1 # Title page

2 ## Title:

3 A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy

4 cattle

5 ## Authors:

6 Young-Lim Lee[1]*, Haruko Takeda[2], Gabriel Costa Monteiro Moreira[2], Latifa Karim[3], Erik Mullaart[4], Wouter Coppieters[2,3], The

7 GplusE consortium[5], Ruth Appeltant[2], Roel F. Veerkamp[1], Martien A. M. Groenen[1], Michel Georges[2], Mirte Bosse[1], Tom

8 Druet[2], Aniek C. Bouwman[1], Carole Charlier[2]

9 ## Affiliations:

10 [1] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, the Netherlands

11 [2] Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Liège, Belgium

12 [3] GIGA Genomics Platform, GIGA Institute, University of Liège, Liège, Belgium

13 [4] CRV B.V., Arnhem, the Netherlands

14 [5] http://www.gpluse.eu/

15

16 * younglim.lee@wur.nl

17

Abstract

Clinical mastitis (CM) is an inflammatory disease occurring in the mammary glands of lactating cows. CM is under genetic control, and a prominent CM resistance QTL located on chromosome 6 was reported in various dairy cattle breeds. Nevertheless, the biological mechanism underpinning this QTL has been lacking. Herein, we mapped, fine-mapped, and discovered the putative causal variant underlying this CM resistance QTL in the Dutch dairy cattle population. We identified a ~12 kb multi-allelic copy number variant (CNV), that is in perfect linkage disequilibrium with a GWAS lead SNP, as a promising candidate variant. By implementing a genome-wide association study (GWAS) and through expression QTL mapping, we showed that the group-specific component gene (*GC*), a gene encoding a vitamin D binding protein, is an excellent candidate causal gene for the QTL. The multiplicated alleles are associated with increased *GC* expression and low CM resistance. Ample evidence from functional genomics data supports the presence of an enhancer within this CNV, which would exert *cis*-regulatory effect on *GC*. We observed that strong positive selection swept the region near the CNV, and haplotypes associated with the multiplicated allele were strongly selected for. Moreover, the multiplicated allele showed pleiotropic effects for increased milk yield and reduced fertility, hinting that a shared underlying biology for these effects may revolve around the vitamin D pathway. These findings together suggest a putative causal variant of a CM resistance QTL, where a *cis*-regulatory element located within a CNV can alter gene expression and affect multiple economically important traits.

Author summary

36   Clinical mastitis (CM) is an inflammatory disease that negatively influences dairy production and compromises animal welfare.

37   Although one major genetic locus for CM resistance was mapped on bovine chromosome 6, a mechanistic description of this

38   association has been lacking. Herein, we report a 12-kb multiallelic copy number variant (CNV), encompassing a strong

39   enhancer for group-specific component gene (*GC*), as a likely causal variant for this locus. This CNV is associated with high *GC*

40   expression and low CM resistance. We speculate that upregulation of *GC* leads to a large amount of vitamin D binding protein,

41   which in turn, reduces biologically available vitamin D, resulting in vitamin D deficiency and low CM resistance. Despite the

42   negative effect on CM resistance, the CNV contributes to increased milk production, hinting at balancing selection. Our results

43   highlight how multiplication of a regulatory element can shape economically important traits in dairy cattle, both in

44   favourable and unfavourable directions.

45

Introduction

Clinical mastitis (CM) is an inflammation in the mammary glands. This condition is often seen in dairy cattle and the repercussions of CM include production loss, use of antibiotics, and compromised animal welfare [1]. CM resistance has a genetic component, with estimated heritabilities ranging between 0.01 and 0.10 [2–6]. Genome-wide associations studies (GWASs) identified several quantitative trait loci (QTL) associated with CM resistance or somatic cell score (SCS), an indicator trait of CM. For instance, the six most significant CM resistance QTLs together capture 8.9% of the genetic variance in Danish Holstein Friesian (HF) cattle, underlining the polygenic nature of CM resistance [7]. Of these six QTLs, the most significant QTL mapped near 88 Mb on the Bos taurus autosome (BTA) 6 was repeatedly reported in various dairy cattle populations [8–12]. Several fine-mapping studies, using imputed whole genome sequence (WGS) variants, reported non-coding candidate causal SNPs at the group-specific component (*GC*) gene [7,12–15]. One of these studies investigated, albeit unsuccessfully, whether one of the non-coding candidate SNP obtained from the GWAS was associated with *GC* expression, leaving the functional mechanisms underlying this association elusive [13]. Interestingly, the proposed candidate SNPs showed antagonistic allele effects for milk yield (MY) (the high CM resistance allele was linked to low MY, and vice versa [7,13,16]), and one study concluded that a single pleiotropic variant regulates both of the traits [17]. Furthermore, this locus harbours QTL for many traits including body conformation, fertility, and longevity [12,15,18–22], implying pleiotropy, which remains to be investigated. Until now, *GC*, a gene that encodes the vitamin D binding protein (DBP), has been considered the most promising candidate gene for the CM resistance QTL on BTA 6 [13,14]. A growing body of literature underpins the importance of DBP, which acts as a macrophage activating factor, modulates immune responses [23], and is central to the vitamin D pathway [24]. For instance, polymorphisms in *GC* have been shown to cause vitamin D deficiency and inflammatory diseases in humans [25]. Moreover, a therapeutic use of vitamin D in lactating, CM-infected cows reduced inflammation, implying a link between vitamin D and inflammation [26,27]. Yet, the GWAS lead SNP was not associated with *GC* expression or alternative transcription, leaving the functional mechanism elusive [13]. Some researchers hypothesized that a copy number variant (CNV) might be the causal variant underlying this QTL [7]. Indeed, a CNV in high linkage disequilibrium (LD) with the GWAS lead SNP was found in the 3' alternative exon of *GC*, however, the functional role of this CNV was not well characterized [13].

In this study, we aimed at confirming the presence of the prominent CM resistance QTL in the Dutch HF population, and identifying a candidate causal gene and a variant that may explain the functional mechanism(s) of the QTL. Our findings strongly suggest that (1) the CM resistance QTL on BTA 6 is present in the Dutch HF population; (2) a 12-kb multi-allelic CNV, encompassing the 3' alternative exon of *GC,* harbours a putative enhancer, which exerts *cis*-regulation on the candidate causal gene, *GC*; (3) a haplotype associated with the CNV allele is strongly selected for, and the CNV has pleiotropic effects on MY, body conformation and fertility traits. These findings together highlight a functional CNV that contains a *cis*-regulatory element, affecting gene expression and subsequently altering the economically important traits.

4

Results

A major CM resistance QTL on BTA 6 segregates in the Dutch HF cattle population

A CM resistance QTL has been identified on BTA6 in several cattle populations [7–10,12–14]. We first aimed at confirming presence of this QTL in the Dutch HF population. All analyses were performed according to the Bovine genome assembly UMD3.1 [28], unless noted otherwise. To map the QTL, we performed a GWAS using 4,142 progeny tested bulls. These animals were genotyped using a custom 16K array, and imputed sequentially, firstly to the Illumina Bovine 50K array, and then to higher density (770K). CM resistance was recorded as a binary trait, depending on the disease status registration done by farmers [29]. The estimated breeding values of CM resistance were de-regressed and used as phenotypes in a single SNP GWAS performed using an additive model, and accounting for genetic relationships. The strong association signal near BTA 6:88.6 Mb found in Danish HF and Norwegian Red populations was replicated in the Dutch HF population ($-\log_{10}P=7.35$; Fig 1A, S1 Fig). This association signal was found downstream of *GC*, a reverse-oriented gene located at BTA 6:88.68-88.74 Mb. We fine-mapped this QTL, using 45,782 imputed WGS level variants present in a 10-Mb window encompassing the QTL (BTA 6:84-93 Mb), aiming at (1) confirming the *GC* as a positional candidate gene and (2) identifying a candidate causal variant. This 10-Mb window contained the previously reported CNV, which was in high LD with the GWAS lead SNP [13]. Thus, we kept SNPs located within the CNV in the WGS imputation panel, as they could possibly tag the CNV. The association signal peaked in a 200-Kb region (BTA 6:88.5-88.7 Mb), spanning over *GC* and the region downstream of *GC*. The lead SNP, rs110813063, located at BTA 6:88,683,517 ($-\log_{10}P=9.59$) was ~4 kb downstream of *GC* (Fig 1B). The T allele of the lead SNP (allele frequency (AF) = 0.58), was associated with low CM resistance, whereas the C allele was associated with high CM resistance (AF=0.42). Finally, a conditional analysis was performed by including the lead SNP as a covariate in the GWAS model. SNPs in the association peak were no longer significant, with the exception of a minor signal at the left side of the association peak (Fig 1C). Our results confirmed the presence of the CM resistance QTL in the Dutch HF population, however, as with previous studies [7,13–15], the non-coding lead SNP (rs110813063) did not have any evidence of functional role. There were no coding variants amongst the variant in high LD with the lead SNP ($r^2>0.9$). Provided that the strong association signals do not necessarily indicate causality, due to confounding factors including LD and limited sample sizes [30], we further characterized the lead SNP using WGS data.

A multi-allelic CNV is in high LD with the GWAS lead SNP for CM resistance QTL on BTA 6

By manually inspecting the WGS data, we observed that the GWAS lead SNP (rs110813063) was located within a ~12 kb CNV encompassing the 14th exon of *GC* (Fig 2A). This CNV, present in both dairy and beef cattle populations [31], was reported to be in high LD with the candidate SNP for CM resistance QTL in a Norwegian Red population [13]. Thus, we hypothesized that

5

107 the CNV might be the causal variant underlying this QTL. To confirm this hypothesis, we characterized the CNV using WGS

108 data of 266 HF animals. Our CNV calling pipeline, exploiting split-read, pair-end mapping, and read depth evidence, confirmed

109 the presence of the ~12 kb CNV at BTA 6:88,681,767-88,693,553 (hereafter referred to as GC CNV; S2 Fig). The 14th exon of

110 *GC*, encompassed by the CNV, is a 3' alternative exon, only accounting for minority of the total *GC* expression [13]. Sequenced

111 individuals presented either normal read depth (no CNV) or a highly inflated coverage, 2.5 to 5 times higher than expected

112 (CNV carriers; Fig 2B), implying multiple copies. To characterize whether different copy numbers (CN) were present at the *GC*

113 CNV locus, sequencing read depth information was used to obtain CN per individual (diploid CN). The distribution of this

114 diploid CN in the sequenced population suggested segregation of four different alleles corresponding to haploid CNs 1, 4, 5,

115 or 6 copies (Fig 2C). Based on these alleles and observed individual CN, we determined corresponding CN genotypes (e.g.,

116 individuals with CN 2 and 6 would be respectively carriers of alleles CN1/CN1 and CN1/CN5). The genotypes remained

117 ambiguous only for individuals with 10 copies, that could be either CN4/CN6 or CN5/CN5. In all sequenced duos or trios, these

118 inferred genotypes were compatible with Mendelian segregation rules (Fig 2D), and no genotype incompatibility was

119 observed. Genotypes from relatives of individuals with 10 copies allowed us also to deduce that all carriers of 10 copies were

120 CN4/CN6. Overall, these results suggest that the four alleles (CNs 1, 4, 5, and 6) are truly segregating in the population, rather

121 than the result of noise associated with depth estimation. Carriers of alleles CN5 or CN6 were restricted to a limited number

122 of families. In these families, the segregation of these two alleles perfectly matched haplotype transmission from parents to

123 offspring (S3 Fig), further supporting the presence of multiple alleles. An analysis of homozygosity-by-descent (HBD) in all

124 sequenced individuals revealed that alleles CN 4, 5 and 6 share a common haplotype, identical-by-descent for at least 200 kb

125 (≥600 SNPs; S1 Table). The HBD segments were rare at the CNV position in CN 1 individuals, presenting more haplotypic

126 diversity. Only two out of 88 heterozygous individuals were IBD in the studied position for an extremely short segment (12

127 kb, 93 SNPs), indicating that the CN 1 allele was associated with different haplotypes. In summary, we identified four alleles

128 at the GC CNV locus: CN 1 corresponds to a single copy and considered wildtype (Wt) given the high haplotypic diversity,

129 whereas alleles CNs 4-6 correspond to multiple copies (Mul), with four to six copies (Fig 3A). In our population, CN 1 and CN

130 4 were the most frequent alleles (0.39 and 0.54, respectively), while CN 5 and CN 6 were rare (0.03 and 0.05, respectively;

131 Fig 3A). Furthermore, the GC CNV, coded as a biallelic variant where CN 1 (Wt) was the reference allele and CNs 4-6 (Mul)

132 were grouped together as the alternative allele, was in perfect linkage (r^2=1) with the GWAS lead SNP (rs110813063). The T

133 allele, associated with low CM resistance, tagged CNs 4-6, whereas the C allele tagged CN 1. Thus, this SNP was used as a

134 surrogate marker for GC CNV in subsequent analyses. The GC CNV contained five tagging SNPs, including the GWAS lead SNP

135 (r^2≥0.98; Fig 3B). These five tagging SNPs showed an allelic imbalance pattern in WGS data of Wt/Mul individuals (Fig 3C),

136 due to a disproportionally high number of reads, supporting alternative alleles on the duplication haplotypes. There was no

137 SNP uniquely tagging CNs 4-6 separately, due to their similar and recent haplotypic background.

6

138

Recent positive selection strongly favoured a haplotype harbouring the GC CNV

140 In livestock breeding, artificial selection for economically important traits (i.e. high CM resistance) potentially drives desired

141 alleles to fixation and removes alleles with negative effects (i.e. low CM resistance) from the population. Thus, it is intriguing

142 that the allele associated with low CM resistance (CNs 4-6) is highly frequent in Dutch HF cattle (combined AF=0.58). We

143 postulated two alternative hypotheses: (1) GC CNV is pleiotropic, conferring positive effects on different traits under selection,

144 contrary to a negative effect on CM resistance, or (2) GC CNV is in high LD with a causal variant of a strongly selected trait,

145 and therefore, genetic hitch-hiking increased the frequency of the low CM resistance allele. In both cases, the GC CNV would

146 be associated with a selected haplotype. Thus, BTA 6 was scanned for signatures of selection based on integrated Haplotype

147 Score (iHS [32,33]), using WGS haplotypes from the 266 HF animals. Of the two strong signals of selection identified near the

148 79 and 89 Mb regions ($-\log_{10}P>5$; S4 Fig), the latter was only ~200 kb away from the GC CNV, and thus was further inspected

149 (Fig 4A). The extended haplotype homozygosity (EHH), centered at the iHS lead SNP (BTA 6:88,861,709, AF=0.28) revealed a

150 strongly favoured haplotype, which extended outwards further than the non-selected haplotypes (Fig 4B). As expected, CNs

151 4-6 were located in the strongly selected haplotype, whereas CN 1 was in the non-selected haplotypes. In addition, this finding

152 was in line with our HBD results, where homozygous CNV carriers (CNs 4-6) shared long HBD haplotypes, whereas

153 homozygous non-CNV carriers (CN 1) did not. These findings supported our hypothesis that a strong positive selection acted

154 upon the region containing the GC CNV. Given the antagonistic effects of the CM resistance QTL on MY [13] and relevance to

155 dairy cattle breeding [34], we deemed MY as a potential target of the selection signature we identified. To confirm whether

156 CM resistance and MY are modulated by the same variant (pleiotropy) or two different variants (LD), the 10-Mb window (BTA

157 6:84-93 Mb) harbouring the GC CNV was fine-mapped for MY. A strong association signal appeared in 89.08 Mb region, ~400

158 kb away from the GC CNV ($-\log_{10}P=7.5$; Fig 4A). The lead SNP was found at 89,077,838-bp and the G allele was associated

159 with high MY, whereas the T allele was associated with low MY (AF=0.45). Regardless of the ~400-Kb distance, the MY lead

160 SNP and the GC CNV were in high LD ($r^2 = 0.88$). Of note, the iHS lead SNP was located between the association signals for CM

161 resistance and MY (Fig 4A). We re-evaluated the EHH results and found that the strongly selected haplotype harbours CNs

162 4-6 alleles of the GC CNV and the G allele of the MY lead SNP, implying that the strong selection resulted in low CM resistance

163 and high MY (Fig 4B). We sought to disentangle these two QTL further, to elucidate whether they are in LD or pleiotropy.

