Genomic Characterisation of New Zealand Working Farm Dogs

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

in

Animal Science

at Massey University, Manawatū, New Zealand.

Florence Smith

2025

Abstract

The majority of working farm dogs in New Zealand belong to either the Huntaway or Heading Dog breeds. These breeds are unique to New Zealand and are vital to efficient livestock farming, yet no large-scale DNA sequencing studies have been conducted on them to date. The Right Dog for the Job project aims to improve the health and performance of the farm dog population by analysing the genetic and phenotypic information from a sample of 2400 Huntaways and Heading Dogs, including 250 whole genome sequences. As the first use of this dataset, this project aimed to characterise the genetic variation in the sample of whole genome sequences. To this end, sequence bioinformatics tools were used to align sequence reads, then call, filter, and annotate 20 million genetic variants. The dataset was intersected with 395 OMIA-reported functional loci to identify 27 previously described Mendelian variants segregating in the population. Five of these (in the *SOD1*, *VWF*, *CUBN*, *CLN8*, and *SGSH* genes) were highlighted as compelling candidates for diagnostic selection. Next, the dataset was surveyed for high-impact variants segregating within 132 genes previously shown to harbour phenotype-causing variants. The aim here was to use variant effect prediction to identify novel causal variants within functionally relevant genes. This analysis yielded nine causal hypotheses (in the *CNGB1*, *ABCA4*, *CNP*, *SLC3A1*, *CCDC66*, *GLB1*, *CYP1A2*, and *STK36* genes) for future association testing.

The second stage of the project leveraged a sample of 299 dogs that were genotyped with the Axiom™ Canine HD Array. Of these dogs, 188 had also been whole genome sequenced. A linkage disequilibrium analysis was conducted to identify predictive markers on the array for several of the 27 Mendelian variants, with a view to supporting future marker-assisted selection. Next, the whole genome sequences were used as a reference to impute missing genotypes in the sample of 111 dogs genotyped with only the Axiom SNP chip, expanding the genomic dataset. Finally, GWAS were performed to identify genetic associations with four body size traits of interest: height, length, chest circumference, and muzzle circumference. Height was significantly associated with a region near *LCORL*, a gene known to regulate body size in various species.

Despite representing one of the first analyses of the ‘Right Dog’ research programme, this project has identified genes and variants that should be of immediate practical use to working dog breeders and owners. The project has also helped to generate an extensive genomic dataset that will underpin future research as part of the broader programme, and ultimately contribute to the health, welfare, and performance of New Zealand’s iconic Huntaway and Heading Dog breeds.

Acknowledgements

I would like to express my gratitude to my primary supervisor, Professor Matt Littlejohn. Matt has provided guidance, advice, and support throughout my entire project. He has gone above and beyond as a mentor, from helping me publish my first paper to facilitating my stipend from Massey University and ensuring that I pursue my dream PhD project. I am grateful to my secondary supervisor, Dr Thomas Lopdell, for his vital assistance throughout this project. His extensive knowledge of bioinformatic programming and readiness to answer my many questions has been greatly appreciated. Thank you to Dr Nick Sneddon, my other secondary supervisor, for providing me with feedback and support.

My deepest thanks go to my parents who are my biggest cheerleaders and who are always willing to listen and offer advice. I want to thank Milly Henry, who has been by my side throughout the course of our studies to discuss ideas, relate to my challenges, and work alongside me. Special thanks are due to Livestock Improvement Corporation (LIC) for supporting my studies with the Pat Shannon Scholarship and everyone in the LIC R&D department for answering my questions and welcoming me into their team. Thank you to everyone in the Right Dog for the Job project for inviting me into the group and providing me with suggestions. Finally, I am sincerely thankful to Massey University for supporting my studies with the Massey University Master's Research Scholarship and RMS Stipend.

Preface

Because this thesis was completed as part of the wider Right Dog for the Job project at Massey University, I would like to preface it by clarifying my role and contributions within the project. My research began during the sample collection phase of the Right Dog project, where a substantial proportion of biological samples had already been collected. Sample processing was a collaborative effort, though I was involved in the preparation of DNA kits for distribution, as well as the recording of data from surveys. I conducted all of the bioinformatic analyses described in this thesis, with the majority of software packages applied as analysis pipelines scripted and executed by me. However, I would like to acknowledge Thomas Lopdell, who wrote some of the custom scripts that I used. Namely, Thomas wrote the pipeline for BWA-MEM sequence alignment, the scripts to align RNA-Seq reads with STAR, the custom pearl script used to add the 'RNABamDepth' annotation to variants, and the LiftOver code for converting OMIA variants between reference genomes. I would also like to recognise Matt Littlejohn, who created the canFam4-specific database that was referenced in the SnpEff variant effect prediction. Finally, I would like to acknowledge Pam Stephen Photography for producing the images used in Figures 1 and 2.

All experiments were performed in strict accordance with the rules and guidelines outlined in the New Zealand Animal Welfare Act 1999. Samples were gathered in accordance with protocols approved by the Massey University Animal Ethics Committee, Palmerston North, New Zealand (approval MUAEC 23/37).

Table of Contents

List of abbreviations

Abbreviation	Definition
µl	Microlitre
AC	Allele count
AF	Allele frequency
A_{jk}	Pairwise relatedness score
ALS	Amyotrophic lateral sclerosis
AN	Allele number
ANOVA	Analysis of variance
BAM	Binary alignment map
bp	Base pair
BV	Breeding value
cDNA	Complementary DNA
cEDS	Classical Ehlers-Danlos syndrome
Chr	Chromosome
CNV	Copy number variation
D	Linkage disequilibrium coefficient
DBVDC	Dog Biomedical Variant Database Consortium
DM	Degenerative myelopathy
DNA	Deoxyribonucleic acid
Dog10K	Dog10K Consortium
DR^2	Dosage R-squared
DSD	Differences in sex development
eQTL	Expression quantitative trait locus
F	Inbreeding coefficient
FS	Fisher strand
GDV	Gastric dilation-volvulus
GLM	Generalised linear model
GOF	Gain of function
GRM	Genomic relationship matrix
GTF	General transfer format
GVCF	Genomic variant call format
GWAS	Genome-wide association study (or studies)
HD	High density
HMM	Hidden Markov model
HWE	Hardy-Weinberg equilibrium
IBD	Identical by descent
IGV	Integrative Genome Viewer
Indel	Small insertion or deletion
kb	Kilobase

kg	Kilogram
LAD	Lethal acrodermatitis
LD	Linkage disequilibrium
lncRNA	Long non-coding RNA
LOF	Loss of function
LPPN	Laryngeal paralysis and polyneuropathy
LS	Lundehund syndrome
M	Million
MAC	Minor allele count
MAF	Minor allele frequency
Mb	Megabases
MLMA-LOCO	Mixed linear model association with leave one chromosome out
mm	Millimeter
MPS	Mucopolysaccharidosis
MQ	RMS mapping quality
MQRankSum	Mapping quality rank sum test
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NCL	Neural ceroid lipofuscinosis
N_e	Effective population size
ng	Nanogram
NZ	New Zealand
OMIA	Online Mendelian Inheritance in Animals
PC	Principal component
PCA	Principal component analysis
PCD	Primary ciliary dyskinesia
PCR	Polymerase chain reaction
PPD	Preaxial polydactyly
PRS	Polygenic risk score
Q1	First quartile
Q3	Third quartile
QC	Quality control
QD	Quality by depth
QTL	Quantitative trait locus
R^2	Squared correlation coefficient
ReadPosRankSum	Read position rank sum test
Right Dog project	Right Dog for the Job project
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
SAM	Sequence alignment map

SE	Standard error
SNP	Single nucleotide polymorphism
SOR	Strand odds ratio
SRA	Sequence read archive
SV	Structural variant
TNS	Trapped neutrophil syndrome
UTR	Untranslated region
VCF	Variant call format
VEP	Variant effect predictor
VQSR	Variant quality score recalibration
vWD	von Willebrand disease
WES	Whole exome sequencing
WGS	Whole genome sequencing
WT	Wild type
ZRS	ZPA regulatory sequence
β	Estimated effect size

List of Figures

List of Tables

Chapter 1: Literature Review and Introduction

1.1 Literature Review

1.1.1 Introduction

The Right Dog for the Job (Right Dog) project is a three-year Massey University-led project co-funded by the Ministry for Primary Industries through the Sustainable Food and Fibre Futures fund. This large project has the overarching aim of improving the health, wellbeing, and performance of the working farm dog population in New Zealand (NZ) by investigating genetic, genomic, and phenotypic data. This project is the first large-scale genetic study of the Huntaway and Heading Dog breeds to date, with the aim of whole genome sequencing 250 dogs and genotyping 2400 dogs with the Axiom™ Canine HD Array (SNP chip) from Thermo Fisher. As an early stage of the larger Right Dog project, this research thesis used bioinformatic techniques to functionally characterise the genetic variation present in the Huntaway and Heading Dog breeds.

The following literature review summarises what is currently known about the NZ Huntaway and Heading Dog breeds and highlights some of the major genetic studies that have previously been conducted on dogs. Additionally, it describes bioinformatic techniques that are commonly used for variant discovery, many of which were employed in the following thesis. Key concepts for understanding genomic data are also explained, including linkage disequilibrium and variant annotation.

1.1.2 Overview of Dogs

Canis lupus familiaris (domestic dogs) are descended from the grey wolf and were the first animal to be domesticated by humans (1). They are thought to have arisen in East Asia around 33,000 years ago (2). Most modern dog breeds, however, did not exist until relatively recently (in the last 200 years), driven by artificial selection for different purposes. There are around 400 modern breeds that are characterised by their wide diversity in morphological and behavioural traits. Of all vertebrates, dogs have the

greatest intra-species phenotypic diversity but also exhibit large behavioural variation (3). Dogs have historically been selected for extremely diverse roles including hunting, herding, transportation, environmental protection, law enforcement work, assistance, and companionship (4). This differential selection led to extensive genetic and phenotypic variation between breeds. However, intense artificial selection of dogs has also led to increased inbreeding and a series of bottle necks, and therefore limited genetic variation is observed within breeds (5). Because of this, purebred dogs tend to have a higher prevalence of recessive genetic diseases than crossbred dogs (mutts). Interestingly, a phylogenetic analysis performed by Dutrow et al. (6) indicated that working farm dog breeds were more genetically diverged from wolves than were other breed lineages.

1.1.2.1 Selective Breeding in Working Dogs

Working farm dogs are an integral part of farming livestock in New Zealand (7). There are approximately 200,000 working farm dogs in New Zealand, most of which belong to one of two breeds: Huntaway (see Figure 1) and Heading Dog (see Figure 2). These dogs move livestock across large areas of hilly land in all conditions; therefore, they must be extremely fit, resilient, and strong. Other useful qualities include stamina, agility, speed, willingness to follow commands, ability to work with other dogs, and the ability to suppress predatorial instincts by not attacking livestock. It is important that working dogs have the correct body size and build to work efficiently on farm (8). For example, working dogs should have large enough chests to allow for sufficient lung capacity, but not so large that it interferes with gait. Similarly, they should be tall enough to run with sufficient speed, but extremely tall dogs carry too much added weight. Boyko et al. (9) estimated that a small number of quantitative trait loci (QTLs) explain the majority of phenotypic variation in dog morphology, suggesting that selection for a small number of genetic markers has led to rapid change.

Farmers rely heavily on working dogs to efficiently herd sheep and cattle, especially on rough New Zealand terrains that could not be accessed otherwise (7). As such, shepherds invest large amounts of their time and money to train their dogs. Historically,

breeding dogs have been selected based on working performance. However, the genetic gain achieved through this type of phenotypic selection is slow, since an individual's phenotype is not, on its own, the best predictor of their genetic merit (10). Genetic and genomic selection have been proven to successfully accelerate genetic change in many livestock improvement programmes, as well as in an American population of seeing eye dogs (11,12). A major advantage of genomic selection is that the generation interval can be shortened, since breeding animals can be selected before exhibiting phenotypes (10). Because no large-scale DNA sequencing studies have been conducted on NZ farm dogs to date, little is known about their genetic make-up. This limits the potential for future improvement in these breeds, since advantageous and deleterious variants cannot be selected for or against.

1.1.2.2 Overview of New Zealand Huntaway and Heading Dog Breeds

While Huntaways and Heading Dogs represent most farm dogs in NZ, they are little-known elsewhere (13). Huntaways are thought to have descended from herding dogs and are traditionally bred for their deep bark and endurance. They are considered a large breed, with an average size of ~28kg and usually have black and tan coats (see Figure 1) (14). Because they are traditionally bred for their working ability, rather than their appearance, Huntaway morphology is highly variable. Huntaways can perform many roles on farms, including heading, hunting, and yard work. They are most well known for their loud, deep barks that allow them to drive stock away from the handler and work from large distances. Heading Dogs are smaller than Huntaways, averaging ~19kg and tend to be black and white with smooth coats (see Figure 2) (13). Unlike Huntaways, they are discouraged from barking and instead are bred to stare down livestock in close quarters. Heading Dogs are thought to have descended from the Border Collie and are used in a similar way on farms (i.e. to control animals from the front and sides).

Disease susceptibility in dogs tends to be breed specific, both due to genetic variation and breed-specific environmental factors. Isaksen et al. (7) performed a longitudinal study in the South Island of NZ and determined that 74% of the 641 surveyed farm dogs presented with clinical abnormalities. Most of these abnormalities were diagnosed as

musculoskeletal illnesses/injuries or skin traumas. An earlier survey conducted on working dogs that presented to veterinary clinics across NZ determined that the diseases that caused the most loss were gastric dilation-volvulus (GDV), degenerative joint diseases, mammary neoplasia, female reproductive diseases, and cardiac diseases (13). Compared to other breeds, Huntaways have a high prevalence of hip dysplasia, cardiomyopathy, and GDV (a.k.a twisted gut), which is thought to be associated with their large size and deep chests (13,15–17). Heading Dogs are overrepresented among cases of trauma induced injuries like hip luxation, Archilles tendon, and tarsal injuries (13). It is not known whether this is due to environmental or genetic factors. Most of the health surveys conducted on NZ working dogs, with the exceptions of Isaksen et al. (7) and Cave et al. (13), focus on single diseases and are biased toward unusual cases, rather than common diseases (15). Additionally, farm dogs live in rural areas and therefore may visit veterinary clinics less often than companionship breeds (7). Because of this, they are often culled or lost on farm without an official diagnosis, making it difficult to estimate the true prevalence of diseases in such breeds.

Figure 1. New Zealand Huntaway. (Pam Stephen Photography, 2024)

Figure 2. New Zealand Heading Dog. (Pam Stephen Photography, 2024)

1.1.3 Functional Variants

Genetic variants (a.k.a mutations) are defined as regions of the genome that differ between individuals in a population, resulting in alternative alleles (18). Most variants do not affect phenotypes (a.k.a traits) due to large intergenic regions and the redundancy of the genetic code (19). The neutral theory of molecular evolution states that while many mutations are deleterious, these are rapidly removed from populations by natural selection (20). However, mutations can cause molecular and phenotypic effects that may be advantageous, deleterious, or neutral to the host organism. Types of variants include single nucleotide polymorphisms/substitutions (SNPs), where a single nucleotide base is swapped, small insertions and deletions (indels), copy number variations (CNVs) such as duplications, and large structural variants (SVs) such as translocations, inversions, and chromosome rearrangements. While large variants are more likely to have effects, small variants can have large impacts, for example if they interrupt a coding sequence such that the protein function is lost (LOF) or a new function is gained (GOF) (21).

Functional variants are commonly defined as those that have a specific action on or affect the molecular function of a protein (e.g. gain, loss, or change of protein function) (18,22). Exons are the protein-coding regions of genes and tend to be the focus of functional studies (23). Coding-region variants can be synonymous, meaning they do not change the amino acid (protein) sequence, or non-synonymous, meaning they do (22). Synonymous variants are more commonly observed than non-synonymous variants because they are less likely to be removed from the population via natural or artificial selection. However, not all non-synonymous changes are functional and not all synonymous changes are non-functional (24).

Variants can have major phenotypic effects without directly changing protein sequences; however, these effects are less well understood by geneticists. For example, non-coding variants may affect protein expression by altering regulatory elements such as promoters, enhancers, and inhibitors (25,26). There are large intergenic regions of the genome that have no known function and, as such, have been termed ‘junk’ DNA (27). The role of this DNA is debated since some intergenic elements like retrotransposons and lncRNAs can

be transcribed and have important roles (28,29). Because exons make up a very small proportion of the genome (e.g. 3% in the canFam4 reference assembly), functional analyses can be greatly simplified by removing non-exonic regions while still retaining the majority of high-impact variants (30,31).

1.1.4 Marker-Assisted Selection

Marker-Assisted selection refers to the selective mating of individuals based on alleles associated with desirable or deleterious traits (22). The most rapid phenotypic change occurs when high-impact, dominantly inherited variants that have large effects on traits are selected (for or against) (32). For example, single genes/variants that control traits through simple inheritance patterns are commonly referred to as Mendelian variants (33). Discovery of such variants is of particular interest because their selection could lead to rapid genetic gain in populations. Complex traits are affected by many genes of small effect, so variants in these genes are unlikely to have major impacts on the overall phenotype (34). Low heritability traits are influenced more by the environment than by genetic factors and so variants in these genes are also less likely to have major impacts on the overall phenotype. The effect size of a variant, being the contribution of the variant to the genetic variance of the phenotype, is determined by the underlying biology (34). Dominant traits are observed when the causative allele is heterozygous or homozygous, while recessive traits are only observed when the allele is homozygous. Additive traits are observed in both cases, but the phenotype differs by the number of allele copies. Therefore, dominant traits spread rapidly throughout populations, while recessive traits can exist silently in the population without causing a phenotype until two carriers are mated.

Inbreeding can be defined and quantified at an individual or population level but commonly refers to the mating of individuals that share recent common ancestry (35). This practice leads to an excess of homozygosity, a decrease in genetic diversity, and a loss of fitness in populations, all of which contribute to inbreeding depression. Because of the increase in homozygosity, the likelihood of two recessive disease carriers mating is increased by inbreeding (32). ‘Popular sire phenomenon’ is an example of inbreeding that

occurs in artificial selection when a particular individual is seen as desirable and therefore sires a large proportion of the next generation (36). This phenomenon reduces genetic diversity and if the sire carries heritable defects, then the population frequency of these variants will increase. This is particularly bad if the trait is dominant. For example, a dominant genetic syndrome was rapidly spread across the New Zealand dairy cattle population when semen from a particular bull and his sons was used widely (32). These bulls had a missense mutation (*de novo* in the sire) in the prolactin (*PRL*) gene, encoding the prolactin hormone. This resulted in a hairy phenotype, along with heat stress and major lactation defects in many dairy cows. Outbreeding refers to the mating of unrelated individuals, often across populations or breeds (37). This practice increases heterosis, reducing the likelihood of two deleterious recessive carriers mating. This also explains why purebred dogs tend to be more susceptible to recessive genetic disease than crossbreeds (37).

1.1.5 DNA Sequencing

DNA sequencing is a laboratory technique whereby the order of nucleotides in a linear string of DNA is determined with varying levels of resolution based on the technology used (38). Whole genome sequencing (WGS) provides the most comprehensive and high-resolution view of genomes by capturing each base. By contrast, whole exome sequencing (WES) captures only the protein-coding regions of the genome. This greatly reduces the cost of sequencing while still capturing the most functionally relevant regions of the genome. Targeted sequencing technologies sequence specific regions of the genome that are of interest with high coverage (a.k.a read depth), meaning the same sequence is read many times to increase accuracy and throughput of samples. The first widely used DNA sequencing technology was developed in 1977 by Fred Sanger, with modern 'next-generation sequencing' taking its place in the early 2000s (39,40).

Before it is sequenced, DNA is usually extracted from a tissue sample, fragmented into smaller reads, and amplified (38). Sequence reads must then be reassembled into the correct order to determine which part of the genome they represent. Typically, reads are aligned to a species- or breed-specific reference assembly based on their matching

sequence. However, for species that do not have reference genomes, sequences can be assembled in a *de novo* manner based on the overlap of adjoining reads. *De novo* assembly is more computationally demanding but can identify novel sequences that reference-based approaches can't (41). Short-read technologies offer higher accuracy and lower cost over long-read technologies; however, they are usually limited to ~100-300bp reads that are not able to capture large SVs (41,42). Long-read technologies on the other hand can sequence reads larger than 10kb, capturing entire SVs and improving *de novo* genome assembly. Short-read sequencing is more error-prone than long-read sequencing in repetitive regions as reads may map anywhere in the repeat if the start and/or end are not captured (43).

1.1.5.1 Variant Calling

Variant calling is the computational process of identifying all genetic variants in a given nucleotide sequence (44). The accuracy of this process relies on the quality of DNA sequence data and requires sufficient sequencing coverage to ensure all variants are identified. Due to the inevitability of sequencing errors and misalignments, variant calling will almost certainly result in false positives and therefore must be followed by quality control (QC) processes. Additionally, false positives occur due to duplications during PCR amplification, obscuring allele frequencies (AFs). PCR duplications can be detected and removed during data pre-processing; however, this can result in the removal of real duplications/repetitive regions. Rare variants may be removed as false negatives in studies with small sample sizes or insufficient coverage due to their low frequencies. Depending on the sequencing technology and pipeline used, variant calling can be particularly error prone for large SVs, hypervariable regions, and repetitive regions. Additionally, false calls are made more frequently near indels than SNPs (45). The final output of variant calling is usually a population variant call format (VCF) file, which stores information about variants in each individual and the sample as a whole (46).

Many tools exist for variant calling that offer different advantages, but GATK HaplotypeCaller (47) is one of the most widely used. Following GATK best practices for germline SNP and indel discovery, HaplotypeCaller calls SNPs and indels simultaneously

via *de novo* assembly (48). Another widely used tool for variant calling is BCFtools (49). BCFtools was first released in 2010 as a subpackage of SAMtools, a software package for manipulating binary alignment map (BAM) and sequence alignment map (SAM) files. However, BCFtools is now its own package that was specifically constructed for variant calling and filtering. In the BCFtools pipeline, mpileup is used to determine genotype likelihoods based on nucleotide and read qualities, then likelihoods are evaluated based on Hardy-Weinberg equilibrium (HWE) and AFs. In a benchmarking study, Lefouili & Nam (50) concluded that mpileup had a better recovery rate and higher accuracy than HaplotypeCaller. However, GATK had a lower false positive rate when used in conjunction with BWA-MEM sequence alignment (51). Out of 15 recent studies comparing the tools, ten concluded that GATK was better, while three concluded that BCFtools was better and two were inconclusive.

Both GATK and BCFtools are alignment based, meaning they call variants based on direct alignment to a reference genome. FreeBayes (52) on the other hand, another common variant caller, uses a more generalised model to call variants based on haplotypes. The advantage of haplotype-based callers over alignment-based callers is that they are not affected by identical sequences having multiple possible alignments. However, Freebayes has shown reduced sensitivity and specificity when compared to GATK and BCFtools (53). Many modern variant calling pipelines combine multiple methods with the aim of increasing accuracy and identifying complex and large SVs. However, this leads to complex analyses. Comparative studies indicate that new deep-learning technologies, such as DeepVariant, may exhibit higher accuracy than all aforementioned tools (54,55). Despite the constant development of new variant callers, GATK and BCFtools remain the most widely used in small variant discovery pipelines (53).

1.1.5.2 Variant Filtering

Quality control filters are applied after variant calling with the aim of removing as many false positives as possible, while retaining true positives. GATK best practices outline two different methods for variant filtering: hard filtering, and variant quality score recalibration (VQSR) (48,56). With GATK hard filtering, numeric thresholds are set for six

quality metrics; quality by depth (QD), fisher strand (FS), strand odds ratio (SOR), RMS mapping quality (MQ), Mapping quality rank sum test (MQRankSum), and read position rank sum test (ReadPosRankSum). GATK recommends thresholds, but emphases the need for these to be tailored to the distribution of the research sample. VQSR uses machine learning to create a model that attempts to distinguish true variants from false positives. It usually considers five to eight dimensions and calculates a quality score per variant (variant quality score log-odds). Variants are then filtered based on their quality score.

VQSR has been shown to be slightly more accurate than hard filtering, however, it is difficult to apply to less well-studied species (57). This is because it requires a large database of high-confidence variant calls known as a 'truth set', and a list of lower confidence variants upon which the model can train. Such datasets exist for well-studied species, like humans, but not for less well-studied species, like dogs. It is possible to generate high-confidence datasets, but these must be derived from large amounts of high-quality species-specific data. Therefore, GATK does not recommend VQSR for studies of non-model organisms and/or small sample sizes. Regardless of the method used, GATK recommends filtering indels and SNPs independently. At the time of writing, GATK is developing a new method for classifying variants that utilises sophisticated machine learning with the aim that it will be more widely applicable than VQSR (56).

Quality filters are routinely applied prior to sequence alignment to increase the accuracy of mapping (51). This includes removing sequence reads that contain a high number of unknown bases, are too short, have been duplicated during PCR, or are too repetitive. Additionally, genotype-specific filters may be applied after variant calling/filtering to remove individuals and variants with high missing call rates, variants with low allele frequencies, and/or variants that deviate significantly from Hardy-Weinberg equilibrium (44). Carson et al. (58) determined that applying variant and genotype filters prior to VQSR, especially genotype quality, read depth, and call rate, improves the quality of results. This is because VQSR only considers metrics such as mapping quality and strand bias, not specific genotypes. The effectiveness and applicability of filters depends on the specific study objectives, for example, studies that aim to identify rare variants should omit AF

filters. More stringent filters lead to less noisy data but increase the risk of losing true variants, favouring specificity over sensitivity.

1.1.6 Variant Annotation

Genome annotation gives meaning to DNA sequences by identifying and characterising genetic elements (59). Types of genome annotation include structural (identification of DNA features such as exons and promoters), comparative (comparison of genes between species to identify evolutionary patterns), and functional (combining structural and biological annotation to predict the function of genetic elements) (60). Variant annotation involves the comparison of genetic variants to annotated reference assemblies to predict the impact of variants on the organism. Variant effect predictors (VEPs) are computational algorithms that use genome annotations to predict the impacts of specific variants (61). They offer a fast and effective method for interpreting variants and can be used in conjunction with experimental evidence and/or association studies to verify variant effects. The effect of a variant depends on the molecular change it causes and its position in the genome relative to a protein. For example, variants within introns are less likely to affect phenotypes than those within exons, truncating variants that affect the start of a protein often have larger effects than those near the end, and variants in regulatory elements can change the expression of the protein without directly altering the protein sequence. Variant effect predictors can aid in the detection of natural selection signals, identification of enrichment patterns in populations, and are critical to the prioritisation of candidate causal variants.

Various computational approaches are used for variant effect prediction and Liu et al. (62) categorises these into six groups: homologous sequence-based predictors, structure-based predictors, sequence and structure combination-based predictors, meta-predictors, combining population data, and disease-, phenotype-, or gene-specific predictors. Homology sequence-based prediction uses comparative genomics and relies on the assumption that variants in regions that are more well conserved are more likely to be deleterious. Structure-based predictors are based on the principle that protein structure is more well conserved than sequence and so variants that alter structure are

more likely to be deleterious. Sequence and structure combination-based predictors combine those two principles and have shown improved performance over using each individually. Meta-predictors produce high-accuracy classifications by evaluating and leveraging the complementarity of multiple predictors; however, these carry a high risk of circularity bias. Comparing population allele frequencies to make predictions relies on the assumption that lower frequency variants tend to be more deleterious. Because allele frequencies differ between populations, population stratification can hinder this method. Finally, targeted sequencing can be used to identify causal variants for specific diseases, phenotypes, or genes of interest, but this relies on prior knowledge of their genetic influences.

A recent advancement in the field of variant effect prediction is the introduction of machine-learning strategies. There are four broad categories of machine-learning strategies for predicting the pathogenicity of variants (63). The first trains directly on datasets of known causal variants. This exploits prior knowledge but can introduce bias from *in-silico* predictions and training sets often overlap test data (63). The second category uses weak labels in training (i.e. lenient labels with low certainty), relying on the assumption that common variants are likely benign while rare variants are likely pathogenic. This reduces circularity bias but can be noisy since rare variants may be benign (63). The third category uses unsupervised machine learning with no labels, instead it predicts pathogenicity based on the dependencies between amino acids. The fourth strategy predicts pathogenicity based on the predicted change in protein structure caused by the variant.

1.1.6.1 Examples of Variant Effect Predictors

Ensembl VEP (64) and SnpEff (65) are examples of computational VEPs. They work via similar algorithms and can annotate SNPs, indels, and SVs in both coding and non-coding regions. Their algorithms combine transcript, protein, non-coding, frequency, phenotype, and citation annotation to predict the impact on the protein. Both SnpEff and Ensembl VEP take VCF files as input and return VCF files with additional information for each variant. Ensembl VEP is more popular than SnpEff due to its integration with

Ensembl and accessible user interface, but both tools are well documented (64,65). Location annotations include ‘intronic’, ‘untranslated region’, ‘upstream’, ‘downstream’, ‘splice region’, and ‘intergenic region’. Predicted protein impacts include ‘high’, ‘moderate’, ‘low’, and ‘modifier’. Each tool has built in reference databases for many organisms but can also be supplied additional or alternative databases to compare variants to. Despite the continuous development of new prediction algorithms that use deep-learning and complex models, SnpEff and Ensembl VEP remain widely used for fast and accurate variant effect prediction (61).

In 2023, Cheng et al. (66) developed AlphaMissense, an algorithm that combines unsupervised machine learning with structural biology to classify human missense variants as likely benign or pathogenic. This is based on a deep learning artificial neural network called AlphaFold that predicts the 3D structure of proteins (67). AlphaFold trains on the evolutionary conservation of protein structures and amino acid properties to predict novel structures with accuracy comparable to that of experimental structure determination. This relies on the assumption that amino acids in contact in the 3D structure of a protein are likely to co-mutate. The AlphaMissense model avoids circularity bias by using weak labels and outperformed all other VEPs in a benchmarking study (68). It can also evaluate the essentiality of genes by calculating their average pathogenicity score. In addition to their open-source code, Cheng et al. (66) released predictions for 71 million missense SNPs in the human proteome and 60,000 alternative human transcripts and isoforms. This tool is a promising advancement in the field of variant effect prediction but is currently limited to chimpanzee and human missense variants.

1.1.6.2 Accuracy of Variant Effect Prediction

There are several caveats of computational variant effect prediction that limit its utility. Naturally, the accuracy of variant annotation relies heavily on the quality of genome annotation, where incomplete reference assemblies can lead to inaccurate functionality estimates (69). Variant effect predictors also tend to make inaccurate predictions for variants in less well-annotated non-exonic regions, meaning that variants within these regions may be excluded from the pipeline (5). This simplifies the analysis and improves

the average accuracy while maintaining the most functionally relevant variants (70). Variant effect predictors are almost certain to result in false predictions and because they use discordant databases, different VEPs may provide different predictions for the same variants. This can lead to opposing conclusions in some cases, suggesting that incorporating multiple tools may increase sensitivity for detecting functional variants but will complicate the analysis. In a comparison study, Grimm et al. (71) described two types of circularity bias that plague most VEPs; type one occurs when there is overlap between the test data and training data and type two occurs because all variants within the same gene tend to be classified the same way. It is important to note that computational variant effect prediction is not a replacement for experimental evidence or association studies and should be interpreted as additional evidence to support other functional and statistical analyses.

1.1.6.3 Using RNA Sequencing to Improve Variant Annotation

RNA sequencing (RNA-Seq) can improve functional annotation by revealing the protein-coding regions that are expressed in a particular tissue, as well as lncRNAs that are only transcribed (72). This deep-sequencing technique profiles all of the mRNA that is transcribed in a cell (a.k.a the transcriptome) (72). To achieve this, the cell's RNA is extracted, reverse transcribed into cDNA, amplified with PCR, fragmented, and then sequenced and reassembled. Like DNA, RNA reads can be reassembled though alignment to a reference genome (genome-based) or *de novo* assembly (transcriptome-based). However, unlike DNA, single RNA reads frequently exclude intronic regions and therefore are reassembled using different tools. STAR is one of the most widely used tools for RNA-Seq genome-based alignment (73). It uses a two-step process to split reads and match their sequences to exons. *De novo* RNA assembly is necessary when no reference sequence exists; however, this is more computationally demanding than reference-based alignment. RNA-Seq data is commonly used to compare gene expression between tissue types, different species, and in case-control studies, such as between diseased and healthy tissues. It can also be used to identify expression quantitative trait loci (eQTLs) and genes of strong biological relevance (74). Regions of the reference genome that are annotated as genes or exons but are not expressed in relevant tissues are likely

to be incorrectly annotated. RNA-Seq expression data can therefore expose these regions and increase the accuracy of functional annotation by excluding variants with erroneous functional classifications (75).

1.1.7 Microarrays

While DNA sequencing technologies are becoming increasingly popular due to their decreasing cost and the comprehensive view they offer, microarrays are still considerably cheaper and remain widely used for genetic analysis (76). Microarrays, commonly referred to as SNP chips, differ from sequencing in that they use probes to extract genotypes at specific sites of an organism's genome. The probes are designed and constructed by geneticists to target up to millions of sites of interest. Many SNP chips are commercially available for genetic testing and they can be species-, breed-, or phenotype-specific. For example, the Axiom™ Canine HD SNP chip from Thermo Fisher captures >710,000 sites in the dog genome that have been previously reported to vary between individuals. Interestingly, commercial genetic testing companies that use arrays are not required to report the positions of the variants they genotype, nor the clinical eligibility of disease-related variants (77). Because microarrays only capture a modest proportion of known variant sites, they are not able to identify novel functional variants like sequencing technologies can. However, an important application of arrays is the identification of genetic markers for traits of interest. This is possible due to a concept called linkage disequilibrium (LD), which is explained in detail below. As a result of LD, genetic markers near causative variants can be correlated with traits, even if they do not themselves influence these traits and so can be useful in diagnostics and/or genetic selection regardless. Pérez-Enciso et al. (76) determined that selection based on WGS does not increase the rate of genetic improvement compared to what can be achieved from microarray-based selection.

1.1.8 Linkage Disequilibrium (LD)

Linkage disequilibrium refers to the non-random association between alleles at different locations in the genome (loci) (78). This concept of genetic correlation is important for understanding the structure and evolution of the genome and can be used for association mapping. The LD coefficient (D) measures the amount of correlation between two alleles and compares the observed frequency of co-occurrence to the frequency that is expected under an assumption of independence. If loci always co-segregate, they are said to be in 'complete' or 'perfect' LD. By contrast, if the frequency of their co-occurrence does not differ significantly from expected, they are said to be in equilibrium. The LD coefficient is often normalised into the squared correlation coefficient (R^2). This equates to 0 when alleles are in equilibrium and 1 when they are in complete disequilibrium. Because recombination is less likely to split loci that are physically close in a genome, LD is negatively correlated with the physical distance between sites. However, linkage between sites can occur over large distances if it is deleterious for the haplotype to be split. For example, loci may be distant in the primary DNA sequence but encode amino acids that bind in the 3D protein structure and therefore co-evolve (79). Patterns of LD are complex, noisy, and differ between populations (80). Additionally, certain regions of the genome exhibit higher levels of LD than others. It has been proposed that some of this structure could be described with models based on haplotype blocks.

1.1.8.1 Haplotypes

Haplotypes are segments of chromosomes that are inherited from a single parent (i.e. a haploid genotype) (80). This term is used in several contexts across the literature, from describing large stretches of DNA that have been passed on by a distant ancestor, to describing the four possible combinations of alleles that could be observed at a pair of biallelic sites. Theoretically, recombination could occur anywhere in the genome and create new combinations of alleles (haplotypes). Under this assumption, millions of haplotypes would be observed in the population over small segments of the genome. In practice however, genetic linkage means that certain alleles are less likely to be split up

by recombination, and considerably fewer haplotypes are observed in populations than are theoretically possible. Regions with little evidence of historical recombination and few distinct haplotypes (limited haplotype diversity) are referred to as haplotype blocks. There is a lack of consensus in the literature for the precise definition of haplotype blocks. Generally, they are described as regions of the genome with shared patterns of ancestry that may be separated by segments of extensive historical recombination (recombination hotspots). Historically, it was thought that populations evolve through the changing frequencies of single alleles; however, it is more likely that evolution occurs due to the changing frequencies of haplotypes (81). Because of this, haplotype-based analyses should be more powerful than SNP-based analyses at detecting phenotypic associations (82).

1.1.8.2 Phasing and Imputation

Most sequencing technologies do not resolve haplotypes, meaning it must be ascertained from experimental or computational evidence which alleles were inherited on the same chromosome (83). This process is called phasing and is required for genotype imputation, identification of compound heterozygotes, and association testing. Phased genotypes create a better understanding of population history by revealing signatures of selection, sites of recombination, and cis regulation of gene expression (83). Although experimental phasing is more accurate, it is laborious, expensive, and therefore rarely performed in large-scale studies. Computational phasing, however, is widely used to resolve haplotypes and is made possible by the fact that LD limits the number of haplotypes observed in populations. Cohorts of related individuals can be phased using regions that are identical by descent (IBD), while unrelated cohorts must be phased by modelling the frequencies of common haplotypes. When variants are close or long-read data is available, read-based phasing can be performed (84). This resolves haplotypes by identifying multiple variants in single reads. In general, larger samples are required to phase unrelated cohorts with the same accuracy as related cohorts. Other factors that influence phasing accuracy are marker density, genotype accuracy, degree of relatedness, and AF.

Genotype imputation is the process of predicting alleles in individuals with limited genotype data based on a phased reference panel (85). This is commonly used to increase the density or coverage of genotype calls, identify causal alleles at ungenotyped loci, combine data from different genotyping arrays to perform meta-analyses, or combine WGS data with genotype array data. The accuracy of imputation is influenced by the size of the reference panel, array density, population structure, effective population size (N_e), minor allele frequency (MAF), and the relationship between the reference and target populations. Hidden Markov models (HMMs) are commonly used to perform both phasing and imputation, although other methods have also been developed (86).

One of the most widely used software tools for phasing and imputation is called Beagle (85,87). It applies an HMM with localised haplotype clustering to phase the reference panel in an iterative and progressive manner (87). Because WGS data contains many variants with low MAFs, two-stage phasing is applied to WGS reference panels and larger samples take longer to phase. Beagle applies a forward-backward HMM to impute genotypes, where the probability of each allele in the target is the sum of the state probabilities for each observed allele in the reference (85). Under an HMM, initial probabilities are equal for all observed states, transitional probabilities model historical recombination, and emission probabilities relate the model to observed data. Beagle models the uncertainty of allele predictions with a probabilistic output called dosage R-squared (DR^2). Dosage R-squared is interpreted as the correlation between imputed genotype dosage and true genotypes but is calculated without knowing the true genotype or discordance rate. Unlike R^2 values, DR^2 incorporates information about allele frequency (dosage), which has a large impact on imputation accuracy.

1.1.9 Genome-Wide Association Studies (GWAS)

First performed in 2005, genome-wide association studies (GWAS) accelerated the rate of gene and variant discovery in the 2000s and are still routinely used to identify associations between genotypes and phenotypes (88,89). In these studies, variants across the genome (usually hundreds of thousands of SNPs from a microarray) are

independently tested for systematic differences in allele frequencies between individuals who differ phenotypically. As well as identifying genetic markers for Mendelian traits, GWAS can lead to the development of prediction models like breeding values (BVs), or polygenic risk scores (PRSs), for complex traits by identifying QTLs. The advantage of this approach is that it does not assume any prior knowledge about the trait's genetic influences (77). Because the entire genome is screened, novel discoveries can be made, potentially by relying on LD if the true causative variant is not on the SNP chip used for the analysis.

In a GWAS, each queried SNP is assigned a p-value, indicating the statistical significance of its association with the trait, and an effect size, indicating the magnitude of impact (88). The results of a GWAS are usually visualised using Manhattan plots, which display the negative log p-value (y-axis) of each SNP across the genome (x-axis). The characteristic 'peaks' of Manhattan plots represent regions where correlated SNPs are all significantly associated with the trait of interest. In addition to acting as biological markers of phenotypes, significantly associated SNPs are likely to be genetically linked to causal variants, if not biologically linked themselves (90). However, the interpretation of these peaks is complicated by the fact that the lowest p-value is not necessarily the causal SNP, even when the causal SNP is on the test panel, and the nearest gene to the associated region is not necessarily the causal gene (88). Because of this, post-GWAS analysis is required to discover causal variants. This may involve the imputation or sequencing of variants, *in silico* functional annotation, and LD analyses. To prove causality, experimental evidence is usually required. This could include reporter assays to show changes in gene expression or cell or animal-based engineered models to show recapitulation of the phenotype.

The statistical models used to test for associations differ depending on whether the trait of interest is binary or quantitative, where linear and logistic regression models are most commonly applied (91). Most GWAS use additive association models, where the phenotype is assumed to be proportional to the number of allele copies (92). Non-additive association models, that consider other modes of inheritance such as dominance, can be applied but require more statistical power and are therefore rare (90). Because so many sites can be tested, there is a large multiple testing burden generated

from GWAS. This burden is addressed by adjusting the significance threshold, usually through the Bonferroni or Benjamini-Hochberg procedures. Genome-wide association studies also require large sample sizes to obtain enough statistical power to generate reproducible results. Additionally, GWAS results are more informative for traits with high heritability, where heritability is generally defined as the proportion of phenotypic variance of a trait that can be explained by genetic variation (90,93). This is because the more that traits depend on environmental factors, the less accurately they can be predicted by genetic factors.

Raw genotyping data is likely to contain many errors that may impact the accurate detection of associations in GWAS (90). Therefore, QC measures are required to produce reliable results. Marees et al. (90) suggest the removal of individuals with large amounts of missing data, an excess or lack of heterogeneity, high relatedness to other individuals in the cohort, or discrepancies between reported and genetic sex. They also recommend the removal of SNPs with large amounts of missing data across samples, low allele frequencies, or deviations from HWE. Depending on the study cohort and research question, different QC measures may be more or less important.

Another key step when running a GWAS in genetically diverse populations is to account for population stratification (88). Population stratification causes systematic bias in GWAS and can lead to false discoveries. While known structures, such as age and sex, can be included as factors in the model to reduce their effect, the most robust way to account for stratification is to use dimensionality reduction. For example, principal component analysis (PCA) can be performed on the genotypes to identify groups within the population that are genetically similar, and principal components (PCs) can be fit as covariates. If there are two distinct groups (e.g. breeds) in a sample, the first PC often accounts for this variation. Because of LD, SNPs are not independent of each other, which can cause collinearity in PCA. This can be corrected by pruning 'redundant' SNPs in LD and retaining only one representative SNP per haplotype block. Population structure can also be corrected by fitting a genomic relationship matrix (GRM) as a covariate. This estimates the variance explained by a set of SNPs in the GWAS model and is implemented in many analysis packages for performing GWAS, such as GCTA software (94).

1.1.10 Previous Genetic Studies in Dogs

The first canine reference genome assembly, canFam1.0, was released in 2005 (95). It utilised whole-genome shotgun sequencing and was based on a female Boxer canine. The canFam1.0 assembly was adapted into canFam2.0, which had slight improvements including larger contigs and fewer gaps. In 2014, canFam3.1 was released, which utilised Sanger sequencing of the same dog. This reference had improved annotations and introduced 85Mb of novel sequence compared to previous references. Many of the current dog variant panels and annotation resources reference this assembly. In late 2020, long-read sequencing technology was used to update canFam3.1 into canFam6 (Dog10K_Boxer_Tasha_1.0), which had fewer gaps and improved annotation and the canFam5 (UMICH_Zoey_3.1) assembly was generated from sequencing a Great Dane in 2019 (96,97). In early 2020, canFam4 (UU_Cfam_GSD_1.0) was assembled from a female German Shepherd (98). The canFam4 reference had 55-fold increased contiguity when compared to canFam3.1 and was annotated using RNA-Seq, miRNA-Seq, and ATAC-Seq data. CanFam4 is now a widely used version of the reference, but the 'best' assembly depends on the target breeds of individual studies. Even the most recent versions of the genome are error-prone and lack functional annotation compared to assemblies from model organisms like humans and mice (99).

Since the first dog was sequenced, numerous association studies have revealed functional genes and variants in dogs that were previously undescribed (100). Modern dog breeds exhibit strong signatures of selection compared to wild canine species (101). This is consistent with artificial selection causing a series of genetic bottlenecks and increasing inbreeding, especially in purebred dogs. Therefore, dogs exhibit extensive LD and large haplotype blocks (102). This provides an advantage for studying their genomes by increasing power for association testing and imputation, since fewer SNPs are required to represent the genome. However, it also makes it difficult to fine-map regions of interest and identify causal variants in downstream analyses.

1.1.10.1 Genetic Resources Currently Available for Dogs

The Dog10K Consortium (Dog10K) is the largest variation database available for dogs to date by number of samples (103). This ongoing study began in 2016 with the aim of sequencing 10,000 canines and canids. As of 2025, they had sequenced approximately 2000 samples from 261 breeds with 20× read coverage. They used GATK best practices to identify 14.4 million indels, 34 million SNPs, and 144,000 SVs (101). Additionally, they created an interactive online tool for studying the dog genome that is integrated with LiftOver so that coordinates can be converted to other genomes. In 2019, Plassais et al. (5) sequenced the genomes of 722 canines, including 144 modern dog breeds. They created a database of more than 91M variants that is available on the National Human Genome Research Institute website. This study substantially increased the number of known genetic variants in dogs.

Because dogs and humans share thousands of orthologous genes and are subject to similar selection pressures, many genetic diseases that are observed in dogs are also observed in humans (104). As a result, dogs are used as a model organism for studying human disease and many canine studies have been medically motivated. For example, the Dog Biomedical Variant Database Consortium DBVDC was formed in 2013 with the aim of compiling canine genetic data from a variety of projects and resources to inform medical research (105). They obtained the genome sequences of 582 dogs from 126 breeds and followed GATK best practices to identify 30M variants.

Other well-known online databases, like European Variation Archive and Online Mendelian Inheritance in Animals (OMIA), catalogue dog-specific variants from a variety of studies. Such archives have been formed over time and describe the findings from diverse analyses, with evolving evidence of causation, and are aligned to alternative reference assemblies, with evolving extents of annotation. The OMIA, for example, describes variants previously reported in dogs that are primarily Mendelian/monogenic in effect (106).

Existing databases may provide sequence, variant, annotation, and frequency information that can be extremely useful for subsequent studies. For example, existing genotype panels can be used for comparative genomics to understand the evolution of

dogs and annotation resources can be leveraged to predict the impact of candidate variants. However, the aforementioned resources are small compared to those available for other model organisms. For example, the largest human genome dataset, the UK Biobank, comprises the whole genomes from over 500,000 individuals (107). DNA expression, epigenetic, and phenotypic data is also available through this resource. The Encyclopedia of DNA Elements project aims to map all functional elements in the human genome (108). This publicly available resource includes information about chromatin state, DNA methylation, and protein expression across the genome for almost 1M humans and over 300,000 mice. AlphaMissense is another example of an extensive functional annotation resource available for humans (66). There are no equivalent resources for dogs. In fact, functional annotation of dog genomes often relies on data from humans and mice, which is informative because many elements are conserved. However, this also means canine-specific features are excluded. Some recent projects, like BarkBase (109), aim to improve the annotation of the canine genome, but the resources available for other model organisms are still considerably more extensive. Additionally, phase-resolved genome assemblies, which are useful to understand the effects of combinations of variants, are not available for canines.

1.1.10.2 Previous Genome-Wide Association Studies in Dogs

Dog GWAS tend to investigate single breed cohorts to reduce the effect of population stratification (110). This means that while their results may inform future studies, they may not be directly applicable to other breeds. Morphological and aesthetic traits like size and coat colour are the most genetically well-understood traits in dogs. For example, Plassais et al. (5) performed GWAS for 16 phenotypes, identifying several novel and previously described associations with morphological traits. Combined with previous findings, their results were able to explain 90% of body size variation in dogs. Complex traits that are controlled by a large number of genes are more difficult to study since individual variants only explain a small proportion of their heritability. For example, behavioural traits are complex, difficult to phenotype accurately, and strongly influenced by environmental factors, making them particularly difficult to characterise genetically. However, the development of GWAS has made this possible. Shan et al. (111) performed

GWAS for four behavioural traits: trainability, herding, predation, and temperament. They studied multiple breeds and incorporated variant effect prediction to identify significant associations for three of these traits around neurologically relevant genes. In humans, GWAS have been used to develop PRSs for complex traits that can assess an individual's genetic risk of developing complex trait phenotypes. Monogenic, or Mendelian, traits are less common in nature; however, because they are simple to study, the catalogue of known causal variants in dogs is biased towards this class of variants.

1.1.10.3 Previous Genetic Studies in Huntaway and Heading Dog Breeds

Very few research studies have been conducted on NZ working farm dogs, and the majority are not genetic studies. Genetic analyses that have been conducted focus on single genes or disease traits, with modest sample sizes or single cases. For example, Yogalingam et al. (112) identified a single Huntaway with mucopolysaccharidosis IIIA and sequenced the candidate gene to identify the causal variant in that individual. They then screened 203 Huntaways for the genotype and identified 15 carriers. Gedye et al. (113) surveyed a sample of 189 Huntaways for a variant in *ABCB1* that had been shown to protect the brain from xeno-biotic toxicity in Border Collies. They concluded that the prevalence of this variant was likely low in the population, with only two carriers identified in the sample. The few functional variants that are known to segregate in the NZ working dog population can be tested for. However, because dogs exhibit extensive interbreed genetic variation, it cannot be assumed that functional variants in one breed are also present in another. Given the current lack of understanding of Huntaway and Heading Dog genetics, it is therefore not known which genetic conditions discovered in other breeds may be relevant to the NZ population. It would be useful to identify which known dog variants segregate in the population to reveal which heritable traits can be genetically tested for and which disease traits may be of concern.

1.1.11 Conclusion

This review outlined what is currently known about Huntaways and Heading Dogs, breeds that make up the majority of working farm dogs in New Zealand but are not well-

understood genetically. The review also discussed bioinformatic techniques that are commonly used to characterise genetic variants in a population and identify potentially functional variants. These techniques include sequencing, variant calling, variant effect prediction, imputation, and genome-wide association studies. Many of these bioinformatics methods rely heavily on the concept of genetic linkage (or linkage disequilibrium), which, among other advantages, enables the inference or prediction of unobserved variants. As the cost of genetic technologies is decreasing, the amount of available genetic data is increasing, and new tools are constantly being developed for its interpretation. Despite this influx of genetic data, dog genomes are still understudied compared to other model organisms. In particular, the genetics of the New Zealand Huntaway and Heading Dog breeds has not been studied on a large scale to date. Because of this, the genetic makeup of these dogs, including their relation to other dog breeds, the amount of inbreeding present, and what advantageous, deleterious, or disease-causing variants they carry, is largely unknown. This limits the potential for genetic gain in a population that is vital to agriculture, New Zealand's largest industry.

1.2 Project Aims

The overall aim of the Right Dog for the Job project is to improve the health and performance of the New Zealand working farm dog population. As an early stage of this project, the following research aimed to characterise the genetic variation in Huntaways and Heading Dogs by mapping, calling, and functionally annotating the genetic variants in a sample of 250 whole genome sequences. Secondly, the project aimed to identify previously reported and/or novel functional variants segregating in the population, with a particular focus on disease variants that may aid in diagnostic selection. Moreover, it aimed to determine whether Axiom™ Canine HD Array genotypes could be used to predict the genotypes of functional variants, either through marker-based prediction or imputation. The final aim of this research was to perform GWAS to identify genetic loci associated with body size phenotypes, representing the first phenotypic data gathered as part of the Right Dog programme.

1.3 Scope and Limitations

The dogs included in this study were predominantly from the New Zealand Huntaway and Heading Dog breeds. Therefore, any conclusions can only be made for this population. A small minority of genotyped and/or sequenced dogs were crosses or from alternative breeds. These dogs were included to increase statistical power, and to create a more realistic representation of the farm dog population. However, none of the secondary breeds made up a large enough proportion of the sample that strong conclusions can be made about them. It should also be noted that while dogs were selected from a range of geographic regions to represent the population, the convenience of sampling was also a factor (i.e. individuals from the same farms and owners were included). This pragmatic sampling approach reduces the cost and time of sample collection but means that the sample is not truly random.

Because the Right Dog project is ongoing, the number of dogs with genotype and phenotype data was limited and evolved throughout the current study. This meant that the majority of the analyses were limited to genomic data and largely excluded phenotypic measurements. Conclusions that could be made pertaining to functionality were therefore limited, since causality for novel variants could not be verified. The project instead relied heavily on computational prediction and functional evidence from previous studies. This also meant that the sample size was relatively small for the imputation and GWAS, limiting their statistical power.

While variants in non-coding regions can have phenotypic effects, the functional prioritisation of variants in this study was simplified by omitting those in non-coding regions. This undoubtedly led to the exclusion of some causal variants, but also represents a pragmatic approach since it refines the scope of the analysis while retaining focus on the categories of variants that have the largest effects (23). Because short-read sequencing was performed, large SVs, which can only be accurately characterised with long-read sequencing data, were generally omitted (41). Instead, the study focused primarily on SNPs and indels.

1.3 Significance of Research

By identifying disease variants segregating in the population, this project enables the genetic testing of New Zealand farm dogs, and therefore the early diagnosis, prevention, and/or treatment of these diseases. Selection for disease markers and the avoidance of carrier-carrier matings could prevent dog owners from spending time and money training dogs that go on to develop genetic disease. The project also provides a more accessible means of genetic testing by identifying predictive markers on the Axiom™ Canine HD Array.

