THE GENETIC ARCHITECTURE OF VARIATION IN HUMANS AND DOGS

by

ELDON GOODWIN PRINCE

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2014

Acknowledgements

I am grateful for the guidance of my two advisors at UT Arlington, Trey Fondon and Jeff Demuth. I appreciate Jeff's mentoring and patience during the first two years of my degree. Trey created an intellectually demanding environment that has helped me to grow personally and professionally; I am grateful for his insights that heavily influence this work, and hope to have assimilated some of his patterns of critical thinking during the past few years. I thank my I-Engage students Emil Burdeos, Courtney Elledge, and Bethany Minter for their hard work and participating in many enlightening and fun conversations. Kristen Correll and Michael Chau contributed substantially to this work by helping to measure and genotype mixed-breed dogs. Shreyas Krishnan and many other graduate students and lab members have been genuine friends to me and my family; your kindness will not soon be forgotten. In addition, I thank Esther Betrán, André Pires da Silva, Todd Castoe, and Cédric Feschotte for valuable instruction and feedback. I also appreciate the Biology Department office staff for their assistance. Throughout this process I have received constant inspiration from my supremely patient and dedicated wife, Stephanie Prince. I could not have completed this work without her listening ear, lucid questions, and detailed proofreading. I thank my children Sophie, Elinor, and Goodwin for their inspiration and being willing to live so long without a backyard. I am grateful for the continuous support I have received from my siblings Edd, Patti, Sarabeth, John, Daniel, Joseph, Dallan, and Pauline. I am indebted to David and Pam Richards for their love and for letting me take their daughter to Texas. My parents Leon and Leah Prince have been pillars of strength and I thank them for their encouragement. Finally, I recognize the omnipresent role my grandparents Douglas and Beth Prince and Edd and Pauline Goodwin have played in shaping who I am.

April 18, 2014

Abstract

THE GENETIC ARCHITECTURE OF VARIATION IN HUMANS AND DOGS

Eldon Goodwin Prince, Ph.D.

The University of Texas at Arlington, 2014

Supervising Professor: John W. Fondon III, Ph.D.

Genetic architecture is broadly defined as the structure of how genes come together to produce phenotypes. Primary aspects of genetic architecture include how many and which genes contribute to phenotypic variation. The genetic architecture of human height has been studied for over a century; indeed it is the classic quantitative trait with hundreds of contributing variants. As genome-wide studies of genetic architecture are extended beyond just humans, the genetic basis of polygenic traits like height can be compared between species. Such interspecies comparisons reveal how many of the same loci contribute to variation within each species. The extent to which the same loci contribute to intraspecific variation depends on species relatedness and reflects underlying constraints on genetic variability and variation.

In this study genome-wide associations are compared between humans and dogs to estimate how many of the same loci contribute to intraspecific height variation. Due to the highly polygenic nature of height variation, one might predict that relatively few loci will be shared between species as distantly related as the human and dog. Contrary to this prediction, I find that at least 25 orthologous regions contribute to intraspecific height variation in humans and dogs, indicating perhaps less obvious constraints on genetic variability and variation.

Height is decomposed in dogs using genome-wide associations to identify loci that are associated with limb, torso, and neck variation. To extend this approach, several height QTLs are correlated with bone measurements in an independent panel of mixed-breed dogs. The prevailing interpretation that morphological traits are genetically simple in dogs relative to humans is then tested. Central to the interpretation of genetic simplicity in dogs is the story of *IGF1*, a gene thought to explain the majority of size variation. The QTL effect size of *IGF1* is tested in the aforementioned panel of mixed-breed dogs and I find that it explains much less variation than previously reported. This experiment and others call into question the Mendelian effect size previously attributed to *IGF1* and the associated interpretation of genetic simplicity for dog morphology.

One of the evolutionary forces that can impact the genetic architecture of traits is meiotic recombination. Preceding the exchange of genetic material between homologous chromosome pairs, double-strand breaks occur via proteins like Spo11 in yeast. Since all crossovers are the result of double-strand breaks, and these breaks are non-randomly distributed throughout the genome, many researchers have sought to understand the process that regulates where double-strand breaks occur. In addition, although not all double-strand breaks result in genetic crossover, the DNA repair process to rejoin them can be both biased and mutagenic. The protein PRDM9 is associated with almost all meiotic double-strand breaks in mice and is thought to play a similarly central role in humans, although the protein is absent in canids. Curiously, the loss of PRDM9 in the canid lineage also coincides with a genome-wide destabilization of repetitive GC content. I conclude this work with a study of the consequences of losing the meiotic recombination-associated protein PRDM9 and the mutagenic role this loss has likely had in shaping the canid lineage.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

The Extent of Height QTL Sharing in Humans and Dogs

Introduction

Genetic architecture is commonly defined as the structure of how genes come together to produce phenotypes. This metaphor can be instructive in some contexts, although in others it is admittedly inadequate and confounding. Thomas Hansen offers a somewhat more comprehensive definition: "Genetic architecture refers to the pattern of genetic effects that build and control a given phenotypic character and its variational properties" (Hansen, 2006). This definition can be broken into two parts: 1) the genes that build and control a phenotypic character and 2) the genetic effects that influence its variational properties. While it may be instructive and of interest to understand all of the genes that build and control phenotypes, geneticists are primarily concerned with genetic and phenotypic variation. For the scope and purpose of this dissertation, Hansen's definition of genetic architecture is modified and limited as follows: the pattern of genetic effects that influence the variational properties of a phenotype.

Aspects of genetic architecture can be studied at multiple levels, including within and among species and populations. Comparisons of genetic architecture within species and among populations may uncover the recent action of evolutionary forces like natural selection (Pickrell et al., 2009), whereas comparisons of genetic architecture among more distant taxa generally focus on convergent or parallel evolutionary patterns that span tens of millions of years (Conte et al., 2012).

Trudy Mackay provides the following checklist that must be completed in order to understand the genetic architecture of quantitative traits (Mackay, 2001):

> (a) the numbers and identities of all genes in the developmental, physiological and/or biochemical pathway leading to the trait phenotype
>
> (b) the mutation rates at these loci

(c) the numbers and identities of the subset of loci that are responsible for variation in the trait within populations, between populations, and between species

(d) the homozygous and heterozygous effects of new mutations and segregating alleles on the trait;

(e) all two-way and higher-order epistatic interaction effects

(f) the pleiotropic effects on other quantitative traits, most importantly reproductive fitness

(g) the extent to which additive, dominance, epistatic, and pleiotropic effects vary between the sexes, and in a range of ecologically relevant environments

(h) the molecular polymorphism(s) that functionally define QTL alleles

(i) the molecular mechanism causing the differences in trait phenotype

(j) QTL allele frequencies

She then quickly adds that this daunting list has yet to be completed for any trait. Under the revised definition of genetic architecture I provided earlier, requirement (a) is unnecessary because only the loci with variation within and among populations and species are immediately relevant. An argument can also be made that (i) the molecular mechanism causing the differences in trait phenotype is not necessary to describe the genetic architecture of phenotypes. A couple of other aspects of genetic architecture that could be added to this list include the molecular structure of chromosomes and genetic elements, patterns of linkage and linkage disequilibrium, historical and current population size and structure, and environmental interactions. While genetic architecture includes many components, depending on one's background and resources, some are more interesting and tractable than others. In this dissertation aspects of genetic architecture are studied in humans and dogs that range from shared QTLs in Chapter 1 to QTL effect size distributions in Chapter 2, and finally mutation and recombination patterns in Chapter 3.

Components of Genetic Architecture

The Number of Loci That Contribute to Variation

Phenotypic distributions broadly reflect the number of loci that underlie character variation. For example, Gregor Mendel could classify phenotypes in pea plants categorically because he was tracking a few large effect size loci (Mendel, 1865). Human height, a character with a normal distribution and infinite gradations, is at the other end of the spectrum of polygenicity. Based on these observations, if multiple species exhibit similar phenotypic variation for a given character, one might conclude that the number of underlying variants is reasonably conserved. This prediction of a similar number of influential variants implicitly assumes that other components of genetic architecture are also largely the same, an assumption that does not always hold.

Repetitive DNA sequences represent one example of how different mutational spectra can alter the number of loci that influence a character shared among species. Repetitive DNA arranged in tandem is highly mutagenic and can confer incremental phenotypic variation if sequences are in either cis-regulatory or protein coding sequences (Gemayel et al., 2010). For example, the age of onset for Huntington's disease, a debilitating human neurodegenerative disorder, is correlated with the length of a trinucleotide repeat in the *HTT* gene (Andrew et al., 1993). Incremental mutations at a few tandem repeat loci can lead to incremental variation that mimics the continuous variation often assumed to be the consequence of many influential loci (Kashi et al., 1997). Tandem repeats can cause differences in the number of influential loci for a shared character if there are differences among species in the genetic regulatory network, epistatic network, the collection of causative tandem repeats, or the mutation rate of repeats.

Divergent population histories among species can also cause differences in the number of variants that affect a given shared character. For example, evolutionary forces such as selection or genetic drift could fix alleles at multiple loci, thus changing the total number of variable loci within a given species. Large deviations in the number of variants for a given character will likely have a substantial impact on phenotypic variation, perhaps calling into question whether the character can even be classified as shared among species. Thus trait definition and decomposition are vital to comparisons of genetic architecture, particularly as taxonomic distances increase. While the number of loci that contribute to phenotypic variation can vary for the same character shared among species, large deviations are unlikely because the character would no longer be recognized as the same or shared.

When the Same Loci Contribute to Variation

Prior to discussing how often the same loci influence character variation within multiple species, it is helpful to more precisely define the variants that impact phenotypes. Quantitative traits and the loci that influence their variation can be defined as follows:

> A quantitative trait is one that has measurable phenotypic variation owing to genetic and/or environmental influences. This variation can consist of discrete values...or can be continuous, such as measurements of height, weight and blood pressure. A QTL [quantitative trait locus] is a genetic locus, the alleles of which affect this variation (Complex Trait Consortium, 2003).

This highly inclusive definition is prone to debate as some might argue that discrete characters should not be classified as quantitative, and others may confine QTLs to loci derived from linkage studies. Despite these objections, the definition of quantitative traits and QTLs given by the Complex Trait Consortium will be used in this dissertation. If orthologous loci influence the variation of a commonly held character in multiple species, they are described as shared QTLs. Characters whose variation can be described by only a few large effect loci are generally considered to have a simpler genetic architecture relative to characters influenced by hundreds of variants. These simple characters are often suitable for candidate gene studies where conserved genetic architecture and gene function are inherently assumed.

A recent study provides a rough estimate of the expected rate of QTL sharing for simple characters by estimating the rate of QTL sharing from candidate gene studies of parallelism and convergence (Conte et al., 2012). The authors find that parallel or convergent phenotypes are influenced by the same QTLs 0.55 ± 0.08 s.e. of the time. As might be expected, as taxonomic distances increase, the rate of QTL sharing also decreases. The rate of QTL sharing is 0.8 for the closest taxa considered, and between 0.1 and 0.4 for the most distantly related taxa in the study (Conte et al., 2012).