164 However, due to a high LD and limitation in the data set (only 13 out of ~4,000 bulls carrying favourable recombinant

165 haplotypes), this was not possible with the current data.

166 Additionally, we fine-mapped other dairy cattle traits, in an attempt to identify other potential selection target trait(s). Our

167 results showed that the GC CNV was the lead variant for body condition score (BCS) and calving interval (CI; $-\log_{10}P=21.4$ and

168 6.6, respectively). Notably, GC CNV had a stronger association signal for BCS than it did for MY (S5 Fig). The CNs 4-6 allele,

169 which was associated with low CM resistance, was correlated with low BCS (meaning low body fat content) and longer CI

170 (meaning low fertility). The SNP association p-values obtained from either CM resistance and BCS or CM resistance and CI,

171 clearly colocalized, underlining that these QTL are driven by the same variant, which is likely to exert pleiotropic effects on

172 each of these traits (S4 Fig and S2 Table).

173

174 **GC is the most functionally relevant gene underlying the CM resistance QTL on BTA 6**

175 Our GWAS results hinted at transcriptional regulation as an underlying mechanism(s) of the CM resistance QTL, as our GWAS

176 lead SNP was found in non-coding region. Thus, we mapped cis-expression QTL (further referred to as eQTL), to (1) identify

177 shared variant(s) that are driving both GWAS and eQTL signals, and (2) to corroborate the causality of the candidate gene GC

178 in the major CM resistance QTL. Prior to eQTL mapping, we firstly determined the most biologically relevant tissue(s) for our

179 investigation. In the human transcriptome database [35–40], GC is predominantly expressed in the liver, whereas breast

180 tissue showed no expression (S3 Table). Also, previous dairy cattle transcriptome studies showed that GC is expressed in the

181 liver, kidney, and cortex, but not in the mammary gland [12,13,41]. Therefore, we performed eQTL mapping within the cis-

182 regulatory range (+/- 1Mb region from the GC CNV), using liver RNA-seq data and Bovine HD genotype of lactating HF cows

183 (n=175). Since GC CNV is not in the BovineHD array, the GC CNV genotypes were obtained by (1) imputing BovineHD

184 genotypes to WGS level variants and (2) genotyping the GC CNV directly. Direct genotyping was done by targeting six

185 polymorphic sites (CN 1 as reference allele and CNs 4-6 as alternative allele) in the GC CNV (S4 Table). The best probe

186 (BTA6:88,683,517) among the six showed 100% compatibility with the imputed GC CNV genotypes, underlining that our

187 imputation approach was robust. Hence, we used the imputed genotypes (BTA6:87-89 Mb) to map eQTL further. The RNA-

188 seq data was mapped using a reference guided method, where both reference annotated transcripts and novel transcripts

189 can be discovered. Our RNA-seq data detected two different GC transcript isoforms: a canonical and an alternative transcript

190 (Fig 5A). The former consisted of 13 exons and accounted for a majority of overall GC expression (~98%). The latter shared

191 the first 12 exons with the canonical transcript, however, it used an alternative 3' exon, the 14th exon, which is located inside

192 the GC CNV. Expression of the alternative transcript was relatively low (~2% of the total GC expression). The reference gene

193 set used for transcript assembly included the canonical form, but not the alternative transcript. Thus, GC gene-level eQTL

194 mapping was done on the canonical form, and we additionally mapped transcript-level eQTL for the alternative GC transcript.

195 Of the 13 genes annotated within the cis-regulatory range of CNV (+/- 1 Mb), GC was abundantly expressed (≥ 5,000

196 transcripts per million (TPM)) and other genes were either lowly expressed (0.1 < TPM < 50), or not expressed (Fig 5B). Our

197 gene-level eQTL mapping results discovered highly significant cis-eQTL for GC and $SLC4A4$. The GC cis-eQTL was found in the

198 BTA 6:88.68-88.89 Mb region ($-\log_{10}P > 23$; Fig 5D). The GC CNV was one of the top variants within the eQTL peak ($-\log_{10}P=24.4$)

199 and was in high LD with the GC eQTL lead SNP ($-\log_{10}P=25.4$; $r^2=0.88$). P-values for GC expression and CM resistance were

8

200   highly correlated (ρ=0.68; Fig 5E), where CNs 4-6 were associated with increased *GC* expression and low CM resistance (Fig

201   5F). The *SLC4A4 cis*-eQTL was found in BTA 6:88.67-89.07 Mb (-$\log_{10}$P > 7, S6 Fig). The lead SNP for *SLC4A4* eQTL (BTA

202   6:88,672,979, -$\log_{10}$P=7.75) was found ~9 kb away from GC CNV, which also showed high significance (-$\log_{10}$P=7.1). The GC

203   CNV and the lead SNP for *SLC4A4* eQTL were in high LD with GC CNV (r^2=0.99). P-values obtained from *SLC4A4* eQTL mapping

204   and CM resistance were highly correlated (ρ=0.82), and CNs 4-6 were associated with increased *SLC4A4* expression (S6 Fig).

205   Additionally, we detected a transcript-level eQTL for the alternative *GC* transcript. Intriguingly, the eQTL signal was driven by

206   GC CNV tagging SNPs, followed by the second most significant variant, GC CNV (-$\log_{10}$P=16.9 and 16.4, respectively; Fig 5G).

207   P-values for alternative *GC* transcript expression and CM resistance were correlated even stronger than the canonical

208   transcript (ρ=0.74; Fig 5H), where CNs 4-6 corresponded to an increased expression of the alternative *GC* transcript (Fig 5F).

209   Our results indicate *GC* as a promising candidate gene, given the strong eQTL signal. On the contrary, high LD between GC

210   CNV and the *SLC4A4* eQTL lead SNP (r^2 =0.99), implied that *SLC4A4* could be a candidate gene. To prioritize between the two

211   candidate genes *GC* and *SLC4A4*, we used summary data-based Mendelian randomization (SMR) analysis [42], which

212   estimates associations between phenotype and gene expression, aiming at identifying a functionally relevant gene,

213   underlying GWAS hits. Our result showed *GC* to be the only gene whose expression was significantly associated with CM

214   resistance association (-$\log_{10}$P=6), whereas *SLC4A4* was below the statistical threshold (-$\log_{10}$P=4.9; S5 Table). A subsequent

215   analysis, heterogeneity in dependent instruments (HEIDI), was conducted to test whether the significant association between

216   GWAS hit and eQTL shown for *GC* was induced by a single underlying variant or two variants that are in LD (i.e. GWAS hit

217   variant is in LD with eQTL lead SNP). The result suggested a single underlying variant modulating both CM resistance GWAS

218   and *GC* eQTL (P_{HEIDI} = 0.06). Thus, we confirmed *GC* as the most promising causal gene underlying CM resistance QTL, whose

219   expression affects the CM resistance phenotype.

220

221   A putative enhancer located in the *GC* CNV likely modulates the level of *GC* expression

222   We subsequently exploited epigenomic data sets to infer the functions of candidate variants in the 200-kb *GC* eQTL region

223   (BTA 6:88.68-88.89 Mb; Fig 6A). Bovine liver epigenomic data, interrogating two histone modifications (ChIP-seq for H3K27ac

224   and H3K4me3 marks) [43] and open chromatin regions (ATAC-seq) [manuscript in preparation] were investigated to infer

225   functional contexts (i.e. active promoters and enhancers). The ChIP-seq data predicted one active promoter and three active

226   enhancers, overlapping with ATAC-seq peaks, supporting the active status of the regulatory elements along *GC* (Fig 6A). Of

227   the three putative enhancers identified, *GC* harbours two putative enhancers: one located inside the GC CNV (further referred

228   to as GC CNV enhancer), and could be considered highly active, given the strong H3K27 acetylation mark. According to the

229   comparative genomics catalogue of regulatory elements in liver [43], the GC CNV enhancer is cattle-specific, whereas the

230   intronic enhancer is conserved between human, mouse, cow, and dog. These two putative enhancers were both supported

231   by corresponding ATAC-seq signals, with a stronger peaks for the GC CNV enhancer (Fig 6A). Additionally, this strong ATAC

232   signal overlapped with a repetitive element, MER115, from the DNA transposon family, hAT-Tip100 (Fig 6B). Some DNA

233   transposons obtain enhancer function during evolution [44], and hAT-Tip100 was shown to function as enhancer in humans

234   [43]. However, although the human orthologous region harbours MER115, this repeat did not show enhancer activity in

235   humans [40], underlining cattle-specific enhancer activity. Finally, we scanned the ATAC peak region inside the GC CNV,

236   searching for transcription factor binding (TFB) motifs using Homer [45]. We found strong evidence for five motifs, including

237   transcriptional enhancer factor (*TEAD*) and hepatocyte nuclear factor 4 alpha (*HNF4A*), supporting for the presence of the

238   active enhancer within the GC CNV (Fig 6C).

239

240

Discussion

In this study, we dissected a prominent CM resistance QTL in Dutch HF cattle, by integrating GWAS, eQTL, and functional (epi)genomics data. Our findings revealed that the GWAS lead variant is the GC CNV, a 12-kb multi-allelic CNV, which harbours a putative *cis*-regulatory element which targets *GC*. This CNV drives both the CM resistance association and the *GC* eQTL signals, underscoring *GC* as the likely causal gene underlying this QTL.

Identifying causal variants from GWAS can be challenging, since, often non-causal SNPs in high LD with the causal variant appear as lead SNPs [30]. Notably, our GWAS candidate SNP (rs110813063) has not been considered a strong candidate in other fine-mapping studies [7,13,14]. The function of our lead SNP, located 4-kb downstream of *GC*, was equally elusive, compared to other candidate SNPs that were located in the intronic region of or upstream of *GC* [7,13–15]. Nonetheless, the allelic imbalance pattern in our candidate SNP (Fig 3C), together with a previous report about a CNV in high LD with the GWAS candidate SNP [13], motivated us to speculate that the CNV might be the causal variant. As expected, we corroborated that our GWAS lead SNP, rs110813063, is a perfect tag SNP of the GC CNV (r^2=1). There are several explanations why previous fine-mapping studies missed rs110813063. Possibly, this SNP was absent in the GWAS variant set, as the standard SNP quality control (QC) criteria (i.e. removing SNPs with high depth and/or unbalanced allelic ratio) tend to eliminate SNPs inside CNVs. Alternatively, rs110813063 was present, but wrongly genotyped, due to highly disproportional allelic depth (Fig 3C). Hence, one may wonder whether a SNP-based GWAS approach, relying on a stringent QC, is sufficient in identifying causal variants, in a form other than point mutations. The answer might depend on the type of variant: studies showed that most deletions are well captured by tagging SNPs, whereas most duplications and multi-allelic CNVs are poorly tagged [46,47]. Thus, an exploratory check for presence of CNVs might be beneficial for fine-mapping studies.

To connect the discovery from statistical associations to the molecular function, we showed that the association mapping signal was driven by the underlying molecular signal, expression of *GC* (Fig 5). Many heritable diseases are manifested in a tissue-specific manner [48]. Our trait of interest, CM, is manifested in the mammary glands, and hence it was considered a biologically relevant tissue for eQTL mapping. However, both large-scale human transcriptome databases (S2 Table) and cattle transcriptome studies indicated liver as the main organ of *GC* expression [1–3]. This discrepancy shows that prior knowledge of tissue-specific manifestation of a trait is crucial for elucidating its molecular basis. In cattle, eQTL data was generated from diverse tissues (adrenal gland, blood, liver, mammary gland, milk, and muscle) [49–55], and some studies utilized the data sets in confirming causality of candidate genes [49–51,53]. We expect that the availability of eQTL data sets of diverse tissues and cell types will lead to rapid discovery and confirmation of candidate genes in farm animals in the future.

Bovine epigenomic data sets were utilized to prioritize candidate variants for the CM resistance QTL and eQTL for *GC* expression. The ChIP-seq and ATAC-seq data supported three active enhancers and one active promoter in the region of interest (Fig 6A). Of these regulatory elements, the GC CNV enhancer is considered the most likely causal variant, as multiple

11

272   copies of enhancers can increase the target gene expression [56]. Hence, we propose that altered *GC* expression, mediated

273   via multiplicated enhancers, is likely the key regulatory mechanism for this QTL. Interestingly, candidate causal variants

274   reported by previous studies [13,14] were found in the 1st intron of or upstream of *GC*, where an active enhancer and an

275   active promoter were found (Fig 6A). In humans, tissue-specific enhancers outnumber protein coding genes [40,57]. Hence,

276   a gene can be regulated by more than one enhancer, and a secondary enhancer is referred to as a shadow enhancer [58]. A

277   recent study showed that redundant enhancers function in an additive manner, thus conferring phenotypic robustness (e.g.

278   activities of multiple enhancers act together, thus removal of an enhancer still results in discernible phenotypes) [59]. In light

279   of this finding, we speculate that the two enhancers located in *GC*, might have additive effects on *GC* expression.

280   Mammalian enhancers evolve rapidly, compared to promoters [43]. Also, rapidly evolving enhancers were exapted from

281   ancestral DNA sequences and associated with positively selected genes [43]. The GC CNV enhancer is likely one of these

282   rapidly evolving enhancers, given that (1) it exapted MER115, which does not function as enhancer in other species, (2) it is

283   found in a selective sweep which harbors *GC*, and (3) it is cattle-specific. These findings strongly suggest that utilizing

284   epigenomic data of species other than the one of interest, can be misleading, as it lacks species-specific regulatory elements.

