Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Characterising a large bovine lactating mammary ATAC-seq dataset

Thomas J Lopdell

thomas.lopdell@lic.co.nz

Livestock Improvement Corporation https://orcid.org/0000-0002-7684-4870

Alexander J Trevarton

University of Auckland https://orcid.org/0000-0002-1404-0199

Janelle Moody

University of Auckland

Russell G Snell

University of Auckland https://orcid.org/0000-0002-8166-4014

Mathew D Littlejohn

Livestock Improvement Corporation https://orcid.org/0000-0001-9044-047X

Research Article

Keywords: ATAC-seq, chromatin accessibility, RNA-seq, gene expression, eQTL, caQTL

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10670646/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: TL and ML are employees of Livestock Improvement Corporation, a commercial provider of bovine germplasm. No other authors declare any conflicts of interest.

Abstract

Background

Many studies of genome-wide transcriptional regulation, or expression quantitative trait loci (eQTL), have been performed in multiple species to help improve our understanding of gene expression and regulation. However, many regulatory elements in the DNA sequence that control gene expression remain poorly known in agricultural species and humans alike. Here, we present a large-scale study using the Assay for Transposase Accessible Chromatin using Sequencing (ATAC-seq) method to detect chromatin accessibility QTLs (caQTL) in the lactating bovine mammary gland, and examine their correlations with gene expression to identify sequence motifs associated with higher or lower expression.

Results

Sequencing ATAC-seq libraries made from lactating mammary biopsies of 199 cows, we discovered 119,318 chromatin accessibility peaks, with 114,910 located on the autosomes. Across all samples, the average fraction of reads in peaks (FRiP) score was 0.493, above the ENCODE recommended threshold. Using high-depth mammary RNA-seq data representing duplicate biopsies of the same cows, significant ($p < 1 \times 10^{-6}$) positive correlations with peak height were seen with the expression of 1805 genes, and negative correlations with 793 genes. Significant *cis*-caQTL were observed for 19,339 peaks, and *cis*eQTL for 5003 genes. Positive correlations between eQTL and caQTL marker effects were observed for 6379 peaks, alongside 5356 for negative correlations, indicating the involvement of chromatin accessibility in the regulation of gene expression. Prediction of transcription factor binding sites in sequences under chromatin accessibility peaks showed differences in transcription factor frequencies between peaks with positive or negative effects on expression, with more STAT-domain containing factors in positive peaks, and more CTCF in negative peaks and insulator elements that decrease expression by blocking nearby enhancer elements.

Conclusion

These findings highlight mechanistically diverse modulation of gene expression in the lactating mammary gland and demonstrate the benefit of combining multiple omics datasets to elucidate the genetic control of gene expression in dairy cattle.

Background

Many studies of expression quantitative trait loci (eQTL) have been performed in dairy and beef cattle, with the aim of unravelling the genetic underpinnings of multiple traits of interest [1–3], culminating in the CattleGTEx [4] and FarmGTEx [5] projects. While this effort has been successful at finding the genes

that underlie numerous QTLs, the genetic architectures and causative variants that in turn control the expression of these genes (eQTLs) remain largely unknown. Previously, we have applied omics methods such as chromatin immunoprecipitation with sequencing (ChIP-seq [6]) to explore the landscape of histone modifications surrounding eQTL tag regions, with a view to identifying putative promoter and enhancer regions [7]. Another technique for assaying chromatin state more directly is the Assay for Transposase Accessible Chromatin using Sequencing (ATAC-seq) method [8]. This method highlights regions of the genome that are in a euchromatin state, and therefore accessible to transcription factors (TFs) and DNA-to-RNA polymerases. To date, several ATAC-seq studies in cattle have been conducted that each examined many tissues (e.g. [9–11]), but only sampled from small numbers of animals (frequently only one). Here, we present an ATAC-seq study of cattle that includes a sufficiently large number of animals (n = 199) to allow the detection of chromatin accessibility QTLs (caQTL).

Methods

Sample collection and sequencing

Lactating mammary tissue samples were collected for RNA-seq in three batches from 411 cows, as part of previously described studies ([12, 13]). The first batch of 22 animals comprised F2 cows collected as part of a large Holstein-Friesian×Jersey (FJX) crossbreeding trial [14]. The remaining two batches were sampled in 2013 and 2014 from cows on a research farm for use in an RNA-seq and eQTL trial. Two biopsy samples were collected concurrently from each animal, with the first sample used for the RNA-seq experiment, and the second frozen in liquid nitrogen before being stored at − 80°C for future analyses. A total of 199 of these secondary biopsies were subsequently assayed to detect open chromatin regions using the ATAC-seq method [8] for the current analysis. ATAC-seq libraries were prepared using the commercial ATAC-Seq Kit from Active Motif (Carlsbad, USA). Four samples were sequenced as a pilot project using a HiSeq X machine (Illumina Inc, USA), and the remainder were sequenced using the Novaseq 6000 genome analyser (Illumina Inc, USA) using a combination of SP and S4 flowcells in three batches to target a minimum of 25 million 150 bp paired-end reads per sample. The Active Motif method is based on a combination of the original ATAC-seq method by Buenrostro et al.[8] and the Omni-ATAC method by Corces et al.[15].

Sequence informatics

ATAC-seq reads were processed using Trimmomatic (v0.39; [16]) to remove adapter sequences using the provided NexteraPE paired-end adapter sequences, and to remove low quality bases from the sequence, using the settings LEADING:20 TRAILING:20 SLIDINGWINDOW:5:15 MINLEN:70. Reads that remained paired after this step were then mapped to a masked bovine reference sequence (based on ARS-UCD1.2, as described by [7]) using BWA-MEM (v0.7.17; [17]), and subsequently filtered to remove secondary and supplementary alignments, as well as any with mapping scores under 40 (out of 60), using SAMtools (v1.13; [18]). To facilitate phenotyping the same set of peaks across all samples, a

consensus BAM file was produced for peak calling by down-sampling 5% of the reads from each sample, including only read pairs where both reads in the pair were mapped with a mapping score of at least 50. The individual subsamples were then concatenated and sorted to form the consensus BAM file. Subsampling, concatenating, and sorting were all performed using SAMtools (v1.13). Peak calling was performed on the consensus BAM file using MACS (v3.0.0a7; [19]) in narrow peak calling mode with the Benjamini-Hochberg q-value threshold set to 0.01, as recommended by the MACS documentation for ATAC-seq calling.

RNA-seq reads were trimmed to remove low-quality sequences using Trimmomatic (v0.39) with settings LEADING:20 TRAILING:20 SLIDINGWINDOW:3:15 MINLEN:50, then mapped to the same masked reference genome using STAR (v2.7.0; [20]) using a two-pass approach. In the first pass, a set of splice junctions was extracted from NCBI assembly GCF 002263795.1 (release 106) and reads mapped to identify novel splice junctions. Novel junctions with at least five uniquely mapped supporting reads were identified and used to augment the junction set that was then used to remap the reads during the second pass to produce the final alignments.

Phenotypes for chromatin accessibility QTL (caQTL) were calculated as described previously [12]. Briefly, read counts were generated for both ATAC-seq and WGS sequence data using the featureCounts function of the Subread software package (v1.5.3; [21]). These counts were then adjusted for sequencing depth, and filtered to remove low ATAC-seq coverage peaks and high WGS coverage peaks to remove those caused by reference genome artefacts. The ATAC-seq counts for the remaining peaks were then fitted against the WGS counts (both on a log scale), and the residual used as the phenotype. Phenotypes for gene expression were also created as described previously [12]. Briefly, RNA-seq read depths were counted using featureCounts to produce a matrix of count data for each of the two datasets. These were subsequently loaded into the R package (v4.3.1; [22]) DESeq2 (v1.42.1; [23], part of the Bioconductor project (v3.18) [24]). The DESeq2 function DESeqDataSetFromMatrix was applied to adjust for sequencing batch effects, with the read counts further processed using the variance stabilising transformation (VST) function to produce the final gene expression phenotypes.

To examine the phenotypic correlations between chromatin accessibility (ATACseq read depth) and gene expression (RNA-seq read depth), new phenotypes were created for chromatin accessibility using the same method as described above for gene expression. This was done to ensure that both gene expression and chromatin accessibility phenotypes were on similar scales, with the aim of improving the interpretability of the results. Outlier detection and removal were applied using principal component analysis (PCA) as described below. Correlations between the phenotypes for gene expression and chromatin accessibility were calculated using R between all genes and all ATAC-seq peaks within 1 Mbp of the transcription start site of each respective gene.

Genotypes and QTL analyses

The 22 cows that were part of the FJX trial described above had been genotyped using the Illumina Bovine50k chip panel. These genotypes were imputed to the higher density Illumina BovineHD 777k panel (HD), which had also been used to genotype the remaining cows, using Beagle version 5.0 [25].

Both the RNA-seq and ATAC-seq datasets were analysed using PCA on the VSTtransformed gene expression or peak height values to remove outlier animals in any of the first six PCs, as previously described [1, 12]. Briefly, outlier samples were removed from each dataset at a threshold of > 4 standard deviations from the mean on any of the first six components. This left a dataset comprising 375 animals with gene expression phenotypes and 193 with ATAC-seq peak phenotypes.

