Genetic influences of reproductive events in cattle

By

Beth M. Lett

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Animal Science)

at the

UNIVERSITY OF WISCONSIN-MADISON

2022

Data of final oral examination: 05/31/2022

The dissertation is approved by the following members of the Final Oral Committee:

Brian W. Kirkpatrick, Advisor & Professor Molecular Genetics, Animal and Dairy Sciences

Hasan Khatib, Associate Chair & Professor Genetics & Epigenetics, Animal and Dairy Sciences

Irene Ong, Assistant Professor, Obstetrics and Gynecology & Bioinformatics

Kent Weigel, Department Chair & Professor Breeding and Genetics, Animal and Dairy Sciences

DEDICATION

I dedicate this to me. For the struggles I over came to make it here. The past me for making the choice and sticking through all the headaches to get to this point. To the me that battled from the ledge numerous times and is still battling every day. To the me that has changed several times over the years and keeps striving to improve.

ACKNOWLEDGEMENTS

First and foremost, I must thank my faith, my family, and my friends who became family without whose support and strength I would not have finished this degree. My faith for seeing me through dark days and family for continuing to tell me I can do this. And my friends, Nicole Gross, Hadjer Namous, Kim Reuscher, Verik (Vera and Erick) Akin, all of whom became my family here in Madison, Wisconsin. Thank you for all the love, support, and help during the last few years.

I need to also thank my advisor Brian Kirkpatrick for accepting me as a Master's student initially, for sticking with me during my head injury, and having faith in me to switch to a PhD. To my committee members Kent Weigel and Hasan Khatib for also sticking with me when I switched to the PhD and supporting me at different conferences. And to Irene Ong for both instilling a true love for bioinformatics in 576 and agreeing to join my committee for my PhD.

Last but not least have been all those who have helped me over the years. Specifically, members of the fourth and fifth floors of the animal science building. Several long discussions on projects, prelim and defense practices, and social gathers helped us all to grow and will be missed. To my long-distance friends, Celina and Jenna thank you so much for being you and even though we do not live close always being but an instant message away. Finally, a thank you to my Pink Sisters for the support and knowledge as I found myself again after the mTBI.

TABLE OF CONTENTS

List of Tables by Chapter

List of Figures by Chapter

CHAPTER 1

Literature Review and Problem Identification

I. General overview

In cattle production, reproduction is a key component for success whether it's in beef or dairy production. Over the years, progress has been made in understanding physiology of the reproductive cycle in cattle but some of the more molecular aspects remain a mystery. Genetics particularly in dairy has been revolutionary in progressing herd productivity and selecting for more desirable phenotypes. The realm of genetics has exploded with improvements and capability of sequencing technologies. This has led to improvements in reference genomes, increased number of genetic markers, and ability to detect different genetic variants. The scope of this work is to investigate different reproductive stages and identify genetic variations linked to different phenotypic traits at that stage of the cattle reproduction cycle.

II. Review of cattle reproductive physiology

2.1 Folliculogenesis

In mammals' reproduction starts with development of the Graafian follicle developing in a process called folliculogenesis (Figure 1.1). Initially the cohort of follicles are dependent on follicle-stimulating hormone (FSH) (Figure 1.1A) [1–4]. A surge of FSH recruits follicles from the cohort to start growing (Figure 1.1B) [1,4–6]. The majority of follicles at this stage undergo atresia. Typically, in cattle, deviation or selection occurs when the largest follicle reaches 8 mm in diameter. Selection begins in response to increases in estradiol and inhibin and decreases of FSH, resulting in the largest follicle becoming dominant and the remaining subordinate follicles undergoing atresia (Figure 1.1C) [1,2,6]. The dominant follicle further increases estradiol and inhibin while decreasing FSH, thus preventing additional follicles from reaching this stage (Figure 1.1D) [1,2,7–9]. A key feature of the dominant follicle is it is no longer dependent on FSH to grow and has developed receptors for luteinizing hormone (LH) [1,3,4,6]. There are two paths left

for the dominant follicle. Like other follicles, most dominant follicles undergo atresia leading to negative feedback on estradiol and inhibin and increasing FSH production (Figure 1.1E). The final option would be developing to a preovulatory stage that would eventually be ovulated (Figure 1.1F) [3,4,7]. This sets the stage for ovulation rate **(OR)**, the number of follicles that mature and ovulate during an estrous or menstrual cycle.

In cattle, folliculogenesis occurs in a wave pattern where follicles begin to grow and regress until the dominant follicle(s) are ovulated. The number of waves varies from cow to cow ranging from two to four waves per cycle (Figure 1.2). Waves occurring at the beginning to mid-cycle correspond with the luteal phase and end with atresia [2,4]. The final wave corresponds to the follicular phase and ends with ovulation [2,4]. During a cycle each wave is initiated by a surge of FSH, then either a steady amount of progesterone during the luteal phase, or a decrease of progesterone and LH surge during the follicular phase (Figure 1.2)[2–5,7]. The LH surge triggers release of the oocyte from the follicle [4,6,10].

Two theories have been proposed to explain how more than one dominant follicle is selected and ovulated. These theories are centered around the period prior to deviation when FSH is above a threshold preventing atresia. In one theory, the increased OR is caused by follicles developing LH receptors at a smaller size and allowing them to survive FSH depletion[1,3,7]. The other is the decreasing FSH levels are not sufficient at deviation, instead they stay high enough to allow more follicles to reach dominance [3,7]. Either or both of these aspects could lead to multiple follicles being ovulated.

2.2 Embryo survivability

While ovulation rate sets the upper limit for number of offspring, embryo survivability **(ES)** ultimately decides if a pregnancy ends with a multiple, single, or loss. The first obstacle is fertilization. Roughly 10 to 20% of oocytes are not fertilized [11,12]. After conception, 10 to 80% of embryos are lost within the first few weeks, with higher percentages being seen in lactating dairy cows [11–13]. On many dairy farms pregnancy checks are conducted at around 28 to 30 days after insemination and around days 42 to 60. During the time between pregnancy checks, embryo loss occurs at ~12% [12]. After day 60 there are only moderate losses of 1 to 3%[12]; however, in cows bearing multiple embryos, a large proportion (75%) are lost between days 60 to 90 [12,14,15]. This is caused primarily by overcrowding in the uterus as the fetuses grow and develop [6,7,12,16]. Loss of the embryo post-fertilization leads to economic loss for the producer (loss of calf, cost of re-breeding, etc.) and impact the reproductive efficiency of the cow.

III. Reproductive events

3.1 Embryo and fetal lethality

Embryo survivability and development and thus fetal survivability is an interplay between the composition of the fertilized oocyte and the uterine environment of dam. As the embryo develops there are several key developmental stages that impact the progress towards fetal development and subsequent parturition of a live calf[17–19]. At any point during development, programmed cell death may occur and trigger an early abortion of the embryo/fetus. Termination of pregnancy caused by the developmental insufficiency contributes to the percentage of unknown causes of cow abortion and may be linked to the genetic make-up of the embryo. Identification of these developmental defects would help improve reproductive efficiency. These effects can be linked both to the genetic composition of the embryo and the quality of the oocyte during ovulation[19]. As far as genetic composition, this can be attributed to

lethal mutation phenotypes, spontaneous mutations or errors in DNA repair, or segmental chromosome rearrangements[20,21].

The other component to embryo survivability lies with the dam's uterine environment. As previously mentioned in multiple births and embryo survivability overcrowding is one example of how that environment contributes to fetal death. But even before embryo elongation and development the uterine environment must change to sustain pregnancy[11,19,22]. This is tied with a cascade of different biological and hormonal pathways and disruption of any stage could lead to abortion of the embryo/fetus. One source of negative environmental impact is lactation status of the cow. Lactation triggers its own pathway cascades that impede and trigger opposite regulatory pathways to those needed for uterine environmental sustainability. Multiple studies have shown that higher producing cows showed decreased reproductive efficiency when compared with lower producing and non-lactating counterparts[21,23]. Nutrition also plays a role in these pathway cascades and controls the uterine environment as the embryo/fetus develops[21,22,24]. The health status of the cow during her previous partition and current gestation impacts the uterine environment. With previous partition affecting her return to cyclicity and early development and current gestation impacting more the later stages[25,26]. Genetic composition of the dam impacts her ability to sustain a viable environment for the embryo/fetus.

3.2 Multiple births and ovulation rate

The term multiple births refers to a pregnancy resulting in two or more offspring being gestated and born. Cattle, like most primates, are typically monotocous – producing one offspring at a time; however, on occasion, more are produced. Typically, these pregnancies result in twins. There are two forms: 1) monozygotic **(MZ)** twins, or identical twins, where the

developing fetuses are from the same ovum fertilized by the same sperm, and 2) dizygotic **(DZ)** twins, or fraternal, which are fetuses that develop from different ovum and sperm.

While MZ pairs must be of the same sex, DZ have a ratio for male-male, male-female, and female-female pairs of 1:2:1 [27–29]. The proportion of MZ in cattle is low at 0.13 – 0.74% of all births [28–34]. Thus, twins are typically assumed to be the result of DZ pregnancies. There are potentially incidences of greater than two however this happens for approximately 0.015% of total births [29]. Because of this, twin is used interchangeably to refer to the phenomenon of multiple births in cattle interchangeably throughout this dissertation. Typically, this trait is measured by the number of calves born and measured as a frequency termed twinning rate **(TR)**. Multiple factors can affect multiple births both intrinsically and extrinsically to the cow herself.

Firstly, the cow's production type (dairy vs beef) and her breed composition may affect her TR [29,30,35–37]. In dairy, the frequency is increasing. The prominent dairy breed, Holstein, has shown a change from 2.0% of all births in 1932 [27] to an average 3.5 to 5.0% of all births from years 1975 to 2018 [32,36–38]. Twining rates in other dairy breeds include ~2.7% in Jersey, ~3.2% in Guernsey, ~2.0% in Ayrshires, and 4.6 to 8.9% in Brown Swiss [30,32,36,37]. Beef cattle tend to have a lower incidence ranging from 0.2 to 4.6% [35,37,39]. A beef breeds also show variation like Angus (0.4 to 1.6%) [35,37,39], Charolais (1.63 to 1.7%) [35,39] and Simmental (2 to 4.6%) [35,37,39].

Additionally, cow age impacts her chances of multiple births. Across multiple species, including humans, age and twinning share a positive correlation. Potentially, this is caused by increases of FSH and/or uterine capacity coinciding with increases in age and number of times giving birth [15,34,40–42]. Age relates both to the chronological age or parity and the number of times one gives birth as they are highly correlated [27].

Another intrinsic factor that has been linked to causing increased TR primarily in dairy cattle has been milk production. Higher producing dairy cows show increased chances of double ovulations and twinning [6,33,43–45]. A hypothesis for this association is higher producing dairy cows have greater feed intakes and subsequently greater steroid metabolism. This causes decreases in circulating estrogen and progesterone that, in turn, increases multiple ovulations [6,46]. Further, the effect of feed intake and feed quality on ovulation rate, referred to as "flushing" has been seen in beef cattle and sheep [15,47].

The final factor unique to the cow is her genetics and repeatability of the trait. Twinning rate tends to be lowly heritable. In Holstein dairy cattle, heritability ranges from 1.7 to 9.0% for linear models and 8.0 to 14.2% with threshold models [32,36,38,48–50]. Repeatability estimates from linear models range from 0 to 6.3%[36,48,51] with one estimate of 28.6% using a threshold model [36]. Ovulation rate has slightly higher heritability ranging from 3 to 40% and repeatability ranging from 10 to 32.6% [7,35,52,53]. TR and OR are highly correlated (0.66 to 0.90) [35,52,53]. However, while heritability is low, genetic variation exists that can be exploited by selection, as seen in the USDA Meat Animal Research Center (MARC) twinning herd. In their final report, the population had attained an annual TR of 60% [16]. Additionally, there are reported cases of unique individuals or families with high fecundity. In the 1800s and 1900s, six cows have been documented for having multiple sets of twins and/or triplets[37]. A family, referred to as the Trio family, has been identified in the last 20 years that has increased frequency of twins and triplets. Multiple studies with the Trio family show a Mendelian inheritance of a trait for increased OR [54,55].

As with any organism, the environment where it is cultured or raised impacts its growth, development, and reproductive performance. For cattle this can be narrowed to herd.

Management decisions influence the incidence of disease, calf survivability, lifetime production, and TR on a farm. One major difference in management style is nutrition. Feed quality varies from location to location and year to year. A hypothesized increase in TR is caused by higher quality feed stuffs. This leads to higher energy intake stimulating folliculogenesis and increased ovulation rates (discussed earlier). The culling practices of a farm influence TR. Cows calving twins have a higher risk of being culled due to fertility or health issues [33,56,57]. This limits chances for repeated incidences and number of offspring that are predisposed to produce multiple births. Additionally, reproductive protocols such as *in vitro* fertilization **(IVF)** and embryo transfer **(ET)** impact ovulation rate. Hormones such as those used in synchronization protocols and antibiotics, have been indicated as influencing TR [33,34,43]; however, in recent work it seems that milk production is the main contributing factor [58–60]. Additionally, cows with ovarian cysts or lacking corpora lutea (CL) have higher risks of double ovulations when subjected to Ovsynch synchronization protocol [43,58].

The next layer of environmental effects comes at the year and season level. Conditions encompassing a year influence not only cow productivity but herd management. For example, drought years cause issues with feed quality of both pasture-based beef operation and mixed rations in dairy cattle. Year to year changes may reflect changes in a herd's health status; healthier herds produce more, are more reproductively fit, leading to more effective feed intake.

A year can also be broken down into components of season. Each season affects calving or conception. Most dairy operations breed for year-round calving, while beef limit to spring and/or fall calving – matching pasture growth. Calving during the summer and early fall (conception fall to early winter) shows increased incidences of multiple births [32,35–37,61,62]. Studies measuring OR saw a trend of increased ovulations during cooler seasons (fall and winter

months) [15,35,60,63]. Heat stress during the summer months could impact both ovulation and embryo survivability, while higher quality fall pastures or supplements increase OR and ES[15,33,60,62].

IV. Genes associated with folliculogenesis, twinning, and embryo survivability

4.1 Transforming growth factor beta super family

Over the years one super family of proteins has been heavily linked to folliculogenesis and twinning. This family is the transforming growth factor-beta **(TGF-β)** superfamily encompassing more than 30 structurally related proteins involved in numerous pathways affecting cell differentiation, proliferation, apoptosis, development, reproductive processes, and many more [64–71]. Members of this superfamily are found in various tissues and their activation typically is cell or tissue specific. The superfamily splits mostly into two subfamilies: 1) the bone morphogenetic proteins **(BMP)** and growth differentiation factor **(GDF)** including > 20 members comprised of the BMP proteins (excluding *BMP1*), GDF proteins, and anti-Mullerian hormone **(AMH)** and 2) the activin/TGF-β subfamily including ~ 8 members comprised of TGF-β proteins, activin, and inhibins [64,68–70,72–74]. The structure and functionality of this super family have been extensively reviewed elsewhere[69,70,74,75].

Briefly it is comprised of the ligand proteins, receptors, and signal transducers – members of the mother against decapentaplegic **(SMAD)** family. The ligands make up most of the superfamily and bind to the receptors with preferences to a specific receptor pair. The receptors break down into type I and type II groups and comprise a smaller group than the ligands. In mammalian species, there are seven known types I receptors and five type II [66–68,70,71,73]. Typically, the ligand influences the receptor pairs forming a heteromeric complex and which is

bound first (type I vs. type II). Members of the activin/TGF-β subfamily tend to have affinity for type II receptors, while BMP/GDF subfamily members tend to have affinity for type I receptors [64,68–71,73]. Additionally, the type I receptor has a preference for the type II receptor they binds, limiting the number of possible combinations. Once a heteromeric receptor/ligand complex is created type II receptors cause phosphorylation of the type I receptor which activates the SMAD family[64,65,68,70,74]. There are eight known SMADs affecting the TGF-β superfamily and they split into three functional groups. SMADs directly affected by the type I receptor phosphorylation are the regulatory SMADs **(R-SMADs)**. R-SMADs include *SMAD1, 2, 3, 5,* and *9* [66,67,69,71,74,76]. Like receptor binding, R-SMADs are influenced by the type I receptor. Typically the BMP/GDF subfamily activates *SMAD1, 5,* and *9* and the activin/TGF-β subfamily activates *SMAD2* and *SMAD3* [65–70].

Once activated, the R-SMADs interact with *SMAD4*, the common SMAD, creating either heterodimers (an R-SMAD and *SMAD4*) or heterotrimers (two R-SMADs and *SMAD4*) [64–66,71]. These complexes allow translocation into the nucleus and subsequent regulation of transcription of target gene(s) either directly or indirectly. They are regulated by SMAD-interacting proteins and/or inhibitory SMADs **(I-SMAD)** [64–69,71,76]. I-SMADs include *SMAD6* and *SMAD7*. *SMAD6* preferentially inhibits the BMP/GDF subfamily by competing with *SMAD4* for binding with R-SMADs as well as competing with BMP/GDF type I receptors and R-SMAD binding. *SMAD7* inhibits both subfamilies by competing with the R-SMADs for binding to the intracellular portion of the type I receptor [64–68,70,71,73,76].

The relation of TGF-β superfamily to multiple births has been widely studied. This family has been indicated as a major regulatory signaling pathway for female reproduction[8,77,78]. Activins promote and inhibins inhibit FSH production in feedback loops during folliculogenesis

[3,78]. Ligands *BMP15* and *GDF9* are oocyte-specific proteins that control follicular development and ovulation [79,80]. Multiple members of the superfamily have been investigated for expression and contribution to litter size (LS), OR, and TR [72,81,82].

These investigations identified multiple members expressed in ovarian tissue, granulosa cells, and theca cell. These include *BMP2, 3, 4, 6, 7, 15, GDF9, 10, AMH, BMPR1A, BMPR1B, TGFBR1, and BMPR2* [3,7,69,70,72,73,77,83–85]. Expression of *BMP2, 4, 6, 7, 15*, *BMPR1A* and *B, BMPRII, SMAD1, 4, 5, GDF9, TGFBR1*, and *INHBB* was measured in Hu high fecundity **(HF)** and low fecundity **(LF)** sheep. *BMP15* expression was lower in HF while *BMP4, BMPRIB, BMPRII, SMAD4, GDF9,* and *TGFBR1* were higher [85].

Unique phenotypes of increased OR and LS, and in some cases infertility, in sheep led to investigation of major genes affecting LS and OR in sheep breeds [86,87]. Nomenclature for these genes is **Fec** standing for fecundity, a number or capital letter standing for chromosome location or gene (e.g. **X** for chromosome X), and in the case of multiple different locations a superscript designating breed or researcher (e.g. $FecX^{I}$). The first identified major gene, called the Booroola gene (identified first in Booroola sheep) or FecB, corresponds to a nonsynonymous mutation in *BMPR1B*. It caused an additive increase in OR based on number of mutated alleles (WT [+/+]< *BMPR1B* [+/-] < *BMPR1B* [-/-]) [86,88–93] (Table 1.1).

Additionally, multiple mutations in oocyte-specific proteins *BMP15* and *GDF9* were identified. To date, nine and six different variants, respectively, have been reported for these genes. Three of the reported variants in *BMP15* produced premature stop codons ($FecX^{H}$ [86,89,94], $FecX^{G}$ [86,89,91,95], and $FecX^{R}$ [86,96,97]), while the others are different non-conservative mutations ($FecX^{O}$ [86,98], $FecX^{Gr}$ [86,98], $FecX^{B}$ [89,95], $FecX^{I}$ [86,88,89,94], and $FecX^{L}$ [86,89,99], $FecX^{Bar}$ [100]). $FecX^{Bar}$ also contains a 3 bp deletion and frameshift insertion. Of these $FecX^{O}$ and $FecX^{Gr}$ show similar

phenotypes as the Booroola gene, but with homozygous mutant individuals having a minimal increase over heterozygous carriers. All other known variants of *BMP15* in sheep cause increased OR in heterozygote carriers and infertility in homozygous mutant individuals (Table 1.1). Two of the *GDF9* variants (FecG^{I} [101] and FecG^{F} [102,103]) are conservative, missense mutations and were both identified by Hanrahan et al[95], but at the time were not associated with LS or OR. Both were subsequently shown to result in increased OR in heterozygous carriers but FecG^{I} homozygous mutants have lower OR than WT while FecG^{I} heterozygotes have increased OR. The remaining four variants are non-conservative mutations. FecG^{H} [86,89,95] and FecG^{T} [86,104] mutations result in phenotypes of increased OR in heterozygote carriers and infertility in homozygous mutant individuals. While FecG^{E} [86,105] causes phenotypes like FecX^{O}, FecX^{Gr}, and FecB (Table 1.1). A novel synonymous mutation in *GDF9* was significantly associated with LS in Hu sheep and has an additive effect with FecB and dominate *TGFBR2* variant. It is a candidate gene mutation for this breed affecting LS[85] (Table 1.1) . Authors point out that reports have shown synonymous mutations may affect mRNA splicing and stability, protein translation and folding, and involved in regulating microRNA meditated genes. They give no additional speculations on why these synonymous mutations are associated with LS.

Knowledge of these major genes, as well as known roles the TGF-β superfamily members play in reproduction, other mutations have been indicated as candidates for OR, TR, or LS in different species. These include *BMP4* [106,107], *BMP7* [108], *BMP15* [106], *GDF9* [109–111], *BMPR1B* [112], and *TGFBR2* [85]. Additionally, mouse and rat knockout **(KO)** and conditional knockouts **(cKO)** have been produced that cause infertility (*GDF9* [113,114], *BMPR1B* [115]*)* or subfertility (*ACVR2* [116], *SMAD3* [117,118], *double cKO SMAD2/SMAD3* [118], *BMP15* [114], and *BMP7* [119]) in these rodents (Table 1.1).

In cattle, TGF-β member variants of *GDF9* and overexpression of *ACVR2A* have been indicated as potential candidates for fertility traits [120,121]. A nonsynonymous mutation of *GDF9* was identified as a potential candidate gene for TR in Maremmana cattle as it was only identified in cows producing twins [109] (Table 1.1). Utilization of the Trio family has led to the identification of a major gene affecting ovulation rate in cattle [54,55,122–126]. No causative polymorphism has been identified yet, the gene (Trio Allele) has been mapped to a 1.2 Mb region of bovine chromosome 10 (BTA10) [54]. Heterozygous and homozygous individuals of the Trio allele generate ~2.6 more follicles per ovulation than non-carriers [54,124–126]. Three positional candidate genes (*SMAD3*, *SMAD6*, and *IQCH*) were initially proposed [54] and screened for polymorphisms in coding regions and 5' and 3' flanking regions (Table 1.3). No plausible causative variants were identified, suggesting the causative variant is within regulatory sequence farther from the gene than the originally considered 5' and 3' flanking sequences. Gene expression analysis showed only S*MAD6* having differential expression between carriers and non-carriers indicating that changes in regulation of *SMAD6* drive the differences in phenotypes (Table 1.1) [122,123].

A recent whole genome sequencing analysis indicated *BMP5*, *BMP6*, *ACVR1*, and *TGFBR2* as candidate genes for reproductive traits in pigs [127]. Additionally, quantitative trait locus (**QTL**) and genome-wide association study (**GWAS**) analysis implicated TGF-β superfamily members as candidate genes for OR, TR, or LS in sheep and swine. These include *TGFBR1* [128,129], *BMP7* [127,128], *INHBA*[127,128], *INHBB* [87], and *SMAD1*[87] (Table 1.4). A peak SNP *rs80956812* (CHR 1: 164674664 from Swine genome assembly Sscrofa 11.1) associated with LS is located in *SMAD6* gene in swine with a second peak SNP rs80912860 (CHR 1: 164637575 from Sscrofa 11.1) upstream from *SMAD6* [130] (Table 1.4). The updated bovine genome

assembly[131] and further evaluation of GWAS-based analysis for TR in cattle done by E.S. Kim et al. in 2009[132] implicate *TGFB1* and *BMP6* (Table 1.3).

4.2 Other influential genes

One source of information of genes and genetic locations has been the QTL database **(Animal QTLdb).** Utilizing information from this database poses challenges related to differing methods used in studies summarized. For this review location of QTLs not already mapped to ARS-UCD 1.2 were approximated by dividing the cM location by 1,000,000 to estimate a physical (bp) location. On BTA5 there are three genes of interest including *SOCS2, NR1H4,* and *IGF1* (Tables 1.2 and 1.3). *IGF1*, insulin-like growth factor I, has roles including hormone activity and has been implicated in multiple studies as a candidate gene. *SOCS2*, suppressor of cytokine signaling 2, functions include IGF receptor binding. Seven QTLs span BTA7 and three contain genes that are part of signaling pathways involved in embryogenesis, the Wnt (*WNT3A* and *WNT9A*) and the fibroblast growth factors (*FGF1)* pathway [78,133] (Table 1.2). The last region of interest is on BTA21 containing *IGF1R* the receptor for *IGF1* (Table 1.2). There could be potentially numerous additional candidate genes as some regions span large segments of a chromosome.

While microsatellite data is challenging to update, SNP-based data is relatively easy. A new database consortium sponsored by the FAANG projects was developed, called FAANGMINE. This database includes primary livestock species consolidating information from multiple sources. The limitations of this resource are utilization of a single reference for each species and that QTLs must have SNP reference locations passing quality filters. This database allowed for further screening of QTLs for TR and OR in cattle (Table 1.3) and two closely related species (Table 1.4). Genes of interest found within this curation of cattle include one

gene, *NDF1P2,* that interacts indirectly with the TGF-β superfamily signaling pathway, four genes, *AKR1D1, NR1H4, CYP2S1,* and *END1*, related to hormone activity, and four genes, *SPATA3, IGCH, PAFAH1B2,* and *MAK,* related to spermatogenesis.

Using both the Animal QTLdb and FAANGMINE, QTLs in porcine and ovine species were also screened for genes of interest in respect to the traits of LS, TR, and OR (Table 1.4). In addition to *IGF1* being implicated in cattle its paralog *IGF2* [134] and insulin like growth factor binding protien (*IGFBP2*)[135] are candidate genes in pigs. Like IGF, growth hormone (*GH*) is involved with folliculogeneis[136] making the *GH* receptor (*GHR*) identified in sheep a good candidate gene[87]. QTLs in both pigs and sheep share common genes of estrogen receptors 1 and 2 (*ESR1*[87,128,137] and *ESR2*[87,138]). These two genes form heterodimers and are needed for reproductive function. Two other reproducitve receptors, gonadotropin relaeasing hormone receptor (*GNRHR*) and follicule stimulating hormone receptor (*FSHR*) are candidate genes in swine [139] and sheep [140] QTLs, respectively. *NCOA1* is a coactivator for hormone receptors located in an ovine QTL[87]. *ZP3,* zona pellucida qlycoprotein 3, and *VPM1,* vacuole membrane protein 1, play roles during embryo development and implantation. Two genes with indirect implications are transcription factor 12 (*TCF12*) [141] and zinc finger FYVE-type containing 9 (*ZFYVE9*) [127]. *TCF12* is related to the Wnt pathway, while *ZFYVE9* interacts with the TGF-β superfamily (Table 1.4). Gene information, functions, and locations (*Bos taurus)* were obtained from NCBI annotation release 106 and human gene card. QTL regions were viewed using NCBI's genome browser and genes documented using references ARS-UCD 1.2 for cattle, Oar Rambouillet v1.0 for sheep, and Sscrofa 11.1 for pig.

V. Sequencing technologies

5.1 First generation sequencing: Sanger

One of the first methods of sequencing originating back in the 1970s, Sanger Sequencing is still a widely used method. This method sequences using a primer to locate a specific region of interest in the template DNA, labeled dideoxyribonucleotides **(ddNTP)** missing the 3' hydroxyl group which prevents extension of the DNA strand, and capillary electrophoresis. The use of the ddNTPs missing the 3' hydroxyl group, where N stands for any one of the four different nucleotides (A, C, T, G), is why this technique is also referred to as chain-termination. Each of the different nucleotides are labeled with a different fluorescent dye and when separated by fragment size via electrophoresis, the resulting sequence can be inferred by the fluorescent label. Though shorter than the other methods of sequencing (~1,000 bp) and limited to single fragments, the accuracy is extremely high (99.99 %)[142,143]. While initially used in genome assembly, the multiple step process to generate a single fragment makes it less ideal for large scale assemblies. It does, however, have strength in low and medium target sequencing projects particularly for visualization and validation of different variant detection events.

5.2 Next generation sequencing: Illumina pair-end reads

The objective of developing low cost, high-throughput sequencing gave rise to a new generation of sequencing that had the capability of increasing the quantity of reads through parallelization[144]. This allowed for massive amount of data being generated. While multiple platforms exist for generating next generation sequencing (454, Solid, Illumina, etc.), the focus here will be on the Illumina platform. Still using the sequence by synthesis methodology as Sanger, this method takes it a step further by using reversible dye termination and flow cells with multiple lanes containing different oligonucleotide anchors[142,145]. Initially, a sequencing library is prepared where DNA is fragmented, adaptors are added to them, and potentially PCR amplification to generate more copies of the adaptor fragments. The library is added to the flow

cell where the oligonucleotide anchors complement the adaptors. Upon amplification fragments will then create bridge clusters between two anchors. The resulting clusters are then sequenced using reverse terminator method which detects single bases incorporated into the DNA template by emission of the different fluorescently labeled nucleotides. Output from this step is numerous copies of either single-end or pair-end reads[142,143]. The advantage of pair-end reads is the ability to sequence both ends of the DNA fragment giving positional information, which increases the ability and accuracy to align the reads to a reference and detect different genetic variants. The depth of sequencing coverage, also provides valuable information on quality and ability to detect certain variants events such as deletions and duplications[142,143]. Even though the reads are slightly less accurate than Sanger sequencing, the advantage of volume, pair-end information, knowledge of the library size, and low cost makes them advantageous[145]. However, the short length of the reads generates challenges with repetitive genomes during assembly/alignment stages and may generate errors of misplacing the reads.

5.3 Third generation sequencing: Long reads

Unlike first and second generation sequencing the defining line between third and second generation is harder to establish. One uniform underlying component for this generation is the capability to generate the sequencing based on a single molecule rather than relying heavily on DNA amplification. The ultimate goal for this generation of sequencing was to generate long reads that have the potential to span full genomes in a few reads in the case of small genomes or complete chromosomes in the case of larger (e.g., human and cattle). Like the other generations, there have been multiple different platforms and advantages to them. Here the focus will be on two widely used platforms of Pacific Biosciences (PacBio) single-molecule real-time (SMRT) and Oxford Nanopore technologies (ONT)[143]. These two platforms utilize different concepts to

generate their respective reads. The SMRT method still uses sequence-by-synthesis and different fluorescent tags on the different nucleotides[146]; however, the key feature is the sequencing is done in real time. This imaging is achieved through utilizing zero-mode waveguide (ZMW) regions[146,147]. The advantage for SMRT is that the single molecule itself will not degrade, but the disadvantage is that read length will be dependent on the ability of the polymerase[142,146]. Currently the average read length is 10-20 Kb with some reads being over 50 Kb[143]. In contrast, the ONT method does not rely on fluorescent labelling but instead on ion currents generated as the nucleic acid cross the membrane and through pores[143,148]. As the DNA passes through the pore, the base generates a change to ion current that reflects which nucleotide is next in the sequence[148,149]. The average read length of ONT depends on the sample preparation, but 10-20 Kb average reads or greater can be routinely obtained with reports of 2.3 Mbp long capabilities[143,148]. In the past a disadvantage of these long-read methods has been the high error rates however this has been reduced for both and can be further improved based on the sequencing depth. The current disadvantage of these technologies is the cost and computing resources to handle the longer reads. But the advantages of being able to span longer segments of a genome make this a promising and exciting generation of sequencing[150].

VI. Genetic variant classifications

6.1 Single nucleotide polymorphisms

One of the most widely used variants has been single nucleotide polymorphisms **(SNP)** which as the name implies are single base changes from one nucleotide to another. They have been widely used in breeding programs as genetic markers because of the abundance throughout the genome, stability, and capability to utilize high-throughput automation. Throughout the years the number of known variants has grown[151]. In cattle the progress of identifying SNPs has

followed a similar trend as sequencing technologies. As the methods of detection gets better and less expensive, the number of SNPs detected increases and knowledge of those that are included on SNP chips. Initially SNP chips started with small numbers < 10k and have progressed to 80k standards. Additionally, whole genome levels can be achieved. The ability to impute from lower density SNP data to higher also increases the amount of data available for studies. This allows for large meta-analysis of trait associations to detect genomic regions of interest to a trait or traits[152]. Typically, these regions of interest are referred to as quantitative trait loci (QTL).

The Cattle QTL database is a curation of QTLs and associations spanning 1,000s of studies. These can be further broken down by different traits including reproduction. Multiple studies have identified associations between SNPs and traits such as early embryonic survival, reproductive efficiency, embryonic mortality, ovulation rate, twinning, and many others. These studies have been used to develop selection indexes for producers and utilized to help improve traits. Examples of those found in cattle involved with twinning and ovulation rate can be seen in Table 1.3.

6.2 Insertion-deletion mutations

Another classification of smaller variants is the Insertion-deletion mutation **(InDel)** which are addition or removal of one or more base pairs. There are variable definitions for the size of these events and cutoffs are arbitrary. One such definition is InDels are deletions or insertions < 1 Kb or more recently those < 50 Bp[153,154]. These differences may reflect the progression, accuracy, and methods involved in both sequencing and variant detection within a sample. For this work we will define an InDel as those ≤ 50 bp and those > 50 bp as structural variants (discussed next). Abundance of these variants is second to SNPs and detection is limited based on the quality of the alignment[153,155].

In cattle numerous InDels have been detected particularly as the 1000 Bull Genomes Project grows. Examples include a 17 bp insertion InDel that affects cattle growth in Chinese cattle[156]. Interestingly, this variant is found within *SMAD3* and affects the transcription levels of the gene. Another example is an 11 bp deletion Indel found in the *MSTN* gene and its connection to double muscling in Belgium Blue cattle[157]. This deletion shortens the amino acid chain, introduces premature stop codons, and leads to *MSTN* protein fragmenting. Thus, it disrupts expression of the normal *MSTN* gene product allowing for hypertrophic muscle growth and the double muscling phenotype.

6.3 Structural Variants

As previously alluded to, structural variants **(SV)** are large (> 50 bp) changes to a genome and are classified into various types. These include the widely studied sub-class called copy number variants **(CNV**). Copy number variants are deletions (losses) or duplications (gains) of a region's copy number relative to a reference genome assembly. Other structural variants include a) inversions - regions of the genome inverted in orientation compared with a reference, b) translocations – regions of the genome that have moved either within the same chromosome or to a different chromosome, and c) insertions – addition of "novel" regions. While less numerous than the smaller variant types (SNP and InDel), SVs have the potential to cause the greatest variation and impact because of their size[150,154].

Detection of these variants varies slightly depending on sequencing information; however, in all cases the sample genome sequence is compared with a reference genome assembly. Previously, detection was conducted utilizing quantitative allele intensity data from SNP chips and primarily focused on CNVs. With the advancements to sequencing technology to include short and long reads, detection of these large structural variants has improved.

Challenges remain due to the size of different events, and SV types can have similar patterns of detection, for example, distinguishing a novel insertion from a tandem duplication[150]. Four main detection strategies have been employed to call variants. The most computationally taxing method is assembly-based where the individual of interest is first *de novo* assembled and then compared to another assembly[150,158]. While the other methods rely heavily on information from the reads themselves, particularly with paired reads where orientation or spacing is abnormal. These methods include read depth (RD), which can measure copy number based on changes in the depth relative to the surrounding areas[158]. It is limited to large size events and CNVs. The other two methods look at disagreements in read mapping of pairs and the expected reference. They include read-pair (RP) and split-read (SR). Read pair methods look for disagreement in read insert size of the sequenced reads versus the expected[158]. Similarly, SR looks for read disagreement but in the form of one pair member mapping and the other entirely or partially failing to align. The unaligned read becomes the site of a potential break point and allows for more base pair level detection[158]. Each method on their own has limitations that can be bolstered by combining the different methods to detect variants.