285   This provides a compelling reason to build species-specific epigenome maps, as already embarked upon by the international

286   Functional Annotation of Animal Genomes (FAANG) project [60,61]. With this community effort, a wider range of species-

287   specific regulatory elements underlying economically important QTL is expected to be unraveled in the future.

288   The major CM resistance QTL is known to have antagonistic effects for MY [13,14] and our results confirmed this. The trade-

289   off between CM resistance and MY suggests that this locus might be under balancing selection [62]. One of the drivers of

290   balancing selection is strong directional selection, which is common in livestock breeding [63]. Of the two traits of interest,

291   MY has been the primary goal of dairy cattle breeding [34], and hence it seems plausible to assume that this locus is under

292   balancing selection. Next to this, we had a particular interest in understanding the genetic basis of the antagonistic effects.

293   The genetic modes considered were (1) a single pleiotropic variant affecting two traits and (2) two independent causal

294   variants for each trait, in high LD. In case of pleiotropy, a particular genomic region cannot enhance both traits simultaneously

295   through breeding. However, in case of LD, selection for recombinant haplotypes, containing favourable alleles for both traits,

296   enables simultaneous improvement on the two traits. Only a small number (0.3%) of the studied animals had recombinant

297   haplotypes (13 out of 4,142 bulls), and hence our data set lacks sufficient power to distinguish LD from pleiotropy. Future

298   studies might consider two approaches for discerning this issue. The first is to exploit a daughter design by obtaining sufficient

299   daughters of the 13 recombinant haplotype carrier Dutch HF bulls [64]. Another possibility would be to harness data from

300   different breed(s). A dairy cattle breed that has both MY and CM resistance recorded, yet with low LD in the QTL region,

301   would be most useful. Otherwise, a meta-GWAS of multiple cattle populations can aid in distinguishing between LD and

302   pleiotropy.

303   We attempted to integrate the findings from the current study and further speculate on how *GC* expression and CM resistance

304   are linked, assuming that the level of *GC* expression and DBP is correlated; meaning absence of post-transcriptional

305   /translational regulation (Fig 7). DBP binds to and transports vitamin D [65], yet it plays additional roles in bone development,

306   fatty acid transport, actin scavenging, and modulating inflammatory responses (see [23,66] for review). To start with, DBP

307   modulates immunity as a macrophage activating factor (DBP-MAF), when deglycosylated via glycosidases of T- and B- cells

308   [67,68] and therefore enhances the immune response. Accordingly, we could hypothesize that CNV carriers have larger

309   amounts of DBP, and subsequently improved immunity, leading to higher CM resistance. However, this hypothesis contradicts

310   with our findings, as CNV carriers were shown to have lower CM resistance. Alternatively, DBP regulates the amount of freely

311   circulating vitamin D metabolites [69]. According to the free hormone hypothesis, only free vitamin D metabolites are able

312   to cross the cell membrane, and are thus biologically available [70]. In humans, only 0.03 % of 25(OH)D, an indicator of vitamin

313   D, is free, whereas the majority of vitamin D is bound either to DBP (85%) or to albumin (15%) [69]. Under these circumstances,

314   CNV carriers, having a large pool of DBP, can be hypothesized to have lower levels of biologically available vitamin D, as

315   postulated in human studies [71,72]. Thus, presumably, if CNV carriers have low amounts of free vitamin D, which may be

316   termed as 'vitamin D deficiency', low CM resistance in these animals seems like a logical consequence (Fig 7).

317   Although vitamin D has been considered crucial in bone health [24], a recent review showed an inverse correlation between

318   vitamin D concentrations and an extensive range of ill-health outcomes in humans [73]. Given the pervasive association

319   between vitamin D and health conditions, there may be other associated traits induced by vitamin D deficiency. Indeed, we

320   found strong association signals for BCS and CI, exerted by the GC CNV, implying its pleiotropic effects for multiple traits (S5

321   Fig). The CNs 4-6, which were associated with low CM resistance, were associated with longer CI, meaning poor fertility. CI, a

322   measure of duration from one calving to the next, is an indicator of fertility issues such as perturbations in the oestrous cycle,

323   and/or anovulation [74]. Human studies reported ovarian disfunction induced by vitamin D deficiency [75]. This finding

324   provides convincing evidence that the GC CNV might be the underlying variant inducing vitamin D deficiency and suboptimal

325   female fertility. Furthermore, pleiotropic effects of the GC CNV indicated that CNs 4-6 were associated with low CM resistance

326   and low BCS. This finding fits with the results found in HF cattle where low BCS was genetically correlated with high disease

327   incidence [76], although the causal relationship between CM resistance and BCS remains unknown. Intriguingly, human

328   studies showed contradictory results: obesity, a condition analogous to high BCS, predisposes patients to vitamin D deficiency,

329   leading to disease susceptibility [51]. These opposing consequences of body composition traits possibly hint at an 'optimum'

330   body fat amount, where an organism can function well without compromising its health.

Conclusions

332   In this study, we dissected the major CM resistance QTL on BTA 6, integrating GWAS, CNV calling, eQTL mapping, and

333   functional prioritization of candidate variants. We revealed a multi-allelic CNV harbouring a strong enhancer targeting *GC*, as

334 the likely causative variant. Our findings revealed that the candidate causal gene *GC* is likely regulated by an enhancer located

335 in the GC CNV. We speculate that GC CNV carriers which were shown to have high *GC* expression would have a larger amount

336 of DBP, and by extension low amount of biologically available vitamin D. This physiological condition probably puts animals

337 into a state analogous to vitamin D deficiency and leads to low CM resistance. Moreover, we report evidence of pleiotropic

338 effects of the GC CNV for other economically important traits as BCS, CI, and MY, which revolves around the vitamin D

339 pathway. The current study provides a novel example of multiplicated *cis*-regulatory elements playing pleiotropic roles on

340 various polygenic traits in dairy cattle.

341

342

14

Materials and methods

Whole genome sequencing and variant discovery

Whole genome sequence data

The genomes of 266 Dutch HF animals were sequenced. These 266 animals were closely related animals, where 240 were forming parents-offspring trios. The biological materials were either from sperm (males) or whole blood (females and males). Whole genome Illumina Nextera PCR free libraries were constructed (550bp insert size) following the protocols provided by the manufacturer. Illumina HiSeq 2000 instrument was used for sequencing, with a paired end protocol (2x100bp) by the GIGA Genomics platform (University of Liège). The data was aligned using BWA mem (version 0.7.9a-r786) [77] to the bovine reference genome UMD3.1 , and converted into bam files using SAMtools 1.9 [78]. Subsequently, the bam files were sorted and PCR duplicates were marked with Sambamba (version 0.4.6) [79]. All samples had minimum mean sequencing depth of 15X and the mean coverage of the bam files was 26X.

SNP calling and imputation panel construction

Variant calling was done using GATK Haplotype caller in N+1 mode. We applied Variant Quality Score Recalibration (VQSR) at truth sensitivity filter level of 97.5 to remove spurious variants. Using the trusted SNP and indel data sets which are explained elsewhere [80], we observed that the VQSR step filtered out GC CNV tagging SNPs, possibly due to the overwhelmingly high depth in this region. Yet, GC CNV tagging SNPs were considered crucial in our research, as we considered that they could tag the GC CNV. Therefore, the genotypes of the GC CNV tagging SNPs obtained from the raw variant calling format (VCF) file were inserted in the VQSR filtered VCF file. While inspecting the CNV tagging SNPs, we discovered genotyping errors (three errors among 5 tagging SNPs of 266 individuals), where Ref/Alt was wrongly genotyped as Alt/Alt, due to severe allelic imbalance (where the number of alternative allele supporting reads is predominantly high). Using parent-offspring relationships available in our data set, we confirmed that these were true errors, and hence they were manually corrected. Finally, the VCF file of sequence level variants in BTA 6:84-94Mb, containing manually corrected GC CNV tagging SNPs was obtained. This VCF file, consisting of 45,820 variants of 266 animals was used as an imputation panel for fine-mapping analyses. For analyses related to haplotype segregation among the 266 sequenced animals, haplotype sharing among carriers and identification of selection signatures, we further applied stringent variant filters on the VCF file as explained in [81] to conserve only highly confident variants and to remove spurious genotyping and map errors.

Genome-wide association studies

Phenotype and genotype data

We obtained genotype and phenotypic data of 4,142 progeny tested HF bulls from Dutch HF cattle breeding programme (CRV B.V., Arnhem, the Netherlands). The phenotype data was consisting of 60 traits, which are routinely collected in the breeding

15

373   programme, including clinical mastitis resistance (S6 Table). The estimated breeding values (EBVs) were de-regressed to

374   correct for the contribution of family members and de-regressed EBVs were used as phenotypes in GWAS. The effective

375   daughter contributions of the 4,142 bulls for CM resistance ranged between 25 and 971.3, with an average of 204.3. The

376   4,142 bulls were genotyped with a low density genotyping array (16K). Afterwards, the 16K genotype data was imputed in

377   two steps, firstly to 50K density, based on two private versions of a Bovine 50K genotyping array, and the panel was consisting

378   of 1,964 HF animals. It was further imputed to a higher density, using a panel of 1,347 HF animals genotyped with Illumina

379   BovineHD BeadChip (770K). Subsequently, the data was imputed to sequence level using the HF WGS imputation panel

380   described above. The imputation was done with Beagle 4 [82], and variants with low minor allele frequency (MAF < 0.025)

381   and low imputation accuracy (allele R^2 < 0.9) were filtered out.

382   **Association model and conditional analysis**

383   To confirm the presence of CM resistance QTL on BTA 6 in the Dutch HF population, we performed association mapping using

384   imputed high density genotypes (BovineHD (770K) reference panel). The association mapping was performed SNP-by-SNP

385   with a linear mixed model in GCTA [83]. The following model was fitted :

386    $$\mathbf{y} = \mathbf{1}\mu + \mathbf{Xb} + \mathbf{g} + \mathbf{e}$$

387   where **y** is the vector of phenotypes (de-regressed EBVs), **1** is a vector of ones, **μ** is the overall mean, **X** is a vector of SNP

388   genotypes coded as biallelic variant (0, 1, or 2), **b** is the additive effect of the SNP which is being tested for association, **g** is

389   the polygenic effect captured by a genomic relationship matrix (GRM; random effect), and **e** is the residual. The GRM was

390   built with GCTA, based on 50K genotypes to avoid losing statistical power by including causal markers [84]. The model

391   assumed equal residual variances. A SNP was regarded significantly associated with CM, when $-\log_{10}P$ value was above 6.45

392   (chromosome-wide Bonferroni multiple-testing correction for 28,669 tests), at a nominal significance of p = 0.01.

393   Subsequently, we repeated the association analysis in the BTA 6:84-94 Mb region, using the imputed sequence level variants

394   (n=45,820) to fine-map the QTL. The association model used was same as described above. Finally, to test if the lead SNP

395   explained the QTL signal completely we ran a conditional association analysis for the imputed sequence variants in BTA 6:84-

396   94 Mb region with the lead SNP as a covariate in the model.

CNV discovery and characterization of GC CNV

398   The CNV was called from the WGS data of 266 animals, using the Smoove pipeline (https://github.com/brentp/smoove). This

399   pipeline collects split and discordant reads using Samblaster [85], and then calls CNVs using Lumpy [86]. Afterwards, the CNV

400   sites were genotyped using SVTyper (https://github.com/hall-lab/svtyper). Additionally, we used Duphold [87] to calculate

401   read depth of the CNs at the GC CNV locus. Duphold exploits read depth between copy number variable regions and normal

402   regions and calculates the ratio between these two. The integer diploid CNs obtained based on the Duphold coverage ratio

403   values were assigned to the 266 animals. The CN distribution showed peaks at diploid CNs of 2, and 5-10, implying four

404 haploid CNs (1, 4, 5, and 6) segregating at the GC CNV locus. Thus, possible haplotypic combinations for the diploid CNs would

405 be: 2 (1/1), 5 (1/4), 6 (1/5), 7 (1/6), 8 (4/4), 9 (4/5), 10 (4/6 or 5/5). There was only one haplotypic combination possible for

406 all the diploid CNs, except 10, were either (4/6) or (5/5) could form diploid CN 10. We confirmed the true presence of these

407 CN alleles by taking advantage of our family structure. First, we verified that observed genotyped followed the Mendelian

408 segregation rules. Next, we also checked that CN alleles transmission within the pedigree was in agreement with haplotype

409 transmission. Haplotypes were reconstructed using familial information and linkage information using LINKPHASE3

410 programme [88]. The program estimates also, at each marker position and for each parent-offspring pair, the probability that

411 a progeny inherited the paternal or the maternal haplotype for its parents.

412 To study the relationship between different CN alleles, we estimated homozygous-by-descent (HBD) probabilities at the CNV

413 locus, and measured length of identified HBD segments. To that end, we ran with RZooRoH [89] a multiple HBD-class model

414 described in [90], with four HBD classes and one non-HBD class with rates equal to 10, 100, 1000, 10,000 and 10,000,

415 respectively. As this approach compares haplotypes within individuals, it does not require haplotype reconstruction and is

416 not affected by eventual phasing errors.

Selection signature analyses

418 The BTA 6 was scanned for haplotype based selection signatures observed in the sequence variants from the 266 animals.

419 We used the integrated haplotype homozygosity score (iHS [33]) using 'rehh' R package [32] for within-population analysis of

420 recent selection signatures. In order to unravel the selection target trait(s) we identified a number of traits within the

421 routinely collected catalogue of 60 traits showing QTL signals in the 84-94Mb region on BTA 6 based on 16K GWAS results

422 (EuroGenomics custom SNP chip [91]). Hence, we performed GWAS for these traits (BCS, CI, and MY) in BTA 6:84-94Mb region

423 based on imputed WGS variants to fine-map those QTL. The input files and the association model for the GWASs were the

424 same as described above, except that phenotypes used de-regressed EBVs of BCS, CI, and MY, respectively. The GWAS results

425 of these traits were plotted against the GWAS result of CM resistance to characterize the colocalization of QTL signals, using

426 the R package "LocusComparer" [92].

eQTL mapping and Summary data–based Mendelian randomization analysis

RNA-seq data and eQTL mapping

429 We used RNA-seq data produced by the GplusE consortium (http://www.gpluse.eu/; EBI ArrayExpress: E-MTAB-9348 and

430 9871)(Wathes et al; accepted but not online yet). Briefly, liver biopsy samples were collected from ~14 day post-partum HF

431 cows (n=178). The procedures had local ethical approval and complied with the relevant national and EU legislation under

432 the European Union Regulations 2012 (S.I. No. 543 of 2012). RNA-seq libraries were constructed using Illumina TruSeq

433 Stranded Total RNA Library Prep Ribo-Zero Gold kit (Illumina, San Diego, CA) and sequenced on Illumina NextSeq 500

434 sequencer with 75-nucleotide single-end reads to reach average 32 million reads per sample. The reads were aligned to the

17

435   bovine reference genome UMD3.1 and its corresponding gene coordinates from UCSC as a reference using HISAT2 [93].