Association analyses were performed to identify expression QTL (eQTL) using the RNA-seq phenotypes, and chromatin accessibility QTL (caQTL) using the ATACseq phenotypes. The HD genotypes were imputed to whole-genome sequence (WGS) resolution using Beagle version 5.3 with a reference population of 1300 cattle. After filtering to remove variants with dosage-R^2 < 0.9 or Hardy-Weinberg Equilibrium (HWE) p-values < 1×10^{-20} (suggesting genotypes that were likely poorly imputed), these filters yielded a dataset containing 14,900,070 variants across the autosomes. The HD genotypes were also used to calculate a genomic relationship matrix (GRM) for the cows using the GCTA software package (version 1.93.2) [26]. The association analyses were performed using the mixed-linear model association (MLMA) method with leave-one-chromosome-out (LOCO). For the eQTL and caQTL analyses, the most significant effects were assumed to be proximal, so a subset of the imputed WGS variants was extracted surrounding the corresponding ATAC-seq peak or gene, where the variants sat within 1Mbp of the middle of the peak for caQTL, or within 2Mbp of the centre of the gene for eQTL. The narrow windows were used for caQTL due to computational time and space constraits imposed by the much larger number of peaks compared to genes.

To examine the genotypic correlations between caQTL and eQTL, pairs of caQTL and eQTL were identified where i) both QTL had at least one tag variant with $p < 1\times10^{-4}$, and ii) the ATAC-seq peak was located within 100kbp of the most significant eQTL variant position, or the median of the positions in the case that multiple variants were equally significant. Variants within 100kbp of the central point for each pair of QTL were extracted, where the central point was defined as the median position of the top tag variants for both QTL. Pairs of QTL with fewer than 50 variants in this window were excluded to reduce spurious correlations between QTL. Pearson correlations between the variant effects (βs) were calculated using R function cor.test for the remaining pairs of caQTL and eQTL.

Motif detection and transcription factor binding site prediction

Peaks showing a significant ($p < 1 \times 10^{-6}$) phenotypic correlation with the expression of at least one gene were identified, and categorised into one of four effect categories (ECs): Singleton (S), where the peak was correlated with only a single gene; Multiple-Positive (MP), where at least 80% of correlations observed were positive; Multiple-Negative (MN), where at least 80% of correlations were negative; and

Multiple-Mixed (MM), for the remaining peaks. Within each EC, the FASTA sequences underlying each peak were extracted from the reference genome and analysed using the TFBSTools (version 1.32.0) R package with the JASPAR2024 CORE database to identify sequence motifs that are predicted to correspond to transcription factor binding sites (TFBSs). The predicted binding sites were subsequently filtered to produce a set of high-confidence sites by removing any with relScore < 0.95 or absScore < 15. To test for differences in the proportions of TFs and TF families regulating the genes correlated with peaks in different ECs, two statistical analyses were performed to compare TF and TF class frequencies between the four peak ECs: i) a four-way test between all the ratios, and ii) a two-way test between Singleton peaks and all others.

Results

Quality statistics

An average of 75.03 million ATAC-seq read pairs were mapped per sample to the ARS-UCD 1.2 bovine reference genome (SD = 43.91 million). Of these, an average of 9.04% (SD = 5.06%) mapped to the mitochondrial chromosome. After consensus peak calling and excluding the mitochondria, a total of 119,318 chromatin accessibility peaks were detected, comprising 114,910 autosomal peaks, 2964 X chromosome peaks, and 1444 peaks on other (unmapped) contigs. This is above the current ENCODE minimum standard for replicated peak calling (100,000; https://www.encodeproject.org/atac-seq/), though below their recommended number of peaks (> 150,000).

As a further quality check, the fraction of reads in peaks (FRiP) score [27] was calculated, where higher values indicate greater enrichment of reads within peaks, as a measure of the signal to noise ratio. Here the mean FRiP score was 0.493 (SD = 0.101), with ten of 197 samples giving FRiP scores < 0.3 (the ENCODE recommended threshold) and only two samples < 0.2, the minimum threshold (Fig. 1A). Another quality statistic recommended for ATAC-seq by ENCODE is the transcription start site (TSS) enrichment score. Across all animals, the average enrichment was 8.09 (SD = 1.78), with five animals giving values < 5. Thresholds for acceptable values vary by species and reference genome build. No thresholds are provided for cattle genomes, but the values given for the human GRCh38 reference are 5–7 for acceptable and > 7 for ideal. Using these thresholds gives sample counts of 56 animals between five and seven, and 135 animals greater than seven, indicating that the majority of samples are of ideal quality based on this metric (Fig. 1B).

The expression of many genes correlates with neighbouring ATAC-seq peaks

Chromatin accessibility is expected to affect the ability of transcription factors and RNA polymerase to bind to the DNA and initiate gene expression. Therefore, we would expect that gene expression should be correlated with chromatin accessibility in neighbouring ATAC-seq peaks, and moreover that this

correlation will be positive more often than negative. Overall, the transcription levels of 1805 genes had at least one positive correlation with a nearby ATAC-seq peak at $p < 1 \times 10^{-6}$, compared to only 793 with a negative correlation of the same significance. Additionally, the strongest positive correlations were stronger than the strongest negative correlations, with a correlation of 0.924 seen between peak 2744 and the gene *USP13*, compared to − 0.688 for peak 79523 and *PICALM*. Table 1 lists the pairs of genes and neighbouring peaks with the highest phenotypic correlations. Figure 2 illustrates the genes *CSTB* and *PICALM*, showing the correlations between gene expression and chromatin accessibility at nearby peaks.

Table 1
Ten strongest positive (top) and negative (bottom) phenotypic correlations observed between ATAC-seq peak read count (chromatin accessibility) and RNAseq gene read count (gene expression).

Peak	Peak location	Score	Gene	Correlation	Pvalue
2744	1:87,411,874–87,412,534	2021	USP13	0.924	9.91×10^{-83}
77540	28:12,154,533–12,155,825	7569	CHRM3	0.922	2.23×10^{-81}
67263	24:12,649,956–12,650,804	7346	SYT4	0.918	1.56×10^{-79}
65070	23:27,126,553–27,127,781	1799	LOC615223	0.917	7.82×10^{-79}
93646	5:32,110,996–32,112,591	4657	CCDC184	0.907	2.13×10^{-74}
4468	1:143,821,993–143,824,141	2235	CSTB	0.905	1.84×10^{-73}
65070	23:27,126,553–27,127,781	1799	LOC786065	0.902	3.20×10^{-72}
81966	3:1,675,407–1,676,105	3197	GPA33	0.895	1.56×10^{-69}
31148	15:80,663,947–80,664,557	440	LOC520402	0.892	1.53×10^{-68}
37589	17:56,046,000–56,046,449	275	CCDC60	0.888	3.61×10^{-67}
112967	9:64,059,645–64,060,222	1104	NT5E	−0.626	1.28×10^{-22}
72385	25:41,010,206–41,011,370	3009	MRM2	−0.627	1.06×10^{-22}
113889	9:87,096,806–87,097,423	348	LOC100336795	−0.629	$7.76 \times 10 - 23$
25531	14:23,232,132–23,233,163	1200	PLAG1	−0.640	8.07×10^{-24}
65188	23:27,927,226–27,927,719	604	BOLA	−0.660	1.03×10^{-25}
31482	16:2,161,714–2,162,424	3080	LOC101907913	−0.661	7.01×10^{-26}
65188	23:27,927,226–27,927,719	604	LOC788634	−0.662	5.79×10^{-26}
84441	3:54,096,842–54,097,280	671	LOC786372	−0.675	2.99×10^{-27}
9225	10:78,271,153–78,271,758	6682	CCDC196	−0.677	1.54×10^{-27}
79523	29:9,513,896–9,514,409	1723	PICALM	−0.688	1.12×10^{-28}

Peak 65070 was correlated with both *LOC615223* and *LOC786065*. Interestingly, both genes encode members of the butyrophilin family: butyrophilin subfamily 3 member A2 and butyrophilin-like protein 1. This family of proteins is involved in regulating both the immune system and the secretion of milkfat droplet from mammary epithelial cells [28, 29]. Peak 65118, mapping in the same region, was correlated

with both *BOLA* and *LOC788634*, where the latter is a pseudogene that encodes BOLA class I histocompatibility antigen, alpha chain BL3-7-like. The bovine leukocyte antigen (BoLA) region is orthologous to the major histocompatibility complex (MHC) region in humans.

Shared regulation of chromatin accessibility and gene expression

ATAC-seq peaks can be associated with either proximal promoter elements or with more distal enhancer or silencer elements. Looking at only the most strongly correlated ATAC-seq peak for each gene, a negative correlation of − 0.291 was seen between the absolute value of the peak height-gene expression correlation and the distance between the ATAC-seq peak and the gene TSS, i.e, more distant pairs showed weaker correlations. When only significant peak height-gene expression correlations ($p < 1 \times 10^{-6}$) were considered, the correlation with distance became more negative, at − 0.315. This suggests, perhaps not surprisingly, that chromatin accessibility regions with bigger effects on gene expression tend to be closer to the TSS, and so are more likely to be promoter elements.

Many ATAC-seq peaks were identified to have a significant *cis*-caQTL, with 19,339 having a tag variant with a p-value $< 1\times10^{-6}$, of which 10,504 were also significant at a threshold of 1×10^{-10}. The 105,357 peaks for which the QTL discovery analysis was run (the remainder had too few variants nearby) covered a total of 74.9 Mbp (3.01% of the genome) and covered 396,007 of the 14.99 million variants (2.64%). The lower percentage of variants within the peaks compared to the percentage of the genome covered by the peaks suggests that the sequences underlying the peaks may be evolutionarily conserved. However, when considering only the 19,339 peaks with significant caQTL, 117,319 markers (0.78%) were contained within only 14.25 Mbp (0.57% of the genome), implying that these peaks are under less evolutionary constraint than the remaining peaks.