The genomic resources produced here will directly enable subsequent studies within the Right Dog for the Job project and will soon become available for public use. These datasets can be used: to better understand the relationship between breeds; to better understand the evolution of breeds; to identify additional deleterious and/or advantageous variants; and potentially to develop a breed-specific SNP chip. Furthermore, the functional candidate variants identified in this project represent hypotheses, where further testing could lead to the discovery of novel causal variants. Overall, the purpose of this project has been to create a greater understanding of Huntaways and Heading Dogs, and expand the genomic information available for dogs.

Chapter 2: Methods

2.1 Sample Collection

Qualified veterinarians collected blood samples from 250 dogs for whole genome sequencing. This cohort included 130 Huntaways, 104 Heading Dogs, and 16 Huntaway/Heading Dog intercrosses or crosses of other working breeds, with breed declared by participating dog owners. Sampling of close relatives was avoided and animals were sourced across several NZ regions (see Table 1). Saliva was collected and genotyped from 299 dogs according to the Ancestry 'Know Your Pet DNA' product kit instructions. This sample included 102 Huntaways, 139 Heading Dogs, and 58 dogs of other working breeds from across NZ (see Table 2). Huntaway/Heading Dog crosses and missing breeds were classified as 'other' in this count. The total genomic dataset comprised 188 individuals that had been both sequenced and SNP chip genotyped, 62 that had been sequenced but not SNP chip genotyped, and 111 that had been SNP chip genotyped but not sequenced. All genetic data was stored in NeSI (New Zealand eScience Infrastructure), a platform for high-performance computing and data analytics (project ID = massey04036).

Tape measures were used to collect morphological phenotypes such as height (withers to floor) (mm), length (base of skull to crest of the ilium) (mm), chest circumference (behind the elbow) (mm), and muzzle circumference (mm). Participating dog owners completed surveys pertaining to other health, performance, and behavioural traits of interest (see Appendix A.1). All phenotype data was stored on The Helical Platform, an online tool for managing and analysing genetic data, provisioned by the Helical Company Ltd. The survey was extensive and provides the basis for all phenotypic analyses to be conducted within the wider Right Dog project. The following thesis primarily focuses on characterising genomic variation and therefore excludes most of the phenotypic information in the survey; however, the aforementioned morphological measurements and owner-declared breed were analysed.

Table 1: WGS sample count by region

	Huntaways [a]	Heading Dogs [a]	Other [b]	Total
Hawkes Bay	44	29	2	75
Otago	17	13	2	32
Southland	43	39	3	85
Waikato	18	18	3	39
Manawatu-Whanganui	0	0	6	6
Other	8	5	0	13
Total	130	104	16	**250**

[a] *Huntaways or Heading Dogs reported to be purebred.*

[b] *Includes Huntaway/Heading Dog crosses and other breeds.*

Table 2: SNP chip sample count by region

	Huntaways [a]	Heading Dogs [a]	Other [b]	Total
Hawkes Bay	30	24	2	56
Otago	14	12	1	27
Southland	32	30	3	65
Waikato	23	71	16	110
Manawatu-Whanganui	1	0	23	24
Canterbury	0	0	12	12
Other	2	2	1	5
Total	102	139	58	**299**

[a] *Huntaways or Heading Dogs reported to be purebred.*

[b] *Includes Huntaway/Heading Dog crosses, other breeds, and missing breeds.*

2.2 Whole Genome Sequence Data Processing

Variant data was generated from whole genome sequences according to the pipeline shown in Figure 3. This outlines the major procedures, tools, and software packages used to generate a 20 million-variant VCF file from 250 blood samples.

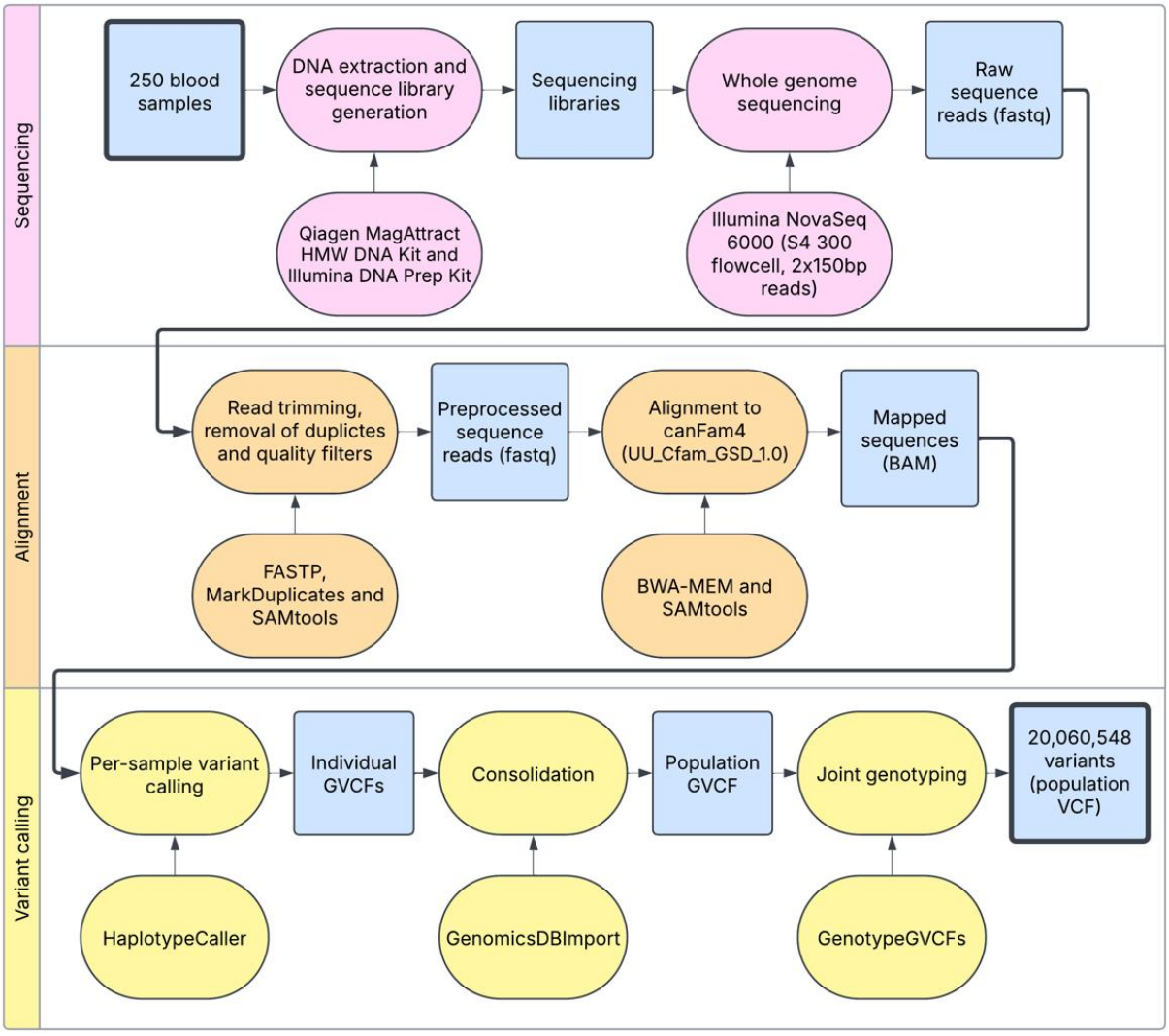

Figure 3. Flowchart of variant data generation from WGS sample. The generation of the WGS variant dataset occurred in three major stages: sequencing, alignment, and variant calling. Blue squares represent the products from each data processing step, where the final product was a 250-sample population VCF file.

DNA was extracted from 200 µl of whole blood using a Qiagen MagAttract HMW DNA Kit (Cat67563 Qiagen NZ), and sequencing libraries were produced from 350 ng of DNA using an Illumina DNA Prep Kit (Illumina, NZ). Paired-end whole genome sequencing was

performed on an Illumina NovaSeq 6000 using an S4 300 flowcell (2×150bp reads; conducted at GeneMark, Hamilton, NZ). Samples were processed by GeneMark into raw fastq sequence reads, generating an average of approximately 430 million sequence reads per sample. Fastq files were pre-processed using FASTP (v0.23.4) (114) to trim poor-quality sequences and artefacts from the sequencing process. Reads were removed if: more than 30% of bases had a quality score of less than 15 (low quality), they were shorter than 40bp, less than 30% of their bases differed from the previous base (low complexity), or they had more than three unknown bases (too many Ns). Poly-G tails, which are long G repeats caused by a lack of signal in the sequencing process, and tails with quality scores less than 30 were trimmed from each read. Artificially duplicated reads, which arise during PCR amplification and bias allele depth calculations, were marked and removed with GATK (v4.5.0) (48). Read mapping was performed using BWA-MEM (v0.7.17) (51) referencing the canFam4 (UU_Cfam_GSD_1.0) genome (98) to produce eight BAM files per sample (forward and reverse reads from each of four flow cell lanes). Post alignment, SAMtools (v1.16.1) (49) was used to flag and remove reads that were unmapped, had an unmapped pair, failed vendor quality checks, mapped with a quality score of less than 40, were supplementary, and/or were not primary. Finally, the eight BAM files were merged, yielding per-sample BAM files with an average read depth of 22.88× and average read quality of 35.68. Other summary statistics from pre-processing and alignment, including error rate and the proportion of reads that were properly mapped and paired, were calculated with SAMtools and averaged across all samples using RStudio (v4.1.2; R Core Team 2021).

2.3 Variant Calling and Filtering in the WGS Dataset

The GATK best practices workflow was followed to call and filter variants in the 250 whole genome sequences (115). HaplotypeCaller (47) was used to call variants per sample, resulting in intermediate individual genomic variant call format (GVCF) files. Individual GVCFs were merged across all samples with GenomicsDBImport, producing an intermediate population GVCF file. Finally, GenotypeGVCFs was used to perform joint

genotyping on the population GVCF, yielding a population VCF file that described 20,060,548 genetic variants (14,469,479 SNPs and 5,981,466 indels).

Quality control of variant calls was performed using the GATK-recommended hard variant thresholds rather than VQSR filtering. This is because dogs are a non-model organism and there are limited high-confidence datasets that could be used to train a VQSR model. SelectVariants was used to subset the population VCF file into SNPs and indels and VariantFiltration was used to remove SNPs and indels that did not meet the criteria outlined in the first filtering step of Figure 4. Before filters were applied, the distribution of each quality metric was visualised in RStudio with density plots to ensure the thresholds were appropriate for the data (see Figure 5). In order to favour sensitivity over specificity, thresholds were deemed appropriate if they intersected with the tails but did not approach the body of the distribution. As seen in Figure 5, the GATK-recommended thresholds were relatively lenient for this dataset, meaning they did not risk removing many true variants and were therefore applied without modification. 12,267,705 SNPs and 4,827,258 indels passed hard variant filters and were concatenated into a single VCF file of approximately 17 million variants.

Next, BCFtools (v1.19) (49) was used to apply additional per-sample genotype quality thresholds that are outlined in the second filtering step in Figure 4. These thresholds were also lenient, favouring sensitivity (116). This yielded the working variant dataset of 16,678,350 variants (12,059,693 SNPs and 4,618,657 indels), which will be referred to as the WGS data hereafter. A subset of 11,521,531 WGS variants was created by removing variants with high missing call rates (> 90%) and low MAF (< 1%). This subset is useful for analyses requiring highly filtered variants but may exclude some true calls and so was not used for most exploratory analyses.

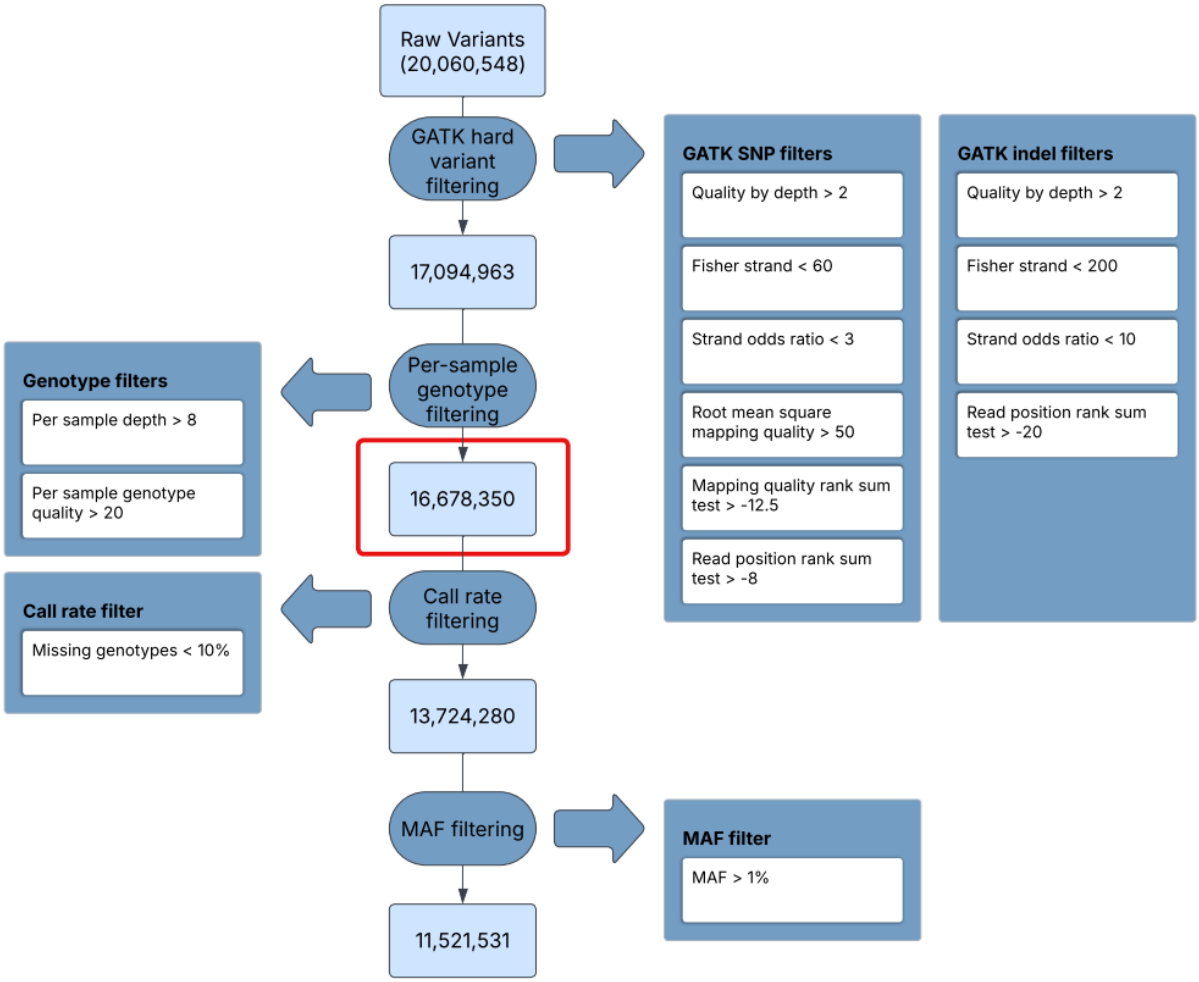

Figure 4. Flowchart of variant filtering pipeline. Variant filtering occurred in four steps: GATK-recommended hard variant filtering, per-sample genotype filtering, call rate filtering, and MAF filtering. Variants were retained if they passed the thresholds shown in each filtering step, where light blue boxes represent the number of variants retained after each step. The red box outlines the number of variants in the working WGS dataset.

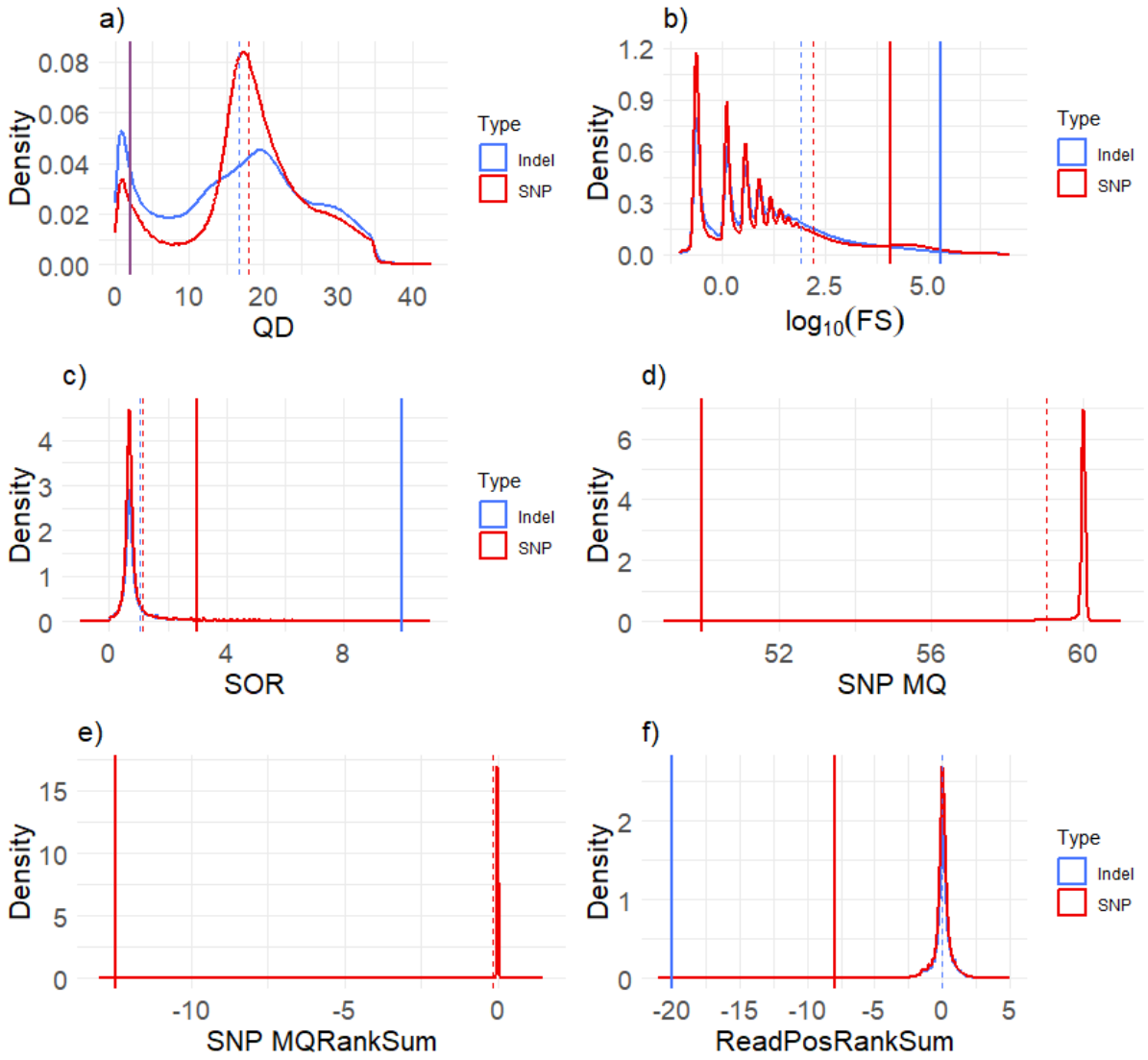

Figure 5. Density plots of GATK-recommended filtering metrics. **a)** Quality by depth (QD), **b)** $\log_{10}$ (Fisher strand (FS)), **c)** Strand odds ratio (SOR), **d)** SNP mapping quality (MQ), **e)** SNP mapping quality rank sum test (MQRankSum), **f)** Read position rank sum test (ReadPosRankSum). Dashed lines represent sample means and solid lines represent hard filtering thresholds. As per GATK best practices, only SNPs were filtered by MQ and MQRankSum.

2.4 Annotation of WGS Variants

2.4.1 Variant Effect Prediction

Computational variant effect prediction was performed with both SnpEff (v5.1) (65) and Ensembl VEP (v107.0) (64) annotation tools. A canFam4-specific SnpEff database was built from general transfer format (GTF) files (assembly = GCA_011100685.1) according to the SnpEff documentation. General transfer format files store information about genetic features such as genes, transcripts, and exons. Referencing this database, SnpEff was used with default options to annotate the WGS VCF file. This predicted the impacts and consequences of variants based on the transcripts they affect. High- and moderate-impact variants were considered 'functional', acknowledging that this classification overlooks the many thousands of regulatory variants with difficult to predict, yet nonetheless functional, effects. A subset of SnpEff-predicted functional variants was created with SnpSift (117) by filtering the output for high- and moderate-impact annotations. Ensembl VEP provides cache files containing all transcript models, regulatory features, and variant data specific to canFam4, meaning a new database did not need to be built. These cache files were downloaded and referenced when Ensembl VEP was used with default options to annotate the same population VCF file. BCFtools was used to create a subset of Ensembl VEP-predicted functional variants by filtering for high- and moderate-impact annotations. The two subsets were intersected with BCFtools to generate a list of putative functional predictions (i.e. variants predicted by both tools to be functional). Because individual VEPs utilise different algorithms and databases that may each have their own inherent biases, the use of multiple tools leads to a more accurate final list of functional predictions (62).

2.4.2 Comparison of Variant Effects in Huntaways and Heading Dogs

The following analysis aimed to investigate whether variant effects in the WGS variant dataset differed between Huntaways and Heading Dogs. Three of the largest genetic studies previously conducted in dogs, Dog10K (101), DBVDC (105), and Plassais et al. (5), annotated variants with SnpEff. Therefore, to ensure the findings from the following

analysis were comparable to those studies, SnpEff annotations were used to make the comparisons.

Excluding all cross breeds, BCFtools was used to split the working variant dataset into breed-specific VCF files. The +fill-tags plugin was applied to update INFO fields (AC, AF, AN, MAF, and F_MISSING) for the respective breeds, since these are not automatically updated. Variants with a MAF of 0 in the respective breeds were subsequently removed. The final Huntaway VCF file represented 14,878,265 variants from 130 individuals and the final Heading Dog VCF file represented 14,024,707 variants from 104 individuals. The resulting files were intersected to determine how many variants were shared between breeds and annotated with SnpEff. Applying the vcfEffOnePerLine.pl script from SnpSift, the EFFECT (predicted functional consequence of a variant on a gene given the location and type of sequence change) and IMPACT (predicted severity of the variant's effect on the gene) field annotations were extracted from each file. Because single variants can affect multiple transcripts and have multiple effects per transcript, the output files contained 52,862,464 and 48,914,922 effect annotations for Huntaways and Heading Dogs respectively. The proportion of each effect category (splice region, start or stop lost or gained, frameshift, in-frame indel, synonymous, intron, downstream or upstream, inter- or intragenic, missense, non-coding transcript, untranslated region (UTR), and 'other') and impact (high, moderate, low, and modifier) was calculated in RStudio. Only the most severe predicted effect of each variant was included in effect counts, while all predicted impact annotations were included in impact counts. Appendix A.2 outlines the specific SnpEff-annotated effects included in each effect category and the SnpEff documentation (118) describes what each effect annotation means. Chi-squared tests were used to compare the proportion of 'functional' variants and the proportion of coding-region variants between the two breeds. Coding-region variants included splice region, start or stop lost or gained, frameshift, in-frame indel, synonymous, start or stop retained, missense, and 'other' effects. 'Functional' variants included high and moderate impacts. Counts and proportions were calculated from the number of variant annotations, rather than the number of variants.

2.4.3 Improving Variant Annotation with RNA-Seq Expression Data

Publicly available RNA-Seq data was used to annotate putative functional predictions based on the level of gene expression at each variant locus. RNA-Seq samples were selected to represent a range of healthy tissue types. The 17 selected samples, described in Table 3, were produced by the DBVDC (105) in 2019 using Illumina paired-end RNA sequencing and were obtained from four breeds of male dogs aged between two and 16 years. The raw fastq files were downloaded from the sequence read archive (SRA) using SRAtoolkit (v3.0.2) (119) and reads were aligned to the canFam4 reference genome using STAR (v2.7.9a) (73). The GenomeGenerate option in STAR was used to generate canfam4 genome indexes from GTF and fasta files. Referencing the index directory, 10 of the RNA-Seq samples were mapped to identify novel splice junctions (i.e. those not annotated in canFam4). GenomeGenerate was used again to generate new canFam4 indexes that incorporated the novel splice junctions and, finally, all 17 RNA-Seq samples were mapped referencing the new index directory, yielding per-sample BAM files.

BCFtools was used to merge BAM files into a single file that represented the expression data from all 17 tissues. A custom PERL script was used to calculate the average RNA-Seq read depth (from this file) for the 10bp region surrounding each variant in the VCF file of putative functional predictions. This value, named 'RNABamDepth', was added to each variant as an INFO field annotation. Variants with RNABamDepth < 5 were assumed to be in falsely annotated genes and were removed from the final list. This somewhat arbitrary but conservative threshold was determined through a visual inspection of mapped RNA-Seq data in Integrative Genome Viewer (IGV) (120). The loci of 50 randomly selected functional predictions were classified as coding or non-coding based on the relative RNA-Seq read depth of the associated gene, since the introns of some genes may be expressed, but not to the same extent as those genes' exons. The loci of 31 variants were classified as coding and had a mean RNABamDepth of 2232, the loci of 19 variants were classified as non-coding (i.e. falsely annotated) and had a mean RNABamDepth of 4. The RNABamDepths of all 50 loci as well as IGV screenshots showing examples of expressed and not expressed transcripts are shown in Appendix A.3. It was concluded based on this assessment that true protein-coding regions had RNABamDepths of at least 5 (and usually much larger) in the 17-tissue BAM file.

Table 3: Description of RNA-Seq samples

Tissue	Accession number	Breed	Age
Liver	SRR8996977	Yorkshire Terrier	11
Bone marrow	SRR5889310	Labrador	16
Lung	SRR5889338	Labrador	16
Occipital cortex	SRR5889306	Labrador	16
Cerebellum	SRR8997035	Belgian Malinois	2
Skeletal muscle	SRR8997023	Belgian Malinois	2
Skin	SRR8996988	Belgian Malinois	3
Pancreas	SRR8996967	Newfoundland	11
Right ventricle	SRR8996969	Newfoundland	11
Small intestine	SRR8996987	Belgian Malinois	3
Spleen	SRR8996986	Belgian Malinois	3
Kidney cortex	SRR8997001	Newfoundland	11
Kidney medulla	SRR8997002	Newfoundland	11
Adrenal gland	SRR8997019	Belgian Malinois	3
Stomach	SRR8997045	Newfoundland	11
Adipose tissue	SRR8997052	Belgian Malinois	2
Lymph node	SRR5889339	Labrador	16

These 17 RNA-Seq samples were used to annotate putative functional predictions with 'RNABamDepth' and remove variants in falsely annotated genes.

Note: *All data were extracted from the SRA and produced by the DBVDC project (105).*

2.5 Survey of Previously Reported Functional Mendelian Variants in NZ Farm Dogs

The Online Mendelian Inheritance in Animals (OMIA) database (106) was leveraged to identify known functional variants segregating in the New Zealand farm dog population. The positions of 449 Mendelian variants were retrieved from the OMIA database, including all Mendelian variants reported as likely causal in dogs as of September 2023. Where alternative genomes were referenced, 100bp flanking sequences were extracted using the faidx method from SAMtools (49) and remapped to canFam4 (98) using BWA-MEM (51). Large SVs were omitted from the analysis due to the difficulty in systematically characterising this category of mutation from short-read data. This excluded 53 OMIA variants, including any gross insertions or deletions, haplotypes, large inversions or duplications, or complex rearrangements. Nine additional variants had no reported positions in OMIA and so were excluded. The remaining 395 Mendelian variants were intersected with the WGS VCF file using BCFtools (49) to identify those segregating in the sample. Allele frequencies were calculated from the total WGS sample and for each breed.

Quality control was performed through a visual inspection of sequence data at each segregating locus in IGV (120). Variants were classified as false calls if they fit any of the following criteria: they occurred in highly repetitive or error-prone regions where non-affected samples contained alternative reads, the number of reads carrying the alternative allele was considerably less than half, the total read depth at the locus was extremely low or there was a clear annotation error in the assembly causing the call (e.g. a 1bp deletion was called in every individual at a locus where there was a 1bp intron annotated in the reference genome). By contrast, variants were classified as true calls if they occurred in high-depth regions, approximately half of the sequence reads contained the alternative allele in heterozygous samples, approximately all of the sequence reads contained the alternative allele in homozygous samples, and alternative allele reads were absent from (or extremely rare in) unaffected samples.

Because the OMIA database represents historical variant discoveries with evolving evidence of causality, the literature pertaining to each segregating variant was evaluated

to assess their functional candidacy. True variants were classified as: 'functional' if they had a known phenotypic effect in one of the breeds of interest, 'likely functional' if there was strong evidence of a phenotypic effect in other dog breeds, 'possibly functional' if there was weak evidence of a phenotypic effect, or 'unlikely to be functional' if there was a lack of evidence of a phenotypic effect or follow-up studies refuted the proposed effect.

To provide additional evidence of causality, the gene expression at each variant locus was verified through visual inspection of the aforementioned public RNA-Seq data in IGV (120). Section 4.4 of this thesis describes the functional evaluation of each variant in more detail. One major aim of this review was to highlight variants that had been previously implicated in disease and were therefore compelling candidates for use as selection diagnostics in Huntaways and Heading Dogs. A short communication based on this analysis has been submitted to Animal Genetics (121).

2.6 Survey of High-Impact Variants within Genes of Interest in NZ Farm Dogs

The next analysis aimed to identify novel functional variants segregating in the farm dog population by surveying the WGS sample for high-impact variants within candidate functional genes. The names of 132 genes of interest were extracted from the OMIA database (see Appendix A.4) (106). These included all genes in the database that contained other LOF variants that were proposed as causing functional effects in dogs as of September 2023. SnpSift was used to filter the WGS VCF file for SnpEff-predicted high-impact variants within genes of interest. OMIA reports gene names for each variant, and Ensembl gene IDs were extracted from the National Center for Biotechnology Information (NCBI) database for each gene. Because SnpEff can call effects for multiple genes for a particular variant (e.g. in the exon of one gene, but in the intron of another gene on the opposite strand), the SnpEff result set was manually filtered using the Ensembl IDs, to retain only variants whose high-impact prediction corresponded to one of the genes of interest from the OMIA reports. This analysis relied on the assumption that high-impact (disruptive) variants within previously reported functional genes are likely to cause

similar phenotypic effects. Moderate-impact predictions were excluded since this category consists of mostly missense variants, the functions of which are more difficult to predict than the nonsense variants captured by the high-impact category.

The above workflow generated a list of variants segregating in the WGS animals that were predicted to disrupt genes that had previously been reported in OMIA. Allele frequencies were calculated from the total WGS sample. Each variant was evaluated according to the following two-step process. First, sequence reads were visually inspected in IGV to attempt to distinguish true variant calls from errors using the criteria outlined above (see Section 2.5). Based on these criteria, the classification of a small minority of variants was unclear. These possible variant calls represented fixed alleles that were most likely annotation errors but could not conclusively be classified as such. Second, the literature pertaining to each gene of interest was reviewed to assess the potential functionality of true variants, sequence reads were visualised in IGV to verify allele frequencies, and the aforementioned public RNA-Seq data was visualised in IGV to verify gene expression. True and possibly true variants were classified as functional candidates if there was strong evidence of comparable variants within the gene causing phenotypic effects and unlikely to be functional otherwise. Section 4.5 of this thesis describes the functional evaluation of each variant in more detail.

2.7 SNP Chip Data Processing

Saliva samples from 299 working farm dogs across NZ (see Table 2) were genotyped on the Thermo Fisher Axiom™ Canine HD Array by AncestryDNA. Genotypes were processed into PLINK text format (ped and map files) by AncestryDNA. These were then converted into PLINK binary format (bed, bim, and fam files) and uploaded to NeSI. A total of 719,182 sites across the canFam4 reference genome were genotyped on the array (SNP chip) and 717,915 of those positions were reported. This sample of genotypes will be referred to as the SNP chip dataset in the following thesis. To assess the quality of the SNP chip genotypes, per-sample and per-variant summary statistics were calculated with PLINK (v1.9) (122) and visualised in RStudio. The quality metrics that were analysed included missing call rate (proportion of SNPs with unknown genotypes in a given

sample), missing genotype rate (proportion of individuals with unknown genotypes at a given locus), MAF (the frequency of the least common allele in the population at a given locus), and inbreeding coefficient (F) (the ratio between observed and expected homozygosity assuming HWE). For quality control, samples with call rate < 0.1 and variants with MAF < 1% or missing genotype rate > 0.1 were excluded from most of the following analyses. The equation used to calculate inbreeding coefficient in PLINK is $\frac{(<observed\ hom.\ Count> - <expected\ count>)}{(<total\ observations> - <expected\ count>)}$ (122).

2.8 Genetic Verification of Matching WGS/SNP Chip Samples

SNP chip samples were identified by unique vial ID codes associated with the Ancestry Know Your Pet DNA Kit, while WGS samples were identified by dog name and a number when multiple dogs had the same name. Since the following LD analysis relied on the overlap between the two genetic datasets, it was important to confirm that the vial IDs and names of the dogs present in both, which will be referred to as the overlapping sample (n = 188), were correctly matched. Therefore, two methods were employed to assess the genetic similarity between SNP chip samples and WGS samples, with the aim of distinguishing genetically identical pairs from genetically distinct pairs. In reality, genotyping errors create some discordance between the technologies, however, an individual compared to themselves should still have considerably higher rates of concordance than an individual compared to another individual.

A subset of the WGS variants was created with BCFtools (49) by retaining only the genotypes at Axiom™ Canine HD Array loci and the SNP chip genotypes were converted from PLINK format into a VCF file with PLINK (v1.9) (122). The resulting VCF files described genotypes at the same loci but from two genotyping technologies. These were merged into a single VCF file containing 548 unique variant IDs, where 187 animals were represented twice, once by their WGS genotypes and once by their SNP chip genotypes. One overlapping individual was excluded due to a high SNP chip missing call rate. VCFtools (v0.1.15) (46) was used to calculate a relatedness matrix between all WGS samples and all SNP chip samples (n = 74,500 comparisons). Pairwise relatedness was

calculated as the adjusted A_{jk} coefficient (123), an estimate of the amount of genome shared between each pair of individuals averaged over all variant positions in the genome.

The A_{jk} statistic is calculated with the following equation: $A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}$ (94), where x_{ij} = the i^{th} SNP of the j^{th} individual, x_{ik} = the i^{th} SNP of the k^{th} individual, p_i = the frequency of the reference allele for the i^{th} SNP, and N = the total number of SNPs.

Theoretically, genetically identical samples have A_{jk} scores of 1, first degree relatives like full siblings have A_{jk} scores close to 0.5, and unrelated individuals within a population have A_{jk} scores of 0. The matrix was split into comparisons within individuals (individuals' relatedness to themselves) and comparisons between individuals (individuals' relatedness to other individuals) and a two-sample T-test in RStudio was performed to compare the mean A_{jk} score of within-individual comparisons to the mean A_{jk} score of between-individual comparisons. The aim of this was to confirm that within-individual comparisons could be distinguished from between-individual comparisons based on their relatedness, and therefore incorrectly matched overlapping samples could be detected by their A_{jk} score.

The second metric used to assess the genetic similarity between pairs of WGS and SNP chip samples was discordance rate. The merged 548-sample VCF file was converted into PLINK binary format and the discordance (number of genotypes that differ) between each pair of WGS and SNP chip samples was calculated with PLINK (v2.0) (124). A two-sample T-test was performed to compare the mean discordance rate (the number of differences between samples divided by the number of sites genotyped in both samples) of within-individual comparisons to the mean discordance rate of between-individual comparisons. The aim of this analysis was to confirm that within-individual comparisons could be distinguished from between-individual comparisons based on their discordance rates, and therefore incorrectly matched overlapping samples could be detected by this metric.

2.9 Identification of Predictive Markers for Mendelian Variants of Interest

A linkage disequilibrium (LD) analysis was carried out to identify single loci on the Axiom™ Canine HD Array that could act as predictive markers for the OMIA-derived Mendelian variants that segregated in the sample (see Table 6). Retaining only the 184 samples that were genetically confirmed to be overlapping (see Sections 2.8 and 3.7), the SNP chip PLINK files were merged with PLINK files that represented the OMIA-derived variants segregating in WGS sample. WGS animal IDs were relabelled to match their SNP chip ID before the merge, resulting in PLINK binary files with 184 unique animal genotypes at the SNP chip loci and the Mendelian variant loci (n = 719,208 loci). Under default PLINK parameters, R^2 allele count correlations were calculated between Mendelian sites and SNP chip sites. By default, PLINK ignores missing genotypes and only reports pairs that correlate with an $R^2 > 0.2$. The SNP chip locus with the highest R^2 correlation to each Mendelian variant was defined as the candidate marker for that variant. If multiple markers correlated with the same R^2 value, the closest marker by physical distance to the Mendelian variant was defined at the candidate marker. To enable the detection of markers for rare variants, no MAF filter was applied. Consequently, it was important to assess each candidate marker and ensure correlations were not due to chance alone. Genotype plots were generated with RStudio to visualise correlations and classify candidate markers as strong, moderate, or weak. Strong candidate markers included those that were perfectly correlated to Mendelian variants, provided they had missing call rate < 0.1 and minor allele count (MAC) ≥ 5, or were at the same genomic position as the Mendelian variant (125). Moderate candidate markers included those that correlated with an $R^2 \geq 0.8$, provided they had missing call rate < 0.1 and MAC ≥ 5. Weak candidate markers included those with $R^2 < 0.8$, high missing call rates, or MAC < 5.

2.10 Imputation of Missing Genotypes in the SNP Chip Sample

A widely used imputation software package, Beagle (v5.4) (85,87), was used to predict (impute) missing genotypes in the SNP chip-only sample (individuals that had only been

genotyped with the array) at all polymorphic sites identified from the WGS sample. The aim of this analysis was to determine whether imputation referencing the WGS sample could be used to predict genotypes in the SNP chip sample. Of particular interest to the current project was the accuracy of imputation at the Mendelian variant loci outlined in Tables 6 and 7 that could not be accurately predicted by single SNP markers on the array (see Table 9). The Beagle phasing algorithm was applied to generate a panel of genotypes by haplotype phasing the 250-sample WGS variant dataset (87). Referencing this panel, missing genotypes in the target sample were phased and imputed (85). This was performed in two stages as phasing the reference panel prior to imputation has been shown to increase accuracy (126). A reference panel of 250 samples is relatively small for imputation; however, the N_e of Huntaways and Heading Dogs is probably small. Therefore, a sample of 250 may be large enough to represent most of the haplotypes in the population. Additionally, the high quality of the WGS data suggests that high-accuracy imputation is achievable, provided the sample is representative of the population. The accuracy of Beagle imputation at each marker is estimated by DR^2. DR^2 ranges from 0 to 1 and is a measure of the squared correlation between the estimated and the true allele dosage at a given marker (85). Imputation was repeated under four sets of criteria to optimise the average DR^2 across all markers.

In the first imputation run, default Beagle parameters were applied and the reference panel comprised the phased WGS VCF file (n = 16,678,350 variants). No genetic map was specified, as this was not available for canFam4, nor an N_e, as this is not known for NZ working dogs. When N_e is not specified, Beagle estimates it for each burnin iteration based on genetic diversity (85). However, Beagle has been shown to grossly overestimate N_e from small sample sizes, which can considerably reduce accuracy (127). Results from the first run supported those findings, with Beagle estimating N_e > 30,000 in several iterations. Therefore, the second imputation run referenced the same genotype panel but an N_e of 266 was specified. This was the average N_e across several breeds according to the findings from Leroy et al. (128). Given only dogs with the highest working ability tend to be bred, the N_e of NZ working dogs is unlikely to exceed that of other dog breeds. Therefore, while 266 may still be an overestimate, this is probably closer to the true N_e of NZ working dogs than 30,000.

In addition to N_e, the accuracy of imputation is highly affected by the quality of the reference panel genotypes (127). In the next two imputation runs, the reference panels comprised more stringently filtered subsets of the WGS variant dataset. For the third run, variants with missing genotype rate > 0.1 were excluded, reducing the size of the reference panel from approximately 16.7M to 13.7M variants. For the fourth run, variants with missing genotype rate > 0.1 and MAF < 1% were excluded from the reference panel, reducing its size to approximately 11.5M variants. This meant rare variants, including some Mendelian variants of interest would not be imputed. However, several studies have suggested the overall accuracy of Beagle imputation increases when rare variants are excluded (127). Like the second imputation run, an N_e of 266 but no genetic map was specified in the latter two runs. The mean DR^2 of each imputation run was calculated to determine which run yielded the highest average accuracy across all sites.

Because imputation of the Mendelian variant loci was of particular interest for this project, a statistical analysis was performed to determine which imputation run yielded the highest average accuracy across these sites. Retaining only Mendelian variant loci, subsets of imputed genotypes from imputation runs 2, 3, and 4 were created with BCFtools (49). The DR^2 value and imputed allele frequency of each marker was extracted and analysed in RStudio. A one-way repeated measure ANOVA was performed to compare the mean DR^2 across imputation runs. This test assumes dependence between the runs and pairs DR^2 values from the same loci, meaning only sites that were imputed in all three imputation runs could be compared. Hence, variants with MAF < 1% or call rate < 0.9 in the WGS sample were excluded (n = 2 Mendelian variant loci). The exclusion of such loci makes the statistical interpretation of the ANOVA difficult, as one must consider the drawback of fewer sites being imputed in the latter runs. However, an independent three-way ANOVA would not have been appropriate for the data, since the assumption of independence is violated.

2.11 Genome-Wide Association Studies for Four Morphological Traits

The final analysis of the project aimed to identify genetic associations with four morphological phenotypes through GWAS. The phenotypes of interest (height (withers to

floor) (mm), length (base of skull to crest of the ilium) (mm), chest circumference (behind the elbow) (mm), and muzzle circumference (mm)) were obtained from all individuals with a tape measure, recorded on paper surveys, and sent via mail to the Al Rae Centre before being entered into the Helical database. Retaining only the measurements from the SNP chip sample, this information was extracted and loaded into NeSI for further analysis. Because sampling for the wider Right Dog project is ongoing, additional SNP chip genotypes had become available at the time of this analysis and were included to increase statistical power. A further 265 individuals had been genotyped with the Axiom™ Canine HD Array meaning that, excluding individuals with no Huntaway or Heading Dog heritage or missing phenotype information, there were 432 dogs in the SNP chip sample. Dogs that were younger than one year of age at the time of sampling were excluded (n = 2), since they may not have reached physical maturity and this could alter the results because their growth is incomplete. The total sample for the GWAS therefore comprised 430 dogs with genotype and phenotype information. Because few estimates for the traits of interest exist for Huntaways and Heading Dogs, breed-specific density plots were created in Rstudio to assess their distributions and means (see Appendix B.6). The samples from which these means were calculated comprised 261 Huntaways and 255 Heading Dogs. Two-sample T-tests revealed that the mean height, length, chest circumference, and muzzle circumference differed significantly between Huntaways and Heading Dogs ($p < 2.2\times10^{-16}$), highlighting the importance of accounting for breed in the GWAS model.

In order to obtain the most accurate results, both genotypes and phenotypes were quality filtered prior to running the GWAS. The sample distribution of each phenotype of interest was inspected in RStudio and outliers were removed. Outliers were defined as values more than three standard deviations from the mean (see Figure 6). There were two height measurements, three length measurements, one chest circumference measurement, and one muzzle circumference measurement that fit this definition. Additionally, two individuals were missing height measurements. Because of rounding, the four quantitative traits investigated here may not be truly continuous. This is most obvious in Figure 6d, where muzzle circumference is separated into distinct bins. However, this is unlikely to have a major impact on the analysis because they are approximately

continuous (129). SNPs with MAF < 1% or missing genotyping rate > 0.1 were excluded from the genotype dataset. SNP chip genotypes, rather than WGS genotypes, are commonly utilised in GWAS since an extremely large sample is required to obtain enough statistical power to detect associations from the number of variants in WGS data (88). Consequently, GWAS are designed to identify associated variants and post-GWAS analysis is required to identify causal variants.

Two popular software packages, PLINK (124) and GCTA (94), were employed to fit additive genetic association models between SNPs and phenotypes of interest. Recessive and dominant genetic models were not fit since 430 is an extremely small sample size for a GWAS and non-additive models require a large amount of statistical power (88).

PLINK (v2.0) (124) was used to fit a generalised linear model (GLM) ($\boldsymbol{y} = \mathbf{Z}\boldsymbol{\beta}_G + \mathbf{X}\boldsymbol{\beta}_X + \boldsymbol{e}$) for each SNP, where $\boldsymbol{y}$ = one of the four phenotypes of interest, $\mathbf{Z}$ = genotype dosage matrix for test SNP (0, 1, or 2), $\boldsymbol{\beta}_G$ = estimated fixed effect of the SNP, $\mathbf{X}$ = fixed factor matrix, $\boldsymbol{\beta}_X$ = estimated effect of the factor variable, and $\boldsymbol{e}$ = random residual effects ($\boldsymbol{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance). In this model, owner-reported breed and sex were fitted as factor variables to account for population stratification. However, studies have shown that fixed variables are insufficient at accounting for population structure (88). Therefore, GCTA (v1.94.1) (94) was used to fit an MLMA-LOCO (mixed linear model association with leave one chromosome out). In this model, the genetic association is still modelled as a fixed effect for each SNP. However, in addition to this, the polygenic effect (i.e. the accumulated effect of all other SNPs, excluding those on the same chromosome) is fit as a random effect: $\boldsymbol{y} = \mathbf{Z}\boldsymbol{\beta}_G + \boldsymbol{g} + \boldsymbol{e}$, where $\boldsymbol{g} \sim \mathrm{N}(0, \mathbf{G}\sigma_g^2)$, with $\mathbf{G}$ representing the genomic relationship matrix (GRM) calculated for all autosomal variants not on the current chromosome, and σ_g^2 representing the (additive) genetic variance (94,130). The GRM consists of A_{jk} scores (see Section 2.8) between pairs of individuals and accounts for population stratification in the model. Because the GRM excludes all variants in the chromosome on which the test SNP is located, the SNP being tested is only analysed once, since it, and any SNPs in LD with it, are excluded from the GRM.

To increase accuracy, the GRM was generated from a highly filtered and pruned subset of SNP chip genotypes. This subset was generated as follows: PLINK (v2.0) was used to remove SNPs with a MAF < 10%, call rate < 0.9, or HWE exact test p-values < 10^{-6}. Because LD can confound genetic relatedness, redundant SNPs in high LD ($R^2 > 0.75$ within 100kb blocks) were pruned out with PLINK (v2.0). The remaining 248,017 variants were included in the GRM generation with GCTA. The subset was split by chromosome and a chromosome-specific GRM was generated for each autosome. An MLMA was fit for all SNPs with MAF > 1%, call rate > 0.9, and HWE exact test p-values > 10^{-6} (n = 541,538 SNPs). Finally, the output files were combined across chromosomes and the qqman package (131) from RStudio was used to create Manhattan plots of Bonferroni-corrected -$\log_{10}$ p-values (132). NCBI Genome Data Viewer was used to determine which genes were within 100kb of significantly associated SNPs.

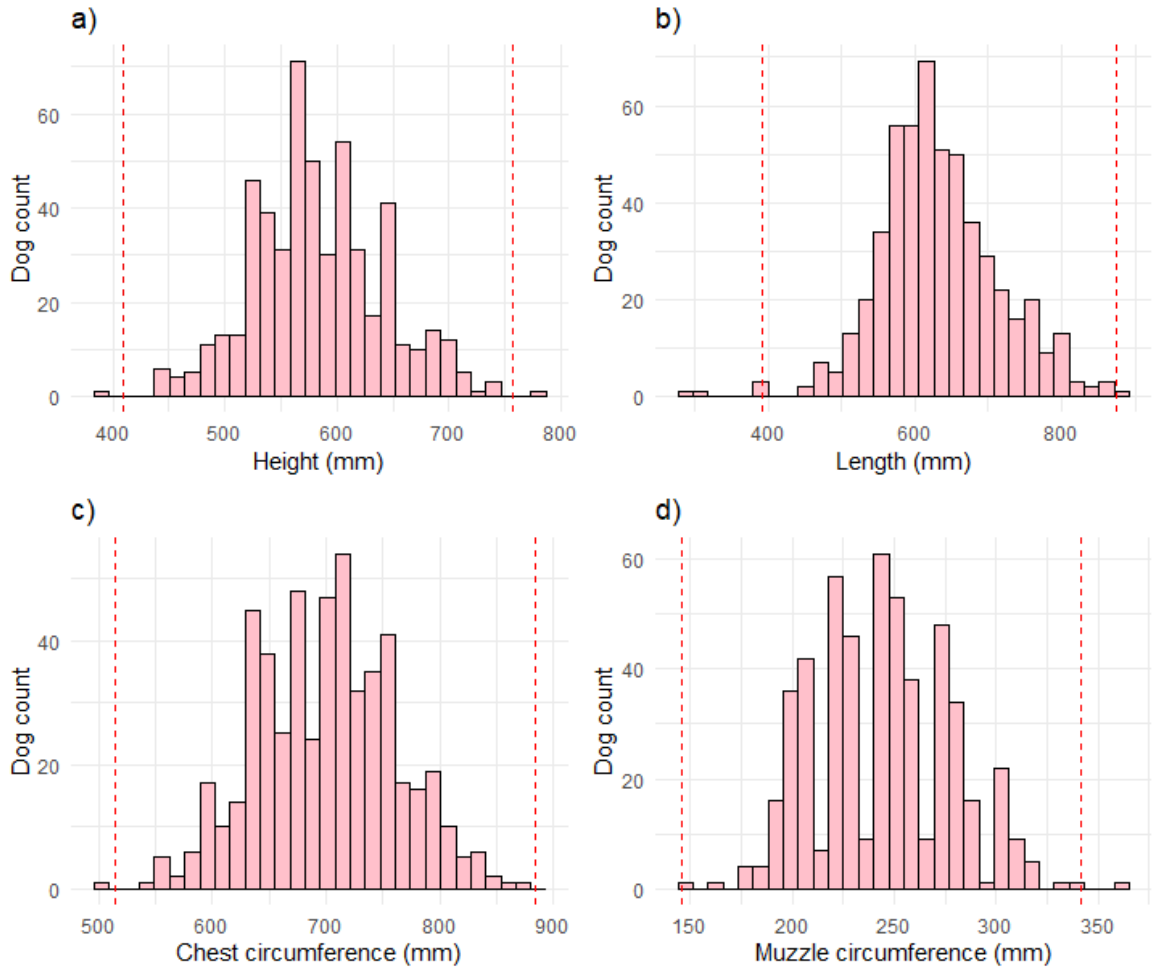

Figure 6. Histograms of GWAS sample phenotypic distributions. **a)** Height, **b)** Length, **c)** Chest circumference, **d)** Muzzle circumference. Dashed lines indicate outlier thresholds (± 3SD from mean) and measurements beyond this range were excluded from the GWAS.

Chapter 3: Results

3.1 Sequencing Summary Statistics

The whole genomes of 250 Huntaways and Heading Dogs were sequenced with an Illumina NovaSeq 6000 and aligned to the canFam4 (UU_Cfam_GSD_1.0) reference assembly (98) with BWA-MEM (51). Summary statistics from the preprocessing and alignment of reads were analysed in RStudio. A per-sample mean of 2.56% of reads were removed due to low quality, 0.02% had too many Ns, 0.03% were too short, 0.02% had low complexity, and 0.04% were duplicates. Figure 7 displays the average read depth, average read quality, and error rate across samples. The average read depth (number of bases mapped divided by the length of the reference genome) per sample ranged from 10.68× to 46.38×, with a mean of 22.88×. The average read quality (ratio between the sum of base qualities and read length) per sample ranged from 35 to 35.9, with a mean of 35.68. The error rate (number of mismatches divided by the number of bases mapped) per sample ranged from 0.43% to 0.71%, with a mean of 0.51%. Of the reads that passed all pre-processing filters, the percentage that were properly mapped and paired per sample ranged from 99.6% to 99.8%, with a mean of 99.71%. Appendix B.1 shows the distribution of the proportion of reads that were properly mapped and paired across samples.

