



OPEN ACCESS

EDITED BY
Guillermo Giovambattista,
CONICET Institute of Veterinary Genetics
(IGEVE), Argentina

REVIEWED BY
Andrea Delledonne,
University of Milan, Italy
Oshin Togla,
National Dairy Research Institute, India

*CORRESPONDENCE
Yu Wang,
✉ yu.wang@lic.co.nz

[†]PRESENT ADDRESS
Tony Nugroho,
NetValue Ltd, Hamilton, New Zealand;
Christine Couldrey,
Independent Science Consultant,
Hamilton, New Zealand

RECEIVED 15 April 2026
REVISED 11 May 2026
ACCEPTED 18 May 2026
PUBLISHED 11 June 2026

CITATION
Wang Y, Nugroho T, Johnson TJJ,
Couldrey C and Harris BL (2026)
Detection and evaluation of copy number
variation using both linked-read and
short-read sequencing in New Zealand
dairy cattle.
Front. Genet. 17:1856199.
doi: 10.3389/fgene.2026.1856199

COPYRIGHT
© 2026 Wang, Nugroho, Johnson,
Couldrey and Harris. This is an open-
access article distributed under the terms
of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that
the original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Detection and evaluation of copy number variation using both linked-read and short-read sequencing in New Zealand dairy cattle

Yu Wang*, Tony Nugroho[†], Thomas J. J. Johnson,
Christine Couldrey[†] and Bevin L. Harris

Research & Development, Livestock Improvement Corporation, Hamilton, New Zealand

In recent years, genetic studies have made significant progress in identifying single-nucleotide polymorphisms (SNPs) associated with cattle health and production traits. However, it is still challenging to identify and validate more complicated forms of variation, such as copy number variation (CNV) and other types of structural variation (SV). In this study, SV regions were identified using 37 New Zealand dairy cattle with linked-read sequence data. A transmission-based framework was used to validate these variants at the population scale. 62,438 putative autosomal SV regions were identified with the LongRanger pipeline following the 10x Genomics recommendations. Copy number states for these regions were subsequently estimated via a read-depth based genotyping method using CNVpytor in a population-representative cohort of 2306 animals using Illumina short-read sequencing technology. Mendelian inheritance of copy number states was assessed using linear mixed models incorporating pedigree information, and transmission levels were used to quantify the biological validity of each CNV region. Transmission levels ranged widely, with a mean of 0.5162 across all regions, where higher transmission levels were proportionally enriched for larger SVs. A total of 7218 CNV regions exhibited high transmission levels (>0.9), indicating strong evidence of inheritance. Among these, 7136 overlapped CNV regions reported in one or more public datasets, while 82 high-confidence regions represent previously unreported variants. High-transmission CNV regions tended to show clear, discrete inheritance patterns in trio families, providing the biological evidence that these CNVs are inherited within the population. Together, these results demonstrate that integrating linked-read sequencing with population-scale transmission-based validation provides a robust framework for identifying high-confidence CNV regions. This catalogue of validated CNV regions represents an important resource for downstream functional analyses and the incorporation of structural variation into genomic selection and breeding programs.

KEYWORDS

cattle, copy number variation, structural variant, transmission level, whole genome sequencing

Introduction

Structural variation (SV) represents a major source of genetic diversity and contributes substantially to phenotypic variation in mammals, complementing the effects of single-nucleotide polymorphisms (SNPs) in shaping complex traits (Zhang et al., 2009; Chen et al., 2024). Among different classes of SVs, copy number variations (CNVs), defined as genomic gains and losses typically exceeding 50 bp, are of particular interest due to their direct influence on gene dosage (Rice and McLysaght, 2017) and regulatory elements (Haraksingh and Snyder, 2013). CNV regions have been shown to affect gene expression and biological pathways, which relate to numerous economically important traits (Liu, 2025). In dairy cattle, CNV regions have been associated with production, fertility, health, and longevity traits, highlighting their potential importance for both biological discovery and applied breeding programs (Butty et al., 2021; Oliveira et al., 2024; Ladeira et al., 2025).

Despite their functional relevance, CNV regions remain challenging to detect and interpret accurately (Liu, 2025). Most large-scale cattle breeding programs rely on SNP-based genotyping platforms, where analytical pipelines and genomic prediction models are well established (VanRaden, 2020). CNV regions have also been discovered using SNP genotyping arrays in dairy cattle, either through intensity-based CNV calling (Hou et al., 2011) or via linkage disequilibrium between SNPs and CNVs (Kadri et al., 2012; Xu et al., 2014). However, array-based approaches provide limited resolution for complex and multi-allelic CNV regions (Bickhart and Liu, 2014). In contrast, accurate CNV detection typically requires whole-genome sequencing and computationally intensive analyses, with performance strongly influenced by accuracy of genome mapping and sequencing depth (Liu et al., 2024). Read-depth-based approaches are scalable but prone to false positives and imprecise copy number estimates, particularly in GC-biased or highly repetitive genomic regions (Bickhart et al., 2012; Duan et al., 2013). As a result, there is no consensus on the best method of CNV identification and many detected CNV regions lack robust biological validation, which limits their utility in downstream analyses and applications in the breeding programs. In addition, studies comparing sequencing platforms have demonstrated that long-read technologies identify substantially more SVs than short-read approaches, particularly for larger variants (Gao et al., 2022). Long-read sequencing technologies provide an improved resolution for CNV discovery by capturing long-range genomic information and enabling more accurate breakpoint inference (Lin et al., 2023). However, the cost, data volume, and limited throughput of these technologies currently preclude their routine deployment at the population level (Nguyen et al., 2023).

An additional challenge in CNV analysis arises from population structure, breed difference, and reference genome bias. Structural variation is increasingly recognised as partly breed-specific, reflecting demographic history, selection pressures, and divergence between sampled individuals and the reference genome (Chen et al., 2021). Studies have reported significant differences in detected number of CNV regions, size distribution, and genomic location between major dairy breeds such as Holstein-Friesian and Jersey, affecting both discovery power and cross-study concordance (Reynolds et al., 2018). Reference genomes derived

from a single breed may further bias detection toward variants present in the reference haplotypes, reducing sensitivity for structurally divergent regions in other populations (Leonard et al., 2022). These factors complicate the interpretation of CNV catalogues across different or admixed dairy cattle populations.