The renowned geneticist and mathematician R.A. Fisher demonstrated that continuously distributed characters could be produced by the combined actions of many Mendelian loci (Fisher, 1918). While 55% may serve as a rough estimate for the rate of QTL sharing for characters described by a small number of large effect size variants, it is unknown whether this estimate applies to characters influenced by hundreds of varying loci. Estimating the extent of QTL sharing between species for a highly polygenic character is the primary objective of Chapter 1.

To make this estimate, the ideal character to compare would be continuously distributed and shared between distantly related species, thus making the character likely to be influenced by many QTLs that independently vary within each species. It is easy to identify distantly related species; the challenge is in identifying distantly related taxa that share a character with comparable phenotypic variation and readily available genetic data. Two species that fit this description are humans and dogs that are separated by roughly 95 million years of evolution (Rosindell and Harmon, 2012). The character height varies extensively in both species and has conveniently been studied in fairly comparable genome-wide associations.

Underlying Causative Mutations

Before proceeding with the comparison of the genetic basis of human and dog height variation, it is useful to delineate the difference between QTLs and the causative mutations, or quantitative trait nucleotides (QTNs) that underlie trait variation. Linkage and association studies are well-suited for identifying regions of the genome that influence trait variation, but additional work is typically required to narrow down the precise molecular variant (Schumacher et al., 2005; Yamamoto et al., 1998). Shared QTLs describe regions of the genome that contribute to intraspecific variation among species; how often do identical or shared mutations underlie instances of QTL sharing?

It is commonly held that mutations are random with respect to phenotypes, although mutation is not a random molecular process (Fitch, 1967; Li et al., 1984; Martincorena et al., 2012). While selection and drift operate on variation, mutational patterns, or variability, describes what varies in the first place (Wagner and Altenberg, 1996). Evolutionary forces and network properties like selection, drift, mutation, recombination, epistasis, and pleiotropy can individually and in concert contribute to shared, yet independent causal mutations among individuals, populations, and species. Shared causative mutations are expected to be more common when evolutionary and network constraints are stronger.

For example, in autosomal dominant achondroplasia, the most common form of short-limbed dwarfism in humans (Figure 1-1), the causative mutation (G380R) not only recurs at the same nucleotide position of *FGFR3*, but mutates from G-to-A in almost all individuals (Bellus et al., 1995). Constraint or purifying selection can explain why only this particular nucleotide position is mutated, although mutational bias best explains why the transition from G-to-A is the most common.

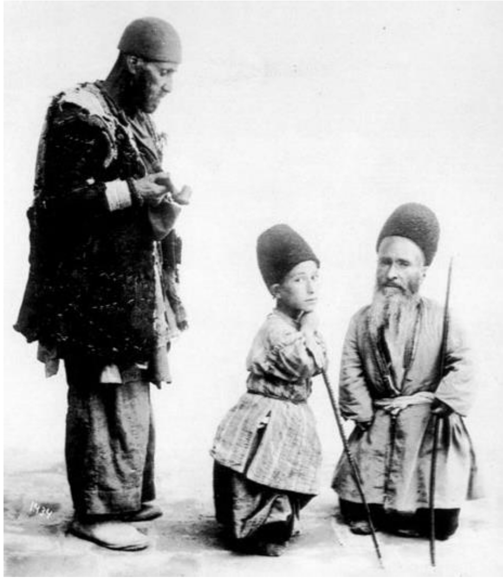

Figure 1-1 Human achondroplasia (Credit: Antoin Sevruguin)

Interestingly, like some other congenital conditions, achondroplasia has a paternal age effect where older men are roughly 10 times more likely to have a child with the condition (Risch et al., 1987). A recent study suggests that this paternal age effect is due to selection where stem cells with the G380R mutation outcompete normal spermatogonial cells (Shinde et al., 2013). The interplay between mutation and selection is nuanced and will likely always merit further study. As is the case with G380R and achondroplasia, biased or constrained evolutionary forces can lead to shared causal mutations with independent origins.

If selection or mutation only weakly influence the genetic basis of trait variation, shared QTNs are less likely, although shared QTLs may still exist. For example, non-synonymous mutations of *MC1R* are associated with dark-colored pigment (melanism) within populations of many species, although the precise causal mutation often varies

(Guernsey, et al., 2013; Majerus and Mundy, 2003). In addition to protein coding mutations, cis-regulatory changes that modify how genes are expressed are thought to have major evolutionary relevance (Wray, 2007). Cis-regulatory mutations that underlie QTL sharing are less likely to be constrained than protein coding mutations because in theory many mutations can be cis-regulatory, and intergenic DNA sequence conservation is markedly lower than coding sequences (Kent et al., 2002). This means that although the same QTL may be responsible for trait variation, derived alleles from cis-regulatory changes could differ in both direction and magnitude.

Comparisons of the QTNs that underlie QTLs shared among species would further clarify the extent of selective and mutational constraints on the genetic basis of traits, but genetic marker-based studies are limited to identifying regions of the genome associated with trait variation, not the causal mutations. Despite this limitation, QTL studies do allow the extent of QTL sharing between species to be estimated.

Setting Up the Comparison Between Humans and Dogs

Human Height Variation

Francis Galton and Karl Pearson were the first to formally characterize human height variation. The casual observer of height can see that it roughly follows a normal distribution (Schilling et al., 2002), a pattern illustrated by a plot of Pearson's data on height (Figure 1-2).

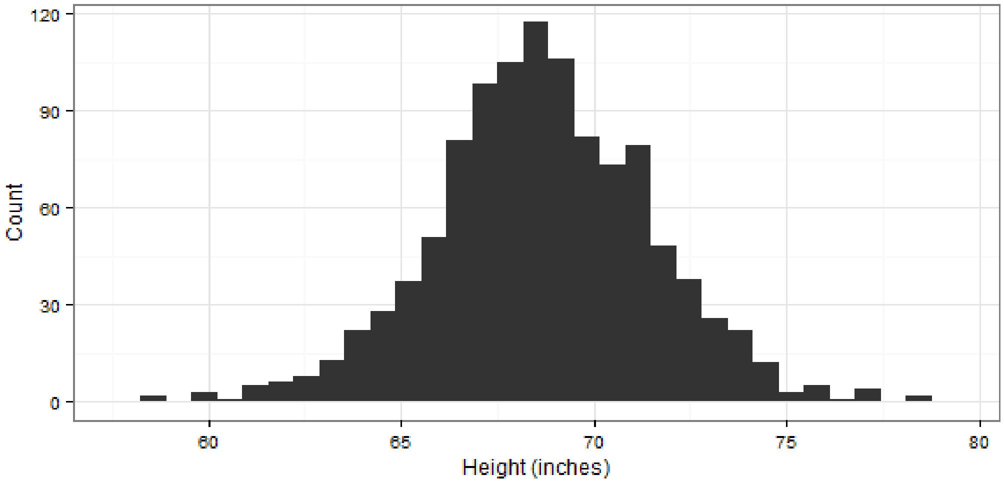

Figure 1-2 Histogram of human height for 1078 males illustrates the normal distribution. Data from (Pearson and Lee, 1903).

Galton, the mentor of Karl Pearson, noted that while height tends to regress towards mediocrity or the mean, it is highly predictive from one generation to the next (Figure 1-3) (Galton, 1886).

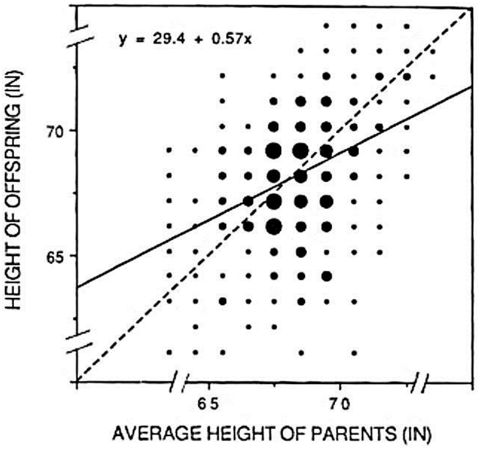

Figure 1-3 Predictability of human height and regression towards the mean. Height for 930 individuals and their parents from Galton's published data (Galton, 1886). Dotted line represents predicted height of offspring based on mean parent height (female height is multiplied by 1.08). Solid line represents regression line for actual offspring height.

Human height is normally distributed within populations and is a reliable predictor of future offspring height, but how much of its variation is due to genetic factors? Broad-sense heritability (H^2) is the proportion of phenotypic variance that can be explained by genetic factors. The most common estimate of heritability (h^2) is the proportion of phenotypic variance explained by additive genetic variance, defined as the average effect of substituting one allele for another. By definition, narrow-sense heritability (h^2) excludes the effects of dominance (allelic interactions at the same locus) and epistasis (allelic interactions at different loci) (Charlesworth and Charlesworth, 2010). The most current estimate of narrow-sense heritability for human height is around 80% (Macgregor et al., 2006; Visscher, 2008). Heritability of 80% means that variation in genotypes among humans is primarily responsible for observed height differences (Visscher et al., 2008).

How much height variation exists between human populations? While countries are not isolated populations, comparisons of average height between countries can still be informative. A plot of such comparisons shows that average height is lowest in Southeast Asia and parts of Southern America and highest in Europe (Figure 1-4) (Appendix A) (Abdulrazzaq et al., 2008; Australian Bureau of Statistics, 1998; Bogin, 1999a; Cavelaars et al., 2000; Chile, 2010; Connor Gorber et al., 2008; Corbett et al., 2009; Dettwyler, 1992; Deurenberg et al., 2003; Dusko Bjelica, 2012; El-Zanaty and Way, 2009; Food and Nutrition Research Institute, 2003; France, 2006; Frankenberg and Jones, 2003; Garcia and Quintana-Domeque, 2007; Haghdoost et al., 2008; Helsedirektoratet, 2009; Herpin, 2003; Van Hung and Sunyoung, 2008; Instituto Brasileiro de Geografi a e Estatística, 2010; Istat, 2011; Japan, 2011; Jordan et al., 2012; Jureša et al., 2012; Kamadjeu et al., 2006; Kułaga et al., 2011; Lim et al., 2000; Mamidi et al., 2011; Meisel and Vega, 2004; Mexican Business Web, 2012; Moosa, 2002; Msamati and Igbigbi, 2000; National Center for Health Statistics, 2008; National Statistics England, 2011; NSO Malta, 2003; Okosun et al., 1998; Ozer, 2008; Peltonen et al., 2008; del Pino et al., 2005; Ranasinghe et al., 2011; Schönbeck et al., 2013; Schultz, 2005; Shields et al., 2011; So et al., 2008; Starc and Strel, 2011; Statistics Netherlands et al., 2012; Statistisches Bundesamt, 2009; Stevo Popovic, 2013; Subramanian et al., 2011; Tawfeek, 2002; Tutkuviene, 2005; Velarde, 2006; Venkaiah et al., 2002; Vignerová et al., 2006; Welsh Assembly Government, 2010; WHO, 2007; Yang et al., 2005). Of the populations studied in the United States, height is lowest for Mexican Americans and highest for White Americans.

