METHODOLOGY ARTICLE

Dissecting closely linked association signals in combination with the mammalian phenotype database can identify candidate genes in dairy cattle

Zexi Cai[*] , Bernt Guldbrandtsen, Mogens Sandø Lund and Goutam Sahana

Abstract

Background: Genome-wide association studies (GWAS) have been successfully implemented in cattle research and breeding. However, moving from the associations to identifying the causal variants and revealing underlying mechanisms have proven complicated. In dairy cattle populations, we face a challenge due to long-range linkage disequilibrium (LD) arising from close familial relationships in the studied individuals. Long range LD makes it difficult to distinguish if one or multiple quantitative trait loci (QTL) are segregating in a genomic region showing association with a phenotype. We had two objectives in this study: 1) to distinguish between multiple QTL segregating in a genomic region, and 2) use of external information to prioritize candidate genes for a QTL along with the candidate variant.

Results: We observed fixing the lead SNP as a covariate can help to distinguish additional close association signal(s). Thereafter, using the mammalian phenotype database, we successfully found candidate genes, in concordance with previous studies, demonstrating the power of this strategy. Secondly, we used variant annotation information to search for causative variants in our candidate genes. The variant information successfully identified known causal mutations and showed the potential to pinpoint the causative mutation(s) which are located in coding regions.

Conclusions: Our approach can distinguish multiple QTL segregating on the same chromosome in a single analysis without manual input. Moreover, utilizing information from the mammalian phenotype database and variant effect predictor as post-GWAS analysis could benefit in candidate genes and causative mutations finding in cattle. Our study not only identified additional candidate genes for milk traits, but also can serve as a routine method for GWAS in dairy cattle.

Keywords: Dairy cattle, Milk traits, GWAS, Closely linked association signals, Candidate genes

Background

Over the last decade, the development of high density single nucleotide polymorphism (SNP) arrays and next-generation sequencing (NGS) technologies have made genome wide marker sets available in many organisms [1, 2]. Combining these with phenotypic records on many individuals, genome wide associate studies (GWAS) present a powerful tool to undercover genetic variants associated with variation in the phenotype [3]. In human, numerous studies successfully identified causal variants for traits such as height [4], bodyweight [5] as well as several complex diseases [6]. However, in livestock, long range linkage disequilibrium typically results in imprecise determination of quantitative trait loci (QTL) locations and the associated genomic regions containing several positional candidate genes. In addition, two or more QTL located close to each other may be misidentified as one QTL. In such situations, additional analyses need to be performed to distinguish multiple QTL located close to each other.

To resolve these issues, we need additional information over and above association statistics. For traits with Mendelian inheritance, techniques such as homozygosity mapping and studies of recombinant haplotypes provide

* Correspondence: zexi.cai@mbg.au.dk
Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830, Tjele, Denmark

Cai *et al. BMC Genetics* (2018) 19:30

Page 2 of 12

important clues due to the unambiguous association of at least some genotypes with phenotypic differences. For quantitative traits, no such close associations exist. However, genomic information of various types do allow relative prioritization among candidate variants. The challenges are which information to consider post-GWAS and how to combine them with GWAS statistics. Expression quantitative trait loci (eQTL) mapping can help; expression profiles as the dependent trait in a GWAS have identified causal genes in some studies [7]. Nevertheless, eQTL studies are time consuming and expensive. Therefore, alternative approaches to incorporate gene expression data into GWAS are needed. Other sources of additional information like variants' annotation [8] and evolutionary conservation scores [9] have been used. Unfortunately, these analyses need to be designed on a case-by-case basis [10]. Their implementation is challenging in livestock due to the sparsity of annotation data.

In this study, we used an approach to separate multiple closely linked QTL in dairy cattle by fixing the lead SNP as a covariate. The analysis detects QTL chromosome by chromosome, and generates a list of lead SNPs for each QTL. The method is demonstrated by application to three milk yield traits in Nordic Holstein cattle. Many previously identified loci were also confirmed here. Furthermore, we used the mammalian phenotype database to help find the candidate genes and Variant Effect Predictor (VEP) annotations to screen for possible causal mutations.

Results

We applied a GWAS analysis approach that automatically and iteratively accounts for the effects of QTL identified in previous iteration(s), a similar approach to conditional analysis implemented in GCTA [11]. The impact of pre-correction on type I error rate was assessed by analyzing simulated data with the effect of a quantitative trait nucleotide (QTN) added to the real phenotypic data (for details on simulation, see Method section). The candidate genes were picked as the closest genes to the lead SNP and listed in Tables 1, 2 and 3. The search for candidate genes started with the top SNP location. However, the whole genomic region showing strong associations with the trait was searched, as the top SNP may not be always located closest to the causal gene due to differences in: LD, imputation accuracy and minor allele frequency. Therefore, we included discussion on other relevant genes (based on association results, known gene function etc.) which could be candidate genes underlying the QTL.

Our approach of including associated SNPs as covariates in subsequent rounds of analyses did not increase the type I error rates. We simulated one SNP as a QTN

and considered ten other SNPs with different levels of LD (r^2) with the QTN in order to test whether our method introduces type I error into analysis when fixing lead SNPs detected in previous iterations as covariates [12]. We generated new phenotypes from the real phenotypic value plus the simulated QTN effects. The QTN's contribution to individuals' phenotypes was obtained by multiplying the genotype dosage of the QTN with the allele substitution effect which was drawn from a normal distribution with a mean 20% of the standard deviation (SD) of the phenotype and variance as 1% of the phenotypic variance. The simulation was repelicated 100 times. We detected the simulated QTN as the lead SNP in the first round of all 100 replicates. When the simulated QTN was included in the model as a covariate, we did not observed any of the 10 SNPs in LD with QTN to be significant (i.e., no false positives detected).

The GWAS of fat yield

Analyzing milk fat yield, our approach detected seven additional QTL over and above the QTL detected in the first round (Fig. 1 and Table 1). In Table 1, the first SNP on each chromosome is the lead SNP from the first round of GWAS analysis, the rest are the additional SNP(s) detected on a chromosome. Sixteen SNPs on chromosome 14 have the same *P*-value in the first round, and these SNPs are in high LD with the two known causative polymorphisms in *DGAT1* [13], BTA14: 1802265 (rs109234250) and BTA14: 1802266 (rs109326954) (Additional file 1: Figure S1). The variant effect predictor (VEP) [14] annotation showed these two variants in *DGAT1* are missense mutations. The second strongest association signal was located on chromosome 5 with lead SNP, BTA5: 93948357 (rs209372883) located within the intron of *MGST1*. *MGST1* was previously reported associated with the milk fat content [15]. On chromosome 26, our lead SNP pointed to *COX15*. In a human study, this gene was proposed involved in biosynthesis of heme A [16]. Even though this gene is a promising positional candidate gene, no biological information currently links this gene to milk fat yield. Another gene known to affect milk fat content is *SCD1* [17] located at chromosome 26: 21141592 ~ 21,148,318. Our lead SNP on chromosome 26 (BTA26:20547445, rs136702635) is located close to it. We estimated the variance explained by QTL. The QTL (16 QTL) found from the first round explained 22.77% of the variance of de-regressed proof breeding value (DRP) for fat yield and all QTL (23 QTL) explained 25.12% of the DRP variance (Table 4).

