J. Dairy Sci. 106:7799–7815
https://doi.org/10.3168/jds.2022-23095

Identification of candidate novel production variants on the *Bos taurus* chromosome X

H. Trebes,* Y. Wang, E. Reynolds, K. Tiplady, C. Harland, T. Lopdell, T. Johnson, S. Davis, B. Harris, R. Spelman, and C. Couldrey
Research and Development, Livestock Improvement Corporation, Hamilton 3240, New Zealand

ABSTRACT

Chromosome X is often excluded from bovine genetic studies due to complications caused by the sex specific nature of the chromosome. As chromosome X is the second largest cattle chromosome and makes up approximately 6% of the female genome, finding ways to include chromosome X in dairy genetic studies is important. Using female animals and treating chromosome X as an autosome, we performed X chromosome inclusive genome-wide association studies in the selective breeding environment of the New Zealand dairy industry, aiming to identify chromosome X variants associated with milk production traits. We report on the findings of these genome-wide association studies and their potential effect within the dairy industry. We identify missense mutations in the *MOSPD1* and *CCDC160* genes that are associated with decreased milk volume and protein production and increased fat production. Both of these mutations are exonic SNP that are more prevalent in the Jersey breed than in Holstein-Friesians. Of the 2 candidates proposed it is likely that only one is causal, though we have not been able to identify which is more likely.
Key words: X chromosome, genome-wide association study, genomic selection, milk production

INTRODUCTION

Genome-wide association studies have become a popular method for examining genomes and identifying QTL and causal variants for target phenotypes. Until recently, chromosome X (**ChrX**) has been largely excluded from these GWAS due to the complications it introduces. In cattle, the inclusion of ChrX could be particularly important because ChrX is the second largest chromosome, representing approximately 6% of the genome in females. In Taurine cattle (*Bos taurus*) ChrX is a large chromosome (139 Mb), containing 1,222 genes, with a gene/Mbp ratio of 8.79 (Czech et al., 2020; Rosen et al., 2020). The inclusion of ChrX in genetic studies is important because of the core principles and methods of genetic improvement in the New Zealand dairy industry. For dairy farmers, while many factors contribute to profit generated from cows, phenotypes associated with milk production are some of the most important.

Expression of genes from ChrX is more complicated than from autosomes, and unlike ChrX inheritance patterns, dosage compensation is hard to predict. Females of XY sex-determined species undergo ChrX inactivation (**XCI**) to compensate for the female having 2 whole copies of ChrX while the male has only one. The XCI is the process whereby one of the female's ChrX is condensed to the point where (most of) the genes on the chromosome are not expressed (Heard et al., 1997). In the past this was believed to be a random process, with a 50/50 chance in every cell of either the maternal or paternal ChrX being inactivated. However, research shows that there are several instances where, for a variety of reasons, this is not the case (Hatakeyama et al., 2004; Migeon, 1998). For example, ChrX imprinting occurs in certain cell and tissue types, meaning that in that cell type only the maternal or paternal copy of certain genes on ChrX can be active (Sado and Ferguson-Smith, 2005). It has also been shown that in some tissues (e.g., mammary), one or the other ChrX is favored depending on the inactivation status of the original blastocyst cell(s) which that tissue grew from (Couldrey et al., 2017). While XCI applies across the entire chromosome and is maintained during cell division, some genes escape inactivation and are expressed even when they are on the inactivated copy of ChrX, resulting in females potentially having a double dose of ChrX product from that region compared with males. Patterns of escape from XCI are not equivalent across species or even across an individual animal, and certain genes are more pre-disposed to it than others, including genes in the ChrX pseudoautosomal region and genes with Y-homologs (Posynick and Brown, 2019).

Received December 1, 2022.
Accepted April 26, 2023.
*Corresponding author: hannah.trebes@lic.co.nz

In spite of the challenges in working with ChrX genotypes, specialized software toolkits similar to XWAS (Gao et al., 2015) have been developed to facilitate genetic studies of ChrX. We make use of genetic analysis programs such as PLINK (Purcell et al., 2007) and BOLT-LMM (Loh et al., 2018) which have had features added to allow them to cope with or at least recognize sex chromosome input.

GWAS of the cattle genome have been used to identify (among other things) milk composition QTL (Littlejohn et al., 2016, p. 1), milk production candidate genes and loci (Liu et al., 2020), novel cattle syndromes (Reynolds et al., 2021), candidate white spotting mutations (Jivanji et al., 2019), fertility variants in US Holstein Dairy bulls (Pacheco et al., 2020), and recurrent infertility in Japanese Black cattle (Arishima et al., 2017). Due to the 1,222 genes encoded on ChrX, it was expected that the chromosome would likely play a role in the expression of dairy production phenotypes. The recent imputation of DNA sequence variants on ChrX by Wang et al. (2021) allowed us to include ChrX in large scale sequence level GWAS and other genetic analyses.

Our main aim was to determine whether any significant relationships existed between ChrX variants and milk production phenotypes. We also wanted to determine whether a representative GWAS including ChrX could be achieved when separate GWAS were performed for individual breeds. The results of this study have the potential to affect and influence the breeding objectives and methods of the NZ dairy industry.

MATERIALS AND METHODS

No human or animal subjects were used, so this analysis did not require approval by an Institutional Animal Care and Use Committee or Institutional Review Board.

Study Population

The animals examined in this study are derived from a set of 91,872 female cattle born in New Zealand between 1990 and 2018. We considered animals from 3 breeds; Holstein-Friesian, Jersey, and crosses of these 2 breeds (henceforth referred to as **HFxJ**). To classify breeds, we used the NZ animal evaluation classification (DairyNZ, 2021) based on breed 16ths from a 4-generation family pedigree. Animals with at most 13/16ths of either Holstein-Friesian or Jersey are classified as HFxJ (n = 41,218). When specifically analyzing purebred Holstein-Friesian (n = 15,835) or purebred Jersey (n = 11,734) animals we are examining the "purebred" animals with 16/16 of their pedigree belonging to one breed. This leaves animals with 14/16ths or 15/16ths of their pedigree belonging to Holstein-Friesian (n = 3,186) or Jersey (n = 1,672) which are included in the analyses of all animals but not in that of purebred animals. It is worth noting that in the case of the 16/16ths purebred animals this pedigree-based breed classification does not necessarily represent a genetic purebred. Due to the speed of generation turnover in the NZ dairy industry an animal could very conceivably have all animals on its 4-generation pedigree belong to one breed classification, but a previous crossbred could have contributed DNA that persisted through the generations. A full breakdown of the number of animals with data for each phenotype can be seen in Supplemental Table S1 (https://data.mendeley.com/datasets/5yzy4p5zr5/1; Trebes, 2023).

Genotypes

Consolidation of SNP-Chip Panels for Sequence Imputation. According to Mao et al. (2016), the imputation accuracy of the X chromosome was improved if the pseudoautosomal (**PAR**) region was treated as autosomal and both non-PAR and PAR regions were imputed separately. Thus, in all sets, genotype data were separated into non-PAR (ChrX: 0–133300518) and PAR (ChrX:133300518–139009144) regions (Johnson et al., 2019).

The genotyping panels were grouped into 3 sets: GGP panels (GGPv3, GGPv3.1, and GGPv4, totaling 1,577 ChrX variants), 50K panels (BovineSNP50v1 and BovineSNP50v2, totaling 1,177 ChrX variants), and GGP50k panels (GGP50kv1 and GGP50kv1.1, totaling 2,275 ChrX variants). All 3 sets were imputed separately to the BovineHD genotype panel, which contained 3,769 animals genotyped on Illumina BovineSNPHD Beadchip (23,539 ChrX variants). The imputed and physically genotyped panels were then combined and imputed to sequence resolution using the sequence-based imputation reference population described below. Beagle 5.1 (Browning et al., 2018) was used for all phasing and imputation steps with the default parameters except for the effective population size that was set at 400, and a window size of 20 Mb.

Reference Population for Sequence-Based Imputation. Whole-genome sequence data were available for 1,298 animals which include 306 Holstein-Friesian, 219 Jersey, and 717 HFxJ crossbred animals or other breeds and crossbreeds (n = 56). The raw sequence data were mapped to the bovine genome build ARS-UCD1.2 bovine reference genome (Rosen et al., 2018) using BWA-MEM algorithm (Li, 2013). The average sequence coverage across the complete reference population was 15.17 fold. 850 animals with a depth above

10 were treated as high-depth sequence animals. The detected variants were then filtered using BCFtools (Li, 2011) and Plink 1.9 for high-depth sequenced animals, retaining variants with allele count of 2 (-i 'AC > 1', '–max-alleles 2'), variants with map quality higher than 50 (-i 'MQ > = 50'), and mendelian error rate below 5%. In the end, singletons and nonvariants were removed again which resulted in 589,484 variants for the X chromosome, in which 492,072 variants are located in the non-PAR region and 97,412 are located in the PAR region. The genotypes at the positions of those filtered variants were then extracted from the sequence data of all 1,298 animals and were phased using the software Beagle 5.1 to generate the sequence-based imputation reference panel.

The Final Genotype Set. A final stage of quality control was performed on the imputed genotypes, where all SNP with a minor allele frequency (**MAF**) below 0.005 or dosage R^2 below 0.9 were removed to bring ChrX more in-line with the autosomes. This resulted in a final imputed set of 16,122,289 autosomal variants and 368,647 ChrX variants. As we are examining only female animals in this study, we were able to re-join the imputed non-PAR and PAR and treat X as an autosome in our analyses.

Phenotypes

We examined the association between ChrX SNP and 12 production phenotypes. Throughout this report the nomenclature for these phenotypes is made up of 2 parts. First is the type of phenotype, this can be "volume" referring to milk volume (liters), "fat" referring to milk fat yield (kilograms), or "protein" referring to milk protein yield (kilograms). The trait is followed by the lactation (parity) denoted by the numbers "1" through to "4" where lactations 1, 2, and 3 represent lactations following a cow's first, second or third calf, while "lactation 4" is a binned aggregate of the fourth lactation and all latter lactations.

For each phenotype being studied, the animals with a phenotype record were selected from the total set of all valid genotyped animals (not all animals had records for all phenotypes, see Supplemental Table S1). Phenotype values were adjusted from the raw yield deviation based on the methods described in Reynolds et al. (2021). Briefly, the raw yield deviation phenotypes were adjusted using the national (New Zealand) genetic evaluation models. This method includes adjustments for age at calving, stage of lactation, record type (lactation traits may be recorded as a.m. milkings, p.m. milkings, or both), effect of induced calving, and pairwise heterosis between breeds.

