52,304 bp) is illustrated in Supplemental Figure S1 (<https://doi.org/10.3168/jds.2016-12199>) with an average length of 7,968 bp; 7 of the sequences are greater than 50,000 bp in length. The BWA-MEM mapping run using the PacBio option resulted in 93.2% of the PacBio reads mapping to the UMB 3.1 bovine assembly and 99.4% of nucleotides covered by at least 1 read. Sequence coverage distribution is illustrated in Supplemental Figure S2 (<https://doi.org/10.3168/jds.2016-12199>).

Sniffles software identified a total of 38,709 putative SV, of which 19,797 were CNV (deletions, n = 18,577; duplications, n = 1,220) in a PacBio sequence generated from Esteem’s DNA. Of the 3,532 CNV (deletions, n = 3,055; duplications, n = 477) that remained after filtering, 899 (deletions, n = 740; duplications, n = 159) and 126 (deletions, n = 76; duplications, n = 50) overlapped by at least 1 bp with Ensembl (build 84; <http://mar2016.archive.ensembl.org/index.html>) annotated genes and exons, respectively (Table 1).

The filtered CNV sizes (Figure 1) identified by Sniffles from Esteem’s PacBio data ranged from 1 to 79,450 bp and had a median size of 321 bp (mean size = 818 bp). The CNV smaller than 100 bp (n = 869) were excluded from further analysis because copy number could not be accurately predicted by CNVnator for CNV of this size. A total of 2,661 CNV remained after this final filtering step.

Illumina

Illumina HiSeq sequencing of Esteem generated 261,839,475 sequences (32,729,934,375 bp) representing approximately 12× coverage of his genome. The BWA-MEM mapping resulted in 99.06% of the reads

Table 1. Number of copy number variants (CNV) called in the Holstein-Friesian bull Esteem from Illumina and PacBio sequencing platforms from an equivalent amount of sequence

Item	~30× Illumina ¹	~30× PacBio ²	Overlap
Total CNV	43,708	19,799	ND ³
CNV in a single UMD 3.1 contig	31,510	3,532	650 (18%)
Overlap with annotated gene	9,781	899	198 (22%)
Overlap with annotated exon	6,277	126	27 (21%)

¹Illumina Inc., San Diego, CA.

²Pacific Biosciences, Menlo Park, CA.

³ND = not determined.

mapping to the UMB 3.1 bovine assembly, giving a depth coverage distribution (Supplemental Figure S3; <https://doi.org/10.3168/jds.2016-12199>) similar to 556 other individuals also sequenced on the Illumina HiSeq platform.

CNVnator software identified a total of 43,708 putative CNV (deletions, $n = 37,376$; duplications, $n = 6,332$) in the Illumina sequence generated from Esteem's DNA. Of these CNV, 31,510 (deletions, $n = 26,644$; duplications, $n = 4,866$) were present in a single UMB 3.1 contig, with 9,781 (deletions, $n = 8,219$; duplications, $n = 1,562$) and 6,277 (deletions, $n = 5,337$; duplications, $n = 940$) overlapping by at least 1 bp with Ensembl (build 84) annotated genes and exons, respectively.

The CNV sizes (Figure 1) identified by CNVnator from Esteem's Illumina data that were contained within a single contig ranged from 900 to 135,400 bp and had a median size of 8,999 bp (mean size = 10,670 bp). The size range of CNV detected in Esteem was contained within the size range observed in all 1,489,819 CNV (contained within a single contig) cumulatively detected in the other 556 animals sequenced (300–357,450 bp; median size = 3,450 bp; mean size = 6,070 bp).

PacBio–Illumina Overlap. Of the 3,532 PacBio CNV located within a single contig, 650 overlapped with CNV identified from Illumina sequence data from the same bull. Of these 650 CNV, 198 and 27 CNV

overlapped with annotated genes and exons, respectively.

CNV Transmission Level. Putative CNV from the PacBio and Illumina sequences from Esteem showed a wide range of transmission levels (Figure 2). The distribution of transmission levels was different for the 2 sequencing and CNV-calling technologies. A greater proportion of putative CNV called from the PacBio sequence showed high transmission levels (30% >0.75) compared with read depth-based CNV called from Illumina sequencing (3% >0.75). However, given the greater number of putative CNV called from Illumina sequencing, the number of CNV with transmission levels greater than 0.75 was similar (792 PacBio vs. 856 Illumina). A greater proportion of the CNV that were identified in both the PacBio and Illumina sequences displayed high transmission level (44% >0.75).

