Leveraging Sequence Data for High-Density Imputation and Genetic Defect Mapping in Ruminants

A Thesis Presented for the

Master of Science

Degree

University of Tennessee, Knoxville

Kenzy Ashton Hoffmann

May 2025

ACKNOWLEDGEMENTS

Thank you to all of my family and friends who have supported me throughout my life. To my parents, thank you for all of your love and support, I would not be where I am without everything you have done for me. There are not enough words to thank you both for your unconditional support, I am the person I am because of you both. To my sister, Carly I could not have asked for a better sister, role model, or best friend. Thank you for always being a listening ear and helping push me along the way. Proud does not begin to describe how I feel about having you as my big sister, and I know you will do great things. To my brother, Trey, thank you for being my best friend and always finding a way to make me laugh. I am so proud of you and excited to see where you go on your academic journey.

To my fellow lab mates and graduate students, thank you for welcoming me and making me feel at home. I am grateful for the friendships I have made throughout my time here. I look forward to seeing where everyone's careers take them. I know everyone is going to accomplish great things!

Thank you to my committee members, Dr. Jon Beever and Dr. Brynn Voy. I am grateful for your guidance and willingness to answer any questions I had.

Lastly, thank you to my mentor, Dr. Troy Rowan for all of your support and for giving me every opportunity to learn and grow as a professional. The experiences I have gained thanks to you throughout my time here have had a profound impact on me.

ABSTRACT

The recent increase in popularity of Single Nucleotide Polymorphism (SNP) chip technology has led to advancements in genomic prediction accuracy across many livestock species. Our ability to leverage genomic information when making predictions for animal performance has led to those predictions increasing in accuracy and driving more efficient genetic gain. While SNP chips are tremendously useful for that purpose, more dense genotyping that would be useful in mapping studies is costly and impractical for routine use in populations. This thesis focuses on optimizing a genotype imputation pipeline for increasing the density of genetic markers for use in downstream Genome Wide Association Studies (GWAS). This allows allow for lower density commercially-available SNP chips such as those with < 50,000 markers to be used in analyses that help fine-map complex trait associations. Another piece of our work leveraged whole genome sequencing to determine possible genetic causes of cleft palate in Boer goats. Sequencing unaffected parents and affected kids can help in identifying variants potentially linked to the defect. We worked to identify candidate variants and verify their impact on the phenotype to support the development of a genetic test for producers. Both studies used whole genome-sequencing data to help inform producer breeding decisions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One: Literature Review

Introduction

Gains in beef cattle production have been largely driven by increases in efficiency that have also resulted in producing more economically valuable animals. This has led to a 20 - 30% increase in yield through advances in genetics, disease control, and nutrition (Thornton, 2010). Efficiency varies greatly depending on the producer, but for most beef cattle producers it means having cattle that reach harvest weight faster by growing efficiently. In recent years there has been a flurry of new and novel statistical methods used to generate estimated breeding values (EBVs; Walsch & Lynch, 2018). The breeding value (BV) of an animal for a trait represents its additive genetic contribution, half of which will be passed on to its offspring (Wray et al., 2019). An EBV is a statistical estimate of an animal's true, but unknown BV. These EBVs allow producers to make more accurate selection decisions when breeding, which leads to improvements in traits over generations. Typically, EBVs are calculated utilizing phenotypic data and a best linear unbiased prediction (BLUP; Henderson 1975).

In beef cattle populations, EBVs are commonly reported as Expected Progeny Differences (EPDs), or one half of an animal's EBV (Garrick, 2011). This allows the resulting value to be interpreted as the sire's contribution to a trait. An EPD cannot exceed half of the EBV as a parent can only pass on one half of their genetic information. Since EPDs were first utilized by Dr. Richard Willham in 1971, they have been widely utilized for breeding decisions across the beef cattle industry (Thompson, 2008; Garrick, 2011). As these tools become more widely used, research and industry groups are interested in increasing the accuracy of EBVs. One way of increasing the accuracy is with the inclusion of marker information into genomic best linear unbiased prediction (GBLUP). GBLUP helps to resolve genomic relationships that are not accurately represented by pedigree estimates. The addition of genomic information has been

shown to increase genetic gain by 8-38% (Meuwissen and Goddard, 1996). Utilizing this marker information is facilitated in complex traits thanks to the phenomenon of linkage disequilibrium (LD). In the case of complex traits, many small effect loci may be closely positioned near one another, allowing for a small subset of markers to represent their full contribution to the trait (Ziang, 2021). It is thanks to LD that we are able to utilize genomic information such as a 50K SNP chip. Even if the causal variant for a trait was not directly measured by the SNP chip, linked marker variants can still be used to estimate the aggregate genotype's effect (Meuwissen et al., 2001; Tyler et al., 2016). This is one of the reasons marker-assisted selection (MAS) is limited by the amount of genetic variation explained by a quantitative trait loci (QTL; Meuwissen et al., 2001).

Continued improvements in EBV accuracy will likely require a more precise understanding of trait-variant associations. One tool for exploring these in large datasets is the genome wide association study (GWAS). These statistical methods are aimed at identifying associations of genotypes with phenotypes by testing for differences in the allele frequencies of genetic variants between individuals who are ancestrally similar but differ phenotypically (Uffelmann, 2021). Through GWAS and other similar genomic studies, loci associated with traits can be identified. The power to detect real associations using GWAS is limited by the size of genotyped/phenotyped individuals, and the resolution of variants being tested. Larger sample sizes can identify smaller-effect associations, while higher-resolution genotypes allow for stronger associations between causal variants and tested markers.

Loci identified by GWAS will undergo validation in varying populations to confirm their impact on the given trait. The validated loci can then be incorporated into statistical genomic prediction models such as GBLUP (Meuwissen and Goddard, 2001). The refined models are

then applied to breeding programs to help select animals that are the most genetically valuable. By using genomic information in the calculation of EPDs, the accuracy of those values should increase, even in the absence of progeny records.

Genomic Prediction

Historically, most genetic improvement in beef cattle has been generated by phenotypic selection of higher performing individuals. Those individuals are the ones selected to be the parents of the next generation. However, this phenotypic selection can only generate a limited amount of genetic gain, as its accuracy is limited by the heritability of a trait. This is further complicated due to delayed expression of certain phenotypes. The most extreme examples of this are female fertility and other reproductive traits. Genetic improvement in fertility traits has been slow due to the binary nature of their occurrences, their low heritability, and the delayed expression (Morris et al., 1993). These traits are delayed due to the nature of a beef cow's breeding season and the age at which they reach sexual maturity. The issue of determining more accurate female fertility traits is further hindered by the reliance on whole-herd reporting systems which require the breeder to record and input data (Rust and Groeneveld 2001; Middleton and Gibb 1991; Morris et al. 1993).

Moreover, the polygenic nature of these complex traits means that phenotypic selection will not accurately capture the full underlying genetic variation of a trait (Garrick, 2011). For example, another complex trait that can be difficult to accurately select for based on phenotype is feed efficiency as there is not an easy and cheap way to measure the trait available to producers. In a study performed by Seabury and others they found that using a 778K SNP panel could identify genetically superior animals for which phenotypic selection alone could not account for (Seabury et al. 2017).

The complex nature of many economically relevant traits is not only impacted by the effect of many genes but also by an animal's environment. Phenotypic selection will not always include the genotype by environment interaction seen in beef cattle. This interaction is important to take into consideration when choosing which breed of cattle to raise in certain climates as well as management practices a producer might want to implement. As production potential can differ due to a breed's natural adaptation toward a certain environment or situation (Vercoe and Frisch, 1984). This is why the formation of contemporary groups is crucial for calculating EPDs in beef cattle. Contemporary groups (CG) are groups of individuals who have been exposed to the same environment, so an example would be animals that are born the same year, season, and have the same breeder. This helps remove phenotypic variation that is due to the environment so the resulting values depict the estimated genetic component. This formation of contemporary groups is utilized in the calculation of EPDs to help further remove environmental variation, to further control for differences across breeds, conversions must be performed.

The complex nature of economically important traits involving their genetic diversity and environmental effect has led to a focus on trying to predict an animal's progeny's performance through previously mentioned tools like EPDs. Before beginning to predict an animal's phenotype based upon their genotype it was necessary to understand how many loci will impact a complex trait. This was informed by one of the fundamental theories in quantitative genetics, Fisher's Infinitesimal Model (Fisher 1918). Fisher's (1918) model postulates that many loci have a small effect on a trait. Fisher's (1918) model assumes that the combined effect of many loci will be normally distributed regardless of allele distribution (Fisher 1918). Understanding this has allowed for other models to increase in complexity by examining dominance and epistatic effects, rather than purely additive ones like Fisher's. The Infinitesimal Model's assumption that

many loci have small effects aligns well with the use of SNP chips in genomic selection (Fischer 1918). By capturing the effects of numerous SNPs spread across the genome, genetic predictions can account for a significant portion of the genetic variance. These markers can then be used to estimate the genetic merit of an individual more accurately, thereby enhancing the prediction of EBVs and consequently EPDs (Meuwissen & Goddard, 2001; VanRaden, 2008). It was because of the foundation that Fisher (1918) provided that paved the way for the development of other genomic prediction methods such as BLUP and eventually GBLUP. Without the assumption of small genetic effects on many loci, utilizing SNP chip technology for genomic predictions would not be feasible due to the limited number of SNPs typed by an array. If only a handful of variants affect a trait, the construction of SNP arrays would likely not be necessary.

This foundation has led to further research into increasing the accuracy of EBVs by including genotypic, phenotypic, and pedigree data. With the further understanding of complex traits, other models came out, those being either nonlinear or linear models such as BLUP and all of its variations (Meuwissen et al. 2001). As previously mentioned, it is because of the foundational idea of Fisher's Infinitesimal Model that the creation of the GBLUP was possible (Fischer 1918). GBLUP utilizes genomic information derived from single-nucleotide polymorphisms (SNPs) to calculate a genetic relationship between individuals commonly referred to as a genomic relationship matrix (GRM) (Van Raden 2008). It is possible to calculate a BLUP without genomic information by using a pedigree relationship matrix. These pedigree relationships represent the likely shared genetics between relatives based on assumed coefficients of relationship (Henderson 1975). Pedigree-based genetic evaluations were the standard method for estimating genetic merit of animals until commercially-available SNP arrays were delivered to the industry in the late 2000s. Another method for integrating genomics into

EBV prediction is single-step GBLUP (ssGBLUP) which is a method that can increase genomic prediction accuracy because it jointly analyzes both genotyped and non-genotyped animals rather than relying on a blending technique of genomic and non-genomic EBVs (Aguilar et al., 2010, Christensen and Lund, 2010).

In a landmark paper, Meuwissen and others paved the way for the future use of SNP technology allowing genomic information to be more effectively utilized in genetic predictions. They postulated that using genotypes of individual chromosome fragments, characterized by SNP genotypes or haplotypes could be used to calculate an EBV. In the paper by Muewissen and others, they compared least squares (LS), BLUP, and Bayesian for predicting a total breeding value (TBV). In simulations, they found that BayesA and BayesB were able to estimate the effect of large QTL, with BayesB outperforming BayesA for estimating large QTL effects due to BayesBs ability to model the possibility that some SNPs have zero effect. However, both methods had limitations in accurately estimating the effect of small QTLs due to the assumption of normally distributed effects. However, they found that the contribution of these small QTLs was similar for both Bayesian methods and BLUP. This meant that Bayesian methods could still function as viable prediction methods in cases where large QTL existed. Both Bayesian and BLUP provided accurate EBV values, whereas LS had overestimated values and a low accuracy. This was due to LS requiring haplotype effects to be estimated simultaneously. However, one of their more impactful conclusions was that genomic selection could significantly increase genetic gain in livestock populations (Meuwissen et al., 2001).

Since the initial introduction of Bayesian methods they have continued to evolve, as evidenced by the development of the Bayesian Alphabet, which includes a range of Bayesian approaches for genomic selection (Gianola et al. 2009). The Bayesian Alphabet consists of many

Bayesian statistical models including BayesA, BayesB, BayesC, BayesCπ Pi, BayesD, BayesDπ, and BayesRC (Gianola et al. 2009). BayesA and BayesB are the earliest methods first introduced by Muewissen and others. mentioned above (Meuwissen et al., 2001). Further expansion on the Bayesian approach took place in the form of BayesC and ByesCπ with BayesC refining BayesB by introducing a common variance for SNP effects helping to simplify the computation. BayesCπ showed comparable, if not superior, accuracy in determining genomic estimated breeding values (GEBVs) when compared to other methods, especially in scenarios with high LD. This improvement is due to the method's inclusion of prior probability (π) as unknown, allowing it to adapt better to different genetic architectures and further enhancing prediction accuracy (Habier et al., 2011). Bayes RC takes the use of prior probabilities a step further by integrating prior biological knowledge into genomic predictions. This is done by using biological information to classify variants into different categories which are each assigned different probabilities to help determine their likelihood of affecting the trait. This focus on higher priority variants helps increase the models precision and accuracy (MacLeod et al., 2016).

Implementing the methods proposed by Meuwissen, Hayes, and Goddard was not possible at the time of their publication. No method existed for assaying all of the haplotype blocks in large groups of animals. The development of the SNP microarray in 2008 made the estimation of genomically-enhanced breeding values possible (Matukumalli et al. 2009). The advent of genomic selection in livestock populations led to a flurry of innovations aimed at increasing prediction accuracies. This increase in interest with MAS led to researchers looking into how to fine map traits of interest and then finding a way to include the information learned in EBVs.

Identification of functional genetic markers

When attempting to integrate causal or functional markers into genetic predictions, it is essential that we first identify the locations of those variants or nearby variations that are near-perfectly associated. The Functional Annotation of Animal Genomes (FAANG) consortium focuses on creating high-quality functional annotations of animal genomes, including cattle (Guiffra et al. 2019). FAANG performed annotations on functional elements within genomes such as genes, regulatory regions and more (Guiffra et al. 2019). Understanding the biology that underlies complex traits via regulatory variation has been shown to increase prediction accuracy in many contexts (Xiang et al. 2019)

In addition to using biological information, statistical genotype-phenotype associations from GWAS can be used to help improve genomic predictions. These mapping studies rely on fine mapping to refine associations from large haplotype blocks down to individual variants. Fine mapping is a computational approach used to identify the most likely causal variants affecting a given phenotype with each of the genetic loci identified by research such as a GWAS. This is done by leveraging patterns of LD and association statistics to prioritize variants (Raychaudhuri et al., 2011; Schaid et al., 2018). There are several factors that influence one's ability to map complex traits including marker density, QTL detection, crossover density, and the molecular architecture of the QTL (Georges 2007). Fine mapping these complex traits is made difficult due to complex traits having many causal variants that each have a small effect (VanRaden 2017). The impact of this has been lessened by fitting all variants simultaneously (MacLeod et al., 2016). To further help identify the causal variant of a trait it is useful to map more than one trait as some traits will have similar variants shared between them (Bolormaa 2014). Lastly, filtering variants can be helpful in identifying the causal variants.

Xiang (2021) and others fine-mapped variants from GWAS, variant clustering, and Bayesian models to filter variants, all with the goal of improving genomic predictions via functional information. There were five steps to their experiment, the first being the ranking of variants based upon GWAS p-values and Functional-And-Evolutionary Trait Heritability (FAETH) scores. FAETH scores were first used by Xiang (2019) and others are calculated by first classifying the variants by partitioning the genome, then GRMs are calculated for each partition of the genome. Then variance is estimated utilizing a restricted maximum likelihood (REML), this variance is then divided by the number of variants for the given trait to determine a per-variant heritability; these values are then averaged for each trait. The scores are meant to provide a way to prioritize variants based on their potential impact on a trait, the higher the FAETH the more functional the variant is predicted to be (Xiang 2019). Continuing with their fine mapping, they clustered certain variants and a BayesRC model was used. Next, the variance of GEBVs was calculated for each of the partitions and variants were ranked based on how much they contributed to the variance. Lastly, they successfully designed an enhanced 50K Genotyping panel with informative markers based on their research findings (Xiang 2021). Genomic predictions using this panel had significantly better accuracies than with standard 50K SNP arrays.