The GWAS for protein yield

We ran the analysis on the milk protein yield (Fig. 2), and found 33 lead SNPs (Table 2), 12 of which were detected in the second or third round. The strongest

Cai *et al. BMC Genetics* (2018) 19:30

Page 3 of 12

Table 1 Lead SNPs from genome-wide associated regions for fat yield in Nordic Holstein cattle. Base positions are given as position in UMD 3.1.1 [49]

BTA	base position	Imputation accuracy	Effect	$-\log_{10}(p)$	Region	Gene	Annotation
2	126979882	0.9972	−1.31	11.46	126041707~ 127230070	*PIGV* (near)	Downstream
2	85991577[b]	0.9542	1.30	8.91	85042155~ 86241732	*ANKRD44*	Intron
3	7226390	0.9998	−1.09	9.01	6264604~ 7476473	*NOS1AP*	Intron
5	93948357	0.9906	3.28	62.41	93698481~ 94198475	*MGST1*	Intron
5	20284735[b]	0.9692	−1.30	9.79	20035379~ 20534779	5S_rRNA	intergenic
6	95497933	0.9996	−1.45	14.76	95248213~ 95747954	*PAQR3*	intergenic
6	32950721[b]	0.4975	6.33	11.39	32367171~ 33200834	ENSBTAG00000047255	Intron
7	57287990	0.8807	−1.66	20.11	57038213~ 57538309	*KCTD16*	Intron
9	38715137	0.9809	−1.47	8.89	38345408~ 38965425	*LAMA4*	Intron
11	88771449	0.9876	1.16	10.43	88521462~ 89021477	ENSBTAG00000047976	intergenic
11	15323223[b]	0.8962	−1.32	9.81	14855568~ 15573444	*TTC27*	Intron
11	55681193[c]	0.9948	−1.60	9.91	55423855~ 55931229	*CTNNA2*	Intron
12	68965758	0.9957	−1.10	8.93	68502223~ 69216445	ENSBTAG00000045195	intergenic
14[a]	1802265	0.9398	−6.93	240.56	1549133~ 2049435	*DGAT1*	missense
14[a]	1802266	0.9362	−6.93	240.56	1549133~ 2049435	*DGAT1*	missense
15	65891100	0.9992	1.50	12.99	65363764~ 66141839	*ELF5*	intergenic
15	25044706[b]	0.9908	−1.17	9.80	24795472~ 25295470	*ZBTB16*	Intron
17	62543160	0.9898	1.14	10.49	62224291~ 62793298	*TBX5*	Intron
18	18970551	0.9442	−1.19	10.30	18341203~ 19220732	*NKD1*	intergenic
19	27522927	0.8500	−1.32	10.86	26625240~ 27773016	*ASGR1*	intergenic
20	22609736	0.9813	1.53	14.23	21664412~ 22859809	*MAP3K1*	intergenic
26	20547445	0.9993	−1.76	21.46	20297497~ 20797570	*COX15*	Intron
26	42408595[b]	0.9998	−1.21	10.30	41409014~ 42658925	*TACC2*	Intron
29	23609412	0.7717	2.06	10.73	22613737~ 23859451	ENSBTAG00000047094	intergenic
	Total number of significant SNPs					54,435	

[a] Fourteen additional SNPs on chromosome 14 located near DGAT1 gene had same highest P value (details on those not presented). [b] indicated this SNP was found on second round, [c] indicated this SNP was found on third round

association signal for protein yield was on BTA14 with lead SNP BTA14:1835440 (rs208567981), located within *BOP1*. The annotation of BTA14:1835440 (rs208567981) is a missense mutation, and the SIFT annotation is tolerant. However, this signal is most likely due to the known mutation in *DGAT1*. The lead SNP (rs208567981) was in strong LD with SNPs located within *DGAT1* and the $-\log_{10}(P)$ value of these 19 SNPs within *DGAT1* were larger than 47.99 (including two known causative variants in *DGAT1*, Additional file 1: Figure S2). This result shows that the causal mutation may not necessarily be the SNP in highest association. The second lead SNP of this analysis is BTA6: 88477501, which is located near the well-studied casein genes *CSN1S1*, *CSN1S2*, *CSN3* and *CSN2* [18]. We estimated the variance explained by QTL. The QTL (21 QTL) found only from the first round explained 10.85% of the DRP variance for protein yield and all QTL (33 QTL) explained 15.34% of the DRP variance (Table 4).

The GWAS for milk yield

We applied our analysis to milk yield (Fig. 3). A total of 26 lead SNPs (Table 3) were detected, out of which six were detected in the second or third round. The most significant association signal was in the *DGAT1* gene. The second most significant association signal was at BTA20:29996719 (rs43116343), which is close to *MRPS30*. A recent study showed *MRPS30* to be associated with lactation persistence in Canadian Holstein cattle [19]. This lead SNP is also close to the growth hormone receptor, *GHR* [20]. The causative mutation of *GHR* is BTA20:31909478 (rs385640152), and is known to affect milk yield [20]. The third strongest lead SNP was BTA5:93953487 (rs210234664). This SNP is close to *MGST1*. A previous eQTL study showed *MGST1* may affect milk composition [21]. With our approach, we detected BTA6: 38027010 (rs43702337) in the third round, located in *ABCG2*. *ABCG2* was previously reported to affect milk yield in dairy cattle [22]. This SNP is a

Cai *et al. BMC Genetics* (2018) 19:30

Page 4 of 12

Table 2 Lead SNPs from genome-wide associated regions for protein yield in Nordic Holstein cattle. Base positions are given as position in UMD 3.1.1 [49]