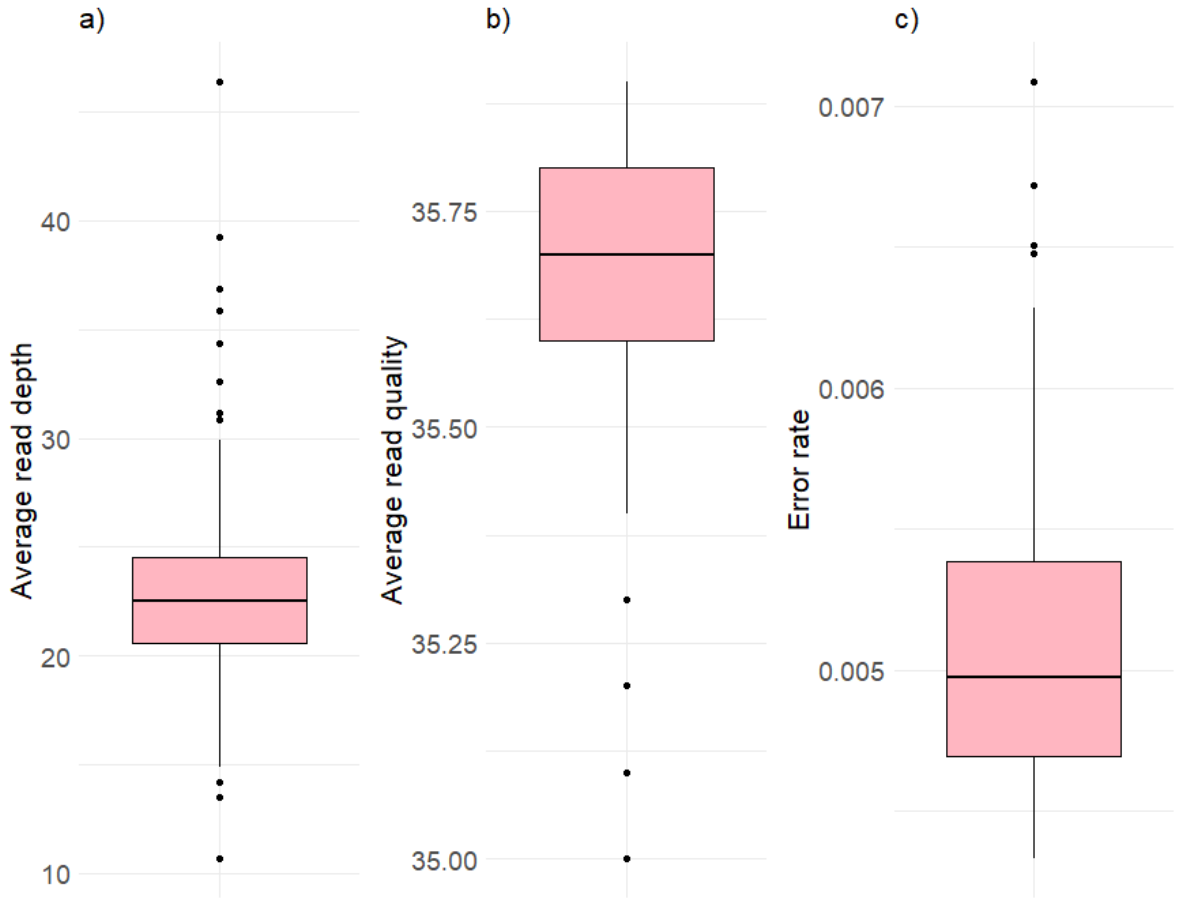

Figure 7. Boxplots of sequence alignment summary statistics. **a)** Average read depth across samples, **b)** Average read quality across samples, **c)** Error rate across samples.

3.2 Annotation of WGS Variants

GATK best practices was used to call and filter genetic variants in the 250 whole genome sequences, where approximately 20 million variants were called and 16.7 million passed quality filters. The filtered variants made up the current WGS variant dataset. Variants were annotated with SnpEff, which predicted 58,485 variants (0.35%) to be 'functional' (high or moderate impact) (see Table 4). It should be noted that this value is slightly less than the sum of high and moderate annotations shown in Table 4 because single variants can affect multiple transcripts with alternative impacts. The WGS dataset was also annotated with Ensembl VEP, which predicted a very similar number of variants (58,684) in these impact categories. The intersect (overlap) between the variants predicted by the respective tools to be functional was 40,067, where 18,418 variants were predicted by SnpEff but not Ensembl VEP to be functional and 18,617 variants were predicted by Ensembl VEP but not SnpEff to be functional. The vast majority of variants (approximately 90%) were predicted by both tools to be modifiers, meaning there is no evidence that they impact proteins. A comparatively small proportion of high-impact variants (~40%) were predicted by both tools to cause high impacts, while a large proportion of modifier variants (>99%) were predicted by both tools to be modifiers (see Table 4). This suggests high-impact predictions are more likely to be false positives and reinforces the rationale of using multiple VEPs to filter for functionality.

RNA-Seq expression data was used to verify gene expression at the loci of variants predicted by both tools to be functional. Publicly available RNA-Seq data from 17 tissues was aligned to the canFam4 assembly with STAR and an 'RNABamDepth' annotation was added to each variant. Variants with RNABamDepth < 5 (n = 2606) were assumed to be in falsely annotated genes. This analysis yielded a final list of 37,461 functional predictions (i.e. variants predicted by both VEPs to directly impact protein sequence at expressed loci). These annotations will be used in future projects to prioritize likely functional variants.

Table 4: Summary of SnpEff- and Ensembl VEP-predicted impacts

	High	Moderate	Low	Modifier
SnpEff	10,106	48,690	79,860	16,603,999
Ensembl VEP	10,128	49,271	76,548	16,599,316
Both [a]	4173	35,809	58,345	16,574,803

Number of WGS variants predicted to have each level of impact by two computational VEPs.

[a] *Number of variants predicted by both SnpEff and Ensembl VEP to have each impact.*

3.3 Predicted Variant Effects in Huntaways and Heading Dogs

The WGS variant dataset was subset into the two breeds of interest to determine whether there was a difference in the proportion of types of variants between them. The Huntaway dataset comprised 14,878,265 variants from 130 dogs and the Heading Dog dataset comprised 14,024,707 variants from 104 dogs. Approximately 12.8 million variants segregated in both Huntaways and Heading Dogs, 2.1 million segregated in only Huntaways, and 1.2 million segregated in only Heading Dogs. In other words, 86% of the 14.9 million variants segregating in Huntaways also segregated in Heading Dogs, and 91% of the 14 million variants segregating in Heading Dogs also segregated in Huntaways. The proportion of each effect category (splice region, start or stop lost or gained, frameshift, in-frame indel, synonymous, intron, downstream or upstream, inter- or intragenic, missense, non-coding transcript, UTR, and other) was extremely similar across breeds (see Appendix B.2). The proportion of predicted effects amongst variants in coding regions is displayed in Figure 8 and the proportion of predicted effects amongst variants in non-coding variants is displayed in Figure 9. Of variants intersecting coding regions, approximately 40% were synonymous, 30% were missense, and 15% were in splice regions, with other categories making up small minorities (Appendix A.2 outlines which SnpEff 'effect' annotations were included in each effect category and Appendix B.2 displays the counts and proportions of each effect by breed). The proportion of predicted impacts (high, moderate, low, and modifier) was also extremely similar between breeds (see Appendix B.2).

The proportions of coding-region and 'functional' annotations in each breed is summarised in Table 5. It should be noted that functional prediction counts represent total annotations, not variants, since single variants can have several consequences. In both breeds, over 99% of annotations were in non-coding regions, with approximately 45% in introns or UTRs and 55% in intergenic regions, intragenic regions, or non-coding transcripts. Similarly, over 99% of variant annotations in both breeds were low or modifier impacts. A chi-squared test provided no evidence of a significant difference in the proportion of 'functional' predictions between Huntaways and Heading Dogs ($p = 0.47$). There was a significant difference in the proportion of coding-region variants between Huntaways and Heading Dogs ($p = 3.27e^{-4}$), however, this difference was small (0.01%).

Table 5: Summary of SnpEff annotations by breed

	Huntaways		Heading Dogs	
	Count	**Proportion**	**Count**	**Proportion**
Functional [c]	136,552	0.003	126,002	0.003
Non-functional	52,725,911	0.997	48,788,919	0.997
Coding region [d]	125,911	0.0084	122,633	0.0083
Non-coding region	14,778,641	0.9916	13,927,330	0.9917

Note: *Counts and proportions of functional predictions consider the impact annotations for all transcripts that each variant affects; counts and proportions of coding region predictions consider only the most severe effect annotation of each variant.*

[c] *High and moderate impacts.*

[d] *Splice region, start or stop lost or gained, frameshift, in-frame indel, synonymous, start or stop retained, missense, and other extreme effects.*

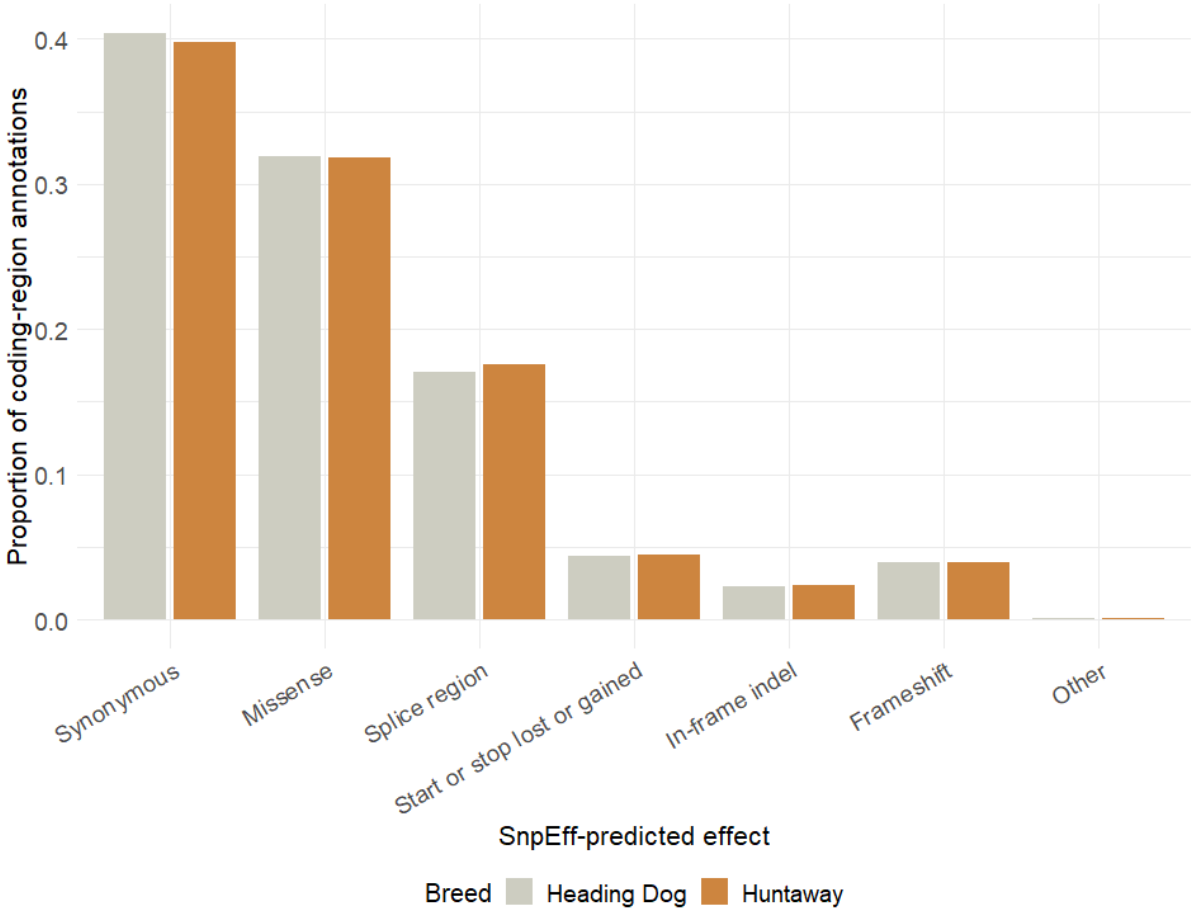

Figure 8. SnpEff-predicted effects of coding-region variants by breed. **Note:** Proportions are calculated from the most severe predicted effect of each variant; Appendix A.2 outlines which 'effect' annotations are included in each effect category; splice region annotations include all variants within 3 bases of exons, within 8 bases of introns, and at branch points (118); 'other' annotations include bidirectional gene fusion, exon loss variant, gene fusion, and transcript ablation.

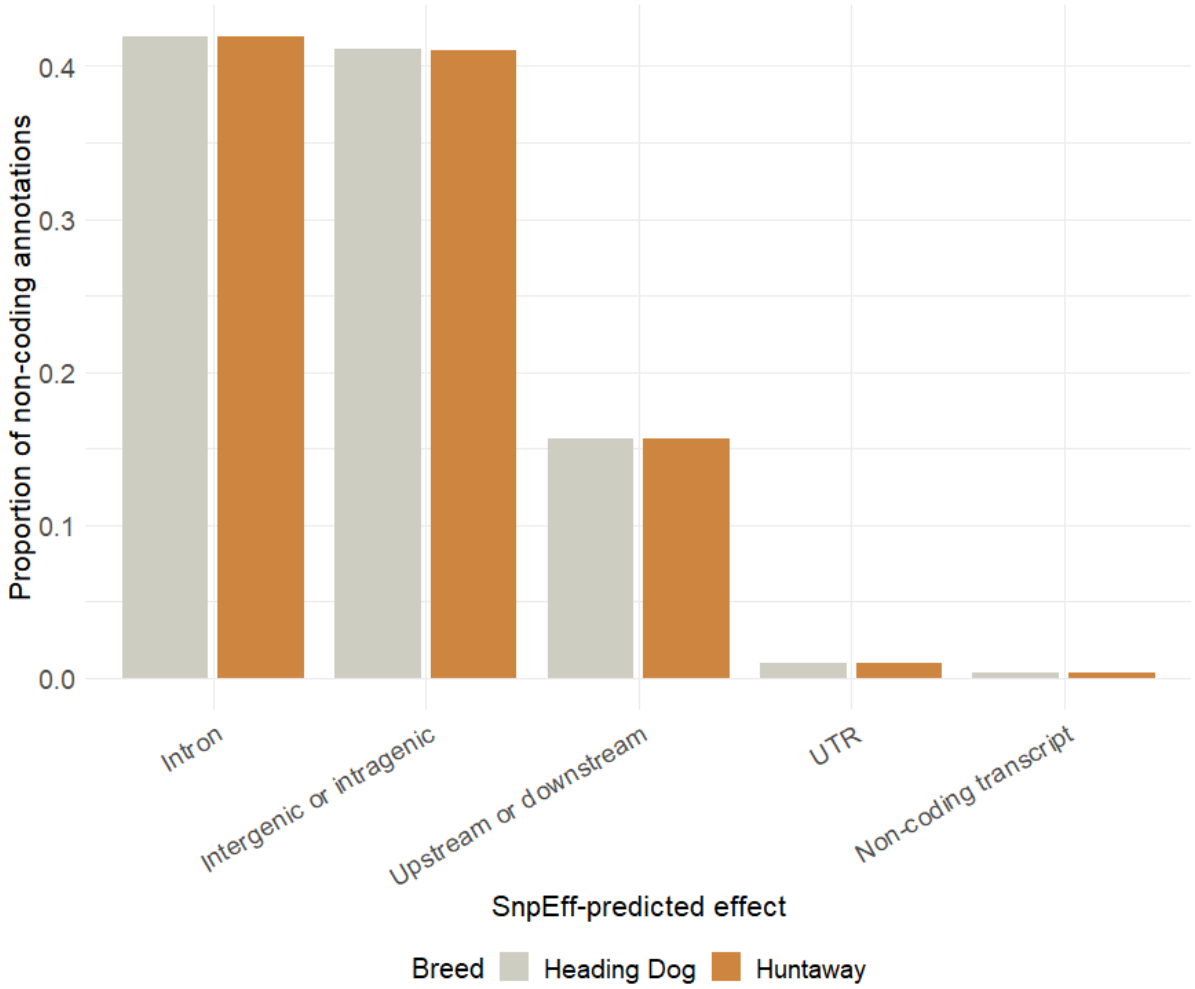

Figure 9. SnpEff-predicted effects of variants in non-coding regions by breed. **Note:** Proportions are calculated from the most severe predicted effect of each variant; Appendix A.2 outlines which 'effect' annotations are included in each effect category; SnpEff defines intragenic variants as those that hit a gene, but no transcripts within the gene (118).

3.4 Previously Reported Mendelian Variants Segregating in the WGS Sample

The WGS dataset was intersected with a list of 395 variants reported by OMIA (106) to be likely causal in dogs. Of these variants, which were predominantly Mendelian in effect, 27 were identified as segregating in at least one dog in the WGS sample. The breed-specific and overall allele frequencies of each variant in the WGS sample are described in Table 6 and the genotype frequencies are described in Table A6 (see Appendix B.3). There were 368 Mendelian variants from the OMIA database that did not segregate in the WGS sample. These are outlined in Table A7. Based on OMIA-reported phenotypes, 14 of the variants segregating in the sample impacted morphological or aesthetic traits (e.g. coat colour/patterning), while 13 have been previously implicated in disease.

All 27 variants passed the quality filters outlined in Figure 4 and sequence data (BAM files) were inspected in IGV to verify the accuracy of each variant call. Based on this inspection, all were classified as true calls, however, the *ABCB1* variant mapped upstream of the gene in the canFam4 reference assembly and the *MFSD12* variant appeared to be misaligned (see Table 7). Gene expression at all loci was verified with publicly available RNA-Seq data (see Table 3), and the literature pertaining to each of the 27 Mendelian variants was critically reviewed to assess their functional candidacy. One variant was classified as functional, 13 were classified as likely functional, nine were classified as possibly functional, and four were classified as unlikely to be functional in the NZ population (see Table 7). Highlighted in blue in Tables 6 and 7 are five variants that were considered strong candidates for use as selection diagnostics. These include a frameshift deletion in *CUBN* (p.Gln2798fs), a nonsense SNP in *CLN8* (p.Trp195*), a frameshift insertion in *SGSH* (p.Tyr229fs), a missense SNP in *SOD1* (p.Glu40Lys), and a splice site SNP in *VWF* (p.Ser2479Ser). See Section 4.4 for a more detailed description of each variant's functional evaluation.

Table 6: Summary of Mendelian variants segregating in a sample of NZ farm dogs

				Allele frequency		
OMIA ID [a]	**Gene** [a]	**Chr**	**Variant** [b]	**Huntaways** [c]	**Heading Dogs** [c]	**Total** [d]
1603	*MC5R*	1	g.24541931C>T	0.638	0.476	0.568
447	*CUBN*	2	g.18932445del	0	0.029	0.012
343	*MC1R*	5	g.64186728G>A	0.050	0.067	0.06
34	*MC1R*	5	g.64186854C>T	0.953	0.589	0.8
577	*SGSH*	9	g.2406797insA	0.004	0	0.01
851	*BTBD17*	9	g.6924623insG	0.102	0.238	0.162
1273	*CNTNAP1*	9	g.20172413C>T	0	0.005	0.002
31	*TYRP1*	11	g.33376317T>A	0.042	0	0.026
267	*TYRP1*	11	g.33385200C>T	0.05	0.005	0.028
796	*TYRP1*	11	g.33385242del3	0.042	0	0.02
442	*ABCB1*	14	g.13720387A>C	0.310	0.389	0.33
1422	*IGF1-AS*	15	g.41511739C>T	0.050	0.353	0.186
1444	*LMBR1*	16	g.20112737C>T	0.109	0.113	0.108
458	*CBD103*	16	g.55468988del3	0.118	0.279	0.184
1522	*RETN*	20	g.52842420C>T	0.077	0	0.046
1081	*MFSD12*	20	g.56247895C>T	0.289	0.345	0.32
106	*ATP7B*	22	g.196868G>A	0.058	0.014	0.038
30	*ASIP*	24	g.23906214C>T	0.020	0	0.012
360	*MLPH*	25	g.48403161G>A	0.050	0.197	0.108
401	*VWF*	27	g.7140281C>T	0.035	0	0.022
35	*KRT71*	27	g.44113063G>A	0.012	0.053	0.032
274	*CYP1A2*	30	g.38261635C>T	0.438	0.101	0.302
36	*SOD1*	31	g.27123057G>A	0.217	0.019	0.126
48	*FGF5*	32	g.35494497C>A	0.097	0.059	0.096
103	*P3H2*	34	g.22231216C>G	0	0.014	0.006
338	*CLN8*	37	g.30769171G>A	0.027	0.063	0.042
612	*KCNJ10*	38	g.22970389del	0.968	0.861	0.918

Note: *Chr = chromosome; ins = insertion; del = deletion; Allele frequencies refer to the alternative allele based on canFam4.*

[a] *As reported by the OMIA database.*

[b] *Genomic positions in canFam4.*

[c] *Breed-specific allele frequencies in the WGS dataset.*

[d] *Allele frequencies in the total WGS dataset.*

Table 7: Functional assessment of Mendelian variants segregating in a sample of NZ farm dogs

OMIA ID [a]	Gene [a]	Associated phenotype [a]	Inheritance [a]	Variant call validation [b]	Functional assessment [c]
1603	*MC5R*	Reduced hair shedding	Likely multifactorial	True variant	Possibly functional
447	*CUBN*	Intestinal cobalamin malabsorption	Autosomal recessive	True variant	Likely functional
343	*MC1R*	Red/yellow coat	Autosomal recessive	True variant	Likely functional
34	*MC1R*	Black melanistic mask	Autosomal dominant	True variant	Possibly functional
577	*SGSH*	Mucopolysaccharidosis IIIA	Autosomal recessive	True variant	Functional
851	*BTBD17*	Lethality	Autosomal recessive lethal	True variant	Unlikely to be functional
1273	*CNTNAP1*	Laryngeal paralysis and polyneuropathy	Autosomal recessive	True variant	Possibly functional
31	*TYRP1*	Brown	Autosomal recessive	True variant	Likely functional
267	*TYRP1*	Brown	Autosomal recessive	True variant	Likely functional
796	*TYRP1*	Brown	Autosomal recessive	True variant	Likely functional
442	*ABCB1*	Adverse reaction to certain drugs	Autosomal incomplete dominance	Misaligned	Unlikely to be functional
1422	*IGF1-AS*	Height	Multifactorial	True variant	Likely functional
1444	*LMBR1*	Dew claws	Autosomal dominant	True variant	Likely functional
458	*CBD103*	Coat colour, dominant black	Autosomal dominant	True variant	Possibly functional, incomplete penetrance
1522	*RETN*	Modifier of copper toxicosis	Unknown	True variant	Unlikely to be functionality
1081	*MFSD12*	Coat colour, white or cream	Autosomal recessive	Misaligned	Possibly functional, incomplete penetrance
106	*ATP7B*	Wilson disease	Autosomal recessive	True variant	Possibly functional
30	*ASIP*	Black	Autosomal recessive	True variant	Likely functional, possible incomplete penetrance

360	*MLPH*	Dilute	Autosomal recessive	True variant	Likely functional
401	*VWF*	Von Willebrand disease I	Autosomal dominant or recessive	True variant	Likely functional, incomplete penetrance
35	*KRT71*	Curly coat	Unknown	True variant	Possibly functional
274	*CYP1A2*	Metabolizer of a cognitive enhancer	Unknown	True variant	Likely functional
36	*SOD1*	Degenerative myelopathy	Autosomal recessive	True variant	Likely functional, incomplete penetrance
48	*FGF5*	Long hair	Autosomal recessive	True variant	Possibly functional
103	*P3H2*	Lundehund syndrome	Autosomal recessive	True variant	Possibly functional, incomplete penetrance
338	*CLN8*	Neuronal ceroid lipofuscinosis	Autosomal recessive	True variant	Likely functional
612	*KCNJ10*	Ataxia, cerebellar	Autosomal recessive	True variant	Unlikely to be functional

[a] *As reported by the OMIA database.*

[b] *Reclassification of variant calls in WGS sample based on a visual inspection of sequence reads in IGV.*

[c] *Functional classification of variants based on a literature review and RNA-Seq data (does not consider any phenotypic evidence from the current sample).*

3.5 High-Impact Variants within Genes of Interest Segregating in the WGS Sample

With the aim of identifying novel functional candidates, the WGS dataset was queried for high-impact variants within 132 genes previously shown to harbour phenotype-causing LOF variants. In the WGS sample, 92 variants within 50 genes of interest were predicted by SnpEff to disrupt protein sequence. These are described in Appendix B.4. Sequence read data was visualised in IGV to classify 25 of these as true calls, four as possible calls, and 63 as likely false positive calls. The functional candidacy of true and possible calls was determined based on their allele frequencies and a review of the literature pertaining to each gene. On this basis, 14 variants within 12 unique genes were highlighted as functional candidates (see Table 8), while 15 variants within nine genes were classified as unlikely to be functional in the NZ population. It should be noted that five of the functional candidates had been previously described by OMIA (highlighted in blue in Table 8), meaning nine were novel. See Section 4.5 for a more detailed description of functional evaluations.

Table 8: Summary of high-impact candidate functional variants segregating in a sample of NZ farm dogs

Gene of interest	Candidate function [a]	Chr	Observed variant	Protein consequence	AF [b]	Variant call validation [c]
CUBN	Cobalamin malabsorption	2	g.18932444del	p.Gln2798fs	0.012	True variant
CNGB1	Retinal atrophy	2	g.57909801del	p.Ser1160fs	0.006	True variant
ABCA4	Stargardt disease	6	g.55659310del2	p.Leu1818fs	0.006	True variant
SGSH	Mucopolysaccharidosis	9	g.2406797ins	p.Tyr229fs	0.01	True variant
CNP	Lysosomal storage disease	9	g.20768542ins	p.Arg137fs	0.002	True variant
SLC3A1	Cystinuria	10	g.47766395G>A	Splice donor	0.002	True variant
TYRP1	Brown, liver colour	11	g.33385200C>T	p.Gln992*	0.028	True variant
CCDC66	Retinal atrophy	20	g.33996814del4	p.Arg30fs	0.03	True variant
GLB1	Gangliosidosis	23	g.4009430del	p.Arg642fs	0.002	True variant
CYP1A2	Metabolizer of a cognitive enhancer	30	g.38260059del	p.Asn253fs	0.062	True variant
CYP1A2	Metabolizer of a cognitive enhancer	30	g.38261635C>T	p.Arg373*	0.302	True variant
CYP1A2	Metabolizer of a cognitive enhancer	30	g.38263975ins10	p.Gly492fs	0.074	True variant
STK36	Primary ciliary dyskinesis	37	g.25198749G>A	p.Trp1290*	0.194	True variant
CLN8	NCL	37	g.30769171G>A	p.Trp195*	0.042	True variant

The high-impact variants in this table were classified as functional candidates based on an inspection of sequence data, their allele frequencies, and a review of the literature.

Note: *Variants previously described in OMIA are highlighted in blue; chr = chromosome; AF = allele frequency; fs = frameshift.*

[a] *OMIA-reported phenotypes from LOF variants within genes of interest.*

[b] *Allele frequencies of predicted high-impact variants in the WGS dataset.*

[c] *Reclassification of variant calls in WGS sample based on a visual inspection of sequence reads in IGV.*

3.6 SNP Chip Data Quality

To assess the quality of the SNP chip genotypes, the sample distributions of four quality metrics were analysed in RStudio (see Figure 10). Figure 10a displays a histogram of missing call rate per sample (n = 299 samples). This distribution is right-skewed and centres around a mean of 0.017. Only two samples had a missing call rate ≥ 0.1, which is a common QC threshold (122). The distribution of per-sample inbreeding coefficients (F statistics) is displayed in Figure 10b. It is roughly normally distributed around a mean of 0.047. According to a one-tailed T-test, the mean F statistic was significantly greater than 0 ($p < 2.2\times10^{-16}$), indicating the presence of inbreeding. According to a two-sample T-test, there was no evidence of a significant difference in the mean F statistic between Huntaways and Heading Dogs (p = 0.28), where both breed-specific means (0.02 and 0.03 respectively) were lower than the overall mean. Several individuals had F statistics of less than 0, but one individual had an extremely low value compared to the rest of the sample (F = -0.35), suggesting the individual may be genetically distinct or the sample was of poor quality. This individual also had the highest missing call rate (0.12) and was classified as an outlier.

Figure 10c displays a histogram of missing genotype rate per SNP (n = 719,182 SNPs). Like per-sample missing call rate, this distribution is skewed to the right and centres around a mean of 0.017. Approximately 45% of SNPs had a missing genotype rate of 0 (n = 322,364 SNPs), and only 3% had a missing genotype rate ≥ 0.1 (n = 20,025 SNPs). There were 89,316 loci on the array that were monomorphic, meaning the alternative allele was either absent or fixed in the current sample. Figure 10d displays the distribution of MAFs across polymorphic sites (n = 629,866 SNPs). Approximately 82% of total variants had MAF > 1% (n = 588,148 SNPs) and an enrichment of rare variants was observed. It was concluded based on these metrics that the SNP chip dataset was of high quality.

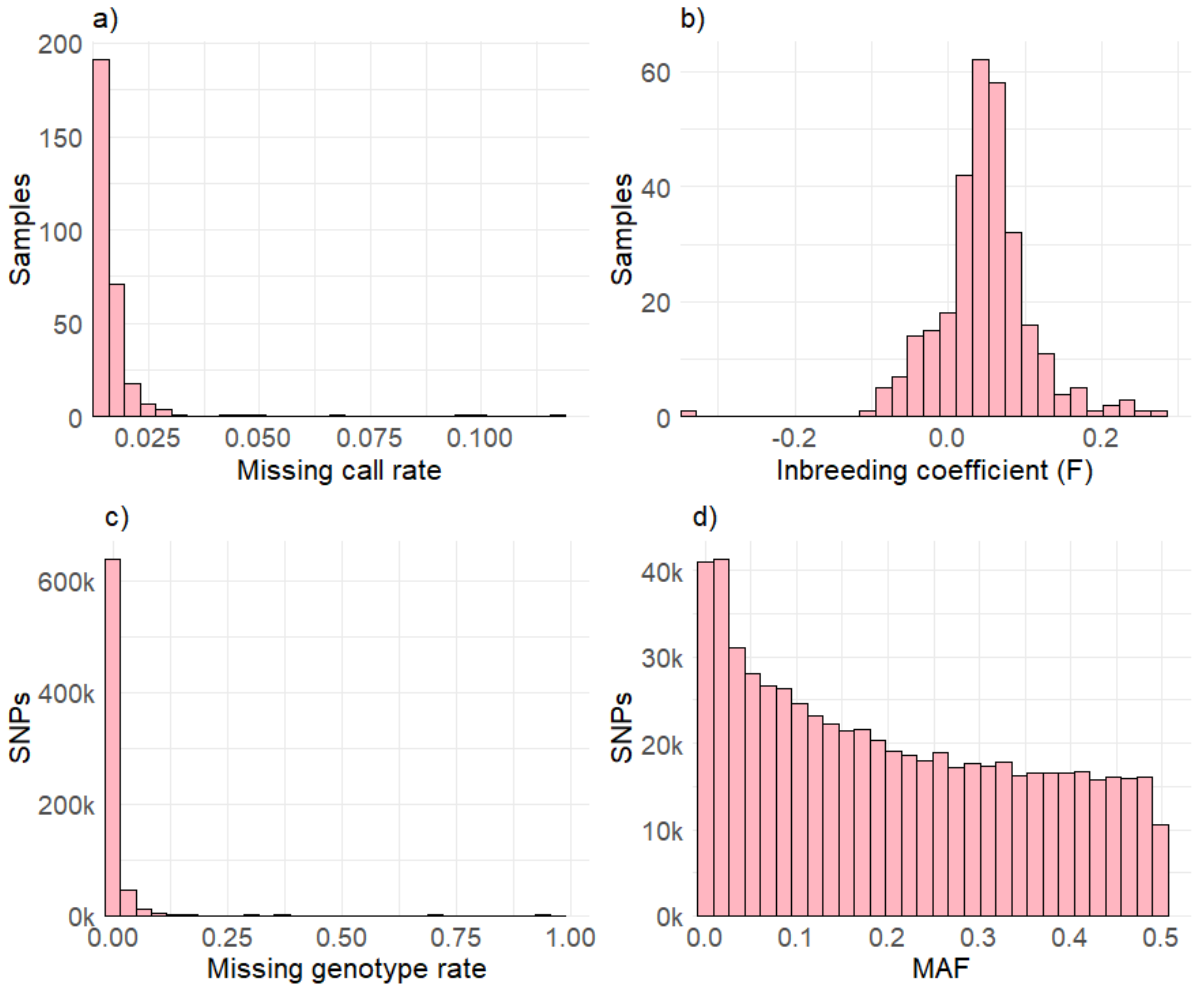

Figure 10. Histograms of quality metrics in the SNP chip dataset. **a)** Missing call rate per sample, **b)** Inbreeding coefficient per sample, **c)** Missing genotype rate per SNP, **d)** MAF per SNP across polymorphic sites.

3.7 Genetic Verification of Matching WGS/SNP Chip Samples

The identification of predictive markers for functional variants in the following analysis relied heavily on the genetic information from individuals who had been both sequenced and SNP chip genotyped. Therefore, it was important to verify that this subset, referred to as the overlapping sample, was correctly labelled (i.e. WGS genotypes that were assigned to the same animal as SNP chip genotypes were sourced from genetically identical DNA). To this end, A_{jk} scores and discordance rates of genotype pairs were compared to determine whether within-individual comparisons could be distinguished from between-individual comparisons using these metrics.

Figure 11 compares box plots of A_{jk} scores from between-individual comparisons (i.e. individuals' WGS genotypes compared to other individuals' SNP chip genotypes) and within-individual comparisons (i.e. overlapping individuals' WGS genotypes compared to their own SNP chip genotypes). A large difference in A_{jk} scores was observed between the two groups, with means of –0.007 and 0.978 respectively. This difference was found to be highly significant according to a two-sample T-test ($p < 2.2\times10^{-16}$), providing evidence that A_{jk} score could be used to distinguish genetically identical pairs from genetically distinct pairs. The range of A_{jk} scores from within-individual pairs was small excluding three pairwise comparisons that yielded A_{jk} scores of 0.203, 0.568, and 0.689. These scores are considerably lower than would be expected for genetically identical samples. The range of A_{jk} scores was larger in the between-individual group than in the within-individual group, with pairs clustering around 0, 0.25, and 0.5. It should be noted that the number of comparisons was considerably greater in the between-individual group (n = 74,312 pairs) than in the within-individual group (n = 187 pairs), contributing to the larger range. Usually, A_{jk} scores range from 0 to 1, and scores beyond this range may indicate that the sample is inbred (133). In the current study, scores above 1 and below 0 were observed. This is likely a result of individuals being compared to themselves, creating an artificially inbred population. Additionally, the highest expected A_{jk} score for non-identical animals is approximately 0.5 (for full sibling pairs); however, one between-individual comparison yielded an extremely high A_{jk} score of 0.787.

A comparison of discordance rates between within-individual pairs and between-individual pairs supported the above findings (see Figure 12). The range of discordance rates from within-individual pairs was extremely small, excluding three outliers, and centred around a mean of 0.005. The fact that there was a number of differences between WGS and SNP chip genotypes in identical samples indicates discordance between the technologies. The range of discordance rates from between-individual pairs was wider and centred around a mean of 0.393. Again, the wider range of discordance rates amongst between-individual comparisons could be a result of the much larger sample size. The difference between these two means was highly significant according to a two-sample T-test ($p < 2.2\times10^{-16}$), indicating discordance rate could also be used to distinguish genetically identical pairs from genetically distinct pairs. Three within-individual pairs yielded considerably higher discordance rates than would be expected for genetically identical samples (discordance rates = 0.4, 0.2, and 0.1). These were the same three pairs with greater-than-expected A_{jk} scores, providing further evidence that they were not genetically similar enough to have been from identical DNA. On this basis, they were excluded from the overlapping sample in the following LD analysis, leaving 184 confirmed overlapping individuals.

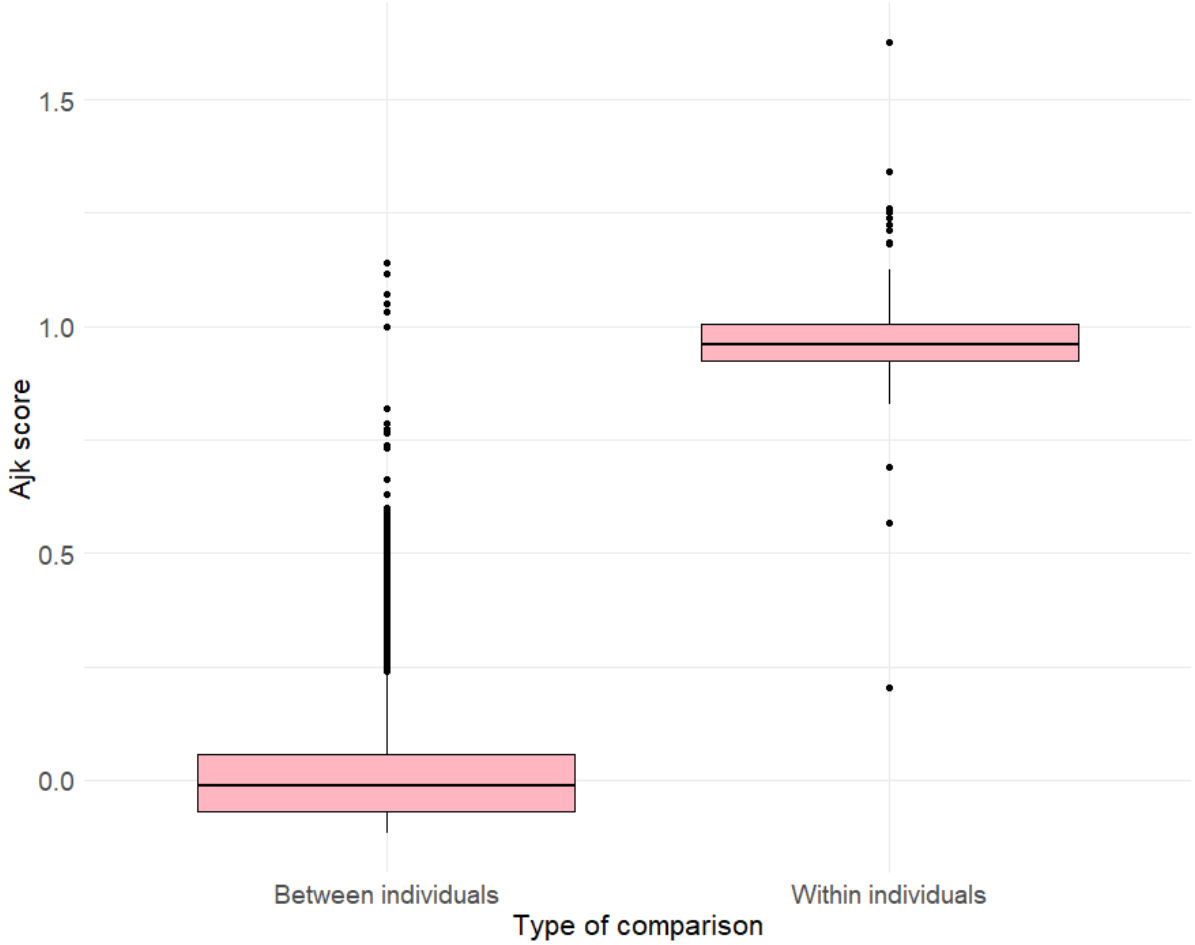

Figure 11. Box plots of A_{jk} scores in between- and within-individual comparisons. **Note:** Between-individual comparisons = WGS genotypes compared to SNP chip genotypes from other individuals (n = 74,312 pairs); within-individual comparisons = WGS genotypes compared to SNP chip genotypes from the same individual (n = 187 pairs).

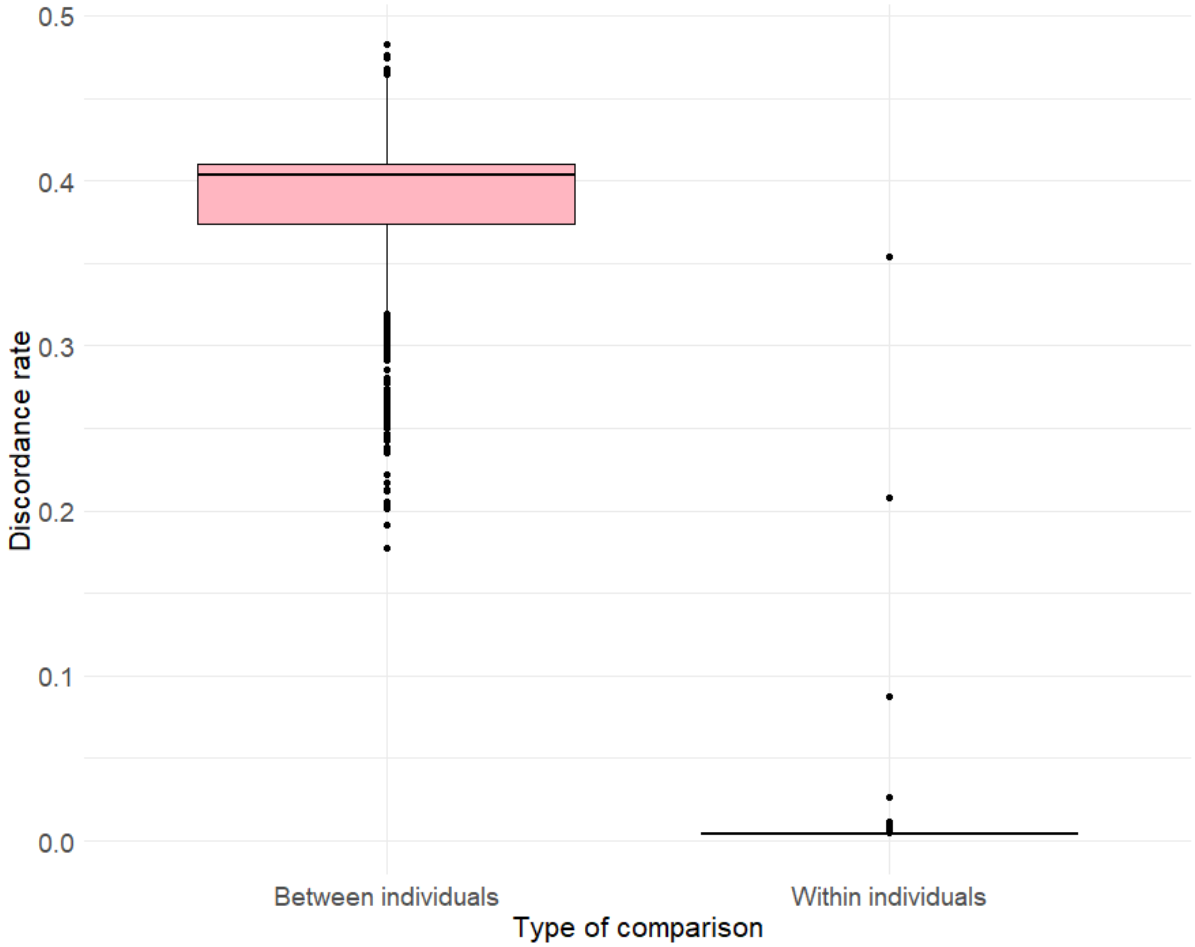

Figure 12. Box plots of discordance rates in between- and within-individual comparisons. **Note:** Between-individual comparisons = WGS genotypes compared to SNP chip genotypes from other individuals (n = 74,312 pairs); within-individual comparisons = WGS genotypes compared to SNP chip genotypes from the same individual (n = 187 pairs).

3.8 Predictive Markers for Mendelian Variants of Interest

With the aim of identifying predictive SNP markers for functional Mendelian variants, the R^2 allele count correlation was calculated between Mendelian variants of interest (see Table 6) and all loci on the Axiom™ Canine HD Array. Only the 184 individuals that were genetically confirmed to be both SNP chip genotyped and whole genome sequenced were included in this analysis (see Section 3.7). There were 66 positions on the array that correlated with 18 of the Mendelian variants with an $R^2 > 0.2$. These loci are summarised in Appendix B.5. Table 9 describes candidate markers, which were defined as the highest correlated locus on the SNP chip to each Mendelian variant. The genomic positions of Mendelian variants in canFam4 are also outlined in Table 9. In Figure 13, Mendelian variant genotypes are regressed against candidate marker genotypes, where each point represents an individual in the overlapping sample. Five Mendelian variants could be directly genotyped with the SNP chip, since they were on the array. These included variants in the *FGF5, P3H2, CYP1A2, MFSD12*, and *IGF1-AS* genes. There were also SNPs on the array that correlated perfectly ($R^2 = 1$) with the Mendelian variants in *RETN*, *SOD1*, and *KCNJ10*, despite them not being on the array. These were classified as strong candidate predictive markers. Moderate candidate markers were identified for Mendelian variants in *MC1R* and *MPLH*, since they were in high but not perfect LD ($1 > R^2 > 0.8$). Weak candidate markers were identified for Mendelian variants in *MC5R, MC1R, BTBD17, ABCB1, LMBR1, CBD103, ATP7B*, and *KRT71*, since they were in moderate LD ($0.2 < R^2 < 0.8$). All candidate markers had low missing genotype rates (see Table 9), with MACs ranging from 6 to 186.

Table 9: Candidate predictive markers for Mendelian variants of interest

Mendelian variant [a]			Candidate marker [b]					
OMIA ID	**Position**	**Gene**	**Array ID**	**Position**	**MAF [c]**	**MAC [c]**	**Missing rate [c]**	**R^{2} [d]**
1603	Chr1: 24541931	*MC5R*	AX-167319705	Chr1: 24542076	0.363	133	0.005	0.59
343	Chr5: 64186728	*MC1R*	AX-168202811	Chr5: 64190343	0.160	59	0.000	0.23
34	Chr5: 64186854	*MC1R*	AX-167328853	Chr5: 64156061	0.196	72	0.000	0.83
851	Chr9: 6924623	*BTBD17*	AX-168006758	Chr9: 6923237	0.225	82	0.011	0.70
442	Chr14: 13720387	*ABCB1*	AX-168108844	Chr14: 13735843	0.261	96	0.000	0.76
1422	Chr15: 41511739	*IGF1-AS*	AX-167351336	Chr15: 41511739	0.204	75	0.000	1
1444	Chr16: 20112737	*LMBR1*	AX-167191180	Chr16: 20109337	0.209	77	0.000	0.47
458	Chr16: 55468987	*CBD103*	AX-167495458	Chr16: 55466525	0.294	108	0.000	0.52
1522	Chr20: 52842420	*RETN*	AX-167560012	Chr20: 52822752	0.038	14	0.000	1
1081	Chr20: 56247895	*MFSD12*	AX-168153313	Chr20: 56247895	0.321	118	0.000	1
106	Chr22: 196868	*ATP7B*	AX-167215741	Chr22: 218897	0.030	11	0.005	0.36
360	Chr25: 48403161	*MLPH*	AX-167774718	Chr25: 48428512	0.131	47	0.022	0.80
35	Chr27: 44113063	*KRT71*	AX-167751297	Chr27: 44098284	0.141	52	0.000	0.21
274	Chr30: 38261635	*CYP1A2*	AX-168196296	Chr30: 38261635	0.285	105	0.000	1
36	Chr31: 27123057	*SOD1*	AX-167251279	Chr31: 27133605	0.106	39	0.000	1
48	Chr32: 35494497	*FGF5*	AX-167434147	Chr32: 35494497	0.093	34	0.005	1
103	Chr34: 22231216	*P3H2*	AX-167299299	Chr34: 22231216	0.008	3	0.000	1
612	Chr38: 22970388	*KCNJ10*	AX-167998824	Chr38: 22968796	0.087	32	0.000	1

Note: *MAF = minor allele frequency; MAC = minor allele count; chr = chromosome.*

[a] OMIA-derived variants segregating in the WGS sample.

[b] SNPs on the Axiom™ Canine HD Array in the highest LD with Mendelian variants.

[c] Calculated from the overlapping sample (n = 184).

[d] Squared allele count correlations between Mendelian variants and candidate markers.

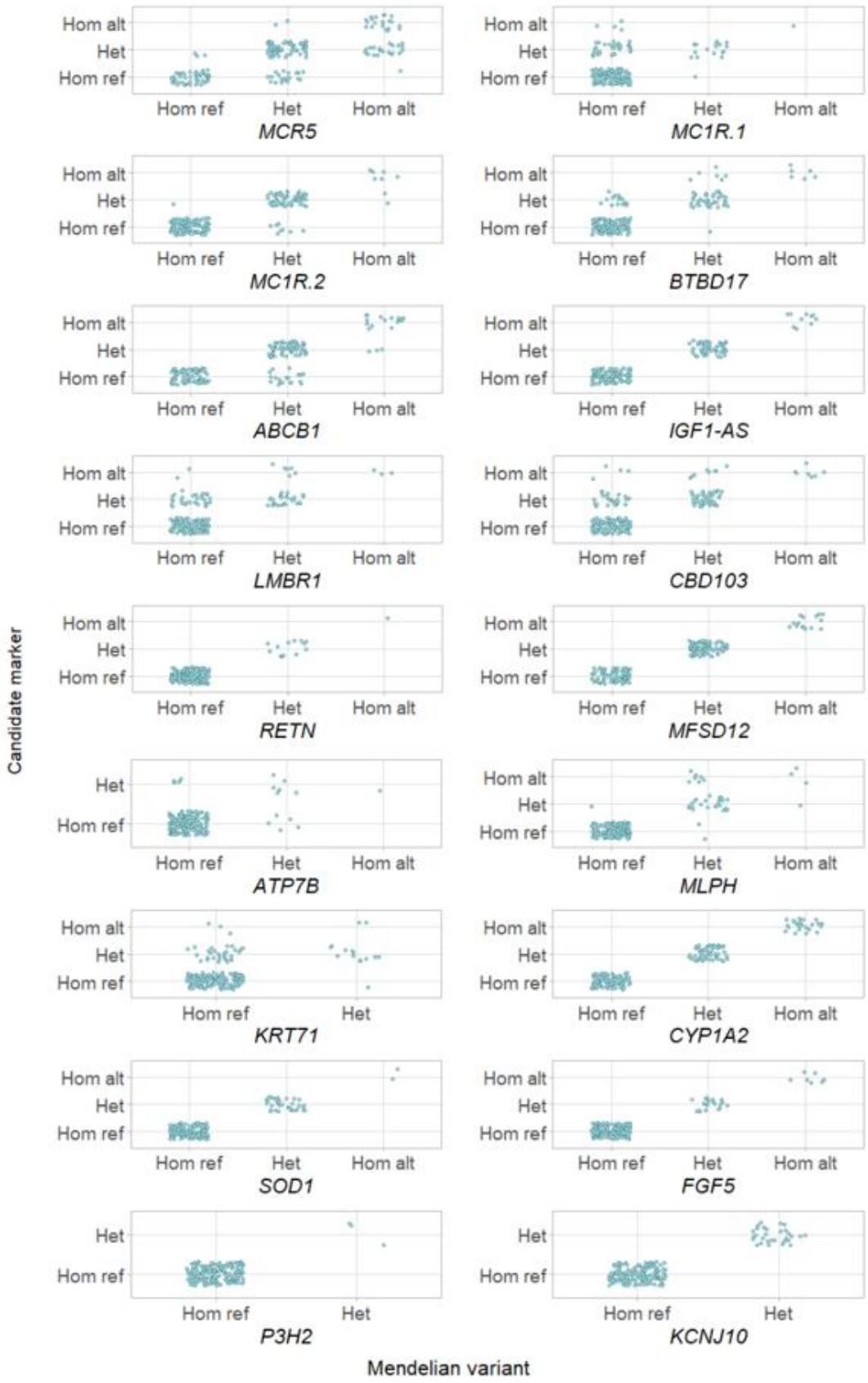

Figure 13. Genotype correlation plots for Mendelian variants and candidate SNP markers. Mendelian variant genotypes are on the x axis (labelled by the associated gene) and candidate marker genotypes are on the y axis (see Table 9). **Note:** *MC1R.1* = OMIA variant 343; *MC1R.2* = OMIA variant 34; hom = homozygous; het = heterozygous; ref = reference; alt = alternative.

3.9 Imputation of Missing Genotypes in the SNP Chip Sample

Referencing the 250-sample WGS variant dataset, missing genotypes were imputed across all variants (n = 20,763,647) in the 111-sample SNP chip-only variant dataset. It should be noted that the number of variants in the reference panel is greater than the number of variants in the working WGS dataset because this count separates multiallelic sites into multiple biallelic variants at the same position. The mean DR^2 value, which reflects the accuracy of imputed genotypes, across all variants was 0.69 for the first imputation run (see Table 10). In this run, no N_e was specified and the working WGS dataset was referenced. The mean DR^2 increased to 0.76 when the same panel was referenced, but an N_e of 266 was specified (imputation run 2). As expected, the highest average accuracies were observed when genotypes were imputed from more stringently filtered reference panels. Imputation run 3 referenced a panel that was filtered for missing call/genotype rate < 0.1 and yielded a mean DR^2 of 0.84, while imputation run 4 referenced a panel that was filtered for missing call/genotype rate < 0.1 and MAF > 1% and yielded a mean DR^2 of 0.93. Table 10 clearly shows that, while the more stringent reference panels yielded higher average accuracies, they imputed fewer genotypes, meaning some variants of interest, such as rare variants, were excluded. Imputation run 2 yielded the highest number of imputed genotypes with DR^2 values equal to 1 and greater than 0.8.

Table 10: Summary of genotype imputation accuracy

	DR^2 summary statistics				Reference	Genotypes	Genotypes
Run	Q1	Median	Q3	Mean	panel size [e]	with $DR^2 = 1$	with $DR^2 > 0.8$
1 [a]	0.36	0.91	0.98	0.69	20,763,647	3,087,735	12,814,720
2 [b]	0.68	0.95	0.99	0.76	20,763,647	4,121,737	14,515,582
3 [c]	0.90	0.98	1	0.84	14,686,056	4,087,680	11,843,441
4 [d]	0.95	0.98	1	0.93	11,910,373	3,944,783	10,761,203

[a] *Referenced the working WGS dataset and did not have a specified N_e.*

[b] *Referenced the working WGS dataset and had a specified N_e of 266.*

[c] *Referenced the call rate-filtered WGS dataset and had a specified N_e of 266.*

[d] *Referenced the call rate- and MAF-filtered WGS dataset and had a specified N_e of 266.*

[e] *Number of variants in the reference panel, where multiallelic variants were counted as multiple biallelic variants.*

A major aim of the imputation analysis was to determine whether the genotypes at Mendelian loci of interest (see Table 6) could be predicted from SNP chip genotypes. In particular, the aim was to predict genotypes at Mendelian loci that did not have strong candidate single SNP markers (see Section 3.8). Table 11 shows the DR^2 value at each Mendelian locus for the latter three imputation runs. All 27 Mendelian variants were imputed in runs 2 and 3, however, only 25 were imputed in run 4 due to the removal of variants with MAF < 1% from the reference panel. With the exception of loci that were on the SNP chip (such as the *P3H2* variant), sites with low allele frequencies were imputed with lower confidence than sites with high allele frequencies. The genotypes of 17 Mendelian variants were able to be imputed with $DR^2 > 0.9$ in at least one imputation run, and 11 of these were not able to be predicted using R^2 with single SNP markers (highlighted in blue in Table 11). The mean DR^2 values from the latter three imputation runs across Mendelian variants were 0.75, 0.76, and 0.80 respectively. However, according to a one-way repeated measure ANOVA test, there was no evidence of a significant difference between these means ($p = 0.282$).