436   Transcript assembly was conducted with StringTie [94], using reference-guided option for transcript assembly, which enables

437   discovery of novel transcripts that are not present in the reference gene set. Reads were counted at gene- or transcript-level

438   using StringTie. After data normalization using DESeq2 [95], we performed principal component (PC) analyses and removed

439   outliers (PC > 3.5 standard deviations from the mean for the top four PCs, n=2). Subsequently, the gene expression levels

440   were corrected with Probabilistic Estimation of Expression Residuals (PEER) [96]. All animals were genotyped using Illumina

441   BovineHD Genotyping BeadChip (770K). The genotype data was imputed to WGS level, using the imputation panel explained

442   above, using Beagle 4 [82] and variants with low minor allele frequency (MAF < 0.025) and low imputation accuracy (allele R^2

443   < 0.9) were filtered out. Finally, the PEER corrected normalized gene expression was associated with the imputed WGS

444   variants for 175 samples, using a linear model in R package "MatrixEQTL" [97].

445   **Prioritizing the causal gene**

446   We prioritized the most functionally relevant gene for the CM resistance QTL on BTA 6, using SMR (version 1.03) [42], to

447   estimate association between phenotype and gene expression. Input data required were summary statistics from CM

448   resistance GWAS and eQTL mapping results explained above. Additionally, the program requires plink format genotype data

449   to analyse LD in the region of interest, for which the imputed WGS variants (BTA 6:84-94 Mb) of 4,142 bulls were used. A

450   subsequent analysis, heterogeneity in dependent instruments (HEIDI), was conducted to test whether the significant

451   association between GWAS hit and eQTL shown for GC was induced by a single underlying variant or two variants that are in

452   LD (i.e. GWAS hit variant is in LD with eQTL lead SNP). The statistical thresholds for SMR ($-\log_{10}P>5$) and HEIDI (P > 0.05) were

453   benchmarked from the original paper [42].

454   **Functional genomics assay data**

455   The human transcriptome data bases were examined to find out in which tissue *GC* is highly expressed via Ensembl website

456   (release 101; [98]). We downloaded liver ChIP-seq data (H3K27ac and H3K4me3) generated from four bulls from ArrayExpress

457   (E-MTAB-2633 [43]). This ChIP-seq data was aligned to the bovine reference genome UMD3.1 using Bowtie2 [42] and peaks

458   were called using MACS2 [99]. A catalogue of mammalian regulatory element conservation

459   (https://www.ebi.ac.uk/research/flicek/publications/FOG15) was used to infer the conservation of the regulatory elements

460   predicted from the ChIP-seq data sets. Next, ATAC-seq data was explored to see if chromatin accessible regions coincided

461   with the histone marks obtained from the ChIP-seq data. We obtained a male calf's liver ATAC-seq data from the GplusE

462   consortium (ArrayExpress accession number: E-MTAB-9872). Data was analysed by following the ENCODE Kundaje lab ATAC-

463   seq pipeline (https://www.encodeproject.org/pipelines/ENCPL792NWO/). Sequences were trimmed using Trimmomatic [41]

464   and aligned on the bovine reference genome UMD3.1 using Bowtie2 [42]. After filtering out low quality, multiple mapped,

465   mitochondrial, and duplicated reads using SAMtools [43] and the Picard Toolkit (http://broadinstitute.github.io/picard/),

466 fragments with map length ≤ 146 bp were kept as nucleosome-free fraction. Genomic loci targeted by TDE1 were defined

467 as 38-bp regions centered either 4 (plus strand reads) or 5-bp (negative strand reads) downstream of the read's 5'-end. ATAC-

468 seq peaks were called using MACS2 [99] (narrowPeak with options --format BED, --nomodel, --keep-dup all, --qvalue 0.05, --

469 shift -19, --extsize 38). We inspected the presence of an enhancer in the human orthologous region of the GC CNV, using

470 ENCODE data [40] in the UCSC genome browser [100]. Transcription factor (TF) binding motifs in ATAC-seq peak regions were

471 discovered using Homer [45]. The TF motifs that are expressed in bovine or human liver are kept and shown in the figure

472 [101,102].

473

EuroGenomics custom array genotyping

475 We attempted to directly genotype the GC CNV in a biallelic mode (CN 1 as reference allele and CNs 4-6 as alternative allele).

476 Probes targeting six polymorphic sites were designed in Illunima DesignStudio Custom Assay Design Tool were added to

477 custom part of the EuroGenomics array [91] (S4 Table). DNA was extracted from the same biological material used for RNA-

478 seq used for eQTL mapping (described above), and genotyped for the EuroGenomics array by the GIGA Genomics platform

479 (University of Liège). The SNP genotypes were assessed using Illumina GenomeStudio software. Average call rate per probe

480 was calculated to assess the quality of the probes. Finally, genotypes obtained from the highest quality (BTA6:88,683,517)

481 was compared to the imputed GC CNV genotypes.

482

Acknowledgements

488

Author contribution

490 **Conceptualization**: Michel Georges, Carole Charlier, Aniek Bouwman

491 **Data Curation**: Wouter Coppieters, Latifa Karim, Gabriel Costa Monteiro Moreira, Haruko Takeda, Erik Mullaart, Ruth

492 Appeltant

493 **Formal analysis**: Young-Lim Lee, Tom Druet, Haruko Takeda

494 **Funding Acquisition**: Roel Veerkamp, Michel Georges

495 **Investigation**: Young-Lim Lee, Tom Druet, Haruko Takeda

496   **Methodology**: Aniek Bouwman, Tom Druet, Haruko Takeda

497   **Resources**: Erik Mullaart

498   **Supervision**: Carole Charlier, Mirte Bosse, Tom Druet, Aniek Bouwman, Michel Georges, Martien Groenen, Roel Veerkamp

499   **Visualization**: Young-Lim Lee

500   **Writing – original draft preparation**: Young-Lim Lee

501   **Writing – review & editing**: Young-Lim Lee, Carole Charlier, Mirte Bosse, Tom Druet, Aniek Bouwman, Michel Georges,

502   Martien Groenen, Roel Veerkamp, Wouter Coppieters, Latifa Karim, Gabriel Costa Monteiro Moreira, Haruko Takeda, Erik

503   Mullaart, Ruth Appeltant

504

505

References

1. Halasa T, Huijps K, Østerås O, Hogeveen H. Economic effects of bovine mastitis and mastitis management: A review. Vet Q. 2007;29(1):18–31.

2. Zwald NR, Weigel KA, Chang YM, Welper RD, Clay JS. Genetic Selection for Health Traits Using Producer-Recorded Data . I . Incidence Rates , Heritability Estimates , and Sire Breeding Values. J Dairy Sci [Internet]. 2004;87(12):4287–94. Available from: http://dx.doi.org/10.3168/jds.S0022-0302(04)73573-0

3. Bloemhof S, Jong G De, Haas Y De. Genetic parameters for clinical mastitis in the first three lactations of Dutch Holstein cattle. Vet Microbiol. 2009;134:165–71.

4. Negussie E, Lidauer M, Nielsen US, Aamand GP. Combining Test Day SCS with Clinical Mastitis and Udder Type Traits: A Random Regression Model for Joint Genetic Evaluation of Udder Health in Denmark, Finland and Sweden. In: Interbull Bulletin. 2010. p. 25–32.

5. Jamrozik J, Koeck A, Miglior F, Kistemaker GJ, Schenkel FS, Kelton DF, et al. Genetic and Genomic Evaluation of Mastitis Resistance in Canada. In: Interbull Bulletin. 2013. p. 43–51.

6. Pritchard T, Coffey M, Mrode R, Wall E. Genetic parameters for production, health, fertility and longevity traits in dairy cows. Animal. 2013;7(1):34–46.

7. Sahana G, Guldbrandtsen B, Thomsen B, Holm LE, Panitz F, Brøndum RF, et al. Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. J Dairy Sci [Internet]. 2014;97(11):7258–75. Available from: http://dx.doi.org/10.3168/jds.2014-8141

8. Abdel-Shafy H, Bortfeldt RH, Reissmann M, Brockmann GA. Short communication: Validation of somatic cell score-associated loci identified in a genome-wide association study in German Holstein cattle. J Dairy Sci [Internet]. 2014;97(4):2481–6. Available from: http://dx.doi.org/10.3168/jds.2013-7149

9. Sahana G, Guldbrandtsen B, Thomsen B, Lund MS. Confirmation and fine-mapping of clinical mastitis and somatic cell score QTL in Nordic Holstein cattle. Anim Genet. 2013;44(6):620–6.

10. Sodeland M, Kent MP, Olsen HG, Opsal MA, Svendsen M, Sehested E, et al. Quantitative trait loci for clinical mastitis on chromosomes 2, 6, 14 and 20 in Norwegian Red cattle. Anim Genet. 2011;42(5):457–65.

11. Veerkamp RF, Bouwman AC, Schrooten C, Calus MPL. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. Genet Sel Evol. 2016;48(1):1–14.

12. Freebern E, Santos DJA, Fang L, Jiang J, Parker Gaddis KL, Liu GE, et al. GWAS and fine-mapping of livability and six disease traits in Holstein cattle. BMC Genomics. 2020;21(1):1–11.

13. Olsen HG, Knutsen TM, Lewandowska-Sabat AM, Grove H, Nome T, Svendsen M, et al. Fine mapping of a QTL on bovine chromosome 6 using imputed full sequence data suggests a key role for the group-specific component (GC) gene in clinical mastitis and milk production. Genet Sel Evol. 2016;48(1):1–16.

14. Cai Z, Guldbrandtsen B, Lund MS, Sahana G. Prioritizing candidate genes post-GWAS using multiple sources of data for mastitis resistance in dairy cattle. BMC Genomics. 2018;19(1):1–11.

15. Tribout T, Croiseau P, Lefebvre R, Barbat A, Boussaha M, Fritz S, et al. Confirmed effects of candidate variants for milk production, udder health, and udder morphology in dairy cattle. Genet Sel Evol [Internet]. 2020;52(1):1–26. Available from: https://doi.org/10.1186/s12711-020-00575-1

16. Koivula M, Mäntysaari EA, Negussie E, Serenius T. Genetic and phenotypic relationships among milk yield and somatic cell count before and after clinical mastitis. J Dairy Sci. 2005;88(2):827–33.

17. Cai Z, Dusza M, Guldbrandtsen B, Lund MS, Sahana G. Distinguishing pleiotropy from linked QTL between milk production traits and mastitis resistance in Nordic Holstein cattle. Genet Sel Evol [Internet]. 2020;52(1):19. Available from: https://gsejournal.biomedcentral.com/articles/10.1186/s12711-020-00538-6

18. Jiang J, Ma L, Prakapenka D, VanRaden PM, Cole JB, Da Y. A large-scale genome-wide association study in U.S. Holstein cattle. Front Genet. 2019;10(MAY).

19. Abo-Ismail MK, Brito LF, Miller SP, Sargolzaei M, Grossi DA, Moore SS, et al. Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle. Genet Sel Evol. 2017;49(82):1–29.

20. Nayeri S, Sargolzaei M, Abo-Ismail M, Miller S, Schenkel F, Moore S, et al. Genome-wide association study for lactation persistency, female fertility, longevity, and lifetime profit index traits in Holstein dairy cattle. J Dairy Sci [Internet]. 2017;100(2):1246–58. Available from: http://dx.doi.org/10.3168/jds.2016-11770

21. Pausch H, Emmerling R, Schwarzenbacher H, Fries R. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. Genet Sel Evol. 2016;48(1):1–9.

22. Xiang R, van den Berg I, MacLeod IM, Daetwyler HD, Goddard ME. Effect direction meta-analysis of GWAS identifies extreme, prevalent and shared pleiotropy in a large mammal. Commun Biol [Internet]. 2020;3(1):1–14. Available from: http://dx.doi.org/10.1038/s42003-020-0823-6

23. Gomme PT, Bertolini J. Therapeutic potential of vitamin D-binding protein. TRENDS Biotechnol Biotechnol. 2004;22(7).

24. Horst RL, Reinhardt TA, Reddy GS. Vitamin D Metabolism. In: Feldman D, Pike JW, Adams JS, editors. Vitamin D. 2nd ed. 2005. p. 15–36.

