Molecular QTL are enriched for structural variants in a cattle long-read cohort

Xena Marie Mapel[1,*], Alexander S. Leonard[1,*], Hubert Pausch[1,#]

[1] Animal Genomics, ETH Zurich, Universitaetstrasse 2, 8092 Zurich, Switzerland

[*] equal contribution

Corresponding author: hubert.pausch@usys.ethz.ch

Abstract

Sequencing cohorts with long-read technology is crucial to understand the impact of structural variants (SVs) on complex traits. Here, we obtained 4.86 terabases of HiFi reads with an average read N50 of 16.3 Kb from 120 *Bos taurus taurus* bulls, yielding a mean coverage depth of 13.5-fold. We genotyped 23.8 M small variants (SNPs and short INDELs) and 79.3 k SVs to perform association testing with molecular phenotypes derived from a subset of 117 bulls with total RNA sequencing data from testis tissue. We identified 27.3 k molecular QTL (molQTL) including 316 for which SVs were the most significant variant. This corresponds to a 2.1- and 5.6-fold enrichment of SVs among expression and splicing QTL, respectively. When considering SVs in perfect LD with the lead small variant, the enrichment increased to 6.1- and 12-fold for expression and splicing QTL in testis, respectively. Imperfect genotyping for SVs limited our ability to detect all SV molQTL, suggesting that the true enrichment of SVs among molQTL may be even higher. These results demonstrate that SVs have a profound impact on gene expression and splicing variation but highlight the necessity of improved SV genotyping to fully leverage long-read sequencing cohorts for dissecting complex traits.

Introduction

Understanding the genetic basis of complex traits and diseases requires comprehensive exploration of genomic variation within and between individuals[1,2]. Genome-wide association studies (GWAS) have identified numerous loci linked to specific traits but require accurate genotypes to detect putative causal variants and estimate allele substitution effects[3]. Single nucleotide polymorphisms (SNPs) derived from microarrays or short-read sequencing serve as the predominant source of genomic data for GWAS[4–6]. Structural variants (SVs),

specifically, insertions, deletions, inversions, and duplications larger than 50 bp, contribute more to genomic variation between individuals than SNPs and other small variants[7]. Short-read sequence data has inherent limitations in detecting and genotyping SVs and thus, until recently, a substantial portion of genomic variation remained undercharacterized[8–10].

Long-read sequencing enables the analysis of DNA fragments orders of magnitude longer than short reads, improving SNP calling in challenging regions[11] and offering the resolution necessary to identify and genotype SVs[12–14]. Studies in humans and plants have begun to explore SV diversity at the cohort level using long-read sequencing[15–17]. While these efforts remain limited in number, they have provided compelling evidence that SVs disproportionately contribute to phenotypic variation. These studies also demonstrated that accurately genotyping large and complex SVs remains challenging, even with long and highly accurate reads[18,19]. However, benchmarking results from human studies may not generalize to other species, since most of the widely applied variant genotyping algorithms are trained on human data and may perform suboptimally elsewhere[11,20,21].

Research on SVs in cattle has primarily relied on pangenome analyses[22–24], short-read sequencing[25–27], and quantitative trait loci (QTL) fine-mapping[28,29]. These studies have revealed several SVs with large impacts on phenotypic traits including polledness[30], coat colour variation[29,31–33], milk production[34], mastitis susceptibility[28], and fertility[35]. The discovery of such impactful SVs highlights their biological and economical relevance underscoring the critical need to systematically investigate them in genome-wide analyses. Association testing with molecular phenotypes and SVs called from moderate-coverage short-read sequence data (average of 13.8-fold) has identified hundreds of SVs that impact gene expression and splicing in cattle testis tissue[23,25]. However, short-read datasets are limited in their ability to accurately identify and genotype SVs, particularly large or multiallelic insertions and deletions[25–27]. Cohort-scale long-read sequencing in cattle has yet to be conducted to reveal the true impact of SVs on gene expression and splicing.

Here, we assess genetic variation in a cohort of 120 bulls from moderately covered HiFi reads. We identify and genotype 23.8 M small variants and 79.3 k SVs larger than 50 bp representing the vast majority of genetic variants segregating in that population. We perform molecular QTL (molQTL) mapping with deeply sequenced total RNA from testis tissue of the same individuals, identifying more than 27 k gene expression and splicing QTL. Our findings show that molQTL are enriched for SVs but also highlight that obtaining accurate genotypes for structural variants remains challenging.

Results

HiFi sequencing, alignment, and variant calling of 120 cattle generates an exhaustive set of variants

We resequenced a biobanked cohort of 120 *Bos taurus taurus* samples of primarily Braunvieh (BV) ancestry with existing short read sequencing data[36] with long reads to assess both small variant and SV diversity. Short-read data from the same cohort were previously utilized to genotype small variants and SVs for association analyses with molecular phenotypes[25,36,37]. We collected 4.86 terabases of HiFi reads sequenced from 49 and 41 Sequel IIe 8M and Revio 25M SMRT cells, respectively. The mean depth of coverage was 13.8-fold and 13.5-fold for the existing Illumina and newly created HiFi reads, respectively (Figure 1a), enabling a fair comparison for variant discovery, although coverage differed substantially between both sequencing technologies for some individuals. The mean read N50 was 16.3 Kb with a mean Phred quality score of 33.7, demonstrating the expected strong negative correlation between read length and quality for HiFi reads (Pearson's r: -0.80, $p=5.87 \times 10^{-37}$; Supplementary Figure 1).

We aligned the 120 HiFi and Illumina samples to the ARS-UCD2.0 *Bos taurus* reference genome, which included the new telomere-to-telomere assembly of a Y chromosome from a Wagyu bull, whereas the remaining genome is from a Hereford cow[38]. Alignment depth of both the short and long-read sequencing was comparable across the autosomes, with only minor increase in the fraction of autosomal sequence covered by alignments suitable for variant calling for HiFi reads (99.4% versus 99.0%; Figure 1b). Conversely, the alignment improvement was noticeable for the X chromosome (90.0% versus 86.0%) and substantial for the Y chromosome (58.9% versus 32.0%). In the former case, this was largely due to the improved mapping of HiFi reads distinguishing high sequence similarity of X:19309231-28499819 to the unplaced contig NKLS02002208.1, while the latter was due to a large fraction of sequence containing higher order repeats, which was almost completely inaccessible with short-read sequencing (Supplementary Figure 2a,b). However, the male-specific region of the Y chromosome (MSY) still displayed uneven and inconsistent alignment with the HiFi reads due to the repeat structure typically exceeding the length of the sequenced long reads which poses substantial challenges to unambiguous read alignment.