Table 11 compares the allele frequencies for variants of interest between the WGS sample and the imputed sample. Allele frequencies tended to be consistent across imputation runs when the accuracy was high, but variation between runs was observed for variants with less confidence (e.g. the *CUBN* variant). The imputed allele frequencies of some high-accuracy variants, such as the *MFSD12* variant, differed substantially from the WGS allele frequencies. This could indicate that there is sample variability between the WGS animals and the SNP chip-only animals.

Table 11: Genotype imputation for Mendelian variants of interest

		WGS	Run 2 [a]		Run 3 [b]		Run 4 [c]	
OMIA ID	**Gene**	**AF**	**AF**	**DR^2**	**AF**	**DR^2**	**AF**	**DR^2**
1603	*MC5R*	0.568	0.513	0.96	0.523	0.97	0.509	0.97
447	*CUBN*	0.012	0.002	0.22	0.008	0.61	0.001	0.03
343	*MC1R*	0.06	0.160	0.97	0.162	0.99	0.162	0.97
34	*MC1R*	0.8	0.771	0.98	0.777	0.98	0.776	0.99
577	*SGSH*	0.01	0.001	0.10	0.009	0.11	0.012	0.83
851	*BTBD17*	0.162	0.267	0.98	0.264	0.98	0.266	0.99
1273	*CNTNAP1*	0.002	0.015	0.53	0.0001	0.01	-	-
31	*TYRP1*	0.026	0.059	1.00	0.055	0.94	0.058	1.00
267	*TYRP1*	0.028	0.067	0.98	0.071	0.95	0.070	0.93
796	*TYRP1*	0.02	0.049	0.98	0.044	0.92	0.046	0.98
442	*ABCB1*	0.33	0.375	0.98	0.390	0.97	0.389	0.97
1422	*IGF1-AS*	0.186	0.342	1.00	0.342	1.00	0.342	1.00
1444	*LMBR1*	0.108	0.152	0.96	0.146	0.99	0.158	0.97
458	*CBD103*	0.184	0.175	0.34	0.171	0.40	0.116	0.49
1522	*RETN*	0.046	0.014	1.00	0.014	1.00	0.014	0.99
1081	*MFSD12*	0.32	0.284	1.00	0.284	1.00	0.284	1.00
106	*ATP7B*	0.038	0.041	0.88	0.054	0.95	0.054	1.00
30	*ASIP*	0.012	0.009	0.92	0.026	0.97	0.028	0.95
360	*MLPH*	0.108	0.000	0.00	0.000	0.00	0.000	0.00
401	*VWF*	0.022	0.023	0.42	0.057	0.32	0.054	0.36
35	*KRT71*	0.032	0.022	0.37	0.026	0.31	0.030	0.32
274	*CYP1A2*	0.302	0.149	1.00	0.149	1.00	0.149	1.00
36	*SOD1*	0.126	0.069	0.87	0.059	0.96	0.059	0.96
48	*FGF5*	0.096	0.048	0.41	0.181	0.59	0.098	0.52
103	*P3H2*	0.006	0.014	1.00	0.014	1.00	-	-
338	*CLN8*	0.042	0.022	0.48	0.034	0.79	0.034	0.87
612	*KCNJ10*	0.918	0.925	0.89	0.921	0.90	0.932	0.86
Mean DR^2:				**0.75**		**0.76**		**0.80**

Note: *All runs (2, 3, and 4) specified and N_e of 266 but no genetic map; variants that did not have strong single SNP candidate markers but were able to be imputed with $DR^2 > 0.9$ are highlighted in blue; AF = allele frequency; Allele frequencies refer to the alternative allele based on canFam4.*

[a] *Referenced the working WGS dataset.*

[b] *Referenced the genotype rate-filtered WGS dataset.*

[c] *Referenced the genotype rate- and MAF-filtered WGS dataset.*

3.10 Genome-Wide Association Studies for Four Morphological Traits

Genome-wide association studies were performed using GCTA's MLMA-LOCO method with the aim of identifying genetic markers associated with four morphological traits of interest: height, length, chest circumference, and muzzle circumference. Figures 14-17 display Manhattan plots of the Bonferroni-corrected $-\log_{10}$ p-values from each SNP tested across the genome. Prior to this, simple GLM GWAS were also performed with PLINK and the Manhattan plots from these tests are displayed in Appendix B.7.

Excluding outliers, height ranged from 450 to 740mm, with a mean of 583mm (see Figure 6). There were two SNPs (see Table 12) that were significantly associated with height according to the MLMA, one in chromosome 3 (g.91717919G>A) and one in chromosome 5 (g.11544449C>T) (see Figure 14). The former was 20kb upstream of *LCORL* (ligand dependent nuclear receptor corepressor like), a known body size QTL in dogs and cattle (5,134). The estimated effect size (β) of the SNP was 30.08mm, meaning that for each copy of the alternative allele, height is estimated to increase by 30.08mm. The height GLM yielded a strong peak in this region, meaning several SNPs in LD were significantly associated with the phenotype (see Figure A6). However, the MLMA GWAS yielded only a single SNP associated with height in this region. Another SNP in chromosome 5 was significantly associated with height according to the MLMA GWAS. This was located in an exon of *UBASH3B* and was within 100kb of the *CRTAM* (cytotoxic and regulatory T cell molecule) and *JHY* (junctional cadherin complex regulator) genes. Like the marker on chromosome 3, this was the only SNP in the region that was significantly associated with height and it was not within a strong peak.

More variation in length than height was observed, with length ranging from 400 to 870mm (mean = 633mm). No SNPs were significantly associated with length according to the MLMA (see Figure 15). According to the GLM, single SNPs on chromosomes 1, 28, and 31 showed evidence of a significant association with length (see Figure A7). The mean chest circumference was 700mm, with a wide range from 540 to 880mm. Muzzle circumference ranged from 150 to 340mm and had a mean of 244mm. A single SNP (see Table 12) in chromosome 32 (g.9992943A>G) was found to be significantly associated with both chest and muzzle circumference (see Figures 16 and 17). This SNP also

approached the significance line in the height MLMA (see Figure 14) and was in the last exon of the *CASP6* (caspase 6) gene. It was within 100kb of *MCUB* (mitochondrial calcium uniporter dominant negative beta subunit), *PLA2G12A* (phospholipase A2 group XIIA), *CFI* (complement factor I), *GAR1* (GAR1 ribonucleoprotein), and *RRH* (retinal pigment epithelium-derived rhodopsin homolog). According to the GLM, several SNPs in LD in this region were significantly associated with chest and muzzle circumference; however, these models yielded much stronger peaks in chromosome 20. The lead (most significantly associated) SNPs of these peaks were near a well-known coat colour gene, *MITF* (melanocyte inducing transcription factor) (see Figures A8 and A9) (135).

Table 12: Summary of SNPs significantly associated with body size

Array ID	Position	Genes [a]	AF [b]	Phenotype	β	SE	p-value
AX-168204191	Chr3: 91717919	*LCORL*	0.22	Height	30.08	5.39	2.4×10^{-8}
AX-167320198	Chr5: 11544449	*UBASH3B*, *CRTAM*, *JHY*	0.14	Height	34.63	5.86	3.4×10^{-9}
AX-167270920	Chr32: 9992943	*CASP6*, *MCUB*, *PLA2G12A*, *CFI*, *GAR1*, *RRH*	0.14	Chest circumference	34.71	5.79	2.0×10^{-9}
				Muzzle circumference	16.13	2.74	3.8×10^{-9}

SNPs that were significantly associated with one of the four morphological phenotypes of interest according to an MLMA-LOCO GWAS.

Note: *AF = allele frequency; β = Estimated effect size of the genotype on the phenotype; SE = standard error.*

[a] *Genes within 100kb of the significant marker.*

[b] *Allele frequency in the GWAS sample (n = 430).*

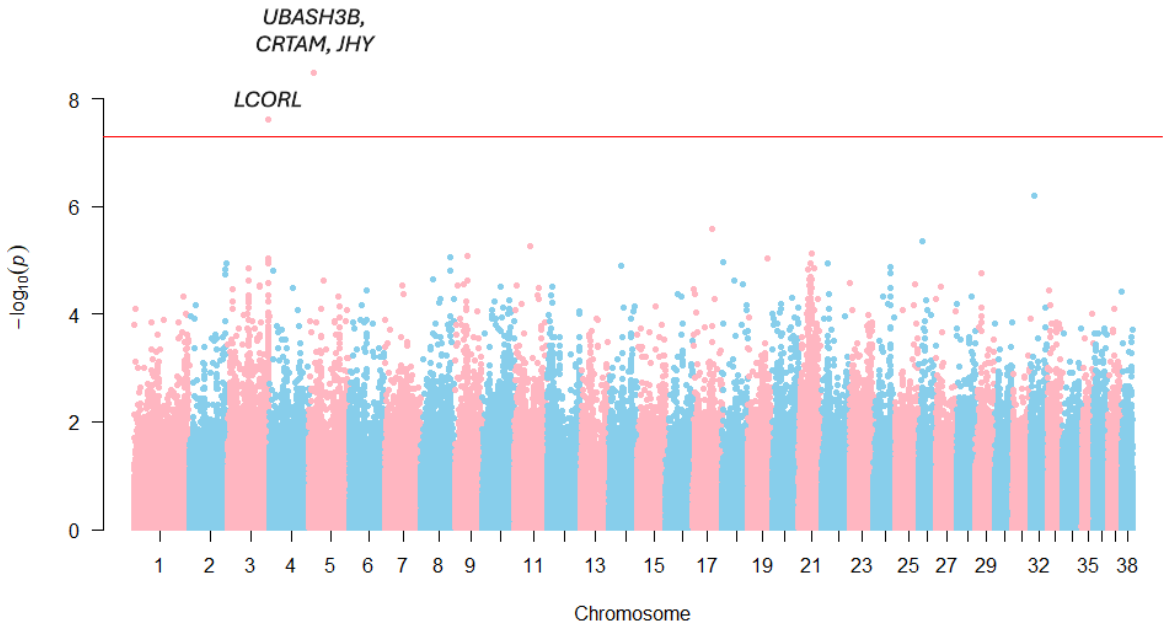

Figure 14. Manhattan plot from a height MLMA-LOCO GWAS. There was one significant SNP near the *LCORL* gene and a second significant SNP near the *UBASH3B*, *CRTAM*, and *JHY* genes. **Note:** The red line indicates the genome-wide significance threshold ($-\log_{10}(p) = -\log_{10}(5\times10^{-8})$).

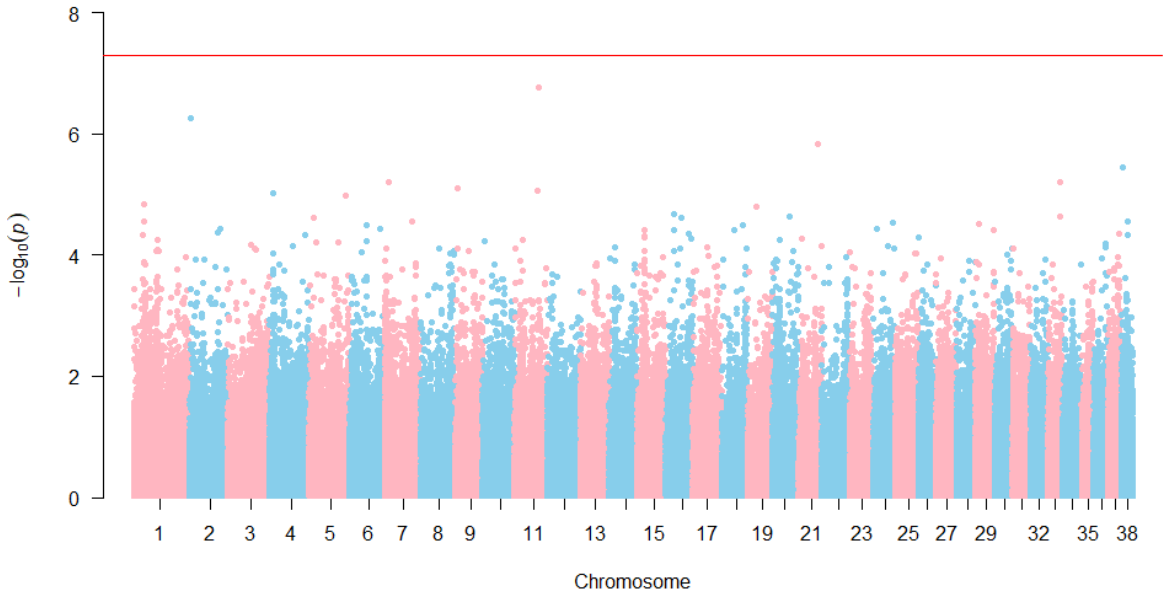

Figure 15. Manhattan plot from a length MLMA-LOCO GWAS. There were no significant SNPs. **Note:** The red line indicates the genome-wide significance threshold ($-\log_{10}(p) = -\log_{10}(5\times10^{-8})$).

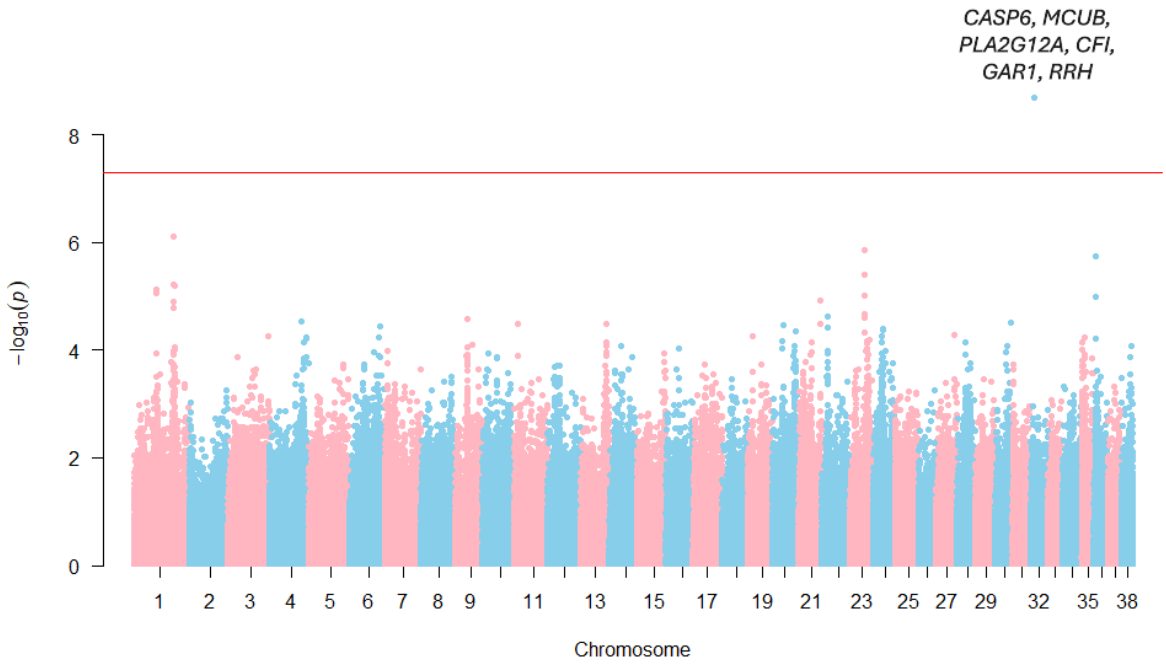

Figure 16. Manhattan plot from a chest circumference MLMA-LOCO GWAS. There was one significant SNP near the *CASP6*, *MCUB*, *PLA2G12A*, *CFI*, *GAR1*, and *RRH* genes. **Note:** The red line indicates the genome-wide significance threshold ($-\log_{10}(p) = -\log_{10}(5\times10^{-8})$).

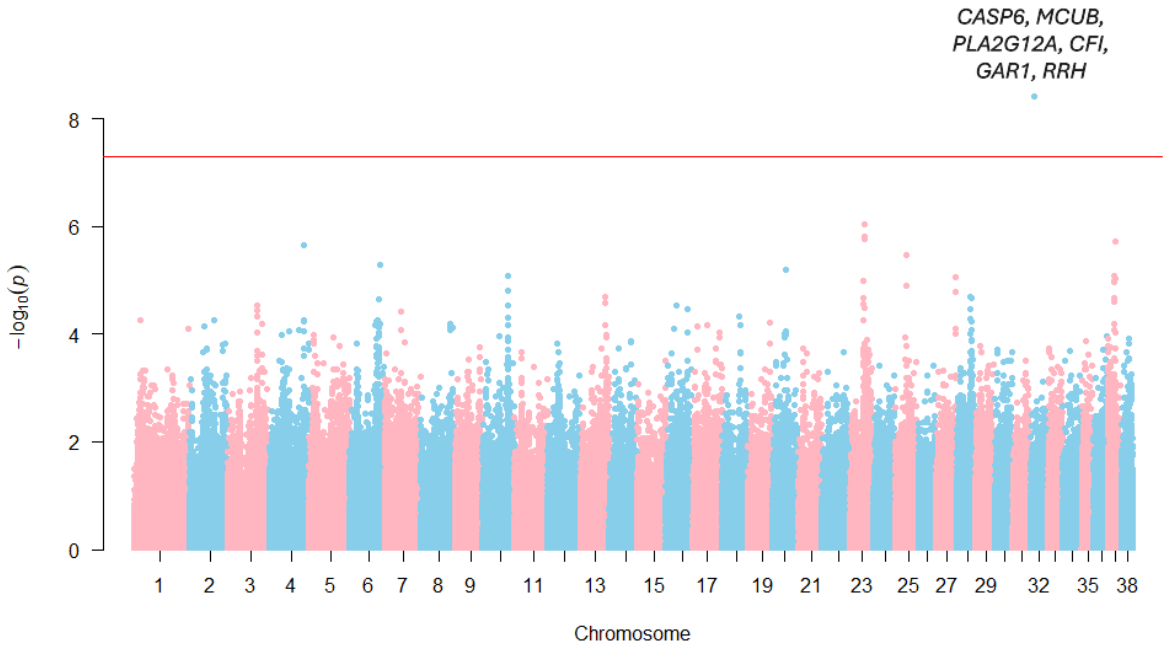

Figure 17. Manhattan plot from a muzzle circumference MLMA-LOCO GWAS. There was one significant SNP near the *CASP6*, *MCUB*, *PLA2G12A*, *CFI*, *GAR1*, and *RRH* genes. **Note:** The red line indicates the genome-wide significance threshold ($-\log_{10}(p) = -\log_{10}(5\times10^{-8})$).

Chapter 4: Discussion

4.1 Summary of Main Findings

This thesis describes the generation and first use of an extensive dataset containing millions of high-quality variant calls from 250 New Zealand farm dogs' whole genome sequences. The genetic variation in the population was characterised through sequence alignment and variant calling, filtering, and effect prediction. Additionally, variant effect prediction and RNA-Seq expression data was used to prioritize the dataset for identifying causal variants, generating a list of 37,461 functional predictions. This pragmatic classification of functional variants omitted those in non-coding regions, acknowledging that thousands of regulatory variants with functional effects were therefore excluded. The OMIA public database was leveraged to address the second aim of the research, where 27 previously reported functional variants and nine novel functional candidates were identified as segregating in the population.

In addition to using WGS data, this project leveraged SNP chip genotypes from 299 NZ working dogs to achieve three primary goals. First, single SNP predictive markers were identified on the array for eight functional variants, providing an opportunity for marker-assisted selection based on SNP chip genotypes. Second, missing genotypes in the SNP chip-only sample were accurately imputed across almost all missing genotypes by referencing the WGS dataset. This enabled genotype prediction for 11 additional Mendelian variants. Third, GWAS were performed for four body size phenotypes, revealing three significant associations. Among these was a known body size gene, providing strong evidence that it contributes to size variability in the population.

4.2 Implications of the WGS Dataset for Future Research

The genomic resources produced in this study will be used for future research within and beyond the Right Dog project to gain a greater understanding of the New Zealand Huntaway and Heading Dog breeds. The whole genomes from 250 dogs were aligned to the canFam4 reference genome, yielding an average read depth of more than 22× and an

average read quality of almost 36. Jiang et al. (136) suggested that a read depth of at least 10× is required to accurately detect variants, while Kishikawa et al. (137) indicated that a read depth of 15× can achieve accurate SNP genotyping. Hon et al. (138) sequenced the genomes of five organisms and yielded average read qualities between 28 and 35, all of which were considered to be highly accurate. Therefore, the mapped reads produced here are of high quality and will be extremely useful. For example, they could be combined with existing datasets to perform phylogenetic and comparative genomic analyses to determine the relationship between Huntaways, Heading Dogs, and other dog breeds. Furthermore, this WGS data provides a basis for genomic selection, and future association studies will utilise it to fine-map causal variants for health- and performance-related traits.

Using whole genome sequences, VEPs, and expression data, an extensive dataset of 20 million functionally annotated variants was generated. These annotations were used in the current thesis to identify candidate functional variants (see Section 3.5) and will be an extremely useful tool to prioritise causal variants in future studies. For example, they could be used to develop a breed-specific SNP chip that captures the most functionally relevant loci in Huntaways and Heading Dogs. Given that the largest known WGS database available for dogs comprises ~2000 genomes, the 250 generated here make up a sizable contribution that will help create a better understanding of dog genetics (101).

4.2.1 Challenges in the Development of the WGS Dataset

One of the challenges in large-scale sequencing studies is the trade-off between sensitivity and specificity during quality control and filtering steps (58). In the current analysis, sensitivity was favoured over specificity, meaning quality filters were lenient and even the most strictly filtered datasets produced will contain false positives. Because of this, annotations and predictions cannot be deterministic of causation, and most of the 'functional' variants identified here, other than those that have been previously proven as causal, require further testing. While the final list of 'functional'

predictions will be extremely useful for prioritisation and providing additional evidence of causation, it should still be applied with this caveat.

Where possible, GATK recommends VQSR filtering for small SNPs and indels, since it has outperformed hard filtering in the differentiation between true and false calls (57). However, VQSR requires multiple high confidence datasets of species-specific variant calls, which are not available for dogs. It is possible to create such datasets; for example, the Dog10K Consortium applied strict filters to commercial array genotypes and trained a VQSR model on this dataset (101). However, this can lead to circularity bias if the training data is built from the working dataset. Additionally, self-generated truth sets are unvalidated, creating uncertainty. While VQSR leads to a lower false discovery rate, it can be less sensitive than hard filters, meaning it creates a greater risk of removing true calls (57). For these reasons, the current study did not apply VQSR filtering, instead applying hard variant filters and favouring sensitivity.

4.2.1.1 Systematic Difference in Quality Between Sequencing Plates

Although it was not discussed in the results section of this thesis, an analysis of the proportion of reads that were properly mapped and paired revealed a potential difference in the quality between sequencing plates (see Appendix B.1). Shown in Figure A3a, the proportion of properly mapped and paired reads was distributed bimodally amongst unfiltered reads, with peaks at 0.8 and 0.84. Conversely, the proportion of filtered reads that were properly mapped and paired more closely resembled a normal distribution, with one distinct peak at 0.997 (see Figure A3b). Initially, it was hypothesised that the bimodal distribution could be explained by breed, since approximately half of the samples were from Huntaways and half were from Heading Dogs. This hypothesis assumes that one breed is more genetically similar to the reference than the other, meaning a greater proportion of reads would map correctly. Talenti et al. (139) demonstrated that breed is an important factor in determining mapping quality, by incorporating breed diversity into the cattle reference genome to increase mapping rates.

However, Huntaways and Heading Dogs were evenly distributed between the two peaks, meaning the difference was not due to breed.

Next, it was hypothesised that a difference between sequencing plates may explain the bimodal distribution, since sequences were processed on five distinct plates and this could lead to a batch effect (140). Figure A4a shows the proportion raw reads that were properly mapped and paired, coloured by sequencing plate. This distribution clearly shows that plates 1 and 3 had considerably lower proportions than plates 2, 4, and 5. The peak at 0.8 comprised the majority of samples from plates 1 and 3, while the peak at 0.84 comprised the majority of samples from plates 2, 4, and 5. As seen in Figure A4b, the difference between these plates was not observed amongst filtered reads, meaning filters successfully corrected for the variation. This suggests that there may have been a systematic difference in the quality between sequencing plates (i.e. a batch effect), but the final dataset was unaffected by this bias as a result of quality control. It should be noted that no statistical tests were performed to determine whether the difference between plates was significant and even before filtering, the difference was small.

4.3 Functional Annotation of WGS Variants

4.3.1 Comparison of Predicted Variant Effects in Different Datasets

All variants in the WGS dataset were annotated with VEPs to aid in their interpretation and to generate a list of functional predictions. This prioritisation method excludes functional variants in non-coding regions since their effects tend to be less accurately predicted by computational VEPs (118). Because VEPs use alternative databases and therefore can lead to conflicting conclusions, both Ensembl VEP and SnpEff were used to functionally annotate variants (141). By only retaining the variants that were predicted by both tools to be functional (i.e. high and moderate impacts), the list was reduced from approximately 58,500 to approximately 40,000 variants. However, only SnpEff annotations were considered in the comparison of variant effects between breeds. This was so the findings would be comparable to those from Wang et al. (142), Plassais et al. (5), and Jagannathan et al. (105), who all used SnpEff to classify variants.

The proportion of SnpEff-predicted effects and impacts was calculated across all samples. As expected, the vast majority (>99%) of variants were within non-coding regions, and of variants in coding regions, less than 40% were predicted to be 'functional' (i.e. disrupt the sequence of the translated protein). Three percent of the canFam4 reference genome is annotated to be protein-coding and theoretically, though not always in reality, mutations are equally likely to occur at all positions in the genome (22,30,31). This would mean that approximately 3% of variants would be expected to map within coding regions, rather than the 0.8% that was observed. However, variants in coding regions, especially those that are functional, are more likely to be removed from the population by purifying selection than those in non-coding regions, since they are more likely to be deleterious (22). Therefore, coding regions are more conserved than non-coding regions and proportionally fewer variants are observed in these regions.

In the WGS dataset, 0.35% of variants were predicted to be high or moderate impact by SnpEff (see Table 4). The DBVDC reported that an average of 2% of variants per sample were predicted to be in coding regions and 1% were predicted to be high or moderate impact (105). The lower proportion of coding-region variants observed in the current study could be due to alternative definitions of coding regions. For example, the current study excluded UTRs from the count, but the DBVDC may have included them. Additionally, the DBVDC referenced the canFam3.1 assembly, rather than the canFam4 assembly, meaning results may not be directly comparable to the current study. Other dog studies have also reported greater proportions of functional predictions than that reported here. Upon the release of the canFam4 assembly, Wang et al. (142) analysed 22 million variants from 19 dog breeds and estimated 1.4% of variants to be functional (i.e. directly influence gene products). The lower proportion of functional predictions observed here may reflect the use of more stringent quality filters, the smaller number of breeds and samples considered, or a true underrepresentation of functional variants in the current sample population. It would be interesting to test the latter hypothesis in future studies.

In the current study, 40% of variants were intronic and 55% were in intergenic, intragenic, or non-coding regions. A previous study of 722 canid genomes by Plassais et al. (5) found that 33% of variants were intronic and 53% were intergenic according to SnpEff

annotations (5). Similarly, Megquier et al. (109) annotated ~8 million canine variants with SnpEff and classified 54% as intergenic and 32% as intronic. The slightly higher proportion of intronic variants observed here could suggest that different criteria were used to define the genetic features in Plassais et al. (5) and Megquier et al. (109). In Plassais et al. (5), 39% of coding-region variants were nonsynonymous and 7% were high impact. Similar proportions were observed here, with 43% of coding-region variants annotated as nonsynonymous and 8% as high impact. The (assumedly) improved annotation of the canFam4 assembly may account for the small differences observed, but further investigation is required to make this conclusion.

4.3.2 Comparison of Predicted Variant Effects in Huntaways and Heading Dogs

Variant effect prediction was used to determine whether variants differed between Huntaways and Heading Dogs. Based on this comparison, there was little evidence of a difference in the types of variants segregating in the two breeds (see Figures 8 and 9). This is not surprising, since approximately 90% of variants were observed in both breeds, and so most of the effect annotations in the two datasets overlapped. Because of this large overlap, the genetic dataset could be analysed in most of the following analyses as a single sample, increasing statistical power.

To gain a better understanding of the difference between the breeds, the proportion of predicted effects should be compared between the variants that were unique to each breed. Additionally, a gene enrichment analysis could determine which genetic pathways are over- or underrepresented and a comparison of variant classes (like inversions and transitions) could reveal a difference in the types of variants observed in each breed. These approaches, along with more sophisticated methods like PCA, would provide insight into the genetic and evolutionary relationship between Huntaways, Heading Dogs, and other dog breeds.

Because only a small number of dogs from each breed were included in the comparison analysis (n = 104 and 129), strong conclusions about the genetic distance between the

two breeds cannot be made. This is also true because, although dogs were sampled to represent multiple geographic regions, the sample was not truly random. Multiple dogs were sampled from the same owner and farm for convenience, meaning the true sampling unit was not the individual dog, making the sample size even smaller than that mentioned. Additionally, the large overlap of variants observed between the two breeds meant the assumption of independence was violated in the chi-squared test, and its results should be interpreted with caution.

4.3.3 The Use of RNA-Seq Data to Improve Functional Annotation

The list of predicted functional variants was reduced from 40,067 to 37,461 using publicly available expression data. This equates to a 6.5% reduction and meant that the final list of predicted functional variants represented just 0.19% of the total variants called. RNA-Seq data reveals which genes are expressed in certain tissues and therefore which genes may be falsely annotated. It is not uncommon for reference genomes in non-model organisms to contain false annotations since they often utilise annotations from more well-studied organisms and therefore lack breed- and species-specific features (143). This study demonstrated how RNA-Seq data can be used to generate a more manageable dataset, while retaining the most functionally relevant loci. In addition to refining the list of predicted functional variants, RNA-Seq data was used to verify expression at candidate functional loci, providing further evidence of causality (see Sections 4.4 and 4.5).

A lenient but relatively arbitrary threshold was applied to remove variants in falsely annotated genes. This threshold was selected to remove as many false positives as possible without removing many true positives (i.e. genes expressed at a low levels). It would have been too lenient to only remove loci with RNA-Seq read depths of 0, since mapping errors, sequencing errors, and the low expression of some non-coding regions mean that low numbers of RNA-Seq reads often map to non-coding regions (144). The threshold of 5 was especially lenient because the BAM file used to count depth included 17 RNA-Seq files. A more conservative threshold could be applied to filter for highly expressed genes in cases where they may of most interest. For example, if a threshold of

10 had been applied to the current dataset, ~11% of high- and moderate-impact variants would have been removed, yielding 35,777 'functional' predictions. Alternatively, RNA-Seq data could have been used to expand the list of functional predictions by identifying novel transcripts. This would be especially beneficial if the canFam4 reference assembly was found to exclude a large number of genes, or particular genes of interest.

4.3.3.1 Limitations of Using RNA-Seq Data

The RNA-Seq analysis relied on the assumption that all coding genes are expressed in at least one of the 17 tissue samples obtained. This is an unrealistic but necessary assumption, since some genes are highly tissue specific (i.e. only expressed in one or two tissues) and some genes are only expressed at certain stages of development or in response to particular stimuli. In reality, expression data from all cell types would be required to ensure all genes are captured; however, it was not feasible to obtain this amount of data, therefore, some real genes are certain to have been excluded (145). This is similar to the way VEPs fail to identify functional variants in genes that are not annotated in the reference genome. Because differential gene expression plays an important role in evolution, it is possible that the transcriptome of the sample population was misrepresented by the use of RNA-Seq data from non-target breeds (146). However, studies that profiled the dog transcriptome using RNA-Seq concluded that, while gene expression is highly variable between species and tissue types, there was no evidence of variation between individuals in the same species, nor any correlation with breed, size, or age of dogs (99,109). Therefore, the use of RNA-Seq samples from different breeds, while potentially suboptimal, should not be detrimental to the efficacy of this analysis.

4.4 Previously Reported Mendelian Variants Segregating in NZ Farm Dogs

The WGS sample was queried for 395 Mendelian variants reported by the OMIA to be likely causal in dogs. This revealed 27 variants segregating in the NZ working dog population (see Tables 6 and 7). Given that dogs were selected to represent multiple

geographic regions, the 368 non-segregating variants (see Table A7) are likely absent from Huntaways and/or Heading Dogs, or present at low enough frequencies that no carriers were sampled. Of the 27 segregating variants, nine coat colour variants were identified in six coat colour genes (*MC1R*, *TYRP1*, *ASIP*, *MLPH*, *CBD103*, and *MFSD12*). Another five variants associated with other coat-related or morphological traits were identified in *MC5R*, *IGF1-AS*, *LMBR1*, *KRT71*, and *FGF5*. The 13 remaining variants had previously been implicated in disease, highlighting potential diseases of concern in the current population. The OMIA database catalogues variants with differing levels of evidence of causality. Therefore, the literature was reviewed to assess the functionality of each variant and the potential use of each disease variant for genetic testing and/or selection in the population. A summary of this review follows.

4.4.1 Functional Assessment of Mendelian Variants Associated with Coat Colour

4.4.1.1 MC1R:g.64186728G>A (p.Arg306*) and *MC1R*:g.64186854C>T (p.Met264Val)

MC1R (melanocortin 1 receptor) is a well-known coat colour gene in dogs and several other species (147). It is involved in regulating the production of melanin pigment and causes a black/brown phenotype when active. Two known functional variants within this gene were observed in the WGS sample. One was a G > A SNP discovered by Newton et al. (148) to truncate the gene and cause a red/yellow phenotype (OMIA ID = 343). They sequenced the candidate gene in 194 dogs and observed complete concordance between the A genotype and red coats. Now known as the e^1 locus, the role of this nonsense variant has been validated by subsequent studies, and it segregates in many breeds (149). In the WGS sample, three homozygotes and 24 heterozygotes were observed (allele frequency = 6%). There is strong evidence that this variant leads to the red/yellow coat in NZ breeds, since it does in several other breeds. Coat colour was recorded as a phenotype in the Right Dog survey (see Appendix A.1), meaning the function of this variant, and all other coat colour variants discussed, can be verified with phenotypic data in future studies. At the time of writing, coat colour information had not

been curated beyond raw image data and paper surveys (i.e. recorded in Helical and categorised for different coat features), precluding that analysis.

The second *MC1R* variant observed in the WGS sample was a C > T missense SNP that was discovered by Schmutz et al. (150) to cause a black melanistic mask (OMIA ID = 34). They sequenced the candidate gene and found an association between the T genotype and the mask phenotype in a sample of 52 dogs. They proposed a dominant mode of inheritance; however, not all dogs that carry the genotype display the mask phenotype, likely due to epistatic interactions with other coat colour genes. Subsequent studies have characterized the variant in other breeds, and it has been termed the E^M locus (149). The T genotype was observed at a high frequency in the WGS sample (allele frequency = 80%) and was almost fixed in the Huntaway sample (allele frequency = 95.3%). While dark pigmentation around the muzzle is not abnormal in the Huntaway breed, a light/tan muzzle with dark colouration around the eyes and ears is more commonly observed. Similarly, Heading Dogs tend to have light masks. In future work, it would be useful to investigate how the E^M locus is associated with a black melanistic mask in the NZ population, and whether epistatic interactions may weaken the association between the proposed causal allele and the mask phenotype.

4.4.1.2 TYRP1:g.33376317T>A (p.Cys41Ser), *TYRP1*:g.33385200C>T (p.Gln331*), and *TYRP1*:g.33385242del3 (p.Pro345del)

Three variants in another well-known coat colour gene, *TYRP1* (tyrosinase related protein 1), segregated in the WGS sample (147). The *TYRP1* gene is commonly known as the brown locus, since it is involved in the synthesis of eumelanin. A missense T > A SNP, known as b^c (OMIA ID = 31), a nonsense C > T SNP, known as b^s (OMIA ID = 267), and a 3bp deletion, known as b^d (OMIA ID = 796) were observed in the current sample (151). The variants were first discovered by Schmutz et al. (151) to cause a brown coat and the locus has been well studied in more recent publications (147). Several other variants within the gene have been reported to cause the brown phenotype and epistatic interactions have been reported between *TYRP1* and *MC1R* (152,153). The allele frequencies of the b^c, b^s, and b^d genotypes in the WGS sample were 2.6%, 2.8%, and 2.0% respectively, with no homozygotes observed. There is strong evidence that these variants affect coat colour in

several breeds, but an association study would be beneficial to assess their effect in the current population.

4.4.1.3 CBD103:g.55468988del3 (p.Gly78del)

Black coat colour is regulated by a Dominant Black (K) locus in dogs, which promotes eumelanin synthesis by inhibiting the antagonism of *MC1R* (154). In 2007, Candille et al. (155) determined that the *CBD103* (canine beta-defensin 103) gene was the K locus based on its ability to bind the *MC1R* receptor. In the same study, they determined that the proposed K^b allele was a 3bp deletion in *CBD103* (OMIA ID = 458). This deletion causes a black coat phenotype when the *MC1R* gene is WT (wild type). The variant has been characterized in several breeds since it was first discovered and an allele frequency of 18.4% was observed in the current sample (147,156,157). Again, an association study should be performed to assess the functionality of the variant.

4.4.1.4 MFSD12:g.56247895C>T (p.Arg51Cys)

Hédan et al. (158) determined that a C > T SNP affecting the 51st amino acid in exon 1 of *MFSD12* (major facilitator superfamily domain containing 12) causes a phaeomelanin dilution and a white/cream coat colour in dogs (OMIA ID = 1081). Sponenberg and Rothschild (159) first hypothesised that some I locus resulted in the pure white phenotype by diluting phaeomelanin when the E locus (*MC1R*) has the e/e genotype. Hédan et al. (158) performed two GWAS to map the position of the proposed I locus and used WGS to identify a SNP upstream of *MFSD12*, which was the most significantly associated variant with the diluted phaeomelanin phenotype. This gene has been associated with skin pigmentation in humans and mice, and *in silico* evidence was used to identify the functional C > T variant, which was associated with the phenotype. However, incomplete penetrance was observed and more recently, Slavney et al. (160) proposed that the I locus is complex, comprising at least five genes.

An IGV inspection of sequence data revealed that the C > T SNP was not within the *MFSD12* gene, but upstream of it. The 51st amino acid of *MFSD12* was in exon 1 and contained a C > T SNP at an allele frequency of 32%. This was assumed to be the variant described by Hédan et al. (158) and was positioned at chr20:56252402 in canFam4. Supporting this hypothesis, the position of the variant was updated in the OMIA database

in 2024 to this position. While the variant in the updated position may affect coat colour, it exhibits incomplete penetrance, and further analysis is required to determine its effect in the current population. It should be noted that the LD and imputation analyses were performed prior to the discovery of the correct position, meaning they were applied to the incorrect variant.

4.4.1.5 ASIP:g.23906214C>T (p.Arg96Cys)

There were six heterozygotes in the WGS sample of a missense C > T SNP in the well-characterised coat colour gene, *ASIP* (agouti signalling protein) (OMIA ID = 30) (147). Kerns et al. (161) discovered this variant in German Shepherds and were the first to characterise the gene in dogs. *ASIP* is an antagonist of the aforementioned *MC1R* gene, and the T allele is associated with a recessive black phenotype. Subsequent studies found that this variant was rare in several breeds, including German Shepherds, Shetland Sheepdogs, and Australian Shepherds (147,162). Instead, several other variants have been associated with the black phenotype in these breeds. A low allele frequency (1.2%) of the T allele was also observed in the current WGS sample. Because it is rare, it would be difficult to assess the functionality of the variant through an association test in the current sample.

4.4.1.6 MLPH:g.48403161G>A (splicing)

A G > A SNP at the 3' splice site of the first exon of *MLPH* (melanophilin) segregated at an allele frequency of 10.8% in the WGS sample (four homozygotes and 36 heterozygotes) (OMIA ID = 360). In coat colour genetics, *MLPH* is known as the dilution (d) locus and multiple variants within the gene have been reported to dilute coat colour in dogs (147). Drögemüller et al. (163) discovered the G > A splicing variant by sequencing *MLPH*, which was a candidate gene because it is implicated in a homologous phenotype in humans and mice. The A allele was predicted *in silico* to reduce splicing affinity and was perfectly associated with the phenotype in seven breeds. Additionally, dogs that were homozygous for the A allele were found to produce only 25% of the *MLPH* mRNA transcript that WT dogs did. In subsequent studies, some breeds, like Chow Chows and Chihuahuas, did not carry the splicing variant, but alternative variants within the gene were found to cause the dilution phenotype in these breeds (164,165). As with other well-characterised coat

colour variants, there is strong evidence that this variant is functional, but an association analysis should be performed to assess this in the target breeds.

4.4.2 Functional Assessment of Mendelian Variants Associated with Coat-Related and Morphological Traits

4.4.2.1 MC5R:g.24541931C>T (p.Ala237Thr)

A missense C > T SNP in the second exon of the *MC5R* (melanocortin 5 receptor) gene segregated in the WGS sample at an allele frequency of 56.8% (OMIA ID = 1603). It was discovered by Hayward et al. (166) to associate with fur length and reduced hair shedding in two independent GWAS. *MC5R* has a role in the production of sebum and the T genotype was predicted *in silico* to damage protein structure. The gene is expressed in the sebaceous gland and an inspection of RNA-Seq data revealed that it was also expressed in skin tissue. While the variant was associated with the phenotype, Hayward et al. (166) acknowledged that further evidence is required to determine whether it is the functional variant. To date, no further studies have investigated the functionality of this variant nor its allele frequency in other populations. Although fur length was not measured quantitatively as part of the Right Dog project, an association test could be performed to determine whether the variant is associated with fur type (e.g. curly or straight, rough or smooth), which was queried in the Right Dog survey. Again, that analysis was not performed here since the coat data had not yet been curated at the time of writing.

4.4.2.2 IGF1-AS:g.41511739C>T (regulatory)

Plassais et al. (167) discovered a T > C SNP in a lncRNA that is antisense to the *IGF1* (insulin-like growth factor 1) gene (OMIA ID = 1422). The *IGF1* gene is estimated to control 15% of the size variation observed between dog breeds and the antisense lncRNA is reported to interact with it (5). The C allele was associated with reduced IGF-2 serum levels and short height in a sample of 1431 dogs from 13 breeds. It was speculated that the variant may work by some regulatory mechanism due to the interaction between *IGF1* and *IGF1-AS*; however, no further studies have investigated this. The allele frequency of the C genotype in the WGS sample was 18.6% (13 homozygotes and 67 heterozygotes),

whereas the allele frequency in Huntaways was 5%. If the variant is functional, or linked to the functional variant, the low frequency of the 'short' allele observed in Huntaways makes sense, since they are a large breed. However, height is a complex trait, meaning it is affected by many environmental and genetic factors, and further analysis is needed to characterise the effect of the variant in the NZ population. A follow-up analysis revealed that this variant was removed from OMIA's list of likely causal variants, possibly due to the characterisation of the height phenotype as multifactorial, making it inappropriate for a database that catalogues Mendelian variants. This finding shows how public databases are constantly revised over time, which can alter the findings of analyses that use them.

4.4.2.3 LMBR1:g.20112737C>T (regulatory)

Two C > T SNPs (DC-1 and DC-2) located in *LMBR1* (limb development membrane protein 1) are reported by OMIA to cause preaxial polydactyly (PPD), a.k.a dew claws, in dogs (168). Polydactyly is caused by variants in a regulatory sequency of *SHH* (sonic hedgehog signalling molecule) in several species (169). This regulatory sequence, named ZRS (ZPA regulatory sequence), is within the fifth intron of *LMBR1*, where DC-1 and DC-2 are located (170). The DC-1 variant was discovered in 2008 to cause dew claws in several Asian breeds but was absent from Western breeds, including the current sample (168). The DC-2 variant (OMIA ID = 1444) was discovered to cause dew claws in Western breeds. The T allele of the DC-2 variant was observed at a frequency of 10.8% (four homozygotes and 46 heterozygotes) in the current sample. Park et al. (168) showed that the T allele caused a >50% reduction in the enhancer activity of ZRS when introduced into a luciferase reporter construct of the ZRS region, providing good evidence that it is causal. This variant was excluded from the list of candidate disease variants for diagnostic selection, since the presence of dew claws is not considered a disease phenotype. However, because working dogs are extremely active, dew claws may infer increased risk of injury in working dogs. This is because the extra digits could be caught during strenuous activities like jumping over fences. Therefore, this likely causal variant may be of interest for selection. The presence of dew claws was queried in the Right Dog survey (see Appendix A.1), meaning the association between DC-2 and dew claws can and will be tested as part of the future project. Such an analysis will produce more robust results once a larger number of samples has been collected.

4.4.2.4 KRT71:g.44113063G>A (p.Arg151Trp)

There were 16 dogs in the WGS sample that were heterozygous for a missense G > A SNP in the second exon of the *KRT71* (keratin-71) (OMIA ID = 35). *KRT71* encodes an intermediate filament protein that is expressed in hair follicles and is important for the structural integrity of epithelial cells (171). Cadieu et al. (172) determined that the *KRT71* missense SNP was associated with curly coats through multiple independent GWAS, fine-mapping, and sequencing. No further analyses have provided functional or experimental evidence to support this association, but variants within *KRT71* have been reported to cause curly coats in other mammals (173). The SNP was found to be absent from other breeds, such as Retrievers, and an alternative variant within the gene was found to cause some, but not all, curly coat phenotypes in these breeds (171,174). The frequency of the A allele in the current sample was 3.2%. Further investigation is required to assess the functionality of the variant in the population.

4.4.2.5 FGF5:g.35494497C>A (p.Cys95Phe)

Housley et al. (175) sequenced a candidate gene, *FGF5* (fibroblast growth factor 5), to discover a missense C > A SNP in exon 1 that was completely concordant with hair length in at least four dog breeds (OMIA ID = 48). The A allele was fixed in long-haired breeds and absent from short-haired breeds (175). The *FGF5* gene is a negative regulator of hair growth and variants within the gene have been reported to cause abnormally long hair in mice (176). In a GWAS with 80 dog breeds, Cadieu et al. (172) determined that this missense *FGF5* variant was the most significantly associated with hair length, although, alternative variants in *FGF5* have been reported to cause long hair in other breeds (177). The A allele had a frequency of 9.6% in the current sample (seven homozygotes and 34 heterozygotes). While there is strong evidence that the *FGF5* gene is responsible for a long hair phenotype in some breeds, further analysis is required to determine whether this variant is functional in NZ breeds. As stated, hair length was not directly recorded in the Right Dog survey.

4.4.3 Functional Assessment of Mendelian Variants Associated with Disease Traits

4.4.3.1 BTBD17:g.6924623insG (intronic)

A 1bp insertion in an intron of *BTBD17* (BTB domain containing 17) was found by Meyers-Wallen et al. (178) to associated with lethality due to XX DSD in German Shorthaired Pointers (OMIA ID = 851). However, Meadows et al. (101) determined that the insertion was common amongst a population of 1172 dogs from various breeds (allele frequency = 21.78%). Therefore, OMIA reported that the association is likely to be breed-specific and the variant is unlikely to be causal. The current study supports this finding, with several homozygotes and heterozygotes observed (allele frequency = 16.2%). Based on its intronic location and high frequency in the general population, this variant is unlikely to be causal.

4.4.3.2 CNTNAP1:g.20172413C>T (p.Gly937Glu)

Letko et al. (179) discovered a missense C > T SNP in the *CNTNAP1* (contactin associated protein 1) gene that was associated with LPPN (laryngeal paralysis and polyneuropathy) (OMIA ID = 1273). LPPN is a degenerative disease that causes muscle weakness, airway restriction, and usually affects older, larger dogs. *CNTNAP1* has also been associated with neuropathy in humans. Letko et al. (179) performed a GWAS that revealed a significant association between the region and early onset LPPN in large dog breeds, then performed targeted sequencing to identify the causal variant. The T genotype was predicted *in silico* to disrupt a highly conserved sequence, and it was the most significantly associated SNP in the target region. However, not all affected dogs were homozygous for the variant, meaning that even if the T genotype is causal, additional causal variants exist. Furthermore, high allele frequencies were observed in some breeds, ranging from 6.6% is Leonbergers to 46.6% in English Bulldogs, meaning the effect may be breed specific (179). A single Heading Dog heterozygote was observed in the current study (allele frequency = 0.2%). This frequency is similar to that observed in the original study's control sample, 0.4% (179). Further research is needed to determine whether the SNP is causal in NZ breeds, however, the frequency is too low in the current sample to perform a meaningful analysis.

4.4.3.3 ABCB1:g.13720387A>C (regulatory)

An A > C SNP reported by Alves et al. (180) to affect the regulation of *ABCB1* (ATP binding cassette subfamily B member 1) was observed at an allele frequency of 33% in the WGS sample (OMIA ID = 442). Also known as *MDR1* (multidrug resistance protein 1), this gene encodes a P-glycoprotein drug transporter that leads to drug resistance when non-functional (181). A 4bp deletion within the gene had previously been shown to cause drug resistance in dogs by introducing a premature stop codon (182). Therefore, Alves et al. (180) performed an association analysis to identify additional causal variants within *ABCB1* and identified a significant association between drug resistance in Border Collies with epilepsy and the C genotype of the aforementioned SNP. However, subsequent studies have reported high frequencies of the SNP in the general population, and did not validate the association. For example, Gagliardo et al. (183) performed a multibreed case-control study and detected no association (p = 0.69). The high allele frequency observed in the present study provides further evidence that the SNP is not functional. Interestingly, the SNP is annotated upstream of the gene, rather than in the first intron of the gene in the canFam4 reference genome. However, an inspection of RNA-Seq data showed evidence of an exon upstream of the variant, indicating it could be intronic. Due to its high frequency, this variant is unlikely to be functional in Huntaways or Heading Dogs.

4.4.3.4 ATP7B:g.196868G>A (p.Arg1384Gln) and *RETN*:g.52842420C>T (p.Leu7Phe)

Two variants that have been previously implicated in copper toxicosis segregated in the WGS sample. One was a missense G > A SNP in *ATP7B* (ATPase copper transporting beta) (OMIA ID = 106), and the other was a missense C > T SNP in *RETN* (resistin) (OMIA ID = 1522). Copper toxicosis is a genetic disorder caused by the accumulation of hepatic copper and can lead to liver cirrhosis (184).

The allele frequency of the *ATP7B* SNP in the WGS sample was 3.8% (one homozygote and 17 heterozygotes). Fieten et al. (184) discovered this variant by performing a GWAS to identify the region of interest, then genotyping variants in two candidate genes. *ATP7B* was a strong candidate because it has a role in maintaining copper levels and has been implicated in Wilson disease in humans. The A allele was associated with increased

hepatic copper levels and the locus was predicted to explain 8.2% of the phenotype's genetic variation. However, the causality of the variant has been questioned by subsequent findings. Haywood et al. (185) observed a significant association between the SNP and copper toxicosis in Bedlington terriers, but they also found that the variant was widespread in several breeds with no reports of the disease. Langlois et al. (186) also observed a significant association but noted that several controls were homozygous for the disease allele while several cases were not. If this allele is functional, it does not explain all disease cases and other genetic and environmental factors are likely implicated.

The allele frequency of the *RETN* SNP in the WGS sample was 4.6% (one homozygote and 21 heterozygotes). Because previously associated variants only partially explained the phenotypic variation of toxic copper levels, Wu et al. (187) aimed to identifying additional genetic factors contributing to the disease. Therefore, they performed targeted sequencing of candidate modifier genes in 95 Labrador Retrievers. They observed a significant association between the T allele and increased hepatic copper score; however, the biological mechanism by which *RETN* may affect this phenotype is unknown. Additionally, Langlois et al. (186) observed an allele frequency of 13% and no association with hepatic copper level. It is therefore unlikely that this variant is functional.

4.4.3.5 CYP1A2:g.38261635C>T (p.Arg373*)

A C > T nonsense SNP in *CYP1A2* (cytochrome P450, family 1, subfamily A, polypeptide 2) has been associated with the poor metabolism of certain pharmaceuticals (OMIA ID = 274) (188). *CYP1A2* encodes an important metabolising enzyme in the liver and has been implicated in drug metabolism in humans (189). Mise et al. (190) hypothesised that a genetic polymorphism in *CYP1A2* was responsible for the variable metabolism of a novel cognitive enhancer drug named AC-3933. This was based on *in vitro* evidence that poor metabolisers showed reduced AC-3933 hydroxylation due to a *CYP1A2* inhibitor. Additionally, the *CYP1A2* protein was absent from poor metabolisers and prominent in extensive metabolisers. In a follow-up study, Tenmizu et al. (188) discovered the aforementioned C > T nonsense variant that caused the LOF. Other studies have investigated the role of the *CYP1A2* LOF in other metabolism phenotypes with varying results (191–193). The variant is therefore likely to affect the metabolism of some

pharmaceuticals, but the extent of this effect is not well understood. The frequency of the T allele in the current sample was high (30.2%) but similar to that observed by Tenmizu et al. (39%) (188). This suggests that poor metabolisers of AC-3933 exist in the NZ dog population. This is not strictly a disease variant, however, it may be of interest for testing, since homozygotes may be less likely to respond to other pharmaceuticals.

