Genome variants altering gene splicing in bovine are extensively shared between tissues

Ruidong Xiang[1,2,*], Ben J. Hayes[2,3], Christy J. Vander Jagt[2], Iona M. MacLeod[2], Majid Khansefid[2], Phil J. Bowman[2,6], Claire P. Prowse-Wilkins[2], Coralie M. Reich[2], Brett A. Mason[2], Josie B. Garner[4], Leah C. Marett[4], Yizhou Chen[5], Sunduimijid Bolormaa[2], Hans D. Daetwyler[2,6], Amanda J. Chamberlain[2] and Michael E. Goddard[1,2]

[1]Faculty of Veterinary & Agricultural Science, University of Melbourne, Parkville, VIC 3010, Australia. [2]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC 3083, Australia. [3]Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, University of Queensland, St. Lucia, QLD 4067, Australia. [4]Agriculture Victoria, Dairy Production Science, Ellinbank, VIC, 3821, Australia. [5]Elizabeth Macarthur Agricultural Institute, New South Wales Department of Primary Industries, Camden, NSW 2570, Australia. [6]School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia.

*Corresponding author: Dr. Ruidong Xiang;

Email: ruidong.xiang@unimelb.edu.au

Abstract

Background: Mammalian phenotypes are shaped by numerous genome variants, many of which may regulate gene transcription or exon splicing. We aim to identify splicing quantitative trait nucleotides (sQTN) in bovine, an important economic and model species, then compare sQTN with well-reported gene expression geQTN and lesser-reported exon expression eeQTN in different tissues.

Results: Using whole genome and RNA sequence data from four tissues of over 200 cattle, sQTN identified using exon inclusion ratios were verified by matching their effects on adjacent intron excision ratios. sQTN had the highest percentage of intronic variants and the lowest percentage of intergenic variants, compared to eeQTN and geQTN. For sQTN and eeQTN, there were more cases where the same variant was detected (FDR<0.01) in multiple tissues, compared to geQTN. To verify such expression QTN sharing between tissues, variants surrounding (±1Mb) the exon or gene were used to build local genomic relationship matrices and estimated genetic correlations between tissues. For many exons, the splicing and expression level was determined by the same *cis* additive genetic variance in different tissues. Thus, an effective meta-analysis combining information from three tissues increased power to detect and validate sQTN. sQTN and eeQTN together were more enriched for variants associated with cattle complex traits, compared to geQTN. Several potentially causal mutations were identified, including an sQTN at Chr6:87392580 within the 5th exon of kappa casein (*CSN3*) affecting milk production traits.

Conclusions: We have substantially improved the knowledge of genome regulation for an important mammalian species and provided novel insights into the analytics of expression QTN.

Keywords: RNA splicing, gene expression, eQTL, sQTL, genetic correlations, meta-analysis, bovine

Background

Cattle are an important source of meat and dairy products for humans worldwide. Also, cattle can be used as clinical models to study genetic causes of human diseases[1]. To improve productivity, health performance and efficiency of cattle, traditional selective breeding has been widely used. In the last decade, the genomic selection technique, originally developed in cattle breeding, has further increased the rate of genetic improvement of complex traits of all livestock species [2, 3]. However, genomic selection commonly uses random single nucleotide polymorphisms (SNPs) and, with a few exceptions, we do not know the genes or the sites causing genetic variation in important traits. Knowledge of the genes involved and polymorphic sites would increase our understanding of the biology and may further increase the rate of genetic improvement [4].

Many of the sequence variants that affect complex traits (quantitative trait nucleotides or QTN) are not coding variants and are presumed to influence the regulation of gene expression, that is to be expression QTN (eQTN) [5]. A regulatory mutation might affect expression of the gene in general, which commonly refers to a gene eQTN or geQTN. In humans, geQTN show significant enrichments for mutations associated with diseases and complex traits [5, 6].

A regulatory mutation might change the expression of one particular or several exons within a gene, which can be defined as an exon eQTN or eeQTN. Furthermore, after transcription, RNAs are spliced by intron removal and exon ligation to create various mature transcripts. A mutation altering the RNA splicing by changing the balance between various transcripts from the gene, is then defined as a splicing QTN or sQTN. If a mutation changes the expression ratio of an exon to the gene, it implies that it alters the splicing pattern. In humans, sQTN were studied by inferring the individual splicing ratio from RNA sequence data[7]. More

recently, sQTN identified using intron information extracted from RNA sequence data were demonstrated to have fundamental links with human diseases [8, 9].

Despite vastly reported human regulatory QTN information, the knowledge of large animal regulatory mutations is minimal. In this study, we aim to map bovine cis splicing QTN using counts of genes, exons and introns of RNA sequence data from hundreds of animals and multiple tissues along with imputed whole genome sequences. We examined the extent to which sQTN affected splicing in white blood cells, milk cells, liver and muscle transcriptomes and the extent to which sQTN are enriched for variants affecting complex traits. We also used the counts of genes and exons to map cis eeQTN and geQTN, then analysed their relationships with sQTN in different tissues.

Results

In total, we analysed 378 transcriptomes of 19 tissue types from 214 cattle generated from four experiments covering major dairy and beef cattle breeds (Table 1, Figure 1a and Additional file 1: Supplementary Methods). Overall, samples from the same or similar tissues clustered together rather than clustering by experiments, based on exon expression levels (Figure 1a). Consistent with previous reports [10], milk cells and mammary gland transcriptomes were closely related.

Table 1. Summary of experiments, data and analyses.

Tissue splicing[a]	Breed splicing[b]	sQTN[c]	Experiment	Tissue type	Tissue No.	Breed	Individual No.
✓			I	18 various[d]	54	Holstein	1
✓			II	milk cells & mammary	12	Holstein	6
✓		✓	III	white blood cells	105	Holstein	105[e]
✓	✓	✓	III	milk cells	131	Holstein & Jersey	105[e] & 26
✓		✓	IV	liver	35	Angus	35
✓		✓	IV	muscle	41	Angus	41

[a]: Differential splicing caused by tissue types estimated using RNA sequence data from all experiments. [b]: Differential splicing caused by Holstein and Jersey breeds estimated using RNA sequence data of milk cells from experiment III. [c]: cis splicing quantitative trait nucleotides, sQTN estimated using imputed whole genome and RNA sequence data from experiment III and IV. Data from experiment III and IV were also used to estimate exon expression eeQTN and gene expression geQTN. [d]: 18 tissues from [11]. [e]: The same 105 Holstein cattle, each of which had both white blood and milk cell transcriptomic data.

Differential splicing between tissues and breeds

We primarily defined a differentially spliced gene as a gene which contained exons whose inclusion ratios (exon expression divided by gene expression) were significantly affected by tissue or breeds (FDR<0.1). To verify the significantly spliced exons, we imposed a requirement that at least one adjacent intron had an excision ratio [8, 9] that was also significantly (FDR<0.1) affected by tissue or breeds (See methods). The overall FDR of such exon splicing events was considered as 0.01 as the exon and intron analyses were independent. The agreement of genes displaying differential splicing from exon and intron analyses were shown and examined in 2 × 2 tables by Chi-square tests (Additional file 2: Supplementary Table S1). Overall, the results from exon and intron analyses were significantly overlapped.

Using data from all experiments (Table 1), there were 8,657 genes with at least one differential spliced exon caused by differences between tissue types. A list of these genes

with the significances of differential splicing for the exons and introns were shown in Additional file 3: Supplementary Table S2. The top 10% of these significantly differentially spliced genes had a GO term enrichment (FDR<0.01) of 'regulation of cellular process', suggesting very general roles of these genes in cell functionalities. A much smaller number of genes (148) had differential splicing events in the milk cell transcriptome between breeds (Table 1, Additional file 4: Supplementary Table S3). While these genes did not show any significant GO term enrichments, they included the milk protein gene *CSN3* [12, 13], where the 5th exon (6:87,392,578-87,392,750) was more commonly included in the transcript in Holstein cattle than in Jersey cattle (Figure 1b).

cis splicing quantitative trait nucleotides (sQTN)

The sQTN mapping was based on data from 312 transcriptomes data generated from experiments III and IV, covering four major tissues (Table 1). In total 207 individuals had imputed whole genome sequence data and in experiment III, 105 genotyped cattle had both white blood and milk cell transcriptome data (Table 1). Similar to differential splicing analyses described above, a cis-acting sQTN was defined as a SNP significantly (FDR<0.1) affecting the inclusion ratio of the exon (up to 1Mb away) and significantly (FDR<0.1) affecting at least one adjacent introns' excision ratio [8, 9] . When analysed separately, the sQTN found by exon and intron analyses were also significantly overlapped (Additional file 2: Supplementary Table S1). After requiring that the inclusion/excision ratios for adjacent exons and introns were both affected by the same SNP, 138,796, 28,907, 11,544 and 5,783 sQTN were called in white blood cells, milk cells, liver and muscle, respectively (Figure 2, Additional file 2,5: Supplementary Table S1 and 4).

The significant sQTN in the white blood and milk cells were mapped to 929 and 283 genes, respectively (Table 2). For instance, in the milk cell transcriptome, the fifth exon of *CSN3*

(6:87392578-87392750), which as described above was differentially spliced in Holstein and Jersey cattle (Figure 1b), was significantly affected by a sQTN (Chr6:87392580, Additional file 5: Supplementary Table S4). The B allele of this SNP increased the expression of the 5th exon and had a higher allele frequency among Holstein cattle than Jerseys cattle (0.79 vs 0.02), which explains the difference between the breeds. This sQTN was also predicted to be a splice site by Ensemble [14]. Much smaller numbers of significant sQTN were detected in liver (11,544 SNPs) and muscle (5,783 SNPs) (Figure 2c,d). This was probably due to the smaller sample size (Table 1) and lower sequence depth of liver and muscle from the experiment IV (Additional file 1: Supplementary Methods) than that of blood and milk from the experiment III (Figure 2a,b, Table 2).

Table 2. Summary of QTN detection. cis sQTN: significant (FDR<0.01) SNPs within ± 1Mb of the exon, affecting its inclusion ratio and also affected the excision of an adjacent intron at the same significance level. cis eeQTN: significant (FDR<0.01) SNPs within ± 1Mb of the exon, affecting its expression level. cis geQTN: significant (FDR<0.01) SNPs within ± 1Mb of the gene affecting its expression level.

cis QTN type	Tissue	QTN number	Gene number
sQTN	white blood cells	138,796	929
	milk cells	28,907	283
	liver	11,544	49
	muscle	5,783	76
eeQTN	white blood cells	802,685	6,446
	milk cells	100,844	2,102
	liver	37,322	346
	muscle	64,675	1,267
geQTN	white blood cells	96,530	842
	milk cells	4,099	99
	liver	39,306	419
	muscle	57,054	1,150

Comparing sQTN with exon expression eeQTN and gene expression geQTN

Many more significant exon expression (eeQTN) than sQTN were detected in all tissues studied (Table 2). For the gene expression (geQTN), in white blood and milk cells, the

number of significant SNPs was smaller than the number of significant sQTN in white blood and milk cells (Table 2).