573   25.   Jolliffe DA, Walton RT, Grif CJ, Martineau AR. Single nucleotide polymorphisms in the vitamin D
574   pathway associating with circulating concentrations of vitamin D metabolites and non-skeletal health
575   outcomes : Review of genetic association studies. J Steroid Biochem Mol Biol. 2016;164:18–29.
576   26.   Poindexter MB, Kweh MF, Zimpel R, Zuniga J, Lopera C, Zenobi MG, et al. Feeding supplemental 25-
577   hydroxyvitamin D 3 increases serum mineral concentrations and alters mammary immunity of lactating
578   dairy cows. J Dairy Sci. 2020;103:805–22.
579   27.   Merriman KE, Powell JL, Santos JEP, Nelson CD. Intramammary 25-hydroxyvitamin D3 treatment
580   modulates innate immune responses to endotoxin-induced mastitis. J Dairy Sci [Internet].
581   2018;101(8):7593–607. Available from: http://dx.doi.org/10.3168/jds.2017-14143
582   28.   Zimin A V., Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the
583   domestic cow, Bos taurus. Genome Biol. 2009;10(4).
584   29.   CRV. Breeding value Udder Health (Manual Quality, Chapter E-27) [Internet]. 2020 [cited 2020 Aug
585   25]. Available from: https://cooperatiecrv-be6.kxcdn.com/wp-content/uploads/2020/04/E_27-
586   Uiergezondheid-April-2020-Engels.pdf
587   30.   Gallagher MD, Chen-plotkin AS. The Post-GWAS Era : From Association to Function. Am J Hum Genet
588   [Internet]. 2018;102(5):717–30. Available from: https://doi.org/10.1016/j.ajhg.2018.04.002
589   31.   Kommadath A, Grant JR, Krivushin K, Butty AM, Baes CF, Carthy TR, et al. A large interactive visual
590   database of copy number variants discovered in taurine cattle. Gigascience. 2019;8(6):1–12.
591   32.   Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect
592   positive selection from haplotype structure. Mol Ecol Resour. 2017;17(1):78–90.
593   33.   Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome.
594   PLoS Biol [Internet]. 2006;4(3):e72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16494531
595   34.   Miglior F, Fleming A, Malchiodi F, Brito LF, Martin P, Baes CF. A 100-Year Review: Identification and
596   genetic selection of economically important traits in dairy cattle. J Dairy Sci [Internet].
597   2017;100(12):10251–71. Available from: http://dx.doi.org/10.3168/jds.2017-12968
598   35.   Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue
599   Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science.
600   2015;348(6235):648–60.
601   36.   Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene
602   expression levels in mammalian organs. Nature. 2011;478(7369):343–8.
603   37.   Lizio M, Abugessaisa I, Noguchi S, Kondo A, Hasegawa A, Hon CC, et al. Update of the FANTOM web
604   resource: Expansion to provide additional transcriptome atlases. Nucleic Acids Res. 2019;47(D1):D752–
605   8.
606   38.   Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human
607   tissue-specific expression by genome-wide integration of transcriptomics and antibody-based
608   proteomics. Mol Cell Proteomics. 2014;13(2):397–406.
609   39.   Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, et al. Expression Atlas: Gene
610   and protein expression across multiple studies and organisms. Nucleic Acids Res. 2018;46(D1):D246–
611   51.
612   40.   The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome.
613   Nature. 2012;489(7414):57–74.
614   41.   Fang L, Sahana G, Su G, Yu Y, Zhang S. Integrating Sequence-based GWAS and RNA-Seq Provides
615   Novel Insights into the Genetic Basis of Mastitis and Milk Production in Dairy Cattle. Sci Rep.
616   2017;7(45560):1–16.
617   42.   Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from
618   GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48(5).
619   43.   Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer evolution across 20
620   mammalian species. Cell [Internet]. 2015;160(3):554–66. Available from:
621   http://dx.doi.org/10.1016/j.cell.2015.01.006
622   44.   Cao Y, Chen G, Wu G, Zhang X, McDermott J, Chen X, et al. Widespread roles of enhancer-like
623   transposable elements in cell identity and long-range genomic interactions. Genome Res.
624   2019;29(1):40–52.
625   45.   Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-
626   Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell
627   Identities. Mol Cell [Internet]. 2010;38(4):576–89. Available from:
628   http://dx.doi.org/10.1016/j.molcel.2010.05.004
629   46.   Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of
630   structural variation in 2,504 human genomes. Nature. 2015;526(7571):75–81.
631   47.   Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic
632   copy number variations in humans. Nat Genet [Internet]. 2015;47(3):296–303. Available from:
633   http://dx.doi.org/10.1038/ng.3200
634   48.   Hekselman I, Yeger-Lotem E. Mechanisms of tissue and cell-type specificity in heritable traits
635   and diseases. Nat Rev Genet [Internet]. 2020; Available from:
636   http://www.ncbi.nlm.nih.gov/pubmed/31913361
637   49.   Littlejohn MD, Tiplady K, Lopdell T, Law TA, Scott A, Harland C, et al. Expression variants of the
638   lipogenic AGPAT6 gene affect diverse milk composition phenotypes in Bos taurus. PLoS One.
639   2014;9(1):1–12.
640   50.   Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, et al. Sequence-based Association
641   Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. Sci Rep.
642   2016;(May):1–14.

51. Kemper KE, Littlejohn MD, Lopdell T, Hayes BJ, Bennett LE, Williams RP, et al. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. BMC Genomics. 2016;17(1):1–9.

52. Brand B, Scheinhardt MO, Friedrich J, Zimmer D, Reinsch N, Ponsuksili S, et al. Adrenal cortex expression quantitative trait loci in a German Holstein × Charolais cross. BMC Genet. 2016;17(1):1–11.

53. Lopdell TJ, Tiplady K, Struchalin M, Johnson TJJ, Keehan M, Sherlock R, et al. DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. BMC Genomics. 2017;18(1):1–18.

54. Leal-Gutiérrez JD, Elzo MA, Mateescu RG. Identification of eQTLs and sQTLs associated with meat quality in beef. BMC Genomics. 2020;21(1):1–15.

55. Van Den Berg I, Hayes BJ, Chamberlain AJ, Goddard ME. Overlap between eQTL and QTL associated with production traits and fertility in dairy cattle. BMC Genomics. 2019;20(1):1–18.

56. Ngcungcu T, Oti M, Sitek JC, Haukanes BI, Linghu B, Bruccoleri R, et al. Duplicated Enhancer Region Increases Expression of CTSB and Segregates with Keratolytic Winter Erythema in South African and Norwegian Families. Am J Hum Genet [Internet]. 2017;100(5):737–50. Available from: http://dx.doi.org/10.1016/j.ajhg.2017.03.012

57. Long HK, Prescott SL, Wysocka J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell [Internet]. 2016;167(5):1170–87. Available from: http://dx.doi.org/10.1016/j.cell.2016.09.018

58. Scholes C, Biette KM, Harden TT, DePace AH. Signal Integration by Shadow Enhancers and Enhancer Duplications Varies across the Drosophila Embryo. Cell Rep [Internet]. 2019;26(9):2407-2418.e5. Available from: https://doi.org/10.1016/j.celrep.2019.01.115

59. Osterwalder M, Barozzi I, Tissières V, Fukuda-yuzawa Y, Mannion BJ, Afzal SY, et al. Enhancer redundancy provides phenotypic robustness in mammalian development. Nat Publ Gr. 2018;

60. Giuffra E, Tuggle CK. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annu Rev Anim Biosci. 2019;7:65–88.

61. The FAANG Consortium, Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 2015;16(1):4–9.

62. Hedrick PW. Heterozygote Advantage : The Effect of Artificial Selection in Livestock and Pets. J Hered. 2015;106(2):141–54.

63. Georges M, Charlier C, Hayes B. Harnessing genomic information for livestock improvement. Nat Rev Genet [Internet]. 2019;20(3):135–56. Available from: http://dx.doi.org/10.1038/s41576-018-0082-2

64. Georges M. Mapping , Fine Mapping , and Molecular Dissection of Quantitative Trait Loci in Domestic Animals. Annu Rev Genomics Hum Genet. 2007;8:131–62.

65. Daiger SP, Mel S, Cavalli-Sforza LL. Group-Specific Component (Gc) Proteins Bind Vitamin D and 25-Hydroxyvitamin D. Proc Nati Acad Sci. 1975;72(6):2076–80.

66. Bouillon R, Schuit F, Antonio L, Rastinejad F. Vitamin D Binding Protein: A Historic Overview. Front Endocrinol (Lausanne). 2020;10(January):1–21.

67. Yamamoto N, Kumashiro R. Conversion of vitamin D3 binding protein (group-specific component) to a macrophage activaating factor by the stepwise action of beta-galactosidase of B cells and sialidase of T cells. J Immunol. 1993;151:2794–802.

68. Swamy N, Dutta A, Ray R. Roles of the structure and orientation of ligands and ligand mimics inside the ligand-binding pocket of the vitamin D-binding protein. Biochemistry. 1997;36(24):7432–6.

69. Bikle DD, Schwartz J. Vitamin D binding protein, total and free Vitamin D levels in different physiological and pathophysiological conditions. Front Endocrinol (Lausanne). 2019;10(MAY):1–12.

70. Chun RF, Peercy BE, Orwoll ES, Nielson CM, Adams JS, Hewison M. Vitamin D and DBP: The free hormone hypothesis revisited. J Steroid Biochem Mol Biol [Internet]. 2014;144(PART A):132–7. Available from: http://dx.doi.org/10.1016/j.jsbmb.2013.09.012

71. Sinotte M, Diorio C, Berube S, Pollak M, Brisson J. Genetic polymorphisms of the vitamin D binding protein and plasma concentrations of 25-hydroxyvitamin D in premenopausal women. Am J Clin Nutr. 2009;25(3):634–40.

72. Lauridsen AL, Vestergaard P, Hermann AP, Brot C, Heickendorff L, Mosekilde L, et al. Plasma concentrations of 25-Hydroxy-Vitamin D and 1,25-Dihydroxy-Vitamin D are Related to the Phenotype of Gc (Vitamin D-Binding Protein): A Cross-sectional Study on 595 Early Postmenopausal Women. Calcif Tissue Int. 2005;25:15–22.

73. Autier P, Boniol M, Pizot C, Mullie P. Vitamin D status and ill health : a systematic review. Lancet Diabetes Endocrinol. 2014;2(January):76–89.

74. Santos JEP, Bisinotto RS, Ribeiro ES. Mechanisms underlying reduced fertility in anovular dairy cows. Theriogenology [Internet]. 2016;86(1):254–62. Available from: http://dx.doi.org/10.1016/j.theriogenology.2016.04.038

75. Irani M, Merhi Z, D M. Role of vitamin D in ovarian physiology and its implication in reproduction : a systematic review. Fertil Steril [Internet]. 2014;102(2):460-468.e3. Available from: http://dx.doi.org/10.1016/j.fertnstert.2014.04.046

76. Dechow CD, Rogers GW, Sander-Nielsen U, Klei L, Lawlor TJ, Clay JS, et al. Correlations among body condition scores from various sources, dairy form, and cow health from the United States and Denmark. J Dairy Sci [Internet]. 2004;87(10):3526–33. Available from: http://dx.doi.org/10.3168/jds.S0022-0302(04)73489-X

77. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr arXiv [Internet]. 2013;00(00):3. Available from: http://arxiv.org/abs/1303.3997

78. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

23

714   79.   Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba : fast processing of NGS alignment
715   formats. Bioinfo. 2015;31(February):2032–4.
716   80.   Kadri NK, Harland C, Faux P, Cambisano N, Karim L, Coppieters W, et al. Coding and noncoding
717   variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B affect recombination rate in cattle.
718   Genome Res. 2016;1323–32.
719   81.   Kadri N, Charlier C, Cambisano N, Deckers M, Mullaart E. High resolution mapping of cross-over events
720   in cattle using NGS data. In: Proceedings of the World Congress on Genetics Applied to Livestock
721   Production. 2018. p. 11.808.
722   82.   Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference
723   Panels. Am J Hum Genet [Internet]. 2018;103(3):338–48. Available from:
724   https://doi.org/10.1016/j.ajhg.2018.07.015
725   83.   Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. Am J
726   Hum Genet [Internet]. 2011;88(1):76–82. Available from:
727   http://dx.doi.org/10.1016/j.ajhg.2010.11.011
728   84.   Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of
729   mixed-model association methods. Nat Genet. 2014;46(2).
730   85.   Faust GG, Hall IM. SAMBLASTER : fast duplicate marking and structural variant read extraction.
731   Bioinformatics. 2014;30(17):2503–5.
732   86.   Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY : a probabilistic framework for structural variant
733   discovery. Genome Biol. 2014;15(R84):1–19.
734   87.   Pedersen BS, Quinlan AR. Duphold : scalable , depth-based annotation and curation of high-confidence
735   structural variant calls. Gigascience. 2019;(March):1–5.
736   88.   Druet T, Georges M. LINKPHASE3: An improved pedigree-based phasing algorithm robust to genotyping
737   and map errors. Bioinformatics. 2015;31(10):1677–9.
738   89.   Bertrand AR, Flori L, Druet T. RZooRoH : An R package to characterize individual genomic autozygosity
739   and identify homozygous-by-descent segments. Methods Ecol Evol. 2019;2019(10):860–6.
740   90.   Druet T, Gautier M. A model-based approach to characterize individual inbreeding at both global and
741   local genomic scales. Mol Ecol. 2017;26:5820–41.
742   91.   Boichard D, Boussaha M, Capitan A, Rocha D, Sanchez MP, Tribout T, et al. Experience from large scale
743   use of the EuroGenomics custom SNP chip in cattle. In: 11th World Congress on Genetics Applied to
744   Livestock Production. 2018. p. 1–6.
745   92.   Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene
746   expression complicate GWAS follow-up. Nat Genet [Internet]. 2019;51(May). Available from:
747   http://dx.doi.org/10.1038/s41588-019-0404-0
748   93.   Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with
749   HISAT2 and HISAT-genotype. Nat Biotechnol [Internet]. 2019;37(August). Available from:
750   http://dx.doi.org/10.1038/s41587-019-0201-4
751   94.   Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved
752   reconstruction of a transcriptome from RNA-seq reads. Nat Biotech. 2016;33(3):290–5.
753   95.   Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with
754   DESeq2. Genome Biol. 2014;15(12):1–21.
755   96.   Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals
756   (PEER) to obtain increased power and interpretability of gene expression analyses. 2012;7(3):500–7.
757   97.   Shabalin AA. Matrix eQTL : ultra fast eQTL analysis via large matrix operations. Bioinformatics.
758   2012;28(10):1353–8.
759   98.   Hunt SE, Mclaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, et al. Ensembl variation
760   resources. 2018;(8):1–12.
761   99.   Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-
762   Seq (MACS). Genome Biol. 2008;(9):R137.
763   100. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser
764   at UCSC. Genome Res. 2002;12(6):996–1006.
765   101. de Souza MM, Zerlotini A, Geistlinger L, Tizioto PC, Taylor JF, Rocha MIP, et al. A comprehensive
766   manually-curated compendium of bovine transcription factors. Sci Rep [Internet]. 2018;8(1):1–12.
767   Available from: http://dx.doi.org/10.1038/s41598-018-32146-2
768   102. Liu X, Yu X, Zack DJ, Zhu H, Qian J. TiGER: A database for tissue-specific gene expression and
769   regulation. BMC Bioinformatics. 2008;9:1–7.
770   103. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps.
771   Bioinformatics. 2005;21(2):263–5.
772

Figures

Figure 1. Association mapping of clinical mastitis resistance on BTA6

(A) Association mapping performed with imputed BovineHD variants on BTA6. The association signal near BTA 6:88.6 Mb, shown in other dairy cattle populations was replicated in the current Dutch HF population. (B) Association mapping performed with imputed WGS variants in BTA 6:84-93 Mb region. A strong association signal was shown in a 200 Kb region (BTA 6:88.5-88.7 Mb), spanning over *GC* gene. Our lead SNP (rs110813063, marked with green and vertical dotted line) has not been reported as a candidate SNP in other CM fine-mapping studies. CM candidate SNPs from other fine-mapping studies are marked as yellow. (C) Conditional analyses including GC CNV as a covariate nullify the association signal.