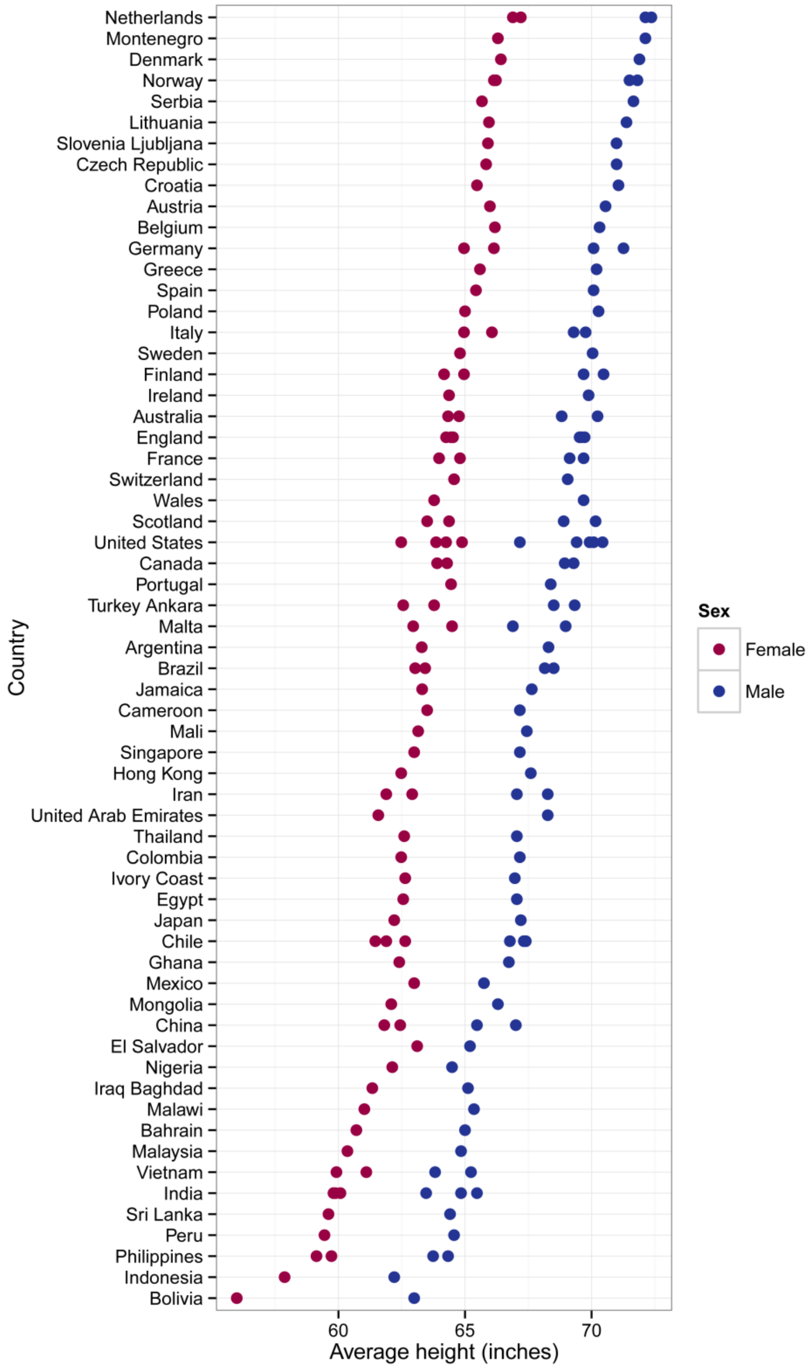

Figure 1-4 Adult male and female average height by country. Multiple data points for the same country represent different ethnic groups or age ranges.

Gustafsson and Lindenfors 2004 provide a more detailed characterization of human height variation by population and region (Figure 1-5) (Gustafsson and Lindenfors, 2004). They find that the tallest populations are the Netherlands and Fiji-Melanesia while the shortest are the Mbuti and Ituri pygmies. Inspection of height by country and by population reveals that males tend to be taller than females and that the magnitude of the difference is fairly consistent across the range of height variation (Figure 1-4, Figure 1-5). These observations are not consistent with Rensch's rule that states when the male is the larger sex, size dimorphism increases with body size (Rensch, 1950). While Rensch's rule applies across species within the primate lineage (Fairbairn, 1997), it does not appear to apply across human countries or populations.

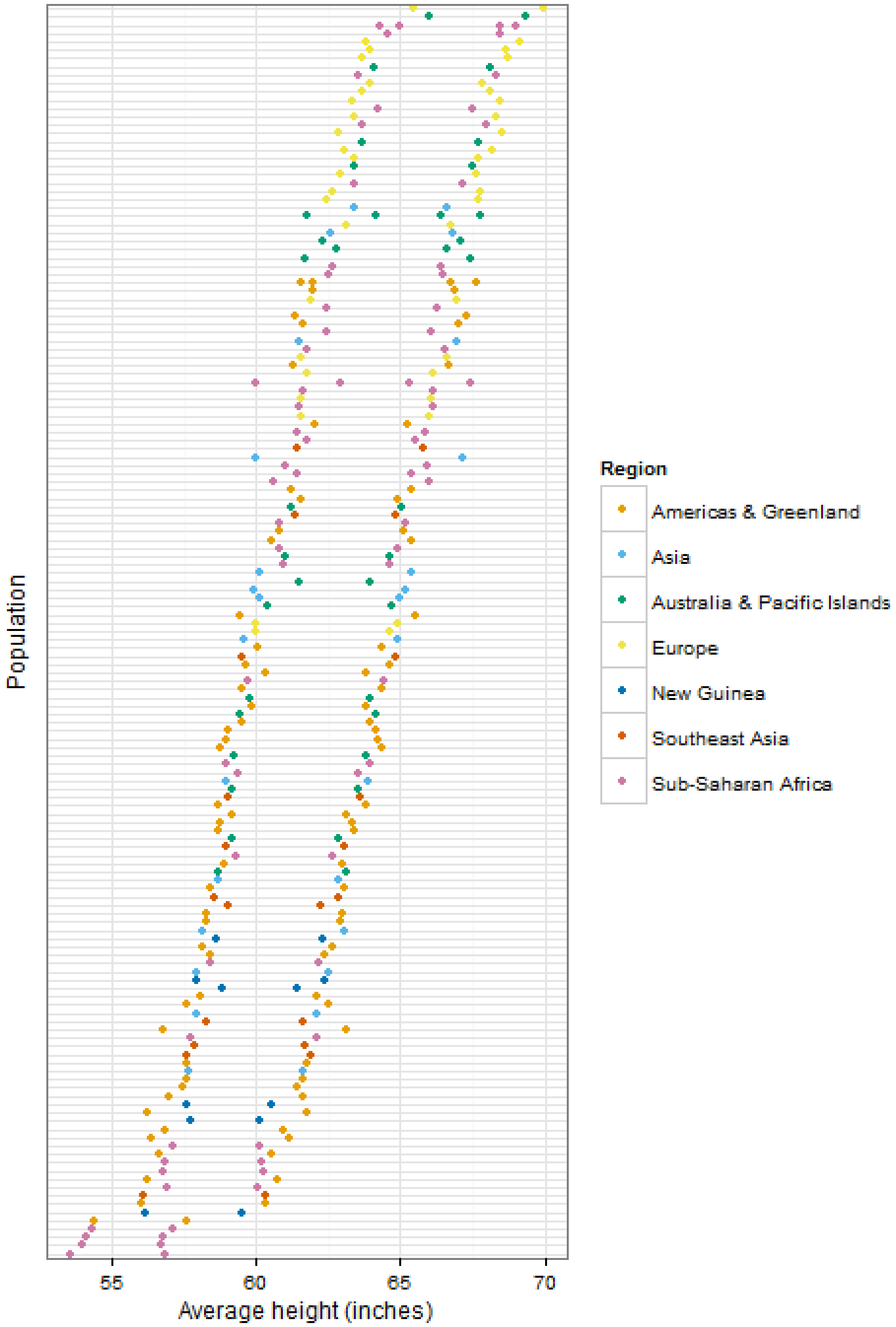

Figure 1-5 Adult male and female average height by population, colored by region. Figure based on data from Appendix I in (Gustafsson and Lindenfors, 2004).

Dog Height Variation

One of the most common breed-defining traits in dogs is size. Dogs exhibit huge extremes in height variation from the tiny Chihuahua to the giant Great Dane. Since the heritability for the same trait tends to be fairly similar among populations and even between species, the heritability of dog height is likely similar to humans, around 80% (Visscher et al., 2008). Height variation between breeds is obvious to any observer, but breeders and researchers alike also recognize substantial variation within breeds. A plot of dog height by breed reveals how some breeds have an expansive range of height while others tend to be more tightly defined (Figure 1-6) (Alderton, 2008). The word *defined* is used here because breed standards define preferred levels of variation; they do not describe existing variation. Thus, ranges given in a plot such as Figure 1-6 are best interpreted as definitions, not descriptions of breed variation. The distribution of height variation within breeds is likely much broader than given by defined ranges. Another important observation from Figure 1-6 is that relative to ancestral gray wolves, the majority of dog breeds are shorter.