Significant *cis*-eQTL were identified for 5003 genes at a significance level of 1×10^{-6}, of which 2717 remained significant at a threshold of 1×10^{-10}. Similar to the phenotypic correlations described in the previous section, if a gene with an eQTL is being regulated by an enhancer or silencer element linked to a caQTL, we expect to see a correlation between SNP effects for eQTL and caQTL. Table 2 lists the pairs of gene and neighbouring peak with the highest variant effect correlations, and Fig. 3 illustrates two correlated eQTL-caQTL pairs. As was seen for the phenotypic correlations, more eQTLs had at least one significant ($p < 1\times10^{-6}$) positive correlation with at least one caQTL (n = 6379), compared to having at least one negative correlation of equivalent significance (n = 5356). The neighbouring genes *PIGP* and *TTC3* both had significant genotypic correlations with peak 4929, and the genes *GPHN* and *CCDC196* were likewise both correlated with peak 9225.

Table 2
Ten strongest positive (top) and negative (bottom) variant effect (β) correlations observed between ATAC-seq caQTL (chromatin accessibility QTL) and RNAseq eQTL (gene expression QTL). All p-values reported here were < 1×10^{-100}.

Peak	Peak location	Score	Gene	Correlation	*n* SNPs
31148	15:80,663,947–80,664,557	440	LOC520402	0.997	1339
77349	28:7,266,571–7,266,983	684	TARBP1	0.995	1366
29624	15:50,358,656–50,359,450	2049	TRIM68	0.995	2949
37589	17:56,046,000–56,046,449	275	CIT	0.994	1613
4929	1:149,304,374–149,304,873	627	PIGP	0.994	1266
4929	1:149,304,374–149,304,873	627	TTC3	0.994	1266
25994	14:36,466,145–36,466,490	217	SBSPON	0.993	898
13192	11:61,090,723–61,090,980	501	EHBP1	0.993	854
104808	7:51,814,967–51,815,184	410	HARS	0.993	1444
58257	21:34,027,112–34,028,114	4668	EDC3	0.993	592
45026	19:27,193,919–27,194,638	4326	LOC100568284	−0.980	435
104260	7:43,473,219–43,474,053	8219	EFNA2	−0.980	150
9225	10:78,271,153–78,271,758	6682	GPHN	−0.980	1583
29642	15:50,735,171–50,735,425	111	OR52I2	−0.982	2367
91236	4:98,848,415–98,849,614	16,930	CYREN	−0.983	657
95378	5:75,299,961–75,300,963	3008	CSF2RB	−0.985	1669
58797	21:55,140,572–55,141,745	3395	ELL3	−0.985	1608
64033	23:9,969,618–6,671,522	12,613	SRPK1	−0.990	662
84441	3:54,096,842–54,097,280	671	LOC786372	−0.990	2268
9225	10:78,271,153–78,271,758	6682	CCDC196	−0.992	1583

Transcription factors differ between peaks with positive or negative correlations

ATAC-seq peaks showing statistically significant phenotypic correlations ($p < 1\times10^{-6}$) with the transcription levels of at least one gene were categorised into one of four categories: Singleton (S), where the peak was correlated with a single gene; MultiplePositive (MP), where the peak was correlated with two or more genes, of which at least 80% were positively correlated; Multiple-Negative (MN), which

is similar but for negative correlations, or Multiple-Mixed (MM), containing correlations with a higher proportion of opposite sign effects (all other peaks). The majority of peaks fell into the Singleton category (n = 6205), followed by MP (n = 1095), MN (n = 710), and MM (n = 499). Within the Singleton peaks, 4822 (77.7%) showed a positive correlation with the expression of the corresponding nearby gene. A small but highly-significant ($p < 1\times10^{-300}$) negative effect (− 0.000614 per kilobase) was also seen between the correlation and distance to the TSS of the gene, suggesting that, for Singleton peaks, chromatin accessibility peaks that are closer to the gene's TSS tend to have stronger effects on expression.

Within each of the four categories, the sequences underlying each peak were analysed to predict transcription factor binding sites (TFBSs). After filtering (absScore ≥ 15), TFs were classified based on the protein domains contained in their structures. C2H2 zinc fingers were the TF class with the most commonly predicted binding sites, accounting for approximately 50% of all predicted TFBSs. This remained true in all categories when peaks were categorised as MP, MN, MM, or S. The second most frequently predicted class for MP peaks was the STAT domain containing factors, followed by Tryptophan cluster factors. By contrast, the second most common TF class for MN peaks was Basic helix-span-helix (bHSH), followed by Basic leucine zipper factors (bZIP). The corresponding frequency order of TF classes for both MM and S peaks was bZIP followed by STAT domain factors. To test the statistical significance of these differences, the frequencies of these TF classes were compared across peak categories using a four-sample test for equality of proportions (Table 3). The most significant differences were seen for the C2H2 zinc finger and Basic helix-spanhelix (bHSH) factors, which were both predicted at significantly higher proportions in MN peaks than in MP or Singleton peaks. Several other classes, including the STAT domain and Homeo domain containing factors, by contrast, were predicted at significantly lower rates in MN peaks.

Table 3

Proportions of transcription factors (TFs) predicted by class, with significant differences in frequency between four categories of peaks: Multiple-Positive (MP), Multiple-Negative (MN), Multiple-Mixed (MM), and Singleton (S); see text for definitions. P-values are shown following Bonferroni correction, and only classes where the corrected p-value was < 0.05 are shown.

TF Class Name	**MP (%)**	**MN (%)**	**MM (%)**	**S (%)**	**P (Bonf)**
C2H2 zinc finger	50.19	58.66	49.16	49.21	5.07×10^{-44}
Basic helix-span-helix (bHSH)	3.22	5.90	4.72	3.34	4.79×10^{-27}
STAT domain	6.65	4.15	6.30	7.45	8.01×10^{-25}
Homeo domain	4.26	1.79	3.22	3.92	5.88×10^{-20}
Basic leucine zipper (bZIP)	6.18	5.61	8.93	7.46	3.93×10^{-13}
Tryptophan cluster	6.56	3.98	6.10	5.61	4.45×10^{-11}
C2CH THAP-type zinc finger	0.35	0.07	0.13	0.08	4.48×10^{-9}
Fork head/winged helix	2.62	1.50	2.61	1.30	3.91×10^{-8}
SMAD/NF-1 DNA-binding domain	0.54	1.08	1.19	0.67	1.04×10^{-6}
Basic helix-loop-helix (bHLH)	5.79	4.62	4.27	5.60	7.38×10^{-5}
MADS box	0.45	0.30	0.79	0.71	6.69×10^{-4}
High-mobility group (HMG) domain	0.73	0.41	0.58	0.83	1.12×10^{-2}
TF Name	MP (%)	MN (%)	MM (%)	S (%)	P (Bonf)
MAZ	3.08	5.13	4.27	3.29	4.81×10^{-13}
Thap11	0.35	0.07	0.13	0.08	5.54×10^{-8}
TFAP2C	1.07	2.03	1.60	1.14	2.05×10^{-7}
TFAP2A	1.08	1.90	1.46	1.04	6.65×10^{-7}
TFAP2B	1.07	1.98	1.66	1.16	2.78×10^{-6}
ZNF148	1.42	2.52	2.11	1.64	2.72×10^{-5}
ELK3	0.48	0.14	0.16	0.18	5.62×10^{-5}
ZNF740	0.49	1.13	0.40	0.67	2.56×10^{-4}
Irf1	0.58	0.12	0.36	0.57	5.25×10^{-4}

TF Class Name	MP (%)	MN (%)	MM (%)	S (%)	P (Bonf)
HES5	0.07	0.11	0.04	0.01	1.09×10^{-3}
SP1	3.91	5.23	3.85	3.93	1.71×10^{-3}
STAT1	1.77	1.28	1.75	2.09	2.58×10^{-3}
Stat5a::Stat5b	0.87	0.55	0.83	1.12	3.21×10^{-3}
HEY1	0.08	0.12	0.09	0.02	1.42×10^{-2}
CREM	0.19	0.08	0.18	0.05	1.74×10^{-2}

Looking at individual TFs, the most commonly predicted binding sites across all categories were for the TFs SP1 (MA0079.5), MAZ (MA1522.2), KLF15 (MA1513.2), and PATZ1 (MA1961.2). MAZ also showed the most significant difference in binding site presence amongst peak categories, with a significantly higher proportion of its predicted TFBSs mapping to ATAC-seq peaks in the MN category (Table 3). Other TFs showing a similar pattern of enrichment under peaks in the MN class included members of the TFAP2 family (TFAP2A [MA0003.5, MA0810.2, MA0872.1], TFAP2B [MA0811.2, MA0812.2, MA0813.1], and TFAP2C [MA0524.3, MA0814.3, MA0815.1]), the zinc finger ZNF148 (MA1653.2), and SP1. Another important C2H2 zinc finger is CCCTC-binding factor (CTCF), with a primary role in maintaining chromatin loop structures and topologically associated domains (TADs) [30, 31]. CTCF binding sites (MA0139.2) were also predicted significantly ($p = 0.019$) more frequently in MN peaks (4.1% of peaks) than in MP (2.7%) or S (2.4%), suggesting that higher chromatin accessibility at these binding sites is associated with lower expression of nearby genes. By contrast, binding sites for the TFs ELK3 (MA0759.3) and Irf1 (MA0050.4) were more commonly predicted in MP peaks, as were the STAT domain factors STAT1 (MA0137.4) and Stat5a::Stat5b (MA0519.2).