GWAS Preparation, Process, and Parameters

GWAS were performed using the BOLT-LMM package (version 2.3.2, **BOLT-LMM**) using a leave one segment out approach, with the size of the segment to be left out set to 5Mbp, as in Tiplady et al. (2021). Infinitesimal mixed model association statistics were evaluated to assess the additive effect of each SNP. A set of autosomal SNP to be used in the BOLT-LMM mixed model (parameter: modelSnps) to account for population structure in the GWAS was generated based on the 50K SNP-chip imputation reference as described in Tiplady et al. (2021). To control genomic inflation ($\boldsymbol{\lambda}$) and account for population structure on ChrX we selected 2,222 ChrX SNP to include in the modelSnps file. These ChrX SNP consist of 698 SNP from the Illumina 50K SNP-chip reference with a MAF >0.1 in the reference population, this was supplemented with 1,524 ChrX SNP from the imputed variants with a MAF >0.1. The number of SNP to include from ChrX (2,222) was chosen based on the size of ChrX being similar to the size of chromosome 2 which was represented by 2,221 SNP in the modelSnps file.

All phenotypes were first analyzed with a BOLT-LMM GWAS including all animals with no fixed effects applied, henceforth the (12) base GWAS. A second GWAS was then run for each phenotype using BOLT-LMM parameters to apply breed as a fixed effect. The purebred breed classification was also used to separate Holstein-Friesians and Jerseys and perform a third GWAS for each purebred breed with no fixed effects.

All phenotypes were then run through iterative GWAS as described in Tiplady et al. (2021) to attempt to distinguish between multiple QTL. Briefly, the iterative GWAS ran multiple rounds of GWAS for a phenotype and in each round included the most significant ChrX SNP from the previous GWAS round(s) as a fixed effect, this continued until there were no longer any significant ($P <$ 5E-8) association effects observed on ChrX. The first iteration used the output of the base GWAS described above as input, i.e., iteration 1 occurred after the base GWAS (see Table 1) and accounted for the most significant ChrX SNP from that GWAS.

It must be noted that the GWAS significance threshold referred to throughout this study is at a *P*-value of 5E-8 and that this is not necessarily the best significance threshold for all or any of these phenotypes. The 5E-8 significance threshold was proposed by Risch and Merikangas (1996) for human GWAS and it has previously been applied to cattle studies (Jivanji et al., 2019; Xiang et al., 2021). We have used this *P*-value significance threshold to allow easy com-

Table 1. Most significant chromosome X (ChrX) and production peak SNP from the GWAS of all animals, for each phenotype with no fixed effects (base GWAS)[1]

Phenotype	ChrX SNP position	MAF	GWAS SNP *P*-value	Gene	Class
Fat lactation 1	34002922	0.00863	3.30E-19	Intergenic	
Fat lactation 1*	18479617	0.04707	4.60E-12	FAM122B	Intronic
Fat lactation 2	34002922	0.00936	1.50E-11	Intergenic	
Fat lactation 2*	18475739	0.0496	8.60E-10	FAM122B	Intronic
Fat lactation 3	34002922	0.00938	2.90E-13	Intergenic	
Fat lactation 3*	18501994	0.06073	2.90E-13	Intergenic	
Fat lactation 4	18475739	0.05143	6.20E-19	FAM122B	Intronic
Volume lactation 1	18584126	0.0469	1.50E-23	MOSPD1	Exonic
Volume lactation 2	18584126	0.04929	8.00E-37	MOSPD1	Exonic
Volume lactation 3	18584126	0.04976	3.00E-32	MOSPD1	Exonic
Volume lactation 4	18476384	0.05708	5.00E-20	FAM122B	Intronic
Protein lactation 1	34002922	0.00863	4.30E-21		Intergenic
Protein lactation 1*	18584126	0.0469	2.90E-16	MOSPD1	Exonic
Protein lactation 2	18348881	0.05476	5.90E-25		Intergenic
Protein lactation 3	18584126	0.04976	2.70E-18	MOSPD1	Exonic
Protein lactation 4	16697687	0.01076	1.00E-12	HS6ST2	Intronic
Protein lactation 4*	18476384	0.05708	1.30E-10	FAM122B	Intronic

[1]Some phenotypes are shown on 2 lines, one with * to denote that line as the most significant production peak SNP, in these cases the SNP shown was either the second or third most significant ChrX SNP. Gene and class values determined from manual inspection of SNP location using IGV (Thorvaldsdóttir et al., 2013). Minor allele frequency shown in MAF column rounded to 5 decimal places. The MAF is specific to the phenotype or GWAS run and is not necessarily equal to the MAF in all imputed animals.

parison with other publications, but this was always used in conjunction with a visual assessment of the GWAS results.

Analysis of Areas of Interest

Regions of interest identified by the GWAS were examined and analyzed with a variety of approaches as described below. As the production phenotypes of milk volume, fat and protein are known to be highly correlated we employed a somewhat holistic approach to classify our candidate SNP of interest. For a SNP to be classified as a candidate it had to meet the below conditions across the phenotypes and lactations (though not necessarily in all phenotypes and lactations). First the SNP must be within the region of interest identified by the GWAS and must have a significant association in several of the 12 phenotypes. Second the SNP must be in high linkage disequilibrium (LD $R^2 > 0.85$) with the SNP at the top of the production peak. Finally, the SNP must have a deleterious effect as classified by the Ensembl Variant Effect Predictor (described below).

To further assess the quality of the imputed sequence around regions of interest, and to confirm that the SNP we saw were not due to genotyping or sequencing artifacts we examined the region of interest using the Integrated Genomics Viewer (**IGV**) (Thorvaldsdóttir et al., 2013). We know from other research in our group that there are regions of the genome that are poorer quality which could potentially have been enriched by the genotyping and imputing to sequence pipelines. Therefore, inspecting the genome with IGV is standard practice for us to mitigate these risks. The IGV was also used to identify recombinants.

We then used the Ensembl Variant Effect Predictor (**VEP**) tool (McLaren et al., 2016) to analyze the potential effects of the SNP around and including the SNP of interest. The VEP was also used in this manner to access the SIFT score (Vaser et al., 2016) for SNP to assess whether a SNP was likely to have a deleterious effect. Homologene (Sayers et al., 2021) was used to assess the homology and pairwise alignment scores of the genes of interest. SNAP2 scores (Hecht et al., 2015) were used to assess the effect of sequence variants on the resulting protein.

While GWAS detect associations between all genotypes and a phenotype (similar to an ANOVA), we were also interested in whether there are differences in phenotype between specific genotypes. For example, knowing if there is a significant difference between heterozygotes and either homozygote could indicate potential escape from ChrX inactivation. To determine if these significant differences between genotypes exist we used the R function "TukeyHSD" to calculate Tukey Honest Significant Differences (**Tukey HSD**; Miller, 1981; Yandell, 2017) based on ANOVA results generated by the R "aov" function (Chambers et al., 1992). These tests were performed for all animals, and for the purebred subsamples separately.

To further test whether the SNP of interest could be causing the production peak we performed a round of iterative GWAS for each phenotype, with one of the SNP of interest forced to be the first SNP accounted for as a fixed effect.

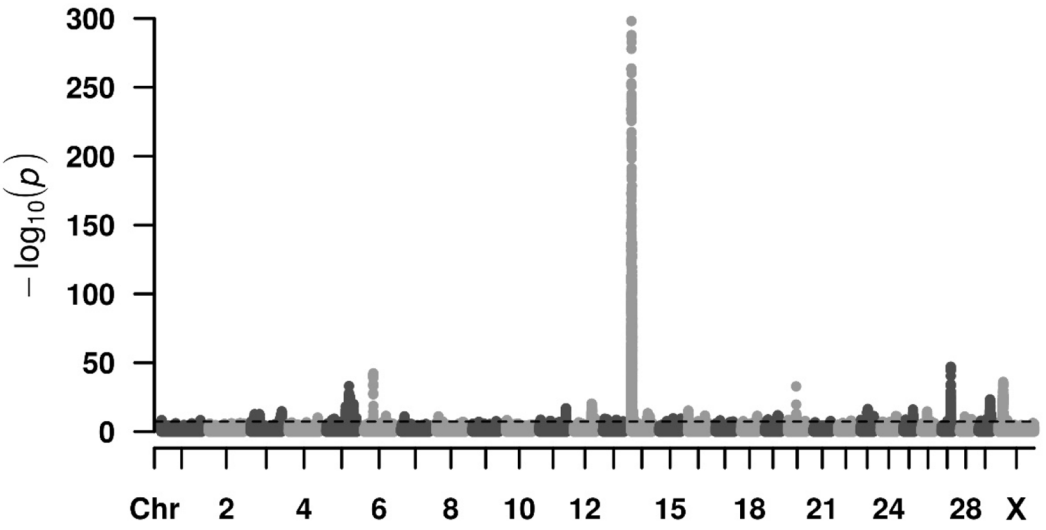

Figure 1. Manhattan plot of volume lactation 2 no-fixed effect GWAS *P*-values. ChrX is shown in light gray on the far right. GWAS performed with BOLT-LMM including animals of all breeds. Plot generated using R CMplot package. This phenotype was chosen to exemplify the ChrX production peak as it has the most significant *P*-value of the no-fixed effect all-animals GWAS and is strongly related to both the other phenotypes.

As part of investigating genes of high interest, we used the same bovine lactating mammary RNaseq data set (n = 423) as Davis et al. (2022) to perform a GWAS of ChrX for expression phenotypes (eQTL analysis) of genes of interest using the methods described in Prowse-Wilkins et al. (2022) employing GCTA (Yang et al., 2011) leave one chromosome out analysis. All animals used for the eQTL analysis were female, and all 3 breeds were represented.

All SNP named in this study are listed in Supplemental Table S2 (https://data.mendeley.com/datasets/5yzy4p5zr5/1; Trebes, 2023) along with their RefSNP numbers (Cezard et al., 2022) if applicable.

RESULTS

Genome-Wide Association Studies

Our GWAS results showed significant associations between ChrX SNP and all phenotypes and lactations of the 12 base GWAS. In all the base GWAS a peak can be seen centered roughly around ChrX:18000000 (see Figures 1 and 2). The peak formed in this region is henceforth referred to as the "production peak." The full range of most significant production peak SNP from these base GWAS can be seen in Table 1. Of these 12 base GWAS, the most significant production peak SNP falls in an exon of MOSPD1 5 times (always ChrX:18584126C > G), an intron of FAM122B 5 times (see Table 1), and at 2 separate intergenic loci (ChrX:18501994A > C, ChrX:18348881C > T). The *P*-values for the most significant production peak SNP ranged between 1.3E-10 (protein lactation 4) and 8E-37 (Volume lactation 2). There were 3 SNP that were significant in all 12 of the base GWAS, 2 of which were intergenic (ChrX:18501994A > C and ChrX:18504680C > T) and one within an exon of MOSPD1 (ChrX:18584126C > G). This is likely a result of high LD between these SNP as the 2 intergenic SNP (ChrX:18501994A > C and ChrX:18504680C > T) have an LD R^2 of 0.997 and the same MAF of 0.057, the LD R^2 between both these SNP and ChrX:18584126C > G is 0.803 (3dp).