Given these constraints, robust validation strategies are essential to distinguish biologically meaningful CNV regions from technical artefacts. Pedigree-based analyses exploiting Mendelian transmission provide a powerful framework for evaluating the biological validity of CNV regions at a population level. Couldrey et al. (2017) demonstrated that CNV regions showing consistent inheritance patterns within families are more likely to represent true structural variants than those detected solely based on read-depth or positional overlap. Transmission-based validation therefore offers a biologically grounded approach that complements discovery-oriented methods and helps bridge the gap between structural variant research and practical breeding applications.

In this study, we integrated the linked-read sequencing-based SV discovery and copy number estimation using short-read sequencing data with population-scale validation in New Zealand dairy cattle. By combining high-resolution discovery in a limited number of animals with transmission-based evaluation across a large, representative cohort, we aim to generate a high-confidence catalogue of CNV regions with strong evidence of inheritance.

Materials and methods

Linked-read sequencing and structural variants discovery

Linked-read sequencing of 37 New Zealand dairy bulls (18 Holstein-Friesian, 17 Jersey, and 2 Holstein-Friesian Jersey crossbred) was undertaken following the standard protocols by 10x Genomics (Marks et al., 2019). These bulls were born between 2006 and 2013, which were either among the most frequently used sires in the New Zealand dairy population during the timeframe or selected for their potential research significance. The influence of these bulls is reflected in the number of daughters attributed to each, which ranged from 220 to 137,288, with an average of 24,705. The details of the sequencing procedure were described in Keehan and Couldrey, (2018). The long-range context is provided by using microfluidics to segment and barcode high-molecular-weight DNA without needing true long reads. We aligned the reads using the LongRanger (v2.2.2) pipeline following 10x Genomics recommendations. All reads were mapped against the cattle genome reference ARS-UCD 1.2 (Rosen et al., 2020) and underwent subsequent SV detection using the `analyze_sv_calls` function from the LongRanger pipeline with the recommended settings. SVs were considered for further analysis if they were larger than 50 bp in length, consistent with common definitions of structural variation. Quality control measures excluded any detected regions that lacked start or end positions or were located outside autosomes. After the data cleaning, the called SV regions were then merged across individuals based on genomic coordinate overlap using the `bcftools -merge` command (Li, 2011), generating a set of non-redundant SV regions.

Short-reads sequencing and CNV genotyping

Read-depth-based CNV genotyping analysis was undertaken in a population of 2306 animals representing the population structure of New Zealand dairy cattle. This cohort included 730 Holstein-Friesian, 468 Jersey, 1,069 Holstein-Friesian × Jersey crossbred, 3 Ayrshire animals, and 36 individuals from mixed or minor breeds. The breed composition of this population closely reflects the current New Zealand dairy population, ensuring that the findings are broadly representative (Harris, 2022). Among the sequenced animals, 1,532 were bulls, many of which are widely used sires that have contributed substantially to the national dairy herd. In addition, 774 cows were included in the study, some of whom are the dams of the sequenced bulls. The animals were sequenced on Illumina HiSeq 2000 instruments targeting 100bp paired-end reads. The raw sequence data were mapped to the ARS-UCD1.2 bovine reference genome (Rosen et al., 2020) using the Burrows-Wheeler alignment algorithm version 0.7.17 (Li, 2013). The copy number states of the identified regions were then estimated in each of the short-read sequenced animals using the genotyping function from CNVpytor version 1.3.1 (Suvakov et al., 2021), a Python version of its previous software, CNVnator (Abyzov et al., 2011). Standard CNVpytor quality control and GC bias correction were applied prior to copy number estimation. The read-depth signals were extracted and compared to the expected distribution for different copy number states for each provided region by dividing the genome into segments of 200bp. For each putative SV region, copy number estimates from constituent bins were aggregated to assign a single copy number state per animal per region. Due to the lack of SV type annotation in the LongRanger pipeline, we implemented a classification method based on the copy number status. The state of each region was categorized into loss, normal, and gain state via copy number estimates per individual. Because copy number estimates derived from read depth are continuous and subject to technical variation, thresholds allowing for uncertainty around the diploid state were applied. Copy number values lower than or equal to 1.5 were classified as losses, values greater than or equal to 2.5 as gains, and intermediate values as normal (Friedrich et al., 2020). The frequency of each state was calculated across all individuals for all regions. Regions were classified as deletion or duplication CNV regions when the corresponding state was present in at least 50% of individuals. Regions showing both loss and gain in at least 5% of individuals were classified as complex CNV regions. Regions with low-frequency copy number changes (<5%) were classified as rare CNV regions, while regions without a dominant state and exhibiting high variance were labelled as uncertain.

Transmission-based validation

Mendelian inheritance of these copy numbers was then estimated using a univariate animal model, which was described in detail in Couldrey et al. (2017). The model is specified as:

$$y_i = \mu + a_i + e_i$$

where y_i is the estimated copy number state of a given region for the individual i , μ is the overall mean, a_i is the additive genetic effect of

the individual i and e_i is the residual error. The additive genetic effects are assumed to follow: $a \sim N(0, A\sigma_a^2)$, where A is the relationship matrix derived from pedigree traced for 3 generations and σ_a^2 is the additive genetic variance. The residual errors are assumed to be independently and normally distributed: $e \sim N(0, I\sigma_e^2)$, where I is the identity matrix and σ_e^2 is the residual variance. Transmission level therefore reflects the proportion of variance in estimated copy number attributable to additive genetic effects. Transmission level ranged from 0 to 1, with 0 indicating the region was a false positive discovery or a sequence artifact and 1 indicating the copy number is inherited following Mendelian inheritance. The APEX linear model suite (www.ghpc.ai) was used to efficiently solve these equations. Copy number inheritance was additionally examined and visualized in 600 trio families to assess Mendelian transmission patterns.