BTA	base position	Imputation accuracy	Effect	$-\log_{10}(p)$	Region	gene	Annotation
1	63177947	0.9885	−1.94	12.35	61838881~ 63178271	ENSBTAG00000046854 (near)	intergenic
2	124837669	0.9886	1.59	12.63	124834921~ 124837676	PTPRU	Intron
3	17160521	0.9717	−1.15	8.76	15377852~ 17160986	S100A12 (near)	Upstream
5	93511826	0.8626	−1.37	14.25	93087740~ 93511841	LMO3 (near)	intergenic
5	21792183[a]	0.9813	−1.37	10.39	20072875~ 21792190	SNORD107 (near)	intergenic
5	87923795[b]	0.9926	1.50	8.97	85996611~ 87924188	ETNK1 (near)	intergenic
6	88477501	0.9962	−2.60	25.98	87154594~ 88477524	SLC4A4	Intron
6	48477272[a]	0.7329	1.49	13.59	46907022~ 48477298	ENSBTAG00000045570 (near)	intergenic
6	88749792[b]	0.3222	−1.66	12.13	86831016~ 88749854	GC (near)	intergenic
7	41372989	0.9999	−1.54	18.14	41085164~ 41373119	MGAT1 (near)	intergenic
7	72100619[a]	0.9077	1.59	13.29	70118741~ 72100721	EBF1 (near)	intergenic
8	93065787	0.8573	1.65	10.07	91065857~ 93066321	GRIN3A	Intron
8	31538155[a]	1.0000	1.91	9.62	30388755~ 31538625	LURAP1L (near)	intergenic
9	33267855	0.8655	−1.46	11.96	32627954~ 33267856	SLC35F1 (near)	intergenic
10	93933304	0.8370	−1.36	9.90	92043775~ 93933391	SEL1L	Intron
11	35512708	0.9999	−1.45	11.82	33534073~ 35512724	ENSBTAG00000027786 (near)	intergenic
13	37208792	0.9279	−1.69	10.90	35572625~ 37208843	MKX (near)	intergenic
13	77657858[a]	0.9906	1.17	9.52	76677111~ 77908632	PREX1	intron
14	1835440	0.7471	2.84	48.66	1448510~ 1836166	BOP1	Intron
14	67981742[a]	0.7652	1.78	11.60	65984180~ 67981772	STK3 (near)	intergenic
16	32262983	0.9290	−1.52	12.79	30519873~ 32263130	SMYD3	Intron
18	57015407	0.9754	2.56	17.71	55015676~ 57015473	POLD1	Intron
18	15272231[a]	0.6697	−1.16	9.53	15032157~ 15272234	SNORA11(near)	downstream
19	27522927	0.8500	−1.42	12.55	26422519~ 27522980	RNASEK (near)	downstream
19	61014793[a]	0.8505	−1.08	8.65	60995058~ 61014874	KCNJ2 (near)	intergenic
20	69006609	0.9920	−1.29	11.27	68120719~ 69006661	IRX1 (near)	intergenic
20	9282667[a]	0.6747	1.71	10.93	7765154~ 9282923	MRPS27	intron
23	10974968	0.9304	−1.18	10.68	9127211~ 10975139	FGD2 (near)	intergenic
25	36403719	1.0000	1.33	10.25	36112575~ 36403849	EPO (near)	intergenic
26	37695494	0.9122	−1.41	14.76	36684176~ 37695588	KIAA1598 (near)	intergenic
27	36304978	0.9834	1.06	8.52	35875452~ 36305040	ANK1	intron
29	17620617	0.9576	1.47	10.37	15650574~ 17620644	NARS2	intron
29	35459126[a]	0.9999	1.61	10.11	33464929~ 35459620	NTM	intron
	Total number of significant SNPs					38,439	

[a]indicated this SNP was found on second round, [b] indicated this SNP was found on third round

missense variant; its SIFT annotation is "deleterious" and has previously been proposed as a causative mutation [23]. We estimated the variance explained by QTL. The QTL (20 QTL) found from the first round explained 18.85% of the DRP variance for milk yield and all QTL (26 QTL) explained 21.29% of the phenotypic variance (Table 4).

Post-GWAS analysis using the mammalian phenotype database

The criteria for selecting positional candidate genes was the gene located closest to the lead SNP. For future identification and research on genes biologically associated with milk traits, we tried to find whether there are other genes which should be considered as potential candidate genes other than the candidate gene list (Tables 1-3). Considering the high LD structure of cattle population, the causal genes may be located within the genome region in LD with lead SNPs. One source of additional information that may help to prioritize genes, is to find the link between the gene and the possible function in the mammalian phenotype database related to milk and milk-organ related traits [24]. Therefore, we extracted genes which overlap with the LD region of the lead SNP and search them in the mammalian phenotype

Cai *et al. BMC Genetics* (2018) 19:30

Page 5 of 12

Table 3 Lead SNP from genome-wide associated regions for milk yield in Nordic Holstein cattle. Base positions are given as position in UMD 3.1.1 [49]

BTA	base position	Imputation accuracy	Effect	$-\log_{10}(p)$	Region	Gene	Annotation
2	80753895	0.9454	1.13	9.95	79587952~ 80754103	*NABP1* (near)	intergenic
3	56402959	0.9308	−1.36	11.68	56392727~ 56402961	ENSBTAG00000001873 (near)	intergenic
4	101547644	0.7008	−1.66	12.65	100921921~ 101547648	*CHRM2* (near)	upstream
5	93953487	0.9726	−2.10	29.52	91953587~ 93953619	*MGST1* (near)	upstream
5	23022794[a]	0.7617	1.38	12.93	21101742~ 23022797	*EEA1* (near)	intergenic
5	85080296[b]	0.7619	−1.28	11.24	84425435~ 85080331	ENSBTAG00000009778 (near)	intergenic
6	88847595	0.9009	−1.78	21.61	88612186~ 88847596	*GC* (near)	intergenic
6	46901490[a]	0.7413	−1.28	11.45	46181675~ 46901554	*SEL1L3* (near)	intergenic
6	38027010[b]	0.9950	−4.75	9.47	36909885~ 38027583	*ABCG2*	missense
7	65370850	0.9848	−1.36	13.58	65256765~ 65370922	*GLRA1* (near)	intergenic
8	73877814	0.8453	−1.37	11.14	71877875~ 73877845	ENSBTAG00000010829 (near)	upstream
8	42062591[a]	0.9595	−1.27	10.07	40245362~ 42062776	*KCNV2*	intergenic
9	33478527	0.8801	−1.25	9.23	31790030~ 33478670	*NUS1* (near)	intergenic
10	1989907	0.9469	−1.15	9.92	448434~ 1990092	ENSBTAG00000047622 (near)	intergenic
13	36822330	0.9933	−1.66	10.74	36663680~ 36822395	*MPP7*	Intron
14[#]	1802667	0.7975	5.98	178.35	1702853~ 1797137	*DGAT1*	Intron
14	67577503[a]	0.8898	−2.16	11.04	66624772~ 67828111	*OSR2*	intergenic
15	54392611	0.9577	1.57	16.58	52771707~ 54393036	*PPME1*	Intron
16	28384260	0.9984	1.64	10.50	28012864~ 28384605	*CNIH3* (near)	Intergenic
17	66510224	0.9438	1.83	11.63	66119023~ 66510712	*CORO1C*	Intron
18	46583346	0.9829	1.86	11.97	44583383~ 46583963	*UPK1A* (near)	upstream
19	27442452	0.7904	−1.26	9.71	26592355~ 27442492	bta-mir-497 (near)	upstream
20	29996719	0.9580	−2.95	31.02	27997007~ 29996870	*MRPS30* (near)	intergenic
23	25076472	0.9797	−1.34	9.23	23690289~ 25076491	*GCM1*	Intron
26	37716420	0.9790	−1.43	12.28	36730021~ 37966463	*TRPA1*	intergenic
28	34972377	0.9991	−1.29	9.81	33464705~ 34972672	*ZMIZ1*	intergenic
	Total number of significant SNPs					57,808	