Figure 2. Discovery of multiallelic GC duplication using deeply sequenced genomes and familial structure

(A) Schematic overview showing the GWAS lead SNP and ~12 kb CNV overlapping with *GC*. *GC* is a reverse oriented gene, consisting of 14 exons, of which two last exons are non-coding. Five CNV tagging SNPs were present within the GC CNV and marked with black asterisks (the middle asterisk covers three tagging SNPs). Among them, the first SNP, which was also the GWAS lead SNP, was in perfect LD with the GC CNV ($r^2 = 1$), whereas the rest were in high LD ($r^2 > 0.98$). The hash marks at the upstream and the intronic region of *GC* indicate CM resistance candidate SNPs reported by others [13,14]. (B) Sequencing depth difference between the CNV region and normal region was used to infer copy numbers. (C) A Histogram of read depth values shows that majority of animals fall into diploid copy number of 2, 5 and 8, and some minor peaks occur at diploid copy number of 6, 7, 9 and 10. Based on this diploid CNs, we inferred haploid CNs of 1, 4, 5, and 6. We showed possible allelic combination(s) above each diploid CN. The diploid CN10 could be comprised of CN5/CN5 and CN4/CN6, however, our results showed that it was always CN4/CN6. (D) Familial information and background haplotypes were used to phase the copy number and thus revealed how the CNV segregates in trios. The upper family tree shown with animal signs stands for diploid copy numbers, whereas the lower tree shows haploid copy numbers (the phase results of the diploid CNs).

Figure 3. Characterization of the GC CNV tagging SNPs and allelic imbalance pattern

(A) A schematic overview of four structural haplotypes and the five tagging SNPs inside the GC CNV, shown together with allele frequencies. (B) Positions, rs ID, alleles, location within GC of GC CNV tagging SNPs. (C) Allelic imbalance pattern shown in Wt/Dup animals. Animals will get more supporting reads for alternative alleles for the five CNV tagging SNPs, thus the tagging SNPs will be called as heterozygous but with high allelic imbalance

Figure 4. Selection signature scan and trait association (clinical mastitis resistance and milk yield) GWAS plot

(A) A 10-Mb region with a strong selective signature signal was zoomed in (BTA 6: 84-93 Mb). Association mapping results from imputed WGS variants on CM resistance (dark blue) and MY (yellow) are shown in the upper panel; iHS results are shown in the lower panel. The CM resistance GWAS peak occurs at the left side of the iHS peak, whereas MY GWAS peak appears on the right side of the iHS peak. The red vertical line marks GC CNV. A 1-Mb region covering GC CNV, iHS lead SNP, and MY lead SNP are marked with translucent blue. (B) The extended haplotype homozygosity of the 1-Mb region marked in panel (A) is shown, together with four genes annotated in this region (top of the figure). The major haplotype shown in the upper part (black) branches outwards, implying recent positive selection acted upon this haplotype. The non-selected haplotype, shown in the lower side (blue) rapidly breaks down from the iHS lead SNP. (C) Pairwise D' and r^2 values between GC CNV, iHS lead SNP, and MY lead SNP in the ~4,000 daughter proven bulls. A screenshot made from Haploview software [103].

Figure 5. eQTL mapping and GWAS-eQTL colocalization results for *GC* and the non-coding RNA

(A) A Schematic overview of the *GC* gene structure and position of the GC CNV. Our data detected two *GC* transcripts, where the canonical form account the majority of the expression (98%) and an alternative form only counting for minor expression (2%). (B) eQTL was mapped for the genes located in a 2-Mb bin (BTA6:87.68-89.68). Of the 13 genes annotated in this bin, *GC* showed predominantly high expression (5,000 TPM <), whereas the rest were lowly expressed or not expressed at all. The eQTL were mapped for *GC* and *SLC4A4*. (C) CM resistance GWAS results were shown for the 2-Mb bin, where eQTL was mapped. The color scale indicates the degree of pair-wise LD (r^2) between the GC CNV and other SNPs. Annotation of genes in this region is drawn as black bars. Six genes on the left part are *AMBN, JCHAIN, RUFY3, GRSF1, MOB1B*, and *DCK*. (D) eQTL mapping results for *GC* (canonical transcript). (E) P-values obtained from CM resistance GWAS and GC (canonical transcript) eQTL mapping were correlated. The GC CNV is located in the right upper corner ($\rho=0.68$), showing that it is significant for both GWAS and eQTL mapping. (F) The box plot shows altered *GC* (canonical transcript) expression depending on GC CNV genotypes. (G) eQTL mapping result for GC (alternative transcript). (H) P-values obtained from CM resistance GWAS and GC (alternative transcript) eQTL mapping were correlated. The GC CNV is located in the right upper corner ($\rho=0.74$), showing that it is significant for both GWAS and eQTL mapping. (I) The box plot shows altered *GC* (alternative transcript) expression depending on GC CNV genotypes. Panels C-E, G, H were made with LocusCompare programme [92]

Figure 6. Inspection of functional elements near the GC CNV

Functional elements were inspected in *GC* eQTL region using ChIP-seq (H3K27ac and H3K4me3) and ATAC-seq data. (A) The *GC* eQTL region was zoomed in. In this region, *GC* is the only annotated gene. The GC CNV is marked with the translucent blue, and CM resistance candidate SNPs reported by other studies [13,14] are marked with translucent yellow. Other significant eQTL lead SNPs in this region are marked with translucent pink. We overlaid ChIP-seq data to identify putative enhancers and promoters (ChIP-seq tracks; red). Furthermore, liver ATAC-seq data revealed highly accessible chromatin regions, supporting the regulatory elements discovered by ChIP-seq data sets (ATAC-seq tracks; blue). (B) We further zoomed in to the ATAC peak within the GC CNV, and

25

838   discovered that the ATAC peak overlaps with MER 115. The predicted hepatic transcription factor binding sites are marked with
839   translucent grey. (C) Transcription factor binding motifs are shown together with the ATAC signal located inside the GC CNV.
840
841   **Figure 7. Summary of the key findings and hypothesis of physiological aspects linking *GC* expression and CM resistance**
842   A schematic overview summarizing the allele effects of wildtype (CN 1) and multiplicated (CNs 4-6) alleles of the GC CNV, a likely
843   causal variant for the major CM resistance QTL. The two alleles at the GC CNV locus lead to altered *GC* transcription, where the
844   multiplicated alleles correspond to high *GC* expression. On the bottom shows the phenotypic association between the GC CNV and
845   CM resistance, where the multiplicated allele is associated with low CM resistance. Finally, the area marked with grey shade shows
846   our hypotheses that the amount of DBP is positively related with the GC expression. Further, we speculated that the amount of DBP
847   and free vitamin D is inversely correlated, as long as vitamin D is bound by DBP, it is not biologically available. The solid arrows
848   indicate the relations based on our findings. The dotted arrows indicate the relations based on our speculation.
849
850   **Fig 1**
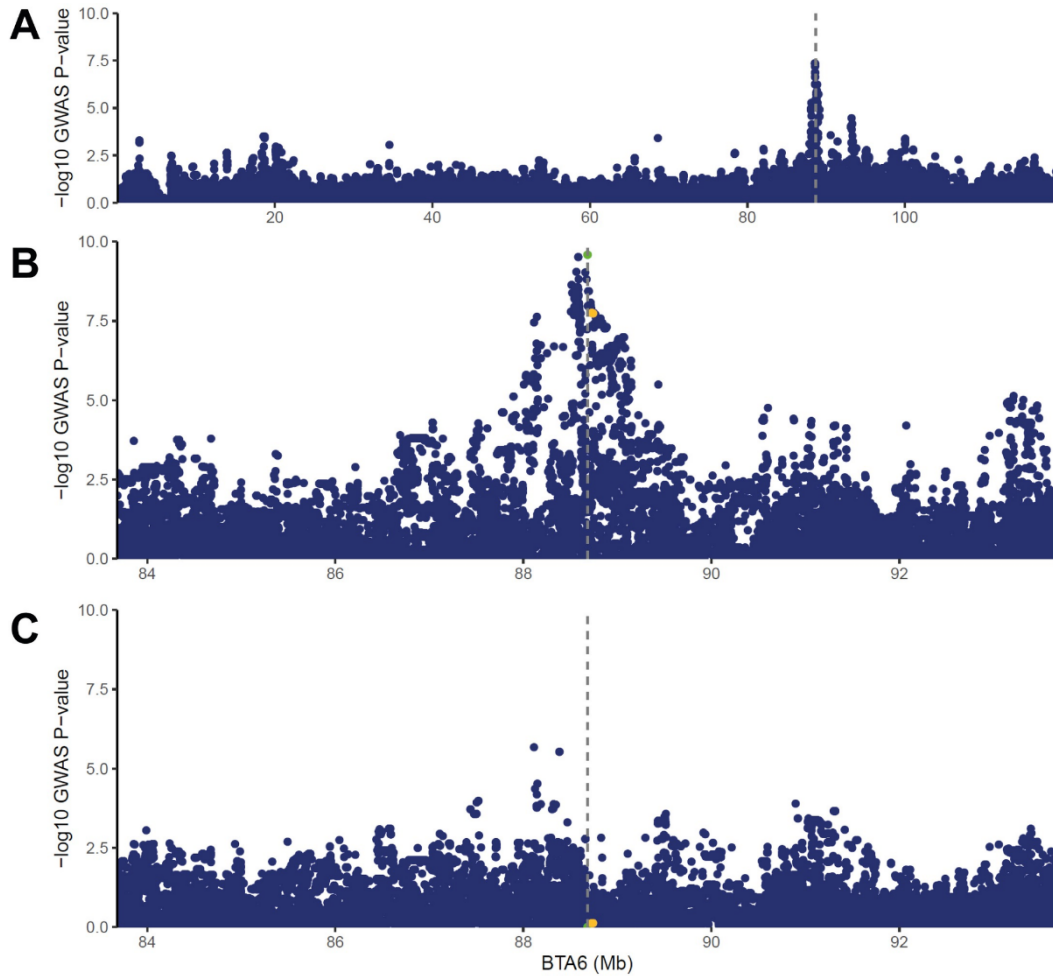
851
852
853   **Fig 2**
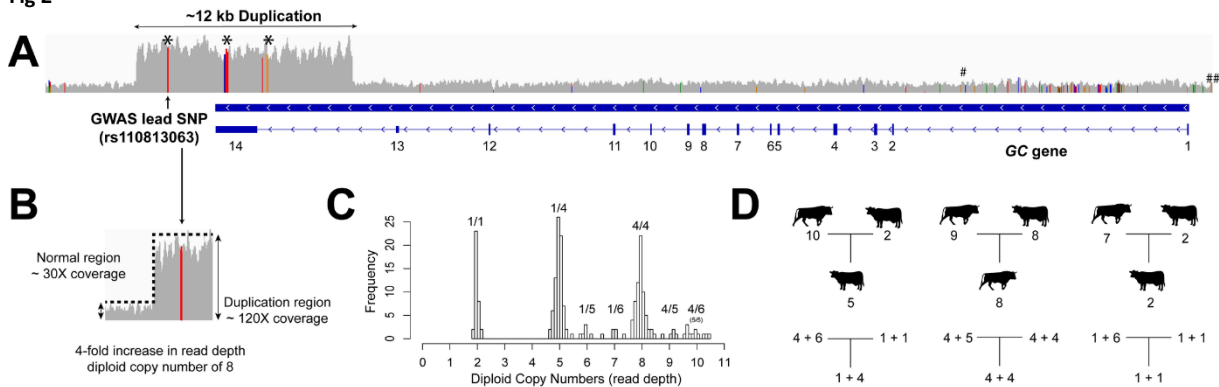
854
855
856   **Fig 3**
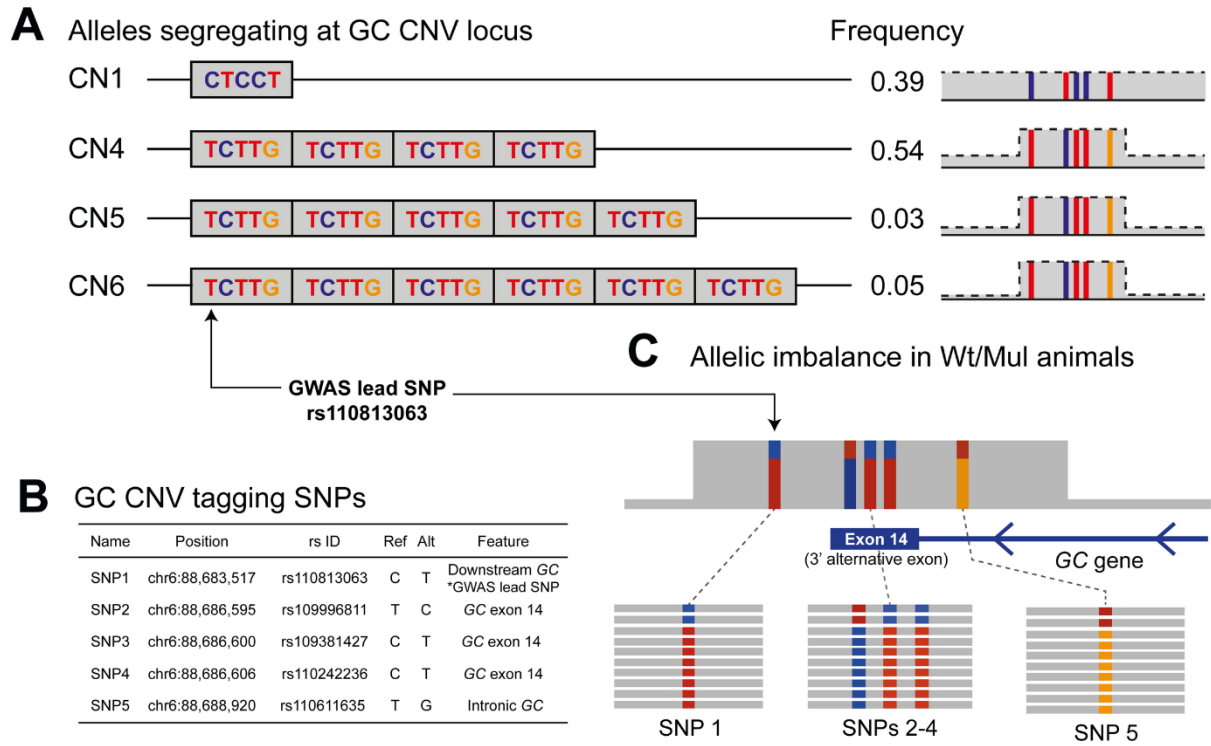
857
858
859 **Fig 4**