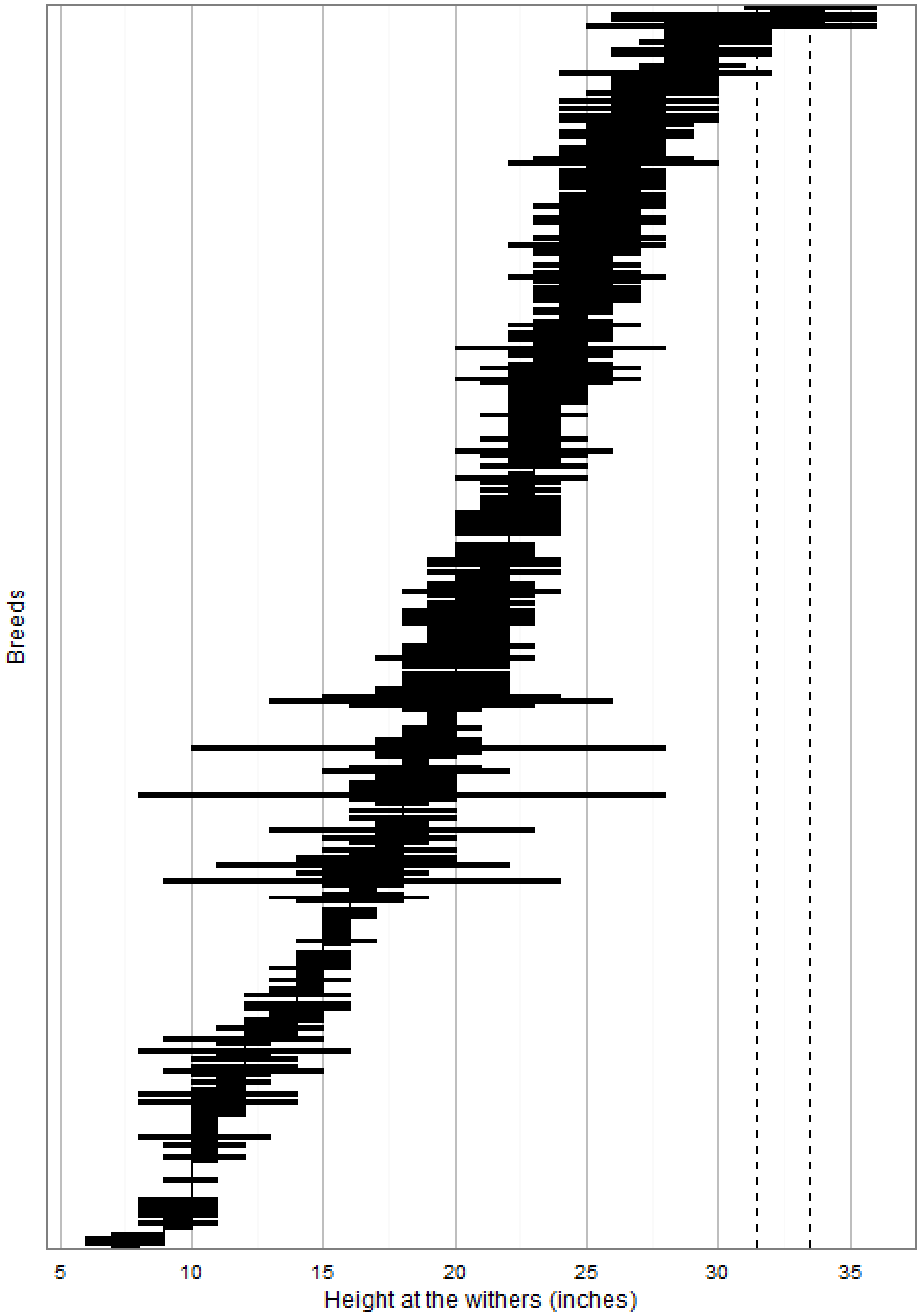

Figure 1-6 Dog height at the withers by breed. Dotted lines represent the likely height range of the ancestral gray wolf. Height range data published in (Alderton, 2008).

Height is measured to the top of the head in bipedal humans. In quadruped animals like the dog, height is traditionally measured to the top of the shoulder, or withers (Figure 1-7). Dog height does not include the spinal column or head; it is essentially a measure of limb length. While height at the withers in dogs is anatomically similar to human arm length, not leg length, it is the preferred measure because relative to hind limbs, height at the withers is less influenced by posture, a potentially confounding factor.

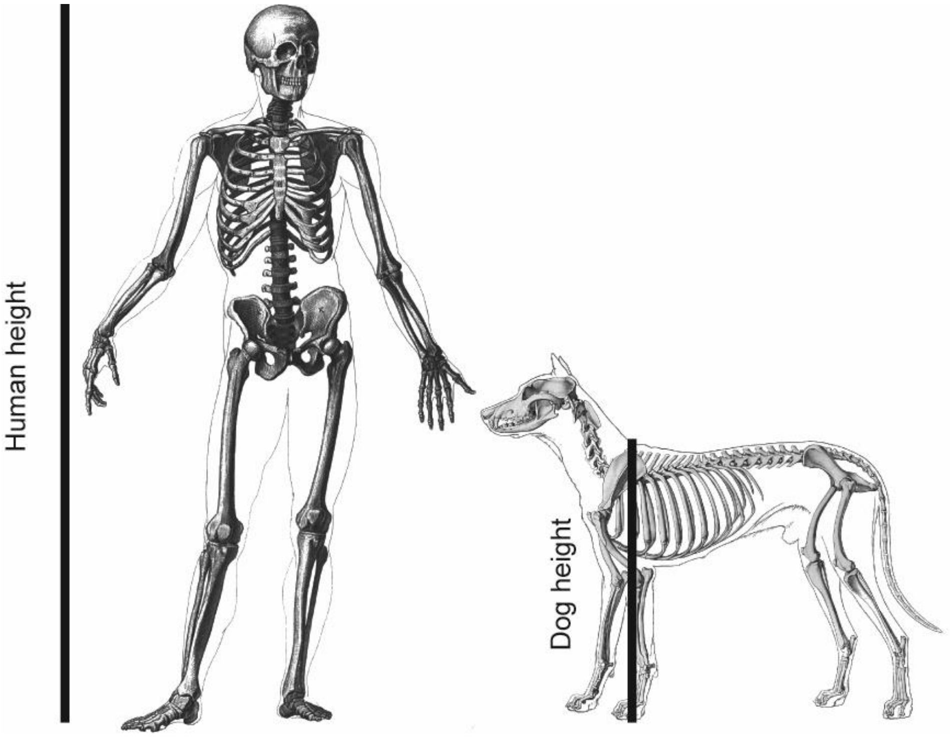

Figure 1-7 Human and dog height measurements.

Like humans, males tend to be larger than females in dogs, and the difference appears fairly consistent across the range of size. This suggests that Rensch's rule also does not apply between dog breeds (Sutter et al., 2008).

History of Height in Humans

As might be expected, height in humans is closely intertwined with the story of our evolution. When the ancestors of modern humans transitioned to a bipedal lifestyle, body proportions also evolved (Ruff, 2002; Schmitt, 2003; Thorpe et al., 2007). Humans evolved shorter forelimbs (arms) and longer hindlimbs (legs) relative to ancient human fossils. By roughly 1.5 million years ago body proportions had evolved to within the range of modern humans (Ruff and Walker, 1993).

Recent height increases in countries around the world have led to the popular idea that humans are taller than we used to be (Blue, 2008; Cohen, 2011). While this is likely true for many populations around the world relative to the past century or so, what about further back in time? Estimates of ancient human height provide a baseline and context for more recent changes in body size.

Estimates of ancient human height from the fossil record must be viewed with caution for at least four reasons: 1) Incomplete skeletons require making assumptions about body proportions because height is extrapolated. For example, the height of individuals from East Africa are thought to have been initially overestimated because the extrapolation was based on European body proportions (Allbrook, 1961; Ruff, 2002). 2) Skeletal remains often only represent a single individual, making it impossible to know the population's height distribution. This means there is no way to tell if an individual was small, average, or tall relative to his peers. 3) Skeletal remains may not belong to the direct lineage that gave rise to modern humans. In other words, the height of a distant cousin may not be representative of our lineage. 4) Skeletal remains may not be of an adult, requiring adult height to be estimated from assumed growth trajectories that are inherently problematic (Bogin, 1999b).

With these challenges in mind, the first human ancestor to consider is *Homo erectus* that lived between 1.89 million and 143,000 years ago (Antón, 2003; Dubois, 1894; Smithsonian Institution, 2010a). The most well-known and complete *Homo erectus* fossil is Turkana Boy (KNM-WT 15000). There has been substantial discussion from researchers about the predicted adult height of Turkana Boy, with the latest trend being somewhere around 5' 4", although previous estimates are as high as 6' 1" (Gibbons, 2010; Hawks, 2010; Ohman et al., 2002). Some argue that Turkana Boy was extraordinarily tall relative to his peers, while others claim he would not have likely grown to be 6' 1" as an adult (Hawks, 2010). The Smithsonian takes a safe approach and publishes the whole range of reported heights for *Homo erectus* as 4' 9" to 6' 1" (Smithsonian Institution, 2010a). Perhaps surprisingly, the range of *Homo erectus* height is somewhat similar to extant human populations. Depending on which estimate one trusts, *Homo erectus* is either in the bottom quarter or top quarter of height based on current estimates of extant humans (Figure 1-8).

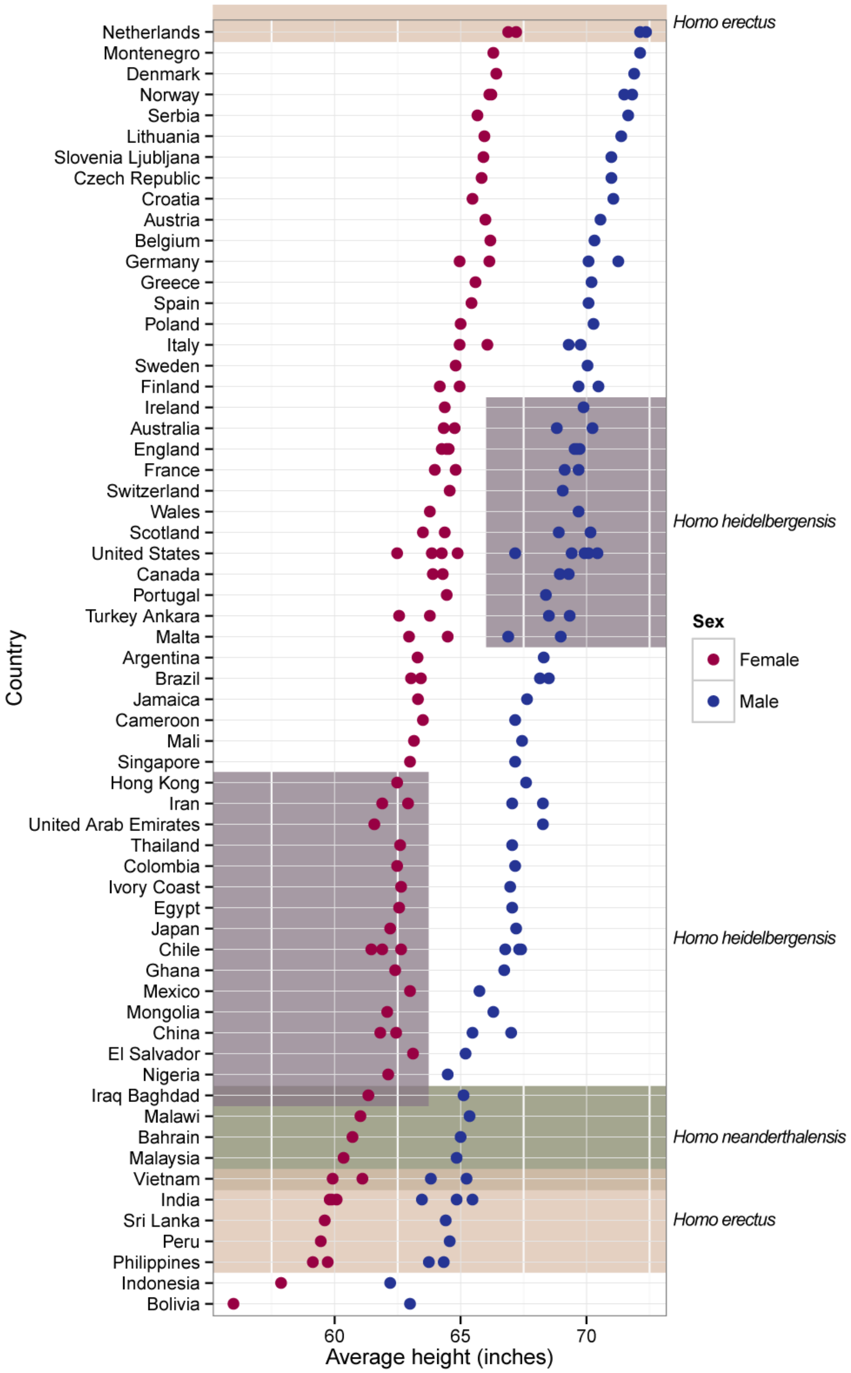

Figure 1-8 Ancient height (shaded blocks) relative to modern adult male and female average height.