Comparison with existing dairy cattle CNV datasets

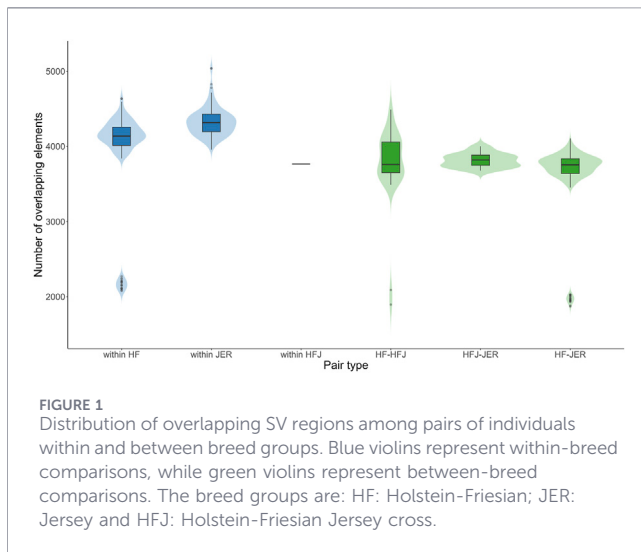
Four previously published CNV datasets related to dairy cattle were selected for comparison with the copy number variants identified in this study. These datasets represent most of the structural variant research in dairy cattle for which relevant information and data have been made publicly available. Supplementary Table S1 provides an overview of these studies, including data type, sample size, breeds, analytical methods, number of detected regions, and database availability. CNV regions archived in the Database of Genomic Variants Archive (DGVa) were obtained from eight studies and remapped to the ARS-UCD1.2 assembly to ensure consistency with the reference genome used in this study. Additional CNV datasets were obtained from Lee et al. (2023), Bhati et al. (2023), Grant et al. (2024). Overlapping CNV regions among datasets were identified using the *GenomicRanges* R package (Lawrence et al., 2013). An overlap of at least 10 bp was required to classify equivalence between our detected region with a publicly available region. Genome feature annotations, including gene and exon coordinates, were downloaded from Ensembl (*Bos_taurus*.ARS-UCD1.2.110.gtf.gz) for downstream annotation of overlapping regions.

Results

In this study we utilized linked-read sequencing to discover potential SV regions. These putative SVs regions were then validated in short-reads sequenced animals by evaluating the copy number status and the estimating the transmission levels. Together, these analyses define a genome-wide set of CNV regions with strong population-level evidence of inheritance.

Structural variants inference

All 37 linked-read sequenced animals have an average sequence depth between 29x and 43x, except for a bull, “Esteem”, which has a sequence depth of 68.89x with the highest number of detected reads. The relationship between average sequence depth and number of structural variants detected can be seen in Supplementary Figure S1. While some variation is observed, SV discovery does not scale



linearly with the sequence depth, and three animals with extreme values are highlighted. The number of reads detected among animals ranged between 634,723,606 and 1,573,807,188 (“Esteem”) and the mapping rate ranged between 83.75% and 96.95%. On average, 850,231,229 reads were generated by 10x genomics, representing 39.7x coverage of the genome. Approximately 7000 SVs were discovered in each animal. One bull (“Lonestar”) showed substantially fewer SVs detected (5916) and lower mapping rate (83.75%) than other animals, resulting in fewer shared SV regions (~2000), consistent with reduced discovery power. On average, the number of regions intersecting among individuals was around 4000. Three animals that have over 5000 overlapping regions originated from a duo family. The number of overlapping SV regions between pairs of individuals sampled were also investigated within the same breed and between different breeds among all 37 animals, which can be seen in [Figure 1](#). The distributions show consistently higher overlap among individuals from the same breed than between breeds. The number of overlapping regions between the bull “Lonestar” and others was around 2000 across all breed combinations, which is significantly lower than the other comparands.

After merging the output from the LongRanger pipeline for all animals, 62,438 putative autosomal SVs over 50bp were identified from 37 linked-read sequenced animals. The relationship between autosomal chromosome length and the number of detected SV regions is shown in [Supplementary Figure S2](#). Each point represents one chromosome, and a strong positive correlation ($r = 0.843$) was observed between chromosome length and the number of detected CNV regions. Chromosome 1, the longest chromosome of the cattle genome, has the most SVs detected with 3498 SVs. Meanwhile, chromosome 25, the shortest chromosome of the cattle genome, has the least SVs detected with 903 SVs. Overall, the cumulative length of SV regions is 33,002,890 bp which accounts for 1.33% of the total length of the genome. [Figure 2](#) presents the size and CNV type distribution of putative SV regions detected across the 37 linked-read sequenced animals. The size of the detected SVs ranged between 50bp and 29,904bp, with an average of 1854bp. Most detected SVs fall into smaller size categories, with progressively fewer SVs detected as size increases.



Copy number evaluation and transmission level estimation

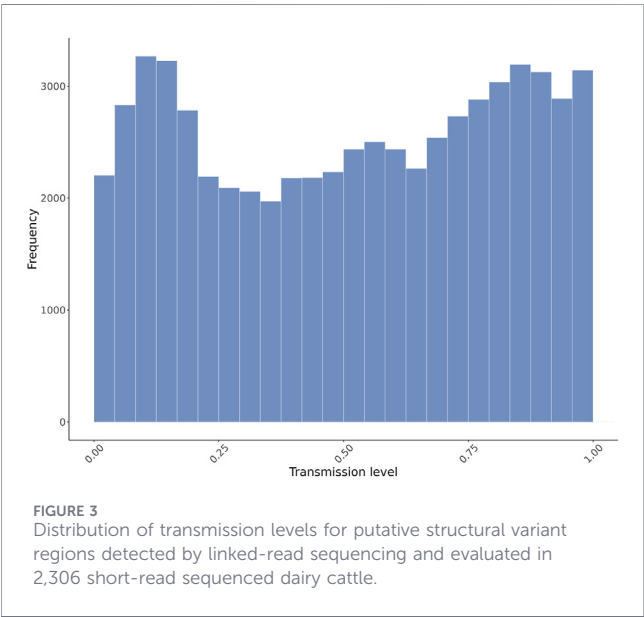
The CNV type for each detected SV region was determined by estimating the proportion of individuals exhibiting gain, loss, or normal states. In summary, most regions were classified as complex (40.69%), deletion (32.07%), or uncertain (18.88%). Duplication accounted for 6.14% of detected regions, while 2.23% were considered rare. The average sizes of regions across most categories were comparable, except for the rare category, which exhibited significantly larger average sizes at 12,895 bp (see [Table 1](#)).

To assess the biological validity of the linked-read-discovered SVs, transmission levels were estimated using population-scale short-read sequencing and pedigree information. The 62,438 putative SV regions discovered in the linked-read sequence data showed a wide range of transmission levels on the 2306 short-read sequenced animals (see [Figure 3](#)). The average transmission level was 0.5162 with a standard deviation of 0.3016. CNV regions were fairly evenly spread across each transmission level bin split by 0.1, with the counts in these bins ranging from 4949 to 7629. This represents between 7.93% and 12.22% of the total number of CNV regions (refer to [Table 2](#)). While small CNV regions are numerically dominant overall, [Figure 4](#) shows that higher transmission levels are proportionally enriched for larger CNV regions, whereas lower transmission level bins are dominated by smaller CNV regions. Notably, regions designated as deletions demonstrated a higher transmission level relative to other categories, averaging 0.71 (see [Table 1](#)).

