J. Dairy Sci. 102:1–5
https://doi.org/10.3168/jds.2018-15638

Short communication: Identification of the pseudoautosomal region in the Hereford bovine reference genome assembly ARS-UCD1.2

T. Johnson,[1] M. Keehan,[1] C. Harland,[1] T. Lopdell,[1] R. J. Spelman,[1] S. R. Davis,[1] B. D. Rosen,[2] T. P. L. Smith,[3] and C. Couldrey[1]*
[1]Research and Development, Livestock Improvement Corporation, Hamilton 3240, New Zealand
[2]Animal Genomics and Improvement Laboratory, Agricultural Research Service USDA, Beltsville, MD 20705
[3]US Meat Animal Research Center, Agricultural Research Service USDA, Clay Center, NE 68933

ABSTRACT

In cattle, the X chromosome accounts for approximately 3 and 6% of the genome in bulls and cows, respectively. In spite of the large size of this chromosome, very few studies report analysis of the X chromosome in genome-wide association studies and genomic selection. This lack of genetic interrogation is likely due to the complexities of undertaking these studies given the hemizygous state of some, but not all, of the X chromosome in males. The first step in facilitating analysis of this gene-rich chromosome is to accurately identify coordinates for the pseudoautosomal boundary (PAB) to split the chromosome into a region that may be treated as autosomal sequence (pseudoautosomal region) and a region that requires more complex statistical models. With the recent release of ARS-UCD1.2, a more complete and accurate assembly of the cattle genome than was previously available, it is timely to fine map the PAB for the first time. Here we report the use of SNP chip genotypes, short-read sequences, and long-read sequences to fine map the PAB (X chromosome:133,300,518) and simultaneously determine the neighboring regions of reduced homology and true pseudoautosomal region. These results greatly facilitate the inclusion of the X chromosome in genome-wide association studies, genomic selection, and other genetic analysis undertaken on this reference genome.
Key words: chromosome X, pseudoautosomal region boundary, cattle

Short Communication

In mammals, sex determination is based on heteromorphic sex chromosomes designated X and Y. Females have 2 of the large, gene-rich X chromosomes, whereas males have a single X chromosome and a small, gene-poor Y chromosome. Although the X and Y chromosomes differ significantly in gene content, they do share a relatively small region of sequence homology known as the pseudoautosomal region (**PAR**). In males, recombination of the X and Y chromosome-specific regions of the sex chromosomes is not possible. In contrast, pairing and recombination within the PAR of X and Y chromosomes is a critical step for faithful segregation of the sex chromosomes during male meiosis (Ellis and Goodfellow; 1989; Ellis et al., 1989). Due to the sequence similarity and recombination that occurs between X and Y chromosomes in the PAR, this region of these chromosomes are inherited in an autosomal pattern.

In most mammals PAR are typically a few hundred kilobases to several megabases in length (Raudsepp and Chowdhary; 2008; Das et al., 2009). The physical domain of the PAR is demarcated by the pseudoautosomal boundary (**PAB**), where the sequence similarity between X and Y chromosomes decreases from ~100% to between 80 and 50% in a region referred to as the region of reduced homology, which is a transition to the purely sex specific sequence (Das et al., 2009). Despite the vital role of PAR recombination during male meiosis, the PAR is evolving, in terms of gene structure and DNA sequence variation, at a considerably faster rate than both adjacent chromosomal regions and autosomes (reviewed by Galtier, 2004; Katsura et al., 2012). The rapid rate of PAR evolution has presented challenges in between-species comparison of molecular organization, and only a few reports of PAR comparisons between species other than human and mouse have been undertaken (Raudsepp and Chowdhary, 2008; Das et al., 2009), with these comparisons limited to rough mapping.

The past 10 yr has seen the rise of genome-wide association studies (**GWAS**) in a wide range of species, including cattle, and has resulted in remarkable