In cattle these large variants have been shown to influence phenotypic appearances, disease resistance, reproductive health, and between-breed differences. The well-known phenotypic trait of horns or polled (no horns) is linked currently to four different structural variants on BTA1. In two of the four it's a large duplication event, and in the other two it's a complex insertion-deletion event[159]. These variants have also been shown to differentiate between sub-species (e.g., taurus vs. indicus) and breeds (e.g., Angus vs. Holstein)[160–162]. In Nordic Red dairy cattle, a 660 Kb deletion has been indicated as causative for low fertility and potentially for embryo lethality[163,164].

VII. Problems identified

7.1 – Negative association of multiple births (Dairy)

Twinning, particularly in dairy cattle, is viewed negatively. This is due to greater risks associated with multiple births compared to potential rewards. A major risk category is health issues including early abortions, difficulties during calving, calves born dead or dying shortly after, retained placenta **(RP)** and metabolic disorders like displaced abomasum **(DA)** and ketosis [165–174]. Consequently, these risks, as well as the pregnancy itself, impact subsequent lactations for a dam. This can include increases in calving interval (**CI**, time between calvings), days to first service, first service conceptions rates, days from calving till conception, and reduced mean lifetime production [57,169,175–177]. Another concern for replacement females is freemartins. Freemartins refers to females that are co-gestated with males and have a <10% chance of being fertile [27,165,178]. In dairy the associated negative cost ranges from $50 to $250 per twin birth[165,179]. With ~5% incidence seen in Holstein and a national herd average of 9 million cows this would be an annual loss of $22.5 to $112.5 million [36].

7.2 – Positive association of multiple births in cattle (Beef)

Profitability in beef production is impacted by weight of calf weaned per cow exposed. A way to increase this and utilize resources is by increasing the number of calves born per cow bred. Potential exists for 104 kg to 186.0 kg increase per cow exposed with twin production [16,180]. The U.S. Meat Animal Research Center **(MARC)** twinning herd found an average total weight weaned per cow of 217.7 kg ± 2.5 for singles, 328.3 kg ± 3.2 for twins, and 378.4kg ± 15.0 for triplets[16].

Beef cattle sustainability is measured by resource management. In a simulation study, a prediction of a 3.2% to 9% reduction in land and water use and reduction in greenhouse gas emissions when utilizing multiple births in beef production[181]. This system additionally allows spreading the per cow maintenance cost between two calves rather than one. Utilizing multiple births in beef cattle could accomplish net gains[168,180,182].

7.3 – Reproductive inefficiency and cost of abortions

A large economic investment for producers is in reproduction costs. In 2006 a study by A. De Vries[183] looked at estimating costs associated with pregnancy. On average the cost of a new pregnancy is $278 with an average cost of $555 associated with pregnancy loss. With ranges being predicted from $90 – $2,000 depending on factors such as days in milk, stage in gestation, and cow value influencing the total loss[183–185]. Abortions also impact the cow as well increasing her time to return to estrus and chance of being culled. Additionally, this may increase chances of reproductive disease depending on stage of gestation. There are several factors that may impact reproductive efficiency and embryo/fetal mortality, one of which is genetics. Previous studies have demonstrated that certain haplotypes were associated with embryonic death and shown to not appear in the population in homozygous states[184,186,187]. Additionally, as previously mentioned in section 6.3 there has been a large deletion found in Nordic Red cattle that influences embryo loss.

VIII. The present studies

8.1 – Twinning in North American Holsteins

Objectives

1. Estimate heritability and repeatability of twinning rate in recent North American Holstein calving records.

2. Conduct genome wide association to identify genomic regions with replicated effects across timeframes.

3. Test if inclusion of whole-genome sequencing showed SNPs of greater association

4. Identify gene pathways with greater association with twinning rate

5. Evaluate the potential for genomic selection in future genetic improvement programs

Rational and hypothesis

Previous studies have looked at identifying QTL and locations associated with twinning in dairy cattle[188–190]. However, there is minimal agreement between them on locations, genes, and pathways. Based on previous reviews and discussions this trait is a negative and frustrating trait for producers[57,165,175]. Identifying genetic markers and developing genomic selection against twinning would be a means of reducing incidence of twin birth. In addition, identifying underlying mechanisms of genes involved may help future research in understanding the biology behind twinning. We hypothesize that the transforming growth factor-beta (TGF-β) superfamily contributes to genetic variation for twinning rate in Holstein cattle.

8.2 – Exploration of embryo lethal candidates in Jersey cattle using large structural variant detections.

Objectives

1. Identify copy number variants in Jersey cattle using short read whole genome sequencing.

2. Validate predicted CNVs focusing on those located within genes.

3. Identify deletions as candidates for embryo lethality potential.

4. Test frequency of embryo lethal candidates in general population of Jersey cattle.

Rational and hypothesis

Jersey cattle are the second most popular breed of dairy cattle in the United States. Reproductive inefficiency leads to a large economic cost to producers and is one of the top reasons for culling cows from a herd[183–185]. Copy number variants, while low in frequency, have greater potential impact phenotype because of their size (> 50 bp)[191]. Large deletions are potential candidates for embryo lethality because of their size and removing sections of DNA needed for developmental progression[163,192]. Detection of deletions that are only present in the heterozygous state would be a strong indicator that the deletion had embryo lethality potential. Providing information on such deletions could help producers in making breeding decisions and improve reproductive efficiency. We hypothesize that CNVs in gene regions with embryo lethal potential cause absence of homozygote individuals.

8.3 – Identification of causative mutation in Trio allele

Objectives

1. Produce a *de novo* assembly of a homozygous individual for the Trio allele

2. Identify candidate variants that may be causing the Trio allele high ovulation phenotype

3. Test the concordance of candidate variant(s) with inferred Trio allele genotype

4. Determine the frequency of the candidate variant(s) in two populations outside the Trio descendants which are most likely to harbor the Trio allele

Rational and hypothesis

The Trio allele is a major gene which produces high ovulation rate in carrier females and has been studied to understand one mechanism for variation in ovulation rate[193]. This work has identified no differences in follicular waves but difference in hormone and follicular development between normal and Trio allele carrier females[124,126]. Previous work has shown that the Trio allele is associated with over-expression of the gene *SMAD6*, which is part of a pathway of genes involved in folliculogenesis[122,123]. This signaling pathway has been previously implicated in contributing to variation in ovulation rate and litter size in sheep[86]. The exact variant causing the *SMAD6* over-expression and high ovulation rate has not been identified. Understanding the mechanism behind this phenotype has the potential to be utilized in different reproductive technologies. The Trio allele itself has the potential to be used to increase twinning rate while limiting some of the negative aspects of the trait. Given a strong pressure to remove cows by involuntary (abortion/death offspring and death of dam) and/or voluntary (producer culling) selection we hypothesize that this mutation is recent and unique to the Treble/Trio family. Under this assumption, candidate variant(s) would be exceedingly rare in the general cattle population, in essence not observed outside of Trio descendants. Within Trio descendants, to be considered as potentially causal, a candidate variant would necessarily show perfect concordance with true Trio allele genotypes. Thus, these two conditions can be employed in screening variants for those which are putatively causal.

CHAPTER 2

Heritability of twinning rate in Holstein Cattle

I. Preface

At the time of submission this chapter was published in the *Journal of Dairy Science* (Accepted: December 11, 2017)

> Lett, B. M. & Kirkpatrick, B. W. Short communication: Heritability of twinning rate in Holstein cattle. *Journal of Dairy Science* **101,** 4307–4311 (2018).

Formatting and reference style were changed for consistency throughout the thesis. Figures and tables have been updated and assigned based on location within thesis and references included in full reference section of thesis. Lastly the acknowledgements have been removed. All other aspects are consistent with the published manuscript.

II. Abstract

Multiple births or twinning in cattle is a naturally occurring reproductive phenomenon. For dairy cattle, twinning is considered a detrimental trait as it can be harmful to cow and calf as well as costly to the producer. The objective of this study was to examine recent US calving records for the Holstein breed to determine a current estimate of heritability for twinning rate along with effects of season and parity. Two models were used in this study: a linear sire model and a binary threshold-logit sire model. Both were mixed models considering fixed effects and random effects. Analyses were conducted using a restricted maximum likelihood method. Heritability estimates were 0.0192 ± 0.0009 and 0.1420 ± 0.0069 for the linear and threshold models, respectively. Repeatabilities from the linear and threshold-logit models were 0.0443 ± 0.0012 and 0.2310 ± 0.0072, respectively. The nonzero estimates of heritability indicate the potential to select against this trait for genetic improvement of Holstein cattle.

Key words: twinning, heritability, dairy cattle

III. Introduction

Twinning in cattle, especially dairy cattle, is viewed as a negative trait. This is due to the negative ratio of risk to reward in having twin births. There is still a debate about whether there is a benefit of increased milk production from dams that deliver multiple calves in a single calving[43,56,169,194] or whether there is a negative effect on milk production[33,165,177]. However, even with increased milk production as a possible positive effect, it is associated with multiple negative effects such as abortion, dystocia, stillbirth, retained placenta, metabolic disorders, displaced abomasum, and ketosis[165,166]. Additionally, there is the effect on future calving for the dam as twinning increases the calving interval and the time period between calving and

conception and reduces mean lifetime production[57]. Twinning is neutral with regard to the proportion of fertile replacement female calves to infertile freemartin females from mixed-sex twin births[166]. The neutrality is from the offset of fertile females from same-sex twin births, though the absolute number would be lessened owing to the lower perinatal survival rate of twins. Even though there is a chance to reduce some of these negatives with adjustments to management (e.g., changes to diet for twin-bearing cows, additional labor), these adjustments would result in increased costs. A range of approximately $50 to $250 loss per twin birth puts a large economic expense on twinning with very minimal, if any, benefits for twins in the dairy industry[165,179]. Frequency of twinning in US Holstein cattle has been approximately 5%[32,38], meaning an annual loss to the industry of $22.5 to $112.5 million assuming a national herd of 9 million cows.

Known contributing factors for twinning need to be accounted for when analyzing heritability. One such factor, as reviewed in 1975 by Rutledge[37], is the seasonal effect. This has been expressed as either the month or season of calving or conception, with conception months of September through October and March through April having the highest incidences. Another factor is parity of the dam. As the parity increases so does the chance of twinning, with the largest increase happening between the first and second parities[37,38]. These effects are presumably effects on ovulation rate, as incidence of monozygotic twinning is low and ovulation rate and twinning rate have a high genetic correlation[52,195]. Additionally, sires can have an influence on twinning rates in their daughters[37], suggesting a genetic component to variation. In 2001 Johanson et al[38] grouped Holstein sires based on birth year and demonstrated that the sire group from the most current time points had more daughters with higher incidences of twinning. As with heritability, repeatability of twinning tends to be low[37,48]. The objective of the current

study was to estimate heritability and repeatability of twinning rate in US Holstein cattle using current calving record information for this population.

IV. Material and methods

Calving records from the years 2010 to 2016 were obtained from AgSource Cooperative Services (Verona, WI). Initially, we obtained more than 2.9 million records from all breeds. Available information included cow, sire, and dam identification (ID), herd, birthdate, calving date, parity, and multiple birth code. After editing, as described below, only the Holstein breed had sufficient records for estimation of heritability using a sire model. All other breeds had fewer than 20 sires represented in edited data versus more than 2,000 for the Holstein breed (Table 2.1). Consequently, efforts focused on data from Holsteins and records were removed if they were not Holstein for the cow, sire, and dam. Additionally, records were excluded if the cow, sire, or dam were missing part or all of their ID code or had codes indicating unknown ID. Records were also excluded if a cow had discrepant sire, dam, and birthdate information between records. Duplicate calving entries were eliminated such that there was 1 record per calving. Two sires and corresponding records were removed due to being listed both as a sire and as a cow. Suspected embryo transfer calvings were likewise excluded (indicated by multiple calvings within the same year for a cow). To increase the reliability of the genetic evaluation, only sires with ≥100 daughter records were included in the final data set (n = 2,223; Figure 2.1). Likewise, only herds with ≥100 records were included (n = 1,748; Figure 2.1). After editing, 1,440,540 records with 658,436 cows remained for use in the analysis. Preliminary fixed effects analysis was done using SAS 9.4 (SAS Institute Inc., Cary, NC). Variance components for heritability and repeatability calculations were estimated using Asreml 4.1 (VSN International, Hemel

Hempstead, UK). Two models were used in alternative analyses: a linear sire model (LM) and a binary threshold-logit sire model (TLM). The general form of the model in matrix notation was

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zs} + \boldsymbol{Ep} + \boldsymbol{e},$$

where y is a vector of calving phenotypes (singles or twins) for the LM and a vector of unobserved liabilities for twinning for the TLM; **X** is an incidence matrix relating phenotypes to fixed effects of herd, year, season, and parity; **b** is a vector of fixed effects; **Z** is a matrix relating phenotypes to sire genetic effects; s is a vector of sire additive genetic effects; **E** is a matrix relating phenotypes to permanent environmental effects; p is a vector of animal environments; and e is a vector of random residuals. The sire, permanent environment, and residual effects were random effects and herd, year, season, and parity were fixed effects. It was assumed for the LM that the random effects followed normal distributions of $s \sim N(0, \mathbf{A}\sigma_s^2)$, p$\sim N(0, \mathbf{I}\sigma_p^2)$, and e$\sim N(0, \mathbf{I}\sigma_e^2)$ where σ_s^2, σ_p^2, and σ_e^2 represent additive genetic, permanent environment, and residual variances, respectively; **A** represents the numerator relationship matrix for sires; and I is the identity matrix. For the TML, random effects were assumed to have followed normal distributions of $s \sim N(0, \mathbf{A}\sigma_{s_T}^2)$, p$\sim N(0, \mathbf{I}\sigma_{p_T}^2)$, and e$\sim N(0, \mathbf{I}\frac{\pi^2}{3})$ where $\sigma_{s_T}^2$, $\sigma_{p_T}^2$, and $\frac{\pi^2}{3}$ represent additive genetic, permanent environment, and residual variances, respectively. Parity for this data was categorized into 4 groups: parity 1 (n = 563,942), parity 2 (n = 415,867), parity 3 (n = 251,010), and parity 4 and above (n = 209,721). Preliminary analysis indicated that means for parities 4 and above were not significantly different ($P < 0.05$), and the number of calving records per parity rapidly diminished with parity 3 and greater (Figure 2.1). Calving dates were categorized as being in 1 of 4 seasons based on month: season 1 = December through February, season 2 = March through May, season 3 = June through August, and season 4 = September

through November. In these data, year is the 7 yr from 2010 to 2016 in which a calving was recorded. Estimates of genetic effects for sires on twinning rate were regressed on year of birth for each sire to evaluate genetic trend for twinning rate. Heritability was estimated using variance components for sire, animal permanent environmental, and residual effects

$$h^2 = \frac{\sigma_s^2 x4}{\sigma_s^2+\sigma_p^2+\sigma_e^2}.$$

For TLM, the underlying scale for the residual variance was $\frac{\pi^2}{3} \sim 3.3$[196] for the logit link and was used as a coefficient to return residual variance to approximately 1 for heritability analysis. Additionally, the variance components for the TLM where $\sigma_{s_T}^2$, $\sigma_{p_T}^2$, and $\frac{\pi^2}{3}$, respectively. Both LM and TLM models were executed as restricted maximum likelihood models[196]. Repeatability was estimated using the variance components for sire, animal permanent environmental, and residual effects as

$$r = \frac{(\sigma_s^2 x4) + \sigma_p^2}{\sigma_s^2+\sigma_p^2+\sigma_e^2}.$$

Similarly, for heritability estimation from the TLM model, the corresponding variance components were $\sigma_{s_T}^2$, $\sigma_{p_T}^2$, and $\frac{\pi^2}{3}$, respectively.

VI. Results and discussion

Our results showed strong evidence ($P < 0.001$) for fixed effects of herd, year, season, and parity on twinning rate. Twinning rate increased with parity (Table 2.2), with the largest change in effect between parity 1 (0.3%) and parity 2 (4.3%). Previous studies with Holstein cattle have shown increases in twinning with parity, with twinning rates between 0.6 and 1.63% in parity 1 and ranging between 3.01 and 6.48% in later parities[32,38,62,197]. Additionally, in 2000

Karlsen et al.[197] and in 2001 Johanson et al.[38] each reported the largest change between parity 1 and parity 2, with increases from 0.7 to 2.8% and 1.63 to 5.22%, respectively. June through August (season 3) had the highest twinning rate among seasons (Table 2.2), which coincides with a conception period of September through November. In 1975 Rutledge[37], reviewing previous reports on twinning, stated that 2 periods of conception were associated with increased twinning rate, one of which (September to October) coincides with our findings. Cady and Van Vleck (1978)[32] found peak of twins in calving between May and July (conceived between September and November), overlapping with the highest twinning rate season observed in the current analysis. Herd effects can vary due to a contribution of multiple management factors such as nutrition, culling practices, and reproductive protocols[37,43].

Previous average twinning rates ranged from 4.82 to 5.02% in the US Holstein breed[32,38]. Average twinning rate in the current data was 4.8%, comparable with these previous reports. A genetic trend for increasing twinning was evidenced by a significant regression of sire genetic effects on birth year ($P < 0.001$), with a regression coefficient of 0.0003 ± 0.0001. Estimates of heritability were 0.0192 ± 0.0009 and 0.1420 ± 0.0069 for the LM and TLM, respectively. This indicates a lowly heritable trait, which concurs with previous findings. The higher heritability estimate from TLM versus LM was expected due to the different underlying distributions for the models. Heritability estimates for LM are in good agreement with previous reports of linear model heritability estimates, which ranged from 1.7 to 9.0%, whereas the TLM is slightly higher compared with previous threshold model estimates, ranging from 8.0 to 10.5%[32,38,48,49,52].

Repeatabilities of twinning rate from the LM and TLM were 0.0443 ± 0.0012 and 0.2310 ± 0.0072, respectively. Our result on the LM scale is within the range of previous findings (0–6.3%)[48,51]. Few published estimates of repeatability from threshold model analyses are available

in the scientific literature. Moioli et al. (2017)[198] reported a repeatability estimate of 0.286 ± 0.012 from a threshold model analysis of twinning rate in Maremmana cattle, and Wolc et al. (2006)[199] reported a repeatability estimate ranging from 0.33 to 0.34 in threshold model analyses of twinning rate in Thoroughbred horses. The results reported here indicate that twinning rate is a lowly heritable and lowly repeatable trait that is affected by herd, year, season, and parity. This data set is one of the largest used thus far in estimation of heritability and repeatability of twinning rate and provides current estimates of each as well as fixed effects of season, parity, and year. The nonzero estimate of heritability and repeatability for twinning rate suggests the opportunity to improve this trait (i.e., reduce twinning rate) by selection.

CHAPTER 3

Identifying genetic variants and pathways influencing daughter averages for twinning in North American Holstein cattle and evaluating the potential for genomic selection

IV. Preface

At the time of submission this chapter was published in the *Journal of Dairy Science* (Accepted: March 4, 2022)

> Lett, B. M. & Kirkpatrick, B. W. Identifying genetic variants and pathways influencing daughter averages for twinning in North American Holstein cattle and evaluating the potential for genomic selection. J. Dairy Sci. (2022). Doi:10.3168/jds.2021-21238

Formatting and reference style were changed for consistency throughout the thesis. Figures and tables have been updated and assigned based on location within thesis and references included in full reference section of thesis. Lastly the acknowledgements have been removed. All other aspects are consistent with the published manuscript.

II. Abstract

Multiple birth in dairy cattle is a detrimental trait both economically for producers and for animal health. Genetics of twinning is complex and has led to several quantitative trait loci regions being associated with increased twinning. To identify variants associated with this trait, calving records from 2 time periods were used to estimate daughter averages for twinning for Holstein bulls. Multiple analyses were conducted and compared including GWAS, genomic prediction, and gene set enrichment analysis for pathway detection. Although pathway analysis did not yield many congruent pathways of interest between data sets, it did indicate two of interest. Both pathways have ties to the strong candidate region on BTA11 from the genome- wide association analysis across data sets. This region does not overlap with previously identified quantitative trait loci regions for twinning or ovulation rate in cattle. The strongest associated SNPs were upstream from 2 candidate genes LHCGR and FSHR, which are involved in folliculogenesis. Genomic prediction showed a moderate correlation accuracy (0.43) when predicting genomic breeding values for bulls with estimates from calving records from 2010 to 2016. Future analysis of the region on BTA11 and the relation of the candidate genes could improve this accuracy.

Key words: twin, cattle, folliculogenesis, genomic prediction

III. Introduction

Twinning in dairy cattle is viewed as an undesirable trait. This is due to the negative association of the trait with calving issues (abortion, dystocia, and calf loss), and cow health and performance (retained placenta, metabolic disorders, increased calving interval, and reduced mean lifetime production)[57,165,175]. The frequency of having twins is referred to as daughter averages for twinning (TW) and is a complex trait influenced by multiple genetic and

environmental factors[37,200,201]. Multiple studies have identified QTL for TW in dairy cattle; however, those identified seldom overlap regarding region[132,188–190,202–204]. With minimal overlap of results between TW studies, this results in multiple QTLs listed in the Animal QTL database (Animal QTL Database (animalgenome.org))[205] and shows the complex and polygenetic nature of twinning.

Recently, a newer method for GWAS called single- step GWAS (ssGWAS), and by extension weighted ss- GWAS (WssGWAS), has been used in various animal genetic studies[206–210]. This method allows for the combination of nongenotyped and genotyped animals with phenotypes to be used in the analysis with pedigree information used to relate nongenotyped animals with genotyped. Further exploration of GWAS results comes from the expansion of gene set enrichment analysis, originally developed for gene expression data[211–213]. Utilization of these methods allows for associating genes and pathways with traits of interest to add more depth to standard GWAS results. Identification of gene sets more involved with TW would lead to a greater understanding of the trait, the biological mechanisms involved, and potentially a better understanding of previously dissimilar results. The low consensus from previous QTL mapping and GWAS studies owes partly to differences in experimental design (granddaughter and daughter designs vs. GWAS, different breeds, sample number, and statistical power) genotype data available (microsatellites vs. SNP chips, and low-density vs. high-density chips), and modeling methods (single SNP vs. simultaneous fit). Although this causes SNPs of interest to vary, genes associated with them may belong to the same pathway. One such example is the transforming growth factor-β superfamily that has members implicated in changes in litter size and fertility in sheep[7,86,214] and increased ovulation rate (OR) in a unique cattle family[54,122,123].

The objective of this study was to conduct GWAS using the methodology of ssGWAS to identify genomic regions with replicable effects using 2 data sets, test if the inclusion of whole-genome sequencing showed SNPs of greater association, test if any gene pathways had greater enrichment for association with twinning and evaluate the potential for genomic selection in future genetic improvement programs.

IV. Materials and methods

No animals were used in this study, and ethical approval for the use of animals was thus deemed unnecessary.

Data

Estimations of sire phenotypes, such as daughter averages, were available from previous studies using calving records from 1994 to 1998[38], and from 1999 to 2008[190]. Estimates for bulls represented in both data sets were combined by taking a weighted average based on number of daughters in the respective data sets. The combined data set is referred to herein as data set B (DSB, n = 8,589). Calving records from 2010 to 2016 were obtained from AgSource Cooperative Services (Verona, WI) and Dairy Records Management Systems (Raleigh, NC). Data were cleaned as described in 2018 by Lett and Kirkpatrick[36] briefly, the data sets were checked for duplicated records, missing or partial IDs, and discrepancies in age to calving date. They were then merged and narrowed to only include sires with ≥100 daughters and herds with ≥100 cows. A total of 4,361,165 calving records encompassing 2,143,606 cows with single or repeated records were used to generate sire phenotypes (daughter averages). Sires remaining after cleaning comprised data set A (DSA, n = 3,154). Although calving records were unique to each data set, 706 bulls were represented in both DSA and DSB meaning that the data sets are not

fully independent. However, we chose to analyze data from these data sets separately to provide validation of results and in recognition of differences between data sets in calculation of phenotypes. Additionally, they are treated as older (DSB) and newer (DSA) for training and testing of genomic prediction. Phenotypes for DSA bulls were estimated based on an average of daughters' twinning rates. Daughter TW was calculated by estimating least squares means (LSM) using R v3.5.1 (R: The R Project for Statistical Computing (r-project.org)) with the model: $y = Xb + \varepsilon$, where y is the vector of multiple birth codes of 0 and 1 for single and multiple birth, X is an incidence matrix relating phenotypes to fixed effects of herd (levels = 5,311), calving year (levels = 7), season (levels = 4), and parity (levels = 4); b is the vector of fixed effects, and ε is the vector of random residual effect. The LSM estimates for herd, year, and season were expressed as deviations from the average for each effect. The LSM estimates for parity were expressed as deviations from mature equivalent (parity 4+). The TW estimate for each record was calculated by summing the multiple birth code (0,1) with the corrected LSM values for fixed effects. An average for each daughter was calculated and then finally an average for each sire. Given the variation in number of daughters per sire, sire phenotypes were weighted in the subsequent GWAS analyses following Garrick et al. (2009)[215]. A pedigree was developed for each data set using a cross reference file from National Association of Animal Breeders (Madison, WI). For each bull with a phenotype the matching ID was found in the reference file and its sire, dam, and year of birth was extracted. Any bull without a matching ID was considered a founder and a 0 was placed in position of sire, dam, and year of birth. A total of 158 bulls in DSB and 47 in DSA were considered founders.

Low-density (LD) 60K SNP data for the bulls were obtained from the Cooperative Dairy DNA Repository (n = 2,974) and Council on Dairy Cattle Breeding (CDCB, n = 2,267) for a

total of 4,720 bulls. Imputation to HD was done using Beagle v4.1[216–218] with a Holstein reference group of 736 animals previously genotyped with Illumina BovineHD (HD) and GeneSeek Genomic Profiler GGP F-250 (F250) chips. This was followed with imputation from HD to whole-genome sequence (WGS) using Fimpute version 3[219] with a population-based model and a reference population of 700 Holstein-Friesian animals from Run 7 Tau of the 1000 Bull Genomes project[220]. Variants were also pruned based on linkage disequilibrium using PLINK v1.9 with window set to 50, step set to 5, and r^2 set to 0.98 to eliminate variants which were proxies for others. Quality control included excluding variants failing a Hardy Weinberger Equilibrium (HWE) test at $P < 1 \times 10^{-10}$ or minor allele frequency <0.02. From 60,026,972 initial variants, 7,994,662 variants remained after quality control. A total of 1,897 DSA and 3,037 DSB bulls were available with both genotypes and phenotypes. This corresponds to 60.15% of the total number of phenotyped bulls for DSA (1,897 of 3,154) and 35.36% for DSB (3,037 of 8,589). Genomic locations of all variants are based on bovine reference assembly ARS-UCD1.2[221]. Quality control for HWE, minor allele frequency, and linkage disequilibrium pruning was done using PLINK v1.9[222]. Phenotype estimation of LSM and imputation used computational resources and assistance of the UW-Madison Center for High Throughput Computing in the Department of Computer Sciences.

Genome-Wide Association Study with WGS

GWAS was conducted using the BLUPF90 family of software[223] including RE-NUMF90 v1.145, BLUPF90 v1.68, and POSTGSF90 v1.68. Single-step GWAS was implemented as in Wang et al. (2012, 2014)[206,224] and Aguilar et al. (2019)[225]. Briefly, RENUMF90 was used to convert the files to correct format, BLUPF90 calculates EBV, and POSTGSF90 estimates SNP-effects, P-values, and percentage of genetic variance explained by

overlapping 500 kb sliding windows with a step size of one SNP. Windows are calculated by dividing the genetic variance of the region by the total genetic variance[206]. A total of 7,994,662 variants were used in this analysis and was conducted on each data set individually. The parameter files used in ssGWAS were set up following the BLUPF90 manual[223] and used an animal model including additive animal effects:

$$y_{ij} = \mu + a_i + \varepsilon_j,$$

where y = daughter averages for twinning, $\mu = 1$, and a and ε are random variables for additive genetic effect and residual for animal i. Covariance between related individuals was accounted for by using an H matrix[226], which uses pedigree information for ungenotyped individuals and pedigree information and genotype data for genotyped individuals in determining relationship. Residual variance and genomic variance were estimated using AIREMLF90 using the LD genotype data. Both BLUPF90 and POSTGSF90 parameter files included the options weightedG, snp_p_value, and no_quality_control with the BLUPF90 also including sol se and POSTGSF90 also including window_variance_mbp 0.5 and windows_variance_type 1. Multiple correction testing was done using false discovery rate (FDR) calculated using the R package qvalue[227]. A threshold of $Q < 0.01$ was set to identify variants of interest and $Q < 0.001$ for those of strong association. Window variance thresholds were set to identify for the top 99.9% and 99.99% of the data. The P-values from DSA and DSB were combined for a meta-analysis using a weighted z transformation method via combine.test from package survcomp v1.38.0[228,229] in R v4.0. 1 (R: The R Project for Statistical Computing (r-project.org)). Weighting equaled the inverse of the prediction error variance of the SNP effect. Manhattan plots were generated using R v4.0.1 and the qqman v0.1.4 package[230]. All results were further compared to a list of genes of interest previously curated by literature review.

Pathway Analysis

Gene set enrichment analysis was conducted using 2 types of software. Initially, SNPs were annotated to genes and gene enrichment was conducted using MAGMA v1.09[231]. Default settings were used with the addition of 2 flags nonhuman and window = 500 (annotates SNPs to genes going ± 500 kb from the start and end of the gene). The annotation used a list of genes downloaded from Ensembl release 96. Gene enrichment was analyzed using the SNP P-value data generated from ssGWAS analysis for DSA and DSB separately. This task is used to generate a degree of association each gene has with the phenotype. Initially this is done by looking at the individual SNPs in a gene and combining the resulting P-values into a gene test-statistic. The software was set to run 2 base models, mean and top, and used to generate a gene-level P-value based off the gene test-statistic for that model (Z or X). Additionally, the software aggregates the resulting model P-values into one. Because P-values were used, additional genotype data were needed to calculate and account for linkage disequilibrium between the SNPs. This consisted of all the bulls before splitting (n = 4,377) and the SNPs used in the GWAS (n = 7,994,662). The gene sets and gene set analysis were conducted using the R package SetRank[232,233]. SetRank compiles information from different publicly available databases and eliminates overlapping gene sets. Gene set databases were generated following manual documentation to obtain information of Bos taurus and included (1) Kyoto Encyclopedia of Genes and Genomes (**KEGG**; KEGG: Kyoto Encyclopedia of Genes and Genomes)[234], (2) BioCyc (BioCyc Pathway/Genome Database Collection)[235], (3) Gene Ontology (**GO**) biological pro- cesses, cellular components, molecular function (Gene Ontology Resource)[236,237], (4) Reactome (Home – Reactome Pathway Database)[238], (5) WikiPathways (WikiPathways – WikiPathways)[239], and (6) a user-curated database. The user-curated database was based on prior

knowledge and literature review of involvement in folliculogenesis and previous indication in other species of influencing litter size (Supplemental Table S3.1, Supplementary Table. Twinning rate pathway analysis (figshare.com)). Implementation of SetRank used default settings. Genes were ranked based on the P-values from MAGMA gene enrichment. Results were compared between DSA and DSB to identify areas of overlap between the results. Sets of highest interest were identified by passing additional optional filters of corrected and adjusted P-value <0.001 or pSetRank <0.05[240].

Genomic Prediction

Genomic prediction was carried out using the PREDF90 v1.12 program from the BLUPF90 suite. This was done using DSB as training and using a subset of the genotyped bulls from DSA as testing. The subset (n = 1,340) was selected to include only those individuals with genotypes, those that are not part of DSB, and are not sires of individuals in DSB. Genotype data used in this analysis were those from the HD imputation and a subset of the imputed HD data that corresponded to the 79K SNPs currently used in genomic prediction by CDCB (personal communication, G. Wiggans, CDCB, Bowie, MD). After quality control as described previously (HWE $P < 1 \times 10^{-10}$ and minor allele frequency < 0.02) a total of 75,598 SNPs were considered for LD and 640,966 for HD. Implementation involved running WssGWAS as described by Zhang et al. (2016)[241] and Fragomeni et al. (2019)[242] on DSB for each genotype level (LD and HD). The difference between ssGWAS and WssGWAS is that BLUPF90 and POSTGSF90 are repeated iteratively with SNP weights being updated each iteration. A threshold for stopping the number of iterations was based on the change in correlation of EBV from BLUPF90 and estimated phenotype between iterations. The model and parameter file options were the same as used during ssGWAS (described above) with one addition to the post POSTGSF90 file of

which_weight. In this analysis, the threshold was set to 0.0001 and weighting during POSTGSF90 used method nonlinear A[242,243]. Then after WssGWAS was completed results from POSTGSF90 were used by PREDF90 to predict genomic EBV (GEBV) in the testing set. Accuracy was calculated as the correlation between phenotype (daughter average) and the GEBV from PREDF90. Bias was evaluated by regressing daughter average on GEBV as recommended by Daetwyler et al. (2013)[244] and the slope was reported for each regression.

V. Results and discussion

Genome-Wide Association Study with WGS

Results from ssGWAS using WGS were looked at each data set separately and in conjunction. The residual variance estimates were $8.27 \times 10^{-5} \pm 6.16 \times 10^{-6}$ for DSA and $1.82 \times 10^{-4} \pm 1.03 \times 10^{-5}$ for DSB and genetic variance was estimated at $1.82 \times 10^{-4} \pm 1.12 \times 10^{-5}$ for DSA and $3.77 \times 10^{-4} \pm 1.25 \times 10^{-5}$ for DSB. False discovery rate in the form of q-values showed no SNPs surpassing $q < 0.05$ in either DSA or DSB alone. In each data set the minimal q-value was 0.2775 DSA (Figure 3.1A) and 0.1467 DSB (Figure 3.1B). Results from window variance of DSA showed 10 different chromosomes with windows >99.9 percentile for a total of 11 regions (Figure 3.2A and Table 3.1); DSB also showed 11 different regions on 10 different chromosomes (Figure 3.2B and Table 3.1). There were 3 regions that had overlap between the 2 data sets found on BTA1, 11, and 21. In 2 of these regions, genes of interest partially or fully spanned the window which included *FSHR* and *LHCGR* (BTA11) and *ZSCAN2* (BTA21). A DSA window on BTA24 also partially or fully spanned the gene *TAF4B* (Table 3.1). Combining the P-values of the 2 data sets using weighted Z transformation indicated 1,214 SNPs with q-value <0.01 and 34 had q-values <0.001. Comparing results with the curated list of candidate genes showed 3 different genes (*GDF9*, *IGFBP2*, and *BMP15*) with at least one variant ± 500 Kb from the

beginning or end of the gene with a q-value <0.01 and 2 genes (*LHCGR*, and *FSHR*) with at least one variant within 500Kb at q-value <0.001 (Figure 3.1C). Additionally, 13 significant (q <0.01) variants of the almost 8 million used in this analysis were seen in *LHCGR*, 16 were in *FSHR*, and one was in *GDF9*. A total of 11 different chromosomes contained associated SNPs with q-values <0.01 (Table 3.2). The strongest peak was found on BTA11 with all q-value <0.001 variants located in this region (Figure 3.3A). The peak containing the most significant SNPs ranged from 28 to 32 Mb with the candidate genes *LHCGR* and *FSHR* located at 30.98 to 31.04 Mb and 31.26 to 31.45 Mb, respectively (Figure 3.3B).