CNV regions with higher transmission levels tended to follow a multimodal distribution of copy number estimates among animals and the inheritance can be clearly observed in the trio families. In [Figure 5](#), we illustrated a CNV region with a transmission level of one. This CNV was selected as a representative example of high-transmission deletions observed genome-wide. From the distribution of the copy number of this CNV region ([Figure 5a](#)), most data points clustered around integer CNV values (0, 1, 2). CNV

TABLE 1 Number, average size and average transmission level of CNV regions of each SV type.

CNV type	Total number	Size in bp (mean ± sd)	Transmission level (mean ± sd)
Deletion	20,022 (32.07%)	1664 ± 3345	0.714 ± 0.241
Duplication	3828 (6.13%)	2201 ± 4071	0.537 ± 0.202
Complex	25,406 (40.69%)	740 ± 1912	0.398 ± 0.272
Rare	1391 (2.23%)	12,895 ± 8346	0.241 ± 0.198
Uncertain	11,791 (18.88%)	3164 ± 4757	0.460 ± 0.311



values of 0 and 1 are more frequent than 2, indicating this copy number variant is likely to be a deletion. Figure 5b illustrates copy number inheritance across sire, dam, and offspring for the same CNV region across 600 trio families. The clustering at integer values suggests that copy number variations are mostly discrete among all individuals. A positive correlation can be observed between parents' copy numbers and their offspring's copy number. When both sire and dam have a copy number of 0, their offspring also tend to have a copy number of 0, indicating a zero-copy deletion being inherited by the offspring.

Overlap with other publicly available databases

We compared our putative CNV regions with those in four public datasets. The number of CNV regions reported in each of these public datasets ranged from 13,732 to 65,550 which can be seen in Table 3. A substantial proportion of previously reported CNV regions were rediscovered in our dataset, with over half of regions from DGVA, Lee et al. (2023), Bhati et al. (2023) overlapping with regions identified in this study. In contrast, overlap with CNV regions reported by Grant et al. (2024) was lower, with only 44.15% of regions using a Menta pipeline and 25.79% of regions using a Smoove pipeline. Inversely, the public dataset that contained the highest number of regions found in this study was DGVA with

more than 86% of our regions overlapping. The public dataset that contained the lowest number of regions found in this study was Bhati et al. (2023) with less than 39% of our regions overlapping. 7218 CNV regions found in this study exhibited high transmission levels (>0.9). Among these, we identified 5267 deletions, 196 duplications, 994 uncertain, 728 complex and 33 rare variants. 82 of these regions represent novel variants that were not reported before, while the remaining regions overlapped one or more published datasets. CNV regions identified consistently across multiple studies tended to show higher transmission levels. 6734 high-transmission CNV regions overlapped regions in DGVA, representing the largest concordant set among all comparisons, whereas fewer high-transmission CNV regions were overlapped with Bhati et al. (2023). 1561 of the regions that have a high transmission level are locating on the region of genes, in which 357 of them were located on an exon.

Discussion

In this study we developed a pipeline combining SV discovery using linked-read sequence data, copy number estimation and population scale transmission-based validation using short-read sequencing data in dairy cattle. Firstly, the SV regions were identified in 37 animals with linked-read sequence data. The copy number status for these regions was then assessed on 2306 animals with Illumina short-read sequencing information. We then evaluated the transmission level for each region to assess biological evidence for validation. These regions were also compared with several publicly available databases. Structural variant discovery has commonly relied on whole-genome sequence data, yet the effectiveness of the analysis depends on the chosen sequencing technology with each offering its own set of trade-offs (Lavrichenko et al., 2021). Conventional short-read whole-genome sequencing provides high base accuracy and scalability (Zhao et al., 2021), but its limited read length constrains the ability to resolve larger SVs, particularly those spanning repetitive or hard-to-map regions. In addition, SV detection from short-read data relies largely on indirect signals such as read depth, paired-end discordance, and split reads, which are sensitive to mapping artefacts and often yield imprecise breakpoints. These limitations contribute to both reduced sensitivity for larger variants and an elevated false-positive rate, especially in complex genomic regions. In contrast, long-read sequencing platforms, such as PacBio (Byrsk-Bishop et al., 2022) and Oxford Nanopore (Wang et al.,

TABLE 2 Number and proportions of CNV regions in each transmission level group.

Transmission level group	Number of CNV regions	Proportion of CNV regions
[0,0.1)	6302	10.09%
[0.1,0.2)	7527	12.06%
[0.2,0.3)	5178	8.29%
[0.3,0.4)	4949	7.93%
[0.4,0.5)	5283	8.46%
[0.5,0.6)	5864	9.39%
[0.6,0.7)	5844	9.36%
[0.7,0.8)	6644	10.64%
[0.8,0.9)	7629	12.22%
[0.9,1)	7218	11.56%

2021), represent the current gold standard for SV discovery. They offer the ability to span entire variants and resolve breakpoints at base-pair resolution. Long-reads consistently outperform short-read approaches in detecting large, complex, and nested SVs, including insertions and rearrangements in highly repetitive regions. However, long-read sequencing remains constrained by higher costs, lower throughput, and practical challenges in generating population-scale datasets, limiting its routine application. In this study, we utilized the linked-read sequencing technology, which partially overcomes the challenges of providing long-range information to short reads and can detect large SVs at a lower cost than long-reads sequencing technology (Spies et al., 2017; Karaoğlu et al., 2020). While LongRanger itself is no longer actively developed, the transmission-based validation framework presented here is independent of the discovery technology and is directly applicable to SVs detected using long-read, pangenome, or graph-based approaches.

A key challenge in structural variant discovery is distinguishing true inherited variants from technical artefacts generated by sequencing and alignment errors. This challenge is amplified as sequencing technologies with higher discovery sensitivity, such as linked-read sequencing, increase the number of candidate SV regions. In this study, the number of putative SVs identified per animal using linked-read data ranged between 5916 and 7981 among all 37 individuals. These results are comparable to Gao et al. (2022), where 8315 SVs were detected in an individual with a sequencing depth of 55x with linked-read sequencing data. However, the overall number of detected SV regions are significantly larger compared to most of the other studies using short-read sequencing technology (see Supplementary Table S1). These counts substantially exceed those typically reported from short-read only pipelines, which generally detect fewer and smaller SVs due to limitations in breakpoint resolution and mapping ambiguity.