4.4.3.6 P3H2:g.22231216C>G (p.Glu617Gln)

Metzger et al. (194) discovered a missense C > G SNP in exon 13 of *P3H2* (prolyl 3-hydroxylase 2) (a.k.a *LEPREL1* (leucine proline-enriched repeat-like protein 1)) (OMIA ID = 103). Prolyl 3-hydroxylase 2 is an enzyme that modifies collagen and has been implicated in some human diseases. Metzger et al. (194) performed a GWAS to reveal a significant association between a region near *P3H2* and the Norwegian Lundehund breed predisposition for Lundehund syndrome (LS). LS is an inherited gastrointestinal disease that is highly prevalent in the Lundehund breed and causes malabsorption, protein-losing enteropathy, and inflammatory bowel disease. No variants within the associated region were predicted *in silico* to be functional; however, the aforementioned missense SNP in *P3H2*, which was proximal to the region and also associated with LS, was predicted to be damaging (194). Several mutant homozygotes exhibited no symptoms, meaning if the variant is functional, it is not completely penetrant. Melis et al. (195) reported that few individuals in the breed did not carry the G allele, meaning selection would not be feasible. Three heterozygotes (all Heading Dogs) were observed in the WGS sample (G allele frequency = 0.6%). Prior to the current study, the variant had not been reported in any breeds other than the Lundehund. This could indicate that the three carriers are not purebred and carry some Lundehund ancestry, or more likely that the variant is present in multiple breeds. Regardless, further analysis should be conducted to determine the functionality of the variant.

4.4.3.7 KCNJ10:g.22970389del (regulatory)

A 1bp insertion near the end of the third exon of *KCNJ10* (potassium inwardly-rectifying channel, subfamily J, member 10) was proposed to cause ataxia (OMIA ID = 612) (196). Ataxia describes a group of neurodegenerative diseases defined by a lack of coordination that causes uncontrolled movements. *KCNJ10* encodes a protein involved in the

regulation of potassium ion flow across cells and multiple variants in the gene have been implicated in ataxia (197). Gast et al. (196) determined that a previously reported causal variant of ataxia in *KCNJ10* did not explain all cases in a sample of terriers. Therefore, they sequenced the *KCNJ10* exome in affected terriers to identify alternative causal variants. They detected 45 variants, one of which was a 1bp insertion predicted *in silico* to influence the regulation of gene expression. It was therefore proposed that the insertion may be causal in the sample; however, some clinically affected dogs had neither variant. Additionally, no experimental evidence was obtained to support the hypothesis. This same insertion allele is part of the canFam4 reference genome, meaning the variant was called as a 1bp deletion in the current sample at an allele frequency of 91.8%. A total of 41 dogs carried the proposed disease allele, which is an extremely large number considering the severity of the phenotype. The variant is therefore unlikely to cause the proposed effect in the current population.

4.4.4 Functional Assessment of Mendelian Disease Variants that are Strong Candidates for Selection

4.4.4.1 CUBN:g.18932445del (p.Gln2798fs)

A 1bp deletion in the 53rd exon of *CUBN* (cubilin) leads to the early truncation of this gene (OMIA ID = 447) (198). Cubilin forms an essential protein complex for the uptake of cobalamin (a.k.a B12) and the truncation leads to a potentially life-threatening malabsorption of the vitamin. This variant was discovered by Owczarek-Lipska et al. (198). They performed a GWAS in Border Collies with seven cases and seven controls to identify the 'critical interval' and used sequence data to identify the deletion as the causal variant. Further genotyping showed the deletion was perfectly associated with the phenotype in a sample of 200 Border Collies and subsequent studies have validated this finding (199,200). Based on its high molecular impact to a relevant gene and strong association with the phenotype, this is a very likely causal variant of intestinal malabsorption. Additional variants within *CUBN* have been reported to cause intestinal malabsorption in other dog breeds (201,202). Six Huntaways in the present sample were heterozygous for the deletion discovered by Owczarek-Lipska et al. (198) (allele frequency = 2.9%). This

frequency is high compared to that in a cohort of 500 Japanese Border Collies (allele frequency = 1.5%) (203). Given Border Collies are known to be close relatives, it is likely that the variant is also functional in Huntaways, making the variant a strong candidate for future selection.

4.4.4.2 SGSH:g.2406797insA (p.Tyr229fs)

There were five Huntaways in the WGS sample that were heterozygous for a 1bp insertion in the *SGSH* (heparan sulfate sulfamidase) gene (OMIA ID = 577). The SGSH enzyme, which degrades heparan sulfate, is deficient in homozygotes and this leads to an untreatable neurodegenerative disorder called Mucopolysaccharidosis (MPS) IIIA (112). The insertion was discovered in Huntaways in 2002 as one of few previous genetic studies of the breed (112). The onset of MPS in Huntaways begins from 18 months of age, with progressive symptoms that resemble cerebellar disease. Huntaways exhibit a severe form of the disease compared to other breeds due to a complete loss of *SGSH* function. By contrast, Dachshunds exhibit a less severe form of MPS that can be explained by partial LOF of the gene (204). Yogalingam et al. (112) estimated the allele frequency of the insertion to be 3.8% in NZ Huntaways. In the current study, an allele frequency of 1.9% was observed in Huntaways; however, four of the five carriers were associated with a dog colony where the disease was previously studied. Removing these dogs, the observed frequency was low in the current sample (one carrier). However, the severity of the disease in Huntaways suggests that routine genetic testing would be beneficial for early diagnosis and selective mating.

4.4.4.3 VWF:g.7140281C>T (p.Ser2479Ser)

A G > A SNP at the 3' spice site of the 43rd exon of *VWF* (von Willebrand factor) was found to decrease splicing efficiency and activate a cryptic splice site upstream of the exon (OMIA ID = 401) (205,206). *VWF* encodes a multimeric plasma glycoprotein involved in platelet adhesion and aggregation. Deficiencies in this protein lead to various severities of a bleeding disorder called von Willebrand disease (vWD). Type 1 is the least severe form and causes mild to moderate bleeding, Type 2 is more severe and causes moderate to severe bleeding, and Type 3 can lead to no detectable protein. The G > A splicing variant was found to cause a frameshift that led to the early truncation of the gene and has been

reported to associate with Type 1 vWD (205,206). Subsequent studies have validated this associated in several breeds, but have reported incomplete penetrance. For example, Crespi et al. (207) reported that 46% of homozygous mutants, 22% of heterozygotes, and no WT homozygotes showed clinical signs of the disease in a sample of Doberman Pinschers. Similarly, Segert et al. (208) determined that 45.5% of homozygous mutants, 23.2% of heterozygotes, and no WT homozygotes showed clinical signs of the disease in a sample of Kromfohrländers. An allele frequency of 2.2% (one homozygote and nine heterozygotes) was observed in the current sample. Although the variant is known to exhibit incomplete penetrance, there is strong evidence that it infers increased risk of developing Type 1 vWD. Because the physical demands of work mean that injuries are common amongst farm dogs, and vWD increases risk of prolonged bleeding, testing to assess the severity of risk in the NZ population and the avoidance of carrier-carrier matings would be beneficial (7).

4.4.4.4 SOD1:g.27123057G>A (p.Glu40Lys)

Awano et al. (209) identified a missense G > A SNP in the second exon of *SOD1* (superoxide dismutase 1) that was associated with degenerative myelopathy (DM) (OMIA ID = 36). DM is an adult-onset degeneration of the spinal cord that causes paraplegia and is a model for amyotrophic lateral sclerosis (ALS) (209). It is untreatable and affected dogs are often euthanised. Awano et al. (209) performed a GWAS to determine that a region containing three genes, including *SOD1*, was significantly associated with DM. *SOD1* was considered a strong causal candidate, since superoxide dismutase breaks down superoxide radicals and 20% of ALS cases have been attributed to variants within the gene (210). The gene was sequenced in multiple independent samples and the missense G > A SNP was found to be significantly associated with DM in all cohorts, with 96% of cases and 34% of controls being homozygous for the A allele (209). This association has been validated in several subsequent studies, but the high frequency of the A allele in controls indicates incomplete penetrance (211,212). Awano et al. (209) hypothesised that the variant led to the aggregation of the protein and this hypothesis was supported by Draper et al. (212) and Tanaka et al. (213) with *in vitro* evidence, explaining the degenerative nature of the disease. Therefore, the SNP is very likely to infer risk for developing DM but is not fully penetrant. The frequency of the A allele in the

present sample was 12.6% (53 heterozygotes and four homozygotes). The allele frequency in Huntaways (21.7%) was high but comparable to that reported in German Shepherds (22%), while the frequency in Heading Dogs (1.9%) was low compared to other breeds (214). The allele frequency in a cohort of 500 Border Collies, Heading Dogs' closest known relatives was 8% (203). Because the disease allele is commonly observed in controls, it is not recommended to exclude carriers from breeding. However, given the high frequency amongst Huntaways, further testing would be beneficial to assess risk in this breed and carrier-carrier matings should be avoided.

4.4.4.5 CLN8:g.30769171G>A (p.Trp195*)

A nonsense G > A SNP in the third exon of *CLN8* (ceroid-lipofuscinosis, neuronal 8) has been shown to cause neural ceroid lipofuscinosis (NCL) in at least two dog breeds (OMIA ID = 338) (215,216). NCL is a severe neurodegenerative disorder characterised by loss of motor functions, seizures, and blindness (215). This disease has no treatment, and affected dogs are usually euthanised early in life. *CLN8* is one of at least 14 genes known to be implicated in the disease in humans, and other variants within *CLN8* have been shown to cause NCL in dogs (217–219). The nonsense SNP was first reported in an Australian Shepherd/Blue Heeler cross and causes a complete LOF (215). In a follow-up study, Guo et al. (216) sequenced the whole genomes of sibling German Shorthaired Pointers, one of which had NCL and was homozygous mutant, the other was clinically normal and heterozygous. Archived DNA sequences from the parental breeds showed the variant was extremely rare, with allele frequencies of 0%, 0.67%, and 0% in Blue Heelers, Australian Shepherds, and German Shorthaired Pointers respectively (215,216). Three of the four Australian Shepherds that were homozygous for the mutant allele were known to exhibit symptoms of NCL. Like the *CUBN* variant, the allele frequencies in the current study's breeds were considerably higher than in the breeds of first discovery. The frequencies of the A allele were 6.3% and 2.7% in Heading Dogs and Huntaways respectively. In the total WGS sample, 21 heterozygotes and no mutant homozygotes were observed (allele frequency = 4.2%). Although no studies have reported the prevalence of NCL in NZ dogs, cases can be assumed to exist in appreciable numbers, highlighting the *CLN8* variant as an obvious candidate for future testing.

4.4.5 Summary

Of the 13 Mendelian disease variants segregating in the sample population, five are highlighted here as compelling candidates for use as selection diagnostics, representing variants in the *CUBN*, *CLN8*, *SGSH*, *SOD1*, and *VWF* genes. Testing for these variants would benefit the working dog population by allowing early treatment where treatments exist, and/or the avoidance of carrier-carrier matings. In particular, the *CLN8* variant segregated at a high frequency in the current sample compared to the breeds of discovery, indicating Huntaways and Heading Dogs are at a high risk of developing NCL. Similarly, the *CUBN* variant segregated at a high frequency in Huntaways compared to previously studied cohorts, indicating the breed may have a high risk of developing cobalamin malabsorption. The comparatively high frequencies of these variants in NZ breeds could be a result of a genetic bottleneck or some other form of inbreeding, like popular sire phenomenon. The *SOD1* and *VWF* variants are known to exhibit incomplete penetrance, meaning future studies should assess the severity of risk in the current population; however, there is strong evidence that they infer some level of increased risk. Although the variants in *LMBR1* and *CYP1A2* are not technically disease variants, they may be of interest for selection, since the *LMBR1* variant may lead to an increased risk of injury due to dew claws, and the *CYP1A2* variant may reduce dogs' responses to pharmaceuticals. By contrast, there was a lack of evidence to suggest that the 'disease' variants in *BTBD17*, *CNTNAP1*, *ABCB1*, *RETN*, *ATP7B*, *P3H2*, and *KCNJ10* are functional. The current study provided further evidence that the *BTBD17* and *KCNJ10* variants are not functional, since they supposedly cause severe phenotypes but were observed at high frequencies in dogs that are considered generally healthy.

Meadows et al. (101) performed a similar analysis to that conducted here, where the Dog10K Consortium was queried for 337 variants aggregated from the OMIA database. They detected 76 variants in at least one dog in the sample, including 58 disease variants. While a smaller number of variants were queried in their study compared to the present study, the Dog10K Consortium is a much larger sample (n = 1987 individuals). It therefore makes sense that more variants were detected by Meadows et al. (101). Additionally, Donner et al. (220) showed that the OMIA database exhibits ascertainment bias, since a greater proportion of causal variants are observed in popular breeds (like many of the

breeds in the Dog10K sample). Unsurprisingly, several variants detected in the current study were also detected in the Dog10K dataset. These included variants in *MC1R*, *BTBD17*, *CNTNAP1*, *TYRP1*, *IGF1-AS*, and *CBD103*. The frequency of the *BTBD17* disease allele was also high in the Dog10K Consortium (22%), indicating is does not cause a lethal phenotype. While a small minority of the variants detected here may have been added to the OMIA database after Meadows et al. (101) performed this analysis, the fact that only six variants were detected in both studies suggests that the current sample may be genetically distant from the breeds analysed by Meadows et al. (101). The WGS dataset produced here could therefore contain several unique variants of interest that are not described in the Dog10K Consortium.

4.4.6 Limitations of the Mendelian Variant Survey

While OMIA (106) describes variants that are primarily Mendelian in effect, the inheritance patterns of many variants in the database deviate from Mendel's laws. Because of this, caution needs to be taken when diagnosis and breeding decisions are made on the basis of risk alleles. While it is acknowledged that 'Mendelian' is an imperfect term, variants retrieved from the OMIA have been referred to as Mendelian throughout the current thesis for simplicity and to distinguish them from other variants discussed.

A short communication has been submitted to Animal Genetics on the basis of this analysis (121). During the revision process of the paper, another limitation was highlighted. Namely, the list of likely functional variants was obtained from the OMIA database in September 2023. Between then and the submission of the paper (March 2025), eight variants originally queried had been removed from the list and 111 variants had been added. Excluding large SVs, this meant that 56 SNPs and indels from the OMIA list of functional variants were not queried. Although it is not described in the results section of this thesis, the analysis was rerun, and the entire list of 443 likely functional variants reported by the OMIA as of March 2025 was queried. It was found that one variant (in *IGF1-AS*) originally detected in the sample had been removed from OMIA. This emphasises a limitation of using old datasets, since they are subject to change.

While rerunning the analysis in March 2025, an error in the code used to convert variant positions to the canFam4 reference genome was identified. Namely, a certain class of variant was being misinterpreted (i.e. where an indel existed between the reference genomes upstream of the variant, it was being misaligned). This was corrected by adding a check for indels before coordinates were updated and led to the detection of a 4bp deletion in *VPS13B* (p.Val5951fs) that had previously not been captured (OMIA ID = 478). The deletion was observed at an allele frequency of 2.3% in Huntaways (six carriers) and 1% in Heading Dogs (two carriers). First discovered in Border Collies in 2011, this allele causes trapped neutrophil syndrome (TNS), a deficiency of segmented neutrophils that leads to a compromised immune system, pyrexia, and lameness (221). While some treatments can prolong life, affected dogs usually die or are euthanised young (222). Variants in *VPS13B* have been shown to cause Cohon syndrome in humans, which is homologous to TNS (221). TNS was first described in a population of Australian and NZ Border Collies. It is therefore not surprising that the causal variant was detected in a closely related, and likely crossbred, population. A low allele frequency was observed in the current study compared to populations of Border Collies in Japan (7%) and Norway (8%). However, the severity of TNS makes this variant a strong candidate for testing and selection.

4.4.7 Future Directions for Research

It is important to note that no phenotypic evidence was considered in the functional assessment of Mendelian variants. Many of the phenotypes implicated here have been queried within the Right Dog project (see Appendix A.1) but were not yet curated beyond raw data at the time of writing. Instead, assessment relied solely on *in silico* analysis and previously published evidence from the literature. This limits the conclusions that can be made for the less well-studied variants, since *in silico* evidence is less reliable than association and experimental data (223). However, the variants highlighted here as functional candidates represent well-founded hypotheses that will be tested in future association studies within the Right Dog project. Additionally, dogs homozygous for proposed disease alleles will be followed up to determine if they develop symptoms.

Diagnosis and selection based on the compelling markers highlighted here is likely to decrease frequency of disease, improving the health of the NZ farm dog population. While classical breeding values may be of greater use within a breeding programme, marker-assisted selection can be utilised to remove animals from the mating population quicker than traditional breeding values. Mendelian phenotypes are unique in that rapid genetic gain in the population can be achieved by selecting for or against single alleles (32). Where disease allele frequencies are high, it is not feasible to exclude all carriers from mating, and Donner et al. (220) showed that with the increasing number of causal variant discoveries, it is more common for dogs to carry at least one disease allele than not. However, disease frequency can still be reduced through the avoidance of carrier-carrier matings. In order for marker-assisted selection to be widely implemented, cheaper and more accessible genotyping technologies are required than WGS (see Sections 3.8 and 3.9).

4.5 High-Impact Variants within Genes of Interest Segregating in NZ Farm Dogs

The WGS sample was queried for additional high-impact variants within 132 previously reported functional genes. This yielded a list of 92 variants within 50 genes that were segregating in the sample; however, an inspection of sequence data in IGV revealed that 63 were most likely false calls (see Appendix B.4). The remaining 29 variants, including two in coat colour genes and 27 in genes implicated in disease, were classified as true or possible true variants based on sequence data and were investigated further to assess their functional candidacy. Variants were either classified as candidate functional variants or unlikely to be functional based on a review of the literature and a summary of this review follows.

4.5.1 Assessment of Candidate Functional High-Impact Variants

4.5.1.1 CNGB1:g.57909801del (p.Ser1160fs)

A 1bp frameshift deletion in exon 33 of *CNGB1* (cyclic nucleotide gated channel subunit beta 1) segregated in the WGS sample at an allele frequency of 0.6% (three carriers). Ahonen et al. (224) reported a complex variant (1bp deletion and 6bp insertion) in *CNGB1* that caused an early truncation and was associated with progressive retinal atrophy in Papillon and Phalène breeds. Progressive retinal atrophy is an eye disease that causes deterioration of light-sensing cells. *CNGB1* has a role in photoreceptor cell function and multiple variants within the gene have been implicated in the degeneration of the retina in mice and humans (224). Given the strong functional candidacy of the associated gene, this frameshift deletion is highlighted as a functional candidate, where further testing to assess its association with the disease would be beneficial.

4.5.1.2 ABCA4:g.55659310del2 (p.Leu1818fs)

There were three carriers of a 2bp deletion within *ABCA4* (ATP binding cassette subfamily A member 4) in the WGS sample (allele frequency = 0.6%). The observed deletion was located at the left splice site of exon 40 and was predicted to cause a frameshift. Mäkeläinen et al. (225) identified a 1bp insertion in *ABCA4* that lead to a complete LOF and was associated with retinal atrophy (Stargardt disease) in Labrador Retrievers. The gene has also been associated with retinal atrophy in humans. Given its location and predicted effect, the deletion is highlighted as a functional candidate. The severity of the disease means that the variant would be a strong candidate for future selection if found to be causal.

4.5.1.3 CNP:g.20768542ins (p.Arg137fs)

A 1bp insertion in exon 4 of *CNP* (2',3'-cyclic nucleotide 3' phosphodiesterase) was heterozygous in a single Heading Dog in the WGS sample (allele frequency = 0.2%). *CNP* encodes an enzyme involved in the central nervous system and two variants within the gene in dogs have been reported to cause a late-onset lysosomal storage disorder similar to NCL (226,227). These include a 1bp frameshift deletion in Dalmatians and a missense

SNP in Weimaraners. Because the observed variant causes a frameshift in a gene implicated in a severe disease, it is a strong functional candidate for future testing.

4.5.1.4 SLC3A1:g.47766395G>A (splicing)

A single Huntaway was heterozygous for a G > A SNP within *SLC3A1* (solute carrier family 3 member 1) that was predicted to alter splicing, being located at the 3' splice donor site of exon 2. *SLC3A1* encodes a heteromeric amino acid transporter and is associated with cystine urolithiasis (a.k.a kidney stones) (228–230). Previously, a nonsense SNP, deletion, and haplotype variant in *SLC3A1* have been reported to disrupt the gene (228–230). It is not known how alternative splicing may influence the expression of this gene; therefore, it would be valuable to investigate the potential effect of this variant further.

4.5.1.5 CCDC66:g.33996814del4 (p.Arg30fs)

Three high-impact variants within *CCDC66* (coiled-coil domain containing 66) were observed, including an A > G SNP (allele frequency = 54.4%), a C > A SNP (allele frequency = 3%), and a 4bp deletion (allele frequency = 3%). *CCDC66* encodes a microtubule binding protein and has been associated with progressive retinal atrophy in Schapendoes and Portuguese Water Dogs (231,232). Affected dogs were homozygous for a 1bp insertion that caused a complete LOF. The A > G SNP (g.33965111A>G) observed in the current study was predicted to cause the loss of a stop codon. However, an inspection in NCBI Genome Data Viewer revealed that the exon was annotated in few transcripts in the Ensembl assembly and no transcript in the RefSeq assembly. It was also not expressed according to the NCBI RNA-Seq coverage. Therefore, it is unlikely to disrupt the protein sequence of most isoforms. The latter two variants were located three nucleotides apart and were perfectly linked (i.e. all individuals were heterozygous for both, homozygous for both, or homozygous for neither). Because the SNP (g.33996811C>A) was directly downstream of the deletion (i.e. in the soft clipping region) it was assumed to be an artifact of the deletion, rather than a separate variant. The 4bp deletion is highlighted as a functional candidate due to its predicted frameshift effect. There were 13 carriers and one homozygote (Huntaway) observed in the current study. If the deletion is functional, it would be of interest for testing, since the associated phenotype is severe.

4.5.1.6 GLB1:g.4009430del (p.Arg642fs)

There was a single carrier of a 1bp frameshift deletion in *GLB1* (galactosidase beta 1) (allele frequency = 0.2%). *GLB1* encodes a protein involved in carbohydrate metabolism and is associated with gangliosidosis in several dog breeds (233–235). Dogs with gangliosidosis are unable to fully degrade oligosaccharides, which leads to proportional dwarfism, weight loss, and ataxia. The associated variant in Mame Shibu Inus was a nonsense 1bp deletion with an estimated allele frequency of 0.246% (234). The deletion observed in the current study is within the last 100 nucleotides of the 80kb coding sequence, meaning the majority of the codons remain unchanged in the truncated gene. Therefore, the observed deletion is a functional candidate, but its proposed effect may be less severe than that of the previously reported nonsense variant, which is 92 codons upstream of the deletion (234).

4.5.1.7 CYP1A2:g.38260059del (p.Asn253fs) and *CYP1A2*:g.38263975ins10 (p.Gly492fs)

Described above, a nonsense C > T SNP within *CYP1A2* was observed at an allele frequency of 30.2% (OMIA ID = 274). This SNP is associated with poor metabolism of certain pharmaceuticals (188,192). Two additional high-impact variants were observed within this gene. There were 31 carriers of a 1bp frameshift deletion in exon 2 (allele frequency = 6.3%) and 37 carriers of a 10bp frameshift insertion in exon 7 (allele frequency = 7.4%). Based on the likely effect that *CYP1A2* has on metabolism, both observed variants are highlighted as functional candidates. It should be noted that because the insertion is located within the last 150bp of the 1.5kb coding sequence, the predicted frameshift effect may cause a less severe phenotype than a frameshift early in the gene.

4.5.1.8 STK36:g.25198749G>A (p.Trp1290*)

Eight individuals were homozygous and 81 were heterozygous for a G > A SNP predicted to cause a premature stop codon in exon 2 of *STK36* (serine/threonine kinase 36) (allele frequency = 19.4%). This gene is associated with primary ciliary dyskinesia (PCD) in dogs, mice, and humans, which causes an impairment of cilia movement (236). One subsequent health complication of PCD is the impaired clearance of mucus, which leads to frequent respiratory infections. In dogs, the disease was previously associated with a

splice site variant that led to the skipping of exon 20 and an early truncation (236). Because the observed variant was predicted to truncate the gene at the second exon, it is highlighted as a functional candidate. It should be noted that while PCD is not directly life threatening, it can have serious consequences, meaning its high frequency in the current population could provide evidence against its functionality. Having said that, if it is functional, its high frequency makes it an obvious candidate for future selection.

4.5.2 Assessment of Predicted High-Impact Variants that are Unlikely to be Functional

4.5.2.1 SLC7A10:g.119762155del (p.Pro10fs) and *SLC7A10*:g.119762157del (p.Gly11fs)

Two 1bp frameshift deletions within *SLC7A10* (solute carrier family 7 member 10) segregated in the WGS sample at allele frequencies of 8.6% and 91.4%. They were positioned two nucleotides apart and all individuals were either heterozygous for both variants or homozygous for one of them. *SLC7A10* encodes a protein that enables transmembrane transport of amino acids and a nonsense SNP in this gene has been reported to associate with paradoxical pseudomyotonia in English Spaniels, a disease that causes muscle stiffness after exercise (237). The high frequencies of the observed deletions in a population of highly active working dogs makes it unlikely that they cause a stiff-muscle phenotype.

4.5.2.2 CUBN:g.18781859ins84 (p.Asn997fs)

Two variants within the *CUBN* gene were predicted to be high impact. Discussed above, one was a deletion that causes the malabsorption of vitamin B12 (OMIA ID = 447) (199). The other was annotated as an 84bp insertion and segregated at an allele frequency of 9% in the population (four homozygotes and 37 heterozygotes). The insertion was predicted to truncate the gene at exon 21, upstream of previously reported LOF variants in this gene. Although large SVs can be difficult to characterize from short-read sequence data, an inspection in IGV revealed that the locus was typical of a tandem duplication (i.e. affected animals exhibited sudden increases in read depth, with affected reads displaying soft clipping on either side of the locus, and the inserted sequence was identical to the

sequence directly upstream) (43). Because the duplication spans a splice donor site, the sequence is repeated within the intronic region, and both the translated sequence and splice site remain intact. Therefore, the coding sequence is not disrupted as predicted, and the variant was classified as unlikely to be functional.

4.5.2.3 COL5A1:g.50881947del (p.Cys1087fs) and *COL5A1*:g.50881951del (p.Cys1087fs)

Two 1bp deletions were observed within *COL5A1* (collagen type V alpha 1 chain) in the WGS sample and were predicted to cause frameshifts. Several variants within this gene have been reported to cause classical Ehlers-Danlos syndrome (cEDS) in different dog breeds (238,239). cEDS is a disorder that causes joint hypermobility and leads to fragile connective tissue. The observed variants were positioned four nucleotides apart and were perfectly linked. All samples were either heterozygous for both or homozygous for one, with allele frequencies of 23.8% and 76.2%. The ubiquitous presence of at least one of these deletions in the WGS sample makes it extremely unlikely that they result in a disease like cEDS, which would impact working ability.

4.5.2.4 NDRG1:g.30235410C>T (p.Trp279*), *NDRG1*:g.30249959C>T (p.Trp37*), and *NDRG1*:g.30235368del (splicing)

Three high-impact variants, including two C > T SNPs and a 1bp deletion, were observed within *NDRG1* (N-myc downstream regulated 1) (allele frequencies = 2.4%, 2.2%, and 2.8%). *NDRG1* is involved in signal transduction and is associated with polyneuropathy in Greyhound and Alaskan Malamute breeds (240,241). Both observed SNPs were predicted to induce early stop codons and the 1bp deletion was predicted to intersect a splice donor site. However, further investigation revealed that all three variants were likely within introns, since few Ensembl transcripts annotated the regions as exons, they were annotated as introns in the RefSeq assembly, and they were not expressed according to the RNA-Seq coverage in NCBI. These variants were therefore classified as unlikely to be functional since they do not disrupt coding sequences or splice sites as predicted.

4.5.2.5 MKLN1:g.5767250T>C (p.Met1?)

A T > C SNP within *MKLN1* (muskelin 1) segregated in the WGS sample at an allele frequency of 17.8% (six homozygotes and 77 heterozygotes). The SNP was predicted to

cause the loss of a start codon. *MKLN1* encodes an intracellular protein and has been associated with lethal acrodermatitis (LAD) (242). LAD is an immune deficiency disorder that causes poor growth, skin lesions, and usually death within the first two years of life. The *MKLN1* variant that was previously reported to cause LAD resulted in the loss of exon 4 and had an extremely low allele frequency (242). The high frequency of the observed SNP in the WGS sample means it cannot cause lethality. An inspection of the reference sequence revealed a second start codon three amino acids downstream of the variant. It is likely that the second start codon is able to compensate for the loss of the first one (or is the majority use codon), explaining how the frequency of this seemingly disruptive variant is so high.

4.5.2.6 PTPRQ:g.23018717T>C (p.Met1?)

A T > C SNP in exon 8 of *PTPRQ* (protein tyrosine phosphatase receptor type Q) was also predicted to cause the loss of a start codon. The SNP segregated at an allele frequency of 65.8%, with more than 100 homozygotes observed in the current sample. *PTPRQ* enables protein binding and a nonsense variant in this gene was reported to associate with deafness in Doberman Pinschers (243). An inspection of the reference sequence revealed that the SNP was in the middle of exon 8 rather than where transcription is initiated. This locus was the start of the coding sequence in a single transcript; however, the high prevalence of the SNP and the presence homozygotes in the sample suggests that the protein is able to function without the disrupted transcript (if this annotation is indeed correct). Additionally, the study population has been selected for their ability to listen to commands, making it unlikely that a common variant causes deafness.

4.5.2.7 MLPH:g.48436229del2 (p.Thr381fs)

There were two carriers of a 2bp frameshift deletion in *MLPH*, a gene associated with reduced pigmentation in humans, mice, and dogs (164). For the same reasons as the *NDRG1* variants, an inspection of NCBI Genome Data Viewer revealed that the deletion was within an intron (see Section 4.5.2.4) and was therefore unlikely to affect the protein sequence.

4.5.2.8 ADAMTS20:g.35899564del13 (p.Trp1842fs) and *ADAMTS20*:g.35899577C>T (p.Trp1842*)

Two variants within the *ADAMTS20* (ADAM metallopeptidase with thrombospondin type 1 motif 20) gene were observed in the WGS sample, including a 13bp deletion (allele frequency = 5.6%) and a C > T SNP (allele frequency = 35%). *ADAMTS20* is associated with cleft lips in dogs (244). Again, for the reasons outlined in Section 4.5.2.4, it was determined that both variants were within introns and were therefore unlikely to disrupt the coding sequence.

4.5.3 Summary and Implications

This analysis exploited existing WGS data, the OMIA database, and SnpEff variant effect prediction to identify nine previously uncharacterised high-impact variants within genes of known functional importance (see Table 8). In total, 14 variants were highlighted as functional candidates; however, five of these had been previously reported by the OMIA. Although further investigation is required to evaluate the effect of novel candidates, this method demonstrated how deleterious variants can potentially be highlighted without the use of phenotypic data. Additionally, potential disease phenotypes of interest in the NZ farm dog population were identified. The novel variants within *CNGB1*, *ABCA4*, *CNP*, and *CCDC66* were particularly strong functional candidates, since they were predicted to cause molecular effects at least as severe as known causal variants within those genes. The functional candidates within *CNGB1*, *ABCA4*, *CNP*, *SLC3A1*, *CCDC66*, *GLB1*, and *CYP1A2* are also of particular interest, since they may be implicated in severe diseases.

Unlike the previous analysis, most of the variants described here have never been characterised. Therefore, the *in silico* evidence provided is not robust enough to diagnose or select on functional candidates without further analysis. Instead, each novel candidate represents a causal hypothesis, providing an opportunity for future association testing and/or experimental studies. Novel causal candidates were identified for seven disorders: retinal atrophy, Stargardt disease, lysosomal storage disease, cystinuria, gangliosidosis, variable metabolism of pharmaceuticals, and PCD. These represent possible disorders of concern in the Huntaway and Heading Dog breeds, where

future studies could investigate their prevalence in the population and test their association with the candidate causal variants highlighted here. Dogs that were homozygous for the candidate disease variants will be followed up to determine if they develop symptoms.

4.5.4 Areas for Improvement in the High-Impact Variant Survey

This was an exploratory analysis and, as such, the methodology was imperfect. The analysis was limited to high-impact variants in order to investigate the most likely causal variants while retaining a manageable dataset. Because of this, several functional variants affecting the genes of interest were not assessed. For example, only five of the Mendelian variants described in Tables 6 and 7 were detected, since the others were not considered to have high impacts by SnpEff, despite many being functional. Additionally, it cannot be concluded that no high-impact variants were present in the genes of interest where they were not detected. This is because, although both Ensembl IDs and gene names were queried, some genes may have been annotated by different IDs/codes.

A large proportion of the high-impact variants were false positives. This is not surprising, since high-impact variants tend to be deleterious and so are selected against in populations (22). Because of this, a greater proportion of annotated high-impact variants are likely to be sequencing/variant calling errors than low-impact variants. Of the variants that were assumed to be real (i.e. polymorphic in the population), more than half were classified as unlikely to be functional. Again, this is not surprising as functional variants tend to be selected against and are therefore rare. For example, several variants that were predicted to cause frameshifts and/or truncations were reclassified as intronic based on an inspection in NCBI Genome Data Viewer. Interestingly, all but two of the variants discussed here were implicated in disease. This reflects the overrepresentation of disease variants in the OMIA database (220).

4.6 SNP Chip Data Quality Assessment

Based on the distribution of four important metrics (missing call rate, missing genotype rate, F statistic, and MAF), the majority of the SNP chip data was of high quality (see Figure 10). There was a small proportion of missing data, with over 97% of SNPs having missing genotype rate < 0.1, and all but two samples having missing call rate < 0.1. These are common QC thresholds and were applied to the SNP chip data for most downstream analyses (245). Often, it is important to remove samples and SNPs with large amounts of missing data, since missingness is usually non-random and can therefore lead to bias (246). Commonly, SNP chip data is also filtered for departures from HWE, as deviations can represent genotyping assay errors. However, variants, especially those that are rare, can deviate from HWE for a variety of reasons, including population substructure and inbreeding. Therefore, an HWE filter was omitted from the LD and imputation analyses to favour sensitivity. By contrast, SNPs included in the MLMA GWAS were filtered for HWE to generate a dataset less likely to generate false positive association signals.

Approximately 12% of SNPs on the Axiom™ Canine HD Array were not polymorphic in the current sample. The SNP chip therefore captured approximately 630,000 SNPs across the genome, rather than >710,000. This is not surprising, since the array was built from a sample that included many companion dogs and other breeds that were not necessarily relevant to the current study. As stated, the WGS data and functional annotations produced by the current study could be used to develop a Huntaway and/or Heading Dog breed-specific genotype panel to capture the most functionally relevant sites in the population. In a study that genotyped 471 canids from 30 breeds with the 170k Illumina Canine HD Array, a similar proportion (10%) of sites were monomorphic (247). While the array used here was not a perfect representation of the target breeds, it still captured an extremely large number of SNPs compared to other arrays. For example, the 170k array, which genotypes less than a third of the sites captured here, is considered high density (248). The high resolution of the SNP chip data used here likely contributed to the success of imputation by enabling accurate haplotype reconstruction (85). Additionally, GWAS rely on test SNPs being in LD with causal variants, therefore significant signals are more likely to be detected from high-density data (88).

As expected, the distribution of MAF revealed an excess of rare variants, where 18% of variants (including monomorphic sites) had MAF < 1% (249). This is a result of purifying selection, which reduces MAF by removing deleterious variants from the population. It can be assumed that the distribution of MAF amongst functional variants would be even more skewed, since functional variants are more likely to be deleterious, and therefore are more likely to be selected against.

The mean inbreeding coefficient (F) was significantly greater than 0, indicating there is inbreeding in the population. This is not surprising since dogs are known to be inbred (250). The mean F value across 227 dog breeds in a previous study was 0.25, suggesting that the current sample may be comparatively less inbred than some other dog populations (250). There was no evidence of a significant difference in the mean F statistic between Huntaways and Heading Dogs; however, the mean in both breeds was lower than the overall sample mean. While this is most likely due to random error, it could suggest that the minor breeds represented in the sample have greater inbreeding coefficients on average than the breeds of interest. It should be noted that because multiple dogs from the same farms and owners were sampled, highly related individuals presumably exist in the dataset. Additionally, population structures, like multiple breeds, have been shown to distort individual inbreeding, for example through the Wahlund effect (133). Because the sample is not random, and population structures exists, strong conclusions about the extent of inbreeding in the NZ farm dog population cannot be made on the basis of this analysis.

Some individuals had an excess of heterozygosity compared to the reference population, meaning they were outbred (F < 0) (133). This could indicate that these individuals were crossbreeds or not from the target breeds. However, one individual had a substantially lower F value than all other individuals in the sample (F = –0.35). While this dog's breed was not recorded, it was speculated to belong to a genetically distant breed. This would explain its low F value, since it would be extremely outbred by comparison. Follow-up investigation revealed that the dog had been sired via artificial insemination from a dog that died 20-30 years prior. This could explain the dog's genetic distance from the sample, since it contains DNA from many generations ago. Alternatively, if the sample was of

inadequate quality (i.e. due to contamination or poor swabbing), this could cause an artificially low F value resulting from error.

4.7 Genetic Verification of Matching WGS/SNP Chip Samples

One important result of this work was using correlations between the Mendelian variants and SNPs on the chip to identify predictive markers that could be genotyped by using a cheaper, more routinely applied technology than WGS. To do this accurately, it was first required to verify that the samples for each dog matched between the WGS and chip genotype datasets, and to remove any dogs where the samples were mismatched; this was done using two metrics: genetic relatedness (A_{jk} score) and discordance rate. The substantial difference observed for both metrics in within-individual pairs of samples compared to between-individual pairs meant that identical pairs could be distinguished from non-identical pairs, allowing the identification of mismatched samples. Based on their extreme values, three of the 187 pairs assigned to the same individual were found to be too genetically dissimilar to have been derived from identical DNA.

As expected, most between-individual pairs yielded A_{jk} scores of approximately 0, indicating that they were unrelated (133). However, a number of pairs had A_{jk} scores of ~0.5 or ~0.25, representing first- and second-degree relatives (see Figure 11). This supports the previously stated hypothesis that highly related individuals exist in the dataset, likely as a result of convenience sampling of all dogs on the same farm or belonging to the same shepherd. Most identical pairs had low discordance rates (i.e. below 0.01) and genetic relatedness scores near 1, indicating that there was high overall concordance between the genotyping technologies (WGS and SNP chip). However, the fact that some discordance was observed in genetically identical samples shows that WGS and SNP chip genotypes are not interchangeable. Where differences arise, it can usually be assumed that the SNP chip-reported allele is correct, since the quality of SNP chip genotypes tends to be much higher than that of WGS genotypes (76). However, this can depend on the genetic distance between the target population and the population used to build the SNP chip. Because the WGS-reported alleles for candidate markers may

be inaccurate, only dogs present in both the WGS and array datasets were used to calculate the LD between variants of interest and SNP chip variants.

4.7.1 Deviations from Expectation

The WGS and SNP chip genotypes from three within-individual comparisons were too dissimilar to have been derived from the same DNA and were therefore assumed to come from distinct individuals. One of these comparisons yielded an A_{jk} score of 0.2 and a discordance rate of 0.35, values that are consistent with between-individual comparisons. Upon further investigation, it was determined that the WGS sample was incorrectly assigned to a different dog's SNP chip sample. The second comparison yielded an A_{jk} score of 0.6 and a discordance rate of 0.2, while the third comparison yielded an A_{jk} score of 0.7 and a discordance rate of 0.1. An obvious labelling error could not be identified for the latter two comparisons. Interestingly, their scores were consistent with comparisons between first degree relatives. It is therefore speculated that the dogs may have been incorrectly assigned to the corresponding sample from a first degree relative, but it is also possible that their samples were cross contaminated.

One between-individual comparison yielded an A_{jk} score of 0.8 and a discordance rate of 0.19. While this discordance rate is consistent with a comparison between first degree relatives, the A_{jk} score of 0.8 is considerably greater than 0.5, which is the expected relatedness between first degree relatives. This could be a result of animals inbreeding over several generations, meaning the pair shared more than 50% of their DNA (251). Alternatively, because the two dogs in this pair had the same owner, their samples may have been cross contaminated.

4.7.2 Shortcomings in the Calculation of Genetic Similarity

Theoretically, comparisons between unrelated individuals yield A_{jk} scores of approximately 0 and comparisons between identical individuals yield A_{jk} scores of approximately 1. In the current analysis, several A_{jk} scores beyond this range were observed, with multiple within-individual pairs scoring above 1 and between-individual

pairs scoring below 0. This can occur when the sample is extremely small or highly related (133). The inclusion of genetically identical pairs created artificially high inbreeding in the current sample, likely resulting in the unusually high and low relatedness scores that were observed. This is because the relatedness between each pair is relative to the average relatedness in the sample, which is artificially high here. When two individuals share a large number of rare alleles, which is clearly the case when individuals are identical, estimates can exceed 1. By contrast, scores below 0 indicate that pairs are less related to each other than the average relatedness in the sample. These values cannot be directly interpreted. Therefore, while this analysis was successful in distinguishing identical pairs from distinct pairs, strong conclusions cannot be made about the extent of relatedness between individual pairs. In the GWAS analysis, the GRM was recalculated between only non-identical pairs.

The sample size of the between-individual comparison group was almost 400 times larger than that of the within-individual comparison group. This largely explains the wider range of the relatedness metrics observed in the between-individual group compared to the within-individual group. Additionally, it limits the interpretation of the two-sample T-tests, since the assumption of equal variance is violated (252). Again, this means results may not be widely applicable, but the analysis was extremely useful for removing mislabelled pairs from the overlapping sample.

4.8 Identification of Predictive Markers for Mendelian Variants of Interest

The R^2 correlation was calculated between the 27 functional Mendelian variants of interest (see Section 3.4) and all variants on the Axiom™ Canine HD Array to identify potential predictive markers. The aim of this analysis was to enable marker-assisted selection without the need for WGS or specialised genetic tests. Highly correlated predictive markers on the SNP chip were identified for eight of the Mendelian variants, where five were able to be directly genotyped and three were in perfect LD ($R^2 = 1$) with alternative SNPs on the array (Table 9 outlines the genomic positions of these variants

and their predictive markers). The Mendelian loci that could be directly genotyped with the array were within the *FGF5*, *P3H2*, *CYP1A2*, *MFSD12*, and *IGF1-AS* genes and the Mendelian variants in perfect LD with alternative SNPs on the array were within the *RETN*, *SOD1*, and *KCNJ10* genes. The number of genotypes available for these variants in the Right Dog project is therefore 2400 (i.e. the number of dogs that will be SNP-chip genotyped by the end of the project), rather than 250 (i.e. the number of dogs that were sequenced), creating a larger sample for future analyses. Ten other Mendelian variants correlated with SNPs on the array; however, these markers are less informative due to their weaker correlations ($0.9 > R^2 > 0.2$) (253).

The low MACs of the candidate markers for the *RETN*, *SOD1*, and *KCNJ10* variants created modest sample sizes for the calculation of R^2 (MAC = 14, 39, and 32). This may be of concern since low frequency markers can create spurious correlations that do not reflect true genetic linkage (254). Therefore, while these predictive markers are informative, they may not be deterministic beyond the current sample and extra caution should be taken in their use as diagnostics (253). It should be noted that the position of the *MFSD12* variant was updated in 2024, meaning the marker tested here corresponded to the incorrect position. The LD between the correct position and SNP chip loci was not calculated within this study, since the analysis was performed prior to the discovery of the position change.

4.8.1 Advantages of Predictive Marker-Assisted Selection

SNP chip genotyping is cheaper, faster, and more readily accessible to dog owners than WGS. Therefore, in the absence of a custom chip that captures all causal variants, it is extremely beneficial to use predictive markers on existing chips to select for traits. Studies have shown that the same amount of genetic gain can be achieved through SNP chip-based selection as can be achieved through WGS-based selection (253). However, caution is required when using predictive tags in a clinical setting, since they are not direct tests and false positives/negatives can be damaging. In many animal breeding programs, each individual's genetic worth is determined by the average phenotype of their progeny. Marker-based selection provides the opportunity for a targeted approach

to selection that does not rely on an individual's phenotypes, or the phenotypes of their offspring. This has economic impacts because less money is spent rearing animals that won't be bred. Additionally, genetic disease markers can be detected in asymptomatic carriers and before affected animals develop symptoms, allowing early diagnosis, preventative treatment, and/or selective mating. Although genomic selection based on BVs is becoming more widely used in animal breeding programmes, marker-based selection has been extremely successful at rapidly decreasing the frequencies of severe diseases through the avoidance of carrier-carrier matings (12). Additionally, most implementations of genomic selection only consider additive genetic effects and are therefore not as effective at removing rare recessive disease variants from the population as marker-assisted selection is (255).

4.8.2 Challenges in the Identification of Predictive Markers

Most association studies that aim to identify predictive markers rely on phenotypic data, since the causal variant is usually not known. This LD analysis is unusual in that rather than searching for markers correlated with phenotypes of interest, it searched for markers correlated with causal variants of interest, and no phenotypic data was utilised. Therefore, alternative considerations were required.

Initially, with the aim of retaining all potential markers, no call rate filters were applied. This led to several spurious perfect correlations, since SNPs with low call rates had extremely low MACs and therefore were perfectly correlated with variants of interest by chance. This was corrected by applying missing call rate and missing genotype rate filters of 0.1, and by excluding interchromosomal comparisons. Additionally, 1267 SNPs on the array were excluded since their positions were not reported, and therefore their distance from variants of interest could not be used to assess their candidacy. This is an important consideration because LD tends to be negatively correlated with physical distance between loci due to the increased probability of recombination occurring between loci that are far apart (78). Having said that, LD can span large distances, for example, if two amino acids are bound in the 3D structure of the protein, it is advantageous for them to coevolve (79). In this way, loci that are far apart in the primary sequence can still be linked.

It is difficult to accurately calculate LD between rare variants. Some studies suggest that variants with MAF < 5% should be excluded from LD analyses to avoid spurious correlations (125). However, MAC is a more direct reflection of the amount of variability available to calculate LD from, as it doesn't rely on sample size. The current study had a particularly small sample size, meaning rare variants were not well represented and tagging markers could not be identified for them. This also meant that, even for common variants, some markers may have been correlated due to chance rather than true linkage.

In a similar study conducted in chickens, Geibel et al. (256) were able to identify markers on a 600k SNP chip for 90% of causal variants, where markers had $R^2 > 0.75$ and were within 15kb of the variant of interest. In the current study, markers that fit this description were only identified for the eight aforementioned variants (i.e. 30% of the variants of interest). There are several possible explanations for the lower success rate in the current study, for example, more stringent marker filtering criteria may have been applied here, or the array used by Geibel et al. (256) may have been biased toward positions near variants of interest. Retrospectively, the exclusion of the SNPs with unreported positions was likely a large contributor to the small number of detected predictive markers. This is because many of these SNPs would have been genetic disease tests protected by patents and may have been diagnostic of the Mendelian disease variants of interest. This analysis could be repeated per chromosome, where the unmapped SNPs are assigned dummy positions on each chromosome in turn. Because the SNP positions are unknown, however, it would be particularly difficult to distinguish true linkage from random chance given the small sample size and low frequencies of many variants of interest.

4.9 Imputation of Missing Genotypes in the SNP Chip Sample

Using the WGS dataset as a reference, missing genotypes across millions of variants were imputed in 111 animals that had been genotyped with only the SNP chip. These high-accuracy imputed genotypes expand the genomic dataset without the need for further sequencing. In 2025, the estimated cost of WGS per animal is between $500 and $1000 NZD, depending on the coverage, sample size, and size of the genome (257,258). By contrast, the Ancestry Know Your Pet DNA genotyping kit, which uses the 700k Canine

HD Array, sells commercially for around $170 NZD per animal. This huge difference in price shows the economic value of imputation, since it means dogs can be genotyped for millions of variants without sequencing. However, because imputation is probabilistic, there is uncertainty in imputed genotype calls (85). Like the WGS and SNP chip genotypes, it is therefore unlikely that the imputed genotypes will be used for clinical diagnostics. Instead, imputed genotypes are most useful in research, since they increase statistical power with minimal extra cost and most analyses are robust to the small amount of introduced error. Imputed genotypes can be used to extend GWAS, fine-map causal variants, combine datasets from alternative genotyping technologies, and decrease the cost of genetic/genomic selection (86).

Of the 19 Mendelian variants of interest that were not in high LD with single SNP markers (see Section 3.8), 11 could be imputed with high accuracy ($DR^2 > 0.9$). These variants were within *MC5R*, *MC1R*, *BTBD17*, *TYRP1*, *ABCB1*, *LMBR1*, *ATP7B*, and *ASIP*. This provides an opportunity for marker-based selection without the need for WGS or specialised tests; however, as stated, caution is required for diagnostics. Imputed allele frequencies tended to stay relatively consistent across imputation runs but sometimes differed substantially from the WGS allele frequencies (see Table 11). This could suggest that the WGS sample does not represent the SNP chip sample well due to substantial genetic variation between them. This potential difference could be a result of convenience sampling and would reduce imputation accuracy if the reference sample and target sample are not genetically matched. Unsurprisingly, Mendelian variants that were imputed with low DR^2 values did not show consistent allele frequencies across runs, providing further evidence that their imputed genotypes are inaccurate. Like in the LD analysis, imputation was performed prior to the discovery that the *MFSD12* variant position was incorrect, meaning the imputed *MFSD12* genotype described in Table 11 corresponds to the incorrect variant.

4.9.1 Challenges and Limitations of Imputation

Imputation was run four times, where each run referenced a more strictly filtered reference panel. The aim of this was to achieve the highest possible accuracy. As

expected, the mean accuracy increased with each run, but naturally, the number of genotyped sites decreased (see Table 10). It is not known whether the increased average DR^2 is simply due to the exclusion of low accuracy genotypes or due to the accuracy of imputed genotypes themselves increasing. For the 25 Mendelian variants of interest that were imputed in all runs, there was no evidence of a significant difference in mean DR^2 between runs. Future studies could investigate whether there was a significant difference in accuracy between runs across all variants that were imputed in all runs. It should be noted that only the effects of N_e, MAF, and missing call rate were adjusted to optimise accuracy. Other contributing factors like GC content and recombination rate were not considered here (86).

The reference panel used in the current study was small (n = 250 individuals). This can negatively impact imputation since rare variants and haplotypes tend to be represented poorly in small samples (86). It is therefore not surprising that variants with low MAF tended to have lower DR^2 values, and removing variants with MAF < 1% from the reference increased the average DR^2 by 0.09. The accuracy of rare variant imputation in small samples can be increased by incorporating family-based approaches or tailored reference panels (259). However, as stated in Section 2.10, a small sample may contain a representative proportion of the population's haplotypes so long as the sample is representative of the population and the effective population size is small, which is likely true here (128). For example, the first study published by the 1000 Bull Genomes project used a 234-sample WGS reference panel to impute missing genotypes in SNP chip-genotyped animals with a high average accuracy (260). To retain maximum statistical power, the current study included Huntaways and Heading Dogs in the same reference panel. Daetwyler et al. (260) demonstrated that when a small reference sample was used to impute missing genotypes in cattle, a higher accuracy was achieved when all breeds were included in the reference than when breed-specific imputation was performed.

4.10 Genome-Wide Association Studies for Four Morphological Traits

The final analysis of the project aimed to identify genetic associations with height, length, chest circumference, and muzzle circumference in NZ working dogs. The size and body

structure of workings dogs are important factors in determining their working ability and have been linked to disease susceptibility (8). Body size is more variable in dogs than any other mammal (250). Despite this, it has been estimated that seven SNPs in six genes explain half of the phenotypic variability of size in dogs (261). More recently, Plassais et al. (5) estimated that 95% of size variability in purebred dogs can be explained by just 14 genes: *IGF1R*, *LCORL*, *STC2*, *GHR*, *SMAD2*, *HMGA2*, *ZNF608*, *IGF1*, *R3HDM1*, *ADAMTS9-AS*, *HNF4G*, *ACSL4*, and *IGSF1*. Many previous dog GWAS have used weight and/or height as a proxy for body size, with few previous GWAS for length, chest circumference, or muzzle circumference (5,110,247). Momozawa et al. (110) performed a GWAS to identify genetic associations with chest circumference but yielded no significant results.

While several studies have characterised the average weight in the NZ farm dog population, there are very few estimates available for the four aforementioned measurements (7,262). Dogs New Zealand (263) estimate Huntaways to be between 560 and 660mm tall, but do not document any estimates for Heading Dogs. In a review of working dog body structure, Zink and Schlehr (8) state that most working dogs from the US are between 540mm and 650mm tall, and the 'ideal' working dog is slightly longer than they are tall. The height of most dogs in the current sample was also within this range, with a mean of 583mm (see Figure 6), and dogs tended to be longer than they were tall, with an average (mean) per-sample difference between height and length of 72mm. Breed-specific density plots were generated for the four phenotypes of interest (see Appendix B.6). As expected, the mean height, length, chest circumference, and muzzle circumference were greater in Huntaways (means = 615mm, 652mm, 731mm, and 270mm respectively) than in Heading Dogs (means = 551mm, 600mm, 664mm, and 224mm respectively), and this difference was statistically significant. Given dogs were selected to represent multiple geographic regions, these means are likely representative of the population.