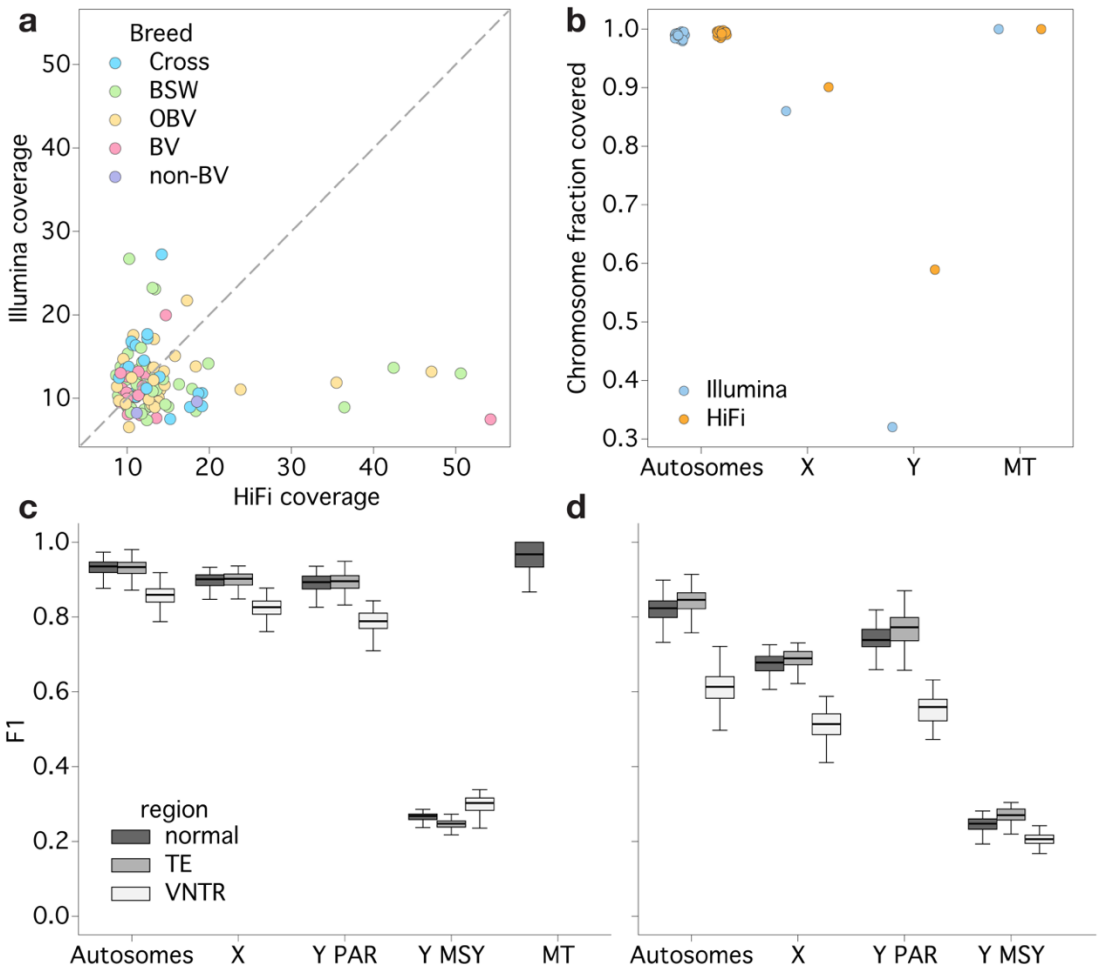

***Figure 1**. Comparison of alignment coverage and variant accuracy between Illumina and HiFi reads. **(a)** HiFi and Illumina sequencing depth was roughly comparable across bulls with breeds assigned as Original Braunvieh (OBV), Brown Swiss (BSW), ambiguous Braunvieh ancestry (BV), crosses between Braunvieh and a non-Braunvieh breed (Cross), or animals without Braunvieh ancestry (non-BV). **(b)** Fraction of chromosomes covered by at least two reads with MAPQ≥5 to enable variant calling. F1 score for SNPs **(c)** and indels **(d)**, taking the HiFi variants as truth. Variants are stratified by regions annotated as tandem repeats (VNTR), transposable elements (TE; SINE/LINE/LTR), or neither (normal). The Y chromosome is separated by the pseudo-autosomal region (Y PAR) and the male-specific Y (Y MSY).*

We called small variants with DeepVariant separately for HiFi and Illumina reads, again finding minor differences in the number of detected variants and called genotypes on autosomes and larger differences on sex chromosomes (Table 1). A substantially higher proportion of singletons among the Illumina-only variants may indicate an elevated variant discovery error rate in short-read alignments (Supplementary Figure 2c,d). The increased number of variants called on X from the HiFi alignments was evenly distributed across the chromosome, while the larger number of variants discovered on Y primarily resulted from the increase in bases covered with HiFi reads in the MSY, with a substantial increase in variants called in newly accessible regions that were almost entirely uncalled from short read sequencing. Given all the samples are male, we specified that variants could have

heterozygous genotypes in the pseudo-autosomal region (PAR) of the Y chromosome, while they could only be hemizygous in the X or MSY regions. However, there were “real” signals of heterozygous variants in the hemizygous region, primarily due to copy number variants of amplicon genes (e.g., *TSPY*, *HSFY2*, and *RBMY*) relative to the reference, where the variant allele frequency could estimate the copy number in addition to coverage-based estimates (Supplementary Figure 3).

We hypothesized that the improved mappability of HiFi reads and the larger number of called variants provide access to previously unidentified genetic variants that may alter gene products or are deleterious to protein function. We assessed potential functional consequences of small variants with the Ensembl Variant Effect Predictor (Table 1), finding 247 and 166 biallelic SNPs with “HIGH” impacts respectively private to HiFi- and Illumina-based variant calls, with 2,013 common to both. After manual inspection of 15 HIGH impact variants private to each group (Supplementary Table 1), we identified poor mapping of Illumina reads as the primary cause for the discrepancy between the variants called from short and long reads. All but one Illumina-only HIGH impact variant were likely genotyping artifacts as evidenced by implausible insert sizes and interchromosomal-mapping reads, with the sole exception of a singleton variant missed by the HiFi reads due to an uneven balance of alleles (7:1 read depth for the alleles). Conversely, all HiFi-only HIGH impact variants appeared correct, mostly missed by Illumina reads due to mapping quality of 0 or highly diverged regions potentially leading to local reference bias (Supplementary Figure 4).

***Table 1**. Variants called by DeepVariant from the HiFi or Illumina alignments. The Y chromosome is split into the Y pseudo-autosomal region (Y PAR; where heterozygous genotypes are allowed) and male-specific Y (MSY; where only hemizygous genotypes are allowed). The number of variants identified by the Ensembl Variant Effect Predictor as HIGH impact are given in parentheses.*

	Autosomes	**X**	**Y PAR**	**MSY**	**MT**
Illumina	22,636,961	383,835	102,570	20,177	379
	(2,198)	(20)	(4)	(0)	(8)
HiFi	23,163,376	444,306	112,420	69,402	386
	(2,273)	(24)	(6)	(1)	(8)
Increase	2.3%	15.8%	9.6%	244.0%	1.8%
	(3.4%)	(20.0%)	(50.0%)	(NA)	(0.0%)

Previous evaluations suggest that, for small variants called from samples with moderate sequencing depth (8-15x), the misgenotyping rate may be as high as 2-5%[1,39,40]. For the present dataset, none of the 120 samples had microarray-called genotypes, thereby precluding a formal assessment of genotyping accuracy against a truth set. However, since

previous research showed that DeepVariant calls cattle genotypes accurately from short reads[21], we considered the short-read variants as a truth and the HiFi variants as query. We found generally high genotyping concordance for SNPs on the autosomes (F1=0.93±0.02), although there was a moderate drop (F1=0.86±0.03) in genomic regions classified as tandem repeats (Figure 1c). There was no such drop for regions containing transposable elements, suggesting that there was sufficient variation within different transposable elements for short reads to correctly align, while they struggle in long tandem repeats with little motif variation. Variants genotyped on the X chromosome (F1=0.84±0.04) and the PAR of the Y chromosome (F1=0.85±0.02) were less concordant across all types of regions, while the limited number of short-read-based variant calls in the MSY region led to a substantial drop in F1 score (F1=0.27±0.02). Indels behaved similarly, with an even more pronounced drop in accuracy around tandem repeats (Figure 1d). Given the observed improvements in long-read alignments and more small variants called genome-wide, particularly in challenging regions, we only retained the HiFi-based small variants for downstream use.

We also called structural variants from the HiFi read alignments with sniffles2, finding 75,164 and 4,111 SVs on the autosomes and sex chromosomes, respectively. We found approximately 23.9±0.6 k SVs per genome, with only on the order of hundreds of novel SVs for each additional individual after 100 samples (Supplementary Figure 5). This suggests that we likely captured the bulk of non-rare SVs present in the Braunvieh breed. SVs were overrepresented in regions annotated as tandem repeats, totalling just 3.2% of the genome, with 30.2% of SVs having at least half of their sequence contained within these regions. We observed a disproportionately large number of SVs of length ~300 bp, ~1.3 Kb, ~5.5 Kb, and ~8.4 Kb (Figure 2a), corresponding to known transposable elements like Bov-A2 (SINE), BTLTR1B (LTR), and BovB (LINE). Given the average read N50 of 16 Kb for this cohort, there is likely a small number of longer SVs that remains unidentified (Supplementary Figure 6). We further found that 7,944 SVs (10.02%) were present in only one sample (6,988 singletons and 956 doubletons), ranging between 6 and 573 SVs unique to each sample. We found 37,380 novel autosomal SVs compared to a previous autosome-only, pangenome-derived SV panel constructed from one Piedmontese and 15 Braunvieh haplotypes not included in the HiFi-cohort analysed here, including 19,507 SVs with a minor allele frequency (MAF) between 1% and 15% (Figure 2b), as well as 376 duplications and 422 inversions which cannot be easily genotyped through the previously applied *k*-mer approach[37]. The 11,280 SVs present only in the pangenome-panel are likely private to those animals or related to differences in assembly-versus-alignment based SV calling.