To address the inherent uncertainty of LongRanger’s breakpoint-centric output, we implemented a robust copy-number-based classification pipeline that translates abstract breakpoints into biologically interpretable structural variants. By

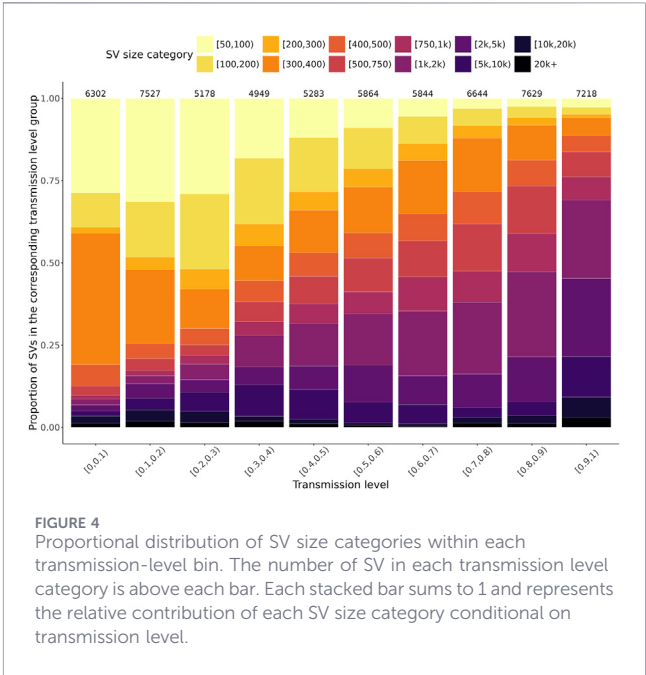


FIGURE 4 Proportional distribution of SV size categories within each transmission-level bin. The number of SV in each transmission level category is above each bar. Each stacked bar sums to 1 and represents the relative contribution of each SV size category conditional on transmission level.

utilizing depth-of-coverage from short-read data, a conservative thresholding strategy (copy number ≤ 1.5 for losses and copy number ≥ 2.5 for gains) was applied. In diploid genomes, the expected baseline copy number for autosomal loci is two. However, read depth-based copy number estimates are continuous values subject to sampling variance, GC bias, mapping bias, and normalization artifacts. Therefore, many studies do not use an exact cutoff at 2.0 but instead define buffered intervals around the diploid state to reduce false positives. Furthermore, stratifying variants into deletion/insertion ($\geq 50\%$ prevalence), rare ($\leq 5\%$ prevalence), and complex ($\geq 5\%$ prevalence) categories enables the distinction between stable polymorphisms and multi-allelic “hotspots” of genomic instability. Rare CNVs are low-frequency copy number changes found in few individuals, likely due to recent mutations or selection, and were analysed separately as they provide limited population-level insight. Highly variable regions were labelled as uncertain, since such patterns can result from true multi-allelic loci or technical artefacts, especially in repetitive areas. These rare and uncertain CNVs were kept separate from deletion, duplication, and complex CNV categories, which required consistent, interpretable patterns across the population. Complex CNVs were defined as regions exhibiting both copy number losses and gains across individuals, reflecting multi-allelic or heterogeneous copy number patterns. This classification is based on population-level copy number distribution and does not necessarily indicate structurally complex rearrangements within individual genomes. In summary, this thresholding approach balances biological plausibility with measurement uncertainty, thereby improving robustness in large-scale CNV classification analyses.

Despite increasing sensitivity of SV discovery technologies, there remains no widely adopted population-scale framework to distinguish inherited CNV regions between technical artefacts and biological evidence. Transmission-level analysis provides a biologically grounded approach to address indeterminate validity

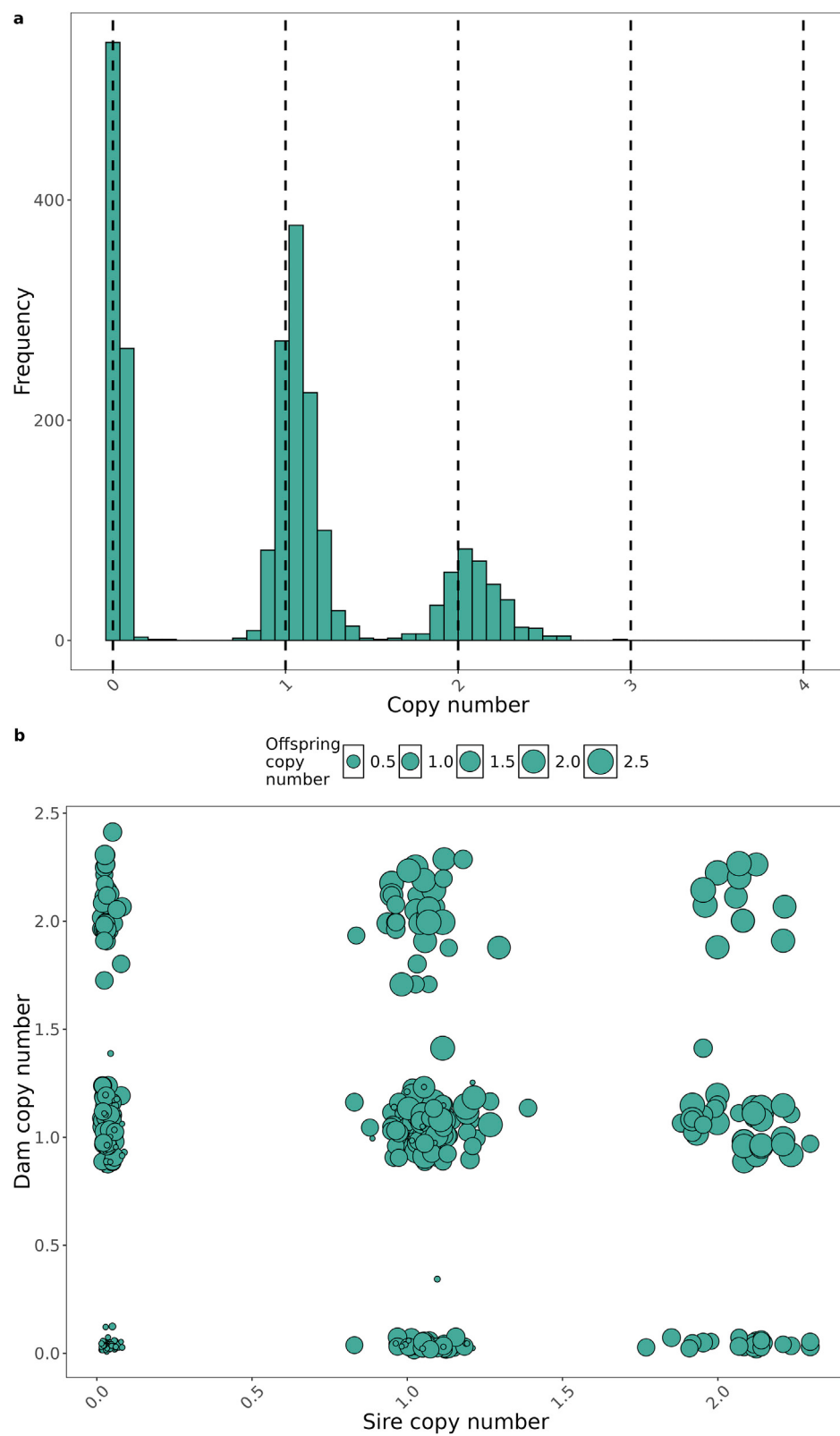


FIGURE 5
Distribution of copy number states **(a)** and inheritance patterns among sire, dam, and offspring across 600 trio families **(b)** for a representative CNV with complete Mendelian transmission (transmission level = 1).

