BMC Genomics

RESEARCH Open Access

Bridging GWAS to genes: an integrative multi-omics approach using cattle data

Mohammad Ghoreishifar[1,2*], Iona M. Macleod[1,2], Tuan Nguyen[1], Thomas J. Lopdell[3,4], Mathew D. Littlejohn[3,4], Ruidong Xiang[1,5], Amanda J. Chamberlain[1,2], Jennie E. Pryce[1,2] and Michael E. Goddard[1,5]

Abstract

Background Genome-wide association studies (GWASs) have identified thousands of loci for complex traits, but pinpointing causal variants and linking them to target genes remains challenging. Several strategies have been proposed to address these challenges, e.g., comparisons across the genome, using larger and multi-breed datasets, multi-trait analyses, leveraging multi-omics data, etc.

Results We used a multi-breed dataset of over 81,000 cows from Australia, including Holstein, Jersey, and Australian Red, with phenotypes for milk lactose percentage (LP) and imputed sequence genotypes. LD pruning excluded SNPs with r2 > 0.95. We used BayesR to estimate SNP effects for LP (~ 1.1 million SNPs remained after LD pruning); These SNP effects were used to predict local genomic breeding values (GEBVs) for ~ 400 mammary RNA-sequenced cows from New Zealand. Then, genetic score omics regression (GSOR) was applied to test associations between observed gene expression and local GEBVs, identifying 711 significant genes (FDR ≤ 0.1) out of 12,000 genes expressed in the mammary gland. We developed a window-based test to investigate the significance of colocalization between GSOR results and GWAS summary statistics obtained from an independent study. We found 30 windows containing both GWAS signals and GSOR-significant genes (i.e., 34 genes); this overlap was significantly higher than chance expectation ($P_{Fisher} = 2.96 \times 10^{-9}$). Among the 34 genes analyzed, 20 contributed to the significantly enriched gene ontology term 'transmembrane transport' and its child terms (FDR < 0.05). These terms are relevant to the physiology of lactose production in the mammary gland.

Conclusions We hypothesized that the 20 genes are the most likely causal genes for the trait because: mammary expression of these genes was associated with GEBV for the trait, they were significantly colocalized with GWAS signals, and they were enriched in gene ontology terms relevant to physiology of the trait. Our approach provides strong evidence for causal genes supported by multiple lines of evidence (GWAS, GSOR, and functional enrichment) and demonstrates the power of multi-omics data integration.

*Correspondence:
Mohammad Ghoreishifar
mohammad.ghoreishifar@agriculture.vic.gov.au
[1]Agriculture Victoria Research, AgriBio Centre for AgriBioscience, Bundoora, VIC 3083, Australia
[2]School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia
[3]Research and Development, Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand
[4]AL Rae Centre for Genetics and Breeding, Massey University, Hamilton, New Zealand
[5]Faculty of Veterinary & Agricultural Science, University of Melbourne, Parkville, VIC 3010, Australia

Introduction

Genome-wide association studies (GWASs) have identified numerous genetic variants (such as single nucleotide polymorphisms or SNPs) associated with complex traits, yet the biological mechanisms connecting these variants to phenotypes remain elusive. Determining which variants are truly causal and which genes they affect is complicated by the extensive linkage disequilibrium (LD) surrounding lead SNPs [1]. In addition, because the majority of GWAS signals fall within noncoding regions of the genome [2], directly linking these variants to their target genes remains a significant challenge. Furthermore, the intricate regulatory mechanisms of genes, combined with the possibility that multiple genes within a single locus contribute to the trait, make identifying the true causal genes even more challenging [3].

To address these complexities, integrative approaches such as multi-trait multi-omics fine-mapping can help identify the genes through which SNPs influence quantitative traits. Many SNPs exhibit pleiotropic effects, influencing various biologically significant complex trait phenotypes [4, 5]. Consequently, employing multi-trait fine-mapping techniques that analyse several traits at once can enhance the power of fine-mapping. When examining two traits, one being a complex trait of interest and the other a molecular trait, like the expression of a particular gene [6], multi-trait fine-mapping resembles colocalization analysis [7]. This analysis evaluates the genetic connection between the traits by investigating whether they share the same causal variants at a specific locus [7]. An illustrative case of pleiotropy arises when a SNP associated with a complex trait through GWAS also influences gene expression, thereby acting as both a trait QTL and an expression QTL (eQTL) [6]. Such SNPs highlight the genetic link between gene expression and phenotypic variation [6]. In particular, genes through which QTL act can be inferred by integrating multi-omics (or multi-trait) data analyses: (i) trait-associated QTL can be identified from genomic data, allowing the assignment of nearby candidate genes to them; and (ii) genes located near trait-associated QTL that also show correlated expression with trait variation (i.e., *cis*-regulated by nearby variants) can be inferred from combined transcriptomic and genomic data [6].

The functional knowledge of genes serves as another source of evidence for mapping GWAS loci (QTL) to their target genes through which they operate [6, 8]. This information, organized in databases like the Gene Ontology (GO) resources [9], aids in pinpointing genes associated with specific biological functions. The functional enrichment of a subset of identified candidate genes within trait-relevant categories serves as additional evidence that the identified genes are the most likely causal genes [6]. In this study, we use milk lactose percentage (LP) as a model trait. LP is a useful model for dissecting the genetic architecture of complex traits because its synthesis is governed by a relatively simple and well-characterized pathway, mainly involving lactose synthase activity in the mammary gland [10, 11]. This feature makes lactose an ideal trait for testing integrative genomic approaches to identify causal genes.

Each of the three types of evidence including proximity of genes to trait-associated SNPs, correlation of gene's expression with the phenotype of interest, and functional enrichment of genes in a biological pathway—has limitations and can result in false positives or negatives [6]. Nonetheless, when multiple methods identify the same genes, our confidence in their biological significance increases. Expanding on our earlier integrative approach [6], we now apply this methodology to a different trait using a much larger multi-breed dataset to enhance statistical power and improve the robustness and generalizability of our findings.

The objectives of this study are (i) to perform gene-based association tests to identify significant gene expression–trait associations using genetic score omics regression (GSOR), introduced by Xiang et al. [12], (ii) to test the significance of window-based co-localization between GSOR-identified genes and GWAS loci reported by Lopdell et al. [13], and (iii) to obtain a list of candidate genes from the co-localization of GSOR-identified genes with GWAS loci for functional enrichment analyses. Overall, we aim to demonstrate the utility of combining GSOR with GWAS and functional enrichment analyses to map candidate genes to QTL, with milk LP as a model trait.

Results

Figure 1 illustrates Manhattan (Miami) plot showing mammary GSOR-identified genes in this study as well as trait associated SNPs identified through GWAS for milk lactose percentage (LP) reported by Lopdell et al. [13]. A total of 12, 237 genes were expressed in the mammary RNA-seq dataset [13–15], of which 711 were significantly associated with local GEBVs for milk LP in the GSOR analysis (FDR ≤ 0.10) (Additional file 1). Among the 12,237 expressed genes, 242 genes were located within 100 kb windows that contained at least one GWAS locus. Of those 242, 34 genes were also significant based on the GSOR analysis (Additional file 1). We also used a white blood cells (WBC) RNA-seq dataset [16–18]. Descriptive statistics about the GSOR analyses in both mammary and WBC datasets are presented in Table 1. A total of 12,536 genes were expressed in the WBC RNA-seq dataset, of which 986 were significantly associated with local GEBVs for milk LP in the GSOR analysis. Among the 12,536 expressed genes in WBC, 242 genes were located within 100 kb windows that contained at least one GWAS locus.