These results show a strong association with BTA11 and the 28 to 32 Mb region. Although neither dataset alone showed q-values <0.05 the combined values showed a very strong association with TW in this region. Additionally, both data sets showed an overlapping window within this region. The window was the most associated window explaining 0.752% of the variance and falling within the 99.99 percentile, whereas for DSA it was the third-most associated window at 0.3516% variance explained and just outside the 99.99 percentile. The association observed for this region in the current study is supported by a recent report of significant association for twinning rate in a similar genomic region in the Swiss Holstein population[245]. Found within the region are 2 genes, LHCGR and FSHR, involved in folliculogenesis making it a strong positional candidate region. Being the gonadotropin receptors to key hormones, luteinizing hormone (LH) and follicle- stimulating hormone (FSH), they play an important role in follicular development and ovulation[4,10]. Mutations involving these genes have been shown to have detrimental effects on human reproduction[246,247]. Studies in sheep have also found these genes as candidates for association with litter size[86,140]. A study in Holstein heifers suggested a missense mutation in *LHCGR* is associated with superovulation traits in 127

animals[248]. This mutation corresponds with SNP rs41256848, which is located on BTA11 at 30978812. The variant sharing the same location in WGS data (11_30978812), had a low association (P = 0.067 DSA, P = 0.139 DSB, P = 0.034 and FDR = 0.92 combined P-values) indicating it is not the causative variant for the association seen in this region. Looking at the variants used in the GWAS located within *LHCGR* with Ensembl variant effect prediction indicated the predicted consequences are related to introns and noncoding transcript variant[249]. Additionally, this genomic region does not overlap with any previously documented twinning rate or OR QTL in the QTL database. The other strong peak on BTA11 (85.3 to 88.8 Mb) does not overlap with previous QTLs or genes from the candidate gene list. However, the most associated SNP, 11_86316394 (P = 6.38 × 10−8, FDR = 0.0019), in this region falls in the gene *GREB1*, which is an estrogen response gene that has primarily been associated with endometritis and breast cancer in humans[246,250]. It could have an unknown role in bovine reproduction and relation to TW. Another set of strong peaks are on BTA19 at location 26.9 to 34.9 Mb and 44.7 to 52.7 Mb. Although the first peak does not overlap with a previously indicated QTL, the second overlaps with the start of a previously identified QTL region spanning 98 to 126 centimorgans[251]. The last chromosome harboring 2 peaks is BTA25. Comparison between data sets in both P-values and window variance only indicated DSA having a strong association with these regions. They also do not overlap with any other previously reported QTLs for twinning or OR. Comparison with 2 previously fine-mapped chromosomes (BTA5 and BTA14) yielded little overlap. BTA5 showed an overlap with a previous QTL region for OR[251] in combined P-value but neither DSA or DSB alone showed strong support for this. It is surprising not to see any SNPs significant at a suggestive level in the *IGF1* region of BTA5 (~66.2 Mb) that has been previously associated with twinning rate in Holstein as well as other cattle Populations[202,203,252–

[254]. Given the objective to identify variants and regions associated across the different data sets, the strongest candidates would be the two regions seen on BTA11.

Pathway Analysis

The gene ranking results from MAGMA showed 62 genes in DSA and 46 in DSB with an aggregated P- value <0.0001. Of these only 11 overlapped between them and only 2 of these had a rank within ±10 of each other. The gene FOXN2 had the same rank within both data sets. Three genes overlapped between DSB and the curated genes of interest used in comparisons for the GWAS. Included in the top 3 were *FSHR* and *LHCGR*. Data set A showed no overlap with the genes of interest.

In gene set enrichment, a total of 1,928 sets were considered significant for DSA and after correction a total of 116 sets remained. Although for DSB, 1,723 sets were initially considered, a total of 107 sets remained after correction. A total of 7 sets are shared between the 2 data sets (Table 3.3), and 2 of these passed the additional filter.

Given the set analysis is based on rank of the genes, it is not surprising that there are few overlapping pathways between data sets (~6% of the total possible). A contributing factor to this is also a limitation of pathway analysis which is the quality of the gene annotation. Genes may be missed, improperly called, or not called in a specific pathway which influences the results. Thus, SNPs are not always annotated correctly to genes, affecting the scoring of a gene and its pathway. Similarly, there were a low number of pathways that showed agreement. One of the gene sets that also passed further filter was MutSalpha complex. This gene set contains 2 genes, *MSH2* located on BTA11 at 29.8 to 29.89 Mb and *MSH6* located also on BTA11 at 30.12 to 30.14 Mb. Both genes are part of the DNA mismatch repair system[255]. Proximity of both genes

to the region of interest in the GWAS may be influencing the association given to these genes and thus this pathway. However, in various species studies, including human, DNA repair genes have been shown to be overexpressed in oocytes and may play a role in oocyte quality[256–258]. The other gene set that passed further filtering was a broad GO class called cellular process involved in reproduction in multicellular organisms. It included a total of 135 different genes of which includes *MSH2* as well as 2 genes of interest, *LHCGR* and *FSHR*. Of the remaining 5 sets overlapping, the only one that stands out is the user defined pathway #5 prolactin signaling pathway, which was included for containing candidate gene *LHCGR*.

Several pathways have been implicated in influencing folliculogenesis and potentially it is a combination of these that contributed to increased [2597,72,106]. Additionally, the above pathways all have interactions with the transforming growth factor-β superfamily which is known to be involved with folliculogenesis and variation in multiple birth in multiple species[259]. A possible future direction would be to create a reference set of genes to help eliminate any source bias, for example genes expressed in the hypothalamus, anterior pituitary, and ovary representing the hypothalamic-pituitary-ovarian axis which regulates female reproduction.

Genomic Prediction

Genomic prediction used SNP predictions of a training set to predict GEVB in a testing set. Evaluation of this was done using results from DSB WssGWAS with LD and HD genotypes as training and tested by predicting GEBV of 1,340 nonoverlapping genotyped bulls in DSA. Correlation between daughter average and GEBV ranged from 0.4235 (iteration 1) to 0.4244 (iteration 3) for LD, and from 0.4272 (iteration 1) to 0.4287 (iteration 4) for HD. Estimate bias was assessed by regression of daughter averages on GEBV. The slopes ranged from 0.7171

(iteration 1) to 0.6934 (iteration 3) for LD and from 0.7303 (iteration 1) to 0.7048 (iteration 4) in HD (Table 3.4, Supplementary table S3.2 shows 5 fold cross-validation).

Results indicate that the prediction estimates change only slightly with iterations of WssGWAS. Across genotype levels the accuracy increased marginally, whereas the slope decreased by iteration. The decrease in slopes indicated that iterations biased the estimates. When comparing LD prediction to HD, the prediction benefited in both accuracy and slope from higher SNP number. Accuracy of these estimates is lower than the average reliability across multiple production traits (net merit, milk, fat, protein, productive life, SCS, daughter pregnancy rate) for Holstein dairy cattle (71 vs. 38.16–56.28%)[260]. Previous reports showed accuracies of 0.14 unadjusted daughter averages, 0.34 for PTAs[190] and 0.39[261]. The differences in methods make these not perfectly comparable, but the results presented here are similar to those previously reported using PTAs. One contributing factor to a lower accuracy is low heritability, which is seen with twinning rate. Previous heritability estimates range between 0.017 to 0.09 using linear models and 0.08 to 0.14 with threshold models[36,38,63]. Another is the moderate sample size of the genotyped individuals making up the reference set (DSB). Although the total number of phenotyped bulls was 8,589, only a small fraction of these were genotyped (2,961). Increasing the number of bulls with both genotypes and phenotypes should lead to improved accuracy of SNP effect estimates and subsequent genomic prediction. Additionally, an unaccounted-for and unidentified environmental factor could be contributing. Not accounting for unidentified environmental factors could cause a decrease in accuracy leading to the moderate accuracy of the prediction seen. Lastly, the correlation between the estimated TW for sires in DSB and the EBV from BLUPF90 was only 0.7, whereas the correlation between estimated TW in DSA bulls and EBV was 0.9. This indicates that the difference in TW estimation could also

have affected the predictions. One potential difference was that in DSA a criterion of ≥100 calving records per herd was imposed, which would affect the accuracy of the LSM calculated for herds. Ideally, a more accurate measurement could be used. One such measure is OR by counting corpora lutea. Genetic correlation between OR and TW is 0.66 to 0.90 meaning selection based on OR could indirectly select for TW[35,52]; however, measuring OR is more labor intensive requiring more equipment, training, and time, and is not routinely performed on farms. Calving data including factors such as herd, calving year, calving season, parity, and birth type typically are routinely recorded and thus more available than OR. These records also more closely reflect what farmers use when making management decisions on farm.

Genomic selection for health traits using producer- recorded health trait data has been proposed[262] and recently implemented in the United States[263]. Twinning has quantifiable economic costs, which should make it feasible to readily incorporate into a selection index for dairy health traits by CDCB. Further evidence of the feasibility of genomic selection for twinning is the recent commercial development of this approach[264].

VI. Conclusions

Utilizing calving records from 2 distinct time frames allowed for comparison between them to identify regions associated with daughter averages for twinning. The strongest associated SNPs are upstream from 2 genes, *LHCGR* and *FSHR*, which are involved in folliculogenesis and are key gonadotropin receptors. Additionally, the window spanning part of each gene accounted for the most variance in DSB and was among the top 3 in DSA. Their involvement with follicular development makes them strong candidates for future analysis, both to identify causal genetic variants and to understand the mechanism underlying genetic variation in twinning. Pathway analysis further supports this region by implicating sets that have genes found within

the area. There was limited correspondence between genomic locations associated with TW in these analyses and previous QTL-mapping studies. The correlation between daughter average and GEBV (0.421 LD and 0.426 HD) in genomic prediction analysis indicates the potential of genomic selection for reducing twinning in the dairy cattle industry.

CHAPTER 4

Identification of copy number variations in Jersey cattle using whole genome sequencing.

I. Preface

At the time of submission this chapter was ready for submission to Animal Genetics but has not been submitted.

Authorship is as follows: **Beth M. Lett, Xian Qiao, and Brian W. Kirkpatrick.**

Acknowledgment and thanks to Taylor Schaefer for assistance with laboratory CNV validation and Alex Bagnato and Maria Strillacci for the discussion on copy number variants.

Data availability set to release at time of publication or October 2022 whichever comes first: Novel CNVs compared with the dVGA database (2020) were deposited in the European Variation Archive (EVA)[265] at EMBL-EBI under accession number PRJEB52447 (https://www.ebi.ac.uk/eva/?eva-study=PRJEB52447). Sequencing data was submitted to the SRA database for the 20 Jersey AI bulls and four MARC Twinner Sires with project accession number: PRJNA826358.

Formatting of figures, tables, and references align with the remainder of the thesis.

II. Summary

Copy number variants have the potential to cause greater phenotypic change due to their size. Knowledge of the location of these variants will facilitate future research examining their direct effects on trait variation. In this study we looked at identifying variants specifically found in Jersey cattle when compared to a genetically diverse group with no Jersey breed background. Consensus CNV calls were retained from four different detection methods with a 93.02% accuracy of detection based on literature review and/or PCR-based genotyping for validation. A total of 1,269 different CNVs were detected with 740 being specific to the Jersey samples used in this study. Additionally, 648 of the 1,269 were novel when compared to the variant archive database. Screening for deleterious recessive deletion alleles looked at the 1,210 CNVs found in at least one Jersey and identified deletions (171) for which none of the initial 20 sires were homozygous. These deletions were further screened in an additional 36 publicly available sequenced Jersey animals. Absence of deletion homozygotes would provide evidence of embryo lethality. Four of the resulting candidates were PCR genotyped in a random sample of Jersey cows, and deletion homozygotes were observed for all four. Additionally, there are 33 deletions with absence of homozygotes not tested. Overall using multiple methods to detect structural variants in short read data yields high accuracy but misses some positives in favor of limiting false positives. Breakpoint resolution remains a challenge even when using short read sequencing data.

III. Introduction

Copy number variations are a type of structural variant **(SV)** that present abnormal changes in number of genome copies inherited. These are typically associated with losses, also

referred to as deletions (**DEL**), or gains, also called duplications **(DUP**). While SVs are less common than SNPs, they present potential for greater functional impact due to their larger size (> 50 bp[191]). Studies have looked at utilizing these regions to pinpoint breed differences[266]. Further, across various species CNVs have been implicated as involved in various phenotypic traits and disease responses. These range from changes in coat color and pattern[267,268] to extremes in function such as embryo and fetal lethality[164,269].

Initially, microarray data was used to infer CNVs and detect regions of the genome that were more susceptible to copy number changes. Previous studies in cattle using microarray data could not always pinpoint exact locations leading to identification of large sections spanning in some cases most of a chromosome[160,162,266]. With the development of next generation sequencing **(NGS,** i.e. millions of short, paired-end reads of ~150 bp**),** methods of detection have been developed to better fine map and identify CNVs. These methods utilize different aspects of NGS technology to predict SV and each has their own limitations. Briefly these include: 1) read depth **(RD)** which utilizes depth of coverage, 2) read-pair **(RP)** that utilizes information of read-pair insert size, orientation, and alignment to infer changes compared to a reference, 3) split-read **(SR)** which, similar to RP, uses information gained from read-pairs when one read maps reliably to the reference sequence while the other does not, 4) assembly **(AS)** which compares a *de novo* assembly of reads to reference, and 5) combined approaches **(CA)** which takes various methods in combination[158,270]. Since the introduction of NGS multiple studies have investigated detection of CNVs using NGS in cattle[163,271,272]. Many of these studies used earlier bovine reference genome assemblies and only used one or two methods individually to detect CNVs.

In addition, few studies have taken this information further to investigate potential negative impacts these large segmental deletions have on the cattle population. One utilization of

this information is identification of deletions that appear without homozygotes in the sampled population suggesting the potential for embryo or fetal lethality. Previously, a similar study was conducted in Belgian beef and New Zealand dairy cattle to screen for genetic variations that are candidates for embryo lethality by absence of homozygotes[192]. Using similar logic, a study looking at Holstein, Jersey, and Nordic Red cattle scanned NGS for large deletions (100 bp to 1 Mb in size) to pinpoint potential embryonic lethal deletions in this population[163]. They tested the deletion lethality potential by comparing cow and known mouse lethal genes.

The following study had two objectives. The first objective was identification of putative CNVs in the US Jersey population using NGS data and validating them by PCR-based assays or other independent means. The second objective was to identify deletions which are potentially associated with embryonic or fetal lethality in Jersey cattle based on absence of homozygous genotypes in a sample from the Jersey population. The goal was to provide additional information about CNVs in cattle, particularly the Jersey breed, and provide information regarding potential deleterious alleles that impact reproductive efficiency.

IV. Materials and methods

DNA and sequencing

DNA was extracted from semen of twenty US Jersey artificial insemination **(AI)** sires, four USDA MARC Twinner bulls with no Jersey background, and extracted from ear punch tissue of one mixed heritage bull with Jersey in the composition, C041, from a unique cattle family at the University of Wisconsin – Madison. The MARC Twinner bulls are from a genetically diverse population with twelve different breeds represented[49]. Illumina paired-end short reads were generated at about 15x coverage for the Jersey and Twinner sires and 50x for

C041 by Beijing Genomics Institute (BGI). The short reads were then aligned to the current bovine reference assembly ARS-UCD1.2 following the 1000 Bull Genomes Project[220,273].

Analysis 1: CNV detection and validation

CNV detection utilized three different publicly available bioinformatic software implemented in four different methods. These methods included: 1) CNVnator[274], a read depth (**RD**) method that uses a mean-shift approach with read coverage to predict DUP and DEL in individual samples, 2) DELLY[275], a CA using RP and SR sequentially to detect SVs in individual samples, 3) LUMPY[276] single sample method **(LS)** which is another CA using RP and SR information but analyzing concurrently instead of sequentially, and 4) LUMPY population method **(LP)** which uses LS methods along with additional steps to combine information across multiple samples to generate a single output. LS and LP were run using the *smoove* docker as recommended by LUMPY authors[277]. All 25 samples were used to generate the population prior knowledge in LP. For the remainder of detection and the other methods, each sample was analyzed individually. All CNV detection methods were run using computational resources and assistance of the UW-Madison Center for High Throughput Computing **(CHTC)** in the Department of Computer Sciences.

Consensus within and between samples was generated by SURVIVOR, a SV toolset allowing for merging and comparing SVs across different pieces of information[278]. To generate within-sample consensus, CNVs were retained if: A) the predicted size was > 30 bp, B) all four methods agreed on the type and strand called and C) the predicted start and end of each method were within 1 kb of each other. To generate a list of CNVs across the samples SURVIVOR was run on the 24 within consensus results with the only change being that the methods no longer had to agree with type and strand. C041 was excluded from the remaining analysis due to mixed

breed background containing Jersey. This list of CNVs was then split based on if the variant was only found in the 20 Jersey bulls (**JE**), only the four MARC Twinner bulls (**TR**), and the remainder being found in all (**ALL**).

Accuracy of CNV detection methods was validated by a subset of putative CNVs which were observed specifically in the Jersey bulls. These CNVs were annotated to genes using MAGMA[231] annotation with a gene map generated based on an ENSEMBL release 96 gene file for bovine assembly ARS-UCD1.2[279]. The CNV list was then reduced to those in functional gene regions based on location spanning at least a coding sequence, exon, 5' or 3' UTR, start or stop codon, and/or transcripts as specified by the ENSEMBL database (.GFT file, release 96). CNVs spanning, overlapping, or located within functional regions were validated using a two-step approach. In the first step CNVs were compared with three different sources to see if they had been previously detected. This was conducted with a python script (Python v3.6) that compared each CNV to a previously detected CNV of the same type and allowed ± 5,000 bp in CNV start or end. Sources of previous CNVs included Ensembl structural variations which includes Database of Genomic Variants Archive **(dVGA)** and dbSNP (accessed 03/12/2020) [279], and two previous reports in 2017 by Chen et al.[271] and in 2019 by Kommadath et al.[272]. Any CNV detected in one of these sources was considered validated and not tested further. Copy number variants not previously reported were validated by PCR-based assays. Primers were designed to allow for unique products for alternative CNV alleles. This was done with one pair spanning the full size of the variant and the third primer located in the middle. If the normal product was small enough < 1000 bp and a second reverse/forward primer could not be designed, only a single pair of primers was used. Primers (Supplementary Table S4.1) were designed using NCBI Primer-BLAST and unique products were checked using NCBI ePrimer in Primer-BLAST[280].

Using the same script as the literature review overlap, comparison of all detected and retained CNVs was done with the dVGA (accessed 3/12/20220). Those that were not found in the dVGA within ± 5,000 bp of the start/end were considered novel to the database. Novel variants were submitted to the dVGA for inclusion in future releases.

Analysis 2: Deleterious deletion identification and validation

Identification of embryonic lethal candidates for Jersey cattle involved first genotyping all the deletions in the Jersey bulls. Genotyping was done using SVTyper[281] and a list of CNVs found in at least one Jersey from this study (n = 1,210 CNVs). Resulting genotype calls consisted of NN (normal), ND (single copy of deletion), DD (homozygous for deletion), and UNK (not enough support to call the genotype). From these genotype calls a list of deletions with zero UNK and DD genotypes was generated. To narrow the list further, the variants were genotyped in a group of Jersey animals available in the NCBI Sequence Read Archive **(SRA**)[282]. This data was downloaded (accessed June – July 2020), aligned to bovine reference ARS-UCD1.2, and genotyped in the same manner as described earlier. Only samples with depth of coverage greater than 10x based on GATK DepthOfCoverage[283] command were used. The genotype calls from the twenty sires were combined with the open-source data. As before, the list was narrowed to deletions with no DD genotypes but included those with <10% of the samples having UNK. Frequency of the deletion allele was calculated and those with a frequency between 0.15 and 0.45 were considered in testing absence of deletion homozygotes. The lower threshold was chosen to attain a probability of ~0.01 or less for absence of homozygotes by chance in a sample of ~200 individuals. The upper threshold was arbitrary and reflected an expectation that a true deleterious allele would not be expected to occur at a frequency exceeding this magnitude.

Testing of deletions as deleterious embryo lethal candidates involved genotyping the deletion in a large sample of Jersey cows. A total of 192 Jersey cows, whose DNA was available from previous studies[284], were chosen for this purpose based on each having a unique sire. Genotyping utilized 96 of the 192 samples initially, with the remainder available for additional genotyping should no deletion homozygotes be observed in the first set of 96. Initially each deletion had primers designed as described before in analysis 1 validation. Following primer design visual inspection using Golden Helix GenomeBrowse[285] was conducted to verify that the primers would span the deletion, that the genotype calls were correct, and to check breakpoint accuracy. To generate additional information on breakpoints, Sanger sequencing was generated from a ND genotype bull for each candidate deletion. This was conducted by first performing PCR in separate reactions for alternative alleles and then either gel purifying (QIA quick gel extraction kit, QIAGEN LLC, Germantown, MD) or directly cleaning (QIAquick PCR purification kit, QIAGEN LLC, Germantown, MD) the PCR product prior to Sanger sequencing. Purified PCR products were then used as template for Sanger sequencing using a BigDye (ThermoFisher Scientific, Waltham, MA) protocol with initial incubation at 96 °C for 1 min and 40 cycles of 20s at 96 °C, 30s at 50 °C and 4 min at 60 °C. The sequencing reaction used 5 picomoles of primer in a 15 µl reaction with the amount of PCR product used as template varying depending on PCR product length (~10 ng/100 bp). Clean up and capillary electrophoresis were performed by the University of Wisconsin-Madison Biotechnology Center. Sequencing results were trimmed, aligned to bovine reference ARS-UCD1.2, and viewed using Geneious prime 2021.2.2 (http://www.geneious.com). Information from both Sanger sequencing and GenomeBrowse were used to correct the predicted size of the PCR products (Table 4.3).

PCR amplification used a basic protocol with a 1.5 min initial denature step, followed by a series of cycles of 30s at 95 °C, 30s at an ideal annealing temperature, and 1 min extension phase at 72 °C, and then a final extension step of 5 min at 72 °C. The number of cycles and optimal annealing temperature depended on primer pair (Table 4.3) and was determined empirically by preliminary PCR analyses using a thermal cycler with annealing temperature gradient (RoboCycler, Stratagene Inc.). Following PCR, products were run on a 1% agarose gel and genotypes were called from visual inspection of the gel.

V. Results

Sequencing data

Next-generation paired-end sequence data was successfully generated for 25 animals (Table 4.1). Read depth after processing and variant calling per the 1000 Bull Genomes Consortium pipeline ranged from 12.55 to 17.31-fold depth of coverage, except for sample C041 which was sequenced at greater depth and had 46.93-fold depth of coverage.

Analysis 1: CNV detection and validation

Of the four methods employed, only CNVnator was limited to detection of CNVs (DEL and DUP) while, the other three methods report multiple types of SVs. Reports herein will consider only CNVs (Table 4.2). For CNVnator the average number of CNVs detected was 3,686 with an average size of 12,737 bp and median of 4,000 bp. This method detected an average of 2,843 DEL and 843 DUP across the 24 samples (excluding C041). DELLY detected the most SVs for the individual sample runs, discovering an average of 19,317 SV across all samples of which an average 9,783 were CNVs. The average size for this method was 1,247,920 bp with the size median being 2,741 bp. The 9,783 CNVs included an average of 7,133 DEL and

2,650 DUP. The LS method identified an average of 5,430 CNVs from a total of 7,651 SV, on average. This method had an average size of CNVs of 123,390 bp and a median size of 660 bp. An average of 4,935 DEL and 494 DUP were observed. The LP method reports only a single file result across the 25 samples. It detected a total of 32,200 SV with a subset of 18,442 being CNVs. This method detected the greatest number of putative SV and CNVs of the four methods. The average size of CNVs by this method was 280,760 bp and the median size was 787 bp. Total CNVs included 15,235 DEL and 3,207 DUP.

Within-sample consensus across methods showed an average of 295 CNVs across the samples (Table 4.2). The smallest number of CNVs (n=207) was found in a non-Jersey bull, while the largest number (n=426) was found in a Jersey sire. In the individual consensus the average size of CNV was 1,287 bp. Like the individual methods, the consensus found on average a greater number of DEL (n=284) than DUP (n=11). Between-sample analysis detected 1,269 unique CNVs. Of these, a total of 1,214 were DEL with the remainder being DUP (n=55), following the pattern of the within-sample methods. In the Jersey bulls (n=20), results showed 1,210 different CNVs. For the non-Jersey bulls, 529 CNVs were observed across four samples. Comparison between the different subsets showed a total of 740 CNVs unique to this sample of Jersey bulls. The remaining 529 CNVs split into those found specifically in the non-Jersey bulls (n =59) and those shared between the groups (n= 470). The distribution of CNV types in the Jersey bulls consisted of 699 DEL and 41 DUP.

For validation, a total of 158 CNVs found only in Jersey's were identified as being within genes and being characterized (gene description is known, e.g. not uncharacterized LOC404063, Figure 4.1). These CNVs overlapped 159 different genes with one DUP predicted to be 85,551 bp, spanning two genes. Deletions accounted for 92.4% and duplications 7.6% of the CNVs

identified. This list was further narrowed to those that spanned functional regions, as described earlier, leaving a remainder of 86 (Figure 4.1). When comparing the variant by type and location with database and literature review, 73 (84.88%) overlapped. Validation of the remaining 13 CNVs was attempted using PCR. For three of these (one DEL and two DUP), primers producing unique products were unattainable due to repetitive DNA sequences. For the remaining ten, seven were validated by PCR and three failed. One failed to produce a normal product and upon closer inspection (visualization on Golden Helix GenomeBrowse and genotyping) appears to be a fixed deletion in this population sample. The other two failed due to off-target amplification making results ambiguous. Thus, of the CNVs examined, 93.02% were validated (Figure 4.1). Comparison of the 1,269 CNVs with the dVGA database (downloaded 3-13-2020) showed 648 were novel when allowing ± 5,000 bp for the start/stop location.

Analysis 2: Potential deleterious deletion candidates' identification and validation

From the list of 1,210 CNVs found in at least one Jersey bull, 171 deletions lacked individuals homozygous for the deletion allele (genotype DD). These were further genotyped in an additional 36 Jersey bulls from the SRA database with coverage > 10x. With the addition of these samples the set of 171 was reduced to 104 deletions with no DD genotypes and <10% UNK genotype. The frequency of the normal and deletion alleles was calculated for the remainder using the 36 SRA bulls and the 20 Jersey AI sires (n = 56). The distribution of the D allele ranged from 0.027 to 0.5 with a median of 0.263 and mean of 0.264. From these a subset was chosen with a D allele frequency between 0.15 and 0.45 (n = 66). Primers were designed as described for CNV detection validation for 34 of these 66. These 34 were selected based on prioritizing three aspects: those found within genes, those with higher frequency of the deletion allele, and utilization of default parameters during primer design of repeat filter and avoiding low

complexity filter. Visual inspection using GenomeBrowse, and Sanger Sequence showed only four of the 34 candidate deletions having ≤ 1 Jersey AI sires showing zero read pile up in the region of interest (Table 4.4 and Figure 4.2). Further examination of these four CNVs by genotyping Jersey cows showed four or more animals with genotypes homozygous for the deletion allele for all CNVs (Table 4.4).

II. VI. Discussion

Analysis 1: CNV detection and validation

Each method yielded varying number of CNVs with various average sizes. Of the single sample methods DELLY predicted the most with 9,783 being the average with an average size of 1,245,181 bp, by far the largest average size (next largest was 280,760 bp in the population-based method). The variation in number and size explains why only a small portion of these (295 CNVs average) were identified as a consensus. It is concerning that only a small number CNVs overlapped between methods even though the inputs for single-step methods were the same and the one additional sample included in the population implementation was one that should increase detection being of greater coverage. Despite the minimal overlap of detection methods, validation based on literature review, database search, and PCR confirmed 93.02% as true positives indicating that consensus from four different methods limits the number of false positives detected. As expected, each method detected more deletions than duplications. This is due to limitations of detection methods typically caused by repeat regions in the genome. One solution is the use of long-read sequence data; as this approach becomes more affordable and read lengths become longer (now tens of thousands of bp on average)[150,286], using such technologies may improve detection of segment duplications since the technology has the capability to read through large repeats.

A total of 159 genes overlapped with at least one CNV. This accounted for 0.58% of the total genes available from Ensembl (n = 27,233). This number does not take into account the genotype state of the variant (i.e., if it is homozygous or heterozygous). Additionally, the function of the gene overlapped by the CNV may be compensated by actions of another gene. Alternatively, the gene function may simply be nonessential for life; distinguishing between redundance and non-essential function is not simple. A recent study reported 167 natural gene knockouts in cattle that were present in the homozygote state, indicating that the genes were nonessential for life[163] Since gene function and expression is beyond the scope of this work, it is unknown what effects the reported variants have.

In depth review of the overlapping dVGA variants showed that a majority of CNVs overlapped with those reported by a 2018 study conducted by Mesbah-Uddin et al.[163]. This is not surprising given that the study conducted by Mesbah-Uddin et al. utilized next-generation sequence data as in the current study, whereas many of the previous studies submitted to the database were conducted using array data. Those reported using array data are typically reported as copy number regions which are regions that tend to have greater number of copy number changes rather than defined individual events. The advantage of the short read data has been the ability to resolve the breakpoints. While 48.93% overlapped with the database, there were still 648 novel variants added that could be examined in the future. As sequencing technologies improve, the ability to identify and validate SV will increase as will the ability to further study the effects of these large genome variations.

Analysis 2:

While the 34 deletions selected for genotyping were observed to have no homozygotes in the initial NGS samples, only a fraction of them had concordance regarding the genotype

determined by different methods (e.g. ND by SVTyper but DD by PCR genotype or visual inspection). A large part of this is due to breakpoint resolution of the original CNV detection. Examination of PCR product sequencing revealed that most deletion size predictions were greater than the actual size. Typically, either one or both breakpoints were incorrectly identified. Since SVTyper uses read support to agree or disagree with the different alleles of the SV, if the deletion starts or ends in a different location than predicted, reads would be found matching the reference supporting that the genotype should by NN or ND rather than DD. The other factor effecting genotyping is the reference. As the predictions are based on the reference a miss-called genotype may reflect a variation of the reference that is not seen in population sequenced. This may be the case where closer inspection of the VCF file of the deletions indicates the bovine reference allele as "N" and the alternative as "Del" wherein the 20 Jersey AI bulls show no coverage through a portion of the predicted variant but were genotyped "ND" (n=19) or "NN" (n=1). Again, the predicted start and stop did not match visual location on either Geneious prime analysis of Sanger sequence data nor Golden Helix GenomeBrowse of the aligned short reads. This is another aspect where long read sequencing would be advantageous to correct and refine breakpoint detection.

For the four deletions that were subjected to genotyping in 96 Jersey cows, all showed at least four homozygotes. In all cases this was despite the frequency of the deletion being low (< 0.26) in both the next-generation sequence data (n=56) and the PCR genotyped samples. Given the low frequency it is not surprising that there had been no homozygotes in the original sample of 20 bulls and there may have been miss-genotyping in the SRA data (visual inspection was not performed on those). For two of these four CNVs < 5% of the animals were homozygous for the deletion. There are 33 other deletions that were not extensively examined that may still be of

interest. In addition, the stringent parameters used to limit false positive SV calls may have eliminated some true positives that would have been of interest.

VII. Limitations

A major limitation of this study are the numbers of DNA samples used for the initial CNV detection and genotyping. Both in the initial detection and follow up screening. While these numbers were low there were still 51.06% novel variants detected and given the high accuracy (93%) of those validated the likelihood these are true CNVs remains high. Regarding screening for CNVs with potential embryo lethality the limitation is both in the low numbers (n = 56) for generating the putative candidate CNV list and in the stringent criteria of retainment of CNVs matching across methods. Utilization of the SRA database to increase both the number of Jersey and non-Jersey samples at the start of CNV detection would have been advantageous. In addition, principal component analysis could be used to assess breed make-up of the samples. An additional limitation of this study is the inherent weakness of short-read sequence data for CNV detection in regions with repetitive sequence motifs vs long-read sequence data.

VIII. Conclusions

Sequencing data provides deeper insights into structural variants. As the technology improves and resolution increases, the accuracy of variant detection will also increase. For now, utilization of consensus across multiple methods of detection in short reads yields high true positive rates (93.02% validated). This type of calling does risk excluding variants in favor of avoiding false positives. Though those intensively examined here did not yield candidates for embryonic lethality, screening for deletions that do not appear in the homozygous state does have the potential to identify deleterious variants. However, to more effectively detect deleterious variants

will require greater numbers of animals in the initial screening at the sequence level due to most cases of homozygote absence simply being a reflection of low allele frequency. Additionally, future CNV detection would benefit from long read sequencing technology to improve accuracy of the breakpoint identification and genotype calls. As a result of this work 648 CNVs have been added to the dVGA providing additional resources for future studies.

CHAPTER 5

Screening for novel causative mutation candidates for the Trio allele

II. I. Preface

At the time of submission, this chapter contains the initial strict screening of the region for variants of interest; prior to publication a less strict screening of variants found in either homozygous individual should be conducted.

Authorships at this stage: Beth M. Lett, Dean M. Sanders, Alvaro Garcia-Guerra, Ricky L. Monson, Milo C. Wiltbank and Brian W. Kirkpatrick

Formatting of figures, tables, and references aligns with the remainder of the thesis.

II. Abstract

Increased ovulation rate lends to increased opportunities for birth of multiple calves, which is potentially beneficial to beef producers. A region on BTA10 was previously indicated as the location harboring a variant causing increased ovulation rate in a cattle family. This allele, called the Trio allele, caused ~ 3x greater ovulation per cycle in carriers compared to non-carrier siblings. The objective of this study was to identify candidates for the causative mutation responsible for the Trio allele phenotype. The breed in which the mutation occurred was determined using 10 members of the Trio family and 28 SNPs within a 1.2 Mb positional candidate region on BTA10. Comparing this haplotype with the Bovine HapMap project genotype data the breed with the highest frequency (0.3) of the haplotype was Hereford. Variant detection utilized short and long read sequence data generated from two individuals homozygous for the Trio allele. Screening the 1.2 Mb positional candidate region for variants and comparing 1) with the 1000 Bull Genomes project, and 2) between the two homozygous individuals indicated only one SNP both common to the two individuals and unique to them as well. Genotyping of this SNP indicated 100% correspondence with the inferred Trio allele genotypes of 85 individuals. Additionally, this variant was absent in two non-Trio populations with high likelihood to possess the allele: Hereford (n= 98) and USDA MARC Twinner herd (n=78). In this strict variant screening, this SNP is a strong variant for further functional testing. An effect of the variant on *SMAD6* overexpression would need to be documented to make a case for causality.

III. Introduction

In beef cattle production increasing the number of calves born alive and reared to market would increase producer profit[180]. Breeding for increased ovulation rate (OR) would improve the

odds to produce multiple calves. In 1996 a commercial New Zealand beef cow who produced three sets of triplets was identified from which a son, Trio, was kept and used in breeding within a twinning selection study at AgResearch (Hamilton, New Zealand). Trio's daughters exhibited increased frequency of multiple births in a proportion (30%)[55]. This was the first indication of Mendelian inheritance of a genetic variant within this family. Imported semen from Trio was subsequently used in matings at the University of Wisconsin – Madison to establish a herd of high ovulation individuals. Ovulation rate phenotypes, genotyping (Bovine 3K SNP chip, Illumina Inc., San Diego, CA) and linkage analysis confirmed Mendelian segregation of a major gene for ovulation rate (Trio allele) and identified a genomic region associated with high ovulation[54]. Initial candidate gene screening did not yield a causative mutation within candidate gene coding sequences or proximal 5' and 3' gene regions, but a haplotype based on three polymorphisms within the candidate gene region was identified which could be used in identifying individuals as carriers and non-carriers of the Trio allele.

Subsequent work has determined that one of the positional candidate genes, *SMAD6*, is approximately nine-fold overexpressed in granulosa cells (GC) from Trio allele carrier females, while other positional candidate genes are not differentially expressed[123]. *SMAD6* is an inhibitor of the transforming growth factor-beta, bone morphogenetic signaling pathway, the same signaling pathway for which receptor[90,287,288] and ligand[94,95] mutations which cause high ovulation rate and litter size in sheep have been identified. Additionally, work was conducted looking at the physiological and hormonal difference between Trio carriers and non-carriers[124–126]. In these studies, carriers were found to have similar follicular waves, concentrations of progesterone, estradiol, and total volume of dominant follicle and luteal tissues as non-carrier females. The key difference was the size of the follicles. Trio carrier follicles grew at a slower

rate and reached deviation at roughly 1/3 the size of non-carrier controls by spherical volume. These works hypothesized the mechanism of higher ovulation rate in Trio allele carriers is influenced by follicle stimulating hormone (FSH) remaining at high enough levels to prevent atresia until multiple smaller follicles (~3-4) attain dominance. Lastly, GC of carriers showed overexpression of *SMAD6* regardless of follicle growth stage compared with non-carriers[124].