Received September 2, 2018.
Accepted December 4, 2018.
*Corresponding author: Christine.couldrey@lic.co.nz

advances being made in the interrogation of complex traits (reviewed by Visscher et al., 2017). However, to date, researchers have almost exclusively used additive models in GWAS. Additive models are often of limited use for analysis of the X chromosome unless the population under investigation is entirely female (so that all individuals carry 2 copies of the X chromosome). Even in cases where all individuals studied are female, the use of additive models may not be appropriate given the inactivation of 1 X chromosome in each cell. Although X inactivation is largely thought to be random in nature (Dindot et al., 2004; Chen et al., 2016), skewed inactivation has been widely reported (Belmont, 1996; Thorvaldsen et al., 2012; Calaway et al., 2013; Couldrey et al., 2017). The first step required for the X chromosome to be included in GWAS in a meaningful manner is fine mapping of the PAR and, more importantly, accurate identification of the PAB. Identification of the PAB then allows the PAR to be analyzed as an autosome separately from the sex-specific region that will require specialized statistical models. Similarly, identification of the PAB is required to undertake other common or important steps in genetic analysis, including accurate imputation and use of X chromosome genotypes to identify or confirm the sex of a sample (Zhang et al., 2016). Before the release of the first complete bovine reference genome assembly (The Bovine Genome Sequencing and Analysis Consortium et al., 2009), sequencing of bacterial artificial chromosomes had been used to determine sequence around the PAB (Van Laere et al., 2008). However, positioning of this sequence onto the subsequent genome assembly (UMD3.1; https://www.ncbi.nlm.nih.gov/assembly/GCA_000003055.5) is not published, perhaps due to the complicated structure of the assembled PAR in UMD3.1 (represented in multiple fragments; Couldrey et al., 2017).

The objective of our study was to finely map the PAR and identify the exact location of the PAB in the newly released bovine reference genome assembly (ARS-UCD1.2, https://www.ncbi.nlm.nih.gov/assembly/GCA_002263795.2; GenBank accession NKLS00000000.2). Data sets used to fine map the PAB were (1) Illumina (San Diego, CA) 150-bp paired end read whole-genome sequence from 2 bulls and 2 cows (~30× coverage/animal), representative of the New Zealand dairy population (Holstein Friesian and Jersey); (2) Illumina BovineHD SNP chip genotypes from 578 bulls and 3,306 cows; and (3) whole-genome long-read sequences (Pacific Biosciences, Menlo Park, CA; ~60× coverage) from 1 New Zealand Holstein Friesian bull and 1 New Zealand Jersey bull.

Illumina sequence and SNP chip positions were mapped using BWA-MEM (Li and Durbin, 2009) to ARS-UCD1.2 both in the presence and absence of a Y chromosome assembly. Given that the ARS-UCD1.2 reference genome has been assembled from a cow, no Y chromosome sequence is available from this animal. The best available Y chromosome assembly generated from the reference cow's sire (Btau_5.0.1; GenBank accession number AAFC00000000) was therefore used (the small PAR sequence present in this assembly was masked) to identify the region of the X chromosome where all males were homozygous and females were heterozygous (sex-specific region) and to identify a sudden change in read depth in bulls (but not cows) indicating the end of the PAR. However, as previously reported (Couldrey et al., 2017), use of SNP chip genotypes allows only a rough estimation of the PAB. Similarly, utilizing sequence coverage with the aim of observing a 2-fold difference in coverage provided only a rough guide as to PAB location due in part to the presence of repetitive elements on both the X and Y chromosomes (data not shown; Katsura et al., 2012).

The utility of long-read DNA sequences has been well documented with regard to de novo genome assembly (Gordon et al., 2016; Bickhart et al., 2017; Jiao et al., 2017; Korlach et al., 2017). Given the challenges in accurately identifying the PAB using short-read sequences and SNP chip genotypes, we used long DNA sequence reads generated from both Pacific Biosciences (**PacBio**) RS II (Holstein Friesian; Cold Spring Harbor Laboratory, Cold Spring Harbor, NY) and PacBio Sequel (Jersey; USDA, Clay Center, NE) instruments. The PacBio sequences were mapped to the ARS-UCD1.2 assembly of the bovine genome plus Btau_5.0.1 chromosome Y (PAR masked) using minimap2 (Li, 2018). Visualization of the mapped sequences was undertaken using Integrated Genome Viewer (Thorvaldsdóttir et al., 2013). Analysis of the region identified as most likely to contain the PAB (ChrX:133,280,000–133,310,000) based on genotypes and short-read sequences clearly revealed the PAR, PAB, and the region of reduced homology (Figure 1; region of reduced homology sequence Supplemental Figure S1, https://doi.org/10.3168/jds.2018-15638). The presence of numerous soft clipped reads (indicated by sequential colored nucleotides in mapped sequences; Figure 1), where clipping had occurred within a narrow window, clearly indicated the presence of DNA sequence reads originating from the Y chromosome. The Y chromosome derivation of these soft clipped reads was confirmed by determining that all these soft clipped regions mapped to a Y chromosome location. Furthermore, sequence reads that had undergone soft clipping contained a specific haplotype distinguishing them from reads that were not soft clipped. This haplotype represents the region of reduced homology, where sequence similarity between the

