# Investigation of Causal Mutations in Livestock and Companion Animals Using Whole-Genome Sequencing (WGS) and Bioinformatics

by

Nahid Bandi Kenari

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Animal Science

Department of Agricultural, Food and Nutritional Science
University of Alberta

# Abstract

In this thesis project, three studies were conducted, aimed at identifying the causal variants responsible for phenotypic/genotypic sex discordance in cattle, copper toxicosis in Dalmatian dogs, and low fertility/infertility in cattle. In all three studies, whole-genome sequencing (WGS) and bioinformatics were used to detect candidate causal mutations.

The first study involved samples flagged during a national genetic testing program in Ireland due to the inconsistencies in genotype-test assigned sex compared to farmer-reported phenotypic sex. In all, cows were initially classified as XY male genotypically. Re-testing confirmed this classification, leading to the hypothesis of Swyer syndrome in these cases. First described in humans, individuals with Swyer syndrome possess XY sex chromosomes but develop female reproductive organs. Study samples were characterized by WGS, and data was analyzed alongside fertile XX and XY control animals. Sex chromosome complement was established using WGS data and confirmed that the genotyped cows were XY. Short sequence read alignments with the reference genome were examined for known disorders of sex development (DSD) genes. In seven out of eight affected individuals, an absence of reads aligning to the *SRY* gene was observed, leading to a designation of *SRY*-negative and explaining the cause of the DSD. To identify candidate causal mutations for the one *SRY*-positive affected individual, two variant calling pipelines for SNPs/indels and structural variants (SVs) along with variant filtering and annotation analysis were used. While no candidate causal variants in known DSD genes were identified, a small number (125) of potential causative mutations elsewhere in the genome were found. Due to the lack of known roles established for the genes affected by these variants in sex determination and development, further refinement of this list is needed to pinpoint a true causative mutation, for

example through the characterization of additional samples identified through ongoing genetic testing.

The second study sought to find the cause of an apparent autosomal recessive condition occurring in Dalmatian dogs and other breeds, termed canine copper toxicosis (CT). Characterized by excessive copper in the liver, the disease ultimately causes liver damage and if untreated, potentially death. Dalmatian samples (n = 10) identified through pedigree analysis as presumptive cases, carriers, and controls were characterized using WGS. Three samples from Bedlington Terriers were included, two believed to have CT. SNPs, indels, and SVs were identified in all samples and filtered using various criteria, considering pedigree information, predicted functional impact, and overlap with known CT genes. This approach ruled out the involvement of the *ATP7B* or *COMMD1* genes, both associated with CT in humans and dogs. One of the affected Bedlington carriers was confirmed to be missing the *COMMD1* gene. Regarding the disease's cause in Dalmatians, no strong candidate mutations could be proposed from the identified variants. Nevertheless, after applying additional filtering steps to the genotype scenario output, we identified 7 SNPs and 1 indel with predicted functional impacts on genes, alongside 3 indels overlapping with candidate genes. However, due to the limited knowledge of the gene's role in copper metabolism, refining this list to pinpoint a specific causal mutation remains challenging.

In the third study, we investigated 20 animals with suboptimal fertility using WGS to identify recessive lethal mutations potentially causing embryo loss. We identified SNPs and indels and applied various strategies to isolate sites with heterozygous genotypes in the sequenced animals, never observed in homozygous form in living cattle. Our approach leveraged WGS results from over 2,703 additional cattle collected through the 1000 Bull Genomes Project. Additional filtering based on predicted functional impact led to the identification of 391 candidate causal

SNPs and 3,527 candidate causal indels potentially linked to subfertility/infertility in beef cattle. The large number of variants reflects the challenge of distinguishing rare or false-positive variants from true recessive lethals. Further experiments are required to substantiate the true recessive lethality of any identified candidate mutations.

As a result of this work, updates have been made to the genotyping chip used in genetic evaluations in Ireland to improve sex assignment. Additionally, a follow-up long-read study has begun, using additional Dalmatian samples to enhance the characterization of genomes in affected, carrier, and control animals.

**Keywords:** Swyer syndrome, copper toxicosis, genetic variations, structural variants, subfertility/infertility, recessive lethal

# Preface

This thesis is an original work by Nahid Bandi Kenari. Chapter 2 has been published as Berry, D. P., Herman, E. K., Carthy, T. R., Jennings, R., Bandi-Kenari, N., O'Connor, R. E., Mee, J.F., O'Donovan, J., Matthews, D., & Stothard, P. Characterization of eight cattle with Swyer syndrome by whole-genome sequencing. Animal Genetics, 2023 Apr;54(2):93-103. EKH executed both variant-calling pipelines developed for the fourth chapter using the dataset specific to this chapter. NBK ran and troubleshooted both workflows and analyzed the data. PS supervised data analysis and integration and edited the chapter.

The study in Chapter 3 was designed by PS and Georgina MacIntyre. EKH ran the two developed variant calling pipelines for SNPs/indels and SVs. NBK analyzed the data. PS and GM supervised data analysis and integration and edited the chapter. This study received ethical clearance from the University of Alberta Animal Care and Use Committee (AUP 2165) following Canadian Council on Animal Care (CCAC) guidelines. In the US, it was approved by the Colorado State University Institutional Animal Care and Use Committee (15-5632A and KP IACUC Initial 4138), adhering to Federal Regulations and Guidelines.

The study in Chapter 4 was designed by PS, Changxi Li, and John Basarab. NBK executed and troubleshooted both variant calling workflows developed by EKH, conducting data analysis. PS supervised data analysis and integration and edited the chapter. The study utilized samples collected from the beef cattle herd at the Lacombe Research and Development Centre (LRDC) under the RFI Methane project, supervised by PI JB. The collection followed the University of Alberta Animal Care and Use Protocol (AUP00001140), approved by the LRDC Animal Care Committee. Additionally, the study adhered to Canadian Council on Animal Care (CCAC)

guidelines for the care and use of farm animals in research, education, and evaluation (CCAC, 2009).

# Dedications

*For my dear family, especially my beloved sister Nasrin and my wonderful husband Aaron who has always been there for me through thick and thin.*

*For the courageous Iranian women and men, who are risking their lives to stand up for their freedom and against oppression and injustice.*

*For "**Woman, Life, Freedom**"*

# Acknowledgments

Words cannot express my gratitude to my supervisor, Dr. Paul Stothard for his full support, patience, mentorship, and advice throughout my master's project, even during the challenging time of the pandemic.

I would like to thank my committee chair, Dr. Changxi Li, and Dr. John Basarab as well as my arms-length examiner, Dr. Anne Laarman, and committee chair, Dr. Cameron Carlyle, for their invaluable insights and feedback. I also could not have undertaken this journey without my defense committee, my arms-length examiner, Dr. Anne Laarman, and my committee chair, Dr. Cameron Carlyle, for their invaluable insights. Special appreciation to Georgina MacIntyre for her oversight and assistance in the Dalmatian study.

Additionally, this endeavor would not have been possible without the generous support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the American Kennel Club (AKC) which financed my research.

I am also grateful to my classmates and office mates, including Dr. Emily Herman and Dr. Jason Grant for their guidance, helpful comments, and feedback. Thanks, should also go to the librarians, and research assistants, who impacted and inspired me.

Lastly, I would be remiss in not mentioning my family, especially my sister Nasrin Bandi, my husband Aaron James McCaw, and my in-laws. Their belief in me has kept my spirits and motivation high during this process.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

BAM: Binary Alignment Map

CDS: Coding DNA sequence

CNVs: Copy number variations

DELs: Deletions

DSDs: Disorders or differences of sex development

DUPs: Duplications

FXR: Farnesoid X nuclear receptor

GATK: The Genome Analysis Toolkit

GWAS: Genome-wide association studies

GUI: Graphical user interface

HMG: High mobility group

INDELs: insertions/deletions

IGV: Integrative Genomics Viewer

INSs: Insertions

INVs: Inversions

MD: Menkes disease

MQ: Mapping Quality

NGS: Next-generation sequencing

QD: Quality by depth

SAM: Sequence Alignment Map

SNP: Single nucleotide polymorphism

SNV: Single nucleotide variant

*SRY*: Sex-determining region Y

SV: Structural variant

TDF: Testis-determining factor

TRAs: Translocations

VCF: Variant call format

VEP: Variant Effect Predictor

WD: Wilson disease

WGS: Whole-genome sequencing

WHW: West Highland White

Yp: Y chromosome short arm

# CHAPTER 1. General Introduction

## 1.1. The Importance of Studying Genes and Genetic Variants Impacting Traits

Genetic variants are changes within the DNA sequence that comprise a gene. Formerly referred to as gene mutations, these types of genetic alterations are now more accurately described as "gene variants", because they do not always result in disease. Variants can affect one or several nucleotides and can be categorized depending on their characteristics, such as SNPs, indels (insertion-deletion), and structural variants (SVs) (Lek et al., 2016). Genetic diversity can originate from variants and from the rearrangement of genetic material that occurs during reproduction (recombination[1] and independent assortment[2]). Gene variants that affect how genes operate, or protein's function may lead to new characteristics in an organism. However, many other variants will not affect gene or protein function and are called "neutral variants." A significant focus of research efforts is distinguishing these neutral variants from those that influence gene function and phenotypes (Nachman & Crowell, 2000).

Gene variants can be categorized into two types: inherited and de novo. Inherited (or hereditary) variants are passed down from parent to offspring and can be found in all body cells. Such variants are inherited by offspring when the egg and sperm cells unite during fertilization. Consequently, in the case of inherited variants, the offspring's cells will carry alleles observed in the parent's DNA (Choi et al., 2009). On the other hand, de novo variants appear for the first time in the progeny and are not observed in the parent's genotype (although the mutation can occur in the parent). Such variants can arise through mutation in a germ cell or the fertilized egg during

[1]. It is also known as crossing over, and it is the process in which DNA sequences are rearranged through a combination of chromosome or chromosome segment breakage, rejoining, and copying (Carroll, 2001).
[2]. The law of independent assortment, proposed by Gregor Mendel, explains that during meiosis, the separation of gene variants for different traits is independent of each other. This leads to various combinations of these gene variants in gametes, contributing to genetic diversity (Gottlieb et al., 2023).

early embryogenesis. De novo variants can also occur at a later stage, such that they affect a small proportion of body cells in the individual and are not passed on to progeny. Such variants are called somatic variants. De novo mutations may arise due to environmental factors or errors in DNA replication during cell division (García-Nieto et al., 2019; Mohiuddin et al., 2022; Veltman & Brunner, 2012). When seeking to understand the genetic basis of a phenotype, particularly a newly observed or rarely observed phenotype, it can be essential to consider both inherited and de novo variants. For example, in livestock, de novo mutations have been identified as a cause of several severe dominant conditions such as glass-eyed albino (GEA), neurocristopathy (NC), osteogenesis imperfecta type 2 (OI), and bulldog calf syndrome (BD1, BD2, and BD3) (Bourneuf et al., 2017).

In some instances, genetic alterations that occur during growth give rise to a phenomenon known as mosaicism. Mosaicism refers to the situation where specific cells in the body possess a slightly different genetic makeup than others. The genetic modifications responsible for mosaicism occur later in life, from embryonic stages through adulthood, rather than in the parent's gametes or the zygote. During cell division, the cells originating from a cell that contains a genetic modification will also possess that variation. In contrast, the cells that come from a cell without the alteration will not have the variant. When some of the body's cells (somatic cells) contain a genetic variation while others do not, it is called "somatic mosaicism." The impact of somatic mosaicism on an individual's health can vary depending on the specific genetic variant and the percentage of cells affected by the variant. When some male or female gametes contain a genetic variation while others do not, this is known as "germline mosaicism." In cases of germline mosaicism, a parent who does not show symptoms of a genetic condition may still pass it on to their offspring, as the genetic variant is present in some of the parent's egg or sperm cells. Germline

mosaicism can explain the puzzling situation where a parent produces multiple offspring suffering from a dominant disorder, yet the parent does not exhibit the phenotype (Mohiuddin et al., 2022).

Most genetic variations do not significantly impact an individual's health, and those that do are relatively uncommon. However, some variations are prevalent enough in the population to be considered common genetic variations. They can account for differences in physical traits such as coat color, horned or polled, and ear shapes in livestock and coat color and pattern, tail length, and body size in dogs. While many of these common variations do not pose any harm to an individual's health, some may increase the likelihood of developing certain health conditions or disorders.

## 1.2. Approaches for Identifying Causative Variants

### 1.2.1. Genetic linkage mapping

Genetic linkage refers to the tendency of certain genes or genetic markers to be inherited together because they are located near each other on the same chromosome. Linkage mapping is a genetic method that uses the frequency of recombination between genetic markers to determine the relative location of genes on a chromosome (Liu, 2017). This method is an effective means of gene mapping since it identifies the chromosomal position of genes responsible for traits and aids in mapping quantitative trait loci (QTLs), regions on the genome linked with variations in a specific trait. This method is recognized as one of the earliest and most extensively employed approaches and has significantly contributed to identifying numerous genes that influence traits. This approach relies on the principle that DNA segments located in close proximity on a chromosome tend to be inherited together during the meiosis stage. This is due to their lower

likelihood of being separated into different chromatids during chromosome crossing-over. In other words, different genes/loci alleles located closer together on a chromosome have a lower probability of recombining and a higher chance of being inherited together (Fay, 2006). In linkage mapping, the presence of a trait is related to the presence of DNA markers, with the aim of identifying a DNA marker that is tightly associated with the unknown causative marker or the causative variant itself.

#### 1.2.1.1. Family-based linkage analysis

Linkage analysis looks for co-segregation of an allele and a phenotype through a pedigree and is often conducted within families and termed family-based linkage analysis (Cavalli-Sforza & Edwards, 1967; Lathrop et al., 1984). This approach applies data from families containing information on hereditary modes and the existence of the trait or disease in family members to detect areas of the genome potentially associated with the trait (Hill, 1974; Hill & Robertson, 1968). The application of linkage disequilibrium (LD) in family-based linkage analysis has been explored and debated in various influential studies (Kruglyak, 1999). Family-based linkage analysis is established by creating a population from a few recent ancestors, in which linkage disequilibrium (LD) is preserved over longer genetic distances within a family and through a few generations (Ben-Ari & Lavi, 2012).

In a research conducted by van de Sluis et al. (1999) using fluorescence in situ hybridization (FISH)[1], they localized the copper toxicosis-related locus in humans to the region of

[1]. It is a laboratory method employed for the precise identification and mapping of a particular DNA sequence situated on a chromosome. In this procedure, the complete set of chromosomes from an individual is attached to a glass slide and subsequently exposed to a "probe," which is a small section of purified DNA labeled with a fluorescent dye. This technique allows for the visualization and localization of the target DNA sequence through fluorescence microscopy (Huber et al., 2018).

chromosome CFA10q26 on the canine chromosome (van de Sluis et al., 1999), which corresponds to the area on human chromosome HSA2p13-21. This localization enabled the exclusion of any candidate gene known to be the cause of the disease (Dagenais et al., 1999; Nanji & Cox, 1999; van de Sluis et al., 1999). Afterward, in further research by van de Sluis et al. in 2002, by narrowing down the positions of the copper toxicosis gene to a section of less than 500 kb on the chromosome using DNA samples from privately owned dogs and linkage disequilibrium mapping, it was discovered that the deletion of exon 2 of the copper metabolism domain containing 1 (*COMMD1*) gene (previously known as *MURR1*) is the primary cause of copper toxicosis in Bedlington Terriers (BT) (Forman et al., 2005; van De Sluis et al., 2002). In another genetic linkage study by Yuzbasiyan-Gurkan et al. (1997), it was found that there was a linkage between copper toxicosis disorder and a microsatellite marker (C04107) in Bedlington Terriers (Yuzbasiyan-Gurkan et al., 1997).

### 1.2.2. Association studies

Genetic association studies involve analyzing the correlations/associations between genetic variants/markers and a particular trait or disease by comparing the genotypes of affected individuals to those of unaffected individuals without the need for a known family history (Lewis & Knight, 2012; Litonjua & Celedón, 2006). Such studies rely on linkage disequilibrium (LD) across families, which refers to the association or connection between genetic variations where the alleles of neighboring genetic markers located on the same chromosome tend to occur together more frequently within a specific population than if they were not linked (Holloway & Prescott, 2017). LD across families or historical LD is crucial in association studies as it allows the identification of genetic markers that effectively represent or "tag" the true underlying causal

variants (Joiret et al., 2019). This relationship enables researchers to leverage the information from these linked markers to gain insights into the associations between specific genetic variants and traits or diseases of interest. Examining genetic variants in tight LD saves time, money, and computational resources as it reduces the number of variants that need to be investigated, streamlining the process efficiently (Motsinger et al., 2007).

In their most basic form, association studies investigate whether a specific genetic marker is found more frequently in individuals with a particular trait or disease than those without (Litonjua & Celedón, 2006). Genetic association studies aim to identify candidate genes or regions of the genome that may be related to a disease by examining the correlation between the disease state and gene variation (Lewis & Knight, 2012). Association studies can be carried out with linkage studies or on their own, either de novo with candidate genes or genome-wide association studies (GWAS) (Mason et al., 2015). A greater frequency of a single-nucleotide polymorphism (SNP) allele or genotype in a group of individuals with a disease can be taken as particular evidence that the variant being studied increases the chance of developing that disease (Lewis & Knight, 2012). In association studies, SNPs are the most commonly analyzed markers; however, markers such as microsatellites, insertions and deletions, VNTR[1]s, and CNV[2]s are also used (Lewis & Knight, 2012). Association studies are essential for detecting genes that increase susceptibility to complex disorders. These disorders are considered "complex" because a combination of multiple genes and environmental factors influences their susceptibility rate. A considerable amount of research on genetic studies for numerous complex disorders verifies that

[1] . Variable number of tandem repeats (VNTR) refers to a location, within DNA where a short sequence of nucleotides is arranged as a repeat, in tandem. These repeats can vary in length among individuals (Guilherme et al., 2011).
[2]. CNV represents a category of genome structural changes. It involves a DNA segment whose copy number varies compared to a standard genome, with sizes ranging from 1 kilobase to multiple mega-bases (Stachowiak et al., 2016).

multiple genetic variations play a role in disease risk; however, each one has only a small impact (Lewis & Knight, 2012).

#### 1.2.2.1. Genome-wide association studies (GWAS)

Genome-wide association studies (GWAS) are large-scale association studies that involve genotyping hundreds of thousands of single nucleotide polymorphisms (SNPs) across many sample genomes to identify genetic variants that are associated with a particular trait or disorder (Uffelmann et al., 2021). For example, it has been established that copper toxicosis can be linked to a genetic mutation in the *COMMD1* gene (*COMMD1* $^{del/del}$); however, it has also been observed that Bedlington Terriers can develop copper toxicosis without this mutation (non-*COMMD1* $^{del/del}$) (Coronado et al., 2003; Haywood et al., 2001; Hyun et al., 2004; van De Sluis et al., 2002). Haywood et al. (2016) conducted a research project to discover any other genetic variations (besides *COMMD1* $^{del/del}$) that may be connected to copper toxicity in Bedlington terriers using a combination of genome-wide association studies followed by deep sequencing of the candidate area. Blood samples for DNA analysis and liver tissue for diagnosis were taken from 30 Bedlington terriers without the *COMMD1* $^{del/del}$ mutation, including the same numbers of dogs affected by copper toxicity and healthy control dogs. The DNA was first analyzed using GWAS and then re-sequenced more thoroughly to locate the potential abnormal gene (Haywood et al., 2016). They have discovered a strong association between a specific region on chromosome 37 and copper toxicosis in Bedlington Terriers that do not have the *COMMD1* $^{del/del}$ mutation. This region possesses multiple SNPs that are closely linked to the condition. Additionally, this chromosomal region harbors the *ABCA12* gene, which functions similarly to the *ATP-ase 7B* (*ATP7B*) gene associated with Wilson's disease in humans (Haywood et al., 2016).

#### 1.2.2.2. Candidate gene studies

Candidate gene association studies focus on analyzing the genetic variations linked to a particular disease or trait within a limited set of predetermined genes (Modena et al., 2019, p. 18). In other words, this approach involves studying specific genes hypothesized to be connected to a particular trait based on previous research. These studies usually adopt a case-control design, where individuals with the disease (cases) and those without it (controls) are identified, and their genetic differences are assessed, such as identifying variants like SNPs, haplotypes[1], indels, and CNVs that are more frequent in one group compared to the other. The findings of these studies are typically presented as odds ratios (ORs)[2] or relative risks (RRs)[3], depending on whether the research is structured as a case-control or cross-sectional study, respectively (Modena et al., 2019, p. 18). For example, the findings of an investigation by Dirksen et al. (2017) comprise both a targeted candidate gene approach and an analysis of the transcriptome (Dirksen, Spee, et al., 2017). Their study's outcomes were crucial in elucidating various facets of copper's involvement as an initiating factor of hepatitis in Dalmatian, introducing *APP* as a potential candidate gene for copper-associated chronic hepatitis, and providing insights into the molecular basis of the disease in Dalmatian dogs (Dirksen, Spee, et al., 2017).

[1]. A haplotype is a collection of genes that are in close proximity to each other on a chromosome and are inherited together as a single biological unit (Fuchs, 2018).

[2]. An odds ratio (OR) is a statistical measure used to assess the association between an exposure and an outcome in research or epidemiology. It quantifies the likelihood of an outcome happening when a specific exposure is present, relative to the likelihood of the outcome occurring when that exposure is absent. In other words, it helps us understand how the odds of an event or outcome change based on whether a particular exposure is present or not (Persoskie & Ferrer, 2017).

[3]. Relative risk (RR) is the proportion of the chances of an event occurring in the exposed group compared to those in the non-exposed group. RR is computed by dividing the risk of death or disease in a particular population subset (Group A) by the risk in all other groups combined (Persoskie & Ferrer, 2017).

### 1.2.3. Whole-genome sequencing (WGS) of a small number of cases and controls

Whole-genome sequencing of a small number of cases and controls can provide valuable insights into the genetic basis of a disease or condition. By comparing the genomes of individuals with the disease (cases) to those without the disease (controls), researchers can identify genetic variations that are more common in one group compared to the other. These variations may then be further studied to determine their potential role in causing or predisposing to the disease. WGS allows researchers to analyze an individual's entire genome, including all their genes and non-coding regions. This provides a more comprehensive view of genetic variations than other genetic analysis methods, such as genotyping or exome sequencing, which only analyze specific genome regions (Q. Wang et al., 2015). In a study, Holopainen et al. (2017) combined SNP-based homozygosity mapping of two Dalmatian dogs affected with acute respiratory distress syndrome (ARDS) with whole-genome sequencing (WGS) of an additional affected dog to uncover the underlying genetic mechanism responsible for ARDS in Dalmatians. They identified a case-specific homozygous nonsense[1] variant, c.31C>T; p.R11*, in the *ANLN* gene that was unique to the affected individuals. Their research sample consisted of eleven Dalmatians, including two affected siblings, one healthy carrier, one healthy sibling, one healthy grandparent, and six other healthy dogs. All eleven dogs were subjected to genotyping using Illumina's CanineHD SNP array[2] (Holopainen et al., 2017).

[1]. A nonsense mutation takes place when a codon originally encoding an amino acid is altered to a stop codon (Clark & Pazdernik, 2013).

[2]. The Illumina CanineHD SNP Array is a high-density genotyping array specifically designed for studying the genetic makeup of dogs (canines). It is a microarray-based technology that allows researchers to analyze a large number of genetic markers across the genome of a dog. The CanineHD SNP Array offered by Illumina is used for various applications in canine genetics, including identifying genetic variants associated with specific traits, diseases, and behaviors in dogs (*CanineHD Whole-Genome Genotyping BeadChip*, n.d.).

## 1.3. Next-Generation Sequencing (NGS)

"Next-generation sequencing" (NGS) is a transformative high-throughput DNA sequencing technology widely used in research and diagnostics across human health, animal science (Xing et al., 2016), and veterinary studies. It has revolutionized various areas of biotechnology (Jiang et al., 2019; Oetting, 2012) and emerged as a leading tool in genomics and bioinformatics (Dunislawska et al., 2017; Gupta & Gupta, 2020). NGS has also significantly advanced molecular biology techniques and animal genetics, providing numerous benefits, including a comprehensive understanding of genetic information (Dunislawska et al., 2017).

NGS technology's emergence is rapidly transforming how researchers conduct genetic studies and the routine work of their studies. Unlike traditional sequencing techniques, NGS encompasses a range of high-throughput, large-scale, and parallel DNA sequencing methods. The term "NGS" does not refer to a single approach but rather a group of technologies that have emerged several decades after the development of the Sanger sequencing technique in 1977. Sanger sequencing was the primary sequencing method for nearly 30 years before the advent of NGS (Kishikawa et al., 2019; Kulski, 2016; Mardis, 2008; Sanger et al., 1992).

NGS encompasses a variety of innovative sequencing technologies such as sequencing-by-synthesis (Mardis & Wilson, 2009), sequencing-by-ligation, ion semiconductor sequencing, and single-molecule sequencing. However, it is worth mentioning that Illumina's sequencing-by-synthesis approach is currently the predominant technology utilized in genetics and related research fields (Kishikawa et al., 2019). Although different NGS platforms employ diverse sequencing techniques, they all provide much more sequence information than was previously attainable. Furthermore, using these sequencing platforms' unique features, such as extraordinary

data processing ability, scalability, and fast data production, scientists and researchers can study biological systems to a degree of detail and precision that has never been imagined (Gupta & Gupta, 2020).

With the rapid development of NGS techniques and concurrent progress in bioinformatics software, programs, and methods, it is now possible for all research laboratories and groups with varying budget levels to generate remapped or even de novo assemblies of whole-genome sequences for any organism of interest (Gonzaga-Jauregui et al., 2012; Kulski, 2016; Lam et al., 2012; Pabinger et al., 2014). NGS technologies have been commercially available for nearly 15 years. During this period, there have been remarkable improvements in the speed, read length, and output size and significant decreases in the cost per base (Deng et al., 2016; Rantsiou et al., 2018; Van Dijk et al., 2014). Furthermore, following the first large-scale genetic variation research on the human genome, known as the "1000 Genomes Project" (1000 Genomes Project Consortium et al., 2010), other genomic studies on a larger scale have also been initiated. These include sequencing thousands of genomes, as seen in the Genome 10K Community of Scientists (2009) (Genome 10K Community of Scientists, 2009), and even millions of genomes, as in the case of the China National GeneBank Million Genomes Project (https://en.genomics.cn/).

Large-scale sequencing efforts have been conducted or are underway in livestock and dogs, including the 1000 Bull Genomes Project and the Canine Genome Sequencing Project. The 1000 Bull Genomes Project comprises complete genetic sequences obtained from 2,703 individuals, encompassing a substantial portion of global cattle diversity. The project has identified an extensive genetic diversity, including 84 million single-nucleotide polymorphisms (SNPs) and 2.5 million small insertion deletions within this dataset. This represents a remarkably high level of genetic variation captured by the project (Hayes & Daetwyler, 2019). The Dog10K project, also

known as the "10,000 Dog Genomes Project," is an initiative aimed at sequencing and analyzing the genomes of 10,000 canines from various breeds. The project aims to establish a comprehensive catalogue of genetic variation in dogs, studying both purebred and mixed-breed individuals. By examining the genetic diversity within the dog population, researchers hope to gain insights into the genetic basis of different traits, diseases, and behaviours in dogs. The Dog10K project can help further our understanding of canine evolution, domestication, and genetic factors underlying various canine traits and health conditions (Ostrander et al., 2019). These large-scale genomic research projects have significantly expanded our understanding of the relationships between genomic variations and traits of interest (Kilpinen & Barrett, 2013; Van Dijk et al., 2014).

Generally, after sequencing, bioinformatics tools are used to characterize an individual's DNA sequence reads by mapping them to a reference genome (Behjati & Tarpey, 2013). With these deep sequencing technologies, each of the billions of DNA bases in a genome is sequenced several times, providing high read depth and precise data that can reveal unexpected DNA variations (Behjati & Tarpey, 2013). The advantages of today's next-generation sequencing (NGS) technologies have transformed genomic studies and made whole-genome sequencing (WGS) a powerful means for this research area (Straiton et al., 2019). NGS has accelerated the discovery of genetic variants and the development of superior diagnostics, treatments, and breeds (Gupta & Gupta, 2020).

### 1.3.1. NGS-based whole-genome sequencing (WGS)

As a result of the advancements made in NGS technology, the feasibility and practicality of several previously complicated procedures have significantly increased. One of the most extensively used applications of NGS currently is whole-genome sequencing (WGS) (Mardis,

2008; Rantsiou et al., 2018; Ståhl & Lundeberg, 2012). WGS refers to the discovery of the arrangement of all nucleotides within an organism's entire DNA sequence in one procedure (Mardis, 2008; Rantsiou et al., 2018; Ståhl & Lundeberg, 2012).

An NGS-based assessment's utility arises from its ability to confidently find the differences between an individual's genome and a reference genome (Muzzey et al., 2015). Although this technique has been widely applied to human genome sequencing, the adaptable and scalable nature of NGS makes it equally advantageous for sequencing the genomes of other organisms, such as agriculturally important livestock, plants, or disease-causing microorganisms. The wide-ranging assessment of the entire genome provided by the WGS method is a crucial feature that makes it ideal for various detection purposes, such as identifying causative mutations of all types. Using the high-quality data generated by WGS can conceivably lead to the detection of all the genetic variations (P. Wu et al., 2019). The depth of sequencing is a crucial factor in the research design of WGS studies, as it determines the accuracy of variant calling and the cost of the study. Several studies have proposed optimal depths of sequencing that balance accuracy and cost (Ajay et al., 2011; Bentley et al., 2008; Fang et al., 2014; Jiang et al., 2019). According to Jiang et al. (2019) findings, a sequencing depth of 10X provides an optimal balance for practical purposes, ensuring sufficient coverage and reliable detection of variants. This depth allowed them to achieve over 99% genome coverage, ensuring accurate results and comprehensive variant discovery (Jiang et al., 2019). The recommended sequencing depths were determined by assessing the number of called variants and the level of agreement with reference datasets (Kishikawa et al., 2019).

## 1.4. DNA Variant Identification and Analysis

DNA variant identification refers to the process of detecting and analyzing genetic variations or mutations in an organism's genome. Identifying genetic variations is crucial to next-generation genome sequencing data analysis and whole-genome sequencing studies (Sathirapongsasuti et al., 2011). These DNA variations can result in trait differences, for example susceptibility to disease development. There are several steps involved in DNA variant identification and analysis. Some general steps include sequencing, read mapping, variant calling, genotype assignment, variant annotation, variant filtering, and validation. It is worth noting that the specific methods and tools used for DNA variant identification and analysis can vary depending on the research question, organism under study, available resources, and the latest advancements in genomics technologies. In other words, choosing the appropriate methodologies for genotype-calling, variant identification, and structural variation studies is directly linked to data management and research objectives (Sathirapongsasuti et al., 2011). Therefore, it is essential to design the study precisely since it fundamentally influences the analysis and experimental approaches (S. Y. Kim et al., 2010).

Bioinformatics is essential in managing large genomic data and is fundamental at every stage of the variant identification process (Muzzey et al., 2015). A key output of most variant detection and genotyping pipelines is a BAM file, which illustrates the relationships between the sequenced reads and a reference genome. BAM files can be interpreted and analyzed using automated approaches or examined manually to assess the genotype of a sample. For instance, a heterozygous SNP is identified when approximately half of the reads match the reference genome while the other half suggests the presence of a different (non-reference or alternate) allele. Using

a BAM visualization tool like IGV, SNPs can be detected as colored highlights in the read coverage section (see Figure 1.1) (Muzzey et al., 2015).

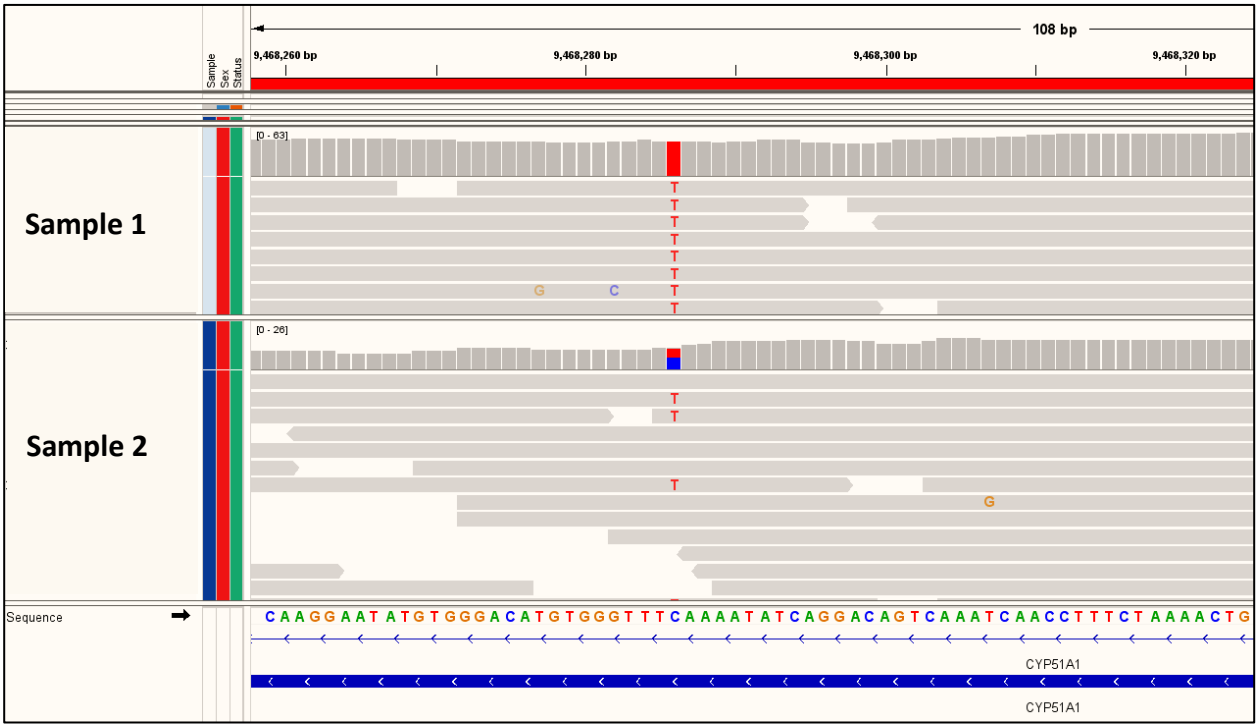


**Figure 1.1.** Investigation of a site of sequence variation in two samples through data visualization. Genomes from two samples aligned to a reference genomic region were visualized using Integrative Genomics Viewer (IGV) software (Thorvaldsdóttir et al., 2013). Differences between the aligned reads (gray horizontal bars) and the reference genome (sequence near the bottom of the figure) are shown using colored letters in the reads. Based on the visible sequence data plausible genotypes are T/T and T/C for the top and bottom samples, respectively.

### 1.4.1. Types of genetic variations

Genetic or DNA variants can be categorized in various ways based on their characteristics. Three commonly used categories of variants are single nucleotide polymorphisms (SNPs), insertion-deletion polymorphisms (indels), and structural variants (SVs) (Modena et al., 2019).

Single nucleotide polymorphisms (SNPs) are genetic variations characterized by substituting a single nucleotide base. In humans, these variations occur at an approximate frequency of one in every 1000 nucleotides. This results in about 3 million common SNPs being inherited across generations (*A global reference for human genetic variation*, 2015; Fareed & Afzal, 2013; Gabriel et al., 2002). In Bos taurus, SNPs occur at an approximate frequency of one SNP per every 700 base pairs (bp) (B. H. Consortium et al., 2009). Meanwhile, dogs have accumulated over two million SNPs due to significant changes in their genetic makeup during their evolutionary process. These genetic markers play a crucial role in understanding genetic diversity within and across different breeds (D'Agaro et al., 2021). The combination of an individual's SNPs or alleles is often referred to as their genotype, however it should be noted that the other types of variants also contribute to genotype.

Insertion–deletion mutations (indels) are genetic variations that can either insert and/or delete one or more nucleotide base pairs, and they are often referred to as variants involving less than 50 bp (1 to 49 bp) (Auton et al., 2015; Ebert et al., 2021). Although there are fewer indels than SNPs in the human genome, they can still cause significant variation in terms of base pair changes (Mullaney et al., 2010).

Structural variants (SVs) are genetic changes that involve alterations in the structure of DNA molecules, typically defined as exceeding 50 base pairs (≥50 bp) (Alkan et al., 2011; Ebert et al., 2021; Sedlazeck et al., n.d.; Sudmant et al., 2015) to approximately 1 kb or sometimes even larger (Freeman et al., 2006; *Overview of Structural Variation*, n.d.). SVs can be classified into several categories based on their characteristics, including deletions (DELs), insertions (INSs), duplications (DUPs), inversions (INVs), and translocations (TRAs) (Abyzov et al., 2015) (Figure 1.2). DUPs can be further classified as tandem or interspersed types based on the distance between

the duplicated copies (Alkan et al., 2011; Stankiewicz & Lupski, 2010). These mutations involve various combinations of gains, losses, or rearrangements of DNA sequences (Alkan et al., 2011; Sedlazeck et al., n.d.; Sudmant et al., 2015). INVs and TRAs are considered balanced forms of structural variants, which means they do not result in any net change in the content of the genome (i.e., the amount of genetic material). In contrast, unbalanced SVs (DELs, DUPs, and INSs) lead to either a net loss or gain of genetic material relative to other individuals. For this reason, these types of SVs are also known as copy number variations (CNVs) because their presence leads to differences in the abundance of DNA segments between individuals (Carvalho & Lupski, 2016).

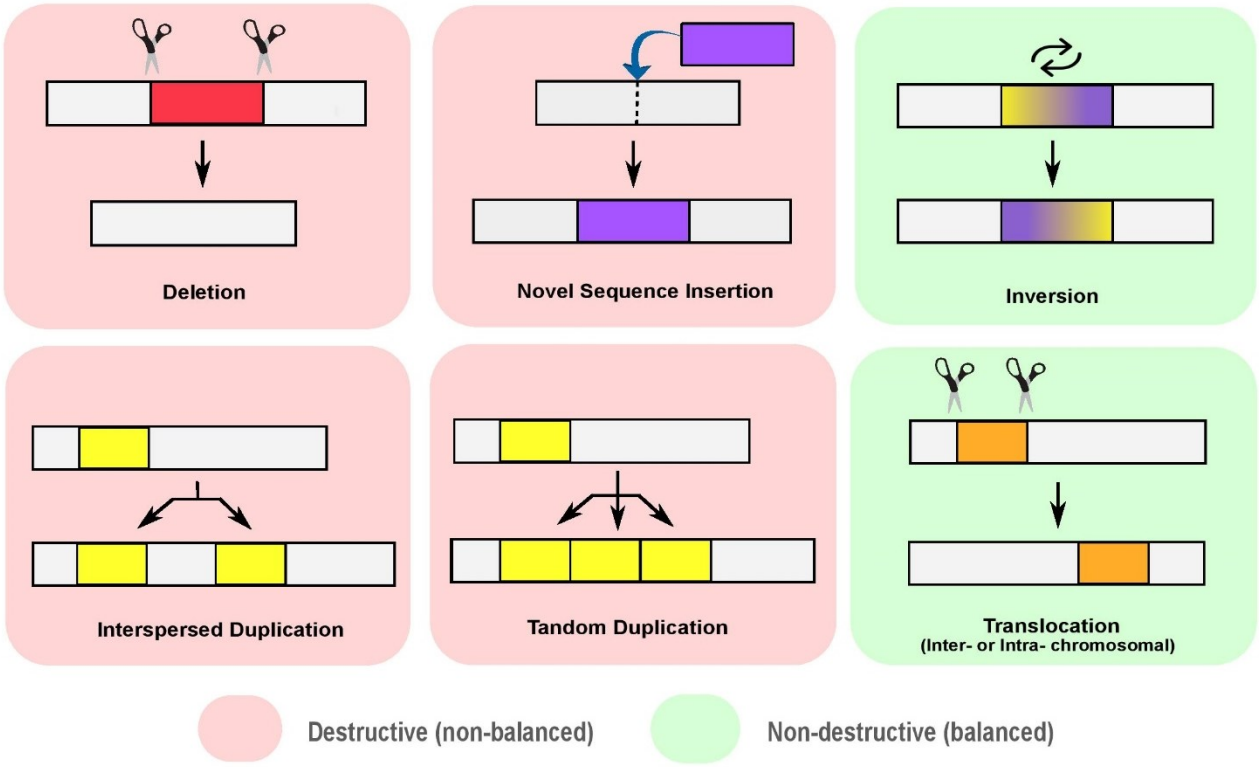


**Figure 1.2.** Different types of SVs. (Figure was adapted from Hall, 2021 (Hall, 2021) and was regenerated).

Studies have identified SVs as a potential cause of variation in various complex traits across different livestock breeds (L. Chen et al., 2021). For example, SVs have been associated with milk production and fertility in Nordic Red cattle (Kadri et al., 2014), nematode resistance in Angus cattle (Hou et al., 2012), coat texture or fur type in several dog breeds (Salmela et al., 2019), and abdominal fat content in chickens (Zhang et al., 2014). Furthermore, it has been demonstrated that SVs can negatively impact fertility in dairy cattle (Kadri et al., 2014). Also, multiple studies have highlighted the significant impact of SVs on functional changes within populations (Jeffares et al., 2017; Perry et al., 2007; Sudmant et al., 2015) and species (Kronenberg et al., 2018). Although SVs occur less frequently than SNPs, they can also have large impacts on cellular functions and gene expression (Chiang et al., 2017), which may affect traits of interest on a larger scale. SVs are commonly treated as distinct from small variants (such as SNPs and indels) as they require specific software tools for detection and are more challenging to detect reliably. As a result, there has been relatively less research on the importance of SVs compared to the vast number of studies on SNPs.

### 1.4.2. Bioinformatic considerations and challenges

Numerous analytical tools are available for scientists and researchers to conduct NGS analysis. However, integrating these techniques to obtain meaningful results can still be demanding and time-consuming, even for experienced researchers. Therefore, to facilitate this process, analytical workflows are practical alternatives that can analyze all stages, from raw data analysis to investigating a set of identified and annotated mutations. Regardless of whether an existing or custom workflow is used or if the analytical steps are performed separately, there are

several key steps in analyzing WGS data: variant calling, variant annotation, and variant filtering (Figure 1.3).