[#]Eight additional SNPs on chromosome 14 had same highest P value. [a] indicated this SNP was found on second round, [b] indicated this SNP was found on third round
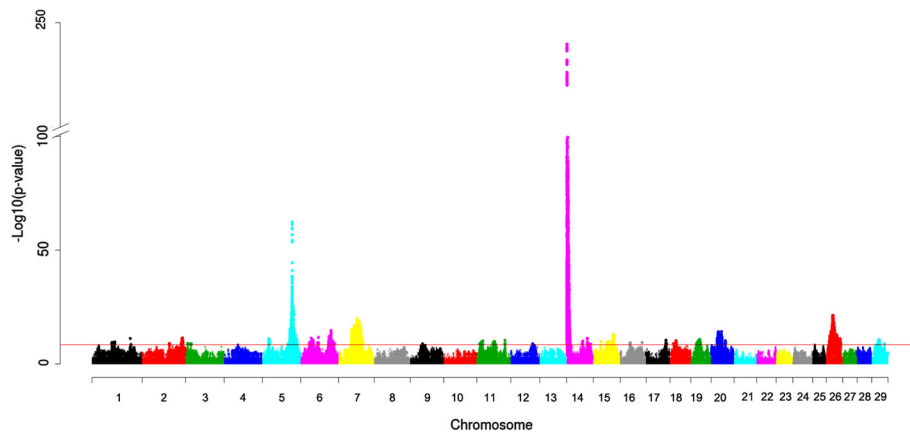

Fig. 1 Manhattan plot for association of SNP with fat yield in Nordic Holstein cattle. Red horizontal line indicates genome-wide significance level [−log10(P) = 8.5]

Cai *et al. BMC Genetics* (2018) 19:30

Page 6 of 12

Table 4 The genetics variants explained by QTL and the rest of SNPs

	Number of QTL	V(G1)/Vp[b] (%)	V(G2)/Vp[c] (%)
Fat1[a]	16	22.77	62.59
Fat2[a]	23	25.12	60.01
Prot1[a]	21	10.85	74.05
Prot2[a]	33	15.34	68.89
Milk1[a]	20	18.85	66.67
Milk2[a]	26	21.29	63.97

[a] Fat means the trait of fat yield, Prot means the trait of protein yield, Milk means the trait of milk yield; 1 indicate the lead SNP list only included the lead SNP from the first round, 2 indicated the lead SNP list included all lead SNP found by our approach. [b] means the percentage of genetics variants explained by the QTL, [c] means the percentage of genetics variants explained by the rest of SNP other than QTL

database [24]. We only paid attention to two kinds of phenotypes: "abnormal mammary gland development" or "abnormal milk composition". Eight genes from the GWAS hits were also annotated as related to these two types of phenotype. This annotation appears to have biological relevance, although the enrichment of these 8 genes in the mammalian phenotype database analyzed by Fishers' exact test was not significant. The results showed four genes were reported to be related with "abnormal milk composition" (Table 5). Out of this list, *CSN1S1*, *CSN2*, *CSN3* and *DGAT1* were reported in dairy cattle and also identified in the present study. Furthermore, we identified five genes related to "abnormal in mammary gland development" (Table 6) in mammalian phenotype database. In this list *DGAT1* showed abnormal phenotype in both kinds of phenotype description we searched. In addition to the well-studied genes (*CSN1S1*, *CSN2*, *CSN3* and *DGAT1*), the remaining four genes are *ELF5*, *CAT*, *STK3* and *CHUK*. *ELF5* is one of the candidate genes proposed by the

closest genes to lead SNP (BTA15: 65891100) associated with the fat yield (Table 1). *ELF5* was previously found related to mouse mammary development [25] and may also influence the milk content through milk protein synthesis in cattle [26]. *CAT* is also located close to the same lead SNP as *ELF5*. *CAT* is involved in several biological processes including GO term 'responds to fatty acid' [27]. *CHUK*, close to BTA26: 20547445, is associated with fat yield (Table 1). This gene is known as a key gene involved in mammary development in mice [28]. *STK3* is the nearest gene to the second lead SNP (BTA14: 67981742) on the same chromosome associated with milk protein yield (Table 2). This gene was found to play a pivot role in controlling cell proliferation [29] and tumor suppression [30] in human studies.

Annotation of SNPs in LD with lead SNPs

As shown before, the causative mutation maybe located in the neighboring region of the lead SNP. Therefore, we extracted all SNPs in LD with leading SNPs ($r^2 > 0.2$) and annotated them using VEP [14]. We extracted 190,130 SNPs and obtained 203,120 annotations (because some genes or transcripts overlap). The majority of these SNPs are intergenic variants or intron variants (Fig. 4a). Among the SNPs that changed the coding sequence of the protein, most of them were synonymous variants (Fig. 4b). Using this result, we checked if we could prioritize candidate mutations in the candidate genes. For example *GHR*, the well-known causative mutation for *GHR* is BTA20:31909478 (rs385640152, F279Y) [20]. The annotation for this SNP is a missense mutation and the SIFT score is 0.02 which is 'deleterious'.

Further, we checked whether we can detect some candidate mutations in the new candidate genes. Four genes (*CSN1S1*, *CSN2*, *CSN3* and *DGAT1*) were found related

Fig. 2 Manhattan plot for association of SNP with protein yield in Nordic Holstein cattle. Red horizontal line indicates genome-wide significance level [−log10(P) = 8.5]