Since the original candidate gene screening, the cost, availability, and quality of sequencing methods have improved. These methods include the next generation sequencing of paired-end short reads and third generation long reads. Using a homozygous individual and the longer sequencing methods would potentially generate sequence information spanning the entire region of interest. Long-read sequencing particularly, Oxford Nanopore, with average lengths of 10-20 Kb with reports of 2.3 megabase read length capabilities[143,148] and thus has the ability to potentially generate sequence contigs spanning the whole region of interest. Additionally, assuming this is a rare mutation, screening the identified variants against 1000 Bull Genomes project variant data to identify unique variants would further narrow the number of variants of interest. Thus, the goal of this work was to identify genetic variants potentially associated with high ovulation in the Trio family assessing their frequency and association with the Trio allele genotype. Based on multiple ovulations leading to potential for multiple births there is strong negative selection pressure from involuntary (abortion/death offspring and/or dam) and voluntary (producer selection) culling of cows that frequently bear multiple offspring with negative consequences. Thus, we hypothesize that the Trio allele is a rare variant, unique to the Treble/Trio family[55].

IV. Materials and methods

Breed background of Trio allele

Identification of the Trio allele's breed background involved determining the haplotype of the positional candidate region for the Trio allele and individuals in the Bovine HapMap dataset[289] with the same haplotype. Trio, two dams, and seven offspring, previously genotyped with the BovineHD chip (Illumina, San Diego, CA), provided information to determine the haplotype across the 1.2 Mb positional candidate region. The Bovine HapMap genotype data[289] containing 1,543 animals representing 134 different breeds was used for comparison. Both datasets were narrowed down to BTA10 and the region between 13,629,354 to 14,817,47 bp, the region containing the Trio allele based on previous analyses[54]. Genotypes from the two animal groups were merged using PLINK 1.9[222]. Missing genotypes were set to 0 using the sed command in Unix. A python script was implemented to convert plink ped format to the fastPhase programs format. Construction of haplotypes was done using the fastPhase program with default parameters[290]. Another python script was then used to convert the output of fastPhase into a three columned file with each line corresponding to a different individual and column two and three corresponding to the different haplotypes. Alternative haplotypes in Trio were directly deduced by separate analysis of Trio, two mates, and six offspring with assignment of the haplotypes as either Trio allele or non-Trio allele-associated based on the ovulation rate record of the offspring. Bovine HapMap individuals were screened for the Trio allele-associated haplotype those possessing that haplotype were identified as to breed.

Generation of Trio allele homozygotes

Potentially homozygous embryos for the Trio allele were produced in July, 2014 by breeding super-ovulated Trio allele carrier females to a carrier bull. Embryos were collected by flushing and transferred singly to recipient females, with calves born in Spring 2015. Presumed Trio allele genotype was determined using the three-variant haplotyping as previously

described[54]. Homozygosity across the positional candidate region was confirmed by genotyping individuals with the BovineHD SNP chip (Illumina, San Diego, CA).

Sequence data

Whole genome sequencing was obtained for two individuals (C041 and C069) homozygous for the Trio allele. Illumina paired-end short reads at 50x coverage were produced for C041, a bull, by Beijing Genomics Institute (BGI), while Oxford Nanopore long reads were generated for C069, a cow, at 14.8x coverage by the University of Wisconsin-Madison Biotechnology Center. DNA from C041 was initially used as template for long read sequencing using the Pacific Biosystems platform. However, sequence quality and read length were insufficient for creation of a quality *de novo* sequence assembly. Being unable to obtain additional usable DNA from C041 (animal deceased and previously collected semen not suitable for extraction of high molecular weight DNA for long reads), DNA was obtained from another individual homozygous for the Trio allele which was still available, C069.

Illumina short reads were aligned to bovine reference ARS-UCD 1.2 following the 1000 Bulls Genomes Project[220,273] guidelines. These reads were also used as needed to polish the long-read data within and around the positional candidate region. Oxford Nanopore Technology (ONT) long reads were used to generate a *de novo* assembly conducted by University of Wisconsin-Madison Biotechnology Center Bioinformatics resource center. Briefly, Flye[291] and Canu[292] were each used to generate a *de novo* assembly for C069 which were compared to each other using Quast[293] and the reference: ARS-UCD 1.2[131,294]. Additionally, 3x iterative contig polishing was performed with Racon[295] with Illumina data from C041 on the C069 contigs (contig_553, contig_1505) representing the 1.2 Mb region of interest.

Variant detection

Both homozygous animals had the following types of variants detected: single nucleotide polymorphisms (SNPs), small insertions and deletions (InDels), and large (> 50 bp) structural variants (SV). The software from Genome Analysis Toolkit (GATK) was used to identify SNPs and InDels from both, while various methods were used for SV detection. Based on Lett et al. 2022 (Lett, 2022 Thesis chapter 3), LUMPY single sample method[296] was used to detect structural variants in C041 and was implemented using the docker *smoove* as recommended by LUMPY authors[277]. Structural variant detection in the long-read data was done using SVIM-asm v1.0.2[297,298]. Results were then compared between the two to identify those that were in common between them. Deletions were visually inspected using Golden Helix GenomeBrowse[285] and Integrative Genomic Viewer (IGV)[299] for copy number. The variants were further screened to identify those SV that overlapped with the 1.2 Mb region on BTA10 (13,629,354-14,817,470 bp).

Small variant detection with GATK v 4.2.3.0 followed both the GATK best practices[300] and a variation for non-model organisms[301]. The short read used the GVCF file created by following the 1000 Bull Genomes project guidelines[220] and the long read data used the 1164_1165 bam file provided by the UW-Madison Biotech Center after narrowing to BTA10 and further polishing with the short reads. A pre-step to create the GVCF file was done on the long-read file using GATK v4.2.3.0 HaplotypeCaller. The remainder of the steps were run on both individuals. Initially, GATK v4.2.3.0 GenotypeGVCFs was run with BTA10 being selected for the short read data. These were then split into SNP and InDels files using GATK v4.2.3.0 SelectVariants and converted into tables with GATK v4.2.3.0 VariantsToTable. An Rscript was run based on the non-model documentation to generate distributions to select the values to use in

filtering. Filtering was conducted using GATK v4.2.3.0 VariantFilteration and based on results from the R plots. Lastly, data was converted from VCF to text and specific information was extracted (BTA, position, type, counts of the different genotypes - heterozygous, homozygous reference, and homozygous variant, the reference and alternative, and the genotype of individual) using GATK v4.2.3.0 VariantsToTable. These variants were then narrowed to those that were homozygous for the alternative allele and were in both C041 and C069. Further, the variants were narrowed to those found within the 1.2 Mb region BTA10:13,629,354-14,817,470 bp and compared with 3,093 bulls from the 1,000 Bull Genome Project run 7.

Variant genotyping

A variant located within the 1.2 Mb positional candidate region of potential interest was genotyped using a PCR-RFLP assay. The primer pairs (Forward = 5'-TGTTCATCATGGGCTTGTCAT-3' and Reverse = 5'-ACACCCAAACCAGACAAAGAC-3') were designed to span a segment of 390 bp centered on the base change. The variant in question altered the restriction site for restriction enzyme, *Bsa*BI. PCR was performed with a touchdown protocol in which annealing temperature was reduced by 0.5°C starting from 63°C for the first 10 PCR cycles following which an annealing temperature of 58°C was used. Restriction digestion of the PCR product used 5 µl of PCR product and 5 units of *Bsa*BI (New England Biolabs, Ipswitch, MA) in a 20 µl reaction with incubation at 60°C for 6 h followed by heat inactivation at 80°C for 20 min. Visualization of products was by gel electrophoresis and ethidium bromide staining using 1.5% agarose gels run for 350 volt-hours at 125 volts. Primer designs were made using NCBI Primer-BLAST and unique products were checked using NCBI ePrimer in Primer-BLAST[280].

Genotyping was conducted to determine concordance of variant genotype with inferred Trio allele genotype and to determine allele frequency in two populations most likely to possess the Trio allele. Trio descendants generated at the University of Wisconsin-Madison were used to determine concordance of variant and inferred Trio allele genotypes. All were females from whom ovulation rate information had been collected over four estrous cycles. An average ovulation rate ≥ 2.5 ova per cycle inferred carrier genotype for the Trio allele, as they had triple ovulations or greater on at least two of four estrous cycles. Animals with single ovulations in all four estrous cycles were inferred to be homozygous normal. Animals with phenotypes intermediate between these thresholds were considered indeterminant and excluded. Genotyping was conducted using Trio descendants (n=85), random Hereford samples (n=98), and a sample of cattle from the USDA Meat Animal Research Center twinning population (MARC Twinner; n=78). Trio descendants' ovulation data was obtained by ultrasound visualization of CL, as described previously[54,125] and DNA was extracted from ear punches. Hereford DNA was extracted from semen samples (n=26) obtained from AI studs or a Hereford breeder between 1992 and 2019 and from hair (n=72) obtained from cattle shown at the 2021 Wisconsin and Iowa State Fairs. Samples were chosen to avoid including close relatives (e.g. half-siblings) and typically no more than three samples were obtained from an individual breeder. The MARC Twinner DNA was obtained from ear punches or semen. Extractions were conducted using a proteolytic digestion and organic extraction protocol[203].

V. Results

Breed background of Trio allele

A total of 28 SNPs in common between the HD SNP data for the Trio family and the 50K SNP data for the HapMap samples spanned the 1.2 Mb positional candidate region of interest.

Alternative haplotypes for Trio were deduced from the family genotype data and the haplotype associated with the Trio high ovulation rate allele identified (Trio allele; Table 1). After identification of the Trio allele haplotype, it was compared with the 1,543 Bovine HapMap individuals. Twenty-seven individuals were observed to have the haplotype associated with the Trio allele based on the Trio family analysis (Table 2). Within those 27 individuals a total of nine different breeds were represented with the greatest proportion (9/27) of the individuals being of Hereford origin. These nine individuals with the Trio allele made up 45% of the Herefords in the data and had a haplotype frequency of 0.3 for the Trio allele haplotype. Beefmaster, Normande, and Charolais each had four individuals possessing the Trio allele haplotype, representing 20% of the animals for their respective breeds. The remaining five breeds included Finnish Ayrshire with two (11% of the individuals) and Belgian Blue (25% of the individuals), White Park (20% of the individuals), Beefalo (100%), and Maine-Anjou (5%) with one individual each (Table 2).

Generation of Trio allele homozygotes

In the spring of 2015, 26 calves from carrier x carrier mating and embryo transfer were born. Of 24 which were genotyped to determine Trio allele inheritance, nine were non-carriers, ten were carriers and five were Trio allele homozygotes (not significantly different from expectations under the assumption of Mendelian inheritance, $p>0.05$). Homozygosity across the positional candidate region was confirmed using high density SNP genotyping for all five predicted Trio allele homozygotes (Figure 5.1).

Assembly

Alignment of the Illumina short reads (C041) to the reference genome (ARS-UCD1.2) yielded depth of coverage averages of 48.07 with 81.9% of the bases above 40x coverage. The

de novo long read assemblies of C069 totaled 2.76 Mbp in 9563 contigs using Canu and 2.63 Mbp present in 1455 contigs using Flye. The Flye assembly N50 outperformed Canu with N50 of 27.6 Mb to just 0.5 Mb in Canu. Additionally, the longest contig assembled by Canu was only 3.18 Mb as compared to Flye which generated a maximum contig length of 119 Mb. The Flye assembly was clearly more contiguous, so we pursued all downstream analyses with it.

Given our interest in the genetic variation present at the Trio allele, we polished the nearby contigs (contig_553, contig_1505) using Illumina data from C041 to remove errors introduced by assembling with Oxford Nanopore reads. To understand the variation relative to a known reference we realigned those two polished contigs using Minimap2 with relaxed conditions (-ax asm20) to ARS-UCD 1.2 given the inherent inaccuracy of ONT reads. Variant detection was conducted on this alignment and C041 alignment.

Variant detection

Variants were detected in each sample individually before being compared between samples. Alignment of the *de novo* C069 assembly with ARS-UCD 1.2 detected 18,427 SV events by SVIM-asm with 816 of them being on BTA10. In C041, LUMPY detected 14,489 SV with 558 being on BTA10. Of the 558 SV, 125 genotyped as homozygous alternatives to the reference ARS-UCD1.2. Comparison between C041 and C069 showed 18 SV that overlapped with each other nearby but not within the 1.2 Mb region of interest (Table 3).

Detection of smaller variants indicated 269,360 SNPs and 36,088 InDels in C041 before quality filtering on BTA10 and 205,470 SNPs and 35,145 InDels after. From these, homozygous alternative genotypes were selected with 93,155 SNPs and 15,387 InDels remaining. InDels split into 7,222 deletions and 8,165 insertions. Less variants were detected in C069, with a total of

6,030 SNPs and 1,600 InDels prior to quality filtering on BTA10. After filtering there were 3,247 SNPs of which 2,611 were genotyped as homozygous alternative and 1,259 InDels remained after filtering with 948 having homozygous alternative genotypes. These split into 797 deletions, and 151 insertions. Narrowing to the 1.2 Mb region showed that C041 had 739 SNPs and 149 InDels while C069 had 214 SNPs and 113 InDels (Table 3). Comparison between these variants and those obtained from the *Bos taurus* run 7 of the 1000 Bull Genomes project involved looking at BTA10 in 3,093 animals. This left 11 SNPs and 20 InDels for C041 and 8 SNPs and 91 InDels for C069. When compared between each other only one SNP remained as novel and no InDels overlapped (Table 3).

The novel SNP is located at 13828552 bp on BTA10 and is a change from A to G. Running Ensembl variant effect prediction (VEP) on this variant indicated that it had an impact listed as modifier and consequence of intergenic variant. Selecting ± 100 bp of sequence surrounding this variant and running BLAST showed 27 alignments with > 50% coverage overlap. Percent identify ranged from 75.26 to 98.51 with a median of 78.24. The highest identify alignment was to *Bos mutus.* Looking at the variant location within ARS-UCD1.2 using the UCSC Genome Browser, the variant location is in a GeneScan predicted gene (chr10:13797154-13889348 bp) and within a LINE element L1M3.

Variant genotyping

The PCR amplicon had a length of 390 bp with the variable base located at base 210 from the 5' end of the forward primer. Presence of A at the variable base creates a *Bsa*BI restriction site, yielding fragments of 207 and 183 bp upon digestion, versus 390 bp for the uncut G allele. Successfully genotyped Trio descendants showed perfect concordance between SNP genotype

and inferred Trio allele genotype (Table 4). In the other two populations, all individuals were homozygous for the A allele corresponding to the non-Trio allele.

VI. Discussion

Breed background of Trio allele

Comparison of Trio's genotypes with daughters and two mothers of daughters allowed for deduction of alternative Trio haplotypes and identification of the haplotype associated with high OR. Frequency of this haplotype in the HapMap data was low, occurring in 27 of 1,543 individuals. Three of twenty individuals of Hereford origin were homozygous and the remaining 25 individuals were heterozygous for the haplotype giving a frequency estimate of 0.30. Considering breeds with more than one sample, the breed with most individuals possessing the haplotype and with highest frequency was Hereford. One of the next highest breeds for haplotype frequency was Beefmaster which is a composite breed including Hereford, Shorthorn and Brahman ancestry. Given the relatively high haplotype frequency in Hereford and a Hereford composite breed and that Treble, the source of the allele and matriarch of the family, was born in New Zealand where Hereford is a common beef breed, we hypothesized that the Trio allele mutation originated in a Hereford haplotype.

Variant detection and genotyping

Only one novel SNP was found in both homozygous individuals and the 1.2 Mb positional region of interest (Figure 5.2). Results from the PCR-RFLP genotyping of this SNP showed perfect concordance between this variant and the Trio allele-inferred genotypes. All individuals with genotype AG were Trio descendants with inferred genotype of carrier, whereas all individuals with AA genotypes had inferred genotypes of non-carrier. When tested in two

populations most likely to possess the allele either due to breed origin (Hereford) or ovulation rate phenotype (MARC Twinner), the Trio allele (G) was not observed. This pattern is consistent with a putative causative variant that is unique to the Trio family, but it is not proof of causation. Predictions of the variant's effect were listed as modifier for which Ensembl defines as either non-coding, affecting non-coding genes or where prediction is difficult or the is no evidence of impact. Given the Trio allele's effect on *SMAD6* expression, causation could potentially be tested *in vitro* by editing the mutation into normal granulosa cells and assessing *SMAD6* expression in the edited cells.

Looking at the structural variants only the long-read data showed variants within the 1.2 Mb region. This may in part be due to the use of the FASTA alignment to detect variants and it illustrates the increased capability to detect structural variants when using long reads. This was also indicated by the long reads initially detecting greater number of SV events on BTA10 from the outset. In contrast, for smaller variant detection the short read alignment yielded more variants. But neither yielded high numbers of novel SNPs (1.4% short reads and 3.7% long reads), while novel InDels ranged from 13.4% for the short read alignment to 80.5% for the long read. The discrepancy maybe due to location difference not being the same between short read and long reads. The 1000 Bull Genomes Project data is based on short read data (mostly) and it could be the predicted variant location is incorrect, thus a greater number of variants are considered novel in the analysis. The fact that none overlap between the two and remain novel is of interest, but this again maybe a factor of discrepancies in location of the call between a region that may have a read spanning the variant fully and one broken between multiple reads.

VII. Conclusions

Based on the haplotyping results the breed background for the Trio allele is most likely Hereford in origin. Comparison of two Trio descendants homozygous for the Trio allele revealed numerous variant overlaps on BTA10 however only 15 InDels and 174 SNPs corresponded within the 1.2 Mb positional region as being homozygous in both C041 and C069. Comparison of the 189 variants to the 1000 Bull genome project data showed only one SNP (10:g.13828552A>G) as novel. Genotyping supported this variant being a strong candidate as all non-carrier Trio descendants and individuals unrelated to Trio were homozygous for the non-Trio allele. Future work to test if and how this variant impacts expression of *SMAD6*, by editing the variant into normal granulosa cells and assessing gene expression, is needed. Additionally, exploration of the difference between InDel size may indicate if the difference in numbers is caused by location difference between read lengths or differences in assembly/alignment quality.

CHAPTER 6

Communicating science to a non-scientific audience

I. Preface

I wrote this chapter because the field of animal and dairy science relies heavily on the committee of farmers, consumers, advocates, and researchers. There is a larger portion that do not have a deep background in the technical side of research or specialization in a topic that we, scientists, tend to forget. Effective communication of science to the stockholders and beneficiaries of our research, has always been at the forefront of my mind while conducting my research. I am thankful to Wisconsin Initiative for Science Literacy (WISL) at UW Madison for providing the opportunity and support to create a chapter designed to help communicate science. Gratitude and many thanks to Professor Bassam Shakhashiri, Elizabeth Reynolds, and Cayce Osborne for helping provide feedback, support, and this opportunity.

II. Introduction

When a lot of people think of genetics, animals, and computers it conjures up thoughts of the movie Jurassic Park. Thankfully, my work does not involve prehistoric creatures that will most likely eat me, but it does involve a lot of computer work, some DNA, and trying to improve cows' reproductive health.

Like humans, cows have roughly a nine-month pregnancy, monthly cycles, and complications that arise at various stages of these processes. Unlike humans, a cow cannot speak of issues during pregnancy, and so noticing and addressing problems that arise falls to their owner and care-givers. Part of what my work focuses on is helping improve the health of the cow through genetics to limit early pregnancy loss or pregnancy loss due to overcrowding before these issues arise.

Each of my studies looks at different events (twinning, embryo lethality, and ovulation rate) in cattle reproduction. Two of them looked to help improve negative pregnancy outcomes and one looked to explore a deeper understanding of mechanisms driving ovulation rate and its links to reproduction. This work includes looking at limiting twins, which are hard on cows, identifying sources of early term abortions, and identifying a potential cause of increased ovulation rate.

III. When too many is a bad thing

The saying "less is more" rings true in cattle when it comes to multiple births. And in my time working and talking with dairy producers this comes up as a problem again, and again. This is because these pregnancies are more demanding both on the cow and the farmer. These super moms produce several pounds of milk a day that takes a toll on the body. During pregnancy, the

toll of carrying multiple babies is heavy, just like it is in pregnant women, and it causes risks to both the mom and the unborn children, which sometimes ends in tragedy.

This makes multiple births an undesirable trait for farmers, and they look to find ways to prevent this from happening. One way is usually removing that cow and her offspring from the herd because multiple birth pregnancies is a genetic trait. Other ways are more invasive and usually lead to termination of the pregnancy. Because of this, I was very eager to start my research career as a master's student looking at ways to solve this problem.

Since we know the incidence of twins is linked to family lines, one noninvasive way to limit multiple births is through genetic selection. Now this sounds easy enough but to effectively implement selection we need to know what DNA changes influence twinning. We also need to know the current values of heritability, the probability of a trait being passed from parent to offspring, and repeatability, the probability of a trait/event happening again, of twinning in a group of cows. And the best source of this information would be from producer records. We obtained calving records from 2010 – 2016 from a dairy record management group called AgSource CRI.

Using producer records however comes with a frustrating challenge. They are noisy and not what some researchers would call "clean". This is because of user errors. Unlike studies that generate their own data and have rigorous guidelines, this data source relies heavily on what the producer records and tends to include more user errors, specifically missing or misentered information.

I first cleaned the records by removing any with missing ID information of the cow, her dad (sire), and her mom (dam). Then I had to correct and remove any date issues. The most interesting date issue I encountered was the cow that was born after she gave birth. And lastly, I

sorted through to match and remove duplicate records that carried over from an old herd to a new herd.

This task was the most time consuming and really what most computer work ends up being. The challenge I faced was the sheer number of records. Initially, I had to sort through 2.9 million records, and programs like Excel cannot perform corrections on that magnitude of data let alone open it. But by writing a few of my own computer scripts, I was able to clean my records and restrict them to ensure that each herd and sire had at least 100 records.

I used the cleaned data, about 1.4 million records, in a heritability and repeatability analysis. I estimated these values using a software, AIREML, that was designed to perform such genetic analysis, and implemented using a model equation. The results showed a low heritability and repeatability. Most reproductive traits show low heritability, and we did not expect the value to be too high. These estimates were in range of previous ones and indicated no drastic increase overtime to twinning in this cow population.

Even though these numbers were low, they were still greater than zero. This told us there is a potential to generate selection tools for producers using genetics. The next step would be to incorporate genotype data in the form of SNPs (single DNA nucleotide changes). An individual's genotype is their genetic make-up. For this study, specific SNPs across the genome are genotyped using a chip panel that has known information (genotyping is the process of detecting genetic differences in individuals). These small changes from one DNA nucleotide to another can cause large changes to how the DNA sequence is read and interpreted into a phenotype, or observable characteristics. Because genotypes are more regularly generated on male cattle, we decided to convert the data into sire-daughter averages and obtain genotypes from the repository of dairy cattle genotypes.

I used another model to estimate values of the factors that can influence twinning. Then I corrected the individual calving records, averaged them per cow, and finally averaged that per sire. These values would serve as the phenotype (or observable characteristic) in the next analysis – genome wide association study.

Genome wide association studies, or GWAS, are a widely used method of looking for association between a phenotype and genotypes. They help to identify regions of interest in the DNA that influence a particular trait. And with the improvement in genetic technologies, we can even locate single DNA base changes for researchers to investigate further.

I initially ran my GWAS using a program called GenABLE. In addition to using the records from 2010-2016, previous estimates were available from previous studies done in my lab group and in collaboration with another group at Iowa State University. Using these time points from 1994-1998 and 1999-2008 as well, I was able to compare across the different datasets to identify genetic regions that showed association with twinning in all three.

My pilot study showed that chromosome 11 in all three datasets had a peak comprised of genotypes that were strongly associated with the twinning phenotype. Unlike humans, cattle have 29 autosomal chromosomes (humans have 22) but have the same number of sex chromosomes (X and Y) . Additionally, two of the three datasets shared the exact same genomic region of interest. What made this region even more exciting was the presence of two genes involved in the female reproductive cycle.

Unfortunately, before I could look deeper, my research took an unexpected health break. This did not stop my interest in the subject, but rather changed my path from a Master's to a PhD. When my path changed it opened new doors on this project. Door one was more calving records

from another record management system. Secondly, another option for conducting GWAS that would allow me to increase the number of bulls I could use in the analysis and test all the SNPs at once rather than individually. These small base changes can either cause large effects on their own or in combination with each other, so by looking at only one at a time you lose information on how they interact together. And lastly, a new and more correct reference genome for cattle and access to the 1000 bulls project data was made available.

In genetic studies, the reference genome is the gold standard from which other individuals of that species are compared. Because genetic technologies keep improving, the quality of the references improves too. In terms of GWAS this means improved knowledge of SNP locations, improved quality of the genotype types from SNP chips, and increased amount of SNPs we can detect.

My initial GWAS only had ~60,000 genotypes. This means across the genome (~3 billion bases) we have detected only 60,000 different base locations. We did impute, that is, utilize information from higher density genotype animals to infer the missing genotypes in the lower density genotyped animals, up to ~ 600,000. This still leaves a lot of missing information and spaces untested. By the lab being part of the 1000 bulls project, we had access to whole genome level SNP data. This data was used to impute the 600,000-genotype data to just under 8 million.

I cleaned the new calving records like I did the original and merged the two newest calving record sources. Because I already had code written, this process was faster, however there were a few more challenges to correct. But by combining these datasets, the total number of records increased to ~4.4 million. Further, I improved the herd and sire restrictions to include only those with at least 100 unique cows per herd and at least 100 unique daughters per sire. This would improve both the quality of the herd effects prediction and the estimates per sire.

Previously, when I ran the GWAS, I could only use sires with both known phenotypes and genotypes. This cut the number of bulls in the study down by about half for 2010-2016 calvings and by ¾ in the older datasets. By using the new program single-step GWAS we could use all the individuals with estimated phenotypes in the association study. We also combined the older two datasets into one since there was greater overlap between bulls and the phenotype estimates.

I conducted the single-step GWAS for each time point, sire phenotype estimates based on 1994 – 2008 calvings and 2010-2016, separately and then combined the data together. Individually, the datasets did not show any significant results but the older (and largest dataset) did show a tendency for association between the sire daughter averages for twinning with chromosome 11 still. When the results were combined the strongest association was the same as the pilot study! And it still included the previous two genes of interest.

Because I switched to the PhD I was able to also use the pilot data in a gene set analysis, which is an analysis to see what genes or gene pathways (set of genes that work together to turn on or off different biological functions) are more involved with your data. I implemented this analysis on the new results from GWAS. It did not yield the results I had hoped for. There was no pathway associated with the twinning rate phenotype that contained the two genes of interest, but there was one pathway that was ranked high in both analyses. When I looked at that pathway, I found that it included only two genes that were also found in the region implicated in GWAS. There are two possibilities: these genes are indeed of interest and they influence the trait, or they are just near the genes of interest. This will be for someone else to test and decide.

My last task for this work was to look at genomic prediction, the ability to use current information to predict breeding values in another individual. Genomic prediction is widely used by producers now and heavily influences the breeding and retention decisions of a farm. By

generating this prediction we can start to provide producers with the knowledge to select for or against twinning in their herds. For studies on prediction, you first need to build a model and train it, and then you need to test the accuracy of the prediction. This is like how google/Pandora "learns" your preference on a specific news or song choices. It makes future decisions based on your previous choices. In this case we use the genotype records to train the model and predict breeding values that can be tested against our phenotypes.

To accomplish this, I split the data into a testing and training set, which was easy for this study as the newer and older records made for a good split. The older data would serve as a training set to generate the values used to predict the values in a new set of animals. The new set would be a portion of the newer data. I removed all bulls that appeared in both data sets, so the testing set was completely new sires compared with the training.

The reliability of my model in prediction was about 42% which may not seem high but is the highest thus far in the literature. This means there is the possibility to utilize this for selection purposes. It also tells me there is room for improvement.

All research is constrained by time, money, and resources available, which puts limits on what can be done. The main limitation I faced was in numbers. Genetics studies, particularly GWAS, are dependent on the number of samples used. As mentioned, I only had genotype information for portions of my data. The other part is that in an ideal situation combining the raw calving records across all the time periods would have been done. However, we were unable to obtain access to the older calving data original records.

In the end of this project, there are still more answers to be sought and questions to be asked. But I can say I have indicated a region of interest that is strongly associated with twinning and two

genes of interest. I also found a pathway containing two additional genes that may be connected to the trait. And I estimated a genomic prediction reliability greater than zero, and generated information that can be provided to producers for selection purposes.

IV. Loss hurts, unknown loss hurts and confuses

Loss of anything is upsetting and frustrating. For farmers, a cow losing a pregnancy is both an emotional and financial burden. In addition to affecting the owner it also impacts the cow, both in her short-term and long-term production.

Pregnancy can be thought of like a chain of events, each connected to another, and breaking any link will impact the results. The earlier in the chain an event occurs, the harder it is to identify the causes. To better understand early losses, a source to study is the DNA and how those building blocks are set up.

As you may know from Jurassic Park and Mr. DNA, DNA is the building block of all things. Breaks or changes in the code can disrupt its function causing cell death, changes in expression of genes and thus phenotypes, or no effects at all. There are multiple sources of genetic variations that affect DNA, ranging from changes to small single bases (SNPs) to large sequence rearrangements (structural variants). A class of structural variations are CNVs, or copy number variants.

Now as the name would imply, CNVs are changes in copy number. By copy number I am referring to the number of times a base or sequence of bases are inherited. In normal diploid individuals, like humans and cows, two copies of a chromosome are inherited. One copy comes from mom and the other comes from dad. A chromosome can be thought of as being made up of several blocks of DNA or chunks. In a normal perfect case these blocks would be continuous

strings of DNA sequence and you would not see blocks. But with a copy number variation you may be going along the string and find a block missing or an extra one added on (Figure 6.1).

While duplications (extra copies) maybe impactful, they are harder to detect in the type of DNA sequences I had available, so our focus was looking at losses (deletions). Losses also carry potential to be destructive as they may remove a DNA block needed for a gene to function. The key first step would be to detect CNVs in a group of samples.

In this case we had 25 samples with pair-end short read sequence data. Short reads are fragmented pieces of an individual's genome about 200-500 bases in length, depending on the method used to obtain them. Being pair-end refers to both ends of the fragment being sequenced rather than just one. These reads can be pieced together using an assembler program or by matching the sequence reads to an already assembled reference submitted and maintained in a database location such as National Center for Biotechnology Information (NCBI).

Jerseys are the second most popular breed of dairy cattle in the dairy industry. Even though they are second in total numbers few studies have been done in just them. Thus, we wanted to focus our work to identify CNVs in Jersey cattle, and then screen these samples to provide Jersey producers with deleterious variants to avoid in breeding programs. Within our 25 sequenced bulls, 20 were purebred Jerseys and one was a mixed breed including Jersey.

To detect CNVs we utilize the information provided by the sequencing fragment (reads) focusing on two concepts. Reads can be thought of like puzzle pieces that fit together in a specific way and the final picture is a genome. The starting image we work from is the reference and we compare the pieces to it. Sometimes they do not match, and this one source is used to detect structural variants. The catch with this puzzle is that it is a 3D puzzle whose height is dependent

on the depth at which the sequences are generated (i.e. 10x coverage means on average you would expect the height of the 3D puzzle to be 10 pieces). This lends to the second source used to identify CNVs, changes in the depth compared to what is expected (Figure 6.2). In most cases the depth is good at finding deletions but has problems with duplications. Using the piece orientation and how it matches to the main piece orientation and matching to the main image, the software has a better chance at finding the exact start and stop of CNVs. For my detection, I used several detection software to predict potential CNVs within a sample. I then used another program to merge the different CNVs from each method to generate a consensus. This method found CNVs in a similar location, size, and type, for each sample. I used the same software to merge the sample consensus CNVs into different groups Jersey, non-Jersey, and both.

Our next step was to find deletion with embryo lethal potential, meaning they could be contributing to early pregnancy loss under the assumption that within the population there is an absence of homozygotes (individuals with two deletion copies). This goes back to the block inheritance concept. Normally you would inherit two copies of the block (NN). In the case of a deletion one parent or both have that block deleted and you would inherit a single copy (ND) or two copies of the deletion (DD). Now not all deletions are harmful and thus are passed from one generation to the next. But some are, and these would not be seen in the population in the DD state due to inability to produce offspring that survive.

So, I took my CNV list and pulled out the deletions and screened them to locate any that did not have DD individuals. I further found additional open-source Jersey sequence data and genotyped my deletions of interest in these animals as well. This helped narrow the list and I designed primers, small sequences that are paired as a forward (start) and reverse (end), that target a

specific target location in the genome. In the end, I was only able to visually match the predicted genotypes and design primers for four from a list of 32.

In designing a three-primer system, we could genotype the deletions in multiple samples with the possible results in image 2. Unlike the image we expect to see only the NN and ND outcomes if the deletion truly has embryo lethal potential. We performed genotyping using 96 Jersey cows that had DNA extracted from a previous study for the four deletions indicated earlier. Like SNP genotyping a PCR assay allows replication of DNA at a specified location and can reveal if an individual has one, two, or no copies of a deletion. The resulting PCR assay image showed DD individuals within the 96 cows for all four deletions tested (Figure 6.3. This disproved the idea that these deletions have embryo lethal potential and are just rarely seen. There are, however, 33 other deletions indicated as being absent DD individuals that maybe of interest for future screening. An initial first step should be improving breakpoint detection (CNV start and stop locations).

While the four deletions I tested did not show embryo lethality, I did however find 468 CNVs that were not previously in the variant database. I also, through using multiple tools, figured out which I would consider ideal for CNV detection and would use in future studies. Like the twinning study, a limiting factor was number of samples. The other limiting factor for this study is the data source. Short read sequences, while cost effective, are prone to alignment mistakes when trying to compare the sample pieces to a reference map. This is particularly challenging given the fact that genomes tend to have a lot of short, repeated pieces (Figure 6.2B). This leads to duplications being hard to differentiate. The new read technology, long reads, allows for generating sequences that span 10,000 – 20,000 bases on average compared to the short reads

(max ~500 bp). I look forward to long reads reaching the affordability of short reads or the day a long read can sequence an entire human/animal chromosome in one entire strand.

V. Understanding a phenomenon

Like humans, multiple sets of twins in cattle are amazing. But multiple sets of triplets is phenomenal. A New Zealand cow named Treble did just that, had not one, not two, but three sets of triplets. It was at this point that she drew the interest of her owner and two researchers invested in studying multiple births in cattle. These researchers, along with Brian Kirkpatrick, used Treble's son Trio to produce granddaughters.

When they looked at the granddaughter calving records, they found that 30% had incidences of multiple birth pregnancies. This was a phenomenon and raised the question of whether there is a genetic component to it. Brian Kirkpatrick imported semen from Trio to generate a herd of cows. To identify a more quantitative measure for the phenotype, Dr. Kirkpatrick used ultrasound to count corpus luteum (CL), a structure that forms on the ovary after an oocyte (egg) ovulates.

In cattle, typically only one egg ovulates at a time producing one CL. When multiple eggs ovulate, there are more CLs present and these structures can be counted using the ultrasound images. In the 2015 study, Dr. Kirkpatrick and Dr. Morris found a portion of the Trio daughters had > 3 eggs per cycle compared to normal. This created a measurable phenotype to split the herd into – high ovulation rate and normal.

The next step involved generating SNP genotype information on the daughters and Trio to help decide if there was a genetically inherited mutation causing this phenomenon. By using two groups, high and normal/low ovulation rate (number of eggs released at a time), they could compare the genotypes between the groups. This allowed them to locate a region on

chromosome 10 that is positionally of interest, meaning they located a segment of change between the high and low groups and proved there was indeed a genetic component. Due to the sparsity of the SNP data at the time, they could only narrow the region to 1.2 mega bases (1,200,000 bases) but were able to use this information to create a genotype assay to distinguish between the carriers and non-carriers.

They called the genetic trait the Trio allele. An allele refers to the genetic change that has alternative forms when inherited. In this case a cow may inherit two normal alleles (normal ovulation rate), one normal allele and one Trio allele (high ovulation rate), or two Trio alleles (same as single copy). However, the exact causative mutation and mechanisms remained unknown.