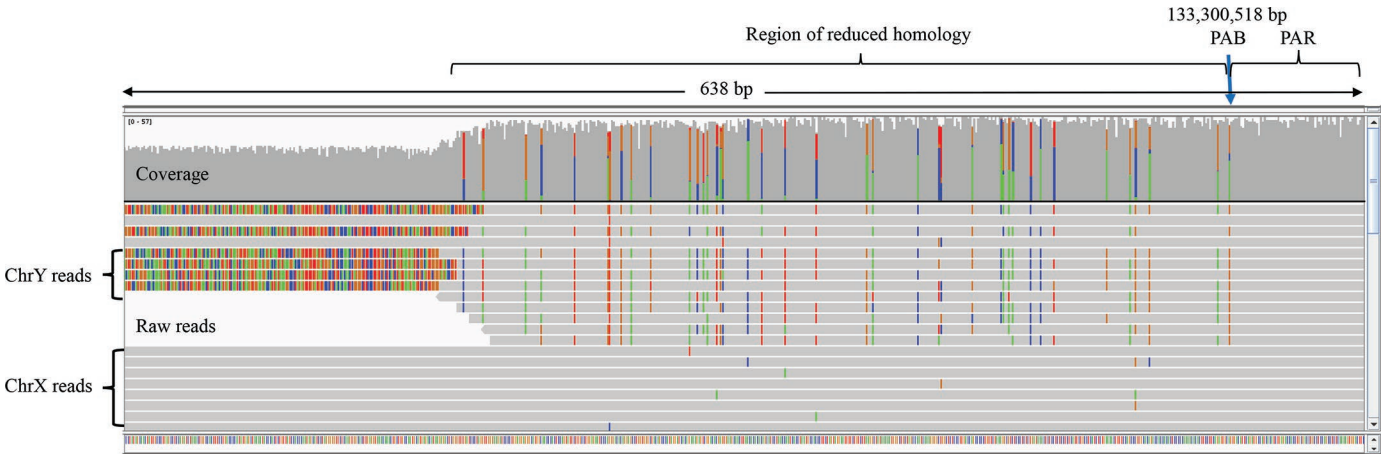

Figure 1. Integrated genome view (IGV) screenshot (ChrX:133,299,946–133,300,584) displaying alignment of raw Pacific Biosciences (Menlo Park, CA) sequence reads and depth coverage across the region containing the pseudoautosomal boundary (PAB), located at ChrX:133,300,518. Color on the 5′ end of sequences indicates soft clipping, other colors indicate variants relative to the reference genome (blue = C, green = A, gold = G, red = T). Proportion of color at variants in the coverage track indicates allelic proportions (excluding soft clipped regions). IGV set to visualize only indels of greater than 30 bp. Examples of reads from the X and Y chromosomes are labeled (ChrX reads and ChrY reads, respectively). PAR = pseudoautosomal region.

X and Y chromosomes is sufficiently high for mapping algorithms to map reads with a considerable number of variants highlighted. The 2 individual bulls for which PacBio sequence was available displayed 40 identical variants, with the Holstein Friesian bull carrying an additional 4 variants in this region of reduced homology. The sequence similarity observed in this region, and the mapping of Y chromosome sequences to the X chromosome here, are confounding factors when attempting to use short- and long-read sequence data to identify a 2-fold change in coverage at the PAB. Because of this sequence similarity, the expected 2-fold drop in sequence depth beginning at the PAB was not observed in PacBio data. Rather, a gradual tapering of read depth was seen across the length of this region, with a more marked drop where the true sex specific sequence begins (and difference in coverage was approximately half that of the PAR).