Figure 3a showed that sQTN had a median distance of about 200 kb from the transcription start site (TSS) and were slightly closer to the TSS than eeQTN and geQTN. For Ensembl predicted SNP function categories [14], all 3 classes of eQTN had a lower percentage of intergenic SNPs and a higher percentage of intronic and coding SNPs, including splice sites than the same categories across all SNPs analysed (Figure 3b). Specifically, sQTN had the highest percentage of intronic SNPs, compared to eeQTN and geQTN. However, no consistent ranking of concentrations of 'Splice' SNP category for sQTN, eeQTN and geQTN were observed in different tissues (Figure 3b).

Shared genetic influences between cis QTN types

Within each tissue, the sharing of SNPs between all three eQTN types was significantly more than expected by chance (Figure 4). However, in the white blood and milk cells, which had relatively large sample size (n>=105, Table 1), the largest absolute amount of SNP sharing appeared to be between sQTN and eeQTN (Figure 4). This was followed by the amount of SNP sharing between eeQTN and geQTN (Figure 4a,b). In liver and muscle tissue which had relatively small sample size (n<=41), the largest absolute amount of SNP sharing was between eeQTN and geQTN, followed by the amount of SNP sharing between sQTN and eeQTN (Figure 4c,d).

To further examine the relationship between sQTN and eeQTN, a two by two table of sQTN and eeQTN count in white blood and milk cells, which had comparable sample sizes, was created (Additional file 6: Supplementary Table S5). This suggested that when a sQTN was found, it was highly likely to be also identified as an eeQTN. For example, for 138,796 sQTN

found in the white blood cells, while 109,155 of them were also blood eeQTN, only 21,766 of them were identified as blood geQTN. Again, for these 138,796 blood sQTN, although only 18,005 and 25,932 of them were milk sQTN and eeQTN, respectively, an even smaller number, 720, of them were identified as milk cell geQTN.

Shared genetic influences between tissues

Within each QTN type between different tissues, the majority of the significant QTN sharing were observed between white blood and milk cells and between liver and muscle (Figure 5a). This not unexpected since the most of the white blood and milk cells came from the same lactating cows and the muscle and liver from different growing Angus bulls. The largest amount of cross-tissue QTN sharing was observed in eeQTN, followed by sQTN and geQTN (Figure 5). Where a SNP significantly affected expression in two tissues, the direction of effect was usually the same in both tissues (Additional file 10: Supplementary Figure S1). The correlation of QTN effects between white blood cells and milk cells (Additional file 10: Supplementary Figure S1a,c,e) was stronger than that between liver and muscle (Additional file 10: Supplementary Figure S1b,d,f). The SNP level sharing between white blood cells and milk cells and between liver and muscle were also evident at the exon (splicing quantitative trait loci sQTL and exon expression eeQTL) and gene (geQTL) level (Figure 5b).

The QTN sharing between tissues was further examined for all types of eQTN by using a less stringent p-value ($p<0.05$) to test their effect (Additional file 10: Supplementary Figure S2). This showed that the QTN sharing between tissues were stronger for sQTN and eeQTN (Additional file 10: Supplementary Figure S2a-d), than the sharing for geQTN (Additional file 10: Supplementary Figure S2e-f). Again, more QTN sharing was found between white blood cells and milk cells than between liver and muscle. For instance, 75% of eeQTN

significant in the white blood cells at $p<0.002$ were significant in milk cells at $p<0.05$ (Additional file 10: Supplementary Figure 2c).

The correlation between estimated SNP effects on gene splicing and expression in different tissues are lower in magnitude than the true correlation between SNP effects, because the effects are estimated with error and these errors are independent between tissues. To estimate the true correlation, we computed the genetic correlation between SNP effects in two different tissues by GREML[16] using a genomic relationship matrix or GRM built from 1 Mb surrounding the exon or gene (Figure 5c). Out of 1,145 analysed sQTL (inclusion ratio of the exon), eeQTL (expression level of the exon) and geQTL (expression level of the gene) between tissues, 598 had genetic correlations significantly ($p<0.05$) different from 0, out of which 561 had genetic correlations insignificantly ($p >= 0.05$) different from 1. That is, in many cases, the same additive genetic variance based on the *cis* polymorphism (or polymorphisms) was driving the expression in both tissues.

Often both splicing events and exon expression within a gene were highly correlated between white blood and milk cells, for instance *DDX19B*, *CTSD* and *EFF1A1* (Figure 5c). In liver and muscle, exons from *HLA-DQA1* encoding major histocompatibility complexes [15] also showed significant genetic correlations between tissues based on both exon expression and splicing. There were more cases of eeQTN than sQTN and geQTN and so there were more estimates of genetic correlations between white blood and milk cells in Figure 5c. The genetic correlations between eeQTN in white blood and milk cells show a range from +1 to -1 although most are close to +1. Exons with negative genetic correlations of expression between white blood cells and milk cells were mapped to *SART1* [15], a post-transcriptional regulator in epithelial tissues and *TTC4* with potential to mediate protein-protein interactions [15]. These negative genetic correlations imply that there are mutations that increase the expression of the exon in milk cells but decrease it in white blood cells. An exon within

S100A10, a cell cycle progress regulator, showed negative genetic correlation of expression between liver and muscle.

Genetic correlations between exon expression in two tissues can be different between exons within the same gene. For example, only the 2nd exon (5:93,942,055 -93,942,195) of *MGST1* (which affects dairy cattle milk fat yield [13, 16]) had a significant genetic correlation of expression between white blood cells and milk cells (Figure 5c). This was largely due to a few eeQTN with very strong effects on the expression levels of the 2nd exon in the milk cells and a similar but smaller effect on the 2nd exon expression in white blood cells (Additional file 10: Supplementary Figure S3).

Multi-transcriptome meta-analysis to increase power of expression QTN detection

Based on shared genetic effects of all types of eQTN across tissues, a multi-transcriptome meta-analysis was introduced to increase the power to detect sQTN, eeQTN and geQTN (Figure 6, Table 3). For sQTN, eeQTN and geQTN that had significant effects ($p < 0.05$) in all of white blood cells, milk cells and muscle transcriptomes, their standardised effects (signed t values) in each transcriptome were simply combined and tested for significance against a χ^2 distribution with 1 degree of freedom. Overall, the multi-transcriptome meta-analysis based on summary statistics substantially increased the power of QTN detection (Figure 6). The significance of multi-transcriptome QTN were compared with their significance in the liver transcriptome (Figure 6, Table 3). For a criteria where the expression QTN had both multi-transcriptome meta-analysis $p < 1\times10^{-5}$ and liver transcriptome analysis $p < 0.05$, all types of expression QTN had significant overlaps of the SNPs between the meta-analysis and the single transcriptome analysis in liver. In fact most of the significant sQTN, eeQTN and geQTN detected by the meta-analysis were also detected by the liver analysis but at a much higher p-value (Table 3).

Table 3. The agreement between the multi-transcriptome meta-analysis of three tissues and the single transcriptome results of liver

sQTN		Meta-analysis[b]		Overlap p
		+	-	
Liver[a]	+	768,650	2,842,552	p < 2e-16
	-	139,551	4,301,536	

eeQTN		Meta-analysis[b]		Overlap p
		+	-	
Liver[a]	+	1,052,041	3,186,839	p < 2e-16
	-	140,033	3,575,617	

geQTN		Meta-analysis[b]		Overlap p
		+	-	
Liver[a]	+	178,093	473,232	p < 2e-16
	-	43,578	6,663,855	

sQTN: splicing quantitative trait nucleotide; eeQTN: exon expression QTN; geQTN: gene expression QTN. [a]QTN with single-transcriptome effects $p < 0.05$ in each tissue of white blood cells, milk cells and muscle. [b]QTN with multi-transcriptome effects χ^2 $p < 1e-05$. '+' indicates the number of SNPs met the significance criteria while '–' indicates number of SNPs failed to meet the significance criteria.

Overlap between QTN and QTL for dairy and beef traits

We examined whether cis sQTN, eeQTN and geQTN were significantly enriched amongst SNPs associated with cattle economic traits. Pleiotropic SNPs significantly (FDR<0.01) associated with more than one of 24 dairy traits [17] and with more than one of 16 beef traits [18] were used to overlap with detected sQTN, eeQTN and geQTN (Figure 7, Additional file 7,8: Supplementary Table S6,7). Overall, sQTN, geQTN and eeQTN identified in white blood and milk cells had higher enrichment of SNPs associated with dairy and beef traits than sQTN, geQTN and eeQTN identified in liver and muscle (Figure 7). sQTN in white blood and milk cells were significantly enriched for dairy cattle pleiotropic SNPs, including SNPs

from the *CSN3* loci on chromosome 6 (Figure 7a,b). eeQTN in the white blood cells had the largest absolute amount of SNPs overlapping with dairy cattle pleiotropic SNPs (Figure 7b) and was the only QTN type with significant enrichment with beef cattle pleiotropic SNPs (Figure 7b). eeQTN in milk cells and liver also had significant enrichment for dairy cattle pleiotropic SNPs (Figure 7a). Of the geQTN, only those from white blood cells had a significant enrichment with dairy cattle pleiotropic SNPs (Figure 7a).

An example of an eeQTN and milk production QTN overlap is for *MGST1*, where effects of milk cell eeQTN were highly significantly associated with their effects on milk fat yield [17] (Figure 7c). Specifically, some QTN with large effects on the milk cell expression levels of exon 2 (Chr5:93,942,195-93,942,055) and exon 3 (Chr5: 93,939,244-93,939,150) of *MGST1* (Additional file 10: Supplementary Figure S3) also had large effects on milk fat yield (Figure 7c). Littlejohn et al.[16] identified SNPs associated with milk yield percentage and *MGST1* expression in the mammary gland, including 17 potential causal variants. Most SNPs identified by Littlejohn et al. originated from whole genome sequence and so were not present on the high density SNP chip data we analysed for dairy cattle pleiotropy [17] (Additional file 9: Supplementary Table S8). However, 53 significant milk cell eeQTN identified by the current study overlapped with the top 200 SNPs from Littlejohn et al [16] (Additional file 9: Supplementary Table S8), which was significantly more than expected by random chance. The overlapping milk cell eeQTN included the SNP suggested as a potential causal candidate (Chr5:93945738) [16], which significantly affected the expression level of the third exon (5:93939150-93939244) of *MGST1* in milk cells (Additional file 9: Supplementary Table S8). No milk cell geQTN was called for *MGST1*, as all of them had weak effects on the whole *MGST1* gene expression in milk cells, thus showed large FDR (Additional file 9: Supplementary Table S8).