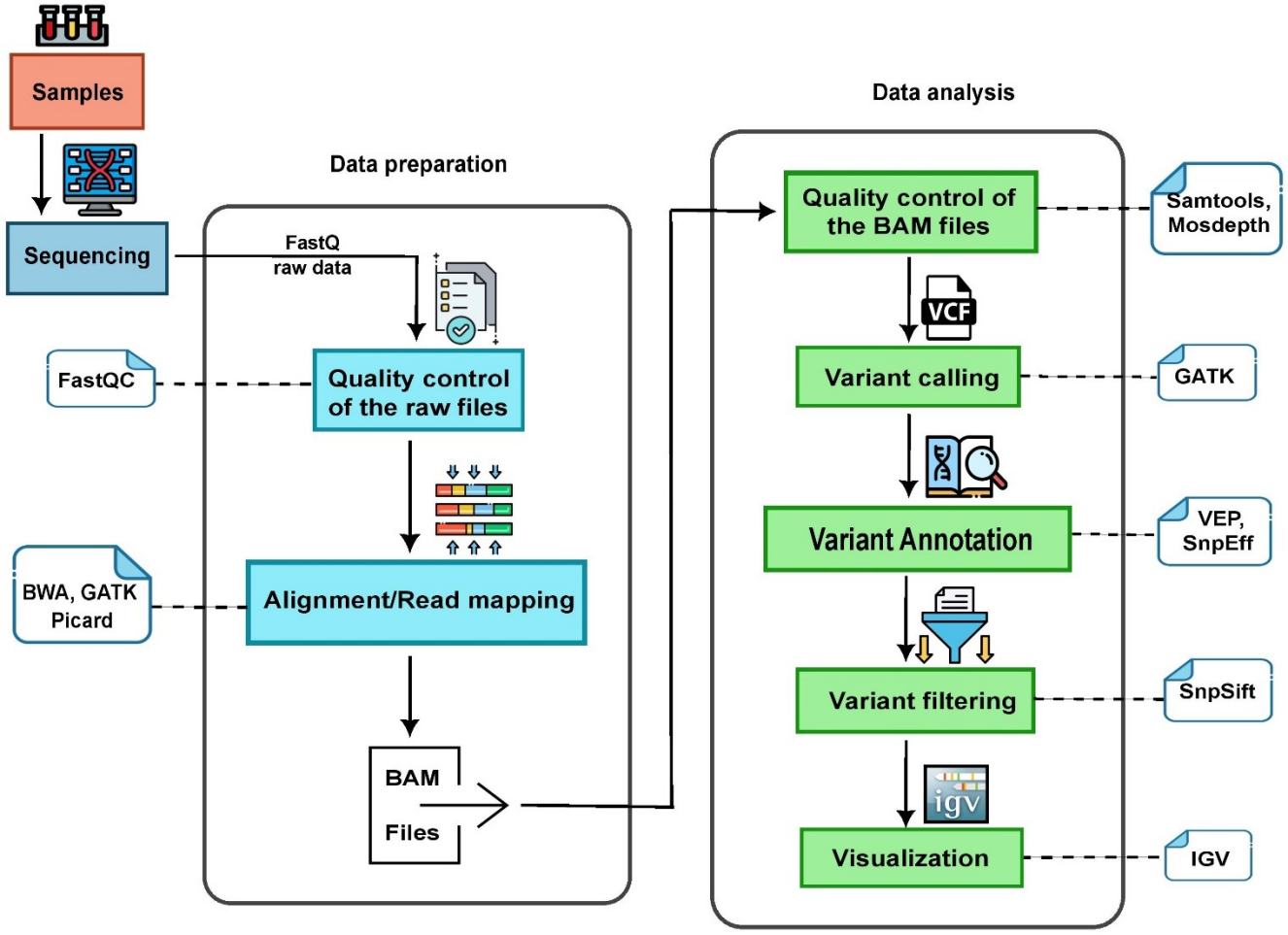


**Figure 1.3.** A diagram of the WGS analysis steps (Figure was adapted from Wolf et. al, (2015) (Wolf et al., 2015) and was regenerated).

### 1.4.2.1. What is variant calling?

The process of detecting genetic variants from raw sequenced data is known as "variant calling" (Muzzey et al., 2015). This process generally involves the following steps: 1) conducting whole-genome sequencing of the samples to produce FASTQ files, 2) aligning the read sequences

in the FASTQ files to a reference genome to generate alignment (BAM) files, and 3) identifying the positions where the aligned reads differ from the reference genome and storing those sites in a VCF (Variant Call Format) file (Muzzey et al., 2015).

Regardless of the type of sample and sequencing method used, the goal of variant calling is to accurately characterize the genotype. The extensive data produced by whole-genome sequencing has led to significant progress in bioinformatics techniques, aiming to accurately identify true variants while minimizing the occurrence of false positives and false negatives (Pabinger et al., 2014; Porterfield, 2015). Variant calling involves programmatically assessing the sequence reads aligned with the reference genome assembly to identify differences that suggest sequence variation. For short variants (SNPs and indels), mismatches between aligned reads, and the reference can reveal variants and be used to determine genotype, provided that read depth is sufficient. For SVs, various additional criteria are used, depending on the variant type. For example, unusually high or low read depth can signal the presence of CNVs. Other methods for identifying SVs include examining split reads (reads with portions aligned to distinct loci in the reference) and discordant read pairs (reads originating from the same physical fragment of DNA but mapping to distinct loci in the reference). Variant calling remains a challenge with short-read sequence data, especially for low-complexity or repetitive regions of the genome where it is difficult to align reliably.

#### 1.4.2.2. Variant annotation

The importance of automatically predicting the functional effects of variants has increased due to the large volume of data generated by NGS platforms. Using computer programs to annotate variants enables study teams to filter and prioritize potential causal variants for further analysis.

Bioinformatics programs use different approaches to annotate variants, many focusing on SNPs and indels, which are easily detectable and evaluated. Some programs, for example, VEP (McLaren et al., 2016), can annotate SNPs, indels, and SVs. Despite the potential role of SVs in influencing traits, their accurate functional annotation remains a significant challenge (Han et al., 2020). Variant annotation programs use multiple techniques, from simple sequence and location analysis to predicting the functional effects of variants on proteins. The results of this analysis usually assign a functional category to each variant. Examples include “nonsynonymous” for a variant that changes the translation of a coding sequence and “intronic” for a variant located within an intron and not expected to affect gene splicing. A standard set of terms for annotating variants is provided through the Sequence Ontology (Cunningham et al., 2015). Variants may also be classified into a few broader categories of functional impact, for example, HIGH, MODERATE, LOW, and MODIFIER in the case of the annotation program SnpEff (Cingolani, Platts, et al., 2012) (for more details, see the Table 1.1). Some tools have more detailed risk categories or scores that indicate the probability of a negative impact (Pabinger et al., 2014).

**Table 1.1.** Types of functional impacts of annotated variants by SnpEff annotation tool

| Impact | Description | Consequence examples |
|---|---|---|
| HIGH | The variant is presumed to exert a significant disruptive effect on the protein, potentially leading to protein truncation, loss of function, or activating a mechanism known as nonsense mediated decay. | Stop_gained[1], Frameshift_variant[2] |
| MODERATE | A variant that is not disruptive but has the potential to alter the effectiveness of the protein. | Missense_variant[3], Inframe_deletion[4] |
| LOW | It is presumed to be mostly harmless or unlikely to alter the behavior of the protein significantly. | Synonymous_variant[5] |
| MODIFIER | Typically, these are non-coding variants or variants that affect non-coding genes. Predicting their impact is challenging, or there is insufficient evidence to suggest any significant effect. | Downstream_gene_variant[6] |

1. A sequence alteration in which at least one base within a codon is modified, causing the emergence of an early stop codon and ultimately resulting in a shortened polypeptide chain (*The MISO Sequence Ontology Browser*). 2. It is a genetic alteration brought about by the insertion or deletion of a specific number of nucleotides. This modification disrupts the reading frame, the way codons are grouped, leading to a profoundly different translation outcome compared to the original sequence (N. Li et al., 2011). 3. It refers to a point mutation where a single nucleotide alteration occurs within a codon, leading to the encoding of a different amino acid (Karch, 2016). 4. It occurs when the gene's reading frame remains intact and undisturbed, allowing to produce a partially truncated protein (King et al., 2014). 5. It alters the DNA sequence of a gene while keeping the amino acid sequence of the resulting protein unchanged (Dhindsa et al., 2022). 6. A sequence variation situated downstream (in the 3' direction) of a gene (Dhindsa et al., 2022).

Many computational variant annotation tools have been created that generate reports which can be used for subsequent rule-based filtering. SnpEff (Cingolani, Platts, et al., 2012), ANNOVAR (K. Wang et al., 2010), and Variant Effect Predictor (VEP) (McLaren et al., 2016) are popular variant annotation programs. Notably all three can be applied to a diverse range of genomes / species, including those of livestock species and of companion animals (of relevance to this thesis).

#### 1.4.2.3. Variant filtering

NGS's large volume of data presents new challenges and opportunities, requiring innovative computational and theoretical approaches to identify causal variants for a disease/trait of interest. While workflows and associated software have greatly improved the processing of raw data and production of high-confidence variant calls, filtering tens of thousands of candidates to identify a relevant subset for a specific study remains a complex task best suited for bioinformaticians (Sefid Dashti & Gamieldien, 2017). Variant filtering is a crucial secondary analysis step in NGS that involves identifying highly confident variants and eliminating falsely called ones (Insights, 2021).

Filtering serves two primary roles in the context of WGS sequence variants. The first is to remove false-positive variants. For example, the VariantFiltration tool from the GATK package is a commonly used to filter variants and remove potential false-positive calls. This tool plays a crucial role in the post-processing of variant calls to improve the accuracy and quality of the results. It is designed to apply filters to variant calls in Variant Call Format (VCF) files, allowing users to remove or flag variants based on specific criteria. VariantFiltration provides flexibility in filtering variants by allowing users to define custom filter expressions based on various annotations and variant attributes. Common annotations include quality scores, allele depth, strand bias, and allele frequency information. By specifying filtering criteria and thresholds, VariantFiltration enables users to retain variants of interest and remove potential false positives or artifacts.

The second primary role is to prioritize the true variants in terms of which are most likely to be contributing to the phenotypic variation of interest. Both typically employ automated approaches but can involve manual visualization and investigation steps. The prioritization of

variants in particular requires biological or biomedical reasoning, and biologists and clinicians are increasingly motivated to do the task themselves. Existing knowledge regarding the molecular and cellular mechanisms underlying a particular disease or phenotype can be highly beneficial when prioritizing probable candidates from a list of genes containing functional variants. Rarity/novelty predicted deleterious effects and expected segregation among affected/cases, and unaffected/control individuals in the study are the main criteria used to generate a list of candidate variants. Variants should not be dismissed as irrelevant solely because the knowledge filter does not reveal any disease, phenotypic, or functional links (Sefid Dashti & Gamieldien, 2017).

## 1.5. Molecular Methods for Validation of the Causality of Identified Candidate Mutations

Verifying that identified DNA variants or mutations are causative for a particular phenotype or disease often involves conducting molecular or functional assays to demonstrate their impact on gene expression, protein function, or other relevant cellular processes. Some of the commonly used methods for molecular and functional confirmation include reporter assays, gene editing, gene knockout or knockdown, and expression studies.

Reporter assays involve fusing the gene of interest (containing the identified mutation) with a reporter gene, such as luciferase or green fluorescent protein (GFP). Changes in the activity of the reporter gene can indicate the functional impact of the mutation. In this method, the wild-type and mutant versions of the gene are separately cloned into plasmids containing the reporter gene. The plasmids are then transfected into cells, and the reporter gene activity is measured and compared.

Gene-editing methodologies, such as CRISPR-Cas9, precisely incorporate the identified mutation into the genome and monitor the resulting effects. The identified mutation is integrated into the cellular or model organism genome, and then functional consequences are evaluated which provides direct evidence of the mutation's causality.

Gene knockout or knockdown refers to complete elimination (knockout) or reduction (knockdown) of the expression of the gene of interest in a model organism or cell line and observing the resulting phenotype. In this method, technologies like CRISPR-Cas9 can be used to create a knockout or knockdown of the gene. Subsequently, the impact of the absence or declined expression of the gene on cellular or organismal function are evaluated.

Expression studies involve analyzing the levels of gene expression in the presence of the identified mutation compared to the wild-type gene. Methods such as quantitative PCR, RNA sequencing, or protein expression assays can be used to assess changes in gene expression associated with the mutation.

## 1.6. Overarching Research Hypothesis and Objectives

This thesis encompasses three research projects concerned with the identification of sequence variants responsible for phenotypic variation, using NGS-based whole-genome sequencing. Subsequent chapters provide relevant background information on each project. Here I provide a summary of the projects in terms of the main hypotheses and objectives:

1) Swyer syndrome in cattle
    - Hypothesis: The Swyer syndrome cases in Irish beef cattle are caused by inherited or de novo sequence variants that can be detected by WGS.

- Objectives: Use WGS data from affected and control animals as well as existing knowledge of genes implicated in disorders (or differences) of sex development (DSD) to identify the cause of Swyer syndrome in beef cattle; determine whether inherited or de novo variants are involved; propose strategies to identify Swyer animals in the future using routine genomic evaluations.

2) Copper toxicosis in Dalmatian
    - Hypothesis: Copper toxicosis (CT) in the Dalmatian breed is caused by a recessive mutation that can be discovered through the WGS of presumptive affected, carrier, and non-carrier animals.
    - Objectives: Use WGS from cases and controls as well as knowledge of CT mechanisms in other species and breeds to identify the cause of CT in Dalmatians; develop low-cost genetic test to identify carriers.

3) Infertility/subfertility in cattle:
    - Hypothesis: Cattle exhibiting low fertility carry recessive lethal alleles causing embryo loss that can be detected by WGS.
    - Objectives: Use WGS from cattle exhibiting low fertility to identify candidate recessive lethal mutations that may explain reduced fertility in cattle.

## 1.7. References

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human

genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Abyzov, A., Li, S., Kim, D. R., Mohiyuddin, M., Stütz, A. M., Parrish, N. F., Mu, X. J., Clark, W., Chen, K., & Hurles, M. (2015). Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications*, *6*(1), 1–12.

Ajay, S. S., Parker, S. C. J., Abaan, H. O., Fajardo, K. V. F., & Margulies, E. H. (2011). Accurate and comprehensive sequencing of personal genomes. *Genome Research*, *21*(9), 1498–1505. https://doi.org/10.1101/gr.123638.111

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), Article 7571. https://doi.org/10.1038/nature15393

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood. Education and Practice Edition*, *98*(6), 236–238. https://doi.org/10.1136/archdischild-2013-304340

Ben-Ari, G., & Lavi, U. (2012). 11—Marker-assisted selection in plant breeding. In A. Altman & P. M. Hasegawa (Eds.), *Plant Biotechnology and Agriculture* (pp. 163–184). Academic Press. https://doi.org/10.1016/B978-0-12-381466-1.00011-0

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter,

R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bourneuf, E., Otz, P., Pausch, H., Jagannathan, V., Michot, P., Grohs, C., Piton, G., Ammermüller, S., Deloche, M.-C., Fritz, S., Leclerc, H., Péchoux, C., Boukadiri, A., Hozé, C., Saintilan, R., Créchet, F., Mosca, M., Segelke, D., Guillaume, F., … Capitan, A. (2017). Rapid Discovery of De Novo Deleterious Mutations in Cattle Enhances the Value of Livestock as Model Species. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-11523-3

*CanineHD Whole-Genome Genotyping BeadChip*. (n.d.). Retrieved August 31, 2023, from https://www.illumina.com/products/by-type/microarray-kits/caninehd.html

Carroll, D. (2001). Genetic Recombination. In S. Brenner & J. H. Miller (Eds.), *Encyclopedia of Genetics* (pp. 841–845). Academic Press. https://doi.org/10.1006/rwgn.2001.0543

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews. Genetics*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, *19*(3 Pt 1), 233.

Chen, L., Pryce, J. E., Hayes, B. J., & Daetwyler, H. D. (2021). Investigating the Effect of Imputed Structural Variants from Whole-Genome Sequence on Genome-Wide Association and Genomic Prediction in Dairy Cattle. *Animals : An Open Access Journal from MDPI*, *11*(2), 541. https://doi.org/10.3390/ani11020541

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., & Montgomery, S. B. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, *49*(5), 692–699.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., & Sanjad, S. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*, *106*(45), 19096–19101.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Clark, D. P., & Pazdernik, N. J. (2013). Chapter 23—Mutations and Repair. In D. P. Clark & N. J. Pazdernik (Eds.), *Molecular Biology (Second Edition)* (pp. 721–766). Academic Press. https://doi.org/10.1016/B978-0-12-378594-7.00023-8

A global reference for human genetic variation, 526 Nature 68 (Nature Publishing Group 2015).

Consortium, B. H., Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., Green, R. D., Hamernik, D. L., & Kappes, S. M. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*, *324*(5926), 528–532.

Coronado, V. A., Damaraju, D., Kohijoki, R., & Cox, D. W. (2003). New haplotypes in the Bedlington terrier indicate complexity in copper toxicosis. *Mammalian Genome*, *14*(7), 483–491.

Cunningham, F., Moore, B., Ruiz-Schultz, N., Ritchie, G. R., & Eilbeck, K. (2015). Improving the Sequence Ontology terminology for genomic variant annotation. *Journal of Biomedical Semantics*, *6*(1), 32. https://doi.org/10.1186/s13326-015-0030-4

D'Agaro, E., Favaro, A., Rosa, D., D'Agaro, E., Favaro, A., & Rosa, D. (2021). Canine Genetics and Genomics. In *Canine Genetics, Health and Medicine*. IntechOpen. https://doi.org/10.5772/intechopen.95781

Dagenais, S. L., Guevara-Fujita, M., Glover, T. W., Burgess, A. C., Loechel, R., Brewer, G. J., Yuzbasiyan-Gurkan, V., & Miller, D. E. (1999). *The canine copper toxicosis locus is not syntenic with ATP7B or ATX1 and maps to a region showing homology to human 2p21*.

Deng, X., den Bakker, H., & Hendriksen, R. (2016). Genomic Epidemiology: Whole-Genome-Sequencing—Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annual Review of Food Science and Technology*, *7*. https://doi.org/10.1146/annurev-food-041715-033259

Dhindsa, R. S., Wang, Q., Vitsios, D., Burren, O. S., Hu, F., DiCarlo, J. E., Kruglyak, L., MacArthur, D. G., Hurles, M. E., & Petrovski, S. (2022). A minimal role for synonymous variation in human disease. *The American Journal of Human Genetics*, *109*(12), 2105–2109. https://doi.org/10.1016/j.ajhg.2022.10.016

Dirksen, K., Spee, B., Penning, L. C., Ingh, T. S. G. A. M. van den, Burgener, I. A., Watson, A. L., Koerkamp, M. G., Rothuizen, J., Steenbeek, F. G. van, & Fieten, H. (2017). Gene expression patterns in the progression of canine copper-associated chronic hepatitis. *PLOS ONE*, *12*(5), e0176826. https://doi.org/10.1371/journal.pone.0176826

Dunislawska, A., Lachmańska, J., Slawińska, A., & Siwek, M. (2017). Next generation sequencing in animal science-a review. *Animal Science Papers & Reports*, *35*(3).

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, *372*(6537), eabf7117. https://doi.org/10.1126/science.abf7117

Fang, H., Wu, Y., Narzisi, G., O'Rawe, J. A., Barrón, L. T. J., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M. C., & Lyon, G. J. (2014). Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Medicine*, *6*(10), 89. https://doi.org/10.1186/s13073-014-0089-z

Fareed, M., & Afzal, M. (2013). Single nucleotide polymorphism in genome-wide association of human population: A tool for broad spectrum service. *Egyptian Journal of Medical Human Genetics*, *14*(2), 123–134. https://doi.org/10.1016/j.ejmhg.2012.08.001

Fay, D. (2006). Genetic mapping and manipulation: Chapter 1-Introduction and basics. *WormBook*, *17*, 1–12.

Forman, O. P., Boursnell, M. E. G., Dunmore, B. J., Stendall, N., van den Sluis, B., Fretwell, N., Jones, C., Wijmenga, C., Rothuizen, J., van Oost, B. A., Holmes, N. G., Binns, M. M., & Jones, P. (2005). Characterization of the COMMD1 (MURR1) mutation causing copper toxicosis in Bedlington terriers. *Animal Genetics*, *36*(6), 497–501. https://doi.org/10.1111/j.1365-2052.2005.01360.x

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., & Hurles, M. E. (2006). Copy number variation: New insights in genome diversity. *Genome Research*, *16*(8), 949–961.

Fuchs, E. (2018). Chapter 106—Haploidentical Hematopoietic Cell Transplantation. In R. Hoffman, E. J. Benz, L. E. Silberstein, H. E. Heslop, J. I. Weitz, J. Anastasi, M. E. Salama, & S. A. Abutalib (Eds.), *Hematology (Seventh Edition)* (pp. 1617-1632.e6). Elsevier. https://doi.org/10.1016/B978-0-323-35762-3.00106-2

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., & Altshuler, D. (2002). The Structure of Haplotype Blocks in the Human Genome. *Science*, *296*(5576), 2225–2229. https://doi.org/10.1126/science.1069424

García-Nieto, P. E., Morrison, A. J., & Fraser, H. B. (2019). The somatic mutation landscape of the human body. *Genome Biology*, *20*(1), 1–20.

Genome 10K Community of Scientists. (2009). Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of Heredity*, *100*(6), 659–674. https://doi.org/10.1093/jhered/esp086

Gonzaga-Jauregui, C., Lupski, J. R., & Gibbs, R. A. (2012). Human genome sequencing in health and disease. *Annual Review of Medicine*, *63*, 35–61.

Gottlieb, S. F., Gulani, A., & Tegay, D. H. (2023). Genetics, Meiosis. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK482462/

Guilherme, L., Köhler, K. F., & Kalil, J. (2011). Chapter 2 - Rheumatic heart disease: Mediation by complex immune events. In G. S. Makowski (Ed.), *Advances in Clinical Chemistry* (Vol. 53, pp. 31–50). Elsevier. https://doi.org/10.1016/B978-0-12-385855-9.00002-3

Gupta, A. K., & Gupta, U. D. (2020). Next generation sequencing and its applications. In *Animal Biotechnology* (pp. 395–421). Elsevier.

Han, L., Zhao, X., Benton, M. L., Perumal, T., Collins, R. L., Hoffman, G. E., Johnson, J. S., Sloofman, L., Wang, H. Z., Stone, M. R., CommonMind Consortium, Brennand, K. J., Brand, H., Sieberts, S. K., Marenco, S., Peters, M. A., Lipska, B. K., Roussos, P., Capra, J. A., … Ruderfer, D. M. (2020). Functional annotation of rare structural variation in the human brain. *Nature Communications*, *11*(1), 2990. https://doi.org/10.1038/s41467-020-16736-1

Hayes, B. J., & Daetwyler, H. D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual Review of Animal Biosciences*, *7*, 89–102. https://doi.org/10.1146/annurev-animal-020518-115024

Haywood, S., Boursnell, M., Loughran, M. J., Trafford, J., Isherwood, D., Liu, X., Olohan, L., & Carter, S. D. (2016). Copper toxicosis in non-COMMD1 Bedlington terriers is associated with metal transport gene ABCA12. *Journal of Trace Elements in Medicine and Biology*, *35*, 83–89. https://doi.org/10.1016/j.jtemb.2016.01.015

Haywood, S., Fuentealba, I. C., Kemp, S. J., & Trafford, J. (2001). Copper toxicosis in the Bedlington terrier: A diagnostic dilemma. *Journal of Small Animal Practice*, *42*(4), 181–185.

Hill, W. G. (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, *33*(2), Article 2. https://doi.org/10.1038/hdy.1974.89

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, *38*(6), 226–231. https://doi.org/10.1007/BF01245622

Holloway, J. W., & Prescott, S. L. (2017). Chapter 2—The Origins of Allergic Disease. In R. E. O'Hehir, S. T. Holgate, & A. Sheikh (Eds.), *Middleton's Allergy Essentials* (pp. 29–50). Elsevier. https://doi.org/10.1016/B978-0-323-37579-5.00002-7

Holopainen, S., Hytönen, M. K., Syrjä, P., Arumilli, M., Järvinen, A.-K., Rajamäki, M., & Lohi, H. (2017). ANLN truncation causes a familial fatal acute respiratory distress syndrome in Dalmatian dogs. *PLOS Genetics*, *13*(2), e1006625. https://doi.org/10.1371/journal.pgen.1006625

Hou, Y., Liu, G. E., Bickhart, D. M., Matukumalli, L. K., Li, C., Song, J., Gasbarre, L. C., Van Tassell, C. P., & Sonstegard, T. S. (2012). Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Functional & Integrative Genomics*, *12*, 81–92.

Huber, D., Voith von Voithenberg, L., & Kaigala, G. V. (2018). Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering*, *1*, 15–24. https://doi.org/10.1016/j.mne.2018.10.006

Hyun, C., Lavulo, L. T., & Filippich, L. J. (2004). Evaluation of haplotypes associated with copper toxicosis in Bedlington Terriers in Australia. *American Journal of Veterinary Research*, *65*(11), 1573–1579.

Insights, Q. D. (2021, February 17). Tips for Variant Filtering & Prioritization. *QIAGEN Digital Insights*. https://digitalinsights.qiagen.com/news/blog/clinical/challenges-in-variant-filtering/

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*, 14061. https://doi.org/10.1038/ncomms14061

Jiang, Y., Jiang, Y., Wang, S., Zhang, Q., & Ding, X. (2019). Optimal sequencing depth design for whole genome re-sequencing in pigs. *BMC Bioinformatics*, *20*(1), 556. https://doi.org/10.1186/s12859-019-3164-z

Joiret, M., Mahachie John, J. M., Gusareva, E. S., & Van Steen, K. (2019). Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Mining*, *12*(1), 11. https://doi.org/10.1186/s13040-019-0199-7

Kadri, N. K., Sahana, G., Charlier, C., Iso-Touru, T., Guldbrandtsen, B., Karim, L., Nielsen, U. S., Panitz, F., Aamand, G. P., & Schulman, N. (2014). A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: Additional evidence for the common occurrence of balancing selection in livestock. *PLoS Genetics*, *10*(1), e1004049.

Karch, S. B. (2016). Autopsy: Molecular. In J. Payne-James & R. W. Byard (Eds.), *Encyclopedia of Forensic and Legal Medicine (Second Edition)* (pp. 290–296). Elsevier. https://doi.org/10.1016/B978-0-12-800034-2.00041-0

Kilpinen, H., & Barrett, J. C. (2013). How next-generation sequencing is transforming complex disease genetics. *Trends in Genetics: TIG*, *29*(1), 23–30. https://doi.org/10.1016/j.tig.2012.10.001

Kim, S. Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Pedersen, O., Wang, J., & Nielsen, R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. *Genetic Epidemiology*, *34*(5), 479–491. https://doi.org/10.1002/gepi.20501

King, R. C., Mulligan, P. K., & Stansfield, W. D. (2014). In-frame mutation. In *A Dictionary of Genetics*. Oxford University Press.

https://www.oxfordreference.com/display/10.1093/acref/9780199766444.001.0001/acref-9780199766444-e-3381

Kishikawa, T., Momozawa, Y., Ozeki, T., Mushiroda, T., Inohara, H., Kamatani, Y., Kubo, M., & Okada, Y. (2019). Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Scientific Reports*, *9*(1), 1–10.

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., Munson, K. M., Hastie, A. R., Diekhans, M., Hormozdiari, F., Lorusso, N., Hoekzema, K., Qiu, R., Clark, K., Raja, A., … Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science (New York, N.Y.)*, *360*(6393), eaar6343. https://doi.org/10.1126/science.aar6343

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, *22*(2), Article 2. https://doi.org/10.1038/9642

Kulski, J. K. (2016). Next-generation sequencing—An overview of the history, tools, and "Omic" applications. *Next Generation Sequencing-Advances, Applications and Challenges*, *10*, 61964.

Lam, H. Y., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'huallachain, M., Dewey, F. E., Habegger, L., Ashley, E. A., & Gerstein, M. B. (2012). Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology*, *30*(1), 78–82.

Lathrop, G. M., Lalouel, J. M., Julier, C., & Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences*, *81*(11), 3443–3446.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), Article 7616. https://doi.org/10.1038/nature19057

Lewis, C. M., & Knight, J. (2012). Introduction to genetic association studies. *Cold Spring Harbor Protocols*, *2012*(3), pdb-top068163.

Li, N., Zhou, Y., Du, L., Wei, M., & Chen, X. (2011). Overview of Cytochrome P450 1B1 gene mutations in patients with primary congenital glaucoma. *Experimental Eye Research*, *93*(5), 572–579. https://doi.org/10.1016/j.exer.2011.07.009

Litonjua, A. A., & Celedón, J. C. (2006). GENETICS | Gene Association Studies. In G. J. Laurent & S. D. Shapiro (Eds.), *Encyclopedia of Respiratory Medicine* (pp. 240–247). Academic Press. https://doi.org/10.1016/B0-12-370879-6/00165-4

Liu, B. H. (2017). *Statistical genomics: Linkage, mapping, and QTL analysis*. CRC press.

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, *9*(1), 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359

Mardis, E. R., & Wilson, R. K. (2009). Cancer genome sequencing: A review. *Human Molecular Genetics*, *18*(R2), R163–R168.

Mason, R. J., Slutsky, A., Murray, J. F., Nadel, J. A., & Gotway, M. B. (2015). *Murray & Nadel's Textbook of Respiratory Medicine E-Book*. Elsevier Health Sciences.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Modena, B. D., Doroudchi, A., Patel, P., & Sathish, V. (2019). Chapter 18—Leveraging genomics to uncover the genetic, environmental and age-related factors leading to asthma. In G. S. Ginsburg, H. F. Willard, E. L. Tsalik, & C. W. Woods (Eds.), *Genomic and Precision Medicine (Third Edition)* (pp. 331–381). Academic Press. https://doi.org/10.1016/B978-0-12-801496-7.00018-6

Mohiuddin, M., Kooy, R. F., & Pearson, C. E. (2022). De novo mutations, genetic mosaicism and human disease. *Frontiers in Genetics.-Place of Publication Unknown*, *13*, 1–20.

Motsinger, A. A., Reif, D. M., Fanelli, T. J., Davis, A. C., & Ritchie, M. D. (2007). Linkage Disequilibrium in Genetic Association Studies Improves the Performance of Grammatical Evolution Neural Networks. *Proceedings of the ... IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, *2007*, 1–8.

Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*(R2), R131–R136.

Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the Basics of NGS: From Mechanism to Variant Calling. *Current Genetic Medicine Reports*, *3*(4), 158–165. https://doi.org/10.1007/s40142-015-0076-8

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297–304.

Nanji, M. S., & Cox, D. W. (1999). The copper chaperone Atox1 in canine copper toxicosis in Bedlington terriers. *Genomics*, *62*(1), 108–112.

Oetting, W. S. (2012). Exome and genome analysis as a tool for disease identification and treatment: The 2011 Human Genome Variation Society scientific meeting. *Human Mutation*, *33*(3), 586–590. https://doi.org/10.1002/humu.22018

Ostrander, E. A., Wang, G.-D., Larson, G., vonHoldt, B. M., Davis, B. W., Jagannathan, V., Hitte, C., Wayne, R. K., & Zhang, Y.-P. (2019). Dog10K: An international sequencing effort to advance studies of canine domestication, phenotypes and health. *National Science Review*, *6*(4), 810–824. https://doi.org/10.1093/nsr/nwz049

*Overview of Structural Variation*. (n.d.). Retrieved May 17, 2023, from https://www.ncbi.nlm.nih.gov/dbvar/content/overview/

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, *15*(2), 256–278. https://doi.org/10.1093/bib/bbs086

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, *39*(10), 1256–1260. https://doi.org/10.1038/ng2123

Persoskie, A., & Ferrer, R. A. (2017). A Most Odd Ratio: Interpreting and Describing Odds Ratios. *American Journal of Preventive Medicine*, *52*(2), 224–228. https://doi.org/10.1016/j.amepre.2016.07.030

Porterfield, A. (2015, May 15). *A Guide to Genetic Variants*. https://bitesizebio.com/23996/whats-so-important-about-variants/

Rantsiou, K., Kathariou, S., Winkler, A., Skandamis, P., Saint-Cyr, M. J., Rouzeau-Szynalski, K., & Amézquita, A. (2018). Next generation microbiological risk assessment: Opportunities of whole genome sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment. *International Journal of Food Microbiology*, *287*, 3–9. https://doi.org/10.1016/j.ijfoodmicro.2017.11.007

Salmela, E., Niskanen, J., Arumilli, M., Donner, J., Lohi, H., & Hytönen, M. K. (2019). A novel *KRT71* variant in curly-coated dogs. *Animal Genetics*, *50*(1), 101–104. https://doi.org/10.1111/age.12746

Sanger, F., Nicklen, S., & Coulson, A. R. (1992). DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology (Reading, Mass.)*, *24*, 104–108.

Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J., & Nelson, S. F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics (Oxford, England)*, *27*(19), 2648–2654. https://doi.org/10.1093/bioinformatics/btr462

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (n.d.). *Piercing the dark matter: Bioinformatics of long-range sequencing and mapping | Nature Reviews Genetics*. Retrieved May 30, 2022, from https://www.nature.com/articles/s41576-018-0003-4

Sefid Dashti, M. J., & Gamieldien, J. (2017). A practical guide to filtering and prioritizing genetic variants. *BioTechniques*, *62*(1), 18–30. https://doi.org/10.2144/000114492

Stachowiak, M., Szczerbal, I., & Switonski, M. (2016). Chapter Eight—Genetics of Adiposity in Large Animal Models for Human Obesity—Studies on Pigs and Dogs. In Y.-X. Tao (Ed.), *Progress in Molecular Biology and Translational Science* (Vol. 140, pp. 233–270). Academic Press. https://doi.org/10.1016/bs.pmbts.2016.01.001

Ståhl, P. L., & Lundeberg, J. (2012). Toward the Single-Hour High-Quality Genome. *Annual Review of Biochemistry*, *81*(1), 359–378. https://doi.org/10.1146/annurev-biochem-060410-094158

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, *61*, 437–455.

Straiton, J., Free, T., Sawyer, A., & Martin, J. (2019). *From Sanger sequencing to genome databases and beyond*. Future Science.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

THE BOVINE HAPMAP CONSORTIUM, Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., Green, R. D., Hamernik, D. L., Kappes, S. M., Lien, S., Matukumalli, L. K., McEwan, J. C., Nazareth, L. V., Schnabel, R. D., Weinstock, G. M., Wheeler, D. A., Ajmone-Marsan, P., Boettcher, P. J., … Dodds, K. G.

(2009). Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science*, *324*(5926), 528–532. https://doi.org/10.1126/science.1167936

*The MISO Sequence Ontology Browser—STOP_GAINED*. (n.d.). Retrieved August 30, 2023, from http://www.sequenceontology.org/browser/current_svn/term/SO:0001587

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), Article 1. https://doi.org/10.1038/s43586-021-00056-9

van de Sluis, B. J., Breen, M., Nanji, M., van Wolferen, M., de Jong, P., Binns, M. M., Pearson, P. L., Kuipers, J., Rothuizen, J., & Cox, D. W. (1999). Genetic mapping of the copper toxicosis locus in Bedlington terriers to dog chromosome 10, in a region syntenic to human chromosome region 2p13–p16. *Human Molecular Genetics*, *8*(3), 501–507.

van De Sluis, B., Rothuizen, J., Pearson, P. L., van Oost, B. A., & Wijmenga, C. (2002). Identification of a new copper metabolism gene by positional cloning in a purebred dog population. *Human Molecular Genetics*, *11*(2), 165–173. https://doi.org/10.1093/hmg/11.2.165

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*(9), 418–426.

Veltman, J. A., & Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics*, *13*(8), 565–575.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, Q., Lu, Q., & Zhao, H. (2015). A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Frontiers in Genetics*, *6*, 149. https://doi.org/10.3389/fgene.2015.00149

Wu, P., Wang, K., Zhou, J., Chen, D., Yang, Q., Yang, X., Liu, Y., Feng, B., Jiang, A., & Shen, L. (2019). GWAS on imputed whole-genome Resequencing from genotyping-by-sequencing data for farrowing interval of different parities in pigs. *Frontiers in Genetics*, 1012.

Xing, K., Zhu, F., Zhai, L., Chen, S., Tan, Z., Sun, Y., Hou, Z., & Wang, C. (2016). Identification of genes for controlling swine adipose deposition by integrating transcriptome, whole-genome resequencing, and quantitative trait loci data. *Scientific Reports*, *6*, 23219. https://doi.org/10.1038/srep23219

Yuzbasiyan-Gurkan, V., Blanton, S. H., Cao, Y., Ferguson, P., Li, J., Venta, P. J., & Brewer, G. J. (1997). Linkage of a microsatellite marker to the canine copper toxicosis locus in Bedlington terriers. *American Journal of Veterinary Research*, *58*(1), 23–27.

Zhang, H., Du, Z.-Q., Dong, J.-Q., Wang, H.-X., Shi, H.-Y., Wang, N., Wang, S.-Z., & Li, H. (2014). Detection of genome-wide copy number variations in two chicken lines divergently selected for abdominal fat content. *BMC Genomics*, *15*, 1–12.

# CHAPTER 2. Identification of Candidate Variants Causing a Disorder of Sex Development (DSD) Called Swyer Syndrome in Cattle.

## 2.1. Introduction and Literature Review

Swyer syndrome in humans is a rare congenital condition characterized by the atypical development of sexual organs. It is classified as a disorder (or difference) of sex development (DSD) (Baetens et al., 2018). This condition affects sexual traits and is characterized by the abnormal growth of sex chromosomes, gonads, or anatomy. Swyer syndrome is also called XY complete gonadal dysgenesis, where "dysgenesis" denotes an abnormal development (Swyer Syndrome, 2022). Typically, in individuals without Swyer syndrome, sex development proceeds through a standard process that depends on the sex chromosomes they possess. However, individuals with Swyer syndrome experience altered sex development due to genetic mutations that cause either a gain or loss of function in genes responsible for gonad development or steroidogenesis[1] (Eggers et al., 2015). Females usually possess two X chromosomes (female cattle karyotype: 60, XX), while males have one X and one Y chromosome (male cattle karyotype: 60, XY). However, in Swyer syndrome, even though individuals have X and Y chromosomes like males, they develop abnormal female reproductive organs. As a result, these XY individuals develop female external genital organs and some female internal reproductive compositions with defects, leading to infertility (Ostrer, 2019). Initially described in humans by Gim Swyer in 1955 (Banoth et al., 2018), Swyer syndrome has been described in four species of the family Canidae (the dog, red fox, arctic fox and Chinese raccoon dog) (Nowacka-Woszuk & Switonski, 2009) and other species including buffalo (L. Iannuzzi et al., 2001), and cattle (Hare et al., 1994).

[1]. Steroidogenesis refers to the intricate series of biochemical processes wherein cholesterol is transformed into steroid hormones, necessitating the involvement of transport proteins, enzymes, redox partners, and cofactors (Miller & Auchus, 2011).

Studies have shown that various genes are associated with Swyer syndrome in humans, some of which play a role in the formation and functioning of gonads and determining the male sex through the development of testes. Some identified genes linked to Swyer syndrome include *SRY*, *SOX9*, *SF1*, *WT1*, and *DHH* (Délot & Vilain, 2019). Mutations in these genes can result in the development of non-functioning gonads and the inability to produce viable gametes. About 50% of people with Swyer syndrome cannot be explained by the currently available molecular approaches, suggesting that several undiscovered genes might be involved in this disorder (Arboleda & Vilain, 2016; Baetens et al., 2018; Délot & Vilain, 2019).

Our study was initiated after the observation of a mysterious condition in multiple cows in Ireland. To improve the genetic merit of the country's beef cattle population and increase productivity, the Irish Cattle Breeding Federation has initiated a national genetic testing program, which the Irish government and the European Union support (Szenthe, 2022). In Ireland, the average pregnancy rate for mated cows is 87% (Irish Cattle Breeding Federation (ICBF), 2022), implying that 130,000 cows out of a herd of 1 million are not giving birth to any offspring. As a result, improving the pregnancy rate benefits farmers financially and enhances environmental sustainability (Szenthe, 2022). Consequently, farmers were urged to provide DNA samples from their female cattle and potential breeding heifers to improve their herd's genetic merit. These samples were then investigated using a large-scale SNP genotyping method to identify the superior females for breeding (genomic selection). More than 1 million samples were gathered in total. The issue that led to this study was that genetic testing reports returned to the farmers, in a small number of cases, the reported genetic sex for an individual was inconsistent with the phenotypic sex that the farmers had observed. Initially they thought that it was due to sample acquisition or handling results, however, multiple new re-tests led to the same outcomes. All cases involved a

phenotypically female animal receiving an SNP-based sex assignment of male, prompting an investigation into whether Swyer syndrome-type mutations were involved.

## 2.2. Outlines of the Research Objectives

a) Identification and genotyping of SNPs and indels.

b) Identification and genotyping of structural variants.

c) Determination of sex chromosome complement.

d) Identification of known genes linked to Swyer syndrome in other species through literature review.

e) Detection of candidate causal variants by variant filtering.

f) Detection of candidate causal variants from read coverage and read alignment analysis.

## 2.3. Methods/Experimental Design and Materials

### 2.3.1. Animal samples for sequencing

Sequence data for eight affected Irish animals and one unaffected dam was collected prior to our involvement in the study (Table 2.1). Animal ethics approval was not sought for these samples as they were already obtained through ongoing routine genetic testing or processing of commercial cattle. For some analyses, sequence data from non-Swyer XX and XY animals from the recessive lethal study chapter (chapter 4) was used. Animal ethics information for those samples is provided in that chapter.