4.10.1 Height GWAS

A single SNP near *LCORL* in chromosome 3 was significantly associated with height according to the MLMA-LOCO GWAS. Typically, regions that are truly associated with a

trait result in several significantly associated SNPs in LD with the causal variant (88). These form the characteristic 'peaks' of a Manhattan plot. However, the SNP's proximity to a known body size gene, *LCORL*, suggests that the association is real, despite it not being within a peak. Additionally, a strong peak was observed in this region in the GLM GWAS for height (see Figure A6). There were no genes, other than *LCORL*, within 100kb of the significantly associated region, providing further evidence that it is the causal gene. The estimated effect size of the associated SNP was 30.08mm (i.e. correcting for relatedness, height is predicted to increase by 30.08mm for every copy of the risk allele that a dog carries). This large effect, along with the findings from Hayward et al. (166) and Plassais et al. (5), suggests that *LCORL* could explain a large proportion of body size variability in the population (167). Plassais et al. (5) predicted *LCORL* to explain almost 16% of the phenotypic variance of height in a cohort of purebred dogs. The fact that the current study was able to detect this association also implies the effect of *LCORL* is large, since there was a lack of statistical power to detect small effects here.

LCORL has been associated with height and/or body size in several mammals including humans, cattle, and dogs, but no causative variants have been definitively demonstrated for these effects (5,110,166,247,264). Through the use of WGS data, Plassais et al. (5) identified a 1bp insertion in the last exon of *LCORL* as a candidate causal variant (rs3327936124; ROS_Cfam_1.0 position chr3:92262279 T > TA). The frequency of this allele in small, medium, and large breeds was 0%, 1%, and 68% respectively. In the current sample, the frequency of this allele was 2% in Heading Dogs (one homozygote and five heterozygotes), and 51% in Huntaways (33 homozygotes and 66 heterozygotes). The position of this variant in canFam4 is chr3:91872822. Interestingly, the variant was called as a deletion in canFam4, suggesting that the German Shepherd (large dog), which the genome was built on, was homozygous for the insertion. The higher observed frequency in Huntaways, a large breed, compared to Heading Dogs, a medium breed, would make sense if the nonsense variant is functional. Overall, this result provides strong evidence that *LCORL* is implicated in height/body size in the current population and could therefore be selected on.

A second SNP was also significantly associated with height according to the MLMA-LOCO GWAS. The associated locus was within 100kb of *JHY*, *UBASH3B*, and *CRTAM*.

None of these genes have known associations with body size phenotypes. *JHY* is thought to be involved in axoneme assembly and brain development and probably works upstream of many processes (265). *UBASH3B* has been found to inhibit the endocytosis of growth factor receptors, which could potentially influence body size; however, its role in mammalian development is unknown (266). *CRTAM* encodes a transmembrane protein involved in the regulation of T cells (267). Because only a single SNP at this locus was significant, and no previous studies have associated the region with body size, the association cannot be ruled out as a false positive. To determine whether this signal reflects a true association, the GWAS should be repeated when more samples become available.

Interestingly, no peaks were observed near the most well-characterised height/body size gene, *IGF1*, or any other known body size genes (167). This is most likely a reflection of the lack of statistical power in the current study, rather than them not having effects in the population. It is possible that causal variants in this gene do not segregate at high enough frequencies in the population for their effect to have been detected here.

4.10.2 Length GWAS

The length MLMA-LOCO GWAS yielded no significant results. Again, this is likely due to a lack of statistical power. The GLM GWAS resulted in four associations that were significant (see Figure A7); however, none of these were near known body size loci. Length has been associated with *IGF1* in a previous dog GWAS, but no signal was detected in this region in the current study with either model (110). While length was normally distributed in the sample, it is speculated based on individual length to height ratios that some participants included head in the length measurement, despite the instruction that it be measured from the withers. This would create some noise in the phenotypic data that makes it more difficult to detect associations. A larger sample size would decrease the effect of this random noise and increase the power to detect length associations.

4.10.3 Chest and Muzzle Circumference GWAS

In all GWAS models run here, chest circumference and muzzle circumference yielded extremely similar results, indicating they are highly correlated. This could suggest that in the current sample, chest circumference reflects the bone structure of the ribs rather than fat, since this is likely to be correlated with the bone structure of the jaw. This would make sense because working dogs tend to be extremely fit and therefore don't carry excess fat around the chest. According to the MLMA-LOCO GWAS, a single SNP on chromosome 32 was significantly associated with both chest and muzzle circumference. The significant locus was within 100kb of six genes (*CASP6*, *CFI*, *GAR1*, *RRH*, *MCUB*, and *PLA2G12A*), none of which appear to have been associated with body size. *CASP6* plays a central role in cell apoptosis, *CFI* regulates the complement cascade, GAR1 is a ribonucleoprotein involved in ribosome biosynthesis, and *RRH* has been implicated in pigmentation and hair loss (268–271). None of these functions have an obvious connection to body size. *MCUB* is a negative regulator of calcium importation in the mitochondrion, and deletions within this gene have been shown to cause fat accumulation (272). Similarly, *PLAG2G12A* has been shown to protect against obesity and insulin resistance (273). The functions of the latter two genes provide a potential mechanism by which they could affect chest circumference. However, as stated, working dogs don't tend to accumulate fat around the chest or muzzle meaning that neither of these phenotypes is likely to reflect fat accumulation as much as bone structure. Like those significantly associated with height, this SNP was not within a peak. Therefore, given the lack of statistical power in the current study, it would not be surprising if it represents spurious association. The GWAS should be repeated when more samples become available in the Right Dog project.

Although not observed in the MLMA-LOCO GWAS, the GLM GWAS yielded highly significant peaks in chromosome 20 for both chest and muzzle circumference (see Figures A8 and A9). In a body size GWAS of ~1800 dogs from multiple breeds, Hayward et al. (166) observed a significant peak in the same region, which is near a well-known coat colour gene, *MITF*. Because coat colour differs between breeds, it seems likely that this association reflects some breed effect, especially because it disappeared when population structure was accounted for with a GRM. However, if this was the case, one

would expect to observe peaks across the entire genome reflecting all variants that segregate perfectly with breed. Instead, a single prominent peak was observed, indicating the association likely reflects a true phenotypic effect. *MITF* has been associated with body weight in quails and mice, but there is little evidence that it affects body size in dogs (274,275). It should also be noted that there are several other genes in this region, although none have been previously associated with body size. Further testing is required to determine whether *MITF* or another gene in this region effects the variation of chest and muzzle circumference in the current population.

4.10.4 Areas for Improvement in the GWAS

The obvious major limitation of this analysis is the lack of statistical power due to a small sample size. Uffelmann et al. (88) suggest that at least 10,000 samples are required to yield reproducible GWAS results for most traits. However, this depends on many factors such as the genetic architecture of the studied species, mode of inheritance, effect size, and allele frequency. Non-additive GWAS require even larger sample sizes to achieve the same statistical power as additive GWAS. Because of this, only additive models were run here, meaning any non-additive effects would not be detected. The analysis was also restricted to autosomes, since alternative considerations are required for non-autosomal chromosomes (94). At this early stage in the Right Dog project, there were not enough dogs with disease phenotypes to perform case-control studies. This meant that the potential phenotypes of interest identified in Sections 3.4 and 3.5 could not be tested.

Population structure can lead to false positive associations in GWAS (88). For example, when multiple breeds are included, significant associations may reflect arbitrary breed differences, rather than true phenotype effects. While optimising the current analysis to account for this, several models were run. The most significant associations were observed when a GLM was run, and a fixed factor for breed was fitted to account for population structure. This is not surprising because fixed factors do not strongly penalise population structures and many of the significant associations from these models are certain to be false positives (see Appendix B.7). By contrast, when population structure was accounted for by fitting a GRM as a random effect, very few significant associations

were observed (see Figures 14-17). This is to be expected as GRMs penalise population structure to a greater degree to remove spurious associations (130).

Because the *LCORL* peak is assumed to be real based on previous studies, it should not have been removed by the GWAS model. It is therefore speculated that the MLMA-LOCO model penalised the SNPs in the peak too hard, resulting in false negatives. This can occur when phenotypes are perfectly confounded by breed and suggests the penalty term in the model is too stringent (130). Additionally, it has been advised that a sample size of at least 3000 is required to precisely (SE < 0.1) estimate the amount of genetic variation explained by common variants (i.e. the GRM) (93,276). In an attempt to correct this, several additional models were trialled that fitted different combinations of fixed and random variables, such as a PCA GWAS where various numbers of PCs were fitted as covariates. However, all these attempts yielded extremely similar results to the MLMA-LOCO GWAS. Because of this, it cannot be determined whether other GLM peaks that were minimised in the MLMA-LOCO GWAS were true or false positives. To eliminate breed confounding, it would be useful to perform GWAS separately in the Huntaway and Heading Dog breeds. The current sample size is not large enough to perform such analyses meaningfully, though should be possible at later stages of the Right Dog programme.

4.10.5 Future GWAS in the Right Dog for the Job Project

As mentioned above, the four GWAS will be repeated once a larger sample (n > 2000) is available as part of the Right Dog project. This will enable more reproducible results by increasing statistical power and will mean that breed-specific and case-control studies can be meaningfully performed. In particular, GWAS should be performed for the potential phenotypes of interest identified in Sections 3.4 and 3.5. GWAS can also be performed for phenotypes that were not discussed in this thesis to identify novel candidates for selection. Additionally, imputed variants should be included in the GWAS to perform fine-mapping and identify causal variants, rather than just associated variants (277). While a sample size of ~2000 is still relatively small for detecting non-additive and rare associations, these alleles can sometimes have especially strong effects, increasing

the likelihood that they are detected (278). GWAS rely heavily on the existence of strong LD between the SNPs on the array and causal variants (279). Studies have shown that haplotype markers have more power in association studies than single variant markers because single variants do not consider LD information of flanking markers. Haplotypes are also more powerful for fine mapping causal variants for this reason. Therefore, future GWAS in the Right Dog project could consider haplotype markers, rather than single SNP markers.

Due to the unique genetic architecture of dogs, there is an opportunity to detect genetic associations with traits that are complex in most species. Strong selection for desirable phenotypes has led to severe inbreeding and consequently, diseases in dogs are often caused by few variants with large effects (280). Furthermore, these diseases can often be attributed to identical variants within breeds that were inherited from a common ancestor. By contrast, most diseases in less inbred species, like humans, are caused my many variants with small effects, where shared diseases may be attributed to several distinct variants. This may be due to relaxed negative selection in dogs and/or deleterious alleles 'hitchhiking' with variants under selection (280). It can be speculated that other complex traits, like working behaviour, may exhibit the same pattern in dogs, which provides a unique opportunity to detect signal for traits that may be too complex to observe in GWAS in other species. For example, Mahmoodi et al. (281) performed GWAS for nine behavioural traits in dogs and identified 41 significant signals. More recently, Jeong et al. (282) performed a GWAS in Border Collies and linked *EPHB1* (ephrin type-B receptor 1) with herding behaviour.

The Right Dog for the Job project has surveyed participating dog owners on multiple behavioural traits in their dogs (e.g. trainability, predator drive, and barking) (see Appendix A.1). The aim here is to perform GWAS to detect variants that influence important working traits like the ability to learn commands and perform specific herding tasks (4,282). Selection for such variants would allow farmers to breed and train only the dogs with the highest genetic potential. Similarly, the identification of genetic markers for disease traits, like GDV, would prevent farmers from investing their time and money on dogs that will ultimately develop a disease and be unable to work (13). Therefore, the data produced in this project will contribute to improving the population and increasing farming efficiency.

Chapter 5: Conclusions

This thesis conducted several key experiments as part of the Right Dog for the Job project, where a variety of bioinformatic techniques were applied to characterise and functionally annotate the genetic variation in the New Zealand Huntaway and Heading Dog breeds. After aligning the whole genomes of 250 working dogs, 20 million variants were called, filtered, and annotated to generate an extensive database, including 37,500 functional predictions. These aligned sequences and annotated variants comprise the largest dataset for Huntaways and Heading Dogs to date, and likely one of the largest working dog genomic datasets in the world. This foundational research will be invaluable for future research projects.

The first use of this dataset was demonstrated by surveying the sample for previously reported functional variants, leading to the detection of 27 known Mendelian variants segregating in the population. Five of these have been highlighted as strong candidates for disease diagnostics and selection, including those within *SOD1*, *VWF*, *CUBN*, *CLN8*, and *SGSH*. A survey of high-impact variants within known functional genes led to the detection of nine novel candidate functional variants within *CNGB1*, *ABCA4*, *CNP*, *SLC3A1*, *CCDC66*, *GLB1*, *CYP1A2*, and *STK36*. The functional assessment of these variants excluded phenotypic data and was instead limited to *in silico* predictions and previously published literature. Therefore, it is proposed that future research projects test and validate their phenotypic effects through association studies and/or functional experiments.

In the second stage of this project, the WGS variant dataset was combined with a 299-sample dataset of SNP chip genotypes to identify predictive markers for functional Mendelian candidates and impute missing genotypes. Of the 27 Mendelian variants of interest, 19 genotypes were able to be accurately predicted with SNP markers or imputation, providing further opportunity for marker-assisted selection. This analysis demonstrated how accurate imputation can expand genomic datasets; however, both the LD estimation and imputation of rare variants were limited by the small sample size. Finally, GWAS detected a significant association between height and a region containing a well-known body size gene, *LCORL*, where the proposed causal SNP segregated in the

sample. Additionally, GWAS showed some evidence of an association between chest and muzzle circumference and a region near *MITF*. These studies were limited to an extremely small sample size yet demonstrate interesting preliminary findings that can be tested more robustly in the future.

Future research should test the functional hypotheses developed here with the ultimate aim of providing additional opportunities for genetic testing/selection. The WGS dataset should be leveraged to compare Huntaways and Heading Dogs with other dog breeds, determine their evolutionary history, and characterise their genetic divergence. Imputation, variant effect prediction, and/or Bayesian statistical models can be applied to fine map causal variants in regions that were significantly associated with body size. Once more phenotype data becomes available, further GWAS will be performed in the Right Dog project to test for additional associations with health and performance traits. In particular, the genetic architecture of dogs may enable the detection of genetic associations with behaviour traits, representing a significant scientific prize given the difficulty of studying these traits in species such as humans.

References

1. Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, et al. Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. Cell Rep. 2017 Apr 25;19(4):697–708.

2. Wang GD, Zhai W, Yang HC, Wang L, Zhong L, Liu YH, et al. Out of southern East Asia: The natural history of domestic dogs across the world. Cell Res. 2016 Jan 1;26:21–33.

3. Jagannathan V, Drögemüller C, Leeb T, Aguirre G, André C, Bannasch D, et al. A comprehensive biomedical variant catalogue based on whole genome sequences of 582 dogs and eight wolves. Anim Genet. 2019 Dec 1;50(6):695–704.

4. Bray EE, Otto CM, Udell MAR, Hall NJ, Johnston AM, MacLean EL. Enhancing the Selection and Performance of Working Dogs. Front Vet Sci. 2021 May 12;8.

5. Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, et al. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nat Commun. 2019 Dec 1;10.

6. Dutrow E V., Serpell JA, Ostrander EA. Domestic dog lineages reveal genetic drivers of behavioral diversification. Cell. 2022 Dec 8;185(25):4737-4755.e18.

7. Isaksen KE, Linney L, Williamson H, Cave NJ, Beausoleil NJ, Norman EJ, et al. TeamMate: A longitudinal study of New Zealand working farm dogs. I. Methods, population characteristics and health on enrolment. BMC Vet Res. 2020 Feb 17;16(1).

8. Zink C, Schlehr MR. Working Dog Structure: Evaluation and Relationship to Function. Front Vet Sci. 2020 Oct 20;7.

9. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, et al. A simple genetic architecture underlies morphological variation in dogs. PLoS Biol. 2010;8(8):49–50.

10. Chen FL, Zimmermann M, Hekman JP, Lord KA, Logan B, Russenberger J, et al. Advancing Genetic Selection and Behavioral Genomics of Working Dogs Through Collaborative Science. Front Vet Sci. 2021 Sep 6;8.

11. Leighton EA, Holle D, Biery DN, Gregor TP, McDonald-Lynch MB, Wallace ML, et al. Genetic improvement of hip-extended scores in 3 breeds of guide dogs using estimated breeding values: Notable progress but more improvement is needed. PLoS One. 2019 Feb 1;14(2).

12. Brito LF, Bedere N, Douhard F, Oliveira HR, Arnal M, Peñagaricano F, et al. Review: Genetic selection of high-yielding dairy cattle toward sustainable farming systems in a rapidly changing world. Animal. 2021 Dec 1;15.

13. Cave NJ, Bridges JP, Cogger N, Farman RS. A survey of diseases of working farm dogs in New Zealand. N Z Vet J. 2009;57(6):305–12.

14. Dalton C. Heading dogs, huntaways and all-purpose dogs – Te Ara Encyclopedia of New Zealand [Internet]. 2009 [cited 2024 Aug 29]. Available from: https://teara.govt.nz/en/farm-dogs/page-1

15. Sheard H. Demographics and Health of New Zealand Working Farm Dogs: A survey of dogs on sheep and beef farms in New Zealand in 2009. [Palmerston North, NZ]: Institute of Veterinary, Animal and Biomedical Sciences Massey University; 2014.

16. Hughes PL. Hip Dysplasia in the New Zealand Huntaway and Heading Dog. N Z Vet J. 2001;49(4):138–41.

17. Munday JS, Dyer CB, Hartman AC, Orbell GMB. A possible predisposition to dilated cardiomyopathy in Huntaway dogs. N Z Vet J. 2006;54(5):231–4.

18. Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GRS, Creixell P, Karchin R, et al. Computational approaches to identify functional genetic variants in cancer genomes. Nat Methods. 2013 Aug;10(8):723–9.

19. Bohry D, Ramos HCC, dos Santos PHD, Boechat MSB, Arêdes FAS, Pirovani AAV, et al. Discovery of SNPs and InDels in papaya genotypes and its potential for marker assisted selection of fruit quality traits. Sci Rep. 2021 Dec 1;11.

20. Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968 Feb 17;217:624–6.

21. Marian AJ. Clinical Interpretation and Management of Genetic Variants. JACC Basic Transl Sci. 2020 Oct 1;5(10):1029–42.

22. Vihinen M. Systematics for types and effects of DNA variations. BMC Genomics. 2018 Dec 28;19.

23. Hoeppner MP, Lundquist A, Pirun M, Meadows JRS, Zamani N, Johnson J, et al. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. PLoS One. 2014 Mar 13;9(3).

24. Davis SR, Ward HE, Kelly V, Palmer D, Ankersmit-Udy AE, Lopdell TJ, et al. Screening for phenotypic outliers identifies an unusually low concentration of a β-lactoglobulin B protein isoform in bovine milk caused by a synonymous SNP. Genetics Selection Evolution. 2022 Dec 1;54.

25. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015 Oct 15;24(R1):R102–10.

26. Teng J, Gao Y, Yin H, Bai Z, Liu S, Zeng H, et al. A compendium of genetic regulatory effects across pig tissues. Nat Genet. 2024 Jan 1;56:112–23.

27. Palazzo AF, Lee ES. Non-coding RNA: What is functional and what is junk? Front Genet. 2015 Jan 26;6.

28. Murphy SC, Evans JM, Tsai KL, Clark LA. Length variations within the Merle retrotransposon of canine PMEL: Correlating genotype with phenotype. Mob DNA. 2018 Aug 3;9.

29. Wang W, Kirkness EF. Short interspersed elements (SINEs) are a major source of canine genomic diversity. Genome Res. 2005 Dec;15(12):1798–808.

30. Canis lupus familiaris Annotation Report - NCBI - NLM [Internet]. 2017 [cited 2025 May 16]. Available from: https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/Canis_lupus_familiaris/106/

31. Canis lupus familiaris genome assembly UU_Cfam_GSD_1.0 - NCBI - NLM [Internet]. [cited 2025 May 16]. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_011100685.1/

32. Littlejohn MD, Henty KM, Tiplady K, Johnson T, Harland C, Lopdell T, et al. Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. Nat Commun. 2014;5.

33. Gayon J. From Mendel to epigenetics: History of genetics. C R Biol. 2016 Jun 2;339(7–8):225–30.

34. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S A. 2011 Nov 1;108(44):18026–31.

35. Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW. Genomics advances the study of inbreeding depression in the wild. Evol Appl. 2016 Dec 1;9(10):1205–18.

36. Vasiliadis D, Metzger J, Distl O. Demographic assessment of the Dalmatian dog – effective population size, linkage disequilibrium and inbreeding coefficients. Canine Med Genet. 2020 Dec;7.

37. Windig JJ, Doekes HP. Limits to genetic rescue by outcross in pedigree dogs. Journal of Animal Breeding and Genetics. 2018 Jun 1;135(3):238–48.

38. Pei XM, Yeung MHY, Wong ANN, Tsang HF, Yu ACS, Yim AKY, et al. Targeted Sequencing Approach and Its Clinical Applications for the Molecular Diagnosis of Human Diseases. Cells. 2023 Feb 2;12(3).

39. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan 1;107(1):1–8.

40. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463–7.

41. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020 Feb 7;21.

42. Han H, Lee HJ, Kim KS, Chung J, Na HS. Comparison of the performance of MiSeq and NovaSeq in oral microbiome study. J Oral Microbiol. 2024;16(1).

43. Sanford Kobayashi E, Batalov S, Wenger AM, Lambert C, Dhillon H, Hall RJ, et al. Approaches to long-read sequencing in a clinical setting to improve diagnostic rate. Sci Rep. 2022 Oct 9;12.

44. Zverinova S, Guryev V. Variant calling: Considerations, practices, and developments. Hum Mutat. 2022 Aug 1;43(8):976–85.

45. Bush SJ. Generalizable characteristics of false-positive bacterial variant calls. Microb Genom. 2021;7(8).

46. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug;27(15):2156–8.

47. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv. 2018 Jul 24;

48. Van der Auwera G, O'Connor B. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. 1st ed. O'Reilly Media. Sebastopol, CA: O'Reilly Media; 2020.

49. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb;10(2).

50. Lefouili M, Nam K. The evaluation of Bcftools mpileup and GATK HaplotypeCaller for variant calling in non-human species. Sci Rep. 2022 Jul 5;12.

51. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv. 2013 May 26;

52. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv: Genomics. 2012 Jul 20;

53. Liu J, Shen Q, Bao H. Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens. PLoS One. 2022 Jan 31;17(1).

54. Zhao S, Agafonov O, Azab A, Stokowy T, Hovig E. Accuracy and efficiency of germline variant calling pipelines for human genome data. Sci Rep. 2020 Nov 19;10.

55. Lin YL, Chang PC, Hsu C, Hung MZ, Chien YH, Hwu WL, et al. Comparison of GATK and DeepVariant by trio sequencing. Sci Rep. 2022 Feb 2;12.

56. (How to) Filter variants either with VQSR or by hard-filtering – GATK [Internet]. [cited 2024 Aug 27]. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360035531112--How-to-Filter-variants-either-with-VQSR-or-by-hard-filtering

57. GATK TUTORIAL :: Variant Callset Evaluation & Filtering. 2016.

58. Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen JB, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Bioinformatics. 2014 May 2;15.

59. Salzberg SL. Next-generation genome annotation: We still struggle to get it right. Genome Biol. 2019 May 16;20.

60. Ejigu GF, Jung J. Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology (Basel). 2020 Sep 18;9(9).

61. Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V. Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet. 2023 Sep 10;55:1512–22.

62. Liu Y, Yeung WSB, Chiu PCN, Cao D. Computational approaches for predicting variant impact: An overview from resources, principles to applications. Front Genet. 2022 Sep 29;13.

63. Horne J, Shukla D. Recent Advances in Machine Learning Variant Effect Prediction Tools for Protein Engineering. Ind Eng Chem Res. 2022 May 18;61(19):6235–45.

64. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Jun 6;17.

65. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012 Apr 1;6(2):80–92.

66. Cheng J, Novati G, Pan J, Bycroft C, Žemgulyte A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science (1979). 2023 Sep 19;381(6664).

67. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Jul 15;596:583–9.

68. Marsh JA, Teichmann SA. Predicting pathogenic protein variants. Science (1979). 2023 Sep 19;381(6664):1284–5.

69. Frankish A, Uszczynska B, Ritchie GRS, Gonzalez JM, Pervouchine D, Petryszak R, et al. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics. 2015 Jun 18;16(Suppl 8).

70. Choi J, Tantisira KG, Duan QL. Whole genome sequencing identifies high-impact variants in well-known pharmacogenomic genes. Pharmacogenomics Journal. 2019 Apr;19(2):127–35.

71. Grimm DG, Azencott CA, Aicheler F, Gieraths U, Macarthur DG, Samocha KE, et al. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. Hum Mutat. 2015 May 1;36(5):513–23.

72. Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, et al. RNA-seq data science: From raw data to effective interpretation. Front Genet. 2023 Mar 13;14.

73. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

74. Lopdell T, Tiplady K, Littlejohn MD. Using RNAseq data to improve genomic selection in dairy cattle. In: World Congress on Genetics to Livestock Production. 2018.

75. Chen G, Shi T, Shi L. Characterizing and annotating the genome using RNA-seq data. Sci China Life Sci. 2017 Feb;60(2):116–25.

76. Pérez-Enciso M, Rincón JC, Legarra A. Sequence- vs. chip-assisted genomic selection: Accurate biological information is advised. Genetics Selection Evolution. 2015 May 9;47.

77. Baker L, Muir P, Sample SJ. Genome-wide association studies and genetic testing: Understanding the science, success, and future of a rapidly developing field. Vol. 255, Journal of the American Veterinary Medical Association. American Veterinary Medical Association; 2019. p. 1126–36.

78. Gaut BS, Long AD. The Lowdown on Linkage Disequilibrium. 2003 Jul;15(7):1502–6.

79. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A. 2011 Dec 6;108(49):E1293–301.

80. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet. 2003 Aug 1;4:587–97.

81. Liu N, Zhang K, Zhao H. Haplotype-Association Analysis. Adv Genet. 2008;60:335–405.

82. Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: What do we gain? European Journal of Human Genetics. 2001 Apr;9(4):291–300.

83. Browning SR, Browning BL. Haplotype phasing: Existing methods and new developments. Nat Rev Genet. 2011 Oct;12:703–14.

84. Bansal V. Integrating read-based and population-based phasing for dense and accurate haplotyping of individual genomes. Bioinformatics. 2019 Jul 5;35(14):i242–8.

85. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018 Sep 6;103(3):338–48.

86. Treccani M, Locatelli E, Patuzzo C, Malerba G. A broad overview of genotype imputation: Standard guidelines, approaches, and future investigations in genomic association studies. Biocell. 2023 May 22;47(6):1225–41.

87. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 2021 Oct 7;108(10):1880–90.

88. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021 Aug 26;1.

89. Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. Nat Commun. 2020 Nov 19;11.

90. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International Journal of Methods in Psychiatric Researche1608. 2018 Jun;27(2):e1608.

91. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-wide association analysis and post-analytic interrogation. Stat Med. 2015 Dec 10;34(28):3769–92.

92. Guindo-Martínez M, Amela R, Bonàs-Guarch S, Puiggròs M, Salvoro C, Miguel-Escalada I, et al. The impact of non-additive genetic associations on age-related complex diseases. Nature Communication. 2021 Apr 23;12.

93. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and interpretation of SNP-based heritability. Nat Genet. 2017 Aug 30;49(9):1304–10.

94. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

95. Nguyen AK, Blacksmith MS, Kidd JM. Duplications and Retrogenes Are Numerous and Widespread in Modern Canine Genomic Assemblies. Genome Biol Evol. 2024 Jul 1;16(7).

96. Jagannathan V, Hitte C, Kidd JM, Masterson P, Murphy TD, Emery S, et al. Dog10k_boxer_tasha_1.0: A long-read assembly of the dog reference genome. Genes (Basel). 2021 May 30;12(6):847.

97. Halo J V., Pendleton AL, Shen F, Doucet AJ, Derrien T, Hitte C, et al. Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. Proc Natl Acad Sci U S A. 2021 Mar 16;118(11).

98. Wang C, Wallerman O, Arendt ML, Sundström E, Karlsson Å, Nordin J, et al. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. Commun Biol. 2021 Feb 10;4.

99. Son H, Borris M, Aldonza D, Nam AR, Lee KH, Lee JW, et al. Integrative mapping of the dog epigenome: Reference annotation for comparative intertissue and cross-species studies. Sci Adv. 2023 Jul 5;9(27).

100. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005 Dec 8;438:803–19.

101. Meadows JRS, Kidd JM, Wang GD, Parker HG, Schall PZ, Bianchi M, et al. Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of demography, genome function and architecture. Genome Biol. 2023 Jul 15;24.

102. Baker L, Muir P, Sample SJ. Genome-wide association studies and genetic testing: Understanding the science, success, and future of a rapidly developing field. J Am Vet Med Assoc. 2019 Nov 15;255(10):1126–36.

103. Zhou T, Pu SY, Zhang SJ, Zhou QJ, Zeng M, Lu JS, et al. Dog10K: an integrated Dog10K database summarizing canine multi-omics. Nucleic Acids Res. 2024 Jan 6;53(D1):D939–47.

104. Wang GD, Larson G, Kidd JM, Vonholdt BM, Ostrander EA, Zhang YP. Dog10K: The International Consortium of Canine Genome Sequencing. Natl Sci Rev. 2019 May 29;6(4):611–3.

105. Jagannathan V, Drögemüller C, Leeb T, Aguirre G, André C, Bannasch D, et al. A comprehensive biomedical variant catalogue based on whole genome sequences of 582 dogs and eight wolves. Anim Genet. 2019 Dec;50(6):695–704.

106. Lenffer J, Nicholas FW, Castle K, Rao A, Gregory S, Poidinger M, et al. OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D599–601.

107. Conroy MC, Lacey B, Bešević J, Omiyale W, Feng Q, Effingham M, et al. UK Biobank: a globally important resource for cancer research. Br J Cancer. 2023 Feb 16;128:519–27.

108. Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Aken B, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020 Jul 30;583:699–710.

109. Megquier K, Genereux DP, Hekman J, Swofford R, Turner-Maier J, Johnson J, et al. Barkbase: Epigenomic annotation of canine genomes. Genes (Basel). 2019 Jun 7;10(6).

110. Momozawa Y, Merveille AC, Battaille G, Wiberg M, Koch J, Willesen JL, et al. Genome wide association study of 40 clinical measurements in eight dog breeds. Sci Rep. 2020 Apr 16;10.

111. Shan S, Xu F, Brenig B. Genome-Wide Association Studies Reveal Neurological Genes for Dog Herding, Predation, Temperament, and Trainability Traits. Front Vet Sci. 2021 Jul 21;8.

112. Yogalingam G, Pollard T, Gliddon B, Jolly RD, Hopwood JJ. Identification of a Mutation Causing Mucopolysaccharidosis Type IIIA in New Zealand Huntaway Dogs. Genomics. 2002 Feb;79(2):150–3.

113. Gedye K, Poole-Crowe E, Shepherd M, Wilding A, Parton K, Lopez-Villalobos N, et al. Prevalence of the ABCB1-1Δ gene mutation in a sample of New Zealand Huntaway dogs. N Z Vet J. 2023 May;71(3):133–6.

114. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884–90.

115. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013 Oct 15;43(1110).

116. Pedersen BS, Brown JM, Dashnow H, Wallace AD, Velinder M, Tristani-Firouzi M, et al. Effective variant filtering and expected candidate variant yield in studies of rare human disease. NPJ Genom Med. 2021 Jul 15;6.

117. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 2012 Mar 15;3.

118. SnpEff & SnpSift (Input & output files) - GitHub [Internet]. [cited 2025 Jun 9]. Available from: https://pcingola.github.io/SnpEff/snpeff/inputoutput/#vcf-output

119. SRA Toolkit Development Team. SRA-Toolkit - NCBI [Internet]. [cited 2024 Jun 24]. Available from: https://hpc.nih.gov/apps/sratoolkit.html

120. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.

121. Smith F, Lopdell T, Stephen M, Henry M, Dittmer K, Hunt H, et al. Survey of Mendelian-effect functional variants in New Zealand Huntaway and Heading dog breeds. TechRxiv. 2025 Feb 15;

122. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Jul 25;81(3):559–75.

123. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010 Jun 20;42:565–9.

124. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 2015 Feb 25;4.

125. Hawkes G, Chundru K, Jackson L, Patel KA, Murray A, Wood AR, et al. Whole-genome sequencing analysis identifies rare, large-effect noncoding variants and regulatory regions associated with circulating protein levels. Nat Genet. 2025 Feb 24;57:626–34.

126. Niehoff T, Pook T, Gholami M, Beissinger T. Imputation of low-density marker chip data in plant breeding: Evaluation of methods based on sugar beet. Plant Genome. 2022 Dec;15(4).

127. Jiang Y, Song H, Gao H, Zhang Q, Ding X. Exploring the optimal strategy of imputation from SNP array to whole-genome sequencing data in farm animals. Front Genet. 2022 Aug 26;13.

128. Leroy G, Verrier E, Meriaux JC, Rognon X. Genetic diversity of dog breeds: Within-breed diversity comparing genealogical and molecular data. Anim Genet. 2009 Jun;40(3):323–32.

129. Robitzsch A. Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods. Front Educ (Lausanne). 2020 Oct 8;5.

130. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014 Jan 29;46:100–6.

131. D. Turner S. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J Open Source Softw. 2018 May 19;3(25):731.

132. Adebiyi E, Adam Y, Samtal C, Brandenburg J tristan, Falola O. Performing post-genome-wide association study analysis: Overview, challenges and recommendations. F1000Res. 2021 Oct 4;10.

133. Wang J. Marker-based estimates of relatedness and inbreeding coefficients: An assessment of current methods. J Evol Biol. 2014 Jan 21;27(3):518–30.

134. Takasuga A. PLAG1 and NCAPG-LCORL in livestock. Animal Science Journal. 2016 Feb;87(2):159–67.

135. Hartman ML, Czyz M. MITF in melanoma: Mechanisms behind its expression and activity. Cellular and Molecular Life Sciences. 2015 Apr;72(7):1249–60.

136. Jiang Y, Jiang Y, Wang S, Zhang Q, Ding X. Optimal sequencing depth design for whole genome re-sequencing in pigs. BMC Bioinformatics. 2019 Nov 8;20.

137. Kishikawa T, Momozawa Y, Ozeki T, Mushiroda T, Inohara H, Kamatani Y, et al. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. Sci Rep. 2019 Feb 11;9.

138. Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data. 2020 Nov 17;7.

139. Talenti A, Powell J, Hemmink JD, Cook EAJ, Wragg D, Jayaraman S, et al. A cattle graph genome incorporating global breed diversity. Nat Commun. 2022 Feb 17;13.

140. Ibing S, Michels BE, Mosdzien M, Meyer HR, Feuerbach L, Korner C. On the impact of batch effect correction in TCGA isomiR expression data. NAR Cancer. 2021 Mar 11;3(1).

141. Mahmood K, Jung CH, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: Implications for discovery and diagnostics. Hum Genomics. 2017 May 16;11(1).

142. Wang C, Wallerman O, Arendt ML, Sundström E, Karlsson Å, Nordin J, et al. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. Commun Biol. 2021 Feb 10;4.

143. Sundaram A, Tengs T, Grimholt U. Issues with RNA-seq analysis in non-model organisms: A salmonid example. Dev Comp Immunol. 2017 Oct;75:38–47.

144. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Jan 26;17.

145. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57–63.

146. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012 Jul;13(7):505–16.

147. Brancalion L, Haase B, Wade CM. Canine coat pigmentation genetics: a review. Anim Genet. 2022 Feb;53(1):3–34.

148. Newton JM, Wilkie AL, He L, Jordan SA, Metallinos DL, Holmes NG, et al. Melanocortin 1 receptor variation in the domestic dog. Mammalian Genome. 2000 Jan;11(1):24–30.

149. Anderson H, Honkanen L, Ruotanen P, Mathlin J, Donner J. Comprehensive genetic testing combined with citizen science reveals a recently characterized ancient MC1R mutation associated with partial recessive red phenotypes in dog. Canine Med Genet. 2020 Nov 5;7.

150. Schmutz SM, Berryere TG, Ellinwood NM, Kerns JA, Barsh GS. MCIR Studies in Dogs with Melanistic Mask or Brindle Patterns. Journal of Heredity. 2003 Jan;94(1):69–73.

151. Schmutz SM, Berryere TG, Goldfinch AD. TYRP1 and MC1R genotypes and their effects on coat color in dogs. Mammalian Genome. 2002;13(7):380–7.

152. Jancuskova T, Langevin M, Pekova S. TYRP1:c.555T>G is a recurrent mutation found in Australian Shepherd and Miniature American Shepherd dogs. Anim Genet. 2018 Oct;49(5):500–1.

153. Wright HE, Schofield E, Mellersh CS, Burmeister LM. A novel TYRP1 variant is associated with liver and tan coat colour in Lancashire Heelers. Anim Genet. 2019 Dec;50(6):783.

154. Kerns JA, Cargill EJ, Clark LA, Candille SI, Berryere TG, Olivier M, et al. Linkage and segregation analysis of black and brindle coat color in domestic dogs. Genetics. 2007 Jul;176(3):1679–89.

155. Candille SI, Kaelin CB, Cattanach BM, Yu B, Thompson DA, Nix MA, et al. A β-defensin mutation causes black coat color in domestic dogs. Science (1979). 2007 Oct 18;318(5855):1418–23.

156. Hrckova Turnova E, Bielikova M, Kostal V, Turna J, Dudas A. Occurrence of the dominant black KB allele of CBD103 in German Shepherd Dogs. Anim Genet. 2022 Apr;53(2):230–1.

157. Ollivier M, Tresset A, Hitte C, Petit C, Hughes S, Gillet B, et al. Evidence of Coat Color Variation Sheds New Light on Ancient Canids. PLoS One. 2013 Oct 2;8(10).

158. Hédan B, Cadieu E, Botherel N, Citres CD de, Letko A, Rimbault M, et al. Identification of a missense variant in MFSD12 involved in dilution of phaeomelanin leading to white or cream coat color in dogs. Genes (Basel). 2019 May 21;10(5):386.

159. Sponenberg DP, Rothschild MF. The Genetics of the Dog. Ruvinsky A, Sampson J, editors. CABI; 2001. 61–85 p.

160. Slavney AJ, Kawakami T, Jensen MK, Nelson TC, Sams AJ, Boyko AR. Five genetic variants explain over 70% of hair coat pheomelanin intensity variation in purebred and mixed breed domestic dogs. PLoS One. 2021 May 27;16(5).

161. Kerns JA, Newton J, Berryere TG, Rubin EM, Cheng JF, Schmutz SM, et al. Characterization of the dog Agouti gene and a nonagouti mutation in German Shepherd Dogs. Mammalian Genome. 2004 Oct;15(10):798–808.

162. Berryere TG, Kerns JA, Barsh GS, Schmutz SM. Association of an Agouti allele with fawn or sable coat color in domestic dogs. Mammalian Genome. 2005 Apr;16(4):262–72.

163. Drögemüller C, Philipp U, Haase B, Günzel-Apel AR, Leeb T. A noncoding melanophilin gene (MLPH) SNP at the splice donor of exon 1 represents a candidate causal mutation for coat color dilution in dogs. Journal of Heredity. 2007;98(5):468–73.

164. Van Buren SL, Minor KM, Grahn RA, Mickelson JR, Grahn JC, Malvick J, et al. A third MLPH variant causing coat color dilution in dogs. Genes (Basel). 2020 Jun 10;11(6).

165. Bauer A, Kehl A, Jagannathan V, Leeb T. A novel MLPH variant in dogs with coat colour dilution. Anim Genet. 2018 Feb;49(1):94–7.

166. Hayward JJ, Castelhano MG, Oliveira KC, Corey E, Balkman C, Baxter TL, et al. Complex disease and phenotype mapping in the domestic dog. Nat Commun. 2016 Jan 22;7.

167. Plassais J, vonHoldt BM, Parker HG, Carmagnini A, Dubos N, Papa I, et al. Natural and human-driven selection of a single non-coding body size variant in ancient and modern canids. Current Biology. 2022 Feb 28;32(4):889-897.e9.

168. Park K, Kang J, Subedi KP, Ha JH, Park C. Canine polydactyl mutations with heterogeneous origin in the conserved intronic sequence of LMBR1. Genetics. 2008 Aug;179(4):2163–72.

169. Gurnett CA, Bowcock AM, Dietz FR, Morcuende JA, Murray JC, Dobbs MB. Two novel point mutations in the long-range SHH enhancer in three families with triphalangeal thumb and preaxial polydactyly. Am J Med Genet A. 2007 Jan 1;143A(1):27–32.

170. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 2003 Jul 15;12(14):1725–35.

171. Salmela E, Niskanen J, Arumilli M, Donner J, Lohi H, Hytönen MK. A novel KRT71 variant in curly-coated dogs. Anim Genet. 2019 Feb;50(1):101–4.

172. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, Parker HG, et al. Coat variation in the domestic dog is governed by variants in three genes. Science (1979). 2009 Oct 2;326(5949):150–3.

173. Gandolfi B, Alhaddad H, Joslin SEK, Khan R, Filler S, Brem G, et al. A splice variant in KRT71 is associated with curly coat phenotype of Selkirk Rex cats. Sci Rep. 2013;3:2000.

174. Bauer A, Hadji Rasouliha S, Brunner MT, Jagannathan V, Bucher I, Bannoehr J, et al. A second KRT71 allele in curly coated dogs. Anim Genet. 2019 Feb;50(1):97–100.

175. Housley DJE, Venta PJ. The long and the short of it: Evidence that FGF5 is a major determinant of canine 'hair'-itability. Anim Genet. 2006 Aug;37(4):309–15.

176. Carrion EA, Moses MM, Behringer RR. FGF5. Differentiation. 2024 Sep;139.

177. Dierks C, Mömke S, Philipp U, Distl O. Allelic heterogeneity of FGF5 mutations causes the long-hair phenotype in dogs. Anim Genet. 2013 Aug;44(4):425–31.

178. Meyers-Wallen VN, Boyko AR, Danko CG, Grenier JK, Mezey JG, Hayward JJ, et al. XX Disorder of Sex Development is associated with an insertion on chromosome 9 and downregulation of RSPO1 in dogs (Canis lupus familiaris). PLoS One. 2017 Oct 20;12(10).

179. Letko A, Minor KM, Friedenberg SG, Shelton GD, Salvador JP, Mandigers PJJ, et al. A cntnap1 missense variant is associated with canine laryngeal paralysis and polyneuropathy. Genes (Basel). 2020 Nov 27;11(12):1426.

180. Alves L, Hülsmeyer V, Jaggy A, Fischer A, Leeb T, Drögemüller M. Polymorphisms in the ABCB1 gene in phenobarbital responsive and resistant idiopathic epileptic border collies. J Vet Intern Med. 2011 May;25(3):484–9.

181. Mealey KL, Owens JG, Freeman E. Canine and feline P-glycoprotein deficiency: What we know and where we need to go. J Vet Pharmacol Ther. 2023 Jan 1;46(1):1–16.

182. Mealey KL, Northrup NC, Bentjen SA. Increased toxicity of P-glycoprotein-substrate chemotherapeutic agents in a dog with the MDR1 deletion mutation associated with ivermectin sensitivity. J Am Vet Med Assoc. 2003 Nov 15;223(10):1453–5.

183. Gagliardo T, Gandini G, Gallucci A, Menchetti M, Bianchi E, Turba ME, et al. ABCB1 c.-6-180T > G polymorphism and clinical risk factors in a multi-breed cohort of dogs with refractory idiopathic epilepsy. Veterinary Journal. 2019 Nov;253.

184. Fieten H, Gill Y, Martin AJ, Concilli M, Dirksen K, Van Steenbeek FG, et al. The Menkes and Wilson disease genes counteract in copper toxicosis in Labrador retrievers: A new canine model for copper-metabolism disorders. DMM Disease Models and Mechanisms. 2016 Jan;9(1):25–38.

185. Haywood S, Swinburne J, Schofield E, Constantino-Casas F, Watson P. Copper toxicosis in Bedlington terriers is associated with multiple independent genetic variants. Veterinary Record. 2023;193(4).

186. Langlois DK, Nagler BSM, Smedley RC, Yang YT, Yuzbasiyan-Gurkan V. ATP7A, ATP7B, and RETN genotypes in Labrador Retrievers with and without copper-associated hepatopathy. J Am Vet Med Assoc. 2022 Apr 27;260(14):1–8.

187. Wu X, Den Boer ER, Vos-Loohuis M, van Steenbeek FG, Monroe GR, Nijman IJ, et al. Investigation of genetic modifiers of copper toxicosis in labrador retrievers. Life. 2020 Oct 31;10(11):266.

188. Tenmizu D, Endo Y, Noguchi K, Kamimura H. Identification of the novel canine CYP1A2 1117 C > T SNP causing protein deletion. Xenobiotica. 2004 Sep;34(9):835–46.

189. Thorn CF, Aklillu E, Klein TE, Altman RB. PharmGKB summary: Very important pharmacogene information for CYP1A2. Pharmacogenet Genomics. 2012 Jan 1;22(1):73–7.

190. Mise M, Hashizume T, Matsumoto S, Terauchi Y, Fujii T. Identification of non-functional allelic variant of CYP1A2 in dogs. Pharmacogenetics. 2004 Nov;14(11):769–73.

191. Tervahauta T, Uutela P, Koskinen M. Metabolism of ropinirole is mediated by several canine CYP enzymes. Vet Med Sci. 2023 Jul;9(4):1584–91.

192. Tenmizu D, Noguchi K, Kamimura H, Ohtani H, Sawada Y. The canine CYP1A2 deficiency polymorphism dramatically affects the pharmacokinetics of 4-cyclohexyl-1-ethyl-7-methylpyrido[2,3-d]-pyrimidine-2-(1H) -one (YM-64227), a phosphodiesterase type 4 inhibitor. Drug Metabolism and Disposition. 2006 May;34(5):800–6.

193. Uno Y, Morikuni S, Shiraishi M, Asano A, Murayama N, Yamazaki H. Novel Cytochrome P450 2C94 Functionally Metabolizes Diclofenac and Omeprazole in Dogs. Drug Metabolism and Disposition. 2023 May;51(5):637–44.

194. Metzger J, Pfahler S, Distl O. Variant detection and runs of homozygosity in next generation sequencing data elucidate the genetic background of Lundehund syndrome. BMC Genomic. 2016 Aug 2;17(1).

195. Melis C, Billing AM, Wold PA, Ludington WB. Gut microbiome dysbiosis is associated with host genetics in the Norwegian Lundehund. Front Microbiol. 2023 Jun;14.

196. Gast AC, Metzger J, Tipold A, Distl O. Genome-wide association study for hereditary ataxia in the Parson Russell Terrier and DNA-testing for ataxia-associated mutations in the Parson and Jack Russell Terrier. BMC Vet Res. 2016 Oct 10;12(1).

197. Gilliam D, O'Brien DP, Coates JR, Johnson GS, Johnson GC, Mhlanga-Mutangadura T, et al. A homozygous KCNJ10 mutation in jack russell terriers and related breeds with spinocerebellar ataxia with myokymia, seizures, or both. J Vet Intern Med. 2014;28(3):871–7.

198. Owczarek-Lipska M, Jagannathan V, Drögemüller C, Lutz S, Glanemann B, Leeb T, et al. A Frameshift Mutation in the Cubilin Gene (CUBN) in Border Collies with Imerslund-Gräsbeck Syndrome (Selective Cobalamin Malabsorption). PLoS One. 2013 Apr 16;8(4).

199. Fyfe JC, Hemker SL, Venta PJ, Fitzgerald CA, Outerbridge CA, Myers SL, et al. An exon 53 frameshift mutation in CUBN abrogates cubam function and causes Imerslund-Gräsbeck syndrome in dogs. Mol Genet Metab. 2013 Aug;109(4):390–6.

200. Erles K, Mugford A, Barfield D, Leeb T, Kook PH. Systemic Scedosporium prolificans infection in an 11-month-old Border collie with cobalamin deficiency secondary to selective cobalamin malabsorption (canine Imerslund-Gräsbeck syndrome). Journal of Small Animal Practice. 2018 Apr;59(4):253–6.

201. Drögemüller M, Jagannathan V, Howard J, Bruggmann R, Drögemüller C, Ruetten M, et al. A frameshift mutation in the cubilin gene (CUBN) in Beagles with Imerslund-Gräsbeck syndrome (selective cobalamin malabsorption). Anim Genet. 2014 Feb;45(1):148–50.

202. Fyfe JC, Hemker SL, Frampton A, Raj K, Nagy PL, Gibbon KJ, et al. Inherited selective cobalamin malabsorption in Komondor dogs associated with a CUBN splice site variant. BMC Vet Res. 2018 Dec 27;14(1).

203. Mizukami K, Yabuki A, Kohyama M, Kushida K, Rahman MM, Uddin MM, et al. Molecular prevalence of multiple genetic disorders in Border collies in Japan and recommendations for genetic counselling. The Veterinary Journal. 2016 Aug;214:21–3.

204. Aronovich EL, Carmichael KP, Morizono H, Koutlas IG, Deanching M, Hoganson G, et al. Canine Heparan Sulfate Sulfamidase and the Molecular Pathology Underlying Sanfilippo Syndrome Type A in Dachshunds. Genomics. 2000 Aug 15;68(1):80–4.

205. Gentilini F, Turba ME. Two novel real-time PCR methods for genotyping the von Willebrand disease type I mutation in Doberman Pinscher dogs. Veterinary Journal. 2013 Aug;197(2):457–60.

206. Venta PJ, Brewer GJ, Yuzbasiyan-Gurkan V, Schall WD. DNA Encoding Canine von Willebrand Factor and Methods of Use. United States; US 6780583 B1, 2004.

207. Crespi JA, Barrientos LS, Giovambattista G. von Willebrand disease type 1 in Doberman Pinscher dogs: genotyping and prevalence of the mutation in the Buenos Aires region, Argentina. Journal of Veterinary Diagnostic Investigation. 2018 Mar 1;30(2):310–4.

208. Segert JH, Seidel JM, Wurzer WJ, Geretschlaeger AM. vWDI is inherited in an autosomal dominant manner with incomplete penetrance, in the Kromfohrländer breed. Canine Genet Epidemiol. 2019 May 16;6.

209. Awano T, Johnson GS, Wade CM, Katz ML, Johnson GC, Taylor JF, et al. Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. Proc Natl Acad Sci U S A. 2009 Feb 24;106(8):2794–9.

210. Berdyński M, Miszta P, Safranow K, Andersen PM, Morita M, Filipek S, et al. SOD1 mutations associated with amyotrophic lateral sclerosis analysis of variant severity. Sci Rep. 2022 Jan 7;12.

211. Tsai KL, Noorai RE, Starr-Moss AN, Quignon P, Rinz CJ, Ostrander EA, et al. Genome-wide association studies for multiple diseases of the German Shepherd Dog. Mammalian Genome. 2012 Feb;23(1–2):203–11.

212. Draper ACE, Wilson Z, Maile C, Faccenda D, Campanella M, Piercy RJ. Species-specific consequences of an E40K missense mutation in superoxide dismutase 1 (SOD1). FASEB Journal. 2020 Jan 1;34(1):458–73.

213. Tanaka N, Kimura S, Kamatari YO, Nakata K, Kobatake Y, Inden M, et al. In vitro evidence of propagation of superoxide dismutase-1 protein aggregation in canine degenerative myelopathy. Veterinary Journal. 2021 Aug;274.

214. Maki S, Islam MS, Itoh T, Nurimoto M, Yabuki A, Furusawa Y, et al. Molecular Epidemiological Survey for Degenerative Myelopathy in German Shepherd Dogs in Japan: Allele Frequency and Clinical Progression Rate. Animals. 2022 Jun 27;12(13).

215. Guo J, Johnson GS, Brown HA, Provencher ML, da Costa RC, Mhlanga-Mutangadura T, et al. A CLN8 nonsense mutation in the whole genome sequence of a mixed breed dog with neuronal ceroid lipofuscinosis and Australian shepherd ancestry. Mol Genet Metab. 2014 Aug;112(4):302–9.

216. Guo J, Johnson GS, Cook J, Harris OK, Mhlanga-Mutangadura T, Schnabel RD, et al. Neuronal ceroid lipofuscinosis in a German Shorthaired Pointer associated with a previously reported CLN8 nonsense variant. Mol Genet Metab Rep. 2019 Oct 21;21.

217. Katz ML, Khan S, Awano T, Shahid SA, Siakotos AN, Johnson GS. A mutation in the CLN8 gene in English Setter dogs with neuronal ceroid-lipofuscinosis. Biochem Biophys Res Commun. 2005 Feb 11;327(2):541–7.

218. Warrier V, Vieira M, Mole SE. Genetic basis and phenotypic correlations of the neuronal ceroid lipofusinoses. Biochim Biophys Acta Mol Basis Dis. 2013 Nov;1832(11):1827–30.

219. Hirz M, Drögemüller M, Schänzer A, Jagannathan V, Dietschi E, Goebel HH, et al. Neuronal ceroid lipofuscinosis (NCL) is caused by the entire deletion of CLN8 in the Alpenländische Dachsbracke dog. Mol Genet Metab. 2017 Mar;120(3):269–77.

220. Donner J, Freyer J, Davison S, Anderson H, Blades M, Honkanen L, et al. Genetic prevalence and clinical relevance of canine Mendelian disease variants in over one million dogs. PLoS Genet. 2023 Feb 27;19(2).

221. Shearman JR, Wilton AN. A canine model of cohen syndrome: Trapped neutrophil syndrome. BMC Genomics. 2011 May 23;12.

222. Suciu A, Starybrat D, Gil-Morales C, Matson H, Jepson R, Williams M, et al. Clinical findings, treatment and outcome of trapped neutrophil syndrome in Border Collies: 12 cases (2011-2022). Journal of Small Animal Practice. 2024 Jul;65(7):560–8.

223. Domené S, Scaglia PA, Gutiérrez ML, Domené HM. Applying bioinformatic platforms, in vitro, and in vivo functional assays in the characterization of genetic variants in the GH/IGF pathway affecting growth and development. Cells. 2021 Aug 12;10(8):2063.