Discussion

We performed a systematic analysis of transcriptomic cis QTN (<=1Mb) in multiple tissues centred around RNA splicing events, using a large number of RNA and whole genome sequence data from an important domestic animal species. Overall, differential splicing between tissues is ubiquitous and between breeds is common. Differential splicing between individuals due to SNPs (sQTN) occurs for many genes and are enriched with cattle complex trait QTL. Within each tissue, all cis QTN types showed significant overlap. However, SNPs causing an sQTN are also likely to cause eeQTN and, to a lesser extent, geQTN. Between tissues, while all QTN types showed significant overlap between white blood cells and milk cells and between liver and muscle, the strongest cross-tissue sharing appeared to be at the exon level (sQTN and eeQTN). This is supported by many significant tissue pair genetic correlations. Such cross-tissue QTN sharing allowed the multi-transcriptome meta-analysis of QTN effects which substantially increases power to detect significant QTN.

The majority of significant sQTN were detected from white blood and milk cells (Figure 2a,b) which also overlapped with SNP chip based complex trait QTL (Figure 7), compared to sQTN detected from liver and muscle. This is probably due to the larger sample size for white blood and milk cells than for liver and muscle (Table 1) and the higher sequencing depth (Additional file 1: Supplementary Methods). One of the significant white blood cell sQTN (29:44585782) for *CAPN1* is also a SNP chip based significant pleiotropic SNP for 16 beef cattle traits (Additional file 7: Supplementary Table S6) and this SNP is associated with shear force in multiple taurine breeds [19]. This implies that a splicing QTN for *CAPN1* may contribute to beef tenderness variation.

In the milk cell transcriptome, a significant sQTN (Chr6:87392580, Figure 2a) with predicted splicing function[14] within the fifth exon (6:87,392,578-87,392,750) of *CSN3* causes

differential splicing between Holstein and Jerseys (Figure 1b). Variants within *CSN3* have long been found to be associated with milk traits [20, 21] but only recently have potential causal variants been prioritised [12]. The milk cell sQTN 6:87392580 had a perfect linkage disequilibrium (r =1) with the variant 6:87390576 which has been suggested as a possible causal variant for effects on milk protein yield and percentage [12, 13] (Additional file 7: Supplementary Table S6). Given it is at a splicing site, it is likely that 6:87392580 is a causal variant contributing to milk production in dairy cattle by altering exon splicing.

Compared to identified bovine cis geQTN, cis sQTN tended to be closer to the transcription starting site (TSS) and had highest concentrations of intronic SNPs (Figure 3). In humans, cis sQTN [8, 22] also had stronger enrichments of intron SNPs than other types of QTN. However, reports of the distance between human QTN and TSS appear be inconsistent. While no difference in enrichment of SNPs near TSS between sQTN and geQTN were found by the human GTEx project [7], A more recent study[8] found that human geQTN were more enriched near TSS than sQTN. Our results appear to stand in between the results of GTEx project and the later findings from Li et. al. [8], where cattle sQTN were slightly more closer to TSS compared to geQTN. However, this difference is not significant in all tissues (Figure 3a). On the other hand, overall significant overlap between sQTN and geQTN was found in this study (Figure 4) and by the human GTEx project [7]. However, Li et. al. [8] found that human cis sQTN were independent of geQTN. These inconsistent observations were likely to be due to a number of differences between these studies, including definition of sQTN, choice of tissues and populations and computational procedures. Also, these inconsistent observations also suggest that we are still at the very early stage of understanding of sQTN features.

Within each studied bovine tissue, the largest amount of SNP overlap between QTN types were found either between exon expression eeQTN and sQTN or between eeQTN and

geQTN (Figure 4). Further, the largest amount of enrichments of cattle pleiotropic SNPs were found for eeQTN, followed by sQTN and geQTN. The white blood cell eeQTN showed particularly strong enrichments of pleiotropic SNPs for dairy and beef cattle. In a large scale human blood cell QTN study [23], eeQTN also showed the strongest enrichments of GWAS variants, followed by sQTN and geQTN. Thus, focusing on exon-level QTN, including eeQTN and sQTN, could increase the chance of finding regulatory variants for complex traits, as proposed by Guan et. al.[24].

A hypothesis to explain these results is that mutations in regulatory DNA may increase the expression of one or more transcripts from a gene. If they increase expression of one transcript then they may be detected that as an eeQTN for the exons in that transcript, as an sQTN for exons spliced out of that transcript or as a geQTN if this transcript forms a large part of the total transcription from the gene. Thus, there is expected to be overlap between eeQTN, geQTN and sQTN, but at least sQTN and eeQTN should overlap and this is what we found (Figure 4, Additional file 6: Supplementary Table S5). It appears that eeQTN detect the largest proportion of these regulatory polymorphisms provided sequencing depth is high.

In humans, significant cross-tissue sharing of sQTN was reported [7] and so was geQTN in different studies [7, 25]. In our study of cattle, the strongest evidence of QTN sharing appeared to be at the exon level. This included sQTN and eeQTN sharing between white blood and milk cells and between liver and muscle (Figure 5). When extending the examination of QTN to include those with small effects ($p < 0.05$, Additional file 10: Supplementary Figure S2), the exon-level QTN cross-tissue sharing is also the greatest. We highlighted a few examples of cross-tissue shared eeQTN along with the related exons, of which the genetic correlations of the expression and splicing in different tissues were significant (Figure 5c). One of these eeQTN is located in the milk fat yield [13, 16] locus *MGST1* (Figure 5c, Additional file 10: Supplementary Figure S3). For eeQTN associated

with *MGST1*, a strong positive relationship of SNP effects was observed between milk cell eeQTN and dairy milk fat yield SNPs (Figure 7c). Furthermore, the identified milk cell eeQTN overlaps with previously identified causal variants [16] within *MGST1* for milk fat percentage, thus supporting their candidacy. This overlap further supports the top candidacy of SNP 5:93945738 with significant effects on the third exon of *MGST1* (Additional file 9: Supplementary Table S8) for milk fat traits. Overall, our analysis demonstrates the significant potential of using detailed exon analysis to aid in identification of putative causative mutations.

Based on the sharing of QTN between tissues, a multi-transcriptome meta-analysis which simply combined QTN effects to substantially increase the power (Figure 6) was introduced. Using this approach, combined QTN effects of white blood cells, milk cells and muscle were validated in the liver (Table 3). This also demonstrated the significant extent of QTN sharing across tissues. Previously, Flutre et. al. [25] combined data from human fibroblasts, lymphoblastoid cell lines and T-cells and found that up to 88% of geQTN were shared across tissues at FDR<0.05 level. In the current study, when the FDR threshold was lowered to 0.05, the meta-analysis combining SNP effects from tissues of white blood cells, milk cells and muscle identified 585,406 geQTN as significant in more than one tissue. This accounts for 69.2% of total common geQTN (845,431) called in the three tested tissues. Given the differences of tissue/cell type biology and sequencing precision between in vitro and in vivo experiments, we argue that our meta-analysis detected sharing of geQTN are comparable with the results from Flutre et. al. Our meta-analysis is powerful for detecting and validating many eQTN that have an effect in the same direction in multiple tissues, and is simpler to implement than that of Flutre et al.

In conclusion, our findings obtained from one of the earliest comprehensive analyses of whole genome and transcriptome sequence data in animals significantly improved our

understanding of variant impacts at the cellular level. sQTN appeared to be closer to the gene body and transcription start sites and was enriched in cattle complex trait variants. However, when comparing with eeQTN and geQTN in various tissues, we found that the exon-level QTN (sQTN and eeQTN) together can be more informative. The splicing and expression level of many exons was determined by the same *cis* genetic variance in different tissues. The potential of exon-level QTN information was demonstrated by the identification of several strong candidates of causal mutations to follow up: sQTN 6:87392580 within *CSN3* for milk production, sQTN 29:44585782 associated with *CAPN1* for beef tenderness and eeQTN 5:93945738 within *MGST1* for milk fat yield. Based on extensive cross-tissue sharing of QTN, a simple and effective multi-transcriptome meta-analysis was introduced to increase power and to validate QTN effects. Our findings have updated the knowledge of the genetic control of mammalian genome regulations and provided effective solutions for analysing high dimensional genomic and transcriptome data.

Methods

Sample collection. For Experiment I, the sampling of 18 tissues from one lactating Holstein cow followed procedures described by Chamberlain et al [11]. For Experiment II and III, the sampling and processing of all tissues including white blood and milk cells was detailed Additional file 1: Supplementary Methods. Briefly, animals of Experiment II and III were selected from Agriculture Victoria Research dairy herd at Ellinbank, Victoria, Australia and the ethics was approved by the Victoria Animal Ethics Committee (application number 2013-14). In the Experiment II, milk and mammary tissue samples were taken from six Holstein cows. In the Experiment III, milk and blood samples were originally taken from 112 Holstein and 29 Jersey cows, but only RNA sequence data of 105 Holstein and 27 Jersey met

concordant alignment rate [26] >70% were used. For Experiment IV, the sampling of 41 *semitendinosus* muscle and 35 liver from Angus bulls was previously described by [27, 28].

RNA seq data. For Experiment I, RNA extraction and sequencing was followed Chamberlain et al [11]. For Experiment II and III, the RNA extraction and sequencing procedures was detailed in Additional file 1: Supplementary Methods. For Experiment IV, RNA extraction and sequencing is previously described by Khansefid et al. [28]. For all experiments sequence quality was checked and were aligned to the Ensembl UMD3.1 bovine genome assembly using TopHat2 [26]. The RNA sequence data processing and results were detailed in Additional file 1: Supplementary Methods.

Whole genome seq data. Experiments III and IV had whole genome sequence genotypes imputed from the SNP chip genotypes using FImpute [29] based on the 1000 bull genomes project [30]. 50K Illumina genotypes were used for imputation for experiment III with previous protocols [13, 31]. For experiment IV, 800K and 50K Illumina genotypes were used with procedures following [18]. SNPs were filtered for minor allele frequency > 0.01 and resulted in 14,302,604 and 13,632,145 SNPs used in the analysis in experiment III and IV, respectively. There were 10,242,837 SNPs shared between experiment III and IV.

Gene/exon analysis. Gene count data were generated by Python package HTSeq [32] using default settings. The exon count data were generated by Bioconductor package featureCounts [33] in R v3.3.2 [34]. Genes and exons with count per million >0 in more than 40% individuals in each tissue type were used for all following analyses. The exon-based tissue principal components analysis used DEseq2 [35]. The exon inclusion as calculated as the exon to gene expression ratio and the intron excision was calculated as the intron to global intron cluster expression ratio, processed by leafcutter package[8, 9]. Those exons and introns with ratio values <0.001 were removed and the remaining ratio values were transformed to

$\log_2$ scale, then underwent exon/intron –wise quantile normalisation and individual –wise zscore standardisation [36].