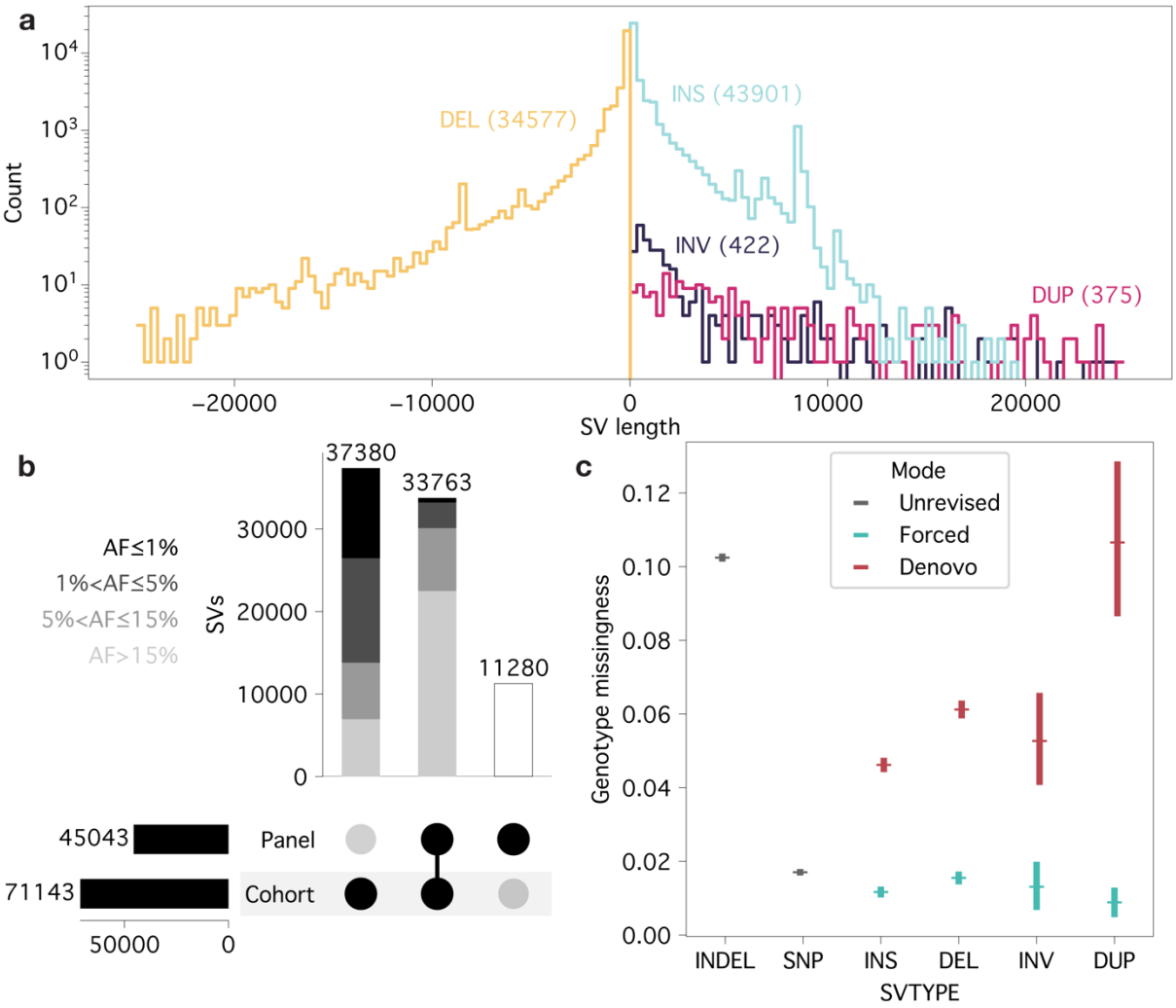

***Figure 2**. Structural variant calling in a long-read cohort. **(a)** SV size distribution across the four SV classes (DEL – deletion, INS – insertion, INV – inversion, DUP – duplication) examined. Large spikes in insertions and deletions correspond to known transposable elements. **(b)** The majority of SVs which were only discovered through the long-read cohort compared to a pangenome panel have low minor allele frequency (AF), whereas those previously discovered are typically common SVs. Allele frequency was calculated from the cohort, and so the panel-only SVs have no assigned allele frequency. **(c)** Forced genotyping of SVs substantially reduces per-site missingness to similar levels identified in SNP variant calling. Markers represent the mean value and error bars indicate the 95% confidence intervals.*

During sniffles2 joint-calling, “similar” SVs were merged into single alleles, collapsing multiallelic SVs. This overwhelmingly homogenised SVs that might only differ by small differences in length or position (Supplementary Figure 7). Approximately 82% of the merged alleles had allele length standard deviations below 10, while only approximately 6% were above 100. However, SVs overlapping tandem repeats were particularly affected by merging alleles, either incorrectly treating different tandem repeat allele lengths as distinct SVs (rather than distinct alleles) or treating different tandem repeat allele lengths as a single

consensus allele (Supplementary Figure 8). Several large tandem repeats contained multiple SVs, primarily differing only by the start coordinate (corresponding to inserting/removing identical motifs at different locations within the tandemly duplicated sequence), which were not merged. On the other hand, smaller tandem repeats tended to be merged too aggressively, picking a consensus allele length which did not accurately reflect the copy-number diversity present in individual, pre-merged alleles.

Joint-calling SVs across the entire cohort resulted in a relatively high proportion of missing genotypes (mean of 5.3%) per-site, with duplications having the highest missingness (mean of 10.6%). Force-calling candidate SVs with sniffles2 from the original alignments substantially reduced the missingness to a mean of 1.3% (with the lowest mean missingness now observed in duplications at 0.9%), slightly lower than the missingness of joint-called SNPs (Figure 2c). Genotypes that remained missing after force-calling were significantly associated with low alignment coverage ($p=5.0 \times 10^{-23}$, Supplementary Figure 9a) and, to a much lesser extent, mean read length ($p=2.9 \times 10^{-6}$) after conducting a Type II ANOVA. Missingness was elevated at the start and end of chromosomes, roughly corresponding to centromeric and telomeric regions in the acrocentric cattle autosomes (Supplementary Figure 9b), with other peaks corresponding to regions known to be highly polymorphic (e.g., BoLA on BTA23) or challenging for alignment (e.g., large segmental duplication on BTA10). Approximately 4% of the SV genotypes (380,911 genotypes across 54,561 unique SVs) changed during force-calling. The overwhelming majority of these changes (89.7%) were newly filled missing values (where there was potentially enough read support to force-genotype the SV despite not initially calling it), while 9.6% were changes between non-missing genotypes, and the remaining 0.7% changed from non-missing to missing genotypes (Supplementary Figure 10a). A disproportionately high number of non-missing genotype changes were within centromeric-like regions (the first 200 Kb of chromosomes, corresponding to 0.2% of the genome), accounting for 11% of these unexpected changes. Other non-missing genotype changes occurred in regions where even manual assignment of genotypes was not obvious from the alignments, with instances of force-calling either improving or worsening genotype accuracy (Supplementary Figure 11). Force-called genotypes have been reported to be slightly less concordant than joint-called genotypes[41], although this negative consequence is very likely outweighed by the vast decrease in missingness.

Given the near-complete catalogue of SVs established for the cohort, we further examined the linkage disequilibrium between small variants and SVs. Even with the improved resolution of small variants within tandem repeats, we still found SVs overlapping tandem repeats as the most poorly tagged class of SVs. Over half of SVs (52%) containing tandem

repeats were not in high linkage disequilibrium (r^2>0.8) with any small variant within a ±1 Mb window, compared to only 17% of SVs which did not contain tandem repeats or transposable elements.

The absence of a validated truth set of SV genotypes prevents a direct assessment of the SV genotyping accuracy achieved with sniffles2. Nonetheless, the autosomal per-individual heterozygosity was significantly higher for small variant (23.99%) than SV genotypes called from the HiFi reads regardless if the de-novo (22.57%, Welch's t-test p-value = 3.45×10^{-11}) or force-called (21.96%, Welch's t-test p-value = 4.99×10^{-19}) SV genotypes were used (Supplementary Figure 10b,c). The discrepancy between small variant- and SV-based heterozygosity is less evident in six samples with more than 30-fold HiFi coverage (de-novo: 23.57% vs. 22.69%, Welch's t-test p-value = 0.07; force-called: 23.57% vs. 22.52%, Welch's t-test p-value = 0.04), suggesting that high sequencing depth is required to comprehensively call heterozygous genotypes with sniffles2 and that heterozygous under-calling is more pronounced in force-called genotypes.

molQTL mapping with HiFi data reveals the influence of structural variation on gene expression and splicing

The contribution of SVs detected from long-read sequencing to complex trait variability has not been investigated in large cohorts of cattle. To do so, we assessed impacts of small variants and SVs on molecular phenotypes through *cis*-molQTL mapping using deeply sequenced total RNA from testis tissue of the same bulls that was available from previous work[36]. After quality control on the RNA sequencing data, we considered 117 bulls for expression QTL (eQTL) and splicing QTL (sQTL) identification. Molecular phenotypes were estimated for 24,281 expressed genes and 16,975 spliced genes (corresponding to 50,853 intron clusters and 207,773 splice junctions), enabling association testing with 20,308,853 small variants (SNPs and small INDELs) and 61,861 SVs that were within 1 Mb of a molecular feature's start site and had a MAF of at least 1% and had non-missing genotypes in more than 90% of the samples (Table 2). The applied filtering substantially altered the characteristics of the variant set used for molQTL mapping, reducing the average per-site missingness to 1.21 ± 1.80 for small variants and 0.03 ± 1.07 for SVs. Among the variants included in the association analyses, 43.1% of small variants and 88.5% of SVs exhibited complete genotype data across all 117 samples.

