Communicating Research to the General Public

The WISL Award for Communicating PhD Research to the Public launched in 2010, and since then over 100 Ph.D. degree recipients have successfully included a chapter in their Ph.D. thesis communicating their research to non-specialists. The goal is to explain the candidate's scholarly research and its significance—as well as their excitement for and journey through their area of study—to a wider audience that includes family members, friends, civic groups, newspaper reporters, program officers at appropriate funding agencies, state legislators, and members of the U.S. Congress.

WISL encourages the inclusion of such chapters in all Ph.D. theses everywhere, through the cooperation of PhD candidates, their mentors, and departments. WISL offers awards of $250 for UW-Madison Ph.D. candidates in science and engineering. Candidates from other institutions may participate, but are not eligible for the cash award. WISL strongly encourages other institutions to launch similar programs.

WISL
Wisconsin Initiative for
Science Literacy

The dual mission of the Wisconsin Initiative for Science Literacy is to promote literacy in science, mathematics and technology among the general public and to attract future generations to careers in research, teaching and public service.

Contact: Prof. Bassam Z. Shakhashiri

UW-Madison Department of Chemistry

bassam@chem.wisc.edu

www.scifun.org

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

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.g. 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 ≤ 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.

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.

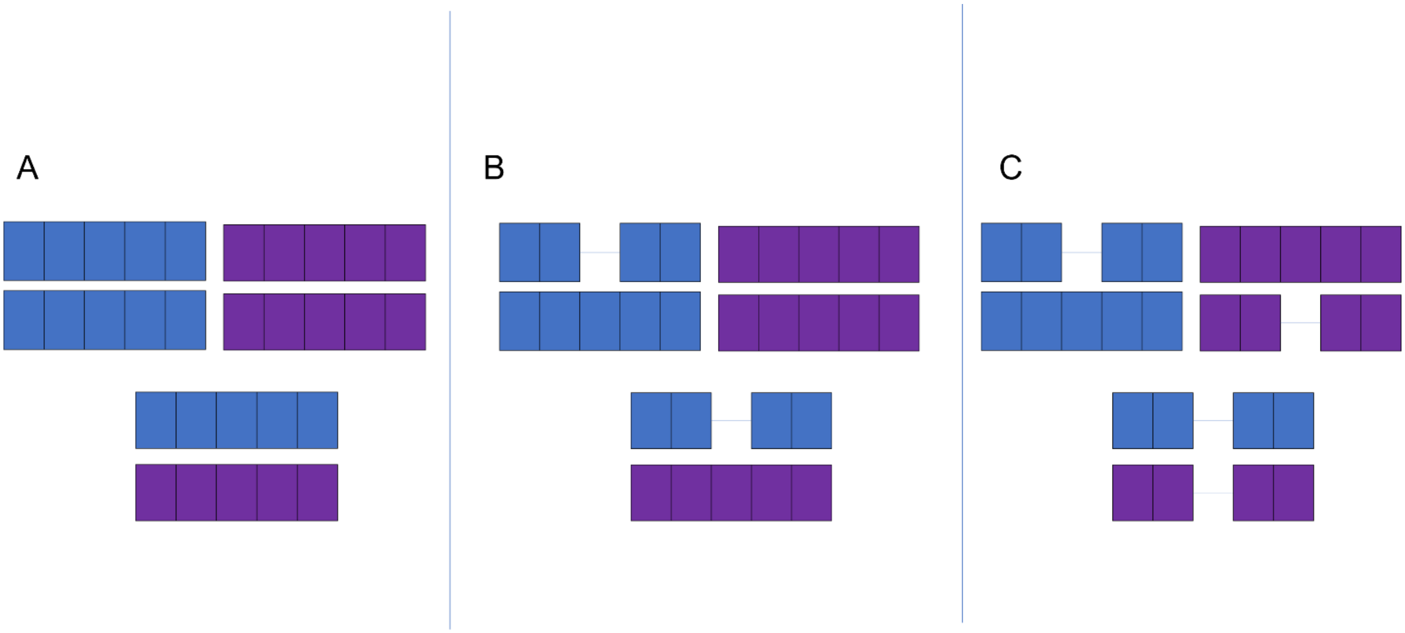
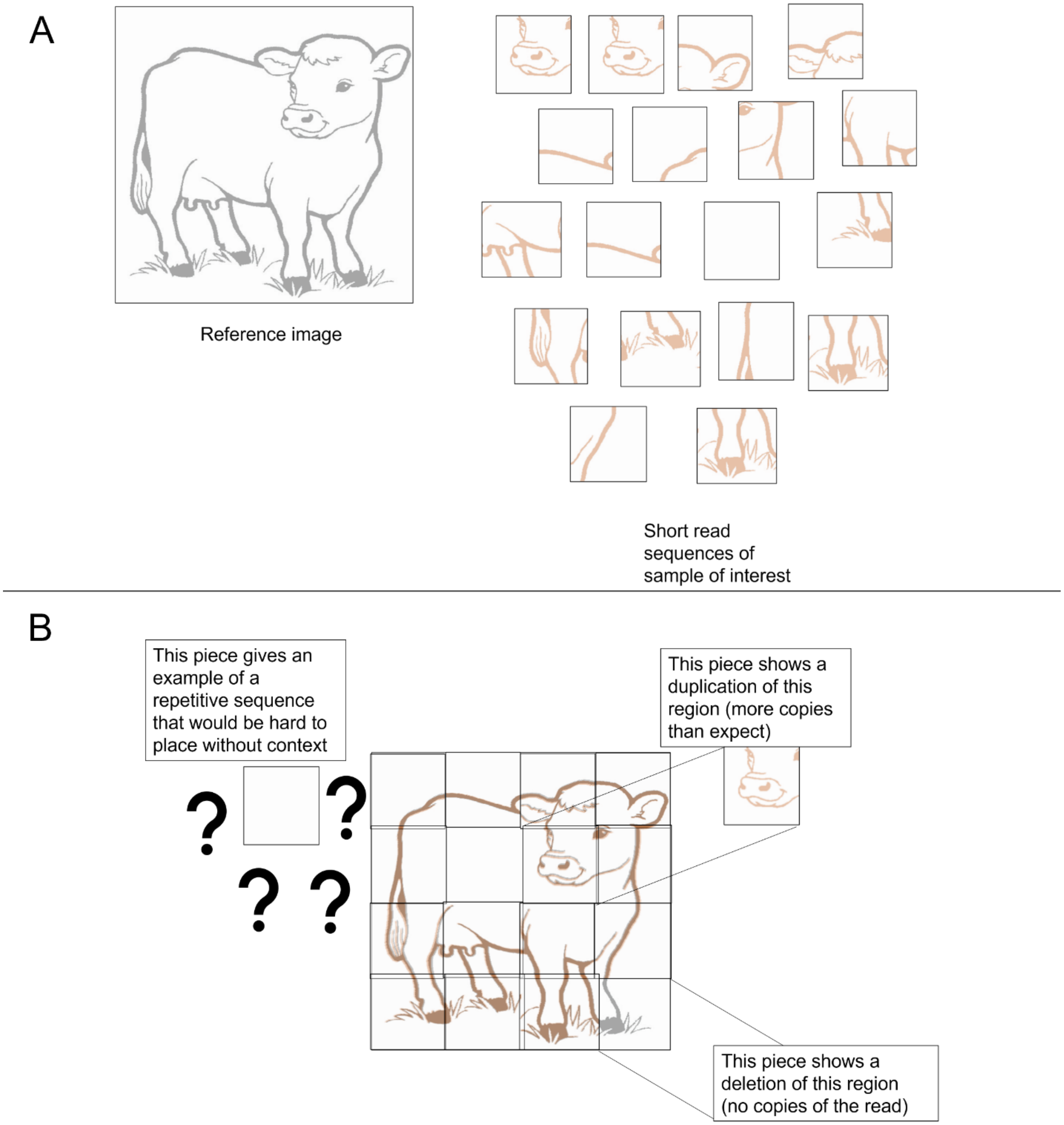


Figure 6.2. Depicts a sequence alignment to a reference as puzzle pieces. In part A, there is a reference image (left) and different reads from a sample individual at 1x depth (right). Part B is those reads aligned to the reference map. It highlights two features and a challenge from dealing with short 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.