In the base GWAS for fat lactations 1 to 3 and protein lactation 1 we observed that the most significantly associated SNP (ChrX:34002922T > C, MAF = 0.009) appears on its own with no other SNP forming a peak up to it (see Figure 2). The highest LD observed between this SNP (ChrX:34002922T > C) and any other was $R^2 = 0.04$. The ChrX:34002922T > C was significant ($P < 5E\text{-}8$) in 11 of the 12 base GWAS but did not show a peak shape in any. A similar SNP was observed topping the production peak for protein lactation 4 at ChrX:16697687C > T. While this SNP is within the bounds of the production peak we have ruled it out from being a true part of the production peak and having an effect on the GWAS itself. This is due to it being in very low LD (highest R^2 with any SNP is 0.002) and having little effect on the iterative GWAS as explained later. For these reasons we have ruled out these SNP (ChrX:34002922T > C and ChrX:16697687C > T) as belonging to the production peak.

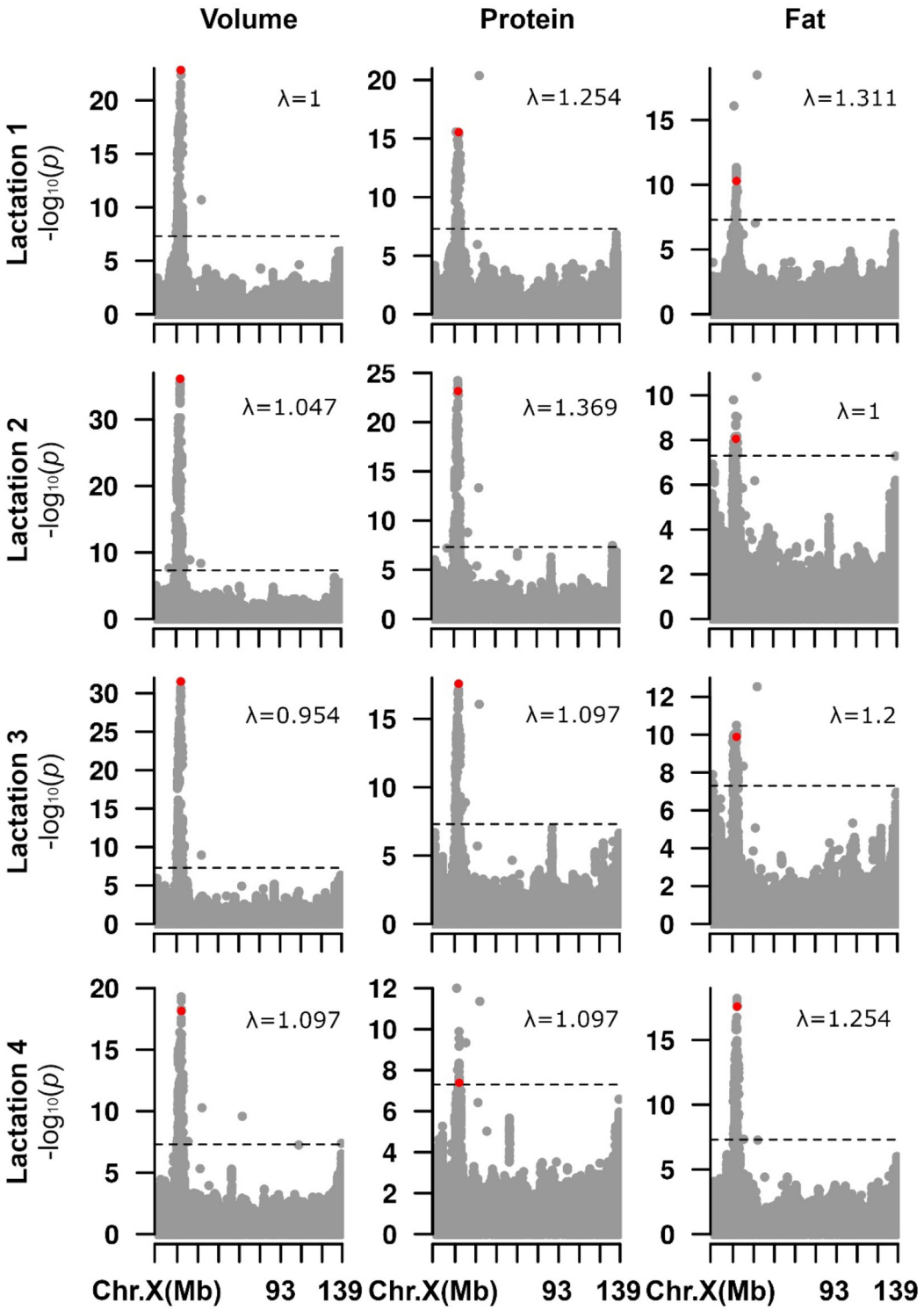

Figure 2. Manhattan plots of ChrX showing no-fixed effect all-animals GWAS *P*-values and genomic inflation parameters (λ) for all phenotypes and lactations. The SNP of interest ChrX:18584126C > G is shown in red to allow easy comparison of peak location. Plots were generated with the R package CMplot.

Figure 1 shows the results of the volume lactation 2 base GWAS across the genome. The Chromosome 14 *DGAT1* peak (Grisart et al., 2002; Spelman et al., 2002) can be clearly seen dominating the genome in Figure 1 (this peak has been truncated for plotting, its true most significant *P*-value is 2.8E-808) with the ChrX production peak (top SNP ChrX:18584126C > G) shown at the far right. Figure 2 shows the *P*-values across ChrX for all lactations or all phenotypes from the base GWAS. The production phenotypes are known to be highly correlated so seeing a common peak across the 3 phenotypes was not surprising.

Iterative GWAS. Performing iterative GWAS indicated that the production peak observed across the Manhattan plots is likely the result of a single QTL. The number of iterations required until no significant ($P < 5E-8$) associations were observed ranged between 2 and 5 and can be seen in Supplemental Table S3 (https://data.mendeley.com/datasets/5yzy4p5zr5/1; Trebes, 2023).

We observed 2 scenarios occurring in the iterative GWAS that enabled us to further define the production peak and surmise that this peak is likely caused by a single QTL. For those phenotypes where the base GWAS most significant ChrX SNP was ChrX:34002922T > C (fat lactation 1 to 3 and protein lactation 1) or ChrX:16697687C > T (protein lactation 4) the first iteration showed the production peak with significance unchanged or increased. In these cases all the most significant iteration 1 SNP were within the production peak in the range ChrX:18475739G > C -ChrX:18584126C > G. For the phenotypes where the base GWAS most significant ChrX SNP was atop the production peak, the first iteration revealed the whole of the production peak had dropped into insignificance and only a few (1–3) lone SNP (no peaks leading up to them) remained above the significance threshold (5E-8) spread across the chromosome. In these cases, the iteration 1 most significant SNP was either ChrX:34002922T > C or ChrX:16697687C > T.

Effect Size of SNP. Figure 3 shows the BOLT-LMM predicted effect size of SNP across ChrX. The SNP found to be strongly associated with volume and protein are predicted (by the GWAS) to have a (relatively) strong negative effect on the phenotype, while the fat effect is predicted to be positive. To provide context, Figure 3 also shows the GWAS predicted effect sizes from chromosome 14 volume lactation 1, as this shows the strong and well-known predicted effects of SNP within the *DGAT1* gene (Grisart et al., 2002; Spelman et al., 2002). The SNP effect sizes shown in Figure 3 are BOLT-LMM predictions of the change to an animals' phenotype for each (active) copy of the alternate allele. Therefore, a positive value indicates a predicted increase in the phenotype while a negative value indicates a decrease.

Covariates. Performing separate GWAS for the purebred breeds showed a clear interaction between breed and the relationship between ChrX and the production phenotypes. Manhattan plots for the GWAS of separate breeds can be seen in Figure 4. We observed that for volume, all lactations showed a clear significant production peak in the Jerseys and a complete absence of this peak in Holstein-Friesians. For protein we observed the tell-tale shape of the production peak across all lactations of Jersey animals; however, this peak was only significant ($P < 5E-8$) in lactations 1 and 2, and again this peak was absent in Holstein-Friesians. For fat the production peak appears to be absent in both breeds for all lactations except in lactation 2 for Holstein-Friesians.

BOLT-LMM flags were also used to run GWAS with breed included in the model as a fixed effect. The results of this can be seen in Supplemental Figure S1 (https://data.mendeley.com/datasets/5yzy4p5zr5/1; Trebes, 2023) and show that there was little difference in the presence and strength of the production peak with or without breed applied as a fixed effect.

Identification of Candidate Causal Variants

Given the significance, novelty, and potential effects of the production peak, we prioritized SNP within this region for further investigation. To identify SNP of interest, we define the production peak as ChrX:15,000,000–20,000,000 as this contains the visible peak on the Manhattan plots (Figure 2) and 98% of all significant ChrX SNP. Of the SNP in this region, 7 were predicted to have deleterious effects (VEP; McLaren et al., 2016), of which 2 (ChrX:18584126C > G and ChrX:18028733A > G, henceforth SNP of interest) were significantly associated with any of the phenotypes. The ChrX:18584126C > G was significantly associated ($P < 5E-8$) with all 12 of the phenotypes base GWAS, and ChrX:18028733A > G was significantly associated ($P <= 5E-8$) with 8 of the 12 phenotypes base GWAS (Volume all lactations, fat lactation 4, and protein lactations 1 to 3).