Cai *et al. BMC Genetics* (2018) 19:30

Page 7 of 12

Fig. 3 Manhattan plot for association of SNP with milk yield in Nordic Holstein cattle. Red horizontal line indicates genome-wide significance level [−log10(P) = 8.5]
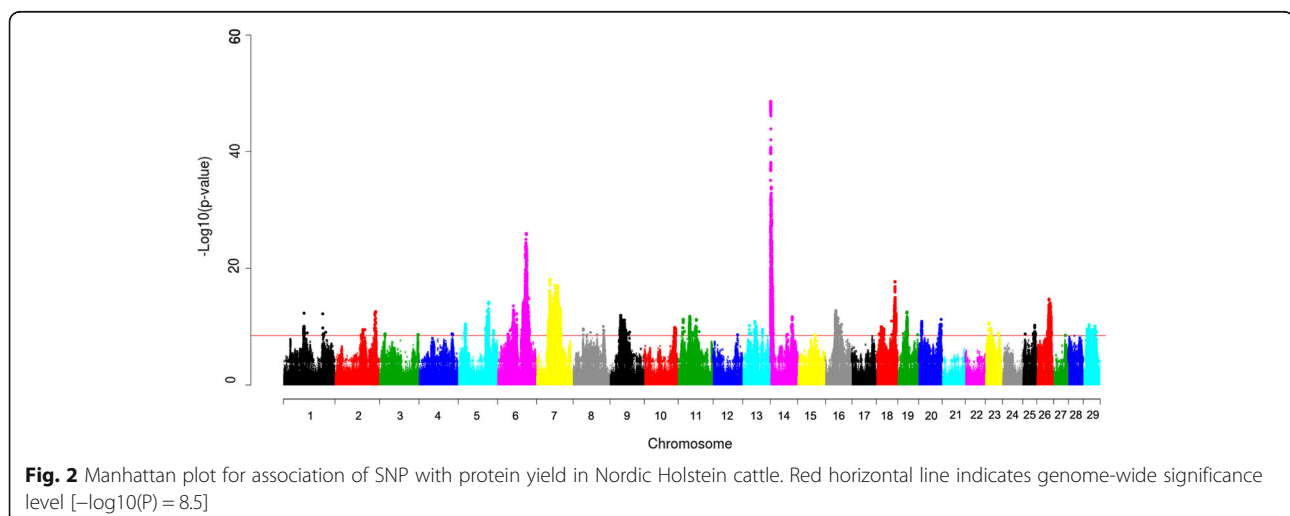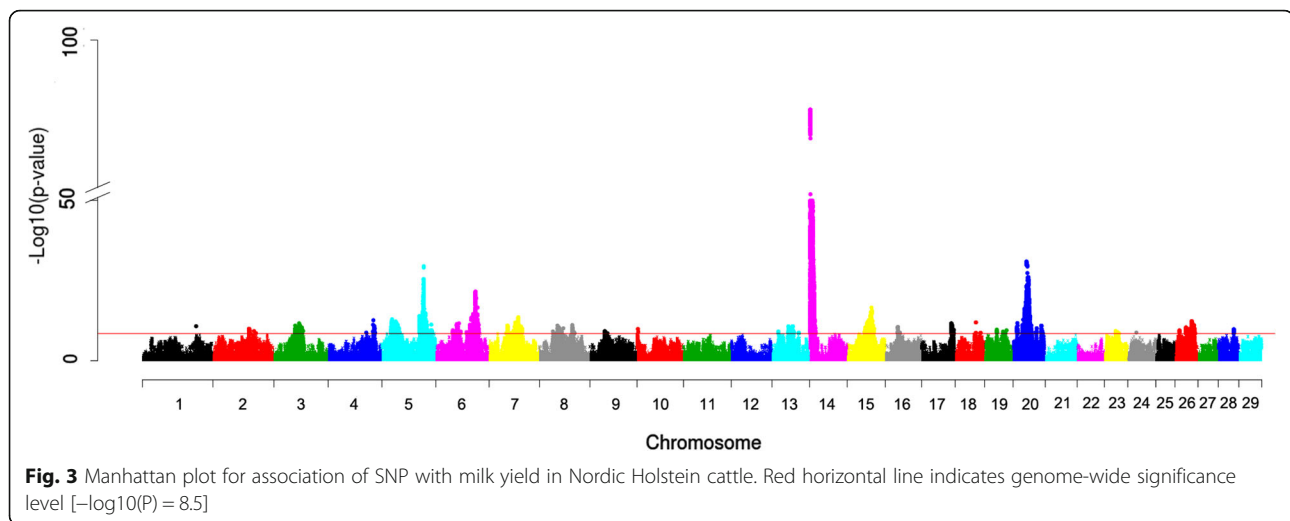
to abnormal milk composition and *DGAT1* related to mammary gland development (Table 5 and Table 6) as reported previously. In addition to *DGAT1*, we found several tolerance missense mutations in *CSN1S1* and *CSN2*. In *CSN3*, we found three tolerance missense mutations and one deleterious mutation (BTA6:87390632, rs43703017). Moreover, in *STK3*, we also found tolerance missense mutations.

Discussion

Although functional gene clustering is weaker in eukaryotes genomes than in prokaryotes genomes, functional grouping of the genes with same or similar function still exists [31]. Therefore, in GWAS analysis, we may fail to detect the nearby genes and may treat them as one significant signal. In our study, we used an analysis approach to detect multiple nearby QTL by iteratively fixing the lead SNP as covariate. However, such an approach can inflate type I error rate [12]. To avoid introducing additional type I errors, we placed a condition that the lead SNPs detected in the second and subsequent rounds must be found to be genome-wide significant in the first round (i.e., significant according to conventional GWAS criteria). In addition, we tested our approach on simulated data with a simulated QTN and multiple SNPs with various levels of LD with the QTN.

In 100 replicates, we found no additional SNP in LD with the QTN other than the simulated causative variants. By using this analysis, we were able to detect multiple QTL (as well as designating the lead SNP for each QTL) on a chromosome automatically. For example, we detected a known QTL on BTA6 (BTA6:38027010, rs43702337) in the third round and also another QTL at 46 Mb (in the second round). This SNP is located in the gene *ABCG2* which was previously reported to affect milk yield in dairy cattle [22] and this lead SNP was the most probable causative mutation [23]. Furthermore, our approach also showed the potential to distinguish closely linked QTL. For example the lead SNPs on chromosome 6 of protein content, we detect the first association signal at BTA6: 88477501 and the third association signal at BTA6: 88749792. Similar conditional analyses were also applied in human and other livestock studies [32–34]. Here, we analyzed one lead SNP at a time, as opposed to Bolormaa et al. [34] who included all lead SNPs simultaneously in the model. We also compared the genetic variants explain by the QTL found by first round and all the QTL found by our approach. The results showed the QTL found at second and third round did explain more phenotype variants (Table 4).

Post GWAS, we face the challenge of identifying the candidate genes. The conventional method is to use the nearest gene, but this may miss the target as many-a-time the lead SNP may not be from the causal gene. This could be due to imputation inaccuracies, multiple QTL in the vicinity or random chance factor. Therefore, we need to use additional information to prioritize the candidate genes. In this study, we used the mammalian phenotype database to search for candidate genes from the genes located in association regions. The mammalian phenotype is based on mouse mutation lines. As a test, we extracted all genes located within LD of the lead

Table 5 Genes related to "abnormal milk composition" phenotype in the mammalian phenotype database [24] overlapped with milk QTL identified in the present study

Gene name	Location	Phenotype
CSN1S1	BTA6: 87,141,556–87,159,096	abnormal milk composition
CSN2	BTA6: 87,179,502–87,188,025	abnormal milk composition
CSN3	BTA6: 87,378,398–87,392,750	abnormal milk composition
DGAT1	BTA14: 1,795,351–1,804,562	abnormal milk composition

Cai *et al. BMC Genetics* (2018) 19:30

Page 8 of 12

Table 6 Genes related to "abnormal of mammary gland development" in the mammalian phenotype database [24] overlapped with milk QTL identified in the present study