The region of reduced homology is 394 bp and, in the 2 bulls analyzed in our study, the similarity between X and Y chromosomes was higher (89–90%) than reported for many species (50–85%; Das et al., 2009). The final position in the region of reduced homology is on the X chromosome (**ChrX**) at 133,300,517, and the PAB was therefore assigned position ChrX:133,300,518. A comparison of sequences around this chromosomal location with sequences surrounding the PAB, as described by Van Laere et al. (2008) before the release of the first bovine reference genome, indicated a perfect match between the PAB (and region of reduced homology) identified in our study using PacBio sequence and sequence comparison of bacterial artificial chromosome sequencing previously reported (Van Laere et al., 2008). Taken together, this presents strong evidence that this position does represent the true PAB.

The size of the PAR on the ARS-UCD1.2 reference genome assembly was 5,708,626 bp. This is notably smaller than the PAR in the previous bovine reference genome (UMD3.1), which consisted of 2 closely located regions with a combined size of approximately 8 Mbp (Couldrey et al., 2017). A comparative analysis of PAR regions in UMD3.1 and ARS-UCD1.2 using map positions of Illumina BovineHD SNP chip markers (Figure 2) indicated approximately one third of the difference in size can be attributed to SNP previously positioned on the first PAR region now mapping to the 900-kb ARS-UCD1.2 “contig_X_unplaced.” The remainder of the first UMD3.1 PAR region now resides within the single PAR in ARS-UCD1.2. Illumina BovineHD SNP chip genotypes from SNP mapping to the “contig_X_unplaced” contig from 578 bulls were reported almost exclusively as homozygous (in reality hemizygous), whereas genotypes from females were a mix of homozygous and heterozygous (data not shown), thereby providing strong evidence that the sequence represented in this contig is not a part of the PAR.

In conclusion, we used long-read PacBio sequences from 2 bulls to fine map the PAR and accurately identify the PAB (ChrX:133,300,518) and region of reduced homology on the X chromosome. These data are important for utilizing the X chromosome in all genetic analyses that involve imputation and GWAS, as it allows the necessary splitting of this chromosome into a region that can be analyzed in a manner similar to autosomes and the larger, sex-specific region that requires different statistical models for interrogation.