This piece gives an example of a repetitive sequence that would be hard to place without context

This piece shows a duplication of this region (more copies than expect)

This piece shows a deletion of this region (no copies of the read)

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.

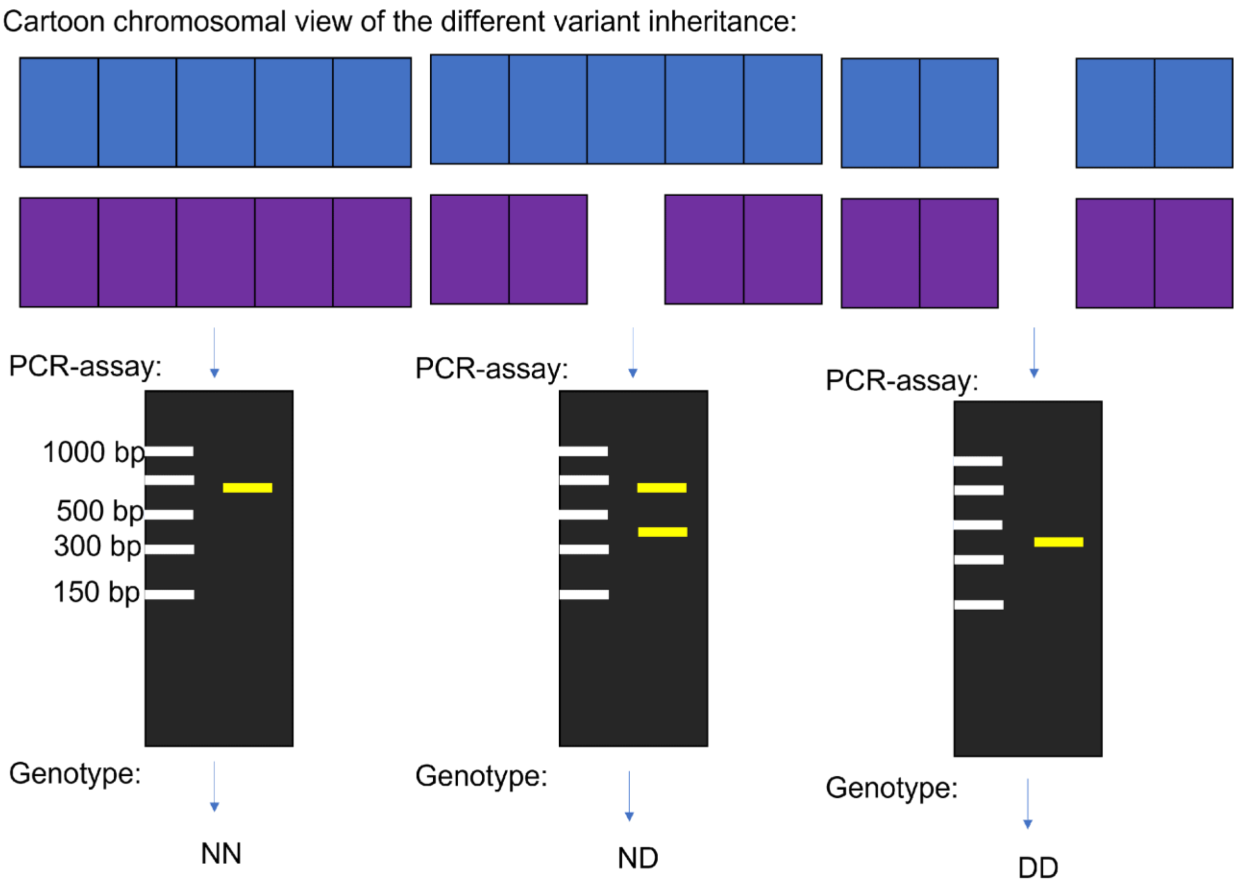
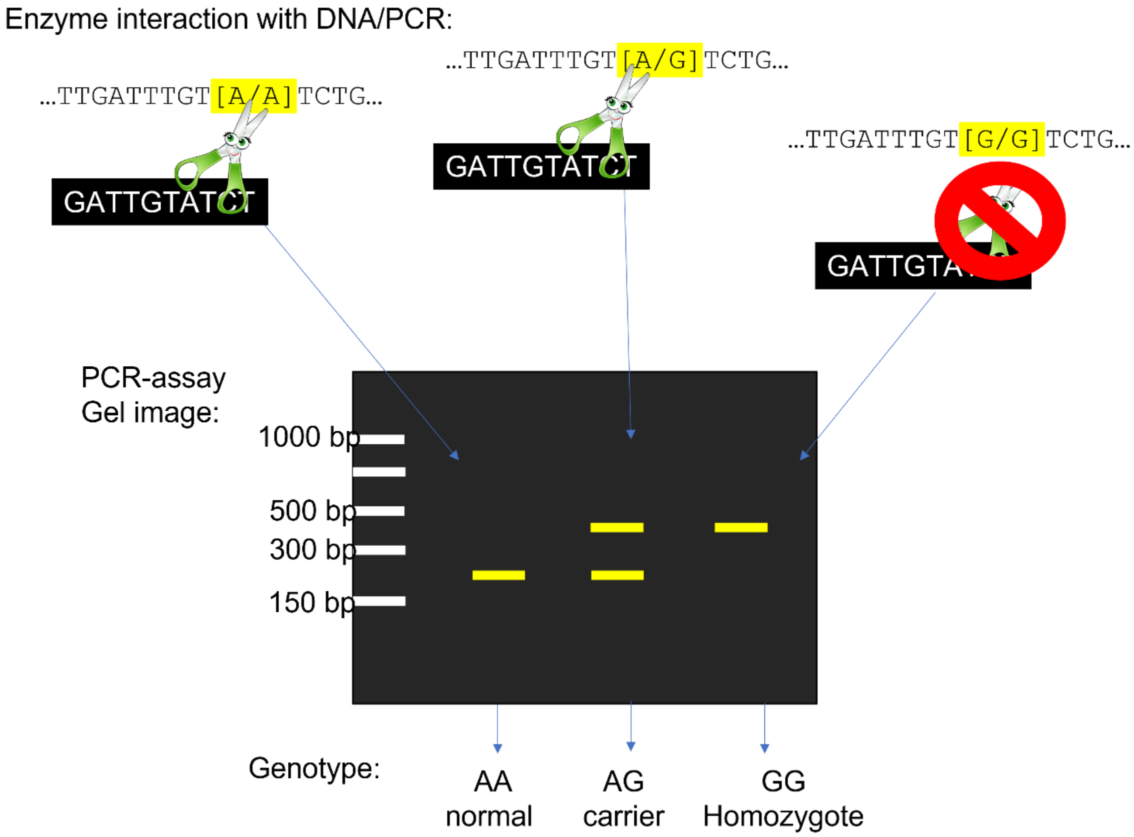


Figure 6.4. Diagram explaining how the restriction digest PCR (RFLP-PCR) assay interacts with the target region containing the SNP of interest. The enzyme (represented as scissors) only cuts where the sequence matches its target site (GATTGTATCT) thus if the variant is nucleotide G it will not cut. Next 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.


Enzyme interaction with DNA/PCR:
...TTGATTTGT[A/A]TCTG...
GATTGTATCT
...TTGATTTGT[A/G]TCTG...
GATTGTATCT
...TTGATTTGT[G/G]TCTG...
GATTGTA
PCR-assay
Gel image:
1000 bp
500 bp
300 bp
150 bp
Genotype:
AA
normal
AG
carrier
GG
Homozygote