Another study sought to characterize the genetic control of gene expression and splicing across tissues in cattle. The Cattle Genotype–Tissue Expression atlas (CattleGTEx) Utilizing these QTLs along with those from FAANG, functional genetic markers in beef cattle were defined and gained insights into the molecular architecture of complex traits (Liu 2022). The identification of functional genetic markers in beef cattle is crucial to our understanding of the genetic basis of complex traits. By utilizing data from the CattleGTEx, FAANG consortium,

and other eQTL studies, scientists will more easily be able to identify causal variants. The identification of functional genetic markers in beef cattle is a complex process that requires the integration of data from multiple sources (Hocquette et al., 2007). A part of identifying causal variants is the application of advanced genomic tools, such as GWAS and variant clustering. These efforts have provided valuable insights into the molecular architecture of complex traits in beef cattle, allowing for improved breeding programs and a better understanding of the genetic basis of these traits. This is all to help increase the accuracy of EBVs and eventually EPDs.

Imputation

The advancement of genomic technologies through the advent of SNP genotyping technology has led to many advancements in our ability to make genomic predictions. While SNP chips have helped these advancements, they represent only a small portion of the genome. SNP arrays are designed to assay high-frequency variants that are evenly spaced across the genome. This allows distinct segments of the genome to be represented by a low number of markers. This relies on statistical associations between nearby loci (i.e., linkage disequilibrium – LD). These assays are useful in genetic evaluation contexts, as they can effectively represent genetic variation between individuals and serve as useful approaches for correcting pedigree relationships, however for mapping causal or functional variants, they lack the necessary resolution. This is particularly problematic for complex traits, which are controlled by many variants of small effects, distributed across the genome. Mapping these variants accurately requires higher resolution than what is possible with standard SNP arrays. Genotype imputation can be used to help fill in those gaps. Imputation uses higher-density reference haplotypes that can be matched to phased low-density samples and missing genotypes filled in using a probabilistic model. The addition of this imputed data can lead to the identification of more

significant associations in GWAS due to increased statistical power (Marchini and Howie, 2010). Imputation enables researchers to genotype large populations at a relatively low density and then use whole-genome resequencing data to infer genotypes at a higher, sequence density. These are detailed in the papers described above (Hoff et al. 2017; Wiggans et al. 2016). This is made possible because of the previously mentioned haplotypes, which are groups of alleles that will be inherited together from a single parent. Performing imputation in a GWAS can lead to a power increase of up to 10% when compared to testing only genotyped SNPs (Spencer et al., 2009), though it is highly dependent on starting array density, population size & structure, and the genetic architecture of the trait.

Imputation is a powerful tool for mapping causal variants because it can allow researchers to analyze non-genotyped SNPs, which enhances the resolution of association signals. This increases the likelihood that a test is directly performed on the causal SNP or a SNP in perfect LD with it. This is an invaluable approach for narrowing down associated regions to pinpoint the causal or perfectly associated variants for a given trait. This is important when studying complex traits as they are impacted by huge numbers of variants. Imputation is possible because it takes advantage of LD allowing SNPs that are not present in the genotyped file to be imputed, thus extending the usable SNPs for association mapping beyond those directly genotyped (Marchini and Howie, 2010). This approach improves the detection of associations by filling in gaps in the genome allowing for a more comprehensive view of the genome prior to performing analysis such as a GWAS (Marchini and Howie, 2010). Moreover, imputation facilitates the study of rare variants, which may be particularly impactful for complex traits. Since rare variants are typically underrepresented in whole genome reference panels, using a

high-density chip intermediate imputation step prior to sequence level imputation allows for those rare variants to be included.

Phased haplotypes are required inputs of imputation. Phasing is a process used to computationally infer which areas of the genome were inherited together. Performing phasing provides a significant boost to imputation accuracy by making the genotype inference haploid rather than diploid. These imputation methods ensure that possible haplotype matches occur from a group of high-density reference haplotypes. There are several statistical methods for performing imputation with the most utilized being a Hidden Markov Model (HMM) which is a class of statistical models that can help to determine any associations between an observed process and an unobserved one. First used by Stephens and others, a HMM is a Bayesian method that uses prior information to reconstruct haplotypes (Stephens et al., 2001). An example of one of the first and more widely used HMM is IMPUTEv1 which fills in the missing genotypes by modeling LD patterns from the reference panel to the study individuals (Marchini et al., 2007). Further advancements have been made in the versions of IMPACT that followed including IMPACTv2 where SNPs were explicitly divided into two sets U and T, where set T are SNPs that are found in both the study dataset and the reference panel while set U are SNPs found in the reference panel. IMPACTv2 is more computationally efficient than v1 because it can employ haploid imputation by determining haplotypes using set T. Moreover, it can alternate between haploid and phasing imputation utilizing a Markov chain Monte Carlo (MCMC) approach (Marchini and Howie, 2010).

Since the initial creation of imputation software several improvements have been made. One of the most recent and widely used iterations of imputation software is IMPUTE5 which is specifically designed to handle large reference panels efficiently. This software improves upon

one of its earlier versions, IMPUTE2 by optimizing the use of a custom subset of haplotypes when imputing each individual. It achieves fast, accurate, and memory efficient imputation by using the Positional Burrows Wheeler Transform (PBWT). PBWT works by transforming haplotypes to a form that makes it easier to find matches between them which can help determine the best matching haplotypes improving speed and accuracy. IMPUTE5 scales sublinearly with reference panel size making it up to 30 times faster than MINIMAC4 and three times faster than BEAGLE5.1 all while using less memory than both (Rubinacci et al., 2020). Both of these are commonly used imputation software programs that IMPUTE5 has managed to outperform leading to the decision to utilize IMPUTE5 in the research performed to optimize imputation. Prior to imputation another process should be performed to help increase speed and accuracy of imputation, that being phasing. The software chosen for this research to perform phasing is SHAPEIT4 which uses PBWT as well, but to quickly identify a set of informative haplotypes for use in imputation (Delaneau et al., 2019).

While these varying software tools produce accurate imputed genotypes for cattle, they were initially designed for use in humans to impute from a high-density genotype panel to full-genome sequence which is not always possible for cattle as most of them are genotyped with lower density SNP chips (Pausch et al. 2017). Imputing from the low density assays used in livestock species directly to full-genome sequence has been shown to be less accurate, even in instances of populations with high LD (MacLeod et al., 2016). To increase accuracy, a two-step strategy can be used which involves imputing from a low-density assay to a high-density assay and lastly to the sequence level (van Binsbergen et al. 2014; Kreiner-Møller et al. 2015). Another factor that can increase the accuracy of imputation is by using a multi-breed reference panel and

incorporating rare variants. Multi-breed reference panels enhance the accuracy of imputation when compared to breed-specific, as does the inclusion of rare markers (Rowan et al., 2019).

Performing imputation with the most up to date software available should help increase its accuracy, but the software do not directly provide informative imputation accuracy measurements. There are several methods for determining the accuracy of imputation, the two most popular methods are concordance rate and Pearson correlation (r). Concordance rate measures the proportion of correctly imputed genotypes and can be useful for quick accuracy checking. Pearson correlation is similar in that it also compares the true genotype to the imputed one. While both methods work for individuals, they overestimate the accuracy of imputation for low MAF variants (Hancock et al., 2012; Lin et al., 2010). Given that some rare variants could have an impact on economically relevant traits it is important to find methods of determining imputation accuracy that could account for alleles with a low MAF. Imputation Quality Score (IQS) can help account for low MAF because it can adjust concordance rates to account for the chance that an imputed genotype is correctly guessed, which is more likely for rare markers (Lin et al., 2010; Rowan et al., 2019). Additionally, IQS and Pearson's correlation penalize more for imputation errors made for rare variants (Rowan et al., 2019). Using all three of these accuracy determinants can further increase confidence in imputation results.

Genome-wide association studies (GWAS)

A genome wide association study (GWAS) aims to investigate the genetic structure of complex traits, identify and map causal variants for complex traits, and predict how an individual's genetics contribute to those traits (Xiang 2021). The results of a GWAS could be used to identify genetic markers or genes associated with traits, and ultimately predict the performance of individuals. This knowledge of the functional causal variants can then be used to

develop GE-EPDs for traits such as birth weight, weaning weight, and yearling weight. A GWAS is a multi-step process that starts with the aggregation of phenotypic data and genotypes from phenotyped animals (Uffelmann et al. 2021). Once animals have been genotyped and phenotypes have been collected a mixed linear model can be used to perform a statistical association test. This is done by performing a linear model:

$y = X\beta + u + e$. Where y represents an adjusted phenotype, X is an animal's genotype which is treated as an independent variable. β represents the allele substitution effect, u is the GRM, and lastly e is a random effect of residuals. This estimates variant effect sizes and their standard deviations. This allows for p-values to be reported as measures of association onfidence. Typically the output of this analysis will be p-values which are visualized using a Manhattan plot. Putative causal variants will be located in the peaks on the Manhattan plot.

GWAS are a helpful tool when it comes to identifying causal variants, but sometimes it can identify loci that are not responsible for variation. This is caused by several factors including that stochastic noise can generate false associations in a small sample (Platt et al 2010). To help prevent this false positive a very large sample size should be utilized. Another potential reason for false positives in GWAS is the potential for confounding due to population structure. This can occur when patterns of correlation among loci and factors responsible for trait variation create indirect associations between markers and traits where no causal relationship exists (Platt et al 2010). To help mitigate the impact of population structure genetic ancestry should be included in the association test model. This helps to control for any confounding effects that may arise from population substructure, ultimately improving the accuracy of the associations identified. Another way of ensuring there is not a falsely identified QTL due to population structure is by validating the identified QTLs in another independent demographically similar

population. If the same variant is identified, then it is very likely to either be the causal variant or be in strong linkage disequilibrium with the causal variant (Saatchi et al 2014).

As previously mentioned, a part of a GWAS is understanding the genetic structure of the complex trait, a part of this is determining the number of variants that impact the trait and their effect sizes (Uffelmann et al 2021).

Conclusion

In beef cattle, research has been focused on making cattle grow as fast and efficiently as possible. A recent avenue to help drive further improvement is the inclusion of causal and functional variants in genetic predictions. Identifying these variants requires statistical association mapping between variants and the trait of interest. Genome-wide association studies are one approach to mapping variants related to complex traits. Using commercial SNP chips directly in GWAS does not provide the necessary resolution to identify causal or functional markers. The cost of genotyping is left solely in the hands of the producer, which is why most genotypes using lower density SNP chips as those are the most cost-effective options. However, as researchers work to leverage commercially-generated genotype data, they have limited resolution for mapping studies. Imputation can help infer missing genotypes and provide greater mapping resolution, resulting in increased statistical power for GWA

Chapter Two: Optimization of Imputation

Abstract

Improvements to single nucleotide polymorphism (SNP) genotyping technology has led to many advancements in our ability to make genomic predictions. While SNP arrays have helped drive these advancements, they only represent a small portion of the diversity represented in the genome. Genotype imputation can be used to increase the resolution of genomic information generated by SNP genotyping platforms. The inheritance of the genome in chunks, known as haplotypes, allows us to infer likely missing markers surrounding those present on genotyping arrays. Our work created an optimized computational pipeline that will perform genotype phasing and imputation using SHAPEIT5 and IMPUTE5, two of the best performing phasing and imputation softwares, respectively. Pre-phasing both low-density genotypes and reference animals is an essential component toward maximizing imputation accuracy. Our implementation of IMPUTE5 involved several processes, the phased samples are broken into chunks, or haplotypes based upon the reference. This process serves to help in parallelization of imputation ensuring faster computational time. We determined the optimal settings for SHAPEIT5 and IMPUTE5 to ensure accurate, efficient imputation. We focused on optimizing imputation using various reference panels and intermediate imputation steps. The reference panels consisted of one made of only high-density (HD) arrays, a combined panel of HD & F250 (COMBO), and the 1000 Bulls reference panel (1K Bulls). The first three reference files were used to perform intermediate imputation, then those imputed files were imputed again to sequence density using the 1000 Bulls reference. This use of intermediate imputation with varying reference panels resulted in significant increases in accuracy across both individual animals and classes of variants. The COMBO reference panel yielded the highest imputation accuracy, with notable improvements across all metrics, particularly for rare variants. It achieved

a concordance of 92.85%, an IQS of 80.70%, and a correlation of 86.79%. The HD panel followed closely with similar values, suggesting that additional functional markers from the F250 had only a small impact on overall imputation accuracy. The least accurate method was imputing straight to the 1000 Bulls reference panel without an intermediate imputation step, which significantly underperformed, especially for IQS (55.27%), highlighting the critical role of intermediate imputation steps for accurate imputation, particularly for low-frequency variants. These results suggest that the use of intermediate imputation with diverse reference panels, particularly the combined HD and F250 panel, is essential for improving imputation accuracy, especially for rare variants. This approach enhances the detection of genetic associations, making it a valuable strategy for genomic analyses and advancing the identification of causal variants in complex traits.

Introduction

Genotype imputation is a vital technique in modern computational genomics, offering a solution to the high cost of whole-genome sequencing by filling in gaps from lower-density SNP arrays. This approach allows for the extension of genomic data at a fraction of the cost of full sequencing, providing valuable insights into complex traits such as growth, disease resistance, and fertility in beef cattle populations. Through imputation, researchers can infer missing genotypes using reference panels, thereby increasing the accuracy of genomic predictions, which is essential for efficient breeding and improving economically important traits.

Despite the demonstrated benefits of imputation, key gaps remain in the optimization of this technology for beef cattle due in part to most software being designed for use in humans. One significant challenge is the accurate imputation of rare variants—alleles with a low minor allele frequency (MAF). These variants, though rare, can have outsized impacts on traits of

economic importance. Imputation accuracy decreases for these low-frequency alleles due to their underrepresentation in both genotyping arrays and commonly used reference panels. This is particularly problematic when studying complex traits controlled by multiple genes, where rare variants may play critical roles in trait expression but are difficult to capture through imputation or genotyping alone.

Another gap lies in the population structure and genetic diversity of beef cattle breeds. Current reference panels may not adequately represent the genomic variation seen across different breeds, particularly when using breed-specific reference panels. Multi-breed reference panels, which include diverse genetic backgrounds, have been shown to improve imputation accuracy, but further refinement is needed to ensure these panels capture the full range of rare and breed-specific variants (Rowan et al., 2019).

Additionally, advances in phasing algorithms, such as SHAPEIT5, and imputation software like IMPUTE5, have made great strides in improving the speed and accuracy of imputation. However, optimizing these tools to handle large reference panels and rare variants across diverse cattle populations remains a challenge. While intermediate imputation steps using high-density reference panels, like those from the HD and F250 arrays, can enhance imputation accuracy, further studies are needed to evaluate the performance of these approaches across different cattle populations. Optimizing imputation from commercial SNP arrays to sequence density can help identify functional variants in association studies, enhancing our understanding of biology and helping to improve the accuracy of genomic predictions.