To further investigate these differences, additional tests were performed to find differences between the TF class proportions in Singleton peaks compared to all others (Table 4). Binding sites for C2H2 zinc fingers, bHSH, C2CH THAP-type zinc finger factors were predicted significantly more frequently in non-Singleton peaks, indicating more generalised regulatory functions, while several other classes, including STAT domain and Homeo domain, were more frequent in the Singleton peaks. Refining the tests to single TFs, the most significant were the C2H2 zinc finger MAZ (MA1522.2) and basic helix-loop-helix (bHLH) factor HES5 (MA0821.2), both of which were observed significantly less frequently in Singleton peaks. Other TFs following this pattern included the bHSH factors TFAP2A, TFAP2B, and TFAP2C and the C2CH THAP-type zinc finger Thap11. By contrast, binding sites for the STAT domain factors STAT1, Stat5a, and Stat5b were all observed more frequently in Singleton peaks.

Table 4
Transcription factor classes (above) and names (below) with significant differences in frequency between Singleton (S) and non-single (MP + MN+MM) peaks (see text for definitions). P-values are shown following Bonferroni correction, and only classes or TFs where the corrected p-value was < 0.05 are shown.

TF Class Name	**MP + MN+MM (%)**	**S (%)**	**P (Bonf)**
C2H2 zinc finger	53.16	49.25	1.67×10^{-19}
STAT domain	5.66	7.46	4.21×10^{-16}
Basic helix-span-helix (bHSH)	4.54	3.35	8.89×10^{-13}
Homeo domain	3.13	3.93	1.12×10^{-5}
Fork head/winged helix	2.04	2.61	2.92×10^{-4}
Basic leucine zipper (bZIP)	6.57	7.47	9.38×10^{-4}
C2CH THAP-type zinc finger	0.20	0.08	1.32×10^{-3}
MADS box	0.47	0.71	9.11×10^{-3}
High-mobility group (HMG) domain	0.58	0.83	1.85×10^{-2}
TF Name	MP + MN+MM (%)	S (%)	P (Bonf)
MAZ	4.19	3.40	1.58×10^{-4}
HES5	0.08	0.008	8.69×10^{-4}
TFAP2A	1.50	1.07	1.34×10^{-3}
Stat5a::Stat5b	0.76	1.16	1.58×10^{-3}
HEY1	0.010	0.019	2.88×10^{-3}
STAT1	1.63	2.15	2.96×10^{-3}
TFAP2C	1.58	1.18	1.10×10^{-2}
Thap11	0.20	0.083	1.79×10^{-2}
Stat5a	0.27	0.49	2.46×10^{-2}
TFAP2B	1.57	1.20	4.18×10^{-2}
CREM	0.15	0.05	4.92×10^{-2}

Links to dairy and beef phenotypes

Many QTL are known in dairy cattle for economically important traits such as milk production or fertility, and a large proportion of these are believed to involve regulation of gene expression [2, 32, 33]. A subset of these regulatory QTL are expected to be driven by chromatin accessibility, and this will also be under genetic control in many cases. Therefore, we should observe cases of both phenotypic correlations between gene expression and chromatin accessibility, and genotypic correlations between eQTL and caQTL. For example, we have previously reported a correlation between the lactoferrin gene (*LTF*) eQTL and a caQTL for a nearby peak [12]. Here, we report additional correlations related to genes that have been linked to lactation (Table 5).

Table 5

Correlations between ATAC-seq peaks and selected lactation-related genes believed to underlie known regulatory QTL. Phenotypic correlations are reported between gene expression phenotypes and chromatin accessibility phenotypes. Genotypic correlations are reported between the associated eQTL and caQTL, calculated using either marker effects (β) or marker log-scaled p-values. Distances are reported in kilobases between the ATAC-seq peaks and the TSS of the gene. Only the phenotypic or genotypic correlations with the largest absolute value within 100kb distance are shown.

Gene	**Peak**	**Dist (kb)**	**Phenotypic Corr**	**Genotypic**		
				Corr (β)	**Corr (log*P*)**	***n* Vars**
ABO	16109	27.00	0.410	0.760	0.890	1979
ANKH	55998	1.57	0.323	0.753	0.626	1424
	56011	72.71	0.373	0.601	0.618	1398
CSN3	100194	15.50	0.349	0.558	0.591	966
	100201	48.71	−0.302	−0.967	0.974	963
CSF2RB	95374	67.48	0.528	0.830	0.969	1682
	95378	48.15	−0.411	−0.985	0.994	1669
DGAT1	24219	10.54	−0.553	−0.878	0.826	266
GPAT4	76604	13.75	0.386	0.694	0.881	923
KCNJ2	48429	89.56	0.299	0.865	0.910	1216
	48433	36.32	0.510	0.887	0.784	1043
LRRC8C	84396	86.28	0.179	0.898	0.764	1682
	84398	73.79	0.319	0.836	0.700	1682
LTF	62755	0.34	0.511	0.950	0.930	1425
	62756	6.95	0.540	0.891	0.912	1336
MGST1	95984	16.73	0.295	0.599	0.902	1335
	95988	34.57	0.319	0.786	−0.422	1673
PAEP	15986	4.34	0.343	0.971	0.975	1048
PICALM	79523	5.23	−0.688	−0.937	0.807	853
	79524	0.60	0.673	0.956	0.864	853
SLC37A1	4356	56.01	0.293	0.747	0.382	1856
SLC50A1	82621	1.93	0.403	0.803	0.927	781

Gene	Peak	Dist (kb)	Phenotypic Corr	Genotypic		
				Corr (*β*)	Corr (log*P*)	*n* Vars
	82622	0.74	0.504	0.807	0.903	781

The gene Phosphatidylinositol-Binding Clathrin Assembly (*PICALM*) has been linked to milk lactose concentration [1]. Two ATAC-seq peaks (79523, 79524) were correlated with *PICALM* expression. Peak 79523 had negative correlations both phenotypically (− 0.689) and genotypically (− 0.937), while peak 79524 had positive correlations (0.673 and 0.956 respectively). However, none of the eQTL tag variants (LD > 0.95 with the most significant marker) map within either peak, with the closest variant landing 3.3kbp from peak 79523 and 8.0kbp from 79524. Another interesting gene is Colony Stimulating Factor 2 Receptor Subunit Beta (*CSF2RB*), which has been linked to several milk composition and yield phenotypes [34]. Two eQTL tag variants, chr5:75281419 and chr5:75281607, both map within ATAC-seq peak 95374, which is the most highly correlated peak with *CSF2RB* gene expression. Neither variant alters any predicted TFBS, although variant rs211440748 maps one base from a predicted binding site for the Homeo domain factor MEIS1 (MA0498.3; absolute score 6.22), which is within an annotated enhancer element (ENSBTAR5_BJXDD).

The gene *DGAT1* shows strong negative correlations both phenotypically and genotypically with ATAC-seq peak 24219. None of the eQTL tag variants map within this peak. Genetic regulation of *DGAT1* expression is complex, with proposed functional variants including a variable number tandem repeat (VNTR) in the promoter region [35] as well as splicing effects via variation in an exon splice enhancer [ESE; 36]. These effects likely dilute or confound any signal caused by chromatin accessibility. A similar dilution effect is seen for other genes where splicing is known to influence gene expression, such as *ABO* [37] and *CSN3* [38], showing similarly low phenotypic correlations. Nevertheless, two tag variants of the *CSN3* eQTL (rs110374385 and rs109499695) are present within ATAC-seq peak 100194, although neither map within a predicted TFBS. Another gene impacted by previously-proposed regulatory variants is *β*-lactoglobulin (*PAEP*), where several promoter and coding variants have been associated with gene expression levels [39, 40]. The caQTL for peak 15986 was strongly correlated with the *PAEP* eQTL, and seven of the eQTL tag variants map within this peak. Of these, rs110935748 mapped within a predicted binding site for the nuclear receptor NR2C1 (MA1535.2; absolute score 10.57), rs5381369303 mapped with predicted TFBSs for several nuclear receptors, including NR2F1 (MA1538.1; absolute score 18.93) and NR2F6 (MA1539.1; absolute score 19.32), and rs110328877 mapped within a predicted TFBS for the C2H2 zinc-finger Zfp335 (MA2002.2; absolute score 10.96).

Beyond lactation-related genes, other examples can be linked to traits studied in beef cattle. The gene Phosphatidylinositol Glycan Anchor Biosynthesis Class P (*PIGP*) has been associated with beef fatty acid composition [41] and carcass weight [42], as well as with feed intake [43] and energy balance [44]. Expression of the *PIGP* gene was correlated with chromatin accessibility at peak 4929 (0.808), and genotypic correlations between the eQTL and caQTL for the same peak was 0.994 across 1266 markers.

None of the eQTL tag variants map within the chromatin peak, with the closest falling approximately 21.6kbp away. Another gene, Cystatin B (*CSTB*), has been linked to oocyte maturation [45] (important for female fertility) and intramuscular fat [46, 47] (important for meat quality). A phenotypic correlation between gene expression and chromatin accessibility of 0.905 was observed for peak 4468, plus a genotypic correlation of 0.990 across 1030 markers for the eQTL and caQTL of the same peak. Five of the tag variants for the *CSTB* eQTL mapped within peak 4468, making these the best candidate causative variants for these QTL. All five variants overlap with predicted TFBSs, with the most significant being the variants rs382131150, which alters a PATZ1 binding site (MA1961.2; absolute score 16.54), and rs208187316, which alters a predicted STAT4 binding site (MA0518.2; absolute score 16.06).

Discussion

The chromatin accessibility data provided by ATAC-seq experiments can be used in several ways. For example, researchers have used these datasets to distinguish between similar-looking cell types, such as pancreatic *α*- and *β*-cells [48], or to profile immune function by identifying the clonal origins of T lymphocytes [49]. Another common use is to study epigenetic changes in chromatin structure, for example in relation to cancer [50]. Another primary use – and the one explored here – is to combine ATAC-seq datasets with gene expression data to investigate the effects of chromatin structure and accessibility on gene expression. Frequently, these multi-omics datasets are combined with genetic information to yield eQTL and caQTL. Most studies to date have been performed using humans or mice (approx 90%) [50]. For example, a recent study by Varshney et al. [51] investigated the genetic regulatory architecture of human skeletal muscle development using single-nucleus eQTLs and caQTLs and found that over 3500 GWAS signals co-localised with these QTL. A second study by Arthur et al. [52] used eQTL, caQTL, and histone acetylation QTL (haQTL) from induced human stem cells to highlight candidate causal variants underlying 296 GWAS QTL using TF binding motifs. A third [53] study examined co-located eQTL and caQTL in 23 cancer types to identify caQTL that influenced cancer prognosis.