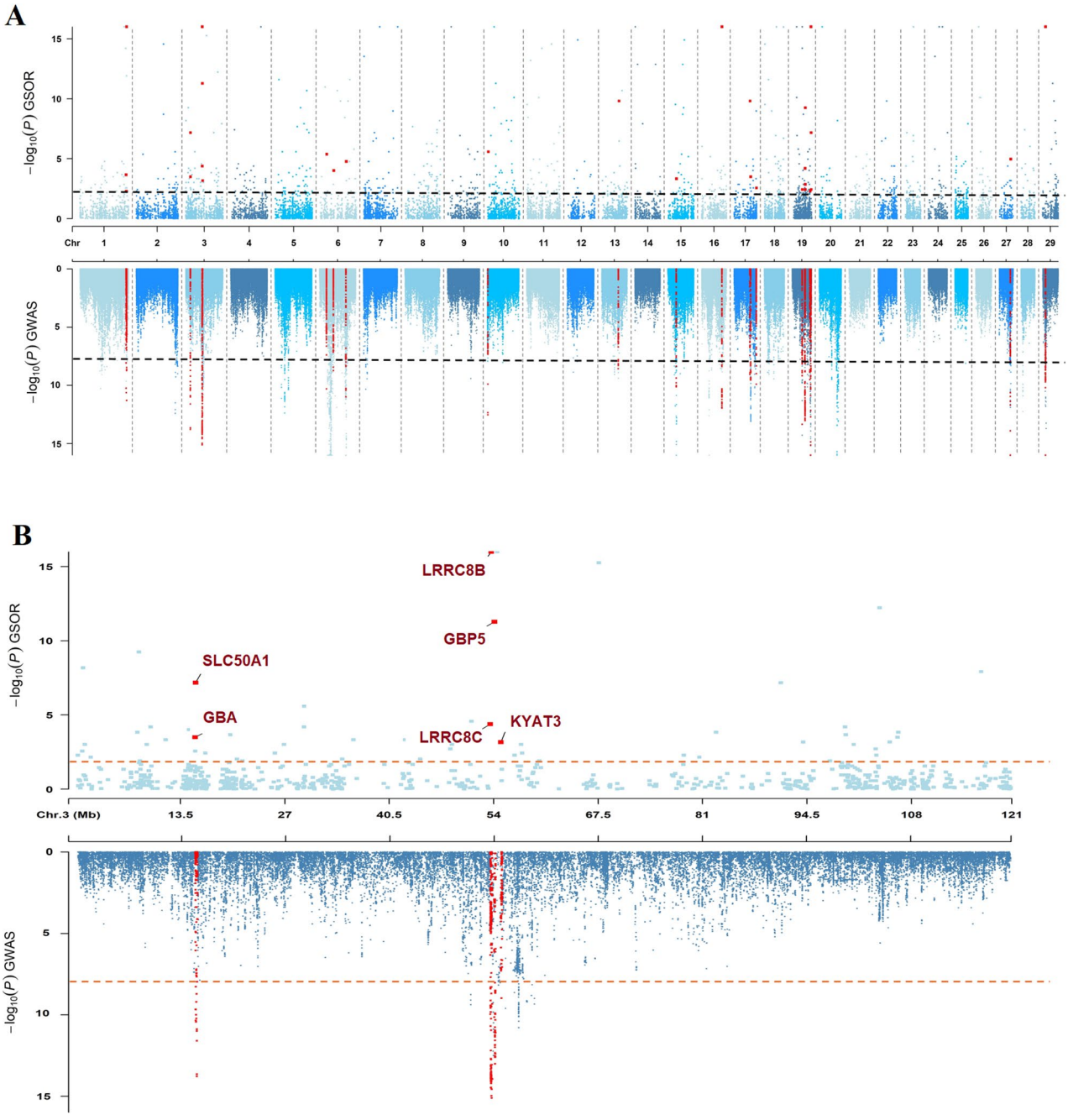

Fig. 1 Miami plot showing GSOR-identified genes (in the mammary RNA-seq data) and GWAS loci: (**A**) genome-wide overview and (**B**) detailed view of chromosome 3. The dashed lines in GSOR represents FDR = 0.01 (*P*-value = 0.005; upper plot) and in GWAS represents *P*-value 1×10^{-8} (lower plot). Overlapping GSOR genes with GWAS loci, defined using 100 kb windows, are highlighted in red. For visualization purposes, the highlighted GWAS regions are displayed larger than 100 kb to ensure they are visible. In addition, SNPs or genes with $p < 1 \times 10^{-16}$ were set to $p = 1 \times 10^{-16}$ for graphical representation

Of the 242 WBC genes, 33 genes were also significant based on the GSOR analysis (Table 1 & Additional file 2). These genes do not completely overlap with those identified using mammary RNA-seq.

Table 1 shows window-based co-localization between GSOR-identified genes and trait-associated SNPs (GWAS loci) obtained using 100 Kb non-overlapping windows. We tested statistical significance of this co-localization, and the results are presented in Fig. 2. Based on Fisher Exact test, our results show that co-localization between GSOR-identified genes and GWAS loci within 100 Kb and 500 Kb windows were significantly greater than expected by chance in both mammary and WBC datasets. For example, using 100 Kb windows with mammary RNA-seq data, we found 30 windows where the GSOR-identified genes shared windows with GWAS loci, 291

Table 1 Descriptive statistics of mammary and WBC GSOR genes and their co-localization with GWAS loci based on 100 Kb windows

RNA-seq data	Total expressed Genes	*N* significant GSOR genes[a]	*N* expressed genes in windows including GWAS loci[b]	*N* significant GSOR genes in windows including GWAS loci[c]
Mammary	12, 237	711	242	34
WBC	12, 536	986	242	33

[a]Number of GSOR-identified genes at FDR ≤ 0.1

[b]Number of expressed genes located within non-overlapping windows of 100 Kb length including at least one GWAS locus; For GWAS, summary statistics for milk LP from an independent study using 12,000 samples was used [13] (see Methods)

[c]Number of GSOR-identified genes located within 100 Kb windows including at least one GWAS loci

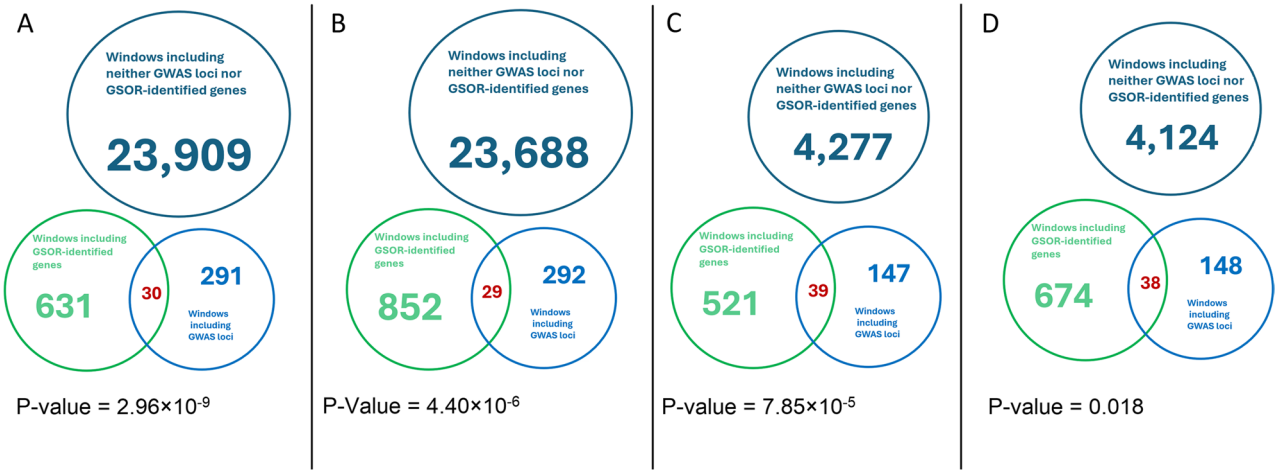

Fig. 2 Venn diagrams showing the co-occurrence between GSOR signals and GWAS loci within non-overlapping genomic windows, assessed using Fisher's exact test. Panels **A**–**D** display the number of non-overlapping windows containing only GSOR-identified gene(s), only GWAS signal(s), both signals, or neither. Panels **A** and **B** use 100 kb windows for mammary and white blood cell (WBC) GSOR analyses, respectively, while Panels **C** and **D** present the same analyses using 500 kb windows. SNPs with P-value ≤ 1×10^{-8} were considered GWAS loci, and genes with FDR ≤ 0.1 in GSOR were classified as significant

windows contained only GWAS loci, 631 included only GSOR-identified genes, while 23,909 windows had neither GWAS loci nor GSOR-identified gene (P_{Fisher} = 2.96×10^{-9} ; odds ratio = 3.9). We found that smaller window size (100 kb) revealed stronger co-localization, as evidenced by more significant *P* value and higher odds ratios (Fig. 2). Furthermore, across the RNA-seq datasets, mammary RNA-seq demonstrated stronger co-localization with GWAS loci than WBC RNA-seq, supported by more stringent *P* values and higher odds ratios (Fig. 2).