Materials and Methods

Testing Dataset

We used 36 registered American Simmental animals from a resequencing project to create a synthetic group of testing individuals. These animals were sequenced to an average depth of 10X. Reads underwent quality control in FastQC, were aligned using BWA, and then had variants called according to GATK best practices (Andrews 2010; Danecek et al., 2021; McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013; Vasimuddin et al., 2019). We used map files from the Bovine SNP50 to extract variants from the resequenced individuals to serve as our testing set (Matukumalli et al., 2009). Positions were determined using the ARS-UCD 2.0 reference genome (Rosen et al. 2020). These were all individually used as the starting point for our imputation test set.

Pipeline Creation and Function

Our imputation pipeline was built using Nextflow, an open source workflow management system designed to allow for the writing of complex pipelines based on the Groovy scripting language (Tommaso et al. 2017). The pipeline built for this project performs several functions prior to phasing and imputation. Starting with genotypes in PLINK format in the form of .bed, .bim, and .fam files which are then converted to Variant Call Format (VCF) files using PLINK (Purcell eat al., 2007). To ensure compatibility with the phasing and imputation software, all input genotype files must be converted to VCF format, followed by a final conversion to Binary Call Format (BCF) using bcftools (Danecek et al., 2021). The pipeline was specifically designed to accommodate input files in any format, including PLINK, VCF, or BCF. Once the genotype files are in VCF format, we execute the final step of converting them to BCF, ensuring they are

properly formatted for downstream processing. Once converted to BCF format, the files are ready for input into the phasing and imputation programs.

The pipeline then processes the BCF file and its index to perform phasing using SHAPEIT5, which estimates the most likely haplotypes for each individual based on the observed genotype data (Delaneau et al., 2019). This phasing step is crucial for increasing the accuracy of subsequent imputation. After phasing, but before imputation, the files are divided into fixed length chunks to help with parallelization of imputation. This chunking process defines both the regions to be imputed and the buffer regions to be considered. Buffer regions are essential for taking linkage disequilibrium (LD) into account, as they extend the range beyond the focal region to enhance imputation accuracy. While only the chunked regions are imputed, the buffer regions provide additional information to improve precision. Performing chunking prior to imputation also allows for parallelization, significantly reducing computational time. The imputation step is carried out using IMPUTE5, and the buffer regions are fine-tuned to balance accuracy and computational efficiency (Rubinacci et al., 2020). A visual representation of the workflow is provided in **Figure 2.1**.

Phasing and Imputation

Upon implementation in Nextflow, we tested the impact of a variety of intermediate imputation on sequence-density imputation accuracy. To perform this analysis, a true genotype file containing 20,100,450 SNPs from 36 individuals was downsampled to 51,008 SNPs, with positions derived from the Bovine 50K SNP map file. To optimize computational efficiency and reduce storage requirements, the analysis was restricted to Chromosome 25, which includes 356,623 SNPs in the true genotype file and 719 SNPs in the 50K downsampled file. This

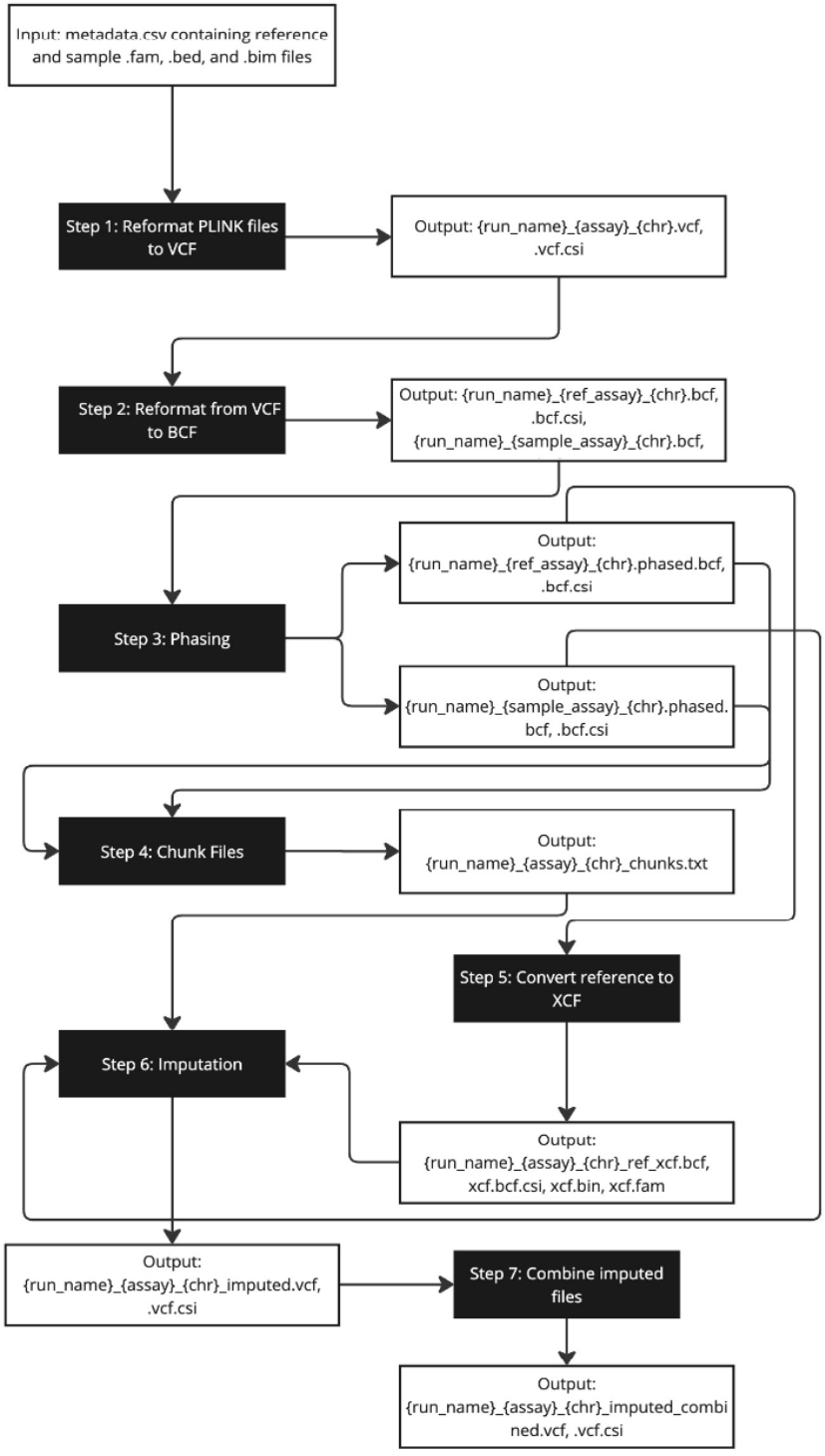

Figure 2.1. Nextflow Pipeline utilized for determining accuracy.
The process includes reformatting, phasing, chunking, converting references to XCF, and imputation, ensuring seamless data flow from input to final imputed BCF files.

approach mirrors strategies used in genomic prediction studies, where chromosomes are often processed individually to reduce memory usage and computational time without sacrificing the accuracy of predictions. For example, VanRaden (2008) demonstrated that processing each chromosome separately significantly reduced memory requirements, allowing efficient genomic prediction across multiple traits. This validates the use of Chromosome 25 as a representative subset of the genome for initial analysis, providing insights while managing computational resources effectively (VanRaden, 2008).

Imputation was performed using IMPUTE5 with phasing performed using SHAPEIT5 as previously described (Rubinacci et al., 2020; Delaneau et al., 2019). The workflow for the pipeline utilized is provided in **Figure 2.1**. Phasing and imputation were performed in two phases with our four classes of reference files. First, files were imputed using higher density chip reference files (700K to 850K SNPs), and then with whole genome sequence data from Run 8: 1000 Bull Genomes Project (Daetwyler et al., 2021). The high-density reference panels were provided phased from the University of Missouri, while the whole genome sequence data from Run 8: 1000 Bull Genomes Project was phased using SHAPEIT5. For Chromosome 25, the HD reference panel consisted of 11,019 SNPs and 22,236 individuals. The combined (COMBO) panel, which included both HD and F250 data, contained 13,943 SNPs and 52,778 individuals. The sequence-based (1K Bulls) panel from the 1000 Bull Genomes Project, which only included variants that passed a quality filter requiring the PASS label, contained 666,312 SNPs and 1,842 individuals. Genotypes were phased using the SHAPEIT5 phase common tool, a method particularly effective for phasing SNP array data with moderate-to-high allele frequencies. The characteristics of these intermediate imputation reference panels are provided in **Table 2.1**.

Table 2.1 Characteristics of Intermediate Imputation References

Intermediate Imputation References	Chr 25 SNP Count	Individuals
HD	11,019	22,236
COMBO	13,943	52,778
1K Bulls	666,312	1,842

This table provides a summary of the SNP count and the number of individuals for the different intermediate imputation references used in this study for chromosome 25. The references include HD, COMBO, and 1K Bulls. Each reference panel varies in both the number of SNPs and individuals, which affects the imputation process.

The high-density SNP chip reference panels used for intermediate imputation included those based on animals genotyped using the HD array only (HD) and a combined reference panel of all animals genotyped using HD and GGP-F250 (COMBO) arrays. The HD and F250 combined reference file was used after a reciprocal imputation between the HD file with the F250. This was done to ensure all SNPs were retained and no missing values were introduced. Imputation was also performed directly using only whole genome sequence calls from Run 8: 1000 Bull Genomes Project (SEQ) which contains 666,312 SNPs and 1,842 individuals after undergoing the PASS filter mentioned above (Daetwyler et al., 2021). To test the impact of intermediate reference panel composition, we evaluated reference panels with varying numbers of animals, SNPs, and densities (**Table 2.1**). The inclusion of the F250 panel allowed us to assess its potential to improve imputation accuracy when combined with the HD panel. Running the pipeline on varying density reference panels and in many stages will allow us to determine the impact of imputing with a higher density reference panel prior to whole sequence data. The impact of inclusion of the F250 panel can also be investigated.

Since we had true variant calls available, we were able to directly compare true and imputed variants to assess accuracy empirically. We quantified imputation accuracy using concordance rate, Pearson correlation (*r*), and IQS with calculations being performed using a custom Python script (Van Rossum and Drake, 2009). In the custom Python script concordance rate, *r*, and IQS were calculated on a per-SNP and per-individual basis. In this script, concordance rate is defined as the proportion of matching genotypes between the imputed and true genotype sets, where a genotype is considered concordant if the true and imputed genotypes were identical. The equation utilized was $Concordance\ rate\ =$

$\frac{Number\ of\ concordant\ genotypes\ (true = imputed)}{Total\ number\ of\ genotypes\ evaluated}$. Pearsons's correlation was calculated as follows: $r = \frac{\Sigma(G_{true} - \underline{G_{true}})(G_{imputed} - \underline{G_{imputed}})}{\sqrt{\Sigma(G_{true} - \underline{G_{true}})^2(G_{imputed} - \underline{G_{imputed}})^2}}$ where the numerator represents the covariance between the true and imputed genotypes, while the denominator is the product of the standard deviations of the true and imputed genotypes. The equation utilized for calculating IQS is $IQS = \frac{C - E}{I - E}$ where C is the observed concordance and E is the expected concordance calculated from the MAF. The use of this kind of equation is what allows IQS to better assess accuracy based on the possibility that rare MAF genotypes are more likely to be correctly inferred on chance alone.

One last metric was calculated, r^2 which is a common metric utilized in quality checking and benchmarking imputation software. r^2 was calculated by comparing the true genotypes to the imputed genotype probabilities output by the imputation software. Specifically, we used the posterior genotype probabilities from the GP field in the outputted imputed files and computed the imputed dosages as $P(1/1) + 0.5 \times P(0/1)$. Then, we calculated the Pearson correlation between the true genotypes and the imputed dosages, squaring this value to obtain r^2 as described by Rubinacci and others (2021).

Results

Average Imputation Accuracy

Imputation on the synthetic 50K SNP chip containing 36 animals was performed using IMPUTE5, after genotypes were pre-phased by SHAPEIT5. **Table 2.2** shows the accuracy measurements for each of the different imputation approaches following imputation to sequence-density via the Run 8: 1000 Bull Genomes Project Reference (Daetwyler et al., 2021). For animals imputed with a HD reference panel and then the 1000 Bull Genomes Project reference

Table 2.2 Final Average Imputation Accuracy Values for Chromosome 25

Intermediate Imputation References	Concordance	IQS	Correlation (*r*)
COMBO	.9616	.7801	.9358
HD	.9629	.7618	.9372
1K Bulls	.8304	.5465	.6823

This table presents the final average imputation accuracy metrics for Chromosome 25 across four intermediate imputation references. It includes key accuracy metrics: concordance, Imputation Quality Score (IQS), and correlation (r), demonstrating higher accuracy for the COMBO panels compared to HD and 1K Bulls. These values were calculated on a per SNP basis and then averaged for the table above.

their accuracies were .9629 for Concordance, .7618 for IQS, and .9372 for Correlation. For animals imputed with the reference panel where the HD and F250 references were combined, accuracies are .9616, .7801, .9358 for Concordance, IQS, and Correlation respectively. Lastly, when imputation did not utilize an intermediate imputation step, (1K Bulls) accuracies were much lower with concordance, IQS, and *r* values of .8309, .5465, and .6823, respectively. The stark contrast in these values compared to the other approaches highlights the significant advantage of including higher-density reference panels or intermediate steps in improving imputation accuracy. All of the accuracy metrics in **Table 2.2** were calculated on a per SNP basis and then averaged to best capture the true accuracy values.