224. Ahonen SJ, Arumilli M, Lohi H. A CNGB1 Frameshift Mutation in Papillon and Phalène Dogs with Progressive Retinal Atrophy. PLoS One. 2013 Aug 28;8(8).

225. Mäkeläinen S, Gòdia M, Hellsand M, Viluma A, Hahn D, Makdoumi K, et al. An ABCA4 loss-of-function mutation causes a canine form of stargardt disease. PLoS Genet. 2019 Mar 19;15(3).

226. Keller SH, Johnson GS, Bullock G, Mhlanga-Mutangadura T, Schwartz M, Pattridge SG, et al. Homozygous CNP Mutation and Neurodegeneration in Weimaraners: Myelin Abnormalities and Accumulation of Lipofuscin-like Inclusions. Genes (Basel). 2024 Feb 15;15(2):246.

227. Bullock G, Johnson GS, Mhlanga-Mutangadura T, Petesch SC, Thompson S, Goebbels S, et al. Lysosomal storage disease associated with a CNP sequence variant in Dalmatian dogs. Gene. 2022 Jul 1;830.

228. Brons AK, Henthorn PS, Raj K, Fitzgerald CA, Liu J, Sewell AC, et al. SLC3A1 and SLC7A9 mutations in autosomal recessive or dominant canine cystinuria: A new classification system. J Vet Intern Med. 2013 Nov;27(6):1400–8.

229. Henthorn PS, Liu J, Gidalevich T, Fang J, Casal ML, Patterson DF, et al. Canine cystinuria: Polymorphism in the canine SLC3A1 gene and identification of a nonsense mutation in cystinuric Newfoundland dogs. Hum Genet. 2000 Oct;107(4):295–303.

230. Harnevik L, Hoppe A, Söderkvist P. SLC7A9 cDNA cloning and mutational analysis of SLC3A1 and SLC7A9 in canine cystinuria. Mammalian Genome. 2006 Jul;17(7):769–76.

231. Murgiano L, Becker D, Spector C, Carlin K, Santana E, Niggel JK, et al. CCDC66 frameshift variant associated with a new form of early-onset progressive retinal atrophy in Portuguese Water Dogs. Sci Rep. 2020 Dec 3;10.

232. Dekomien G, Vollrath C, Petrasch-Parwez E, Boevé MH, Akkad DA, Gerding WM, et al. Progressive retinal atrophy in Schapendoes dogs: mutation of the newly identified CCDC66 gene. Neurogenetics. 2010 May;11(2):163–74.

233. Kreutzer R, Leeb T, Müller G, Moritz A, Baumgärtner W. A duplication in the canine β-galactosidase gene GLB1 causes exon skipping and GM1-gangliosidosis in Alaskan huskies. Genetics. 2005 Aug;170(4):1857–61.

234. Yamato O, Endoh D, Kobayashi A, Masuoka Y, Yonemura M, Hatakeyama A, et al. A novel mutation in the gene for canine acid β-galactosidase that causes GM1-gangliosidosis in Shiba dogs. J Inherit Metab Dis. 2002 Oct;25(6):525–6.

235. Wang ZH, Zeng B, Shibuya H, Johnson GS, Alroy J, Pastores GM, et al. Isolation and characterization of the normal canine β-galactosidase gene and its mutation in a dog model of GM1-gangliosidosis. J Inherit Metab Dis. 2000;23(6):593–606.

236. Christen M, Ludwig-Peisker O, Jagannathan V, Hetzel U, Schönball U, Leeb T. STK36 splice site variant in an Australian Shepherd dog with primary ciliary dyskinesia. Anim Genet. 2023 Jun;54(3):412–5.

237. Van Poucke M, Stee K, Lowrie M, Peelman L. The c.126C>A(p.(Cys42Ter)) SLC7A10 nonsense variant is a candidate causative variant for paradoxical pseudomyotonia in English Cocker and Springer Spaniels. Anim Genet. 2023 Aug;54(4):483–90.

238. Bauer A, Bateman JF, Lamandé SR, Hanssen E, Kirejczyk SGM, Yee M, et al. Identification of two independent COL5A1 variants in dogs with ehlers-danlos syndrome. Genes (Basel). 2019 Sep 21;10(10):731.

239. Bullock G, Jaffey JA, Cohn LA, Sox E, Hostnik ET, Hutcheson KD, et al. Novel COL5A1 variants and associated disease phenotypes in dogs with classical Ehlers-Danlos syndrome. J Vet Intern Med. 2024 Sep;28(5):2431–43.

240. Drögemüller C, Becker D, Kessler B, Kemter E, Tetens J, Jurina K, et al. A deletion in the N-MYC downstream regulated gene 1 (NDRG1) gene in greyhounds with polyneuropathy. PLoS One. 2010 Jun 22;5(6).

241. Bruun CS, Jäderlund KH, Berendt M, Jensen KB, Spodsberg EH, Gredal H, et al. A Gly98Val Mutation in the N-Myc Downstream Regulated Gene 1 (NDRG1) in Alaskan Malamutes with Polyneuropathy. PLoS One. 2013;8(2).

242. Bauer A, Jagannathan V, Högler S, Richter B, McEwan NA, Thomas A, et al. MKLN1 splicing defect in dogs with lethal acrodermatitis. PLoS Genet. 2018 Mar 22;14(3).

243. Guevar J, Olby NJ, Meurs KM, Yost O, Friedenberg SG. Deafness and vestibular dysfunction in a Doberman Pinscher puppy associated with a mutation in the PTPRQ gene. J Vet Intern Med. 2018 Mar;32(2):665–9.

244. Wolf ZT, Brand HA, Shaffer JR, Leslie EJ, Arzi B, Willet CE, et al. Genome-Wide Association Studies in Dogs and Humans Identify ADAMTS20 as a Risk Variant for Cleft Lip and Palate. PLoS Genet. 2015 Mar 23;11(3).

245. Cooper TA, Wiggans GR, VanRaden PM. Relationship of call rate and accuracy of single nucleotide polymorphism genotypes in dairy cattle. J Dairy Sci. 2013 May;96(5):3336–9.

246. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 2010 Sep;34(6):591–602.

247. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Pielberg GR, Sigurdsson S, et al. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 2011 Oct;7(10).

248. Illumina. Data Sheet: DNA Genotyping [Internet]. [cited 2025 Mar 24]. Available from: http://www.ncbi.nlm.nih.gov/genome/guide/dog/

249. Maruki T, Kumar S, Kim Y. Purifying selection modulates the estimates of population differentiation and confounds genome-wide comparisons across single-nucleotide polymorphisms. Mol Biol Evol. 2012 Aug 31;29(12):3617–23.

250. Bannasch D, Famula T, Donner J, Anderson H, Honkanen L, Batcher K, et al. The effect of inbreeding, body size and morphology on health in dog breeds. Canine Med Genet. 2021 Dec 2;8.

251. Vigeland MD. Relatedness coefficients in pedigrees with inbred founders. J Math Biol. 2020 Jun 8;81:185–207.

252. Kim TK, Park JH. More about the basic assumptions of t-test: Normality and sample size. Korean J Anesthesiol. 2019 Aug;72(4):331–5.

253. Zhu D, Zhao Y, Zhang R, Wu H, Cai G, Wu Z, et al. Genomic prediction based on selective linkage disequilibrium pruning of low-coverage whole-genome sequence variants in a pure Duroc population. Genetics Selection Evolution. 2023 Oct 18;55.

254. Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, et al. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. 2014 Jul;24(7):1193–208.

255. Varona L, Legarra A, Toro MA, Vitezica ZG. Non-additive effects in genomic selection. Front Genet. 2018 Mar 6;9.

256. Geibel J, Praefke NP, Weigend S, Simianer H, Reimer C. Assessment of linkage disequilibrium patterns between structural variants and single nucleotide polymorphisms in three commercial chicken populations. BMC Genomics. 2022 Mar 9;23.

257. Genomics gets faster, cheaper, and more accurate - Wellcome Sanger Institute [Internet]. [cited 2025 Jun 9]. Available from: https://sangerinstitute.blog/2024/02/29/genomics-gets-faster-cheaper-and-more-accurate/

258. Whole Genome Sequencing Costs 2024: New Prices and Future Projections - 3billion [Internet]. [cited 2025 Jun 9]. Available from: https://3billion.io/blog/whole-genome-sequencing-costs-2024-new-prices-and-future-projections

259. Asimit JL, Zeggini E. Imputation of rare variants in next generation association studies. Hum Hered. 2012;74(3–4):196–204.

260. Daetwyler HD, Capitan A, Pausch H, Stothard P, Van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014 Aug;46(8):858–65.

261. Rimbault M, Beale HC, Schoenebeck JJ, Hoopes BC, Allen JJ, Kilroy-Glynn P, et al. Derived variants at six genes explain nearly half of size reduction in dog breeds. Genome Res. 2013 Dec;23(12):1985–95.

262. Singh I, Tucker LA, Gendall P, Rutherfurd-Markwick KJ, Cline J, Thomas DG. Age, breed, sex distribution and nutrition of a population of working farm dogs in New Zealand: Results of a crosssectional study of members of the New Zealand sheep dog trial association. N Z Vet J. 2011;59(3):133–8.

263. Dogs New Zealand - NZ Huntaway - Information and NZ Breed Standards [Internet]. [cited 2025 May 27]. Available from: https://www.dogsnz.org.nz/breeds/info/nz-huntaway/535/breed-standard

264. Majeres LE, Dilger AC, Shike DW, McCann JC, Beever JE. Defining a Haplotype Encompassing the LCORL-NCAPG Locus Associated with Increased Lean Growth in Beef Cattle. Genes (Basel). 2024 Apr 30;15(5).

265. Appelbe OK, Bollman B, Attarwala A, Triebes LA, Muniz-Talavera H, Curry DJ, et al. Disruption of the mouse Jhy gene causes abnormal ciliary microtubule patterning and juvenile hydrocephalus. Dev Biol. 2013 Oct 1;382(1):172–85.

266. Mian AA, Baumann I, Liebermann M, Grebien F, Superti-Furga G, Ruthardt M, et al. The phosphatase UBASH3B/Sts-1 is a negative regulator of Bcr-Abl kinase activity and leukemogenesis. Leukemia. 2019 Sep;33(9):2319–23.

267. Leavy O. Polarity and CRTAM: a matter of timing. Nat Rev Immunol. 2008 Apr;8.

268. Wang XJ, Cao Q, Zhang Y, Su XD. Activation and regulation of caspase-6 and its role in neurodegenerative diseases. Annu Rev Pharmacol Toxicol. 2015 Jan 6;55:553–72.

269. Van De Ven JPH, Nilsson SC, Tan PL, Buitendijk GHS, Ristau T, Mohlin FC, et al. A functional variant in the CFI gene confers a high risk of age-related macular degeneration. Nat Genet. 2013 Jul;45:813–7.

270. Fujikane R, Behm-Ansmant I, Tillault AS, Loegler C, Igel-Bourguignon V, Marguet E, et al. Contribution of protein Gar1 to the RNA-guided and RNA-independent rRNA:Ψ-synthase activities of the archaeal Cbf5 protein. Sci Rep. 2018 Sep 14;8.

271. Bellingham J, Wells DJ, Foster RG. In silico characterisation and chromosomal localisation of human RRH (peropsin)--implications for opsin evolution. 2003 Jan 24;4(1).

272. Huo J, Prasad V, Grimes KM, Vanhoutte D, Blair NS, Lin SC, et al. MCUb is an inducible regulator of calcium-dependent mitochondrial metabolism and substrate utilization in muscle. Cell Rep. 2023 Nov 28;42(11).

273. Kinehara M, Fukuda I, Yoshida K ichi, Ashida H. Aryl hydrocarbon receptor-mediated induction of the cytosolic phospholipase A2α gene by 2,3,7,8-tetrachlorodibenzo-p-dioxin in mouse hepatoma Hepa-1c1c7 cells. J Biosci Bioeng. 2009 Oct;108(4):277–81.

274. Minvielle F, Bed B, Coville JL, Ito ichi, Inoue-Murayama M, Gourichon D. The “silver” Japanese quail and the MITF gene: causal mutation, associated traits and homology with the “blue” chicken plumage. 2010 Feb;11.

275. Nariyama M, Kota Y, Kaneko S, Asada Y, Yamane A. Association between the lack of teeth and the expression of myosins in masticatory muscles of microphthalmic mouse. Cell Biochem Funct. 2012 Jan;30(1):82–8.

276. Visscher PM, Hemani G, Vinkhuyzen AAE, Chen GB, Lee SH, Wray NR, et al. Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. PLoS Genet. 2014 Apr 10;10(4).

277. Mortezaei Z, Tavallaei M. Recent innovations and in-depth aspects of post-genome wide association study (Post-GWAS) to understand the genetic basis of complex phenotypes. Heredity (Edinb). 2021 Dec;127(6):485–97.

278. Reynolds EGM, Neeley C, Lopdell TJ, Keehan M, Dittmer K, Harland CS, et al. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. Nat Genet. 2021 Jul;53(7):949–54.

279. Akey J, Li J, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? European Journal of Human Genetics. 2001 Apr;9(4):291–300.

280. Axelsson Id E, Ljungvall I, Conn LB, Muren E, Ohlsson Id Å, Høier L, et al. The genetic consequences of dog breed formation-Accumulation of deleterious genetic variation and fixation of mutations associated with myxomatous mitral valve disease in cavalier King Charles spaniels. 2021;

281. Mahmoodi M, Ayatollahi Mehrgardi A, Momen M, Serpell JA, Esmailizadeh A. Deciphering the genetic basis of behavioral traits in dogs: Observed-trait GWAS and latent-trait GWAS analysis reveal key genes and variants. The Veterinary Journal. 2024 Dec;308:106251.

282. Jeong H, Ostrander EA, Kim J. Genomic evidence for behavioral adaptation of herding dogs. Science Advances . 2025 May 2;11(18):4591.

Appendix A - Supplementary Methods

A.1 Right Dog for the Job Health and Behaviour Trait Survey

Owner name		Dog name		
Email & phone		Dog age		
Owner region		Full time on farm	Yes	No
Recorder name		Vial No		
Hours worked per day (avg)	☐Rarely ☐ Less than 1 hour ☐Between 1-3 hours ☐3 hours+ per day			
Reproduction status	☐ M - Entire ☐ M - Neutered ☐ F - Entire ☐ F - Spayed			
Dogs breed	☐Heading Dog ☐ Huntaway ☐Border Collie ☐Heading Dog/Huntaway ☐Heading Dog/Border Collie ☐Other:			
Dogs colour	☐Tri - Black/Tan/White) ☐ Black/White ☐Black/Tan ☐Brown/White ☐Brown ☐Tan ☐Black ☐White ☐Brindle			
Ear set	☐Pricked(both) ☐ Pricked (1 only) ☐Floppy ☐ Tipped (pricked with folded tips)			
Coat	☐ Smooth ☐ Smooth plus ☐ Medium ☐ Rough ☐ Grizzly/Beardie			
Height (withers to floor, mm)				
Length (base of skull to crest of the illum, mm)				
Chest circumference (behind the elbow, mm)				
Muzzle circumference (mm)				
Jaw Alignment	☐ Normal ☐ Undershot lower jaw ☐ Overshot lower jaw			
Front dew claws	☐Yes ☐ No ☐Surgical removal ☐Not sure			
Hind dew claws	☐Yes ☐ No ☐Surgical removal ☐Surgical removal			

Had surgery for knee ligament (cruciate) damage		☐Yes, 1 ☐Yes, both ☐Maybe ☐No										
Diagnosed with hip dysplasia?	☐Yes, mild one ☐ Yes, mild both ☐ Yes, severe one ☐ Yes, severe both ☐ No, does get sore hips ☐ No lameness noted											
Diagnosed with elbow dysplasia?	☐Yes, mild one ☐ Yes, mild both ☐ Yes, severe one ☐ Yes, severe both ☐ No, does get sore front legs ☐ No lameness noted											
Diagnosed with shoulder OCD?	☐Yes, mild one ☐ Yes, mild both ☐ Yes, severe one ☐ Yes, severe both ☐ No, does get sore front legs ☐ No lameness noted											
Been diagnosed with a heart condition	☐Yes ☐ No											
Had cancer?	☐Yes ☐ No											
Been bred from:	☐Yes ☐ No ☐ Don't know											
Any vision impairment:	☐Yes ☐ No											
Any hearing impairment:	☐Yes ☐ No											
Any allergies:	☐Yes ☐ No											
Had 2+ seizures in its life so far.	☐Yes ☐ No											
Had a twisted gut (GDV):	☐Yes (surgery required) ☐ No (surgery unknown) ☐Yes, preemptive surgery has been done ☐No – no surgery ☐Don't know											
Any siblings had a twisted gut (GDV):	☐Yes (fatal or surgery required) ☐No (surgery unknown) ☐Yes (preemptive surgery has been done) ☐ No surgery) ☐Don't know											
Parents had a twisted gut (GDV):	☐Yes (fatal or surgery required) ☐No (surgery unknown) ☐Yes (preemptive surgery has been done) ☐ No surgery) ☐Don't know											
Get sore feet	(Almost never)	1	2	3	4	5	6	7	8	9	10	(Paws easily damaged)
Have physical stamina	(Tires easily)	1	2	3	4	5	6	7	8	9	10	(Extreme stamina)

General health comments:

Is this dog nervous/ highly strung?												
Not at all	0	10	20	30	40	50	60	70	80	90	100	Extremely nervous
How does this dog interact with unknown people?												
Frightened	0	10	20	30	40	50	60	70	80	90	100	Bold
How does this dog interact with unknown dogs?												
Frightened	0	10	20	30	40	50	60	70	80	90	100	Bold
Is this dog tidy in the kennels?												
Never	0	10	20	30	40	50	60	70	80	90	100	Always
Does this dog create a nuisance in the kennels?												
Never	0	10	20	30	40	50	60	70	80	90	100	Often
Is this dog biddable/trainable?												
Not at all	0	10	20	30	40	50	60	70	80	90	100	Absolutely
Is this dog calm and collected when working stock?												
Busy	0	10	20	30	40	50	60	70	80	90	100	Steady
Is this dog inclined to bite the sheep unnecessarily?												
Not at all	0	10	20	30	40	50	60	70	80	90	100	Absolutely
Was this dog easy to train for sheep work?												
Not at all	0	10	20	30	40	50	60	70	80	90	100	Absolutely
How old was this dog when it started showing interest in working sheep?												
Is this dog confident working at a distance (i.e >800m)?												
Not at all	0	10	20	30	40	50	60	70	80	90	100	Absolutely
Would you trust this dog to be left alone with the sheep?												
Not at all	0	10	20	30	40	50	60	70	80	90	100	Absolutely
Can this dog maintain close contact with sheep without unsettling them?												
Not at all	0	10	20	30	40	50	60	70	80	90	100	Absolutely
How would you rate this dog's ability overall?												
Worst dog ever	0	10	20	30	40	50	60	70	80	90	100	Best dog ever

If dogs breed is Huntaway or Huntaway cross, does this dog have a commanding bark?												
Weak	0	10	20	30	40	50	60	70	80	90	100	Strong
What is the pitch of the dog's bark?												
Yappy	0	10	20	30	40	50	60	70	80	90	100	Deep

General behaviour/working comments:

A.2 Supplementary methods from variant effect prediction

Table A1: SnpEff effect categories

Effect category	SnpEff annotations
Synonymous	Synonymous variant
Missense	Initiator codon variant, missense variant, stop retained variant, start retained variant
Start or stop lost or gained	5 prime UTR premature start codon gain variant, start lost, stop gained, stop lost
Splice region	Splice region variant, splice acceptor variant, splice donor variant
Frameshift	Frameshift variant
In-frame indel	Conservative in-frame deletion, conservative in-frame insertion, disruptive in-frame deletion, disruptive in-frame insertion
Other	Bidirectional gene fusion, exon loss variant, gene fusion, transcript ablation
Intron	Intron variant
Downstream or upstream	Downstream gene variant, upstream gene variant
Inter- or intragenic	Intergenic region, intragenic variant
Non-coding transcript	Non-coding transcript exon variant, non-coding transcript variant
UTR	3 prime UTR variant, 5 prime UTR variant, 5 prime UTR truncation

Description of SnpEff effect annotations included in each effect category.

Note: *SnpEff defines intragenic variants as those that hit a gene, but no transcripts within the gene (118).*

A.3 Supplementary methods from RNA-Seq annotation

Table A2: RNABamDepth of candidate functional loci

Coding status [a]	RNABamDepth [b]	Mean depth
Coding	7, 2664, 151, 57, 59, 3667, 28, 365, 5530, 55, 198, 30, 342, 48, 3196, 753, 2179, 21,365, 11, 10, 11,945, 511, 673, 18, 224, 11,235, 778, 2856, 167, 22, 46	2232
Non-coding	0, 5, 4, 2, 2, 0, 0, 0, 0, 2, 0, 0, 0, 1, 36, 4, 6, 5, 5	4

Assessment of the RNA-seq read depths at 50 candidate functional loci by protein-coding status. This data was used to set a threshold for filtering functional predictions.

[a] Allocated protein coding status of loci based on IGV inspection.

[b] Read depth in a 17-tissue combined RNA-Seq BAM file.

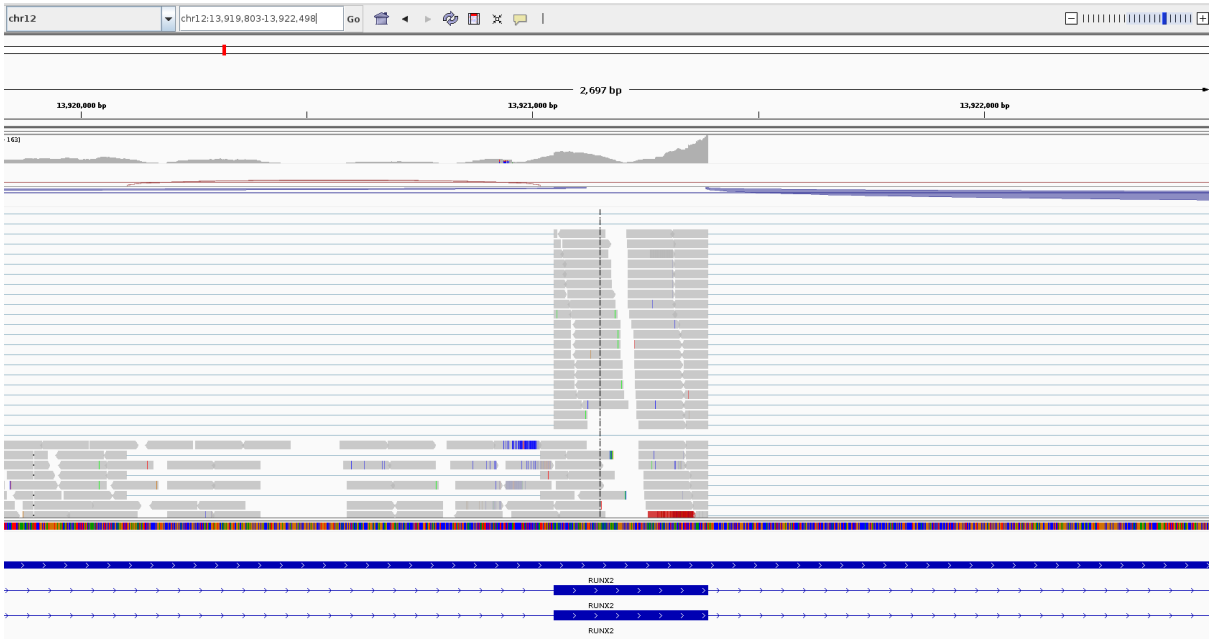

Figure A1. IGV screenshot of RNA-Seq reads at a candidate functional locus that was assigned as protein-coding. The black line represents a variant that was annotated to intersect a coding sequence. Based on the high RNA-Seq read depth (grey bars) in the annotated exon compared to the neighbouring intron, the transcript was presumed to be expressed.

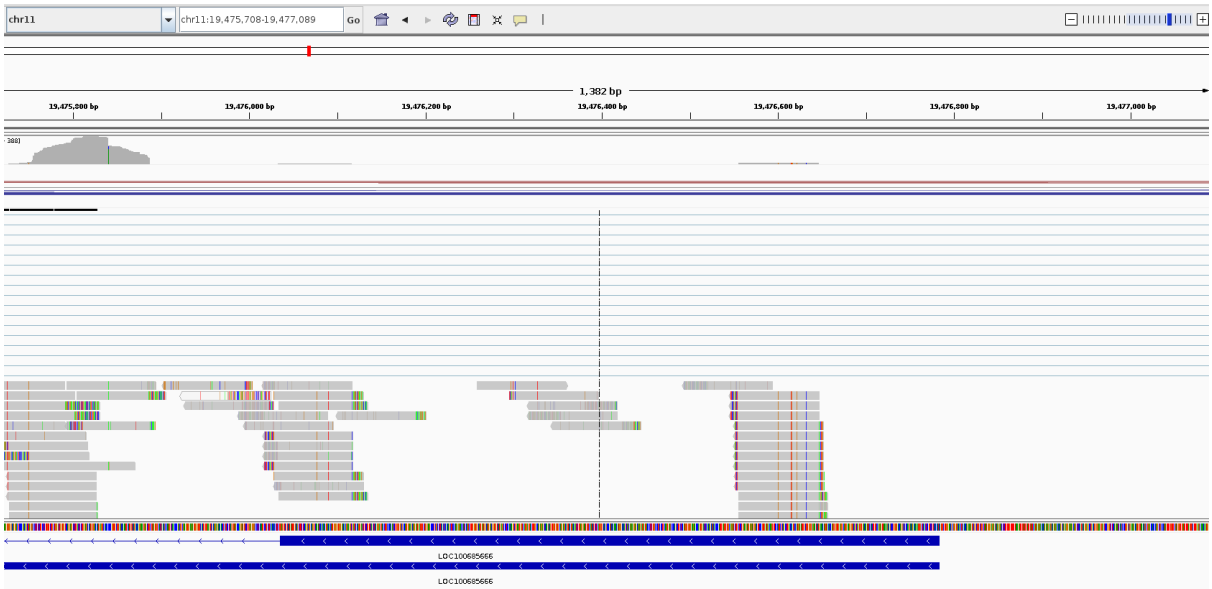

Figure A2. IGV screenshot of RNA-Seq reads at a candidate functional locus that was assigned as non-coding. The black line represents a variant that was annotated to intersect a coding sequence. Based on the low RNA-Seq read depth (grey bars) in the annotated exon compared to the neighbouring intron, the variant transcript was presumed to not be expressed.

A.4 Supplementary methods from high-impact variant survey

Table A3: List of OMIA-derived genes of interest

Gene name	Gene description
ABCA4	ATP-binding cassette, sub-family A (ABC1), member 4
ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1
ABHD5	abhydrolase domain containing 5
ACADVL	acyl-CoA dehydrogenase, very long chain
ADAMTS17	ADAM metallopeptidase with thrombospondin type 1 motif, 17
ADAMTS2	ADAM metallopeptidase with thrombospondin type 1 motif, 2
ADAMTS20	ADAM metallopeptidase with thrombospondin type 1 motif, 20
AGL	amylo-alpha-1,6-glucosidase and 4-alpha-glucanotransferase
AKNA	AT-hook transcription factor
ANLN	anillin, actin binding protein
AP3B1	adaptor-related protein complex 3, beta 1 subunit
ARSB	arylsulfatase B
BBS4	Bardet-Biedl syndrome 4
BEST1	bestrophin 1
C3	complement component 3
CCDC39	coiled-coil domain containing 39
CCDC66	coiled-coil domain containing 66
CHRNE	cholinergic receptor, nicotinic, epsilon (muscle)
CLCN1	chloride channel, voltage-sensitive 1
CLN5	ceroid-lipofuscinosis, neuronal 5
CLN8	ceroid-lipofuscinosis, neuronal 8 (epilepsy, progressive with mental retardation)
CNGA1	cyclic nucleotide gated channel alpha 1
CNGB1	cyclic nucleotide gated channel beta 1
CNP	2',3'-cyclic nucleotide 3' phosphodiesterase
COL1A2	collagen, type I, alpha 2
COL4A5	collagen, type IV, alpha 5
COL5A1	collagen, type V, alpha 1
COL6A1	collagen, type VI, alpha 1
COL6A3	collagen, type VI, alpha 3
COL7A1	collagen, type VII, alpha 1
COL9A3	collagen, type IX, alpha 3
CUBN	cubilin (intrinsic factor-cobalamin receptor)
CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2
CYP27B1	cytochrome P450, family 27, subfamily B, polypeptide 1
DIRAS1	DIRAS family, GTP-binding RAS-like 1

DMD	dystrophin
DSG1	desmoglein 1
DVL2	dishevelled segment polarity protein 2
EDA	ectodysplasin A
EHBP1L1	EH domain binding protein 1-like 1
ENAM	enamelin
EXT2	exostosin glycosyltransferase 2
F8	coagulation factor VIII, procoagulant component
F9	coagulation factor IX
FAM83H	family with sequence similarity 83, member H
FGA	fibrinogen alpha chain
FGF5	fibroblast growth factor 5
FNIP2	folliculin interacting protein 2
FOXI3	forkhead box I3
FUCA1	fucosidase, alpha-L- 1, tissue
FYCO1	FYVE and coiled-coil domain containing 1
GAA	glucosidase, alpha; acid
GLB1	galactosidase, beta 1
GUCY2D	guanylate cyclase 2D, membrane (retina-specific)
HACE1	HECT domain and ankyrin repeat containing E3 ubiquitin protein ligase 1
HES7	hes family bHLH transcription factor 7
HEXB	hexosaminidase B (beta polypeptide)
HPS3	Hermansky-Pudlak syndrome 3
HSD17B3	hydroxysteroid (17-beta) dehydrogenase 3
HSF4	heat shock transcription factor 4
IL2RG	interleukin 2 receptor, gamma
IQCB1	IQ motif containing B1
ITGA10	integrin, alpha 10
ITGA2B	integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
LAMA2	laminin subunit alpha 2
LARGE1	like-glycosyltransferase
LGI2	leucine-rich repeat LGI family, member 2
LRIT3	leucine-rich repeat, immunoglobulin-like and transmembrane domains 3
MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)
MFF	mitochondrial fission factor

MFSD8	major facilitator superfamily domain containing 8
MKLN1	muskelin 1, intracellular mediator containing kelch motifs
MLPH	melanophilin
MOCOS	molybdenum cofactor sulfurase
MPO	myeloperoxidase
MSTN	myostatin
MTM1	myotubularin 1
MYO5A	myosin VA
NAPEPLD	N-acyl phosphatidylethanolamine phospholipase D
NDP	Norrie disease (pseudoglioma)
NDRG1	N-myc downstream regulated 1
NEB	nebulin
NIPAL4	NIPA-like domain containing 4
NKX2-8	NK2 homeobox 8
NME5	NME/NM23 family member 5
NSDHL	NAD(P) dependent steroid dehydrogenase-like
PCARE	photoreceptor cilium actin regulator
PDE6A	phosphodiesterase 6A, cGMP-specific, rod, alpha
PDE6B	phosphodiesterase 6B, cGMP-specific, rod, beta
PDP1	pyruvate dehyrogenase phosphatase catalytic subunit 1
PFKM	phosphofructokinase, muscle
PKD1	polycystic kidney disease 1 (autosomal dominant)
PKLR	pyruvate kinase, liver and RBC
PLEC	plectin
PNPLA1	patatin-like phospholipase domain containing 1
PNPLA8	patatin-like phospholipase domain containing 8
PPT1	palmitoyl-protein thioesterase 1
PRKDC	protein kinase, DNA-activated, catalytic polypeptide
PTPRQ	protein tyrosine phosphatase, receptor type, Q
RAB3GAP1	RAB3 GTPase activating protein subunit 1 (catalytic)
RAG1	recombination activating gene 1
RASGRP2	RAS guanyl releasing protein 2 (calcium and DAG-regulated)
RELN	reelin
RPE65	retinal pigment epithelium-specific protein 65kDa
RPGR	retinitis pigmentosa GTPase regulator
SAG	S-antigen; retina and pineal gland (arrestin)
SBF2	SET binding factor 2
SCARF2	scavenger receptor class F, member 2
SGCA	sarcoglycan, alpha (50kDa dystrophin-associated glycoprotein)
SGCD	sarcoglycan, delta (35kDa dystrophin-associated glycoprotein)
SGK3	serum/glucocorticoid regulated kinase family, member 3

SGSH	N-sulfoglucosamine sulfohydrolase
SH3TC2	SH3 domain and tetratricopeptide repeats 2
SIX6	SIX homeobox 6
SLC12A6	solute carrier family 12 (potassium/chloride transporter), member 6
SLC19A3	solute carrier family 19 (thiamine transporter), member 3
SLC3A1	solute carrier family 3 (amino acid transporter heavy chain), member 1
SLC45A2	solute carrier family 45, member 2
SLC4A3	solute carrier family 4 (anion exchanger), member 3
SLC6A5	solute carrier family 6 (neurotransmitter transporter), member 5
SLC7A10	solute carrier family 7 (neutral amino acid transporter light chain, asc system), member 10
SOD1	superoxide dismutase 1, soluble
SPTBN2	spectrin, beta, non-erythrocytic 2
STK36	serine/threonine kinase 36
TNR	tenascin R
TPO	thyroid peroxidase
TPP1	tripeptidyl peptidase I
TTC8	tetratricopeptide repeat domain 8
TYRP1	tyrosinase-related protein 1
VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor
VLDLR	very low density lipoprotein receptor

Genes previously reported as functional by OMIA (i.e. genes that affect a phenotype when disrupted) that were surveyed for high-impact variants in the WGS dataset.

Note: *All data in this table were obtained from the OMIA database (106).*

Appendix B – Supplementary Results

B.1 Reads properly mapped and paired

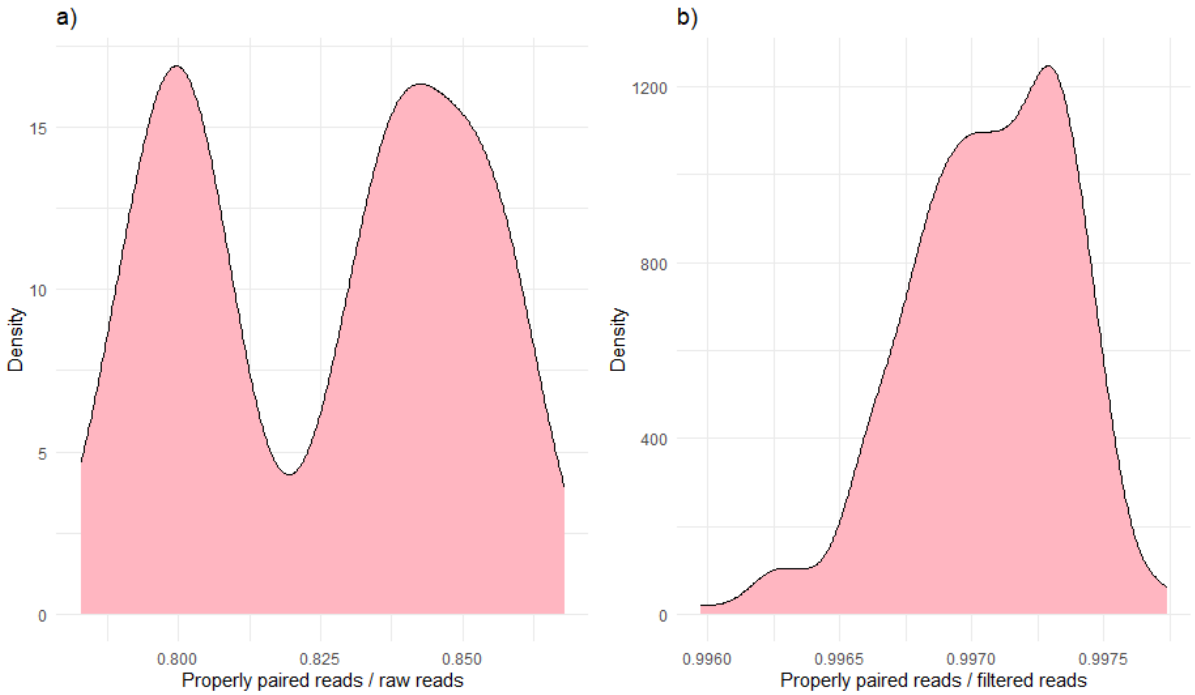

Figure A3. Density plots of the proportion of WGS reads that were properly mapped and paired. **a)** Proportion of unfiltered reads that were properly mapped and paired, **b)** Proportion of filtered reads that were properly mapped and paired.

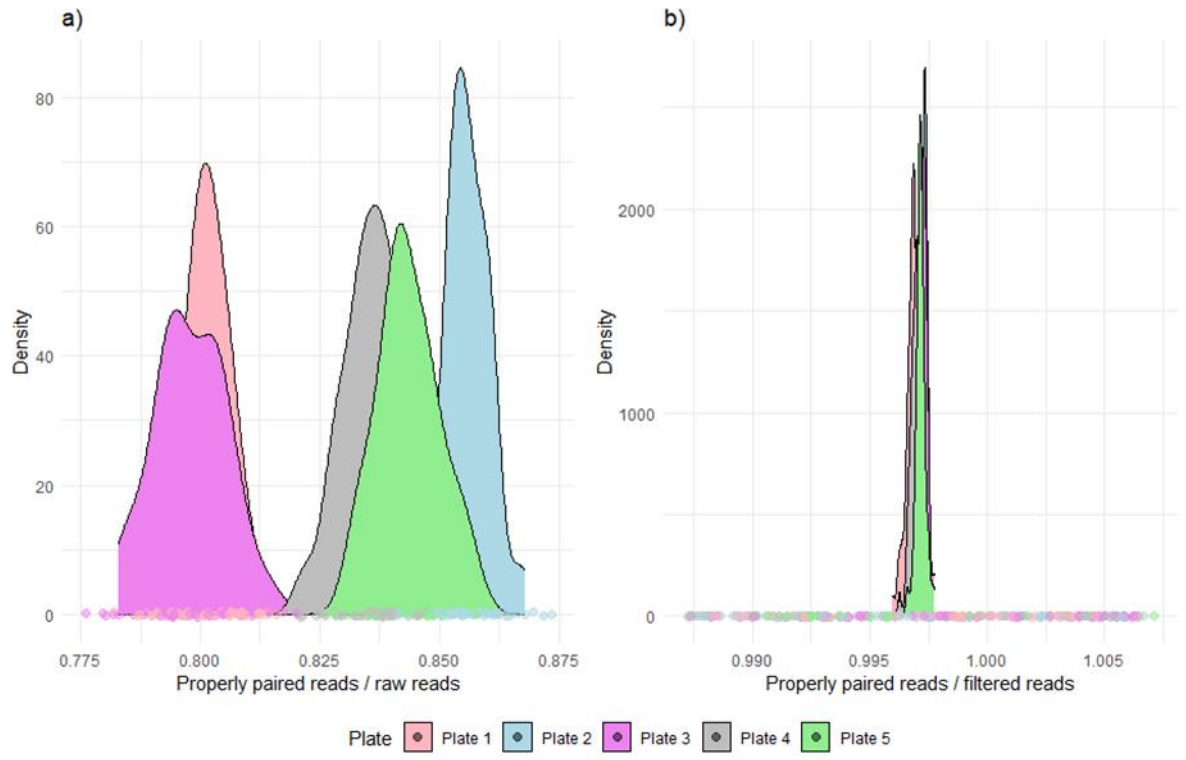

Figure A4. Density plots of the proportion of WGS reads that were properly mapped and paired by plate. **a)** Proportion of unfiltered reads that were properly mapped and paired, **b)** Proportion of filtered reads that were properly mapped and paired.

B.2 Supplementary results from SnpEff annotation

Table A4: SnpEff-predicted effects by breed

	Huntaways		Heading Dogs	
Effect [a]	**Count**	**Proportion**	**Count**	**Proportion**
Synonymous	50,070	0.0034	47,228	0.0034
Missense	40,080	0.0027	37,307	0.0027
Start or stop lost or gained	5604	0.0004	5164	0.0004
Splice region	22,095	0.0015	19,935	0.0014
Frameshift	5019	0.0003	4576	0.0003
In-frame indel	2967	0.0002	2466	0.0002
Other	76	5.10×10^{-6}	77	5.48×10^{-6}
Intron	6,195,845	0.4157	5,836,598	0.4156
Upstream or downstream	2,319,163	0.1556	2,185,242	0.1556
Intergenic or intragenic	6,066,171	0.4070	5,723,501	0.4075
Non-coding transcript	52,037	0.0035	49,763	0.0035
UTR	145,425	0.0098	132,226	0.0094

Note: *Counts and proportions are calculated from the most severe effect annotation for each variant.*

[a] *SnpEff effect categories (see Table A1).*

Table A5: SnpEff-predicted impacts by breed

	Huntaways		Heading Dogs	
Impact [a]	**Count**	**Proportion**	**Count**	**Proportion**
High	18,135	0.0003	16,601	0.0003
Moderate	118,417	0.0022	109,401	0.0022
Low	234,190	0.0044	217,119	0.0044
Modifier	52,491,721	0.9930	48,571,800	0.9930

Note: *Counts and proportions are calculated from the annotated impacts for all transcripts affected by variants.*

[a] *SnpEff impact annotations.*

B.3 Supplementary results from Mendelian variant survey

Table A6: Genotype frequencies of Mendelian variants segregating in the WGS sample

OMIA ID	Gene	Genotype frequencies		
		Huntaways [a]	Heading Dogs [a]	Total [b]
1603	*MC5R*	C/C = 0.15 C/T = 0.42 T/T = 0.42	C/C = 0.15 C/T = 0.42 T/T = 0.42	C/C = 0.15 C/T = 0.42 T/T = 0.42
447	*CUBN*	C/C = 1 C/- = 0 -/- = 0	C/C = 1 C/- = 0 -/- = 0	C/C = 1 C/- = 0 -/- = 0
343	*MC1R*	G/G = 0.02 G/A = 0.05 A/A = 0.92	G/G = 0.02 G/A = 0.05 A/A = 0.92	G/G = 0.02 G/A = 0.05 A/A = 0.92
34	*MC1R*	C/C = 0.02 C/T = 0.08 T/T = 0.92	C/C = 0.02 C/T = 0.08 T/T = 0.92	C/C = 0.02 C/T = 0.08 T/T = 0.92
577	*SGSH*	T/T = 0.99 T/TA = 0.01 TA/TA = 0	T/T = 0.99 T/TA = 0.01 TA/TA = 0	T/T = 0.99 T/TA = 0.01 TA/TA = 0
851	*BTBD17*	C/C = 0.82 C/CG = 0.17 CG/CG = 0.02	C/C = 0.82 C/CG = 0.17 CG/CG = 0.02	C/C = 0.82 C/CG = 0.17 CG/CG = 0.02
1273	*CNTNAP1*	C/C = 1 C/T = 0 T/T = 0	C/C = 1 C/T = 0 T/T = 0	C/C = 1 C/T = 0 T/T = 0
31	*TYRP1*	T/T = 0.92 T/A = 0.08 A/A = 0	T/T = 0.92 T/A = 0.08 A/A = 0	T/T = 0.92 T/A = 0.08 A/A = 0
267	*TYRP1*	C/C = 0.91 C/T = 0.09 T/T = 0	C/C = 0.91 C/T = 0.09 T/T = 0	C/C = 0.91 C/T = 0.09 T/T = 0
796	*TYRP1*	CCT/CCT = 0.92 CCT/- = 0.08 -/- = 0	CCT/CCT = 0.92 CCT/- = 0.08 -/- = 0	CCT/CCT = 0.92 CCT/- = 0.08 -/- = 0
442	*ABCB1*	A/A = 0.47 A/C = 0.45 C/C = 0.08	A/A = 0.47 A/C = 0.45 C/C = 0.08	A/A = 0.47 A/C = 0.45 C/C = 0.08
1422	*IGF1-AS*	C/C = 0.90 C/T = 0.1 T/T = 0	C/C = 0.41 C/T = 0.46 T/T = 0.13	C/C = 0.68 C/T = 0.27 T/T = 0.05
1444	*LMBR1*	C/C = 0.79 C/T = 0.20 T/T = 0.01	C/C = 0.79 C/T = 0.20 T/T = 0.01	C/C = 0.79 C/T = 0.20 T/T = 0.01
458	*CBD103*	GGG/GGG = 0.78 GGG/- = 0.22 -/- = 0.01	GGG/GGG = 0.78 GGG/- = 0.22 -/- = 0.01	GGG/GGG = 0.78 GGG/- = 0.22 -/- = 0.01

1522	*RETN*	C/C = 0.85 C/T = 0.14 T/T = 0.01	C/C = 0.85 C/T = 0.14 T/T = 0.01	C/C = 0.85 C/T = 0.14 T/T = 0.01
1081	*MFSD12*	C/C = 0.48 C/T = 0.44 T/T = 0.08	C/C = 0.43 C/T = 0.45 T/T = 0.12	C/C = 0.45 C/T = 0.45 T/T = 0.10
106	*ATP7B*	G/G = 0.88 G/A = 0.12 A/A = 0	G/G = 0.88 G/A = 0.12 A/A = 0	G/G = 0.88 G/A = 0.12 A/A = 0
30	*ASIP*	C/C = 0.96 C/T = 0.04 T/T = 0	C/C = 0.96 C/T = 0.04 T/T = 0	C/C = 0.96 C/T = 0.04 T/T = 0
360	*MLPH*	G/G = 0.9 G/A = 0.1 A/A = 0	G/G = 0.9 G/A = 0.1 A/A = 0	G/G = 0.9 G/A = 0.1 A/A = 0
401	*VWF*	C/C = 0.93 C/T = 0.07 T/T = 0	C/C = 0.93 C/T = 0.07 T/T = 0	C/C = 0.93 C/T = 0.07 T/T = 0
35	*KRT71*	G/G = 0.98 G/A = 0.02 A/A = 0	G/G = 0.98 G/A = 0.02 A/A = 0	G/G = 0.98 G/A = 0.02 A/A = 0
274	*CYP1A2*	C/C = 0.35 C/T = 0.45 T/T = 0.21	C/C = 0.35 C/T = 0.45 T/T = 0.21	C/C = 0.35 C/T = 0.45 T/T = 0.21
36	*SOD1*	G/G = 0.61 G/A = 0.35 A/A = 0.04	G/G = 0.61 G/A = 0.35 A/A = 0.04	G/G = 0.61 G/A = 0.35 A/A = 0.04
48	*FGF5*	C/C = 0.83 C/A = 0.15 A/A = 0.02	C/C = 0.83 C/A = 0.15 A/A = 0.02	C/C = 0.83 C/A = 0.15 A/A = 0.02
103	*P3H2*	C/C = 1 C/G = 0 G/G = 0	C/C = 1 C/G = 0 G/G = 0	C/C = 1 C/G = 0 G/G = 0
338	*CLN8*	G/G = 0.95 G/A = 0.05 A/A = 0	G/G = 0.95 G/A = 0.05 A/A = 0	G/G = 0.95 G/A = 0.05 A/A = 0
612	*KCNJ10*	C/C = 0 C/- = 0.07 -/- = 0.93	C/C = 0 C/- = 0.28 -/- = 0. 72	C/C = 0 C/- = 0.16 -/- = 0.84

[a] *Calculated from dogs reported as purebred in the WGS sample.*

[b] *Calculated from total WGS sample.*

Table A7: Summary of Mendelian variants not detected in a sample of NZ farm dogs

OMIA ID [a]	Gene [a]	Associated phenotype [a]	Chr	Variant [b]
109	*PIGN*	Dyskinesia, paroxysmal	1	g.14781926C>T
79	*TBXT*	Bob tail	1	g.54742499G>C
1078	*PLN*	Cardiomyopathy, dilated	1	g.59146722G>A
1389	*LAMA2*	Congenital muscular dystrophy	1	g.68441978G>A
1095	*HSD17B3*	Disorder of sexual development, 78, XY, SRY-positive	1	g.71141090del2
917	*VLDLR*	Cerebellar hypoplasia	1	g.91944759insA
739	*JAK2*	Polycythemia	1	g.94091480delins
1045	*ACPT*	Amelogenesis imperfecta	1	g.106718704insC
54	*RYR1*	Malignant hyperthermia	1	g.115393062A>G
1534	*SLC7A10*	Paradoxical pseudomyotonia	1	g.119752573C>A
85	*SLC7A9*	Cystinuria, type II-B	1	g.120056493G>A
529	*CUBN*	Intestinal cobalamin	2	g.18746067AC>A
1036	*CUBN*	Intestinal cobalamin	2	g.18939568G>A
970	*SUV39H2*	Nasal parakeratosis	2	g.20695177del4
86	*SUV39H2*	Nasal parakeratosis	2	g.20695207A>C
1305	*PITRM1*	Epilepsy, mitochondrial dysfunction and neurodegeneration	2	g.31223269del6
463	*HEXB*	Gangliosidosis, GM2, type II	2	g.56476930del
798	*HEXB*	Gangliosidosis, GM2, type II	2	g.56494916del3
918	*CNGB1*	Progressive retinal atrophy	2	g.57893552delins
1391	*BBS2*	Bardet-Biedl syndrome 2	2	g.58970967G>C
461	*FUCA1*	Fucosidosis, alpha	2	g.75060030del14
1021	*ALPL*	Hypophosphatasia	2	g.76963551A>C
1067	*ATP13A2*	Neuronal ceroid lipofuscinosis, 12	2	g.80692306C>T
400	*ATP13A2*	Neuronal ceroid lipofuscinosis, 12	2	g.80694515TG>T
472	*MFN2*	Neuroaxonal dystrophy	2	g.83765720del3
1206	*APC*	Familial Adenomatous Polyposis	3	g.309816del2
859	*ARSB*	Mucopolysaccharidosis VI	3	g.28031844C>T
1258	*ARSB*	Mucopolysaccharidosis VI	3	g.28112904G>A
580	*AP3B1*	Neutropenia, cyclic	3	g.28829080insA
846	*OCA2*	Coat colour, oculocutaneous albinism	3	g.32515468T>C
685	*ADAMTS17*	Glaucoma, primary open angle	3	g.40986895del20
365	*ADAMTS17*	Lens luxation	3	g.41160144G>A
96	*ADAMTS17*	Glaucoma, primary open angle	3	g.41186358G>A
942	*ADAMTS17*	Primary open-angle glaucoma (POAG), primary lens luxation (PLL), or both	3	g.41320121del6
83	*SLC2A9*	Urolithiasis	3	g.70034534G>T
269	*LGI2*	Epilepsy, benign familial juvenile	3	g.85807159A>T
1154	*KCNIP4*	Ataxia, cerebellar	3	g.89488542T>C
911	*IDUA*	Mucopolysaccharidosis I	3	g.92143028C>T
1190	*IDUA*	Mucopolysaccharidosis I	3	g.92143164insGGGGGCCG

582	*PDE6B*	Rod-cone dysplasia 1a	3	g.92356052insACTTCAGG
282	*PDE6B*	Rod-cone dysplasia 1	3	g.92356080C>T
528	*PDE6B*	Cone-rod dystrophy 1	3	g.92356091del3
1230	*PDE6B*	PRA	3	g.92358235insCCAGAA
1496	*CDH23*	Deafness	4	g.23074925C>T
88	*RAB24*	Ataxia, cerebellar	4	g.36942678A>C
563	*NIPAL4*	Ichthyosis	4	g.53624074AT>A
1612	*SGCD*	Limb-girdle muscular dystrophy	4	g.54154870A>G
802	*SGCD*	Muscular dystrophy, limb-girdle, type 2F	4	g.54241112del2
475	*PDE6A*	Rod-cone dysplasia 3	4	g.60060370del
1593	*SH3TC2*	Polyneuropathy, hypomyelinating	4	g.60798310C>T
444	*GDNF*	Acral mutilation syndrome	4	g.71791074C>T
795	*SLC45A2*	Coat colour, albinism, oculocutaneous type IV	4	g.74775374AC>A
92	*SLC45A2*	Coat colour, albinism, oculocutaneous type IV	4	g.74777829G>A
1342	*FAM134B*	Neuropathy, sensory	4	g.89424555G>A
1158	*SLC37A2*	Craniomandibular osteopathy	5	g.9406175C>T
411	*SLC37A2*	Craniomandibular osteopathy	5	g.9406431G>A
995	*VPS11*	Neuroaxonal dystrophy	5	g.14803496T>C
614	*CHRNE*	Myasthenic syndrome, congenital	5	g.31912773insC
804	*CHRNE*	Myasthenic syndrome, congenital	5	g.31915099insG
972	*ACADVL*	Exercise induced metabolic myopathy	5	g.32400577C>A
1056	*DVL2*	Screw tail	5	g.32401932del
1536	*GUCY2D*	Progressive retinal atrophy	5	g.33055062insT
535	*HES7*	Spondylocostal dysostosis	5	g.33151478del
89	*FAM83G*	Hyperkeratosis, palmoplantar	5	g.41270206G>C
77	*FLCN*	Renal cystadenocarcinoma and nodular dermatofibrosis	5	g.42401247A>G
997	*MC1R*	White coat colour	5	g.64186827del2
32	*MC1R*	Grizzle	5	g.64187411C>A
998	*MC1R*	Cream coat colour	5	g.64188072C>G
108	*LOC489707*	Macular corneal dystrophy	5	g.75855833C>A
568	*HSF4*	Cataract, early onset	5	g.82783337insT
456	*HSF4*	Cataract, early onset	5	g.82783337del
58	*GUSB*	Mucopolysaccharidosis VII	6	g.654444G>A
57	*GUSB*	Mucopolysaccharidosis VII	6	g.655445C>T
102	*FAM20C*	Dental hypomineralization	6	g.16600662G>A
1585	*PKD1*	Polycystic kidney disease	6	g.39295382G>T
72	*PKD1*	Polycystic kidney disease	6	g.39301108G>A
466	*AGL*	Glycogen storage disease IIIa	6	g.50508478del
1050	*ABCA4*	Stargardt disease 1	6	g.55639737insC
468	*RPE65*	Leber congenital amaurosis (congenital stationary night blindness)	6	g.77428110del4