Table 2 shows descriptive statistics about the 34 mammary GSOR-identified genes co-localized with GWAS loci. Of the 34 co-localised GSOR-GWAS genes presented in Table 2, 31 were successfully converted to human ortholog and used in functional enrichment analyses, along with the 11,728 background genes that successfully converted (out of 12, 237 expressed genes). This data is presented in Additional file 3. Successfully converted genes were used to conduct enrichment analyses using gprofiler2 [19] R package.

We identified 22 significantly enriched gene ontology biological process (GO: BP) and Reactome pathways, where terms "transport", "transmembrane transport" and their child terms dominated the list (Table 3).

We also performed gene list enrichment analysis using WBC GSOR-identified genes co-localized with GWAS loci in the 100 Kb windows. This analysis revealed five significant GO terms, of which only the GO term "response to growth hormone" (GO:0060416) could be indirectly related to physiology of milk LP. These results are presented in Additional file 4. We found that only two genes, *STAT5B* and *HAP1* shared in both mammary and WBC GSOR-GWAS genes (Table 3 & Additional file 4). According to Wainberg et al. [20], transcriptome-wide association study or TWAS [21] is particularly susceptible to false-positive gene associations when expression data come from irrelevant tissues or cell types. It seems the same may also be true for GSOR [12], as the GSOR-identified genes from mammary gland expression data revealed functionally enriched terms that aligned more closely with milk lactose physiology, whereas the pathways detected from WBC data tended to be broader and less specific.

To assess whether a broader genomic window could improve the detection of relevant signals to LP, we

Table 2 GSOR-identified genes co-localized in 30 windows (100 Kb) with GWAS loci. P (GWAS) represents the most stringent P value within the corresponding window

Chr	Win Start	Win End	SNP	*P* (GWAS)	GSOR-gene	FDR (GSOR)
1	150,530,998	150,630,997	1:150591449	8.12E-09	KCNJ15	0.008
1	151,130,998	151,230,997	1:151180934	2.39E-11	ETS2	0.087
1	152,230,998	152,330,997	1:152324264	2.31E-09	SH3BP5	9.36E-49
3	15,344,539	15,444,538	3:15368547	3.75E-11	THBS3, GBA	0.057, 0.011
3	15,444,539	15,544,538	3:15464742	1.67E-14	SLC50A1	7.62E-06
3	53,544,539	53,644,538	3:53597089	3.28E-15	LRRC8C	0.002
3	53,644,539	53,744,538	3:53674128	7.86E-16	LRRC8B	1.01E-17
3	54,044,539	54,144,538	3:54109337	9.76E-14	GBP5	1.35E-09
3	54,844,539	54,944,538	3:54943906	1.04E-09	KYAT3	0.020
6	22,391,294	22,491,293	6:22469293	4.95E-11	SLC39A8	0.0003
6	45,091,294	45,191,293	6:45129256	3.64E-09	SLC34A2	0.004
6	85,991,294	86,091,293	6:86055882	1.06E-11	ENAM	0.001
10	2,166,988	2,266,987	10:2190824	2.94E-13	NREP	0.0002
13	54,439,729	54,539,728	13:54476595	5.13E-09	SLC17A9	3.07E-08
15	27,470,773	27,570,772	15:27543169	8.40E-11	APOA1	0.015
16	66,207,162	66,307,161	16:66279444	1.11E-12	IVNS1ABP	2.31E-17
17	51,725,017	51,825,016	17:51739903	1.22E-09	DNAH10	3.07E-08
17	53,925,017	54,025,016	17:53934607	2.89E-10	P2RX4	0.011
17	72,325,017	72,425,016	17:72377251	1.26E-10	LRRC74B	0.057
19	32,957,926	33,057,925	19:33056069	4.58E-14	TVP23B	0.072
19	41,857,926	41,957,925	19:41909633	2.90E-11	HAP1	0.003
19	42,157,926	42,257,925	19:42236007	1.53E-13	NKIRAS2, RAB5C, DHX58	0.0032, 0.035, 9.8e-08
19	42,257,926	42,357,925	19:42349652	9.38E-14	GHDC, STAT5B	0.035, 0.087
19	42,357,926	42,457,925	19:42358091	9.38E-14	STAT3	0.072
19	58,057,926	58,157,925	19:58114277	3.08E-10	C19H17orf80	0.087
19	60,557,926	60,657,925	19:60560812	1.23E-24	KCNJ2	3.04E-23
19	61,357,926	61,457,925	19:61408199	2.44E-09	ABCA10	7.62E-06
19	61,457,926	61,557,925	19:61512690	1.66E-09	ABCA9	0.072
27	36,520,984	36,620,983	27:36523102	2.84E-22	GPAT4	0.0007
29	9,447,532	9,547,531	29:9545883	4.03E-35	PICALM	2.63E-45

increased the window size, allowing mammary GSOR-identified genes to fall within 500 Kb of GWAS loci. A total of 22 significant terms were identified (Additional file 5), of which 15 shared with terms identified when using 100 Kb windows (Fig. 3).

Discussion

In our previous study [6], we showed that integrating GSOR, GWAS, and functional enrichment can help prioritize candidate genes for milk composition traits, though inference was limited by small sample sizes. Here, we leveraged a large, multi-breed reference population (> 81,000 cows) to train genomic prediction models, increasing both GEBV accuracy and the power to detect expression–trait associations. This enabled us to identify 20 genes likely to mediate lactose QTL, supported by convergence between GSOR, GWAS, and enrichment results. Because long-range LD, complex regulation, and small QTL effects complicate causal gene discovery, our framework, integrating local GEBV with tissue-relevant expression data, GWAS co-localization, and functional enrichment, may help overcome these challenges and offer a scalable, biologically grounded strategy to dissect the molecular basis of complex traits in livestock and other species.

When the most significant GWAS variant is located within a coding region, the causal gene may be directly implicated [6]. However, the majority of trait QTL reside in non-coding regions [6, 20, 22, 23], and thus, they have unknown functions. Non-coding QTL can impact phenotypes by regulating gene expression through *cis* or *trans* regulatory mechanisms. Even when causal variants are known, it is challenging to determine their functional impact and link them to specific target genes. This is because regulatory elements can act over long genomic distances and their effects can be highly cell-type specific [24]. TWASs were proposed to fill this gap by linking predicted tissue-specific gene expression to observed phenotypes [20, 21, 25]. However, one limitation of TWAS is that they usually rely on a limited set of individuals with assayed expression data and genotypes, which can limit the statistical power and decrease the accuracy of models

Table 3 Gene list enrichment analyses using mammary GSOR-identified genes co-localized with GWAS loci in 100 Kb windows