To the best of our knowledge, no large caQTL studies have previously been conducted in livestock. Previously, some large-scale ATAC-seq experiments in cattle have been undertaken, but these focused on comparing multiple tissues from a small number of animals. For example, Yuan et al. [9] created a catalogue of almost 1 million ATACseq elements from 63 tissues, mostly from a single Holstein animal, then looked at enrichment of TFBSs in both tissue-specific or ubiquitous peaks. Other bovine studies [10, 11] have used ATAC-seq to explore the regulation of skeletal muscle using small numbers of animals. Similarly, the authors have previously used caQTL from mammary tissue to investigate the regulation of the *LTF* locus [12]. Large scale generation of data and QTL for epigenetic modifications, including chromatin accessibility, are scheduled as part of Phase 2 of the FarmGTEX project [5]; however, this is currently planned for 2030–2035.

Open chromatin regions with negative effects on gene expression

Overall, significant positive correlations were observed more frequently than negative between gene expression and chromatin accessibility. Likewise, more positive correlations were observed between eQTL marker effects and caQTL. Similarly, when ATAC-seq peaks with multiple gene expression correlations were categorised as MP or MN, more MP peaks were identified. Together, these observations suggest that increased chromatin accessibility near a gene will generally lead to increased gene expression, as expected. Many of the accessible chromatin windows identified here are proximal to the TSS, so are likely to be promoter elements. This is reinforced by the observation that, the closer a peak is to the TSS, the more positive on average the correlation between gene expression and chromatin accessibility is likely to be.

The nature of the peaks that showed negative correlations with gene expression is unknown. One hypothesis is that these represent silencer elements [54]. A recent study by Doni Jayavelu et al. [55] used Massively-Parallel Reporter Assay (MPRA) data to train a support vector machine algorithm that identified over 1.7 million putative human silencer elements, suggesting that silencers are common in mammalian genomes. These elements are expected to contain binding sequences for TFs such as REST (MA0138.3) [56]. However, the current experiment identified only one REST element across all peaks, and none of the TFs identified by Doni Jayavelu et al. had motifs that were significantly enriched in the MN peaks. A second hypothesis is that these open chromatin regions represent insulator sequences, specifically, enhancerblocking elements interposed between the TSS and a distal enhancer element [57, 58]. These elements are related to 3D chromatin structure and topologically associated domains (TADs) [30]. Activating an insulator element by increasing chromatin accessibility could decrease gene expression by blocking the action of a nearby enhancer element. Binding sites for CTCF were predicted significantly more frequently under MN peaks, suggesting that at least some of these peaks represent insulator elements.

Transcription factor classes

The most commonly predicted TFBSs were those belonging to C2H2 zinc fingers. These are known to be common in mammals, and have numerous and diverse functions [59, 60]. In the current study, the most frequently predicted binding site was for the C2H2 zinc finger MAZ, often binding to the same sites as a second C2H2 factor, SP1 [61]. Functionally, MAZ is required to bypass *MYB* promoter repression and allow cell cycle progression [62], and also increases expression of HBA in humans [63] and activates signalling pathways important for ocular development [64]. Unlike these functions, however, the current study shows that ATAC-seq peaks containing predicted MAZ TFBSs are more frequently associated with reduced gene expression, suggesting a possible role as a transcription silencer. Indeed, MAZ has been shown to suppress the expression of the *MAZ* gene itself, alongside SP1 [65]. This regulation is thought to be due to the recruitment of histone acetylases, deactylases, and methylases to specific transcriptional complexes [65]. MAZ has also been shown to act as an insulator element that can block

the effect of distal enhancer elements on gene expression, suggesting a possible mechanism by which this TF could be associated with decreased gene expression [66].

By contrast, binding sites for STAT-domain containing factors were observed significantly more often in MP peaks compared to MN (Table 3), and in Singleton peaks (Table 4). This finding suggests that these TFs are involved in promoting mammary gene expression. Significantly fewer binding sites were predicted under MN peaks for both STAT5 and STAT1. STAT domain factors act through the JAK-STAT signalling pathway, having general roles in cell proliferation and differentiation, and regulation of apoptosis [67, 68]. STAT5 mediates signals from hormones such as prolactin and growth hormone to promote mammary cell differentiation and milk protein expression [69–72], while STAT3 is linked to mammary gland involution [73–75]. Milk trait QTLs on bovine chromosome 19 have been linked to variants in both STAT3 and STAT5 [76, 77]. The transcription factor STAT1 has been shown to protect against some forms of breast cancer [78], and may have other functions during mammary gland infection, i.e., mastitis [79]. Markers near the *STAT1* gene have also been associated with milk phenotypes [80].

Conclusions

This study presents one of the first large ATAC-seq studies designed to allow the detection of caQTL in livestock. Phenotypic correlations were frequently observed between gene expression and chromatin accessibility in nearby peaks. Strong genotypic correlations between the SNP effects for eQTL and caQTL were also frequently observed, often showing stronger correlations than the corresponding phenotypes. The majority of the correlations observed were positive, showing the expected increase in gene expression associated with an increase in chromatin accessibility. Transcription factor binding site predictions for sequences under ATAC-seq peaks that had positive correlations with gene expression were enriched for STAT-domain factors, a family of TFs that is known to include important regulators of lactation. By contrast, peaks showing negative correlations with gene expression were enriched for CTCF but not for known markers of silencer elements, suggesting that many of these peaks represent insulator elements.

The chromatin accessibility regions identified herein will allow for improved annotation of regulatory regions in cattle, where data are less well represented than in humans or model mammalian species such as mice. This has the potential to improve the prediction of marker effects in regulatory regions, facilitating the inclusion of these variants into genomic prediction for traits of both economic and animal welfare importance. Together, the results presented herein show that these data could be usefully combined with other omics datasets to elucidate the genetic control of gene expression in cattle.

Declarations

Ethics approval and consent to participate.

No animals were sacrificed for this study. All animal experiments were conducted in strict accordance with the rules and guidelines outlined in the New Zealand Animal Welfare Act 1999. Biopsy samples for RNA-seq and ATAC-seq were obtained in accordance with protocols approved by the Ruakura Animal Ethics Committee, Hamilton, New Zealand (approval AEC 12845). These cows were situated on a research farm and permission was sought and obtained to biopsy mammary tissue from the owner of these animals (AgResearch, NZ).

Consent for publication.

Not applicable

Competing interests.

TL and ML are employees of Livestock Improvement Corporation, a commercial provider of bovine germplasm. No other authors declare any conflicts of interest.

Funding.

RNA-seq sampling was supported by the Ministry for Primary Industries (Wellington, New Zealand), who co-funded the work through the Primary Growth Partnership. The ATAC-seq analyses were co-funded by the Ministry for Business, Innovation, and Employment as part of the NextGen GS programme. External funders had no role in the design of the experiment, the collection, analysis or interpretation of the data, or writing the manuscript.

Authors’ contributions.

AJT, JM, and RGS performed the ATAC-seq experiment, with advice from HT. TJL performed the bioinformatic and statisitical analyses. RGS and MDL were involved in supervision of the project. TJL wrote the manuscript, with assistance from MDL, AJT, and RGS. All authors have read and approved the manuscript.

Acknowledgements.

The authors wish to thank Dr Haruko Takeda (University of Liège, Liège, Belgium) for her expertise and assistance with the ATAC-seq experiment. The authors also wish to acknowledge use of the eResearch Infrastructure Platform hosted by the Crown company, Research and Education Advanced Network New Zealand (REANNZ) Ltd., and funded by the Ministry of Business, Innovation & Employment. URL: https://www.reannz.co.nz

Availability of data and materials.

RNA-seq alignments are available on the sequence read archive (SRA) under BioProject PRJNA682457. Peak call locations and statistics are available in Supplementary File 1. Phenotypic with gene expression and genotypic correlations with eQTL are available in Supplementary Files 2 and 3.