The SNP at ChrX:18584126C > G is a missense mutation in the third exon (McLaren et al., 2016) of the gene Motile Sperm Domain-Containing protein 1 (*MOSPD1*). The Arginine to Proline substitution at *MOSPD1* AA 96 caused by ChrX:18584126C > G has a very strong SNAP2 (Hecht et al., 2015) signal for effect of 91. Variant Effect Predictor (McLaren et al., 2016) showed that of the 39 SNP in *MOSPD1*, ChrX:18584126C > G was the only predicted deleterious SNP and had a SIFT score of "0.0" indicating maximum

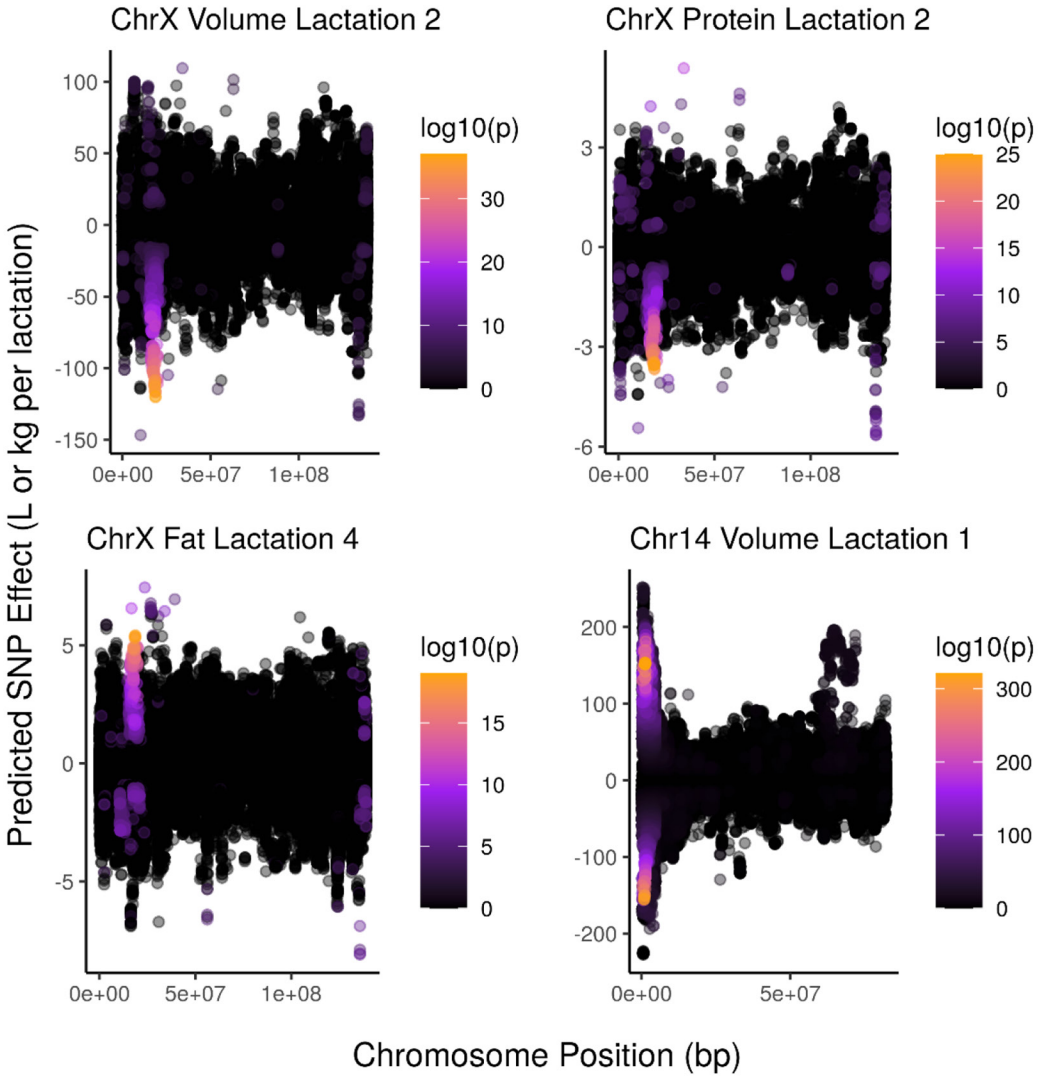

Figure 3. SNP effect sizes predicted by BOLT-LMM for the total GWAS population on ChrX and Chromosome 14 (Chr14). GWAS $\log_{10}$ *P*-value indicated by point color. Chromosome 14 included for comparison with the DGAT1 peak. For each phenotype, the lactation that yielded the most significant ChrX production peak QTL in the base GWAS is shown. Y-axes units are liters/lactation for volume, and kilograms/lactation for fat and protein. A positive effect indicates a predicted increase in phenotype, while a negative effect predicts a decrease in phenotype. Effect is for each active copy of the alternate/minor allele.

confidence in the deleterious prediction (Vaser et al., 2016).

The ChrX:18028733A > G SNP is a missense mutation in the third exon of the gene Coiled Coil Domain-Containing protein 160 (*CCDC160*) (McLaren et al., 2016) and causes a Lysine to Glutamic Acid substitution. The ChrX:18028733A > G SNP is predicted deleterious by VEP (SIFT score = 0.04) with the missense mutation receiving a SNAP2 score of 52 (indicating a strong signal for effect on the resulting protein).

The MAF for all GWAS animals for the ChrX:18028733A > G SNP is 0.101 and for ChrX:18584126C > G is 0.047 (both rounded to 3dp). The MAF of ChrX:18028733A > G and ChrX:18584126C > G in Jerseys are 0.091 and 0.086, respectively, and in Holstein-Friesians are 0.1 and 0.008, respectively. We can see in Figure 5 that the frequency of the alternate alleles for the 2 SNP of interest have been gradually decreasing over time. This coincides with the number of heterozygotes becoming gradually greater through the years compared with the homozygous alternate genotype. The SNP of interest show great variation in their LD between the breed populations. In all the GWAS animals, the LD R^2 between these SNP is 0.393. The LD R^2 between the SNP of interest in the purebred Jersey animals is 0.89 while for Holstein-Friesians the R^2 is 0.055. Table 2

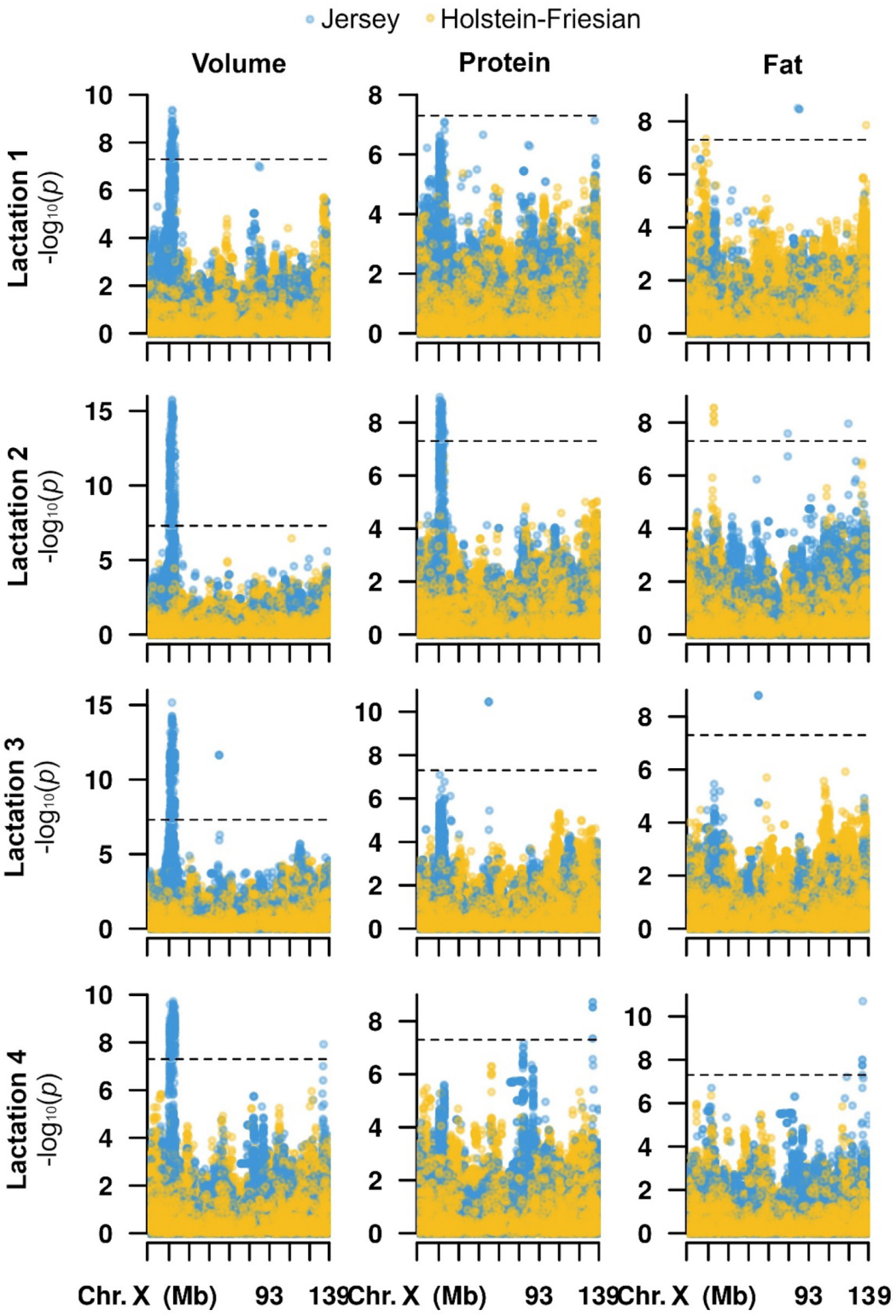

Figure 4. Manhattan plots of ChrX GWAS *P*-values for all phenotypes and lactations when separate GWAS were run for each of purebred Holstein-Friesians and purebred Jerseys. Jersey SNP shown in blue; Holstein-Friesian SNP in yellow.