Visual assessment of histograms displaying the frequency of copy number at the 2,661 PacBio CNV suggested that CNV with higher transmission levels were more likely to follow a multimodal distribution. This multimodal distribution was indicative of distinct copy numbers occurring in the population of sequenced animals (Supplemental Figure S4; <https://doi.org/10.3168/jds.2016-12199>), with the majority of CNV appearing to be biallelic (with 0, 1, or 2 copies indicating deletions) in nature (Figure 3). A visual assessment of BAM files at regions with copy number transmission levels greater than 0.95 revealed clear evidence for the presence of CNV based on combinations of read depth, the insert size, and the presence of split reads.

DISCUSSION

The ability to generate sequencing reads spanning tens of thousands of base pairs from a single molecule using long-read technologies such as PacBio is proving to be invaluable in closing gaps in current sequence references (English et al., 2012; Rhoads and Au, 2015) and for the assembly of complex genomic regions (Rhoads and Au, 2015; English et al., 2015). The ability of this long-read technology to sequence through extended repetitive regions makes it a desirable platform to use for the identification of CNV. We used a combination of PacBio and Illumina sequencing of a prominent New Zealand dairy bull to characterize putative CNV and assess them in a group of animals representative of the wider New Zealand dairy cattle population.

PacBio long-read sequencing identified fewer potential CNV than Illumina sequencing in the genome of Esteem. Moreover, the potential CNV identified from PacBio sequencing were smaller than those identified from Illumina sequencing; many were in the size range considered to be indels (<50 bp). The smaller size

Figure 1. Comparison of copy number variants (CNV) called in the Holstein-Friesian bull Esteem from PacBio (Sniffles; Pacific Biosciences, Menlo Park, CA) and Illumina (Illumina Inc., San Diego, CA) sequences using the read depth-based CNV caller CNVnator. Size distribution of 3,532 and 31,510 putative CNV discovered from PacBio and Illumina sequences are illustrated in the box and whisker plot. The upper, middle, and lower lines in each plot define the upper quartile, the medium, and lower quartiles, respectively. Sizes that fall outside this range are illustrated by circles.

Figure 2. Transmission level of putative copy number variants (CNV) discovered in the Holstein-Friesian bull Esteem. Putative CNV greater than 100 bp discovered from PacBio (Pacific Biosciences, Menlo Park, CA) sequencing (top) and putative CNV discovered from Illumina (Illumina Inc., San Diego, CA) sequencing (bottom), genotyped for copy number in 556 dairy cattle sequenced by Illumina CNVnator.

of the PacBio-derived CNV and the limited overlap between positions of CNV from PacBio and Illumina sequences suggest an inherent difference in the scope of detection between the 2 sequencing technologies. Such an inherent difference is not unexpected given the wide range of nonoverlapping CNV identified from a single data set using a variety of detection algorithms (Duan et al., 2013). The detection of a subset of smaller CNV spans the previous gap between indels (up to 50 bp)

and CNV easily identified by next-generation sequencing [dependent on depth and breadth of coverage but generally >100 bp (Abyzov et al., 2011), with CNV identified from genotyping SNP chip platforms being larger still (tens of thousands of bp, depending on the array used; Zhang et al., 2014)].

Despite the difference in size of CNV detected from the different sequencing platforms, using CNVnator in genotyping mode and a whole-genome Illumina sequence from 556 animals we were able to determine the copy number at all putative CNV locations identified in Esteem's sequence by both CNVnator and Sniffles SV detection software. These CNV genotypes were used as phenotypes to allow the transmission levels of the copy number to be estimated using ASreml in an attempt to make a distinction between real CNV and the many false positives detected when calling CNV from short-read sequencing. In theory, for a biallelic CNV, all positions should have a transmission level of 0 (de-novo CNV or sequencing artifact) or 1 (copy number transmitted in a Mendelian fashion). In practice, we observed transmission levels across the entire spectrum 0 to 1 for both Sniffles- and CNVnator-discovered CNV (Figure 2). This spread, rather than 2 distinct groups around 0 and 1, is likely to arise from at least 3 contributing factors. The first factor is CNV located on chromosome X due to the difference in inheritance of this chromosome in males and females in this mixed-sex population (maximum transmission level would be

Figure 3. Distribution of copy number of a copy number variant (CNV) in dairy cattle. An example of a CNV identified by Sniffles (version 0.0.1; <https://github.com/fritzsedlazeck/Sniffles>) and genotyped for copy number in 556 dairy cattle where pedigree was found to account for all the variation in copy number (transmission level = 1).