Gene name	Location	Phenotype
CAT	BTA15: 65,779,325–65,815,261	decreased mammary gland tumor incidence
ELF5	BTA15: 65,824,442–65,854,386	abnormal mammary gland development
STK3	BTA14: 67,677,676–67,987,801	increased mammary gland tumor incidence
DGAT1	BTA14: 1,795,351–1,804,562	abnormal mammary gland development
CHUK	BTA26: 20,966,010–21,008,277	abnormal mammary gland growth during pregnancy

SNPs for all three milk yield traits and searched for related phenotype terms. Here, we searched for two phenotype terms 'abnormal mammary gland development' and 'abnormal milk composition'. We successfully identified some well-known genes affecting milk related traits in cattle as well as new candidate genes (Table 5 and Table 6). For the term 'abnormal milk composition', we identified four genes. All of which were reported previously in different studies [35, 36], and only *DGAT1* is the nearest gene to the lead SNP on chromosome 14. Another term we searched is 'abnormal in mammary gland development' and found five genes. *CAT* and *CHUK* are not the nearest genes to the lead SNP. However, differences between mice and cattle may introduce some false positives. In all, using this strategy we not only found some well-studied genes missing from the nearest genes method (pick the gene which is nearest to lead SNP as candidate genes), but also identified new candidate genes which may be helpful in finding causal factors.

We also face another challenge of identifying the causative variant once the causal gene is identified as levels of linkage disequilibrium in cattle are high [37]. In many cases the causative variant is not the lead SNP [38] but another SNP hidden within the LD of the lead SNP. In human studies, there are different strategies to prioritize variants [10]. In this study, information from Ensembl [14] was used to prioritize potential causative variants. In our case, the *DGAT1* and *ABCG2* can be detected in our lead SNP list, and the causative mutation of both can be detected in VEP annotation as missense

variants. *GHR* was found nearby the location of lead SNPs. For *ABCG2* and *GHR*, the SIFT score show these mutations as 'deleterious'. For *DGAT1*, even though the SIFT showed these two mutations are tolerated the amino acid of the protein is changed. Therefore, the impact of moderate and high reported by VEP can be considered as possible causative mutations, while SIFT score can be used to provide additional support.

In summary, our analysis approach can distinguish nearby association signals of multiple QTL. In our study, we found candidate genes reported by previous studies. Followed by searching genes within the LD region of the lead SNPs, we can find high confidence candidate genes. Lastly, using VEP can help us to find putative causative mutations within candidate genes and provides a good source for further functional validation. However, our approach will not be able to pinpoint causal variants located in the non-coding and regulatory regions due to lack of annotation of the cattle genome.

Conclusions

In this study, we designed an approach for detecting closely linked multiple association signals and performed the analysis in Nordic Holstein cattle for milk, fat and protein yields. The results showed we not only detected most of the well-known genes affecting these three milk yield traits but also detected additional candidate genes. Post-GWAS, we used information from the mammalian phenotype database and variant effect predictor to confirm known genes and causative mutations. In the meanwhile, we detected additional genes which might be

Fig. 4 The VEP annotation of SNPs in linkage disequilibrium (LD > 0.20) with leading SNPs. **a** The summary of all annotation. **b** The summary of annotation that change the protein coding sequence

Cai *et al. BMC Genetics* (2018) 19:30

Page 9 of 12
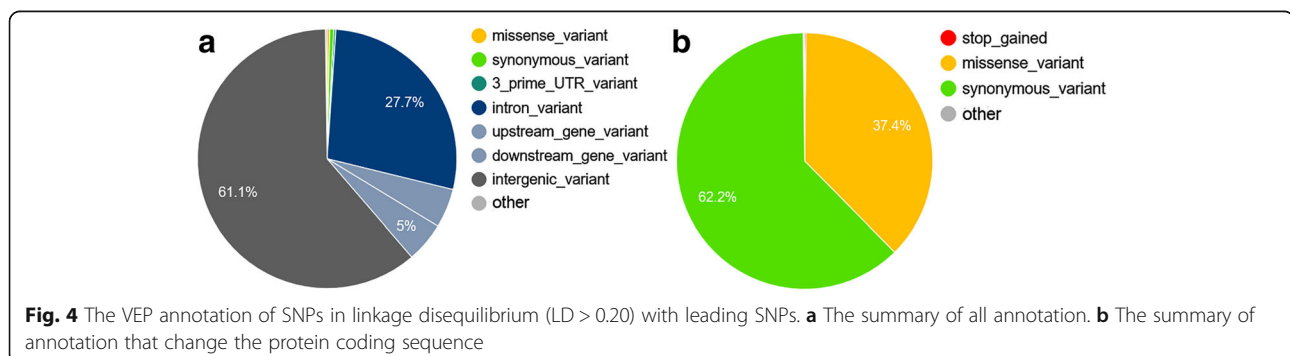
contributing to variation in milk traits in Nordic Holstein cattle. Therefore, we concluded our approach can be used routinely for GWAS studies in dairy cattle.

Methods

Phenotype and genotype data

No animal experiments were performed in this study, and therefore, approval from the ethics committee was not required.

Phenotypic records for Nordic Holstein cattle are kept in a centralized database (Nordic Cattle Genetic Evaluation, NAV. http://www.nordicebv.info/). Breeding values for milk, fat and protein yield (MY, FY and PY) are based on production figures expressed in kilograms taken from routine milk records and then combined into an index for each trait. For details on genetic evaluation for milk yield traits in Nordic countries see (http://www.nordicebv.info/production). The breeding values used for association analysis were de-regressed proof breeding values [39, 40] from the routine genetic evaluation by NAV and were available for 5382 progeny tested Holstein bulls.

The association study was carried out by using imputed WGS data, as previously described by Iso-Touru et al. [41] and Wu et al. [42]. A total of 4921 bulls were genotyped with the Illumina BovineSNP50 BeadChip (54 k) ver. 1 or 2 (Illumina, San Diego, CA, USA). The 54 k genotypes were imputed to WGS variants by a 2-step approach [43]. First, all animals were imputed to the high-density (HD) level by using a multibreed reference of 3383 animals (1222 Holsteins, 1326 Nordic Red Dairy Cattle, and 835 Danish Jerseys), which had been genotyped with the Illumina BovineHD BeadChip. Subsequently, these imputed HD genotypes were imputed to the WGS level by using a multibreed reference of 1228 animals from *Run4* of the 1000 Bull Genomes Project [1] (1148 cattle, including 288 individuals from the global Holstein population, 56 Nordic Red Dairy Cattle, 61 Jerseys, and 743 cattle from other breeds [1] and additional data from Aarhus University (80 individuals, including 23 Holsteins, 30 Nordic Red Dairy Cattle, and 27 Danish Jerseys).