Homo heidelbergensis lived roughly 700,000 to 200,000 years ago and is generally thought to be the predecessor of Neanderthals and modern humans (Mounier et al., 2009; Rightmire, 1998). Males had an average height of 5' 9" and females were 5' 2" (Smithsonian Institution, 2010b). The sexual size dimorphism of 5' 9" compared to 5' 2" is about 2 inches more than what is generally observed in extant humans (Figure 1-8). While it is possible that sexual size dimorphism was greater in *Homo heidelbergensis*, uncertainty and error in height estimates are probably the cause of this inconsistency relative to ancient and modern humans. Depending on if the male or female estimate of height is trusted, *Homo heidelbergensis* is either near the median or lower half of height relative to extant humans, respectively (Figure 1-8).

The last relevant human ancestor to consider is *Homo neanderthalensis* that lived between 200,000 and 28,000 years ago (Delson and Harvati, 2006; Smithsonian Institution, 2010c). Unlike the more slender *Homo erectus* that lived in a tropical climate, stocky Neanderthals lived in a colder European climate (Hawks, 2005; Katzmarzyk and Leonard, 1998; Ruff et al., 2005). While there has been an ongoing debate concerning how much, if any, admixture occurred between Neanderthals and a subset of the modern human lineage (Eriksson and Manica, 2012; Holliday, 1997; Wang et al., 2013), the most recent studies indicate that humans of European descent likely inherit ~1-3% of their genomes from Neanderthals (Callaway, 2014; Sankararaman et al., 2014; Vernot and Akey, 2014). Males averaged around 5' 5" and females around 5' 1", putting Neanderthals within the range of extant human variation, albeit in the bottom quarter (Figure 1-8) (Smithsonian Institution, 2010c).

Despite an inherently sparse and frequently debated fossil record, relative to the countries of the world, ancient humans would likely fit somewhere in the bottom half of height variation. This means that the generalization that humans are taller now than they used to be is not altogether true. Most extant populations are likely taller, but a substantial proportion of humans are about the same height, or shorter than ancient humans (Figure 1-8). Based on this observation it may be tempting to conclude that height hasn't evolved much in the last 1.5 million years, but a brief review of more recent human history reveals a complex story in which environmental and selective forces have likely altered height.

This synopsis is admittedly generalized and European-centric, although it is illustrative of some of the recent evolution of human height. Human height decreased in the Neolithic era when humans transitioned from a hunter and gatherer lifestyle to an agriculturally based society about 12,000 years ago (Özer et al., 2011). Immediately preceding and during the Bronze and Iron ages, human height increased (3,500 - 2,500 years ago), only to sharply decrease with the establishment of the Roman Empire (Jaeger et al., 1998; Ozer, 2008). The collapse of the Roman Empire and a warmer climate correlate with rising height until the early Middle Ages (1400 A.D.) (Ozer, 2008; Steckel, 2004). Height then decreased for a couple hundred years until sometime in the 1700s when it started to increase again (Ozer, 2008; Steckel, 2004). Agricultural, economic, and environmental trends have influenced height in populations around the world. Selection may have also played a major role as has been suggested with African pygmy populations and in Northern and Southern Europeans (Shea and Bailey, 1996; Turchin et al., 2012).

The idea that natural selection is still acting in human populations is at odds with the theory that with the rise of modern humans and the accompanying development of culture and technology, natural selection ceased to operate (Furness, 2013). While culture itself can certainly be under selection, it does not preclude biological adaptation

(Stock, 2008). In fact, a recent book *The 10,000 Year Explosion* argues that civilization actually served to accelerate evolution (Cochran and Harpending, 2009). Despite modern medicine and its ability to allow more people to live, humans are still evolving, and natural selection is still in operation. It is true that the relative intensities of selective forces have changed in many human populations, but as long as heritable variation is impacting reproduction or survival, natural selection will occur. This does not mean that humans will improve and become smarter or stronger over time; natural selection is under no obligation to improve the human race.

In summary, human height has evolved and continues to evolve, although the extent of variation is not drastically beyond the range of ancient humans (Figure 1-8). While selective forces like sexual antagonism and climatic factors have impacted human height variation, human populations have rarely, if ever, experienced the intense directional selection for height that recently occurred in dogs (Connallon and Clark, 2014; Frey et al., 2010; Katzmarzyk and Leonard, 1998).

History of Height in Dogs

A history of height in dogs must begin with their wild ancestor: the gray wolf. When, where, and how many times domestication may have occurred is hotly debated, and is an area of intensive research. Archaeological evidence suggests that dogs may have been associated with humans as early as 30,000 years ago (Germonpré et al., 2009). These ancient bones found in modern-day Belgium are assumed to be archaic dogs because they are different than gray wolves, but this does not exclude the possibility that they belong to a now-extinct species of wolf. It is however clear that dogs were domesticated by at least 12,000 years ago, concurrent with the human shift to an agrarian lifestyle (Morey, 1994). Whether it was 30,000 or 12,000 years ago, other than humans, dogs represent the first domesticated mammal.

With the rise of molecular biology, mitochondrial DNA sequences were leveraged to investigate the timing of domestication. The first mitochondrial-based study suggested that domestication could have occurred multiple times and as early as 135,000 years ago (Vilà et al., 1997). This position was later disputed, once again using mitochondrial DNA, when it was suggested that dogs originated only one time, and from East Asia either 40,000 or 15,000 years ago (Savolainen et al., 2002). The discrepancy between 135,000, 40,000, and 15,000 years underlines the challenge of using mitochondrial DNA to date evolutionary events. The completion of a high-quality draft dog genome sequence offered a nuclear DNA-based estimate of domestication around 18,000 to 27,000 years ago, more consistent with archaeological evidence (Lindblad-Toh et al., 2005).

A microsatellite marker based approach published in 2004 concluded that most dogs had a European, not an East Asian origin (Parker et al., 2004). In 2010 a SNP marker based approach suggested that rather than an East Asian or European origin, most dog breeds came from the Middle East (Vonholdt et al., 2010). This study also reinforced the idea that there are several ancient breeds with different origins than dog breeds created in the 19th century. A few years later the same research group sequenced ancient mitochondrial DNA from Europe and switched to the opinion that most breeds have a European origin from around 18,000 to 32,000 years ago (Thalmann et al., 2013). This study has been particularly controversial because the sampling biases, which plague all domestication origin studies, were particularly extreme (all of the ancient dogs sampled were from Europe). In addition, mitochondrial DNA is inherently problematic (for example, mtDNA provides no evidence of admixture between humans and Neanderthals; nuclear DNA indicates otherwise) (Green et al., 2010; Pennisi, 2013; Serre et al., 2004).

Perhaps the most comprehensive and least biased study of dog origins was a 2012 publication by Kerstin Lindblad-Toh and colleagues which incorporated genetics, archeology, and biogeography (Larson et al., 2012). Citing uncertainty regarding the identity of ancient skeletal material, the authors conclude that domestication was in progress at least 15,000 years ago. They go on to argue that the genetic variation present in contemporary representatives of "ancient" breeds (e.g. Akita, Basenji, Saluki, Shar-Pei) merely reflect a paucity of recent admixture with other breeds. This claim is supported by the finding that none of the archeological evidence for dogs correlate with where ancient breeds supposedly originated (Larson et al., 2012).

Despite the debate on when and where dog domestication occurred, there is little doubt that it preceded the shift to agriculture, meaning that dogs and humans were first companions, and possibly competitors, as hunters, gatherers, and/or scavengers. In addition, dogs were likely domesticated multiple times, suggesting that the gray wolf was for some reason prone to being domesticated. As a bit of an aside, it is generally assumed that humans domesticated dogs, but a recent perspective argues the opposite: dogs domesticated humans (Hare and Woods, 2013a, 2013b). Regardless of who domesticated whom, and where it occurred, humans and dogs have been associated for at least 15,000 years, sharing a similar environment and diet.

Consistent with this idea, dogs evolved an increased ability to digest starch, an adaptation surely useful as humans shifted to agricultural societies (Axelsson et al., 2013). This study is particularly relevant to comparisons of shared QTLs in humans and dogs because in both species, adaptation occurred through an increase in the number of copies of amylase genes (Axelsson et al., 2013; Perry et al., 2007). In fact, human populations and dog breeds that have historically had more starch in their diets also harbor more copies of amylase genes. This example illustrates how traits and their associated QTLs can evolve in parallel in distantly related taxa.

Throughout much of their domesticated history, most dogs were probably not all that different in form from gray wolves. Much of the dog height variation observed among breeds has evolved within the last few hundred years as a result of selective breeding imposed by humans. Based on the phenotypic distribution of the gray wolf, it is clear that most of the size selection in dogs has been to reduce it (Figure 1-6).

Genetic Architecture of Human Height

While Galton and Pearson were the first to formally study the inheritance of human height variation in a quantitative framework, it was R.A. Fisher who described how a continuously distributed trait like height could fit into a Mendelian genetic framework. Fisher showed how, rather than being fundamentally distinct from Mendelian inheritance, quantitative traits were merely governed by the combined action of many contributing Mendelian loci (Fisher, 1918). Technological advances now allow for the identification of height variants at an unprecedented rate.

In the most recent National Human Genome Research Institute (NHGRI) database of genome-wide associations, 421 genes are associated with differences in human height (Hindorff et al., 2014). Loci replicated in at least six studies include the reported genes *ZBTB38*, *HMGA2*, *LCORL*, *HMGA1*, *ADAMTSL3*, and *EFEMP1*. One of the challenges that researchers have faced in the study of human height is understanding why so little of the heritable portion of height variation can be explained by genetic variants. In other words, if height is roughly 80% heritable, why isn’t more of its variation explained from the several hundred loci identified in genome-wide studies?

This is a question that has garnered much discussion because the allure of genome-wide association studies was originally the promise of identifying variants behind common, yet complex human diseases (Hirschhorn and Daly, 2005; McCarthy et al., 2008; Syvänen, 2001). After a few years of studying some of the most relevant diseases and phenotypes with genome-wide associations, it started to become apparent that expectations were not being met (Donnelly, 2008; Maher, 2008). Variants underlying complex traits were being identified, but they had little predictive power in explaining trait variation. Despite the lack of predictive power as some might have hoped, genome-wide associations have been successful in identifying many QTLs that underlie complex traits like height, and more variants will likely be forthcoming as microarray platforms are replaced by high-throughput sequencing.

Experiments in mice that knockout one gene at a time estimate that more than 6,000 genes can impact body size in some way; this is roughly a quarter of the mammalian genome (Reed et al., 2008)! Interestingly, knockouts that confer smaller size are ten times as common as those that increase body size, thus biasing the phenotypic direction of size variants (Kemper et al., 2012; Reed et al., 2008). Based on extensive conservation of gene function between humans and mice, the estimate of >6,000 possible height genes is likely comparable to humans. Although just because over 6,000 genes can in theory impact body size, this does not mean that all of them are varying in human populations, and thus relevant to genetic architecture.

Despite the highly polygenic nature and lack of explained heritability for characters like diabetes, the company 23andMe at one time offered health reports to customers from marker based genetic tests. In November 2013 the Food and Drug Administration sent a letter to 23andMe demanding that they cease selling their genetic test kits because there wasn't sufficient support for their health predictions. In regards to the FDA's response to 23andMe, Robert Klitzman of Columbia University wrote, "Over the past decade, we have heard about the 'fat gene,' the 'diabetes gene,' the 'alcoholism gene,' the 'intelligence gene,' even the 'God gene.' In the end, none of these so-called discoveries proved correct" (Klitzman, 2013). Indeed, complex traits live up to their name and will require much subsequent study before genotypes can reliably predict health outcomes or phenotypes.