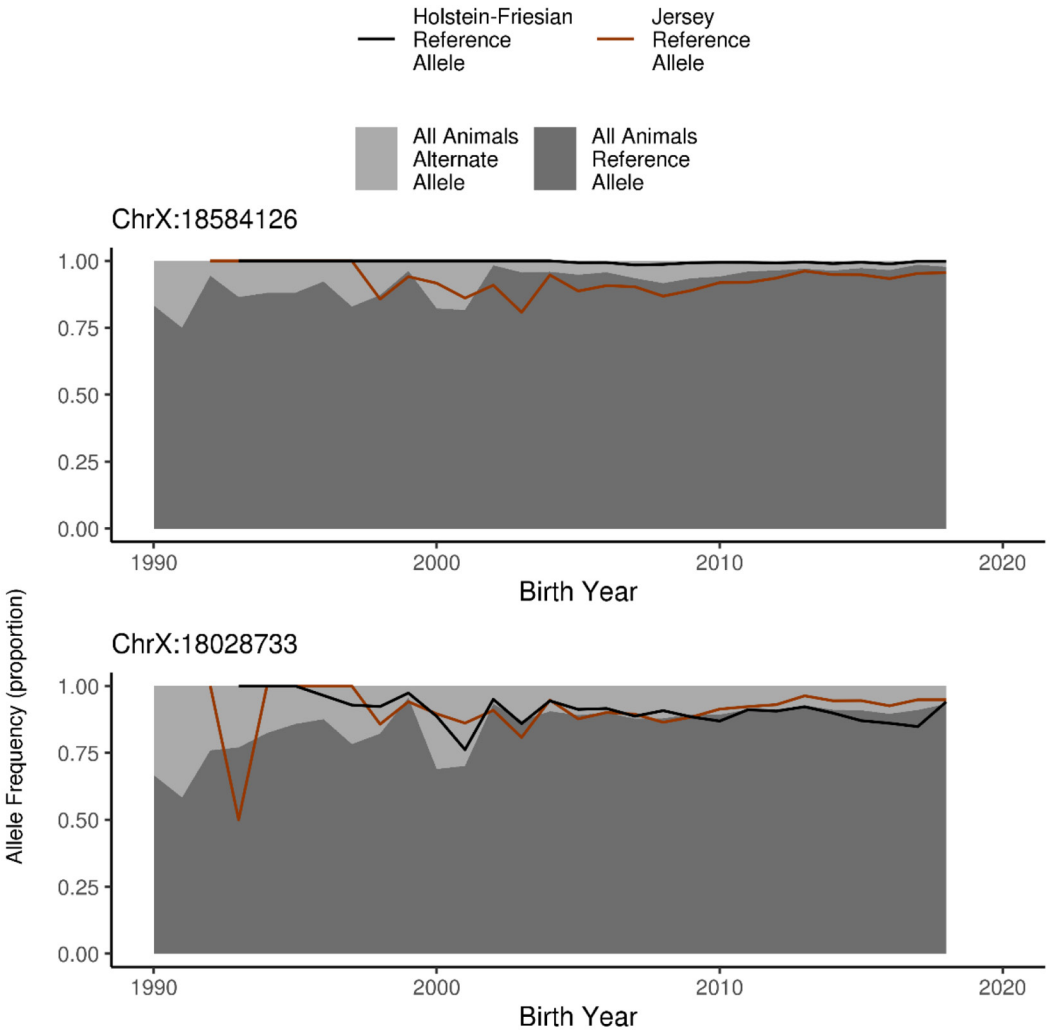

Figure 5. Allele frequency (as a proportion of the GWAS population) for SNP of interest (ChrX:18584126C > G and ChrX:18028733A > G) over time (1990–2018). Filled sections of the plots show the allele frequencies in the total GWAS population, lines show the allele frequencies of the reference allele in the purebred Jersey and Holstein-Friesian animals. Note that before 1995 each year had less than 100 animals.

shows that ChrX:18584126C > G is consistently in high LD ($R^2 > 0.79$) with the most significant production peak SNP in all GWAS animals. Supplemental Table S5 (https://data.mendeley.com/datasets/5yzy4p5zr5/1; Trebes, 2023) shows that ChrX:18028733A > G is consistently in high LD ($R^2 > 0.88$) with the most significant production peak SNP in Jersey cattle, but not in Holstein-Friesians or all GWAS animals (Table 2).

Further to the predicted effect sizes shown across ChrX in Figure 3, Table 3 shows the predicted effect sizes per lactation of the 2 SNP of interest (ChrX:18584126C > G, ChrX:18028733A > G). The proportion of genetic variance and total phenotype variance attributed to each SNP of interest can be seen in Table 4. The proportion of total variance explained by ChrX:18584126C > G ranged between 0.0006 and 0.00234 and for ChrX:18028733A > G ranged between 0.00025 and 0.00121. The proportion of genetic variance explained by each SNP ranged between 0.00276 and 0.00891 for ChrX:18284126 C > G and between 0.0011 and 0.00501 for ChrX:18028733A > G.

As shown in Figure 6 we observed that heterozygotes for the SNP of interest produce less milk, and protein than homozygous reference animals, and that the homozygous alternate animals produce less milk again. These production differences were more obvious at the SNP ChrX:18584126C > G than at ChrX:18028733A > G. We tested the significance of these observed differences using Tukey HSD tests, the results of which are shown in Table 5 and expanded upon in Supplemental Tables S6 and S7 (https://data.mendeley.com/datasets/5yzy4p5zr5/1; Trebes, 2023). For the

Table 2. Linkage disequilibrium (LD) between the SNP of interest and the top SNP of each of the base GWAS, or with the most significant production peak SNP as indicated by an asterisk[1]

			ChrX:18584126C > G		ChrX:18028733A > G	
Phenotype	SNP position	GWAS SNP *P*-value	*P*-value	LD	*P*-value	LD
Fat lactation 1	34002922	3.30E-19	5.10E-11	9.43402E-05	3.40E-05	6.12671E-06
Fat lactation 1*	18479617	4.60E-12	5.10E-11	0.981	3.40E-05	0.394
Fat lactation 2	34002922	1.50E-11	8.80E-09	9.43402E-05	5.70E-04	6.12671E-06
Fat lactation 2*	18475739	8.60E-10	8.80E-09	0.9785	5.70E-04	0.3945
Fat lactation 3	34002922	2.90E-13	1.30E-10	9.43402E-05	2.40E-06	6.12671E-06
Fat lactation 3*	18501994	2.90E-13	1.30E-10	0.7961	2.40E-06	0.3869
Fat lactation 4	18475739	6.20E-19	2.60E-18	0.9785	5.00E-08	0.3945
Volume lactation 1	18584126	1.50E-23	1.50E-23		7.70E-11	0.3888
Volume lactation 2	18584126	8.00E-37	8.00E-37		1.60E-19	0.3888
Volume lactation 3	18584126	3.00E-32	3.00E-32		1.50E-14	0.3888
Volume lactation 4	18476384	5.00E-20	7.10E-19	0.8643	8.20E-09	0.3465
Protein lactation 1	34002922	4.30E-21	2.90E-16	9.43402E-05	1.00E-08	6.12671E-06
Protein lactation 1*	18584126	2.90E-16	2.90E-16		1.00E-08	0.3888
Protein lactation 2	18348881	5.90E-25	7.00E-24	0.8723	3.10E-14	0.4379
Protein lactation 3	18584126	2.70E-18	2.70E-18		4.20E-09	0.3888
Protein lactation 4	16697687	1.00E-12	4.10E-08	7.10629E-05	3.90E-04	3.80929E-05
Protein lactation 4*	18476384	1.30E-10	4.10E-08	0.8643	3.90E-04	0.3465

[1]LD = linkage disequilibrium R^2. This table is expanded and separated into LD between breeds in Supplemental Tables S4 and S5 (https://data.mendeley.com/datasets/5yzy4p5zr5/1; Trebes, 2023).

SNP ChrX:18584126C > G, of the 36 Tukey HSD tests 32 were significant ($P < 0.05$), with the 4 insignificant results coming from the test between homozygous reference animals and heterozygotes for each lactation of the fat phenotype. For ChrX:18028733A > G 26 of the 36 Tukey HSD tests were significant ($P < 0.05$). Again, all the insignificant Tukey HSD results came from the fat phenotype, with only fat lactation 2 showing the homozygous alternate animals had significantly different production to the homozygous reference and heterozygous animals.

To assess whether the genes MOSPD1 and CCDC1160 were undergoing differential expression we performed an eQTL analysis. The eQTL GWAS showed no ChrX peaks, indicating that no variants on ChrX were causing differential expression of either MOSPD1 or CCDC160.

Table 6 shows the top production peak SNP in iteration 1 after fitting the SNP of interest as fixed effects. Fitting ChrX:18584126C > G as a fixed effect caused the production peak to drop into insignificance across all phenotypes and lactations with all *P*-values >6E-7. Fitting ChrX:18028733A > G caused a reduction in the significance of the production peak in all phenotypes and lactations with all *P*-values >1.6E-17. For all the fat lactations and protein lactations 1 to 3, the same SNP topped the production peak after 1 iteration regardless of which of the SNP of interest was fitted as a fixed effect. Fitting ChrX:18028733A > G for volume

Table 3. The BOLT-LMM package GWAS predicted per-lactation effects (SE) of the alternate allele at the MOSPD1 (ChrX:18584126C > G) and CCDC160 (ChrX:18028733A > G) SNP of interest[1]

Item	Fat	Volume	Protein
ChrX:18584126C > G			
All animals	3.50 (0.50)	−106.70 (6.19)	−2.95 (0.37)
Jersey	2.43 (0.90)	−113.23 (10.18)	−2.14 (0.62)
Holstein-Friesian	8.22 (2.62)	−89.59 (37.67)	−2.38 (2.00)
ChrX:18028733A > G			
All animals	1.63 (0.36)	−51.15 (7.09)	−1.47 (0.26)
Jersey	1.92 (0.88)	−97.84 (15.88)	−2.82 (0.60)
Holstein-Friesian	0.36 (0.78)	−14.43 (17.92)	−0.26 (0.60)

[1]Effects displayed are the mean average of the effect of the SNP on the phenotype across all 4 lactations. Fat and protein units are kilograms per lactation, volume unit is liters per lactation. All units rounded to 2 decimal places. A positive value indicates a predicted increase in the phenotype while a negative value indicates a decrease for each active copy of the minor or alternate allele.

Table 4. The proportion of genetic and total variance in a phenotype explained by each of the SNP of interest Chr:18584126 C > G and ChrX: 18028733 A > G[1]

	Proportion of genetic variance		Proportion of total variance	
Phenotype	ChrX:18584126 C > G	ChrX:18028733 A > G	ChrX:18584126 C > G	ChrX:18028733 A > G
Fat lactation 1	0.00309	0.00125	0.00071	0.00029
Fat lactation 2	0.00302	0.00110	0.00069	0.00025
Fat lactation 3	0.00510	0.00279	0.00104	0.00057
Fat lactation 4	0.00891	0.00354	0.00204	0.00081
Volume lactation 1	0.00334	0.00144	0.00101	0.00044
Volume lactation 2	0.00753	0.00389	0.00234	0.00121
Volume lactation 3	0.00775	0.00333	0.00228	0.00098
Volume lactation 4	0.00424	0.00182	0.00131	0.00056
Protein lactation 1	0.00414	0.00207	0.00091	0.00045
Protein lactation 2	0.00869	0.00501	0.00196	0.00113
Protein lactation 3	0.00785	0.00362	0.00158	0.00073
Protein lactation 4	0.00276	0.00117	0.00060	0.00026

[1]Proportions have been rounded to 5 decimal places.

lactations 1 to 3 and protein lactation 1 resulted in the iteration 1 most significant production peak SNP being the other SNP of interest ChrX:18584126C > G.