References

1. Lopdell TJ, Tiplady K, Struchalin M, Johnson TJ, Keehan M, Sherlock R et al (2017) DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. BMC Genomics 18(1):968. https://doi.org/10.1186/s12864-017-4320-3
2. Van Den Berg I, Hayes B, Chamberlain A, Goddard M (2019) Overlap between eQTL and QTL associated with production traits and fertility in dairy cattle. BMC Genomics 20:291. https://doi.org/10.1186/s12864-019-5656-7
3. Cai W, Zhang Y, Chang T, Wang Z, Zhu B, Chen Y et al (2023) The eQTL colocalization and transcriptome-wide association study identify potentially causal genes responsible for economic traits in Simmental beef cattle. J Anim Sci Biotechnol 14:78. https://doi.org/10.1186/s40104-023-00876-7
4. Liu S, Gao Y, Canela-Xandri O, Wang S, Yu Y, Cai W et al (2022) A multi-tissue atlas of regulatory variants in cattle. Nat Genet 54(9):1438–1447. https://doi.org/10.1038/s41588-022-01153-5
5. Fang L, Teng J, Lin Q, Bai Z, Liu S, Guan D et al (2025) The farm animal genotype– tissue expression (FarmGtex) project. Nat Genet 57:786–796. https://doi.org/10.1038/s41588-025-02121-5
6. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T et al (2007) Genomewide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4(8):651–657. https://doi.org/10.1038/nmeth1068
7. Prowse-Wilkins CP, Lopdell TJ, Xiang R, Vander Jagt CJ, Littlejohn MD, Chamberlain AJ et al (2022) Genetic variation in histone modifications and gene expression identifies regulatory variants in the mammary gland of cattle. BMC Genomics 23:815. https://doi.org/10.1186/s12864-022-09002-9
8. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ (2015) ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol 109:21–29. https://doi.org/10.1002/0471142727.mb2129s109
9. Yuan C, Tang L, Lopdell T, Petrov VA, Oget-Ebrad C, Moreira GCM et al (2023) An organism-wide ATAC-seq peak catalog for the bovine and its use to identify regulatory variants. Genome Res 33(10):1848–1864. https://doi.org/10.1101/gr.277947.123
10. Wang J, Li B, Yang X, Liang C, Raza SHA, Pan Y et al (2022) Integration of RNAseq and ATAC-seq identifies muscle-regulated hub genes in cattle. Front Veterinary Sci 9:925590. https://doi.org/10.3389/fvets.2022.925590
11. Cao X, Cheng J, Huang Y, Lan X, Lei C, Chen H (2021) Comparative enhancer map of cattle muscle genome annotated by ATAC-Seq. Front veterinary Sci 8:782409.

https://doi.org/10.3389/fvets.2021.782409

12. Lopdell TJ, Trevarton AJ, Moody J, Prowse-Wilkins C, Knowles S, Tiplady K et al (2024) A common regulatory haplotype doubles lactoferrin concentration in milk. Genet Selection Evol 56:22. https://doi.org/10.1186/s12711-024-00890-x
13. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T et al (2016) Sequence-based association analysis reveals an *MGST1* eQTL with pleiotropic effects on bovine milk composition. Sci Rep 6:25376. https://doi.org/10.1038/srep25376
14. Spelman R, Miller F, Hooper J, Thielen M, Garrick D (2001) Experimental design for QTL trial involving New Zealand Friesian and Jersey breeds. In: Proceedings of the Association for the Advancement of Animal Breeding and Genetics. vol. 14. AAABG; pp. 393–396. Available from: http://livestocklibrary.com.au/handle/1234/5512
15. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S et al (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods 14(10):959–962. https://doi.org/10.1038/nmeth.4396
16. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170
17. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWAMEM. arXiv 13033997. https://doi.org/10.48550/arXiv.1303.3997. [qbio.GN]
18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352
19. Feng J, Liu T, Qin B, Zhang Y, Liu XS (2012) Identifying ChIP-seq enrichment using MACS. Nature Protocols. ;7(9):1728–1740. https://doi.org/10.1038/nprot. 2012.101
20. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635
21. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30(7):923–930. https://doi.org/10.1093/bioinformatics/btt656
22. R Core Team R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.R-project.org/
23. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8
24. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80. https://doi.org/10.1186/gb-2004-5-10-r80
25. Browning BL, Zhou Y, Browning SR (2018) A one-penny imputed genome from next-generation reference panels. Am J Hum Genet 103(3):338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

26. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR et al (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565–569. https://doi.org/10.1038/ng.608
27. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22(9):1813–1831. https://doi.org/10.1101/gr.136184.111
28. Afrache H, Gouret P, Ainouche S, Pontarotti P, Olive D (2012) The butyrophilin (BTN) gene family: from milk fat to the regulation of the immune response. Immunogenetics 64(11):781–794. https://doi.org/10.1007/s00251-012-0619-z
29. Ogg SL, Weldon AK, Dobbie L, Smith AJ, Mather IH (2004) Expression of butyrophilin (Btn1a1) in lactating mammary gland is essential for the regulated secretion of milk-lipid droplets. Proceedings of the National Academy of Sciences. ;101(27):10084–10089. https://doi.org/10.1073/pnas.0402930101
30. Phillips JE, Corces VG (2009) CTCF: master weaver of the genome. Cell 137(7):1194–1211. https://doi.org/10.1016/j.cell.2009.06.001
31. Ong CT, Corces VG (2014) CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet 15(4):234–246. https://doi.org/10.1038/nrg3663
32. Ron M, Israeli G, Seroussi E, Weller JI, Gregg JP, Shani M et al (2007) Combining mouse mammary gland gene expression and comparative mapping for the identification of candidate genes for QTL of milk production traits in cattle. BMC Genomics 8:183. https://doi.org/10.1186/1471-2164-8-183
33. Lopdell TJ (2023) Using QTL to identify genes and pathways underlying the regulation and production of milk components in cattle. Animals 13(5):911. https://doi.org/10.3390/ani13050911
34. Lopdell TJ, Tiplady K, Couldrey C, Johnson TJ, Keehan M, Davis SR et al (2019) Multiple QTL underlie milk phenotypes at the *CSF2RB* locus. Genet Selection Evol 51:3. https://doi.org/10.1186/s12711-019-0446-x
35. Fu¨rbass R, Winter A, Fries R, Kuhn C (2006) Alleles of the bovine *DGAT1* variable number of tandem repeat associated with a milk fat QTL at chromosome 14 can stimulate gene expression. Physiol Genom 25(1):116–120. https://doi.org/10.1152/physiolgenomics.00145.2005
36. Fink T, Lopdell TJ, Tiplady K, Handley R, Johnson TJ, Spelman RJ et al (2020) A new mechanism for a familiar mutation–bovine *DGAT1* K232A modulates gene expression through multi-junction exon splice enhancement. BMC Genomics 21(1):591. https://doi.org/10.1186/s12864-020-07004-z
37. Tiplady KM, Lopdell TJ, Reynolds E, Sherlock RG, Keehan M, Johnson TJ et al (2021) Sequence-based genome-wide association study of individual milk midinfrared wavenumbers in mixed-breed dairy cattle. Genet Selection Evol 53:62. https://doi.org/10.1186/s12711-021-00648-9
38. Xiang R, Hayes BJ, Vander Jagt CJ, MacLeod IM, Khansefid M, Bowman PJ et al (2018) Genome variants associated with RNA splicing variations in bovine are extensively shared between tissues. BMC Genomics 19:521. https://doi.org/10.1186/s12864-018-4902-8

39. Ganai N, Bovenhuis H, Van Arendonk J, Visker M (2009) Novel polymorphisms in the bovine *β-lactoglobulin* gene and their effects on β-lactoglobulin protein concentration in milk. Anim Genet 40(2):127–133. https://doi.org/10.1111/j.1365-2052.2008.01806.x
40. Davis SR, Ward HE, Kelly V, Palmer D, Ankersmit-Udy AE, Lopdell TJ et al (2022) Screening for phenotypic outliers identifies an unusually low concentration of a β-lactoglobulin B protein isoform in bovine milk caused by a synonymous SNP. Genet Selection Evol 54:22. https://doi.org/10.1186/s12711-022-00711-z
41. Kelly M, Tume R, Fortes M, Thompson J (2014) Whole-genome association study of fatty acid composition in a diverse range of beef cattle breeds. J Anim Sci 92(5):1895–1901. https://doi.org/10.2527/jas.2013-6901
42. Putra W (2023) Hartati. Genome-Wide Association Study for Body Weight and Carcass Weight in Sumba Ongole Bulls (*Bos indicus*). Trop Anim Sci J 46(4):389–395. https://doi.org/10.5398/tasj.2023.46.4.389
43. Zhang F, Wang Y, Mukiibi R, Chen L, Vinsky M, Plastow G et al (2020) Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: I: feed efficiency and component traits. BMC Genomics 21:36. https://doi.org/10.1186/s12864-019-6362-1
44. Hu H, Atashi H, Franceschini S, Lemal P, Grelet C, Chen Y et al (2025) Using singlestep genome-wide association analyses to compare predicted negative energy balance and a novel energy deficiency score in early-lactation Holstein cows. JDS Commun 6(6):792–796. https://doi.org/10.3168/jdsc.2025-0778
45. Chen Y, Abudureyimu G, Lin J, Wang L, Ma X, Wu X et al (2025) Effects of CSTB on in vitro maturation of ovine oocytes. Anim Reprod Sci 276:107839. https://doi.org/10.1016/j.anireprosci.2025.107839
46. Wang YH, Bower N, Reverter A, Tan S, De Jager N, Wang R et al (2009) Gene expression patterns during intramuscular fat development in cattle. J Anim Sci 87(1):119–130. https://doi.org/10.2527/jas.2008-1082
47. Peletto S, Strillacci M, Capucchio M, Biasibetti E, Modesto P, Acutis P et al (2017) Genetic basis of Lipomatous Myopathy in Piedmontese beef cattle. Livest Sci 206:9–16. https://doi.org/10.1016/j.livsci.2017.09.027
48. Ackermann AM, Wang Z, Schug J, Naji A, Kaestner KH (2016) Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol Metabolism 5(3):233–244. https://doi.org/10.1016/j.molmet.2016.01.002
49. Satpathy AT, Saligrama N, Buenrostro JD, Wei Y, Wu B, Rubin AJ et al (2018) Transcript-indexed ATAC-seq for precision immune profiling. Nat Med 24(5):580–590. https://doi.org/10.1038/s41591-018-0008-8
50. Luo L, Gribskov M, Wang S (2022) Bibliometric review of ATAC-Seq and its application in gene expression. Brief Bioinform 23(3):bbac061. https://doi. org/10.1093/bib/bbac061