Genetic Architecture of Dog Height

Breed formation in the 19th century marked the beginning of intense directional selection for reduced body size in nearly all dogs (Figure 1-6). The history of dog breeding includes selection for large effect variants as well as selection for variants that offset negative pleiotropic consequences from large effect variants. This pattern of selection should result in a J-shaped or exponential QTL effect size distribution (Figure 1-9) (Kemper et al., 2012; Robertson, 1967).

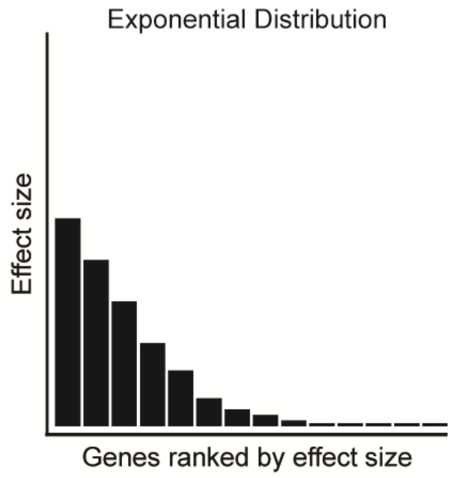

Figure 1-9 Exponential distribution with large, medium, and small QTL effect sizes

Consistent with this expectation, a recent genome-wide association for variation of height at the withers between dog breeds identifies ten significant autosomal loci (Boyko et al., 2010). Prior to this study, six autosomal loci had been associated with height differences between breeds (Jones et al., 2008; Parker et al., 2009). In the tradition of previous genome-wide association studies, the authors of the Boyko et al. 2010 genome-wide association employ a Bonferroni correction, which is the most conservative correction for the problem of multiple testing (Johnson et al., 2010). By excluding all but the strongest associations, the authors of the Boyko et al. 2010 study may have overlooked many relevant loci. This idea is supported by a genome-wide study of height variation in Northern and Southern Europeans that found meaningful associations well below the threshold of genome-wide significance (Turchin et al., 2012).

The strongest association in the Boyko et al. 2010 study is on chromosome 15 for *IGF1*, a large effect size variant previously found to influence overall size differences within and among dog breeds (Chase et al., 2002; Sutter et al., 2007). Another significant association is for a retrogene of *FGF4*, an expressed retrotransposition shown to be responsible for chondrodysplasia in breeds like the dachshund (Parker et al., 2009). For the most part, the other significant variants like *SMAD2* have been either identified in previous dog association studies or are near genes that impact size in either humans or mice (Jones et al., 2008).

In summary, the genetic architecture of dog height variation between breeds appears to reasonably follow the predicted exponential QTL effect size distribution. This means that there are likely many more meaningful associations beyond the ten significant loci originally reported by the study of Boyko et al. 2010.

Comparison of the Genetic Basis of Human and Dog Height Variation

Before proceeding with the comparison of the genetic basis of human and dog height variation, it is critical to consider how these characters are defined. The way human height is defined makes it a highly composite trait. Contrasted with dog height that includes fewer bones and body regions (Figure 1-7), based on trait composition alone, more loci are expected to contribute to human height variation than dog height variation. As traits are decomposed the number of loci that influence trait variation (polygenicity) decreases, but the extent of pleiotropy increases. This is one reason why dog height measured at the withers should be less polygenic than human height: it is a more decomposed trait based on how it is measured. One advantage of decomposing traits is that the loci that impact specific aspects of character variation can be identified. For example, it is more useful to know that a particular variant specifically impacts torso length rather than just knowing it impacts height in general.

Human Height QTLs

The most convenient source for human height variants is the *Catalog of Published Genome-Wide Association Studies* curated by the National Human Genome Research Institute (NHGRI) (Hindorff et al., 2014). This catalog currently includes 421 reported genes from 21 different studies. A manual clustering of genes based on genomic context results in ~263 regions of the human genome associated with height variation (see Appendix A for clustered genes). This collection of associations will be used to investigate height QTL sharing in humans and dogs. Additional analysis is required to obtain a comparable list of QTLs that influence height variation between dog breeds.

Dog Height QTLs

Conveniently, the genome-wide study of Boyko et al. 2010 that associates over 60,000 SNPs with 57 breed measures from 80 domestic breeds includes measurements that can serve as a proxy for human height. Height at the withers, body length, and neck length anatomically capture the majority of what is measured as human height (Figure 1-10). Height at the withers is preferred over height at the tail because it is less influenced by posture. In addition to making the comparison with human height more comparable anatomically, by including three measures of dog size, the number of variants underlying trait variation increases and becomes more similar to humans.

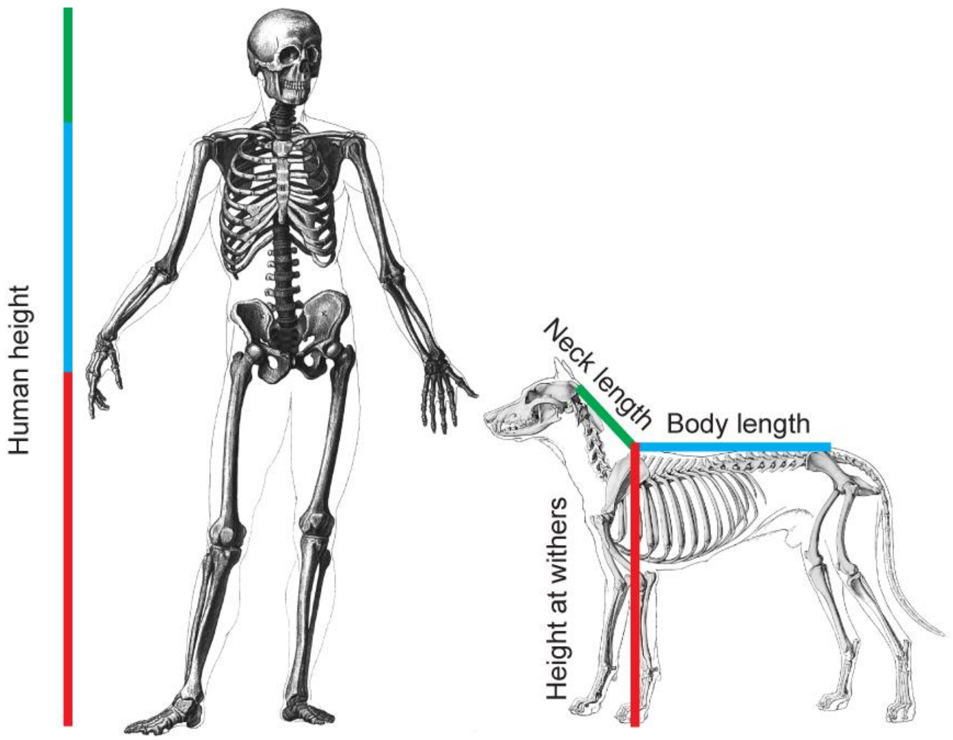

Figure 1-10 Comparable measures of height in humans and dogs

To combine height at the withers, body length, and neck length, I record the most significant association at each autosomal non-zero SNP (43,209). Since none of the current human height associations are on the X chromosome, only autosomal SNPs are considered.

The false discovery rate (*FDR*) provides a consistent approach for reconsidering statistical thresholds for studies like this that need to correct for multiple comparisons as well as combining traits (Storey and Tibshirani, 2003). For this reason false discovery rates are calculated after the three traits are combined, and then three different cutoffs are chosen: one at $FDR < 0.04$ ($P < 0.00225$), a more conservative cutoff at $FDR < 0.02$ ($P < 0.00029$), and the Bonferroni cutoff chosen by the Boyko et al. 2010 study that is $P < 0.00005$ ($FDR < 0.0093$) (Figure 1-11).

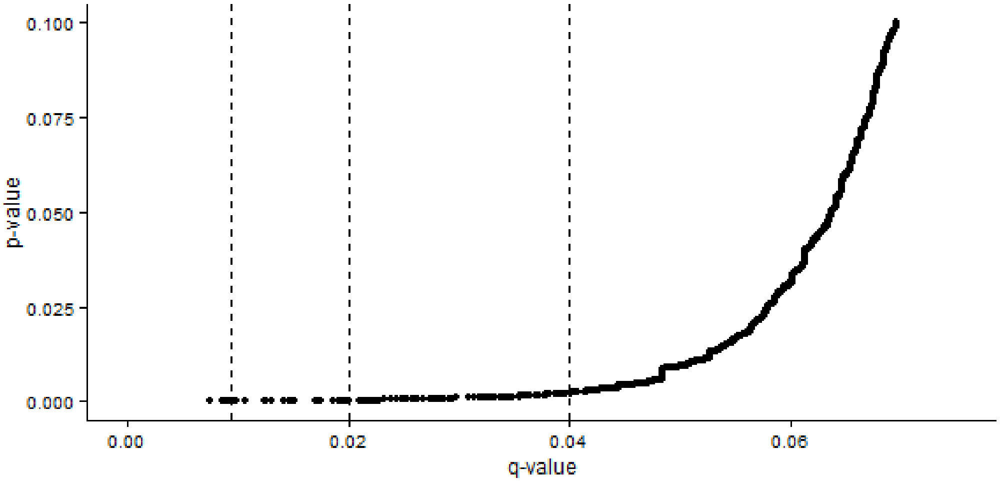

Figure 1-11 False discovery rate (q-value) cutoffs for genome-wide associations of dog height at the withers, body length, and neck length.

These false discovery rate estimates (*FDR* < 0.0093, 0.02, and 0.04) are reasonable for studies concerned with questions regarding the genetic architecture of traits (Barsh et al., 2012). In addition, since each gene on the microarray has many SNPs that are non-independent, the rate of predicted false discoveries per SNP (*FDR* < 0.0093, 0.02, or 0.04) is higher than the false discovery rate per gene.

A combined Manhattan plot of the three traits with *FDR* cutoffs reveals that this composite measure of dog size, comparable to human height, has many significant associations across the genome (Figure 1-12). Inspection of this Manhattan plots shows that although new regions of the genome are associated with this composite measure of height at *FDR* < 0.02 and 0.04, many of the additional SNPs included at these cutoffs simply add support to existing associated regions.