DISCUSSION

In this paper we describe the use of GWAS to identify 2 candidate ChrX variants associated with milk production traits in dairy cattle. We used this study as an opportunity to evaluate different approaches for performing GWAS with an imputed ChrX and found that with female animals ChrX could be treated as an autosome, and that separating breeds can have important effects on the outcome of GWAS analysis of this type.

GWAS in Subsets of the Population

Previous work by our group has indicated significant challenges in accounting for all population structure in the NZ dairy population (B. Harris, personal communication, 2021). It was therefore important to understand whether the base GWAS results were due to population structure or whether the peaks observed are breed specific. We therefore ran separate GWAS for each breed and saw significant differences in the GWAS results with all animals (reflecting the HFxJ population that makes up most of the animals studied) and purebred Jerseys clearly showing a strong peak which was entirely absent in purebred Holstein-Friesians except in fat lactation 2. The differing associations observed between the breeds could be a result of differing allele frequencies of the SNP of interest. The MAF of the SNP of interest ChrX:18584126C > G is 10 times higher in Jerseys than in Holstein-Friesians, with 2020 animals carrying the alternate allele in purebred Jerseys (MAF = 0.086) compared with 2999 in Holstein-Friesians (MAF = 0.008). However, it must be kept in mind that as well as MAF in the GWAS population, the sample sizes also differ between the breeds and phenotypes (see Supplemental Table S1) and that reducing the sample sizes by excluding the HFxJ and separating the purebred animals will have greatly reduced the power of the GWAS. For example, the base GWAS of all animals for fat lactation 2 had data from 65,822 animals, while the purebred analyses had only 12,691 and 9,476 animals for Holstein-Friesians and Jerseys, respectively.

SNP and Genes of Interest

While we have not yet proven that either of the SNP of interest (ChrX:18584126C > G and ChrX:18028733A > G) are causing the effects of the production peak, they are both rare, exonic missense SNP that have a very significant association with the production phenotypes in Jerseys and HFxJ thus making them strong candidates. It is possible that these SNP could be linked to Jersey milk having a higher fat content than Holstein-Friesian milk. Changes in the way NZ farmers are paid for milk, and the relative value placed on fats, proteins and milk volume in the NZ dairy industry could cause a change in the allele frequency of these variants.

All analyses showed a significant GWAS association between ChrX variants and the production phenotypes. Based on the results of the iterative GWAS (drastic drops in production peak significance when SNP from this region are accounted for as fixed effects, and low numbers of iterations required for all SNP to be below the $P = 5E\text{-}8$ significance threshold) it is likely that the production peak as it appears in each individual phenotype is the result of a single, or a very few QTL. Based on observations of ChrX:34002922T > C and

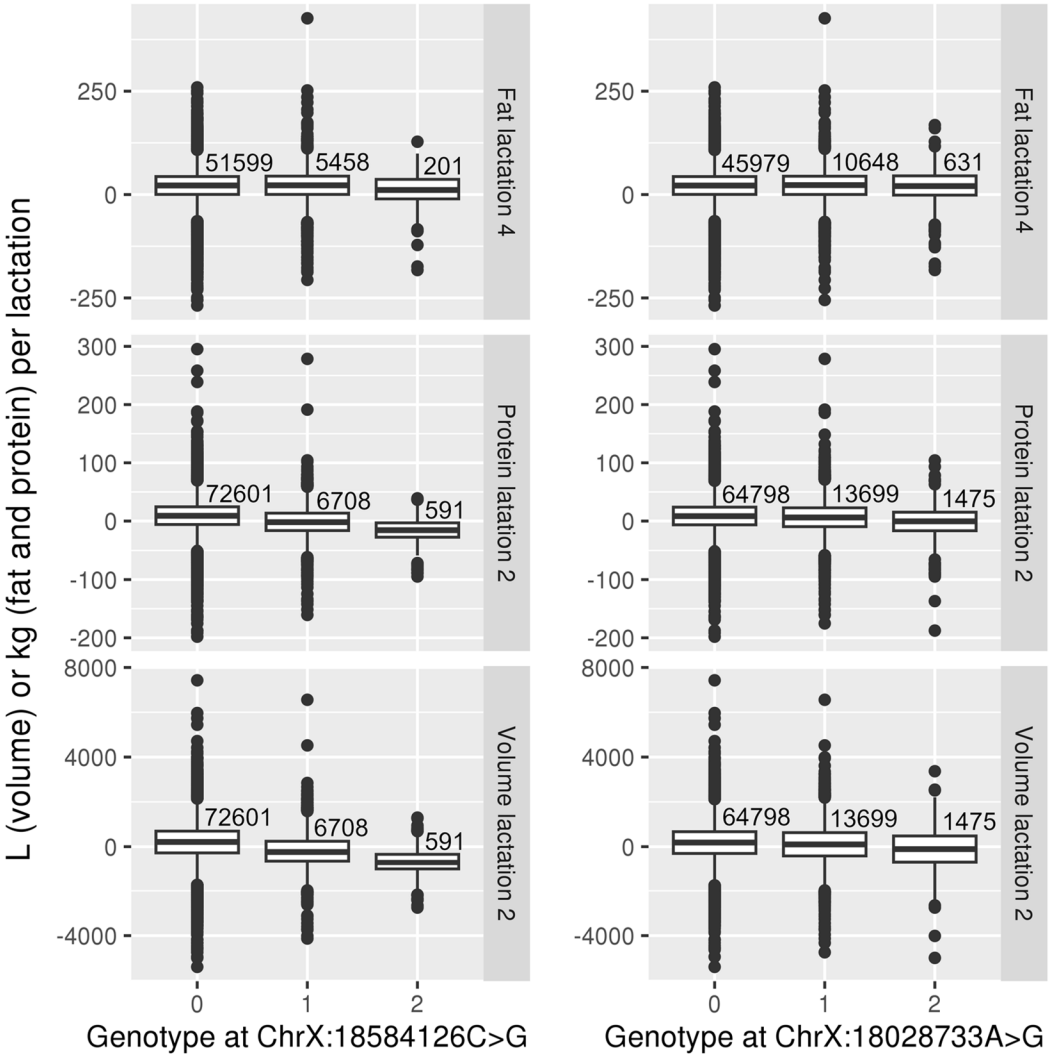

Figure 6. Boxplot of phenotype yields for animals of each genotype at the SNP ChrX:18028733A > G and ChrX:18584126C > G for the 3 phenotypes volume lactation 2, fat lactation 4, and protein lactation 2. The genotype (0, 1, or 2) refers to the number of copies of the minor or alternate allele the animal has at the specified SNP. The lactations shown were chosen as they had the most significant associations in the base GWAS. Numbers above each box indicate the number of animals in that group. Boxes represent the interquartile ranges (IQR), with the bottom of the box being the 25th percentile (Q1) and the top of the box being the 75th percentile (Q3), the median/50th percentile (Q2) is shown by a horizontal line within each box. Vertical lines (whiskers) from the top and bottom of each box represent the ranges Q3 + 1.5 × IQR and Q1 − 1.5 × IQR, respectively. Dots represent potential outliers as defined by values being outside the range Q1 − 1.5 × IQR to Q3 + 1.5 × IQR. These boxplots were generated using the R package ggplot and the geom geom_boxplot, which calculates whether each value falls into the ranges of the boxes, whiskers, or dots shown.

ChrX:16697687C > T from the base GWAS and iteration 1, we believe that these 2 SNP are separate from the production peak. The ChrX:34002922T > C is separate because of its distance from the peak, and the possibility of a sequencing error is ruled out by having ChrX:34002922T > C as a fixed effect in the GWAS having little to no effect on the production peak. The ChrX:16697687C > T is located within the bounds of the production peak, however when it is set as a fixed effect in the GWAS the production peak shows no great change. The inverse for both ChrX:34002922T > C and ChrX:16697687C > T also supports these SNP not belonging to the production peak in that the first GWAS iteration fitting a production peak SNP has little to no effect on the significance of these variants. As both ChrX:34002922T > C and ChrX:16697687C > T appear as lone SNP in the initial base GWAS or throughout the iterations we believe that their significance is an artifact or error with both SNP having very low MAF (0.009 and 0.01 for ChrX:34002922T > C and ChrX:16697687C > T, respectively) and being in very low LD with nearby SNP (highest LD with any

Table 5. Analysis of variance and Tukey honest significant difference (Tukey HSD) *P*-values for phenotype difference between genotype combinations for the 2 SNP of interest[1]

Phenotype or breed	Analysis of variance	Tukey 1–0	Tukey 2–0	Tukey 2–1
ChrX:18584126C > G				
Volume 2	<2E-16	0	0	3.62E-14
Fat 4	9.08E-05	3.26E-01	1.56E-04	6.04E-05
Protein 2	<2E-16	0	6.83E-10	2.90E-04
ChrX:18028733A > G				
Volume 2	<2E-16	2.98E-14	2.48E-14	1.07E-10
Fat 4	<2E-16	2.61E-14	3.45E-14	5.35E-11
Protein 2	0.0759	0.1530316	0.4416646	0.2132943

[1]Phenotypes (rows) denoted followed by lactation number. The lactations shown are those with the most significant ChrX production peak SNP from the base GWAS. Analysis of variance performed using the R "aov" function. Tukey HSD performed with R function "TukeyHSD." Hyphenated numbers (e.g., 1-0, 2-0, and 2-1) indicate the 2 genotypes being tested for significance of difference in phenotype using Tukey HSD. Values of <2E-16 and 0 are the result of the R aov and TukeyHSD packages reporting values so small they were outside the precision capabilities of the package. This table is expanded in Supplemental Table S6 to show the results across all phenotypes, and in Supplemental Table S7 (https://data.mendeley.com/datasets/5yzy4p5zr5/1; Trebes, 2023) to show the results for the phenotypes shown here for the separate purebred breeds.

SNP being 0.4 and 0.002 for ChrX:34002922T > C and ChrX:16697687C > T, respectively). Due to the abnormalities with the LD, low MAF, lack of peak shape (SNP often appeared on their own) and the evidence presented from the GWAS and iterative GWAS, we discounted these 2 SNP as belonging to the production peak and from being SNP of interest.