Imputation Accuracy per Animal

Much like what is demonstrated in **Table 2.2**, the approach using the COMBO reference as an intermediate imputation step provided the highest accuracy sequence-density imputation across individuals. This bears out across individual imputation accuracies where the COMBO imputation generated the highest accuracy for more than half of the animals (**Table 2.3**). The next most accurate was the HD approach, with about a third of all individuals having their best imputation accuracies via that approach. For the COMBO intermediate imputation reference, this is likely due to its ability to more accurately impute for rare variants in the sequence reference as a result of the rare composition of the F250 array and the increased diversity represented in the reference panel. The cross-imputation needed for the COMBO reference may have introduced some errors as compared to only the HD reference, which did not require this step. The most important attribute of the COMBO panel is likely that it contained the largest number of animals. Despite many of these animals being genotyped on the lower-density F250 array, the huge

Table 2.3 Intermediate Imputation References Accuracy Metrics per Individual

	HD			COMBO			1K Bulls		
Individual	Concordance	IQS	Correlation	Concordance	IQS	Correlation	Concordance	IQS	Correlation
ABHA_utia-0152_ARS20	0.97	0.87	0.96	0.97	0.88	0.97	0.85	0.63	0.81
ABHA_utia-0153_ARS20	0.97	0.84	0.97	0.97	0.87	0.97	0.84	0.63	0.82
ABHA_utia-0154_ARS20	0.93	0.72	0.92	0.92	0.68	0.92	0.84	0.57	0.81
ABHA_utia-0155_ARS20	0.93	0.76	0.93	0.93	0.79	0.93	0.85	0.62	0.82
ABHA_utia-0156_ARS20	0.91	0.72	0.90	0.92	0.76	0.90	0.84	0.62	0.80
AHA_utia-0220_ARS20	0.98	0.68	0.96	0.98	0.67	0.96	0.91	0.55	0.85
AHA_utia-0221_ARS20	0.95	0.78	0.95	0.94	0.77	0.94	0.85	0.59	0.82
AHA_utia-0222_ARS20	0.95	0.83	0.94	0.95	0.80	0.94	0.88	0.70	0.83
AHA_utia-0223_ARS20	0.95	0.61	0.91	0.94	0.60	0.90	0.89	0.52	0.79
AHA_utia-0224_ARS20	0.90	0.64	0.86	0.90	0.61	0.86	0.87	0.60	0.80
ASA_jb-1123_ARS20	0.85	0.59	0.83	0.85	0.61	0.84	0.79	0.57	0.76
ASA_jb-1124_ARS20	0.92	0.83	0.92	0.92	0.82	0.92	0.80	0.62	0.78
ASA_jb-1125_ARS20	0.95	0.88	0.95	0.96	0.89	0.96	0.84	0.68	0.83
ASA_jb-1126_ARS20	0.93	0.83	0.93	0.94	0.84	0.94	0.81	0.62	0.80
ASA_utia-0107_ARS20	0.95	0.89	0.95	0.96	0.90	0.96	0.81	0.65	0.82
ASA_utia-0108_ARS20	0.94	0.86	0.95	0.94	0.86	0.95	0.82	0.64	0.79
ASA_utia-0109_ARS20	0.96	0.88	0.96	0.96	0.88	0.97	0.84	0.65	0.83
ASA_utia-0110_ARS20	0.88	0.69	0.88	0.88	0.67	0.88	0.81	0.61	0.79
ASA_utia-0111_ARS20	0.91	0.78	0.91	0.91	0.78	0.91	0.83	0.65	0.82
ASA_utia-0112_ARS20	0.91	0.81	0.92	0.92	0.81	0.92	0.80	0.61	0.80
ASA_utia-0113_ARS20	0.96	0.91	0.96	0.96	0.90	0.96	0.83	0.67	0.83
ASA_utia-0114_ARS20	0.89	0.76	0.90	0.89	0.74	0.89	0.81	0.62	0.80
ASA_utia-0115_ARS20	0.93	0.83	0.94	0.94	0.85	0.94	0.85	0.68	0.84
ASA_utia-0116_ARS20	0.92	0.83	0.92	0.94	0.86	0.94	0.83	0.66	0.84
ASA_utia-0148_ARS20	0.96	0.90	0.97	0.97	0.92	0.98	0.85	0.70	0.85
ASA_utia-0149_ARS20	0.94	0.82	0.94	0.94	0.81	0.94	0.83	0.65	0.83
ASA_utia-0150_ARS20	0.97	0.88	0.98	0.98	0.89	0.98	0.87	0.69	0.88
ASA_utia-0151_ARS20	0.96	0.88	0.96	0.96	0.89	0.96	0.86	0.71	0.86
SXF_calf-1_ARS20	0.89	0.68	0.90	0.89	0.71	0.90	0.81	0.61	0.81
SXF_calf-2_ARS20	0.93	0.86	0.94	0.95	0.88	0.96	0.80	0.62	0.80
SXF_calf-3_ARS20	0.88	0.72	0.89	0.88	0.72	0.88	0.79	0.58	0.77
SXF_h041_ARS20	0.93	0.87	0.94	0.94	0.87	0.95	0.80	0.63	0.82
SXF_j1035_ARS20	0.93	0.85	0.94	0.93	0.84	0.94	0.81	0.63	0.82
SXF_j1041_ARS20	0.92	0.83	0.93	0.91	0.75	0.92	0.82	0.63	0.81
SXF_j1052_ARS20	0.87	0.69	0.87	0.90	0.77	0.89	0.78	0.56	0.76
SXF_z204_ARS20	0.87	0.68	0.87	0.88	0.71	0.88	0.79	0.58	0.78

This table provides the concordance, IQS, and correlation values per individual for four imputation strategies: HD, COMBO, and 1K Bulls. It allows for a detailed comparison of imputation performance at an individual level across different imputation strategies.

increase in haplotypes represented resulted in improved downstream imputation. The animals who performed best with the HD intermediate imputation reference likely had fewer rare variants or were more related to individuals genotyped on the HD panel. The HD reference panel had the second most total animals. 1K Bulls was the lowest across all animals, as there was no intermediate imputation prior to sequence-level.

This pattern is also reflected in **Figure 2.2**, which shows boxplots of the imputation accuracy metrics (Concordance, IQS, and Correlation) for 1K Bulls, COMBO, and HD. The COMBO and HD panels exhibit higher median accuracy values with more variability compared to 1K Bulls, which shows lower medians and more significant outliers. These results further emphasize the consistency and reliability of the COMBO and HD panels, with COMBO consistently outperforming across all metrics. In contrast, the 1K Bulls panel's larger spread and lower accuracy highlight its limitations.

All of this is shown in **Table 2.3**, demonstrating that certain intermediate imputation references perform better with certain individuals. For all the accuracy metrics displayed, after the final step of imputation where the initial HD or another reference was used and then followed by the 1K Bulls reference panel, the COMBO reference consistently outperformed its counterparts. In regard to the range of IQS values, as seen in **Table 2.3**, the HD intermediate imputation reference ranged from 0.59 to 0.91, and the COMBO intermediate imputation reference ranged from 0.61 to 0.98. Finally, the imputation directly from 50K to sequence exhibited the lowest range of accuracy metrics, with the lowest IQS spanning from 0.52 to 0.89. The 1K Bulls imputation reference is the least consistent, with higher variability compared to COMBO and HD, suggesting that some animals may be able to be directly imputed to sequence, provided they have adequate haplotype similarity.

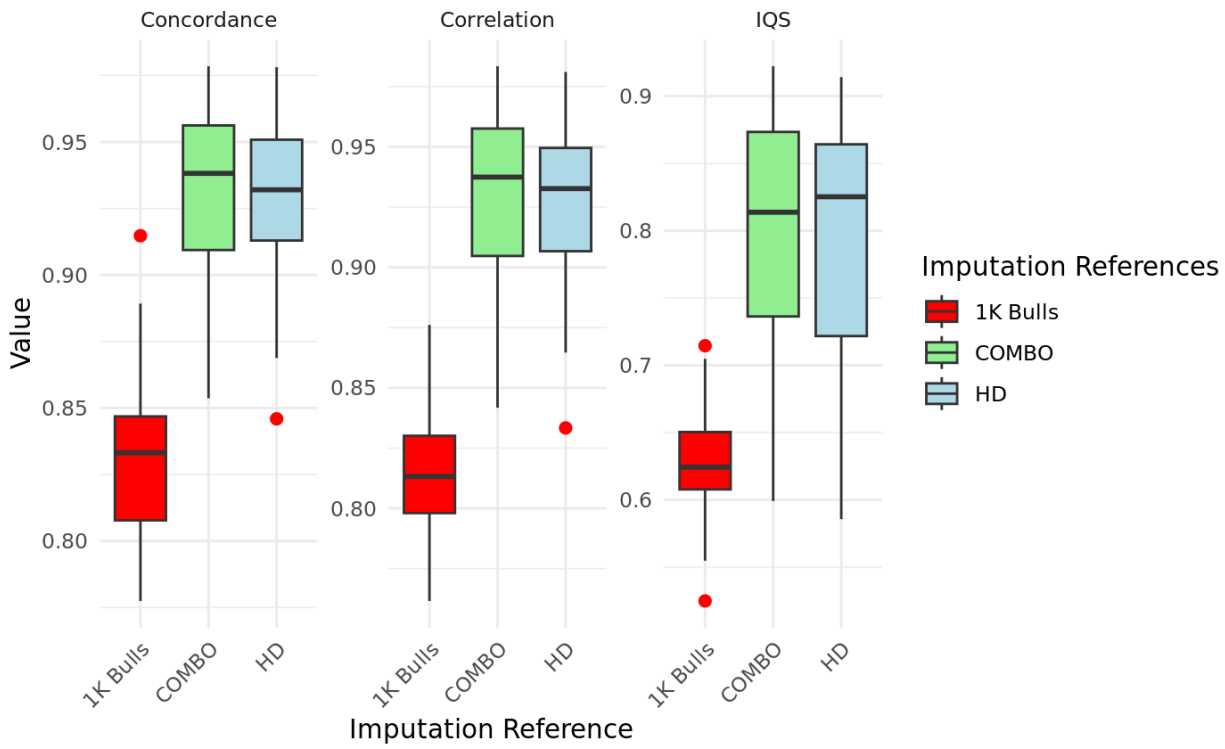

Figure 2.2 Intermediate Imputation References Accuracy Metrics Averaged per Individual
This figure presents individual-level accuracy metrics across three imputation references: 1K Bulls, COMBO, and HD. Metrics include Concordance, Correlation (r), and Imputation Quality Score (IQS), each illustrated through boxplots showing variability and central tendency. Results indicate that COMBO and HD generally yield higher and less variable accuracy values compared to 1K Bulls. These metrics were calculated on a per-individual basis, demonstrating the impact of imputation reference panel choice on imputation accuracy.

Further insights into the accuracy and variability of these panels are provided in **Table 2.4**, which summarizes the mean and standard deviation of concordance, IQS, and correlation values across the three imputation strategies. The COMBO panel demonstrated the highest mean values across all metrics, particularly excelling in IQS (mean = .8226) and correlation (mean = .9287), underscoring its consistent performance. The HD panel followed closely, with a slightly lower mean IQS (mean = .8190) but comparable correlation values (mean = .9266), affirming its robustness for common variants. In contrast, the 1K Bulls panel displayed the lowest mean values across all metrics, with the IQS mean at only .6266, highlighting its limitations as a standalone reference. **Table 2.5** further underscores the statistical significance of these differences, with pairwise comparisons revealing that COMBO significantly outperformed 1K Bulls across all metrics ($p < 0.001$). While the comparisons between HD and COMBO showed similar results for IQS ($p = 0.4644$) and correlation ($p = 0.0695$), the HD and 1K Bulls panels exhibited significant differences, emphasizing the reduced reliability of the 1K Bulls reference. Collectively, these results reinforce the advantages of the COMBO and HD panels, with the COMBO reference consistently achieving the highest accuracy and reliability across all metrics.

Imputation Accuracy dependent on MAF

Figures 2.3 and **2.4** illustrate the relationship between Imputation Quality Score (IQS) and the imputation strategy. The COMBO panel consistently outperformed the other reference panels, maintaining the highest IQS values across all ranges furthering proving its value in imputing low MAF variants. The HD panel followed closely, exhibiting similar performance for common variants but slightly reduced accuracy for those with a low MAF. In contrast, the 1K

Table 2.4 Mean and Standard Deviation (SD) of Intermediate Imputation References Accuracy Metrics per Individual

	Concordance		IQS		Correlation	
Assay	Mean	SD	Mean	SD	Mean	SD
HD	0.9286	0.0321	0.8190	0.0714	0.9266	0.0341
COMBO	0.9311	0.0319	0.8226	0.0740	0.9287	0.0344
1K Bulls	0.8310	0.0306	0.6266	0.0437	0.8127	0.0261

This table summarizes the mean and standard deviation (SD) of imputation accuracy metrics—concordance, imputation quality score (IQS), and correlation—across different intermediate imputation references: HD, COMBO, and 1K Bulls. These metrics were calculated to assess the performance of each reference panel for improving imputation accuracy on a per individual basis.

Table 2.5. P Values of Intermediate Imputation References Accuracy Metrics per Individual

Comparison	Concordance	IQS	Correlation (*r*)
COMBO vs 1K Bulls	1.41E-22	2.41E-17	4.87E-27
HD vs COMBO	0.0233	0.4644	0.0695
HD vs 1K Bulls	3.11E-23	5.76E-18	4.53E-27

This table presents p-values for pairwise comparisons of imputation accuracy metrics (concordance, IQS, and correlation) between different intermediate imputation reference panels. Statistically significant differences are observed in comparisons involving 1K Bulls, which shows the lowest p-values, indicating substantial differences in imputation accuracy compared to the other panels.

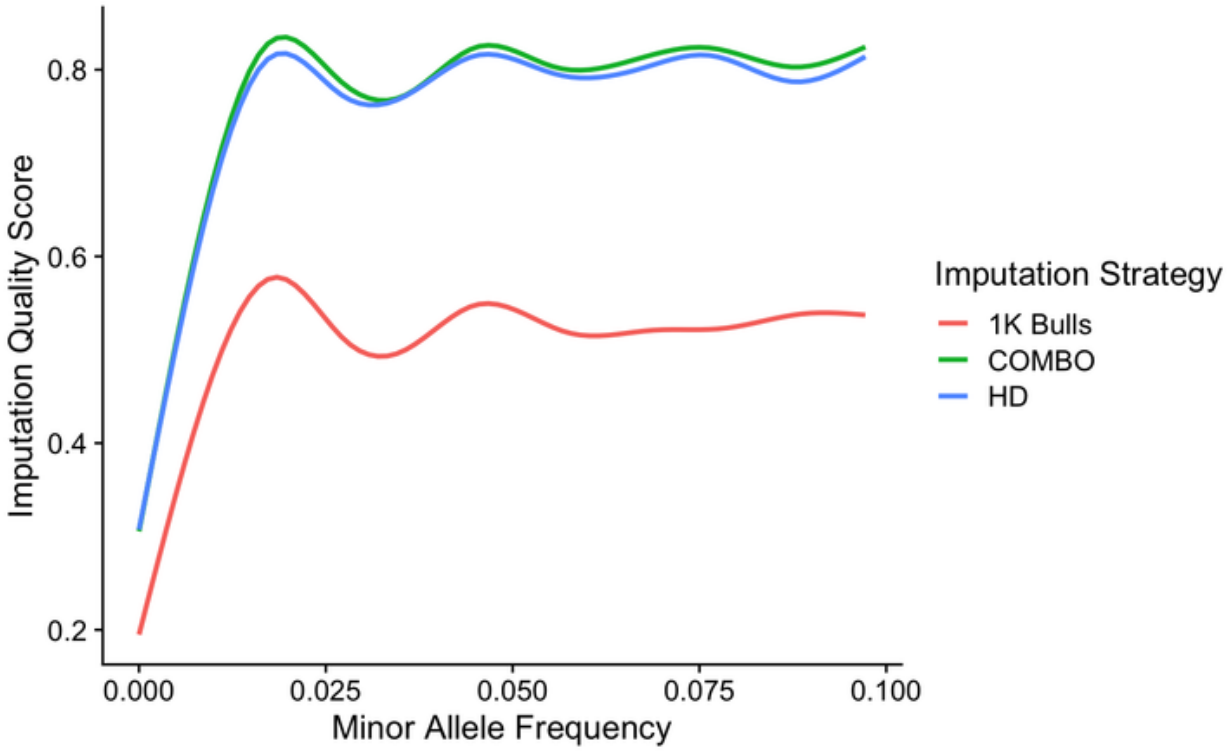

Figure 2.3. IQS & MAF (0 - 0.1) for Different Intermediate Imputation References.
This figure illustrates the relationship between Minor Allele Frequency (MAF) and Imputation Quality Score (IQS) for three imputation strategies: 1K Bulls, COMBO, and HD. The x-axis represents the MAF values ranging from 0.00 to 0.10, while the y-axis shows the IQS. The data indicates that both COMBO and HD panels exhibit significantly higher IQS compared to the 1K Bulls reference across all MAF values. HD and COMBO curves overlap slightly, suggesting similar performance, particularly at higher MAF values, whereas 1K Bulls demonstrate lower imputation quality with more variability.