860
861
862 **Fig 5**

27

Fig 6

Fig 7

28

869
870

29
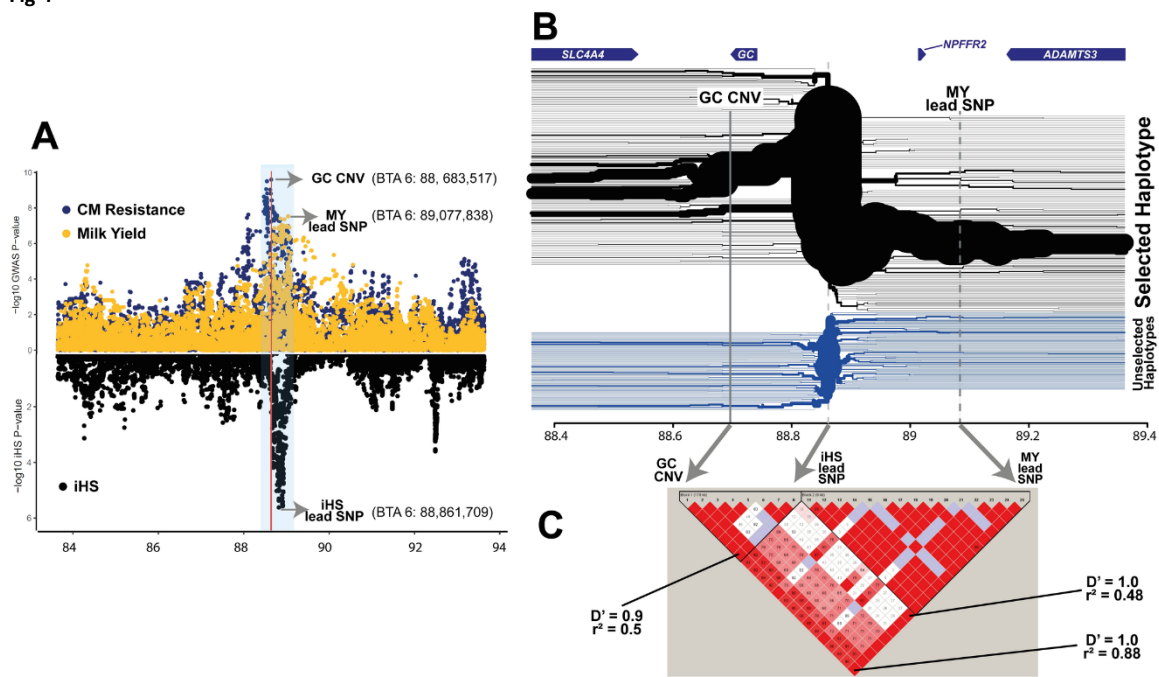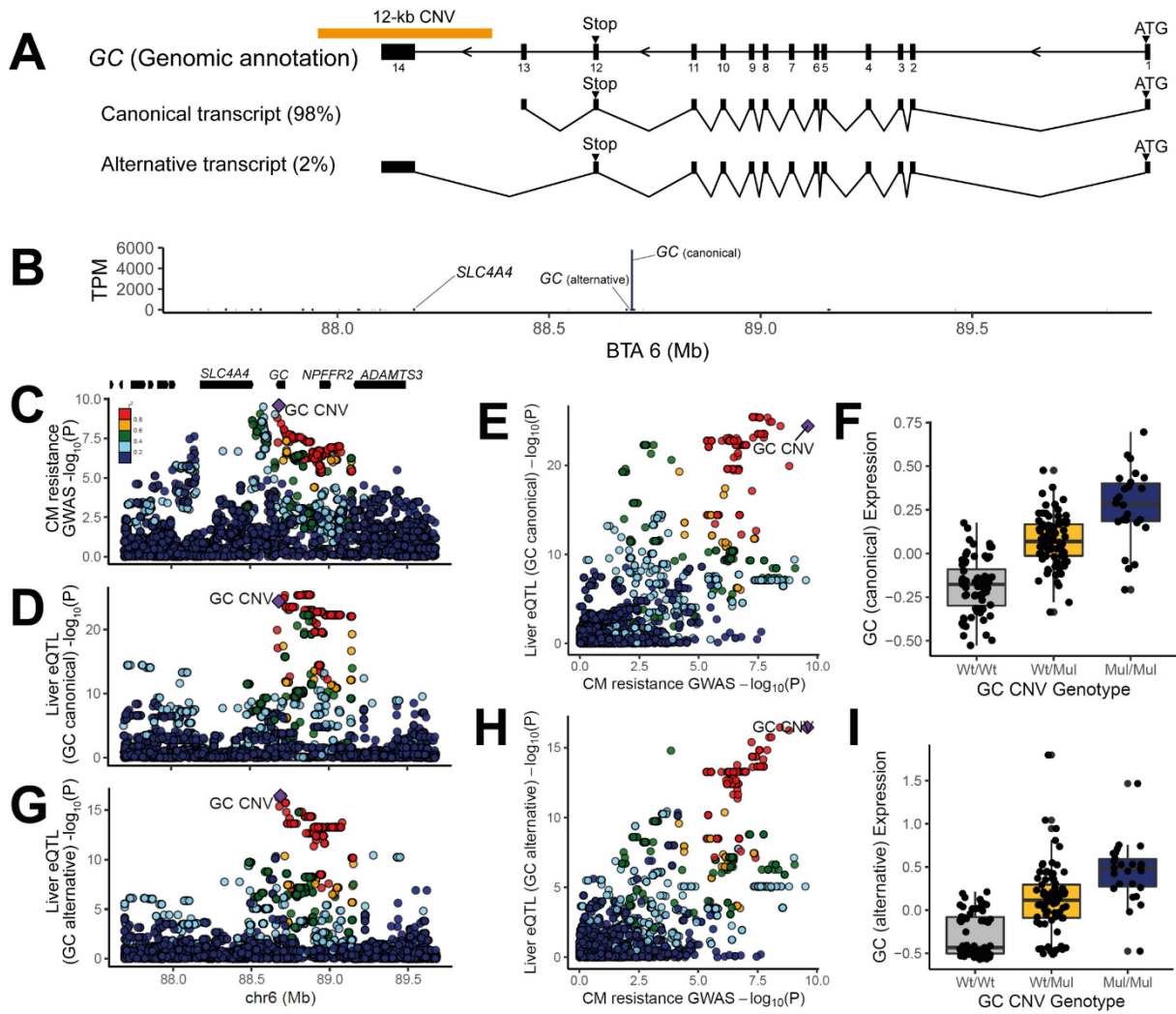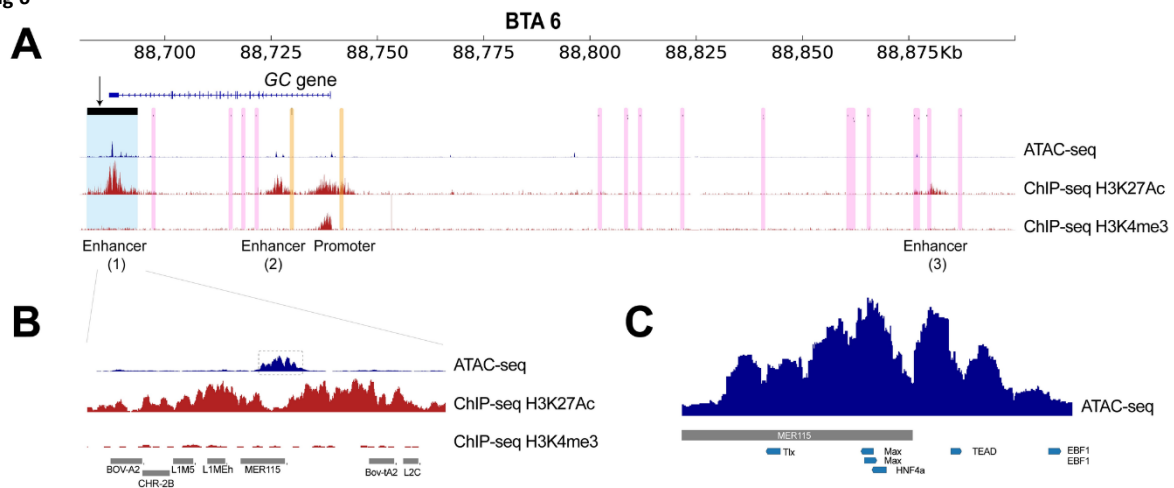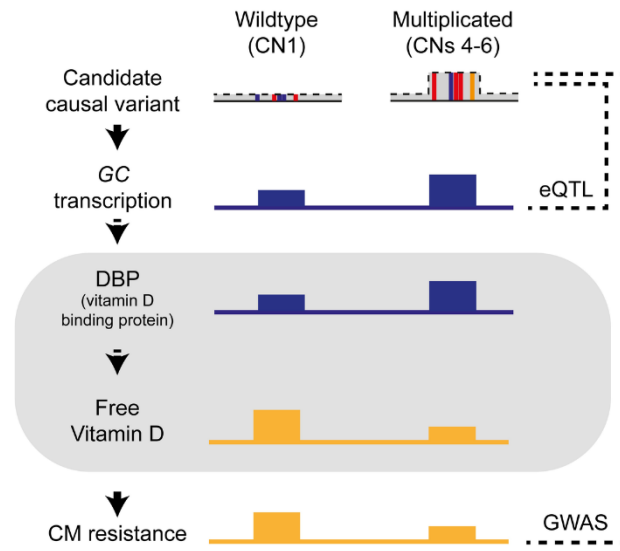
Supporting information

S1 Figure. Manhattan plot for CM resistance GWAS results
Genome-wide association mapping of the CM resistance trait, performed using imputed BovineHD data

S2 Figure. IGV screen shots for the GC CNV and its' breakpoints
(A) The GC CNV (BTA 6:88,681,767-88,693,553) is shown in the IGV screen shot. The grey reads are normally mapped reads, whereas green ones are discordantly mapped reads, providing evidence for a tandem duplication. The sequencing coverage of the CNV region is higher than non-CNV region. (B) The left breakpoint is flanked by MIR repeat. (C) The right breakpoint does not overlap with repeats. (D and E) The left and right breakpoints are zoomed in and the soft-clipped reads (positions where nucleotide sequences are written) information revealed the 5-bp microhomology "CACAT" (marked as yellow) at the breakpoints.

S3 Figure. Family tree of animals having CN 5 and CN6 allele on GC CNV locus
In the panel of 266 animals, we observed CN 5 and CN6 alleles are mostly segregating among a small number of related animals. To show that CN5 and CN6 alleles are truly segregating, transmission probabilities at each marker position were calculated with LINKPHASE3. Transmission probabilities were estimated for paternal haplotypes (1 and 0 indicated transmission of the paternal and maternal allele, respectively). (A) Largest family of CN5 carriers. Each circle indicates one individual and the number inside the circle indicates the GC CNV copy number genotype. The numbers above each individual stand for the transmission probability. Questions marks mean individuals with no copy number information. Male animals are shown as squares, female animals are shown as circles, and animals with unknown gender are marked with diamonds. (B) Largest family of CN6 carriers. The legends are identical to panel (A).

S4 Figure. Chromosome-wide scan of selection signatures using integrated Haplotype Score (iHS) on BTA6
Chromosome-wide scan of integrated Haplotype Score revealed two iHS peaks with high significance ($-\log_{10}P > 5$), near BTA6:78 Mb and BTA6:89 Mb.

S5 Figure. Three traits that showed strong association signals in CM resistance QTL region on BTA 6
Colocalization of fine-mapping p-values of a pair of traits are showing that body condition score (BCS), calving interval (CI), and milk yield (MY) are having association signal near or at the CM resistance QTL region on BTA 6. (A) Colocalization between body condition score and CM resistance is shown in the left panel, together with the separate Manhattan plots for body condition score and CM resistance on the right. (B) Colocalization between calving interval and CM resistance. (C) Colocalization between milk yield and CM resistance. Panel layout of (B) and (C) is the same as panel (A). In each panel, the colours of dots indicate degree of LD (r^2) with GC CNV (colour scale shown in the left upper corner of panel A). The purple diamond marks GC CNV. The figure was made with LocusCompare programme [92].