**Table 2.1.** Sequenced presumptive XY female (Swyer animal's) information.

| Sample | Breed | Phenotypic Sex | Sire breed | Dam breed | NCBI SRA[9] ID |
|---|---|---|---|---|---|
| Control XX female | Limousin Cross | F | LM[1] | LMX[6] | SRS13440364 |
| XY female (Swyer 1) | Limousin Cross | F | LM | LMX | SRS13440365 |
| XY female (Swyer 2) | Belgian Blue cross | F | BB[2] | LMX | SRS13440366 |
| XY female (Swyer 3) | Limousin cross | F | LM | LMX | SRS13440367 |
| XY female (Swyer 4) | Limousin cross | F | LM | HEX[7] | SRS13440368 |
| XY female (Swyer 5) | Charolais cross | F | CH[3] | SIX[8] | SRS13440369 |
| XY female (Swyer 6) | Simmental cross | F | SI[4] | SIX | SRS13440370 |
| XY female (Swyer 7) | Angus | F | AA[5] | AA | SRS13440371 |
| XY female (Swyer 8) | Charolais | F | CH | CH | SRS13440372 |

[1]. LM = Limousin, [2]. BB= Belgian Blue, [3]. CH= Charolais, [4]. SI= Simmental, [5]. AA= Angus, [6]. LMX= Limousin Cross, [7]. HEX= Hereford Cross, [8]. SIX= Simmental Cross, [9]. SRA (Sequence Read Archive): The NCBI Sequence Read Archive is a repository that stores raw sequencing data from various high-throughput sequencing platforms, such as next-generation sequencing (NGS) and Sanger sequencing. Researchers and scientists deposit their sequencing data into the SRA to make it publicly accessible for analysis and research purposes.

### 2.3.2. SNPs and indels identification and genotyping

A bioinformatics workflow was used to call SNPs and indels using a high-performance computing (HPC) accessed through Compute Canada. Before applying the raw data (FASTQ files) to the variant calling analysis, the initial stages of the workflow involved data quality control and pre-processing of the raw reads. These steps were undertaken as part of the variant calling process to ensure data integrity and prepare the data for further analysis. The implementation of quality control measures not only enables the assessment of the quality of sequencing data but also facilitates the acquisition of high-quality data for downstream analysis (He et al., 2020). In addition, these steps are required to convert raw FASTQ files into a BAM file ready for analysis. Quality control was performed using the FastQC v0.11.9 tool (Andrews, 2010) on the raw FASTQ

inputs. The Trimmomatic v0.36 (Bolger et al., 2014) was employed to trim and remove adapters from the raw data. For read mapping, the trimmed reads were aligned against the standard ARS-UCD1.2_Btau5.0.1Y reference genome of Bos taurus using the BWA v0.7.17 software package (H. Li & Durbin, 2009). The ARS-UCD1.2_Btau5.0.1Y reference genome can be accessed at https://sites.ualberta.ca/~stothard/1000_bull_genomes/.

Another preprocessing step of the workflow, was base quality score recalibration (BQSR), which was carried out by the BaseRecalibrator tool of Genome Analysis Toolkit (GATK) v3.8 (McKenna et al., 2010), and a VCF file comprising identified variant locations provided through the 1000 Bull Genomes Project (Hayes & Daetwyler, 2019), which can be accessed at https://sites.ualberta.ca/~stothard/1000_bull_genomes/ARS1.2PlusY_BQSR.vcf.gz. BQSR improves variant calling accuracy by detecting and correcting systematic errors in base call estimation during next-generation sequencing (NGS). This step reduces false-positive and false-negative variant calls, enhancing the reliability and accuracy of variant identification.

Another tool from GATK called AnalyzeCovariates was used to evaluate the quality of the recalibration process. In the next step, GATK HaplotypeCaller was used to identify potential variant positions, and then the generated GVCF output files were combined and genotyped with GATK GenotypeGVCFs. Next, GATK SelectVariants extracted SNPs and indels from the output file. In the final output to soft filter the detected SNPs and indels, GATK VariantFiltration was applied.

### 2.3.3. Structural variants (SVs) identification and genotyping

Smoove v0.2.6 (Pedersen et al., 2018/2020) was used to detect and genotype structural variants (SVs), excluding gapped regions and regions where coverage is more than four standard deviations higher than the mean coverage, calculated across 10,000 bp windows by Mosdepth v0.3.2 (Pedersen & Quinlan, 2018). Genotyping was followed by running Duphold v0.2.3 (Pedersen & Quinlan, 2019) to annotate SV calls with sequence depth information. SnpSift v4.3t (Cingolani, Patel, et al., 2012) was used to soft filter variants that satisfied the pre-defined criteria.

### 2.3.4. Sex chromosome assessment of study samples using sequence data

Multiple approaches were used to infer the sex chromosome complement of the sample animals from the analyzed next-generation sequencing (NGS) outputs.

#### 2.3.4.1. X: Y coverage ratio

The SAMtools v1.12 idxstats command (H. Li et al., 2009) was used to count the number of reads mapped to the X and Y chromosomes in the BAM files. Each count was divided by the length of the respective chromosome, and these normalized counts were used to calculate the final ratio.

#### 2.3.4.2. X: autosome coverage ratio

The SAMtools v1.12 idxstats command (H. Li et al., 2009) was used to count the number of reads mapped to the X chromosome and chromosome 10 in the BAM files. Each count was

divided by the length of the respective chromosome, and these normalized counts were used to calculate the final ratio.

#### 2.3.4.3. Y: autosome coverage ratio

The SAMtools v1.12 idxstats command (H. Li et al., 2009) was used to count the number of reads mapped to chromosome Y and chromosome 10 in the BAM files. Each count was divided by the length of the respective chromosome, and these normalized counts were used to calculate the final ratio.

### 2.3.5. Detection of candidate causal variants from inspection of DSD gene alignments

#### 2.3.5.1. Visualization of BAM files

A compilation of 85 genes associated with disorders of sex development (DSD) in humans was made through literature searches (Table 2.2). The list of 85 DSD gene regions, including 1 kb upstream and 1 kb downstream, in each study sample were inspected using IGV v2.12.3 (Thorvaldsdóttir et al., 2013). The location of the bovine orthologs of these genes, classified as autosomal, X-linked, or chromosome-unassigned, was determined for the ARS-UCD1.2 assembly using Ensembl BioMart (release 105) (Howe et al., 2021). In addition, the location of *SRY* on the Y chromosome of the Btau5.0.1 assembly was retrieved from NCBI (release 105) (Sayers et al., 2021). The process involved loading the BAM files of all study samples into IGV, along with the DSD gene locations that had flanking sequences added in BED format. An attributes file was also supplied to display sample information in IGV. The gene regions were visually inspected for any evidence of low or absent reads and read pairing and insert size information were examined for

any signs of deletions, insertions, inversions, and translocations. The Y chromosome was also analyzed similarly using IGV. In this case, all Y chromosome gene locations were obtained from NCBI and visually inspected to identify regions with missing coverage in the study samples compared to the control samples.

**Table 2.2.** Information of genes previously reported to be associated with DSDs.

| Ensembl Gene ID | Gene name | Chromosome /scaffold name | Gene start (bp) | Gene end (bp) |
|---|---|---|---|---|
| ENSBTAG00000000124 | *DHH* | 5 | 30728408 | 30733209 |
| ENSBTAG00000000730 | *GCG* | 2 | 34292569 | 34302949 |
| ENSBTAG00000001002 | *TCF7* | 7 | 45768913 | 45800957 |
| ENSBTAG00000001315 | *DMRTA1* | 8 | 21730255 | 21735209 |
| ENSBTAG00000001346 | *STRA8* | 4 | 98862169 | 98885072 |
| ENSBTAG00000001649 | *ZFPM2* | 14 | 58819937 | 59239114 |
| ENSBTAG00000001812 | *H1-8* | 22 | 56187762 | 56218695 |
| ENSBTAG00000001992 | *CYP51A1* | 4 | 9459764 | 9476713 |
| ENSBTAG00000002505 | *GPR3* | 2 | 125893878 | 125894870 |
| ENSBTAG00000003027 | *EMX2* | 26 | 37830786 | 37837033 |
| ENSBTAG00000003108 | *SRD5A2* | 11 | 14324110 | 14370453 |
| ENSBTAG00000003837 | *RSPO1* | 3 | 108267191 | 108288596 |
| ENSBTAG00000004257 | *TAF4B* | 24 | 30501605 | 30607271 |
| ENSBTAG00000004886 | *ZAR1* | 6 | 67172309 | 67176434 |
| ENSBTAG00000005370 | *TMTC1* | 5 | 79807657 | 80114645 |
| ENSBTAG00000005425 | *GATA4* | 8 | 7642403 | 7695092 |
| ENSBTAG00000005734 | *GATA6* | 24 | 34163203 | 34192624 |
| ENSBTAG00000006934 | *CYP11A1* | 21 | 34328403 | 34342900 |
| ENSBTAG00000007173 | *PDGFRA* | 6 | 69723655 | 69771549 |
| ENSBTAG00000007605 | *FKBP4* | 5 | 106946944 | 106955086 |
| ENSBTAG00000008062 | *DMRT2* | 8 | 43506897 | 43513871 |
| ENSBTAG00000008871 | *DDX4* | 20 | 23358630 | 23424574 |
| ENSBTAG00000008905 | *FIGLA* | 11 | 13600445 | 13615325 |
| ENSBTAG00000009017 | *NR5A1* | 11 | 95513842 | 95538735 |
| ENSBTAG00000009062 | *DMRTB1* | 3 | 92781366 | 92786271 |
| ENSBTAG00000009478 | *GDF9* | 7 | 44442457 | 44445396 |
| ENSBTAG00000009832 | *NOBOX* | 4 | 107667095 | 107672444 |
| ENSBTAG00000009917 | *PIP5K1B* | 8 | 44808121 | 45153937 |
| ENSBTAG00000009972 | *INHA* | 2 | 107501784 | 107504801 |
| ENSBTAG00000010414 | *DMRTA2* | 3 | 95865909 | 95870988 |
| ENSBTAG00000010885 | *TSPYL1* | 9 | 34430224 | 34433482 |
| ENSBTAG00000010889 | *FSHB* | 15 | 60976952 | 60980910 |
| ENSBTAG00000011173 | *FAM189A2* | 8 | 45469516 | 45532469 |
| ENSBTAG00000011234 | *FOXO3* | 9 | 41522588 | 41620269 |
| ENSBTAG00000012888 | *C2H2orf80* | 2 | 96437069 | 96460618 |
| ENSBTAG00000013499 | *LHX9* | 16 | 76858671 | 76878665 |
| ENSBTAG00000013790 | *MAP3K1* | 20 | 22340163 | 22417428 |
| ENSBTAG00000013801 | *PBX1* | 3 | 4147294 | 4246502 |
| ENSBTAG00000014064 | *FGFR2* | 26 | 41450745 | 41558092 |
| ENSBTAG00000014485 | *POF1B* | X | 66868308 | 66966819 |
| ENSBTAG00000014735 | *HOXA13* | 4 | 68842500 | 68844385 |
| ENSBTAG00000014890 | *CYP19A1* | 10 | 59102240 | 59156607 |
| ENSBTAG00000015132 | *RXFP2* | 12 | 29212314 | 29274747 |
| ENSBTAG00000015532 | *NR0B1* | X | 113283185 | 113289019 |

**Table 2.2.** Information of genes previously reported to be associated with DSDs (continued).

| Ensembl Gene ID | Gene name | Chromosome /scaffold name | Gene start (bp) | Gene end (bp) |
|---|---|---|---|---|
| ENSBTAG00000016465 | *DHCR7* | 29 | 48285096 | 48304702 |
| ENSBTAG00000015894 | *WWOX* | 18 | 5240349 | 6170333 |
| ENSBTAG00000016573 | *LHCGR* | 11 | 30977114 | 31040358 |
| ENSBTAG00000016867 | *HSD17B3* | 8 | 82961057 | 83014357 |
| ENSBTAG00000017082 | *POR* | 25 | 34146939 | 34213388 |
| ENSBTAG00000017800 | *DMRT3* | 8 | 43580997 | 43593395 |
| ENSBTAG00000018224 | *GLP1R* | 23 | 12914475 | 12946096 |
| ENSBTAG00000019145 | *MOS* | 14 | 23299177 | 23300199 |
| ENSBTAG00000019192 | *LHX8* | 3 | 69636164 | 69662240 |
| ENSBTAG00000019273 | *DMRTC2* | 18 | 51367681 | 51372973 |
| ENSBTAG00000019552 | *PGRMC1* | X | 3623055 | 3631983 |
| ENSBTAG00000020414 | *DHX37* | 17 | 50838143 | 50868907 |
| ENSBTAG00000020737 | *SOX8* | 25 | 788526 | 793121 |
| ENSBTAG00000020904 | *JAK3* | 7 | 5306153 | 5323991 |
| ENSBTAG00000021111 | *POU5F1* | 23 | 27982798 | 27987297 |
| ENSBTAG00000021713 | *MAMLD1* | X | 33093515 | 33213252 |
| ENSBTAG00000021827 | *TCF21* | 9 | 72066568 | 72069851 |
| ENSBTAG00000022255 | *AR* | X | 83085618 | 83294907 |
| ENSBTAG00000022564 | *AKR1C4* | 13 | 43616999 | 43626832 |
| ENSBTAG00000023832 | *ADAM8* | 26 | 25733217 | 25745433 |
| ENSBTAG00000024269 | *TGFBR3* | 3 | 51501752 | 51709606 |
| ENSBTAG00000031277 | *FOXL2* | 1 | 130185388 | 130186521 |
| ENSBTAG00000032424 | *FSHR* | 11 | 31255653 | 31450384 |
| ENSBTAG00000032887 | *SUPT3H* | 23 | 18239634 | 18641106 |
| ENSBTAG00000033345 | *STAR* | 27 | 33312972 | 33320397 |
| ENSBTAG00000033352 | *DMRT1* | 8 | 43600804 | 43697170 |
| ENSBTAG00000038306 | *CBX2* | 19 | 52751728 | 52760319 |
| ENSBTAG00000038434 | *ATRX* | X | 74597105 | 74880603 |
| ENSBTAG00000038735 | *LHB* | 18 | 55571842 | 55573733 |
| ENSBTAG00000038815 | *DIAPH2* | X | 45209466 | 45296719 |
| ENSBTAG00000038976 | *ZGLP1* | 7 | 14838239 | 14841025 |
| ENSBTAG00000040585 | *NR6A1* | 11 | 95553002 | 95774429 |
| ENSBTAG00000044105 | *FOXO1* | 12 | 21900230 | 21991693 |
| ENSBTAG00000045782 | *BMP15* | X | 88803967 | 88810596 |
| ENSBTAG00000045824 | *SOX9* | 19 | 58919579 | 58923174 |
| ENSBTAG00000047268 | *WT1* | 15 | 63118358 | 63168835 |
| ENSBTAG00000048237 | *FGF9* | 12 | 35378340 | 35406261 |
| ENSBTAG00000049481 | *SOX3* | NKLS02002208.1 | 4426876 | 4428231 |
| ENSBTAG00000049558 | *CYP17A1* | 26 | 23411942 | 23419174 |
| ENSBTAG00000050504 | *ARX* | X | 118322990 | 118333857 |
| ENSBTAG00000051083 | *WNT4* | 2 | 130696821 | 130711436 |

Ensembl BioMart (release 105) (Howe et al., 2021) was used to determine the genomic location of the bovine orthologs of these genes on the ARS-UCD1.2 assembly.

### 2.3.6. Identification of candidate causal variants by variant filtering

#### 2.3.6.1. Variant annotation

Functional impact predictions for SNPs and indels were assigned using SnpEff v5.0e (Cingolani, Platts, et al., 2012) and the ARS-UCD1.2.99 SnpEff annotation database. In addition, Ensembl VEP (release 105) (McLaren et al., 2016) was used for SVs annotation. However, since these methods could not predict the impact of variants on the Y chromosome (due to the lack of Y chromosome information in the reference databases used for variant annotation), the VCF files were manually checked for variants that overlapped with the *SRY* gene, using gene coordinates obtained from NCBI (release 105).

#### 2.3.6.2. Variant filtering for detecting causal variants for the *SRY*-positive XY female (Swyer 2)

The visualization of sequence read alignments identified *SRY* deletions in all affected animals except for Swyer 2. For this animal additional analyses were conducted. To identify candidate causal variants for the *SRY*-positive XY female (Swyer 2), variant filtering was performed using three filters: "recessive DSD," "dominant DSD," and "dominant de novo." The "recessive DSD" filter, implemented using SnpSift v5.0e (Cingolani, Patel, et al., 2012) , was used to identify high- or moderate-impact variants in DSD genes with a homozygous non-reference-allele genotype in the *SRY*-positive XY female (Swyer 2). Similarly, the "dominant DSD" filter was used to identify high- or moderate-impact variants with a heterozygous genotype in sample Swyer 2. Both the "recessive DSD" and "dominant DSD" filters were applied to SNPs, indels, and SVs. SNPs and indels passing these filters were then assessed in the 1000 Bull Genomes Project

Run 9 data set, consisting of whole-genome SNP and indel genotypes from 5116 animals, to rule out common variants. The "dominant de novo" filter, on the other hand, was only applied to SNPs and indels (because there isn't sufficient population-scale SV data available for determining de novo SVs). It identified high- or moderate-impact variants with a heterozygous genotype in the Swyer 2 sample that were not observed in the 1000 Bull Genomes Project Run 9 (Hayes & Daetwyler, 2019), and that were not present in the VCF file containing known variant sites used for base quality score recalibration. Only variants with a FILTER value of PASS were retained. This filter was applied using a combination of the BCFtools v1.8 isec command (Danecek et al., 2021), SnpSift v5.0e (Cingolani, Patel, et al., 2012) and standard command-line text-processing utilities.

### 2.3.7. Read depth plots

Read depth plots were used to visualize read coverage of whole chromosomes and large chromosomal regions. A Python script was used to count reads at all positions using the SAMtools v1.12 depth command (H. Li et al., 2009) to generate sliding-window read depth plots. The mean depth for each window was then calculated and plotted using matplotlib v3.5.1 (Hunter, 2007). For the full Y chromosome plots, a window size of 10,000 was used, while a window size of 1000 was used for the *SRY*-region plots.

## 2.4. Results

### 2.4.1. Sample's sex chromosome assessments

To check for the presence of X and Y chromosomes in the Swyer samples we used read depth-based metrics to compare control and affected individuals. All three metrics performed as expected in the control animals, clearly separating XX and XY animals. For all three metrics the Swyer animals closely resembled the XY control animals but not the XX control animals (Figure 2.1, Table 2.3). The X: Y coverage ratio was very high for control XX animals (mean = 141.05) and closer to 1 for control XY (mean = 0.93) and Swyer cases (mean = 1.51). The X: autosome coverage ratio was close to 1 for control XX animals (mean = 1.03) and close to 0.5 for control XY (mean = 0.52) and Swyer animals (mean = 0.55). Lastly, the Y: autosome coverage ratio was close to zero for control XX animals (mean = 0.01) and close to 0.5 for control XY (mean = 0.57) and Swyer samples (mean = 0.44).

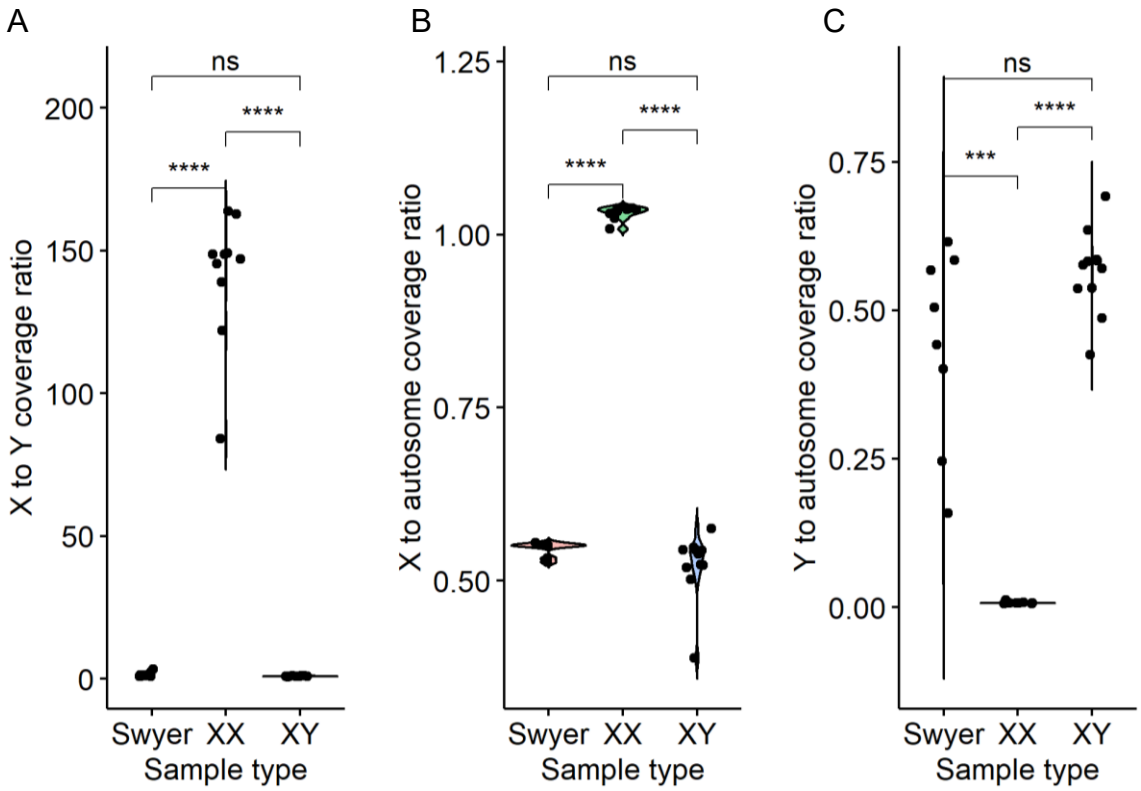


**Figure 2.1.** Violin plots of the sample's sex chromosome assessments using a sequenced dataset of eight Swyer animals and 21 control samples (10 females, labeled as XX; 11 males, labeled as XY). These plots exhibit the allocations in Swyer, XX, and XY samples of X to Y coverage ratio (the average is 1.5 in the Swyer group and 0.9 in the XY group; (A), X to autosome coverage ratio (B), and Y to autosome coverage ratio (C). Welch's t-test was utilized for comparing the means between the three sample groups and the results are displayed as * ($P < 0.05$), ** ($P < 0.01$), *** ($P < 0.001$), **** ($P < 0.0001$), or ns ($P > 0.05$).

**Table 2.3.** Read coverage ratios for genotypic sex assessments.

| Sample type | NCBI SRA ID | X: Y coverage ratio | X to autosome coverage ratio | Y to autosome coverage ratio |
|---|---|---|---|---|
| Control XX | SRS13440364 | 162.83 | 1.01 | 0.01 |
| Control XY | SRS11145576 | 0.94 | 0.54 | 0.57 |
| XY (Swyer 1) | SRS13440365 | 1.24 | 0.55 | 0.44 |
| XY (Swyer 2) | SRS13440366 | 0.91 | 0.53 | 0.58 |
| XY (Swyer 3) | SRS13440367 | 1.37 | 0.55 | 0.40 |
| XY (Swyer 4) | SRS13440368 | 2.24 | 0.55 | 0.25 |
| XY (Swyer 5) | SRS13440369 | 0.90 | 0.55 | 0.62 |
| XY (Swyer 6) | SRS13440370 | 3.32 | 0.53 | 0.16 |
| XY (Swyer 7) | SRS13440371 | 1.10 | 0.55 | 0.50 |
| XY (Swyer 8) | SRS13440372 | 0.97 | 0.55 | 0.57 |
| Control XX | SRS11145583 | 139.00 | 1.03 | 0.01 |
| Control XX | SRS11145582 | 149.13 | 1.03 | 0.01 |
| Control XX | SRS11145584 | 163.75 | 1.04 | 0.01 |
| Control XY | SRS11145580 | 0.94 | 0.54 | 0.58 |
| Control XX | SRS11145585 | 145.33 | 1.04 | 0.01 |
| Control XX | SRS11145588 | 121.92 | 1.04 | 0.01 |
| Control XX | SRS11145586 | 148.67 | 1.04 | 0.01 |
| Control XX | SRS11145587 | 147.10 | 1.03 | 0.01 |
| Control XX | SRS11145589 | 84.00 | 1.02 | 0.01 |
| Control XX | SRS11145581 | 148.75 | 1.04 | 0.01 |
| Control XY | SRS11145574 | 0.97 | 0.52 | 0.54 |
| Control XY | SRS11145571 | 0.73 | 0.50 | 0.69 |
| Control XY | SRS11145572 | 0.82 | 0.52 | 0.63 |
| Control XY | SRS11145575 | 0.93 | 0.54 | 0.59 |
| Control XY | SRS11145573 | 0.94 | 0.55 | 0.58 |
| Control XY | SRS11145577 | 0.93 | 0.54 | 0.58 |
| Control XY | SRS11145578 | 1.07 | 0.58 | 0.54 |
| Control XY | SRS11145579 | 0.91 | 0.39 | 0.43 |
| Control XY | SRS11145590 | 1.08 | 0.52 | 0.49 |

### 2.4.2. Inspection of the candidate gene alignments in IGV

In parallel to the text-based identification and filtering of variants, manual inspection of read alignments was performed using IGV. This analysis was confined to known DSD genes and was conducted for control and Swyer animals, to look for read coverage or alignment anomalies that could be indicative of differences between the samples. An obvious difference in read coverage was observed between the control XY individual and seven out of the eight XY female (Swyer) cattle, involving the *SRY* (sex-determining region Y) gene (Figure 2.2). Among the eight XY females with Swyer syndrome, seven of them exhibited a deletion on the Y chromosome that encompassed the *SRY* gene, resulting in an absence of the *SRY* gene (referred to as *SRY*-negative or *SRY*−).

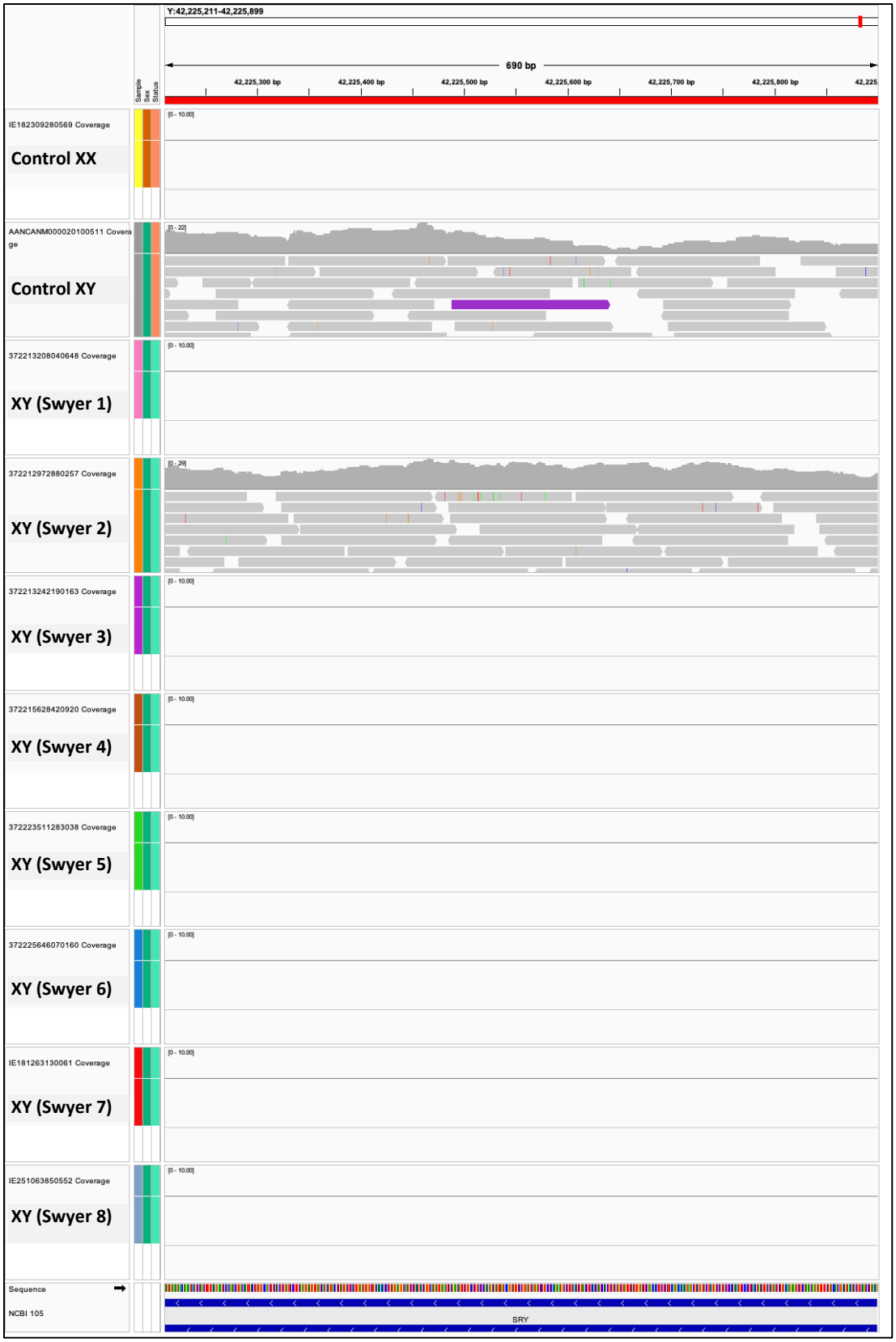


**Figure 2.2.** Integrative genomics viewer (IGV) visualization of the *SRY* gene on the Y chromosome.

Further inspections using IGV confirmed that the affected XY females did indeed have the Y chromosome, as there was high read coverage observed in another region of the Y chromosome (*ZFY* gene) (Figure 2.3), well above what is observed in the XX animal. Thus, as expected based on the previously calculated read coverage metrics, the affected animals possess a Y chromosome but one that is incomplete, missing the *SRY* gene.

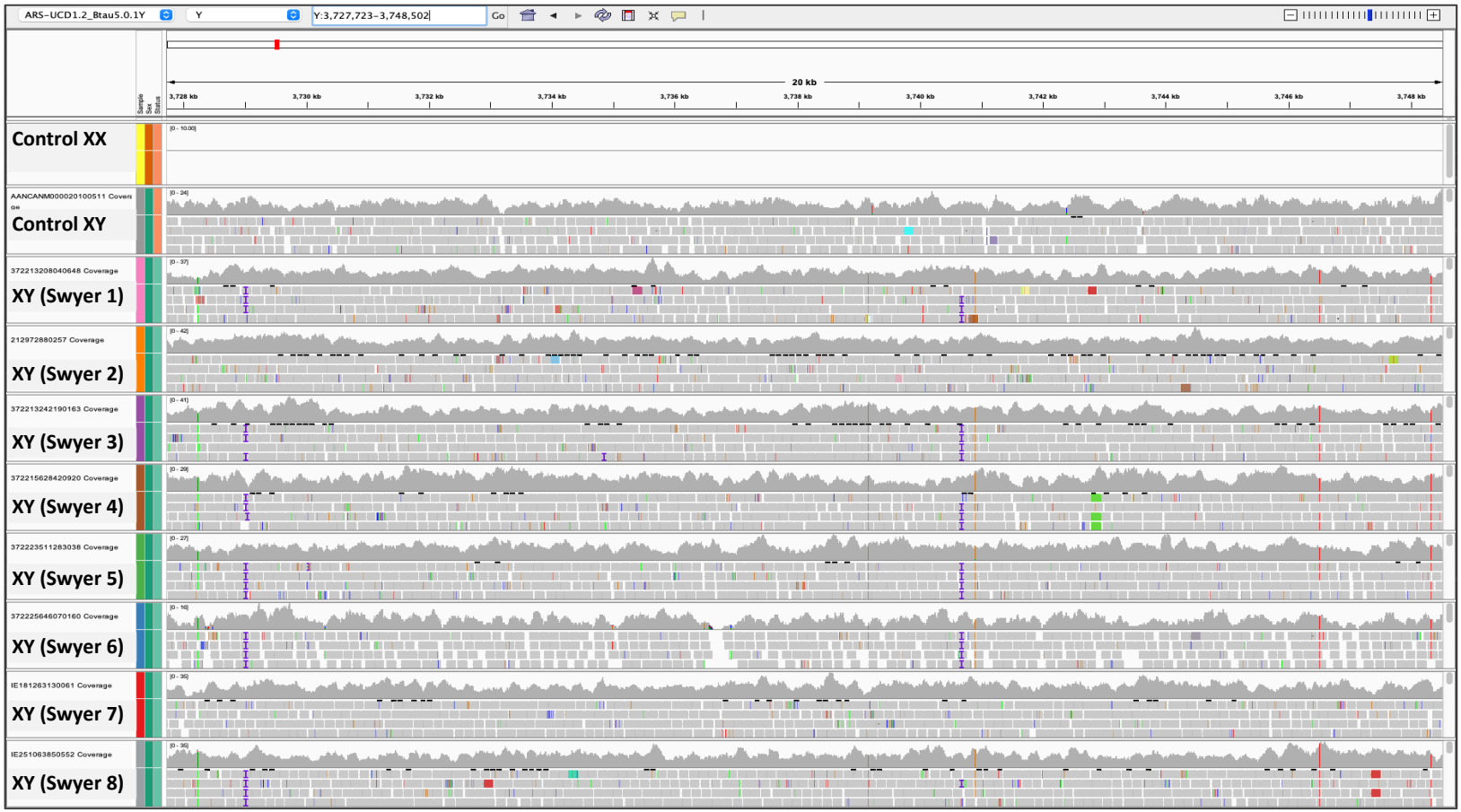


**Figure 2.3.** Integrative genomics viewer (IGV) visualization of *ZFY* gene locus on the Y chromosome in control (XX and XY) and Swyer Individuals.

### 2.4.3. Variant filtering for detecting causal variants for the *SRY*-positive affected sample (Swyer 2)

Since the visual inspection of read alignments for candidate gene regions revealed that *SRY* deletions were responsible for seven of the eight Swyer cases, SNP, indel, and SV filtering was used to identify a causative variant for the one remaining case (Swyer 2). First the initial set of 18,454,171 SNPs/indels were filtered to remove false-positive variants, giving 17,494,752 SNPs/indels. However, the initial set of 19,831 SVs were not subjected to this quality filter because there were no available filtering tools and reference databases for discerning false-positive SVs. Next, three separate (non-sequential) filters were applied, each tailored to a particular variant type: "recessive DSD", "dominant DSD", and "dominant de novo".

The "recessive DSD" filter detected no SVs and only 13 SNPs/indels. However, all these variants were detected at relatively common frequencies within the reference genotype dataset (with allele frequencies $\geq 0.21$). Consequently, they did not emerge as convincing candidate mutations. Similarly, the "dominant DSD" filter identified no SVs and 25 SNPs/indels. Yet, these variants also lacked compelling candidacy as causal mutations due to their prevalence in the reference dataset (allele frequency $\geq 0.003$). The "dominant de novo" filter, on the other hand, produced a list of 125 SNPs/indels. Nonetheless, due to the absence of established roles for the genes affected by these variants in sex determination and development, further refinement of this list to pinpoint a specific causative mutation remains challenging.

### 2.4.4. Read depth plots

Sliding-window plots were used to illustrate the unusual coverage patterns of the Y chromosome in the affected animals, which showed large segments of normal, intermediate, and

large missing coverage (as depicted in Figure 2.4). The absence of coverage in the *SRY* region is interpreted as an absence of the *SRY* gene in these samples, which would explain the DSD phenotype. The reason for the intermediate coverage in most of the Swyer samples for other regions of the Y chromosome is not known (note coverage drop after the 10 Mb position in most Swyer samples). In addition, inspecting the sliding-window coverage plot and reading map alignment in IGV suggests that one affected sample (Swyer 3) has a larger deletion compared to the deletion observed in the other affected samples (as shown in Figure 2.4 and 2.5).

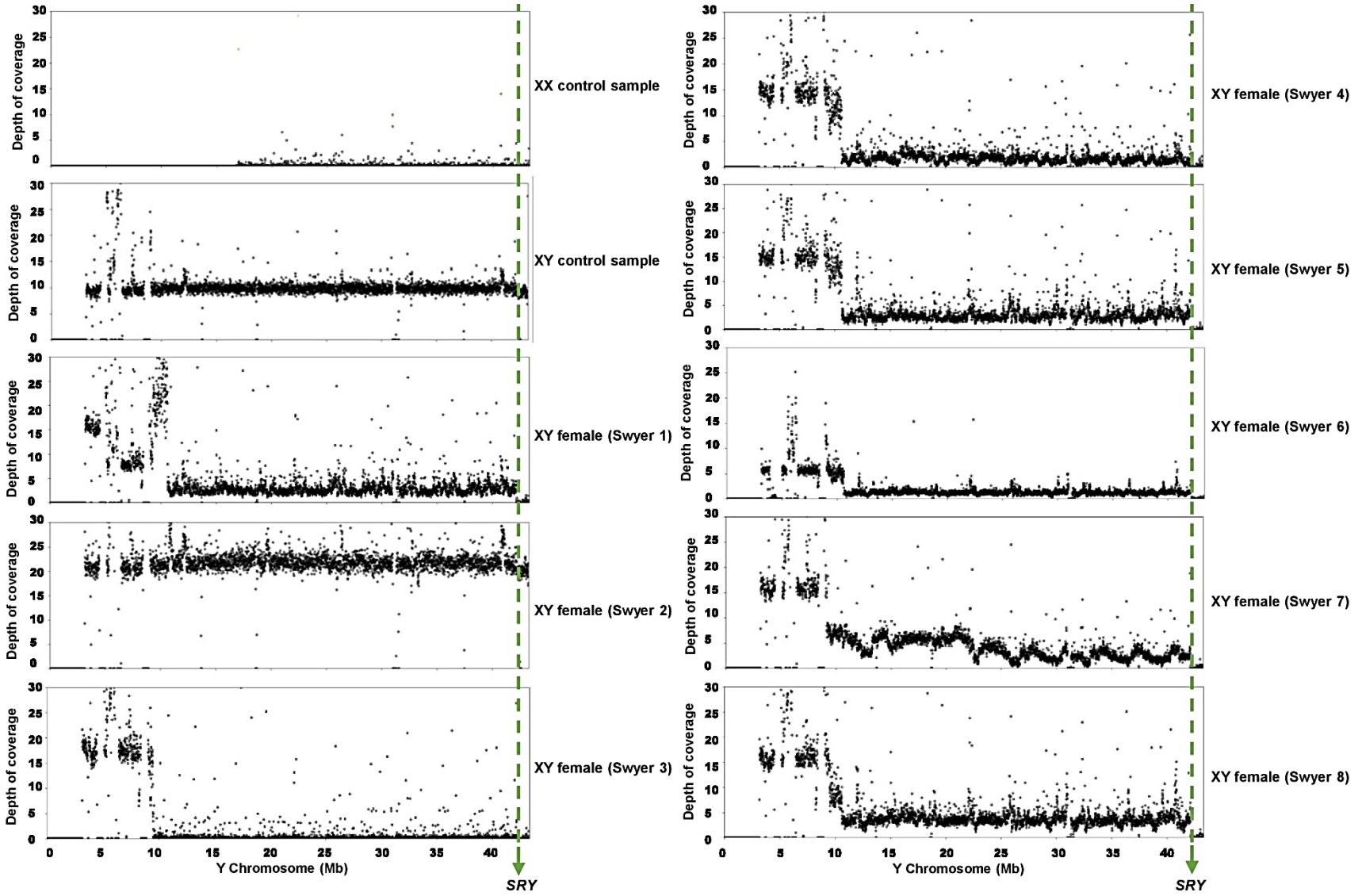


**Figure 2.4.** Sliding-window coverage plots of the Y chromosome of the studied samples. In all affected samples except Swyer 3, the coverage dropped from 42 Mb to the end of the Y chromosome. This region harbors the *SRY* gene (green dashed arrow lines).

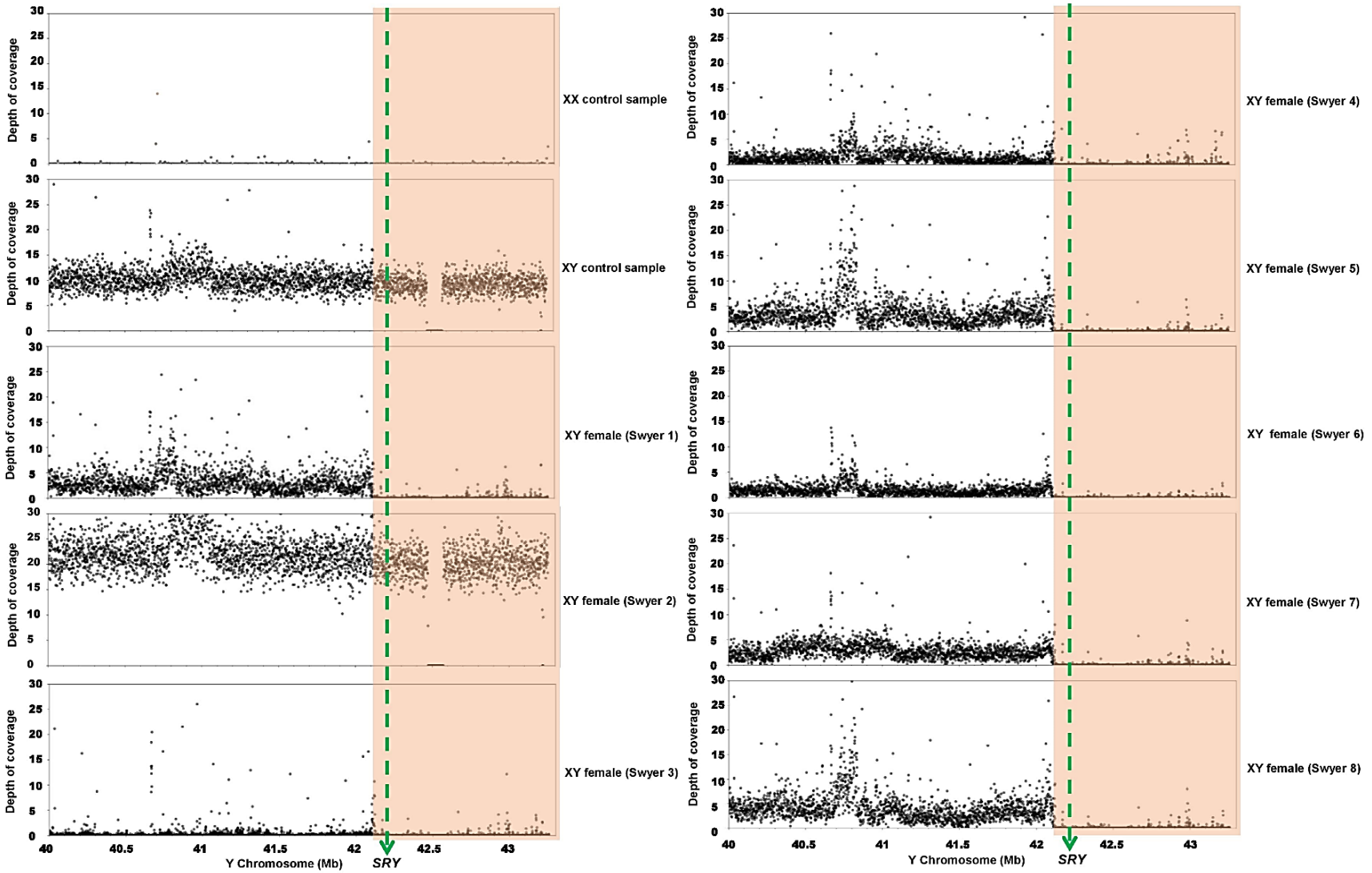


**Figure 2.5.** The sliding-window plot of the right end of chromosome Y in all the samples including those that lack *SRY* (*SRY* position on reference genome highlighted by green dashed arrow lines (42.2 Mb)). Drop in coverage (highlighted in orange) suggests deletion occurs at approximately 42.1 Mb to the end of the chromosome (43.3 Mb) in all the affected samples except Swyer 2 (lacks the *SRY* deletion) and Swyer 3 (the deletion appears to be much larger (from 10 Mb to 43.3 Mb)).