We have identified animals with a recombination event in the production peak between the 2 SNP of interest (ChrX:18584126C > G and ChrX:18028733A > G). At present the number of animals with this recombination is small; however, once more recombinants are identified we may be better able to test if the production peak we observed is the result of a single QTL. To facilitate the discovery of more potential recombinants, we have begun a breeding trial inseminating cattle with one or more of the SNP of interest with sperm from bulls with one or more of the SNP of interest. The offspring resulting from these matings will potentially give us the opportunity to observe cattle with the SNP of interest throughout their lives on a more day-to-day scale than our current data provides, as well as enabling us to potentially examine changes to the traits of animals with a recombination in the production peak region.

The function of the CCDC160 protein that contains the SNP ChrX:18028733A > G is unknown (Swiss Institute of Bioinformatics, 2023) though it is associated with ChrX-linked hypothyroidism and kidney disease. According to Homologene (Sayers et al., 2021) *CCDC160* is conserved across tetrapods with pairwise alignment scores of 70.9% (protein) and 83.3% (DNA) between the *Bos taurus* and human genes.

ChrX:18584126C > G lies within *MOSPD1* which encodes a transmembrane tether protein localized to the Endoplasmic Reticulum (Cabukusta et al., 2020). The protein coded by this gene is proposed to be involved in Epithelial-to-Mesenchymal cell transition (Kara et al., 2015; Thaler et al., 2011) and differentiation or proliferation of mesenchymal stem cells. NCBI's tool Homologene (Sayers et al., 2021) reports *MOSPD1* as conserved in Bilateria (animals with bilaterally symmetric embryos). The pairwise alignment score of the *Bos taurus* and human gene is reported as 97.2% (protein) and 93.6% (DNA; Sayers et al., 2021).

MOSPD1 is known to duplicate in humans as part of Xq25q26 duplication syndrome, potentially causing double outlet right ventricle (Hirota et al., 2017). In our data, duplications could potentially present as males showing heterozygous diploid for SNP on these genes (where they would normally be hemizygous). However, due to the methods with which our imputed genotypes were calculated, males were forced to hemizygosity – so this method would not show in the GWAS genotypes. None of the eQTL Manhattan plots generated for the genes of interest (*MOSPD1*, *CCDC160*) showed significant peaks on ChrX, indicating that if the effects described in this study are caused by either of these genes, it is likely not due to differing gene expression levels. Again, if either of these genes are causative, the lack of eQTL peaks also increases the likelihood the variant/s behind the differences in production are exonic, rather than intronic or regulatory as these would be more likely to effect expression levels and produce significant eQTL peaks. Hence the predicted effects seen in Table 3 are likely the result of the changes to the protein structure resulting from the missense mutations.

As a point of comparison for the ChrX chromosome effect sizes reported in Table 3, the effect size of average allele substitution of 2 DGAT1 SNP on chromosome 14 reported by Spelman et al. (2002) was −134 L/lactation and −110 L/lactation of milk volume for Holstein-Friesians and Jerseys, respectively. Spelman

Table 6. The SNP position and *P*-value for the most significant SNP in the production peak (defined as 15,000,000–2,200,000) in the base GWAS, and after an iterative GWAS fitting the SNP of interest ChrX:18584126 and ChrX:18028733[1]

	Iteration 0 (Base GWAS)		Fitting 18584126C > G		Fitting 18028733A > G	
Phenotype	SNP position	*P*-value	SNP position	*P*-value	SNP position	*P*-value
Fat lactation 1	18479617	4.60E-12	21717458	9.50E-05	21717458	9.50E-05
Fat lactation 2	18475739	8.60E-10	21788247	2.10E-04	21788247	2.10E-04
Fat lactation 3	18501994	2.90E-13	15200522	1.80E-04	15200522	1.80E-04
Fat lactation 4	18475739	6.20E-19	20848836	1.60E-05	20848836	1.60E-05
Volume lactation 1	18584126	1.50E-23	15658109	3.20E-05	18584126	4.50E-12
Volume lactation 2	18584126	8.00E-37	16385790	6.00E-07	18584126	1.60E-17
Volume lactation 3	18584126	3.00E-32	16754640	3.80E-05	18584126	3.00E-17
Volume lactation 4	18476384	5.00E-20	18037102	5.10E-05	18476384	1.10E-10
Protein lactation 1	18584126	2.90E-16	21717458	3.10E-06	18584126	7.00E-10
Protein lactation 2	18348881	5.90E-25	16315118	9.70E-05	16315118	9.70E-05
Protein lactation 3	18584126	2.70E-18	15049968	5.60E-04	15049968	5.60E-04
Protein lactation 4	18476384	1.30E-10	18135792	1.30E-05	18135792	1.30E-05

[1]Separate iterative GWAS performed to account for each of the SNP of interest. A version of this table showing the effect of fitting the SNP of interest on the highest chromosome X (ChrX) SNP can be seen Supplemental Table S8 (https://data.mendeley.com/datasets/5yzy4p5zr5/1; Trebes, 2023).

et al. (2002) also predicted the average change to milk fat yield being 5.76 kg/lactation and 3.3 kg/lactation for Holstein-Friesians and Jerseys, respectively, and milk protein changing by an average of −2.45 kg/lactation and −2.48 kg/lactation for Holstein-Friesians and Jerseys, respectively. These predictions, examining 2 alleles, are comparable in magnitude to the results we obtained for the effects of the SNP of interest reported here as ChrX:18584126C > G and ChrX:18028733A > G in the *MOSPD1* and *CCDC160* genes, respectively.

CONCLUSIONS

We found significant associations between 2 ChrX missense SNP and milk production phenotypes in New Zealand grazing HFxJ and purebred Jersey cattle. We observed differences in the existence/significance of genetic associations with production traits between the breeds, and so believe that for ChrX analysis in this context it is wise to examine the breeds separately as well as together. As no functional trials have been performed, we cannot say whether the SNP identified (ChrX:18584126C > G; *MOSPD1*, ChrX:18028733A > G; *CCDC160*) are indeed responsible for the GWAS production peaks we are seeing. Breeding trials are being carried out to attempt to further assess the actual effects of these SNP, though functional trials will likely also be necessary.

ACKNOWLEDGMENTS

This study received financial support from the NZ Ministry of Primary Industries, SFF Futures Program: Resilient Dairy–Innovative breeding for a sustainable dairy future (grant number: PGP06-17006). The authors wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities, consulting support, or training services as part of this research. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation and Employment's Research Infrastructure program (https://www.nesi.org.nz). At the time of writing, all authors are paid employees of Livestock Improvement Corporation, a breeding company and supplier of bovine germplasm. Livestock Improvement Corporation is also the applicant for a patent related to the variants identified as SNP of interest in this work, with H. T. and C. C. as named inventors. Specifically, this filed patent (789900) relates to the application of genetic markers and methods affecting the *MOSPD1* and *CCDC160* genes. The authors have not stated any other conflicts of interest.

REFERENCES

Arishima, T., S. Sasaki, T. Isobe, Y. Ikebata, S. Shimbara, S. Ikeda, K. Kawashima, Y. Suzuki, M. Watanabe, S. Sugano, K. Mizoshita, and Y. Sugimoto. 2017. Maternal variant in the upstream of FOXP3 gene on the X chromosome is associated with recurrent infertility in Japanese Black cattle. BMC Genet. 18:103. https://doi.org/10.1186/s12863-017-0573-8.

Browning, B. L., Y. Zhou, and S. R. Browning. 2018. A one-penny imputed genome from next-generation reference panels. Am. J. Hum. Genet. 103:338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.

Cabukusta, B., I. Berlin, D. M. van Elsland, I. Forkink, M. Spits, A. W. M. de Jong, J. J. L. L. Akkermans, R. H. M. Wijdeven, G. M. C. Janssen, P. A. van Veelen, and J. Neefjes. 2020. Human VAPome analysis reveals MOSPD1 and MOSPD3 as membrane contact site proteins interacting with FFAT-related FFNT motifs. Cell Rep. 33:108475. https://doi.org/10.1016/j.celrep.2020.108475.

Cezard, T., F. Cunningham, S. E. Hunt, B. Koylass, N. Kumar, G. Saunders, A. Shen, A. F. Silva, K. Tsukanov, S. Venkataraman, P. Flicek, H. Parkinson, and T. M. Keane. 2022. The European Variation Archive: A FAIR resource of genomic variation for all

species. Nucleic Acids Res. 50(D1):D1216–D1220. https://doi.org/10.1093/nar/gkab960.

Chambers, J. M., A. E. Freeny, and R. M. Heiberger. 1992. Analysis of variance; designed experiments. In Statistical Models in S. Routledge.

Couldrey, C., T. Johnson, T. Lopdell, I. L. Zhang, M. D. Littlejohn, M. Keehan, R. G. Sherlock, K. Tiplady, A. Scott, S. R. Davis, and R. J. Spelman. 2017. Bovine mammary gland X chromosome inactivation. J. Dairy Sci. 100:5491–5500. https://doi.org/10.3168/jds.2016-12490.

Czech, B., B. Guldbrandtsen, and J. Szyda. 2020. Patterns of DNA variation between the autosomes, the X chromosome, and the Y chromosome in *Bos taurus* genome. Sci. Rep. 10:13641. https://doi.org/10.1038/s41598-020-70380-9.

Davis, S. R., H. E. Ward, V. Kelly, D. Palmer, A. E. Ankersmit-Udy, T. J. Lopdell, S. D. Berry, M. D. Littlejohn, K. Tiplady, L. F. Adams, K. Carnie, A. Burrett, N. Thomas, R. G. Snell, R. J. Spelman, and K. Lehnert. 2022. Screening for phenotypic outliers identifies an unusually low concentration of a β-lactoglobulin B protein isoform in bovine milk caused by a synonymous SNP. Genet. Sel. Evol. 54:22. https://doi.org/10.1186/s12711-022-00711-z.

Gao, F., D. Chang, A. Biddanda, L. Ma, Y. Guo, Z. Zhou, and A. Keinan. 2015. XWAS: A software toolset for genetic data analysis and association studies of the X chromosome. J. Hered. 106:666–671. https://doi.org/10.1093/jhered/esv059.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. 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. 12:222–231. https://doi.org/10.1101/gr.224202.

Hatakeyama, C., C. Anderson, C. Beever, M. Peñaherrera, C. Brown, and W. Robinson. 2004. The dynamics of X-inactivation skewing as women age. Clin. Genet. 66:327–332. https://doi.org/10.1111/j.1399-0004.2004.00310.x.

Heard, E., P. Clerc, and P. Avner. 1997. X-chromosome inactivation in mammals. Annu. Rev. Genet. 31:571–610. https://doi.org/10.1146/annurev.genet.31.1.571.