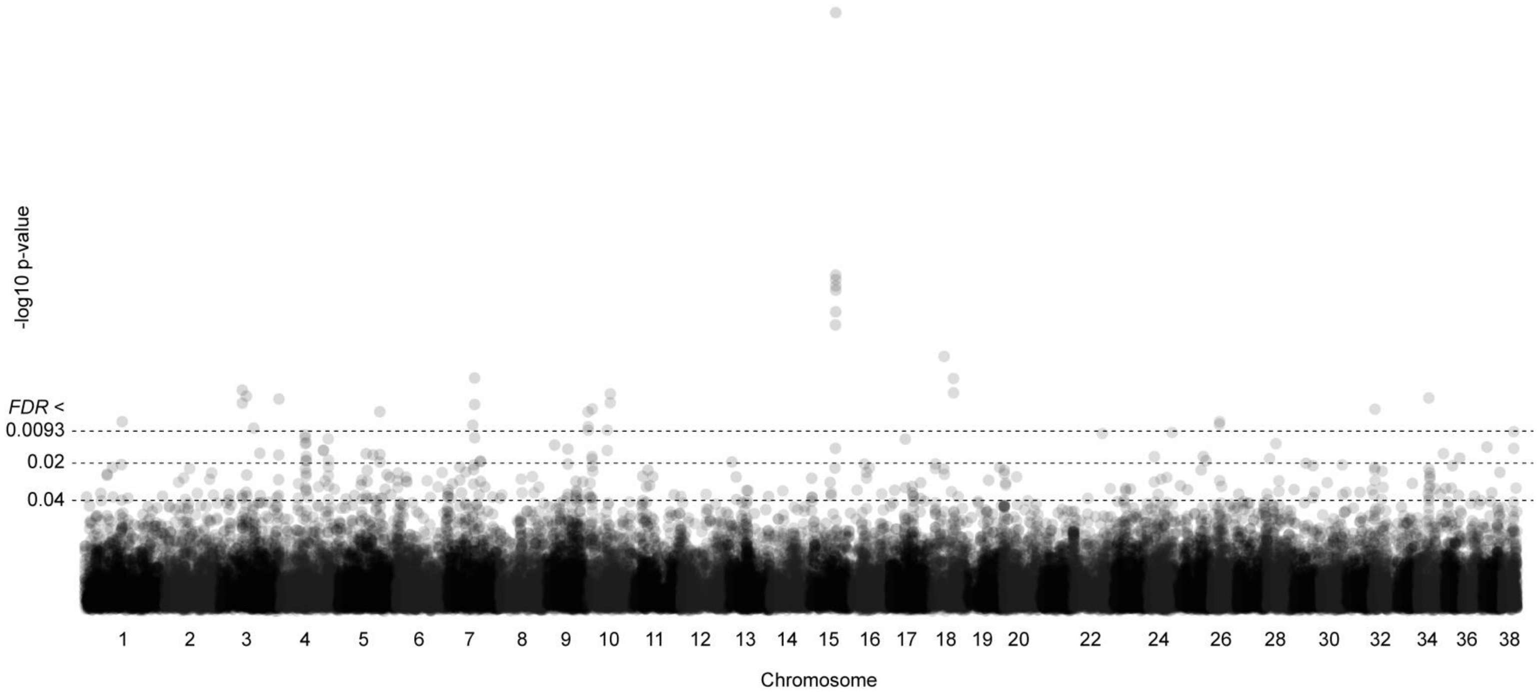

Figure 1-12 Manhattan plot of combined associations for height at the withers, body length, and neck length. Horizontal lines represent *FDR* < 0.0093, 0.02, and 0.0

Decomposition of Height

For consistency, the Bonferroni threshold will now be referred to by its false discovery rate of $FDR < 0.0093$. A total of 31 SNPs are included at $FDR < 0.0093$ for the composite measure of dog height that includes height at the withers, body length, and neck length. 71 SNPs are significant at $FDR < 0.02$, and 289 at $FDR < 0.04$. While there is considerable overlap between height at the withers, body length, and neck length, a considerable portion of SNPs are specific to each trait (Figure 1-13). The Venn diagram in Figure 1-13 identifies trait specific SNPs as those that are only significant for a given trait or traits. While this is one way to identify genes that impact limb length, torso length, or neck length specifically, a more stringent approach is to classify SNPs unique to a given trait if they are at least an order of magnitude more significant than the next best SNP.

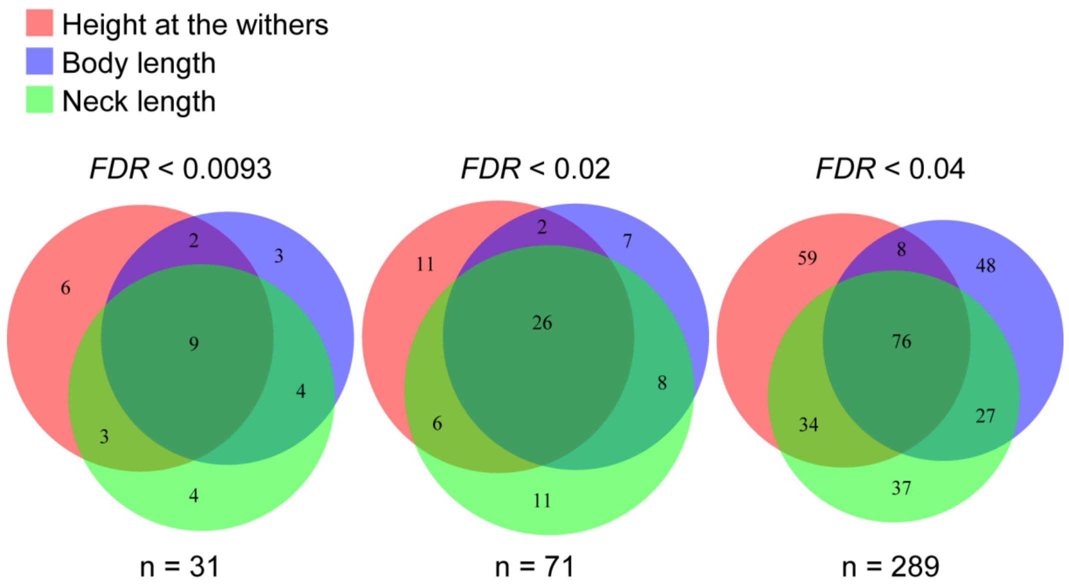

Figure 1-13 The origin of composite dog height SNPs by statistical threshold. Trait specific SNPs represent uniquely significant associations.

For *FDR* < 0.0093 SNPs, this magnitude-based method identifies *IGF1*, *FGF4* retrogene, *MED13L*, *IGF1R*, and *LCORL* as trait specific. While body weight is not part of the composite measure of dog height used in this study, it is useful to consider when decomposing height because it provides a measure of overall body size (Table 1-1).

Table 1-1 Decomposition of dog height

Gene	SNP (CanFam2)	-log10 p-value			
		Body weight	Height withers	Body length	Neck length
IGF1	chr15: 44226659	15.3	10.8	14.3	12.7
FGF4 retrogene	chr18: 23298242	4.4	6.1	2.9	2.8
IGF1R	chr3: 44099822	3.4	5.1	2.9	3.5
MED13L	chr26: 16269905	N/A	4.5	3.1	2.8
LCORL	chr3: 93851186	3.7	3.5	5.1	4.0

IGF1 is more correlated with variation in body length and body weight than with height at the withers or neck length. This observation is consistent with the study in the Portuguese Water Dog that first identified a QTL near *IGF1* that influences overall body

size (Chase et al., 2005). *FGF4* retrogene is most strongly correlated with height at the withers. This is logical because the *FGF4* retrotransposition is known to cause chondrodysplasia in breeds like the dachshund and basset hound (Parker et al., 2009). *IGF1R* is also most strongly correlated with variation in height at the withers. Since *IGF1* impacts overall body size, one might assume that *IGF1R* would also impact overall body size, not height at the withers specifically. Recently a non-synonymous mutation in *IGF1R* was associated with reduced body size in dogs (Hoopes et al., 2012). In this study size was treated as a binary trait, precluding any indication about which aspect of size *IGF1R* might be specifically influencing. Further study is required to understand precisely how the *IGF1R* variant contributes to a reduction in body size. The *MED13L* variant is most strongly associated with height at the withers and could merit fine-scale mapping and further study. Although relatively little is known about *MED13L*, *MED13* is linked to obesity as well as other conditions like type II diabetes and heart failure. In addition, in the mouse, deletion of *MED13* enhances obesity in response to a high-fat diet (Grueter et al., 2012). It is thus plausible that *MED13L* could play a meaningful role in influencing size in dogs. Lastly, *LCORL* is most strongly associated with body length, a particularly interesting finding since in humans *LCORL* is also most strongly associated with trunk length (Soranzo et al., 2009). *LCORL* is associated with height at the withers in horses (Metzger et al., 2013) and height at the hip in cattle (Pryce et al., 2011). It would be interesting to see if perhaps body length has an even stronger correlation in these domesticates. While the magnitude of p-value differences from genome-wide association studies may not always reliably decompose height, at least for *IGF1*, *FGF4* retrogene, and *LCORL*, this decomposition is consistent with other studies of variation at these loci in humans and dogs.

Functional Analysis of Dog Height QTLs

Before comparing dog size QTLs with human height variants, it is prudent to get an indication of the quality of the 31, 71, and 289 significant SNPs. One way to assess the variants that appear to influence size differences between dog breeds is to perform gene ontology analysis. While this type of analysis can be problematic, it does offer cursory evidence of enriched biological processes that can serve as a sort of litmus test to see if SNPs are tagging relevant variants. Since gene ontologies describe the functions and biological processes of genes, SNPs must first be assigned to nearby genes.

SNPs within 200 kilobases (kb) of a gene are assigned to that gene. Where the same SNP can be assigned to multiple genes, all genes are included for analysis. This is done to prevent bias from choosing one particular gene over another. Gene functions are compared for all three levels of significance with the web-based tool DAVID (Huang et al., 2009a, 2009b). Genes found in previous genome-wide association studies of human height are enriched at each level of significance, suggesting substantial overlap between human and dog height QTLs (Table 1-2).

The biological process *regulation of growth* is also enriched at each level of significance. Some of the other enriched processes include: *cell division*, *regulation of cell component size*, *gland development*, *fibroblast proliferation regulation*, *skeletal system development*, and *bone development*. As might be expected if additional SNPs are meaningful, as more SNPs are considered significant, whole molecular pathways become enriched. Indeed, at $FDR < 0.02$ the KEGG pathway *pathways in cancer* becomes enriched with 7 contributing genes. At $FDR < 0.04$ the following KEGG pathways become enriched: *TGF-beta signaling*, *MAPK signaling*, and *Insulin signaling*. These three pathways are thoroughly studied and have known impacts on growth (Chen, 2012; Katz et al., 2007; Siddle, 2011). As a whole, gene ontology analysis supports the notion that a substantial number of SNPs at each level of significance considered are true associations (Table 1-2). A complete table of gene ontology results can be found in Appendix A.