Work done by two previous graduate students, Mamat Kamalludin and Alvaro Garcia-Guerra, provided more insight into this phenomenon. In Kamalludin's work, he showed that the gene *SMAD6* was overexpressed (the protein encoded by this gene is seen in the cells more often than normal) in carriers compared to non-carrier individuals. The positional candidate region, the 1.2 mega-base (Mb) region previously identified, is located near the DNA block encoding this gene, making it the most likely gene of interest. The work Garcia-Guerra performed looked more into the physiology difference between Trio allele carriers and non-carriers. His work showed that the carriers ovulated more eggs at a smaller size even though the timing and hormonal profiles were similar to noncarrier siblings.

Now since the initial genetic screen, genetic sequencing technologies have improved, and the cost has decreased for specific types. This made it possible to produce both short read sequencing and long read sequencing. But the first task was animal selection. For this, having individuals homozygous, those who inherited two copies of the Trio allele, would be

advantageous. The logic is that by sequencing a homozygous individual rather than a heterozygous one (inherited only one copy), you rule out variants not found in that state and simplify the comparison process.

In the end our lab sequenced one cow, C069, and a bull, C041. I performed sequencing alignment like with my CNV project on C041, since his sequencing was done using pair-end short reads. Initial attempts to assemble C041's long read sequencing (average lengths are 10,000 bp rather than 350) failed. I lacked the computing power to perform this task, so we outsourced to the University of Wisconsin Madison Biotech Center Bioinformatics department.

They found that the quality of the original long read sequencing was not adequate to generate a good quality assembly. In comes C069, since DNA was no longer available from C041. These two animals are full siblings meaning their DNA should be similar and both should have inherited two copies of the Trio allele. Newer methods of long read technology were used to generate sequencing results on C069. The hope was to get the 1.2 mb region in one continuous read rather than in chunks.

Once we had an assembly of reads for both C041 and C069, I began to detect multiple genetic variant types in them. These include SNPS (single base changes), InDels (small $\leq$ 50 bp base additions or deletions), and structural variations (large > 50 bp genomic rearrangements). Detection of SNPs and InDels was straightforward, as a program, GATK, is designed for such studies. The structural variant detection was slightly more complicated. In the case of C041, detection was easy since I implemented multiple methods previously and selected the method that performed the best. C069 presented more of a challenge, as the methods I used previously were all designed for short reads.

I found a program online that would allow me to both align the assembly to a reference and perform detection at the same time. The next hurdle was removing all variants that were not homozygous and did not match the bovine reference genome. We assumed the reference contained the normal allele. Thankfully, all the software I used performs genotyping, so I only needed to create a script to pull out those variants based on the genotype.

From there I performed a comparison between variants detected in C041 and C069. This narrowed the results from the thousands down to hundreds for SNPs and less for InDels. In the case of structural variants (SV) only five predicted variants were implicated, and none fell in the region of interest. I am curious about one of the SVs and hope that future work will investigate that more. A total of 15 InDels and 174 SNPs fell within the original 1.2 mb window.

The second assumption we made was that the Trio allele mutation would be rare. Previous work granted us access to 1000 Bull genome data, which is a consortium of DNA sequencing and variant calls similar to the 1000 Genomes project in humans. This provided variant, SNP and InDel data, on over 3,000 bulls from various breed backgrounds. I took my list of SNPs and InDels and compared them with this dataset to locate any novel variants. Of those in the 1.2 Mb region, only one SNP had not been previously detected. Digging into the SNP more, I used a prediction software to see if the DNA change caused an impact. The results indicated it caused no fancy or drastic changes to a gene – this is known. It may be that, since it is novel, the variant causes a change for which the impact is unknown.

A goal was to see if this variant had consistency with the high ovulation phenotype and not with two other cattle populations of interest. These included the MARC Twinner herd, a herd bred to increase twinning and ovulation rate in beef cattle, since more calves in beef is a positive trait. The other population is Hereford cattle. This was selected because previous work looking to

identify the breed background showed Hereford as the most likely breed from which this mutation arose.

Dr. Kirkpatrick performed genotyping using a restriction enzyme digest PCR (PCR-RFLP) assay. Like the CNV work, a segment of DNA is targeted and replicated using PCR. In a PCR-RFLP assay, we must use a restriction enzyme (a reaction catalyst whose property cleaves DNA) to cut the fragment at the base change location only if the individual had the reference allele. The restriction enzyme should ignore the sequence if the individual had the Trio allele (Figure 6.4).

How it works is first PCR is run on the samples to test. They would all produce a ~400 bp band when viewed. Then the PCR is added with the enzyme in a reaction to digest the DNA at the restriction site targeted by the enzyme (Figure 6.4). Next these would be run on a gel using an electric current that pulls the negatively charged DNA down the gel towards the positive end. Each sample has its own lane (Figure 6.4). The DNA travels at different speeds depending on the weight or amount of DNA. These bands correspond to different genotypes.

A single band at ~400 would indicate only the variant was present (G in this case). While a single band at ~200 shows only the normal allele is present (A), and two bands at ~400 and the other at ~200 would indicate both the normal and variant is present (A/G). If this is indeed tied to the Trio allele then we would hope to see A/G for our carrier animals, A for our non-carriers, and G for the homozygous Trio allele individuals (Figure 6.4). We would also hope to only see the A band in the other two populations making it unique to the Trio family.

Running the genotypes was performed by Brian Kirkpatrick and the corresponding genotypes were assessed by both of us. Thus far the genotypes are descriptively in accordance with the Trio allele phenotypes. Additionally, the other populations all show the single band for the A allele,

indicating this may be the causative mutation, but the underlying mechanism is still unknown. Questions remain on if and how it affects *SMAD6,* the candidate gene of interest, and further tests will need to be done to figure that out.

While the exact mutation was not identified, there is a strong possibility that the SNP I found is linked in someway to the Trio allele. Limitations in this work have been mostly time, funding, and computing resources. The COVID-19 pandemic really halted progress and forced extensions on a lot of projects. Time lost figuring out the computing resources that were not there to perform assembly in-house, along with slow run times on some of the variant calling also hindered progress.

VI. Wrapping up

While many of my studies have left doors open, they have helped create bridges for future directions. My work with the twinning project provides an updated estimate of twinning heritability and repeatability. It also provided a location for further study on chromosome 11 and two positional candidate genes. This location also has recently been indicated as an area of interest by another group in a different population of cows. The CNV work provided additional information on limitations and knowledge gaps on software for CNV detection. It also increased the amount of data available in the variant archive and sequencing archives, allowing future researchers to utilize this information for their studies. And lastly the Trio allele work generated one SNP of high interest but also hundreds of other variants that may be looked at in the future. Four of these events stand out because they were not present in the 1000 Bulls genome database but were not looked at further here due to their locations being outside the 1.2 mb region of interest. Genetics influences many aspects of life as its building blocks. In studying its role in reproduction, we move towards improving the health and well-being of our animals. We can

capitalize on non-invasive means of care and increase our own knowledge and understanding of the mechanisms driving the reproductive cycle.

Chapter 7

Conclusions and future directions

Cattle production, both milk and meat, functions as a direct result of successful reproductive events. Looking at potential genetic mechanisms that underline events such as multiple births, embryo lethality, and ovulation rate provide can increase subject knowledge and provide tools for producers to use in making breeding selections. At the conclusion of the twinning research, a new estimate of twining rate heritability and repeatability using 1.44 million North American Holstein calving records were generated using both a linear and threshold model. This is one of the first times repeatability of twinning has been reported using a threshold model. Additionally, a region on BTA11 from 28-32 Mb was identified to be significantly associated with sire daughter averages for twinning rate. Within this region are two candidate genes of interest, *FSHR* and *LHCGR*, which are involved in folliculogenesis. The estimation of genomic predictions showed an accuracy of ~40% when using older sires to predict newer ones. Results from the copy number variant analysis using whole genome sequencing added 468 new variants to the variant archive, and 33 potential deletions to screen further for embryo lethality potential. Finally, the work on the Trio allele showed that haplotype breed background most likely comes from Hereford origins. Comparison of two homozygote individuals showed congruent results between the short and long read data. Screening variants within the 1.2 Mb positional region previously identified against the 1000 Bull Genomes data narrowed the potential candidates down to a single SNP. The exact effect remains unknown, but genotyping showed 100% correspondence between the variant and the inferred genotype of Trio descendants. Further when genotyping non-Trio descendant populations, it was only present in homozygous normal state.

For as much as we learned, there is still more to uncover about each of the mechanisms discussed. Though a region of interest that harbors two genes of interest was identified, more

work is needed to identify how this region influences frequency of multiple births. Further, correspondence of the most significant SNP within the region and how it influences prediction of twinning and the role it can play in selection. Initial deletion screening only implicated four deletions for embryo lethality potential. After further genotyping, none presented with the missing homozygotes. A deeper view into the deletions themselves revealed mis-genotyped variants with incorrect estimation of breakpoints. This highlights an area of further study – improving and benchmarking the quality of current structural variant detection methods used in animal breeding. Additionally, looking further into the variants detected to identify phenotypic or expression changes caused by these large deletions and duplications. Lastly, screening the other 33 deletions not tested due to time and funding constraints would be advantageous to rule them out as embryo lethal candidates. Regarding the Trio allele causative mutation, a novel variant was identified but the connection with *SMAD6* overexpression remains elusive. Looking at inserting this variant into normal granulosa cells and measuring expression of *SMAD6* would indicate if it is indeed the causative mutation. Given the use of two sources of sequencing data (short and long reads) and a contig was able to span the full 1.2 Mb region this screen for variants has been extensive. If this novel variant is not causative, then either looking at one of the 173 SNPs or 15 InDels that overlapped in the 1000 Bull genome data maybe a next step or looking outside the region previously indicated. Utilizing a trio sequence technique (sequence offspring and both parents) in the future, should none of the currently detected variants in the homozygous half-siblings prove causative, may be an advantageous next step. Improvements in sequencing technologies in terms of quality, length and cost would improve detection of variants (SNPs, InDels, and structural) and would help all projects when trying to identify underlying

genetic mechanisms for multiple births, the Trio high ovulation rate phenotype, and screening deletions for embryo lethality potential.

References

1. Findlay, J. K. *et al. Follicle Selection in Mammalian Ovaries*. *The Ovary* (Elsevier Inc., 2018). doi:10.1016/b978-0-12-813209-8.00001-7

2. Vegetti, W. & Alagna, F. FSH and folliculogenesis: From physiology to ovarian stimulation. *Reprod. Biomed. Online* **12**, 684–694 (2006).

3. Scaramuzzi, R. J. *et al.* Regulation of folliculogenesis and the determination of ovulation rate in ruminants. *Reprod. Fertil. Dev.* **23**, 444–67 (2011).

4. Palermo, R. Differential actions of FSH and LH during folliculogenesis. *Reprod. Biomed. Online* **15**, 326–337 (2007).

5. Ginther, O. J. Selection of the dominant follicle in cattle and horses. *Anim. Reprod. Sci.* **60**–**61**, 61–79 (2000).

6. Wiltbank, M. C., Fricke, P. M., Sangsritavong, S., Sartori, R. & Ginther, O. J. Mechanisms that Prevent and Produce Double Ovulations in Dairy Cattle. *J. Dairy Sci.* **83**, 2998–3007 (2000).

7. Vinet, A. *et al.* Genetic control of multiple births in low ovulating mammalian species. *Mamm. Genome* **23**, 727–740 (2012).

8. Knight, P. G. & Glister, C. TGF-beta superfamily members and ovarian follicle development. *Reproduction* **132**, 191–206 (2006).

9. Ginther, O. J., Bergfelt, D. R., Kulick, L. J. & Kot, K. Selection of the dominant follicle in cattle: role of estradiol. *Biol. Reprod.* **63**, 383–389 (2000).

10. Russell, D. L. & Robker, R. L. Ovulation: The Coordination of Intrafollicular Networks to Ensure Oocyte Release. in *The Ovary* 217–234 (Elsevier, 2019). doi:10.1016/B978-0-12-813209-8.00014-5

11. Diskin, M. G., Parr, M. H. & Morris, D. G. Embryo death in cattle: an update. *Reprod. Fertil. Dev.* **24**, 244 (2012).

12. Wiltbank, M. C. *et al.* Pivotal periods for pregnancy loss during the first trimester of gestation in lactating dairy cows. *Theriogenology* **86**, 239–253 (2016).

13. Morris, D. & Diskin, M. Effect of progesterone on embryo survival. *Animal* **2**, 1112–1119 (2008).

14. Lopez-Gatius, F., Santolaria, P., Yaniz, J., Garbayo, J. & Hunter, R. Timing of Early Foetal Loss for Single and Twin Pregnancies in Dairy Cattle. *Reprod. Domest. Anim.* **39**, 429–433 (2004).

15. Echternkamp, S. E., Cushman, R. A., Allan, M. F., Thallman, R. M. & Gregory, K. E. Effects of ovulation rate and fetal number on fertility in twin-producing cattle. *J. Anim. Sci.* **85**, 3228–3238 (2007).

16. Echternkamp, S. E., Thallman, R. M., Cushman, R. A., Allan, M. F. & Gregory, K. E. Increased calf production in cattle selected for twin ovulations1,2. *J. Anim. Sci.* **85**, 3239–

3248 (2007).

17. Schultz, R. M. The molecular foundations of the maternal to zygotic transition in the preimplantation embryo. *Hum. Reprod. Update* **8**, 323–331 (2002).

18. Svoboda, P., Franke, V. & Schultz, R. M. Sculpting the Transcriptome During the Oocyte-to-Embryo Transition in Mouse. in *Current Topics in Developmental Biology* **113**, 305–349 (Academic Press, 2015).

19. Lonergan, P., Fair, T., Forde, N. & Rizos, D. Embryo development in dairy cattle. *Theriogenology* **86**, 270–277 (2016).

20. Valour, D. *et al.* Dairy cattle reproduction is a tightly regulated genetic process: Highlights on genes, pathways, and biological processes. *Anim. Front.* **5**, 32–41 (2015).

21. Diskin, M. G., Waters, S. M., Parr, M. H. & Kenny, D. A. Pregnancy losses in cattle: Potential for improvement. *Reprod. Fertil. Dev.* **28**, 83–93 (2016).

22. Baruselli, P. S. *et al.* Intrinsic and extrinsic factors that influence ovarian environment and efficiency of reproduction in cattle. *Anim. Reprod.* **14**, 48–60 (2017).

23. Berry, D. P., Friggens, N. C., Lucy, M. & Roche, J. R. Milk production and fertility in cattle. *Annu. Rev. Anim. Biosci.* **4**, 269–290 (2016).

24. D'Occhio, M. J., Baruselli, P. S. & Campanile, G. Influence of nutrition, body condition, and metabolic status on reproduction in female beef cattle: A review. *Theriogenology* **125**, 277–284 (2019).

25. Ribeiro, E. S. *et al.* Carryover effect of postpartum inflammatory diseases on developmental biology and fertility in lactating dairy cows. *J. Dairy Sci.* **99**, 2201–2220 (2016).

26. Gilbert, R. O. The effects of endometritis on the establishment of pregnancy in cattle. *Reprod. Fertil. Dev.* **24**, 252–257 (2011).

27. Johansson, I. The sex ratio and multiple births in cattle. *Zeitschrift für Tierzüchtung und Züchtungsbiologie einschließlich Tierernährung* **24**, 183–268 (1932).

28. Silva del Río, N., Kirkpatrick, B. W. & Fricke, P. M. Observed frequency of monozygotic twinning in Holstein dairy cattle. *Theriogenology* **66**, 1292–1299 (2006).

29. Johansson, I., Lindhé, B. & Pirchner, F. Causes of variation in the frequency of monozygous and dizygous twinning in various breeds of cattle. *Hereditas* **78**, 201–234 (1974).

30. Meadows, C. E. & Lush, J. L. Twinning in Dairy Cattle and Its Relation to Production. *J. Dairy Sci.* **40**, 1430–1436 (1957).

31. Erb, R. E. & Morrison, R. A. Effects of Twinning on Reproductive Efficiency in a Holstein-Friesian Herd. *J. Dairy Sci.* **42**, 512–519 (1959).

32. Cady, R. A. & Van Vleck, L. D. Factors affecting twinning and effects of twinning in Holstein dairy cattle. *J. Anim. Sci.* **46**, 950–956 (1978).

33. Nielen, M., Schukken, Y. H., Scholl, D. T., Wilbrink, H. J. & Brand, A. Twinning in dairy cattle: A study of risk factors and effects. *Theriogenology* **32**, 845–862 (1989).

34. Ryan, D. P. & Boland, M. P. Frequency of twin births among Holstein-Friesian cows in a warm dry climate. *Theriogenology* **36**, 1–10 (1991).

35. Fitzgerald, A. M., Berry, D. P., Carthy, T., Cromie, A. R. & Ryan, D. P. Risk factors associated with multiple ovulation and twin birth rate in Irish dairy and beef cattle. *J. Anim. Sci.* **92**, 966–973 (2014).

36. Lett, B. M. & Kirkpatrick, B. W. Short communication: Heritability of twinning rate in Holstein cattle. *J. Dairy Sci.* **101**, 4307–4311 (2018).

37. Rutledge, J. J. Twinning in Cattle. *J. Anim. Sci.* **40**, 803–815 (1975).

38. Johanson, J. M., Berger, P. J., Kirkpatrick, B. W. & Dentine, M. R. Twinning Rates for North American Holstein Sires. *J. Dairy Sci.* **84**, 2081–2088 (2001).

39. Morris, C. A. & Packard, P. M. *Progress with Beefplan*. (1984).

40. Beemsterboer, S. N. *et al.* The paradox of declining fertility but increasing twinning rates with advancing maternal age. *Hum. Reprod.* **21**, 1531–1532 (2006).

41. Broekmans, F. J., Soules, M. R. & Fauser, B. C. Ovarian aging: Mechanisms and clinical consequences. *Endocr. Rev.* **30**, 465–493 (2009).

42. Hoekstra, C. *et al.* Dizygotic twinning. *Hum. Reprod. Update* **14**, 37–47 (2008).

43. Kinsel, M. L., Marsh, W. E., Ruegg, P. L. & Etherington, W. G. Risk factors for twinning in dairy cows. *J. Dairy Sci.* **81**, 989–993 (1998).

44. Lopez, H., Caraviello, D. Z., Satter, L. D., Fricke, P. M. & Wiltbank, M. C. Relationship Between Level of Milk Production and Multiple Ovulations in Lactating Dairy Cows. *J. Dairy Sci.* **88**, 2783–2793 (2005).

45. Kusaka, H., Miura, H., Kikuchi, M. & Sakaguchi, M. Incidence of double ovulation during the early postpartum period in lactating dairy cows. *Theriogenology* **91**, 98–103 (2017).

46. Wiltbank, M., Lopez, H., Sartori, R., Sangsritavong, S. & Gümen, A. Changes in reproductive physiology of lactating dairy cows due to elevated steroid metabolism. *Theriogenology* **65**, 17–29 (2006).

47. Dunn, T. G. & Moss, G. E. Effects of nutrient deficiencies and excesses on reproductive efficiency of livestock. *J. Anim. Sci.* **70**, 1580–1593 (1992).

48. Syrstad, O. Genetic Aspects of Twinning in Dairy Cattle. *Acta Agric. Scand.* **24**, 319–322 (1974).

49. Gregory, K. E. *et al.* Twinning in cattle: I. Foundation animals and genetic and environmental effects on twinning rate. *J. Anim. Sci.* **68**, 1867–1876 (1990).

50. Ron, M., Ezra, E. & Weller, J. Genetic analysis of twinning rate in Israeli Holstein cattle. *Genet. Sel. Evol.* **22**, 349–359 (1990).

51. Bowman, J. C. & Hendy, C. R. C. C. The incidence, repeatability and effect on dam performance of twinning in British Friesian cattle. *Anim. Prod.* **12**, 55–62 (1970).

52. Van Vleck, L. D., Gregory, K. E. & Echternkamp, S. E. Ovulation rate and twinning rate in cattle: heritabilities and genetic correlation. *J. Anim. Sci.* **69**, 3213 (1991).

53. Gregory, K. E., Bennett, G. L., Van Vleck, L. D., Echternkamp, S. E. & Cundiff, L. V. Genetic and environmental parameters for ovulation rate, twinning rate, and weight traits in a cattle population selected for twinning. *J. Anim. Sci.* **75**, 1213–1222 (1997).

54. Kirkpatrick, B. W. & Morris, C. A. A Major Gene for Bovine Ovulation Rate. *PLoS One* **10**, e0129025 (2015).

55. Morris, C. A., Wheeler, M., Levet, G. L. & Kirkpatrick, B. W. A cattle family in New Zealand with triplet calving ability. *Livest. Sci.* **128**, 193–196 (2010).

56. Sawa, A., Bogucki, M. & Głowska, M. Effect of single and multiple pregnancies on performance of primiparous and multiparous cows. *Arch. Anim. Breed.* **58**, 43–48 (2015).

57. Andreu-Vázquez, C., Garcia-Ispierto, I., Ganau, S., Fricke, P. M. & López-Gatius, F. Effects of twinning on the subsequent reproductive performance and productive lifespan of high-producing dairy cows. *Theriogenology* **78**, 2061–2070 (2012).

58. Silva-del-Río, N., Colloton, J. D. & Fricke, P. M. Factors affecting pregnancy loss for single and twin pregnancies in a high-producing dairy herd. *Theriogenology* **71**, 1462–1471 (2009).

59. Fricke, P. M. & Wiltbank, M. C. Effect of milk production on the incidence of double ovulation in dairy cows. *Theriogenology* **52**, 1133–1143 (1999).

60. López-Gatius, F., López-Béjar, M., Fenech, M. & Hunter, R. H. F. Ovulation failure and double ovulation in dairy cattle: Risk factors and effects. *Theriogenology* **63**, 1298–1307 (2005).

61. Del Río, N. S., Stewart, S., Rapnicki, P., Chang, Y. M. & Fricke, P. M. An Observational Analysis of Twin Births, Calf Sex Ratio, and Calf Mortality in Holstein Dairy Cattle. *J. Dairy Sci.* **90**, 1255–1264 (2007).

62. Ghavi Hossein-Zadeh, N., Nejati-Javaremi, A., Miraei-Ashtiani, S. R. R. & Kohram, H. Estimation of variance components and genetic trends for twinning rate in Holstein dairy cattle of Iran. *J. Dairy Sci.* **92**, 3411–3421 (2009).

63. Echternkamp, S. E. *et al.* Twinning in cattle: II. Genetic and environmental effects on ovulation rate in puberal heifers and postpartum cows and the effects of ovulation rate on embryonic survival. *J. Anim. Sci.* **68**, 1877–1888 (1990).

64. Shi, Y. & Massagué, J. Mechanisms of TGF-β Signaling from Cell Membrane to the Nucleus. *Cell* **113**, 685–700 (2003).

65. Massague, J. Smad transcription factors. *Genes Dev.* **19**, 2783–2810 (2005).

66. Derynck, R. & Zhang, Y. E. Smad-dependent and Smad-independent pathways in TGF-β family signalling. *Nature* **425**, 577–584 (2003).

67. Miyazono, K., Kusanagi, K. & Inoue, H. Divergence and convergence of TGF-β/BMP signaling. *J. Cell. Physiol.* **187**, 265–276 (2001).

68. Vukicevic, S. & Sampath, K. T. *Bone Morphogenetic Proteins: Systems Biology Regulators*. (Springer International Publishing, 2017). doi:10.1007/978-3-319-47507-3

69. Bragdon, B. *et al.* Bone Morphogenetic Proteins: A critical review. *Cell. Signal.* **23**, 609–620 (2011).

70. Shimasaki, S., Moore, R. K., Otsuka, F. & Erickson, G. F. The Bone Morphogenetic Protein System in Mammalian Reproduction. *Endocr. Rev.* **25**, 72–101 (2004).

71. Mazerbourg, S. & Hsueh, A. J. W. Genomic analyses facilitate identification of receptors and signalling pathways for growth differentiation factor 9 and related orphan bone morphogenetic protein/growth differentiation factor ligands. *Hum. Reprod. Update* **12**, 373–383 (2006).

72. Fabre, S. *et al.* Regulation of ovulation rate in mammals: contribution of sheep genetic models. *Reprod. Biol. Endocrinol.* **4**, 20 (2006).

73. Miyazono, K., Maeda, S. & Imamura, T. BMP receptor signaling: Transcriptional targets, regulation of signals, and signaling cross-talk. *Cytokine Growth Factor Rev.* **16**, 251–263 (2005).

74. Rider, C. C. & Mulloy, B. Bone morphogenetic protein and growth differentiation factor cytokine families and their protein antagonists. *Biochem. J.* **429**, 1–12 (2010).

75. Hinck, A. P., Mueller, T. D. & Springer, T. A. Structural biology and evolution of the TGF-β family. *Cold Spring Harb. Perspect. Biol.* **8**, (2016).

76. Li, Q. Inhibitory SMADs: Potential Regulators of Ovarian Function1. *Biol. Reprod.* **92**, 1–6 (2015).

77. Juengel, J. L. & McNatty, K. P. The role of proteins of the transforming growth factor-β superfamily in the intraovarian regulation of follicular development. *Hum. Reprod. Update* **11**, 144–161 (2005).

78. Richards, J. & Pangas, S. A. The ovary : basic biology and clinical implications. *J. Clin. Investig.* **120**, 963–972 (2010).

79. De Castro, F. C., Cruz, M. H. C. & Leal, C. L. V. Role of Growth differentiation factor 9 and bone morphogenetic protein 15 in ovarian function and their importance in mammalian female fertility - A review. *Asian-Australasian J. Anim. Sci.* **29**, 1065–1074 (2016).

80. Otsuka, F., McTavish, K. J. & Shimasaki, S. Integral role of GDF-9 and BMP-15 in ovarian function. *Mol. Reprod. Dev.* **78**, 9–21 (2011).

81. Pangas, S. A. & Matzuk, M. M. Genetic models for transforming growth factor β superfamily signaling in ovarian follicle development. *Mol. Cell. Endocrinol.* **225**, 83–91 (2004).

82. Jorgez, C. J., Lin, Y.-N. & Matzuk, M. M. Genetic manipulations to study reproduction.

Mol. Cell. Endocrinol. **234**, 127–135 (2005).

83. Glister, C., Kemp, C. F. & Knight, P. G. Bone morphogenetic protein (BMP) ligands and receptors in bovine ovarian follicle cells: actions of BMP-4, -6 and -7 on granulosa cells and differential modulation of Smad-1 phosphorylation by follistatin. *Reproduction* **127**, 239–254 (2004).

84. Shimasaki, S. *et al.* A functional bone morphogenetic protein system in the ovary. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 7282–7287 (1999).

85. Wang, W. *et al.* The genetic polymorphisms of TGFβ superfamily genes are associated with litter size in a Chinese indigenous sheep breed (Hu sheep). *Animal Reproduction Science* **189**, 19–29 (2018).

86. Abdoli, R., Zamani, P., Mirhoseini, S., Ghavi Hossein-Zadeh, N. & Nadri, S. A review on prolificacy genes in sheep. *Reprod. Domest. Anim.* **51**, 631–637 (2016).

87. Xu, S. S. *et al.* Genome-wide association analyses highlight the potential for different genetic mechanisms for litter size among sheep breeds. *Front. Genet.* **9**, 1–14 (2018).

88. Davis, G. H. *et al.* Investigation of the Booroola (FecB) and Inverdale (FecXI) mutations in 21 prolific breeds and strains of sheep sampled in 13 countries. *Anim. Reprod. Sci.* **92**, 87–96 (2006).

89. Davis, G. H. Major genes affecting ovulation rate in sheep. *Genet. Sel. Evol.* **37**, 11–23 (2005).

90. Souza, C. J. H., MacDougall, C., Campbell, B. K., McNeilly, A. S. & Baird, D. T. The Booroola (FecB) phenotype is associated with a mutation in the bone morphogenetic receptor type 1 B (BMPR1B) gene. *J. Endocrinol.* **169**, (2001).

91. Chu, M. X. *et al.* Mutations in BMPR-IB and BMP-15 genes are associated with litter size in Small Tailed Han sheep (Ovis aries)1. *J. Anim. Sci.* **85**, 598–603 (2007).

92. Mulsant, P. *et al.* Mutation in bone morphogenetic protein receptor-IB is associated with increased ovulation rate in Booroola Merino ewes. *Proc. Natl. Acad. Sci.* **98**, 5104–5109 (2001).

93. Wilson, T. *et al.* Highly Prolific Booroola Sheep Have a Mutation in the Intracellular Kinase Domain of Bone Morphogenetic Protein IB Receptor (ALK-6) That Is Expressed in Both Oocytes and Granulosa Cells1. *Biol. Reprod.* **64**, 1225–1235 (2001).

94. Galloway, S. M. *et al.* Mutations in an oocyte-derived growth factor gene (BMP15) cause increased ovulation rate and infertility in a dosage-sensitive manner. *Nat. Genet.* **25**, 279–283 (2000).

95. Hanrahan, J. P. *et al.* Mutations in the Genes for Oocyte-Derived Growth Factors GDF9 and BMP15 Are Associated with Both Increased Ovulation Rate and Sterility in Cambridge and Belclare Sheep (Ovis aries)1. *Biol. Reprod.* **70**, 900–909 (2004).

96. Monteagudo, L. V., Ponz, R., Tejedor, M. T., Laviña, A. & Sierra, I. A 17 bp deletion in the Bone Morphogenetic Protein 15 (BMP15) gene is associated to increased prolificacy

in the Rasa Aragonesa sheep breed. *Anim. Reprod. Sci.* **110**, 139–146 (2009).

97. Martinez-Royo, A. *et al.* A deletion in the bone morphogenetic protein 15 gene causes sterility and increased prolificacy in Rasa Aragonesa sheep. *Anim. Genet.* **39**, 294–297 (2008).

98. Demars, J. *et al.* Genome-Wide Association Studies Identify Two Novel BMP15 Mutations Responsible for an Atypical Hyperprolificacy Phenotype in Sheep. *PLoS Genet.* **9**, (2013).

99. Bodin, L. *et al.* A novel mutation in the bone morphogenetic protein 15 gene causing defective protein secretion is associated with both increased ovulation rate and sterility in Lacaune sheep. *Endocrinology* **148**, 393–400 (2007).

100. Lassoued, N. *et al.* FecX Bar a Novel BMP15 mutation responsible for prolificacy and female sterility in Tunisian Barbarine Sheep. *BMC Genet.* **18**, 1–10 (2017).

101. Moradband, F., Rahimi, G. & Gholizadeh, M. Association of polymorphisms in fecundity genes of GDF9, BMP15 and BMP15-1B with litter size in iranian Baluchi sheep. *Asian-Australasian J. Anim. Sci.* **24**, 1179–1183 (2011).

102. Mullen, M. P. & Hanrahan, J. P. Direct evidence on the contribution of a missense mutation in GDF9 to variation in ovulation rate of Finnsheep. *PLoS One* **9**, (2014).

103. Våge, D. I., Husdal, M., Kent, M. P., Klemetsdal, G. & Boman, I. A. A missense mutation in growth differentiation factor 9 (GDF9) is strongly associated with litter size in sheep. *BMC Genet.* **14**, 1–8 (2013).

104. Nicol, L. *et al.* Homozygosity for a single base-pair mutation in the oocyte-specific GDF9 gene results in sterility in Thoka sheep. *Reproduction* **138**, 921–933 (2009).

105. Silva, B. D. M. *et al.* A new polymorphism in the Growth and Differentiation Factor 9 (GDF9) gene is associated with increased ovulation rate and prolificacy in homozygous sheep. *Anim. Genet.* **42**, 89–92 (2011).

106. Harris, R. A. *et al.* Evolutionary genetics and implications of small size and twinning in callitrichine primates. *Proc. Natl. Acad. Sci.* **111**, 1467–1472 (2014).

107. Sharma, R. *et al.* Polymorphism of BMP4 gene in Indian goat breeds differing in prolificacy. *Gene* **532**, 140–145 (2013).

108. Feng, X. *et al.* Polymorphisms of the bone morphogenetic protein 7 gene (BMP7) and association analysis with sow productive traits. *Anim. Reprod. Sci.* **142**, 56–62 (2013).

109. Marchitelli, C. & Nardone, A. Mutations and sequence variants in GDF9, BMP15, and BMPR1B genes in Maremmana cattle breed with single and twin births. *Rend. Lincei* **26**, 553–560 (2015).

110. Palmer, J. S. *et al.* Novel variants in growth differentiation factor 9 in mothers of dizygotic twins. *J. Clin. Endocrinol. Metab.* **91**, 4713–4716 (2006).

111. Montgomery, G. W. *et al.* A deletion mutation in GDF9 in sisters with spontaneous DZ twins. *Twin Res.* **7**, 548–555 (2004).

112. Li, W.-T. *et al.* Whole-genome resequencing reveals candidate mutations for pig prolificacy. *Proc. R. Soc. B Biol. Sci.* **284**, 20172437 (2017).

113. Dong, J. *et al.* Growth differentiation factor-9 is required during early ovarian folliculogenesis. *Nature* **383**, 531–535 (1996).

114. Yan, C. *et al.* Synergistic Roles of Bone Morphogenetic Protein 15 and Growth Differentiation Factor 9 in Ovarian Function. *Mol. Endocrinol.* **15**, 854–866 (2001).

115. Yi, S. E. *et al.* The type I BMP receptor BmprIB is essential for female reproductive function. *Proc. Natl. Acad. Sci.* **98**, 7994–7999 (2001).

116. Matzuk, M. M., Kumar, T. R. & Bradley, A. Different phenotypes for mice deficient in either activins or activin receptor type II. *Nature* **374**, 356–360 (1995).

117. Tomic, D. *et al.* Smad 3 May Regulate Follicular Growth in the Mouse Ovary1. *Biol. Reprod.* **66**, 917–923 (2005).

118. Li, Q. *et al.* Redundant Roles of SMAD2 and SMAD3 in Ovarian Granulosa Cells In Vivo. *Mol. Cell. Biol.* **28**, 7001–7011 (2008).

119. Lee, W. S., Otsuka, F., Moore, R. K. & Shimasaki, S. Effect of bone morphogenetic protein-7 on folliculogenesis and ovulation in the rat. *Biol. Reprod.* **65**, 994–999 (2001).

120. Inayah, A., Rahayu, S., Widodo, N. & Prasdini, W. A. A new nucleotide variant G1358A potentially change growth differentiation factor 9 profile that may affect the reproduction performance of Friesian Holstein cattle. *Asian Pacific Journal of Reproduction* **5**, 140–143 (2016).

121. Sasaki, S. *et al.* Genetic variants in the upstream region of activin receptor IIA are associated with female fertility in Japanese Black cattle. *BMC Genet.* **16**, (2015).

122. Kamalludin, M. H., Garcia-Guerra, A., Wiltbank, M. C. & Kirkpatrick, B. W. Proteomic analysis of follicular fluid in carriers and non-carriers of the Trio allele for high ovulation rate in cattle. *Reprod. Fertil. Dev.* **30**, 1643–1650 (2018).

123. Kamalludin, M. H., Garcia-Guerra, A., Wiltbank, M. C. & Kirkpatrick, B. W. Trio, a novel high fecundity allele: I. Transcriptome analysis of granulosa cells from carriers and noncarriers of a major gene for bovine ovulation rate†. *Biol. Reprod.* **98**, 323–334 (2018).

124. García-Guerra, A. *et al.* Trio, a novel bovine high fecundity allele: III. Acquisition of dominance and ovulatory capacity at a smaller follicle size†. *Biol. Reprod.* **0**, 1–16 (2018).

125. García-Guerra, A., Kirkpatrick, B. W. & Wiltbank, M. C. Follicular waves and hormonal profiles during the estrous cycle of carriers and non-carriers of the Trio allele, a major bovine gene for high ovulation and fecundity. *Theriogenology* **100**, 100–113 (2017).

126. Garcia-Guerra, A., Kamalludin, M. H., Kirkpatrick, B. W. & Wiltbank, M. C. Trio a novel bovine high-fecundity allele: II. Hormonal profile and follicular dynamics underlying the high ovulation rate†. *Biol. Reprod.* **98**, 335–349 (2018).