Source	Term ID	Term Name	FDR	Genes
GO: BP	GO:0001408	guanine nucleotide transport	0.001	LRRC8C, LRRC8B, SLC17A9
GO: BP	GO:1,903,790	guanine nucleotide transmembrane transport	0.001	LRRC8C, LRRC8B, SLC17A9
GO: BP	GO:0055085	transmembrane transport	0.001	KCNJ15, SLC50A1, LRRC8C, LRRC8B, SLC39A8, SLC34A2, SLC17A9, P2RX4, HAP1, KCNJ2, ABCA10, ABCA9
GO: BP	GO:0098739	import across plasma membrane	0.005	KCNJ15, LRRC8C, LRRC8B, SLC39A8, KCNJ2
GO: BP	GO:1,901,679	nucleotide transmembrane transport	0.007	LRRC8C, LRRC8B, SLC17A9
GO: BP	GO:1,901,264	carbohydrate derivative transport	0.009	SLC50A1, LRRC8C, LRRC8B, SLC17A9
GO: BP	GO:0015868	purine ribonucleotide transport	0.010	LRRC8C, LRRC8B, SLC17A9
GO: BP	GO:0051503	adenine nucleotide transport	0.011	LRRC8C, LRRC8B, SLC17A9
GO: BP	GO:0072530	purine-containing compound transmembrane transport	0.011	LRRC8C, LRRC8B, SLC17A9
GO: BP	GO:0015865	purine nucleotide transport	0.011	LRRC8C, LRRC8B, SLC17A9
GO: BP	GO:0006862	nucleotide transport	0.016	LRRC8C, LRRC8B, SLC17A9
GO: BP	GO:0034220	monoatomic ion transmembrane transport	0.016	KCNJ15, LRRC8C, LRRC8B, SLC39A8, SLC34A2, P2RX4, HAP1, KCNJ2
GO: BP	GO:0040014	regulation of multicellular organism growth	0.024	STAT5B, STAT3, GPAT4
GO: BP	GO:0042592	homeostatic process	0.042	GBA1, SLC39A8, SLC34A2, P2RX4, HAP1, NKIRAS2, STAT5B, STAT3, KCNJ2, PICALM
GO: BP	GO:0098659	inorganic cation import across plasma membrane	0.042	KCNJ15, SLC39A8, KCNJ2
GO: BP	GO:0099587	inorganic ion import across plasma membrane	0.042	KCNJ15, SLC39A8, KCNJ2
GO: BP	GO:0006811	monoatomic ion transport	0.042	KCNJ15, LRRC8C, LRRC8B, SLC39A8, SLC34A2, P2RX4, HAP1, KCNJ2
GO: BP	GO:0006810	transport	0.042	KCNJ15, SLC50A1, LRRC8C, LRRC8B, SLC39A8, SLC34A2, SLC17A9, P2RX4, HAP1, RAB5C, STAT5B, STAT3, KCNJ2, ABCA10, ABCA9, GPAT4, PICALM
GO: BP	GO:0048871	multicellular organismal-level homeostasis	0.044	GBA1, SLC39A8, P2RX4, NKIRAS2, STAT5B, STAT3, PICALM
REAC	REAC: R-HSA-382,551	Transport of small molecules	0.008	SLC50A1, LRRC8C, LRRC8B, SLC39A8, SLC34A2, ABCA10, ABCA9
REAC	REAC: R-HSA-186,797	Signaling by PDGF	0.008	THBS3, STAT5B, STAT3
REAC	REAC: R-HSA-425,407	SLC-mediated transmembrane transport	0.035	SLC50A1, SLC39A8, SLC34A2

1. GSOR-identified genes were converted to human orthologs before gene list enrichment analyses

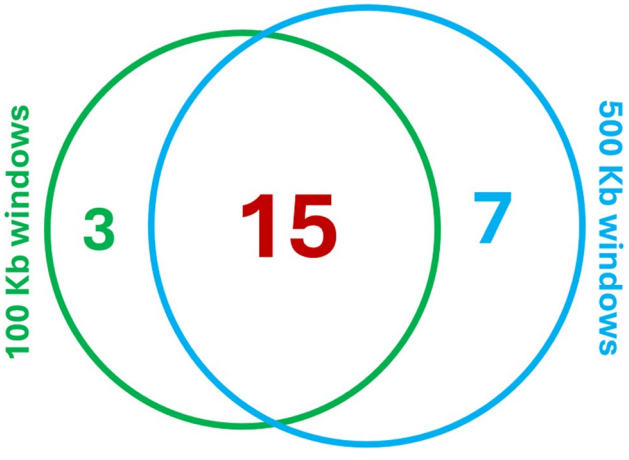

Fig. 3 Venn diagram showing the number of significant GO and KEGG terms identified using GSOR-GWAS gene lists derived from 100 kb versus 500 kb non-overlapping windows in mammary tissue

used for predicting gene expression [12]. In this study, we used GSOR to identify statistically significant gene-trait associations. Xiang et al. [12] found that GSOR is more powerful than TWAS because it uses the contribution to phenotype of the variants close to the gene whose expression is being investigated rather than the complete phenotype. However, the problem of spurious associations due to LD still exists [6]. Gene list enrichment analyses may help overcome LD issues by aggregating signals across functionally related genes, highlighting biologically coherent associations rather than single-gene signals confounded by LD.

As mentioned, GSOR does not provide direct evidence of causality, as correlations between gene expression and the trait can be induced by LD, among other factors. For instance, if two genes share *cis*-eQTLs in LD but only one eQTL is causal, both genes may appear statistically associated with the trait [6]. Therefore, an additional source

of evidence is required to prioritize candidate causal genes.

We evaluated significance of the overlap between GSOR-identified genes and GWAS loci by dividing the genome into non-overlapping windows of 100 Kb. We showed that GSOR-identified genes were significantly enriched within windows containing GWAS loci. This window based co-localization enrichment is consistent with our previous findings using a different milk composition trait [6], suggesting that the expression of these genes may mediate the effects of GWAS loci on complex traits through a *cis*-regulatory mechanism. However, LD can still confound these signals, so functional enrichment analyses are particularly important for highlighting the most relevant GWAS-GSOR genes. Milk LP is a promising target phenotype to test this hypothesis because its underlying biology is relatively well understood, with established roles for specific pathways such as lactose synthesis, ion transport, and hormonal signaling [10, 11, 13], allowing for meaningful interpretation of enrichment results. Of the 31 genes used in functional enrichment analyses, 55% share the term transport (GO:0006810 ; $P = 0.042$) and 39% share its descendant term "transmembrane transport" (GO:0055085; $P = 0.001$).

Lactose is the major osmotic solute in milk, responsible for drawing water into the alveolar lumen and driving milk volume [10, 26]. As a result, LP shows limited variation. However, it is not the only osmolarity regulator and an increase in other osmolytes can change LP. Indeed, studies show that increases in sodium, potassium or chloride in milk are inversely correlated with lactose concentration—highlighting that any process affecting ionic balance or secretion can shift lactose levels via osmotic compensation [27, 28]. We identified monoatomic ion transmembrane transport (GO0034220; FDR = 0.016) enriched with *KCNJ15*, *LRRC8C*, *LRRC8B*, *SLC39A8*, *SLC34A2*, *P2RX4*, *HAP1*, and *KCNJ2* genes. *KCNJ2* (Kir2.1) and *KCNJ15* (Kir4.2) both encode inwardly rectifying potassium channels (Kir) that favor K^+ influx under hyperpolarizing conditions [29]. Kamikawa and Ishikawa [30] identified Kir2.1-like channels — encoded by *KCNJ2* — functionally expressed in secretory mammary epithelial cells of lactating mice. These authors concluded that different types of K^+ channels might play a role in producing species-specific milk, while particular type of K^+ channels might generally express and participate in milk production in mammalian milk secretary cells [30]. An early study on ionic concentrations in milk found a significant association between different ions, such as K^+, and lactose levels in milk [31]. In cattle, a QTL on BTA19 including *KCNJ2* and *KCNJ16* has been associated with lactose concentration [13, 32], protein concentration [33] and milk yield [34]. Therefore, an eQTL that affects the abundance of *KCNJ2* or *KCNJ15* proteins (and K^+ ion transport) might lead to osmotic compensation influencing milk LP.

Maintaining cell volume is essential for mammary epithelial cells to remain functional during the osmotic shifts involved in milk secretion. Voltage-regulated anion channels (VRACs) is one mechanism by which they can do this. VRACs encoded by *LRRC8* proteins A to E, help regulate cell volume by exporting Cl^- ions and small organic anions [35, 36]. *P2RX4* is an ATP-gated cation channel permeable to Ca^{2+}; ATP-triggered *P2RX4* activation modulates VRAC in rat liver cells [37]. While not confirmed in mammary cells, this mechanism may apply during milk secretion. Together, these "ion transporters" help set the ionic gradients that draw water into the alveoli along with lactose.

Lactose, the primary sugar in milk, is synthesized in Golgi vesicles and secreted along with ions and water. In fact, lactose accumulation in secretory vesicles draws water into milk by osmosis – milk volume is almost perfectly correlated ($r \approx 0.99$) with lactose production [38]. This osmotic role of lactose means that many membrane transporters (for solutes and ions) must be active during lactation to supply substrates and maintain ion gradients. Consistent with this, a recent study found that genes involved in membrane transport were enriched among loci affecting milk lactose content [13]. Thus, our finding of enriched GO terms for "transmembrane transport" and "ion transmembrane transport" fits well with known lactation physiology, highlights that genes in these categories can modulate lactose synthesis and milk osmolarity [13].