***Table 2**. Results for molecular QTL mapping with small variants and SVs. e/sGenes (gene that contains at least one eQTL or sQTL) and e/sQTL (independent-acting signals) include both small variants and SV. e/sVariants are defined as a variant that passes the gene's significance threshold, with the total number of unique variants listed in parentheses. Top variants are defined as the most significant variant for each independent-acting eQTL or sQTL, with the number of unique variants listed in parentheses.*

Variant type	Tested variants	eGenes	eQTL	eVariants (unique)	Top variants (unique)	sGenes	sQTL	sVariants (unique)	Top variants (unique)
Small variants	20,308,853	12,584	16,664	5,866,009 (3,612,237)	16,539 (16,334)	7,567	10,611	6,363,647 (2,357,818)	10,400 (10,359)
SVs	61,861			15,218 (9,618)	105 (102)			14,351 (5,535)	211* (176)

* 136 SV sQTL contained an SV with the same p-value as the top small variant

Half of the expressed genes had at least one eQTL (12,584 genes; eGenes), corresponding to 16,644 independent-acting eQTL and 3,621,855 variants that passed the nominal significant threshold (eVariants; Table 2). This included 261 eQTL (218 eGenes) on the X chromosome and 15 eQTL (13 eGenes) on the Y chromosome. Approximately 40% of eGenes had at least one eVariant that was an SV. SVs were lead variants for 105 eQTL (hereafter referred to as SV eQTL; Table 2), i.e. they either had the smallest p-value or the same p-value as the top small variant for an eGene. Three SVs were lead variants for multiple eGenes, including an 85,550 bp duplication on chromosome 8 (8:103486033–103571583) that was associated with *ORM1* and *COL27A1* expression. This SV was previously described as a likely causal variant for an increased expression of *ORM1* in liver tissue[26]. SV eQTL were depleted in intergenic regions and enriched in exons, CpG islands, promoters, and enhancers (Figure 3a; Supplementary Table 2). We identified 23 SV eQTL that were not in strong LD (r^2 >= 0.8) with a small variant, including a deletion that was associated with the expression of *ATXN7L3B* on chromosome 5. Notably, this 218 bp deletion was seven orders of magnitude more significant than the next eVariant (SV nominal p-value = 1.55×10^{-17}; Figure 3b,c).

Nearly half of the alternatively spliced genes possessed sQTL (Table 2). We detected at least one sQTL for 10,388 intron clusters (23,292 splice junctions) within 7,567 genes (sGenes), which totalled to 2,363,353 significant variants (sVariants) and 10,611 independent-acting sQTL. The X chromosome had 80 sQTL for 77 genes while the Y chromosome had 9 sQTL for 7 genes. Approximately 9% of the SVs considered for

association testing were sVariants, totalling to 5,535 SV sVariants. SVs were lead variants (hereafter referred to as SV sQTL; Table 2) for 211 independent-acting sQTL for 173 genes and included 176 unique variants. One SV was associated with multiple intron clusters within the same gene, while 30 SVs were associated with multiple junctions from the same cluster. Only one SVs was associated with multiple sGenes; a 783 bp deletion on chromosome 25 (17,283,696 bp) was an SV sQTL for *VPS35L*, where the SV was located within an intron, and an sQTL for *KNOP1*, which was approximately 23 Kb upstream of the variant. SV sQTL were enriched in introns, exons, and splice sites, but depleted in intergenic regions (Supplementary Table 2). Approximately 16% of SV sQTL were not in strong LD with a small variant (r^2 >= 0.8), though some of these SVs were much more significant than the next sVariant, such as a 514 bp deletion on chromosome 23 that was associated with alternative splicing of *ADGRF1* (Figure 3d,e). One SV sQTL was not tagged (r^2 < 0.2) by any nearby small variants—a 120 bp insertion on chromosome 21 which was the only variant that passed the significance threshold for *LOC112443211*. The significant SV and gene both resided within a complex region near the beginning of the chromosome with over 150-fold coverage (at least 10 times higher than average) (Supplemental Figure 12). The lack of LD with nearby SNPs is almost certainly driven by the SV and gene residing in a poorly resolved region of the reference genome, although the sQTL may still reflect a valid association.

We found that SV eQTL and SV sQTL exhibited similar characteristics. Specifically, low frequency variants had larger effect sizes (Figure 3f,g), as did variants that were within the feature's boundary (gene for eQTL, intron cluster and gene for sQTL; Supplemental Figure 13). Both eQTL and sQTL were underrepresented among low-frequency variants, reflecting the reduced statistical power to detect rare molecular QTLs (Figure 3h). This underrepresentation was stronger for SVs than small variants. We observed no statistical difference in effect size between SVs and small variants for eQTL ($p = 0.23$; Figure 3j), but insertions and deletions had slightly smaller effect sizes than small variants for sQTL ($p = 2.7 \times 10^{-8}$; Figure 3j). Regardless, top variants for eQTL and sQTL were 2.1 ($p = 9.29 \times 10^{-11}$) and 5.6-fold ($p = 1.92 \times 10^{-70}$) enriched for SVs, respectively, highlighting the functional relevance of SVs. Manual inspection of molQTL regions and HiFi alignments confirmed that the SV top variants were genuine compelling causal candidates.

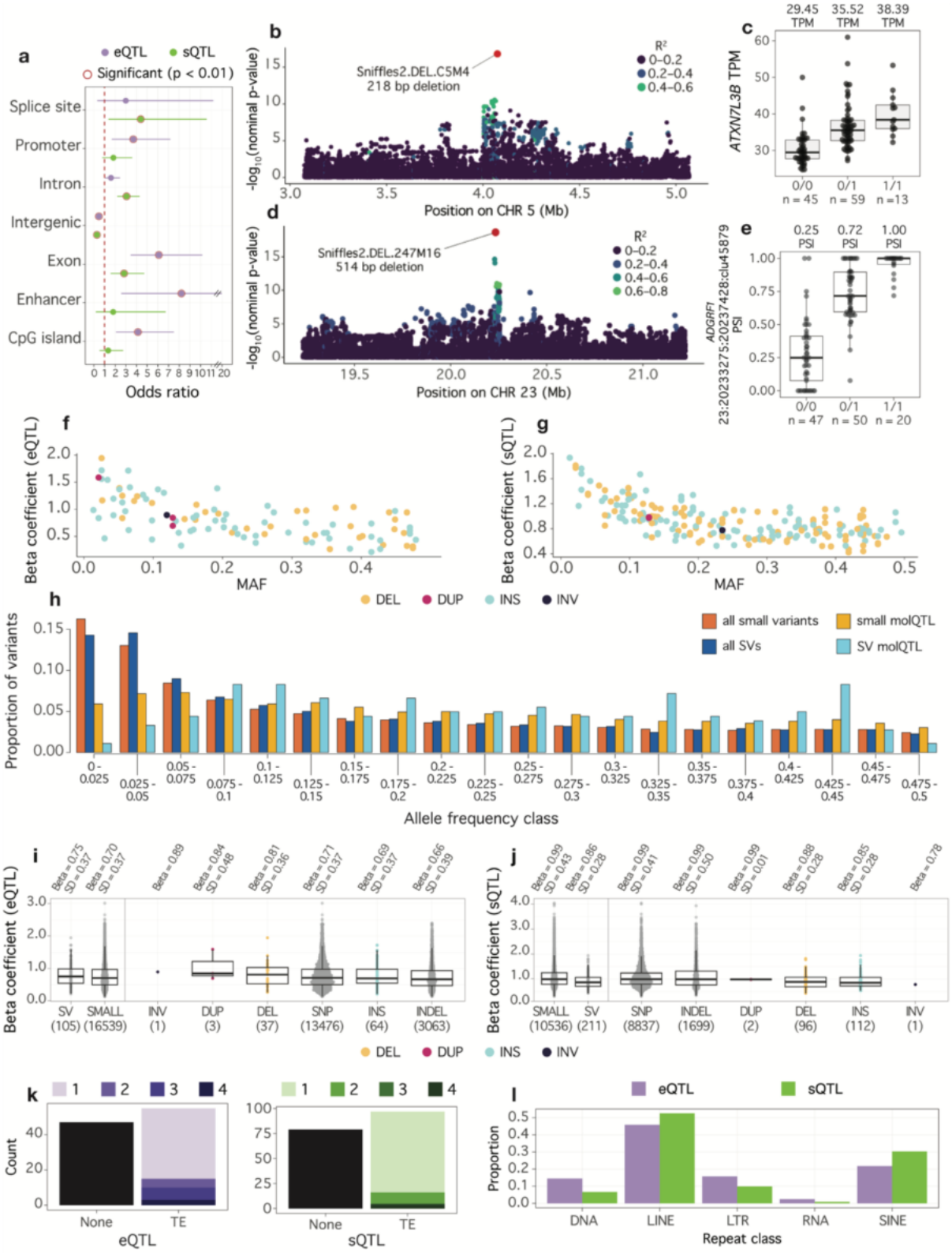

Figure 3. SV molQTL characteristics. ***(a)*** *Enrichment of SV molQTL in functional elements. Odds ratios and p-values were inferred with a Fisher's test, with significant (p < 0.01) observations circled in red.* ***(b)*** *Manhattan plot for a poorly tagged SV eQTL that was associated with expression of* ATXN7L3B. *Variants are coloured by linkage (r^2) with the SV.* ***(c)*** *Boxplot of* ATXN7L3 *expression (TPM) across genotypes of the SV eQTL. Median TPM values for each genotype are reported above, and number of samples belonging to each genotype are reported below.* ***(d)*** *Manhattan plot for a poorly tagged SV sQTL that was associated with splicing of* ADGRF1. ***(e)*** *Boxplot of* ADGRF1 *intron usage (PSI) across genotypes of the SV sQTL. Median PSI values for each genotype are reported above, and number of samples belonging to each genotype are reported below.* ***(f/g)*** *Minor allele*