127. Li, X. *et al.* Whole-genome sequencing identifies potential candidate genes for reproductive traits in pigs. *Genomics* (2019). doi:10.1016/j.ygeno.2019.01.014

128. Schneider, J. F., Nonneman, D. J., Wiedmann, R. T., Vallet, J. L. & Rohrer, G. A. Genomewide association and identification of candidate genes for ovulation rate in swine1,2. *J. Anim. Sci.* **92**, 3792–3803 (2014).

129. Chen, K. *et al.* Association of the Porcine Transforming Growth Factor Beta Type I Receptor (TGFBR1) Gene with Growth and Carcass Traits. *Anim. Biotechnol.* **23**, 43–63 (2012).

130. Sell-Kubiak, E. *et al.* Genome-wide association study reveals novel loci for litter size and its variability in a Large White pig population. *BMC Genomics* **16**, 1–13 (2015).

131. Rosen, B. D. *et al.* Modernizing the Bovine Reference Genome Assembly. *Proc. World Congr. Genet. Appl. Livest. Prod.* 802 (2018).

132. Kim, E.-S., Berger, P. J. & Kirkpatrick, B. W. Genome-wide scan for bovine twinning rate QTL using linkage disequilibrium. *Anim. Genet.* **40**, 300–307 (2009).

133. Logan, C. Y. & Nusse, R. the Wnt Signaling Pathway in Development and Disease. *Annu. Rev. Cell Dev. Biol.* **20**, 781–810 (2004).

134. Stinckens, A. *et al.* Indirect effect of IGF2 intron3 g . 3072G > A mutation on prolificacy in sows. 493–498 (2010). doi:10.1111/j.1365-2052.2010.02040.x

135. An, S. M. *et al.* Effect of Single Nucleotide Polymorphisms in IGFBP2 and IGFBP3 Genes on Litter Size Traits in Berkshire Pigs. *Anim. Biotechnol.* **29**, 301–308 (2018).

136. Silva, J. R. V, Figueiredo, J. R. & van den Hurk, R. Involvement of growth hormone (GH) and insulin-like growth factor (IGF) system in ovarian folliculogenesis. *Theriogenology* **71**, 1193–1208 (2009).

137. Rothschild, M. *et al.* The estrogen receptor locus is associated with a major gene influencing litter size in pigs. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 201–205 (1996).

138. Laliotis, G. P., Marantidis, A. & Avdi, M. Association of BF, RBP4, and ESR2 Genotypes with Litter Size in an Autochthonous Pig Population. *Anim. Biotechnol.* **28**, 138–143 (2017).

139. Jiang, Z., Gibson, J. P., Archibald, A. L. & Haley, C. S. The porcine gonadotropin-releasing hormone receptor gene (GNRHR): Genomic organization, polymorphisms, and association with the number of corpora lutea. *Genome* **44**, 7–12 (2001).

140. Chu, M. X. *et al.* Polymorphism of 5′ regulatory region of ovine FSHR gene and its association with litter size in Small Tail Han sheep. *Mol. Biol. Rep.* **39**, 3721–3725 (2012).

141. Tao, H. *et al.* Associations of TCF12, CTNNAL1 and WNT10B gene polymorphisms with litter size in pigs. *Anim. Reprod. Sci.* **140**, 189–194 (2013).

142. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

143. Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* **18**, 9–19 (2020).

144. RFA-HG-06-020: Revolutionary Genome Sequencing Technologies – The $1000 Genome (R01). 1–30 (2010).

145. Quail, M. *et al.* A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).

146. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (80-.).* **323**, 133–138 (2009).

147. Levene, M. J. *et al.* Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science (80-.).* **299**, 682–686 (2003).

148. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).

149. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).

150. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).

151. Vignal, A., Milan, D., SanCristobal, M. & Eggen, A. A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* **34**, 275 (2002).

152. Khatkar, M. S., Thomson, P. C., Tammen, I. & Raadsma, H. W. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genet. Sel. Evol.* **36**, 163–190 (2004).

153. Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19**, R131–R136 (2010).

154. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).

155. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res.* **23**, 749–761 (2013).

156. Shi, T. *et al.* A novel 17 bp indel in the <i>SMAD3</i> gene alters transcription level, contributing to phenotypic traits in Chinese cattle. *Arch. Anim. Breed.* **59**, 151–157 (2016).

157. Jakaria, J. *et al.* Discovery of SNPs and indel 11-bp of the myostatin gene and its association with the double-muscled phenotype in Belgian blue crossbred cattle. *Gene* **784**, 145598 (2021).

158. Pirooznia, M., Goes, F. & Zandi, P. P. Whole-genome CNV analysis: Advances in computational approaches. *Front. Genet.* **6**, 1–9 (2015).

159. Aldersey, J. E., Sonstegard, T. S., Williams, J. L. & Bottema, C. D. K. Understanding the effects of the bovine POLLED variants. *Anim. Genet.* **51**, 166–176 (2020).

160. Liu, G. E. *et al.* Analysis of copy number variations among diverse cattle breeds. *Genome Res.* **20**, 693–703 (2010).

161. Boussaha, M. *et al.* Genome-wide study of structural variants in bovine Holstein, Montbéliarde and Normande dairy breeds. *PLoS One* **10**, 1–21 (2015).

162. Hou, Y. *et al.* Genomic characteristics of cattle copy number variations. *BMC Genomics* **12**, (2011).

163. Mesbah-Uddin, M. *et al.* Genome-wide mapping of large deletions and their population-genetic properties in dairy cattle. *DNA Res.* **25**, 49–59 (2018).

164. Kadri, N. K. *et al.* A 660-Kb Deletion with Antagonistic Effects on Fertility and Milk Production Segregates at High Frequency in Nordic Red Cattle: Additional Evidence for the Common Occurrence of Balancing Selection in Livestock. *PLoS Genet.* **10**, (2014).

165. Fricke, P. M. Twinning in Dairy Cattle. *Prof. Anim. Sci.* **17**, 61–67 (2001).

166. Komisarek, J. & Dorynek, Z. Genetic aspects of twinning in cattle. *J. Appl. Genet.* **43**, 55–68 (2002).

167. Cammack, K. M., Thomas, M. G. & Enns, R. M. Reproductive traits and their heritabilities in beef cattle. *Prof. Anim. Sci.* **25**, 517–528 (2009).

168. Gregory, K. E., Echternkamp, S. E. & Cundiff, L. V.Twinning in cattle: III. Effects of Twinning on Dystocia, Calf Survival, Calf Growth, Carcass Traits, and C. P. Twinning in cattle: III. Effects of Twinning on Dystocia, Calf Survival, Calf Growth, Carcass Traits, and Cow Productivity. *J. Anim. Sci.* **68**, 3133–3144 (1990).

169. Hossein-Zadeh, N. G. The effect of twinning on milk yield, dystocia, calf birth weight and open days in Holstein dairy cows of Iran. *J. Anim. Physiol. Anim. Nutr. (Berl).* **94**, 780–787 (2010).

170. Mee, J. F., Berry, D. P. & Cromie, A. R. Risk factors for calving assistance and dystocia in pasture-based Holstein–Friesian heifers and cows in Ireland. *Vet. J.* **187**, 189–194 (2011).

171. Pardon, B. *et al.* Left abomasal displacement between the uterus and rumen during bovine twin pregnancy. *J. Vet. Sci.* **13**, 437–440 (2012).

172. Silva-del-Río, N., Fricke, P. M. & Grummer, R. R. Effects of twin pregnancy and dry period feeding strategy on milk production, energy balance, and metabolic profiles in dairy cows. *J. Anim. Sci.* **88**, 1048–1060 (2010).

173. Gregory, K. E., Echternkamp, S. E. & Cundiff, L. V. Effects of twinning on dystocia, calf survival, calf growth, carcass traits, and cow productivity. *J. Anim. Sci.* **74**, 1223–33 (1996).

174. Echternkamp, S. E. & Gregory, K. E. Reproductive, growth, feedlot, and carcass traits of twin vs single births in cattle. *J. Anim. Sci.* **80**, 1–10 (2002).

175. Echternkamp, S. E. & Gregory, K. E. Effects of twinning on postpartum reproductive performance in cattle selected for twin births. *J. Anim. Sci.* **77**, 48 (1999).

176. López-Gatius, F., Andreu-Vázquez, C., Mur-Novales, R., Cabrera, V. E. & Hunter, R. H. . The dilemma of twin pregnancies in dairy cattle. A review of practical prospects. *Livest.*

Sci. **197**, 12–16 (2017).

177. Chapin, C. A. & Van Vleck, L. D. Effects of twinning on lactation and days open in Holsteins. *J. Dairy Sci.* **63**, 1881–1886 (1980).

178. Esteves, A., Båge, R. & Payan-Carreira, R. Freemartinism in cattle. *Ruminants Anatomy, Behav. Dis. Free. Cattle* 99–120 (2012).

179. Beerepoot, G. M., Dykhuizen, A. A., Nielen, Y. & Schukken, Y. H. The economics of naturally occurring twinning in dairy cattle. *J. Dairy Sci.* **75**, 1044–1051 (1992).

180. de Rose, E. P. & Wilton, J. W. Productivity and profitability of twin births in beef cattle. *J. Anim. Sci.* **69**, 3085 (1991).

181. White, R. R., Brady, M., Capper, J. L., McNamara, J. P. & Johnson, K. A. Cow–calf reproductive, genetic, and nutritional management to improve the sustainability of whole beef production systems. *J. Anim. Sci.* **93**, 3197–3211 (2015).

182. Cummins, L. J., Morris, C. A. & Kirkpatrick, B. W. Developing twinning cattle for commercial production. *Aust. J. Exp. Agric.* **48**, 930 (2008).

183. De Vries, A. Economic Value of Pregnancy in Dairy Cattle. *J. Dairy Sci.* **89**, 3876–3885 (2006).

184. Sigdel, A., Bisinotto, R. S. & Peñagaricano, F. Genes and pathways associated with pregnancy loss in dairy cattle. *Sci. Rep.* **11**, 13329 (2021).

185. Cabrera, V. E. Economics of fertility in high-yielding dairy cows on confined TMR systems. *Animal* **8**, 211–221 (2014).

186. Gershoni, M., Ezra, E. & Weller, J. I. Genetic and genomic analysis of long insemination interval in Israeli dairy cattle as an indicator of early abortions. *J. Dairy Sci.* **103**, 4495–4509 (2020).

187. VanRaden, P. M., Olson, K. M., Null, D. J. & Hutchison, J. L. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *J. Dairy Sci.* **94**, 6153–6161 (2011).

188. Kappes, S. M. *et al.* Initial results of genomic scans for ovulation rate in a cattle population selected for increased twinning rate. *J. Anim. Sci.* **78**, 3053–3059 (2000).

189. Weller, J. I., Golik, M., Seroussi, E., Ron, M. & Ezra, E. Detection of Quantitative Trait Loci Affecting Twinning Rate in Israeli Holsteins by the Daughter Design. *J. Dairy Sci.* **91**, 2469–2474 (2008).

190. Bierman, C. D. *et al.* Validation of whole genome linkage-linkage disequilibrium and association results, and identification of markers to predict genetic merit for twinning. *Anim. Genet.* **41**, 406–416 (2010).

191. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Publ. Gr.* (2011). doi:10.1038/nrg2958

192. Charlier, C. *et al.* NGS-based reverse genetic screen for common embryonic lethal

mutations compromising fertility in livestock. *Genome Res.* **26**, 1333–1341 (2016).

193. Kirkpatrick, B. & Morris, C. Discovery of a Major Gene for Bovine Ovulation Rate. *2011 SSR 44th Annu. Meet.* 192 (2011). doi:10.5061/dryad.66s7c.Funding

194. Kay, R. M. Changes in milk production, fertility and calf mortality associated with retained placentae or the birth of twins. *Vet. Rec.* **102**, 477–479 (1978).

195. Young, A. S. & Kirkpatrick, B. W. Frequency of leukochimerism in Holstein and Jersey twinsets1,2. *J. Anim. Sci.* **94**, 4507–4515 (2016).

196. Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J. and Thompson, R. ASReml User Guide. *VSN Int. Ltd, Hemel Hempstead* **HP1 1ES**, **U**, (2015).

197. Karlsen, A., Ruane, J., Klemetsdal, G. & Heringstad, B. Twinning rate in Norwegian cattle: Frequency, (co)variance components, and genetic trends. *J. Anim. Sci.* **78**, 15–20 (2000).

198. Moioli, B., Steri, R., Marchitelli, C., Catillo, G. & Buttazzoni, L. Genetic parameters and genome-wide associations of twinning rate in a local breed, the Maremmana cattle. *animal* **11**, 1660–1666 (2017).

199. Wolc, A., Bresińska, A. & Szwaczkowski, T. Genetic and permanent environmental variability of twinning in Thoroughbred horses estimated via three threshold models. *J. Anim. Breed. Genet.* **123**, 186–190 (2006).

200. Caraviello, D. Z. *et al.* Survey of Management Practices on Reproductive Performance of Dairy Cattle on Large US Commercial Farms. *J. Dairy Sci.* **89**, 4723–4735 (2006).

201. Kirkpatrick, B. W. & Lett, B. M. Short communication: Heritability of susceptibility to infection by Mycobacterium avium ssp. paratuberculosis in Holstein cattle. *J. Dairy Sci.* **101**, 11165–11169 (2018).

202. Meuwissen, T. H. E., Karlsen, A., Lien, S., Olsaker, I. & Goddard, M. E. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**, 373–379 (2002).

203. Cruickshank, J., Dentine, M. R., Berger, P. J. & Kirkpatrick, B. W. Evidence for quantitative trait loci affecting twinning rate in North American Holstein cattle. *Anim. Genet.* **35**, 206–212 (2004).

204. Cobanoglu, O., Berger, P. J. & Kirkpatrick, B. W. Genome screen for twinning rate QTL in four North American Holstein families. *Anim. Genet.* **36**, 303–308 (2005).

205. Hu, Z.-L., Park, C. A. & Reecy, J. M. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res.* **47**, D701–D710 (2019).

206. Wang, H. *et al.* Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Front. Genet.* **5**, 134 (2014).

207. Tiezzi, F., Parker-Gaddis, K. L., Cole, J. B., Clay, J. S. & Maltecca, C. A Genome-Wide

Association Study for Clinical Mastitis in First Parity US Holstein Cows Using Single-Step Approach and Genomic Matrix Re-Weighting Procedure. *PLoS One* **10**, e0114919 (2015).

208. Marques, D. B. D. *et al.* Weighted single-step GWAS and gene network analysis reveal new candidate genes for semen traits in pigs. *Genet. Sel. Evol.* **50**, 40 (2018).

209. Teissier, M., Larroque, H. & Robert-Granié, C. Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: a quantitative trait influenced by a major gene. *Genet. Sel. Evol.* **50**, 31 (2018).

210. Quick, A. E., Ollivett, T. L., Kirkpatrick, B. W. & Weigel, K. A. Genomic analysis of bovine respiratory disease and lung consolidation in preweaned Holstein calves using clinical scoring and lung ultrasound. *J. Dairy Sci.* **103**, 1632–1641 (2020).

211. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11**, 843–854 (2010).

212. Kao, P. Y. P., Leung, K. H., Chan, L. W. C., Yip, S. P. & Yap, M. K. H. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. *Biochim. Biophys. Acta - Gen. Subj.* **1861**, 335–353 (2017).

213. Das, S., McClain, C. J. & Rai, S. N. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. *Entropy* **22**, 427 (2020).

214. Fogarty, N. M. A review of the effects of the Booroola gene (FecB) on sheep production. *Small Rumin. Res.* **85**, 75–84 (2009).

215. Garrick, D. J., Taylor, J. F. & Fernando, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* **41**, 55 (2009).

216. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

217. Browning, B. L. & Browning, S. R. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* **194**, 459–471 (2013).

218. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2008).

219. Sargolzaei, M., Chesnais, J. P. & Schenkel, F. S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, (2014).

220. Hayes, B. J. & Daetwyler, H. D. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu. Rev. Anim. Biosci.* **7**, 89–102 (2019).

221. Rosen, B. D. *et al.* De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9**, 1–9 (2020).

222. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer

datasets. *Gigascience* **4**, 7 (2015).

223. Misztal, I. *et al.* Manual for BLUPF90 family of programs. *Univ. Georg. Athens, USA* (2015).

224. Wang, H., Misztal, I., Aguilar, I., Legarra, A. & Muir, W. M. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res. (Camb).* **94**, 73–83 (2012).

225. Aguilar, I., Legarra, A., Cardoso, F., Masuda, Y. & Lourenco, D. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. *Genet. Sel. Evol.* **51**, 1–8 (2019).

226. Aguilar, I. *et al.* Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* **93**, 743–752 (2010).

227. Storey, J. D., Bass, A. J., Dabney, A., Robinson, D. & Warnes, G. qvalue: Q-value estimation for false discovery rate control. R package version 2.24.0 (2021).

228. Haibe-Kains, B., Desmedt, C., Sotiriou, C. & Bontempi, G. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics* **24**, 2200–2208 (2008).

229. Schröder, M. S., Culhane, A. C., Quackenbush, J. & Haibe-Kains, B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* **27**, 3206–3208 (2011).

230. Turner, S. qqman: Q-Q and Manhattan Plots for GWAS Data. (2017).

231. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).

232. Simillion, C. SetRank: Advanced Gene Set Enrichment Analysis. R package version 1.1.0 (2016).

233. Simillion, C., Liechti, R., Lischer, H. E. L., Ioannidis, V. & Bruggmann, R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bio* **8**, 1–14 (2017).

234. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

235. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways. **20**, 1085–1093 (2019).

236. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

237. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

238. Wu, G. & Haw, R. Functional Interaction Network Construction and Analysis for Disease Discovery. in *Methods in Molecular Biology* **1558**, 235–253 (2017).

239. Martens, M. *et al.* WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–D621 (2021).

240. Simillion, C. SetRank User Manual. *cran.r-project.org* 1–26 (2016). Available at: https://cran.r-project.org/web/packages/SetRank/vignettes/vignette.pdf.

241. Zhang, X., Lourenco, D., Aguilar, I., Legarra, A. & Misztal, I. Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS. *Front. Genet.* **7**, 1–14 (2016).

242. Fragomeni, B. O., Lourenco, D. A. L., Legarra, A., VanRaden, P. M. & Misztal, I. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. *J. Dairy Sci.* **102**, 10012–10019 (2019).

243. VanRaden, P. M. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).

244. Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G. & Hickey, J. M. Genomic Prediction in Animals and Plants :Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics* **193**, 347–365 (2013).

245. Widmer, S. *et al.* A major QTL at the LHCGR / FSHR locus for multiple birth in Holstein cattle. *Genet. Sel. Evol.* **53**, 1–15 (2021).

246. Gajbhiye, R., Fung, J. N. & Montgomery, G. W. Complex genetics of female fertility. *npj Genomic Med.* **3**, 29 (2018).

247. Huhtaniemi, I. & Rivero-Müller, A. Mutations and Polymorphisms, and Their Functional Consequences, in Gonadotropin and Gonadotropin Receptor Genes. in *The Ovary* 127–148 (Elsevier, 2019). doi:10.1016/B978-0-12-813209-8.00008-X

248. Yu, Y. *et al.* Association of a missense mutation in the luteinizing hormone/choriogonadotropin receptor gene (LHCGR) with superovulation traits in Chinese Holstein heifers. *J. Anim. Sci. Biotechnol.* **3**, 35 (2012).

249. Mclaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, (2016).

250. Mohammed, H. *et al.* Endogenous Purification Reveals GREB1 as a Key Estrogen Receptor Regulatory Factor. *Cell Rep.* **3**, 342–349 (2013).

251. Kirkpatrick, B. W., Byla, B. M. & Gregory, K. E. Mapping quantitative trait loci for bovine ovulation rate. *Mamm. Genome* **11**, 136–139 (2000).

252. Lien, S. *et al.* A primary screen of the bovine genome for quantitative trait loci affecting twinning rate. *Mamm. Genome* **11**, 877–882 (2000).

253. Kim, E.-S. *et al.* Refined mapping of twinning-rate quantitative trait loci on bovine chromosome 5 and analysis of insulin-like growth factor-1 as a positional candidate gene1. *J. Anim. Sci.* **87**, 835–843 (2009).

254. Kirkpatrick, B. W., Thallman, R. M. & Kuehn, L. A. Validation of SNP associations with bovine ovulation and twinning rate. *Anim. Genet.* **50**, 259–261 (2019).

255. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinforma.* **54**, 1.30.1-1.30.33 (2016).

256. Zheng, P., Schramm, R. D. & Latham, K. E. Developmental Regulation and In Vitro Culture Effects on Expression of DNA Repair and Cell Cycle Checkpoint Control Genes in Rhesus Monkey Oocytes and Embryos. *Biol. Reprod.* **72**, 1359–1369 (2005).

257. Virant-klun, I., Knez, K., Tomazevic, T. & Skutella, T. Gene Expression Profiling of Human Oocytes Developed and Matured In Vivo or In Vitro. *Biomed Res. Int.* **2013**, 20 (2013).

258. Khokhlova, Evgenia V., Zoia S. Fesenko, Julia V. Sopova, and E. I. L. Features of DNA Repair in the Early Stages of Mammalian Embryonic Development. *Genes (Basel).* **11**, 1138 (2020).

259. Guo, X. & Wang, X.-F. Signaling cross-talk between TGF-beta/BMP and other pathways. *Cell Res.* **19**, 71–88 (2009).

260. VanRaden, P. M. & O'Connell, J. R. Validating Genomic Reliabilities and Gains from Phenotypic Updates. *Interbull Bull.* **53**, (2018).

261. Vukasinovic, N., Gonzalez-Pena, D., Brooker, J., Przybyla, C. & Denise, S. 156. Genomic evaluation for abortions and twinning in dairy cattle. *J. Dairy Sci.* **103**, 60–61 (2020).

262. Gaddis, K. L. P., Cole, J. B., Clay, J. S. & Maltecca, C. Genomic selection for producer-recorded health event data in US dairy cattle. *J. Dairy Sci.* **97**, 3190–3199 (2014).

263. CDCB. *Council on Dairy Cattle Breeding Activity Report 2018 Oct 17/Sep 18*. (2018).

264. Mcgovern, S. P. *et al.* Genomic Prediction for Twin Pregnancies. *Animals* **11**, 843 (2021).

265. Cezard, T. *et al.* The European Variation Archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res.* **50**, D1216–D1220 (2022).

266. Hou, Y. *et al.* Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Funct. Integr. Genomics* **12**, 81–92 (2012).

267. Durkin, K. *et al.* Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* **482**, 81–84 (2012).

268. Guan, D. *et al.* Estimating the copy number of the agouti signaling protein (ASIP) gene in goat breeds with different color patterns. *Livest. Sci.* **246**, 1–5 (2021).

269. Sahana, G. *et al.* A 0.5-Mbp deletion on bovine chromosome 23 is a strong candidate for stillbirth in Nordic Red cattle. *Genet. Sel. Evol.* **48**, 1–12 (2016).

270. Tattini, L., D'Aurizio, R. & Magi, A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front. Bioeng. Biotechnol.* **3**, 1–8 (2015).

271. Chen, L., Chamberlain, A. J., Reich, C. M., Daetwyler, H. D. & Hayes, B. J. Detection and validation of structural variations in bovine whole-genome sequence data. *Genet. Sel. Evol.* **49**, 1–13 (2017).

272. Kommadath, A. *et al.* A large interactive visual database of copy number variants discovered in taurine cattle. *Gigascience* **8**, 1–12 (2019).

273. Ben Hayes, Hans Daetwyler, Ruedi Fries, Paul Stothard, H. P., Rianne van Binsbergen, Roel Veerkamp, Aurélien Capitan, S., Fritz, Mogens Lund, Didier Boichard,, Curt Van Tassell, B. & Guldbrandtsen, Xiaoping Liao, and the 1000 bull genomes consortium. 1000 Bull Genomes Project. Available at: http://www.1000bullgenomes.com/.

274. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).

275. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

276. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **15**, 1–19 (2014).

277. Quinlan, A. R., Layer, R. M., Pedersen, B. & Larson, D. GitHub - arq5x/lumpy-sv: lumpy: a general probabilistic framework for structural variant discovery. *github.com* (2020). Available at: https://github.com/arq5x/lumpy-sv. (Accessed: 4th March 2022)

278. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 1–11 (2017).

279. Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res.* **47**, 745–751 (2019).

280. Ye, J. *et al.* Primer-BLAST : A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, (2012).

281. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).

282. Leinonen, R., Sugawara, H. & Shumway, M. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).

283. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* (2013). doi:10.1002/0471250953.bi1110s43

284. Zare, Y., Shook, G. E., Collins, M. T. & Kirkpatrick, B. W. Genome-Wide Association Analysis and Genomic Prediction of Mycobacterium avium Subspecies paratuberculosis Infection in US Jersey Cattle. *PLoS One* **9**, e88380 (2014).

285. Golden Helix GenomeBrowse ® visualization tool.

286. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).

287. Mulsant, P. *et al.* Mutation in bone morphogenetic protein receptor-IB is associated with increased ovulation rate in Booroola Mérino ewes. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5104–9 (2001).

288. Wilson, T. *et al.* Highly Prolific Booroola Sheep Have a Mutation in the Intracellular Kinase Domain of Bone Morphogenetic Protein IB Receptor (ALK-6) That Is Expressed in Both Oocytes and Granulosa Cells1. *Biol. Reprod.* **64**, 1225–1235 (2001).

289. Decker, J. E. *et al.* Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *PLoS Genet.* **10**, (2014).

290. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–44 (2006).

291. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol. 2019 375* **37**, 540–546 (2019).

292. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

293. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

294. ARS-UCD1.2 - bosTau9 - Genome - Assembly - NCBI. Available at: https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.1/. (Accessed: 13th May 2022)

295. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

296. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **15**, (2014).

297. GitHub - eldariont/svim-asm: Structural Variant Identification Method using Genome Assemblies. Available at: https://github.com/eldariont/svim-asm. (Accessed: 4th April 2022)

298. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).

299. Lummaa, V., Jokela, J. & Haukioja, E. Gender difference in benefits of twinning in pre-industrial humans: boys did not pay. *J. Anim. Ecol.* **70**, 739–746 (2001).

300. About the GATK Best Practices – GATK. Available at: https://gatk.broadinstitute.org/hc/en-us/articles/360035894711-About-the-GATK-Best-Practices. (Accessed: 19th April 2022)

301. GATK: the best practice for genotype calling in a non-model organism - Dmytro Kryvokhyzha. Available at: https://evodify.com/gatk-in-non-model-organism/. (Accessed: 4th April 2022)

302. Sugiura, K., Su, Y.-Q. & Eppig, J. J. Does Bone Morphogenetic Protein 6 (BMP6) Affect Female Fertility in the Mouse?1. *Biol. Reprod.* **83**, 997–1004 (2010).

303. Souza, H. C. J., McNeilly, A. S., Benavides, M. V., Melo, E. D. O. & Ferrugen, M. J. C. Novel GDF9 polymorphism determining higher ovulation rate and litter size in sheep. *Proc. Soc. Reprod. Fertil. Annu. Conf. Edinburgh* 350

304. Prasasya, R. D. & Mayo, K. E. *Regulation of Follicle Formation and Development by Ovarian Signaling Pathways*. *The Ovary* (Elsevier Inc., 2018). doi:10.1016/b978-0-12-813209-8.00002-9

305. Kaivo-Oja, N., Jeffery, L. A., Ritvos, O. & Mottershead, D. G. Smad signalling in the ovary. *Reprod. Biol. Endocrinol.* **4**, 1–13 (2006).

306. Allan, M. F. *et al.* Confirmation of quantitative trait loci using a low-density single nucleotide polymorphism map for twinning and ovulation rate on bovine chromosome 5. *J. Anim. Sci.* **87**, 46–56 (2009).

307. Gonda, M. G., Arias, J. A., Shook, G. E. & Kirkpatrick, B. W. Identification of an ovulation rate QTL in cattle on BTA14 using selective DNA pooling and interval mapping. *Anim. Genet.* **35**, 298–304 (2004).

308. He, L. C. *et al.* Identification of new single nucleotide polymorphisms affecting total number born and candidate genes related to ovulation rate in Chinese Erhualian pigs. *Anim. Genet.* **48**, 48–54 (2017).

309. Drouilhet, L. *et al.* The Highly Prolific Phenotype of Lacaune Sheep Is Associated with an Ectopic Expression of the B4GALNT2 Gene within the Ovary. *PLoS Genet.* **9**, (2013).

310. Guo, X. *et al.* Molecular cloning of the b4galnt2 gene and its single nucleotide polymorphisms association with litter size in small tail han sheep. *Animals* **8**, (2018).

311. Chong, Y., Huang, H., Liu, G., Jiang, X. & Rong, W. A single nucleotide polymorphism in the zona pellucida 3 gene is associated with the first parity litter size in Hu sheep. *Anim. Reprod. Sci.* **193**, 26–32 (2018).

Tables

Table 1.1. TGF-β Superfamily genes with known knockout and point mutations related to litter size, ovulation rate, or twinning rate across different species.

Gene	Species	Mutation Type	Phenotype	Reference
BMP4	Primate	non-synonymous mutation	Candidate increases TR	[106]
	Caprine	non-synonymous mutation	Candidate increases OR/LS	[107]
BMP6	Mouse	Knock out	Decreased OR	[302]
BMP7	Rat	Over expression	Decreased the number of ovulated Oocytes	[119]
	Porcine	Substitution	Candidate polymorphism for LS	[108]
BMP15	Ovine: $FecX^{G}$	premature stop	Increased OR in heterozygotes carriers (+/-) and sterility in homozygote (-/-)	[86,89,91,95]
	Ovine: $FecX^{O}$	non-conservative mutation	Increases OR in homozygotes and heterozygote	[86,98]
	Ovine: $FecX^{Gr}$	non-conservative mutation	Increases OR in homozygotes and heterozygote	[86,98]
	Ovine: $FecX^{B}$	non-conservative mutation	Increased OR in heterozygotes carriers and sterility in homozygote (-/-)	[89,95]

	Ovine: $FecX^{I}$	non-conservative mutation	Increased OR/LS in heterozygotes carriers and sterility in homozygote (-/-)	86,88,89,94
	Ovine: $FecX^{H}$	premature stop	Increased OR in heterozygotes carriers and sterility in homozygote (-/-)	86,89,94
	Ovine: $FecX^{L}$	non-conservative mutation	Increased OR in heterozygotes carriers and sterility in homozygote (-/-)	86,89,99
	Ovine: $FecX^{R}$	17 bp deletion: premature stop	Increased OR in heterozygotes carriers and sterility in homozygote (-/-)	86,96,97
	Ovine: $FecX^{Bar}$	Non-conservative substitution, 3 bp deletion, and frame shift insertion	Increased OR in heterozygotes carriers and sterility in homozygote (-/-)	100
	Primate	non-synonymous mutation	Candidate increases TR	106
	Mouse	Knockout	Subfertility in homozygotes; LS reduction	114
GDF9	Ovine: $FecG^{H}$	non-conservative mutation	Increased OR/LS in heterozygotes and sterility in homozygote carriers	86,89,95
	Ovine: $FecG^{T}$	non-conservative mutation	Increased OR/LS in heterozygotes and sterility in homozygote carriers	86,104
	Ovine: $FecG^{E}$	non-conservative mutation	Major increase in OR in homozygotes and minimal increase in OR in heterozygotes	86,105

	Ovine: FecG^F	Conservative mutation	Increases OR in homozygotes and heterozygote	[102,103]
	Ovine: FecG^V	Non-conservative mutation	Increase OR in heterozygous carriers	[303]
	Ovine: FecG^I	Conservative mutation	Increased OR in heterozygous carriers with lower OR in homozygous carriers	[304]
	Ovine: Novel	Synonymous mutation	Additive candidate mutation to high fecundity in Hu sheep breed	[85]
	Bovine	Nonsynonymous mutation	Candidate increase TR	[109]
	Human	Deletion/insertion (3): stop codon	Candidate for TR	[110,111]
	Human	Missense mutation (4)	Candidate for TR	[110]
	Mouse	Knockout	Infertility in homozygotes	[113,114]
BMPR1B	Ovine: FecB^B	non-synonymous mutation	- Additive increase in OR based on number of mutated alleles	[86,88–93]
	Porcine	Non-coding mutation	Candidate for LS	[112]
	Mouse	Knockout	Infertility – homozygotes (fertilization issues)	[305]
SMAD2/3	Mouse	Conditional Knockout	Reduced OR and LS in cKO of both *SMAD2* and *SMAD3*	[118]
SMAD3	Mouse	Deficient	Reduced fertility in deficient mice	[117]

SMAD6	Bovine	Over expression	Increases OR and TR in homozygotes of the allele and heterozygote carriers	[54,122,123]
ACVR2	Mouse	Knock out	FSH suppression leading to disruption of estrous cycle and increase atresia of follicles	[116]
TGFBR2	Ovine	Synonymous mutation	Additive candidate mutation to high fecundity in Hu sheep breed	[85]

Table 1.2. Documented QTLs in cattle for TR and OR based on microsatellite data from Animal QTL database or literature review.

BTA	Number of TR QTL	Total number of QTLs	QTL by location cM	Candidate gene(s)[1]	References
1	1	1	83.69-83.89		189
5	2	2	ND		188,190
	3	5	24.30-25.80		306
	2	3	29.91-33.92	*SOCS2*	253,306
	4	4	36.07-47.23		253
	7	7	56.63-90.84	*IGF1*	202,203,252–254
	0	1	98.00-126.00		251
6	1	1	26.68-28.88		189
	1	1	43.07-43.27		189
7	0	1	3.01-3.21	*WNT3A, WNT9A*	251
	2	3	27.06-31.60		189,251
	2	2	53.63-71.02	*FGF1*	189
	1	1	116.60-124.90		252
8	1	1	25.89-26.09		189
	1	1	92.70-122.90		204
10	1	1	35.07-44.25		204
12	1	1	0.00-15.11		252
13	0	1	ND		188
14	2	3	51.94-83.93		189,204,307
15	1	2	ND		188,189
19	0	1	48.00-80.00		251
21	1	1	0.00-12.60	*IGF1R*	204

23	1	1	4.70-11.83	[203]
	2	2	20.66-42.44	[203,252]
	1	1	58.19-64.37	[203]
	1	1	74.08-74.30	[189]
29	1	1	29.20-65.64	[204]

ND = not in QTL database

[1] Genes indentified using rough bp location cM/1,000,000 when possible and searched for in ARS-UCD 1.2 and NCBI release 106

Table 1.3. Cattle QTL TR and OR studies based on SNP data found within the FAANGMINE database and updated in ARS-UCD 1.2.

BTA	TR related QTL	Total related QTL	Spanning locations (MB)[1]	Candidate Gene(s)[2]	Reference
1	4	4	3.95 - 151.1		[132]
2	2	2	27.4 - 66.5		[132]
	9	9	112.9 - 113.1		[132]
	2	2	118.8 - 130.7	*SPATA3*	[132]
3	2	2	97.6 - 105.2		[132]
4	2	2	42.9 - 58.1		[132]
	2	2	103.8 - 114.5	*AKR1D1*	[132]
5	4	4	9.1 - 24.5		[132]
	7	7	64.4 - 117.6	*NR1H4*	[132]
6	4	4	12.7 - 23.8		[132]
	2	2	44.7 - 85.5		[132]
7	1	1	59.2 - 64.1		[132]
8	1	1	24.07		[132]
	5	5	52.8 - 63.6		[132]
	1	1	107.3		[132]
9	2	2	34.2 - 34.6		[132]
	2	2	81.6 - 90.3		[132]
10	1	2	13.7 - 17.1	*SMAD6*, SMAD3*, IQCH*	[193]
	1	1	65.8		[132]
11	1	1	0.5 - 0.8		[132]
	3	3	43.0 - 65.6		[132]
	3	3	93.2 - 104.8		[132]

12	5	5	48.1 - 54.6	*NDFIP2*	[132]
13	4	4	6.4 - 9.7		[132]
	1	1	38.9		[132]
14	5	5	12.0 - 20.2		[132]
	3	3	29.3 -37.5	*CYP7B1*	[132]
15	1	1	27.8	*PAFAH1B2*	[132]
	1	1	77.3		[132]
16	2	2	46.7 - 49.2		[132]
18	3	3	6.7-24.7		[132]
	2	2	42.7 - 50.5	*TGFB1**	[132]
20	2	2	22.8 - 37.1		[132]
	2	2	57.1 - 68.8		[132]
21	1	1	44.4		[132]
22	2	2	41.5 - 50.8		[132]
23	1	1	19.2		[132]
	2	2	41.9-48.3	*BMP6*, MAK, NEDD9, EDN1*	[132]
24	1	1	39.5		[198]
	1	1	62		[132]
26	5	5	5.5 - 28.2		[132]
	3	3	38.8 - 45.4		[132]
27	1	1	14.8		[132]
	2	2	22.4 - 23.33		[132]
28	2	2	36.9 - 44.8		[132]

[1] Spanning location includes multiple SNPs all near each other

[2] Genes looked up via NCBI release 106 for functions

* Member of the TGF-β superfamily

\

Table 1.4 Documented QTLs in sheep and swine for twinning rate, litter size, and ovulation rate containing genes involved or implicated in reproductive functions. Shaded rows overlap with cattle QTL regions.