417	*PKP1*	Ectodermal dysplasia/skin fragility_syndrome	7	g.1725249C>G
1239	*LAMB3*	Epidermolysis bullosa, junctionalis	7	g.8075232A>G
1543	*TNR*	Dystonia-ataxia syndrome, paroxysmal	7	g.23940978insC
897	*PKLR*	Pyruvate kinase deficiency of erythrocyte	7	g.42254572TC>T
896	*PKLR*	Pyruvate kinase deficiency of erythrocyte	7	g.42255380C>T
894	*PKLR*	Pyruvate kinase deficiency of erythrocyte	7	g.42255429T>C
895	*PKLR*	Pyruvate kinase deficiency of erythrocyte	7	g.42255675G>A
898	*PKLR*	Pyruvate kinase deficiency of erythrocyte	7	g.42256499insAAAAAA
1314	*LOXHD1*	Nonsyndromic hearing loss	7	g.44796966G>C
1356	*MOCOS*	Xanthinuria, type II	7	g.53959096del
1355	*MOCOS*	Xanthinuria, type II	7	g.53964250C>A
1357	*MOCOS*	Xanthinuria, type II	7	g.53971022A>G
1194	*DSG1*	Hyperkeratosis, palmoplantar	7	g.58140382del5
1324	*LAMA3*	Epidermolysis bullosa, junctionalis	7	g.64474839T>A
622	*NKX2-8*	Spinal dysraphism	8	g.15317007delins
735	*L2HGDH*	L-2-hydroxyglutaricacidemia	8	g.26947409delins
427	*L2HGDH*	L-2-hydroxyglutaricacidemia	8	g.26984190T>C
1098	*SIX6*	Eye malformation, congenital	8	g.35850298C>T
37	*SPTB*	Elliptocytosis	8	g.39446734G>A
28	*SEL1L*	Ataxia, cerebellar, progressive early-onset	8	g.54085954A>G
51	*GALC*	Krabbe disease	8	g.59638635T>G
1607	*GALC*	Krabbe disease	8	g.59642523G>A
949	*TTC8*	Golden Retriever PRA 2	8	g.60419810del
95	*TECPR2*	Neuroaxonal dystrophy, juvenile	8	g.70839678C>T
426	*AMN*	Intestinal cobalamin malabsorption	8	g.71205527C>A
954	*SGSH*	Mucopolysaccharidosis IIIA	9	g.2406847del3
270	*GAA*	Glycogen storage disease II	9	g.2466009C>T
76	*PRCD*	Progressive rod-cone degeneration	9	g.5090913C>T
1130	*TSEN54*	Leukodystrophy	9	g.5896106C>T
67	*ARSG*	Neuronal ceroid lipofuscinosis, 4A	9	g.11855051C>A
1232	*GH1*	Dwarfism, growth-hormone_ deficiency	9	g.15094281del8
114	*GFAP*	Alexander disease	9	g.18459694G>A
80	*ITGA2B*	Thrombasthenia	9	g.18929832G>C
1568	*ITGA2B*	Thrombasthenia	9	g.18932487C>T
369	*ITGA2B*	Thrombasthenia	9	g.18932490dup14
44	*G6PC*	Glycogen storage disease Ia	9	g.20011008C>G
1502	*CNP*	Lysosomal storage disease	9	g.20763542GC>G
936	*KRT16*	Palmoplantar keratoderma, nonepidermolytic, focal 1	9	g.21046467delins
1579	*KRT10*	Ichthyosis, epidermolytic	9	g.21814695G>A

364	*KRT10*	Hyperkeratosis, epidermolytic	9	g.21816726G>T
1280	*SGCA*	Muscular dystrophy, limb-girdle, type R3	9	g.26117203G>A
342	*MPO*	Myeloperoxidase deficiency	9	g.32899265G>A
1512	*VMP1*	Cerebellar abiotrophy	9	g.34218228C>A
1034	*INPP5E*	Diffuse cystic renal dysplasia and hepatic fibrosis	9	g.49068035G>A
608	*LHX3*	Pituitary dwarfism	9	g.49251376insACA
60	*ADAMTSL2*	Musladin-Lueke syndrome	9	g.49987474C>T
1124	*COL5A1*	Ehlers-Danlos syndrome, classic type1	9	g.50865039del
1125	*COL5A1*	Ehlers-Danlos syndrome, classic type1	9	g.50891817G>A
937	*SLC27A4*	Ichthyosis	9	g.55256760C>T
39	*DNM1*	Exercise-induced collapse	9	g.55370803C>A
33	*PSMB7*	Harlequin	9	g.58614853T>G
1540	*SDR9C7*	Ichthyosis, non-epidermolytic	10	g.1471341G>A
1576	*CYP27B1*	Vitamin D-deficiency rickets, type IA	10	g.2182971G>T
850	*CYB5R3*	Methemoglobinaemia	10	g.23789913C>T
967	*CYB5R3*	Methemoglobinaemia	10	g.23793901A>C
113	*PLA2G6*	Neuroaxonal dystrophy	10	g.27545305G>A
55	*MYH9*	May-Hegglin anomaly	10	g.29119875G>A
1371	*LARGE*	Muscular dystrophy-dystroglycanopathy	10	g.31373423C>T
548	*CNGA3*	Achromatopsia-2	10	g.45284311del3
97	*CNGA3*	Achromatopsia-2	10	g.45284974G>A
526	*SLC3A1*	Cystinuria, type I-A	10	g.47761267del
268	*SLC3A1*	Cystinuria, type I-A	10	g.47766325C>T
527	*SLC3A1*	Cystinuria, type II-A	10	g.47785676del6
1542	*SLC3A1*	Cystinuria, type I-A	10	46705989A>G
111	*ASPRV1*	Ichthyosis	10	g.69888218A>G
1117	*ADAMTS2*	Ehlers-Danlos syndrome, type VII (Dermatosparaxis)	11	g.2675690C>T
1514	*ADAMTS2*	Ehlers-Danlos syndrome, type VII (Dermatosparaxis)	11	g.2757936G>A
1150	*GDF9*	Fecundity	11	g.21147009G>A
1096	*NME5*	Ciliary dyskinesia, primary	11	g.25839016del
1282	*TYRP1*	Brown	11	g.33376321G>A
797	*TYRP1*	Brown	11	g.33377856T>G
1113	*TYRP1*	Liver	11	g.33385234T>G
1094	*AKNA*	Recurrent inflammatory pulmonary disease	11	g.68985438del4
78	*COL11A2*	Skeletal dysplasia 2 (SD2)	12	g.2807737C>G
616	*PNPLA1*	Ichthyosis	12	g.5599031delins
64	*HCRTR2*	Narcolepsy	12	g.22748455G>A
368	*HCRTR2*	Narcolepsy	12	g.22850983G>A
415	*SNX14*	Cerebellar cortical degeneration, Hungarian Vizsla	12	g.45749113C>T
1421	*HACE1*	Ataxia	12	g.62609465del

478	*VPS13B*	Trapped Neutrophil Syndrome	13	g.1412654del3
112	*MTBP*	Periodic Fever Syndrome	13	g.19383758G>A
474	*NDRG1*	Polyneuropathy	13	g.30226270del10
73	*NDRG1*	Polyneuropathy	13	g.30249614C>A
460	*FAM83H*	Congenital keratoconjunctivitis sicca and ichthyosiform dermatosis	13	g.38072383del
351	*PLEC*	Epidermolysis bullosa, simplex	13	g.38215387C>T
547	*CNGA1*	Progressive retinal atrophy	13	g.44524171del4
570	*KIT*	Coat colour, white spotting	13	g.47906538insA
464	*KIT*	Gastrointestinal stromal tumor	13	g.47940550del6
465	*KIT*	Gastrointestinal stromal tumor	13	g.47940555del6
459	*KIT*	Coat colour, white spotting	13	g.47941194del3
1044	*ENAM*	Amelogenesis imperfecta	13	g.61004719C>T
452	*ENAM*	Amelogenesis imperfecta	13	g.61005994del5
1604	*IBA57*	Necrotising myelopathy	14	g.479737G>A
976	*MKLN1*	Lethal acrodermatitis	14	g.5560588T>G
469	*ABCB1*	Adverse reaction to certain drugs	14	g.13704487del4
1114	*COL1A2*	Osteogenesis imperfecta	14	g.19857649insGCC
852	*COL1A2*	Osteogenesis imperfecta	14	g.19857851G>A
762	*COL1A2*	Osteogenesis imperfecta	14	g.19877627delins
353	*ANLN*	Respiratory distress syndrome	14	g.47912313C>T
1170	*HIVEP3*	Progressive retinal atrophy, Miniature Schnauzer, type 1	15	g.1501554G>A
423	*PPT1*	Neuronal ceroid lipofuscinosis, 1	15	g.2930621G>A
579	*PPT1*	Neuronal ceroid lipofuscinosis, 1	15	g.2953667insT
988	*PTPRQ*	Deafness, unilateral and vestibular dysfunction	15	g.23227946insA
1336	*FGA*	Afibrinogenaemia	15	g.52601736del
531	*FNIP2*	Hypomyelination of the central nervous system	15	g.56323262del
609	*CLCN1*	Myotonia	16	g.6073325insT
1041	*CLCN1*	Myotonia	16	g.6077507T>A
62	*CLCN1*	Myotonia	16	g.6094735G>A
1570	*CLCN1*	Myotonia		
1364	*CLCN1*	Myotonia	16	g.6097599insGAGA
93	*BRAF*	Invasive transitional cell carcinoma of the bladder	16	g.8296284T>A
1445	*LMBR1*	Dew claws	16	g.20112974C>T
74	*KLKB1*	Prekallikrein deficiency	16	g.45427398A>T
960	*ARHGEF10*	Polyneuropathy	16	g.60206059del10
425	*TPO*	Hypothyroidism	17	g.775431insG
273	*TPO*	Hypothyroidism	17	g.785529C>T
50	*TPO*	Hypothyroidism	17	g.800004C>T
407	*TPO*	Hypothyroidism	17	g.802503T>C
809	*POMC*	Obesity	17	g.19431807del14
583	*PCARE*	Rod-cone dysplasia 4	17	g.22937958insG
1358	*XDH*	Xanthinuria, type I	17	g.24970436C>T
571	*FOXI3*	Ectodermal dysplasia	17	g.38316024insCCGCCCG

336	*ITGA10*	Chondrodysplasia, disproportionate short-limbed	17	g.59361295G>A
1470	*PNPLA8*	Hereditary ataxia	18	g.12379523insTT
1580	*RELN*	Lissenecephaly and cerebellar hypoplasia	18	g.16909943del
986	*NAPEPLD*	Leukoencephalomyelopathy	18	g.17261361insC
985	*NAPEPLD*	Leukoencephalomyelopathy	18	g.17261558G>C
1454	*EPS8L2*	Early onset adult deafness	18	g.26126092del12
284	*RAG1*	Severe combined immunodeficiency disease	18	g.32033642C>A
49	*CAT*	Hypocatalasia	18	g.33795164C>T
980	*EXT2*	Osteochondromatosis	18	g.45534191G>T
66	*CTSD*	Neuronal ceroid lipofuscinosis, 10	18	g.46468292C>T
1596	*KCNQ1*	Long QT syndrome	18	g.47058836C>A
1157	*UNC93B1*	Exfoliative cutaneous lupus erythematosus	18	g.50337804C>A
457	*SPTBN2*	Ataxia, spinocerebellar	18	g.51168595del8
1481	*EHBP1L1*	Dyserythropoietic anemia and myopathy syndrome (DAMS)	18	g.52123539del
1483	*EHBP1L1*	Congenital dyserythropoietic anemia and polymyopathy	18	g.52128140G>A
105	*CAPN1*	Ataxia, spinocerebellar	18	g.52515250C>T
585	*RASGRP2*	Thrombopathia	18	g.52923455insA
477	*RASGRP2*	Thrombopathia	18	g.52923511del3
285	*RASGRP2*	Thrombopathia	18	g.52925445C>T
576	*FERMT3*	Leukocyte adhesion deficiency, type III	18	g.53343333insGGCAGCCGTCTT
737	*BEST1*	Multifocal retinopathy 3	18	g.54470590del
59	*BEST1*	Multifocal retinopathy 2	18	g.55078249C>T
275	*BEST1*	Multifocal retinopathy 1	18	g.55080684G>A
551	*MFSD8*	Neuronal ceroid lipofuscinosis, 7	19	g.13821469del
398	*BIN1*	Inherited myopathy of Great Danes	19	g.24872877A>G
546	*RAB3GAP1*	Polyneuropathy, ocular abnormalities and neuronal vacuolation	19	g.39392722del
961	*NEB*	Nemaline myopathy	19	g.54495515G>T
29	*RHO*	PRA	20	g.5711695G>C
1301	*IFT122*	Progressive retinal atrophy	20	g.5722348C>T
1274	*CCDC66*	Progressive retinal atrophy, early onset	20	g.33966929insT
574	*CCDC66*	Generalized PRA	20	g.33994683insT
357	*COL7A1*	Epidermolysis bullosa, dystrophic	20	g.40940557C>T
38	*COL7A1*	Epidermolysis bullosa, dystrophic	20	g.40946544G>A
1436	*FYCO1*	Juvenile cataract	20	g.42952995del
1037	*SLC5A5*	Congenital dyshormonogenic hypothyroidism with goiter	20	g.45432681C>T
1434	*DNM2*	Centronuclear myopathy 1	20	g.50822943G>A
94	*ATG4D*	Neurodegenerative vacuolar storage disease	20	g.51018057C>T

43	*ADAMTS10*	Glaucoma, primary open angle	20	g.53503415C>T
101	*ADAMTS10*	Glaucoma, primary open angle	20	g.53508972C>T
455	*C3*	C3 deficiency	20	g.53972803del
565	*DIRAS1*	Epilepsy, generalized myoclonic, with photosensitivity	20	g.56946244del4
1591	*MTMR2*	Polyneuropathy, hypomyelinating	21	g.5387227G>A
1247	*TYR*	Himalayan	21	g.10984640C>T
1079	*MYO7A*	Deafness, bilateral, and vestibular dysfunction	21	g.10984641G>A
473	*TPP1*	Neuronal ceroid lipofuscinosis, 2	21	g.33682425C>A
1233	*SBF2*	Polyneuropathy	21	g.43292574del2
1080	*SLC6A5*	Hyperekplexia (Startle disease)	21	g.30662007C>T
279	*CLN5*	Neuronal ceroid lipofuscinosis, 5	22	g.30662319del2
541	*CLN5*	Neuronal ceroid lipofuscinosis, 5	22	g.60736226G>A
40	*F7*	Factor VII deficiency	22	g.2774932del14
1388	*ABHD5*	Ichthyosis	23	g.3949837G>A
41	*GLB1*	Gangliosidosis, GM1	23	g.3991908TC>T
462	*GLB1*	Gangliosidosis, GM1	23	g.3991946insGGATCCCAGACTTGCCCCA
573	*GLB1*	Gangliosidosis, GM1	23	g.33682425C>A
1151	*LOC608697*	Myasthenic syndrome, congenital	23	g.27687716G>A
900	*LOC608697*	Myasthenic syndrome, congenital	23	g.27688894T>C
1215	*HPS3*	Cocoa	23	g.44487038G>A
454	*P2RY12*	Bleeding disorder	23	g.46454857del3
98	*TUBB1*	Thrombocytopaenia	24	g.44509479G>A
81	*TUBB1*	Thrombocytopaenia	24	g.44514320G>A
581	*COL9A3*	Oculoskeletal dysplasia 1	24	g.47503144insG
1092	*COL9A3*	Oculoskeletal dysplasia 1	24	g.47509795C>T
1489	*MFF*	Mitochondrial fission encephalopathy	25	g.40322999Tdelins
578	*SLC19A3*	Necrotising encephalopathy, subacute, of Leigh	25	g.40615336delins
359	*SAG*	Progressive retinal atrophy	25	g.45028946T>C
1208	*COL6A3*	Muscular dystrophy	25	g.48289626C>T
1207	*COL6A3*	Muscular dystrophy	25	g.48296579G>A
1216	*MLPH*	Dilute	25	g.48431713insC
948	*MLPH*	Dilute	25	g.48431759G>C
75	*AGXT*	Primary hyperoxaluria type I	25	g.51027024G>A
1533	*ATP2A2*	Darier disease	26	g.8434781A>C
552	*SCARF2*	Van den Ende-Gupta syndrome	26	g.30405793del2
84	*VWF*	Von Willebrand disease II	27	g.7168042T>G
371	*VWF*	Von Willebrand disease III	27	g.7200327C>A
803	*VWF*	Von Willebrand disease II	27	g.7205292A>G
968	*VWF*	Von Willebrand disease III	27	g.7223626del
479	*VWF*	Von Willebrand disease III	27	g.7244427del
1083	*NECAP1*	Progressive retinal atrophy	27	g.8634115C>A
1574	*ABCC9*	Cardiomyopathy, dilated	27	g.21213534C>T
1172	*YARS2*	Cardiomyopathy and juvenile mortality	27	g.30417127C>A

537	*ADAMTS20*	Cleft lip with or without cleft palate	27	g.35997332del2
422	*ANO6*	Platelet receptor for factor X	27	g.37639307G>T
370	*VDR*	Vitamin D-deficiency rickets, type II	27	g.39700521del
45	*PFKM*	Glycogen storage disease VII	27	g.39963602C>A
271	*PFKM*	Glycogen storage disease VII	27	g.39974411G>T
1583	*LMBR1L*	Hyposegmentation of granulocytes	27	g.41169674C>T
1077	*SCN8A*	Ataxia, spinocerebellar	27	g.43476519G>A
1480	*KRT5*	Epidermolysis bullosa, simplex	27	g.44080887C>T
1043	*KRT71*	Curly coat	27	g.44109043delins
1495	*KRT1*	Ichthyosis	27	g.44229724del3
280	*AMHR2*	Persistent Mullerian duct syndrome	27	g.44857716C>A
61	*CHAT*	Myasthenic syndrome, congenital	28	g.1628555G>A
993	*RBP4*	Microphthalmia, isolated, with coloboma	28	g.8009398del3
283	*PRKDC*	Severe combined immunodeficiency disease, autosomal	29	g.306491C>A
1063	*SGK3*	Hypotrichosis, recessive	29	g.16820435insT
564	*SGK3*	Hypotrichosis, recessive	29	g.16835127del4
27	*CNGB3*	Achromatopsia	29	g.33481715C>T
281	*PDP1*	Pyruvate dehydrogenase deficiency	29	g.39443849C>T
1218	*SLC12A6*	Ataxia, spinocerebellar	30	g.1045146delins
1372	*MYO5A*	Coat colour dilution and neurological defects	30	g.18294886insT
68	*CLN6*	Neuronal ceroid lipofuscinosis, 6	30	g.32679521A>G
26	*HEXA*	Gangliosidosis, GM2, type I	30	g.36282814C>T
356	*BBS4*	Bardet-Biedl syndrome 4	30	g.36504244A>T
1291	*POU1F1*	Pituitary dwarfism	31	g.1099273C>A
1302	*SOD1*	Paroxysmal dyskinesia, juvenile	31	g.27118846delins
87	*SOD1*	Degenerative myelopathy	31	g.27118886A>T
53	*ITGB2*	Leukocyte adhesion deficiency, type I	31	g.38480401C>G
340	*COL6A1*	Muscular dystrophy, Ullrich type	31	g.39284371G>T
1260	*LRIT3*	Night blindness, congenital stationary	32	g.9852193del
1072	*MANBA*	Beta mannosidosis	32	g.15735654T>A
1093	*MANBA*	Beta mannosidosis	32	g.15825842insTCACT
1373	*PRKG2*	Dwarfism, disproportionate	32	g.34701592G>A
862	*BMP3*	Brachycephaly	32	g.5231894C>A
104	*FGF5*	Long hair	32	g.35475211G>T
952	*FGF5*	Long hair	32	g.35475232del16
950	*FGF5*	Long hair	32	g.35475228insCC
418	*FGF5*	Long hair	32	g.35486609A>T
606	*IQCB1*	Cone-rod dystrophy 2	33	g.25712644insC
1515	*PCYT1A*	Skeletal dysplasia 3	33	g.30067814A>G
266	*CCDC39*	Ciliary dyskinesia, primary	34	g.14126119G>A
1168	*LAMP3*	Surfactant metabolism dysfunction, pulmonary	34	g.16264641C>T
1240	*ALDH5A1*	Succinic Semialdehyde Dehydrogenase Deficiency	35	g.23984370G>A

1526	*SCN9A*	Congenital insensitivity to pain	36	g.11652662G>A
1471	*SLC25A12*	Cerebellar Degeneration-Myositis Complex	36	g.16504064G>A
1262	*SLC25A12*	Inflammatory myopathy	36	g.16530334A>G
65	*ATF2*	Neonatal encephalopathy with seizures	36	g.19402063A>C
470	*MSTN*	Muscular hypertrophy (double muscling)	37	g.630059del2
1527	*STK36*	Primary ciliary dyskinesia	37	g.25167072G>A
575	*SLC4A3*	Golden Retriever PRA 1	37	g.26021552insC
971	*CLN8*	Neuronal ceroid lipofuscinosis, 8	37	g.30759848insT
69	*CLN8*	Neuronal ceroid lipofuscinosis, 8	37	g.30760000T>C
1374	*MIA3*	Dental-skeletal-retinal anomaly	38	g.17616670del2
1592	*MPZ*	Polyneuropathy, hypomyelinating	38	g.22037876T>C
945	*KCNJ10*	Ataxia, cerebellar	38	g.22969668C>G
947	*KCNJ10*	Spongy degeneration with cerebellar ataxia 1	38	g.22970027T>C
1614	*DMD*	Muscular dystrophy, Duchenne	X	g.26939052G>A
367	*DMD*	Muscular dystrophy, Duchenne	X	g.27122234C>T
562	*DMD*	Muscular dystrophy, Duchenne	X	g.27615619del7
1616	*DMD*	Muscular dystrophy, Duchenne	X	g.27774668insT
542	*DMD*	Muscular dystrophy, Duchenne	X	g.27778709del
1236	*DMD*	Muscular dystrophy, Duchenne	X	g.27794550G>A
366	*DMD*	Muscular dystrophy, Duchenne	X	g.28126496T>C
480	*RPGR*	Progressive retinal atrophy	X	g.33219538del5
1315	*NDP*	Retinal dysplasia	X	g.38158131insC
1017	*EDA*	X-linked hypohidrotic ectodermal dysplasia	X	g.54703152del
361	*EDA*	Anhidrotic ectodermal dysplasia	X	g.54705078G>A
584	*IL2RG*	Severe combined immunodeficiency disease, X-linked	X	g.55679174insG
476	*IL2RG*	Severe combined immunodeficiency disease, X-linked	X	g.55680372del4
107	*ATP7A*	Modifier of copper toxicosis	X	g.60422217C>T
82	*PLP1*	Tremor, X-linked	X	g.77750792A>C
471	*COL4A5*	Nephritis, X-linked	X	g.82740164del10
276	*COL4A5*	Nephritis, X-linked	X	g.82802716G>T
1039	*F9*	Haemophilia B	X	g.110486026del
467	*F9*	Haemophilia B	X	g.110505881delins
47	*F9*	Haemophilia B	X	g.110515405G>A
1363	*F9*	Haemophilia B	X	g.110516121insA
46	*F9*	Haemophilia B	X	g.110516555G>A
63	*MTM1*	Myotubular myopathy 1	X	g.120191588C>A
91	*MTM1*	Myotubular myopathy 1	X	g.120207756A>C
1459	*MTM1*	Myotubular myopathy 1	X	g.120209973C>T
1243	*NSDHL*	Verrucous epidermal keratinocytic nevi	X	g.122058735G>A
1268	*NSDHL*	Verrucous epidermal keratinocytic nevi	X	g.122058751del5

1588	*F8*	Haemophilia A	X	g.124075495del
350	*F8*	Haemophilia A	X	g.124075774del2
100	*F8*	Haemophilia A	X	g.124091989G>A
99	*F8*	Haemophilia A	X	g.124094178C>T
272	*F8*	Haemophilia A	X	g.124099739G>C

Description of OMIA-derived variants not segregating in the WGS sample.

Note: *Chr = chromosome; del = deletion; ins = insertion; delins = deletion insertion.*

[a] *As reported by the OMIA database.*

[b] *Genomic positions in canFam4.*

B.4 Supplementary results from high-impact variant survey

Table A8: Summary of high-impact variants within previously reported functional genes

Gene	Gene function [a]	Chr	Observed variant	Allele frequency [b]	Variant call validation [c]	Functional assessment [d]
HSD17B3	XY DSD	1	g.71074374ins5	0.244	Likely false variant call	-
SLC7A10	Paradoxical pseudomyotonia	1	g.119761953A>C	0.004	Likely false variant call	-
SLC7A10	Paradoxical pseudomyotonia	1	g.119762155del	0.086	Possible true variant	Unlikely to be functional
SLC7A10	Paradoxical pseudomyotonia	1	g.119762157del	0.914	Possible true variant	Unlikely to be functional
CUBN	Cobalamin malabsorption	2	g.18781859ins84	0.09	True variant	Unlikely to be functional
CUBN	Cobalamin malabsorption	2	g.18932444del	0.012	True variant	Functional candidate
CNGB1	Retinal atrophy	2	g.57885286T>G	0.014	Likely false variant call	-
CNGB1	Retinal atrophy	2	g.57909801del	0.006	True variant	Functional candidate
FUCA1	Fucosidosis	2	g.75059692ins2	0.052	Likely false variant call	-
ARSB	Mucopolysaccharidosis	3	g.28048573ins	0.054	Likely false variant call	-
ARSB	Mucopolysaccharidosis	3	g.28048574ins	0.946	Likely false variant call	-
AP3B1	Neutropenia	3	g.28829080ins	0.006	Likely false variant call	-
SGCD	Muscular dystrophy	4	g.54151040C>T	0.194	Likely false variant call	-
SGCD	Muscular dystrophy	4	g.54545756delins	-	Likely false variant call	-
CHRNE	Myasthenic syndrome	5	g.31914919T>G	0.002	Likely false variant call	-
HES7	Spondylocostal dysostosis	5	g.33149996T>C	0.018	Likely false variant call	-
HSF4	Cataract	5	g.82783329delins	-	Likely false variant call	-
PKD1	Polycystic kidney disease	6	g.39300812A>C	-	Likely false variant call	-
ABCA4	Stargardt disease	6	g.55641338del	0.358	Likely false variant call	-
ABCA4	Stargardt disease	6	g.55641342del	0.642	Likely false variant call	-
ABCA4	Stargardt disease	6	g.55659310del2	0.006	True variant	Functional candidate
TNR	Dystonia–ataxia syndrome	7	g.23985012del	0.986	Likely false variant call	-
PKLR	PK deficiency	7	g.42258229A>C	0.022	Likely false variant call	-

PKLR	PK deficiency	7	g.42258230G>C	0.014	Likely false variant call	-
SGSH	Mucopolysaccharidosis	9	g.2406797ins	0.01	True variant	Functional candidate
CNP	Lysosomal storage disease	9	g.20768542ins	0.002	True variant	Functional candidate
MPO	Myeloperoxidase deficiency	9	g.32906176A>C	0.006	Likely false variant call	-
COL5A1	Ehlers-Danlos syndrome	9	g.50765323ins	0.002	Likely false variant call	-
COL5A1	Ehlers-Danlos syndrome	9	g.50881947del	0.238	Possible true variant	Unlikely to be functional
COL5A1	Ehlers-Danlos syndrome	9	g.50881951del	0.762	Possible true variant	Unlikely to be functional
SLC3A1	Cystinuria	10	g.47766395G>A	0.002	True variant	Functional candidate
TYRP1	Brown, liver colour	11	g.33385200C>T	0.028	True variant	Functional candidate
AKNA	Pulmonary disease	11	g.68972045del2	0.018	Likely false variant call	-
NDRG1	Polyneuropathy	13	g.30235368del	0.028	True variant	Unlikely to be functional
NDRG1	Polyneuropathy	13	g.30235410C>T	0.024	True variant	Unlikely to be functional
NDRG1	Polyneuropathy	13	g.30249959C>T	0.022	True variant	Unlikely to be functional
PLEC	Epidermolysis bullosa	13	g.38219603del	0.752	Likely false variant call	-
MKLN1	Lethal acrodermatitis	14	g.5767250T>C	0.178	True variant	Unlikely to be functional
COL1A2	Osteogenesis imperfecta	14	g.19876612C>A	0.006	Likely false variant call	-
PTPRQ	Deafness	15	g.23018717T>C	0.658	True variant	Unlikely to be functional
PTPRQ	Deafness	15	g.23132132del	0.6	Likely false variant call	-
PTPRQ	Deafness	15	g.23132133del	-	Likely false variant call	-
PTPRQ	Deafness	15	g.23240583G>T	0.06	Likely false variant call	-
FGA	Afibrinogenaemia	15	g.52602217ins6	0.032	Likely false variant call	-
FOXI3	Ectodermal dysplasia	17	g.38315594ins	0.006	Likely false variant call	-
FOXI3	Ectodermal dysplasia	17	g.38315635ins6	0.006	Likely false variant call	-
FOXI3	Ectodermal dysplasia	17	g.38315676ins6	0.03	Likely false variant call	-
FOXI3	Ectodermal dysplasia	17	g.38315812del131	0.002	Likely false variant call	-
PNPLA8	Ataxia	18	g.12386685ins40	0.002	Likely false variant call	-
RELN	Lissenecephaly and cerebellar hypoplasia	18	g.17017866del	0.562	Likely false variant call	-

RELN	Lissenecephaly and cerebellar hypoplasia	18	g.17017868del	0.438	Likely false variant call	-
NAPEPLD	Leukoencephalo myelopathy	18	g.17240285del	0.006	Likely false variant call	-
NAPEPLD	Leukoencephalo myelopathy	18	g.17240287delins	-	Likely false variant call	-
CCDC66	Retinal atrophy	20	g.33965111A>G	0.544	True variant	Unlikely to be functional
CCDC66	Retinal atrophy	20	g.33996811C>A	0.03	True variant	Unlikely to be functional
CCDC66	Retinal atrophy	20	g.33996814del4	0.03	True variant	Functional candidate
COL7A1	Epidermolysis bullosa	20	g. 40922028del12	0.494	Likely false variant call	-
SLC6A5	Hyperekplexia	21	g.43263723delins		Likely false variant call	-
SLC6A5	Hyperekplexia	21	g.43280048G>T	0.006	Likely false variant call	-
GLB1	Gangliosidosis	23	g.4009430del	0.002	True variant	Functional candidate
COL9A3	Oculoskeletal dysplasia	24	g.47506941A>C	0.054	Likely false variant call	-
COL6A3	Muscular dystrophy	25	g.48301196del	0.038	Likely false variant call	-
MLPH	Coat colour dilution	25	g.48436229del2	0.004	True variant	Unlikely to be functional
ADAMTS20	Cleft lip	27	g.35899564del13	0.056	True variant	Unlikely to be functional
ADAMTS20	Cleft lip	27	g.35899577C>T	0.35	True variant	Unlikely to be functional
ADAMTS20	Cleft lip	27	g.35960853ins	0.002	Likely false variant call	-
ADAMTS20	Cleft lip	27	g.36041346del14	0.264	Likely false variant call	-
ADAMTS20	Cleft lip	27	g.36041365del4	0.156	Likely false variant call	-
ADAMTS20	Cleft lip	27	g.36057158del4	0.138	Likely false variant call	-
VDR	Rickets	27	g.39696791del	0.612	Likely false variant call	-
PRKDC	SCID	29	g.285754del	0.062	Likely false variant call	-
PRKDC	SCID	29	g.285756del	0.938	Likely false variant call	-
SLC12A6	Ataxia	30	g.1121671del4	0.002	Likely false variant call	-
MYO5A	Neurological defects, colour dilution	30	g.18470586T>G	0.006	Likely false variant call	-
CYP1A2	Metabolizer of a cognitive enhancer	30	g.38260059del	0.062	True variant	Functional candidate

CYP1A2	Metabolizer of a cognitive enhancer	30	g.38261635C>T	0.302	True variant	Functional candidate
CYP1A2	Metabolizer of a cognitive enhancer	30	g.38263975ins10	0.074	True variant	Functional candidate
SOD1	Degenerative myelopathy	31	g.27125694del	0.268	Likely false variant call	-
STK36	Primary ciliary dyskinesis	37	g.25198749G>A	0.194	True variant	Functional candidate
SLC4A3	Retinal atrophy	37	g.26020887del	0.402	Likely false variant call	-
CLN8	NCL	37	g.30769171G>A	0.042	True variant	Functional candidate
F9	Haemophilia B	X	g.110492444A>G	0.002	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122026888ins66	0.012	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122037290T>C	0.016	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122041237ins	0.004	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122051662delins	-	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122055226delins	-	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122055227T>C	0.002	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122057562ins16	0.002	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122057563delins	-	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122058789ins16	0.002	Likely false variant call	-
NSDHL	CHILD-like syndrome	X	g.122058790T>C	0.004	Likely false variant call	-

Description of all predicted high-impact variants that segregated in the WGS sample within OMIA-derived genes of interest (see Table A3).

Note: *Chr = chromosome; del = deletion; ins = insertion; delins = deletion insertion.*

[a] *OMIA-reported phenotypes of LOF variants within genes of interest.*

[b] *Allele frequencies of predicted high-impact variants in the WGS dataset.*

[c] *Reclassification of variant calls in WGS sample based on a visual inspection of sequence reads in IGV.*

[d] *Functionality classification of high-impact variants based on sample allele frequencies and a review of the literature.*

B.5 Supplementary results from the identification of predictive markers

Table A9: Array SNPs in LD with Mendelian variants of interest

Mendelian variant [a]			Array SNP [b]		
OMIA ID	**Gene**	**Position**	**ID**	**Position**	**R^2 [c]**
1603	*MC5R*	Chr1:24541931	AX-168165074	Chr1: 24529159	0.268208
			AX-168217582	Chr1:24538731	0.351787
			AX-167569213	Chr1:24538991	0.384123
			AX-167275855	Chr1:24541699	0.372851
			AX-167319705	Chr1:24542076	0.591467
			AX-167826052	Chr1:24582879	0.423482
434	*MC1R*	Chr5:64186728	AX-168038747	Chr5:64156728	0.205621
			AX-167210461	Chr5:64158549	0.20696
			AX-168202811	Chr5:64190343	0.228851
34	*MC1R*	Chr5:64186854	AX-167328853	Chr5:64156061	0.829712
			AX-167843375	Chr5:64188925	0.240911
			AX-167695536	Chr5:64195666	0.727896
			AX-168128802	Chr5:64202691	0.246478
851	*BTBD17*	Chr9:6924623	AX-167990882	Chr9:6905339	0.280129
			AX-167798293	Chr9:6909358	0.281276
			AX-167644637	Chr9:6914185	0.294621
			AX-167499207	Chr9:6914652	0.319817
			AX-168167782	Chr9:6918372	0.362917
			AX-167418340	Chr9:6918663	0.366199
			AX-168006758	Chr9:6923237	0.702527
			AX-167947136	Chr9:6925353	0.555025
442	*ABCB1*	Chr14:13720387	AX-167349571	Chr14:13720699	0.272012
			AX-167348610	Chr14:13735369	0.338956
			AX-168108844	Chr14:13735843	0.755786
1422	*IGF1-AS*	Chr15:41511739	AX-167372994	Chr15:41494457	0.212528
			AX-167183545	Chr15:41508682	1
			AX-167556446	Chr15:41510969	0.362584
			AX-167351336	Chr15:41511739	1
			AX-167832588	Chr15:41516150	0.848001
			AX-168167499	Chr15:41521682	0.835984
			AX-167973541	Chr15:41521823	0.24489
1444	*LMBR1*	Chr16:20112737	AX-167191180	Chr16:20109337	0.473416
			AX-167482611	Chr16:20126773	0.473416
			AX-167163431	Chr16:20141105	0.473416
458	*CBD103*	Chr16:55468987	AX-167740876	Chr16:55458108	0.304815
			AX-167495458	Chr16:55466525	0.524995
			AX-167911668	Chr16:55471931	0.233658
1522	*RETN*	Chr20:52842420	AX-167560012	Chr20:52822752	1
			AX-168149666	Chr20:52828694	0.602963
			AX-167474199	Chr20:52851969	0.49436
			AX-167937375	Chr20:52853835	0.683262

1081	*MFSD12*	Chr20:56247895	AX-167305263	Chr20:56231108	0.344169
			AX-167783908	Chr20:56244201	0.327369
			AX-167791764	Chr20:56247719	0.763504
			AX-168153313	Chr20:56247895	1
			AX-167650844	Chr20:56250841	0.701249
			AX-167492583	Chr20:56258073	0.324604
106	*ATP7B*	Chr22:196868	AX-167215741	Chr22:218897	0.358215
360	*MLPH*	Chr25:48403161	AX-168168867	Chr25:48392892	0.390486
			AX-168264347	Chr25:48401715	0.350297
			AX-167400210	Chr25:48413896	0.343477
			AX-167643417	Chr25:48425889	0.475635
			AX-167774718	Chr25:48428512	0.803071
35	*KRT71*	Chr27:44113063	AX-167259693	Chr27:44094380	0.204838
			AX-167751297	Chr27:44098284	0.211862
274	*CYP1A2*	Chr30:38261635	AX-167246204	Chr30:38245067	0.233848
			AX-167363062	Chr30:38258002	0.857936
			AX-168219166	Chr30:38259871	0.326045
			AX-168196296	Chr30:38261635	1
			AX-168139663	Chr30:38263963	0.465848
			AX-167360486	Chr30:38273320	0.462619
			AX-167535335	Chr30:38280024	0.307743
36	*SOD1*	Chr31:27123057	AX-167251279	Chr31:27133605	1
			AX-168070830	Chr31:27162789	0.309912
48	*FGF5*	Chr32:35494497	AX-167434147	Chr32:35494497	1
103	*P3H2*	Chr34:22231216	AX-167299299	Chr34:22231216	1
612	*KCNJ10*	Chr38:22970388	AX-167782545	Chr38:22966455	0.60228
			AX-167998824	Chr38:22968796	1
			AX-167374580	Chr38:22969143	0.961247
			AX-168104897	Chr38:22971237	0.297558

Variants aggregated from OMIA that segregated in the WGS sample and correlated with $R^2 > 0.2$ with a SNP on the Axiom™ Canine HD Array.

Note: *Chr = chromosome.*

[a] *Mendelian variants segregating in the WGS sample.*

[b] *SNPs on the array.*

[c] *R^2 allele count correlations between Mendelian variants and array SNPs.*

B.6 Density plots of body size traits by breed

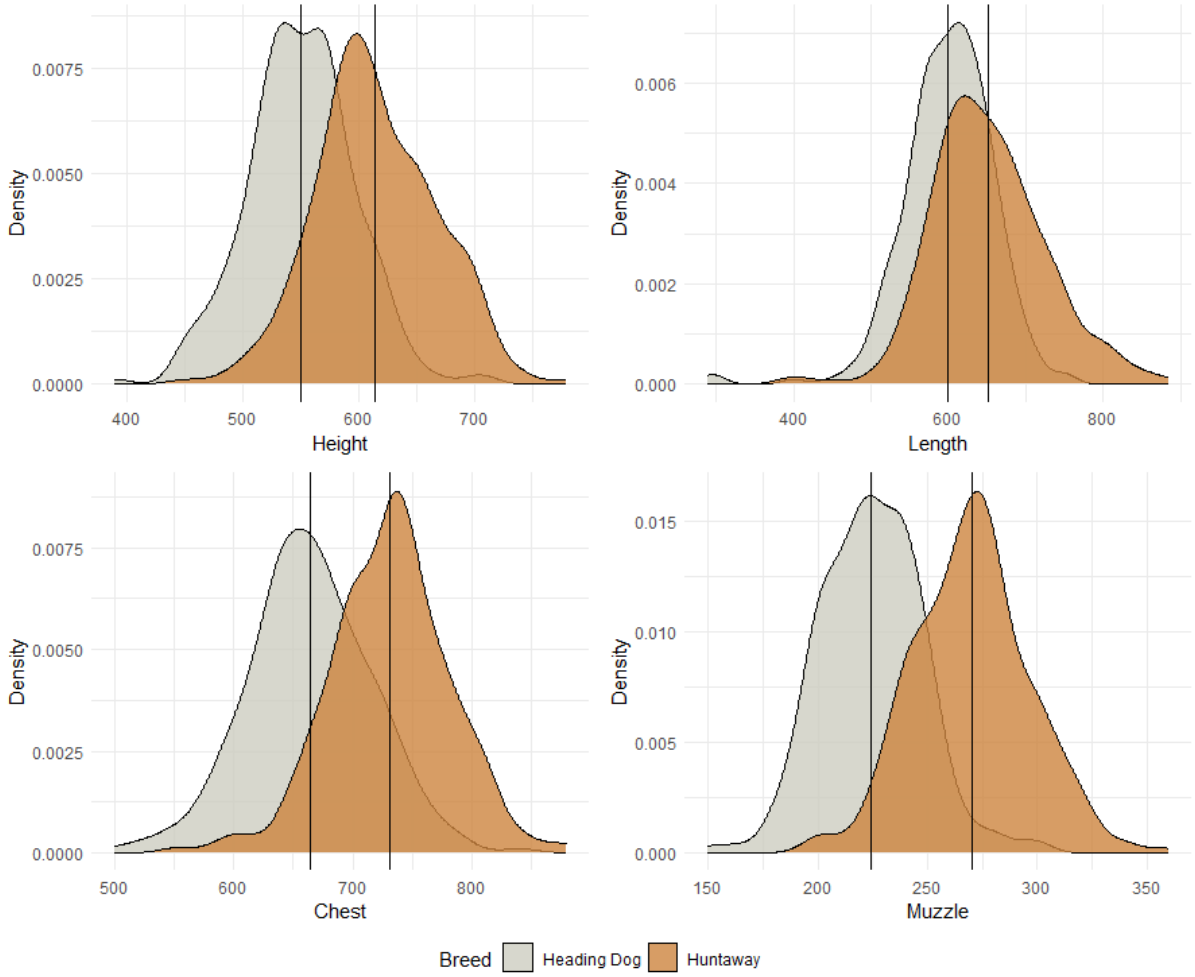

Figure A5. Density plots of body size traits by breed. **a)** Height, **b)** Length, **c)** Chest circumference, **d)** Muzzle circumference. Black lines indicate breed-specific means. These distributions include all purebred Huntaways and Heading Dogs above the age of 1 that had measurements available at the time of the analysis (n = 261 Huntaways and 255 Heading Dogs).

B.7 Manhattan plots from PLINK GLM GWAS

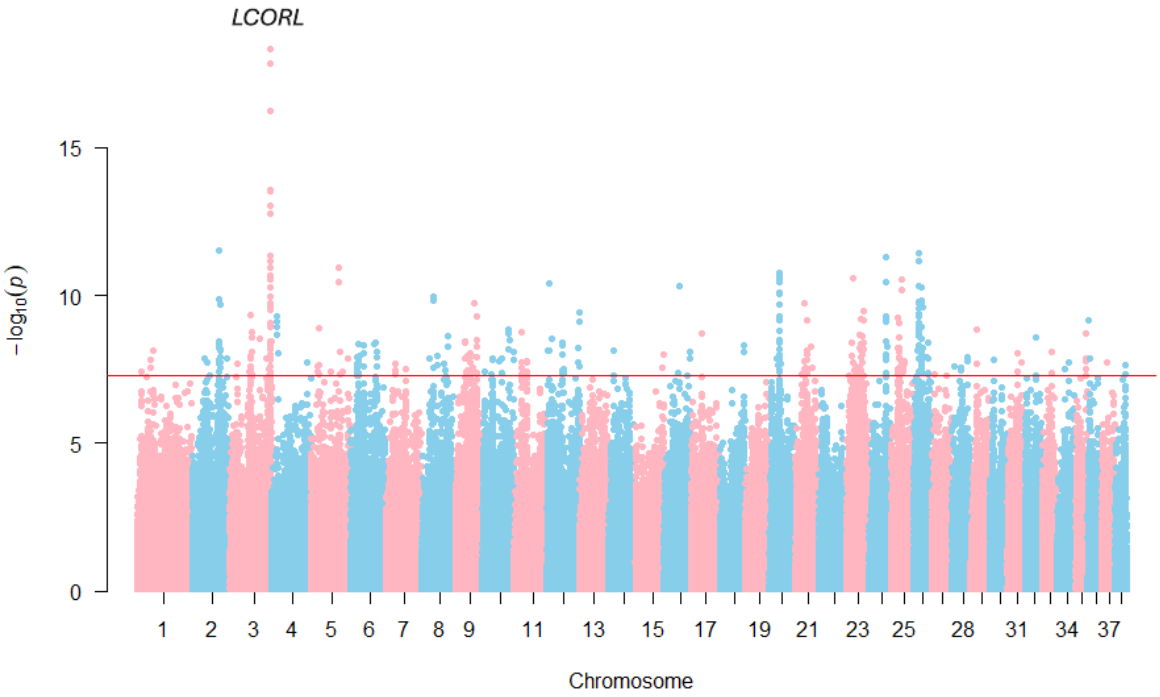

Figure A6. Manhattan plot from a height GLM GWAS. There was a strong peak was near the *LCORL* gene. **Note:** The red line indicates the genome-wide significance threshold ($-\log_{10}(p) = -\log_{10}(5\times10^{-8})$).

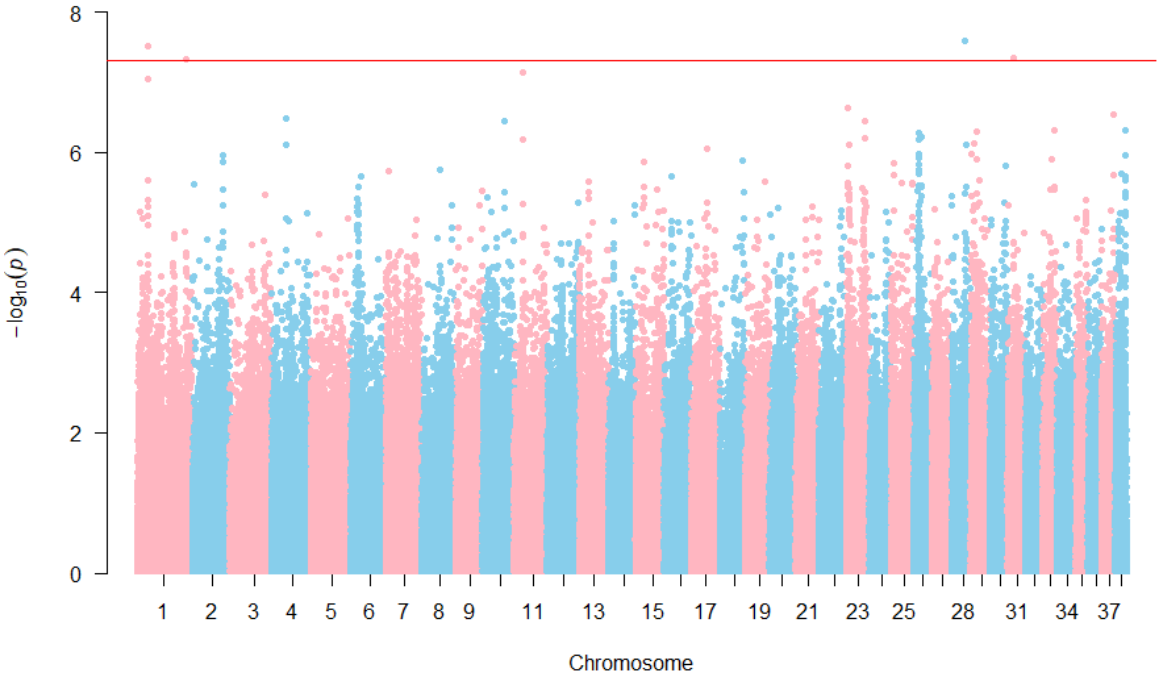

Figure A7. Manhattan plot from a length GLM GWAS. There were no strong peaks. **Note:** The red line indicates the genome-wide significance threshold ($-\log_{10}(p) = -\log_{10}(5\times10^{-8})$).

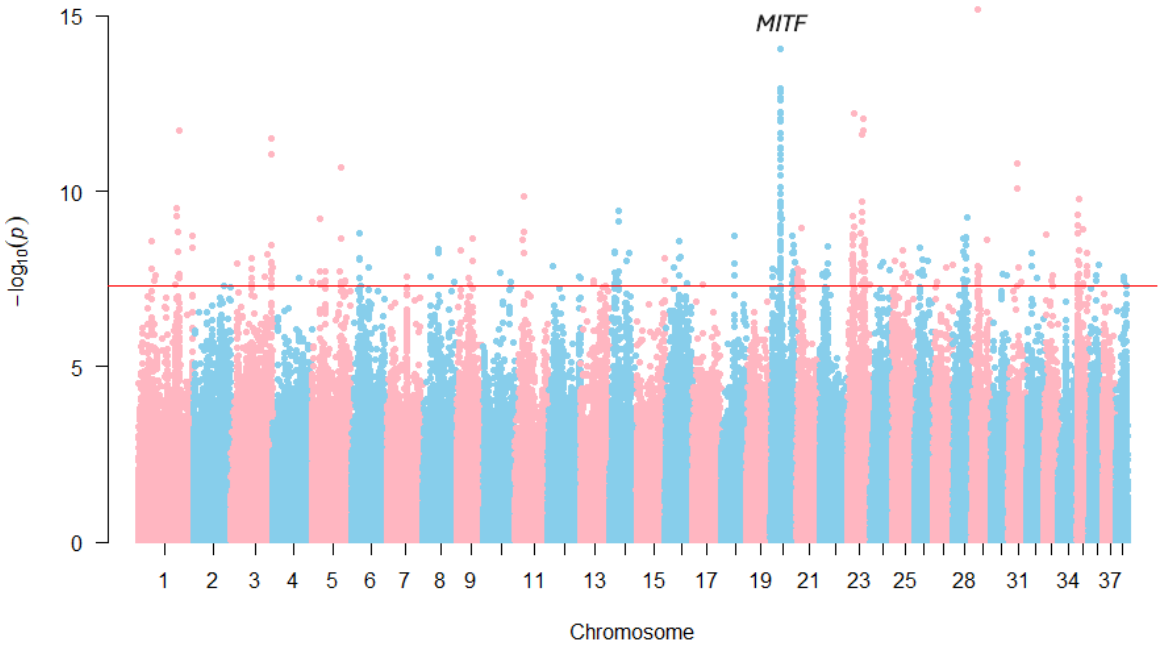

Figure A8. Manhattan plot from a chest circumference GLM GWAS. There was a strong peak near the *MITF* gene. **Note:** The red line indicates the genome-wide significance threshold ($-\log_{10}(p) = -\log_{10}(5\times10^{-8})$).

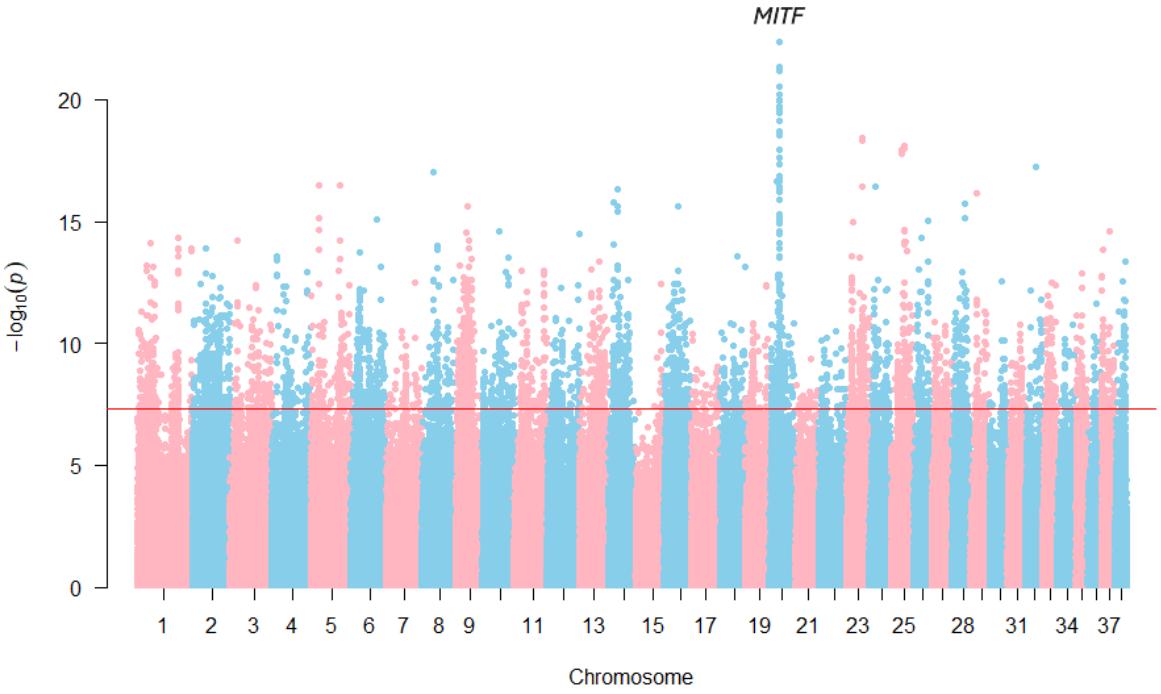

Figure A9. Manhattan plot from a muzzle circumference GLM GWAS. There was a strong peak near the *MITF* gene. **Note:** The red line indicates the genome-wide significance threshold ($-\log_{10}(p) = -\log_{10}(5\times10^{-8})$).