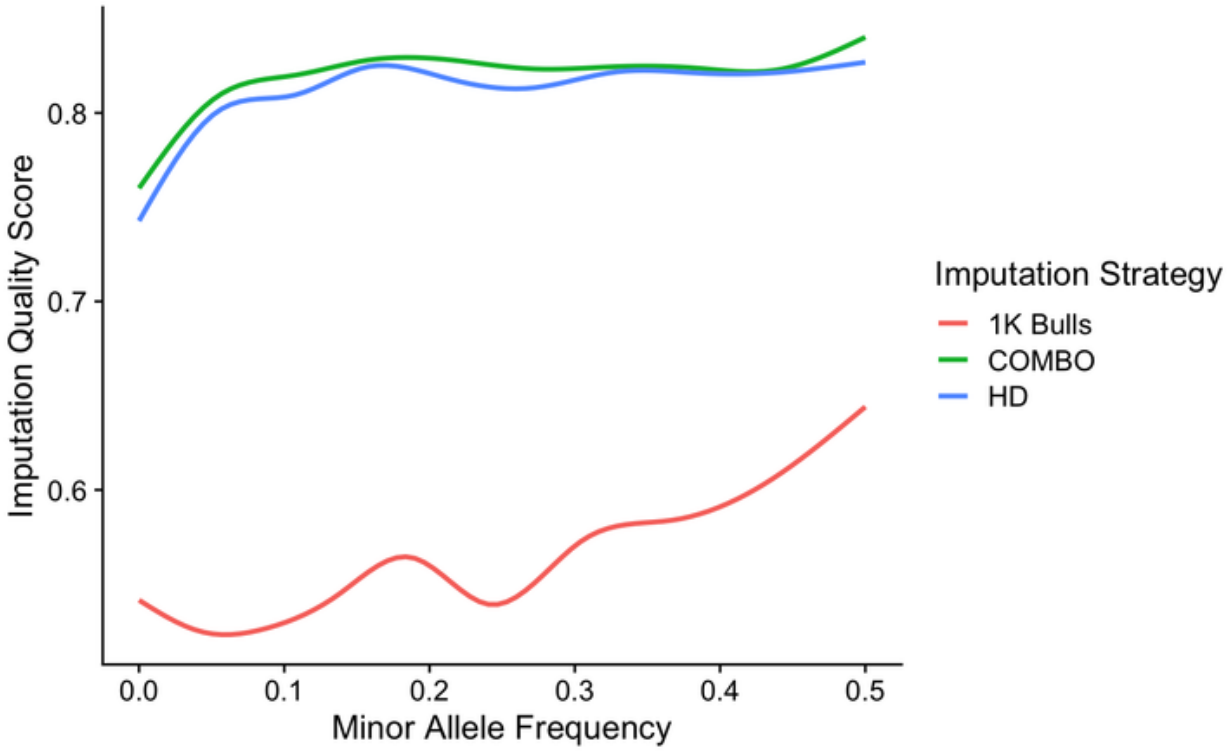

Figure 2.4. IQS & MAF (0 - 0.5) for Different Intermediate Imputation References.
This figure demonstrates the relationship between Minor Allele Frequency (MAF) and Imputation Quality Score (IQS) across three imputation strategies: 1K Bulls, COMBO, and HD. The x-axis represents MAF values (0.0–0.5), and the y-axis displays IQS. The COMBO and HD strategies outperform 1K Bulls across the MAF spectrum, exhibiting higher IQS values with minimal variation. COMBO and HD have nearly identical performance, particularly at lower MAFs, maintaining high accuracy across all allele frequencies. In contrast, 1K Bulls shows substantially lower IQS, particularly for low and intermediate MAF values, reflecting its reduced effectiveness as an imputation reference panel.

Bulls panel demonstrated the lowest IQS values, with significant variability, highlighting its limitations when used as the only reference panel.

Figures 2.5 and **2.6** show the correlation (r) between imputed and true genotypes for the same reference panels. Once again, the COMBO panel achieved the highest and most stable correlation values, outperforming the HD and 1K Bulls panels across all categories. The HD panel displayed strong performance for common variants but experienced a decline in correlation for rarer variants. The 1K Bulls panel performed poorly, showing substantial declines in correlation and increased variability, particularly in scenarios where accuracy was most critical. These findings underscore the advantages of the COMBO panel, which consistently achieved higher accuracy and reliability across metrics. The HD panel also performed well but showed slightly reduced accuracy compared to COMBO in some cases, particularly at lower MAF. By comparison, the 1K Bulls panel struggled to provide reliable results, particularly in scenarios requiring precise imputation.

In summary, the COMBO panel demonstrated superior performance across all metrics, confirming its robustness and reliability in improving imputation accuracy, especially for low MAF variants. The HD panel offered competitive results but was slightly less effective for low MAF variants likely due to the lack of a F250 reference panel. The 1K Bulls panel, while functional, showed substantial limitations for low MAF variants being the least accurate across all metrics further indicating the need to perform intermediate imputation prior to reaching the sequence level.

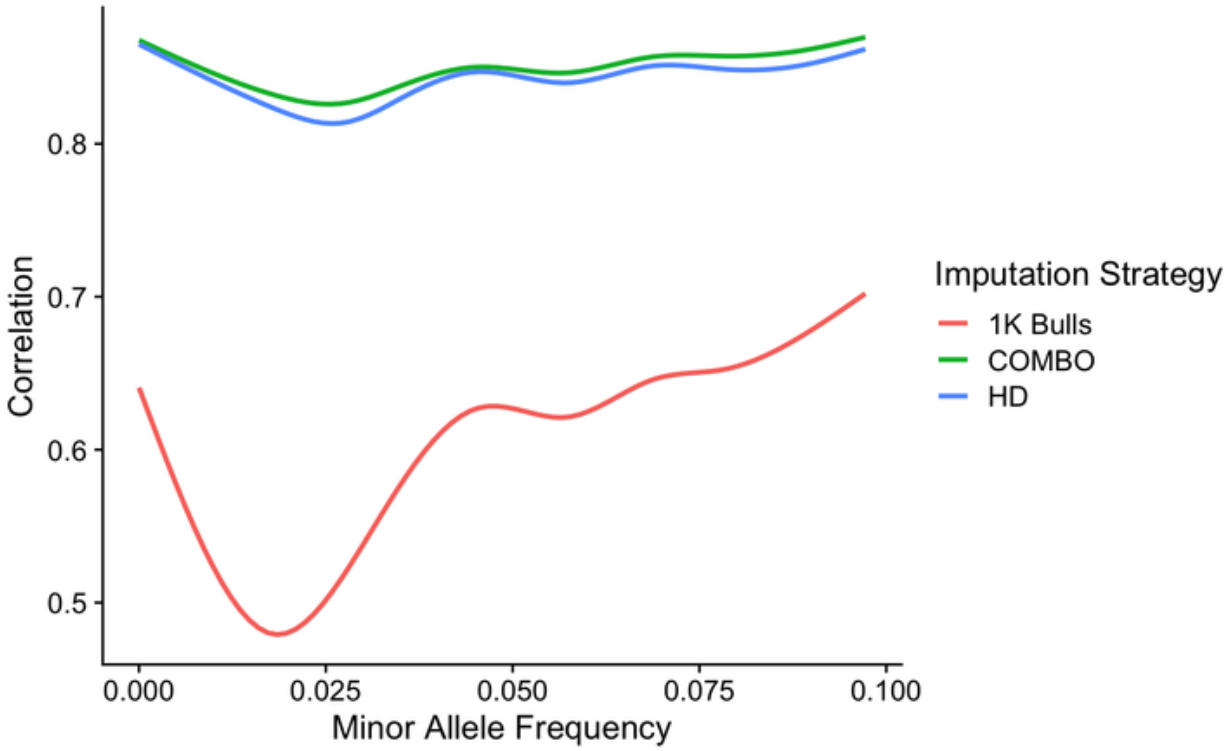

Figure 2.5. r & MAF (0 - 0.1) for Different Intermediate Imputation References.
This figure illustrates the correlation (*r*) between imputed and true genotypes as a function of Minor Allele Frequency (MAF) for three imputation strategies: 1K Bulls, COMBO, and HD, focusing on the MAF range of 0 to 0.1. The COMBO and HD strategies exhibit consistently high correlation values with minimal variability, outperforming the 1K Bulls reference, which shows a significant dip in correlation for very low MAF values. This highlights the limitations of 1K Bulls for rare allele imputation and the robustness of COMBO and HD for accurately predicting genotypes across the low MAF spectrum.

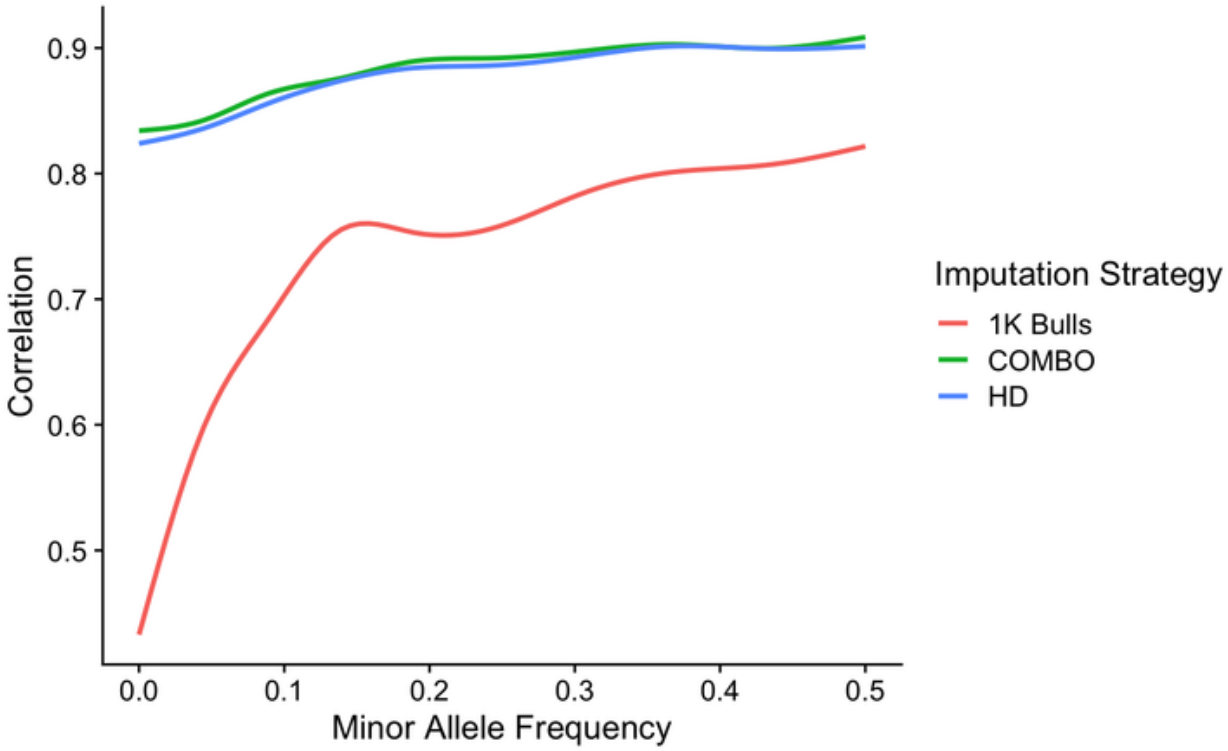

Figure 2.6. r & MAF (0 - 0.5) for Different Intermediate Imputation References.
This figure expands the analysis of correlation (*r*) across a broader MAF range (0 to 0.5) for the same three imputation strategies. The COMBO and HD panels maintain high and stable correlation values across all MAF intervals, with minimal performance differences between the two. In contrast, the 1K Bulls panel shows a sharp increase in correlation with increasing MAF, reaching levels comparable to COMBO and HD only at higher frequencies. These results underscore the superior performance of COMBO and HD panels across all MAF ranges, particularly for low-frequency variants.

Comparing r & IQS Across Different Intermediate Imputation References

The results illustrated in **Figures 2.7**, **2.8**, and **2.9** compare the relationship between *r* and IQS across three reference panels: HD, COMBO, and 1K Bulls. These figures provide insights into the performance and robustness of each imputation strategy, with a focus on how well they handle variants of varying MAF. The observed trends emphasize the importance of using intermediate imputation steps and diverse reference panels to achieve higher imputation accuracy, particularly for rare variants. The COMBO panel consistently delivered the best results, followed by the HD panel, with 1K Bulls lagging significantly in performance. These results are consistent with prior analyses indicating that intermediate imputation, particularly with diverse panels like COMBO, which improves accuracy by providing a broader range of haplotypes and better MAF representation.

Discussion

Our findings indicate that the imputation accuracy for SNP chip-derived genotypes are highly dependent on the intermediate imputation reference panel, especially for low MAF variants. The reference panel COMBO, which included animals genotyped on the Bovine HD and the GGP-F250 arrays consistently provided the highest imputation accuracies across all metrics: concordance rate, IQS, and correlation. This is likely due to the inclusion of rare variants captured within the F250 array, and a large increase in haplotype representation. By performing intermediate imputation using dense reference panels like HD or F250, the accuracy of imputing to sequence level accuracy is increased (Rowan et al., 2019). The inclusion of F250 helps capture more genetic diversity especially for regions with low MAF. The other primary

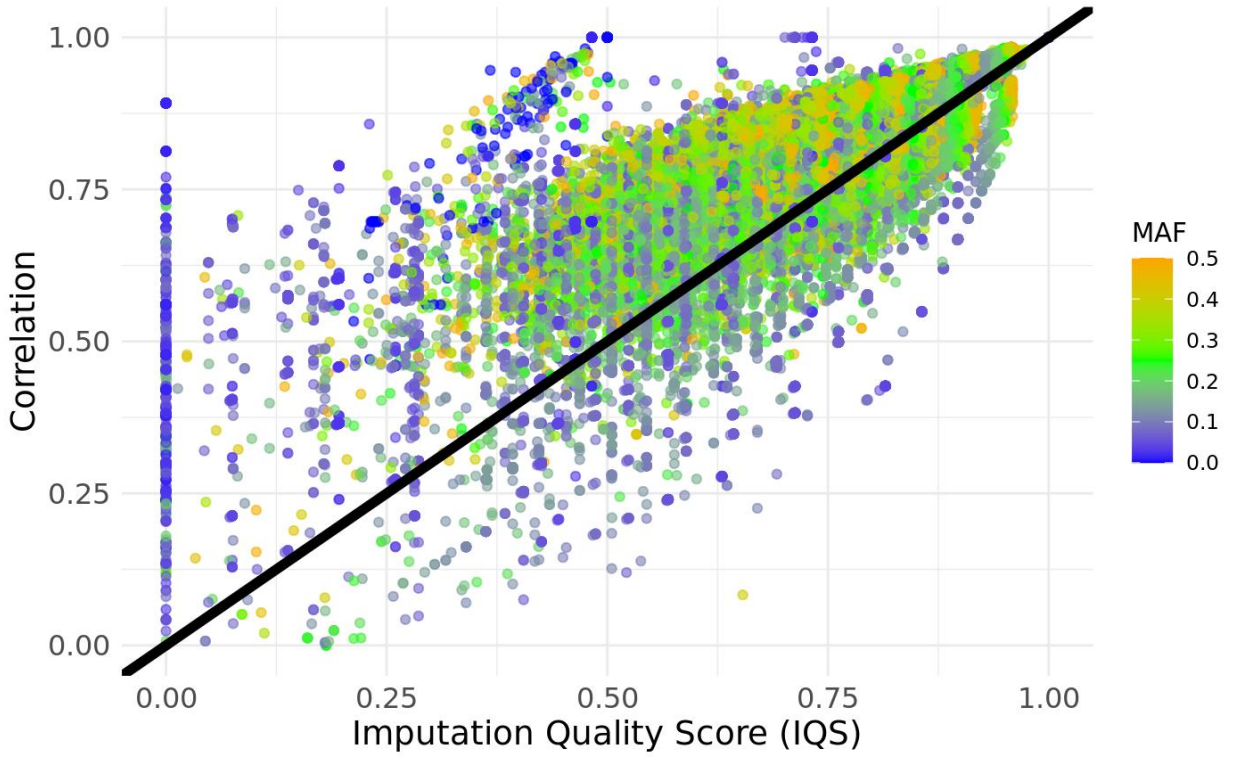

Figure 2.7. HD: r vs. IQS

The HD reference panel demonstrated a strong positive association between r and IQS, as seen in the dense clustering of points near the diagonal reference line. Higher IQS values consistently corresponded to higher correlation, particularly for moderate and high MAF variants. However, deviations from the reference line at lower IQS values suggest reduced accuracy for rare variants, as indicated by the more scattered distribution of blue points (representing low MAF). This aligns with expectations that HD panels, while dense, may have limitations in handling rare variants due to their reliance on high-frequency SNPs.