Gene differential splicing. Both exon inclusions and intron excisions were analysed and used in combination for gene differential splicing for (1) the overall tissue effects and (2) the breed effects. Primarily, differential splicing was defined for the gene containing exons whose inclusion ratios were significantly (FDR<0.1) affected by the tissue or breed variable. To be called as significantly spliced exons, they were required to have at least one adjacent intron whose excision ratios were also significantly (FDR<0.1) affected by the tissue or breed variable. The tissue effects were analysed in a linear mixed model in lme4 [37] in R as:

$$y_{ijk} = b_i + x_j + t_k + e \quad (1)$$

Where y = exon inclusion or intron excision ratios, b_i = the animal random effects (i=214), x_j= the experiments (j=4), t_k = tissue type (k=19), e= random residual terms. The breed effects for the milk transcriptome data were analysed in a linear model in R as:

$$y_l = breed_l + e \quad (2)$$

Where y = exon inclusion or intron excision ratios in the milk cell transcriptome, $breed_l$ = breeds (l=2, Holstein and Jersey). p values of exons/introns for the tissue effects in equation (1) or the breed effects in equation (2) were used to calculate the false discovery rates (FDR) using qvalue [38] in R. As the exon and intron analyses were relatively independent, the overall FDR threshold of such detected exon/intron group was considered as 0.1× 01 = 0.01.

cis expression splicing QTN. Similar to differential splicing analysis described above, a significant (FDR<0.01) cis splicing QTN was demanded to satisfy two conditions simultaneously: (1) a SNP, within or up to ± 1Mb away from the exon, significantly (FDR<0.1) affected its inclusion ratio and (2) the same SNP significantly (FDR<0.1) affected

at least one event of the excision of the intron next to the same exon at the same significance level. Both individual exon inclusion and intron excision values were used as phenotype to map associated QTN with widely used [7] Matrix eQTL[39] package in R. For each cell type of the experiment III (white blood and milk cells) and experiment IV (liver and muscle), SNPs ± 1Mb from the exon or intron were tested for regressions with the exon inclusion or intron excision phenotype. For milk cell transcriptome, breed was fitted as a covariate.

To compare cis sQTN with exon expression cis eeQTN and cis gene expression geQTN, the expression count data were normalised by voom [40] estimating mean-variance relationship to calculate observation-level weighted expression values. Normalised expression values of exons and genes were used as phenotype to map cis QTN (within ± 1Mb) at FDR <0.01 level as described above.

SNP annotation. The gene transcription start site coordinates were downloaded from Ensembl (http://www.ensembl.org) and the absolute difference between the position of a SNP and the transcription start site of the gene were calculated for SNP with significant cis effects. The SNP functional categories were generated using predictions from Ensembl Variant Effect Predictor [14] in conjunction with NGS-SNP [41]. All analysed SNPs were assigned a functional category.

Dairy and beef cattle pleiotropic QTL. To test the significance of overlap between cis QTN and SNPs associated with cattle phenotype, meta-analyses of dairy and beef cattle pleiotropy were performed using single-trait GWAS results from Xiang et al [17] and Bolormaa et al [18]. HD 800K SNP chip genotype were used for trait GWAS. 24 dairy cattle traits with matching phenotype in 9,662 bulls and cows and 16 beef cattle traits with animal numbers >2,000 were selected. Briefly, the multi-trait χ^2 statistic for the *i*th SNP was calculated based on its signed t values generated from each single trait GWAS [18]:

$$\chi^2 = t'_i V^{-1} t_i \quad (3)$$

For dairy cattle, the meta-analysis was based on the weighted SNP effects t_w combining SNP effects calculated separately in bulls and cows. The t_w accounting for the phenotypic error differences between bulls and cows [13] was calculated as:

$$t_w = \frac{B_w}{se_w} = \frac{\dfrac{\dfrac{B_{bull}}{se_{bull}^2} + \dfrac{B_{cow}}{se_{cow}^2}}{\dfrac{1}{se_{bull}^2} + \dfrac{1}{se_{cow}^2}}}{\sqrt{\dfrac{1}{\dfrac{1}{se_{bull}^2} + \dfrac{1}{se_{cow}^2}}}} \quad (4)$$

Where the weighted SNP t value t_w was the quotient of the weighted SNP effects B_w and the weighted effect error se_w. B_{bull} and se_{bull} were the SNP effects and error obtained from single-trait GWAS in bulls and B_{cow} and se_{cow} were the SNP effects and error of cows. Those SNPs which had meta-analysis FDR<0.01 were chosen to be compared with cis QTN. The lead SNP loci were defined as ±1 Mb from the lead SNPs identified in the previous analysis [17] and [18].

The significances of overlap were compared with the expected number using the Fisher's exact test (p) implemented in GeneOverlap [42] in R. The union number of whole genome sequence SNPs with MAF>0.01 in each breed and the bovine high density chip SNPs was used as the background. If QTN categories from different breeds of dairy and beef cattle were tested for the overlap, the number of common SNPs between breeds were used.

The cross-tissue sharing of SNPs were confirmed by bivariate GREML analysis using GCTA [43]. For an exon or a gene of interest, its inclusion ratios or expression levels in two different tissues were treated as two different phenotype, tr_1 and tr_2. The SNPs within 1Mb of this exon or gene were used to make a genomic relationship matrix representing the local

polygenic component *a*. This allowed linear mixed modelling of the additive genetic variances of tr_1, $\mathrm{var_g}(tr_1)$ and of tr_2, $\mathrm{var_g}(tr_2)$ and the additive genetic covariance between t_1 and t_2, $\mathrm{cov_g}(tr_1,tr_2)$ using GREML [43]. This then allowed the estimation of genetic correlation $r_g = \frac{\mathrm{cov}_g(tr_1, tr_2)}{\sqrt{\mathrm{var}_g(tr_1)\,\mathrm{var}_g(tr_2)}}$ (5). Genetic correlations were also tested for their significances of being different from 0 or 1, by fixing the correlation value to 0 and 1 using GCTA [43].

Validation by multi-transcriptome meta-analysis. The validation based on QTN effect commonality across tissues were conducted by comparing the combined QTN effects from white blood cells (experiment III), milk cells (experiment III) and muscle (experiment IV) transcriptomes with their effects in the liver transcriptome (experiment IV). The standardised QTN effects, b/se, signed t values were calculated from single-transcriptome results of white blood cells (t_1), milk cells (t_2) and muscle (t_3). The significance of multi-transcriptome effects of a QTN was tested by χ^2 distribution with 1 degree of freedom:

$$\chi_{(1)}{}^2 = \left[\sum_{n=1}^{N} \frac{t_n}{\sqrt{N}}\right]^2 \quad (6)$$

N = the number of studied tissues (*N*=3 in this case) where the original SNP t values were estimated. The QTN that participated in the validation analysis had single-transcriptome effect $p < 0.05$ in each tissue and the significance of the multi-transcriptome effects was defined as $p < 1e\text{-}05$. Significant multi-transcriptome QTN were compared with the liver single-transcriptome effects at $p < 0.05$ level.

List of abbreviations

sQTN: splicing quantitative trait nucleotides;

geQTN: gene expression quantitative trait nucleotides;

eeQTN: exon expression quantitative trait nucleotides.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The RNA sequence data for experiment I was published [11] (NCBI Sequence Read Archive, SRA, accession SRP042639); For experiment II: SRA accession SRP111067; For experiment III: SRA accession PRJNA305942; For experiment IV: SRA accession PRJNA392196; The imputed whole genome sequence data was part of the 1000 bull genome project [30].

Competing interests

The authors declare that they have no competing interests.

Funding

Australian Research Council's Discovery Projects (DP160101056) supported R.X. and M.E.G. Dairy Futures CRC and DairyBio (a joint venture project between Agriculture Victoria and Dairy Australia) supported the generation of the transcriptome data of experiment I-III. No funding bodies participated in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author contributions

A.J.C., R.X., M.E.G., B.J.H and H.D.D. conceived the experiments. C.P.P., C.M.R., B.A.M., J.B.G., L.C.M., Y.C. and A.J.C. performed sample collections and RNA sequencing experiments. S.B., I.M.M., M.K. and H.D.D. provided data and assisted with study design. R.X., B.J.H., C.J.V., A.J.C., M.E.G and P.J.B. analysed data. R.X. and M.E.G. wrote the paper. R.X., M.E.G., B.J.H, A.J.C., I.M.M and H.D.D.. revised the paper. All authors read and approved the final manuscript.

Acknowledgements

We thank DataGene for access to data used in this study and Gert Nieuwhof, Kon Konstantinov and Timothy P. Hancock for preparation and provision of data.

Reference

1. Bourneuf E, Otz P, Pausch H, Jagannathan V, Michot P, Grohs C, Piton G, Ammermüller S, Deloche M-C, Fritz S: **Rapid Discovery of De Novo Deleterious Mutations in Cattle Enhances the Value of Livestock as Model Species.** *Sci Rep* 2017, **7**.
2. Hayes BJ, Lewin HA, Goddard ME: **The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation.** *Trends Genet* 2013, **29:**206-214.

3. Meuwissen T, Hayes B, Goddard M: **Genomic selection: A paradigm shift in animal breeding.** *Animal frontiers* 2016, **6:**6-14.
4. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW: **Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project.** *Genome Biol* 2015, **16**.
5. Albert FW, Kruglyak L: **The role of regulatory variation in complex traits and disease.** *Nature reviews Genetics* 2015, **16:**197.
6. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M: **Linking disease associations with regulatory information in the human genome.** *Genome Res* 2012, **22:**1748-1759.
7. Consortium G: **The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans.** *Science* 2015, **348:**648-660.
8. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK: **RNA splicing is a primary link between genetic variation and disease.** *Science* 2016, **352:**600-604.
9. Li YI, Knowles DA, Pritchard JK: **LeafCutter: Annotation-free quantification of RNA splicing.** *bioRxiv* 2016**:**044107.
10. Cánovas A, Rincón G, Bevilacqua C, Islas-Trejo A, Brenaut P, Hovey RC, Boutinaud M, Morgenthaler C, VanKlompenberg MK, Martin P: **Comparison of five different RNA sources to examine the lactating bovine mammary gland transcriptome using RNA-Sequencing.** *Sci Rep* 2014, **4**.
11. Chamberlain AJ, Vander Jagt CJ, Hayes BJ, Khansefid M, Marett LC, Millen CA, Nguyen TTT, Goddard ME: **Extensive variation between tissues in allele specific expression in an outbred mammal.** *BMC Genomics* 2015, **16:**993.
12. MacLeod I, Bowman P, Vander Jagt C, Haile-Mariam M, Kemper K, Chamberlain A, Schrooten C, Hayes B, Goddard M: **Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits.** *BMC Genomics* 2016, **17:**1.
13. Kemper KE, Reich CM, Bowman PJ, vander Jagt CJ, Chamberlain AJ, Mason BA, Hayes BJ, Goddard ME: **Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions.** *Genet Sel Evol* 2015, **47:**1.
14. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F: **The Ensembl Variant Effect Predictor.** *Genome Biol* 2016, **17:**1.
15. **National Center for Biotechnology Information (NCBI)** [https://www.ncbi.nlm.nih.gov/]
16. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, Couldrey C, Keehan M, Sherlock RG, Harland C: **Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition.** *Sci Rep* 2016, **6**.
17. Xiang R, MacLeod I, Bolormaa S, Goddard M: **Genome-wide comparative analyses of correlated and uncorrelated phenotypes identify major pleiotropic variants in dairy cattle.** *Sci Rep* 2017, **7:**9248.
18. Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, Tier B, Savin K, Hayes BJ, Goddard ME: **A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle.** *PLoS Genet* 2014, **10:**e1004198.
19. McClure M, Ramey H, Rolf M, McKay S, Decker J, Chapple R, Kim J, Taxis T, Weaber R, Schnabel R: **Genome-wide association analysis for quantitative trait loci influencing Warner–Bratzler shear force in five taurine cattle breeds.** *Anim Genet* 2012, **43:**662-673.
20. Kühn C, Freyer G, Weikard R, Goldammer T, Schwerin M: **Detection of QTL for milk production traits in cattle by application of a specifically developed marker map of BTA6.** *Anim Genet* 1999, **30:**333-339.
21. Glantz M, Gustavsson F, Bertelsen HP, Stålhammar H, Lindmark-Månsson H, Paulsson M, Bendixen C, Gregersen VR: **Bovine chromosomal regions affecting rheological traits in acid-induced skim milk gels.** *J Dairy Sci* 2015, **98:**1273-1285.
22. Takata A, Matsumoto N, Kato T: **Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci.** *Nature Communications* 2017, **8:**14519.

23. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, van't Hof P, Mei H, van Dijk F, Westra H-J: **Identification of context-dependent expression quantitative trait loci in whole blood.** *Nat Genet* 2017, **49:**139-145.
24. Guan L, Yang Q, Gu M, Chen L, Zhang X: **Exon expression QTL (eeQTL) analysis highlights distant genomic variations associated with splicing regulation.** *Quantitative Biology* 2014, **2:**71-79.
25. Flutre T, Wen X, Pritchard J, Stephens M: **A statistical framework for joint eQTL analysis in multiple tissues.** *PLoS Genet* 2013, **9:**e1003486.
26. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14:**R36.
27. Chen Y, Gondro C, Quinn K, Herd R, Parnell P, Vanselow B: **Global gene expression profiling reveals genes expressed differentially in cattle with high and low residual feed intake.** *Anim Genet* 2011, **42:**475-490.
28. Khansefid M MC, Chen Y, Pryce JE, Chamberlain AJ, Vander Jagt CJ, Gondro C, Goddard ME.: **Gene expression analysis of blood, liver and muscle in cattle divergently selected for high and low residual feed intake.** *J Anim Sci* 2017, **In press**.
29. Sargolzaei M, Chesnais JP, Schenkel FS: **A new approach for efficient genotype imputation using information from relatives.** *BMC Genomics* 2014, **15**.
30. Daetwyler HD, Capitan A, Pausch H, Stothard P, Van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C: **Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle.** *Nat Genet* 2014, **46:**858-865.
31. Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, Mason B, Goddard M: **Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels.** *J Dairy Sci* 2012, **95:**4114-4129.
32. Anders S, Pyl PT, Huber W: **HTSeq—a Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31:**166-169.
33. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2013, **30:**923-930.
34. Team RC: **R: A language and environment for statistical computing.** 2013.
35. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15:**550.
36. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, de Leon S, Michelini K, Lewellen N, Crawford GE: **DNase I sensitivity QTLs are a major determinant of human expression variation.** *Nature* 2012, **482:**390-394.
37. Bates D, Mächler M, Bolker B, Walker S: **Fitting linear mixed-effects models using lme4.** *arXiv preprint arXiv:14065823* 2014.
38. Bass AJ, Dabney A, Robinson D: **qvalue: Q-value estimation for false discovery rate control. R package version 2.9.0, http://github.com/jdstorey/qvalue.**, vol. 2017; 2015.
39. Shabalin AA: **Matrix eQTL: ultra fast eQTL analysis via large matrix operations.** *Bioinformatics* 2012, **28:**1353-1358.
40. Law CW, Chen Y, Shi W, Smyth GK: **Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome Biol* 2014, **15:**R29.
41. Grant JR, Arantes AS, Liao X, Stothard P: **In-depth annotation of SNPs arising from resequencing projects using NGS-SNP.** *Bioinformatics* 2011, **27:**2300-2301.
42. Shen L: **GeneOverlap: An R package to test and visualize gene overlaps.** 2014.
43. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: a tool for genome-wide complex trait analysis.** *Am J Hum Genet* 2011, **88:**76-82.

Additional files:

Additional file 1 (DOCX): Supplementary Methods.

Additional file 2 (XLSX): Supplementary Table S1. Agreement between exon and intron analysis.

Additional file 3 (XLSX): Supplementary Table S2. Tissue splicing results.

Additional file 4 (XLSX): Supplementary Table S3. Breed splicing results.

Additional file 5 (XLSX): Supplementary Table S4. sQTN summary.

Additional file 6 (XLSX): Supplementary Table S5. SNP overlap between white blood and milk sQTN and eeQTN.

Additional file 7 (XLSX): Supplementary Table S6. SNP overlap between expression QTN and dairy cattle pleiotropic variants.

Additional file 8 (XLSX): Supplementary Table S7. SNP overlap between expression QTN and beef cattle pleiotropic variants.

Additional file 9 (XLSX): Supplementary Table S8. SNP overlap between blood eeQTN and causal variants identified by Littlejohn et al (2016).

Additional file 10 (PDF): Supplementary Figure S1-3.

Figure Legends:

Figure 1. **a**: Sample principal components clustering based on exon expression. **b**: The significant splicing events between breeds and between genotypes (cis quantitative trait nucleotides, sQTN) agreed on *CSN3* in the milk cell transcriptome. In the upper panel of **b**, the expression of the gene (ENSBTAG00000039787) and the 5th exon (6:87392578-87392750), the inclusion ratio of the 5th exon and the frequency of the B allele of the sQTN (Chr6:87392580) for *CSN3* were compared between breeds. The effects (signed t values, b/se) of the sQTN B allele on the expression of the gene, the 5th exon and the inclusion ratio of the 5th exon were given in the lower panel of **b**.

Figure 2. Manhattan plots of significant cis splicing quantitative trait nucleotides (sQTN, FDR<0.01 and within 1Mb of the exon) in white blood cells (**a**), milk cells (**b**), liver tissue (**c**) and muscle tissue (**d**). A significant sQTN was defined as a SNP affecting the exon inclusion ratio and also at least one excision of an adjacent intron at the same significance level. The input SNPs had significance p<0.0001. sQTN in all tissues with their associated genes and significance were given in Supplementary Table S4.

Figure 3. Features of cis splicing quantitative trait nucleotide (sQTN) compared to exon expression eeQTN and gene expression geQTN. **a**: The distance between the transcription start site (TSS) and the QTN. TSS information were downloaded from Ensembl(bovine reference UMD3.1). **b**: The proportions of QTN annotated as splice, UTR, gene_end, synonymous, missense, intron, intergenic or other. SNP annotations were based on Variant

Effect Predictor. ‘Splice’ included all SNP annotations containing the word ‘splice’. ‘UTR’ included 3’ and 5’ untranslated region. ‘Gene_end’ included upstream and downstream.

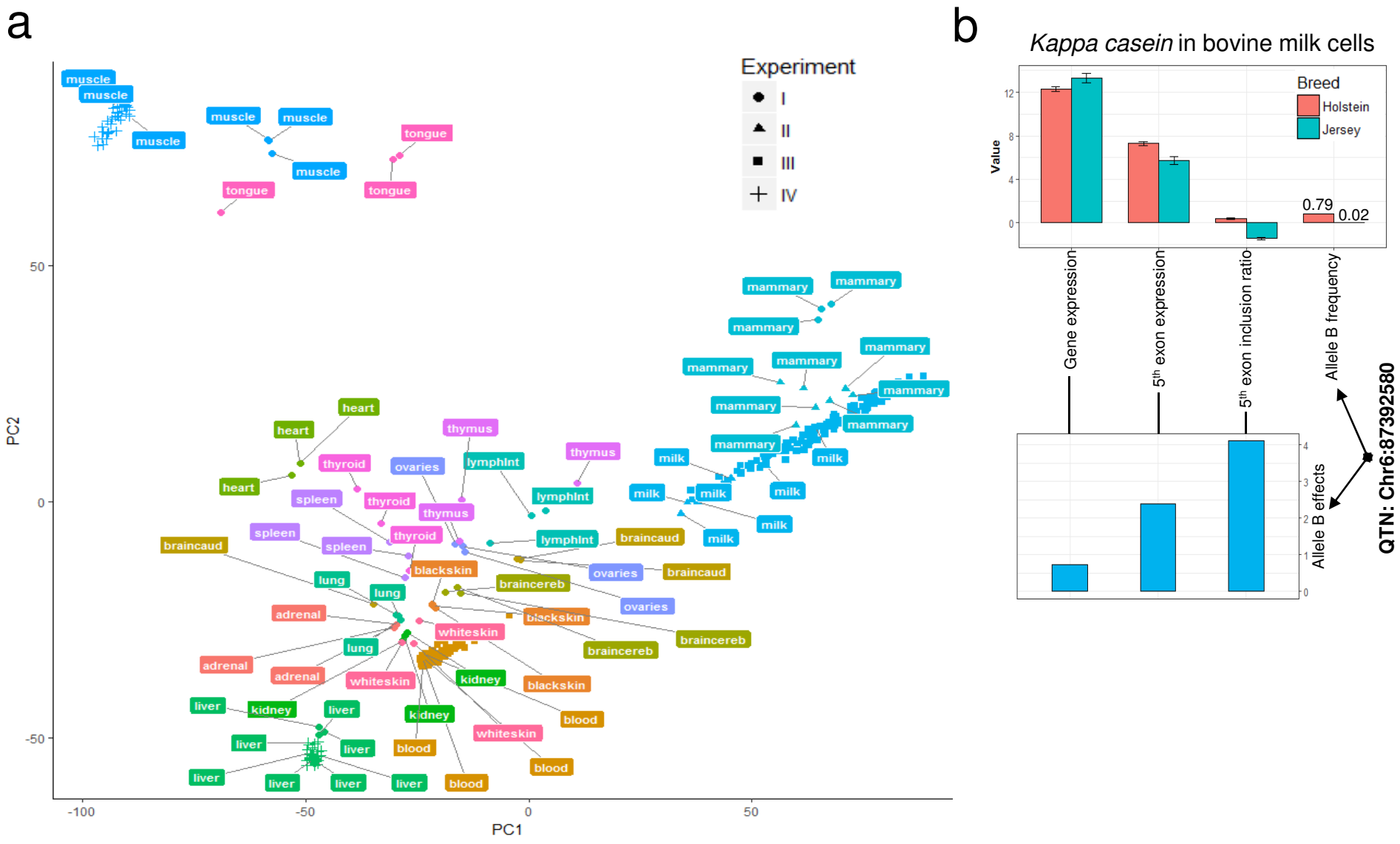
Figure 4. Number of cis splicing sQTN, exon expression eeQTN and gene expression geQTN as well as the number of SNP overlapping between them within white blood cells (**a**), milk cells (**b**), liver(**c**) and muscle (**d**). The green dots indicate the types of significant expression eQTN entered the overlap analysis. Significances of the overlap were based on the Fisher’s exact test, where the darker the colour the higher the significance.