Some genes may affect LP by directly affecting lactose production. *SLC50A1* (SWEET1) is a Golgi-localized sugar transporter that exports glucose (direct precursor for lactose synthesis) into the Golgi lumen [39]. Experimental studies support *SLC50A1* providing glucose for lactose production in mammary cells [39, 40]. Thus, variation in *SLC50A1* could directly affect lactose synthesis by altering sugar supply. More broadly, enrichment of "SLC-mediated transmembrane transport" (Reactome R-HSA-425407) in our data – which includes *SLC50A1*, *SLC39A8*, and *SLC34A2* – points to solute carrier proteins as key players. Other solute carriers in our gene list fit this picture. *SLC39A8* (ZIP8) is a zinc/manganese importer; it moves Mn^{2+} and Zn^{2+} into cells [41]. Manganese is an essential cofactor for many glycosyltransferases, including the β4-galactosyltransferase I subunit of the lactose synthase, and is required for the function of various Golgi enzymes [42]. *SLC34A2* (also known as *NPT2b*) is a sodium-dependent phosphate transporter highly expressed in lactating mammary epithelium [43]. While its precise role in milk synthesis was not directly tested, its known function in phosphate uptake suggests a potential role in supporting ATP and nucleotide sugar

synthesis during lactose production. These transporters illustrate how the enriched GO terms reflect lactation biology: they supply essential substrates (glucose, phosphate) and cofactors (Mn, Zn, Ca) for lactose synthesis, while maintaining ion gradients and water balance critical for milk secretion.

The Reactome pathway "Signaling by PDGF (Platelet-derived growth factor)" involves *STAT5B*, *STAT3*, and *THBS3*. *STAT5* (both *STAT5A* and *STAT5B*) is a key transcription factor for lactation: activated by prolactin via Janus kinases (JAK2), it drives the expression of milk proteins and enzymes, and epithelial cell differentiation during pregnancy [44]. Indeed, genetic knockout of *STAT5A* in mice causes failure of alveolar differentiation and lactogenesis [44] and experiments note *STAT5* as "a primary transcription factor for milk production" [45]. PDGF promotes stromal/epithelial proliferation and survival. Thus, enrichment of this pathway suggests that growth factor-STAT signaling networks influence mammary development, which potentially impact milk LP phenotype. For example, greater *STAT5* signaling would boost lactose synthesis (via more alveolar cells and lactose enzymes), whereas *STAT3* activity would reduce lactose output. In sum, the "Signaling by PDGF" term likely flags a set of regulatory genes (STAT5B/STAT3) that govern cell differentiation, survival, and secretory activity in the udder. Changes in these signals would alter mammary function (and indirectly lactose percentage) even though they are not lactose-specific per se.

In conclusion, we demonstrated the utility of integrating multi-omics data with functional enrichment to identify genes likely regulated by QTL associated with milk LP. We used LP as a test case to assess whether combining gene expression, GWAS, and functional enrichment could pinpoint causal genes for complex traits. To mitigate the confounding effects of LD, we leveraged functional enrichment analyses, which highlighted biologically coherent associations. Our results show that this strategy successfully identified 20 genes with clear mechanistic links to LP, acting primarily through indirect regulatory pathways. Future experimental validations will be needed to confirm these findings.

Methods

Overview

This study used four independent datasets:

(1) ~ 400 New Zealand (NZ) cows with mammary RNA-seq data [13–15]. This data is used for gene-based associations test, specifically genetic score omics regression (GSOR); introduced by Xiang et al. [12].

(2) ~ 400 Australian (AU) cows with white blood cells (WBC) RNA-seq data [16–18]. This data is also used for gene-based associations test (GSOR).

(3) ~ 81,000 AU multibreed cows with lactose percentage (LP) phenotypes and SNP genotypes (referred to as GSOR reference population). This data is used to estimate local GEBVs for LP in the RNA-sequenced cows of NZ and AU.

(4) GWAS summary statistics for milk LP, based on 12,000 NZ cows and reported by Lopdell et al. [13] were used to test for co-localization between significant GSOR genes and GWAS loci.

Phenotypic data for milk lactose percentage

Phenotypic data for milk test-day LP was provided by DataGene (an independent industry-owned organisation that provides the national genetic evaluations for dairy cattle in Australia). Animals born between 2012 and 2024 that had Holstein, Jersey or Australian Red breed codes were retained. Outliers deviating ± 3 SD of the mean phenotypic value of LP were excluded. Test-day records were included if a cow's age at calving was between 18 and 25 months and days in milk (DIM) between 5 and 315 days. Additionally, herds and test dates with less than five observations were excluded from the analysis. The final data contained 4,995,316 test-day records belonging to 477,822 cows. ASReml [46] was used to adjust phenotypes for fixed effects and average them for each cow (i.e. effect of Cow) following the model proposed by [47].

$$\begin{aligned} y_{ijklm} = \mu + H_i TD_j + M_k + pol(DIM, 8) \\ + pol(Age, 2) + Cow_l + e_{ijklm} \end{aligned}$$

where, y_{ijklm} is the test-day record for LP ($N = 4{,}995{,}316$), μ is the effect of overall mean, $H_i TD_j$ is the effect of the i^{th} herd and j^{th} test-day ($N = 82{,}058$); M_k is the effect of the k^{th} calving month ($N = 12$); $pol(DIM, 8)$ and $pol(Age, 2)$ are the regression coefficients of Legendre polynomials of order 1–8 for DIM and of order 1–2 for age at calving in months; Cow_l and e_{ijklm} are the random effects of the l^{th} cow ($N = 477{,}822$) and the random residual term, respectively. The following sections will explain our integrative multi-omics approach in detail. A schematic overview of the method is illustrated in Fig. 4.

Genotypic data for the GSOR reference population

Of the cows with phenotypic data described above, SNP panel genotypes were available for 81,658 cows, comprising 79% Holstein, 16% Jersey and 5% Australian Red. All SNP markers were mapped to the ARS-UCD1.2 reference genome [48], and included autosomal markers as well as those on the non-pseudo autosomal region of the X chromosome. Any raw genotypes with a GenCall score of < 0.6 were set to missing and any marker or animal

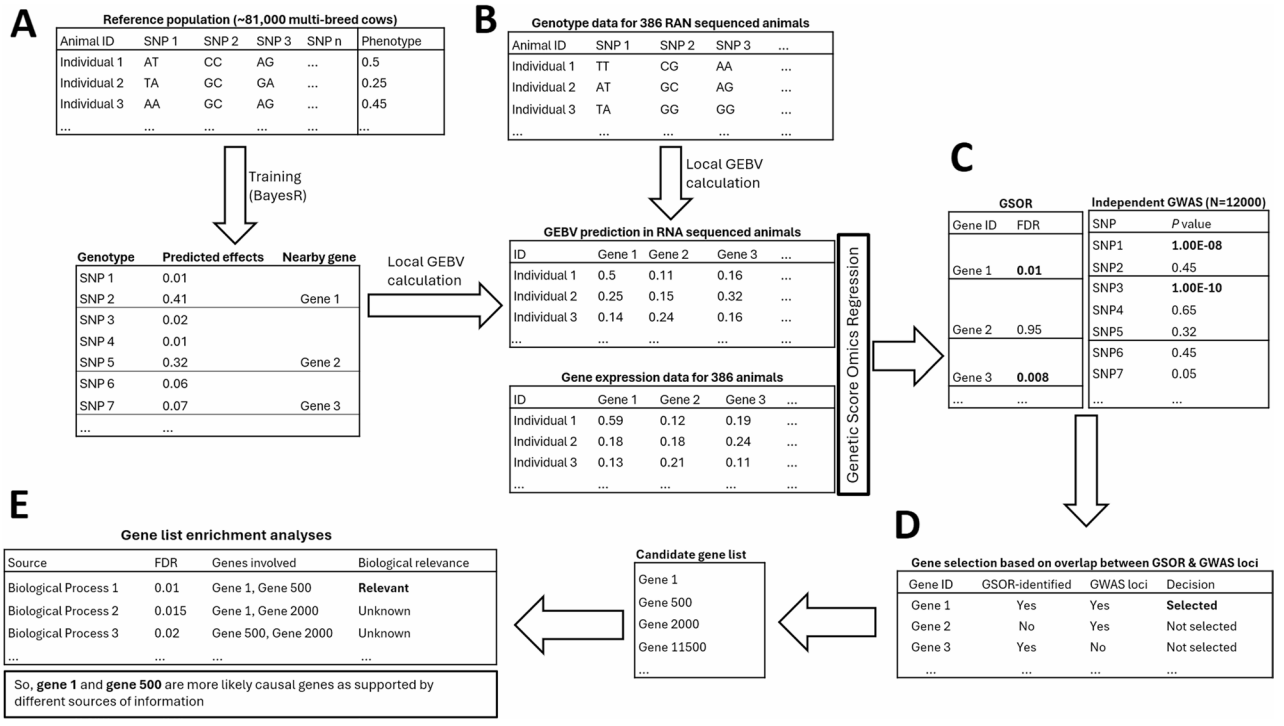

Fig. 4 A schematic overview of the method used in this study. **A** Reference population including animals with both SNP genotype and phenotype of LP were used to predict SNP effects. **B** Predicted SNP effects were used to calculate local GEBV in RNA-sequenced animals, local GEBV and gene expression data were combined in GSOR analyses. **C** GSOR results and an independent GWAS were combined to investigate significance of co-localization between GSOR-identified genes and GWAS loci in 100 Kb windows. **D** Genes were selected if they were both significant in GSOR and co-localized with GWAS loci in 100 Kb windows and (**E**) used in gene list enrichment analyses