## 2.5. Discussion/Conclusions

This study aimed to investigate the potential factors that contribute to Swyer syndrome by examining a range of genetic variations such as SNPs, indels, and SVs using WGS data from affected and control animals. It leveraged existing knowledge of genes linked to sex development disorders to pinpoint the syndrome's underlying cause. In 7 out of 8 affected cases, an absence of the *SRY* gene was detected (seven *SRY*-negative affected individuals). The additional SNPs, indels, SVs filtering did not reveal an obvious cause for that one *SRY*-positive affected individual (Swyer 2). Nevertheless, several potential causative mutations were discovered; however, further investigation is required to conclusively confirm them as true causal variants.

In humans, XY individuals with apparent female physical characteristics (Swyer syndrome) are observed at a ratio of about 1 in 20,000 to 80,000 live-borns (Kane et al., 2023). According to previous reports and studies, approximately 15–20% of Swyer syndrome cases in humans are mostly caused by mutations in the sex-determining region of the Y chromosome (*SRY*) gene (Kane et al., 2023), or due to a significant deletion of the section of the Y chromosome containing the *SRY* gene (Behtash & Karimi Zarchi, 2007; Machado et al., 2014; Michala et al., 2008; Ono & Harley, 2013; Ostrer, 2019). The *SRY* gene is a master sex determining gene that produces a transcription factor belonging to the high mobility group (HMG) box family (Kanai et al., 2005). This protein functions as the testis-determining factor (TDF) and initiates the differentiation of bipotential undifferentiated gonadal tissue into testes in mammals (Kanai et al., 2005). Therefore, any mutations in this gene can have significant implications, leading to the development of abnormal sex glands and XY females, as seen in Swyer syndrome. In several mutational analysis studies in human XY female patients, it has been identified a list of 11 mutations in the DNA-binding high mobility group (HMG)-box of the *SRY* gene, including six

missense mutations, three nonsense mutations, and two deletions causing frameshifts (Berta et al., 1990; Hawkins, Taylor, Berta, et al., 1992; Hawkins, Taylor, Goodfellow, et al., 1992; Hawkins, 1993; Jäger et al., 1990, 1992; K. D. McElreavey et al., 1992; Müller et al., 1992).

In a study by Kawakura et al. (1996), which involved chromosome analysis using PCR, it was reported that in 3 out of 6 cases of bovine XY females banding patterns for the *SRY* gene in the agarose gel was not observed (*SRY*-negative XY females). Thus, for the other three remaining *SRY*-positive XY females, the true cause remained unknown (Kawakura et al., 1996). In their subsequent study (Kawakura et al., 1997), they investigated the structure of the Y chromosome using FISH and G-binding in the three *SRY*-negative cases of bovine XY females. They found that although two Y chromosome-specific probes (BC1.2 and btDYZ-1) hybridized to the short arm of the Y chromosome in a normal bull, they hybridized to both arms of the Y chromosome in the three *SRY*-negative XY females. Therefore, they concluded that the Y chromosome of these XY affected cattle that lack *SRY* had an abnormal structure. Additionally, their findings support the location of *SRY* on the short arm of the bovine Y chromosome (Yp). (Kawakura et al., 1997). Using next-generation sequencing (NGS) in our investigation of eight XY affected females detected through SNP chip/phenotype discordance in Ireland delivers an additional understanding of the occurrence and characteristics of *SRY* deletions in cattle. Additionally, we detected an *SRY*-positive XY female (Swyer 2). Therefore, all but one of the cases in this study can be explained by an *SRY* deletion. According to our study findings, Swyer syndrome in cattle is estimated to occur at a rate of 1 in every 100,000 cases. Our results also indicate that *SRY* deletions are a predominant cause of Swyer syndrome in cattle.

It is believed that defects in genes other than the *SRY* gene can cause Swyer syndrome (Khare et al., 2017). These additional genes are thought to be involved in promoting testicular

development and the differentiation of XY fetuses into males (Ostrer, 2019). One such gene implicated in Swyer syndrome is *Map3K1*, as mutations in this gene are frequently identified as a contributing factor to the condition (Ostrer, 2019; Pearlman et al., 2010). In some instances of Swyer syndrome, affected individuals have been found to carry mutations in the *NROB1* gene on the X chromosome (Du et al., 2014; Kane et al., 2023). Furthermore, investigations have linked some instances of Swyer syndrome to mutations in the desert hedgehog (*DHH*) gene on chromosome 12 (Canto et al., 2004). Mutations in the *DEAH37* gene have also been identified as a common cause of the condition (K. McElreavey et al., 2020). In a few rare cases, mutations in genes such as steroidogenic factor 1 (*SF1* or *NR5A1*) (Köhler et al., 2008), *Wnt-4* (*WNT4*) (Domenice et al., 2004), *CBX2*, *GATA4,* and *WWOX* (Khare et al., 2017) have been associated with Swyer syndrome (Ostrer, 2019). However, it is noted that additional genes not yet identified may be involved in the development of Swyer syndrome (Ostrer, 2019). In some cases, it has been reported that the condition is not hereditary and may be caused due to a new mutation (de novo variant) or abnormality (Berta et al., 1990). Nevertheless, there were reports of instances where individuals affected by Swyer syndrome, caused by a mutation in the *SRY* gene, have fathers (and in some cases, brothers) who also possess the *SRY* mutation on the Y chromosome (Hawkins, Taylor, Goodfellow, et al., 1992). The reasons why these fathers and/or brothers did not experience the development of Swyer syndrome remain unknown. It has been hypothesized that additional genes and/or factors, in conjunction with the *SRY* gene mutation, may be required for the manifestation of Swyer syndrome in these affected individuals.

In the present study, the *SRY* deletion was not detected using the standard (non-visual) programs and pipelines that were used for SV calling and analysis. This is likely attributed to the large size of the deletion, which could present challenges for detection using the specific

methodology employed in the study. However, as previously mentioned, it was through further assessments of read alignments in IGV (Thorvaldsdóttir et al., 2013) and sliding-window read depth plots that we could identify a lack of *SRY* coverage in 7 out of 8 affected cattle. Remarkably, six out of the seven *SRY*-negative XY affected animals appeared to have a similar or identical deletion involving approximately 1.2 Mb at the distal region of the short arm (Yp) of the Y chromosome (Fechner, 1996). It worth mentioning that one individual (Swyer 3) appeared to have a much larger deletion, approximately 42.3 Mb. The observed similarity or identity in deletions among the six animals could be attributed to specific characteristics of the bovine Y chromosome that make it susceptible to this type of deletion. While the presence of the same mutation in different individuals could reflect inheritance of a common mutation (for example, a bull with a mosaic germline could conceivably sire multiple XY females with the same Y chromosome abnormality), this is not the case in this data set as each animal was produced by a different sire.

It is noteworthy that on the sliding-window read depth plots; we observed that the XY control sample and one of the XY affected samples (Swyer 2) exhibited consistent coverage across the Y chromosome, while the other seven XY affected Swyer individuals showed decreased coverage starting around 10 Mb. Additionally, the read coverage near the end of the Y chromosome drops to zero before the *SRY* gene, except for the XY affected sample (Swyer 3), which has a significantly larger deletion starting near 10 Mb. One potential explanation of this situation is that the deletion size varies between cells within each sample, resulting in a mixture of Y chromosome templates being analyzed through sequencing. This mixture may arise because an initial deletion makes cells more likely to acquire additional deletions (Guo et al., 2020). By conducting further investigations that involve karyotyping multiple cells from the same sample or individual, it should be possible to gain additional insights and achieve a deeper understanding of

the underlying mechanisms related to this matter. However, it is worth mentioning that this study is the first to reveal an unusual coverage pattern of bovine Y chromosome, potentially indicating chromosome instability.

The presence of *SRY* deletions in 7 out of 8 affected samples in this comprehensive study of XY females suggests that incorporating *SRY* SNP markers on SNP chips could aid in identifying and characterizing XY females. In cases where these markers are recorded as missing, it could indicate the presence of an *SRY* deletion. As an alternative, the findings of this study suggest that PCR-based methods, including conventional or real-time PCR, could be used to check for the presence of the *SRY* gene and provide valuable information. Despite the low prospect of economic impact due to the rarity and infertility of affected cases, discrepancies between genotype-based sex and phenotypic sex can lead to mistrust among farmers towards genotyping and DNA-based breeding techniques. However, the ability to provide explanations for these incidents could significantly reduce such concerns.

For the one *SRY*-positive XY female, no true causal mutation was identified based on inspections of known DSD genes for dominant or recessive mutations involving SNPs, indels, or SVs. However, further investigation through "dominant de novo" filter resulted in the discovery of a list of 125 de novo variants including 23 high-impact and 102 moderate-impact SNPs/indels candidate mutations. Nonetheless, because the genes impacted by these variants lack established functions in sex determination and development, narrowing down this list to identify a particular causative mutation is still a challenging task.

It is essential to note that the identification and characterization of variants are influenced and be restricted by factors such as the length of reads, sequencing depth, available annotation information, reference sequences, and software tools. Additionally, it is crucial to consider that the

tissue sample sequenced from this individual may not contain the causative mutation. Chimerism, a condition where a mutation occurs in a portion of cells leading to a change in sex, can also impact the results, as observed in some rare cases (Berkovitz & Seeherunvong, 1998; Reindollar et al., 1987). It is worth mentioning that this study represents the first research using WGS for the identification of *SRY* deletions in XY females within the cattle population. Despite the challenges in identifying the causative mutation in this individual, the gathered information, including candidate variants, descriptive phenotypes, and sequence data, could be valuable for future investigations in identifying the mutations associated with this condition in cattle and other species.

For more exploration of the result of our study, researchers in Ireland conducted cytogenetic and necropsy observations of the *SRY*-positive individual (Swyer 2). Karyotype analyses were performed on the *SRY*-positive individual's blood, mammary, and reproductive tract, consistently showing a karyotype of 60, XY. This indicated no mosaicism with 60, XX and confirmed the presence of the *SRY* gene in the individual's genome based on the karyotype analysis.

It is acknowledged that in regards to Swyer syndrome cases in human, even with advancements in genomic technology, almost 50% of Swyer syndrome cases often have no known cause, suggesting that unidentified mutations may play a role (Baetens et al., 2018). The continued large-scale phenotyping and genotyping of cattle could identify further Swyer syndrome as well as other DSD cases, the in-depth characterization of which could lead to the discovery of novel causative mechanisms and new knowledge of the processes that regulate sexual development.

## 2.6. References


Andrews, S. (2010, April 26). *Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arboleda, V., & Vilain, E. (2016). Chapter 19—Disorders of Sex Development. In R. E. Weiss & S. Refetoff (Eds.), *Genetic Diagnosis of Endocrine Disorders (Second Edition)* (pp. 259–278). Academic Press. https://doi.org/10.1016/B978-0-12-800892-8.00019-1

Baetens, D., Güran, T., Mendonca, B. B., Gomes, N. L., De Cauwer, L., Peelman, F., Verdin, H., Vuylsteke, M., Van der Linden, M., ESR2 STUDY GROUP, Atay, Z., Bereket, A., de Krijger, R. R., Preter, K. de, Domenice, S., Turan, S., Stoop, H., Looijenga, L. H., De Bosscher, K., … De Baere, E. (2018). Biallelic and monoallelic ESR2 variants associated with 46,XY disorders of sex development. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *20*(7), 717–727. https://doi.org/10.1038/gim.2017.163

Banoth, M., Naru, R. R., Inamdar, M. B., & Chowhan, A. K. (2018). Familial Swyer syndrome: A rare genetic entity. *Gynecological Endocrinology*, *34*(5), 389–393. https://doi.org/10.1080/09513590.2017.1393662

Behtash, N., & Karimi Zarchi, M. (2007). Dysgerminoma in three patients with Swyer syndrome. *World Journal of Surgical Oncology*, *5*(1), 71. https://doi.org/10.1186/1477-7819-5-71

Berkovitz, G. D., & Seeherunvong, T. (1998). Abnormalities of gonadal differentiation. *Bailliere's Clinical Endocrinology and Metabolism*, *12*(1), 133–142. https://doi.org/10.1016/s0950-351x(98)80512-0

Berta, P., Hawkins, J. R., Sinclair, A. H., Taylor, A., Griffiths, B. L., Goodfellow, P. N., & Fellous, M. (1990). Genetic evidence equating *SRY* and the testis-determining factor. *Nature*, *348*(6300), 448–450. https://doi.org/10.1038/348448A0

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Canto, P., Söderlund, D., Reyes, E., & Méndez, J. P. (2004). Mutations in the desert hedgehog (DHH) gene in patients with 46,XY complete pure gonadal dysgenesis. *The Journal of Clinical Endocrinology and Metabolism*, *89*(9), 4480–4483. https://doi.org/10.1210/jc.2004-0863

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Frontiers in Genetics*, *3*, 35. https://doi.org/10.3389/fgene.2012.00035

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

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

Délot, E. C., & Vilain, E. (2019). Disorders of sex development. In *Yen and Jaffe's Reproductive Endocrinology* (pp. 365–393). Elsevier.

Domenice, S., Correa, R. V., Costa, E. M. F., Nishi, M. Y., Vilain, E., Arnhold, I. J. P., & Mendonca, B. B. (2004). Mutations in the *SRY*, DAX1, SF1 and WNT4 genes in Brazilian sex-reversed patients. *Brazilian Journal of Medical and Biological Research*, *37*, 145–150.

Du, X., Zhang, X., Li, Y., & Han, Y. (2014). 46, XY female sex reversal syndrome with bilateral gonadoblastoma and dysgerminoma. *Experimental and Therapeutic Medicine*, *8*(4), 1102–1104.

Eggers, S., Smith, K. R., Bahlo, M., Looijenga, L. H., Drop, S. L., Juniarto, Z. A., Harley, V. R., Koopman, P., Faradz, S. M., & Sinclair, A. H. (2015). Whole exome sequencing combined with linkage analysis identifies a novel 3 bp deletion in NR5A1. *European Journal of Human Genetics*, *23*(4), 486–493. https://doi.org/10.1038/ejhg.2014.130

Fechner, P. Y. (1996). The role of *SRY* in mammalian sex determination. *Acta Paediatrica Japonica: Overseas Edition*, *38*(4), 380–389. https://doi.org/10.1111/j.1442-200x.1996.tb03512.x

Guo, X., Dai, X., Zhou, T., Wang, H., Ni, J., Xue, J., & Wang, X. (2020). Mosaic loss of human Y chromosome: What, how and why. *Human Genetics*, *139*(4), 421–446. https://doi.org/10.1007/s00439-020-02114-w

Hare, J. E., Baird, J. D., Duignan, P., Saunders, J., Floetenmeyer, R., & Basrur, P. K. (1994). XY gonadal dysgenesis and tetralogy of Fallot in an Angus calf. *The Canadian Veterinary Journal*, *35*(8), 510–512.

Hawkins, J. R. (1993). Mutational analysis of *SRY* in XY females. *Human Mutation*, *2*(5), 347–350. https://doi.org/10.1002/humu.1380020504

Hawkins, J. R., Taylor, A., Berta, P., Levilliers, J., Van der Auwera, B., & Goodfellow, P. N. (1992). Mutational analysis of *SRY*: Nonsense and missense mutations in XY sex reversal. *Human Genetics*, *88*(4), 471–474. https://doi.org/10.1007/BF00215684

Hawkins, J. R., Taylor, A., Goodfellow, P. N., Migeon, C. J., Smith, K. D., & Berkovitz, G. D. (1992). Evidence for increased prevalence of *SRY* mutations in XY females with complete rather than partial gonadal dysgenesis. *American Journal of Human Genetics*, *51*(5), 979–984.

Hayes, B. J., & Daetwyler, H. D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual Review of Animal Biosciences*, *7*, 89–102. https://doi.org/10.1146/annurev-animal-020518-115024

He, B., Zhu, R., Yang, H., Lu, Q., Wang, W., Song, L., Sun, X., Zhang, G., Li, S., Yang, J., Tian, G., Bing, P., & Lang, J. (2020). Assessing the Impact of Data Preprocessing on Analyzing Next Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, *8*. https://www.frontiersin.org/articles/10.3389/fbioe.2020.00817

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Iannuzzi, L., Di Meo, G. P., Perucatti, A., Di Palo, R., & Zicarelli, L. (2001). 50, XY gonadal dysgenesis (Swyer's syndrome) in a female river buffalo (Bubalus bubalis). *Veterinary Record*, *148*(20), 634–635.

Irish Cattle Breeding Federation (ICBF). (2022). *Beef Calving Stats, ICBF Web Application*. https://webapp.icbf.com/v2/app/beef-stats

Jäger, R. J., Anvret, M., Hall, K., & Scherer, G. (1990). A human XY female with a frame shift mutation in the candidate testis-determining gene *SRY*. *Nature*, *348*(6300), 452–454. https://doi.org/10.1038/348452a0

Jäger, R. J., Harley, V. R., Pfeiffer, R. A., Goodfellow, P. N., & Scherer, G. (1992). A familial mutation in the testis-determining gene *SRY* shared by both sexes. *Human Genetics*, *90*(4), 350–355. https://doi.org/10.1007/BF00220457

Kanai, Y., Hiramatsu, R., Matoba, S., & Kidokoro, T. (2005). From *SRY* to *SOX9*: Mammalian testis differentiation. *Journal of Biochemistry*, *138*(1), 13–19. https://doi.org/10.1093/jb/mvi098

Kane, D. T., Dempsey, M. A., Burke, A. L., & Morrison, J. J. (2023). Prenatal Diagnosis and Fetal Sonographic Features of Swyer Syndrome. *Journal of Fetal Medicine*, 36–39. https://doi.org/10.1055/s-0043-57035

Kawakura, K., Miyake, Y.-I., Murakami, R. K., Kondoh, S., Hirata, T.-I., & Kaneda, Y. (1997). Abnormal structure of the Y chromosome detected in bovine gonadal hypoplasia (XY female) by FISH. *Cytogenetic and Genome Research*, *76*(1–2), 36–38. https://doi.org/10.1159/000134508

Kawakura, K., Miyake, Y.-I., Murakami, R.-K., Kondoh, S., Hirata, T. -l., & Kaneda, Y. (1996). Deletion of the *SRY* region on the Y chromosome detected in bovine gonadal hypoplasia

(XY female) by PCR. *Cytogenetic and Genome Research*, *72*(2–3), 183–184. https://doi.org/10.1159/000134182

Khare, J., Deb, P., Srivastava, P., & Reddy, B. H. (2017). Swyer syndrome: The gender swayer? *Alexandria Journal of Medicine*, *53*(2), 197–200. https://doi.org/10.1016/j.ajme.2016.05.006

Köhler, B., Lin, L., Ferraz-de-Souza, B., Wieacker, P., Heidemann, P., Schröder, V., Biebermann, H., Schnabel, D., Grüters, A., & Achermann, J. C. (2008). Five novel mutations in steroidogenic factor 1 (SF1, NR5A1) in 46,XY patients with severe underandrogenization but without adrenal insufficiency. *Human Mutation*, *29*(1), 59–64. https://doi.org/10.1002/humu.20588

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Machado, C., Pereira, A., Cruz, J. M., Cadilhe, A., Silva, A., & Pereira, A. (2014). A novel SRY nonsense mutation in a case of Swyer syndrome. *Journal of Pediatric and Neonatal Individualized Medicine (JPNIM)*, *3*(1), e030107–e030107.

McElreavey, K. D., Vilain, E., Boucekkine, C., Vidaud, M., Jaubert, F., Richaud, F., & Fellous, M. (1992). XY sex reversal associated with a nonsense mutation in SRY. *Genomics*, *13*(3), 838–840. https://doi.org/10.1016/0888-7543(92)90164-n

McElreavey, K., Jorgensen, A., Eozenou, C., Merel, T., Bignon-Topalovic, J., Tan, D. S., Houzelstein, D., Buonocore, F., Warr, N., Kay, R. G. G., Peycelon, M., Siffroi, J.-P., Mazen, I., Achermann, J. C., Shcherbak, Y., Leger, J., Sallai, A., Carel, J.-C., Martinerie, L., … Bashamboo, A. (2020). Pathogenic variants in the DEAH-box RNA helicase DHX37 are a frequent cause of 46,XY gonadal dysgenesis and 46,XY testicular regression syndrome. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *22*(1), 150–159. https://doi.org/10.1038/s41436-019-0606-y

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Michala, L., Goswami, D., Creighton, S. M., & Conway, G. S. (2008). Swyer syndrome: Presentation and outcomes. *BJOG: An International Journal of Obstetrics & Gynaecology*, *115*(6), 737–741.

Miller, W. L., & Auchus, R. J. (2011). The Molecular Biology, Biochemistry, and Physiology of Human Steroidogenesis and Its Disorders. *Endocrine Reviews*, *32*(1), 81–151. https://doi.org/10.1210/er.2010-0013

Müller, J., Schwartz, M., & Skakkebaek, N. E. (1992). Analysis of the sex-determining region of the Y chromosome (SRY) in sex reversed patients: Point-mutation in SRY causing sex-

reversion in a 46,XY female. *The Journal of Clinical Endocrinology and Metabolism*, *75*(1), 331–333. https://doi.org/10.1210/jcem.75.1.1619028

Nowacka-Woszuk, J., & Switonski, M. (2009). Differentiated Evolutionary Conservatism and Lack of Polymorphism of Crucial Sex Determination Genes (SRY and SOX9) in Four Species of the Family Canidae. *Folia Biologica*, *57*(3–4), 171–176. https://doi.org/10.3409/fb57_3-4.171-176

Ono, M., & Harley, V. R. (2013). Disorders of sex development: New genes, new concepts. *Nature Reviews Endocrinology*, *9*(2), 79–91.

Ostrer, H. (2019). Swyer syndrome. *NORD (National Organization for Rare Disorders)*. https://rarediseases.org/rare-diseases/swyer-syndrome/

Pearlman, A., Loke, J., Le Caignec, C., White, S., Chin, L., Friedman, A., Warr, N., Willan, J., Brauer, D., Farmer, C., Brooks, E., Oddoux, C., Riley, B., Shajahan, S., Camerino, G., Homfray, T., Crosby, A. H., Couper, J., David, A., … Ostrer, H. (2010). Mutations in MAP3K1 cause 46,XY disorders of sex development and implicate a common signal transduction pathway in human testis determination. *American Journal of Human Genetics*, *87*(6), 898–904. https://doi.org/10.1016/j.ajhg.2010.11.003

Pedersen, B. S., Layer, R., & Quinlan, A. R. (2020). *smoove: Structural-variant calling and genotyping with existing tools* (0.2.8) [Go]. https://github.com/brentp/smoove/blob/c68ae05317dc5cedcc848b0389fd1a6b9d0fb92e/CITATION.cff (Original work published 2018)

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics (Oxford, England)*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699

Pedersen, B. S., & Quinlan, A. R. (2019). Duphold: Scalable, depth-based annotation and curation of high-confidence structural variant calls. *GigaScience*, *8*(4), giz040. https://doi.org/10.1093/gigascience/giz040

Reindollar, R. H., Byrd, J. R., Hahn, D. H., Haseltine, F. P., & McDonough, P. G. (1987). A cytogenetic and endocrinologic study of a set of monozygotic isokaryotic 45,X/46,XY twins discordant for phenotypic sex: Mosaicism versus chimerism. *Fertility and Sterility*, *47*(4), 626–633. https://doi.org/10.1016/s0015-0282(16)59113-4

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., … Sherry, S. T. (2021). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *49*(D1), D10–D17. https://doi.org/10.1093/nar/gkaa892

*Swyer syndrome: MedlinePlus Genetics*. (2022, February 25). https://medlineplus.gov/genetics/condition/swyer-syndrome/

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

# CHAPTER 3. The Search for Candidate Genomic Mutations Associated with Copper Toxicosis in Dalmatian Dogs.

## 3.1. Introduction and Literature Review

### 3.1.1. Copper toxicosis (CT)

Copper toxicosis (CT) is a complex genetic condition in humans, dogs, and several other species, diagnosed by the excess of the trace element copper in the liver and bloodstream (X. Wu, Chien, et al., 2019; X. Wu, den Boer, et al., 2020). If not promptly treated, this metabolic disorder can be life-threatening and may lead to liver failure and the eventual death of the affected individual (Scheinberg & Sternlieb, 1965). Copper is an essential element critical to various cellular functions, acting as an electron donor or acceptor in diverse reactions. However, an excess of copper ions in cells can result in detrimental effects, producing free radicals and increasing oxidative stress. Copper metabolism in multicellular organisms involves several processes, including absorption, dispersion, retention, and elimination at both the cellular and systemic levels. In mammals, bioavailable dietary copper ions are taken up by enterocytes through a process that relies on the copper transport protein Ctr1 (J. Chen et al., 2020).

Copper is transported from the intestinal tract to the liver for the purpose of redistribution to copper-dependent proteins (Linder et al., 1998). One of these proteins is ceruloplasmin, which requires at least six copper molecules to perform its main role of transporting iron throughout the body (Harris et al., 1999). After copper ions are taken up by cells, they are transported to ATP7A (ATPase copper-transporting α) (Chelly et al., 1993; Mercer et al., 1993; Vulpe et al., 1993), which moves Cu+ from enterocytes into the bloodstream. From there, copper ions are carried to the liver via the portal vein, where they are taken up by hepatocytes through Ctr1 (copper transporter 1). Cu+ can then be secreted into the blood or bile via the Atox1/ATP7B (ATPase copper-transporting β)/ceruloplasmin pathway. In the blood, copper can travel to peripheral tissues and be taken up

again by Ctr1. In peripheral tissue cells, copper ions are either stored by molecules such as metallothionein or directed to utilization pathways by chaperones like Atox1 (Antioxidant 1 Copper Chaperone), Cox17 (Cytochrome c oxidase copper chaperone), and CCS (copper chaperone for Sod1). Excess copper within hepatocytes is sequestered into vesicles for extrusion at the plasma membrane (Figure 3.1.). Maintaining proper regulation of copper metabolism is crucial for homeostasis and disease prevention. Disruptions in copper balance, whether inherited or acquired, such as deficiency, excess, or misplacement, can lead to the development or exacerbation of various diseases (J. Chen et al., 2020).

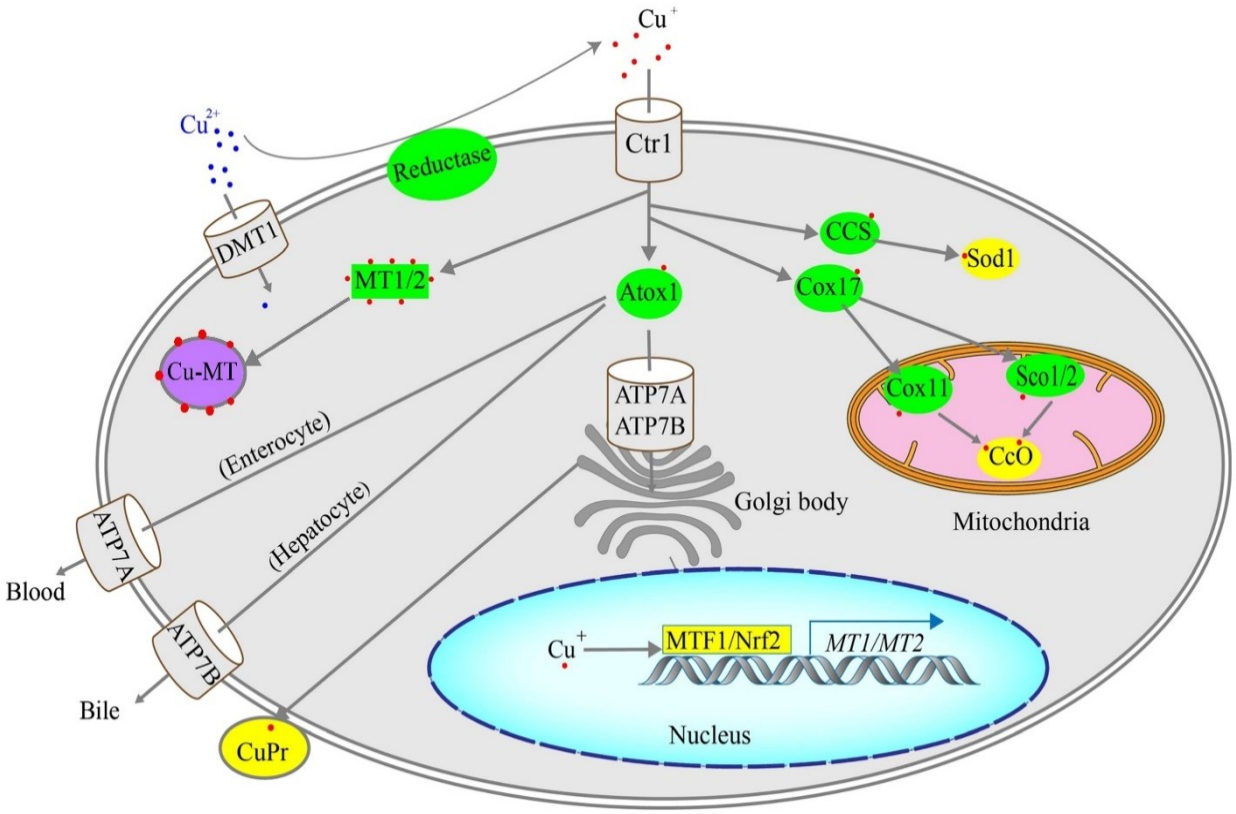


**Figure 3.1.** Copper metabolism in mammals. Cuprous ions (Cu+) undergo various utilization and detoxification pathways within the cell, including the cytosolic, mitochondrial, and Golgi routes. In the cytosol, CCS plays a role in loading Cu+ ions and activating superoxide dismutase 1 (Sod1). The chaperone Cox17 guides Cu+ ions to CcO through Sco1 or Sco2. Atox1 transfers copper to P1B-type ATPases, such as ATP7A and ATP7B, located on the trans-Golgi network (TGN). Cuproproteins (CuPrs) assembled in the Golgi compartment can either be secreted from the cell or sorted into specific organelles. Excessive copper ions in the cytosol can be temporarily bound by metallothioneins (MT1/MT2) and stored as a complex known as Cu-MT. In enterocytes or hepatocytes, ATP7A and ATP7B translocate to the plasma membrane, facilitating the efflux of Cu+ ions into the blood or bile, respectively. Elevated copper levels induce the expression of MT1 and MT2, which are regulated by transcription factors MTF1 and Nrf2, aiding in removing excess Cu+. (This figure was altered with permission from the "Springer Nature" (Chen et al., 2020)).

### 3.1.2. Copper toxicosis in humans (Wilson's disease (WD))

Wilson's disease (WD), a rare autosomal recessive disorder in humans, is the well-studied disease of copper overload that occurs as a result of mutations in the *ATP7B* gene that codes the copper transporter ATP7B protein (Bull et al., 1993; Corbee & Penning, 2021; Fieten et al., 2012; Scheinberg & Sternlieb, 1965; Tanzi et al., 1993; Weiskirchen & Penning, 2021). The ATP7B protein is vital in the excretion of the hepatic (liver) accumulated copper. Despite *ATP7B* gene mutations being recognized as the primary cause of Wilson's disease, there have been reported cases where individuals are diagnosed with the condition despite lacking any mutations in this gene (Kenney & Cox, 2007; Mak et al., 2008; Nicastro et al., 2010; Park et al., 2010). Although over 500 known mutations in the *ATP7B* gene associated with the disease has been recorded so far (https://grenada.lumc.nl/LSDB_list/lsdbs/ATP7B, access date: 17 March 2023), there are suggestions that other genetic variants (genes) may contribute to Wilson's disease onset or symptoms (Coffey et al., 2013).

There are two primary manifestations of Wilson's disease in humans, distinguished by their initial symptoms: one primarily related to the liver, and the other neurological in nature. Initially, it was expected that specific *ATP7B* gene variants would align with each of these clinical presentations, establishing a clear genotype-to-phenotype correlation. However, this anticipated correlation does not seem to hold true. Therefore, researchers have explored other genes or modifier genes that may regulate cell copper levels due to the phenotypic variation observed within the same genotype (to distinguish neurological from hepatic onset) or the lack of a clear correlation between genotype and phenotype in Wilson's disease (WD) (Weiskirchen & Penning, 2021). In dogs, only the hepatic form of Wilson's disease has been identified.

*Modifier genes* are defined as genes that can affect the expression of another gene or the phenotype it presents. Several modifier genes have been proposed as potential candidates for influencing the varied phenotypic expression of WD (Genin et al., 2008; Medici & LaSalle, 2019). Grubenbecher et al. (2006) discovered that a common variation in the *PRNP* gene could affect the clinical progression of neurological Wilson's disease. This genetic variation delays the age at which symptoms manifest and influences the specific symptoms that develop. Moreover, the *PRNP* gene is responsible for the synthesis of prion protein, which is present in various tissues, including the brain, and has been proposed to be involved in copper transport. The interaction among genetic variations in various genes, including *ATP7B* and *PRNP*, can influence the clinical presentation and advancement of Wilson's disease, underscoring the intricate nature of this condition and the significance of genetic factors in its development (Grubenbecher et al., 2006).

Several human diseases have canine equivalents, such as copper toxicosis. The identification of genes and mutations linked to diseases in dogs can hold significant implications for humans as well. Such discoveries can provide valuable insights into the genetic pathways and mechanisms involved in the development and progression of diseases in both dogs and humans. Additionally, the endeavor to tackle canine diseases is driven by a combination of scientific curiosity and the desire to improve the well-being of beloved human companions. In the event that the present research identifies a genetic variant responsible for the disorder in Dalmatians, it could bring the potential to create a genetic test to diminish the number of dogs affected by the condition. This test would empower breeders to identify carriers of the CT variant and adapt their breeding plans accordingly. An illustrative success story in this regard is the reduction in the prevalence of CT among Bedlington terriers through the application of the *COMMD1* gene deletion test.

### 3.1.3. Copper toxicosis in dogs

Several studies have shown similarities and differences in the characteristics of Wilson disease and copper toxicity in certain dog breeds. As in humans, mutations in *ATP7B* have been shown to be linked to the disease. However, in dogs, additional mutations have been associated with the disorder. For example, in Bedlington terriers, copper toxicity is caused by an autosomal recessive defect (a deletion) in exon 2 of the *COMMD1* (Copper Metabolism Domain Containing 1) gene (Coronado et al., 2003; Klomp et al., 2003), while in Labrador retrievers and Dobermann dogs, evidence suggests it is caused by genetic modifications in both *ATP7A* and *ATP7B* genes (Fedoseienko et al., 2014; Fieten et al., 2012, 2016; Forman et al., 2005, p. 1; X. Wu, Mandigers, et al., 2019).

The *COMMD1* gene, also known as *MURR1* and *C2orf5*, encodes the copper metabolism MURR1 domain-containing protein 1 (Favier et al., 2012; Forman et al., 2005; Y. G. Kim et al., 2016; Klomp et al., 2003; Lee et al., 2007; Nakaichi et al., 2021; van De Sluis et al., 2002). The *COMMD1* protein is classified as a housekeeping protein due to its extensive interaction with several other proteins (Fedoseienko et al., 2014). Studies indicated that it plays a vital role in facilitating the recycling of valuable proteins for reuse, such as ATP7B. This protein is involved in copper excretion from hepatocytes, and its absence leads to a progressive accumulation of copper within these cells, resulting in chronic hepatitis and cirrhosis in Bedlington terrier dogs (Favier et al., 2012; Forman et al., 2005; Y. G. Kim et al., 2016; Klomp et al., 2003; Lee et al., 2007; Nakaichi et al., 2021; van De Sluis et al., 2002).

In a genome-wide association study involving 235 Labrador retrievers conducted by Fieten et al. (2016), two chromosome regions were detected as harboring *ATP7B* and *ATP7A*, which were

associated with variations in hepatic copper levels. Through DNA sequence analysis, missense mutations were discovered in both genes. Specifically, the amino acid substitution ATP7B:p.Arg1453Gln (mutation *ATP7B*:c.4358G>A) was found to be linked to copper accumulation, while the amino acid substitution ATP7A:p.Thr327Ile (mutation *ATP7A*:c.980C>T) was observed to provide partial protection against copper accumulation (Fieten et al., 2016). Wu et.al (2019), in a research explored whether these two detected mutations (*ATP7A*:c.980C>T and *ATP7B*:c.4358G>A) in Labrador Retrievers were associated with hepatic copper accumulation in Dobermann pinschers as well. They found a significant association between the *ATP7B*: c.4358G>A mutation and hepatic copper accumulation in Dobermanns (X. Wu, Mandigers, et al., 2019). The *ATP7B* gene encodes the ATPase copper-transporting beta (ATP7B) protein, and it plays a crucial role in regulating copper levels in the liver and facilitating copper excretion into the bile (Bull et al., 1993).

In addition to Bedlington terriers (Twedt et al., 1979), Labrador retrievers, and Dobermann pinschers (Thornburg, 1998), there is evidence for familial hepatic copper toxicosis disorder in a few other dog breeds, including Dalmatians (V. L. Cooper et al., 1997; Corbee & Penning, 2021; Webb et al., 2002), Skye terriers (Askwith & Kaplan, 1998; V. L. Cooper et al., 1997), West-Highland White terriers (Thornburg et al., 1986, 1990), and Keeshonds (Dirksen, Burgener, et al., 2017).

Pedigree studies suggested that copper toxicosis in dogs can be inherited in a complex manner, except for the Bedlington terrier breed (Corbee & Penning, 2021). This means that a single genetic mutation does not solely determine the phenotype of copper toxicosis but is influenced by other factors, such as environmental factors and/or additional genes (Corbee & Penning, 2021). For example, a study by Fieten et al. (2016) suggests that copper toxicosis in dogs is believed to

be inherited through a complex mode of inheritance involving both sex-linked and additive patterns (Fieten et al., 2016). According to a pedigree study conducted by Nakaichi et al. (2022), the findings indicated that copper-associated hepatitis observed in the studied Dalmatians follows an autosomal recessive mode of inheritance (Nakaichi et al., 2022). Regardless of the mode of inheritance, investigating the causes of this disorder in different dog breeds could have significant value in the veterinary field (Corbee & Penning, 2021; X. Wu, Mandigers, et al., 2020).

### 3.1.4. Using whole-genome sequencing in the search for copper toxicosis causes in Dalmatians

Whole-genome sequencing (WGS) has facilitated the identification of mutations accountable for various inherited conditions in dogs (Jagannathan et al., 2019). Notable instances include the identification of a de novo variant in *ASPRV1* linked to ichthyosis (Bauer et al., 2017), variants in *GJA9* associated with polyneuropathy (also known as lower motor neuron disease) (Becker et al., 2017), and *RBP4* involved in a congenital eye defect (Kaukonen et al., 2018). The advantage of WGS is that it can provide a nearly complete characterization of the genomes of affected and unaffected individuals. Through the sequencing of even a small number of individuals, variants unique to or enriched in the affected group can be discovered. Moreover, the impacts of variants on gene products can be reliably predicted using gene location and function information available from gene information databases such as those provide by NCBI and Ensembl. The WGS technology together with bioinformatics approach works particularly well for Mendelian disorders (Teare & Santibañez Koref, 2014) but has been used for complex conditions as well (French et al., 2022; Kanzi et al., 2020).

Given the success of recent work in dogs and other species employing WGS for causative mutation investigation, we sought to identify CT mutations in Dalmatians using WGS. Extensive pedigree analysis conducted for this project (Dr. Georgina Macintyre, Personal Communication, March 01, 2017) and a smaller study by others supports an autosomal recessive inheritance pattern (Nakaichi et al., 2021, 2022), suggesting that a causal mutation might be found through studying the presence and inheritance pattern of variants in affected and unaffected individuals. In this study we analyzed whole-genome sequence data from 13 Dalmatians including presumptive affected, unaffected, and carrier individuals identified through a combination of pedigree analysis, liver biopsy, and owner-provided health information. Three Bedlington Terrier dogs were included in the study as well, including one CT-affected individual known to carry a *COMMD1* mutation, another with CT but with an intact *COMMD1* gene and a surmised unknown CT-associated mutation, and an unaffected individual, also with an intact *COMMD1* gene.

## 3.2. Research Objectives

a) Identification and genotyping of SNPs and indels through whole-genome sequence analysis

b) Identification and genotyping of structural variants through whole-genome sequence analysis

c) Identification of candidate CT-causal variants by variant filtering

d) Detection of candidate CT-causal variants from read coverage and read alignment analysis

## 3.3. Methods/Experimental Design and Materials

### 3.3.1. Samples for sequencing

The present study obtained ethical research clearance from the Committee of Animal Care and Use at the University of Alberta (AUP 2165) under the guidelines set forth by the Canadian Council on Animal Care (CCAC) regarding the care and utilization of animals in research, pedagogy, and assessment (Canadian Council on Animal Care, 2009). In the US, approved by the Colorado State University Institutional Animal Care and Use Committee (15-5632A and KP IACUC Initial 4138), following Federal Regulations and Guidelines.

In total, genomic DNAs from sixteen dogs, including 13 Dalmatians and 3 Bedlington Terriers, were sequenced using the HiSeqX PE150 and NovaSeq 6000 PE150 Illumina platforms (30x coverage). The dog samples for sequencing were carefully chosen based on pedigree information, disease status (healthy, mildly affected and affected). Five Dalmatian families with affected and unaffected littermates were chosen (Figure 3.2). Two of the Bedlington Terriers were affected and the other one was a healthy individual. Further details on the samples can be found in Table 3.1.