TABLE 3 Comparison of CNV regions identified in this study with publicly available CNV datasets of dairy cattle, showing overlap counts and transmission-based validation.

Study	Number of detected CNV regions	Number and percentage of their regions that overlap with ours	Number and percentage of our regions that overlap with theirs	Number of our regions that are overlapped and have a transmission level above 0.9
DGVa	23,053	12,529 (54.35%)	53,859 (86.26%)	6734
Lee et al. (2023)	13,732	10,445 (76.06%)	31,436 (50.35%)	5251
Bhati et al. (2023)	13,942	9438 (67.69%)	24,419 (39.11%)	2963
Grant et al. (2024) - Menta	30,112	13,296 (44.15%)	30,630 (49.06%)	3831
Grant et al. (2024) - Smoove	65,550	16,908 (25.79%)	41,897 (67.10%)	6191

of detected CNV regions by directly evaluating whether inferred copy number states behave as heritable genetic variants at the population level. Unlike commonly used validation strategies based on variant size, frequency, or overlap with external databases, transmission-level analysis quantifies the proportion of variance in copy number estimates attributable to additive genetic effects. Variants with high transmission levels therefore show consistent inheritance across generations, whereas regions with low transmission are more likely to reflect mapping artefacts or stochastic read-depth noise. [Figure 3](#) illustrates that although linked-read sequencing enables the discovery of a large number of putative SVs, these variants span a wide range of transmission levels, with a mean close to 0.5. Our results demonstrate that a substantial fraction of linked-read discovered SVs show weak or negligible evidence of Mendelian inheritance, highlighting that even technologies with enhanced sensitivity remain prone to artefactual calls. This observation underscores that linked-read sequencing, while superior to short-reads for SV discovery, still requires robust downstream validation to distinguish true inherited variants from technical noise.

High-transmission CNV regions identified in this study show strong reproducibility across independent datasets, reinforcing transmission level as a reliable benchmark of biological evidence. Among the 7218 CNV regions with transmission levels greater than 0.9, the vast majority overlapped CNV regions reported in one or more public resources, with more than 86% overlapping regions archived in DGVa and substantial concordance with other large cattle CNV studies. In contrast, CNV regions with lower transmission levels showed reduced overlap with external datasets, suggesting that poor reproducibility reflects technical noise rather than population specificity. The strong composition of high-transmission CNV regions in variants consistently observed across studies supports the principle that reproducibility across populations and platforms is a hallmark of true inherited structural variation. Importantly, the identification of a small set of novel high-transmission CNV regions using these robust inheritance patterns indicates that a lack of prior annotation does not preclude biological authenticity. This depth-based validation is particularly reliable for deletions, which LongRanger detects with

high sensitivity by leveraging barcode-overlap signals ([Marks et al., 2019](#)). This has also been reflected in the transmission level among all five categories in which deletions had the highest average transmission levels. Transmission-based filtering not only validates CNV regions technically but also yields biologically relevant subsets for interpretation. Of the 7,218 highly transmitted CNV regions (>0.9), many overlapped with genes, including exonic sequences, indicating functional importance. Notably, some high-transmission regions found did not overlap existing resources. Their inheritance patterns imply they are genuine variants and valuable for future functional studies as genomic data grows.

Structural variant discovery is inherently influenced by population composition and breed background, reflecting both genetic diversity and divergence relative to reference genome. In this study, SVs detected using linked-read sequencing showed clear evidence of breed and population effects. This is consistent with previous reports that CNV number, size distribution, and genomic location vary substantially between dairy breeds such as Holstein-Friesian and Jersey ([Reynolds et al., 2018](#)). In this study, the linked-read sequenced animals are prominent bulls that have been frequently involved in breeding schemes. Therefore, the population, which includes the short-read sequenced animals, is very similar and closely related to the original animals used to detect SV regions. In contrast, CNV regions that are rare, breed-specific, or poorly represented in the reference haplotypes are more difficult to detect and validate consistently. We found that over 50% of the detected regions overlap with DGVa, [Lee et al. \(2023\)](#), [Grant et al. \(2024\)](#). However, the lower overlap with [Bhati et al. \(2023\)](#) likely reflects breed composition differences, as their study included Fleckvieh animals, which are genetically more distant from the Holstein-Friesian and Jersey populations analysed here. Similarly, [Chen et al. \(2021\)](#) detected more SVs in Holstein animals compared to Jersey animals. These findings highlight that SV catalogues are not universally transferable across populations and emphasise the importance of population-matched discovery panels and breed-aware interpretation when generating and applying structural variant resources in livestock genomics.

A validated set of high-transmission CNV regions provides new opportunities to extend the genetic architecture captured in animal breeding programs beyond SNP variation. CNV regions with strong Mendelian inheritance are more likely to represent stable additive genetic effects, making them suitable candidates for inclusion in genomic prediction models. Incorporating such CNV regions alongside SNPs may improve the proportion of genetic variance explained, particularly for traits influenced by structural variation that is poorly tagged by SNP markers (Chen et al., 2021). High-confidence CNV regions could be modelled as additional loci, grouped into regional genomic features, or used to refine haplotype-based prediction frameworks. Breed-specific CNV catalogues further enable the capture of population-specific genetic effects, reducing reference bias and supporting more accurate estimation of breeding values. Together, these developments highlight the potential of validated CNV regions to improve the performance of genomic selection.