with 10% or more missing genotypes was discarded. The remaining sporadic missing genotypes were imputed with FImpute v.3 software [49]. The genotypes were from a range of SNP panels (≥ 6,000 markers) and were imputed with FImpute v.3 to a custom 74 K SNP genotype panel that is used by DataGene for national genetic evaluations [50]. The imputation reference population for the 74 K SNP panel included over 28,000 animals (Holstein, Jersey and Australian Red breeds). Next, the 74 K SNP genotypes were imputed to the Illumina High Density (HD) Bovine SNP panel that included 714,451 SNP in an imputation reference population of 2,910 animals (breeds as for the 74 K panel). Prior to HD imputation, approximately 20,000 SNP in the custom 74 K set that did not overlap the HD set were removed and then added back in before the final imputation to whole genome sequences (WGS). The sequenced imputation reference population included 5,036 *Bos taurus* cattle from Run9 of the 1000 Bull Genomes project [51]. Following Nguyen et al. [52], sequence variants were pre-filtered (49,114,602 variants remaining) and phased with Eagle v2 [53] before using Beagle v5.2.1 [54] to impute all animals to WGS.

Post-imputation, sequence variants with a Beagle DR2 (estimated imputation accuracy) < 0.9 were excluded, as well as those with minor allele frequency (MAF) < 0.01 and genotype frequencies deviating from Hardy-Weinberg equilibrium ($P \leq 1 \times 10^{-8}$). LD pruning was performed using PLINK v1.9 [55] with parameters --indep-pairwise 5000 500 0.95 to exclude variants that were in high LD (r2 > 0.95). These procedures retained 1,181,628 variants for subsequent analyses. We used this data to train a BayesR model [56] using BayesR3 software [57] to estimate prediction equations (SNP effects) for LP. The model was as follows: $\mathbf{y} = \mathbf{Xu} + \mathbf{Vg} + \mathbf{e}$, where $\mathbf{y}$ is an $n \times 1$ vector of phenotypic records, in which n is the number of animals in the reference population (N=81,658); $\mathbf{X}$ is an $n \times m$ incidence matrix, $\mathbf{u}$ is $m \times 1$ vector of fixed effects and m corresponds to fixed effects including breed effect with three levels; $\mathbf{V}$ is the coded genotype, representing the observed genotypes of each individual; $\mathbf{g}$ is a vector of SNP effects; and $\mathbf{e}$ is the residual term. BayesR3 was run with 50,000 MCMC iterations and 25,000 burn-in. In the BayesR3 model, the SNP effects follow a mixture of four normal distributions with zero mean and additive genetic variances of zero, 0.0001, 0.001, and 0.01 times the genetic variance. Starting values for proportions of the four SNP effect distributions were defined as 0.994, 0.0055, 0.00049, and 0.00001, respectively. Prediction equations were applied to the RNA-sequenced animals to calculate their local GEBVs for LP (described below).

RNA-seq data and gene-based associations test (GSOR)

We analysed two distinct sets of RNA sequencing data from WBC and mammary tissue.

The WBC gene expression data were obtained from 313 lactating cows of multiple breeds from the Agriculture Victoria research farm (Ellinbank Smart Farm). Details of sample processing, RNA extraction, library preparation, and sequencing are provided in [16–18]. For the WBC RNA-seq animals, genotypes were imputed to WGS using Run9 of the 1000 Bull Genomes project [51] as described above for the GSOR reference population.

The mammary gene expression data include 386 lactating NZ cows, including Holstein, Jersey and their crosses. The cows in this dataset, were previously imputed to WGS using 1,298 imputation reference animals, including 306 Holstein-Friesian, 219 Jersey, 717 crossbreds (Holstein-Friesian x Jersey) and 56 other breeds as described in [58]. For the WBC and mammary RNA-sequenced animals, we retained the same 1,181,628 variants that were retained in the GSOR reference population.

Prior to fitting per-gene GSOR models, we calculated local GEBVs for RNA-sequenced animals using the estimated effects for variants located within a ± 1 Mb window centred on the transcription start site of the gene being tested. These local GEBVs served as response variables in the GSOR analysis, in which gene expression levels were tested as predictors to identify genes whose expression is associated with genetically driven, *cis*-regulatory variation in the trait (Fig. 4). The following per-gene GSOR model was applied [6]:

$$\widehat{\mathbf{gebv}}_{\mathbf{local}} = \boldsymbol{\lambda}\, \mathrm{b}_1 + \mathbf{X}\mathbf{b}_2 + \mathbf{g} + \mathbf{e}$$

where $\widehat{\mathbf{gebv}}_{\mathbf{local}}$ is an $\mathrm{m} \times 1$ vector of local GEBVs predicted (in the RNA-sequenced cows) using the SNP effects from a ± 1 Mb window around the gene being tested; $\boldsymbol{\lambda}$ is a $\mathrm{m} \times 1$ vector of tissue-specific expression of the gene across the corresponding RNA-sequenced cows; b_1 is the regression coefficient of the $\widehat{\mathbf{gebv}}_{\mathbf{local}}$ on $\boldsymbol{\lambda}$; $\mathbf{X}$ represents a design matrix for fixed effects (see next paragraph), and $\mathbf{b}_2$ is the vector of fixed effects for the corresponding RNA-sequenced animals; $\mathbf{g}$ is a vector of random polygenic effects across the same RNA-sequenced cows, assumed to follow a normal distribution $\mathbf{g} \sim \mathrm{N}(0, \mathbf{G}\sigma^2_{\mathrm{g}})$, where $\mathbf{G}$ is the genomic relationship matrix [59], and σ^2_{g} is the additive genetic variance explained by the whole genome SNPs; $\mathbf{e}$ is the vector of residuals, assumed to follow a normal distribution $\mathbf{e} \sim \mathrm{N}\left(0, \mathbf{I}\sigma^2_{\mathrm{e}}\right)$, where $\mathbf{I}$ is an identity matrix, and σ^2_{e} is residual variance.

Models for the WBC RNA-seq dataset incorporated the experiment (with five levels corresponding to sampling times), breed (with three levels: Holstein, Jersey, and Australian Red), and herd (by four levels, with Ellinbank SmartFarm representing the majority of samples) as a categorical fixed effect, while days in milk (DIM) was used as a quantitative fixed effect, with a mean and SD of 86 (± 36) days. For the mammary RNA-seq dataset, no fixed effects were necessary [6]; therefore, the previous model was reduced as follows:

$$\widehat{\mathbf{gebv}}_{\mathbf{local}} = \boldsymbol{\lambda}\, \mathrm{b} + \mathbf{g} + \mathbf{e}$$

After applying the per-gene GSOR model to all genes within each RNA-seq dataset, the *P* values were corrected to address the multiple testing issue. Within each dataset, genes with a $\mathrm{FDR}_{\mathrm{GSOR}} \leq 0.1$ were regarded significant (GSOR-identified genes).

GWAS summary statistics

The original SNP coordinates were based on the UMD3.1 bovine reference genome. To ensure consistency with our datasets, we converted these coordinates to the ARS-UCD1.2 [48] reference genome using the UCSC LiftOver tool [60] (https://genome.ucsc.edu/cgi-bin/hgLiftOver), leaving 1,088,337 SNPs for downstream analyses. To account for multiple testing, we regarded SNPs with $P \leq 1 \times 10^{-8}$ as genome-wide significant (GWAS loci).