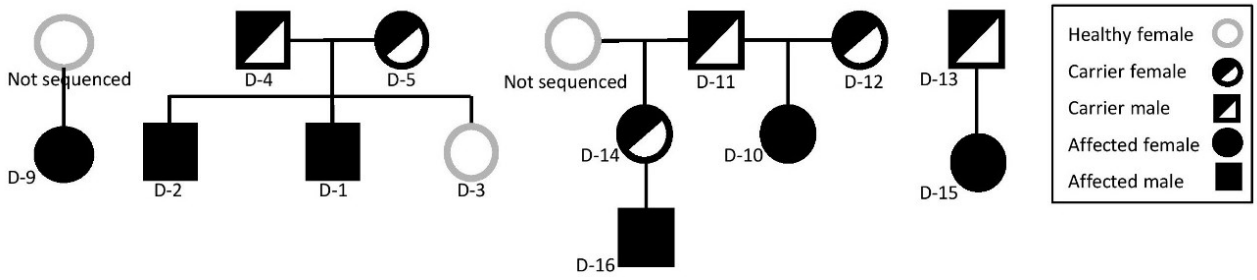


**Figure 3.2.** Pedigree of the Dalmatian families.

**Table 3.1.** Details of sequenced animal samples.

| Sample | Breed | Relationship | Sex | Biopsy | Disease Status |
|---|---|---|---|---|---|
| D-2 | Dalmatian | Littermate of D-1 & D-3 | M | Yes | Affected - Copper toxicosis |
| D-9 | Dalmatian | Not related | F | Yes | Affected - Copper toxicosis |
| D-10 | Dalmatian | Offspring of D-11 & D12 | F | Yes | Affected - Copper toxicosis - copper values not yet provided |
| D-15 | Dalmatian | Offspring of D-13 | F | Yes | Affected - Copper toxicosis |
| D-16 | Dalmatian | Offspring of D-14 | M | Yes | Affected - Copper toxicosis |
| D-3 | Dalmatian | Littermate of D-2 & D-1 | F | Yes | Healthy - No liver copper at 4 years |
| D-4 | Dalmatian | Sire of D1-D3 | M | No | Healthy |
| D-5 | Dalmatian | Dam of D1-D3 | F | No | Healthy |
| D-1 | Dalmatian | Littermate of D-2 & D-3 | M | Yes | Mildly affected - Mild copper-associated hepatopathy |
| D-11 | Dalmatian | Sire of D-10 & D-14 and Great Grand Sire (dam side) of D-16 | M | Yes | Unknown - negative biopsy (anecdotal) |
| D-12 | Dalmatian | Dam of D-10 | F | Unknown | 'Healthy' at 16 years old (anecdotal) |
| D-13 | Dalmatian | Sire of D-15 | M | Unknown | Unknown |
| D-14 | Dalmatian | Dam of D-16 | F | Unknown | Unknown |
| BT-6 | Bedlington Terrier | Not related | M | Yes | Affected - Copper toxicosis |
| BT-7 | Bedlington Terrier | Not related | F | Yes | Healthy |
| BT-8 | Bedlington Terrier | Not related | F | Yes | Affected - Copper toxicosis |

### 3.3.2. Read mapping and genetic variant calling (SNPs, indels, or larger deletions)

A Snakemake workflow (Köster & Rahmann, 2012) was used to detect SNPs and indels in a high-performance computing (HPC) environment provided by Compute Canada. Prior to the variant calling analysis, the initial stages of the workflow involved quality control and preprocessing of the raw sequencing data (FASTQ files). FastQC v0.11.9 and Trimmomatic v0.36

were used to inspect read quality and to trim reads, respectively. BWA v0.7.17 (H. Li & Durbin, 2009) was used to align the trimmed reads to a reference genome of Canis lupus familiaris (ROS_Cfam_1.0). For coordinating, sorting and indexing the generated SAM files to BAM files the Picard v2.18 (Broad Institute, 2018) SortSam tool was used. Picard v2.18 MarkDuplicates (Broad Institute, 2018) was used to combine the BAM files from the same sample and mark suspected read duplicates. Base quality score recalibration (BQSR) was performed using the BaseRecalibrator tool of Genome Analysis Toolkit (GATK) v3.8 (McKenna et al., 2010) to improve the accuracy of variant calling. The quality of the recalibration run was assessed using GATK AnalyzeCovariates tool. Potential variant sites were detected using GATK HaplotypeCaller, and the resulting GVCF files were jointly genotyped with GATK GenotypeGVCFs. SNPs and indels were extracted using GATK SelectVariants. GATK VariantFiltration was employed to generate filters for soft filtering of SNPs and indels.

### 3.3.3. Structural variant calling

Smoove v0.2.6 (Pedersen et al., 2018/2020) was used to detect structural variants (SVs) and genotype samples, excluding gapped regions and regions where coverage is more than four standard deviations higher than the mean coverage, calculated across 10,000 bp windows by Mosdepth v0.3.2 (Pedersen & Quinlan, 2018). Genotyping was followed by running Duphold v0.2.3 (Pedersen & Quinlan, 2019) to annotate SV calls with sequence depth information. Next, SnpSift v4.3t (Cingolani, Patel, et al., 2012) was used to soft filter variants that satisfied the following criteria: for all SVs, a MeanSmooveHetQuality (MSHQ) > 3; in the case of deletions, a fold-change for the variant (H. Li et al., 2009) depth relative to flanking regions (DHFFC) < 0.7; and in the case of duplications, a DHFFC > 1.25.

### 3.3.4. Mapping rate and average genome coverage

Two metrics of mapping rate and average genome coverage were used to check the quality of the read-mapping process and its output BAM files. The mapping rate, also known as the "Uniquely mapped reads percentage," is the ratio of uniquely mapped reads out of all the input reads. A mapping rate exceeding 90% is deemed excellent for variant calling analysis, while mapping rates above 80% are considered good (Dobin & Gingeras, 2015; Mistry et al., 2021). SAMtools v1.12 (H. Li et al., 2009) was used to calculate this metric.

In the concept of whole-genome sequencing, the average genome coverage can be estimated using the parameters including the number of reads (N), the length of the average read (L) and the length of the reference genome (G):

$N \times L/G$ (Sims et al., 2014)

Mosdepth v0.3.2 (Pedersen & Quinlan, 2018) was used for estimating the average genome coverage of the studied sample's BAM files. The acceptable range of the average genome coverage for a reliable variant calling is different depending on various study goals. Although according to studies by Ajay et al. (2011) and Sims et al. (2014) the acceptable range of the average genome coverage for a reliable variant calling should be between 30 to 50 (Ajay et al., 2011; Sims et al., 2014), Jiang et al. (2019) findings indicated that an average genome coverage of 10 provides an optimal balance for practical purposes, ensuring sufficient coverage and reliable detection of variants (Jiang et al., 2019).

### 3.3.5. Verifying sample characteristics

Variant calling results were analyzed to verify sample sex and reported relationships, to highlight sample handling or attribution errors. For sex assessment, X chromosome heterozygous

SNP proportion was calculated from the SNP VCF file using the vcftools v0.1.16 (Danecek et al., 2011). The "het" option in vcftools was used for this calculation, considering only SNPs with a "FILTER" value of "PASS" in the VCF file. Relatedness between samples was calculated using the KING algorithm, implemented in vcftools v0.1.16 (Danecek et al., 2011), and accessed using the relatedness2 option. The analysis was performed on the SNP VCF file, considering only SNPs with a "FILTER" value of "PASS" according to the variant call format (VCF) file used in the study.

### 3.3.6. Identification of candidate causal variants by variant filtering

#### 3.3.6.1. Variant annotation

SNPs and indels were annotated with functional impact predictions using SnpEff v5.0e (Cingolani, Platts, et al., 2012). For SVs annotation, Ensembl Variant Effect Predictor (VEP) (release 105) (McLaren et al., 2016) was used.

#### 3.3.6.2. Variant filtering

Variants were filtered using SnpSift v5.0e (Cingolani, Patel, et al., 2012) based on quality, genotype, predicted functional impact, and overlap with previously proposed candidate copper toxicosis genes. Genotype filtering was used to identify variants that fit with the expected mode of inheritance, the disease status of study animals, and the known relationships among study animals. Multiple filters were applied, to accommodate uncertainties around the status of some samples. Predicted functional impact filtering was used to identify variants that, based on variant annotation, are expected to impact protein function. For filtering based on overlap with candidate genes, a list of 71 genes (as presented in Table 3.2) known to be involved in copper balance

regulation, as well as closely related genes or those potentially involved in copper metabolism or the development of copper metabolism disorders in dog were selected (X. Wu, den Boer, et al., 2020). The VCF outputs obtained through the previous filtering step (genotype-based filtering) were used in this analysis. The functional impact and candidate gene filters were applied separately as well as in succession to the variants passing the genotype filters.

**Table 3.2.** List of candidate genes used for variant filtering.

| Gene | EnsemblID (Gene stable ID) | Chromosome | Gene | EnsemblID (Gene stable ID) | Chromosome |
|---|---|---|---|---|---|
| *GRM1* | ENSCAFG00845013661 | 1 | *STX5* | ENSCAFG00845007709 | 18 |
| *COX6B1* | ENSCAFG00845001977 | 1 | *MGST2* | ENSCAFG00845021915 | 19 |
| *SOD2* | ENSCAFG00845001754 | 1 | *CHCHD4* | ENSCAFG00845023967 | 20 |
| *MT2A* | ENSCAFG00845003852 | 2 | *LTF* | ENSCAFG00845011930 | 20 |
| *MT4* | ENSCAFG00845003871 | 2 | *STXBP2* | ENSCAFG00845027713 | 20 |
| *MT1* | ENSCAFG00845003828 | 2 | *RETN* | ENSCAFG00845027860 | 20 |
| *COMMD3* | ENSCAFG00845008000 | 2 | *ATP7B* | ENSCAFG00845020988 | 22 |
| *PAM* | ENSCAFG00845001601 | 3 | *COMMD6* | ENSCAFG00845024802 | 22 |
| *FAH* | ENSCAFG00845004974 | 3 | *CP* | ENSCAFG00845023529 | 23 |
| *ATOX1* | ENSCAFG00845005141 | 4 | *COMMD2* | ENSCAFG00845024239 | 23 |
| *SCO1* | ENSCAFG00845002812 | 5 | *FOXA2* | ENSCAFG00845027259 | 24 |
| *COX10* | ENSCAFG00845004049 | 5 | *PRNP* | ENSCAFG00845013725 | 24 |
| *CIRH1A* | ENSCAFG00845010015 | 5 | *COMMD7* | ENSCAFG00845021762 | 24 |
| *COX19* | ENSCAFG00845008542 | 6 | *GSS* | ENSCAFG00845017262 | 24 |
| *ALT* | ENSCAFG00845011384 | 6 | *EDN3* | ENSCAFG00845014217 | 24 |
| *NPC1* | ENSCAFG00845014993 | 7 | *CLU* | ENSCAFG00845010565 | 25 |
| *ESCO1* | ENSCAFG00845016045 | 7 | *ACTA2* | ENSCAFG00845029801 | 26 |
| *CLUL1* | ENSCAFG00845001733 | 7 | *KRT18* | ENSCAFG00845023645 | 27 |
| *FOXA1* | ENSCAFG00845002956 | 8 | *DMT1* | ENSCAFG00845015776 | 27 |
| *KTN1* | ENSCAFG00845015557 | 8 | *COX15* | ENSCAFG00845030808 | 28 |
| *ABCD4* | ENSCAFG00845014165 | 8 | *CUTC* | ENSCAFG00845030834 | 28 |
| *COX11* | ENSCAFG00845016207 | 9 | *COMMD4* | ENSCAFG00845019252 | 30 |
| *SURF1* | ENSCAFG00845009922 | 9 | *APP* | ENSCAFG00845015660 | 31 |
| *SCO2* | ENSCAFG00845027878 | 10 | *SOD1* | ENSCAFG00845022725 | 31 |
| *COMMD1* | ENSCAFG00845016604 | 10 | *COX17* | ENSCAFG00845017310 | 33 |
| *COMMD10* | ENSCAFG00845005464 | 11 | *ALS2* | ENSCAFG00845030698 | 37 |
| *CTR2* | ENSCAFG00845014791 | 11 | *CPO* | ENSCAFG00845021788 | 37 |
| *CTR1* | ENSCAFG00845007669 | 11 | *IDH1* | ENSCAFG00845023478 | 37 |
| *PRDM1* | ENSCAFG00845007842 | 12 | *PIKFYVE* | ENSCAFG00845023569 | 37 |
| *COMMD5* | ENSCAFG00845011944 | 13 | *CPS1* | ENSCAFG00845017752 | 37 |
| *COMMD8* | ENSCAFG00845008641 | 13 | *ABCA12* | ENSCAFG00845020774 | 37 |
| *IL6* | ENSCAFG00845008825 | 14 | *SLC11A1* | ENSCAFG00845027087 | 37 |
| *COMMD9* | ENSCAFG00845020681 | 18 | *DNPEP* | ENSCAFG00845030107 | 37 |
| *WT1* | ENSCAFG00845027359 | 18 | *ATP7A* | ENSCAFG00845020199 | X |
| *MTL5* | ENSCAFG00845006803 | 18 | *XIAP* | ENSCAFG00845025680 | X |
| *CCS* | ENSCAFG00845012574 | 18 | | | |

Data in this table is from X. Wu, den Boer, et al., (2020).

### 3.3.7. Detection of candidate causal variants from inspection of copper toxicosis candidate gene alignments using IGV

The list of candidate genes related to copper toxicosis in dogs in the previous step was also used in this step. Ensembl identifiers of candidate copper toxicosis genes were used to obtain the locations of the genes on the ROS_Cfam_1.0 assembly. IGV v2.12.3 (Thorvaldsdóttir et al., 2013) was used to inspect the copper toxicosis-related gene regions, including 1 kb upstream and 1 kb downstream in each study sample. The BAM files from all study samples were loaded into IGV, along with the gene locations and their flanking sequence in BED format. An attributes file was supplied so that sample information could be displayed in IGV. Each gene region was inspected for visual evidence of low or absent reads. Read pairing and insert size information was examined for evidence of deletions, insertions, inversions, and translocations.

## 3.4. Results

### 3.4.1. Sequencing, variant calling, and sample quality control

Alignment of sequence reads from the 16 sequenced individuals yielded an average genome coverage of 38.13 and a read mapping rate of 99.43% (Table 3.3), which is well within the range recommended for variant discovery and genotyping (Ajay et al., 2011; Mistry et al., 2021). SNP/Indel and SV variant calling produced 13,391,948 and 49,429 variants respectively, which were used in downstream variant filtering. Relationship and sex checks conducted using the SNP data were consistent with sample attributes provided prior to sequencing, except for the sex assignments of the Bedlington Terrier samples (data not shown). The cause of this discrepancy is unknown but did not preclude the planned analyses, which focused on the Dalmatians.

**Table 3.3.** Average genome coverage and read mapping rates of the studied samples.

| Sample | Average genome coverage | Mapping rate (%) |
|---|---|---|
| D-2 | 40.16 | 99.03 |
| D-9 | 35.17 | 99.84 |
| D-10 | 37.24 | 99.85 |
| D-15 | 33.70 | 99.84 |
| D-16 | 33.50 | 99.87 |
| D-3 | 43.05 | 98.99 |
| D-4 | 41.33 | 98.84 |
| D-5 | 43.31 | 98.95 |
| D-1 | 38.87 | 99.06 |
| D-11 | 30.15 | 99.89 |
| D-12 | 32.05 | 99.88 |
| D-13 | 37.35 | 99.85 |
| D-14 | 41.50 | 99.86 |
| BT-6 | 42.46 | 98.95 |
| BT-7 | 39.13 | 99.05 |
| BT-8 | 41.03 | 99.05 |
| **Mean** | **38.13** | **99.43** |

### 3.4.2. Variant filtering

The SNP/Indel and SV variant calling workflows generated VCF files containing millions of variants, which is not unexpected. To refine the search for a candidate causal variant, several filtering steps were taken, each reducing the number of remaining variants under consideration (Table 3.4). The first filtering step called “Quality” was applied to remove false-positive variants using commonly used criteria in the variant calling process. This quality filter was not applied to the SVs due to the lack of filtering tools and reference databases for identifying false-positive SVs.

**Table 3.4.** The number of SNPs, indels and SVs that passed downstream variant filtering.

| Input source | Input SNPs | Input Indels | Input SVs | Filters | Output SNPs | Output Indels | Output SVs |
|---|---|---|---|---|---|---|---|
| Variant calling | 9,050,768 | 3,548,482 | 49,429 | Quality | 8,216,377 | 3,480,761 | 49,429 |
| Quality | 8,216,377 | 3,480,761 | 49,429 | Scenario 1 | 3,252 | 1,979 | 17 |
| Scenario 1 | 3,252 | 1,979 | 17 | Function | 7 | 1 | 0 |
| Scenario 1 | 3,252 | 1,979 | 17 | Candidate | 0 | 3 | 0 |
| Scenario 1 | 3,252 | 1,979 | 17 | Function & Candidate | 0 | 0 | 0 |
| Quality | 8,216,377 | 3,480,761 | 49,429 | Scenario 2 | 0 | 0 | 0 |
| Quality | 8,216,377 | 3,480,761 | 49,429 | Scenario 2b | 5 | 17 | 0 |
| Scenario 2b | 5 | 17 | 0 | Function | 0 | 0 | 0 |
| Scenario 2b | 5 | 17 | 0 | Candidate | 0 | 0 | 0 |
| Scenario 2b | 5 | 17 | 0 | Function & Candidate | 0 | 0 | 0 |

After quality control filtering, multiple "filtering by genotypes" approaches were applied, in order to identify variants that fit with the hypothesized autosomal recessive inheritance pattern for a causative allele. Three different genotype scenarios were considered (Table 3.5). Each scenario is defined by the genotype pattern that leads to variants passing the filter (i.e., being considered as candidate causal mutations). Alleles in the genotypes are expressed either as “0” corresponding to the allele observed in the reference genome, or “1” corresponding to the allele not observed in the reference genome (also termed the alternative allele). In the first scenario (Scenario 1), a subset of samples with high-confidence phenotypes and their parents were considered. This filter kept variants where affected individuals D-2, D-9, D-10, D-15, and D-16 have homozygous non-reference genotypes (1/1), and carrier parents D-4 and D-5 are heterozygous (1/0). An unaffected littermate D-3 was permitted to be homozygous reference (0/0) or heterozygous (1/0) (Table 3.5). The genotypes of the remaining samples were not examined in

Scenario 1. The number of variants entering and exiting this, and other filters is shown in Table 3.4 .

An alternate, more restrictive filter was also applied, termed “Scenario 2”. This filter considered the genotypes of all the Dalmatians in the data set, looking for 1/1 genotypes in the case of affected animals, and 1/0 genotypes for parents of affected animals. For this filter, D-2 was assigned a genotype of 0/0 to be consistent with the autosomal recessive model and the lack of disease in this individual at 4 years of age. No variants passed this filter (Table 3.4).

A third filter was applied, termed “Scenario 2b”, that is slightly less restrictive than Scenario 2 in that it allows D-3 to be 1/0 or 0/0 and D-1 to be 1/1 or 1/0 (Table 3.5). These allowances were made to capture uncertainty around these samples’ genotypes, given their phenotypes and /or relationships to other animals in the data set. D-1 dog was mildly affected (diagnosed with mild copper-associated hepatopathy). This could indicate two possibilities that we considered in this filtering scenario: either this individual was a CT-affected Dalmatian that was detected in the early stages of the progression of the disorder, or it is a carrier which influenced by an environmental factor, such as a diet rich in copper, or potentially other genetic and environmental changes. D-3 sample was a healthy individual with no signs of hepatic copper at 4 years. This diagnosis could imply two possibilities that was considered in this filtering scenario: This individual could be a non-carrier and therefore representing a completely healthy genetic control, or it could be a carrier without evident symptoms, which is typically the case. The challenge here arises from our limited understanding of whether carriers can eventually develop symptoms, a question that also poses complexities in humans, where most carriers do not exhibit symptoms, based on our current understanding. In the end, 5 SNPs, 17 indels and no SVs were kept in the resulting file following the implementation of this filtering step (Table 3.4).

**Table 3.5.** Filtering by genotypes scenarios

| Sample | Disease status | Relationship | If autosomal recessive (simple model) | Scenario 1 genotypes | Scenario 2 genotypes | Scenario 2b genotypes |
|---|---|---|---|---|---|---|
| D-2 | Affected - Copper toxicosis | Littermate of D-1 & D-3 | Diseased | 1/1 | 1/1 | 1/1 |
| D-9 | Affected - Copper toxicosis | Distantly related (no parental DNA sequenced) | Diseased | 1/1 | 1/1 | 1/1 |
| D-10 | Affected - | Offspring of D-11 & D12 | Diseased | 1/1 | 1/1 | 1/1 |
| D-15 | Affected - Copper toxicosis | Offspring of D-13 | Diseased | 1/1 | 1/1 | 1/1 |
| D-16 | Affected - Copper toxicosis | Offspring of D-14 | Diseased | 1/1 | 1/1 | 1/1 |
| D-3 | Healthy - No liver copper at 4 years | Littermate of D-2 & D-1 | Healthy (possible carrier, but unlikely) | 1/0 or 0/0 | 0/0 | 1/0 or 0/0 |
| D-4 | Healthy | Sire of D1-D3 | Obligate carrier | 1/0 | 1/0 | 1/0 |
| D-5 | Healthy | Dam of D1-D3 | Obligate carrier | 1/0 | 1/0 | 1/0 |
| D-1 | Mildly affected - Mild copper-associated hepatopathy | Littermate of D-2 & D-3 | Obligate carrier/affected | - | 1/1 | 1/1 or 1/0 |
| D-11 | Unknown - negative biopsy (anecdotal) | Sire of D-10 & D-14 and Great Grand Sire of D-16 | Obligate carrier | - | 1/0 | 1/0 |
| D-12 | 'Healthy' at 16 years old (anecdotal) | Dam of D-10 | Obligate carrier | - | 1/0 | 1/0 |
| D-13 | Unknown | Sire of D-15 | Obligate carrier | - | 1/0 | 1/0 |
| D-14 | Unknown | Dam of D-16 | Obligate carrier | - | 1/0 | 1/0 |
| BT-6 | Affected - Copper toxicosis | Not related | Diseased - fits published gene model - high copper | - | - | - |
| BT-7 | Healthy | Not related | Healthy (possible carrier) | - | - | - |
| BT-8 | Affected - Copper toxicosis | Not related | Diseased - does not fit published gene model - intermediate copper | - | - | - |

Given the large number of variants passing Scenario 1 and 2b, three additional filters were applied. The first filter kept variants predicted to have an impact on the function of a gene or protein and was termed “Function”. The second filter, termed “Candidate”, kept variants that overlap with a proposed copper toxicosis candidate gene. The third filter, termed “Function & candidate” kept variants that passed both function and candidate gene filters. Ultimately, just 7 SNPs and 1 indel from Scenario 1 were identified as having a predicted functional impact on genes, and 3 indels from Scenario 1 overlapped with candidate genes (Table 3.6). No variants from Scenario 2b passed either filter and no variants passed through both filters.

**Table 3.6.** Details of Scenario1 variants that passed final filtrations.

| Variant | Final Filter | Chromosome | Position | ID | Gene Name | Ensembl Gene ID | Impact | Consequence | Gene Information |
|---|---|---|---|---|---|---|---|---|---|
| SNP | Function | 6 | 55489862 | rs24346917 | *ABCA4* | ENSCAFG00845015579 | MODERATE | Missense variant[1] | Description: ATP binding cassette subfamily A member 4 [Source: NCBI]. Gene type: Protein coding |
| SNP | Function | 9 | 36668670 | rs851554803 | *PPM1D* | ENSCAFG00845028053 | MODERATE | Missense variant | Description: protein phosphatase, Mg2+/Mn2+ dependent 1D [Source: HGNC]. Gene type: Protein coding |
| SNP | Function | 9 | 61752385 | rs852337682 | *DAB2IP* | ENSCAFG00845014246 | MODERATE | Missense variant | Description: DAB2 interacting protein [Source: HGNC]. Gene type: Protein coding |
| SNP | Function | 13 | 50140922 | rs852477228 | *ENSCAFG00845014237* | ENSCAFG00845014237 | MODERATE | Missense variant | Description: This is a novel dog gene [Source: Ensembl]. Gene type: Protein coding |
| SNP | Function | 23 | 6948039 | rs851775508 | *TRANK1* | ENSCAFG00845026194 | MODERATE | Missense variant | Description: tetratricopeptide repeat and ankyrin repeat containing 1 [Source: HGNC]. Gene type: Protein coding |
| SNP | Function | 35 | 23807015 | rs23933721 | *ENSCAFG00845024392* | ENSCAFG00845024392 | MODERATE | Missense variant | Description: This is a novel dog gene [Source: Ensembl]. Gene type: Protein coding |
| SNP | Function | 35 | 24024171 | rs852645952 | *ENSCAFG00845025154* | ENSCAFG00845025154 | MODERATE | Missense variant | Description: This is a novel dog gene that has 1 transcript [Source: Ensembl]. Gene type: Protein coding |
| Indel | Function | 10 | 20740068 | . | *PPARA* | ENSCAFG00845010405 | HIGH | Frameshift variant [2] | Description: peroxisome proliferator activated receptor alpha [Source: NCBI]. Gene type: Protein coding |
| Indel | Candidate | 8 | 31692149 | . | *KTN1* | ENSCAFG00845015557 | MODIFIER | intron variant[3] | Description: kinectin 1 [Source: HGNC]. Gene type: Protein coding |
| Indel | Candidate | 24 | 1563989 | . | *FOXA2-PAX1* | ENSCAFG00845027259-ENSCAFG00845027290 | MODIFIER | intergenic region[4] | Note: This variant overlapped with two candidate genes. *FOXA2* description: forkhead box A2 [Source: HGNC]. *PAX1* description: paired box 1 [Source: HGNC]. Both gene's type: Protein coding. |
| Indel | Candidate | 33 | 23542313 | . | *COX17* | ENSCAFG00845017310 | MODIFIER | downstream gene variant[5] | Description: cytochrome c oxidase copper chaperone COX17 [Source: NCBI]. Gene type: Protein coding |

1. A genetic mutation where one amino acid in a protein is replaced by another due to a single nucleotide change. This can affect the protein's function and may or may not be linked to disease. 2. A genetic mutation caused by insertions or deletions of DNA nucleotides, shifting the reading frame during translation. This often results in a nonfunctional or truncated protein and possibly can be associated with severe consequences. 3. A variant occurring within a gene's non-coding intron region, between exons. These variants typically don't impact protein-coding but can affect gene regulation and splicing. 4. A variant in non-coding genomic regions between genes. These variants don't affect gene coding sequences and may have regulatory or structural roles. 5. A variant in the non-coding region just after a gene. These can influence gene regulation and splicing but don't change the gene's coding sequence.

### 3.4.2.1. Inspection of the candidate gene alignments in IGV

The inspection of candidate gene regions in IGV for anomalous read coverage or alignments revealed no differences between cases and controls for the Dalmatian samples. In one of the Bedlington Terrier samples (BT-6), a control known to have the *COMMD1* deletion from previous PCR studies (Coronado et al., 2003; Forman et al., 2005), there was a clear absence of read coverage for the *COMMD1* gene region, consisting with a homozygous deletion (figure 3.3). This does not represent a new finding but demonstrates the ability of whole-genome sequencing to reveal such deletions. Notably no such deletion is present in the Dalmatian samples.


**Figure 3.3.** The observed large deletion in the *COMMD1* gene in one of the Bedlington Terrier samples (BT-6).

## 3.5. Discussion/Conclusions

In the present study, various potential causes of copper toxicosis in a collection of presumptive cases, carriers, and controls, considering SNPs, indels, and SVs were investigated. Despite employing SNP, indel, and SV filtering strategies based on genotype, predicted functional consequence, and candidate gene overlap, no obvious candidate causal variants were identified. Ultimately, it was through multiple additional filtering steps on Scenario 1 output variants that we detected 7 SNPs and 1 indel of the variants (variants passing the genotype filters) as having a predicted functional impact on genes and 3 indels were detected to be overlapping with candidate genes. However, because of the lack of the affected gene's known roles in copper metabolism, it is difficult to refine this list further to produce a specific candidate causal mutation.

It is worth noting that among the three detected indels that passed the candidate filter, one is located downstream (position: 23542313 bp) of the *COX17* gene, which encodes the cytochrome c oxidase copper chaperone COX17 (Banci et al., 2011). This indel is categorized as having a modifier functional impact. Modifier functional impact refers to a genetic variant predicted to have a regulatory or modifying effect on other genetic elements rather than directly altering protein-coding sequences. Such variants can play a role in modulating gene expression or modifying the impact of other genetic variants, hence the term "modifier." While this variant doesn't alter the gene's coding sequence, it may affect gene regulation or splicing processes, potentially influencing *COX17*'s role in copper metabolism and cytochrome c oxidase function. However, it is crucial to emphasize that further experiments/research are needed to provide more robust evidence that any of the identified candidate variants are true candidate mutations.

There are multiple potential reasons for the absence of a true candidate causal mutation in this study. One possible explanation is genetic heterogeneity. Genetic heterogeneity happens when various genetic variations or mutations cause the same condition or disease in a group of affected individuals (Rimoin et al., 2013). In other words, it can occur as a result of different mutations of a gene (allelic heterogeneity) or mutations in different genes (locus heterogeneity) and both instances can cause the same disease in different individuals (Rimoin et al., 2013).

The presence of modifier genes could also present difficulties in the search for a causative mutation using the approaches applied in this study. Unlike disease-causing genes, which result in the direct development of specific conditions through pathogenic mutations, modifier genes are genetic elements that do not cause a particular condition independently, instead they affect the expression of other genes, both in terms of their physical manifestation and at the molecular level. They function as controllers or influencers of the impacts exerted by other genes within the genome and have the ability to influence various traits (Geurts van Kessel et al., 2013).

Another possible reason for not detecting a candidate causal mutation in this study could be due to a situation called phenocopies. Phenocopies are instances where a specific set of observable traits (phenotype) appears in individuals, but it is not attributed to the anticipated genetic mutation or inheritance pattern. Instead, the observed phenotype arises from influences other than genetics, such as non-genetic factors or environmental elements. Phenocopies can occur for various reasons, which include: 1) Environmental factors / phenotypic variability: Environmental factors like diet, toxins, medications, or infections can imitate the effects of specific genetic mutations, resulting in similar observable characteristics. Copper toxicosis in some canine breeds, is believed to be a multifactorial disease involving both genetic and environmental factors (Fieten et al., 2016). For example, copper levels in the liver are influenced by the uptake of copper

through diet (Fieten et al., 2016) and in some cases, dogs with high levels of copper in the liver can be treated with a copper-restricted diet (Fieten et al., 2014). Thus, factors and possible interventions may not always be apparent and could complicate the issue of assignment cases and controls.

Another possible explanation for this could be that the detection of certain gene variants might have been limited because the list of candidate genes selected for filtering in this study were primarily chosen based on their known association with copper metabolism including genes involved in copper balance regulation and closely related genes with potential involvement in copper metabolism disorders in dogs. However, implying new filters that include other pathways, such as vesicular trafficking of proteins, might reveal intriguing gene variants.

Additionally, another contributing factor to this issue might be that some of the affected individuals may not have generated high-quality reads across specific genomic regions, which could result in their exclusion from the analysis, as they needed to meet the criteria of having homozygous non-reference genotype (1/1).

Despite the disappointing lack of a clear candidate casual mutation in Dalmatians from this study the results are still of value. Firstly, the in-depth assessments of CT candidate genes done in this study suggest that mutations in the previously identified genes *COMMD1*, *ATP7A*, and *ATP7B* (Fedoseienko et al., 2014; Fieten et al., 2012; Forman et al., 2005; Klomp et al., 2003) are unlikely to be involved in the disease in Dalmatians. In the case of *COMMD1* it is especially clear from the IGV visualizations that a deletion of this gene is not present in the affected Dalmatians, and that we would detect such a deletion if it were present. Therefore, the findings of this study indicate that the existing *COMMD1* deletion test available to breeders, to identify carriers of the CT-

associated allele, will not be applicable or useful for the Dalmatian breed. The results of the presented study also suggest that the same limitation applies to *ATP7B* as well. The test currently offered for Labrador Retrievers will not be a reliable indicator for the Dalmatian breed. Secondly, the sequence data generated can be used in future studies involving additional samples and employing new bioinformatic methods. For example, new reference genomes, mapping procedures, or variant calling methods could deliver new insights.

## 3.6. References


Ajay, S. S., Parker, S. C. J., Abaan, H. O., Fajardo, K. V. F., & Margulies, E. H. (2011). Accurate and comprehensive sequencing of personal genomes. *Genome Research*, *21*(9), 1498–1505. https://doi.org/10.1101/gr.123638.111

Askwith, C., & Kaplan, J. (1998). Iron and copper transport in yeast and its relevance to human disease. *Trends in Biochemical Sciences*, *23*(4), 135–138. https://doi.org/10.1016/s0968-0004(98)01192-x

Bauer, A., Waluk, D. P., Galichet, A., Timm, K., Jagannathan, V., Sayar, B. S., Wiener, D. J., Dietschi, E., Müller, E. J., Roosje, P., Welle, M. M., & Leeb, T. (2017). A de novo variant in the ASPRV1 gene in a dog with ichthyosis. *PLoS Genetics*, *13*(3), e1006651. https://doi.org/10.1371/journal.pgen.1006651

Becker, D., Minor, K. M., Letko, A., Ekenstedt, K. J., Jagannathan, V., Leeb, T., Shelton, G. D., Mickelson, J. R., & Drögemüller, C. (2017). A GJA9 frameshift variant is associated with polyneuropathy in Leonberger dogs. *BMC Genomics*, *18*(1), 662. https://doi.org/10.1186/s12864-017-4081-z

Broad Institute. (2018). *Picard Tools—By Broad Institute*. http://broadinstitute.github.io/picard/

Bull, P. C., Thomas, G. R., Rommens, J. M., Forbes, J. R., & Cox, D. W. (1993). The Wilson disease gene is a putative copper transporting P-type ATPase similar to the Menkes gene. *Nature Genetics*, *5*(4), 327–337. https://doi.org/10.1038/ng1293-327

Chelly, J., Tümer, Z., Tønnesen, T., Petterson, A., Ishikawa-Brush, Y., Tommerup, N., Horn, N., & Monaco, A. P. (1993). Isolation of a candidate gene for Menkes disease that encodes a potential heavy metal binding protein. *Nature Genetics*, *3*(1), 14–19. https://doi.org/10.1038/ng0193-14

Chen, J., Jiang, Y., Shi, H., Peng, Y., Fan, X., & Li, C. (2020). The molecular mechanisms of copper metabolism and its roles in human diseases. *Pflugers Archiv: European Journal of Physiology*, *472*(10), 1415–1429. https://doi.org/10.1007/s00424-020-02412-2

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Frontiers in Genetics*, *3*, 35. https://doi.org/10.3389/fgene.2012.00035

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Coffey, A. J., Durkie, M., Hague, S., McLay, K., Emmerson, J., Lo, C., Klaffke, S., Joyce, C. J., Dhawan, A., Hadzic, N., Mieli-Vergani, G., Kirk, R., Elizabeth Allen, K., Nicholl, D., Wong, S., Griffiths, W., Smithson, S., Giffin, N., Taha, A., … Bandmann, O. (2013). A genetic study of Wilson's disease in the United Kingdom. *Brain*, *136*(5), 1476–1487. https://doi.org/10.1093/brain/awt035

Cooper, V. L., Carlson, M. P., Jacobson, J., & Schneider, N. R. (1997). Hepatitis and increased copper levels in a dalmatian. *Journal of Veterinary Diagnostic Investigation: Official Publication of the American Association of Veterinary Laboratory Diagnosticians, Inc*, *9*(2), 201–203. https://doi.org/10.1177/104063879700900217

*Copper Poisoning in Dogs—Symptoms, Causes, Diagnosis, Treatment, Recovery, Management, Cost*. (n.d.). Retrieved April 21, 2022, from https://wagwalking.com/condition/copper-poisoning, https://wagwalking.com/condition/copper-poisoning

Corbee, R. J., & Penning, L. C. (2021). COMMD1 Exemplifies the Power of Inbred Dogs to Dissect Genetic Causes of Rare Copper-Related Disorders. *Animals: An Open Access Journal from MDPI*, *11*(3), 601. https://doi.org/10.3390/ani11030601

Coronado, V. A., Damaraju, D., Kohijoki, R., & Cox, D. W. (2003). New haplotypes in the Bedlington terrier indicate complexity in copper toxicosis. *Mammalian Genome*, *14*(7), 483–491.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dirksen, K., Burgener, I. A., Rothuizen, J., van den Ingh, T. S. G. a. M., Penning, L. C., Spee, B., & Fieten, H. (2017). Sensitivity and Specificity of Plasma ALT, ALP, and Bile Acids for Hepatitis in Labrador Retrievers. *Journal of Veterinary Internal Medicine*, *31*(4), 1017–1027. https://doi.org/10.1111/jvim.14716

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *51*, 11.14.1-11.14.19. https://doi.org/10.1002/0471250953.bi1114s51

Favier, R. P., Spee, B., Schotanus, B. A., Ingh, T. S. G. A. M. van den, Fieten, H., Brinkhof, B., Viebahn, C. S., Penning, L. C., & Rothuizen, J. (2012). COMMD1-Deficient Dogs Accumulate Copper in Hepatocytes and Provide a Good Model for Chronic Hepatitis and Fibrosis. *PLOS ONE*, *7*(8), e42158. https://doi.org/10.1371/journal.pone.0042158

Fedoseienko, A., Bartuzi, P., & van de Sluis, B. (2014). Functional understanding of the versatile protein copper metabolism MURR1 domain 1 (COMMD1) in copper homeostasis. *Annals of the New York Academy of Sciences*, *1314*, 6–14. https://doi.org/10.1111/nyas.12353

Fieten, H., Biourge, V. C., Watson, A. L., Leegwater, P. A. J., van den Ingh, T. S. G. A. M., & Rothuizen, J. (2014). Nutritional management of inherited copper-associated hepatitis in the Labrador retriever. *Veterinary Journal (London, England: 1997)*, *199*(3), 429–433. https://doi.org/10.1016/j.tvjl.2013.12.017

Fieten, H., Gill, Y., Martin, A. J., Concilli, M., Dirksen, K., van Steenbeek, F. G., Spee, B., van den Ingh, T. S. G. A. M., Martens, E. C. C. P., Festa, P., Chesi, G., van de Sluis, B., Houwen, R. H. J. H., Watson, A. L., Aulchenko, Y. S., Hodgkinson, V. L., Zhu, S., Petris, M. J., Polishchuk, R. S., … Rothuizen, J. (2016). The Menkes and Wilson disease genes counteract in copper toxicosis in Labrador retrievers: A new canine model for copper-metabolism disorders. *Disease Models & Mechanisms*, *9*(1), 25–38. https://doi.org/10.1242/dmm.020263

Fieten, H., Leegwater, P. A. J., Watson, A. L., & Rothuizen, J. (2012). Canine models of copper toxicosis for understanding mammalian copper metabolism. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, *23*(1–2), 62–75. https://doi.org/10.1007/s00335-011-9378-7

Forman, O. P., Boursnell, M. E. G., Dunmore, B. J., Stendall, N., van den Sluis, B., Fretwell, N., Jones, C., Wijmenga, C., Rothuizen, J., van Oost, B. A., Holmes, N. G., Binns, M. M., & Jones, P. (2005). Characterization of the COMMD1 (MURR1) mutation causing copper toxicosis in Bedlington terriers. *Animal Genetics*, *36*(6), 497–501. https://doi.org/10.1111/j.1365-2052.2005.01360.x

French, C. E., Dolling, H., Mégy, K., Sanchis-Juan, A., Kumar, A., Delon, I., Wakeling, M., Mallin, L., Agrawal, S., Austin, T., Walston, F., Park, S.-M., Parker, A., Piyasena, C., Bradbury, K., Ellard, S., Rowitch, D. H., & Raymond, F. L. (2022). Refinements and considerations for trio whole-genome sequence analysis when investigating Mendelian diseases presenting in early childhood. *Human Genetics and Genomics Advances*, *3*(3), 100113. https://doi.org/10.1016/j.xhgg.2022.100113

Genin, E., Feingold, J., & Clerget-Darpoux, F. (2008). Identifying modifier genes of monogenic disease: Strategies and difficulties. *Human Genetics*, *124*(4), 357–368. https://doi.org/10.1007/s00439-008-0560-2

Geurts van Kessel, A., Venkatachalam, R., & Kuiper, R. P. (2013). Genomic and Personalized Medicine, Chapter 62—Colorectal Cancer. In G. S. Ginsburg & H. F. Willard (Eds.), *Genomic and Personalized Medicine (Second Edition)* (pp. 722–732). Academic Press. https://doi.org/10.1016/B978-0-12-382227-7.00062-8

Grubenbecher, S., Stüve, O., Hefter, H., & Korth, C. (2006). Prion protein gene codon 129 modulates clinical course of neurological Wilson disease. *NeuroReport*, *17*(5), 549. https://doi.org/10.1097/01.wnr.0000209006.48105.90

Harris, Z. L., Durley, A. P., Man, T. K., & Gitlin, J. D. (1999). Targeted gene disruption reveals an essential role for ceruloplasmin in cellular iron efflux. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(19), 10812–10817. https://doi.org/10.1073/pnas.96.19.10812

Jagannathan, V., Drögemüller, C., Leeb, T., & Consortium (DBVDC), D. B. V. D. (2019). A comprehensive biomedical variant catalogue based on whole genome sequences of 582 dogs and eight wolves. *Animal Genetics*, *50*(6), 695–704. https://doi.org/10.1111/age.12834

Kanzi, A. M., San, J. E., Chimukangara, B., Wilkinson, E., Fish, M., Ramsuran, V., & de Oliveira, T. (2020). Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. *Frontiers in Genetics*, *11*. https://www.frontiersin.org/articles/10.3389/fgene.2020.544162

Kaukonen, M., Woods, S., Ahonen, S., Lemberg, S., Hellman, M., Hytönen, M. K., Permi, P., Glaser, T., & Lohi, H. (2018). Maternal Inheritance of a Recessive RBP4 Defect in Canine Congenital Eye Disease. *Cell Reports*, *23*(9), 2643–2652. https://doi.org/10.1016/j.celrep.2018.04.118

Kenney, S. M., & Cox, D. W. (2007). Sequence variation database for the Wilson disease copper transporter, ATP7B. *Human Mutation*, *28*(12), 1171–1177. https://doi.org/10.1002/humu.20586

Kim, Y. G., Kim, S. Y., Kim, J. H., Lee, K. K., & Yun, Y. M. (2016). Prevalence and Clinical Relevance of Exon 2 Deletion of COMMD1 in Bedlington Terriers in Korea. *Journal of Veterinary Internal Medicine*, *30*(6), 1846–1850. https://doi.org/10.1111/jvim.14590

Klomp, A. E. M., van de Sluis, B., Klomp, L. W. J., & Wijmenga, C. (2003). The ubiquitously expressed MURR1 protein is absent in canine copper toxicosis. *Journal of Hepatology*, *39*(5), 703–709. https://doi.org/10.1016/S0168-8278(03)00380-5

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Lee, S.-A., Fillipich, L. J., & Hyun, C. (2007). Prevalence of the exon 2 deletion of the COMMD1 gene in Australian Bedlington terriers. *Journal of Genetics*, *86*(3), 289–291. https://doi.org/10.1007/s12041-007-0039-2

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Linder, M. C., Wooten, L., Cerveza, P., Cotton, S., Shulze, R., & Lomeli, N. (1998). Copper transport. *The American Journal of Clinical Nutrition*, *67*(5 Suppl), 965S-971S. https://doi.org/10.1093/ajcn/67.5.965S

Mak, C. M., Lam, C.-W., Tam, S., Lai, C.-L., Chan, L.-Y., Fan, S.-T., Lau, Y.-L., Lai, J.-Y., Yuen, P., Hui, J., Fu, C.-C., Wong, K.-S., Mak, W.-L., Tze, K., Tong, S.-F., Lau, A.,

Leung, N., Hui, A., Cheung, K.-M., … Chan, Y.-W. (2008). Mutational analysis of 65 Wilson disease patients in Hong Kong Chinese: Identification of 17 novel mutations and its genetic heterogeneity. *Journal of Human Genetics*, *53*(1), Article 1. https://doi.org/10.1007/s10038-007-0218-2

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Medici, V., & LaSalle, J. M. (2019). Genetics and epigenetic factors of Wilson disease. *Annals of Translational Medicine*, *7*(Suppl 2), Article Suppl 2. https://doi.org/10.21037/atm.2019.01.67

Mercer, J. F. B., Livingston, J., Hall, B., Paynter, J. A., Begy, C., Chandrasekharappa, S., Lockhart, P., Grimes, A., Bhave, M., Siemieniak, D., & Glover, T. W. (1993). Isolation of a partial candidate gene for Menkes disease by positional cloning. *Nature Genetics*, *3*(1), Article 1. https://doi.org/10.1038/ng0193-20

Mistry, M., Ho Sui, S., Liu, J., Piper, M., Gammerdinger, W., & Khetani, R. (2021, September 15). *Troubleshooting the ChIP-seq workflow: From sequence reads to peak calls*. Introduction to ChIP-Seq Using High-Performance Computing. https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/troubleshooting_chipseq_partI.html

Nakaichi, M., Iseri, T., Horikirizono, H., Itoh, H., Sunahara, H., Nemoto, Y., Itamoto, K., & Tani, K. (2022). Pedigree study of the heredity of copper-associated hepatitis in Dalmatians in Japan. *The Canadian Veterinary Journal*, *63*(6), 633–636.