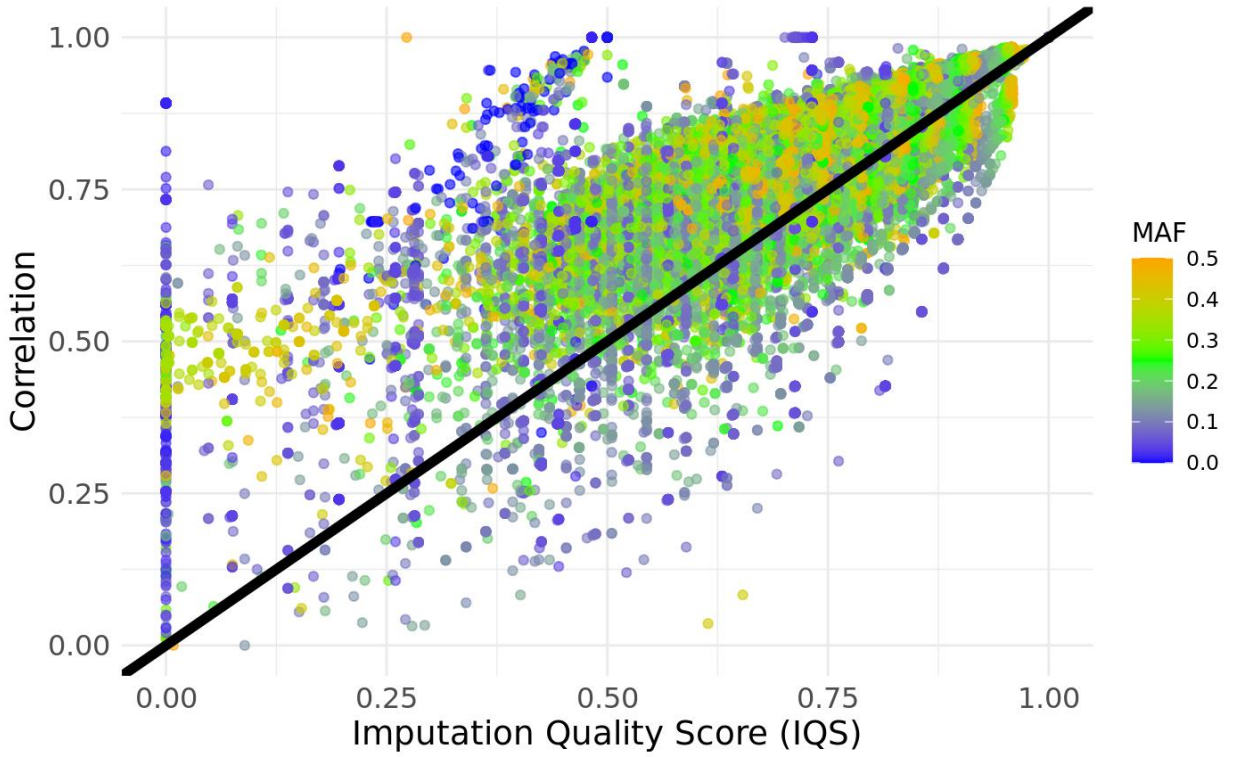

Figure 2.8. COMBO: r vs. IQS

The COMBO reference panel, which integrates HD and F250 data, outperformed the HD-only panel by achieving higher IQS and correlation values across all MAF levels. Points representing rare variants (blue) were more tightly clustered along the diagonal line, indicating improved imputation accuracy for low MAF alleles. This result highlights the benefit of combining reference panels, as the inclusion of F250 data introduces additional genetic diversity and better representation of rare variants. The overall density of points along the reference line further supports the robustness of the COMBO panel.

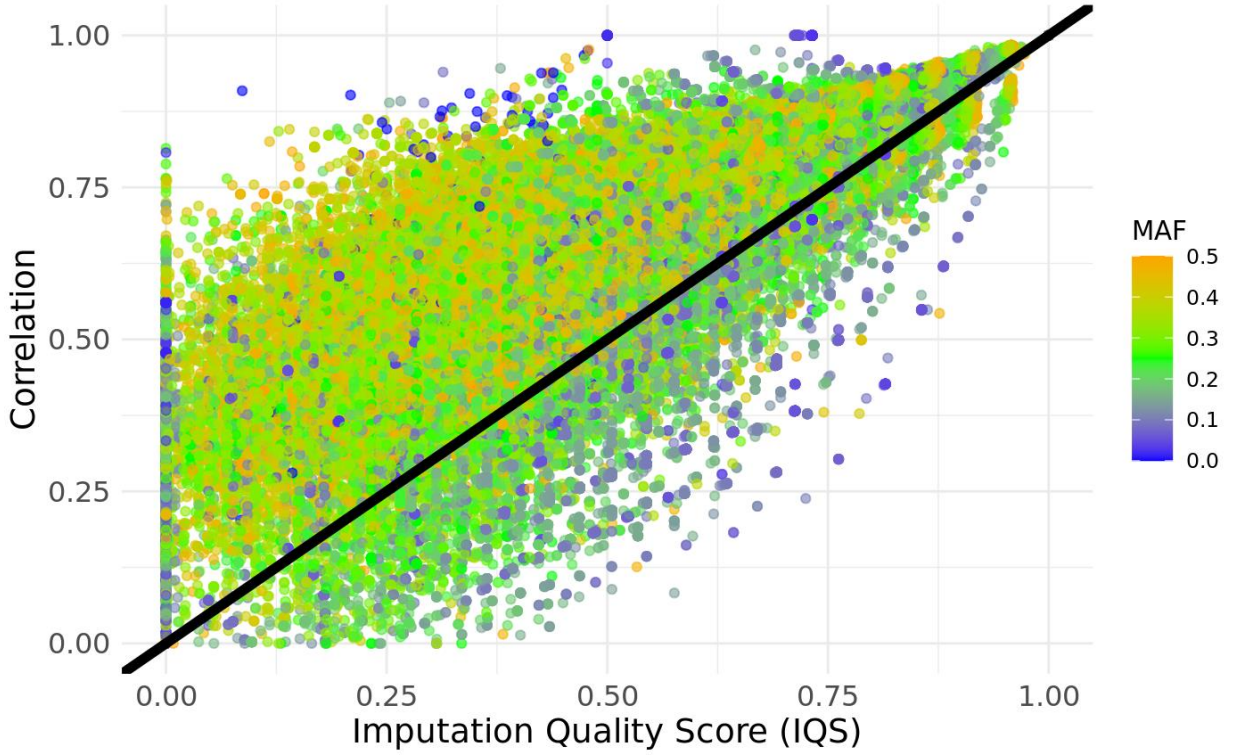

Figure 2.9. 1K Bulls: r vs. IQS

The 1K Bulls reference panel exhibited the weakest performance among the three, with a broader and less dense scatter of points below the diagonal reference line. While high MAF variants (orange and green points) showed moderate imputation accuracy, rare variants were poorly imputed, as evidenced by the significant deviations for low IQS values. This finding underscores the challenges associated with single-step imputation strategies, particularly when intermediate imputation steps are omitted.

reason for the increased accuracy in the COMBO panel is the presence of more animals, which allows for IMPUTE5 to have more haplotypes to choose from for performing imputation. Additionally, as the results showed, the 1K Bulls reference panel alone had the lowest performance in terms of accuracy across all metrics. This is expected, as this imputation references lacked the intermediate imputation steps that helped refine the genotype information before imputing to whole-genome sequence data. Proving that panels that skip intermediate imputation often struggle with imputing low MAF variants.

The performance differences between each of the intermediate imputation references further highlights the importance of selecting an appropriate reference panel depending on the population to be imputed. By ensuring imputation is able to properly handle low MAF, associations with complex traits can be identified. Utilizing a combined reference panel within the intermediate imputation step increases overall imputation accuracy. We would expect further improvements to imputation to be driven mostly by the inclusion of more individuals in the sequence reference panel, particularly from populations that are related to the one being imputed. The use of an intermediate reference panel during imputation prior to performing a GWAS should allow for the more accurate detection of causal variants for complex traits. By ensuring rarer variants are properly imputed, there would be a decreased chance they are missed as the causal variant for a complex trait.

Conclusion

This study highlights the importance of intermediate imputation and careful reference panel selection in enhancing genotype imputation accuracy. Our findings demonstrate that using a combined reference panel of higher density intermediate imputation references, such as COMBO, significantly improves imputation performance, particularly for rare variants. These

results underscore the value of leveraging diverse reference panels and intermediate steps to achieve reliable imputation outcomes, which is crucial for accurately detecting associations with complex traits. Adopting these strategies can lead to more precise genomic analyses, ultimately advancing the field of genomic prediction and association studies.

Chapter Three: Uncovering the genetic basis of cleft palate in Boer goats

Abstract

Cleft palate is an autosomal recessive genetic deformity that can occur in several species, it is the result of the palate of the mouth not fully forming during gestation. This leaves an opening in the mouth that could extend to the nasal cavity, potentially causing the animal to be unable to eat or drink properly. This typically leads to the breeder euthanizing the animal shortly after birth. The objective of this study was to determine the genetic cause of cleft palate in Boer goats so a diagnostic tool could be developed that allows producers to select against it. Fifteen blood samples (n=15) were collected from three Boer goat herds. The samples represented four trios and two quads that included sire, dam, and affected kid(s). All affected kids were sired by the same buck. Genomic DNA was extracted and used to generate libraries for whole genome sequencing. An average of 179.7 million reads were generated across all samples (range from 151.2 million to 224.4 million reads) resulting in approximately 10X genome coverage per sample. After sequencing, quality filtering was performed using a Nextflow pipeline modeled after the germline short variant discovery best practices workflow from the Genome Analysis Toolkit (GATK) software. After filtering, reads were mapped to the ARS_v1.2 reference genome. The pipeline called 18,867,792 variants across the 15 samples. Variants were filtered using Python based on an autosomal recessive mode of inheritance (Van Rossum and Drake, 2009). After variants were found that matched the assumed inheritance pattern Ensembl Variant Effect Predictor (VEP) was used to identify 20 potentially causative variants in which some were found to be in genes HDAC9, ZBTB20, and MEOX2 which have been found to cause craniofacial abnormalities.

Introduction

A total or partial opening of the roof of the mouth, or palate, is known as cleft palate. The palate is the structure that separates the nasal and oral cavities, making it essential for proper breathing, eating, and drinking. When cleft palate occurs in livestock animals they are typically euthanized shortly after birth. In humans, they are among the most common birth defects occurring at a frequency of about 1 in every 500 – 2500 live births (Vanderas 1987; Schutte and Murray 1999; Dixon et al. 2011; Mangold, Ludwig, and Nöthen 2011). There are several methods of classifying cleft palate with the Veau classification focusing on morphological and anatomical characteristics with Veau I being clefts of the soft (muscular) palate and Veau II being the occurrence of cleft palate on the hard (bony) and soft palate. Veau III is also a cleft of both the hard and soft palate but it extends unilaterally through the gums with Veau IV extending bilaterally through the alveolus (Houkes et al. 2023). Cleft palate is a common occurrence due to the complex nature of the formation of the palate, the process of which is known as palatogenesis.

Mammalian palatogenesis is a meticulously regulated morphogenetic process that starts with the merging of three components: the primary palate derived from the frontonasal process and the secondary palate consisting of the two lateral maxillary palatal shelves. The formation of the secondary palate is a complex process starting with the growth of the palatal shelves, their elevation, the fusion of the paired shelves, and the eventual disappearance of the midline epithelial seam. The process is initiated with the formation of the palatal shelf primordia, which begins with mesenchymal cell proliferation within the maxillary processes. This proliferation leads to the appearance of the palatal shelves, which grow vertically beside the tongue. The shelves then undergo a rapid elevation to a horizontal position above the tongue, facilitating their

fusion. The medial edge epithelium of the palatal shelves then fuses at the midline epithelial seam, creating a continuous palate that separates the oropharynx from the nasopharynx (Zhang et al. 2002; Bush and Jiang 2012; Lan et al. 2015). The complex process behind the formation of the palate further illustrates the role that proper cellular maintenance plays in its success. Due to the abnormal communication between the nasal and oral cavities causing cleft palate, breathing and feeding deficits occur, with auditory deficits occurring in humans (Levi et al. 2011). For humans, surgery is often used to correct the condition, but for livestock the procedure can be costly and difficult.

Cleft palate in sheep and goats can also result from certain environmental exposures during gestation. Studies have shown that cleft palate can be induced in fetal goats when their dams consume wild tree tobacco (*Nicotiana glauca*) during day 35 – 41 of gestation. In one study performed by Panter and others they found that goats were more susceptible to cleft palate formation when compared to sheep with a 3% occurrence of the condition as compared to 100% in goats (Panter et al. 2000). The higher susceptibility in goats is believed to be due to an alkaloid-induced reduction in fetal movement during the period of palate closure. Ultrasound imaging of fetuses in does fed with Nicotiana glauca during this critical period showed little space between the chin and sternum due to a tight flexure of the head and neck. This contrasts the usual extension of the head and neck at days 35 – 38 of gestation in goats without cleft palate. The hyper flexed state of the fetal head and neck inhibits the movement of the tongue which explains the mechanical mechanism of alkaloid-induced cleft palate formation in goats (Panter et al. 2000; Panter and Keeler 1992). This proposed mechanism is similar to theories explaining the pathogenesis of cleft palate in Pierre Robin syndrome (PRS). PRS is a condition in humans caused by mispositioning, which prevents the tongue from descending from the nasal

cavity, thereby inhibiting the fusion of the palate at the midline (Rintala et al. 1984). Due to the occurrence of a similar condition in humans when compared to goats it can be believed that the genes that impact the condition are similar between the two. Goats were determined to be a model organism for humans by Weinzweig and others as they found that similar defects occur in goats, which further supports that genes that impact craniofacial development in humans are likely to serve similar functions in goats (Weinzweig's and Weinzweig 2017).

The goat's role as a model organism in humans led to the decision to prioritize genes known for their role in craniofacial or embryonic development in humans and other species. Mice have also been used as a model organism further demonstrating that the genetic and molecular pathways involved in palatogenesis are conserved across species (Houkes et al. 2023). Genes such as IRF6, MSX1, and PAX7 are crucial for craniofacial development in humans and have similar roles in other mammals, including sheep and goats. Understanding these genetic factors is essential for helping to select against the deformity in livestock species and can help inform human medical research (Xu et al. 2016).