*frequency (MAF) of each SV eQTL **(f)** and sQTL **(g)**, and effect size magnitude, coloured by variant type. **(h)** molQTL are depleted for rare variants, particularly for SVs. Colours correspond to variant type (SV or small variant) and whether the variant was a top variant. **(i/j)** Effect size magnitude for eQTL **(i)** and sQTL **(j)** across variant types. Left panel shows all SV (>= 50 bp) and small variant (< 50 bp) molQTL, while the right panel is separated by specific variant types. Median effect size (Beta) and standard deviation (SD) is reported above, and number of variants within each category is below within parentheses. **(k)** Number of SV molQTL with transposable elements (TEs). Proportion is coloured by number of different TE classes for each SV eQTL or sQTL. **(l)** Proportion of TE SV molQTL across repeat classes.*

Approximately 55% of the SV molQTL contained the full or partial sequence of at least one transposable element (Figure 3k). LINEs were the most common transposable element class among SV molQTL; however, they were underrepresented compared to their overall genome-wide abundance (Figure 3l; Supplementary Table 3). LTRs were also depleted when considering their overall abundance, while DNA and RNA transposable elements were slightly enriched among SV molQTL, though not significantly (Supplementary Table 3). There was no statistical difference in effect size across transposable element classes for both SV molQTL; though, when considering classes with more than two observations, median effect size was slightly larger for RNA and LTR SV eQTL and LTR and SINE SV sQTL. Only 8% and 12% of SV eQTL and SV sQTL contained tandemly repeated motifs, which was lower than the overall proportion of SVs containing such repetitive sequence. However, we encountered cases where incorrectly merged multi-allelic tandem repeats were omitted as potential SV molQTL (Supplemental Figure 14).

Imperfect genotyping led to an underestimation of SV molQTL

Inspection of the molQTL summary statistics revealed many SVs that appeared as strong putative candidate causal variants but had slightly larger p-values than the top small variant. There were 167 eQTL and 168 sQTL where a small top variant was in perfect LD ($r^2 = 1.0$) with at least one SV that had a larger p-value (Figure 4a,c). This difference in p-values between two variants in perfect LD was often a result of an unbalanced proportion of missing genotypes. Nearly all (97%) of these 335 molQTL had at least one “missing” call for the most significantly associated small variant. Only 17 of the 335 molQTL had a “missing” call for the SV in LD.

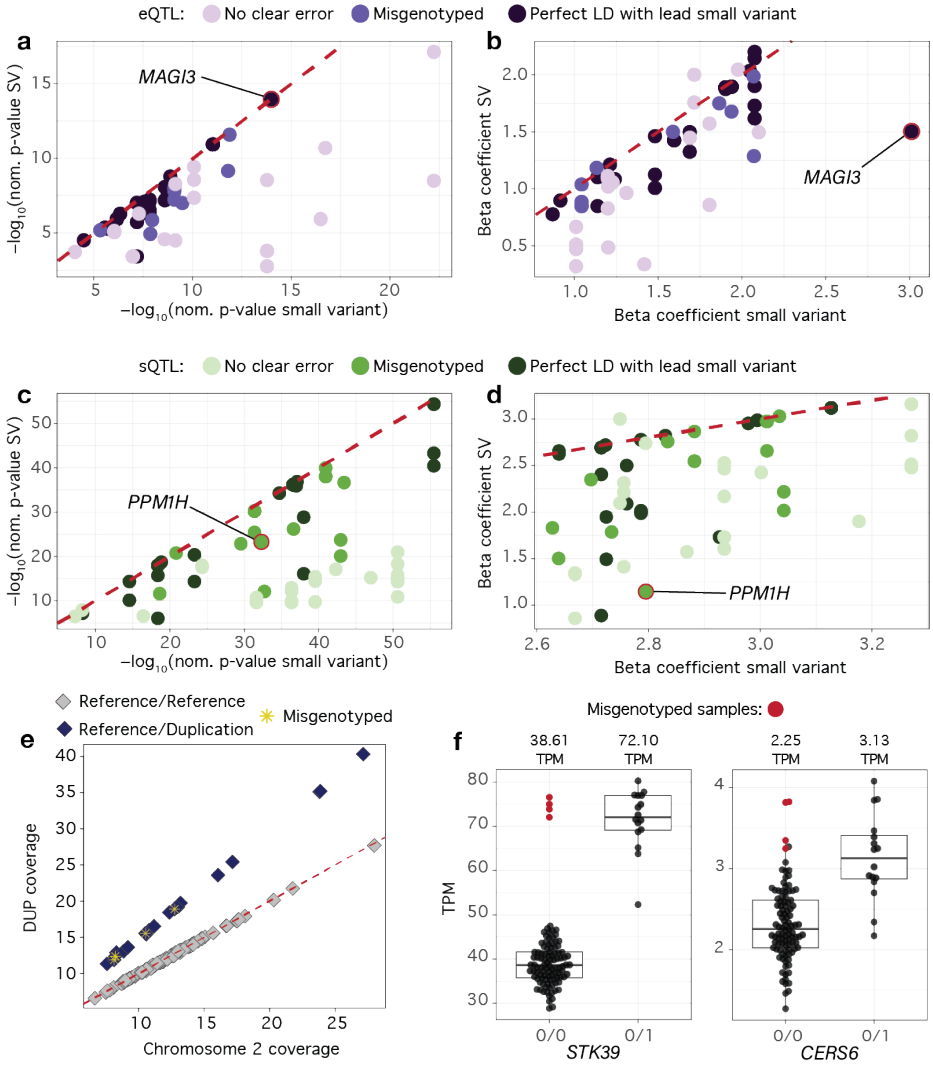

Figure 4. Misgenotyping of SVs impacts SV molQTL identification ***(a/c)*** *Nominal p-value of top small variant and significant SVs for large effect eQTL* ***(a)*** *and sQTL* ***(c)****. Colours correspond to whether the SV was in perfect linkage disequilibrium (LD) with the top small variant, had evidence of misgenotyping or if we did not identify an obvious error.* ***(b/d)*** *Nominal p-value of top small variant and significant SVs for large effect eQTL* ***(b)*** *and sQTL* ***(d)****. Colours correspond to whether the SV was in perfect LD with the top small variant, had evidence of misgenotyping or if we did not identify an obvious error.* ***(e)*** *Comparing sample-wise average coverage across chromosome 2 and coverage of the 681,722 bp duplication clearly separates heterozygous carriers of the duplication (black symbols) from non-carriers (grey symbols). The red dotted line is an identity line. As expected, heterozygotes demonstrated on average an 1.51-fold (between 1.43 and 1.62-fold) coverage increase across the duplication relative to chromosome 2 background coverage. Four heterozygous samples (yellow stars) were misgenotyped as homozygous reference by sniffles2.* ***(f)*** *Boxplots representing the expression of* STK39 *and* CERS6 *in carriers (0/1) and non-carries (0/0) of the 681,722 bp duplication the was misgenotyped for 4 samples (red symbols). Median TPM values for each group are listed above.*

The enrichment of SVs among molQTL was substantially stronger (6.1-fold enrichment for eQTL, $p = 9.38 \times 10^{-130}$; 12.0-fold enrichment for sQTL, $p = 8.04 \times 10^{-256}$) when considering SVs in perfect LD with a top small variant. Furthermore, the proportion of SV eQTL and sQTL increased to 1.4% and 3.6%, respectively, when considering these molQTL. This included an *MAGI3* eQTL for which a SNP (3:29677355_A_G) was a lead variant. This SNP

had a “missing” genotype in one sample and was in perfect LD with a 6,966 bp insertion that was within an intron of *MAGI3* but had non-missing genotypes for all samples. The nominal p-value (p-value = 1.03×10^{-14}) of the SNP was slightly lower than that of the SV (nominal p-value = 1.14×10^{-14}; Supplemental Figure 15), demonstrating that an unequal number of samples with missing genotypes across variants introduces bias in association tests. We also identified instances where “missing” genotypes for SVs resulted in missed SV molQTL. For instance, we observed an 8,790 bp insertion that had “missing” genotypes in four samples, which we identified as homozygous for the reference allele upon manual inspection of the overlapping read alignments (Supplemental Figure 16). This insertion was in complete LD with a small variant sQTL for *LOC768028*. Exclusion of the “missed” genotype samples for this SV during association testing had a substantial impact on the estimated p-value (nominal p-value = 4.97×10^{-12}), which was consequentially fifteen orders of magnitude larger than the p-value of the top SNP (nominal p-value = 3.60×10^{-27}).