Nakaichi, M., Iseri, T., Horikirizono, H., Komine, M., Itoh, H., Sunahara, H., Nemoto, Y., Itamoto, K., & Tani, K. (2021). Copper-associated hepatitis in a young Dalmatian dog in Japan. *The Journal of Veterinary Medical Science*, *83*(6), 911–915. https://doi.org/10.1292/jvms.21-0061

Nicastro, E., Ranucci, G., Vajro, P., Vegnente, A., & Iorio, R. (2010). Re-evaluation of the diagnostic criteria for Wilson disease in children with mild liver disease. *Hepatology*, *52*(6), 1948–1956. https://doi.org/10.1002/hep.23910

Park, H.-D., Park, H.-K., Chung, H.-S., Lee, S.-Y., Kim, J.-W., & Ki, C.-S. (2010). Association of ATP7BMutation Detection Rate with Biochemical Characteristics in Korean Patients with Wilson Disease. *Annals of Clinical & Laboratory Science*, *40*(1), 15–19.

Pedersen, B. S., Layer, R., & Quinlan, A. R. (2020). *smoove: Structural-variant calling and genotyping with existing tools* (0.2.8) [Go]. https://github.com/brentp/smoove/blob/c68ae05317dc5cedcc848b0389fd1a6b9d0fb92e/CITATION.cff (Original work published 2018)

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics (Oxford, England)*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699

Pedersen, B. S., & Quinlan, A. R. (2019). Duphold: Scalable, depth-based annotation and curation of high-confidence structural variant calls. *GigaScience*, *8*(4), giz040. https://doi.org/10.1093/gigascience/giz040

Rimoin, D. L., Pyeritz, R. E., & Korf, B. (2013). *Emery and Rimoin's essential medical genetics*. Elsevier.

Scheinberg, I. H., & Sternlieb, I. (1965). Wilson's disease. *Annual Review of Medicine*, *16*(1), 119–134.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews. Genetics*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

Tanzi, R. E., Petrukhin, K., Chernov, I., Pellequer, J. L., Wasco, W., Ross, B., Romano, D. M., Parano, E., Pavone, L., Brzustowicz, L. M., Devoto, M., Peppercorn, J., Bush, A. I., Sternlieb, I., Pirastu, M., Gusella, J. F., Evgrafov, O., Penchaszadeh, G. K., Honig, B., … Gilliam, T. C. (1993). The Wilson disease gene is a copper transporting ATPase with homology to the Menkes disease gene. *Nature Genetics*, *5*(4), Article 4. https://doi.org/10.1038/ng1293-344

Teare, M. D., & Santibañez Koref, M. F. (2014). Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing. *Briefings in Functional Genomics*, *13*(5), 378–383. https://doi.org/10.1093/bfgp/elu024

Thornburg, L. P. (1998). Histomorphological and immunohistochemical studies of chronic active hepatitis in Doberman Pinschers. *Veterinary Pathology*, *35*(5), 380–385. https://doi.org/10.1177/030098589803500507

Thornburg, L. P., Rottinghaus, G., McGowan, M., Kupka, K., Crawford, S., & Forbes, S. (1990). Hepatic copper concentrations in purebred and mixed-breed dogs. *Veterinary Pathology*, *27*(2), 81–88. https://doi.org/10.1177/030098589002700202

Thornburg, L. P., Shaw, D., Dolan, M., Raisbeck, M., Crawford, S., Dennis, G. L., & Olwin, D. B. (1986). Hereditary copper toxicosis in West Highland white terriers. *Veterinary Pathology*, *23*(2), 148–154. https://doi.org/10.1177/030098588602300207

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Twedt, D. C., Sternlieb, I., & Gilbertson, S. R. (1979). Clinical, morphologic, and chemical studies on copper toxicosis of Bedlington Terriers. *Journal of the American Veterinary Medical Association*, *175*(3), 269–275.

van De Sluis, B., Rothuizen, J., Pearson, P. L., van Oost, B. A., & Wijmenga, C. (2002). Identification of a new copper metabolism gene by positional cloning in a purebred dog population. *Human Molecular Genetics*, *11*(2), 165–173. https://doi.org/10.1093/hmg/11.2.165

Vulpe, C., Levinson, B., Whitney, S., Packman, S., & Gitschier, J. (1993). Isolation of a candidate gene for Menkes disease and evidence that it encodes a copper-transporting ATPase. *Nature Genetics*, *3*(1), 7–13. https://doi.org/10.1038/ng0193-7

Webb, C. B., Twedt, D. C., & Meyer, D. J. (2002). Copper-associated liver disease in Dalmatians: A review of 10 dogs (1998–2001). *Journal of Veterinary Internal Medicine*, *16*(6), 665–668.

Weiskirchen, R., & Penning, L. C. (2021). COMMD1, a multi-potent intracellular protein involved in copper homeostasis, protein trafficking, inflammation, and cancer. *Journal of Trace Elements in Medicine and Biology*, *65*, 126712. https://doi.org/10.1016/j.jtemb.2021.126712

Wu, X., Chien, H., van Wolferen, M. E., Kruitwagen, H. S., Oosterhoff, L. A., & Penning, L. C. (2019). Reduced FXR Target Gene Expression in Copper-Laden Livers of COMMD1-Deficient Dogs. *Veterinary Sciences*, *6*(4), Article 4. https://doi.org/10.3390/vetsci6040078

Wu, X., den Boer, E. R., Vos-Loohuis, M., van Steenbeek, F. G., Monroe, G. R., Nijman, I. J., Leegwater, Peter. A. J., & Fieten, H. (2020). Investigation of Genetic Modifiers of Copper Toxicosis in Labrador Retrievers. *Life*, *10*(11), 266. https://doi.org/10.3390/life10110266

Wu, X., Mandigers, P. J. J., Fieten, H., & Leegwater, P. A. (2020). Evaluation of COMMD1 in copper toxicosis in Labrador retrievers and Dobermans. *Veterinary Journal (London, England: 1997)*, *265*, 105561. https://doi.org/10.1016/j.tvjl.2020.105561

Wu, X., Mandigers, P. J. J., Watson, A. L., van den Ingh, T. S. G. A. M., Leegwater, P. A. J., & Fieten, H. (2019). Association of the canine ATP7A and ATP7B with hepatic copper accumulation in Dobermann dogs. *Journal of Veterinary Internal Medicine*, *33*(4), 1646–1652. https://doi.org/10.1111/jvim.15536

# CHAPTER 4. Detection of Candidate Recessive Lethal Mutations Contributing to Low Fertility in Beef Cattle Using Whole-Genome Sequencing (WGS)

## 4.1. Introduction and Literature Review

The condition known as suboptimal fertility/infertility refers to a decreased or absent capacity to produce progeny that can grow and function as healthy and productive individuals and it leads to significant economic damage to producers. According to the Alberta Cow-Calf Audit (1986-88, 1997-1998), the culling rate in Alberta stands at around 10 to 11% (SD=10%) of the cow herd annually (Alberta Cow-Calf Audit, 2001). Additionally, findings from the 2nd Western Canadian Cow-Calf Survey (WCCCS II aggregate report, 2017) reveal that the average culling rate specifically for breeding females during Fall/Winter 2017/18 was documented at 11.7% (2nd Western Canadian Cow-Calf Survey, 2018). With prevailing weaned calf prices reported at $3.60/lb for a 500 lb calf (Alberta Beef Producers, 2023), each infertile cow represents a loss of $1,800 in income for the cow-calf producer. Additionally, the replacement of an infertile cow incurs a cost of approximately $900 per replacement heifer (factoring in the cost of developing the replacement heifer minus the income from selling the culled cow). Consequently, an incidence rate of 11% for infertile cows (considering an average culling rate of 11%), equivalent to 11 cows in a 100-head herd, could potentially result in a total economic loss of $29,700 annually.

Generally, cattle breeding programs aim is to have 90-95% of their cows and replacement heifers bred in an average 65-day breeding season (Navarre, 2010). However, achieving this goal may encounter challenges due to various issues, including reproductive loss, bull infertility, inadequate nutrition pre-breeding season, drought, improper development of replacement heifers, low heterosis, inadequate vaccination program for reproductive diseases.

Given the low heritability of fertility traits, a significant proportion of the phenotypic variance in this trait is attributed to non-additive genetic effects rather than additive genetic effects.

This phenomenon arises from the unique combination of genes inherited by an offspring through genetic recombination and Mendelian sampling. Notably, heterosis, a non-additive genetic effect, significantly influences female fertility, longevity, and lifetime productivity (Basarab et al., 2018). The underlying biological mechanisms linked to these advantageous outcomes are believed to be associated with dominance and epistasis, which denote interactions between genes.

Reproductive loss pertains to a decrease in the reproductive efficiency of a cattle herd, resulting in a diminished number of offspring being born. Pinpointing the exact cause in each instance can be challenging, given that reproductive loss may arise from various factors, including infertility, abortion, stillbirths, and neonatal mortality (Waldner, 2019). The impact of these losses on the profitability of a cattle operation can be significant, as they result in fewer calves being born (Daly, 2021; Waldner, 2019).

Causes of infertility/suboptimal fertility can be categorized into two groups: infectious and non-infectious (Navarre, 2010). Non-infectious causes of infertility in female cattle are nutrition, stress and genetics and in male cattle are failing to use a bull that has passed a breeding soundness exam, insufficient bull numbers for cow numbers and environment, and bull(s) with poor libido (Navarre, 2010). Inadequate nutrition for cows is one of the leading factors that can cause infertility in beef herds. Although overfeeding, particularly in heifers, can also be problematic, it is not as prevalent. Infertility can be influenced by the cow's body condition prior to calving, after calving, and at breeding. For example, cows that are thin when they calve but gain weight during breeding are more likely to conceive than cows that remain thin. However, cows that maintain optimal body conditions during pregnancy are more likely to become pregnant again (Navarre, 2010).

Stress is another significant risk factor for pregnancy loss during any gestation phase in cattle. Therefore, practicing low-stress handling techniques while processing cattle is crucial for minimizing this issue. Heat stress can also cause infertility, particularly in the initial stages of pregnancy. High humidity can further aggravate heat stress, resulting in a more detrimental effect on conception rates (Navarre, 2010).

The presence of recessive lethal mutations is one of the genetic factors that can contribute to infertility or suboptimal fertility in cattle. When two carriers (heterozygotes) possessing a recessive lethal mutation mate, offspring homozygous for the lethal mutation are not viable and depending on when the lethality occurs this cause of reduced fertility may lead to early embryonic loss, stillbirth, or calf death. A carrier animal will appear normal phenotypically, making it difficult to address the issue through traditional breeding techniques without knowledge of the allele and a means to detect its presence.

Several studies have identified genomic regions that appear to harbor recessive lethal variants, in a variety of breeds including Ayrshires (T. A. Cooper et al., 2014), Brown Swiss (Schwarzenbacher et al., 2012; VanRaden et al., 2011), Holstein (Adams et al., 2016; Agerholm et al., 2006; Charlier et al., 2012; Fritz et al., 2013; Sahana et al., 2013; VanRaden et al., 2011), Jersey (Sonstegard et al., 2013; VanRaden et al., 2011), Montbéliarde (Fritz et al., 2013), and Normande breed (Fritz et al., 2013). Using the above studies results, Valour et al. (2015) in a review article concluded that an overall significant negative impact on the calving rate was observed for 20 haplotypes. Additional studies were conducted for the characterization of six of those areas using whole-genome sequencing data obtained from carrier bulls. These studies identified mutations believed to be responsible for the lethality, in the genes *FANCI* (Brachyspina),

*APAF1* (haplotype H1), *SMC2* (HH3), *GART* (HH4), *CWC15* (JH1), and *SLC37A2* (MH1) (Valour et al., 2015).

VanRaden et al. (2011) found five new lethal recessives by investigating haplotypes with high frequencies among three cattle breeds, including Holstein, Jersey, and Brown Swiss, but they were not observed in homozygous form (VanRaden et al., 2011). Cooper et al (2014) identified a haplotype that affects Ayrshire fertility and traces back in the genotyped population to the bull Selwood Betty's Commander (born in 1953). The haplotype carrier frequency for genotyped Ayrshires was 26.1%. That haplotype, which was designated as AH1, was expected to be homozygous in 14 animals (56 mating of carrier sire by carrier maternal grandsire (MGS) divided by 4). However, none were observed in the genotyped population (T. A. Cooper et al., 2014). In a study by Hoff et al (2017), seven loci that potentially possess recessive lethal haplotypes were found but none of them were in homozygous form despite relatively high expectation of homozygote occurrence due to an adequately high frequency and pedigree-based prediction. They proposed these haplotypes as candidates for carrying autosomal recessive lethal alleles. The direct phenotypic effects of these haplotypes had not been noticed, but could be inconspicuous such as in the event of early embryonic loss or may be unreported defects leading to the loss of the calf (Hoff et al., 2017).

In this study we evaluated an alternative approach for detecting candidate recessive lethal mutations: whole-genome sequencing of 20 beef cattle selected among those that had demonstrated a high infertility/subfertility rate (more than 15% open rate). The rationale for this approach is that WGS provides a near-complete view of the genome that, coupled with bioinformatic tools, can reveal mutations that have a large impact on gene function.

## 4.2. Research Objectives

a) Identification and genotyping of SNPs and indels based on whole-genome sequences of the nine selected beef cows and 11 AI sires.
b) Use of bioinformatic strategies to identify candidate recessive lethal variants among the identified variants.

## 4.3. Methods/Experimental Design and Materials

### 4.3.1. Sampling and whole-genome sequencing (WGS)

The samples used in this study were collected under the RFI Methane project, from the beef cattle herd at the Lacombe Research and Development Centre (LRDC), under the supervision of PI John Basarab, following the University of Alberta Animal Care and Use Protocol (AUP00001140) approved by the LRDC Animal Care Committee. Furthermore, the study adhered to the guidelines established by the Canadian Council on Animal Care (CCAC) concerning the care and utilization of farm animals in research, education, and evaluation (Canadian Council on Animal Care, 2009).

Beef cattle were chosen among those that had demonstrated a high infertility rate (more than 15% open rate). For this matter, 11 mating groups of 15–49 crossbred cows were exposed to purebred black Angus, Red Angus, and black Simmental AI sires during the breeding season of 2017 at the LRDC. The open rate among these mating groups varied from 5 to 22%. Therefore, three criteria were considered for selecting the animals for sequencing: 1) Selecting cows from the mating groups was based on the highest open rate for the 2017 breeding season. 2) Only the

youngest open cows were selected. 3) Bulls common to mating groups with the highest open rates were selected. In total 20 animals, including nine cows and 11 AI sires, were selected for sequencing.

Genomic DNA samples from the 20 selected animals were sent to the Génome Québec Innovation Centre for Illumina HiSeqX PE150 whole-genome sequencing (with 20x sequencing depth). The resulting FASTQ files were used for variant calling and genotyping.

### 4.3.2. SNPs and indels identification and genotyping through variant calling

SNPs and indels calling was performed using a custom pipeline written by Dr. Emily Herman (University of Alberta) using the Snakemake workflow management system (Köster & Rahmann, 2012; Mölder et al., 2021). The workflow was run on a high-performance computing cluster (cedar) accessed through Compute Canada. The pipeline steps can be classified into two main procedures: read mapping and variant calling. Additional information about this pipeline is available from the GitHub repository: https://github.com/stothard-group/variant-calling-pipeline.

#### 4.3.2.1. Read mapping

For sequence read quality assessment, the FastQC v0.11.9 program (Andrews, 2010) was run on all FASTQ files. Trimmomatic v0.36 (Bolger et al., 2014) was used for quality trimming. For mapping the trimmed reads against the reference genome of Bos taurus (ARS-UCD1.2_Btau5.0.1Y), the BWA v 0.7.17 (H. Li & Durbin, 2009) software package was used. To coordinate-sort and to index the created SAM files from the previous step to BAM files, the Picard v2.18 (Broad Institute, 2018) SortSam tool was used. Picard v2.18 MarkDuplicates (Broad

Institute, 2018) was run to aggregate the BAM files from the same sample and mark suspected read duplicates. Base quality recalibration of BAM files was performed using GATK v.3.8 (O'Connor & Van Auwera, 2020) PrintReads tool to create recalibrated bam files.

#### 4.3.2.1.1. Read counts, mapping rates and sequencing coverage/depth

Read counts, mapping rates and average genome coverage/depth were calculated using the BAM files generated for each sample through the read mapping process, along with SAMtools v1.17 (H. Li et al., 2009) and htslib v1.17 (H. Li & Durbin, 2009). The number of sequencing reads generated from a sample is referred to as the "read count." The mapping rate indicates the proportion of reads correctly aligned to the reference genome (Y. Chen et al., 2021; Trapnell et al., 2009). Average genome coverage, also referred to as average genome depth, denotes the mean frequency with which a base in the genome is sequenced.

### 4.3.2.2. Variant calling

The HaplotypeCaller tool from GATK v.3.8 (O'Connor & Van Auwera, 2020), was run on contigs and chromosome together with the recalibrated BAM files generated in the "Read Mapping" procedure. The HaplotypeCaller calls SNPs and indels concurrently through local re-assembly of haplotypes. For genotyping chromosomes and contigs, the GenotypeVCFs tool from the GATK v.3.8 (O'Connor & Van Auwera, 2020) toolkit, was run on the created '.g.vcf' files for each chromosome and the combined contigs, per sample. For extracting SNPs and indels from each chromosome and contigs, per sample, the GATK v.3.8 (O'Connor & Van Auwera, 2020) SelectVariants tool was used. Filtering SNPs and indels from chromosomes and contigs, was

performed using the VariantFiltration tool from the GATK v.3.8 (O'Connor & Van Auwera, 2020) toolkit. Finally, for concatenating and combining chromosome and contig SNPs and indels VCF files, the GatherVcfs (*GatherVcfs (Picard)*, 2019) was used.

### 4.3.3. Assessments of sample's sex chromosomes for quality control of the samples

As part of routine screening for sample handling errors, the sex chromosome complement of samples was assessed using SNP genotypes. Two approaches were applied to high-confidence SNPs (SNPs with a FILTER value of "PASS") to differentiate XX and XY samples: X-chromosome heterozygosity and Y-chromosome missing genotype proportion. The first metric is expected to be higher in XX animals and the latter lower in XY animals. VCFtools v0.1.16 (Danecek et al., 2011) was used to calculate the proportion of heterozygous X-chromosome SNPs in each sample. The Y chromosome missing genotype proportion was calculated using BCFtools v1.8 query command (Danecek et al., 2021)

### 4.3.4. Identification of candidate recessive lethals using variant filtering

To identify candidate recessive lethals related to suboptimal fertility/infertility in cattle, a four-step filtering process was employed to narrow down the large number of variants discovered during the first objective. This process was implemented to identify the most credible candidate lethal haplotypes.

The primary objective of the first filtering step, "Filter 1," was to eliminate sites that had a "FILTER" flag other than "PASS." The filter set defined in the variant calling pipeline employed in this study, which were based on the recommendations of the 1000 Bull Genomes Project, was

utilized for this purpose. The primary purpose of applying this filter was to remove low-quality variants that lacked adequate supporting evidence, thereby reducing the number of false-positive variants in the dataset.

The purpose of the second filtering step, called "Filter 2," was to remove sites where a homozygous non-reference genotype was observed. The reason for applying this filter was that any variants observed in homozygous forms are not reasonable recessive lethal candidate mutations since phenotypically normal cattle were sequenced to generate the VCF files for analysis. However, one challenge with this step was that a site could receive a homozygous genotype in an individual due to low read depth, leading to potential false positives. Therefore, sites that exhibited at least one homozygous non-reference genotype and where that genotype was of good quality (GQ > 10), according to the VCF file, were removed. "GQ" refers to "genotype quality," a measure of confidence in the genotype call at a particular site. Thus, a higher GQ score indicates a higher confidence level in the genotype call. GQ is calculated by the variant calling software based on the read depth and quality of the reads supporting each allele.

The third filtering step, called "Filter 3", was done using the 1000 Bull Genomes Project dataset (Hayes & Daetwyler, 2019). This project aims to provide the bovine research community with an extensive database of genotypes for the imputation of genetic variants for genomic prediction and genome-wide association studies in all cattle breeds. In this step, variants that were homozygous non-reference (1/1) in any animal in the 1000 Bulls data set were removed from the 20-animal data set, with the rationale that such variants are not good candidates as recessive lethals.

For the next filtering step, called "Filter 4," initially, the remaining variants were annotated with predicted functional impacts and then those with high functional impacts were kept. SNPs

and indels were annotated with functional impact predictions using SnpEff v5.0e (Cingolani, Platts, et al., 2012) and the ARS-UCD1.2.99 SnpEff annotation database.

### 4.3.5. Allele frequency (AF) of alternative alleles for candidate causal SNPs and indels

The allele frequency (AF) signifies the proportion or occurrence frequency of the alternate allele in the population/dataset at a specific genomic position. AF for the remaining candidate recessive lethal variants was assessed in the sequenced animals, using INFO tags in the VCF file, to further assess the plausibility of each recessive lethal candidate.

## 4.4. Results

The current research involved analyzing 20 animal's genomes to detect potential recessive lethal mutations associated with subfertility/infertility in beef cattle using two variant calling pipelines. The alignment of sequence reads from the 20 sequenced individuals resulted in an average genome coverage of 18.43 and a read mapping rate of 97.71% (details in table 4.1). These values fall within the recommended range for variant discovery and genotyping, as suggested by Mistry et al. (2021) and koboldt (2020) (Koboldt, 2020; Mistry et al., 2021).

**Table 4.1.** Average genome coverage and read mapping rates of the studied samples.

| Sample | Read counts | Mapping rates (%) | Average genome coverage |
|---|---|---|---|
| C1 | 388,304,443 | 99.70 | 18.57 |
| C2 | 367,379,344 | 99.66 | 16.87 |
| C3 | 401,869,604 | 99.64 | 19.41 |
| B1 | 363,255,910 | 99.34 | 16.87 |
| C4 | 413,252,308 | 97.84 | 19.37 |
| C5 | 457,212,533 | 96.98 | 21.19 |
| C6 | 551,415,804 | 99.68 | 26.28 |
| C7 | 452,295,797 | 99.54 | 21.73 |
| C8 | 437,641,238 | 99.56 | 21.08 |
| C9 | 364,415,007 | 99.34 | 17.50 |
| B2 | 387,655,892 | 99.64 | 18.22 |
| B3 | 337,508,516 | 99.76 | 16.00 |
| B4 | 407,098,910 | 99.68 | 19.05 |
| B5 | 401,277,740 | 99.77 | 18.97 |
| B6 | 344,519,847 | 99.66 | 16.31 |
| B7 | 448,596,000 | 99.27 | 20.87 |
| B8 | 385,603,318 | 99.49 | 17.85 |
| B9 | 468,042,708 | 66.36 | 12.29 |
| B10 | 33,2951,159 | 99.42 | 15.61 |
| B11 | 307,258,012 | 99.77 | 14.49 |
| **Mean** | **400,877,705** | **97.71** | **18.43** |

The findings from the analysis of the sample's sex chromosomes, including X-chromosome heterozygosity and Y-chromosome missing genotype proportion, were consistent with what was anticipated based on the supplied sample information (data not shown).

Following variant calling and sex chromosome assessment, various filters were applied in an attempt to identify a final set of plausible recessive lethal candidate mutations. The successive filter steps reduced the initial set of 17,255,921 SNPs and 2,590,194 indels, to a final set of 391 SNPs, and 3,527 indels, (Table 4.2).

**Table 4.2.** The number of retained SNPs and indels after each step of filtering.

| Filters | Description | Number of retained SNPs after filtering steps | Number of retained indels after filtering steps |
|---|---|---|---|
| **Variant calling** | This shows the initial numbers of SNPs and indels obtained using variant calling process. | 17,255,921 | 2,590,194 |
| **Filter 1** | This filter is for removing low-quality variants (sites that had a "FILTER" flag other than "PASS.") | 16,076,677 | 2,521,814 |
| **Filter 2** | This filter is for removing sites where a homozygous non-reference genotype was observed | 7,734,915 | 1,133,491 |
| **Filter 3** | This filter is for removing variants from the 20-animal data set that were homozygous non-reference (1/1) in any animal in the 1000 Bulls data set | 1,110,503 | 400,760 |
| **Filter 4** | This filter is for keeping variants with high functional impacts | 391 | 3,527 |

Each filtering scenario is characterized by a specific quality criterion, genotype pattern, or functional annotation that results in variants successfully passing the filter, thereby being identified as potential causal mutations. As a result of implementing four filtering steps in this research study, 391 SNPs and 3,527 indels with "HIGH" functional impacts were obtained which are considered potential causative variants. These variants are available at https://doi.org/10.5683/SP3/MZTWVP.

The frequency of the putative causal alleles was assessed in the 20 sequenced animals, the rationale being that true recessive lethal alleles should be rare, and that alleles appearing in several animals are unlikely to be causative alleles. A large portion of the candidate causal SNP alleles (43%) appeared just once in the sequenced animals (figure 4.1), whereas 79% of the candidate cause indel alleles were observed once (figure 4.2).

**Table 4.3.** Allele count (AC) and allele frequency (AF) of genotypes for each alternative allele (ALT) in the 391 candidate causal SNPs.

| Allele count (AC) in genotypes for each ALT allele in the VCF file | Allele frequency (AF) for each ALT allele in the VCF file | The number of locations/sites (variant counts) in the candidate SNPs |
|---|---|---|
| AC= 1 | AF=0.025 | 169 |
| AC= 2 | AF=0.05 | 63 |
| AC= 3 | AF=0.075 | 35 |
| AC= 3,2 | AF=0.075,0.05 | 1 |
| AC= 3,6 | AF=0.075,0.15 | 1 |
| AC= 4,4 | AF=0.111,0.111 | 1 |
| AC= 4 | AF=0.1 | 21 |
| AC= 5,2 | AF=0.125,0.05 | 1 |
| AC= 5 | AF=0.125 | 21 |
| AC= 6 | AF=0.15 | 14 |
| AC= 7,3 | AF=0.194,0.083 | 1 |
| AC= 7 | AF=0.175 | 12 |
| AC= 8, 1 | AF=0.2,0.025 | 1 |
| AC= 8 | AF=0.2 | 4 |
| AC= 9 | AF=0.225 | 7 |
| AC= 10 | AF=0.25 | 3 |
| AC= 11 | AF=0.275 | 4 |
| AC= 12 | AF=0.3 | 6 |
| AC= 13 | AF=0.325 | 7 |
| AC= 14 | AF=0.35 | 4 |
| AC= 15 | AF=0.375 | 1 |
| AC= 16 | AF=0.4 | 1 |
| AC= 17 | AF=0.425 | 3 |
| AC= 18,2 | AF=0.45,0.05 | 1 |
| AC= 18 | AF=0.45 | 2 |
| AC= 19 | AF=0.475 | 1 |
| AC= 20 | AF=0.5 | 5 |
| AC= 21 | AF=0.525 | 1 |

AC and AF for the final candidate recessive lethal SNPs was assessed in the sequenced animals, using INFO tags in the VCF file.

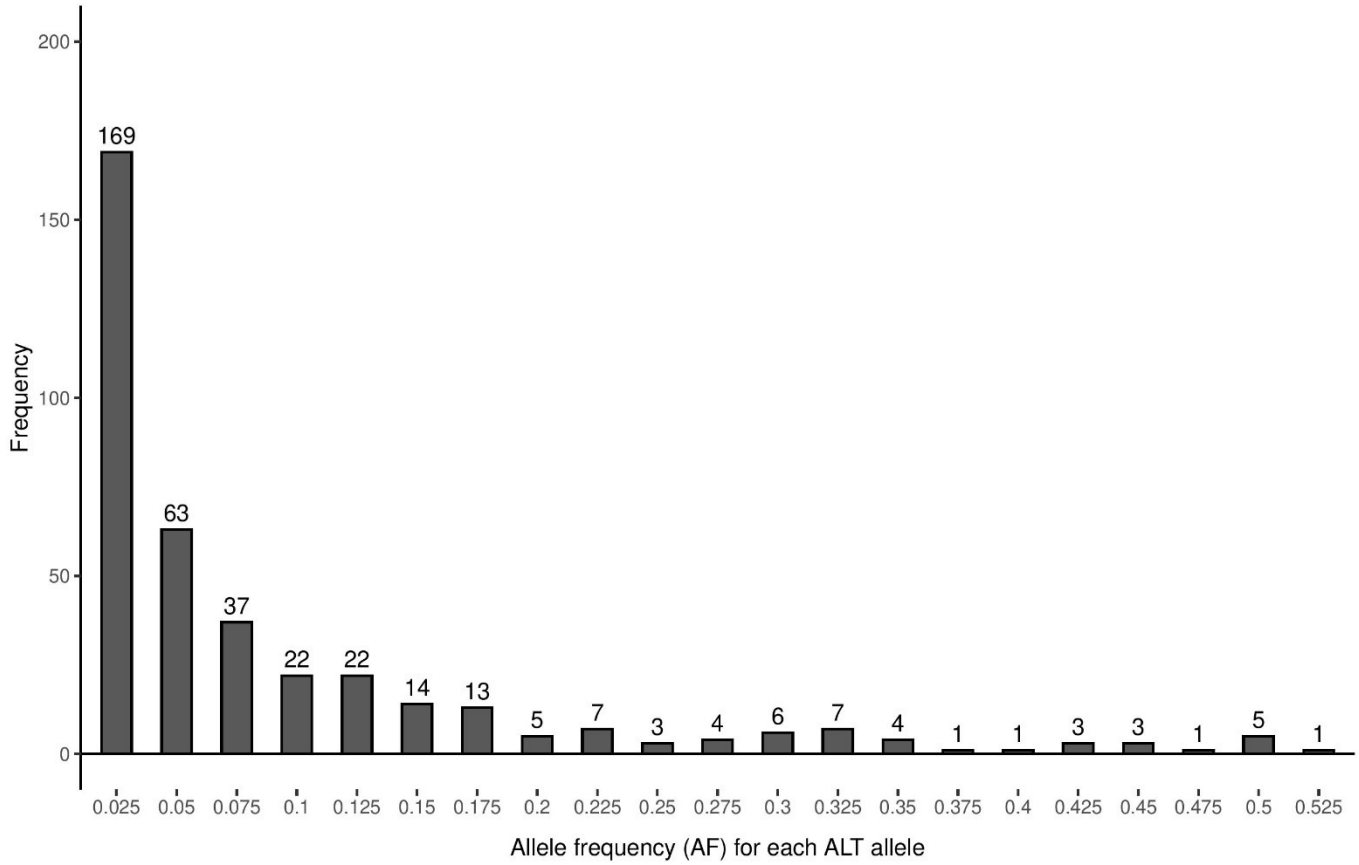


**Figure 4.1.** Prevalence of the various ALT allele frequencies (AF) of the candidate causal SNPs.

**Table 4.4.** Allele count (AC) and allele frequency (AF) of genotypes for each alternate allele (ALT) in the 3,527 candidate causal indels.

| Allele count (AC) in genotypes for each ALT allele in the same order as in the VCF file | Allele frequency (AF) for each ALT allele in the same order as in the VCF file | The number of locations (variant counts) in the candidate indels |
|---|---|---|
| AC=1,1,1 | AF=0.025,0.025,0.025 | 1 |
| AC=1,1 | AF=0.025,0.025 | 22 |
| AC=1,2 | AF=0.025,0.05 | 1 |
| AC=1,5 | AF=0.026,0.132 | 1 |
| AC= 1 | AF=0.025 | 2787 |
| AC=2,1,1 | AF=0.05,0.025,0.025 | 3 |
| AC=2,1 | AF=0.05,0.025 | 6 |
| AC=2,2 | AF=0.05,0.05 | 3 |
| AC=2,3 | AF=0.053,0.079 | 1 |
| AC=2,6 | AF=0.05,0.15 | 1 |
| AC=2,8 | AF=0.05,0.2 | 1 |
| AC= 2 | AF=0.05 | 224 |
| AC=3,1 | AF=0.075,0.025 | 3 |
| AC=3,2 | AF=0.075,0.05 | 2 |
| AC=3,4 | AF=0.075,0.1 | 1 |
| AC=3,5 | AF=0.083,0.139 | 1 |
| AC=3,7 | AF=0.083,0.194 | 1 |
| AC=3,9,1 | AF=0.075,0.225,0.025 | 1 |
| AC= 3 | AF=0.075 | 129 |
| AC=4,1 | AF=0.1,0.025 | 2 |
| AC= 4 | AF=0.1 | 89 |
| AC=5,1 | AF=0.132,0.026 | 1 |
| AC=5,3 | AF=0.139,0.083 | 1 |
| AC=5,5,2,5,2,1 | AF=0.125,0.125,0.05,0.125,0.05,0.025 | 1 |
| AC= 5 | AF=0.125 | 56 |
| AC=6,1,1 | AF=0.15,0.025,0.025 | 1 |
| AC=6,1 | AF=0.15,0.025 | 2 |
| AC=6,3 | AF=0.158,0.079 | 1 |
| AC=6,6 | AF=0.158,0.158 | 1 |
| AC= 6 | AF=0.15 | 33 |
| AC=7,1 | AF=0.175,0.025 | 1 |
| AC=7,4 | AF=0.175,0.1 | 1 |
| AC= 7 | AF=0.175 | 32 |
| AC=8,1 | AF=0.2,0.025 | 1 |
| AC=8,13 | AF=0.2,0.325 | 1 |
| AC=8,3 | AF=0.2,0.075 | 1 |
| AC=8,4 | AF=0.2,0.1 | 1 |
| AC=8,5 | AF=0.2,0.125 | 1 |
| AC= 8 | AF=0.2 | 14 |
| AC= 9 | AF=0.225 | 11 |
| AC=10,2 | AF=0.263,0.053 | 1 |
| AC=10,3 | AF=0.263,0.079 | 1 |

AC and AF for the final candidate recessive lethal indels was assessed in the sequenced animals, using INFO tags in the VCF file.

**Table 4.4.** Allele count (AC) and allele frequency (AF) of genotypes for each alternate allele (ALT) in the 3527 candidate causal indels (continued).

| Allele count (AC) in genotypes for each ALT allele in the same order as in the VCF file | Allele frequency (AF) for each ALT allele in the same order as in the VCF file | The number of locations (variant counts) in the candidate indels |
|---|---|---|
| AC=10,4 | AF=0.25,0.1 | 1 |
| AC=10,5 | AF=0.263,0.132 | 1 |
| AC= 10 | AF=0.25 | 24 |
| AC= 11 | AF=0.275 | 7 |
| AC= 12 | AF=0.3 | 8 |
| AC=13,2 | AF=0.342,0.053 | 1 |
| AC= 13 | AF=0.325 | 9 |
| AC= 14 | AF=0.35 | 8 |
| AC= 16 | AF=0.4 | 4 |
| AC= 17 | AF=0.425 | 2 |
| AC= 18 | AF=0.45 | 4 |
| AC= 19 | AF=0.475 | 4 |
| AC= 20 | AF=0.5 | 9 |
| AC= 21 | AF=0.525 | 2 |

In several cases, the Allele Counts (AC) display multiple values. This happens when various alternative alleles are observed at these specific positions. The naming convention for these alleles adheres to the order in which they are listed in the variant VCF file. For example, when AC=3,9,1, it denotes the presence of 3 instances of ALT allele type 1 (0/1 genotype), 9 instances of ALT allele type 2 (0/2 genotype), and 1 instance of ALT allele type 3 (0/3 genotype).

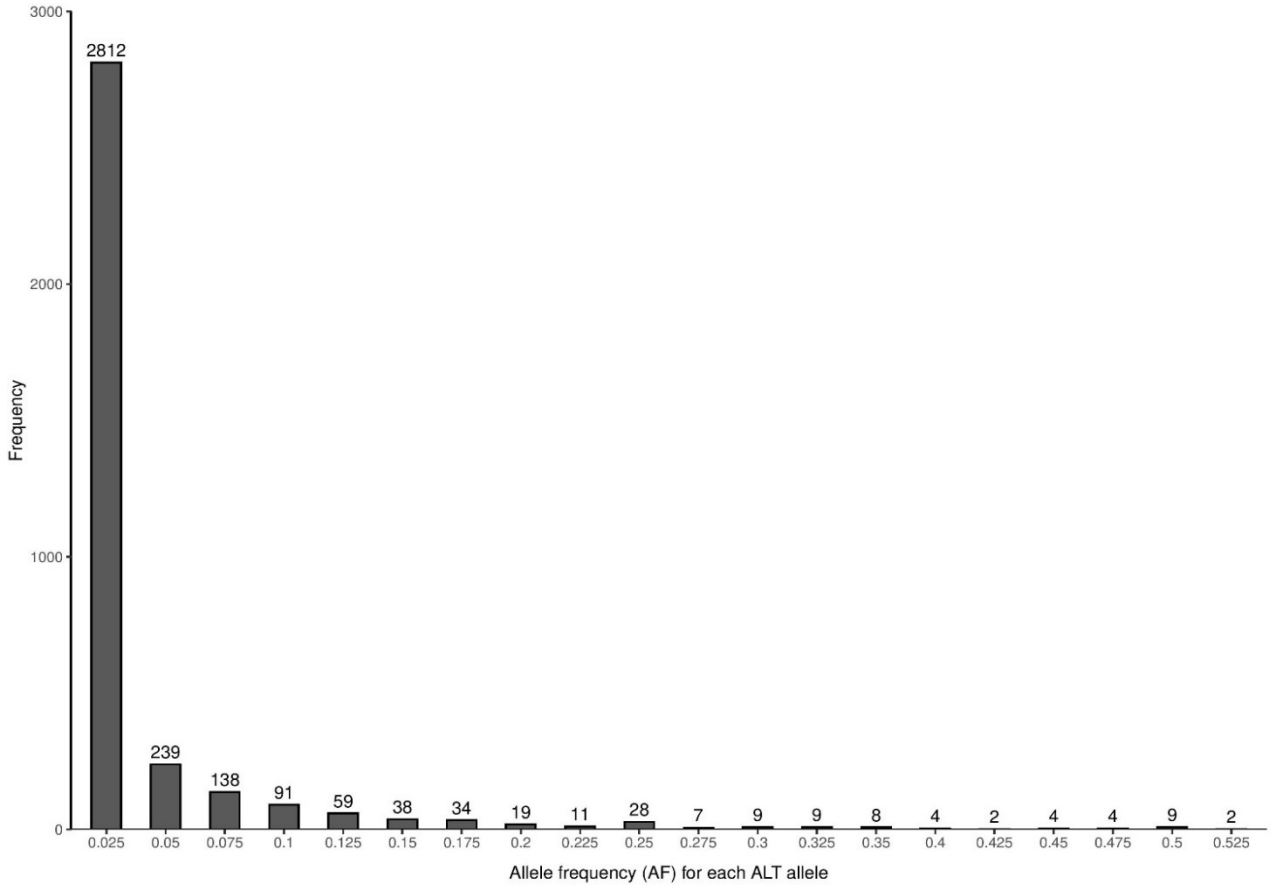


**Figure 4.2.** Prevalence of the various ALT allele frequencies (AF) of the candidate causal indels.

## 4.5. Discussion/Conclusions

In the present study, 20 cattle's genomes were analyzed through whole-genome sequencing (WGS), resulting in the identification of 391 candidate causal SNPs and 3,527 candidate causal indels with "HIGH" functional impacts that could be potentially related to subfertility/infertility in beef cattle. However, the large number of the detected candidate variants indicates that this approach may not be a sufficient method to pinpoint the specific true recessive lethal mutations. Some of the identified candidate recessive mutations were found to be at high frequency among the 20 sequenced individuals, suggesting that in the future additional filtering steps could be added to refine the candidate list further. Many of the candidate causal variants are likely variants of no functional significance or are false-positive variants arising from read alignment errors, which could stem from reference genome issues. The latter issue could explain variants that are found in nearly all animals but never in homozygous form.