Different variant calling pipelines were used for the 1000 Bull Genome Project data and the in-house Nordic data at Aarhus University. The whole genome sequence data at Aarhus University was analyzed as described by Brøndum et al. [44]; while the same for 1000 Bull Genome Project was described by Daetwyler et al. [1]. Detailed guidelines are available at http://www.1000bullgenomes.com. Data from both sources were available as VCF files. The data from the two sources were combined using Picards MergeVCFs (https://broadinstitute.github.io/picard/). As the 1000 Bull Genomes Project only shares data after variant calling, some markers were not called for all animals in the combined dataset. To avoid large gaps of missing markers in the dataset, only markers that were called in both the Nordic and the 1000 Bull Genomes Project datasets were kept. For positions containing both a SNP and an INDEL, the INDEL was discarded, as the imputation methods rely on unambiguous sequences of variants. Positions with disagreements between alleles for sequence and HD data were also deleted. Reference genotype probability data was run through BEAGLE [45] and all markers with an R^2 value (squared correlation between the true and imputed allele dosages) below 0.9 were removed from the original sequence data. This was done in order to remove poorly imputed markers that might have adverse effects on the imputation procedures.

Imputation from 54 k to HD genotypes to HD and imputation to the WGS level were undertaken with IMPUTE2 v2.3.1 [46] and Minimac2 [47], respectively. The imputation to whole genome sequence was done in chunks of 5 Mb with the length of buffer region of 0.25 Mb on either side. A total of 22,751,039 biallelic variants were present in the imputed sequence data. After excluding SNP with a minor allele frequency below 1% or with large deviation from Hardy–Weinberg proportions ($P < 1.0^{-6}$), 15,355,402 SNPs on 29 autosomes in Nordic Holstein cattle were retained for association analyses. The average accuracy (r^2-values from Minimac2) was 0.85 for across breed imputation. Information on the distribution of imputation accuracy as a function of minor allele frequency has previously been published [42].

The methodology of multiple QTL detection

We developed an analysis approach to run the conditional GWAS analysis, similar to the *GCTA-COJO* approach in GCTA [11]. However, GCTA-COJO uses GWAS summary data while we have reanalyzed the data after fitting only the lead SNP(s) on a chromosome. Furthermore, we used imputed dosage data instead of number of copies of the reference allele. This takes account of inaccuracies in genotype imputation. We first performed a single SNP GWAS analysis using GCTA [11] for each chromosome as the first round. Then we ranked the SNP based on their $-\log_{10}P$ value in the GWAS. The SNP with the largest $-\log_{10}P$ value, the lead SNP, within each chromosome was identified. An experiment-wise 0.05 type I error rate after Bonferroni correction for 15,335,402 simultaneous tests corresponds to a threshold of $-\log_{10}P \approx 8.5$. If the $-\log_{10}P$ value of the lead SNP exceeded 8.5; we extracted the lead SNP's genotype dosage, fitted it as a covariate, and scanned the whole chromosome again as the second round. If the result of second round detected another SNP with a $-\log_{10}P$ value exceeding 8.5 and this SNP also was significant in the first round ($-\log_{10}P > 8.5$), we extracted the

Cai *et al. BMC Genetics* (2018) 19:30

Page 10 of 12

allele dosage of this SNP and fixed it as another covariate and scanned the chromosome in a third round. This same procedure was iterated until no additional SNP remained significant. The lead SNP in each round were collected to build a lead SNP list. Moreover, in each round solo SNP, that is, SNP with no other significant SNP within a 1 Mb region were removed. A boundary for each QTL peak was defined as follows: for each QTL, we scanned the 1 Mb region up- and down-stream of each lead SNP, if SNP $-\log_{10}P$ value decreased by more than 3 units compared to the value at the leading SNP and the region is larger than 0.25 Mb we set this SNP as a boundary, otherwise we set ±0.25 Mb as the boundary. The list of candidate genes were generated from the closest annotated genome feature to the lead SNP list.

Testing the type I error rate using simulation data

We used simulated phenotype data to test whether our approach to detecting multiple QTL on a chromosome by incorporating previously identified QTL as covariates, inflates the type I error rates [12]. We selected a SNP randomly from the genome as a causative mutation (QTN) with a MAF (Minor Allele Frequency) between 0.05 and 0.10 and in Hardy Weinberg equilibrium. Ten additional SNP with different levels of LD (linkage disequilibrium, r^2) with the simulated QTN were selected. These 10 SNPs have different r^2 with the QTN as follows: one with 0.9–1, one with 0.8–0.9, one with 0.8–0.7, one with 0.7–0.6, one with 0.6–0.5, one with 0.5–0.4, one with 0.4–0.3, one with 0.3–0.2 and two with less than 0.2. Allele substitution effects at the QTL were sampled from a univariate normal distribution with mean of 20% of the standard deviation of phenotype and variance equal to 1% of the phenotypic variance. We repeated this simulation and applied our analysis 100 times. Lastly, we investigated how many times we found a SNP in LD with the simulated QTN after we fix the simulated causative mutation as a covariate i.e. false positive detection.

LD calculation and annotation

We calculated the pairwise r^2 between lead SNP and all other SNPs on the same chromosome using PLINK [48] and extracted all the SNPs which have $r^2 > 0.2$ with the lead SNP. All these SNPs were annotated by VEP (Variant Effect Predictor) [14]. To find the candidate genes, we extracted all the genes which overlap with LD regions of the lead SNP and searched these gene entries in the Mammalian Phenotype database [24]. We collected all the lead SNPs and calculated the pairwise r^2 with SNPs in the chromosome. The boundary was set to the last SNP that has $r^2 > 0.2$. Then we extracted all the genes overlapping these regions and searched them in the database. We found 411 genes located in the LD regions, of which 375 have

gene symbols. These 375 genes were searched in the database and 364 have mutation lines with phenotype descriptions in the Mammalian Phenotype database. We refined results using two terms for phenotypes: 'abnormal in mammary gland development' and 'abnormal in milk production'.

The genetics variants explained by QTL

We used the lead SNP list to generate the genetic relationship matrix (GRM) as group 1. Then we excluded all flank 2.5 Mb SNPs of the lead SNP from the imputed HD data to generate GRM as group 2. At last, we estimated variance explained by these two groups for each trait. The whole analysis was conducted using GCTA [11].