The effect size distributions for both sQTL and eQTL revealed that the most extreme molecular QTL effects were primarily driven by small variants rather than SVs (Figure 3h,i). We manually examined the top 1% of large effect molQTL with at least one SV passing the gene’s significance threshold to determine whether erroneously genotyped SVs prevented their identification as molQTL. Among the 33 eQTL and 31 sQTL that passed these criteria, we identified genotyping errors for SVs that were strong causal candidates for 9 eQTL and 10 sQTL (27.3% and 32.3% of the manually examined large effect eQTL and sQTL, respectively; Figure 4a,b,c,d). Thus, while nearly a third of these variants were affected by genotyping errors, the extent of misgenotyping per variant was low; for most of the manually examined SVs, only one or two samples showed genotyping errors, corresponding to a per-variant error rate of approximately 2%. Most genotyping errors stemmed from low overall coverage, reads failing to span the length of the variant, and insufficient coverage of one haplotype (often leading to heterozygous genotypes being miscalled as homozygous; Supplemental Figure 17). Large insertions were particularly susceptible to misgenotyping, and we observed multiple instances of insertions exceeding 8 Kb being missed as SV molQTL. Notably, the set of eVariants and sVariants were depleted for insertions between 8–9Kb, highlighting the broader effects of misgenotyping variants of this size (eVariants: OR = 0.41, $p = 5.82 \times 10^{-18}$; sVariants: OR = 0.27, $p = 1.08 \times 10^{-18}$). For example, we identified an 8,573 bp insertion that contained an L1 LINE element and was associated with an alternative splicing event of *PPM1H.* This variant was nearly ten orders of magnitude less significant than the top SNP (SV nominal p-value = 5.20×10^{-24}; SNP nominal p-value = 5.13×10^{-33}; Figure 4c,d), despite being located within the intron that was affected by differential splicing. Inspection of the HiFi alignments revealed that two samples were misgenotyped as

homozygous insertion carriers due to low coverage of the reference haplotype (Supplemental Figure 18). Correcting the two misgenotyped samples resulted in a stronger association between the splicing event and the SV (corrected SV nominal p-value = 1.31×10^{-26}).

Our SV catalogue contained 43 duplications spanning from 3 to 956 Kb that completely overlapped 135 testis-expressed genes, yet only four were identified as molQTL. This low proportion is unexpected, considering that previous studies have highlighted the substantial impact of such duplications on gene expression[26]. Nineteen of these 43 duplications could be confidently verified though coverage-based genotyping, corroborating that sniffles can identify duplications that are much larger than the HiFi reads. Taking the coverage-based genotypes as truth and the sniffles-called genotypes as query, we found that at least one sample was misgenotyped for all nineteen variants examined. Overall, we found that a substantial fraction of the sniffles-called genotypes for these long duplications were erroneous in both lowly and highly covered samples (Supplemental Figure 19). We identified a relatively common 682 Kb duplication on chromosome 2 (2:27313972–27995694), which overlaps the entire coding sequences of *CERS6* and *STK39*. This duplication was called as heterozygous by sniffles in 16 samples; however, the coverage-based genotyping indicated that 4 heterozygous samples were incorrectly genotyped as homozygous reference (i.e., did not have the duplication; Figure 4e; Supplemental Figure 20), despite having moderate coverage. These genotyping errors resulted in this duplication being missed as an SV eQTL for both overlapped genes. We observed that the top small variants for these genes were one to three orders of magnitude more significantly associated than the erroneously genotyped duplication (top small variant p-value *CERS6* = 3.31×10^{-15}, *STK39* = 1.90×10^{-14}; duplication p-value *CERS6* = 8.34×10^{-12}, *STK39* = 7.17×10^{-13}; Figure 4f). However, if genotyped correctly, the duplication is the top variant for both genes (corrected duplication p-value *CERS6* = 1.33×10^{-15}, *STK39* = 3.96×10^{-15}).

Discussion

We have generated and analysed a cohort of Braunvieh cattle that exceeded the effective population size with long and accurate HiFi reads, representing one of the largest long-read datasets available for a non-model organism. Notably, this is the first bovine long-read DNA cohort which also has paired functional data, providing the opportunity to investigate the consequences of small and structural variation on molecular phenotypes. This cohort enabled us to detect and characterize 23,789,890 small variants and 79,275 SVs, including

tens of thousands of previously unidentified SVs. Our collection of SVs is almost twice as large as that identified in a previously built 16-sample cattle pangenome [37] and substantially larger than previously reported in similar-sized, or larger, short-read sequenced cattle cohorts[25–27]. While short-read based SV cohorts are either depleted for insertions or lack them completely[9,25,42], we identified—as expected[23,43]—more insertions than deletions from the long-read alignments, thereby providing comprehensive access to a vastly understudied variant type. A saturation analysis corroborated that our SV set is a near complete catalogue of common SVs segregating within the Braunvieh population. However, we suspect that longer insertions and duplications (>15,000 bp) were missed for samples with shorter average read lengths. Long-read sequence data also improved alignments to the X and Y chromosomes—which are often neglected in GWAS for several reasons, including poor genotyping resulting from alignment errors due to their repetitiveness—and thereby facilitated the identification of over one hundred thousand small variants that were previously inaccessible with short-read sequencing, including 7 SNPs predicted as HIGH impact. However, variants along the MSY remained poorly characterized due to uneven coverage and copy-number complexity, limiting the ability to identify HIGH impact variants and molQTL for genes within this region. We also demonstrated improved small variant calling on the autosomes, particularly in tandem repeat regions, when using long reads over short reads, thereby creating a comprehensive genetic resource for genome-wide association testing.

Our expression and splicing QTL analyses revealed that molQTL are enriched for SVs. Despite being derived from a specific phenotype (testis RNA sequencing data), our findings support the broader importance of SVs in complex trait variation, as previously reported in cattle and other species[15,44–46]. We observed that 0.6% of eQTL and 1.9% of sQTL had an SV as the top variant. These proportions of SV molQTL were between 2.1-times and 14-times greater than those reported in short-read-based studies in bulls[25,37]. More than half of the SV molQTL were due to insertions that were largely inaccessible with short reads. Notably, the number of the SV molQTL nearly doubled when accounting for SVs in complete LD with a lead small variant, increasing to 1.4% of eQTL and 3.6% of sQTL. These estimates are more comparable to short-read- and long-read-based SV molQTL studies that were conducted in humans with higher coverage sequence data and larger sample sizes[46,47]. Nonetheless, confirming a causal role for these SVs requires further functional experiments. Contrary to other SV molQTL studies, we did not observe a statistical difference in effect size magnitude between small variants and SVs[46]. However, we identified multiple large effect molQTL where small variants were prioritized as candidate causal variants despite the presence of compelling candidate causal SVs. We suspect that several large allelic

substitution effects were erroneously attributed to the small variants due to imperfect SV and small variant genotyping, which introduces some bias in this assessment.

The forced genotyping approach substantially reduced the proportion of missing genotypes for all types of SVs. Yet, many SVs were missed as molQTL due to erroneous SV genotypes and an unequal distribution of missing genotypes between SVs and small variants. Errors in SV genotyping can have a substantial impact on QTL detection; for example, Chiang et al [46] reported that a 5% increase in error rate reduced SV eQTL detection by 19.6%. Almost one-third of the manually examined molQTL and all manually examined large duplications were affected by imperfectly genotyped small variants and SVs. The per-variant error rate among the manually examined molQTL was relatively low. However, because our sample size was small, even a few misgenotyped samples could strongly influence molQTL detection. We also observed instances, e.g., for large duplications, where much higher per-variant error rates obscured SV molQTL detection. Using the manually revised genotypes, we identified compelling candidate causal SVs in many of the examined molQTL where the summary statistics from the association tests initially prioritized small variants. We were able to attribute some of the SV genotyping errors to insufficient coverage and the undercalling of heterozygotes. This suggests that an average HiFi coverage higher than the 13.5-fold obtained for our cohort will produce more accurate genotypes for both small variants and SVs which will benefit association testing. Such findings also indicate that SVs are even stronger enriched among molQTLs than reflected by the results from the statistical association testing. Imputation can enhance statistical power and mitigate some of the missing genotype biases in association testing that we observed[48,49]. However, the accuracy of imputation can be much lower for SVs than small variants, which negatively impacts association testing and causal variant identification[50]. Since it remains unclear whether the benefits outweigh the drawbacks, we did not attempt to impute missing genotypes neither for SVs nor for small variants. Generating high-coverage long-read data at the sample-sizes required for detecting QTL is becoming increasingly feasible[17,51], which will reduce the missingness, thereby improving the statistical power of association studies utilising long-reads for variant discovery and genotyping. In the meantime, moderate-coverage long-read sequencing data, when collected in cohorts approximating the effective population size as demonstrated here, will enable the development of population-scale analytical methods such as SV genotype imputation[52].