It should be noted that in order to accurately identify the true recessive lethal alleles among the list of candidates found in this study, more information of the impact of the identified variant on the low fertility or the genes contributions to the low fertility would be required. It is also important to keep in mind that other factors besides recessive lethal alleles may also contribute to low fertility in these 20 animals. One of the factors contributing to low fertility could be the presence of alleles with negative effects on fertility, such as those leading to reduced sperm motility. For instance, a mutation in the *TMEM95* gene, situated on cattle chromosome 19, has been identified as a cause of complete infertility in bulls, resulting in a significantly low insemination success rate of 2% (Pausch et al., 2014; Steckler, 2014).

Moreover, another possible factor contributing to low fertility/infertility in cattle involves various chromosomal abnormalities, including numerical anomalies affecting sex chromosomes and structural (balanced) chromosome aberrations. These abnormalities have been found to have detrimental effects on reproduction, particularly in female carriers. Since breeders cannot spot these chromosomal abnormalities as noticeable phenotype abnormalities, they can persist in the population without being selected against. Therefore, they can evade animal selection without cytogenetic control, ultimately negatively impacting fertility (A. Iannuzzi et al., 2021).

Sex chromosome abnormalities that can cause infertility in cattle encompass X-trisomy, X-monosomy, XXY syndrome, XY Sex Reversal Syndrome, XX Sex Reversal, XX/XY Mosaicism (Free-Martinism), and Diploid-Triploid XX/XXY Mosaicism (Mixoploidy). Additionally, structural chromosome abnormalities in cattle include reciprocal translocations (rcp), Robertsonian translocations (rob), simple translocations, pericentric inversions, tandem fusion (TAN), and cytogenetically detectable deletions and duplications (A. Iannuzzi et al., 2021). In most of the reported cases involving sex chromosome abnormalities, sterility was a common observation. However, in one of the cases documented by Schmutz et al. (1994), a Klinefelter bull was born to a fertile 61,XXX cow (X-trisomy). This is noteworthy because these cows had been previously described as having limited fertility or, more commonly, being sterile (Schmutz et al., 1994). On the other hand, cases involving structural chromosome abnormalities were generally associated with reduced fertility or subfertility which aligns with the focus of the current study.

Another potential explanation for the low fertility in these animals is their low heterosis for fertility, indicating that they possess an unfavorable combination of alleles that adversely impacts this trait. The overall influence of their genetic makeup on fertility could be the underlying cause, rather than a specific recessive lethal allele. Nonetheless, it is essential to consider the influence

of environmental factors as well. Exposure to toxins, nutritional deficiencies, diseases or infections, temperature stress, trauma, and stress can potentially affect specific animal samples, resulting in challenges related to conception or offspring loss.

Some of the candidate SNPs and indels that passed our filtration steps were common among all the 20 sequenced animals, which suggests they are common variants in these breeds or sequencing artifacts and not recessive lethals. New filtration steps could be applied in future studies to address this issue and eliminate these common mutations. During the examination of the final list of candidate recessive lethals, a notable challenge emerged concerning positions with various alternate alleles (genotypes 0/1, 0/2, 0/3, and so on). This complexity could stem from the repeated occurrence of the same base in adjacent positions of the genome, leading to sequencing or alignment errors. Or these sites could truly be highly polymorphic. The accuracy and fidelity of the sequencing process can significantly influence variant detection. Consequently, certain genomic regions might pose difficulties in sequencing, ultimately giving rise to the occurrence of false-positive variants.

The presence of multiple alleles at some sites complicates variant assessment. For example, a variant in the data set, the candidate SNP located on chromosome 5 at 58924076 bp (AC=18, 2 and AF=0.45, 0.05) with eighteen instances of alternate allele type 1 (0/1 genotype) and two instances of alternate allele type 2 (0/2 genotype). The high frequency of alternative allele type 1 renders it unlikely to be a true recessive lethal allele, but alternative allele type 2 shows potential as a candidate causative recessive lethal allele, necessitating further investigation. To enhance the identification of true causative recessive lethal alleles, a filter that removes those variants appearing in more than a certain number of individuals could be implemented in future research endeavors.

Additional experiments are needed to provide more robust evidence that any of the identified candidate mutations are true recessive lethals. The first step would be to verify the presence of the candidate allele in the affected sample(s) using Sanger sequencing. The next step would be to apply a low-cost genotyping assay for the site to many animals in the population, to establish allele frequency and departure from Hardy-Weinberg equilibrium in the form of a notable lack of homozygous genotypes. Another approach would be to cross heterozygous animals for the candidate mutation and observe the frequency of homozygous offspring. If the mutation is a true recessive lethal, viable homozygous offspring would not be observed. With many such matings and an absence of homozygous offspring it would be possible to make a strong case that the mutation is a recessive lethal. Another approach would be to perform genome editing to create homozygous animals for the mutation and observe their viability and fertility. This approach could work to rule out candidate mutations but a failure to obtain homozygous progeny would be difficult to interpret. Moreover, the cost and complexity of this approach makes it unlikely to be used.

For future strategies for finding recessive lethals in this population, a more routine genotype panel approach may be warranted, as it has been used in previous studies. This approach involves analyzing genome-wide genotypes from a large number of animals to highlight regions of the genome that are not observed in homozygous form. Once these regions are identified, targeted sequencing can be employed to pinpoint specific mutations. Currently, mid- or high-density SNP chips are the most suitable genotyping platform for these studies, as they are routinely used for other purposes like GEBV estimation. In the future, sequencing might be feasible provided that a large number of animals can be sequenced so that departures from Hardy-Weinberg equilibrium can be detected for low frequency alleles.

## 4.6. References


2nd Western Canadian Cow-Calf Survey. (2018, Fall). *Western Canadian Cow-Calf Survey*. Western Beef Development Centre (WBDC). http://westernbeef.org/wcccs.htm

Adams, H. A., Sonstegard, T. S., VanRaden, P. M., Null, D. J., Van Tassell, C. P., Larkin, D. M., & Lewin, H. A. (2016). Identification of a nonsense mutation in APAF1 that is likely causal for a decrease in reproductive efficiency in Holstein dairy cattle. *Journal of Dairy Science*, *99*(8), 6693–6701.

Agerholm, J. S., McEvoy, F., & Arnbjerg, J. (2006). Brachyspina syndrome in a Holstein calf. *Journal of Veterinary Diagnostic Investigation*, *18*(4), 418–422.

Alberta Beef Producers. (2023, November 24). Cattle Report. *Alberta Beef Producers*. https://abpdaily.com/cattle-report/

Alberta Cow-Calf Audit. (2001). *Alberta Cow-Calf Audit, 1997/1998: Production indicators and management practices over the last 10 years.* Alberta Agriculture and Forestry, Lacombe Research and Development Centre, Lacombe, AB, Canada. http://westernbeef.org/pdfs/economics/AB_CowCalf_Audit.pdf

Andrews, S. (2010, April 26). *Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Basarab, J. A., Crowley, J. J., Abo-Ismail, M. K., Manafiazar, G. M., Akanno, E. C., Baron, V. S., & Plastow, G. (2018). Genomic retained heterosis effects on fertility and lifetime productivity in beef heifers. *Canadian Journal of Animal Science*, *98*(4), 642–655. https://doi.org/10.1139/cjas-2017-0192

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Broad Institute. (2018). *Picard Tools—By Broad Institute*. http://broadinstitute.github.io/picard/

Charlier, C., Agerholm, J. S., Coppieters, W., Karlskov-Mortensen, P., Li, W., De Jong, G., Fasquelle, C., Karim, L., Cirera, S., & Cambisano, N. (2012). *A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina*.

Chen, Y., Zhang, Y., Wang, A. Y., Gao, M., & Chong, Z. (2021). Accurate long-read de novo assembly evaluation with Inspector. *Genome Biology*, *22*(1), 312. https://doi.org/10.1186/s13059-021-02527-4

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cooper, T. A., Wiggans, G. R., Null, D. J., Hutchison, J. L., & Cole, J. B. (2014). Genomic evaluation, breed identification, and discovery of a haplotype affecting fertility for Ayrshire dairy cattle. *Journal of Dairy Science*, *97*(6), 3878–3882.

Daly, R. (2021, February 10). *Late-Term Reproductive Losses in Beef Cattle: Diagnosing the Cause*. https://extension.sdstate.edu/late-term-reproductive-losses-beef-cattle-diagnosing-cause

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes

Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

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

Fritz, S., Capitan, A., Djari, A., Rodriguez, S. C., Barbat, A., Baur, A., Grohs, C., Weiss, B., Boussaha, M., & Esquerre, D. (2013). Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in GART, SHBG and SLC37A2. *PloS One*, *8*(6), e65550.

*GatherVcfs (Picard)*. (2019, May 25). GATK. https://gatk.broadinstitute.org/hc/en-us/articles/360036803571-GatherVcfs-Picard-

Hayes, B. J., & Daetwyler, H. D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual Review of Animal Biosciences*, *7*, 89–102. https://doi.org/10.1146/annurev-animal-020518-115024

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), 1–11.

Iannuzzi, A., Parma, P., & Iannuzzi, L. (2021). Chromosome Abnormalities and Fertility in Domestic Bovids: A Review. *Animals : An Open Access Journal from MDPI*, *11*(3), 802. https://doi.org/10.3390/ani11030802

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*(1), 91. https://doi.org/10.1186/s13073-020-00791-w

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Mistry, M., Ho Sui, S., Liu, J., Piper, M., Gammerdinger, W., & Khetani, R. (2021, September 15). *Troubleshooting the ChIP-seq workflow: From sequence reads to peak calls*. Introduction to ChIP-Seq Using High-Performance Computing. https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/troubleshooting_chipseq_partI.html

Mölder, F., Jablonski, K. P., & Letcher, B. (2021). *Sustainable data analysis with Snakemake.[version 2; peer review: 2 approved]. F1000Research 10: 33*.

Navarre, C. B. (2010). Infertility in cattle. *Animal Health Notes*.

O'Connor, B. D., & Van Auwera, G. (2020). *Genomics in the cloud: Using Docker, Gatk, and WDL in Terra*. O'Reilly Media Sebastopol, CA, USA:

Pausch, H., Kölle, S., Wurmser, C., Schwarzenbacher, H., Emmerling, R., Jansen, S., Trottmann, M., Fuerst, C., Götz, K.-U., & Fries, R. (2014). A nonsense mutation in TMEM95 encoding a nondescript transmembrane protein causes idiopathic male subfertility in cattle. *PLoS Genetics*, *10*(1), e1004044.

Sahana, G., Nielsen, U. S., Aamand, G. P., Lund, M. S., & Guldbrandtsen, B. (2013). Novel harmful recessive haplotypes identified for fertility traits in Nordic Holstein cattle. *PLoS One*, *8*(12), e82909.

Schmutz, S. M., Barth, A. D., & Moker, J. S. (1994). A Klinefelter bull with a 1;29 translocation born to a fertile 61,XXX cow. *The Canadian Veterinary Journal*, *35*(3), 182–184.

Schwarzenbacher, H., Fuerst, C., Fuerst-Waltl, B., & Dolezal, M. (2012). A genomewide search for harmful recessive haplotypes in Brown Swiss and Fleckvieh cattle In: 63rd EAAP Meeting Book of Abstracts. *Bratislava, Slovakia*.

Sonstegard, T. S., Cole, J. B., VanRaden, P. M., Van Tassell, C. P., Null, D. J., Schroeder, S. G., Bickhart, D., & McClure, M. C. (2013). Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. *PloS One*, *8*(1), e54872.

Steckler, T. (2014, January 21). *Cause of infertility in cattle found: University of Illinois Extension*. https://extension.illinois.edu/blogs/cattle-blog/2014-01-21-cause-infertility-cattle-found

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111.

Valour, D., Michot, P., Eozenou, C., Lefebvre, R., Bonnet, A., Capitan, A., Uzbekova, S., Sellem, E., Ponsart, C., & Schibler, L. (2015). Dairy cattle reproduction is a tightly regulated genetic process: Highlights on genes, pathways, and biological processes. *Animal Frontiers*, *5*(1), 32–41. https://doi.org/10.2527/af.2015-0006

VanRaden, P. M., Olson, K. M., Null, D. J., & Hutchison, J. L. (2011). Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *Journal of Dairy Science*, *94*(12), 6153–6161.

Waldner, C. (2019, November 26). *Preventing Reproductive Failure in Cow-Calf Herds—Beef Cattle Research Council*. Preventing Reproductive Failure in Cow-Calf Herds - Beef Cattle Research Council. http://www.beefresearch.ca/research-topic.cfm

# CHAPTER 5. General Conclusion

In this thesis, three studies were carried out seeking to identify the causal variants contributing to low fertility in cattle, Swyer syndrome in cattle, and copper toxicosis in the Dalmatian dog breed. In each study, whole-genome sequencing (WGS) and bioinformatics techniques were applied based on emerging evidence into the molecular mechanism of these conditions that supports the possible involvement of large-effect genetic variations. Beyond identifying causal variants, I sought to enhance our comprehension and insight into the potential genetic roots of these conditions in the studied species.

In the first study, eight phenotypically female cattle with SNP chip genotypes indicating a male genotype were investigated to understand the genetic basis of this apparent discrepancy between genotype and phenotype. In humans, this condition (Swyer syndrome) has been linked to a variety of mutations and inheritance patterns. Through an in-depth investigation of whole-genome sequence data collected from the affected individuals and controls, all but one case can be attributed to a Y chromosome deletion that encompasses the *SRY* gene. A clear cause could not be identified for the remaining case, despite careful inspections of known DSD genes for impactful dominant or recessive mutations and any other genes for impactful de novo mutations. A list of 125 SNPs/indels was produced using the "dominant de novo" filter for the one remaining case, but more work is needed to narrow this list further. The finding that most cases are related to Y chromosome anomalies leads to recommendations for how large-scale genetic improvement programs can better detect such cases in the future to avoid confusion and technology distrust among farmers.

In the second study, to identify the possible causative variant related to copper toxicosis in 13 Dalmatians and 3 Bedlington Terriers, all the identified variants within genes known to contribute to the human form of the disease and suspected to play a role in certain other dog breeds

were considered. Despite multiple investigations using different variant filtering methods and manual data inspections, no candidate causal variant associated with copper toxicosis in Dalmatians was found. One challenge was that the phenotype (copper levels in the liver) can be affected by diet, environmental contamination, and possibly other non-genetic factors (Gaetke et al., 2014), which may lead to incorrect classification samples. Even though the true causative variant was not discovered in this study, the gathered information, including the dataset, along with the results and outputs, will still be helpful to researchers working in this area. In particular, the results provide strong evidence that previously reported genes *ATP7A*, *ATP7B*, and *COMMD1* are not responsible for copper toxicosis in Dalmatians. However, it must be noted that completely ruling out the involvement of a particular gene is difficult from whole-genome sequence data alone, as uncharacterized sequence differences, even far from the gene, could exist and affect phenotype. Indeed, several factors, such as read length, sequencing depth, annotation information, reference sequences, and software tools, can influence variant detection and characterization.

In the third study, I aimed to identify recessive alleles responsible for low fertility in a sample of 20 cattle. The implicit hypothesis was that recessive lethal variants contributed to low fertility and could be detected through whole-genome sequencing and bioinformatics. Although many candidate recessive lethals were identified, more information on the impact of the identified variant on the disease or the genes' contributions to the disease would be needed to identify a more specific candidate or a true recessive lethal allele. The low fertility observed in our sample animals may not solely be due to the presence of a specific recessive lethal allele. However, it may instead be influenced by a collection of alleles that negatively affect the trait. Additionally, the observed low fertility in our crossbred beef cows may be associated with low heterosis. Other factors, such as environmental conditions or other unknown causes, may have contributed to their inability to

conceive or to lose their offspring. Thus, it is essential to consider all potential factors that could affect fertility in our sample, not just the recessive lethal alleles. Some of the candidate causal mutations that passed our filtration steps appear to be common in our set of 20 sequenced animals, which has lowered our confidence in their potential significance because a true recessive lethal is unlikely to be that common. In future studies, additional filtration steps must be applied to eliminate these common mutations further.

## 5.1. Limitations

Certain limitations in the present research could have impacted on the findings observed.

**Small sample size:** A notable constraint of this study was the relatively limited sample size in each study group, potentially impeding our ability to reliably identify consistent significant or non-significant results. Small sample sizes can arise in various scenarios, such as when the number of cases is low or when obtaining samples proves challenging. For instance, in the Swyer Syndrome study, the condition's rarity naturally leads to a small sample size. Similarly, collecting a large sample size can be challenging in the second study, the copper toxicosis disease, where the number of affected animals is limited. Moreover, in the context of whole-genome sequencing (WGS), the high cost of sequencing can impose constraints on the feasible sample size for a study. Employing filtering methods on a limited sample size may lead to a considerable number of variants passing the filter, potentially resulting in erroneous conclusions and interpretations. This situation arises because small sample sizes lack the statistical power required to distinguish genuine variants from noise or false positives. When applying statistical approaches with small sample sizes, the statistical power diminishes, reducing the likelihood of detecting true

associations. Consequently, false-negative results may occur, leading to the failure to identify true associations and missing crucial genetic variants or pathways.

**Sequencing data:** Another limitation of our study was that the short reads generated by next-generation sequencing technologies could pose challenges in accurately identifying genetic variations. The length of these reads is usually under 300 base pairs, which can be problematic when trying to resolve complex genomic regions or repetitive sequences. This can result in false positives and negatives in variant calling, ultimately missing a true causative variant (Zhou et al., 2022). Furthermore, incomplete reference genomes can also pose challenges for accurate variant discovery and genotyping. The reference genome plays a crucial role in variant calling, serving as a fundamental benchmark for identifying discrepancies between the sample and the reference. It is essential to recognize that the reference genome is typically constructed from a relatively small number of individuals and may not encompass the full diversity of genetic variations present in the population. As a result, this limitation can lead to false negatives, wherein true genetic variations existing in the sample are not identified (Veeckman et al., 2019). In addition, the identification and characterization of variants are also limited by other factors, including sequencing depth, quality of available annotation information, and software tools.

**Phenotyping:** Measuring phenotypes can pose challenges and introduce unreliability due to several factors, including invasiveness, variability, and environmental effects. For instance, in the second study, measuring copper levels in the liver can be invasive, necessitating a liver biopsy, which may not always be feasible or practical. Additionally, copper levels in the liver can be influenced by various factors, such as dietary intake and environmental contamination, potentially involving non-genetic elements (Gaetke et al., 2014), which may lead to incorrect classification of samples. These environmental factors may lead to variation in the phenotype of interest, potentially

obscuring any underlying genetic effects. Additionally, in the third study, environmental conditions may have contributed to their inability to conceive or maintain their offspring. Hence, we must consider all potential factors affecting fertility in our sample, not just the recessive lethal alleles.

**Polygenic and single-gene contributions to phenotype:** the genetic basis of copper toxicosis in dogs is not entirely clear. It is not known if this disorder is caused by a single gene or multiple genes with varying effects. The approach employed in this study will be practical only if the datasets have significant contributions from a single gene with a large effect. Therefore, it is uncertain if a single genetic variant with a large effect size can be identified using our filtering approach. Furthermore, the phenotype associated with fertility may be influenced by multiple genes with minor effects, including recessive lethals. However, it is important to acknowledge that complex phenotypes such as fertility are often polygenic and can be influenced by a combination of genetic and environmental factors, which may not be captured by our approach.

Moreover, in the third study, the reduced fertility (the phenotype) observed in our animal samples might be influenced by other factors rather than entirely affected by a specific recessive lethal allele. In other words, it may be influenced by a combination of unfavorable alleles that affect the trait or may be is due to low heterosis in our sample animals.

## 5.2. Implications of Work

**Swyer syndrome study:** The list of 125 SNPs/indels in SNPs/indels that was compiled by the "dominant de novo" filter the specific *SRY*+ case in this study could potentially hold value for

future research in both cattle and humans, particularly in cases where the underlying causes are still unknown.

**Copper toxicosis study:** The results of our study have proved that specific significant genes previously known to be linked to the disease are not implicated in Dalmatians. The comprehensive information provided in this study can serve as a foundation for future research endeavors focused on unraveling the genetic origins of the disease in Dalmatians as well as other canine breeds.

**Subfertility/infertility study:** This study depicted that the applied approach alone may not be adequate to identify the true causative recessive lethal alleles associated with the subfertility/infertility in beef cattle.

## 5.3. Future Work

**Swyer syndrome study:** Due to Ireland's extensive cattle breeding program, which involves genotyping, they are in a favorable position to detect DSD cases, which could serve as an asset in identifying other genes associated with Swyer syndrome. For example, other cases may arise from the unknown factor in the same *SRY*+ animal, providing an opportunity to employ filters to determine the underlying cause in this specific scenario.

**Copper toxicosis study:** To tackle the challenge of detecting structural variations (SVs) using short-read sequencing in this study, two potential solutions can be carried out in future research. The first option involves performing long-read sequencing to address the issue. The second option is gathering more samples from affected and control groups to facilitate a more comprehensive linkage or genome-wide association (GWAS) analysis. Despite the difficulty of phenotyping, a

larger sample size can allow for the application of statistical methods instead of relying solely on rigid filtering criteria, which may help to mitigate issues arising from misclassified phenotypes.

**Subfertility/infertility study:** The appropriate strategy to adopt for future work depends on the objective. For example, if the primary aim is to identify recessive lethals, employing a "missing haplotypes in homozygous form" research would be advisable using SNP chip genotypes on large numbers of individuals. Conversely, if the goal is to enhance fertility or identify any underlying causal variants, approaches such as estimated breeding values (EBV) and genome-wide association studies (GWAS) would be more suitable.

## 5.4. References


Gaetke, L. M., Chow-Johnson, H. S., & Chow, C. K. (2014). Copper: Toxicological relevance and mechanisms. *Archives of Toxicology*, *88*(11), 1929–1938. https://doi.org/10.1007/s00204-014-1355-y

Veeckman, E., Van Glabeke, S., Haegeman, A., Muylle, H., van Parijs, F. R. D., Byrne, S. L., Asp, T., Studer, B., Rohde, A., Roldán-Ruiz, I., Vandepoele, K., & Ruttink, T. (2019). Overcoming challenges in variant calling: Exploring sequence diversity in candidate genes for plant development in perennial ryegrass (Lolium perenne). *DNA Research*, *26*(1), 1–12. https://doi.org/10.1093/dnares/dsy033

Zhou, Y., Yang, L., Han, X., Han, J., Hu, Y., Li, F., Xia, H., Peng, L., Boschiero, C., Rosen, B. D., Bickhart, D. M., Zhang, S., Guo, A., Van Tassell, C. P., Smith, T. P. L., Yang, L., & Liu, G. E. (2022). Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and

evolutionary history. Genome Research, 32(8), 1585–1601.

https://doi.org/10.1101/gr.276550.122

# Bibliography

2nd Western Canadian Cow-Calf Survey. (2018, Fall). *Western Canadian Cow-Calf Survey*. Western Beef Development Centre (WBDC). http://westernbeef.org/wcccs.htm

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Abyzov, A., Li, S., Kim, D. R., Mohiyuddin, M., Stütz, A. M., Parrish, N. F., Mu, X. J., Clark, W., Chen, K., & Hurles, M. (2015). Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications*, *6*(1), 1–12.

Adams, H. A., Sonstegard, T. S., VanRaden, P. M., Null, D. J., Van Tassell, C. P., Larkin, D. M., & Lewin, H. A. (2016). Identification of a nonsense mutation in APAF1 that is likely causal for a decrease in reproductive efficiency in Holstein dairy cattle. *Journal of Dairy Science*, *99*(8), 6693–6701.

Agerholm, J. S., McEvoy, F., & Arnbjerg, J. (2006). Brachyspina syndrome in a Holstein calf. *Journal of Veterinary Diagnostic Investigation*, *18*(4), 418–422.

Ajay, S. S., Parker, S. C. J., Abaan, H. O., Fajardo, K. V. F., & Margulies, E. H. (2011). Accurate and comprehensive sequencing of personal genomes. *Genome Research*, *21*(9), 1498–1505. https://doi.org/10.1101/gr.123638.111

Alberta Beef Producers. (2023, November 24). Cattle Report. *Alberta Beef Producers*. https://abpdaily.com/cattle-report/

Alberta Cow-Calf Audit. (2001). *Alberta Cow-Calf Audit, 1997/1998: Production indicators and management practices over the last 10 years.* Alberta Agriculture and Forestry, Lacombe

Research and Development Centre, Lacombe, AB, Canada. http://westernbeef.org/pdfs/economics/AB_CowCalf_Audit.pdf

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Andrews, S. (2010, April 26). *Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arboleda, V., & Vilain, E. (2016). Chapter 19—Disorders of Sex Development. In R. E. Weiss & S. Refetoff (Eds.), *Genetic Diagnosis of Endocrine Disorders (Second Edition)* (pp. 259–278). Academic Press. https://doi.org/10.1016/B978-0-12-800892-8.00019-1

Askwith, C., & Kaplan, J. (1998). Iron and copper transport in yeast and its relevance to human disease. *Trends in Biochemical Sciences*, *23*(4), 135–138. https://doi.org/10.1016/s0968-0004(98)01192-x

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), Article 7571. https://doi.org/10.1038/nature15393

Baetens, D., Güran, T., Mendonca, B. B., Gomes, N. L., De Cauwer, L., Peelman, F., Verdin, H., Vuylsteke, M., Van der Linden, M., ESR2 STUDY GROUP, Atay, Z., Bereket, A., de Krijger, R. R., Preter, K. de, Domenice, S., Turan, S., Stoop, H., Looijenga, L. H., De Bosscher, K., … De Baere, E. (2018). Biallelic and monoallelic ESR2 variants associated with 46,XY disorders of sex development. *Genetics in Medicine: Official Journal of the*

*American College of Medical Genetics*, *20*(7), 717–727. https://doi.org/10.1038/gim.2017.163

Banoth, M., Naru, R. R., Inamdar, M. B., & Chowhan, A. K. (2018). Familial Swyer syndrome: A rare genetic entity. *Gynecological Endocrinology*, *34*(5), 389–393. https://doi.org/10.1080/09513590.2017.1393662

Basarab, J. A., Crowley, J. J., Abo-Ismail, M. K., Manafiazar, G. M., Akanno, E. C., Baron, V. S., & Plastow, G. (2018). Genomic retained heterosis effects on fertility and lifetime productivity in beef heifers. *Canadian Journal of Animal Science*, *98*(4), 642–655. https://doi.org/10.1139/cjas-2017-0192

Bauer, A., Waluk, D. P., Galichet, A., Timm, K., Jagannathan, V., Sayar, B. S., Wiener, D. J., Dietschi, E., Müller, E. J., Roosje, P., Welle, M. M., & Leeb, T. (2017). A de novo variant in the ASPRV1 gene in a dog with ichthyosis. *PLoS Genetics*, *13*(3), e1006651. https://doi.org/10.1371/journal.pgen.1006651

Becker, D., Minor, K. M., Letko, A., Ekenstedt, K. J., Jagannathan, V., Leeb, T., Shelton, G. D., Mickelson, J. R., & Drögemüller, C. (2017). A GJA9 frameshift variant is associated with polyneuropathy in Leonberger dogs. *BMC Genomics*, *18*(1), 662. https://doi.org/10.1186/s12864-017-4081-z

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood. Education and Practice Edition*, *98*(6), 236–238. https://doi.org/10.1136/archdischild-2013-304340

Behtash, N., & Karimi Zarchi, M. (2007). Dysgerminoma in three patients with Swyer syndrome. *World Journal of Surgical Oncology*, *5*(1), 71. https://doi.org/10.1186/1477-7819-5-71

Ben-Ari, G., & Lavi, U. (2012). 11—Marker-assisted selection in plant breeding. In A. Altman & P. M. Hasegawa (Eds.), *Plant Biotechnology and Agriculture* (pp. 163–184). Academic Press. https://doi.org/10.1016/B978-0-12-381466-1.00011-0

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Berkovitz, G. D., & Seeherunvong, T. (1998). Abnormalities of gonadal differentiation. *Bailliere's Clinical Endocrinology and Metabolism*, *12*(1), 133–142. https://doi.org/10.1016/s0950-351x(98)80512-0

Berta, P., Hawkins, J. R., Sinclair, A. H., Taylor, A., Griffiths, B. L., Goodfellow, P. N., & Fellous, M. (1990). Genetic evidence equating SRY and the testis-determining factor. *Nature*, *348*(6300), 448–450. https://doi.org/10.1038/348448A0

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bourneuf, E., Otz, P., Pausch, H., Jagannathan, V., Michot, P., Grohs, C., Piton, G., Ammermüller, S., Deloche, M.-C., Fritz, S., Leclerc, H., Péchoux, C., Boukadiri, A., Hozé, C., Saintilan, R., Créchet, F., Mosca, M., Segelke, D., Guillaume, F., … Capitan, A. (2017). Rapid Discovery of De Novo Deleterious Mutations in Cattle Enhances the

Value of Livestock as Model Species. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-11523-3

Broad Institute. (2018). *Picard Tools—By Broad Institute*. http://broadinstitute.github.io/picard/

Bull, P. C., Thomas, G. R., Rommens, J. M., Forbes, J. R., & Cox, D. W. (1993). The Wilson disease gene is a putative copper transporting P-type ATPase similar to the Menkes gene. *Nature Genetics*, *5*(4), 327–337. https://doi.org/10.1038/ng1293-327

*CanineHD Whole-Genome Genotyping BeadChip*. (n.d.). Retrieved August 31, 2023, from https://www.illumina.com/products/by-type/microarray-kits/caninehd.html

Canto, P., Söderlund, D., Reyes, E., & Méndez, J. P. (2004). Mutations in the desert hedgehog (DHH) gene in patients with 46,XY complete pure gonadal dysgenesis. *The Journal of Clinical Endocrinology and Metabolism*, *89*(9), 4480–4483. https://doi.org/10.1210/jc.2004-0863

Carroll, D. (2001). Genetic Recombination. In S. Brenner & J. H. Miller (Eds.), *Encyclopedia of Genetics* (pp. 841–845). Academic Press. https://doi.org/10.1006/rwgn.2001.0543

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews. Genetics*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, *19*(3 Pt 1), 233.

Charlier, C., Agerholm, J. S., Coppieters, W., Karlskov-Mortensen, P., Li, W., De Jong, G., Fasquelle, C., Karim, L., Cirera, S., & Cambisano, N. (2012). *A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina*.

Chelly, J., Tümer, Z., Tønnesen, T., Petterson, A., Ishikawa-Brush, Y., Tommerup, N., Horn, N., & Monaco, A. P. (1993). Isolation of a candidate gene for Menkes disease that encodes a potential heavy metal binding protein. *Nature Genetics*, *3*(1), 14–19. https://doi.org/10.1038/ng0193-14

Chen, J., Jiang, Y., Shi, H., Peng, Y., Fan, X., & Li, C. (2020). The molecular mechanisms of copper metabolism and its roles in human diseases. *Pflugers Archiv: European Journal of Physiology*, *472*(10), 1415–1429. https://doi.org/10.1007/s00424-020-02412-2

Chen, L., Pryce, J. E., Hayes, B. J., & Daetwyler, H. D. (2021). Investigating the Effect of Imputed Structural Variants from Whole-Genome Sequence on Genome-Wide Association and Genomic Prediction in Dairy Cattle. *Animals : An Open Access Journal from MDPI*, *11*(2), 541. https://doi.org/10.3390/ani11020541

Chen, Y., Zhang, Y., Wang, A. Y., Gao, M., & Chong, Z. (2021). Accurate long-read de novo assembly evaluation with Inspector. *Genome Biology*, *22*(1), 312. https://doi.org/10.1186/s13059-021-02527-4

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., & Montgomery, S. B. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, *49*(5), 692–699.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., & Sanjad, S. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*, *106*(45), 19096–19101.

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies

with a New Program, SnpSift. *Frontiers in Genetics*, *3*, 35. https://doi.org/10.3389/fgene.2012.00035

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Clark, D. P., & Pazdernik, N. J. (2013). Chapter 23—Mutations and Repair. In D. P. Clark & N. J. Pazdernik (Eds.), *Molecular Biology (Second Edition)* (pp. 721–766). Academic Press. https://doi.org/10.1016/B978-0-12-378594-7.00023-8

Coffey, A. J., Durkie, M., Hague, S., McLay, K., Emmerson, J., Lo, C., Klaffke, S., Joyce, C. J., Dhawan, A., Hadzic, N., Mieli-Vergani, G., Kirk, R., Elizabeth Allen, K., Nicholl, D., Wong, S., Griffiths, W., Smithson, S., Giffin, N., Taha, A., … Bandmann, O. (2013). A genetic study of Wilson's disease in the United Kingdom. *Brain*, *136*(5), 1476–1487. https://doi.org/10.1093/brain/awt035

A global reference for human genetic variation, 526 Nature 68 (Nature Publishing Group 2015).

Consortium, B. H., Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., Green, R. D., Hamernik, D. L., & Kappes, S. M. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*, *324*(5926), 528–532.

Cooper, T. A., Wiggans, G. R., Null, D. J., Hutchison, J. L., & Cole, J. B. (2014). Genomic evaluation, breed identification, and discovery of a haplotype affecting fertility for Ayrshire dairy cattle. *Journal of Dairy Science*, *97*(6), 3878–3882.

Cooper, V. L., Carlson, M. P., Jacobson, J., & Schneider, N. R. (1997). Hepatitis and increased copper levels in a dalmatian. *Journal of Veterinary Diagnostic Investigation: Official Publication of the American Association of Veterinary Laboratory Diagnosticians, Inc*, *9*(2), 201–203. https://doi.org/10.1177/104063879700900217

*Copper Poisoning in Dogs—Symptoms, Causes, Diagnosis, Treatment, Recovery, Management, Cost*. (n.d.). Retrieved April 21, 2022, from https://wagwalking.com/condition/copper-poisoning, https://wagwalking.com/condition/copper-poisoning

Corbee, R. J., & Penning, L. C. (2021). COMMD1 Exemplifies the Power of Inbred Dogs to Dissect Genetic Causes of Rare Copper-Related Disorders. *Animals: An Open Access Journal from MDPI*, *11*(3), 601. https://doi.org/10.3390/ani11030601

Coronado, V. A., Damaraju, D., Kohijoki, R., & Cox, D. W. (2003). New haplotypes in the Bedlington terrier indicate complexity in copper toxicosis. *Mammalian Genome*, *14*(7), 483–491.

Cunningham, F., Moore, B., Ruiz-Schultz, N., Ritchie, G. R., & Eilbeck, K. (2015). Improving the Sequence Ontology terminology for genomic variant annotation. *Journal of Biomedical Semantics*, *6*(1), 32. https://doi.org/10.1186/s13326-015-0030-4

D'Agaro, E., Favaro, A., Rosa, D., D'Agaro, E., Favaro, A., & Rosa, D. (2021). Canine Genetics and Genomics. In *Canine Genetics, Health and Medicine*. IntechOpen. https://doi.org/10.5772/intechopen.95781

Dagenais, S. L., Guevara-Fujita, M., Glover, T. W., Burgess, A. C., Loechel, R., Brewer, G. J., Yuzbasiyan-Gurkan, V., & Miller, D. E. (1999). *The canine copper toxicosis locus is not syntenic with ATP7B or ATX1 and maps to a region showing homology to human 2p21*.

Daly, R. (2021, February 10). *Late-Term Reproductive Losses in Beef Cattle: Diagnosing the Cause*. https://extension.sdstate.edu/late-term-reproductive-losses-beef-cattle-diagnosing-cause

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

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

Délot, E. C., & Vilain, E. (2019). Disorders of sex development. In *Yen and Jaffe's Reproductive Endocrinology* (pp. 365–393). Elsevier.

Deng, X., den Bakker, H., & Hendriksen, R. (2016). Genomic Epidemiology: Whole-Genome-Sequencing—Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annual Review of Food Science and Technology*, *7*. https://doi.org/10.1146/annurev-food-041715-033259

Dhindsa, R. S., Wang, Q., Vitsios, D., Burren, O. S., Hu, F., DiCarlo, J. E., Kruglyak, L., MacArthur, D. G., Hurles, M. E., & Petrovski, S. (2022). A minimal role for synonymous variation in human disease. *The American Journal of Human Genetics*, *109*(12), 2105–2109. https://doi.org/10.1016/j.ajhg.2022.10.016

Dirksen, K., Burgener, I. A., Rothuizen, J., van den Ingh, T. S. G. a. M., Penning, L. C., Spee, B., & Fieten, H. (2017). Sensitivity and Specificity of Plasma ALT, ALP, and Bile Acids

for Hepatitis in Labrador Retrievers. *Journal of Veterinary Internal Medicine*, *31*(4), 1017–1027. https://doi.org/10.1111/jvim.14716

Dirksen, K., Spee, B., Penning, L. C., Ingh, T. S. G. A. M. van den, Burgener, I. A., Watson, A. L., Koerkamp, M. G., Rothuizen, J., Steenbeek, F. G. van, & Fieten, H. (2017). Gene expression patterns in the progression of canine copper-associated chronic hepatitis. *PLOS ONE*, *12*(5), e0176826. https://doi.org/10.1371/journal.pone.0176826

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *51*, 11.14.1-11.14.19. https://doi.org/10.1002/0471250953.bi1114s51

Domenice, S., Correa, R. V., Costa, E. M. F., Nishi, M. Y., Vilain, E., Arnhold, I. J. P., & Mendonca, B. B. (2004). Mutations in the SRY, DAX1, SF1 and WNT4 genes in Brazilian sex-reversed patients. *Brazilian Journal of Medical and Biological Research*, *37*, 145–150.

Du, X., Zhang, X., Li, Y., & Han, Y. (2014). 46, XY female sex reversal syndrome with bilateral gonadoblastoma and dysgerminoma. *Experimental and Therapeutic Medicine*, *8*(4), 1102–1104.

Dunislawska, A., Lachmańska, J., Slawińska, A., & Siwek, M. (2017). Next generation sequencing in animal science-a review. *Animal Science Papers & Reports*, *35*(3).

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, *372*(6537), eabf7117. https://doi.org/10.1126/science.abf7117

Eggers, S., Smith, K. R., Bahlo, M., Looijenga, L. H., Drop, S. L., Juniarto, Z. A., Harley, V. R., Koopman, P., Faradz, S. M., & Sinclair, A. H. (2015). Whole exome sequencing combined with linkage analysis identifies a novel 3 bp deletion in NR5A1. *European Journal of Human Genetics*, *23*(4), 486–493. https://doi.org/10.1038/ejhg.2014.130

Fang, H., Wu, Y., Narzisi, G., O'Rawe, J. A., Barrón, L. T. J., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M. C., & Lyon, G. J. (2014). Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Medicine*, *6*(10), 89. https://doi.org/10.1186/s13073-014-0089-z

Fareed, M., & Afzal, M. (2013). Single nucleotide polymorphism in genome-wide association of human population: A tool for broad spectrum service. *Egyptian Journal of Medical Human Genetics*, *14*(2), 123–134. https://doi.org/10.1016/j.ejmhg.2012.08.001

Favier, R. P., Spee, B., Schotanus, B. A., Ingh, T. S. G. A. M. van den, Fieten, H., Brinkhof, B., Viebahn, C. S., Penning, L. C., & Rothuizen, J. (2012). COMMD1-Deficient Dogs Accumulate Copper in Hepatocytes and Provide a Good Model for Chronic Hepatitis and Fibrosis. *PLOS ONE*, *7*(8), e42158. https://doi.org/10.1371/journal.pone.0042158

Fay, D. (2006). Genetic mapping and manipulation: Chapter 1-Introduction and basics. *WormBook*, *17*, 1–12.

Fechner, P. Y. (1996). The role of SRY in mammalian sex determination. *Acta Paediatrica Japonica: Overseas Edition*, *38*(4), 380–389. https://doi.org/10.1111/j.1442-200x.1996.tb03512.x

Fedoseienko, A., Bartuzi, P., & van de Sluis, B. (2014). Functional understanding of the versatile protein copper metabolism MURR1 domain 1 (COMMD1) in copper homeostasis.

*Annals of the New York Academy of Sciences*, *1314*, 6–14. https://doi.org/10.1111/nyas.12353

Fieten, H., Biourge, V. C., Watson, A. L., Leegwater, P. A. J., van den Ingh, T. S. G. A. M., & Rothuizen, J. (2014). Nutritional management of inherited copper-associated hepatitis in the Labrador retriever. *Veterinary Journal (London, England: 1997)*, *199*(3), 429–433. https://doi.org/10.1016/j.tvjl.2013.12.017

Fieten, H., Gill, Y., Martin, A. J., Concilli, M., Dirksen, K., van Steenbeek, F. G., Spee, B., van den Ingh, T. S. G. A. M., Martens, E. C. C. P., Festa, P., Chesi, G., van de Sluis, B., Houwen, R. H. J. H., Watson, A. L., Aulchenko, Y. S., Hodgkinson, V. L., Zhu, S., Petris, M. J., Polishchuk, R. S., … Rothuizen, J. (2016). The Menkes and Wilson disease genes counteract in copper toxicosis in Labrador retrievers: A new canine model for copper-metabolism disorders. *Disease Models & Mechanisms*, *9*(1), 25–38. https://doi.org/10.1242/dmm.020263

Fieten, H., Leegwater, P. A. J., Watson, A. L., & Rothuizen, J. (2012). Canine models of copper toxicosis for understanding mammalian copper metabolism. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, *23*(1–2), 62–75. https://doi.org/10.1007/s00335-011-9378-7

Forman, O. P., Boursnell, M. E. G., Dunmore, B. J., Stendall, N., van den Sluis, B., Fretwell, N., Jones, C., Wijmenga, C., Rothuizen, J., van Oost, B. A., Holmes, N. G., Binns, M. M., & Jones, P. (2005). Characterization of the COMMD1 (MURR1) mutation causing copper toxicosis in Bedlington terriers. *Animal Genetics*, *36*(6), 497–501. https://doi.org/10.1111/j.1365-2052.2005.01360.x

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., & Hurles, M. E. (2006). Copy number variation: New insights in genome diversity. *Genome Research*, *16*(8), 949–961.