Window-based co-localization between GSOR-identified genes and GWAS loci

We partitioned the genome into non-overlapping windows of 100 kb and 500 kb in separate analyses, to test whether GSOR-identified genes and GWAS loci co-occur in the same genomic regions more often than expected by chance. For each window size, we classified windows based on the presence or absence of GSOR-identified genes and GWAS loci into four categories, including windows containing: (1) both GSOR-identified genes and GWAS loci; (2) only GSOR-identified gene(s); (3) only GWAS loci; and (4) neither. A GSOR-identified gene was assigned to a window if its transcription start site fell within that window. We then applied Fisher's exact test to assess the statistical significance of the co-localization, considering $P_{\mathrm{Fisher}} \leq 0.05$ as significant.

Gene list enrichment analysis

Candidate genes were defined as GSOR-identified gene(s) located within non-overlapping genomic windows that contained at least one GWAS locus. We used R (v4.4.3) package gprofiler2 (v0.2.3) [19] to convert bovine genes to their human (*Homo sapiens*) orthologs and to perform gene list enrichment analyses. We examined overrepresented Gene Ontology (GO) Biological Process (GO:BP) terms and Reactome pathways among the candidate genes, and terms with $\mathrm{FDR}_{\mathrm{term}} \leq 0.05$ were regarded

significant. All the genes in the corresponding RNA-seq data were used as background genes after being converted to their human (*Homo sapiens*) orthologs.

We hypothesized that the strongest candidate causal genes are those that (i) show tissue-specific expression correlated with local GEBVs (GSOR-identified genes), (ii) are enriched in window-based co-localization with GWAS loci, and (iii) are enriched for trait-relevant physiological functions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12525-0.

Additional file 1: Summary statistics from GSOR analyses of genes expressed in the mammary gland. Legend: Includes effect sizes, standard errors, p-values, and FDR values for all genes expressed in the mammary gland. The last column indicates whether each gene overlaps a QTL identified by GWAS, with 0 and 1 denoting no overlap and overlap, respectively.

Additional file 2: Summary statistics from GSOR analyses of genes expressed in the white blood cells (WBC). Legend: Includes effect sizes, standard errors, p-values, and FDR values for all genes expressed in the WBC. The last column indicates whether each gene overlaps a QTL identified by GWAS, with 0 and 1 denoting no overlap and overlap, respectively.

Additional file 3: Human orthologs of bovine genes expressed in the mammary gland and gene list used for enrichment analyses. Legend: The first sheet lists all bovine genes that expressed in the mammary gland and were successfully converted to human orthologs; these genes used as background in the gene enrichment analyses. The second sheet includes the genes used as input in the gene list enrichment analyses.

Additional file 4: Enriched Gene Ontology (GO) terms among candidate genes identified in the white blood cells (WBC) RNA-seq data. Legend: These candidate genes were both significant in the GSOR analysis and co-localized with GWAS loci within ±100 kb windows.

Additional file 5: Enriched Gene Ontology (GO) terms among candidate genes identified in the mammary gland. Legend: These candidate genes were both significant in the GSOR analysis and co-localized with GWAS loci within ±500 kb windows.

Authors' contributions

Conceptualization: M. Goddard, M. Ghoreishifar. Methodology & Formal Analyses: M. Ghoreishifar. Writing Original Draft & Visualization: M. Ghoreishifar. Writing Reviewing Editing: M. Ghoreishifar, I. Macleod, T. Nguyen, T. Lopdell, M. Littlejohn, R. Xiang, A. Chamberlain, J. Pryce, M. Goddard. Supervision: M. Goddard, J. Pryce, A. Chamberlain. Funding Acquisition: J. Pryce.

Funding

This study was undertaken as part of the DairyBio program, which is jointly funded by Dairy Australia (Melbourne, Australia), Agriculture Victoria (Melbourne, Australia), and The Gardiner Foundation (Melbourne, Australia). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

DataGene Australia (http://www.datagene.com.au/) is the custodian of the raw phenotype and genotype data of Australian farm animals. Access to these data for research requires permission from DataGene under a Data Use Agreement. Other supporting data are shown in the Supplementary Materials of the manuscript. Code and tutorials for GSOR are available at https://github.com/rxiangr/GSOR-and-MTAO. All gene expression data was taken from previously published studies as detailed in the Methods section.

Declarations

Ethics approval and consent to participate

All animal experiments were conducted in strict accordance with the rules and guidelines outlined in the New Zealand Animal Welfare Act 1999. Most data were generated as part of a mammary tissue biopsy experiment, with all samples obtained in accordance with protocols approved by the Ruakura Animal Ethics Committee, Hamilton, New Zealand (approval AEC 12845). No animals were sacrificed for this study. The study is reported in accordance with ARRIVE guidelines.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 23 September 2025 / Accepted: 5 January 2026
Published online: 14 January 2026

References

1. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19(8):491–504.
2. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common Disease-Associated variation in regulatory DNA. Science. 2012;337(6099):1190–5.
3. Qi T, Song L, Guo Y, Chen C, Yang J. From genetic associations to genes: methods, applications, and challenges. Trends Genet. 2024;40(8):642–67.
4. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome–phenome relationship using phenome-wide association studies. Nat Rev Genet. 2016;17(3):129–45.
5. van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of polygenic disease traits: from theory to practice. Nat Rev Genet. 2019;20(10):567–81.
6. Ghoreishifar M, Macleod IM, Chamberlain AJ, Liu Z, Lopdell TJ, Littlejohn MD, Xiang R, Pryce JE, Goddard ME. An integrative approach to prioritize candidate causal genes for complex traits in cattle. PLoS Genet. 2025;21(5):e1011492.
7. Li Z, Zhou X. Towards improved fine-mapping of candidate causal variants. Nat Rev Genet. 2025;26:847–61.
8. Brodie A, Azaria JR, Ofran Y. How Far from the SNP May the causative genes be? Nucleic Acids Res. 2016;44(13):6046–54.
9. Consortium TGO. The gene ontology resource: enriching a gold mine. Nucleic Acids Res. 2020;49(D1):D325–34.
10. Sadovnikova A, Garcia SC, Hovey RC. A comparative review of the cell Biology, Biochemistry, and genetics of lactose synthesis. J Mammary Gland Biol Neoplasia. 2021;26(2):181–96.
11. Strucken EM, Laurenson YC, Brockmann GA. Go with the flow-biology and genetics of the lactation cycle. Front Genet. 2015;6:118.
12. Xiang R, Fang L, Liu S, Liu GE, Tenesa A, Gao Y, Mason BA, Chamberlain AJ, Goddard ME, Consortium C. Genetic score omics regression and multi-trait meta-analysis detect widespread cis-regulatory effects shaping bovine complex traits. PNAS Nexus. 2025;4:pgaf208.
13. Lopdell TJ, Tiplady K, Struchalin M, Johnson TJJ, Keehan M, Sherlock R, Couldrey C, Davis SR, Snell RG, Spelman RJ, et al. DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. BMC Genomics. 2017;18(1):968.
14. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, Couldrey C, Keehan M, Sherlock RG, Harland C, et al. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. Sci Rep. 2016;6(1):25376.
15. Prowse-Wilkins CP, Lopdell TJ, Xiang R, Vander Jagt CJ, Littlejohn MD, Chamberlain AJ, Goddard ME. Genetic variation in histone modifications and gene expression identifies regulatory variants in the mammary gland of cattle. BMC Genomics. 2022;23(1):815.