Genotyping errors within complex regions, such as those containing multiallelic SVs or tandem repeats, and an inappropriate merging of SVs within these regions likely caused the observed underrepresentation of SV molQTL annotated as tandem repeats. In addition, the lack of reads spanning large variants resulted in erroneously genotyped large SVs, including

some molecular phenotype-associated variants, which remained undetected by our association analyses. This contributed to the lack of association of gene-containing duplications and likely contributed to the observed underrepresentation of LINE and LTR transposable element molQTL, which have previously been associated with large impacts on gene expression[53,54]. Further research is warranted to develop refined genotyping and association testing methods to fully capitalize on long-read sequencing data and explore the association between functionally relevant SVs and complex traits[55].

Methods

Ethics statement

The *Bos taurus taurus* tissues used in this study were previously collected from a commercial abattoir in Zürich, Switzerland after regular slaughter. None of the authors were involved in the decision to slaughter the bulls. No ethics approval was required for this study.

Sample extraction and DNA sequencing

Reproductive tissues (including testis and the caput of the epididymis) from 120 mature bulls were sampled by Mapel et al.[36]. Tissue samples were flash frozen in liquid nitrogen and stored at -80 C. Sampling was conducted between 2019 and 2021 and HiFi sequencing was performed from December 2022 to May 2023.

We extracted HMW DNA from flash-frozen testis (N=105) or epididymis (N=15) tissue samples with the Monarch HMW DNA Extraction Kit for Tissue (New England Biolabs). We followed the manufacturer recommended protocol for both tissue types. DNA was shipped on dry ice to PacBio (Rolling Stock Yard, London) for fragment length analysis, library preparation, and sequencing.

DNA read alignment

We aligned the HiFi reads with minimap2 (v2.24)[56] to ARS-UCD2.0 with the X chromosomal pseudo-autosomal region (PAR) hard masked (X:133300518-139009144), using the “map-hifi” preset. Alignments were coordinate sorted with SAMtools (v1.19.2)[57]. The previously collected short reads were aligned with bwa-mem2 (v2.2.1)[58,59] to the same reference, before collating by name, marking duplicates, and coordinate sorting with SAMtools.

We assessed alignment coverage using SAMtools bedcov and BEDtools (v2.30)[60] coverage, using 100 Kb windows generated by BEDtools makewindows. Regions with a minimum primary read depth of 2 and MAPQ of 5 were considered suitable for variant calling.

Small variant calling

We called small variants with DeepVariant (v1.6.0)[61], using the “PACBIO” model for the HiFi reads and “WGS” for the short Illumina paired-end reads, specifying --haploid_contigs "X,Y” and the PAR bed file, producing a gvcf for each sample. Samples were merged into population vcfs using GLnexus (v1.4.1)[61] with the DeepVariant config and revise_genotypes=False. DeepVariant variants larger than 45 bp were removed due to avoid unintentional overlapping with the dedicated structural variant calling.

We classified regions as tandem repeats using the pbsv (v2.9.0) utility findTandemRepeats (https://github.com/PacificBiosciences/pbsv/blob/master/annotations/findTandemRepeats) or as transposable elements using the RepeatMasker (v.4.1.5) (https://www.repeatmasker.org/) utility RM2Bed.py on the UCSC browser repeat file (https://hgdownload.soe.ucsc.edu/hubs/GCF/002/263/795/GCF_002263795.3/GCF_002263795.3.repeatMasker.out.gz) and keeping all SINE/LINE/LTR elements. Regions overlapping in tandem repeats and transposable elements were prioritized into the tandem repeat bed. All remaining regions were assigned as “normal” using BEDtools complement.

We assessed variant calling and genotyping accuracy per sample with hap.py (v0.3.15), removing all reference calls, left-shifting indels, and stratifying with the genomic regions specified above.

Structural variant calling

We called SVs with sniffles2 (v2.2)[62], providing the tandem repeat locations from above. We merged per-sample snf files across the cohort with sniffles2, removing any breakend SVs or SVs larger than 1 Mb, before force-genotyping each sample again with the merged set of SV candidates. We finally merged force-called VCFs with BCFtools merge and removed SVs with more than 10% missingness.

We assessed the repetitive element and tandem repeat content of the SV sequence (the reference for deletions and alternate for insertions) using Repeatmasker and TRF (v4.09.1)[63], respectively.

Variant analyses

We used VEP (release 113)[64] to assess functional impacts and consequences of small variants using the RefSeq (release 106) cattle annotation. We used plink (v1.90b6.26)[65] to calculate the LD between SVs and small variants, using a 1 Mb window and a minimum r^2-threshold of 0.2 and 0.8 to assess weak and strong tagging, respectively. We tested for mutually present SVs between our cohort and the SV pangenome panel from Leonard et al.[37] using Jasmine (v1.1.5)[66], with the flags "--pre_normalize max_dist_linear=0.5 max_dist=250" to allow slight mismatches in SV position or length between the two callsets. We used BEDtools intersect to examine the overlap between SVs and the previously described tandem repeat file.

Coverage analysis

We used mosdepth (version 0.2.2)[67] to calculate the average coverage in 500 bp windows for the HiFi alignments. Heterozygous and homozygous genotypes were assigned to the sniffles2-called duplications when the average coverage over the duplicated sequence was at least 1.3-fold and 1.8-fold higher, respectively, than the average coverage outside the duplication on the chromosome containing the duplication.

Molecular phenotype preparation

Total RNA from the HiFi-sequenced individuals was sequenced previously and made publicly available in Mapel et al.[36] (Supplementary Table 4). We filtered RNA reads with fastp (v0.23.1)[68] to remove adapter sequences, poly-A-tails, poly-G-tails, and low-quality bases.

We used Kallisto (v0.50.0)[69] and the RefSeq annotation for cattle to quantify transcript expression (TPM) and counts, which were aggregated to gene-level with tximport (v1.34.0)[70]. We considered genes with TPM >= 0.1 and at least 6 supporting reads in >=10% of samples for eQTL mapping. Expression values were inverse normal transformed and quantile normalized.

To identify and quantify splicing events, we first aligned cleaned reads to the ARS-UCD2.0 reference and RefSeq annotation with STAR (version 2.7.11a)[71] and included WASP filtering[72] to account for allelic bias, using heterozygous sites called from the HiFi reads. We extracted exon-exon junctions with Regtools (v0.5.2) [73], then used LeafCutter (v0.2.9) [74] to

cluster introns, calculate intron excision ratios, perform filtering, and normalize splicing phenotypes for sQTL mapping.

molQTL mapping

Cis-molQTL mapping was conducted with QTLtools (v1.3.1)[75]. We included variants that were within a 1 Mb window of the feature's start site (gene for eQTL and intron cluster for sQTL), and had MAF >= 1%, and missingness < 10%. We selected covariates for eQTL and sQTL testing with PCAForQTL[76] which implemented an elbow test to estimate the number of principal components (PCs) from the expression or splicing matrix to include as hidden confounders (hereafter, "RNA PCs"). We used the filterKnownCovariates function (R^2 thresholds = 0.5) to remove known covariates that were captured by the RNA PCs. The covariates considered for eQTL mapping included PCs 2-10 of the genome relationship matrix (constructed from 381,111 unlinked variants with plink (v2.00a3.6LM)[65] and --indep-pairwise 1000 5 0.2, RIN, age, and RNA PCs 1-12. The sQTL covariates included PCs 2-10 of the genome relationship matrix, RIN, age, and RNA PCs 1-11. We conducted 1,000 permutations with --permute function in QTLtools to infer beta corrected p-values, then used the qtltools_runFDR_cis.R script to apply a 5% false discovery rate (FDR) and estimate significance thresholds for each gene. sQTL testing include the --grp-best flag to account for multiple intron clusters within a gene. We performed a conditional analysis (as described in Delaneau et al.[75]) to identify independent-acting signals for each gene or intron cluster and the corresponding most significant variant.

We defined "SV molQTL" as QTL for which either an SV had the smallest p-value, or an SV had a p-value that was identical with a lead small variant. Enrichment was inferred with a Fisher's exact test. We used BEDtools to identify SVs that overlapped functional elements. Promoters and enhancer annotations for cattle testis tissue were obtained from Salavati et al.[77], while CpG islands were obtained from the UCSC genome browser[78]. Abbreviations and full descriptions of all discussed genes are listed in Supplementary Table 5.

Data availability

DNA and RNA sequencing data of the analysed cohort are available in the ENA database at the study accessions PRJEB42335 (Long-read sequencing data from cattle for the purpose of de-novo genome assembly), PRJEB28191 (Short read sequencing of cattle) and PRJEB46995 (Testis transcriptome of mature bulls). Accession identifiers for all samples are

available as Supplementary Table 4. Gene expression and splicing matrices, a VCF file of genome-wide small and structural variant genotypes used for e/sQTL mapping, a cross-table to link genotype and transcriptome data as well as results from e/sQTL mapping have been archived at zenodo (https://zenodo.org/records/15431126). The source data behind the graphs in the paper can be found at zenodo (https://zenodo.org/records/15431126).