French, C. E., Dolling, H., Mégy, K., Sanchis-Juan, A., Kumar, A., Delon, I., Wakeling, M., Mallin, L., Agrawal, S., Austin, T., Walston, F., Park, S.-M., Parker, A., Piyasena, C., Bradbury, K., Ellard, S., Rowitch, D. H., & Raymond, F. L. (2022). Refinements and considerations for trio whole-genome sequence analysis when investigating Mendelian diseases presenting in early childhood. *Human Genetics and Genomics Advances*, *3*(3), 100113. https://doi.org/10.1016/j.xhgg.2022.100113

Fritz, S., Capitan, A., Djari, A., Rodriguez, S. C., Barbat, A., Baur, A., Grohs, C., Weiss, B., Boussaha, M., & Esquerre, D. (2013). Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in GART, SHBG and SLC37A2. *PloS One*, *8*(6), e65550.

Fuchs, E. (2018). Chapter 106—Haploidentical Hematopoietic Cell Transplantation. In R. Hoffman, E. J. Benz, L. E. Silberstein, H. E. Heslop, J. I. Weitz, J. Anastasi, M. E. Salama, & S. A. Abutalib (Eds.), *Hematology (Seventh Edition)* (pp. 1617-1632.e6). Elsevier. https://doi.org/10.1016/B978-0-323-35762-3.00106-2

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., & Altshuler, D. (2002). The Structure of Haplotype Blocks in the Human Genome. *Science*, *296*(5576), 2225–2229. https://doi.org/10.1126/science.1069424

Gaetke, L. M., Chow-Johnson, H. S., & Chow, C. K. (2014). Copper: Toxicological relevance and mechanisms. *Archives of Toxicology*, *88*(11), 1929–1938. https://doi.org/10.1007/s00204-014-1355-y

García-Nieto, P. E., Morrison, A. J., & Fraser, H. B. (2019). The somatic mutation landscape of the human body. *Genome Biology*, *20*(1), 1–20.

*GatherVcfs (Picard)*. (2019, May 25). GATK. https://gatk.broadinstitute.org/hc/en-us/articles/360036803571-GatherVcfs-Picard-

Genin, E., Feingold, J., & Clerget-Darpoux, F. (2008). Identifying modifier genes of monogenic disease: Strategies and difficulties. *Human Genetics*, *124*(4), 357–368. https://doi.org/10.1007/s00439-008-0560-2

Genome 10K Community of Scientists. (2009). Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of Heredity*, *100*(6), 659–674. https://doi.org/10.1093/jhered/esp086

Geurts van Kessel, A., Venkatachalam, R., & Kuiper, R. P. (2013). Genomic and Personalized Medicine, Chapter 62—Colorectal Cancer. In G. S. Ginsburg & H. F. Willard (Eds.), *Genomic and Personalized Medicine (Second Edition)* (pp. 722–732). Academic Press. https://doi.org/10.1016/B978-0-12-382227-7.00062-8

Gonzaga-Jauregui, C., Lupski, J. R., & Gibbs, R. A. (2012). Human genome sequencing in health and disease. *Annual Review of Medicine*, *63*, 35–61.

Gottlieb, S. F., Gulani, A., & Tegay, D. H. (2023). Genetics, Meiosis. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK482462/

Grubenbecher, S., Stüve, O., Hefter, H., & Korth, C. (2006). Prion protein gene codon 129 modulates clinical course of neurological Wilson disease. *NeuroReport*, *17*(5), 549. https://doi.org/10.1097/01.wnr.0000209006.48105.90

Guilherme, L., Köhler, K. F., & Kalil, J. (2011). Chapter 2 - Rheumatic heart disease: Mediation by complex immune events. In G. S. Makowski (Ed.), *Advances in Clinical Chemistry* (Vol. 53, pp. 31–50). Elsevier. https://doi.org/10.1016/B978-0-12-385855-9.00002-3

Guo, X., Dai, X., Zhou, T., Wang, H., Ni, J., Xue, J., & Wang, X. (2020). Mosaic loss of human Y chromosome: What, how and why. *Human Genetics*, *139*(4), 421–446. https://doi.org/10.1007/s00439-020-02114-w

Gupta, A. K., & Gupta, U. D. (2020). Next generation sequencing and its applications. In *Animal Biotechnology* (pp. 395–421). Elsevier.

Han, L., Zhao, X., Benton, M. L., Perumal, T., Collins, R. L., Hoffman, G. E., Johnson, J. S., Sloofman, L., Wang, H. Z., Stone, M. R., CommonMind Consortium, Brennand, K. J., Brand, H., Sieberts, S. K., Marenco, S., Peters, M. A., Lipska, B. K., Roussos, P., Capra, J. A., … Ruderfer, D. M. (2020). Functional annotation of rare structural variation in the human brain. *Nature Communications*, *11*(1), 2990. https://doi.org/10.1038/s41467-020-16736-1

Hare, J. E., Baird, J. D., Duignan, P., Saunders, J., Floetenmeyer, R., & Basrur, P. K. (1994). XY gonadal dysgenesis and tetralogy of Fallot in an Angus calf. *The Canadian Veterinary Journal*, *35*(8), 510–512.

Harris, Z. L., Durley, A. P., Man, T. K., & Gitlin, J. D. (1999). Targeted gene disruption reveals an essential role for ceruloplasmin in cellular iron efflux. *Proceedings of the National*

*Academy of Sciences of the United States of America*, *96*(19), 10812–10817. https://doi.org/10.1073/pnas.96.19.10812

Hawkins, J. R. (1993). Mutational analysis of SRY in XY females. *Human Mutation*, *2*(5), 347–350. https://doi.org/10.1002/humu.1380020504

Hawkins, J. R., Taylor, A., Berta, P., Levilliers, J., Van der Auwera, B., & Goodfellow, P. N. (1992). Mutational analysis of SRY: Nonsense and missense mutations in XY sex reversal. *Human Genetics*, *88*(4), 471–474. https://doi.org/10.1007/BF00215684

Hawkins, J. R., Taylor, A., Goodfellow, P. N., Migeon, C. J., Smith, K. D., & Berkovitz, G. D. (1992). Evidence for increased prevalence of SRY mutations in XY females with complete rather than partial gonadal dysgenesis. *American Journal of Human Genetics*, *51*(5), 979–984.

Hayes, B. J., & Daetwyler, H. D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual Review of Animal Biosciences*, *7*, 89–102. https://doi.org/10.1146/annurev-animal-020518-115024

Haywood, S., Boursnell, M., Loughran, M. J., Trafford, J., Isherwood, D., Liu, X., Olohan, L., & Carter, S. D. (2016). Copper toxicosis in non-COMMD1 Bedlington terriers is associated with metal transport gene ABCA12. *Journal of Trace Elements in Medicine and Biology*, *35*, 83–89. https://doi.org/10.1016/j.jtemb.2016.01.015

Haywood, S., Fuentealba, I. C., Kemp, S. J., & Trafford, J. (2001). Copper toxicosis in the Bedlington terrier: A diagnostic dilemma. *Journal of Small Animal Practice*, *42*(4), 181–185.

He, B., Zhu, R., Yang, H., Lu, Q., Wang, W., Song, L., Sun, X., Zhang, G., Li, S., Yang, J., Tian, G., Bing, P., & Lang, J. (2020). Assessing the Impact of Data Preprocessing on

Analyzing Next Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, *8*. https://www.frontiersin.org/articles/10.3389/fbioe.2020.00817

Hill, W. G. (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, *33*(2), Article 2. https://doi.org/10.1038/hdy.1974.89

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, *38*(6), 226–231. https://doi.org/10.1007/BF01245622

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), 1–11.

Holloway, J. W., & Prescott, S. L. (2017). Chapter 2—The Origins of Allergic Disease. In R. E. O'Hehir, S. T. Holgate, & A. Sheikh (Eds.), *Middleton's Allergy Essentials* (pp. 29–50). Elsevier. https://doi.org/10.1016/B978-0-323-37579-5.00002-7

Holopainen, S., Hytönen, M. K., Syrjä, P., Arumilli, M., Järvinen, A.-K., Rajamäki, M., & Lohi, H. (2017). ANLN truncation causes a familial fatal acute respiratory distress syndrome in Dalmatian dogs. *PLOS Genetics*, *13*(2), e1006625. https://doi.org/10.1371/journal.pgen.1006625

Hou, Y., Liu, G. E., Bickhart, D. M., Matukumalli, L. K., Li, C., Song, J., Gasbarre, L. C., Van Tassell, C. P., & Sonstegard, T. S. (2012). Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Functional & Integrative Genomics*, *12*, 81–92.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P.

(2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Huber, D., Voith von Voithenberg, L., & Kaigala, G. V. (2018). Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering*, *1*, 15–24. https://doi.org/10.1016/j.mne.2018.10.006

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Hyun, C., Lavulo, L. T., & Filippich, L. J. (2004). Evaluation of haplotypes associated with copper toxicosis in Bedlington Terriers in Australia. *American Journal of Veterinary Research*, *65*(11), 1573–1579.

Iannuzzi, A., Parma, P., & Iannuzzi, L. (2021). Chromosome Abnormalities and Fertility in Domestic Bovids: A Review. *Animals : An Open Access Journal from MDPI*, *11*(3), 802. https://doi.org/10.3390/ani11030802

Iannuzzi, L., Di Meo, G. P., Perucatti, A., Di Palo, R., & Zicarelli, L. (2001). 50, XY gonadal dysgenesis (Swyer's syndrome) in a female river buffalo (Bubalus bubalis). *Veterinary Record*, *148*(20), 634–635.

Insights, Q. D. (2021, February 17). Tips for Variant Filtering & Prioritization. *QIAGEN Digital Insights*. https://digitalinsights.qiagen.com/news/blog/clinical/challenges-in-variant-filtering/

Irish Cattle Breeding Federation (ICBF). (2022). *Beef Calving Stats, ICBF Web Application*. https://webapp.icbf.com/v2/app/beef-stats

Jagannathan, V., Drögemüller, C., Leeb, T., & Consortium (DBVDC), D. B. V. D. (2019). A comprehensive biomedical variant catalogue based on whole genome sequences of 582

dogs and eight wolves. *Animal Genetics*, *50*(6), 695–704. https://doi.org/10.1111/age.12834

Jäger, R. J., Anvret, M., Hall, K., & Scherer, G. (1990). A human XY female with a frame shift mutation in the candidate testis-determining gene SRY. *Nature*, *348*(6300), 452–454. https://doi.org/10.1038/348452a0

Jäger, R. J., Harley, V. R., Pfeiffer, R. A., Goodfellow, P. N., & Scherer, G. (1992). A familial mutation in the testis-determining gene SRY shared by both sexes. *Human Genetics*, *90*(4), 350–355. https://doi.org/10.1007/BF00220457

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*, 14061. https://doi.org/10.1038/ncomms14061

Jiang, Y., Jiang, Y., Wang, S., Zhang, Q., & Ding, X. (2019). Optimal sequencing depth design for whole genome re-sequencing in pigs. *BMC Bioinformatics*, *20*(1), 556. https://doi.org/10.1186/s12859-019-3164-z

Joiret, M., Mahachie John, J. M., Gusareva, E. S., & Van Steen, K. (2019). Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Mining*, *12*(1), 11. https://doi.org/10.1186/s13040-019-0199-7

Kadri, N. K., Sahana, G., Charlier, C., Iso-Touru, T., Guldbrandtsen, B., Karim, L., Nielsen, U. S., Panitz, F., Aamand, G. P., & Schulman, N. (2014). A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: Additional evidence for the common occurrence of balancing selection in livestock. *PLoS Genetics*, *10*(1), e1004049.

Kanai, Y., Hiramatsu, R., Matoba, S., & Kidokoro, T. (2005). From SRY to SOX9: Mammalian testis differentiation. *Journal of Biochemistry*, *138*(1), 13–19. https://doi.org/10.1093/jb/mvi098

Kane, D. T., Dempsey, M. A., Burke, A. L., & Morrison, J. J. (2023). Prenatal Diagnosis and Fetal Sonographic Features of Swyer Syndrome. *Journal of Fetal Medicine*, 36–39. https://doi.org/10.1055/s-0043-57035

Kanzi, A. M., San, J. E., Chimukangara, B., Wilkinson, E., Fish, M., Ramsuran, V., & de Oliveira, T. (2020). Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. *Frontiers in Genetics*, *11*. https://www.frontiersin.org/articles/10.3389/fgene.2020.544162

Karch, S. B. (2016). Autopsy: Molecular. In J. Payne-James & R. W. Byard (Eds.), *Encyclopedia of Forensic and Legal Medicine (Second Edition)* (pp. 290–296). Elsevier. https://doi.org/10.1016/B978-0-12-800034-2.00041-0

Kaukonen, M., Woods, S., Ahonen, S., Lemberg, S., Hellman, M., Hytönen, M. K., Permi, P., Glaser, T., & Lohi, H. (2018). Maternal Inheritance of a Recessive RBP4 Defect in Canine Congenital Eye Disease. *Cell Reports*, *23*(9), 2643–2652. https://doi.org/10.1016/j.celrep.2018.04.118

Kawakura, K., Miyake, Y.-I., Murakami, R. K., Kondoh, S., Hirata, T.-I., & Kaneda, Y. (1997). Abnormal structure of the Y chromosome detected in bovine gonadal hypoplasia (XY female) by FISH. *Cytogenetic and Genome Research*, *76*(1–2), 36–38. https://doi.org/10.1159/000134508

Kawakura, K., Miyake, Y.-I., Murakami, R.-K., Kondoh, S., Hirata, T. -l., & Kaneda, Y. (1996). Deletion of the SRY region on the Y chromosome detected in bovine gonadal hypoplasia

(XY female) by PCR. *Cytogenetic and Genome Research*, *72*(2–3), 183–184. https://doi.org/10.1159/000134182

Kenney, S. M., & Cox, D. W. (2007). Sequence variation database for the Wilson disease copper transporter, ATP7B. *Human Mutation*, *28*(12), 1171–1177. https://doi.org/10.1002/humu.20586

Khare, J., Deb, P., Srivastava, P., & Reddy, B. H. (2017). Swyer syndrome: The gender swayer? *Alexandria Journal of Medicine*, *53*(2), 197–200. https://doi.org/10.1016/j.ajme.2016.05.006

Kilpinen, H., & Barrett, J. C. (2013). How next-generation sequencing is transforming complex disease genetics. *Trends in Genetics: TIG*, *29*(1), 23–30. https://doi.org/10.1016/j.tig.2012.10.001

Kim, S. Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Pedersen, O., Wang, J., & Nielsen, R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. *Genetic Epidemiology*, *34*(5), 479–491. https://doi.org/10.1002/gepi.20501

Kim, Y. G., Kim, S. Y., Kim, J. H., Lee, K. K., & Yun, Y. M. (2016). Prevalence and Clinical Relevance of Exon 2 Deletion of COMMD1 in Bedlington Terriers in Korea. *Journal of Veterinary Internal Medicine*, *30*(6), 1846–1850. https://doi.org/10.1111/jvim.14590

King, R. C., Mulligan, P. K., & Stansfield, W. D. (2014). In-frame mutation. In *A Dictionary of Genetics*. Oxford University Press. https://www.oxfordreference.com/display/10.1093/acref/9780199766444.001.0001/acref-9780199766444-e-3381

Kishikawa, T., Momozawa, Y., Ozeki, T., Mushiroda, T., Inohara, H., Kamatani, Y., Kubo, M., & Okada, Y. (2019). Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Scientific Reports*, *9*(1), 1–10.

Klomp, A. E. M., van de Sluis, B., Klomp, L. W. J., & Wijmenga, C. (2003). The ubiquitously expressed MURR1 protein is absent in canine copper toxicosis. *Journal of Hepatology*, *39*(5), 703–709. https://doi.org/10.1016/S0168-8278(03)00380-5

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*(1), 91. https://doi.org/10.1186/s13073-020-00791-w

Köhler, B., Lin, L., Ferraz-de-Souza, B., Wieacker, P., Heidemann, P., Schröder, V., Biebermann, H., Schnabel, D., Grüters, A., & Achermann, J. C. (2008). Five novel mutations in steroidogenic factor 1 (SF1, NR5A1) in 46,XY patients with severe underandrogenization but without adrenal insufficiency. *Human Mutation*, *29*(1), 59–64. https://doi.org/10.1002/humu.20588

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., Munson, K. M., Hastie, A. R., Diekhans, M., Hormozdiari, F., Lorusso, N., Hoekzema, K., Qiu, R., Clark, K., Raja, A., … Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science (New York, N.Y.)*, *360*(6393), eaar6343. https://doi.org/10.1126/science.aar6343

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, *22*(2), Article 2. https://doi.org/10.1038/9642

Kulski, J. K. (2016). Next-generation sequencing—An overview of the history, tools, and "Omic" applications. *Next Generation Sequencing-Advances, Applications and Challenges*, *10*, 61964.

Lam, H. Y., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'huallachain, M., Dewey, F. E., Habegger, L., Ashley, E. A., & Gerstein, M. B. (2012). Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology*, *30*(1), 78–82.

Lathrop, G. M., Lalouel, J. M., Julier, C., & Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences*, *81*(11), 3443–3446.

Lee, S.-A., Fillipich, L. J., & Hyun, C. (2007). Prevalence of the exon 2 deletion of the COMMD1 gene in Australian Bedlington terriers. *Journal of Genetics*, *86*(3), 289–291. https://doi.org/10.1007/s12041-007-0039-2

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), Article 7616. https://doi.org/10.1038/nature19057

Lewis, C. M., & Knight, J. (2012). Introduction to genetic association studies. *Cold Spring Harbor Protocols*, *2012*(3), pdb-top068163.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, N., Zhou, Y., Du, L., Wei, M., & Chen, X. (2011). Overview of Cytochrome P450 1B1 gene mutations in patients with primary congenital glaucoma. *Experimental Eye Research*, *93*(5), 572–579. https://doi.org/10.1016/j.exer.2011.07.009

Linder, M. C., Wooten, L., Cerveza, P., Cotton, S., Shulze, R., & Lomeli, N. (1998). Copper transport. *The American Journal of Clinical Nutrition*, *67*(5 Suppl), 965S-971S. https://doi.org/10.1093/ajcn/67.5.965S

Litonjua, A. A., & Celedón, J. C. (2006). GENETICS | Gene Association Studies. In G. J. Laurent & S. D. Shapiro (Eds.), *Encyclopedia of Respiratory Medicine* (pp. 240–247). Academic Press. https://doi.org/10.1016/B0-12-370879-6/00165-4

Liu, B. H. (2017). *Statistical genomics: Linkage, mapping, and QTL analysis*. CRC press.

Machado, C., Pereira, A., Cruz, J. M., Cadilhe, A., Silva, A., & Pereira, A. (2014). A novel SRY nonsense mutation in a case of Swyer syndrome. *Journal of Pediatric and Neonatal Individualized Medicine (JPNIM)*, *3*(1), e030107–e030107.

Mak, C. M., Lam, C.-W., Tam, S., Lai, C.-L., Chan, L.-Y., Fan, S.-T., Lau, Y.-L., Lai, J.-Y., Yuen, P., Hui, J., Fu, C.-C., Wong, K.-S., Mak, W.-L., Tze, K., Tong, S.-F., Lau, A., Leung, N., Hui, A., Cheung, K.-M., … Chan, Y.-W. (2008). Mutational analysis of 65 Wilson disease patients in Hong Kong Chinese: Identification of 17 novel mutations and its genetic heterogeneity. *Journal of Human Genetics*, *53*(1), Article 1. https://doi.org/10.1007/s10038-007-0218-2

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, *9*(1), 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359

Mardis, E. R., & Wilson, R. K. (2009). Cancer genome sequencing: A review. *Human Molecular Genetics*, *18*(R2), R163–R168.

Mason, R. J., Slutsky, A., Murray, J. F., Nadel, J. A., & Gotway, M. B. (2015). *Murray & Nadel's Textbook of Respiratory Medicine E-Book*. Elsevier Health Sciences.

McElreavey, K. D., Vilain, E., Boucekkine, C., Vidaud, M., Jaubert, F., Richaud, F., & Fellous, M. (1992). XY sex reversal associated with a nonsense mutation in SRY. *Genomics*, *13*(3), 838–840. https://doi.org/10.1016/0888-7543(92)90164-n

McElreavey, K., Jorgensen, A., Eozenou, C., Merel, T., Bignon-Topalovic, J., Tan, D. S., Houzelstein, D., Buonocore, F., Warr, N., Kay, R. G. G., Peycelon, M., Siffroi, J.-P., Mazen, I., Achermann, J. C., Shcherbak, Y., Leger, J., Sallai, A., Carel, J.-C., Martinerie, L., … Bashamboo, A. (2020). Pathogenic variants in the DEAH-box RNA helicase DHX37 are a frequent cause of 46,XY gonadal dysgenesis and 46,XY testicular regression syndrome. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *22*(1), 150–159. https://doi.org/10.1038/s41436-019-0606-y

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Medici, V., & LaSalle, J. M. (2019). Genetics and epigenetic factors of Wilson disease. *Annals of Translational Medicine*, *7*(Suppl 2), Article Suppl 2. https://doi.org/10.21037/atm.2019.01.67

Mercer, J. F. B., Livingston, J., Hall, B., Paynter, J. A., Begy, C., Chandrasekharappa, S., Lockhart, P., Grimes, A., Bhave, M., Siemieniak, D., & Glover, T. W. (1993). Isolation of a partial candidate gene for Menkes disease by positional cloning. *Nature Genetics*, *3*(1), Article 1. https://doi.org/10.1038/ng0193-20

Michala, L., Goswami, D., Creighton, S. M., & Conway, G. S. (2008). Swyer syndrome: Presentation and outcomes. *BJOG: An International Journal of Obstetrics & Gynaecology*, *115*(6), 737–741.

Miller, W. L., & Auchus, R. J. (2011). The Molecular Biology, Biochemistry, and Physiology of Human Steroidogenesis and Its Disorders. *Endocrine Reviews*, *32*(1), 81–151. https://doi.org/10.1210/er.2010-0013

Mistry, M., Ho Sui, S., Liu, J., Piper, M., Gammerdinger, W., & Khetani, R. (2021, September 15). *Troubleshooting the ChIP-seq workflow: From sequence reads to peak calls*. Introduction to ChIP-Seq Using High-Performance Computing. https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/troubleshooting_chipseq_partI.html

Modena, B. D., Doroudchi, A., Patel, P., & Sathish, V. (2019). Chapter 18—Leveraging genomics to uncover the genetic, environmental and age-related factors leading to

asthma. In G. S. Ginsburg, H. F. Willard, E. L. Tsalik, & C. W. Woods (Eds.), *Genomic and Precision Medicine (Third Edition)* (pp. 331–381). Academic Press. https://doi.org/10.1016/B978-0-12-801496-7.00018-6

Mohiuddin, M., Kooy, R. F., & Pearson, C. E. (2022). De novo mutations, genetic mosaicism and human disease. *Frontiers in Genetics.-Place of Publication Unknown*, *13*, 1–20.

Mölder, F., Jablonski, K. P., & Letcher, B. (2021). *Sustainable data analysis with Snakemake.[version 2; peer review: 2 approved]. F1000Research 10: 33*.

Motsinger, A. A., Reif, D. M., Fanelli, T. J., Davis, A. C., & Ritchie, M. D. (2007). Linkage Disequilibrium in Genetic Association Studies Improves the Performance of Grammatical Evolution Neural Networks. *Proceedings of the ... IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, *2007*, 1–8.

Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*(R2), R131–R136.

Müller, J., Schwartz, M., & Skakkebaek, N. E. (1992). Analysis of the sex-determining region of the Y chromosome (SRY) in sex reversed patients: Point-mutation in SRY causing sex-reversion in a 46,XY female. *The Journal of Clinical Endocrinology and Metabolism*, *75*(1), 331–333. https://doi.org/10.1210/jcem.75.1.1619028

Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the Basics of NGS: From Mechanism to Variant Calling. *Current Genetic Medicine Reports*, *3*(4), 158–165. https://doi.org/10.1007/s40142-015-0076-8

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297–304.

Nakaichi, M., Iseri, T., Horikirizono, H., Itoh, H., Sunahara, H., Nemoto, Y., Itamoto, K., & Tani, K. (2022). Pedigree study of the heredity of copper-associated hepatitis in Dalmatians in Japan. *The Canadian Veterinary Journal*, *63*(6), 633–636.

Nakaichi, M., Iseri, T., Horikirizono, H., Komine, M., Itoh, H., Sunahara, H., Nemoto, Y., Itamoto, K., & Tani, K. (2021). Copper-associated hepatitis in a young Dalmatian dog in Japan. *The Journal of Veterinary Medical Science*, *83*(6), 911–915. https://doi.org/10.1292/jvms.21-0061

Nanji, M. S., & Cox, D. W. (1999). The copper chaperone Atox1 in canine copper toxicosis in Bedlington terriers. *Genomics*, *62*(1), 108–112.

Navarre, C. B. (2010). Infertility in cattle. *Animal Health Notes*.

Nicastro, E., Ranucci, G., Vajro, P., Vegnente, A., & Iorio, R. (2010). Re-evaluation of the diagnostic criteria for Wilson disease in children with mild liver disease. *Hepatology*, *52*(6), 1948–1956. https://doi.org/10.1002/hep.23910

Nowacka-Woszuk, J., & Switonski, M. (2009). Differentiated Evolutionary Conservatism and Lack of Polymorphism of Crucial Sex Determination Genes (SRY and SOX9) in Four Species of the Family Canidae. *Folia Biologica*, *57*(3–4), 171–176. https://doi.org/10.3409/fb57_3-4.171-176

O'Connor, B. D., & Van Auwera, G. (2020). *Genomics in the cloud: Using Docker, Gatk, and WDL in Terra*. O'Reilly Media Sebastopol, CA, USA:

Oetting, W. S. (2012). Exome and genome analysis as a tool for disease identification and treatment: The 2011 Human Genome Variation Society scientific meeting. *Human Mutation*, *33*(3), 586–590. https://doi.org/10.1002/humu.22018

Ono, M., & Harley, V. R. (2013). Disorders of sex development: New genes, new concepts. *Nature Reviews Endocrinology*, *9*(2), 79–91.

Ostrander, E. A., Wang, G.-D., Larson, G., vonHoldt, B. M., Davis, B. W., Jagannathan, V., Hitte, C., Wayne, R. K., & Zhang, Y.-P. (2019). Dog10K: An international sequencing effort to advance studies of canine domestication, phenotypes and health. *National Science Review*, *6*(4), 810–824. https://doi.org/10.1093/nsr/nwz049

Ostrer, H. (2019). Swyer syndrome. *NORD (National Organization for Rare Disorders)*. https://rarediseases.org/rare-diseases/swyer-syndrome/

*Overview of Structural Variation*. (n.d.). Retrieved May 17, 2023, from https://www.ncbi.nlm.nih.gov/dbvar/content/overview/

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, *15*(2), 256–278. https://doi.org/10.1093/bib/bbs086

Park, H.-D., Park, H.-K., Chung, H.-S., Lee, S.-Y., Kim, J.-W., & Ki, C.-S. (2010). Association of ATP7BMutation Detection Rate with Biochemical Characteristics in Korean Patients with Wilson Disease. *Annals of Clinical & Laboratory Science*, *40*(1), 15–19.

Pausch, H., Kölle, S., Wurmser, C., Schwarzenbacher, H., Emmerling, R., Jansen, S., Trottmann, M., Fuerst, C., Götz, K.-U., & Fries, R. (2014). A nonsense mutation in TMEM95

encoding a nondescript transmembrane protein causes idiopathic male subfertility in cattle. *PLoS Genetics*, *10*(1), e1004044.

Pearlman, A., Loke, J., Le Caignec, C., White, S., Chin, L., Friedman, A., Warr, N., Willan, J., Brauer, D., Farmer, C., Brooks, E., Oddoux, C., Riley, B., Shajahan, S., Camerino, G., Homfray, T., Crosby, A. H., Couper, J., David, A., … Ostrer, H. (2010). Mutations in MAP3K1 cause 46,XY disorders of sex development and implicate a common signal transduction pathway in human testis determination. *American Journal of Human Genetics*, *87*(6), 898–904. https://doi.org/10.1016/j.ajhg.2010.11.003

Pedersen, B. S., Layer, R., & Quinlan, A. R. (2020). *smoove: Structural-variant calling and genotyping with existing tools* (0.2.8) [Go]. https://github.com/brentp/smoove/blob/c68ae05317dc5cedcc848b0389fd1a6b9d0fb92e/CITATION.cff (Original work published 2018)

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics (Oxford, England)*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699

Pedersen, B. S., & Quinlan, A. R. (2019). Duphold: Scalable, depth-based annotation and curation of high-confidence structural variant calls. *GigaScience*, *8*(4), giz040. https://doi.org/10.1093/gigascience/giz040

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, *39*(10), 1256–1260. https://doi.org/10.1038/ng2123

Persoskie, A., & Ferrer, R. A. (2017). A Most Odd Ratio: Interpreting and Describing Odds Ratios. *American Journal of Preventive Medicine*, *52*(2), 224–228. https://doi.org/10.1016/j.amepre.2016.07.030

Porterfield, A. (2015, May 15). *A Guide to Genetic Variants*. https://bitesizebio.com/23996/whats-so-important-about-variants/

Rantsiou, K., Kathariou, S., Winkler, A., Skandamis, P., Saint-Cyr, M. J., Rouzeau-Szynalski, K., & Amézquita, A. (2018). Next generation microbiological risk assessment: Opportunities of whole genome sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment. *International Journal of Food Microbiology*, *287*, 3–9. https://doi.org/10.1016/j.ijfoodmicro.2017.11.007

Reindollar, R. H., Byrd, J. R., Hahn, D. H., Haseltine, F. P., & McDonough, P. G. (1987). A cytogenetic and endocrinologic study of a set of monozygotic isokaryotic 45,X/46,XY twins discordant for phenotypic sex: Mosaicism versus chimerism. *Fertility and Sterility*, *47*(4), 626–633. https://doi.org/10.1016/s0015-0282(16)59113-4

Rimoin, D. L., Pyeritz, R. E., & Korf, B. (2013). *Emery and Rimoin's essential medical genetics*. Elsevier.

Sahana, G., Nielsen, U. S., Aamand, G. P., Lund, M. S., & Guldbrandtsen, B. (2013). Novel harmful recessive haplotypes identified for fertility traits in Nordic Holstein cattle. *PLoS One*, *8*(12), e82909.

Salmela, E., Niskanen, J., Arumilli, M., Donner, J., Lohi, H., & Hytönen, M. K. (2019). A novel *KRT71* variant in curly-coated dogs. *Animal Genetics*, *50*(1), 101–104. https://doi.org/10.1111/age.12746

Sanger, F., Nicklen, S., & Coulson, A. R. (1992). DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology (Reading, Mass.)*, *24*, 104–108.

Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J., & Nelson, S. F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics (Oxford, England)*, *27*(19), 2648–2654. https://doi.org/10.1093/bioinformatics/btr462

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., … Sherry, S. T. (2021). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *49*(D1), D10–D17. https://doi.org/10.1093/nar/gkaa892

Scheinberg, I. H., & Sternlieb, I. (1965). Wilson's disease. *Annual Review of Medicine*, *16*(1), 119–134.

Schmutz, S. M., Barth, A. D., & Moker, J. S. (1994). A Klinefelter bull with a 1;29 translocation born to a fertile 61,XXX cow. *The Canadian Veterinary Journal*, *35*(3), 182–184.

Schwarzenbacher, H., Fuerst, C., Fuerst-Waltl, B., & Dolezal, M. (2012). A genomewide search for harmful recessive haplotypes in Brown Swiss and Fleckvieh cattle In: 63rd EAAP Meeting Book of Abstracts. *Bratislava, Slovakia*.

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (n.d.). *Piercing the dark matter: Bioinformatics of long-range sequencing and mapping | Nature Reviews Genetics*. Retrieved May 30, 2022, from https://www.nature.com/articles/s41576-018-0003-4

Sefid Dashti, M. J., & Gamieldien, J. (2017). A practical guide to filtering and prioritizing genetic variants. *BioTechniques*, *62*(1), 18–30. https://doi.org/10.2144/000114492

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews. Genetics*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

Sonstegard, T. S., Cole, J. B., VanRaden, P. M., Van Tassell, C. P., Null, D. J., Schroeder, S. G., Bickhart, D., & McClure, M. C. (2013). Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. *PloS One*, *8*(1), e54872.

Stachowiak, M., Szczerbal, I., & Switonski, M. (2016). Chapter Eight—Genetics of Adiposity in Large Animal Models for Human Obesity—Studies on Pigs and Dogs. In Y.-X. Tao (Ed.), *Progress in Molecular Biology and Translational Science* (Vol. 140, pp. 233–270). Academic Press. https://doi.org/10.1016/bs.pmbts.2016.01.001

Ståhl, P. L., & Lundeberg, J. (2012). Toward the Single-Hour High-Quality Genome. *Annual Review of Biochemistry*, *81*(1), 359–378. https://doi.org/10.1146/annurev-biochem-060410-094158

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, *61*, 437–455.

Steckler, T. (2014, January 21). *Cause of infertility in cattle found: University of Illinois Extension*. https://extension.illinois.edu/blogs/cattle-blog/2014-01-21-cause-infertility-cattle-found

Straiton, J., Free, T., Sawyer, A., & Martin, J. (2019). *From Sanger sequencing to genome databases and beyond*. Future Science.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi,

X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Swyer, G. I. M. (1955). Male Pseudohermaphroditism: A Hitherto Undescribed Form. *British Medical Journal*, *2*(4941), 709–712.

*Swyer syndrome: MedlinePlus Genetics*. (2022, February 25). https://medlineplus.gov/genetics/condition/swyer-syndrome/

Tanzi, R. E., Petrukhin, K., Chernov, I., Pellequer, J. L., Wasco, W., Ross, B., Romano, D. M., Parano, E., Pavone, L., Brzustowicz, L. M., Devoto, M., Peppercorn, J., Bush, A. I., Sternlieb, I., Pirastu, M., Gusella, J. F., Evgrafov, O., Penchaszadeh, G. K., Honig, B., … Gilliam, T. C. (1993). The Wilson disease gene is a copper transporting ATPase with homology to the Menkes disease gene. *Nature Genetics*, *5*(4), Article 4. https://doi.org/10.1038/ng1293-344

Teare, M. D., & Santibañez Koref, M. F. (2014). Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing. *Briefings in Functional Genomics*, *13*(5), 378–383. https://doi.org/10.1093/bfgp/elu024

THE BOVINE HAPMAP CONSORTIUM, Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., Green, R. D., Hamernik, D. L., Kappes, S. M., Lien, S., Matukumalli, L. K., McEwan, J. C., Nazareth, L. V., Schnabel, R. D., Weinstock, G. M., Wheeler, D. A., Ajmone-Marsan, P., Boettcher, P. J., … Dodds, K. G. (2009). Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science*, *324*(5926), 528–532. https://doi.org/10.1126/science.1167936

*The MISO Sequence Ontology Browser—STOP_GAINED*. (n.d.). Retrieved August 30, 2023, from http://www.sequenceontology.org/browser/current_svn/term/SO:0001587

Thornburg, L. P. (1998). Histomorphological and immunohistochemical studies of chronic active hepatitis in Doberman Pinschers. *Veterinary Pathology*, *35*(5), 380–385. https://doi.org/10.1177/030098589803500507

Thornburg, L. P., Rottinghaus, G., McGowan, M., Kupka, K., Crawford, S., & Forbes, S. (1990). Hepatic copper concentrations in purebred and mixed-breed dogs. *Veterinary Pathology*, *27*(2), 81–88. https://doi.org/10.1177/030098589002700202

Thornburg, L. P., Shaw, D., Dolan, M., Raisbeck, M., Crawford, S., Dennis, G. L., & Olwin, D. B. (1986). Hereditary copper toxicosis in West Highland white terriers. *Veterinary Pathology*, *23*(2), 148–154. https://doi.org/10.1177/030098588602300207

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111.

Twedt, D. C., Sternlieb, I., & Gilbertson, S. R. (1979). Clinical, morphologic, and chemical studies on copper toxicosis of Bedlington Terriers. *Journal of the American Veterinary Medical Association*, *175*(3), 269–275.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), Article 1. https://doi.org/10.1038/s43586-021-00056-9

Valour, D., Michot, P., Eozenou, C., Lefebvre, R., Bonnet, A., Capitan, A., Uzbekova, S., Sellem, E., Ponsart, C., & Schibler, L. (2015). Dairy cattle reproduction is a tightly regulated genetic process: Highlights on genes, pathways, and biological processes. *Animal Frontiers*, *5*(1), 32–41. https://doi.org/10.2527/af.2015-0006

van de Sluis, B. J., Breen, M., Nanji, M., van Wolferen, M., de Jong, P., Binns, M. M., Pearson, P. L., Kuipers, J., Rothuizen, J., & Cox, D. W. (1999). Genetic mapping of the copper toxicosis locus in Bedlington terriers to dog chromosome 10, in a region syntenic to human chromosome region 2p13–p16. *Human Molecular Genetics*, *8*(3), 501–507.

van De Sluis, B., Rothuizen, J., Pearson, P. L., van Oost, B. A., & Wijmenga, C. (2002). Identification of a new copper metabolism gene by positional cloning in a purebred dog population. *Human Molecular Genetics*, *11*(2), 165–173. https://doi.org/10.1093/hmg/11.2.165

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*(9), 418–426.

VanRaden, P. M., Olson, K. M., Null, D. J., & Hutchison, J. L. (2011). Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *Journal of Dairy Science*, *94*(12), 6153–6161.

Veeckman, E., Van Glabeke, S., Haegeman, A., Muylle, H., van Parijs, F. R. D., Byrne, S. L., Asp, T., Studer, B., Rohde, A., Roldán-Ruiz, I., Vandepoele, K., & Ruttink, T. (2019). Overcoming challenges in variant calling: Exploring sequence diversity in candidate genes for plant development in perennial ryegrass (Lolium perenne). *DNA Research*, *26*(1), 1–12. https://doi.org/10.1093/dnares/dsy033

Veltman, J. A., & Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics*, *13*(8), 565–575.

Vulpe, C., Levinson, B., Whitney, S., Packman, S., & Gitschier, J. (1993). Isolation of a candidate gene for Menkes disease and evidence that it encodes a copper-transporting ATPase. *Nature Genetics*, *3*(1), 7–13. https://doi.org/10.1038/ng0193-7

Waldner, C. (2019, November 26). *Preventing Reproductive Failure in Cow-Calf Herds—Beef Cattle Research Council*. Preventing Reproductive Failure in Cow-Calf Herds - Beef Cattle Research Council. http://www.beefresearch.ca/research-topic.cfm

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, Q., Lu, Q., & Zhao, H. (2015). A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Frontiers in Genetics*, *6*, 149. https://doi.org/10.3389/fgene.2015.00149

Webb, C. B., Twedt, D. C., & Meyer, D. J. (2002). Copper-associated liver disease in Dalmatians: A review of 10 dogs (1998–2001). *Journal of Veterinary Internal Medicine*, *16*(6), 665–668.

Weiskirchen, R., & Penning, L. C. (2021). COMMD1, a multi-potent intracellular protein involved in copper homeostasis, protein trafficking, inflammation, and cancer. *Journal of Trace Elements in Medicine and Biology*, *65*, 126712. https://doi.org/10.1016/j.jtemb.2021.126712

Wu, P., Wang, K., Zhou, J., Chen, D., Yang, Q., Yang, X., Liu, Y., Feng, B., Jiang, A., & Shen, L. (2019). GWAS on imputed whole-genome Resequencing from genotyping-by-

sequencing data for farrowing interval of different parities in pigs. *Frontiers in Genetics*, 1012.

Wu, X., Chien, H., van Wolferen, M. E., Kruitwagen, H. S., Oosterhoff, L. A., & Penning, L. C. (2019). Reduced FXR Target Gene Expression in Copper-Laden Livers of COMMD1-Deficient Dogs. *Veterinary Sciences*, *6*(4), Article 4. https://doi.org/10.3390/vetsci6040078

Wu, X., den Boer, E. R., Vos-Loohuis, M., van Steenbeek, F. G., Monroe, G. R., Nijman, I. J., Leegwater, Peter. A. J., & Fieten, H. (2020). Investigation of Genetic Modifiers of Copper Toxicosis in Labrador Retrievers. *Life*, *10*(11), 266. https://doi.org/10.3390/life10110266

Wu, X., Mandigers, P. J. J., Fieten, H., & Leegwater, P. A. (2020). Evaluation of COMMD1 in copper toxicosis in Labrador retrievers and Dobermans. *Veterinary Journal (London, England: 1997)*, *265*, 105561. https://doi.org/10.1016/j.tvjl.2020.105561

Wu, X., Mandigers, P. J. J., Watson, A. L., van den Ingh, T. S. G. A. M., Leegwater, P. A. J., & Fieten, H. (2019). Association of the canine ATP7A and ATP7B with hepatic copper accumulation in Dobermann dogs. *Journal of Veterinary Internal Medicine*, *33*(4), 1646–1652. https://doi.org/10.1111/jvim.15536

Xing, K., Zhu, F., Zhai, L., Chen, S., Tan, Z., Sun, Y., Hou, Z., & Wang, C. (2016). Identification of genes for controlling swine adipose deposition by integrating transcriptome, whole-genome resequencing, and quantitative trait loci data. *Scientific Reports*, *6*, 23219. https://doi.org/10.1038/srep23219

Yuzbasiyan-Gurkan, V., Blanton, S. H., Cao, Y., Ferguson, P., Li, J., Venta, P. J., & Brewer, G. J. (1997). Linkage of a microsatellite marker to the canine copper toxicosis locus in Bedlington terriers. *American Journal of Veterinary Research*, *58*(1), 23–27.

Zhang, H., Du, Z.-Q., Dong, J.-Q., Wang, H.-X., Shi, H.-Y., Wang, N., Wang, S.-Z., & Li, H. (2014). Detection of genome-wide copy number variations in two chicken lines divergently selected for abdominal fat content. *BMC Genomics*, *15*, 1–12.

Zhou, Y., Yang, L., Han, X., Han, J., Hu, Y., Li, F., Xia, H., Peng, L., Boschiero, C., Rosen, B. D., Bickhart, D. M., Zhang, S., Guo, A., Van Tassell, C. P., Smith, T. P. L., Yang, L., & Liu, G. E. (2022). Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. *Genome Research*, *32*(8), 1585–1601. https://doi.org/10.1101/gr.276550.122