51. Varshney A, Manickam N, Orchard P, Tovar A, Ventresca C, Zhang Z et al (2024) Population-scale skeletal muscle single-nucleus multi-omic profiling reveals extensive context specific genetic regulation. bioRxiv 571696. https://doi.org/10.1101/2023.12.15.571696
52. Arthur TD, Nguyen JP, Henson BA, D'Antonio-Chronowska A, Jaureguy J, Silva N et al (2025) Multiomic QTL mapping reveals phenotypic complexity of GWAS loci and prioritizes putative causal variants. Cell Genomics 5:100775. https://doi.org/10.1016/j.xgen.2025.100775
53. Wang D, Wu X, Jiang G, Yang J, Yu Z, Yang Y et al (2022) Systematic analysis of the effects of genetic variants on chromatin accessibility to decipher functional variants in non-coding regions. Front Oncol 12:1035855. https://doi.org/10.3389/fonc.2022.1035855
54. Laimins L, Holmgren-Koenig M, Khoury G (1986) Transcriptional silencer element in rat repetitive sequences associated with the rat insulin 1 gene locus. Proceedings of the National Academy of Sciences. ;83(10):3151–3155. https://doi.org/10.1073/pnas.83.10.3151
55. Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD (2020) Candidate silencer elements for the human and mouse genomes. Nat Commun 11:1061. https://doi.org/10.1038/s41467-020-14853-5
56. Chong JA, Tapia-Ramirez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC et al (1995) REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. Cell 80(6):949–957. https://doi.org/10.1016/0092-8674(
57. West AG, Gaszner M, Felsenfeld G (2002) Insulators: many functions, many mechanisms. Genes Dev 16(3):271–288. https://doi.org/10.1101/gad
58. Gaszner M, Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet 7(9):703–713. https://doi.org/10.1038/nrg1925
59. Iuchi S (2001) Three classes of C_2H_2 zinc finger proteins. Cell Mol Life Sci CMLS 58(4):625–635. https://doi.org/10.1007/PL00000885
60. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A et al (2015) C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat Biotechnol 33(5):555–562. https://doi.org/10.1038/nbt.3128
61. Williams L, Abou-Samra A (2000) The transcription factors SP1 and MAZ regulate expression of the parathyroid hormone/parathyroid hormone-related peptide receptor gene. J Mol Endocrinol 25(3):309–319. https://doi.org/10.1677/jme.0.0250309
62. Alvaro-Blanco J, Urso K, Chiodo Y, Mart´ın-Cort´azar C, Kourani O, Arco PG et al (2017) MAZ induces *MYB* expression during the exit from quiescence via the E2F site in the *MYB* promoter. Nucleic Acids Res 45(17):9960–9975. https://doi.org/10.1093/nar/gkx641
63. Deen D, Butter F, Daniels DE, Ferrer-Vicens I, Ferguson DC, Holland ML et al (2021) Identification of the transcription factor MAZ as a regulator of erythropoiesis. Blood Adv 5(15):3002–3015. https://doi.org/10.1182/bloodadvances. 2021004609
64. Medina-Martinez O, Haller M, Rosenfeld JA, O'Neill MA, Lamb DJ, Jamrich M (2020) The transcription factor Maz is essential for normal eye development. Dis Models Mech 13(8):dmm044412. https://doi.org/10.1242/dmm. 044412

65. Song J, Ugai H, Nakata-Tsutsui H, Kishikawa S, Suzuki E, Murata T et al (2003) Transcriptional regulation by zinc-finger proteins Sp1 and MAZ involves interactions with the same cis-elements. Int J Mol Med 11(5):547–553. https://doi.org/10.3892/ijmm.11.5.547
66. Xiao T, Li X, Felsenfeld G (2021) The Myc-associated zinc finger protein (MAZ) works together with CTCF to control cohesin positioning and genome organization. Proceedings of the National Academy of Sciences. ;118(7):e2023127118. https://doi.org/10.1073/pnas.2023127118
67. Rawlings JS, Rosler KM, Harrison DA (2004) The JAK/STAT signaling pathway. J Cell Sci 117(8):1281–1283. https://doi.org/10.1242/jcs. 00963
68. Hu X, Li J, Fu M, Zhao X, Wang W (2021) The JAK/STAT signaling pathway: from bench to clinic. Signal Transduct Target Therapy 6:402. https://doi.org/10.1038/s41392-021-00791-1
69. Kazansky AV, Raught B, Lindsey SM, Wang Y, Rosen JM (1995) Regulation of mammary gland factor/Stat5a during mammary gland development. Mol Endocrinol 9(11):1598–1609. https://doi.org/10.1210/mend.9.11
70. Liu X, Robinson GW, Hennighausen L (1996) Activation of Stat5a and Stat5b by tyrosine phosphorylation is tightly linked to mammary gland differentiation. Mol Endocrinol 10(12):1496–1506. https://doi.org/10.1210/mend. 10.12.8961260
71. Liu X, Robinson GW, Wagner KU, Garrett L, Wynshaw-Boris A, Hennighausen L (1997) Stat5a is mandatory for adult mammary gland development and lactogenesis. Genes Dev 11(2):179–186. https://doi.org/10.1101/gad.11. 2.179
72. Gallego MI, Binart N, Robinson GW, Okagaki R, Coschigano KT, Perry J et al (2001) Prolactin, growth hormone, and epidermal growth factor activate Stat5 in different compartments of mammary tissue and exert different and overlapping developmental effects. Dev Biol 229(1):163–175. https://doi.org/10.1006/dbio.2000.9961
73. Hennighausen L, Robinson GW, Wagner KU, Liu X (1997) Developing a mammary gland is a st¯at affair. J Mammary Gland Biol Neoplasia 2(4):365–372. https://doi.org/10.1023/A:1026347313096
74. Chapman RS, Lourenco PC, Tonner E, Flint DJ, Selbert S, Takeda K et al (1999) Suppression of epithelial apoptosis and delayed mammary gland involution in mice with a conditional knockout of Stat3. Genes Dev 13(19):2604–2616
75. Hughes K, Watson CJ (2018) The multifaceted role of STAT3 in mammary gland involution and breast cancer. Int J Mol Sci 19(6):1695. https://doi.org/10.3390/ijms19061695
76. Usman T, Wang Y, Liu C, Wang X, Zhang Y, Yu Y (2015) Association study of single nucleotide polymorphisms in *JAK2* and *STAT5B* genes and their differential mRNA expression with mastitis susceptibility in Chinese Holstein cattle. Anim Genet 46(4):371–380. https://doi.org/10.1111/age.12306
77. Ratcliffe L, Mullen M, McClure M, McClure J, Kearney F (2017) Single nucleotide polymorphisms in the signal transducer and regulator of transcription (STAT) genes are associated with milk production, milk composition, and fertility traits in Holstein Friesian cattle. J Anim Sci 95(suppl 4):94–94. https://doi.org/10.2527/asasann.2017.190

78. Chan SR, Vermi W, Luo J, Lucini L, Rickert C, Fowler AM et al (2012) STAT1-deficient mice spontaneously develop estrogen receptor α-positive luminal mammary carcinomas. Breast Cancer Res 14:R16. https://doi.org/10.1186/bcr3100
79. Khan MZ, Khan A, Xiao J, Ma Y, Ma J, Gao J et al (2020) Role of the JAK-STAT pathway in bovine mastitis and milk production. Animals 10(11):2107. https://doi.org/10.3390/ani10112107
80. Cobanoglu O, Zaitoun I, Chang Y, Shook G, Khatib H (2006) Effects of the signal transducer and activator of transcription 1 (*STAT1*) gene on milk production traits in Holstein dairy cattle. J Dairy Sci 89(11):4433–4437. https://doi.org/10.3168/jds.S0022-0302(06)72491-2

Figures

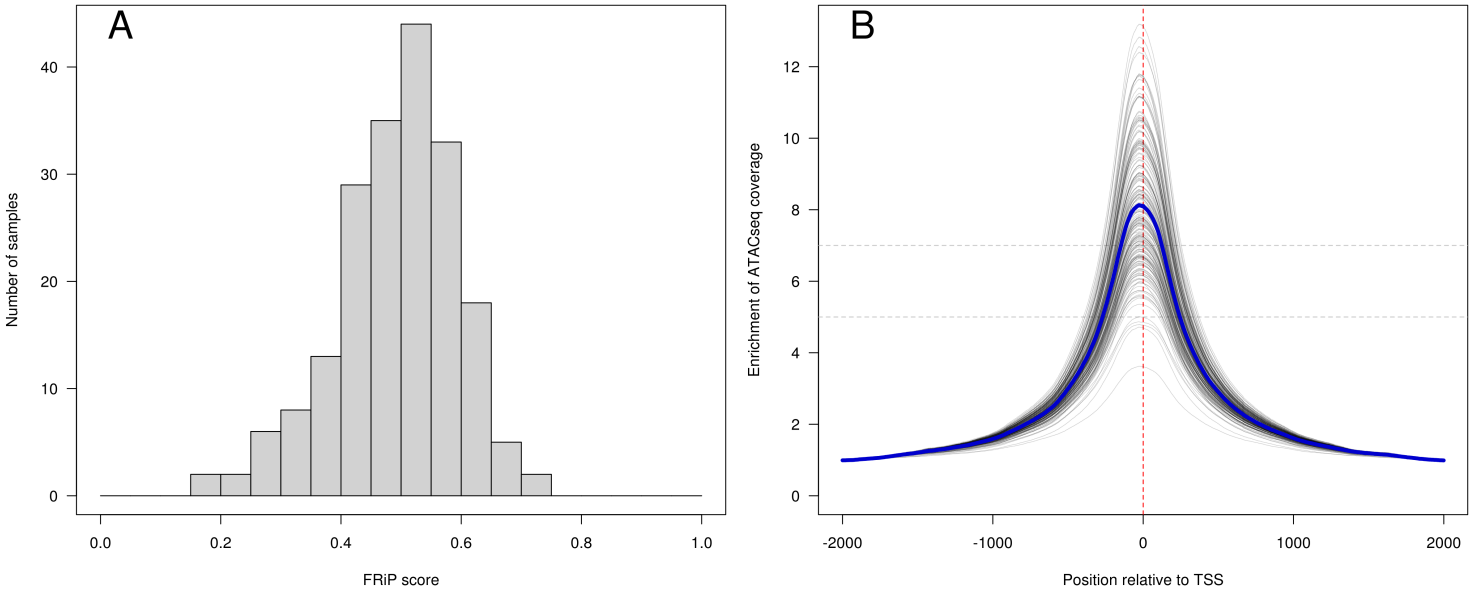


Figure 1

A: the distribution of the fraction of ATAC-seq reads in peaks (FRiP) statistic across all samples. **B**: enrichment of ATAC-seq read coverage by distance from the nearest transcription start site (TSS). Grey lines represent individual animals, and the blue line is the average of all animals at each position.