Several genes have been found to impact craniofacial, embryonic, or skeletal development. Those include the Sonic Hedgehog (SHH) gene plays a significant role in the formation of the palate because its signaling is vital for maintaining the expression of Bone Morphogenetic Protein 2 (BMP2) and Forkhead box (FOX) family transcription factors. Disruption in SHH signaling can lead to downregulation of BMP2 which serves a crucial function in the growth of the anterior palatal shelves, additionally mice lacking FOXF2 exhibited cleft palate further illustrated the role SHH plays in palate development (Han et al. 2021). Another gene with known functions in craniofacial development is Twist Family BHLH Transcription Factor 1 (TWIST1), a helix-loop-helix transcription factor. TWIST1 will interact

with RUNX2 to coordinate the development of the cranial neural crest-derived cells essential for palate myogenesis, the haploinsufficiency of TWIST1 in the presence of a RUNX2 deficiency can help correct cleft palate (Satokata and Maas 1994). Msh Homeobox 1 (MSX1) disruption can lead to insufficient proliferation and differentiation of cranial neural crest cells resulting in the failure of palatal shelves to grow and fuse properly. Loss of MSX1 function leads to reduced BMP4 expression further impairing usual palate formation, causing cleft palate (van den Boogaard et al. 2000; Funato et al. 2009). This loss of MSX1 function is due to TGF-β signaling, which is crucial for the epithelial-mesenchymal transformation (EMT) during palatal fusion, where it mediates the removal of the medial edge epithelium of the palatal shelves (Bush and Jiang 2012; Lan et al. 2015). When Heart And Neural Crest Derivatives Expressed 2 (HAND2) is up or downregulated it will interact with RUNX2 negatively by inhibiting its DNA binding activity causing abnormalities in bone development. Proper inhibition of RUNX2 by HAND2 ensures that the differentiation of osteoblasts occurs at the right time (Bronner and Quail 2019). Due to the complex nature of palatogenesis there are several genes that could potentially impact it causing cleft palate, those above being some of the more intensively studied genes definitively known to cause cleft palate.

The presence of cleft palate is highly deleterious for goats and must be selected against. That is why a genetic test should be formed to allow breeders to select against it. However, first the genetic cause of cleft palate must be determined. We hypothesized that this deformity is hereditary following an autosomal recessive inheritance pattern, and we undertook genetic analysis to identify the causal mutation(s) of cleft palate.

Materials and Methods

Sample Collection

Blood samples were collected from a total of fifteen Boer goats (n=15) from three show herds which consisted of four trios and two quads. Each trio consisted of an unaffected sire, unaffected dam and an affected kid, quads included an unaffected sire, unaffected dam, and two affected kids. All the affected kids were sired by the same buck as shown in **Figure 3.1**.

DNA Extraction and Sequencing

Extraction and sequencing were performed by the University of Tennessee's Ge-nomics Core. Libraries were prepped with the Illumina DNA Prep Tagmentation kit at one fourth reaction volume. Tagmentation fragments DNA and tag it with adapter se-quences [20, 21]. It and subsequent library preparation steps were automated using the Dispendix G Station NGS workstation which includes the I.DOT liquid handler and C.WASH station to ensure human error could not impact results. The final libraries were then pooled and sequenced on the Illumina NovaSeq platform using a 300 cycle S4 flowcell at 2 x 150 base pairs for paired end reads.

Variant Calling

Once sequencing was complete the raw reads were fed into a variant calling Nextflow pipeline. Reads were first trimmed using fastp to help remove low quality reads and adapter sequences (Chen et al., 2018). The reads were then aligned to the reference genome ARS1.2 using bwa-mem2 (Vasimuddin et al., 2019). Alignments are sorted and indexed using samtools and variants are called using Genome Analysis Toolkit (GATK) based on their best practices workflow (Danecek et al., 2021; McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et

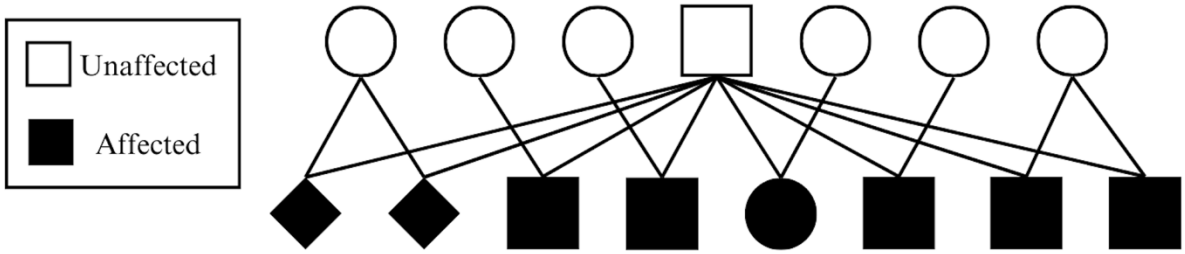

Figure 3.1. Family Pedigree of the Animals in the Dataset

This pedigree chart (Figure 3.1) shows the family lineage of the animals in the dataset, distinguishing between affected and unaffected individuals.Squares indicate males, circles indicate females, and diamonds represent individuals of unknown sex, helping to visually represent the inheritance patterns and family structure within the dataset.

al., 2013). The processes performed in GATK are as follows: GATK HaplotypeCaller was utilized in gvcf mode, the gvcf files were then combined using GenomicsDBImport, and lastly joint genotyping was performed using the GenotypeGVCFs command.

Filtering for Candidate Variants

After variant calling was completed, several filtering methods were utilized to best identify the potentially causative variant(s). Variants were filtered assuming an autosomal recessive mode of inheritance, where the sire and dam are heterozygous, and the kid is non reference homozygous. To account for errors in genotype calls or read depth differences, when variant filtering was performed only 13 animals (n – 2) were required to pass all filters. The kids were required to be homozygous non reference and the sire and dams were required to be heterozygous, both filters were applied using Python (Van Rossum and Drake, 2009). From there regions of interest were identified by determining runs of homozygosity (ROH) in the affected kids utilizing PLINK (Purcell et al., 2007). ROH regions were used to determine regions of interest due to their implications on haplotype inheritance patterns as regions that are conserved across many animals can be indicative of shared ancestry. After filtering was performed, the variants were analyzed using Ensembl Variant Effect Predictor (VEP; McLaren et al., 2016). Further analysis of the list was performed prioritizing those variants that were contained within a gene that could have an impact on craniofacial, skeletal, or embryonic development.

Results

Eight Boer goat kids presented visually with cleft palate, all with a shared sire. All kids were euthanized shortly after birth due to the presence of Veau II cleft palate affecting both the soft and hard palate (see **Figure 3.2**). Pedigree data was obtained for all affected kids and variants

Figure 3.2. Image of an unaffected and affected goat. a. Normal, unaffected. b. Affected with cleft palate.

The image highlights the physical manifestation of the condition in the affected goat, where the cleft palate is clearly visible.

were analyzed and filtered. After variant filtering was completed, 20 potentially causative variants were identified (see **Table 3.1**). The variants were further analyzed and prioritized based on their predicted VEP impact and the roles the genes could play in craniofacial, skeletal, or embryonic development. Some of the variants identified existed in genes that could impact craniofacial development similarly to MSX1, those genes being VCW2 and LTBP3. Each of the variants below were contained within a ROH which can be indicative of similar ancestry and conserved haplotypes amongst individuals. This indicates that these variants could have a significant impact on the phenotypic expression of traits related to craniofacial development, providing valuable insights into the genetic mechanisms underlying these developmental processes and the cause of cleft palate in Boer goats.

Discussion

Of the identified variants, the most probably causative are those that exist in genes zinc finger and BTB domain containing 20 (ZBTB20), histone deacetylase 9 (HDAC9), mesenchyme homeobox 2 (MEOX2), von Willebrand factor C domain containing 2 (VWC2), protocadherin 7 (PCDH7), neurotrimin (NTM), and latent transforming growth factor beta binding protein 3 (LTBP3). While most of these genes are not directly involved in craniofacial development, they could play a role in palatogenesis, embryonic, or skeletal development. While not directly involved in palatogenesis, changes in ZBTB20 have been found to cause Primrose syndrome in humans, where they have skeletal abnormalities which is likely due to ZBTB20s role in glucose metabolism, postnatal growth, and neurogenesis (Cordeddu et al., 2014; Juven et al., 2020). On the other hand, HDAC9 has been found to cause craniofacial abnormalities due to its disruption of TWIST1, a previously mentioned gene known for causing cleft palate (Hirsch et al., 2019).

Table 3.1 Potentially causative variants for cleft palate in Boer goats

Chr	Position (bp)	Reference	Alternative	Variant Annotation	Gene
1	129067721	T	*,C	intron variant, sequence alteration	NMNAT3
1	58788973	GA	G	intron & upstream variant	ZBTB20
4	93206915	A	AT,ATT	intron variant	HDAC9
4	93240287	A	G	intron variant	HDAC9
4	96482765	C	CT	intron variant	MEOX2
4	114421352	G	T	upstream gene variant	VWC2
4	114425395	T	G	upstream gene variant	VWC2
4	114437785	GTA	G	intron variant	VWC2
4	114442569	TA	T,TAA	intron variant, sequence alteration	VWC2
4	114450725	GGGCCCTGAAGGCC	G	intron variant	VWC2
4	114451322	CGCTTGTGA	C	intron variant	VWC2
4	114490555	A	G	intron variant	VWC2
4	114492652	A	G	intron variant	VWC2
4	114503876	A	ATGG	intron variant	VWC2
6	50834343	A	G	intron variant	PCDH7
6	98782802	CAA	CA,C	intron variant, sequence alteration	GPAT3
8	8313430	G	GAA	upstream gene variant	PINX1
8	8313431	C	CTGA,A	upstream gene variant	PINX1
29	34033636	GTA	G	intron variant	NTM
29	44138521	G	A	intron variant	LTBP3

The table provides information on chromosome position, reference and alternative alleles, variant annotations, and corresponding genes for each identified variant. These variants may play a role in the development of cleft palate due to their locations in or near functional gene regions.

Hirsch and others found several variants that could have an impact on transcript regulation of TWIST1, including six intronic HDAC9 variants (Hirsch et al., 2019). Another of the candidate variants is contained within MEOX2, which in mice lacking it leads to weak fusion in the posterior palate which is due to MEOX2s role in TGF-β mediated fusion of the palate during craniofacial development (Smith et al. 2012; Jin and Ding 2006). VCW2 could play a similar role in craniofacial development as MSX1 and SHH due to its interaction with BMP, while the direct impact of VCW2s interaction with BMP has not been studied in relation to craniofacial development, its interaction with the family of proteins known for leading to cleft palate in other genes does support its potential impact. Additionally, VWC2 acts as an antagonist to BMPs, particularly BMP-2 and BMP-4, by binding to these proteins and preventing their interaction with BMP receptors on the cell surface. This inhibition is crucial for ensuring proper BMP signaling during skeletal development (Zhang et al., 2007). PCDH7 could have an impact on cleft palate due to its role in cell recognition and adhesion concentrated in the head (Wang et al., 2020; Xiao et al., 2018). Another of the identified variants that is within a gene dealing with cell by cell adhesion is NTM which is involved in promoting neurite outgrowth and cell adhesion (Chen et al., 2001). Lastly, LTBP3 is a tremendously important regulatory gene due to its activation of TGF-beta which as previously highlighted is extremely important in palatogenesis. In a knockout mice model, LTBP3 caused dental abnormalities which can demonstrate its role in craniofacial development (Huckert et al., 2015).

Further analysis will need to be performed to verify the truly causative variant for cleft palate by using Sanger sequencing or Polymerase Chain Reaction (PCR). Sanger sequencing will provide accurate sequencing of the DNA region containing the variant, while PCR can amplify the specific DNA segment to facilitate detailed analysis. These methods will help in validating

the presence and impact of the identified variants. Once the causative variant is confirmed, it will pave the way for the development of a genetic test for cleft palate. A genetic test for the deformity could allow breeders to make informed breeding decisions against cleft palate. Through genetic selection against the condition, the incidence of it could become a much rarer occurrence. This can lead to less financial losses for those breeders who lose much of the kid crop to the deformity. Additional, further understanding of cleft palate in goats could help in knowledge of other livestock species allowing for the creation of genetic tests for all livestock to further prevent cleft palate.

Conclusion

This study successfully identified several genetic variants as potentially responsible for cleft palate in Boer goats by utilizing whole genome sequencing and rigorous variant filtering methods. Many of the identified variants are within genes known to impact craniofacial, embryonic, or skeletal development including HDAC9, ZBTB20, and MEOX2. Additionally, many regulatory genes such as VCW2, PCDH7, NTM, and LTBP3 were highlighted for their roles in regulation as it could relate to craniofacial development. To confirm the truly causative variant(s) further analysis using Sanger sequencing, PCR, or other genetic isolation methods need to be used. Once confirmed, a genetic test can be developed to enable breeders the ability to make informed decisions regarding selecting against cleft palate. The insights from studying cleft palate in goats can enhance our understanding of similar conditions in other livestock species or humans. Having the ability to select against cleft palate in Boer goats could help with animal welfare and production.

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of dairy science, 93(2), 743-752.

Aho, A. V., Kernighan, B. W., & Weinberger, P. J. (2023). The AWK programming language. Addison-Wesley Professional.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Basta, L. P., Sil, P., Jones, R. A., Little, K. A., Hayward-Lara, G., & Devenport, D. (2023). Celsr1 and Celsr2 exhibit distinct adhesive interactions and contributions to planar cell polarity. Frontiers in Cell and Developmental Biology, 10, 1064907.

Bolormaa, Sunduimijid, et al. "A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle." PLoS genetics 10.3 (2014): e1004198

Bush, J. O., & Jiang, R. (2012). Palatogenesis: morphogenetic and molecular mechanisms of secondary palate development. Development, 139(2), 231-243.

Chen, S., Gil, O., Ren, Y.Q. et al. Neurotrimin expression during cerebellar development suggests roles in axon fasciculation and synaptogenesis. J Neurocytol 30, 927–937 (2001). https://doi.org/10.1023/A:1020673318536

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890. DOI: 10.1093/bioinformatics/bty560.

Chowdhury, R., Wang, Y., Campbell, M., Goderie, S. K., Doyle, F., Tenenbaum, S. A., ... & Temple, S. (2021). STAU2 binds a complex RNA cargo that changes temporally with production

of diverse intermediate progenitor cells during mouse corti-cogenesis. Development, 148(15), dev199376.

Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. Genetics Selection Evolution, 42, 1-8.

Cordeddu, V., Redeker, B., Stellacci, E., Jongejan, A., Fragale, A., Bradley, T. E., Anselmi, M., Ciolfi, A., Cecchetti, S., Muto, V., Bernardini, L., Azage, M., Carvalho, D. R., Espay, A. J., Male, A., Molin, A. M., Posmyk, R., Battisti, C., Casertano, A., Melis, D., … Hennekam, R. C. (2014). Mutations in ZBTB20 cause Primrose syndrome. Nature genetics, 46(8), 815–817. https://doi.org/10.1038/ng.3035

Daetwyler, H. D. (Creator), Capitan, A. (Creator), Pausch, H. (Creator), Stothard, P. (Creator), van Binsbergen, R. (Creator), Brondum, R. F. (Creator), Liao, X. (Creator), Djari, A. (Creator), Rodriguez, S. C. (Creator), Grohs, C. (Creator), Esquerré, D. (Creator), Bouchez, O. (Creator), Rossignol, M. N. (Creator), Klopp, C. (Creator), Rocha, D. (Creator), Fritz, S. (Creator), Eggen, A. (Creator), Bowman, P. J. (Creator), Coote, D. (Creator), Chamberlain, A. J. (Creator), Anderson, C. L. (Creator), Tassel, C. P. (Creator), Hulsegge, B. (Creator), Goddard, M. E. (Creator), Guldbrandsten, B. (Creator), Lund, M. S. (Creator), Veerkamp, R. F. (Creator), Boichard, D. A. (Creator), Fries, R. (Creator), Hayes, B. J. (Creator) (1 Feb 2021). Run8: The 1000 Bull Genomes Project. Wageningen University & Research.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Delaneau, O., Zagury, JF., Robinson, M.R. et al. Accurate, scalable and integrative haplotype estimation. Nat Commun 10, 5436 (2019). https://doi.org/10.1038/s41467-019-13225-y

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics, 43(5), 491-498.