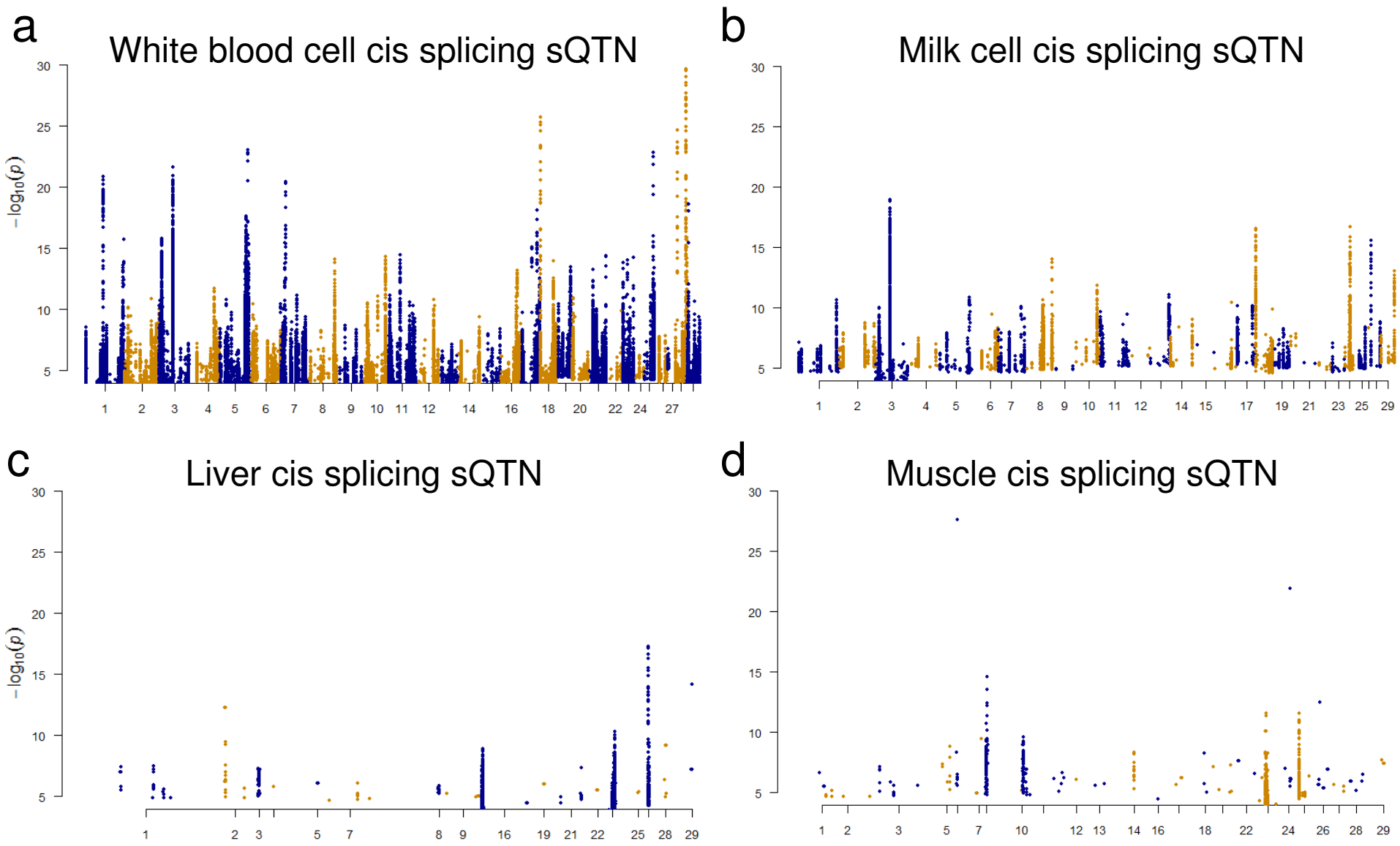
Figure 5. Shared genetic influence on the splicing, exon and gene expression between tissues. Blood refers to white blood cells and milk refers to milk cells. **a**: Each matrix shows the pair wise comparison of the numbers of significant SNP and the total number of significant SNPs detected for each analysis shown in parentheses. The significance of each overlap was tested by Fisher’ exact test, given the total number of SNP analysed and the total number of significant SNP, the result of which is represented by the colour of that position in the matrix. **b**: Each matrix shows the pair wise comparison of the numbers of exon/gene with significant associations and the total number of exon/gene detected with significant assosications for each analysis shown in parentheses. In panel **b**, the numbers were either exon numbers for sQTN (splicing quantitative trait nucleotide) and eeQTN (exon expression quantitative trait nucleotide) or gene numbers for geQTN (gene expression quantitative trait nucleotide. **c**: Between tissue genetic correlations of either the inclusion ratio of the exons, the expression of the exons or the expression of the genes that had significant sharing of QTN in panel **a.** Dot size and transparency were negatively correlated with p value of the significance of the

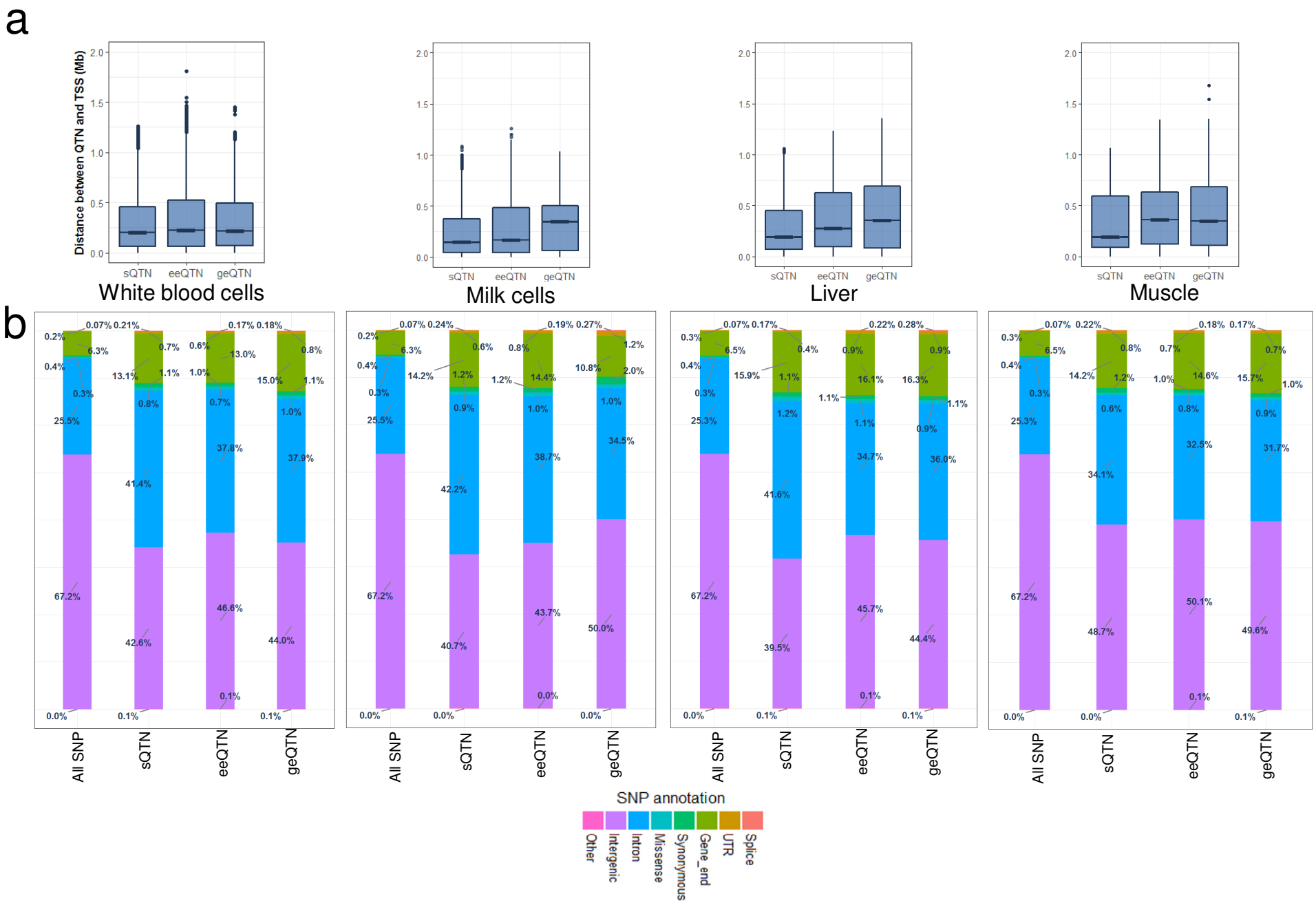
genetic correlation being different from 0. The error bars of the genetic correlation were shown in vertical lines of each dot. Some genes of interests were highlighted.