Species	**Gene**	***Bos taurus* locations (BTA: start bp – end bp)[1]**	**Reference**
Ovine	*INHBB*[2]	2: 72108054 - 72113653	[87]
Porcine	*IGFBP2*	2: 104619962 - 104648632	[135]
Porcine	*ZFYVE9* [3]	3: 94012835- 94201474	[127]
Porcine	*INHBA*[2]	4: 79279914 - 79300035	[127,128]
Porcine	*GNRHR*	6: 83434759 - 83452222	[139]
Porcine	*TGFBR1*[2]	8: 64106861 - 64182858	[128,129]
Porcine, Ovine	*ESR1*	9: 88683050 - 89098471	[87,128,137]
Porcine	*SMAD6*	10: 13544183 - 13622063	[130]
Porcine	*TCF12*	10: 53023354 - 53416751	[141]
Porcine, Ovine	*ESR2*	10: 76393666 - 76460150	[87,138]
Ovine	*FSHR*	11 31255649 - 31450537	[140]
Ovine	*NCOA1*	11: 74447679 - 74666711	[87]
Porcine	*BMP7*[2]	13: 58889622 - 58975046	[127,128]
Ovine	*SMAD1*[2]	17: 12740787 - 12823219	[87]

Porcine	*VMP1*	19: 10674095 - 10803987	[308]
Ovine	*B4GALNT2 (FecL^{L})*	19: 37390785 - 37466628	[87,309,310]
Ovine	*GHR*	20: 31869704 - 32043372	[87]
Ovine	*ZP3*	25: 34446154 - 34453895	[311]
Porcine	*IGF2*	29: 49395153 - 49422469	[134]

[1] Location based on bovine assembly ARS-UCD1.2 and NCBI annotation release 106
[2] Member of the TGF-β superfamily
[3] Involved with the TGF-β superfamily directly
[4] Manual exploration of QTLs from Animal QTL Lab (https://www.animalgenome.org/cgi-bin/QTLdb/index) and FAANGMINE ()

Table 2.1. Breakdown of records, herds, animals, and sires per breed with corresponding average twinning rates.

Breeds	Herds (no.)	Animals (no.)	Sires (no.)	Total records (no.)	Average Twinning (%)
Holstein	3,074	831,579	20,317	1,806,505	-
Holstein[1]	1,748	658,436	2,223	1,440,540	4.8%
Jersey[2]	491	14,938	1,767	29,378	2.7%
Brown Swiss[2]	318	5,466	724	11,725	4.6%
Guernsey[2]	80	2,846	357	6,062	3.2%
Ayrshire[2]	95	983	258	2,151	2.0%
Milking Shorthorn	63	563	193	1,359	5.7%

[1] Number of records after removal of herds and sire with less than 100 records. Data used in the heritability analysis. All other breeds had fewer than twenty sires after similar editing were not used in heritability estimation.
[2] Number of records prior to removal of herds and sires with less than 100 records

Table 2.2. The least squares means (LSM) and standard errors (SE) for parity, season and year fixed effects.

Effect[1]	LSM	SE
Parity		
1	0.0030	0.0004
2	0.0425	0.0004
3	0.0690	0.0005
4	0.0763	0.0005
Season		
1	0.0459	0.0004
2	0.0467	0.0005
3	0.0519	0.0004
4	0.0462	0.0004
Year		
2010	0.0441	0.0008
2011	0.0438	0.0006
2012	0.0470	0.0005
2013	0.0501	0.0005
2014	0.0505	0.0005
2015	0.0486	0.0005
2016	0.0497	0.0005

[1]Effects of season, parity, and year were significant ($P < 0.0001$). Season 1, December-February; Season 2, March-May; Season 3, June-August; Season 4, September-November.

Table 3.1. Genomic 0.5 Mb windows explaining > 99.9% of the dataset's variance.

Dataset - Analysis	BTA	Window region (bp)	% Variance Explained	Positional Candidate Genes
	1	55,834,185 - 56,333,878 [Δ]	0.7992[*]	
	25	21,068,151 - 21,568,055	0.3862[*]	
	11	31,002,263 - 31,502,083 [Δ]	0.3516	*LHCGR*, *FSHR*
	21	3,678,838 - 4,178,477	0.3451	
	15	69,356,243 - 69,855,814	0.3282	
DSA[1]	24	30,219,508 - 30,718,316	0.2934	*TAF4B*
	21	22,103,608 - 22,603,573 [Δ]	0.2924	*ZSCAN2*
	29	13,194,498 - 13,694,492	0.2885	
	23	28,947,216 - 29,447,202	0.2632	
	2	56,060,044 - 56,559,883	0.2617	
	4	3,691,381 - 4,191,265	0.2537	
	11	31,002,263 - 31,502,083 [Δ]	0.7520[*]	*LHCGR*, *FSHR*
	10	27,530,261 - 28,030,117	0.4096	
	18	8,071,002 - 8,570,969	0.3227	
	6	9,607,396 - 10,107,227	0.3125	
	4	113,464,343 - 113,964,263	0.3053	
DSB[2]	24	32,596,784 - 33,096,765	0.2926	
	21	22,103,608 - 22,603,573 [Δ]	0.2757	*ZSCAN2*
	1	55,834,185 - 56,333,878 [Δ]	0.2718	
	14	24,526,962 - 25,026,602	0.2489	
	15	8,959,561 - 9,458,927	0.2466	
	6	8,652,990 - 9,152,983	0.2433	

[1] DSA = dataset A calving records from 2010-2016
[2] DSB = dataset B calving records from 1994-2008
[Δ] Windows overlap between the two datasets
[*] Explained > 99.99% of the dataset's variance

Table 3.2. Most significant SNPs by chromosome from single-step genome-wide association study using whole genome sequence.

BTA	Location (bp)	p-value[1]	FDR	p-value DSA[1]	SNP effect[4] DSA[2]	p-value DSB[2]	SNP effect[4] DSB[3]	Minor allele frequency
2	104826747	4.78×10^{-7}	4.89×10^{-3}	1.20×10^{-4}	-4.65×10^{-7}	5.08×10^{-4}	-4.39×10^{-7}	0.395
5	106037831	1.16×10^{-7}	2.20×10^{-3}	4.06×10^{-4}	4.36×10^{-7}	3.45×10^{-5}	5.28×10^{-7}	0.488
6	25264112	3.29×10^{-7}	4.01×10^{-3}	7.80×10^{-6}	-2.66×10^{-7}	4.46×10^{-3}	-1.79×10^{-7}	0.030
7	44424467	1.36×10^{-7}	2.39×10^{-3}	1.44×10^{-5}	-3.05×10^{-7}	1.31×10^{-3}	-2.23×10^{-7}	0.08
9	39650984	4.28×10^{-8}	1.66×10^{-3}	6.41×10^{-7}	3.81×10^{-7}	3.45×10^{-3}	2.27×10^{-7}	0.103
11*	29977957	5.99×10^{-12}	4.79×10^{-5}	1.43×10^{-6}	-5.65×10^{-7}	4.63×10^{-7}	-6.01×10^{-7}	0.349
11	86316394	6.38×10^{-8}	1.89×10^{-3}	3.73×10^{-5}	-4.96×10^{-7}	2.04×10^{-4}	-4.47×10^{-7}	0.348
11	40142237	$2.01x 10^{-7}$	2.94×10^{-3}	1.68×10^{-4}	-3.30×10^{-7}	1.70×10^{-4}	-3.37×10^{-7}	0.154
11	24578435	3.03×10^{-7}	3.82×10^{-3}	5.75×10^{-5}	-3.93×10^{-7}	7.54×10^{-4}	-3.45×10^{-7}	0.211
14	25256056	1.19×10^{-6}	8.50×10^{-3}	2.40×10^{-3}	-3.71×10^{-7}	6.13×10^{-5}	-5.21×10^{-7}	0.394
15	79242162	4.53×10^{-7}	4.82×10^{-3}	1.48×10^{-6}	6.82×10^{-7}	1.99×10^{-2}	3.37×10^{-7}	0.216
19	48738300	4.66×10^{-9}	1.01×10^{-3}	4.63×10^{-6}	4.01×10^{-7}	1.30×10^{-4}	3.35×10^{-7}	0.129
19	30859152	9.17×10^{-9}	1.32×10^{-3}	4.99×10^{-6}	-4.80×10^{-7}	1.93×10^{-4}	-3.90×10^{-7}	0.245
25	20772567	5.10×10^{-8}	1.73×10^{-3}	1.20×10^{-5}	-4.70×10^{-7}	5.94×10^{-4}	-3.66×10^{-7}	0.253
25	29826449	2.82×10^{-7}	3.64×10^{-3}	7.78×10^{-6}	-3.88×10^{-7}	4.10×10^{-3}	-2.34×10^{-7}	0.114
X	88629215	1.25×10^{-6}	8.86×10^{-3}	2.47×10^{-3}	-5.56×10^{-7}	7.11×10^{-5}	-7.36×10^{-7}	0.434

[1] P-value combined across datasets using a weighted z transformation method via combine.test from package survcomp v1.38.0[228,229]

[2] DSA = dataset A calving records from 2010-2016

[3] DSB = dataset B calving records from 1994-2008

[4] SNP effects ($\hat{u}$) are estimated from GEBV as follows: $\hat{u} = q\mathbf{DZ}'(\mathbf{ZDZ}'q)^{-1}\hat{a}$, where q is weighting factor, **D** is a diagonal weight matrix for SNP, **Z** is a matrix of gene content adjusted for allele frequency (0, 1, 2, for AA, AB, and BB respectively), and $\hat{a}$ is the GEBVs of genotyped animals[206]

* Indicates most significantly associated SNP out 7,994,662

Table 3.3. Pathways identified as overlapping between newer calving records (DSA) and older (DSB) when using the whole genome sequence results from single-step GWAS.

Pathway Identification	Description	Database	Size	Order by P-value[1] (DSA[2])	Order by P-value[1] (DSB[3])
GO:0022412	cellular process involved in reproduction in multicellular organism	GOBP	135	1[4]	2[4]
GO:0032301	MutSalpha complex	GOCC	2	2[4]	4[4]
GO:0016607	nuclear speck	GOCC	124	3	7
bta04150	mTOR signaling pathway	KEGG	152	4	6
WP1069	Integrin-mediated Cell Adhesion	WikiPathways	86	5	5
user5	prolctin 280991	BL_TR	81	6	1[*]
R-BTA-6805567	Keratinization	REACTOME	108	7	3[*]

[1] Ordered first by the P-value of the set rank then the adjusted P-value and lastly the corrected P-value
[2] DSA = dataset A calving records from 2010-2016
[3] DSB = dataset B calving records from 1994-2008
[4] Value of high interest found and found in both datasets
[*] Value was considered highly of interest based on P-value of set rank < 0.05 OR the adjusted and corrected P-values were ≤ 0.001[240].

Table 3.4. Assessment of accuracy and bias of genomic prediction based on correlation and regression analysis of daughter average and genomic breeding value.

	DSA[1] testing, DSB[2] training LD		DSA[1] testing, DSB[2] training HD	
Iteration	r[3] (SE)	β[4] (SE)	r[3] (SE)	β[4] (SE)
1	0.4235 (0.025)	0.7171 (0.042)	0.4272 (0.025)	0.7303 (0.042)
2	0.4242 (0.025)	0.6958 (0.041)	0.4283 (0.025)	0.7075 (0.041)
3	0.4244 (0.025)	0.6934 (0.040)	0.4286 (0.025)	0.7052 (0.041)
4	-	-	0.4287 (0.025)	0.7048 (0.041)

[1]DSA, dataset A, sires with daughters having calving records from years 2010-2016.
[2]DSB, dataset B, sires with daughters having calving records from years 1994-2008.
[3]Correlation of phenotype (daughter average) and GEBV
[4]Slope from regression of phenotype on GEBV.

Table 4.1. Read depth for sequenced DNA samples.

Animal ID	Read depth
JEUSA000000660226	16.18
JEUSA000000654500	13.57
JEUSA000000660675	12.55
JEUSA000000661399	15.56
JEUSA000000662737	12.91
JEUSA000000664195	14.04
JEUSA000067011433	13.71
JEUSA000067282568	14.69
JEUSA000110106571	16.49
JEUSA000110379366	13.19
JEUSA000110641243	15.90
JEUSA000110874946	13.59
JEUSA000110980032	13.29
JEUSA000111142055	15.56
JEUSA000113076851	13.24
JEUSA000116279413	14.55
JEUSA000111080315	14.59
JEUSA000113503201	15.80
JEUSA000111023978	16.61
JEUSA000113672851	15.46
MARC_839802	15.04
MARC_029661	17.18
MARC_039610	17.31
MARC_988688	14.64
C041	46.93

Table 4.2. Average number of Copy Number Variants detected by the methods and consensus. As well as the average and median size (bp), and average number of deletions and duplications per method and consensus.

Method	Average number detected	Average size (bp)	Median Size (bp)	Average number of deletions	Average number of duplications
CNVnator	3,686	12,737	4,000	2,843	843
Delly	9,783	1,247,920	2,741	7,133	2,650
Lumpy – Single sample (LS)	5,430	123,390	660	4,936	494
Lumpy – Populations (LP)	18,442*	280,760	787	15,235*	3,207*
Consensus	296	2,828	1,287	284	11

* Results are exact, not averages

Table 4.3. Primers designed for PCR-assay genotyping

			Amplicon Size			Primer Sequences
BTA	CNV start (bp)	Tm[1] (°C)	F&R (normal template)	F&R (deletion template)	F&D (normal template)	Forward (F) / 1st Reverse (R) / 2nd Reverse (D)
1	21,845,290	62	849	423	533	AGGCTTCACATGGATTACCTC / GAGGAGGAAGTGGCAATCTG / GAAATAGGGAAGCAGCCAGG
5	75,404,077	59	1341	296	562	GAATCCTGAGACCCAAGTCC / AGGAGAGGGGAAACAACCTA / GAGTCCTTGCCCTTAACCAC
18	65,442,186	58	798	355	514	GACCCCCAGTTTGATAGAACA / GGTGTCGTCTGCTAGATTGG / GGCTAGAGAAAGGCTCGATG
23	26,626,756	58	6905	486	575	ACCCTAGCATTCCCAGACAT / GGCCTCAAGAAATCCACCAT / TTGGAGAGTGCTGATGGTTG

[1] Empirically determined optimal PCR annealing temperature.

Table 4.4. Details on the four embryo lethal candidates PCR genotyped in 96 Jersey cows.

BTA	Predicted Start	Predicted Stop	Predicted size	Corrected Start	Corrected Stop	Corrected Size	Number of homozygotes/number genotyped
1	21,845,287	21,845,863	576	21,845,290	21,845,715	426	4/96
5	75,404,076	75,405,121	1,045	75,404,076	75,405,121	1,045	4/96
18	65,442,115	65,442,609	494	65,442,186	65,442,628	445	9/96
23	26,626,756	26,633,190	6,434	26,626,756	26,633,174	6,419	6/96

Table 5.1. Haplotypes in the positional candidate region for Trio, offspring, and mates [a,b,c].

Animal	Genotype	Haplotype A	Haplotype B
TRIO	Carrier	22111221112211112212121121 21	22211221121112122221221111121
FR0015	Dam (U019)	12221122112211122221222212212	12111221122111111121212211112
S070	Dam (U020)	122211221122111112121221121	122211221122111112121221121
		Paternally inherited haplotype	Maternally inherited haplotype
U004	Carrier	22111221112211112212121121 21	22111221112211112212121121 12
U019	Carrier	22111221112211112212121121 21	12221122112211122221222212212
YW059	Carrier	22111221112211112212121121 21	22211221121112122221222111112
OW064	Non-Carrier	22211221121112122221221111121	122211221122111112121221121
U020	Non-Carrier	22211221121112122221221111121	122211221122111112121221121
YW020	Non-Carrier	22211221121112122221221111121	22211112112122122221221112212

[a] 1 corresponds to Illumina 50K genotype chip allele A; 2 corresponds to Illumina 50K genotype chip allele B.
[b] Shaded gray entries are the haplotype containing the Trio allele
[c] Starting SNP, rs29021659 at BTA10:13,644,156 (UMD3.1 location BTA10:13,606,664); ending SNP, rs41613055 at BTA10:14,747,084 (UMD3.1 location 10:14,720,037).

Table 5.2. Breed results for individuals from the Bovine HapMap data with the at least one copy of the haplotype containing the Trio allele after using fastPhase on the merged data of nine members of the Trio family and the bovine HapMap data.

Breed	**Total num. samples of breed**	**Total num. of breed with Trio haplotype**	**Frequency of Trio haplotype in Breed**	**Percentage of samples with Trio haplotype within breed (%)**
Beefalo	1	1	0.500	100
Hereford	20	9	0.300	45
Belgian Blue	4	1	0.125	25
Beefmaster	20	4	0.100	20
Normande	20	4	0.100	20
Charolais	20	4	0.100	20
White Park	5	1	0.100	20
Finnish Ayrshire	18	2	0.056	11
Maine-Anjou	20	1	0.025	5

Table 5.3. The variants (SNPs, InDels, and structural variants) detected in C041 and C069 pre-quality control filtering, post, narrowing to homozygous only, the resulting overlap, overlap in the positional candidate region (BTA10:13629354-14817470) , and final filtering against the 1000 Bull Genomes project data.

	SNP		InDels		Structural Variants	
	C041	C069	C041	C069	C041	C069
Initial detection (BTA10)	269,360	6,030	36,088	1,600	558	816
Quality control filtering	205,470	3,247	35,145	1,259	NA[a]	NA[a]
Homozygous Alternative	93,155	2,611	15,387	948	125	816
1.2 Mb candidate regions	739	214	149	113	0	29
Comparison 1,000 Bull Genomes Run7	11	8	20	91	NA[b]	NA[b]
Overlap	1		0		0	

[a] Quality control filter was applied during the software implementation prior to narrowing to BTA10

[b] Comparison with 1,000 Bull Genomes Run7 was not possible do to no corresponding structural variant information

Table 5.4. SNP (10:g.13828552A>G) genotyping results.

	Trio descendants inferred genotype		Non-Trio relation populations	
Genotype	Carrier	Non-carrier	Hereford	MARC Twinner[1]
GG	0	0	0	0
AG	40	0	0	0
AA	0	45	98	78

[1] USDA Meat Animal Research Center twinning population

Figures

Figure 1.1. Folliculogenesis of a primary follicle till preovulatory stage or atretic. Cycle consists of FSH dependent cohort (A), recruitment (B), selection (C), dominant follicle (D), and either atretic follicle (E) or preovulatory (F). The deviation is the time period of drastic change between the two largest follicles (green dashed line) and leads to the selection of the dominant follicle.

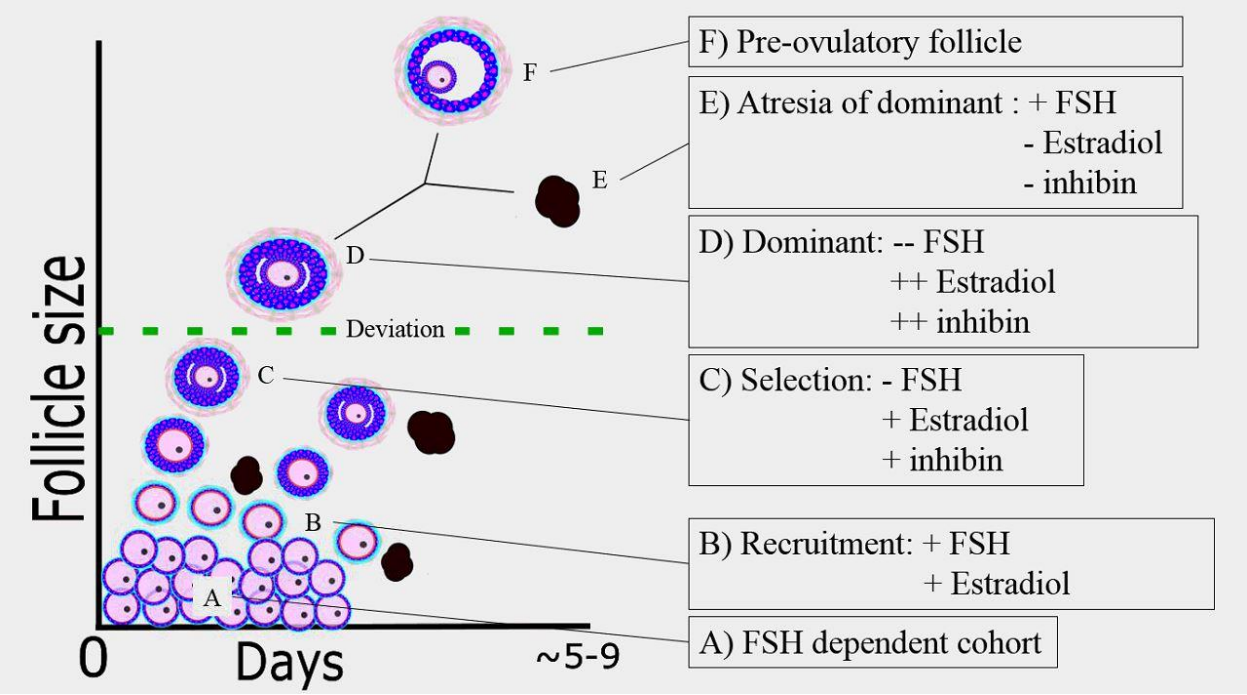


Figure 1.2. The 3-wave cycle in cattle. The blue line indicates progesterone levels, green arrows FSH surges, and the black arrow the LH surge.

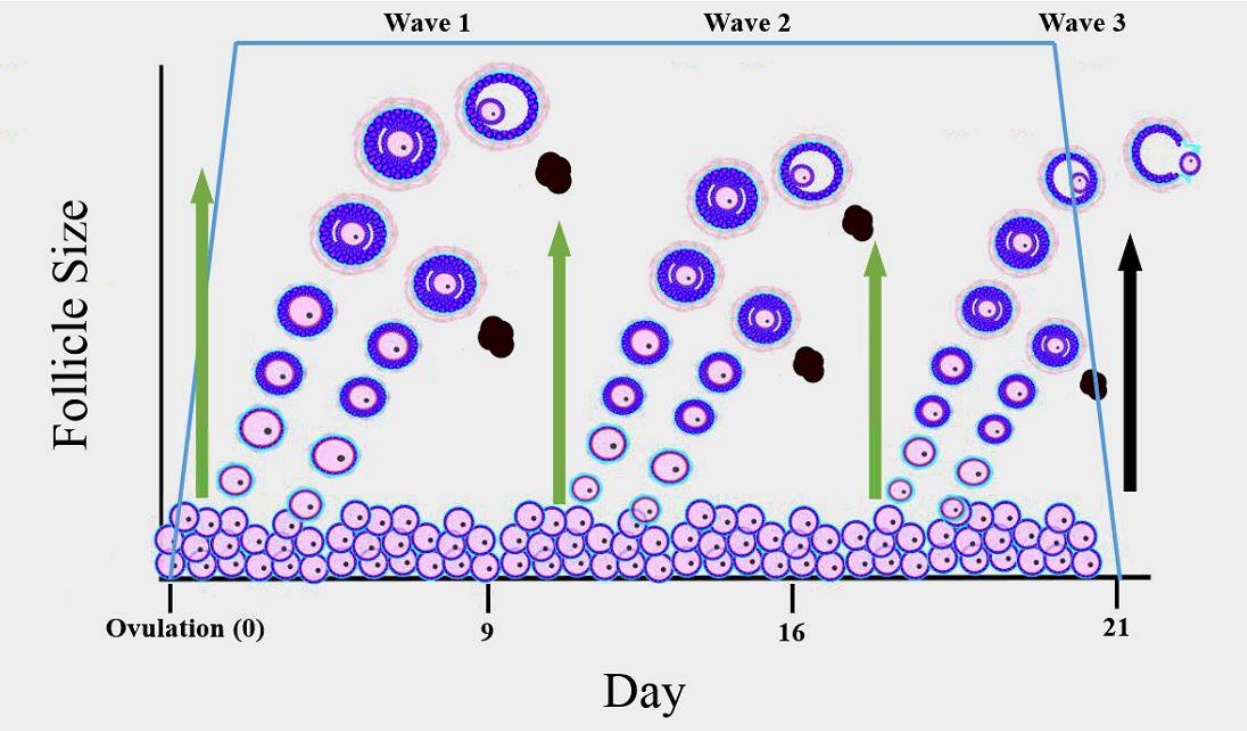


Figure 2.1. (A) Distribution of daughters per sire plotted on a log10 scale. Bars in gray represent sires that were kept after removing sires with fewer than 100 daughter records. (B) Distribution of records per herd plotted on a log scale. Bars in gray represent herds that were kept after removing those with fewer than 100 records. (C) Distribution of the number of calving records by parity for the data set before exclusion for low daughters per sire (<100) or low records per herd (<100) (n = 1,806,505).

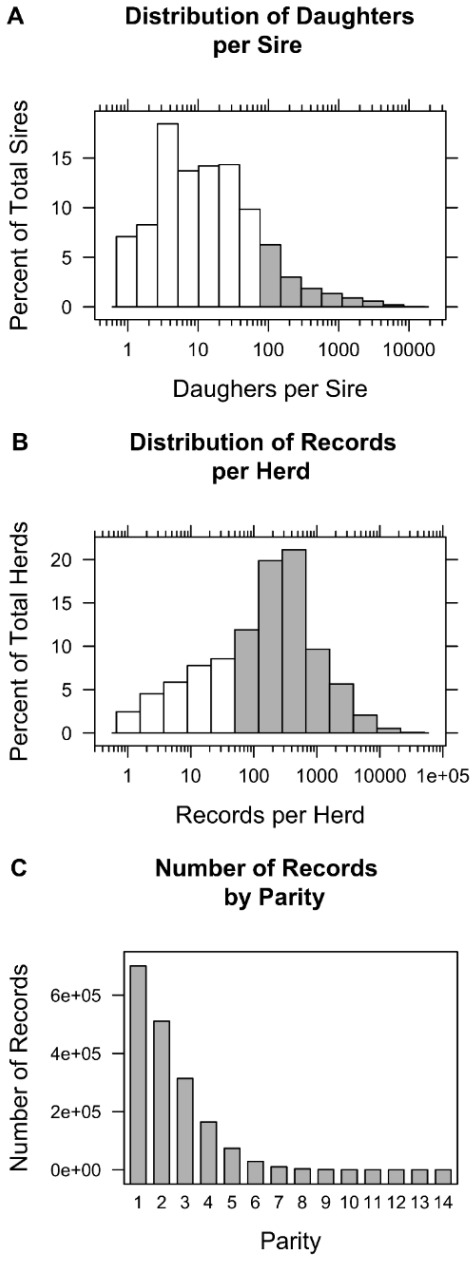


Figure 3.1. Results from single-step GWAS using whole genome sequence data. SNP significance, as -$\log_{10}$ of the P-value, is shown on the y-axis and SNP genomic locations are shown on the x-axis. A) Manhattan plot of -$\log_{10}$(p-value) from DSA, B) Manhattan plot of the -$\log_{10}$(p-value) from DSB, C) Manhattan plot of the -$\log_{10}$(weight z-transformed combined p-values) of DSA and DSB. The top red line indicates false discovery rate (FDR) threshold of 0.001 and lower blue line indicates FDR threshold of 0.01.

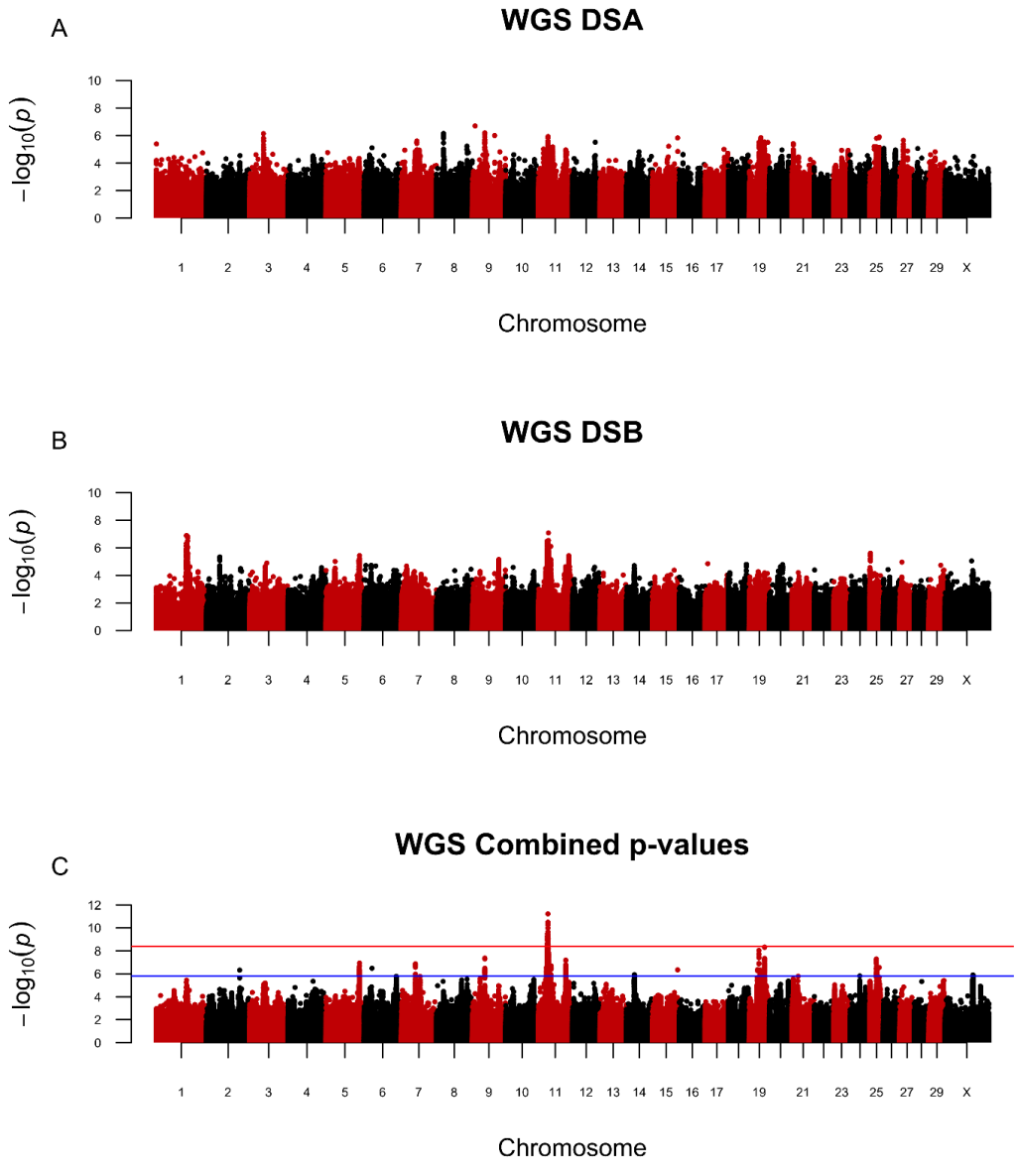


Figure 3.2. Variance attributable to 500-kb overlapping windows variance from single-step GWAS using whole-genome sequence data. The % variance explained is shown on the y-axis and SNP genomic locations are shown on the x-axis. The Blue horizontal line represents the threshold for variance falling in the top 0.001% (99.9th percentile) and the red horizontal line represents the threshold for variance falling in the top 0.0001% (99.99th percentile). A) Results from dataset A calving records, 2010-2016 **(DSA)**. B) Results from dataset B calving records. 1994-2008 **(DSB)**.

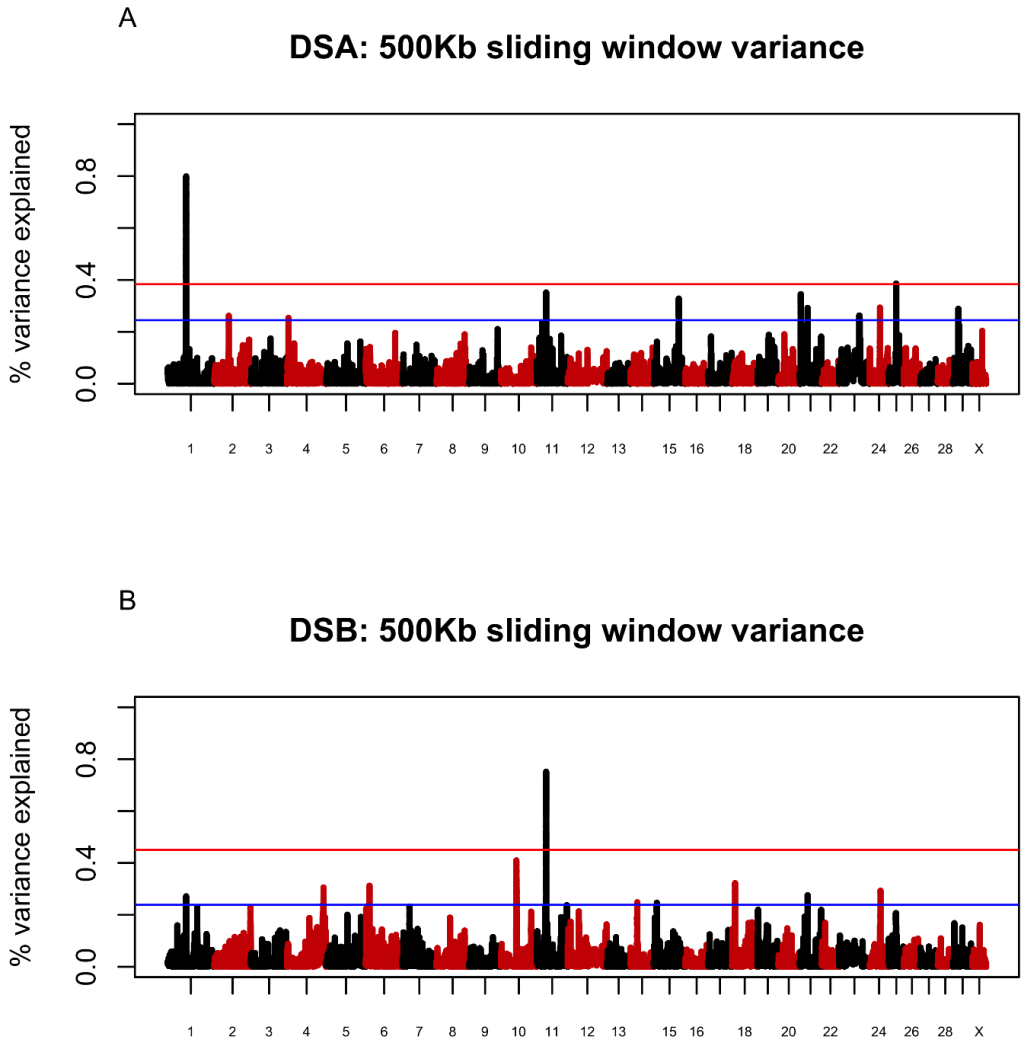


Figure 3.3. Results highlighting the two strongest peaks on BTA11 from the combined p-values of single-step GWAS using whole genome sequence data. SNP significance, as $-\log_{10}$ of the P-value, is shown on the y-axis and SNP genomic locations are shown on the x-axis. Horizontal lines denote false discovery rate (FDR) < 0.01 (blue) and FDR < 0.001 (red) association thresholds. A) Manhattan plot of associations for BTA11 variants, and B) positional candidate gene region (28 to 32 Mb) of BTA11. Variants highlighted by squares fall within the two candidate genes *LHCGR* and *FSHR*. Those colored pink highlight *LHCGR* and those colored purple highlight *FSHR*.