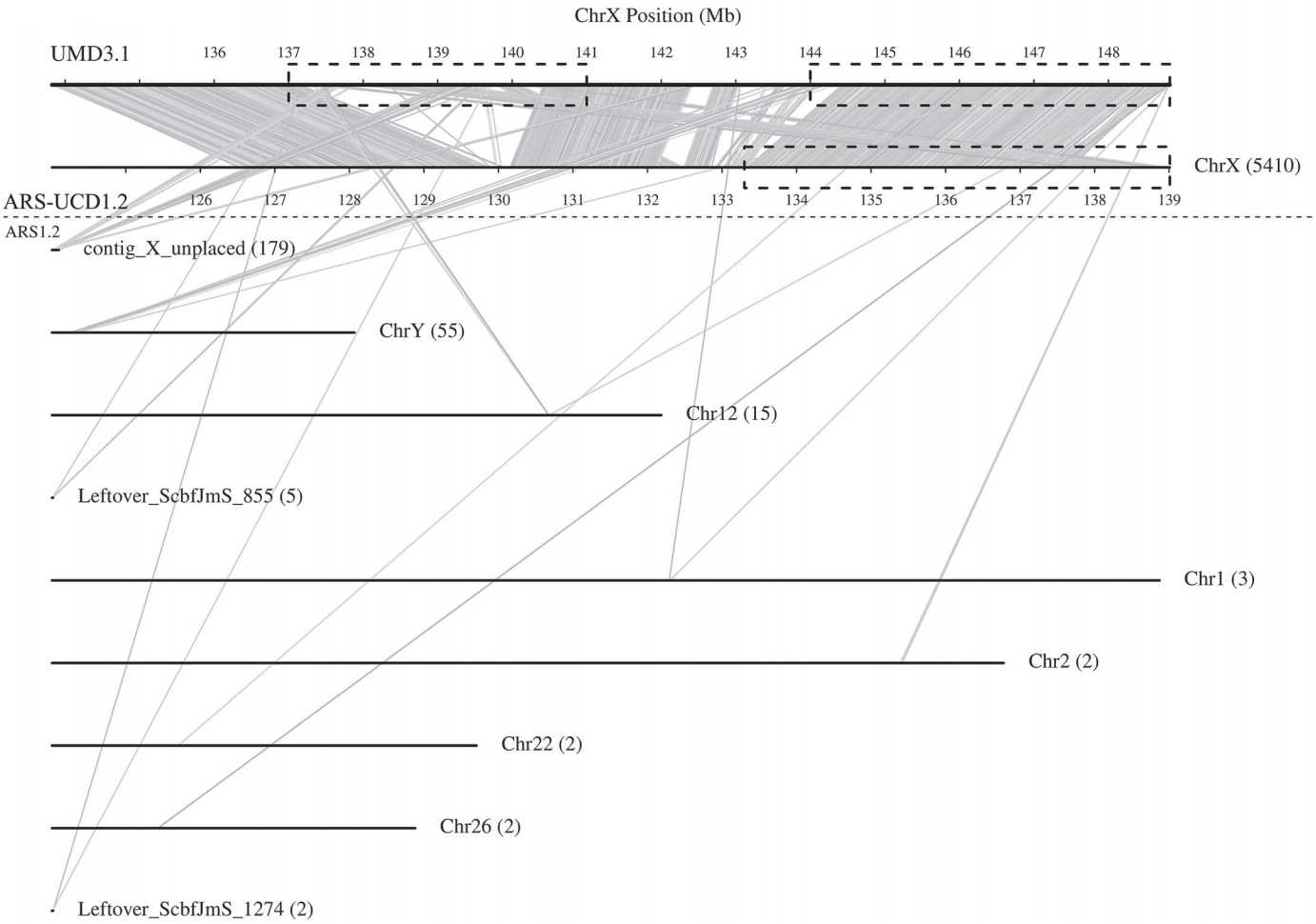

Figure 2. Comparison of Illumina Bovine HD SNP chip (Illumina Inc., San Diego, CA) marker positions on 15 Mbp of the X chromosome (ChrX). The region displayed contains the pseudoautosomal region (PAR) from the previous (UMD3.1; https://www.ncbi.nlm.nih.gov/assembly/GCA_000003055.5) and present (ARS-UCD1.2; https://www.ncbi.nlm.nih.gov/assembly/GCA_002263795.2) bovine genome reference assemblies. Dashed boxes represent PAR regions in UMD3.1 and ARS-UCD1.2. Numbers in brackets indicate the number of SNP from UMD3.1 ChrX reads at 134 to 149 Mbp that map to the chromosome/contig. Contigs with names beginning with "Leftover_ScbfJmS" represent contigs that have not been assembled into a chromosome.

ACKNOWLEDGMENTS

The authors thank Vivienne Bennett (Livestock Improvement Corporation, Hamilton, New Zealand) for critical reading of the manuscript. Funding for this project was provided by the New Zealand Ministry for Primary Industries (Wellington, New Zealand) as a Primary Growth Partnership.

REFERENCES

Belmont, J. W. 1996. Genetic control of X inactivation and processes leading to X-inactivation skewing. Am. J. Hum. Genet. 58:1101–1108.

Bickhart, D. M., B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. T. Sullivan, J. N. Burton, H. J. Huson, J. C. Nystrom, C. M. Kelley, J. L. Hutchison, Y. Zhou, J. Sun, A. Crisà, F. A. Ponce de León, J. C. Schwartz, J. A. Hammond, G. C. Waldbieser, S. G. Schroeder, G. E. Liu, M. J. Dunham, J. Shendure, T. S. Sonstegard, A. M. Phillippy, C. P. Van Tassell, and T. P. Smith. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat. Genet. 49:643–650. https://doi.org/10.1038/ng.3802.

Calaway, J. D., A. B. Lenarcic, J. P. Didion, J. R. Wang, J. B. Searle, L. McMillan, W. Valdar, and F. Pardo-Manuel de Villena. 2013. Genetic architecture of skewed X inactivation in the laboratory mouse. PLoS Genet. 9:e1003853. https://doi.org/10.1371/journal.pgen.1003853.

Chen, Z., D. E. Hagen, J. Wang, C. G. Elsik, T. Ji, L. G. Siqueira, P. J. Hansen, and R. M. Rivera. 2016. Global assessment of imprinted gene expression in the bovine conceptus by next generation sequencing. Epigenetics 11:501–516. https://doi.org/10.1080/15592294.2016.1184805.

Couldrey, C., T. Johnson, T. Lopdell, I. L. Zhang, M. D. Littlejohn, M. Keehan, R. G. Sherlock, K. Tiplady, A. Scott, S. R. Davis, and R. J. Spelman. 2017. Bovine mammary gland X chromosome inactivation. J. Dairy Sci. 100:5491–5500. https://doi.org/10.3168/jds.2016-12490.