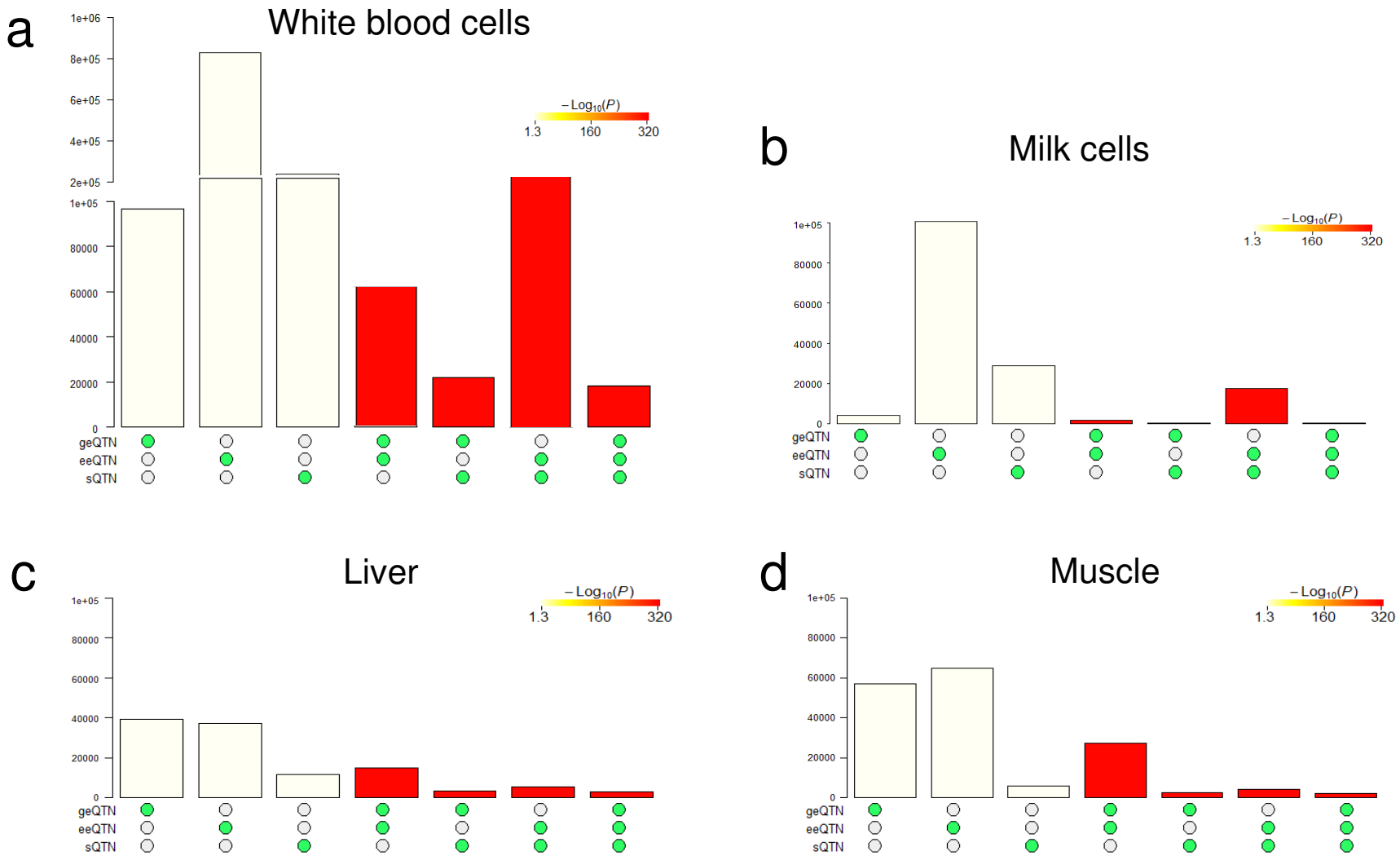
Figure 6. Multi-transcriptome meta-analysis (blood, milk and muscle) for cis splicing sQTN (**a**), exon expression eeQTN (**b**) and gene expression geQTN (**c**). In each panel, the significance of multi-transcriptome effects were tested against a χ^2 with 1 degree of freedom for combined QTN effects (dots in blue and orange). These multi-transcriptome effects were shown together with the single-transcriptome effects in liver of the same QTN (dots in green).

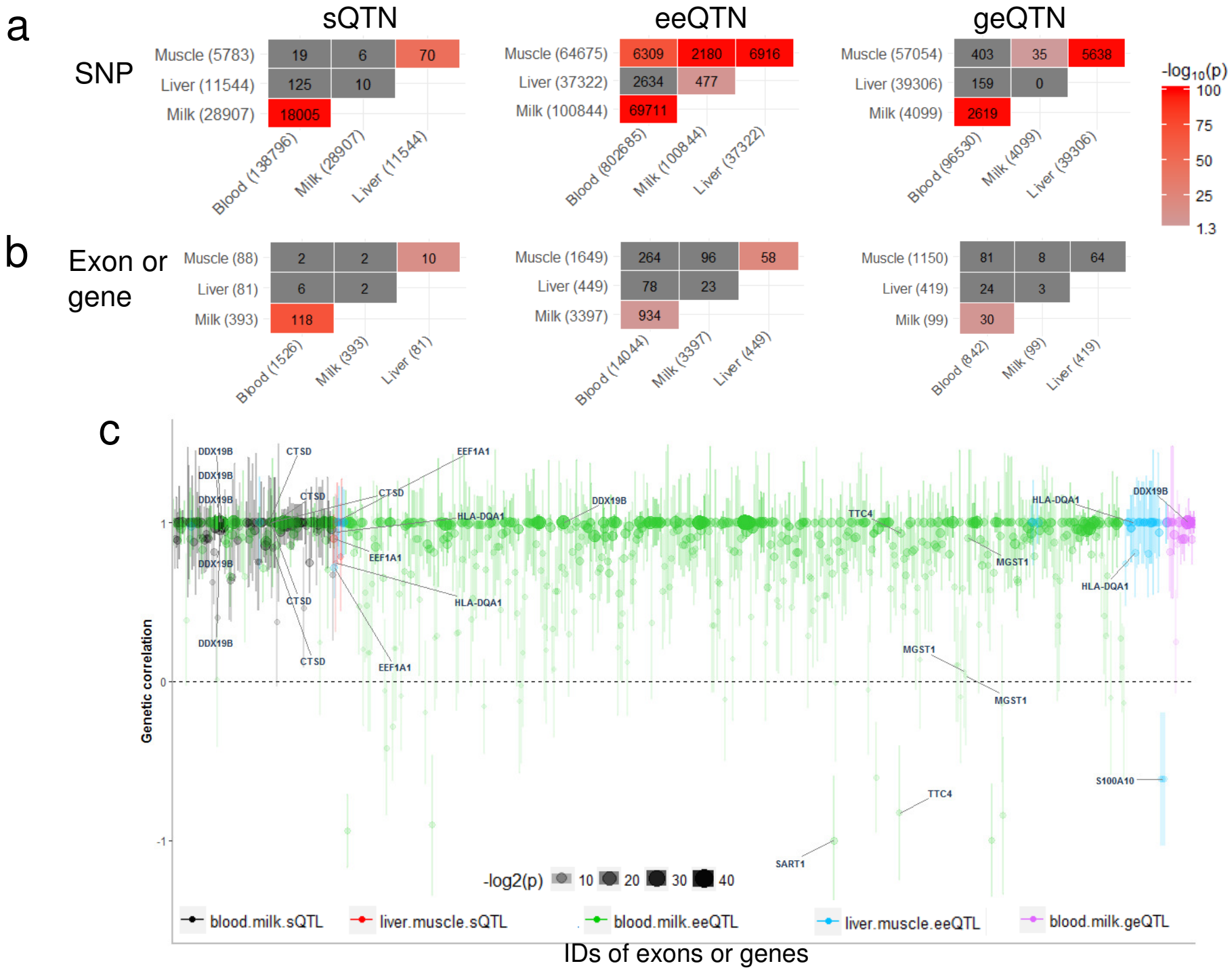
Figure 7. Significance of the overlap, based on the Fisher's exact test, between pleiotropic QTL for a range of traits in cattle (dairy, panel **a** and beef, panel **b**) and cis splicing quantitative trait nucleotide (sQTN), exon expression (eeQTN) and gene expression (geQTN) in all tissues, where the colour represents the significance of the overlap. Blood refers to white blood cells and milk refers to milk cells. Significance of the overlap were based on the Fisher's exact test. Only chromosomes containing overlapped SNPs were shown. **c**) An example of *MGST1* showing the relationship between QTN effects on exon expression in milk cells and their effects on dairy cattle milk fat yield.

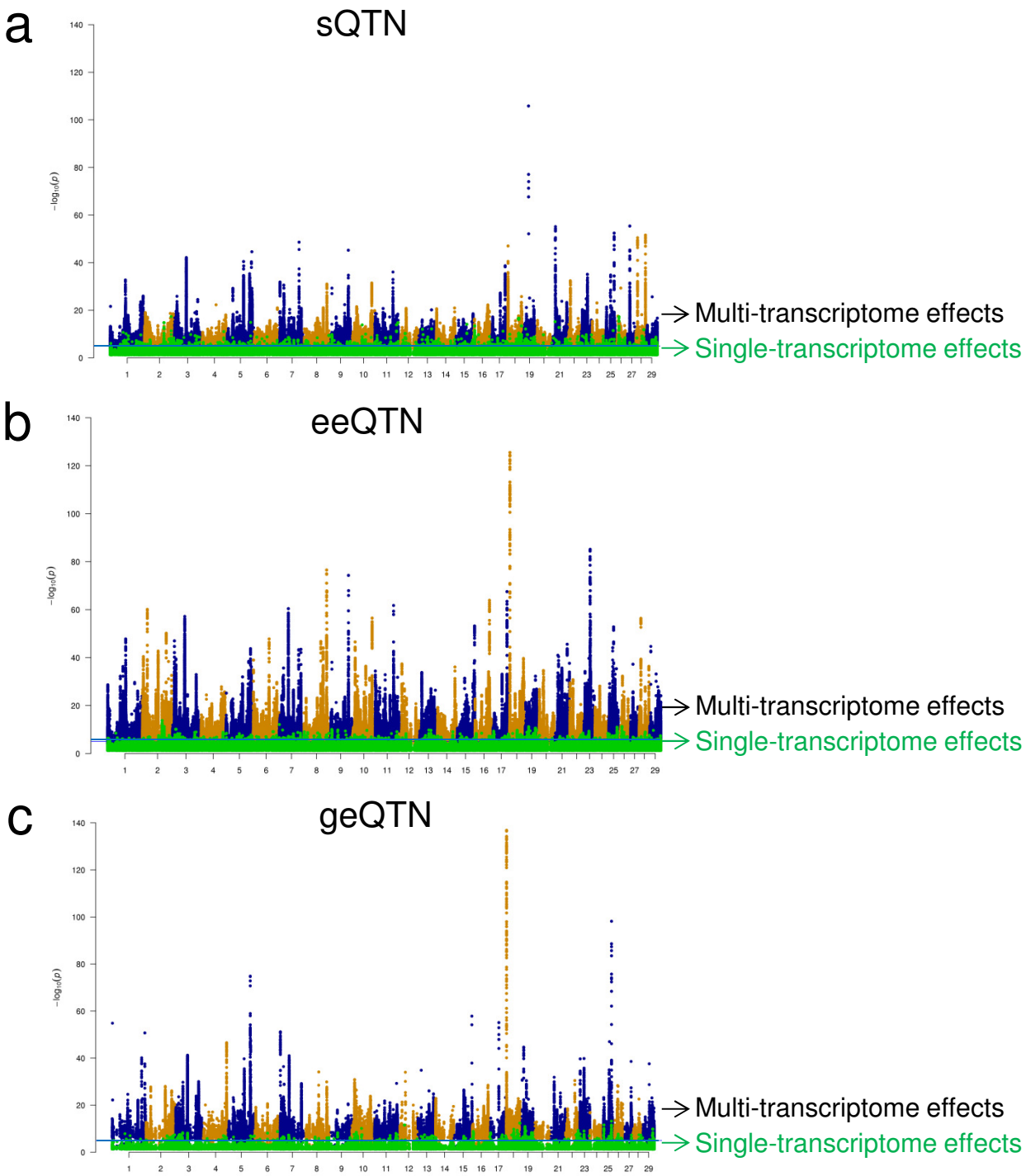

a
Experiment
I
II
III
IV
PC1
PC2
muscle
tongue
mammary
milk
heart
thymus
thyroid
ovaries
lymphInt
spleen
braincaud
blackskin
lung
braincereb
adrenal
whiteskin
kidney
liver
blood
b
Kappa casein in bovine milk cells
Breed
Holstein
Jersey
Value
0.79
0.02
Gene expression
5th exon expression
5th exon inclusion ratio
Allele B frequency
Allele B effects
QTN: Chr6:87392580