Code availability

Computational workflows are available through https://github.com/AnimalGenomicsETH/HiFi_cohort

Acknowledgements

This study was supported by an ETH Research Grant and a grant from the Swiss National Science Foundation (SNSF, grant-ID 204654). The funding bodies were neither involved in the design of the study and collection, analysis, and interpretation of data nor in writing the manuscript. We thank Eirini Lampraki from Pacific Biosciences for DNA fragment analysis and sequencing. We thank Audald Lloret-Villas and Qiongyu He for valuable discussions.

Author contributions

X.M.M. sampled tissue and purified HMW DNA, aligned RNA reads against the reference, developed and applied workflows to quantify gene expression and splicing variation, conducted molecular QTL mapping, interpreted results, and drafted the manuscript; A.S.L. aligned DNA reads against the reference, called variants from short- and long-read alignments, interpreted results, and drafted the manuscript; H.P. conceived the study, contributed to the analysis of expression and splicing QTL, interpreted results, and contributed to the writing of the manuscript. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Daetwyler, H. D. *et al.* Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46**, 858–865 (2014).

2. McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

3. Abecasis, G. R., Cherny, S. S. & Cardon, L. R. The impact of genotyping error on family-based analysis of quantitative traits. *Eur J Hum Genet* **9**, 130–134 (2001).

4. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977–D985 (2023).

5. Hu, Z.-L., Park, C. A. & Reecy, J. M. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Res* **50**, D956–D961 (2022).

6. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat Genet* **50**, 1593–1599 (2018).

7. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**, 125–138 (2013).

8. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461–468 (2018).

9. Cameron, D. L., Di Stefano, L. & Papenfuss, A. T. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun* **10**, 3240 (2019).

10. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* **27**, 677–685 (2017).

11. Kosugi, S. & Terao, C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. *Hum Genome Var* **11**, 18 (2024).

12. Ahsan, M. U., Liu, Q., Perdomo, J. E., Fang, L. & Wang, K. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nat Methods* **20**, 1143–1158 (2023).

13. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* **14**, 1097–1102 (2014).

14. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

15. Alonge, M. *et al.* Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**, 145-161.e23 (2020).

16. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53**, 779–786 (2021).

17. Schloissnig, S. *et al.* Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644**, 442–452 (2025).

18. Duan, X., Pan, M. & Fan, S. Comprehensive evaluation of structural variant genotyping methods based on long-read sequencing data. *BMC Genomics* **23**, 324 (2022).

19. Yang, Q. *et al.* SVLearn: a dual-reference machine learning approach enables accurate cross-species genotyping of structural variants. *Nat Commun* **16**, 2406 (2025).

20. Kalleberg, J., Rissman, J. & Schnabel, R. D. Overcoming limitations to customize DeepVariant for domesticated animals with TrioTrain. *Genome Res* **35**, 1859–1874 (2025).

21. Lloret-Villas, A., Bhati, M., Kadri, N. K., Fries, R. & Pausch, H. Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. *BMC Genomics* **22**, 363 (2021).

22. Leonard, A. S. *et al.* Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat Commun* **13**, 3012 (2022).

23. Leonard, A. S., Crysnanto, D., Mapel, X. M., Bhati, M. & Pausch, H. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. *Genome Biol* **24**, 124 (2023).

24. Talenti, A. *et al.* A cattle graph genome incorporating global breed diversity. *Nat Commun* **13**, 910 (2022).

25. Bhati, M., Mapel, X. M., Lloret-Villas, A. & Pausch, H. Structural variants and short tandem repeats impact gene expression and splicing in bovine testis tissue. *Genetics* **225**, iyad161 (2023).

26. Lee, Y.-L. *et al.* High-resolution structural variants catalogue in a large-scale whole genome sequenced bovine family cohort data. *BMC Genomics* **24**, 225 (2023).

27. Grant, J. R. *et al.* A large structural variant collection in Holstein cattle and associated database for variant discovery, characterization, and application. *BMC Genomics* **25**, 903 (2024).

28. Lee, Y.-L. *et al.* A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle. *PLoS Genet* **17**, e1009331 (2021).

29. Trigo, B. B. *et al.* Variants at the ASIP locus contribute to coat color darkening in Nellore cattle. *Genetics Selection Evolution* **53**, 40 (2021).

30. Rothammer, S. *et al.* The 80-kb DNA duplication on BTA1 is the only remaining candidate mutation for the polled phenotype of Friesian origin. *Genet Sel Evol* **46**, 44 (2014).

31. Milia, S. *et al.* Taurine pangenome uncovers a segmental duplication upstream of KIT associated with depigmentation in white-headed cattle. *Genome Res* **35**, 1041–1052 (2025).

32. Durkin, K. *et al.* Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* **482**, 81–84 (2012).

33. Küttel, L. *et al.* A complex structural variant at the KIT locus in cattle with the Pinzgauer spotting pattern. *Anim Genet* **50**, 423–429 (2019).

34. Kadri, N. K. *et al.* 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 (2014).

35. Venhoranta, H. *et al.* Ectopic KIT copy number variation underlies impaired migration of primordial germ cells associated with gonadal hypoplasia in cattle (Bos taurus). *PLoS ONE* **8**, e75659 (2013).

36. Mapel, X. M. *et al.* Molecular quantitative trait loci in reproductive tissues impact male fertility in cattle. *Nat Commun* **15**, 674 (2024).

37. Leonard, A. S., Mapel, X. M. & Pausch, H. Pangenome genotyped structural variation improves molecular phenotype mapping in cattle. *Genome Res.* gr.278267.123 (2024) doi:10.1101/gr.278267.123.

38. Olagunju, T. A. *et al.* Telomere-to-telomere assemblies of cattle and sheep Y-chromosomes uncover divergent structure and gene content. *Nat Commun* **15**, 8277 (2024).

39. Crysnanto, D., Wurmser, C. & Pausch, H. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. *Genet Sel Evol* **51**, 21 (2019).

40. Jansen, S. *et al.* Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. *BMC Genomics* **14**, 446 (2013).

41. English, A. C., Cunial, F., Metcalf, G. A., Gibbs, R. A. & Sedlazeck, F. J. K-mer analysis of long-read alignment pileups for structural variant genotyping. *Nat Commun* **16**, 3218 (2025).

42. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

43. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).

44. Scott, A. J., Chiang, C. & Hall, I. M. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res* **31**, 2249–2257 (2021).

45. Zhang, Y. *et al.* Structural variation reshapes population gene expression and trait variation in 2,105 Brassica napus accessions. *Nat Genet* **56**, 2538–2550 (2024).

46. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat Genet* **49**, 692–699 (2017).

47. Billingsley, K. J. *et al.* Long-read sequencing of hundreds of diverse brains provides insight into the impact of structural variation on gene expression and DNA methylation. *bioRxiv* 2024.12.16.628723 (2024) doi:10.1101/2024.12.16.628723.

48. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499 (2010).

49. Huang, L., Wang, C. & Rosenberg, N. A. The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am J Hum Genet* **85**, 692–698 (2009).

50. Noyvert, B. *et al.* Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. *eLife* **14**, (2025).

51. Gong, J. *et al.* Long-read sequencing of 945 Han individuals identifies structural variants associated with phenotypic diversity and disease susceptibility. *Nat Commun* **16**, 1494 (2025).

52. Garimella, K. V. *et al.* Population-scale Long-read Sequencing in the All of Us Research Program. 2025.10.02.25336942 Preprint at https://doi.org/10.1101/2025.10.02.25336942 (2025).

53. Tang, L. *et al.* GWAS reveals determinants of mobilization rate and dynamics of an active endogenous retrovirus of cattle. *Nat Commun* **15**, 2154 (2024).

54. Adelson, D. L., Raison, J. M. & Edgar, R. C. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proc Natl Acad Sci U S A* **106**, 12855–12860 (2009).

55. Cui, Y. *et al.* Multi-omic quantitative trait loci link tandem repeat size variation to gene regulation in human brain. *Nat Genet* **57**, 369–378 (2025).

56. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

57. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

58. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* http://arxiv.org/abs/1303.3997 (2013).

59. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (2019). doi:10.1109/IPDPS.2019.00041.

60. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

61. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2021).

62. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* **42**, 1571–1580 (2024).

63. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).

64. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).

65. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

66. Kirsche, M. *et al.* Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* **20**, 408–417 (2023).

67. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).

68. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

69. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* **34**, 525–527 (2016).

70. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).

71. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

72. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061–1063 (2015).

73. Cotto, K. C. *et al.* Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat Commun* **14**, 1589 (2023).

74. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet* **50**, 151–158 (2018).

75. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat Commun* **8**, 15452 (2017).

76. Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol* **23**, 210 (2022).

77. Salavati, M. *et al.* Improving the annotation of the cattle genome by annotating transcription start sites in a diverse set of tissues and populations using Cap Analysis Gene Expression sequencing. *G3 (Bethesda)* **13**, jkad108 (2023).

78. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51–54 (2003).