~0.5 in this population; some examples are highlighted in Supplemental Figure S4, <https://doi.org/10.3168/jds.2016-12199>). Second, the presence of multiallelic CNV and a lack of information of how many copies each parent contributes will cause transmission levels to be skewed from 0 and 1. Finally, and perhaps the largest contributor to the observed distribution of transmission levels, is the variability in sequencing depth encountered in real data and hence difficulties in obtaining accurate copy numbers for some regions of the genome. Despite this latter challenge, highly transmissible CNV were observed from both sequencing technologies. The difference in proportion of CNV with transmission levels greater than 0.75 between CNVnator- and Sniffles-discovered CNV suggests that Sniffles discovers a higher proportion of true CNV from long-read sequences compared with CNVnator and short sequences.

Sorting Sniffles-discovered CNV by level of transmission (Supplemental Figure S4; <https://doi.org/10.3168/jds.2016-12199>) and plotting distribution of copy number in the population indicated a trend of increasing multimodality of copy number with increasing level of transmission. With the exception of those CNV discovered on chromosome X, many of the CNV (in a single contig and >100 bp in size) 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, sequencing insert size, and split reads are consistent with the detection of bona fide CNV; this provided strong evidence that these highly transmissible CNV were likely to be present in our population. The observation that many of these bi- and 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, simply because of the large proportional differences in sequence content for deletions compared with duplications (Abyzov et al., 2011).

The considerable cost of undertaking PacBio long-read sequencing has initially restricted the use of this technology to a single bull. As other technologies (e.g., 10x Genomics) that allow long-range genomic information to be obtained become available, similar analyses will be able to be undertaken on a larger group of animals. In the interim, although the data here do not present a comprehensive list of CNV in the New Zealand dairy cattle population, they do illustrate the potential of long-read single-molecule sequencing as an additional valuable source for identification of CNV. Furthermore, long-read sequence information, combined with independent short-read sequencing and pedigree

information in 556 animals representative of the population, provide compelling evidence of the existence of CNV in our dairy cattle population, are not simply false-positive results, and allow us to begin a catalog of CNV. Characterization of population CNV has 2 major benefits for the cattle breeding industry. First, once identified, CNV may be inexpensively identified alongside SNP simply by adding appropriately designed probes to existing SNP chip genotyping platforms and by including CNV genotype information as an additional source of genetic variation in genomic prediction models. It could be argued that much of this variation is already captured by SNP in linkage disequilibrium with CNV; however, it is unlikely that multiallelic CNV would be accurately tagged using biallelic SNP, and initial reports indicate that around 20% of large CNV identified from SNP chip platforms are not well tagged (Xu et al., 2014). Given that CNV affect large segments of the genome, they also represent a class of variation much more likely to affect phenotypes, and CNV that coincide with genic regions are likely to have profound functional consequences. From a practical perspective, the presence of (unobserved) CNV 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 sequences, understanding where CNV 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 used in genomic prediction models.

CONCLUSIONS

The previously documented challenges involved in CNV detection from next-generation sequencing data were also evident in this study. Despite these challenges, we present a first step toward a gold standard list of CNV in dairy cattle by utilizing both long- and short-read sequencing technologies for CNV identification together with conservative filtering and an easy genome-wide strategy for assessing the transmission of these CNV. Collectively, this methodology provided compelling evidence for determining which SV were segregating in the population. Given the increasing use of imputation strategies in cattle breeding, identification and characterization of CNV (and all classes of SV) will lead to improved imputation accuracy and will ultimately contribute to improved genomic prediction.

ACKNOWLEDGMENTS

We acknowledge New Zealand Genomics Ltd. (Dunedin) and the University of Auckland Centre for Genom-

ics, Proteomics, and Metabolomics (Auckland, New Zealand) for genomic DNA extraction and the Australian Genomics Research Facility (Melbourne) and Illumina FastTrack (San Diego, CA) for genomic DNA sequencing. Funding for this project was provided by the Ministry for Primary Industries (Wellington, New Zealand) as a Primary Growth Partnership.