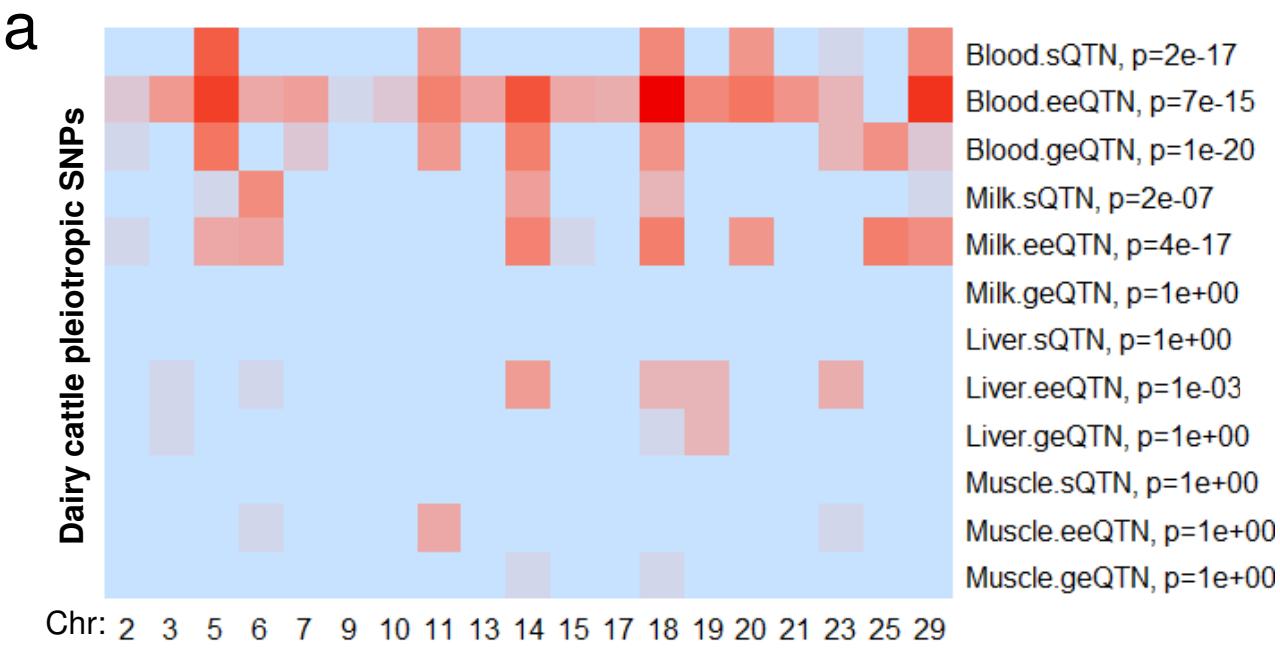

a
White blood cell cis splicing sQTN
b
Milk cell cis splicing sQTN
c
Liver cis splicing sQTN
d
Muscle cis splicing sQTN
$-\log_{10}(p)$

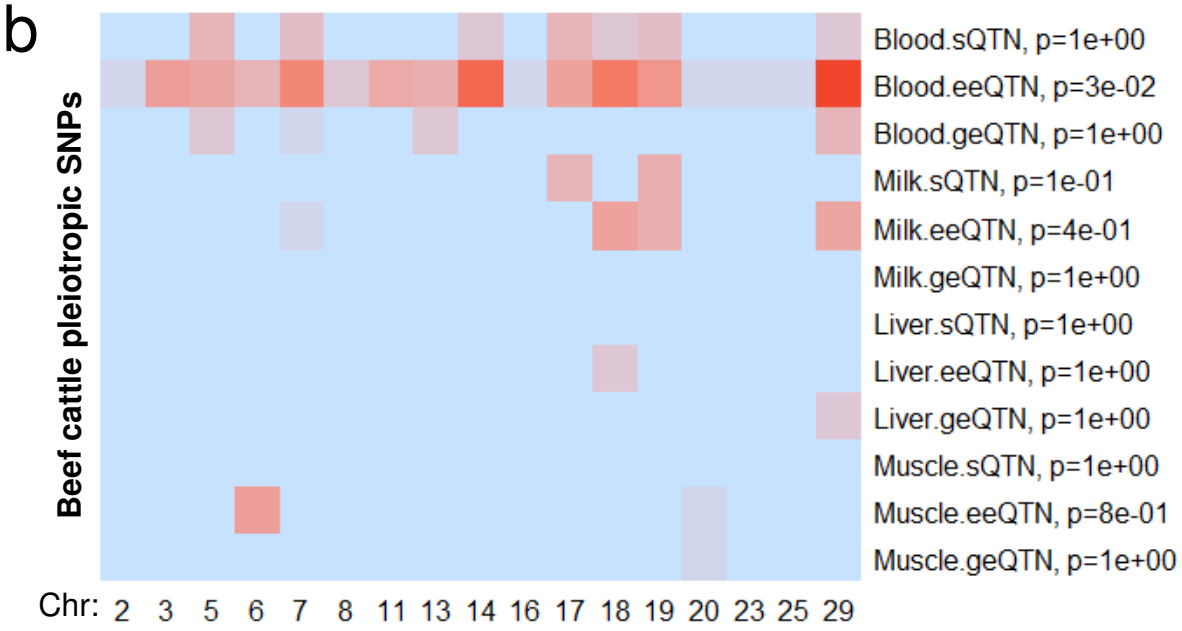

a
Distance between QTN and TSS (Mb)
sQTN
eeQTN
geQTN
White blood cells
Milk cells
Liver
Muscle
b
All SNP
sQTN
eeQTN
geQTN
SNP annotation
Other
Intergenic
Intron
Missense
Synonymous
Gene_end
UTR
Splice

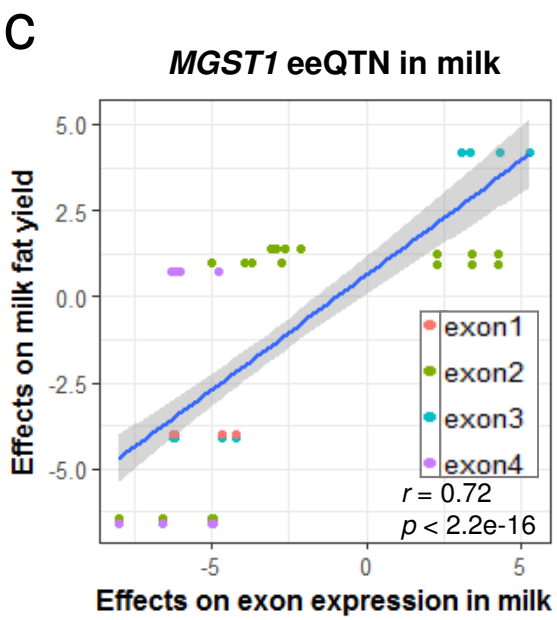

a
White blood cells
b
Milk cells
c
Liver
d
Muscle
$-\mathrm{Log}_{10}(P)$
1.3
160
320
geQTN
eeQTN
sQTN

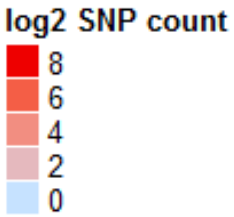

a
SNP
sQTN
Muscle (5783)
Liver (11544)
Milk (28907)
19
6
70
125
10
18005
Blood (138796)
Milk (28907)
Liver (11544)
eeQTN
Muscle (64675)
Liver (37322)
Milk (100844)
6309
2180
6916
2634
477
69711
Blood (802685)
Milk (100844)
Liver (37322)
geQTN
Muscle (57054)
Liver (39306)
Milk (4099)
403
35
5638
159
0
2619
Blood (96530)
Milk (4099)
Liver (39306)
-log10(p)
100
75
50
25
1.3
b
Exon or gene
Muscle (88)
Liver (81)
Milk (393)
2
2
10
6
2
118
Blood (1526)
Milk (393)
Liver (81)
Muscle (1649)
Liver (449)
Milk (3397)
264
96
58
78
23
934
Blood (14044)
Milk (3397)
Liver (449)
Muscle (1150)
Liver (419)
Milk (99)
81
8
64
24
3
30
Blood (842)
Milk (99)
Liver (419)
c
Genetic correlation
1
0
-1
DDX19B
CTSD
EEF1A1
HLA-DQA1
TTC4
MGST1
S100A10
SART1
-log2(p)
10
20
30
40
blood.milk.sQTL
liver.muscle.sQTL
blood.milk.eeQTL
liver.muscle.eeQTL
blood.milk.geQTL
IDs of exons or genes


a
sQTN
b
eeQTN
c
geQTN
Multi-transcriptome effects
Single-transcriptome effects
Multi-transcriptome effects
Single-transcriptome effects
Multi-transcriptome effects
Single-transcriptome effects


a
Dairy cattle pleiotropic SNPs
Blood.sQTN, p=2e-17
Blood.eeQTN, p=7e-15
Blood.geQTN, p=1e-20
Milk.sQTN, p=2e-07
Milk.eeQTN, p=4e-17
Milk.geQTN, p=1e+00
Liver.sQTN, p=1e+00
Liver.eeQTN, p=1e-03
Liver.geQTN, p=1e+00
Muscle.sQTN, p=1e+00
Muscle.eeQTN, p=1e+00
Muscle.geQTN, p=1e+00
Chr: 2 3 5 6 7 9 10 11 13 14 15 17 18 19 20 21 23 25 29


b
Beef cattle pleiotropic SNPs
Blood.sQTN, p=1e+00
Blood.eeQTN, p=3e-02
Blood.geQTN, p=1e+00
Milk.sQTN, p=1e-01
Milk.eeQTN, p=4e-01
Milk.geQTN, p=1e+00
Liver.sQTN, p=1e+00
Liver.eeQTN, p=1e+00
Liver.geQTN, p=1e+00
Muscle.sQTN, p=1e+00
Muscle.eeQTN, p=8e-01
Muscle.geQTN, p=1e+00
Chr: 2 3 5 6 7 8 11 13 14 16 17 18 19 20 23 25 29


c
MGST1 eeQTN in milk
Effects on milk fat yield
Effects on exon expression in milk
5.0
2.5
0.0
-2.5
-5.0
-5
0
5
exon1
exon2
exon3
exon4
r = 0.72
p < 2.2e-16


log2 SNP count
8
6
4
2
0