Conclusion

In this study, we combined linked-read sequencing-based discovery with population-scale validation to generate a high-confidence catalogue of copy number variation in New Zealand dairy cattle. Linked-read sequencing of 37 animals identified 62,438 putative autosomal CNV regions, which were subsequently genotyped in 2306 animals using short-read whole-genome sequencing. By estimating transmission levels using pedigree-based linear mixed models, we quantified Mendelian inheritance and distinguished biologically meaningful CNV regions from technical artefacts. A total of 7218 CNV regions exhibited high transmission levels (>0.9), most of which overlapped previously reported variants, while a small number represented novel regions. These results demonstrate that transmission-based validation provides a robust and complementary framework for CNV discovery and confirms the utility of integrating linked-read and short-read sequencing for structural variation studies in livestock populations. Our results highlight the influence of population composition and reference genome choice on CNV discovery and validation, consistent with growing evidence that structural variation is partly breed-specific. By providing a scalable, biologically grounded validation strategy, this work helps bridge the gap between structural variant discovery and practical application in genomic selection and lays the groundwork for further improving the biological interpretation of structural variation in cattle genomes.

Data availability statement

The list of CNV regions and their associated transmission level estimates generated in this study are provided in the [Supplementary Data Sheet 1](#), including genomic coordinates, CNV classifications, and transmission level values used in all downstream analyses and figures. The linked-read sequencing data used for structural variant discovery are available from the corresponding author upon reasonable request. The Illumina short-read whole-genome sequencing data analysed in this study comprise a larger internal

dataset, of which a subset was previously published by Reynolds et al. (2021) and is publicly available via the NCBI BioProject database under accession number PRJNA656361. Pedigree information used for population-scale validation are subject to confidentiality restrictions and are not publicly available. Requests for access to these data should be directed to the corresponding author and will be considered subject to appropriate data-sharing and confidentiality agreements.

Ethics statement

The requirement of ethical approval was waived by New Zealand Animal Welfare Act 1999 for the studies involving animals because data used in this study was generated as part of routine commercial activities outside the scope of that requiring formal committee assessment and ethical approval (as defined by the above guidelines). The studies were conducted in accordance with the local legislation and institutional requirements.

Author contributions

YW: Formal Analysis, Investigation, Methodology, Resources, Supervision, Visualization, Writing – original draft, Writing – review and editing. TN: Data curation, Formal Analysis, Investigation, Software, Writing – review and editing. TJJ: Data curation, Methodology, Software, Writing – review and editing. CC: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review and editing. BLH: Conceptualization, Funding acquisition, Methodology, Writing – review and editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This study received financial support from the NZ Ministry of Primary Industries, SFF Futures Program: Resilient Dairy–Innovative breeding for a sustainable dairy future (grant number: PGP06-17006).

Acknowledgements

We thank Mike Keehan for contributions to early data preparation. This research was made possible using the supercomputing resources of the New Zealand eScience Infrastructure (NeSI) platform (<https://www.nesi.org.nz>). In addition, we wish to thank the reviewers reviews for their constructive suggestions to improve the final manuscript.

Conflict of interest

Authors YW, TN, TJ, CC and BH were employed by Livestock Improvement Corporation (LIC; Hamilton, New Zealand), a commercial provider of bovine germplasm.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi:10.1101/gr.114876.110
- Bhati, M., Mapel, X. M., Lloret-Villas, A., and Pausch, H. (2023). Structural variants and short tandem repeats impact gene expression and splicing in Bovine testis tissue. *Genetics* 225, iyad161. doi:10.1093/genetics/iyad161
- Bickhart, D. M., and Liu, G. E. (2014). The challenges and importance of structural variation detection in livestock. *Front. Genet.* 5, 37. doi:10.3389/fgene.2014.00037
- Bickhart, D. M., Hou, Y., Schroeder, S., Alkan, C., Cardone, M. F., Matukumalli, L. K., et al. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.* 22, 778–790. doi:10.1101/GR.133967.111
- Butty, A. M., Chud, T. C. S., Cardoso, D. F., Lopes, L. S. F., Miglior, F., Schenkel, F. S., et al. (2021). Genome-wide association study between copy number variants and hoof health traits in Holstein dairy cattle. *J. Dairy Sci.* 104, 8050–8061. doi:10.3168/JDS.2020-19879
- Byrka-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* 185, 3426–3440.e19. doi:10.1016/j.cell.2022.08.004
- Chen, L., Pryce, J. E., Hayes, B. J., and Daetwyler, H. D. (2021). Investigating the effect of imputed structural variants from whole-genome sequence on genome-wide association and genomic prediction in dairy cattle. *Animals* 11, 1–16. doi:10.3390/ani11020541
- Chen, Y., Khan, M. Z., Wang, X., Liang, H., Ren, W., Kou, X., et al. (2024). Structural variations in livestock genomes and their associations with phenotypic traits: a review. *Front. Vet. Sci.* 11, 1416220. doi:10.3389/fvets.2024.1416220
- Couldrey, C., Keehan, M., Johnson, T., Tiplady, K., Winkelman, A. M., Littlejohn, M. D., et al. (2017). Detection and assessment of copy number variation using PacBio long-read and Illumina sequencing in New Zealand dairy cattle. *J. Dairy Sci.* 100, 5472–5478. doi:10.3168/JDS.2016-12199
- Duan, J., Zhang, J.-G., Deng, H.-W., and Wang, Y.-P. (2013). Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLOS ONE* 8, e59128. doi:10.1371/journal.pone.0059128
- Friedrich, S., Barbulescu, R., Helleday, T., and Sonhammer, E. L. L. (2020). MetaCNV - a consensus approach to infer accurate copy numbers from low coverage data. *BMC Med. Genomics* 13, 76. doi:10.1186/s12920-020-00731-y
- Gao, Y., Ma, L., and Liu, G. E. (2022). Initial analysis of structural variation detections in cattle using long-read sequencing methods. *Genes* 13, 828. doi:10.3390/genes13050828
- Grant, J. R., Herman, E. K., Barlow, L. D., Miglior, F., Schenkel, F. S., Baes, C. F., et al. (2024). A large structural variant collection in Holstein cattle and associated database for variant discovery, characterization, and application. *BMC Genomics* 25, 903. doi:10.1186/s12864-024-10812-2
- Haraksingh, R. R., and Snyder, M. P. (2013). Impacts of variation in the human genome on gene regulation. *J. Mol. Biol.* 425, 3970–3977. doi:10.1016/j.jmb.2013.07.015
- Harris, B. L. (2022). Genomic evaluations for crossbred dairy cattle. *JDS Commun.* 3, 152–155. doi:10.3168/jdsc.2021-0176
- Hou, Y., Liu, G. E., Bickhart, D. M., Cardone, M. F., Wang, K., Kim, E., et al. (2011). Genomic characteristics of cattle copy number variations. *BMC Genomics* 12, 127. doi:10.1186/1471-2164-12-127
- Kadri, N. K., Koks, P. D., and Meuwissen, T. H. E. (2012). Prediction of a deletion copy number variant by a dense SNP panel. *Genet. Sel. Evol.* 44, 7. doi:10.1186/1297-9686-44-7
- Karaoglanoglu, F., Ricketts, C., Ebre, E., Rasekh, M. E., Hajirasouliha, I., and Alkan, C. (2020). VALOR2: characterization of large-scale structural variants using linked-reads. *Genome Biol.* 21, 72. doi:10.1186/s13059-020-01975-8
- Keehan, M., and Couldrey, C. (2018). "Validation of synthetic long reads for use in constructing variant graphs for dairy cattle breeding," in *Proceedings of 11th World Congress of Genetics Applied to Livestock Production*.
- Ladeira, G. C., Pinedo, P. J., Santos, J. E. P., Thatcher, W. W., and Rezende, F. M. (2025). Detecting and characterizing copy number variation in a large commercial U.S. Holstein cattle population. *BMC Genomics* 26, 381. doi:10.1186/s12864-025-11536-7
- Lavrichenko, K., Johansson, S., and Jonassen, I. (2021). Comprehensive characterization of copy number variation (CNV) called from array, long- and short-read data. *BMC Genomics* 22, 826. doi:10.1186/s12864-021-08082-3
- Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9, e1003118. doi:10.1371/journal.pcbi.1003118
- Lee, Y.-L., Bosse, M., Takeda, H., Moreira, G. C. M., Karim, L., Druet, T., et al. (2023). High-resolution structural variants catalogue in a large-scale whole genome sequenced bovine family cohort data. *BMC Genomics* 24, 225. doi:10.1186/s12864-023-09259-8
- Leonard, A. S., Crysanto, D., Fang, Z. H., Heaton, M. P., Vander Ley, B. L., Herrera, C., et al. (2022). Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat. Commun.* 13, 3012. doi:10.1038/s41467-022-30680-2
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi:10.1093/bioinformatics/btr509
- Li, H. (2013). *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. doi:10.48550/arXiv.1303.3997
- Lin, J., Jia, P., Wang, S., Kusters, W., and Ye, K. (2023). Comparison and benchmark of structural variants detected from long read and long-read assembly. *Brief. Bioinform* 24, bbad188. doi:10.1093/bib/bbad188
- Liu, G. E. (2025). Exploring cattle structural variation in the era of long reads, pangenome graphs, and near-complete assemblies. *J. Anim. Sci. Biotechnol.* 16, 158. doi:10.1186/s40104-025-01294-7
- Liu, X., Chen, W., Huang, B., Wang, X., Peng, Y., Zhang, X., et al. (2024). Advancements in copy number variation screening in herbivorous livestock genomes and their association with phenotypic traits. *Front. Vet. Sci.* 10, 1334434. doi:10.3389/fvets.2023.1334434