REFERENCES

- Abyzov, A., A. E. Urban, M. Snyder, and M. Gerstein. 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21:974–984. <https://doi.org/10.1101/gr.114876.110>.
- Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity* (Edinb.) 112:39–47. <https://doi.org/10.1038/hdy.2013.13>.
- Duan, J., J. G. Zhang, H. W. Deng, and Y. P. Wang. 2013. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS ONE* 8:e59128. <https://doi.org/10.1371/journal.pone.0059128>.
- English, A. C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid, K. C. Worley, and R. A. Gibbs. 2012. Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* 7:e47768. <https://doi.org/10.1371/journal.pone.0047768>.
- English, A. C., W. J. Salerno, O. A. Hampton, C. Gonzaga-Jauregui, S. Ambreth, D. I. Ritter, C. R. Beck, C. F. Davis, M. Dahdouli, S. Ma, A. Carroll, N. Veeraraghavan, J. Bruestle, B. Drees, A. Hastie, E. T. Lam, S. White, P. Mishra, M. Wang, Y. Han, F. Zhang, P. Stankiewicz, D. A. Wheeler, J. G. Reid, D. M. Muzny, J. Rogers, A. Sabo, K. C. Worley, J. R. Lupski, E. Boerwinkle, and R. A. Gibbs. 2015. Assessing structural variation in a personal genome—Towards a human reference diploid genome. *BMC Genomics* 16:286. <https://doi.org/10.1186/s12864-015-1479-3>.
- 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.
- Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* 36:949–951. <https://doi.org/10.1038/ng1416>.
- Jenkins, G. M., M. E. Goddard, M. A. Black, R. Brauning, B. Auvray, K. G. Dodds, J. W. Kijas, N. Cockett, and J. C. McEwan. 2016. Copy number variants in the sheep genome detected using multiple approaches. *BMC Genomics* 17:441. <https://doi.org/10.1186/s12864-016-2754-7>.
- Jiang, J., J. Wang, H. Wang, Y. Zhang, H. Kang, X. Feng, Z. Yin, W. Bao, Q. Zhang, and J. F. Liu. 2014. Global copy number analyses by next generation sequencing provide insight into pig genome variation. *BMC Genomics* 15:593. <https://doi.org/10.1186/1471-2164-15-593>.
- Kadri, N. K., G. Sahana, C. Charlier, T. Iso-Touru, B. Guldbrandtsen, L. Karim, U. S. Nielsen, F. Panitz, G. P. Aamand, N. Schulman, M. Georges, J. Vilkkii, M. S. Lund, and T. Druet. 2014. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: Additional evidence for the common occurrence of balancing selection in livestock. *PLoS Genet.* 10:e1004049. <https://doi.org/10.1371/journal.pgen.1004049>.
- Littlejohn, M. D., K. Tiplady, T. A. Fink, K. Lehnert, T. Lopdell, T. Johnson, C. Couldrey, M. Keehan, R. Sherlock, C. Harland, A. Scott, R. Snell, S. Davis, and R. J. Spelman. 2016. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. *Sci. Rep.* 5:25376.
- MacDonald, J. R., R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer. 2014. The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res.* 42:D986–D992. <https://doi.org/10.1093/nar/gkt958>.
- Rhoads, A., and K. F. Au. 2015. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 13:278–289. <https://doi.org/10.1016/j.gpb.2015.08.002>.
- Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Mánér, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. 2004. Large-scale copy number polymorphism in the human genome. *Science* 305:525–528. <https://doi.org/10.1126/science.1098918>.
- Sedlazeck, F. J., P. Rescheneder, M. Nattestad, and M. C. Schatz. 2015. Detection of structural variants using third generation sequencing. *Genome Informatics*, Cold Spring Harbor, NY. Accessed Apr. 18, 2017. <http://schatzlab.cshl.edu/publications/posters/2015/2015.GI.Sniffles.pdf>.
- Sudmant, P. H., T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. Paolo Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. Lam, X. Jasmine Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E. W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalina, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel, and 1000 Genomes Project Consortium. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75–81. <https://doi.org/10.1038/nature15394>.
- Weischenfeldt, J., O. Symmons, F. Spitz, and J. O. Korbel. 2013. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nat. Rev. Genet.* 14:125–138. <https://doi.org/10.1038/nrg3373>.
- Xu, L., J. B. Cole, D. M. Bickhart, Y. Hou, J. Song, P. M. VanRaden, T. S. Sonstegard, C. P. Van Tassell, and G. E. Liu. 2014. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC Genomics* 15:683. <https://doi.org/10.1186/1471-2164-15-683>.
- Zarrei, M., J. R. MacDonald, D. Merico, and S. W. Scherer. 2015. A copy number variation map of the human genome. *Nat. Rev. Genet.* 16:172–183. <https://doi.org/10.1038/nrg3871>.
- Zhang, X., R. Du, S. Li, F. Zhang, L. Jin, and H. Wang. 2014. Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinformatics* 15:50. <https://doi.org/10.1186/1471-2105-15-50>.