Das, P. J., B. P. Chowdhary, and T. Raudsepp. 2009. Characterization of the bovine pseudoautosomal region and comparison with sheep,

goat, and other mammalian pseudoautosomal regions. Cytogenet. Genome Res. 126:139–147. https://doi.org/10.1159/000245913.

Dindot, S. V., K. C. Kent, B. Evers, N. Loskutoff, J. Womack, and J. A. Piedrahita. 2004. Conservation of genomic imprinting at the XIST, IGF2, and GTL2 loci in the bovine. Mamm. Genome 15:966–974. https://doi.org/10.1007/s00335-004-2407-z.

Ellis, N., and P. N. Goodfellow. 1989. The mammalian pseudoautosomal region. Trends Genet. 5:406–410.

Ellis, N. A., P. J. Goodfellow, B. Pym, M. Smith, M. Palmer, A. M. Frischauf, and P. N. Goodfellow. 1989. The pseudoautosomal boundary in man is defined by an Alu repeat sequence inserted on the Y chromosome. Nature 337:81–84. https://doi.org/10.1038/337081a0.

Galtier, N. 2004. Recombination, GC-content and the human pseudoautosomal boundary paradox. Trends Genet. 20:347–349. https://doi.org/10.1016/j.tig.2004.06.001.

Gordon, D., J. Huddleston, M. J. Chaisson, C. M. Hill, Z. N. Kronenberg, K. M. Munson, M. Malig, A. Raja, I. Fiddes, L. W. Hillier, C. Dunn, C. Baker, J. Armstrong, M. Diekhans, B. Paten, J. Shendure, R. K. Wilson, D. Haussler, C. S. Chin, and E. E. Eichler. 2016. Long-read sequence assembly of the gorilla genome. Science 352:aae0344. https://doi.org/10.1126/science.aae0344.

Jiao, Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, B. Wang, M. S. Campbell, J. C. Stein, X. Wei, C. S. Chin, K. Guill, M. Regulski, S. Kumari, A. Olson, J. Gent, K. L. Schneider, T. K. Wolfgruber, M. R. May, N. M. Springer, E. Antoniou, W. R. McCombie, G. G. Presting, M. McMullen, J. Ross-Ibarra, R. K. Dawe, A. Hastie, D. R. Rank, and D. Ware. 2017. Improved maize reference genome with single-molecule technologies. Nature 546:524–527. https://doi.org/10.1038/nature22971.

Katsura, Y., M. Iwase, and Y. Satta. 2012. Evolution of genomic structures on mammalian sex chromosomes. Curr. Genomics 13:115–123. https://doi.org/10.2174/138920212799860625.

Korlach, J., G. Gedman, S. B. Kingan, C. S. Chin, J. T. Howard, J. N. Audet, L. Cantin, and E. D. Jarvis. 2017. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. Gigascience 6:1–16. https://doi.org/10.1093/gigascience/gix085.

Li, H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty191.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754–1760.

Raudsepp, T., and B. P. Chowdhary. 2008. The horse pseudoautosomal region (PAR): Characterization and comparison with the human, chimp and mouse PARs. Cytogenet. Genome Res. 121:102–109. https://doi.org/10.1159/000125835.

The Bovine Genome Sequencing and Analysis Consortium, C. G. Elsik, R. L. Tellam, and K. C. Worley. 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science 324:522–528. https://doi.org/10.1126/science.1169588.

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov. 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14:178–192. https://doi.org/10.1093/bib/bbs017.

Thorvaldsen, J. L., C. Krapp, H. F. Willard, and M. S. Bartolomei. 2012. Nonrandom X chromosome inactivation is influenced by multiple regions on the murine X chromosome. Genetics 192:1095–1107. https://doi.org/10.1534/genetics.112.144477.

Van Laere, A. S., W. Coppieters, and M. Georges. 2008. Characterization of the bovine pseudoautosomal boundary: Documenting the evolutionary history of mammalian sex chromosomes. Genome Res. 18:1884–1895. https://doi.org/10.1101/gr.082487.108.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 2017. 10 years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet. 101:5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

Zhang, I. L., C. Couldrey, and R. G. Sherlock. 2016. Using genomic information to predict sex in dairy cattle. Pages 26–30 in Proceedings of the New Zealand Society of Animal Production. New Zealand Society of Animal Production, Hamilton, New Zealand.