16. Chamberlain A, Hayes B, Xiang R, Vander Jagt C, Reich C, Macleod I, Prowse-Wilkins C, Mason B, Daetwyler H, Goddard M. Identification of regulatory variation in dairy cattle with RNA sequence data. In: Proceedings of the 11th World Congress on Genetics Applied to Livestock Production: 2018;2018:11–16.
17. Xiang R, Fang L, Liu S, Macleod IM, Liu Z, Breen EJ, Gao Y, Liu GE, Tenesa A, Mason BA et al. Gene expression and RNA splicing explain large proportions of the heritability for complex traits in cattle. Cell Genomics. 2023;3:100385.
18. Xiang R, Hayes BJ, Vander Jagt CJ, MacLeod IM, Khansefid M, Bowman PJ, Yuan Z, Prowse-Wilkins CP, Reich CM, Mason BA, et al. Genome variants associated with RNA splicing variations in bovine are extensively shared between tissues. BMC Genomics. 2018;19(1):521.
19. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. gprofiler2–an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler. F1000Research. 2020;9:ELIXIR-709.
20. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51(4):592–9.
21. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–8.
22. Ghoreishifar M, Chamberlain AJ, Xiang R, Prowse-Wilkins CP, Lopdell TJ, Littlejohn MD, Pryce JE, Goddard ME. Allele-specific binding variants causing ChIP-seq peak height of histone modification are not enriched in expression QTL annotations. Genet Selection Evol. 2024;56(1):50.
23. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol. 2012;30(11):1095–106.
24. Spitz F. Gene regulation at a distance: from remote enhancers to 3D regulatory ensembles. Semin Cell Dev Biol. 2016;57:57–67.
25. Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am J Hum Genet. 2017;100(3):473–87.
26. Costa A, Egger-Danner C, Mészáros G, Fuerst C, Penasa M, Sölkner J, Fuerst-Waltl B. Genetic associations of lactose and its ratios to other milk solids with health traits in Austrian Fleckvieh cows. J Dairy Sci. 2019;102(5):4238–48.
27. Holt C. Interrelationships of the concentrations of some ionic constituents of human milk and comparison with cow and goat milks. Comp Biochem Physiol Part A: Physiol. 1993;104(1):35–41.
28. Wack RP, Lien EL, Taft D, Roscelli JD. Electrolyte composition of human breast milk beyond the early postpartum period. Nutrition. 1997;13(9):774–7.
29. Kubo Y, Adelman JP, Clapham DE, Jan LY, Karschin A, Kurachi Y, Lazdunski M, Nichols CG, Seino S, Vandenberg CA. International union of Pharmacology. LIV. Nomenclature and molecular relationships of inwardly rectifying potassium channels. Pharmacol Rev. 2005;57(4):509–26.
30. Kamikawa A, Ishikawa T. Functional expression of a Kir2. 1-like inwardly rectifying potassium channel in mouse mammary secretory cells. Am J Physiology-Cell Physiol. 2014;306(3):C230–40.
31. Barry J, Rowland S. Variations in the ionic and lactose concentrations of milk. Biochem J. 1953;54(4):575.
32. Tiplady KM, Lopdell TJ, Reynolds E, Sherlock RG, Keehan M, Johnson TJJ, Pryce JE, Davis SR, Spelman RJ, Harris BL, et al. Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. Genet Selection Evol. 2021;53(1):62.
33. Pedrosa VB, Schenkel FS, Chen S-Y, Oliveira HR, Casey TM, Melka MG, Brito LF. Genomewide association analyses of lactation persistency and milk production traits in Holstein cattle based on imputed Whole-Genome sequence data. Genes. 2021;12(11):1830.
34. Pausch H, Emmerling R, Gredler-Grandl B, Fries R, Daetwyler HD, Goddard ME. Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentages in milk at nucleotide resolution. BMC Genomics. 2017;18(1):853.
35. Voss FK, Ullrich F, Münch J, Lazarow K, Lutter D, Mah N, Andrade-Navarro MA, von Kries JP, Stauber T, Jentsch TJ. Identification of LRRC8 heteromers as an essential component of the Volume-Regulated anion channel VRAC. Science. 2014;344(6184):634–8.
36. Syeda R, Qiu Z, Dubin AE, Murthy SE, Florendo MN, Mason DE, Mathur J, Cahalan SM, Peters EC, Montal M, et al. LRRC8 proteins form Volume-Regulated anion channels that sense ionic strength. Cell. 2016;164(3):499–511.
37. Varela D, Penna A, Simon F, Eguiguren AL, Leiva-Salcedo E, Cerda O, Sala F, Stutzin A. P2X4 Activation Modulates Volume-sensitive Outwardly Rectifying Chloride Channels in Rat Hepatoma Cells *. J Biol Chem. 2010;285(10):7566–74.
38. Sneddon N, Lopez-Villalobos N, Davis S, Hickson R, Shalloo L. Genetic parameters for milk components including lactose from test day records in the new Zealand dairy herd. New Z J Agricultural Res. 2015;58(2):97–107.
39. Wang G, Jin W, Zhang L, Dong M, Zhang X, Zhou Z, Wang X. SLC50A1 inhibits the doxorubicin sensitivity in hepatocellular carcinoma cells through regulating the tumor Glycolysis. Cell Death Discovery. 2024;10(1):495.
40. Chen L-Q, Hou B-H, Lalonde S, Takanaga H, Hartung ML, Qu X-Q, Guo W-J, Kim J-G, Underwood W, Chaudhuri B, et al. Sugar transporters for intercellular exchange and nutrition of pathogens. Nature. 2010;468(7323):527–32.
41. Nebert DW, Liu Z. SLC39A8 gene encoding a metal ion transporter: discovery and bench to bedside. Hum Genomics. 2019;13(Suppl 1):51.
42. Guo M. Chemical composition of human milk. In: Human Milk Biochemistry and Infant Formula Manufacturing Technology. Edited by Guo M: Woodhead Publishing; 2014:327–62.
43. Wang X, Zhang B, Dong W, Zhao Y, Zhao X, Zhang Y, Zhang Q. SLC34A2 targets in Calcium/Phosphorus homeostasis of mammary gland and involvement in development of clinical mastitis in dairy cows. Animals. 2024;14:1275.
44. Liu X, Robinson GW, Wagner KU, Garrett L, Wynshaw-Boris A, Hennighausen L. Stat5a is mandatory for adult mammary gland development and lactogenesis. Genes Dev. 1997;11(2):179–86.
45. Kobayashi K, Wakasa H, Han L, Koyama T, Tsugami Y, Nishimura T. Lactose on the basolateral side of mammary epithelial cells inhibits milk production concomitantly with signal transducer and activator of transcription 5 inactivation. Cell Tissue Res. 2022;389(3):501–15.
46. Gilmour A, Gogel B, Cullis B, Welham S, R T. ASReml user guide release 4.2 structural specification. Hemel Hempstead, UK. In.,: VSN International; 2022.
47. Khansefid M, Pryce JE, Shahinfar S, Axford M, Goddard ME, Haile-Mariam M. Improving accuracy and stability of genetic predictions for dairy cow survival. Anim Prod Sci. 2023;63(11):1031–42.
48. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C et al. De Novo assembly of the cattle reference genome with single-molecule sequencing. GigaScience. 2020;9:giaa021.
49. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15(1):478.
50. van den Berg I, Nguyen TV, Nguyen TTT, Pryce JE, Nieuwhof GJ, MacLeod IM. Imputation accuracy and carrier frequency of deleterious recessive defects in Australian dairy cattle. J Dairy Sci. 2024;107(11):9591–601.
51. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al. Whole-genome sequencing of 234 bulls facilitates mapping of Monogenic and complex traits in cattle. Nat Genet. 2014;46(8):858–65.
52. Nguyen TV, Bolormaa S, Reich CM, Chamberlain AJ, Vander Jagt CJ, Daetwyler HD, MacLeod IM. Empirical versus estimated accuracy of imputation: optimising filtering thresholds for sequence imputation. Genet Selection Evol. 2024;56(1):72.
53. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR HKF, Schoenherr S, Forer L, McCarthy S, Abecasis GR, et al. Reference-based phasing using the haplotype reference consortium panel. Nat Genet. 2016;48(11):1443–8.
54. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84(2):210–23.
55. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
56. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95(7):4114–29.

57. Breen EJ, MacLeod IM, Ho PN, Haile-Mariam M, Pryce JE, Thomas CD, Daetwyler HD, Goddard ME. BayesR3 enables fast MCMC blocked processing for largescale multi-trait genomic prediction and QTN mapping analysis. Commun Biology. 2022;5(1):661.
58. Trebes H, Wang Y, Reynolds E, Tiplady K, Harland C, Lopdell T, Johnson T, Davis S, Harris B, Spelman R, et al. Identification of candidate novel production variants on the Bos Taurus chromosome X. J Dairy Sci. 2023;106(11):7799–815.
59. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565–9.
60. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. The UCSC genome browser database: update 2006. Nucleic Acids Res. 2006;34(suppl1):D590–8.

Publisher's note