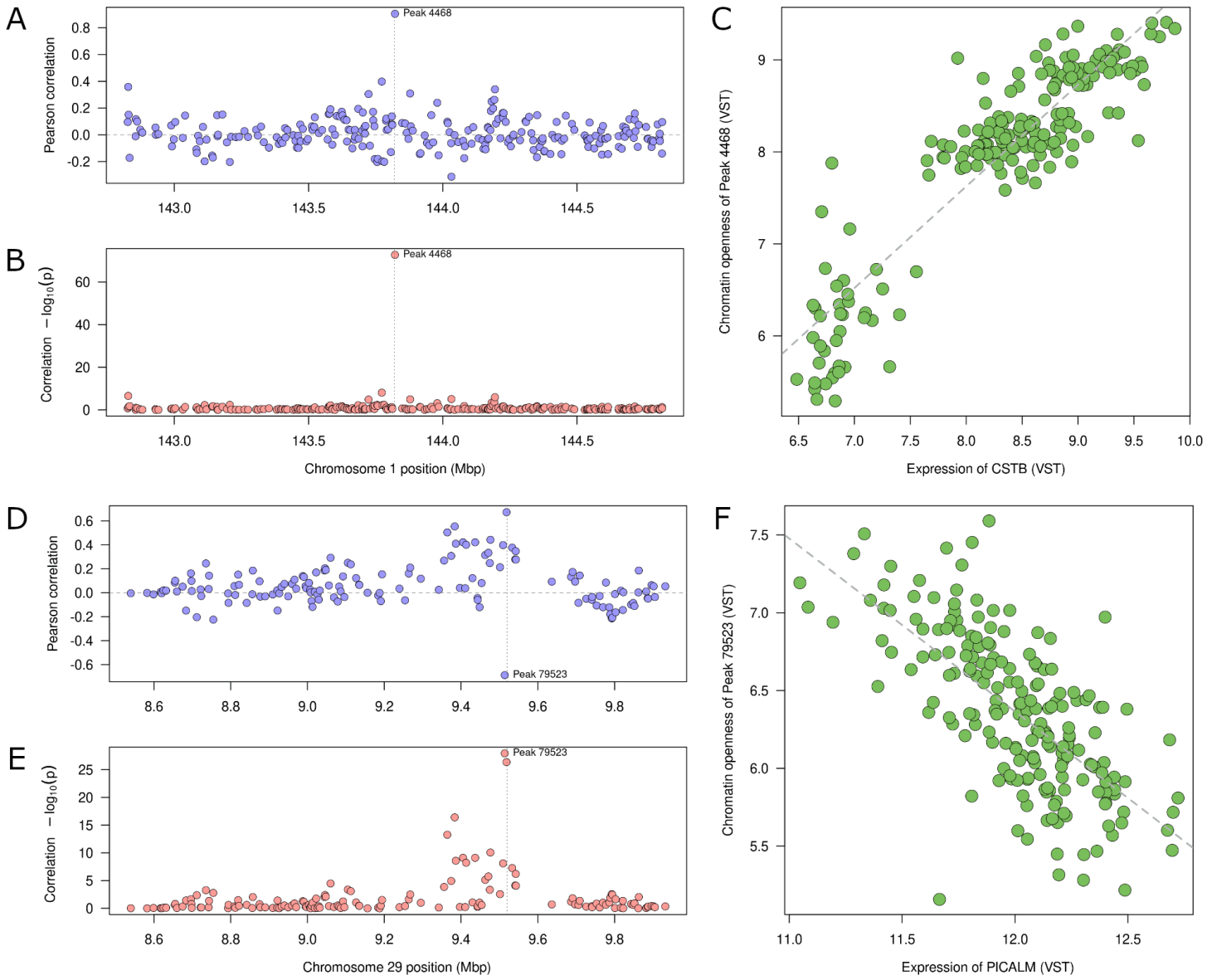


Figure 2

Examples of two genes (CSTB and PICALM) illustrating correlations between gene expression (GE) and chromatin accessibility (CA) at nearby ATAC-seq peaks. The strongest correlations occur close to the gene locus. **A** and **D**: the Pearson correlations between GE and CA, where each dot represents one peak. The vertical grey line shows the gene locus. **B** and **E**: Manhattan plots of the p-values for the correlations, with the null hypothesis H_0: r = 0. **C** and **F**: GE compared to CA for the peak with the largest correlation, showing the linear relationships between the two phenotypes.

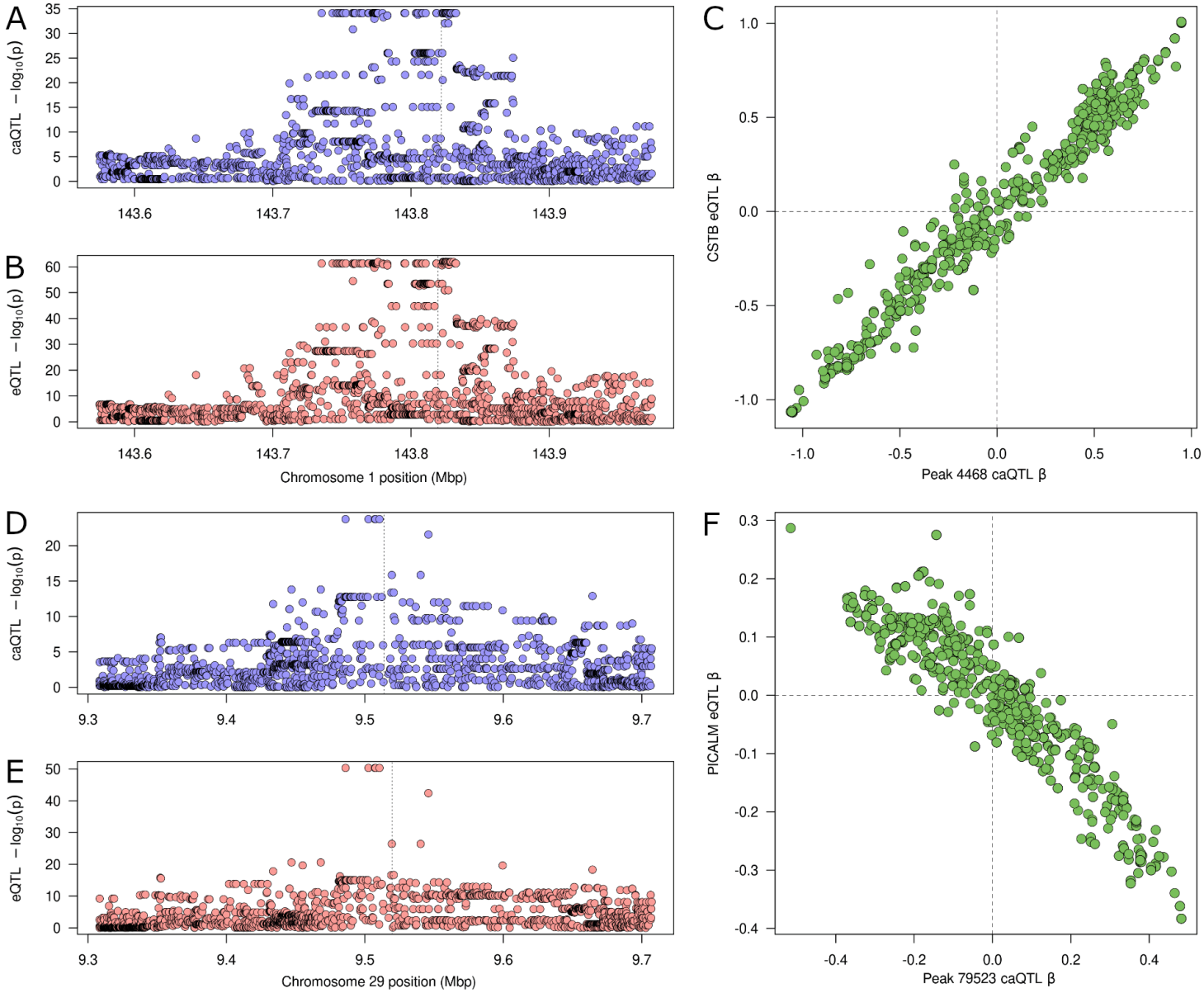


Figure 3

Examples of two genes (CSTB and PICALM) illustrating correlations between SNP effects for eQTL and caQTL. **A** and **D**: Manhattan plots of the caQTL for peaks 4468 and 79523. The grey vertical lines indicate the position of the peak. **B** and **E**: Manhattan plots of the eQTL, with a vertical line showing the gene location. **C** and **F**: SNP effects (βs) for the eQTL plotted against those for the caQTL, showing strong linear relationships between the two QTL, suggesting genetic co-regulation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppfile1.narrowpeak