Dyberg, C., Papachristou, P., Haug, B. H., Lagercrantz, H., Kogner, P., Ringstedt, T., ... & Johnsen, J. I. (2016). Planar cell polar-ity gene expression correlates with tumor cell viability and prognostic outcome in neuroblastoma. BMC cancer, 16, 1-14.

Dixon, M. J., Marazita, M. L., Beaty, T. H., & Murray, J. C. (2011). Cleft lip and palate: understanding genetic and environ-mental influences. Nature Reviews Genetics, 12(3), 167-178.

Fisher R.A.The correlation between relatives under the supposition of Mendelian inheritance Trans. Roy. Soc. Edinburgh, 52 (1918), pp. 399-433

Funato, N., Chapman, S. L., McKee, M. D., Funato, H., Morris, J. A., Shelton, J. M., ... & Yanagisawa, H. (2009). Hand2 controls osteoblast differentiation in the branchial arch by inhibiting DNA binding of Runx2.

Garrick, D.J. The nature, scope and impact of genomic prediction in beef cattle in the United States. Genet Sel Evol 43, 17 (2011). https://doi.org/10.1186/1297-9686-43-17

Georges M. Mapping, fine mapping, and molecular dissection of quantitative trait Loci in domestic animals. Annu Rev Genomics Hum Genet. 2007;8:131-62. doi: 10.1146/annurev.genom.8.080706.092408. PMID: 17477823.

Gianola, D., de Los Campos, G., Hill, W. G., Manfredi, E., & Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. Genetics, 183(1), 347-363.

Giuffra, E., Tuggle, C. K., & Faang Consortium. (2019). Functional annotation of animal genomes (FAANG): current achievements and roadmap. Annual review of animal biosciences, 7, 65-88.

Guo, T., Han, X., He, J., Feng, J., Jing, J., Janečková, E., ... & Chai, Y. (2022). KDM6B interacts with TFDP1 to activate P53 sig-naling in regulating mouse palatogenesis. Elife, 11, e74595.

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics, 12, 1-12.

Hancock, D. B., Levy, J. L., Gaddis, N. C., Bierut, L. J., Saccone, N. L., Page, G. P., & Johnson, E. O. (2012). Assessment of genotype imputation performance using 1000 Genomes in African American studies. *PloS one*, *7*(11), e50610. https://doi.org/10.1371/journal.pone.0050610

Han, X., Feng, J., Guo, T., Loh, Y. H. E., Yuan, Y., Ho, T. V., ... & Chai, Y. (2021). Runx2-Twist1 interaction coordinates cranial neural crest guidance of soft palate myogenesis. Elife, 10, e62387.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics, 31(2), 423–447. https://doi.org/10.2307/2529430

Hirsch, N., Dahan, I., D'haene, E., Avni, M., Vergult, S., Vidal-García, M., ... & Birnbaum, R. Y. (2022). HDAC9 structural variants disrupting TWIST1 transcriptional regulation lead to craniofacial and limb malformations. Genome Research, 32(7), 1242-1253.

Hocquette J-F, Lehnert S, Barendse W, Cassar-Malek I, Picard B. Recent advances in cattle functional genomics and their application to beef quality. animal. 2007;1(1):159-173. doi:10.1017/S1751731107658042

Hoff, J. L., Decker, J. E., Schnabel, R. D., & Taylor, J. F. (2017). Candidate lethal haplotypes and causal mutations in Angus cattle. BMC genomics, 18, 1-11.

Houkes, R., Smit, J., Mossey, P., Don Griot, P., Persson, M., Neville, A., ... & Breugem, C. (2023). Classification systems of cleft lip, alveolus and palate: results of an international survey. The Cleft Palate Craniofacial Journal, 60(2), 189-196.

Huckert, M., Stoetzel, C., Morkmued, S., Laugel-Haushalter, V., Geoffroy, V., Muller, J., ... & Bloch-Zupan, A. (2015). Mutations in the latent TGF-beta binding protein 3 (LTBP3) gene cause brachyolmia with amelogenesis imperfecta. Human molecular genetics, 24(11), 3038-3049.

Illumina DNA Prep Reference Guide (2021). "Illumina DNA Prep Reference Guide." *Illumina, Inc.*

Jin, J. Z., & Ding, J. (2006). Analysis of Meox-2 mutant mice reveals a novel postfusion-based cleft palate. Developmental dynamics: an official publication of the American Association of Anatomists, 235(2), 539-546.

Juven, A., Nambot, S., Piton, A., Jean-Marçais, N., Masurel, A., Callier, P., ... & Faivre, L. (2020). Primrose syndrome: a phenotypic comparison of patients with a ZBTB20 missense variant versus a 3q13. 31 microdeletion including ZBTB20. European Journal of Human Genetics, 28(8), 1044-1055.

Kreiner-Møller, E., Medina-Gomez, C., Uitterlinden, A. G., Rivadeneira, F., & Estrada, K. (2015). Improving accuracy of rare variant imputation with a two-step imputation approach. European Journal of Human Genetics, 23(3), 395-400.

Lan, Y., Xu, J., & Jiang, R. (2015). Cellular and molecular mechanisms of palatogenesis. Current topics in developmental biol-ogy, 115, 59-84.

Levi, B., Brugman, S., Wong, V. W., Grova, M., Longaker, M. T., & Wan, D. C. (2011). Palatogenesis: engineering, pathways and pathologies. Organogenesis, 7(4), 242-254.

Lin, P., Hartz, S. M., Zhang, Z., Saccone, S. F., Wang, J., Tischfield, J. A., ... & COGA Collaborators COGEND Collaborators, GENEVA. (2010). A new statistic to evaluate imputation reliability. *PloS one*, *5*(3), e9697.

Liu, S., Gao, Y., Canela-Xandri, O. et al. A multi-tissue atlas of regulatory variants in cattle. Nat Genet 54, 1438–1447 (2022). https://doi.org/10.1038/s41588-022-01153-5

MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., ... & Goddard, M. E. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC genomics, 17, 1-21.

Mangold, E., Ludwig, K. U., & Nöthen, M. M. (2011). Breakthroughs in the genetics of orofacial clefting. Trends in molecular medicine, 17(12), 725-733.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, *39*(7), 906-913.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. Nature Reviews Genetics, 11(7), 499-511.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., ... & Van Tassell, C. P. (2009). Development and characterization of a high density SNP genotyping assay for cattle. PloS one, 4(4), e5350.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 20(9), 1297-1303.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., ... & Cunningham, F. (2016). The ensembl variant effect predictor. Genome biology, 17, 1-14.

Meuwissen, T. H. E. & Goddard, M. E. (1996). The use of marker haplotypes in animal breeding schemes. Genetics, Selection, Evolution 28, 161–176.

Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001 Apr;157(4):1819-29. doi: 10.1093/genetics/157.4.1819. PMID: 11290733; PMCID: PMC1461589.

Middleton, B. K., & Gibb, J. B. (1991). An overview of beef cattle improvement programs in the United States. Journal of animal science, 69(9), 3861-3871.

Morris, C. A., Baker, R. L., Cullen, N. G., Hickey, S. M., & Wilson, J. A. (1993). Genetic analyses of cow lifetime production up to 12 mating years in crossbred beef cattle. Animal Science, 57(1), 29-36.

Murdoch, J. N., Damrau, C., Paudyal, A., Bogani, D., Wells, S., Greene, N. D., ... & Copp, A. J. (2014). Genetic interactions between planar cell polarity genes cause diverse neural tube defects in mice. Disease models & mechanisms, 7(10), 1153-1163.

Pausch, H., MacLeod, I. M., Fries, R., Emmerling, R., Bowman, P. J., Daetwyler, H. D., & Goddard, M. E. (2017). Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genetics Selection Evolution, 49, 1-14.

P. Di Tommaso, et al. Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316–319 (2017) doi:10.1038/nbt.3820

Platt A, Vilhjálmsson BJ, Nordborg M: Conditions under which genome-wide association studies will be positively misleading. Genetics. 2010, 186: 1045-1052.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics, 81(3), 559-575.

Raychaudhuri, S. (2011). Mapping rare and common causal alleles for complex human diseases. *Cell*, *147*(1), 57-69.

Rowan, T. N., Hoff, J. L., Crum, T. E., Taylor, J. F., Schnabel, R. D., & Decker, J. E. (2019). A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. *Genetics Selection Evolution*, *51*, 1-16.

Rowan, T. N., Durbin, H. J., Seabury, C. M., Schnabel, R. D., & Decker, J. E. (2021). Powerful detection of polygenic selection and evidence of environmental adaptation in US beef cattle. PLoS genetics, 17(7), e1009652.

Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the Positional Burrows Wheeler Transform. PLoS Genet. 2020 Nov 16;16(11):e1009049. doi: 10.1371/journal.pgen.1009049. PMID: 33196638; PMCID: PMC7704051.

Rust, T., & Groeneveld, E. (2001). Variance component estimation on female fertility traits in beef cattle. South African Journal of Animal Science, 31(3), 131-141.

Saatchi, M., Schnabel, R. D., Taylor, J. F., & Garrick, D. J. (2014). Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC genomics, 15(1), 1-17.

Satokata, I., & Maas, R. (1994). Msx1 deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth devel-opment. Nature genetics, 6(4), 348-356.

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491-504.

Schutte, B. C., & Murray, J. C. (1999). The many faces and factors of orofacial clefts. Human molecular genetics, 8(10), 1853-1859.

Seabury, C. M., Oldeschulte, D. L., Saatchi, M., Beever, J. E., Decker, J. E., Halley, Y. A., ... & Taylor, J. F. (2017). Genome-wide association study for feed efficiency and growth traits in US beef cattle. BMC genomics, 18, 1-25.

Smeriglio, P., & Zalc, A. (2023). Cranial neural crest cells contribution to craniofacial bone development and regeneration. Current Osteoporosis Reports, 21(5), 624-631.

Smith, T. M., Lozanoff, S., Iyyanar, P. P., & Nazarali, A. J. (2013). Molecular signaling along the anterior–posterior axis of early palate development. Frontiers in physiology, 3, 488.

Som, P. M., Streit, A., & Naidich, T. P. (2014). Illustrated review of the embryology and development of the facial region, part 3: an overview of the molecular interactions responsible for facial development. American Journal of Neuroradiology, 35(2), 223-229.

Spencer CC, Su Z, Donnelly P, Marchini J: Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 2009, 5(5):e1000477.

Thornton, P. K. (2010). Livestock production: recent trends, future prospects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1554), 2853-2867.

Thompson, S. 2008. No Rocking Chair for Willham. Iowa State University Stories.chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://publications.iowa.gov/18407/1/Stories_2008Fall.pdf

Taylor, J. F., Taylor, K. H., & Decker, J. E. (2016). Holsteins are the genomic selection poster cows. Proceedings of the National Academy of Sciences, 113(28), 7690-7692.

Uffelmann, E., Huang, Q.Q., Munung, N.S. et al. Genome-wide association studies. Nat Rev Methods Primers 1, 59 (2021). https://doi.org/10.1038/s43586-021-00056-9

van Binsbergen, R., Bink, M. C., Calus, M. P., van Eeuwijk, F. A., Hayes, B. J., Hulsegge, I., & Veerkamp, R. F. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution, 46, 1-13.

Vanderas, A. P. (1987). Prevalence of craniomandibular dysfunction and adolescents: a review. Pediatric dentistry, 9(4), 313-319.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. Journal of dairy science, 91(11), 4414-4423.

VanRaden, P. M., Tooker, M. E., O'connell, J. R., Cole, J. B., & Bickhart, D. M. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. Genetics Selection Evolution, 49, 1-12.

van den Boogaard, M. J. H., Dorland, M., Beemer, F. A., & van Amstel, H. K. P. (2000). MSX1 mutation is associated with orofacial clefting and tooth agenesis in humans. Nature genetics, 24(4), 342-343.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 43(1), 11-10.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019, May). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In 2019 IEEE international parallel and distributed processing symposium (IPDPS) (pp. 314-324). IEEE.

Vercoe J.E. and Frisch, J.E., 1992. Genotype (breed) and environment interaction with particular reference to cattle in the tropics-review. Asian-Austral. J. of Anim. Sci., 5:401-409.

Walcsh, B., & Lynch, M. (2018). Evolution and selection of quantitative traits. Oxford University Press.

Wang C, Chen A, Ruan B, et al. PCDH7 inhibits the formation of homotypic cell-in-cell structure. Front Cell Dev Biol. 2020;8:329. doi: 10.3389/fcell.2020.00329

Weinzweig's, Jeffrey, and J. Weinzweig. 2017. "Jeffrey Weinzweig's Experiments on in Utero Cleft Palate Repair in Goats (1999-2002)." https://keep.lib.asu.edu/items/173354.

Wiggans, G. R., Cooper, T. A., VanRaden, P. M., Van Tassell, C. P., Bickhart, D. M., & Sonstegard, T. S. (2016). Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. Journal of Dairy Science, 99(6), 4504-4511.

Wray, N. R., Kemper, K. E., Hayes, B. J., Goddard, M. E., & Visscher, P. M. (2019). Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction. Genetics, 211(4), 1131-1141.

Xiang, R., Berg, I. V. D., MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., ... & Goddard, M. E. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proceedings of the National Academy of Sciences, 116(39), 19398-19408.

Xiang, R., MacLeod, I.M., Daetwyler, H.D. et al. Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. Nat Commun 12, 860 (2021). https://doi.org/10.1038/s41467-021-21001-0.

Xiao H, Sun Z, Wan J, Hou S, Xiong Y. Overexpression of protocadherin 7 inhibits neuronal survival by downregulating BIRC5 in vitro. Exp Cell Res. 2018;366(1):71–80. doi: 10.1016/j.yexcr.2018.03.016

Xu, J., Liu, H., Lan, Y., Aronow, B. J., Kalinichenko, V. V., & Jiang, R. (2016). A Shh-Foxf-Fgf18-Shh molecular circuit regulat-ing palate development. PLoS genetics, 12(1), e1005769.

Zhang, Z., Song, Y., Zhao, X., Zhang, X., Fermin, C., & Chen, Y. (2002). Rescue of cleft palate in Msx1-deficient mice by trans-genic Bmp4 reveals a network of BMP and Shh signaling in the regulation of mammalian palatogenesis.

Zhang, J. L., Huang, Y., Qiu, L. Y., Nickel, J., & Sebald, W. (2007). von Willebrand factor type C domain-containing proteins regulate bone morphogenetic protein signaling through different recognition mechanisms. Journal of Biological Chemistry, 282(27), 20002-20014.

Vita

Kenzy Hoffmann grew up in New Braunfels, TX on her family's farm where they raised Boer goats and Southdown sheep alongside a commercial Brangus cattle herd. After graduating from Canyon Lake High School in 2019, she attended Texas A&M University in College Station, TX. While there, she obtained her Bachelor of Science degree in Animal Science with a minor in Agricultural Leadership and Development while participating in a variety of organizations, internships, and undergraduate research. Kenzy began working on her Master of Science degree in Animal Science in 2023 at the University of Tennessee, Knoxville. She worked with Dr. Troy Rowan and completed her Master of Science degree in Spring 2025.