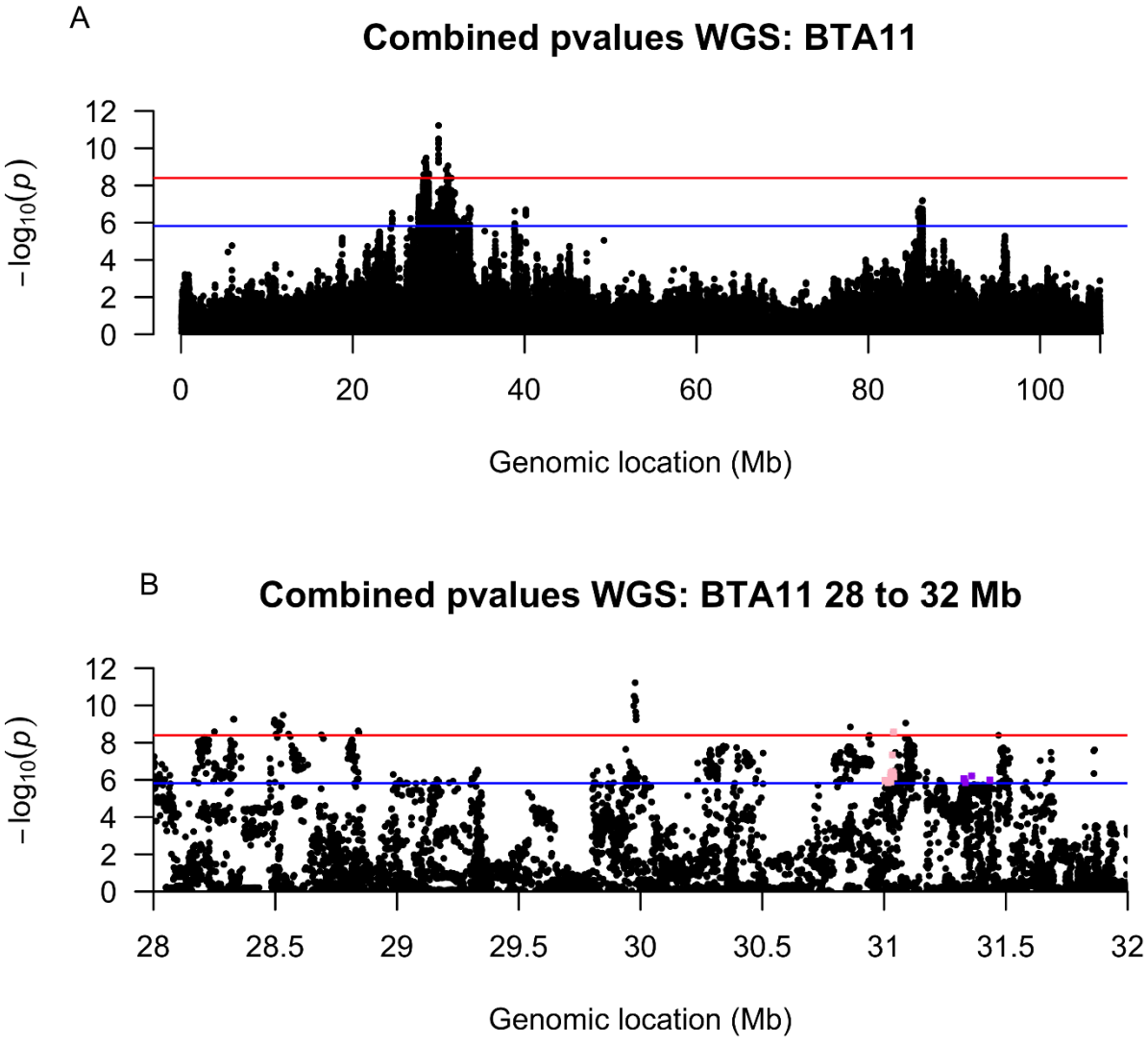


Figure 4.1. Stacked Venn diagram of the variants used to test validation of the consensus method starting from the 740 CNVs found in Jersey cattle and narrowing based on overlapping the genomic region encompassing a predicted or characterized gene, the genomic region of a characterized gene, functional gene regions (coding sequence, exon, 5' or 3' UTR, start or stop codon, and/or processed transcripts as specified), and finally those validated by literature review, database, or PCR.

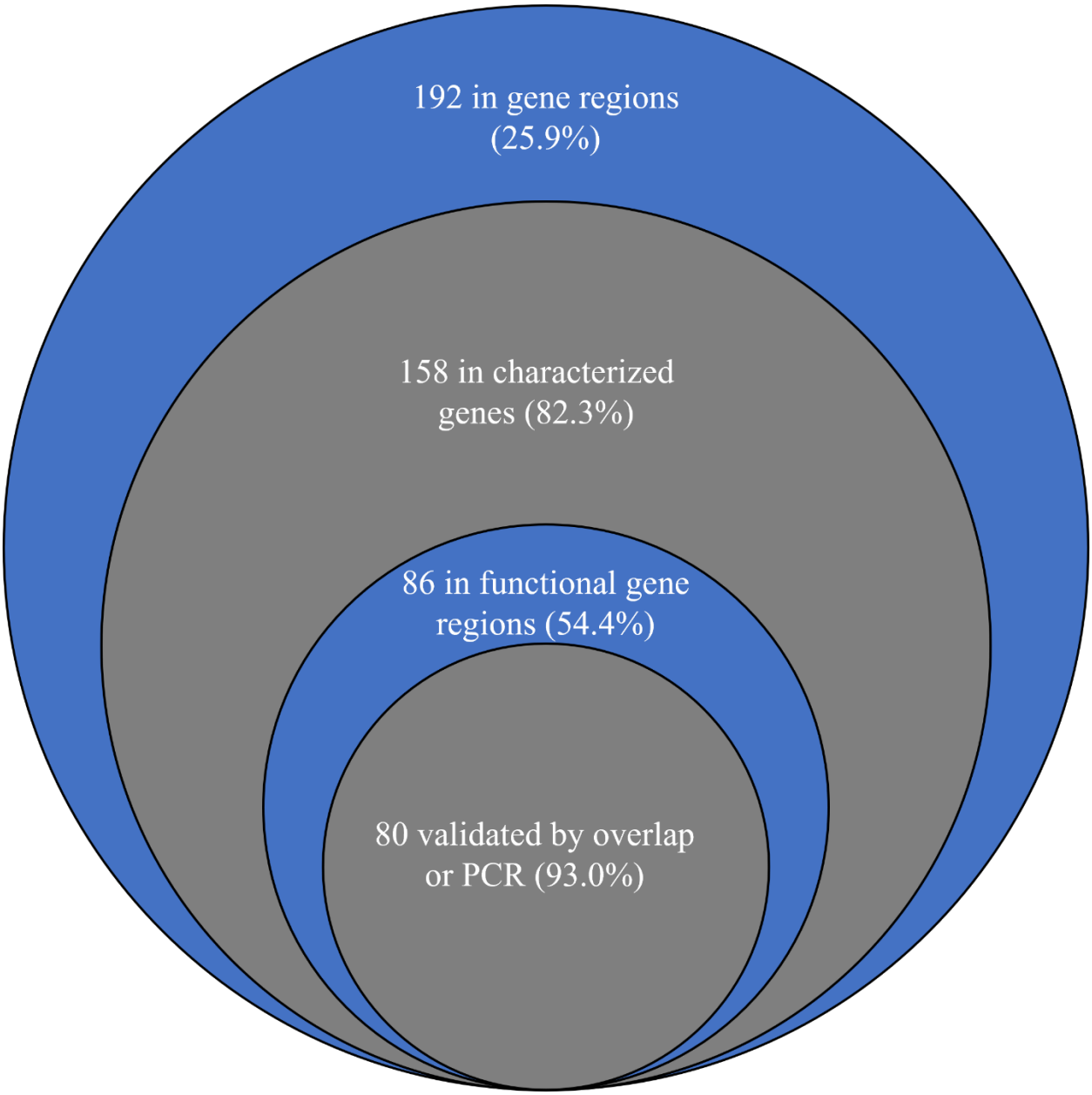


Figure 4.2. Plot views of one of the four deletions subjected to further testing for embryo lethality. A) Is a screenshot from Geneious prime viewer of the Sanger sequencing results aligned to bovine reference ARS-UCD 1.2 for BTA18. This shows the new breakpoints for the deletion indicated by the black bar. B) Is a screenshot from Golden Helix GenomeBrowse visualization tool (Version 3.0.0) of the same region on BTA18 in reference ARS-UCD 1.2. It displays three different bulls sequence coverage and sequence pile up. These bulls correspond to a heterozygote, normal, and homozygote. The heterozygote and normal bull are Jersey while the homozygous bull is a MARC twinner. Black lines indicate where the deletion is located based on Sanger sequence data.

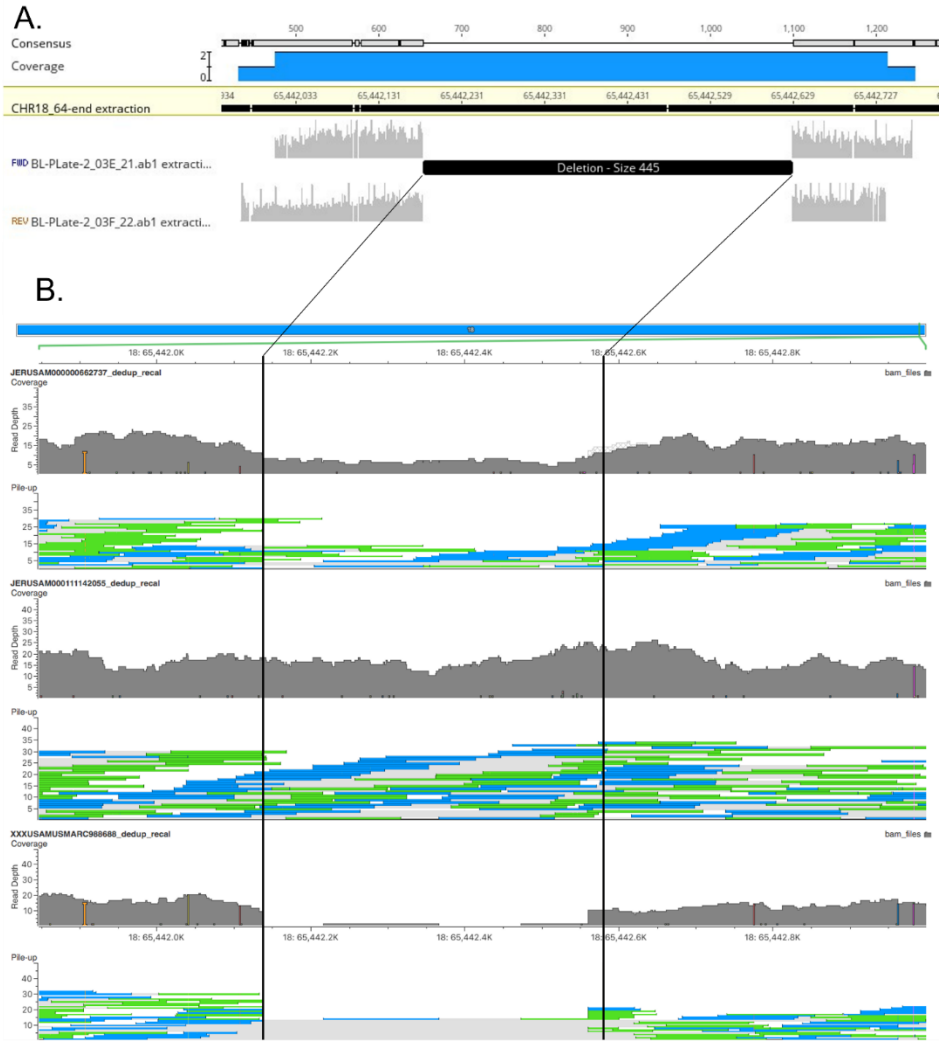


Figure 5.1. Extent of homozygosity on BTA10 including the positional candidate gene region for five individuals homozygous for the Trio allele (sequencing candidates). Genotypes were generated from the BovineHD SNP chip (Illumina, San Diego, CA). Shaded blue/gray area denotes the positional candidate region and solid black lines indicate regions where genotypes were homozygous for each individual.

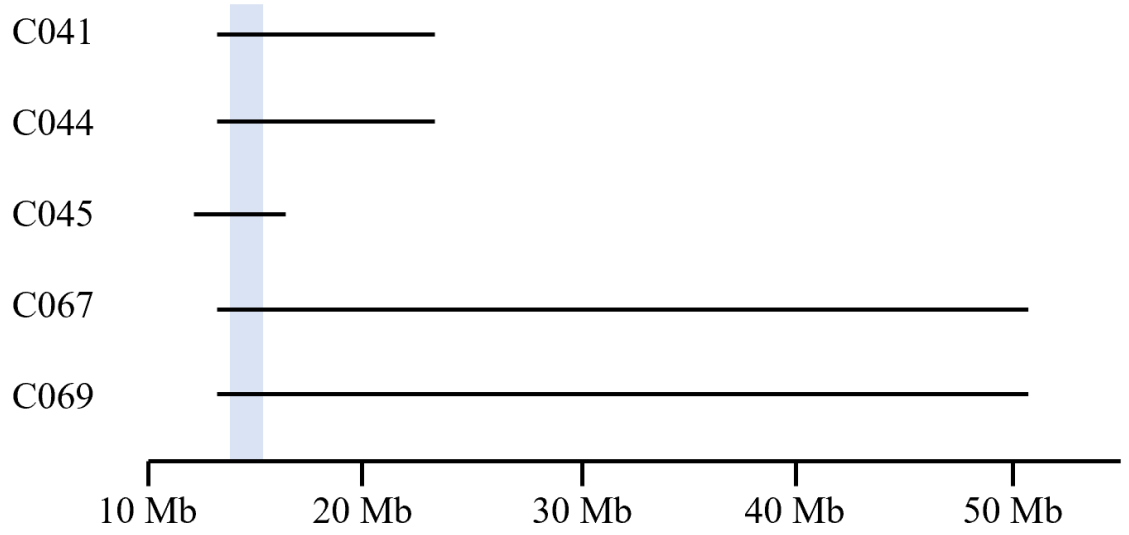


Figure 5.2. Golden Helix GenomeBrowse image of C069 and C041 alignments to ARS-UCD1.2 from 10:13,447,197 to 14,899,170. In coverage view green indicates reverse orientation reads and blue indicates forward oriented reads. C069 only has one read depth being based on the alignment of a FASTA file of a single contig making up this region. The thick single black lines denote the start (13,629,354) and stop (14,817,470) of the 1.2 Mb region of interest and the black box indicates the region for which the haplotype SNPs are located with the vertical lines at the start and stop SNPs. The novel SNP in this region (10:g.13828552A>G) is indicated by the purple line.

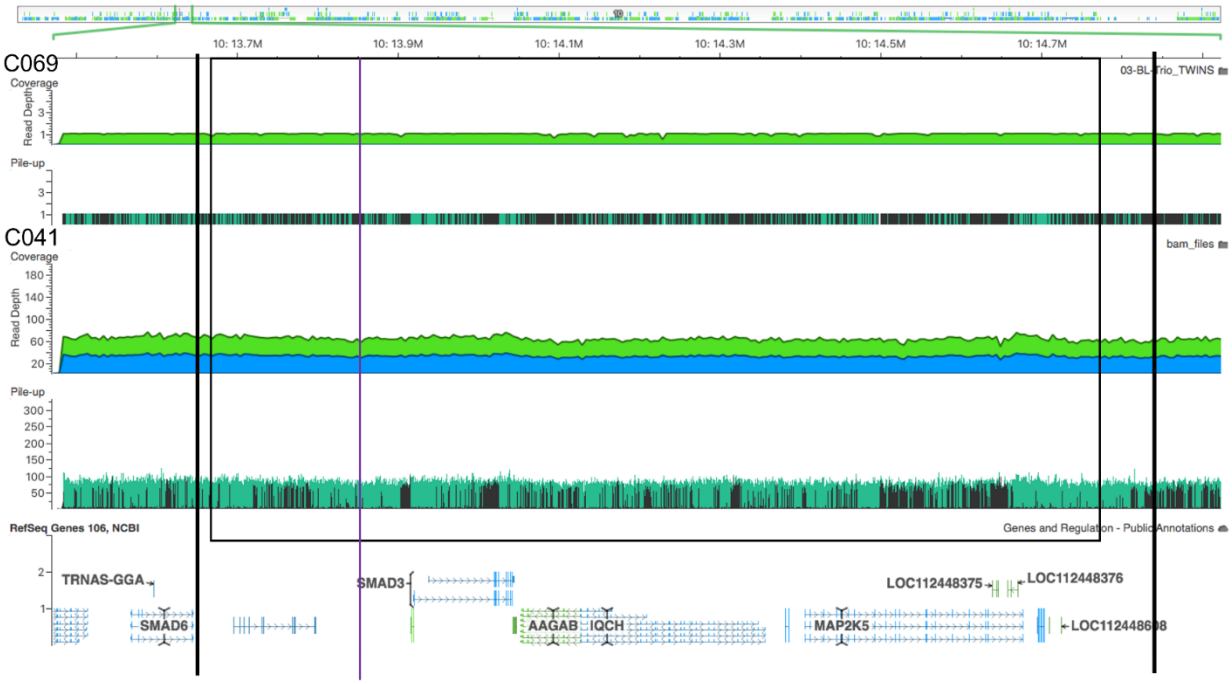


Figure 6.1. The different blocks present different DNA chunks within a chromosome. Blue blocks represent the father, purple the mom, and the combination of colors their offspring. A) Represents no missing blocks – no deletion, B) the father has a missing block that is inherited – single copy of the deletion, and in C) both parents have a block missing that is inherited by the offspring – double copy of the deletion.

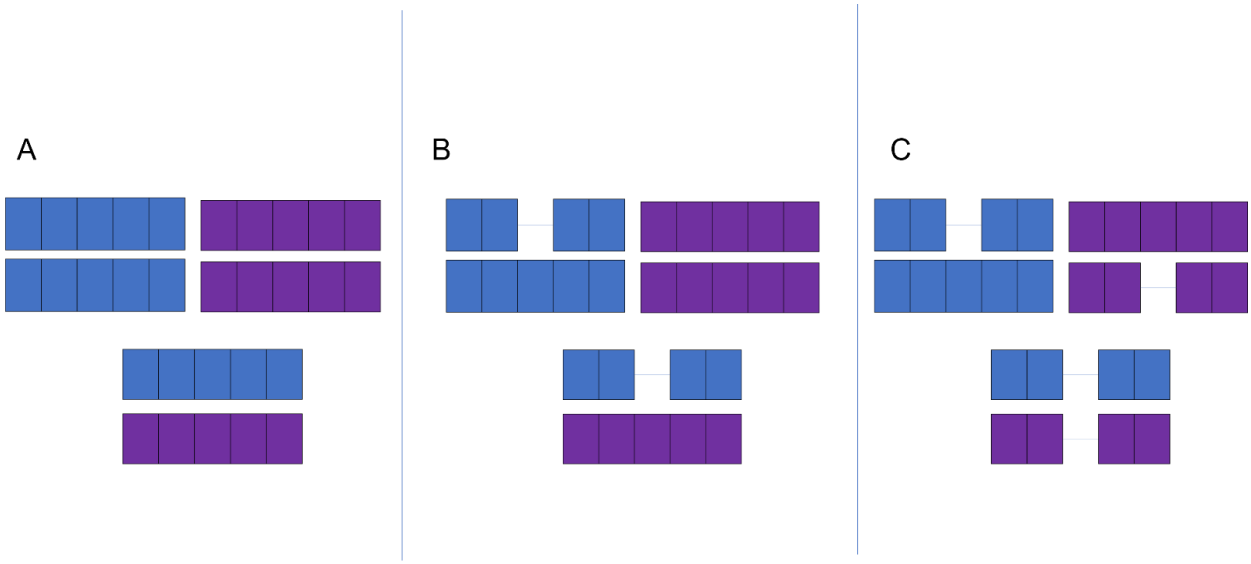


Figure 6.2. Depicts a sequence alignment to a reference as puzzle pieces. In part A, there is a reference image (left) and different reads from a sample individual at 1x depth (right). Part B is those reads aligned to the reference map. It highlights two features and a challenge from dealing with short read alignments: repeat regions where reads are ambiguous; duplication where a sequence appears more than once; and a deletion where the sequence does not appear at all when compared to the reference.

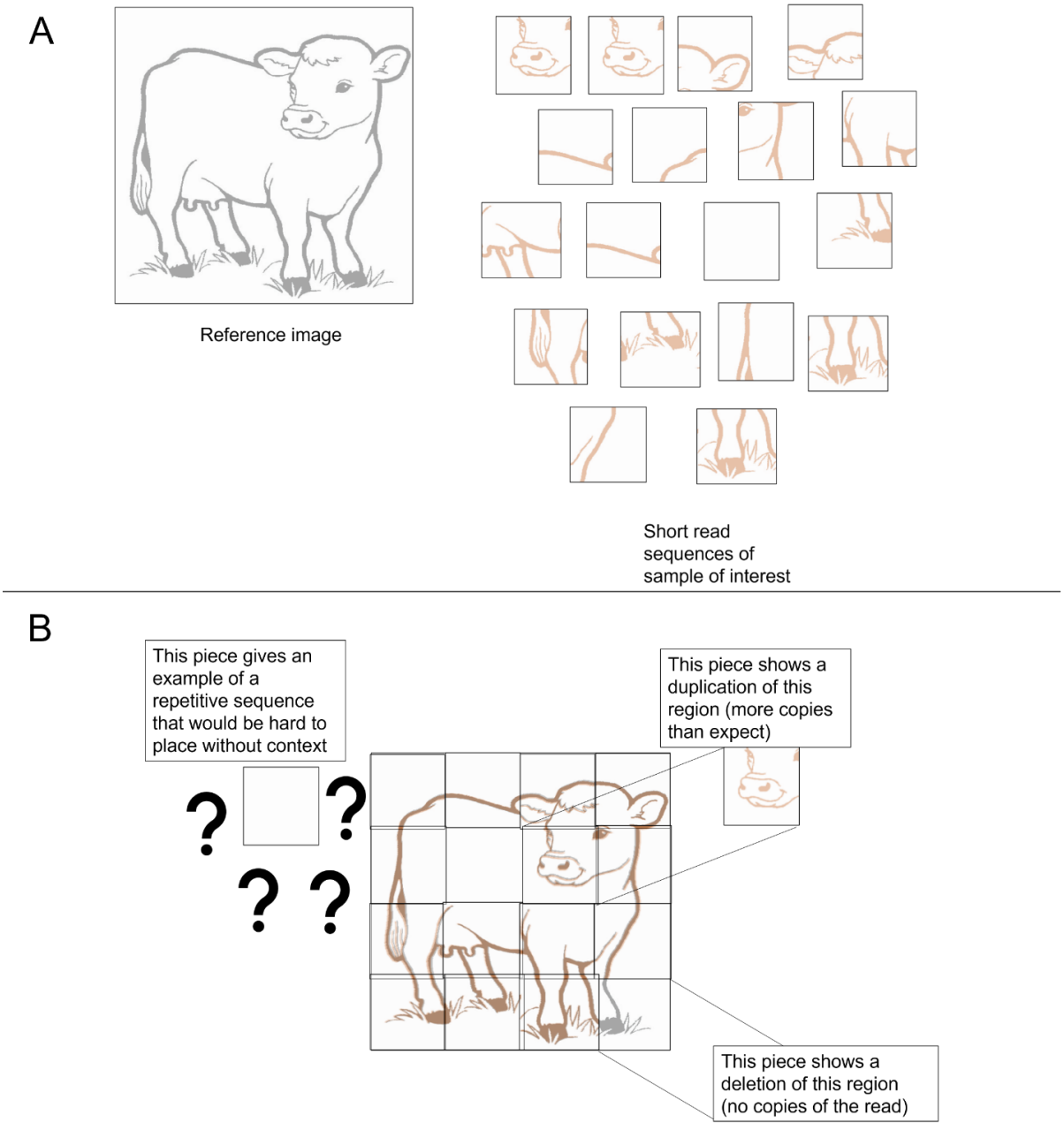


Figure 6.3. Depiction of genotyping deletions using three-primer assay and corresponding genotypes with blue segments representing chromosome segments inherited from father and purple from the mother.

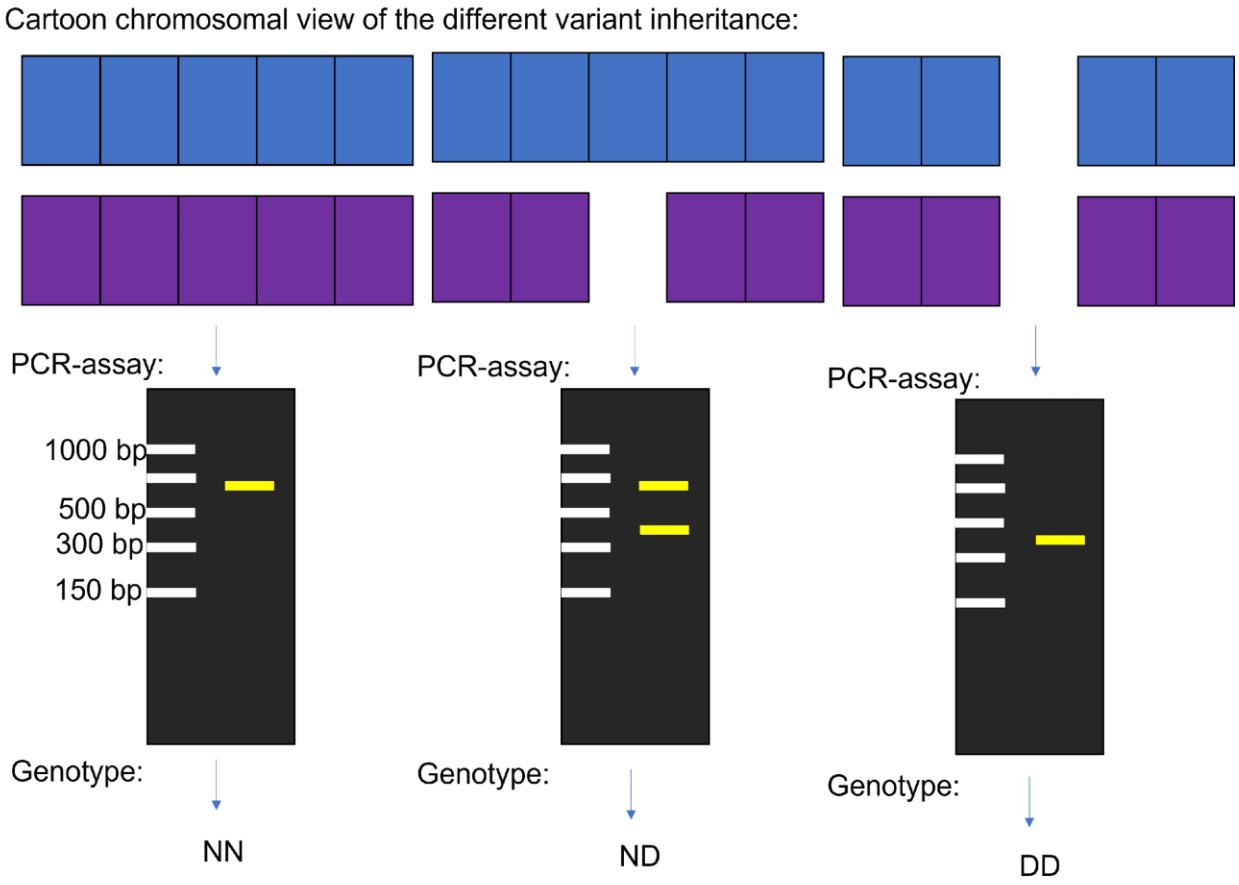
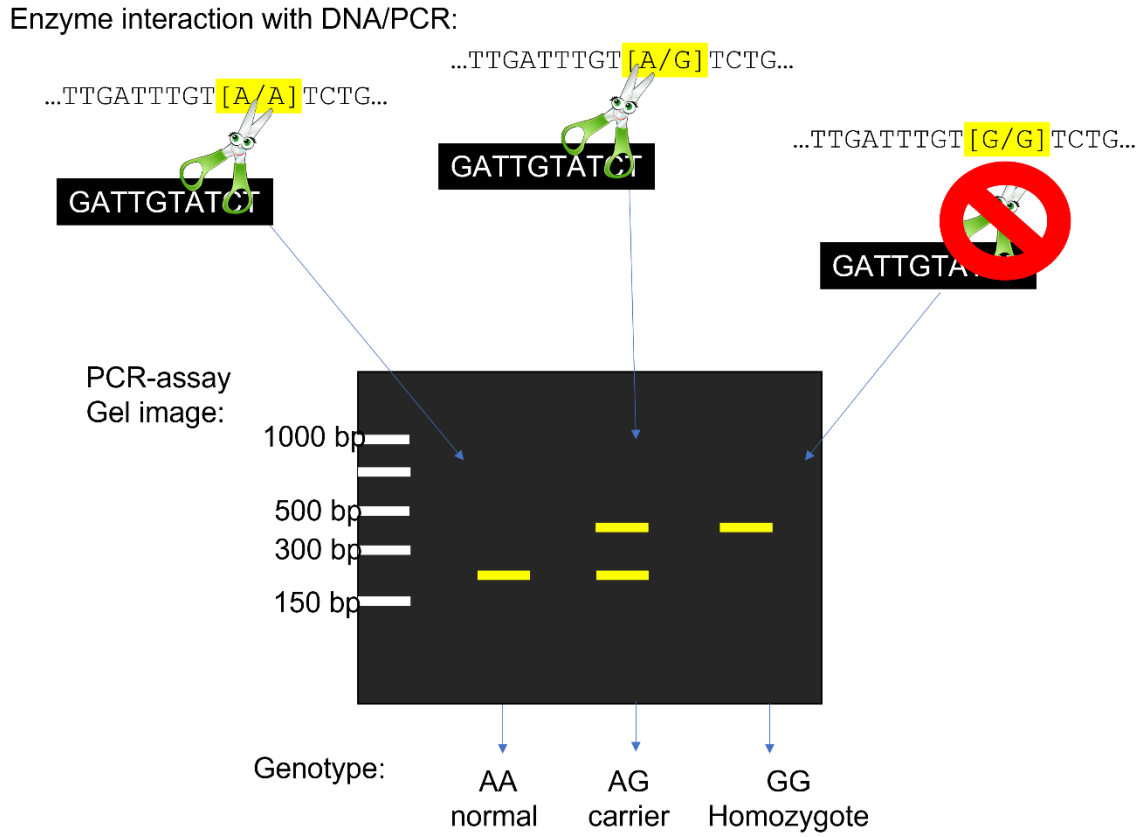


Figure 6.4. Diagram explaining how the restriction digest PCR (RFLP-PCR) assay interacts with the target region containing the SNP of interest. The enzyme (represented as scissors) only cuts where the sequence matches its target site (GATTGTATCT) thus if the variant is nucleotide G it will not cut. Next the results are displayed on a gel where bands correspond to the different alleles (~400 bp = G and ~200 bp = A). Lastly is the resulting genotype calls followed by their characterization to the Trio allele. Highlighted region in the sequence is the SNP of interest.



Apendix

Supplmentary table S3.1: Information regarding the user defined pathways included during the pathway analysis. These are pathways of interest owning to connection with folliculogenesis or the candidate genes of interest.

Pathway Identification	Number of Genes	Description	Source
User 1	53	Ovarian Steroidogenesis – bta04913	BovineMine V 1.6
User 2	54	User 1 plus the gene *GDF9*	
User 3	14	*GDF9* and *BMP15* pathways	Literature review
User 4	66	User 1 and user 3 (without duplicates)	
User 5	81	Prolactin signaling pathway	BovineMine V 1.6

Supplmentary table S3.2 Unpublished results of genomic prediction testing using 5-fold cross valiations.

	Validation method – LD				**Validation method – HD**			
	DSA[1] 5-fold cross-validation		DSB[2] 5-fold cross-validation		DSA[1] 5-fold cross-validation		DSB[2] 5-fold cross-validation	
Iteration	$r^{3,5}$	$\beta^{4,5}$	$r^{3,5}$	$\beta^{4,5}$	$r^{3,5}$	$\beta^{4,5}$	$r^{3,5}$	$\beta^{4,5}$
1	0.555 (0.043)	0.969 (0.074)	0.396 (0.038)	0.950 (0.091)	0.550 (0.043)	0.970 (0.075)	0.395 (0.038)	0.952 (0.091)
2	0.557 (0.043)	0.952 (0.073)	0.398 (0.038)	0.934 (0.089)	0.552 (0.043)	0.953 (0.074)	0.396 (0.038)	0.934 (0.089)
3	0.557 (0.043)	0.950 (0.072)	0.399 (0.038)	0.933 (0.088)	0.552 (0.043)	0.951 (0.073)	0.397 (0.038)	0.933 (0.089)
4	-	-	-	-	-	-	0.373 (0.038)	0.884 (0.091)

[1]DSA, dataset A, sires with daughters having calving records from years 2010-2016.
[2]DSB, dataset B, sires with daughters having calving records from years 1994-2008.
[3]Correlation of phenotype (daughter average) and GEBV
[4]Slope from regression of phenotype on GEBV.
[5]Average across the five testing subsets

Supplmentary table S4.1 Primer design for the 10 validation CNVs including the predicted genomic location of the variant, its type, the expected product lengths and the validation status.

BTA	Predicted variant start (bp)	Predicted variant end (bp)	Type	F&R Product Length (normal)	F&R Product Length (deletion)	F&D Product length (normal)	Validation status	Forward primer sequence (F)	Reverse primer sequence (R)	Second reverse primer sequence (D)
4	16,347,345	16,347,873	DEL	1102	574	NA	Failed - off targets	TCCAGAGACTACAGTGGGAG	CTCCAGTATTGATGCCTGGG	-
6	86,460,133	86,460,644	DEL	1017	506	631	Failed - fixed DD	TGTTAGTTCAGCTGAGGACG	TCACTGCTGAACATTTTGGC	GGTGAAAAGACAGCCTTCAGA
7	82,154,890	82,155,697	DEL	1115	308	NA	Validated	CCTGTGTAGCCACCATTCAA	AGGAAATCCAAAGTCAGGCTAC	-
11	13,095,979	13,096,992	DEL	1378	365	453	Validated	TCTTAATCCGTGGGCTCCTA	CTGGGATTACGGGAGACCTAT	CAGCCATTCTGACTGTGGTA
20	9,931,216	9,932,200	DEL	1254	270	522	Validated	TGTCCTGTATTTGTGTGGGTC	CCAGGCCCTTATTCTGTGAC	CTGCGATGAACATTGGGGTA
20	65,744,143	65,745,214	DEL	1430	359	580	Validated	ACTAAGATCATGGCATCTCATCC	CTTTAGGATGGACTGGGTGGA	TGACTGCTCAAAGGAATCTGT

22	38,000,845	38,027,594	DEL	26934	185	727	Validated	CAACACTTATGAGCTGGGTGA	TCCCAAAGTCATCAGGCAAA	TAAACCAGGCTTGCTTCAGG
23	914,957	916,403	DEL	1980	534	830	Validated	CCAGTAGCAACTTGTCCTCC	AAGCCATAGCAAACCCGTAG	CCTTATGATCCAGCAACCCC
23	44,946,422	44,946,933	DEL	932	421	NA	Validated	CACCCAGGACTGATCTTTAGAA	GAATGGGAAAGACAGGGATCT	-
26	51,809,137	51,809,870	DEL	1210	477	NA	Failed - off targets	ACAAATGGAGACAGAAGAGGC	TTGTGAAAATGAGACCCTGGC	-