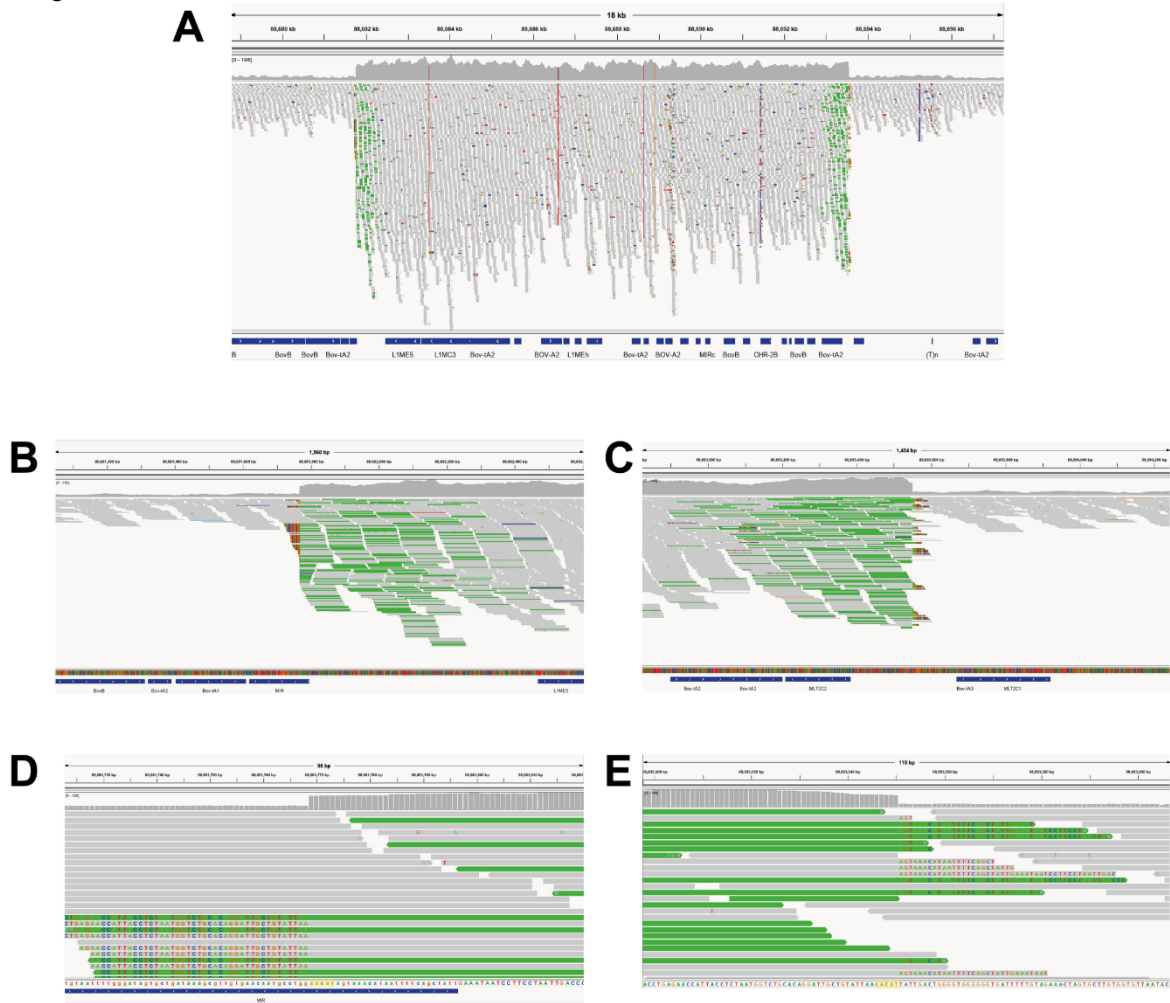**S6 Figure. eQTL mapping and GWAS-eQTL colocalization results for *SLC4A4***
A colocalization plot for CM resistance GWAS and *SLC4A4* eQTL mapping results is shown on the left side. The right upper panel is the CM resistance GWAS results and the right lower panel is the *SLC4A4* eQTL mapping results. In between these two right panels are the genes located in this region. Six genes on the left part are *AMBN, JCHAIN, RUFY3, GRSF1, MOB1B*, and *DCK*. In each panel, the colours of dots indicate degree of LD (r^2) with GC CNV (colour scale shown in the left upper corner of the colocalization figure. The purple diamond marks GC CNV.
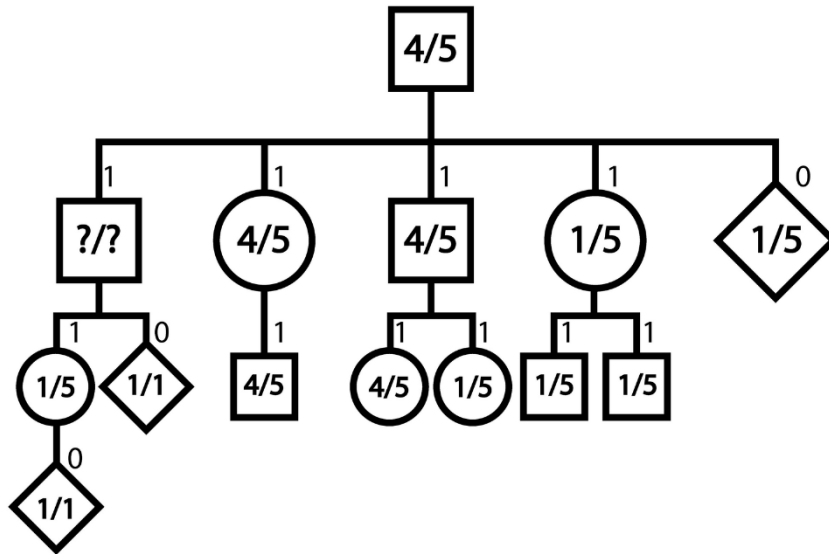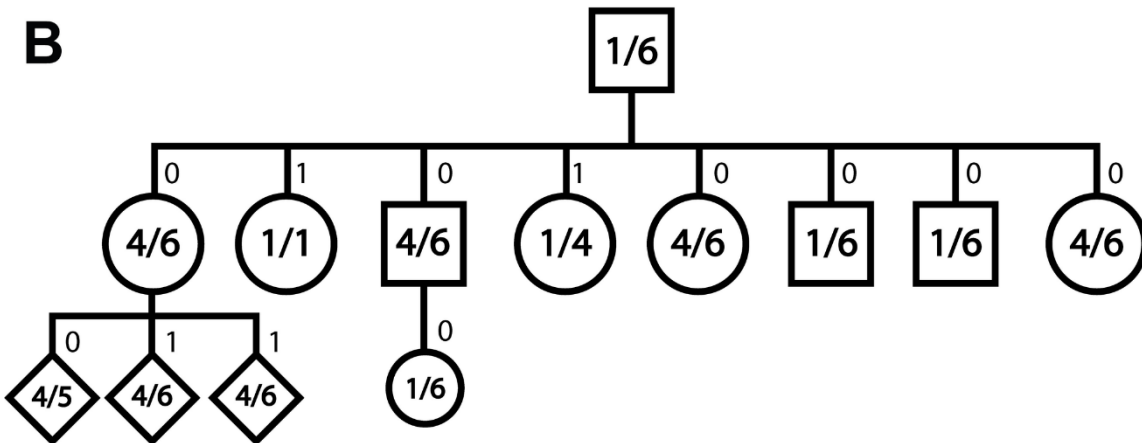
919   **S1 Fig**
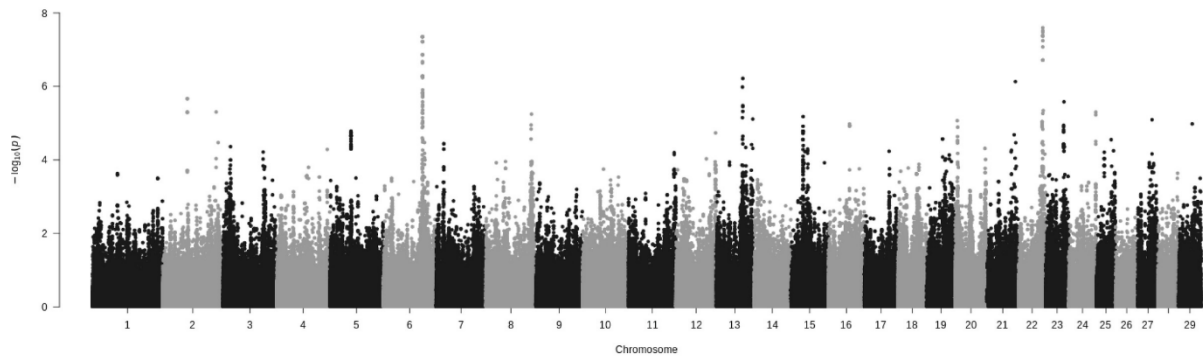
920
921
922   **S2 Fig**

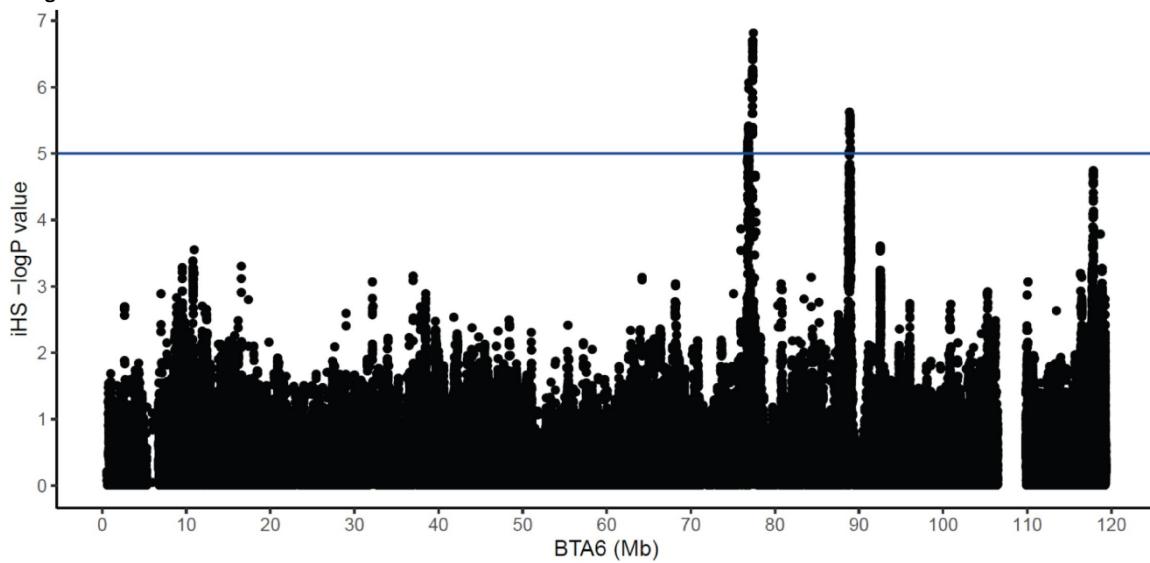
923
924
925   **S3 Fig**

31

926
927
928 **S4 Fig**

929
930
931 **S5 Fig**
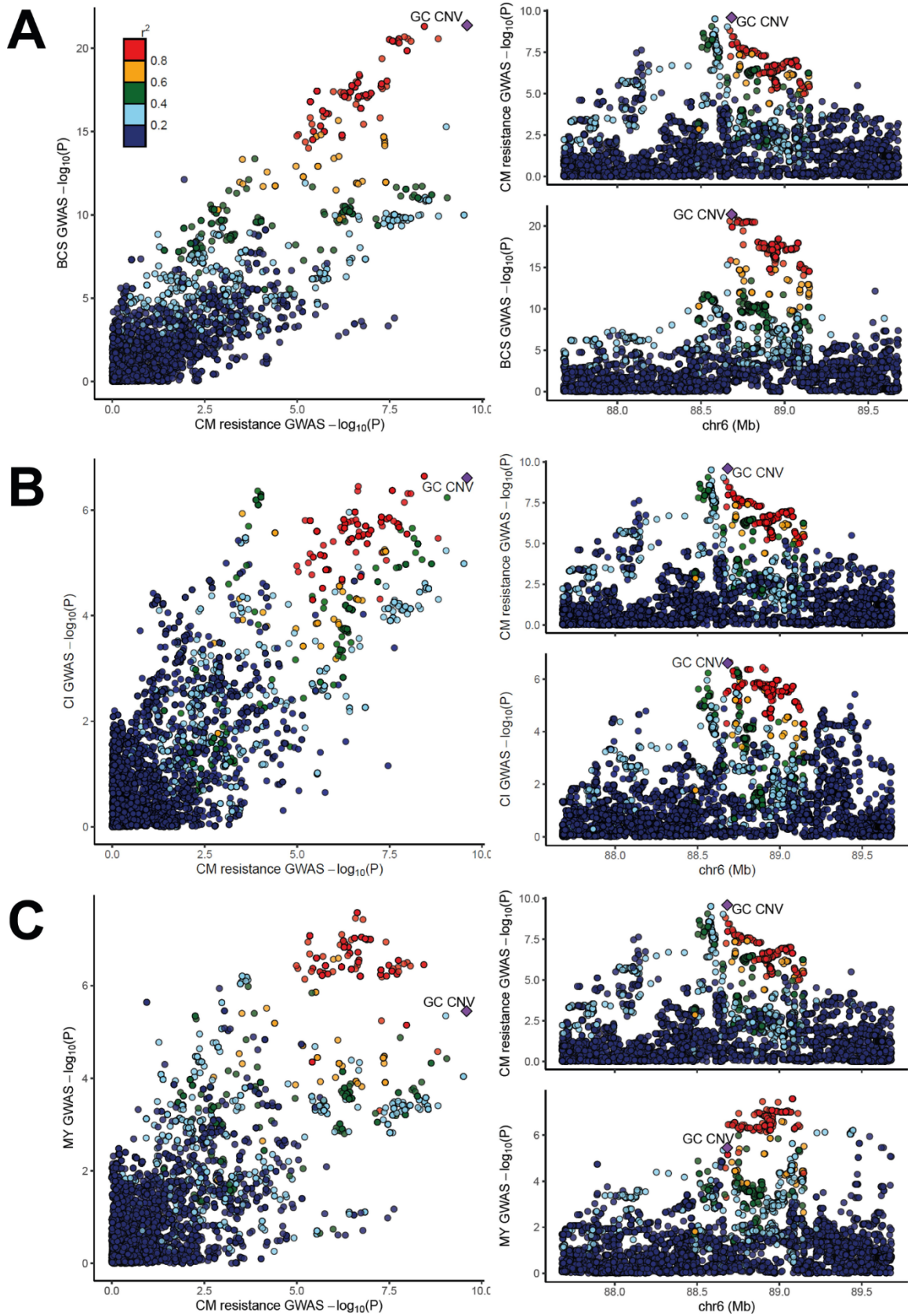
932
933
934 **S6 Fig**
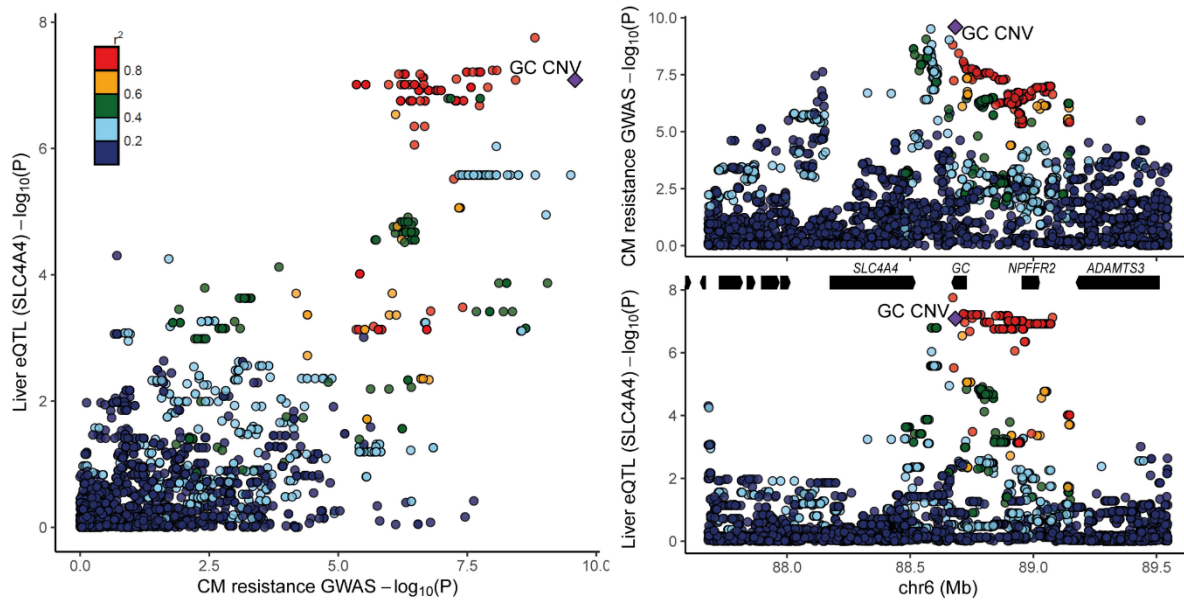
Data reporting

Genome sequence data of the CM resistance QTL region (BTA 6:84-93 Mb) of the 266 Dutch Holstein Friesian animals are deposited in the European Nucleotide Archive under accession XXXX (under preparation). The RNA-seq data is deposited under EBI ArrayExpress accession E-MTAB-9348 and 9871. The genotype data used for eQTL mapping is deposited under European Variant Archive XXXX (under preparation). The ATAC-seq data is deposited under EBI ArrayExpress accession E-MTAB-9872. Genotype and phenotype data used for genome-wise association studies were obtained from CRV B.V., a commercial cattle breeding company. As such, the data are available upon reasonable request and with permission of CRV B.V..

Additional information requested at submission

Competing interests

EM is an employee of CRV B.V., one of the partners of the Breed4Food consortium. All other authors declare that they have no conflict of interest.