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2026.1856199/full#supplementary-material>

SUPPLEMENTARY FIGURE S1

Relationship between the average sequencing depth and the number of detected structural variant regions detected in 37 linked-read sequenced animals. Three bulls with highest/lowest sequence depth and number of detected structural variants are highlighted.

SUPPLEMENTARY FIGURE S2

Relationship between chromosome length and the number of detected CNV regions across the cattle genome.

SUPPLEMENTARY TABLE S1

Summary of publicly available cattle CNV datasets used for comparison, including data type, sample size, breeds, analytical methods, number of detected regions, and database availability.

SUPPLEMENTARY DATA SHEET 1

The list of CNV regions and their associated transmission level estimates generated in this study.

- Marks, P., Garcia, S., Barrio, A. M., Belhocine, K., Bernate, J., Bharadwaj, R., et al. (2019). Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* 29, 635–645. doi:10.1101/gr.234443.118
- Nguyen, T. V., Vander Jagt, C. J., Wang, J., Daetwyler, H. D., Xiang, R., Goddard, M. E., et al. (2023). In it for the long run: perspectives on exploiting long-read sequencing in livestock for population scale studies of structural variants. *Genet. Sel. Evol.* 55, 9. doi:10.1186/s12711-023-00783-5
- Oliveira, H. R., Chud, T. C., Oliveira, G. A., Hermisdorff, I. C., Narayana, S. G., Rochus, C. M., et al. (2024). Genome-wide association analyses reveals copy number variant regions associated with reproduction and disease traits in Canadian Holstein cattle. *J. Dairy Sci.* 107 (9), 7052–7063. doi:10.3168/JDS.2023-24295
- Reynolds, E., Couldrey, C., Keehan, M., Tiplady, K., Johnson, T., Garrick, D. J., et al. (2018). “Discovery of breed-specific genomic content in beef and dairy breeds,” in *Proceedings of 11th World Congress of Genetics Applied to Livestock Production*.
- Reynolds, E. G. M., Neeley, C., Lopdell, T. J., Keehan, M., Dittmer, K., Harland, C. S., et al. (2021). Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. *Nat. Genet.* 53, 949–954. doi:10.1038/s41588-021-00872-5
- Rice, A. M., and McLysaght, A. (2017). Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat. Commun.* 8, 14366. doi:10.1038/ncomms14366
- Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., et al. (2020). *De novo* assembly of the cattle reference genome with single-molecule sequencing. *GigaScience* 9, 1–9. doi:10.1093/gigascience/giaa021
- Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J. M., et al. (2017). Genome-wide reconstruction of complex structural variants using read clouds. *Nat. Methods* 14, 915–920. doi:10.1038/nmeth.4366
- Suvakov, M., Panda, A., Diesh, C., Holmes, I., and Abyzov, A. (2021). CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. *GigaScience* 10, 1–9. doi:10.1093/gigascience/giab074
- VanRaden, P. M. (2020). *Symposium review: how to implement genomic selection*. *J. Dairy Sci.* 103, 5291–5301. doi:10.3168/jds.2019-17684
- Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 39, 1348–1365. doi:10.1038/s41587-021-01108-x
- Xu, L., Cole, J. B., Bickhart, D. M., Hou, Y., Song, J., VanRaden, P. M., et al. (2014). Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC Genomics* 15, 683. doi:10.1186/1471-2164-15-683
- Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481. doi:10.1146/annurev.genom.9.081307.164217
- Zhao, X., Collins, R. L., Lee, W.-P., Weber, A. M., Jun, Y., Zhu, Q., et al. (2021). Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.* 108, 919–928. doi:10.1016/j.ajhg.2021.03.014