Additional file

Additional file 1: Figure S1. The locuszoom figure of previous report causative mutation of *DGAT1* of the genome-wide association result in fat content of milk in Nordic Holstein cattle and **Figure S2**. The locuszoom figure of previous report causative mutation of *DGAT1* of the genome-wide association result in fat content of milk in Nordic Holstein cattle. *BOP1* was not include in USCS refFlat file. (DOCX 551 kb)

Cai *et al. BMC Genetics* (2018) 19:30

Page 11 of 12

References
1. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46(8):858–65.
2. Consortium GP. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.
3. Kruglyak L. The road to genome-wide association studies. Nat Rev Genet. 2008;9(4):314.
4. Visscher PM. Sizing up human height variation. Nat Genet. 2008;40(5):489–90.
5. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, JA L, Mägi R. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010;42(11):937–48.
6. Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med. 2009;360(17):1759–68.
7. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48(5):481–7.
8. Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G, Holm H, Kong A, Thorsteinsdottir U, Sulem P. Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nat Genet. 2016;48(3):314.
9. Nishizaki SS, Boyle AP. Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. Trends Genet. 2017;33(1):34–45.
10. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017;18(10):599-612.
11. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
12. Sahana G, de Koning DJ, Guldbrandtsen B, Sørensen P, Lund MS. The efficiency of mapping of quantitative trait loci using cofactor analysis in half-sib design. Genet Sel Evol. 2006;38(2):167.
13. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 2002;12(2):222–31.
14. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
15. Raven L-A, Cocks BG, Hayes BJ. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC Genomics. 2014;15(1):62.
16. Antonicka H, Mattman A, Carlson CG, Glerum DM, Hoffbuhr KC, Leary SC, Kennaway NG, Shoubridge EA. Mutations in COX15 produce a defect in the mitochondrial heme biosynthetic pathway, causing early-onset fatal hypertrophic cardiomyopathy. Am J Hum Genet. 2003;72(1):101–14.
17. Schennink A, Heck JM, Bovenhuis H, Visker MH, van Valenberg HJ, van Arendonk JA. Milk fatty acid unsaturation: genetic parameters and effects of stearoyl-CoA desaturase (SCD1) and acyl CoA: diacylglycerol acyltransferase 1 (DGAT1). J Dairy Sci. 2008;91(5):2135–43.
18. Sigl T, Meyer H, Wiedemann S. Gene expression of six major milk proteins in primary bovine mammary epithelial cells isolated from milk during the first twenty weeks of lactation. Czech J Anim Sci. 2012;57(10):469–80.
19. Do D, Bissonnette N, Lacasse P, Miglior F, Sargolzaei M, Zhao X, Ibeagha-Awemu E. Genome-wide association analysis and pathways enrichment for lactation persistency in Canadian Holstein cattle. J Dairy Sci. 2017;100(3):1955–70.
20. Rahmatalla SA, Müller U, Strucken EM, Reissmann M, Brockmann GA. The F279Y polymorphism of the GHR gene and its relation to milk production and somatic cell score in German Holstein dairy cattle. J Appl Genet. 2011;52(4):459–65.
21. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, Couldrey C, Keehan M, Sherlock RG, Harland C, et al. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. Sci Rep. 2016;6:25376.
22. Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee J-H, Drackley JK, Band MR, Hernandez A, Shani M. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 2005;15(7):936–44.
23. Olsen HG, Nilsen H, Hayes B, Berg PR, Svendsen M, Lien S, Meuwissen T. Genetic support for a quantitative trait nucleotide in the ABCG2 gene affecting milk composition of dairy cattle. BMC Genet. 2007;8(1):32.
24. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA. The mouse genome database (MGD): mouse biology and model systems. Nucleic Acids Res. 2008;36(suppl_1):D724–8.
25. Zhou J, Chehab R, Tkalcevic J, Naylor MJ, Harris J, Wilson TJ, Tsao S, Tellis I, Zavarsek S, Xu D. Elf5 is essential for early embryogenesis and mammary gland development during pregnancy and lactation. EMBO J. 2005;24(3):635–44.
26. Bionaz M, Loor JJ. Gene networks driving bovine mammary protein synthesis during the lactation cycle. Bioinf Biol Insights. 2011;5(BBI):S7003.
27. Consortium U. Reorganizing the protein space at the universal protein resource (UniProt). Nucleic Acids Res. 2011;40:D71–5.
28. Cao Y, Bonizzi G, Seagroves TN, Greten FR, Johnson R, Schmidt EV, Karin M. IKKα provides an essential link between RANK signaling and cyclin D1 expression during mammary gland development. Cell. 2001;107(6):763–75.
29. Praskova M, Xia F, Avruch J. MOBKL1A/MOBKL1B phosphorylation by MST1 and MST2 inhibits cell proliferation. Curr Biol. 2008;18(5):311–21.
30. Chan EH, Nousiainen M, Chalamalasetty RB, Schäfer A, Nigg EA, Sillje HH. The Ste20-like kinase Mst2 activates the human large tumor suppressor kinase Lats1. Oncogene. 2005;24(12):2076.
31. Thévenin A, Ein-Dor L, Ozery-Flato M, Shamir R. Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. Nucleic Acids Res. 2014;42(15):9854–61.
32. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369–75.
33. Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, Tier B, Savin K, Hayes BJ, Goddard ME. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLoS Genet. 2014;10(3):e1004198.
34. Bolormaa S, Hayes BJ, van der Werf JH, Pethick D, Goddard ME, Daetwyler HD. Detailed phenotyping identifies genes with pleiotropic effects on body composition. BMC Genomics. 2016;17(1):224.
35. Nilsen H, Olsen H, Hayes B, Nome T, Sehested E, Svendsen M, Meuwissen T, Lien S. Characterization of a QTL region affecting clinical mastitis and protein yield on BTA6. Anim Genet. 2009;40(5):701–12.
36. Li H, Wang Z, Moore S, Schenkel F, Stothard P. Genome-wide scan for positional and functional candidate genes affecting milk production traits in Canadian Holstein cattle. In: Proc 9th WCGALP. Leipzig: Germany; 2010. p. 26.
37. Lipkin E, Straus K, Stein RT, Bagnato A, Schiavini F, Fontanesi L, Russo V, Medugorac I, Foerster M, Sölkner J. Extensive long-range and nonsyntenic linkage disequilibrium in livestock populations: deconstruction of a conundrum. Genetics. 2009;181(2):691–9.
38. Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be? Nucleic Acids Res. 2016;44(13):6046–54.
39. Goddard M. A method of comparing sires evaluated in different countries. Livest Prod Sci. 1985;13(4):321–31.
40. Schaeffer L. Model for international evaluation of dairy sires. Livest Prod Sci. 1985;12(2):105–15.
41. Iso-Touru T, Sahana G, Guldbrandtsen B, Lund MS, Vilkki J. Genome-wide association analysis of milk yield traits in Nordic red cattle using imputed whole genome sequence variants. BMC Genet. 2016;17(1):55.
42. Wu X, Guldbrandtsen B, Lund MS, Sahana G. Association analysis for feet and legs disorders with whole-genome sequence variants in 3 dairy cattle breeds. J Dairy Sci. 2016;99(9):7221–31.

Cai *et al. BMC Genetics* (2018) 19:30

Page 12 of 12

43. Brøndum RF, Rius-Vilarrasa E, Strandén I, Su G, Guldbrandtsen B, Fikse W, Lund MS. Reliabilities of genomic prediction using combined reference data of the Nordic red dairy cattle populations. J Dairy Sci. 2011;94(9):4700–7.

44. Brøndum RF, Guldbrandtsen B, Sahana G, Lund MS, Su G. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics. 2014;15(1):728.

45. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.

46. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3: Genes|Genomes|Genetics. 2011;1(6):457–70.

47. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2015;31(5):782–4.

48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–75.

49. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 2009;10(4):R42.