Hecht, M., Y. Bromberg, and B. Rost. 2015. Better prediction of functional effects for sequence variants. BMC Genomics 16(Suppl 8):S1. https://doi.org/10.1186/1471-2164-16-S8-S1.

Hirota, Y., T. Minami, T. Sato, A. Yokomizo, A. Matsumoto, M. Goto, E. Jinbo, and T. Yamamgata. 2017. Xq26.1–26.3 duplication including MOSPD1 and GPC3 identified in boy with short stature and double outlet right ventricle. Am. J. Med. Genet. A. 173:2446–2450. https://doi.org/10.1002/ajmg.a.38297.

Jivanji, S., G. Worth, T. J. Lopdell, A. Yeates, C. Couldrey, E. Reynolds, K. Tiplady, L. McNaughton, T. J. J. Johnson, S. R. Davis, B. Harris, R. Spelman, R. G. Snell, D. Garrick, and M. D. Littlejohn. 2019. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. Genet. Sel. Evol. 51:62. https://doi.org/10.1186/s12711-019-0506-2.

Johnson, T., M. Keehan, C. Harland, T. Lopdell, R. J. Spelman, S. R. Davis, B. D. Rosen, T. P. L. Smith, and C. Couldrey. 2019. Short communication: Identification of the pseudoautosomal region in the Hereford bovine reference genome assembly ARS-UCD1.2. J. Dairy Sci. 102:3254–3258. https://doi.org/10.3168/jds.2018-15638.

Kara, M., R. A. Axton, M. Jackson, S. Ghaffari, K. Buerger, A. J. Watt, A. H. Taylor, B. Orr, W. R. Hardy, B. Peault, and L. M. Forrester. 2015. A role for MOSPD1 in mesenchymal stem cell proliferation and differentiation Stem Cells 33:3077–3086. https://doi.org/10.1002/stem.2102.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping, and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

Li, H. 2013. Aligning sequence reads, clone sequences, and assembly contigs with BWA-MEM. ArXiv:1303.3997v1 [q-Bio.GN].

Littlejohn, M. D., K. Tiplady, T. A. Fink, K. Lehnert, T. Lopdell, T. Johnson, C. Couldrey, M. Keehan, R. G. Sherlock, C. Harland, A. Scott, R. G. Snell, S. R. Davis, and R. J. Spelman. 2016. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. Sci. Rep. 6:25376. https://doi.org/10.1038/srep25376.

Liu, L., J. Zhou, C. J. Chen, J. Zhang, W. Wen, J. Tian, Z. Zhang, and Y. Gu. 2020. GWAS-based identification of new loci for milk yield, fat, and protein in Holstein cattle. Animals (Basel) 10:2048. https://doi.org/10.3390/ani10112048.

Loh, P.-R., G. Kichaev, S. Gazal, A. P. Schoech, and A. L. Price. 2018. Mixed-model association for biobank-scale datasets. Nat. Genet. 50:906–908. https://doi.org/10.1038/s41588-018-0144-6.

Mao, X., A. M. Johansson, G. Sahana, B. Guldbrandtsen, and D.-J. De Koning. 2016. Short communication: Imputation of markers on the bovine X chromosome. J. Dairy Sci. 99:7313–7318. https://doi.org/10.3168/jds.2016-11160.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. The ensembl variant effect predictor. Genome Biol. 17:122. https://doi.org/10.1186/s13059-016-0974-4.

Migeon, B. R. 1998. Non-random X chromosome inactivation in mammalian cells. Cytogenet. Cell Genet. 80:142–148. https://doi.org/10.1159/000014971.

Miller, R. G. 1981. Normal univariate techniques. Pages 37–108 in Simultaneous Statistical Inference. R. G. Miller, ed. Springer. https://doi.org/10.1007/978-1-4613-8122-8_2

DairyNZ. 2021. New Zealand Dairy Statistics 2020–21. Accessed May 19, 2022. https://www.dairynz.co.nz/publications/dairy-industry/new-zealand-dairy-statistics-2020-21/

Pacheco, H. A., F. M. Rezende, and F. Peñagaricano. 2020. Gene mapping and genomic prediction of bull fertility using sex chromosome markers. J. Dairy Sci. 103:3304–3311. https://doi.org/10.3168/jds.2019-17767.

Posynick, B. J., and C. J. Brown. 2019. Escape from X–chromosome inactivation: An evolutionary perspective. Front. Cell Dev. Biol. 7:241. https://doi.org/10.3389/fcell.2019.00241.

Prowse-Wilkins, C. P., T. J. Lopdell, R. Xiang, C. J. Vander Jagt, M. D. Littlejohn, A. J. Chamberlain, and M. E. Goddard. 2022. Genetic variation in histone modifications and gene expression identifies regulatory variants in the mammary gland of cattle. BMC Genomics 23:815. https://doi.org/10.1186/s12864-022-09002-9.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575. https://doi.org/10.1086/519795.

Reynolds, E. G. M., C. Neeley, T. J. Lopdell, M. Keehan, K. Dittmer, C. S. Harland, C. Couldrey, T. J. J. Johnson, K. Tiplady, G. Worth, M. Walker, S. R. Davis, R. G. Sherlock, K. Carnie, B. L. Harris, C. Charlier, M. Georges, R. J. Spelman, D. J. Garrick, and M. D. Littlejohn. 2021. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. Nat. Genet. 53:949–954. https://doi.org/10.1038/s41588-021-00872-5.

Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517. https://doi.org/10.1126/science.273.5281.1516.

Rosen, B., D. Bickhart, R. Schnabel, S. Koren, C. Elsik, A. Zimin, C. Dreischer, S. Schultheiss, R. Hall, and S. G. Schroeder. 2018. Modernizing the bovine reference genome assembly. Proceedings of the 11th World Congress on Genetics Applied to Livestock Production.

Rosen, B. D., D. M. Bickhart, R. D. Schnabel, S. Koren, C. G. Elsik, E. Tseng, T. N. Rowan, W. Y. Low, A. Zimin, C. Couldrey, R. Hall, W. Li, A. Rhie, J. Ghurye, S. D. McKay, F. Thibaud-Nissen, J. Hoffman, B. M. Murdoch, W. M. Snelling, T. G. McDaneld, J. A. Hammond, J. C. Schwartz, W. Nandolo, D. E. Hagen, C. Dreischer, S. J. Schultheiss, S. G. Schroeder, A. M. Phillippy, J. B. Cole, C. P. Van Tassell, G. Liu, T. P. L. Smith, and J. F. Medrano. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9:giaa021. https://doi.org/10.1093/gigascience/giaa021.

Sado, T., and A. C. Ferguson-Smith. 2005. Imprinted X inactivation and reprogramming in the preimplantation mouse embryo. Hum.

Mol. Genet. 14(suppl_1):R59–R64. https://doi.org/10.1093/hmg/ddi117.
Sayers, E. W., J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke, A. Marchler-Bauer, M. Landrum, S. Lathrop, Z. Lu, T. L. Madden, N. O'Leary, L. Phan, S. H. Rangwala, V. A. Schneider, Y. Skripchenko, J. Wang, J. Ye, B. W. Trawick, K. D. Pruitt, and S. T. Sherry. 2021. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 49(D1):D10–D17. https://doi.org/10.1093/nar/gkaa892.
Spelman, R. J., C. A. Ford, P. McElhinney, G. C. Gregory, and R. G. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85:3514–3517. https://doi.org/10.3168/jds.S0022-0302(02)74440-8.
Swiss Institute of Bioinformatics. CCDC160—Coiled-coil domain-containing protein 160—Function. 2023. Accessed Jan. 11, 2022. https://www.nextprot.org/entry/NX_A6NGH7/
Thaler, R., M. Rumpler, S. Spitzer, K. Klaushofer, and F. Varga. 2011. Mospd1, a new player in mesenchymal versus epidermal cell differentiation. J. Cell. Physiol. 226:2505–2515. https://doi.org/10.1002/jcp.22595.
Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov. 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14:178–192. https://doi.org/10.1093/bib/bbs017.
Tiplady, K. M., T. J. Lopdell, E. Reynolds, R. G. Sherlock, M. Keehan, T. J. J. Johnson, J. E. Pryce, S. R. Davis, R. J. Spelman, B. L. Harris, D. J. Garrick, and M. D. Littlejohn. 2021. Sequence-based genome-wide association study of individual milk mid-infrared wavenumbers in mixed-breed dairy cattle. Genet. Sel. Evol. 53:62. https://doi.org/10.1186/s12711-021-00648-9.
Trebes, H. 2023. Identification of candidate novel production variants on the Bos taurus chromosome X - Supplementary Information. Mendeley Data, V1. https://doi.org/10.17632/5yzy4p5zr5.1.
Vaser, R., S. Adusumalli, S. N. Leng, M. Sikic, and P. C. Ng. 2016. SIFT missense predictions for genomes. Nat. Protoc. 11:1–9. https://doi.org/10.1038/nprot.2015.123.
Wang, Y., K. Tiplady, T. J. J. Johnson, C. Harland, M. Keehan, T. J. Lopdell, R. G. Sherlock, A. Wallace, B. L. Harris, M. D. Littlejohn, R. Spelman, D. Garrick, and C. Couldrey. 2021. Investigating the accuracy of imputing variants on chromosome X in admixed dairy cattle using the ARS-UCD1.2 assembly of the bovine genome. In Proceedings from the 38th International Society for Animal Genetics Virtual Conference. https://www.isag.us/Docs/Proceedings/ISAG2021_Proceedings.pdf?v=20211015.
Xiang, R., E. J. Breen, S. Bolormaa, C. J. V. Jagt, A. J. Chamberlain, I. M. Macleod, and M. E. Goddard. 2021. Mutant alleles differentially shape fitness and other complex traits in cattle. Commun. Biol. 4:1353. https://doi.org/10.1038/s42003-021-02874-9.
Yandell, B. S. 2017. Practical Data Analysis for Designed Experiments. Routledge. https://doi.org/10.1201/9780203742563.
Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: A tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

ORCIDS

H. Trebes https://orcid.org/0000-0002-4166-3201
Y. Wang https://orcid.org/0000-0003-3756-6246
K. Tiplady https://orcid.org/0000-0002-3307-9208
C. Harland https://orcid.org/0000-0002-6268-0107
T. Lopdell https://orcid.org/0000-0002-7684-4870
T. Johnson https://orcid.org/0000-0003-1045-456X
S. Davis https://orcid.org/0000-0002-4942-1055
B. Harris https://orcid.org/0000-0003-0844-7539
R. Spelman https://orcid.org/0000-0002-7968-1392
C. Couldrey https://orcid.org/0000-0001-7410-9410