Table 1-2 Functional analysis of genes near SNPs associated with dog height variation

Metric	Description	Gene number (N/A if not enriched)		
		FDR<0.0093 (n = 75)	*FDR*<0.02 (n = 151)	*FDR*<0.04 (n = 701)
OMIM	Genome-wide association analysis identifies 20 loci that influence adult height	3	3	N/A
	Many sequence variants affecting diversity of adult human height	N/A	8	13
Gene ontology biological process	Regulation of growth	6	9	23
	Cell division	5	6	N/A
	Regulation of cell component size	4	N/A	19
	Gland development	N/A	6	13
	Fibroblast proliferation regulation	N/A	N/A	6
	Skeletal system development	N/A	7	24
	Bone development	N/A	4	10
KEGG	Cell cycle	3	N/A	N/A
	Pathways in cancer	N/A	7	19
	TGF-beta signaling pathway	N/A	N/A	8
	MAPK signaling pathway	N/A	N/A	16
	Insulin signaling pathway	N/A	N/A	11

Selective Sweep Analysis of Dog Height QTLs

Another approach for assessing these dog height SNPs is to leverage other studies that use independent samples of dogs to identify regions of the genome that exhibit signatures of selection. If false discovery rate SNPs are meaningful, a portion of them should also appear in selective sweeps from samples of other dogs. This expectation is reasonable because size has been a major focus of selection in dogs. Of course sampling and technical considerations play an important role in determining which loci appear under selective pressure and are identified as height QTLs; thus perfect concordance between false discovery rate SNPs and selective sweep mapping is not expected. Despite this limitation, overlap does provide some indication that it is reasonable to include false discovery rate SNPs to investigate height QTL sharing.

Two different recent studies identify regions of the dog genome that display signatures of selection. The first uses an F_{ST} based approach (*di*) and identifies 155 SNPs that contribute to breed-defining variation (Akey et al., 2010). SNPs within 200 kb of Akey et al. 2010 SNPs are considered under selection. The second study employs a heterozygosity based approach (*Si*) in addition to the F_{ST} approach (*di*) (Vaysse et al., 2011). An *FDR* < 0.1 was applied to the *di*, and *FDR* < 0.05 to the *Si* results of the Vaysse et al. 2011 study to allow for a reasonable number of regions under selection. Of the 31 SNPs included at *FDR* < 0.0093 for composite dog height, 7 (23%) and 12 (39%) have been identified as under selection by the Akey et al. 2010 and Vaysse et al. 2011 studies, respectively (Figure 1-14). For *FDR* < 0.02 the percentage drops to 11% (8 of 71) and 25% (18 of 71) for the two studies. At *FDR* < 0.04 the proportion of height SNPs associated with selective sweep studies drops to 6% (17 of 289) and 15% (42 of 289). Of the 155 regions under selection from the Akey et al. 2010 study, *FDR* < 0.0093, 0.02, and 0.04 SNPs overlap with 2, 3, and 11 regions, respectively. The Vaysse et al. 2011 study includes 583 regions of which 13, 20, and 35 regions overlap with height SNPs at *FDR* < 0.0093, 0.02, and 0.04 levels of significance, respectively. This pattern suggests, as one might expect, as more SNPs are included for analysis, the ratio of signal to noise decreases.

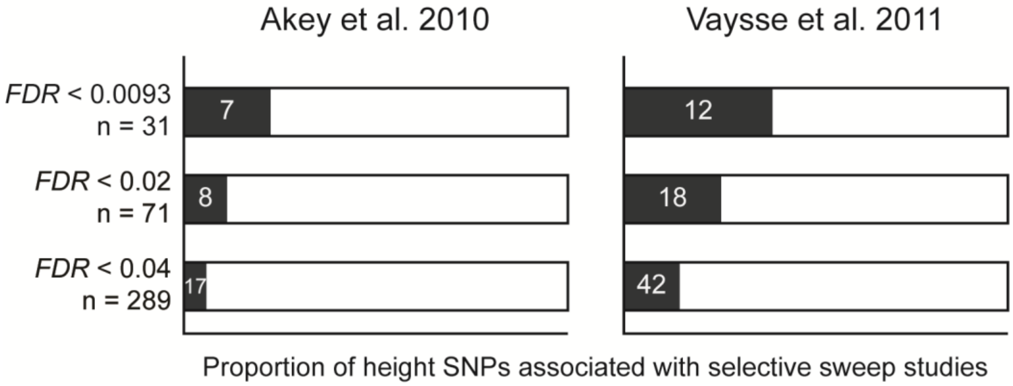

Figure 1-14 Proportion of selective sweep loci that are also composite dog height SNPs

These selective sweep results, combined with the previous gene ontology functional analysis, suggest that the SNPs at *FDR* < 0.02 and 0.04 capture enough signal to justify the cost of added noise in order to more realistically estimate the loci that impact dog height variation. As such, all three levels of significance will be reported when describing the extent of height QTL sharing in humans and dogs.

Identification of Shared Height QTLs

There are two general approaches for identifying shared QTLs. In the SNP-centric approach, associated SNPs from one species are mapped to orthologous loci in another species. For the gene-centric approach, reported genes from one species are mapped to orthologous genes in another species. If the causative mutations for height QTLs tend to be at orthologous DNA sequences, the SNP-centric approach will perform best. If the loci for causative mutations are not well conserved, but still tend to occur in or around the same genes, the gene-centric approach will perform best. Since most human height SNPs are at less conserved loci (32% intergenic and 58% intronic), the gene-centric approach is reported here.

To compare the genetic basis of human and dog height, dog genomic coordinates for human-dog orthologs with HGNC symbols are first retrieved from

Ensembl (version 74 Genes) via BioMart (Durinck et al., 2005, 2009; Kasprzyk, 2011; Vilella et al., 2009). Since dog height associations from Boyko et al. 2010 are given in the CanFam2 assembly, CanFam3 coordinates for orthologs are then mapped to CanFam2 using the UCSC LiftOver tool (Lawrence et al., 2009). Human-dog orthologs from Ensembl are then supplemented with 27 manually annotated orthologs provided in Appendix A. Of the 421 reported human height genes, 378 have orthologs with dogs. Overlap between dog height SNPs and the 378 orthologous human height genes is then computed (Aboyoun et al.; Lawrence et al., 2013). For the Boyko et al. 2010 panel, interbreed linkage disequilibrium (LD) stretches to ~200 kb where mean $R^2 > 0.15$ (Boyko et al., 2010). Based on this estimate of average interbreed LD, human height genes within 200 kb of dog height SNPs are counted as shared QTL. All shared height QTLs in humans and dogs are provided in Table 1-3.

Table 1-3 Human and dog shared height QTLs

Shared QTL	Significance[+]	# Human studies	Additional genes in shared QTL
ACAN	***	5	*POLG*
ANKFN1	*	1	
BMP3	***	2	*PRKG2, RASGEF1B*
BOD1	**	1	*STC2, FBXW11*
C14orf39	*	1	
CDH13	*	1	
DCLK1	*	1	
DTL	*	1	
FGFR3	*	1	*SLBP*
HMGA2	***	10	
IGF1	***	2	*CCDC53, NUP37, C12orf48, PMCH, GNPTAB*
IGF1R	*	1	*ADAMTS17*
IGF2BP2	***	1	
LCORL	***	10	*NCAPG*
NOG	*	3	*DGKE, TRIM25, COIL, RISK*
NPPC	**	3	*PDE6D, COPS7B, DIS3L2, ALPP, PTMA*
NPR3	**	5	*C5orf23*
PBX1	**	1	
PPIF	*	2	
SIX6	*	1	
SMOX	*	1	
SOCS2	*	5	*MRPL42, CRADD, UBE2N*
VGLL2	***	1	
ZFAT	*	1	
ZMIZ1	**	1	

[+] Significance levels: *** for *FDR* < 0.0093, ** for *FDR* < 0.02, and * for *FDR* < 0.04

The 31 *FDR* < 0.0093 composite dog height SNPs overlap with 81 human-dog orthologs, of which 12 genes from 7 regions are shared QTLs (Figure 1-15). The extent of QTL sharing increases at *FDR* < 0.02 with 20 genes from 11 regions shared. At the least conservative level of *FDR* < 0.04, 38 genes from 25 regions influence both human and dog height. This means that at least 10% (25 of 241) of all human height QTLs are shared in dogs. The proportion of shared dog QTLs decreases with additional loci, as does the power to detect true associations. For this reason the proportion of dog shared QTLs is calculated based on the most conservative false discovery rate threshold (*FDR* < 0.0093). Based on this threshold, 44% (7 of ~16 regions) of dog height QTLs are shared;

this is a much higher proportion of sharing than observed in humans. This difference in the proportion of QTL sharing likely reflects the fact that human height QTLs have been much more thoroughly interrogated relative to dog height QTLs. As additional intrabreed and interbreed studies of dog height are completed, the proportion of shared human QTLs should increase from 10%, and the proportion of shared dog QTLs should decrease from 44%. Thus 10% and 44% likely bound the true extent of QTL sharing for height in humans and dogs.

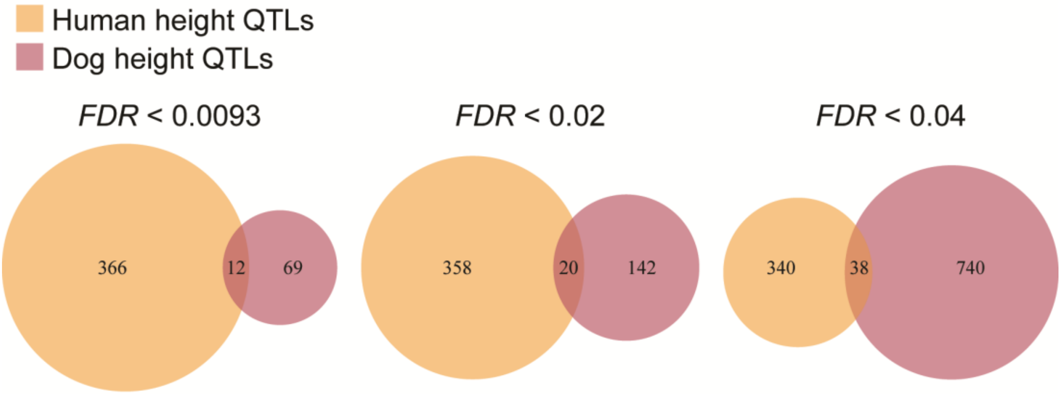

Figure 1-15 Extent of height QTL sharing in humans and dogs

Assessment of the Extent of QTL Sharing

To assess the likelihood of obtaining this extent of QTL sharing by chance, a randomization test is performed similar to a study of height QTL sharing in humans and cattle (Pryce et al., 2011). In the present study there are 378 orthologous human height genes from 241 regions. To assess the extent of QTL sharing, 100,000 random sets of 378 genes are sampled from a collection of 18,853 human-dog orthologs. The number of dog height genes contained within each random set of orthologs is then counted. This forms a distribution that reflects QTL sharing when random orthologs are chosen rather than the reported human height genes. This distribution is compared to the number of regions identified as shared QTL at the various levels of dog height significance (*FDR* <

0.0093, 0.02, and 0.04). The number of shared QTLs at each false discovery rate level changes because what is considered a significant dog height gene changes. Since the 241 regions are still not necessarily independent because they depend on manual delineations, the distribution of random ortholog sets (each containing 378 genes) is compared to shared QTL regions (n=241) rather than genes (n=378) to account for any dependence among the 241 regions and to ensure that the randomization test is conservative.

The extent of QTL sharing at the *FDR* < 0.0093 level of dog height significance (7 regions) is highly unlikely to occur by random chance alone (*P* < 0.0002). At *FDR* < 0.02 only 9 random sets of genes out of 100,000 have more shared QTLs than the 11 observed (*P* < 0.00009). The likelihood of 25 shared QTL at *FDR* < 0.04 is *P* < 0.0078. These results show that the extent of QTL sharing is more than expected by chance at every level of significance considered (Figure 1-16).

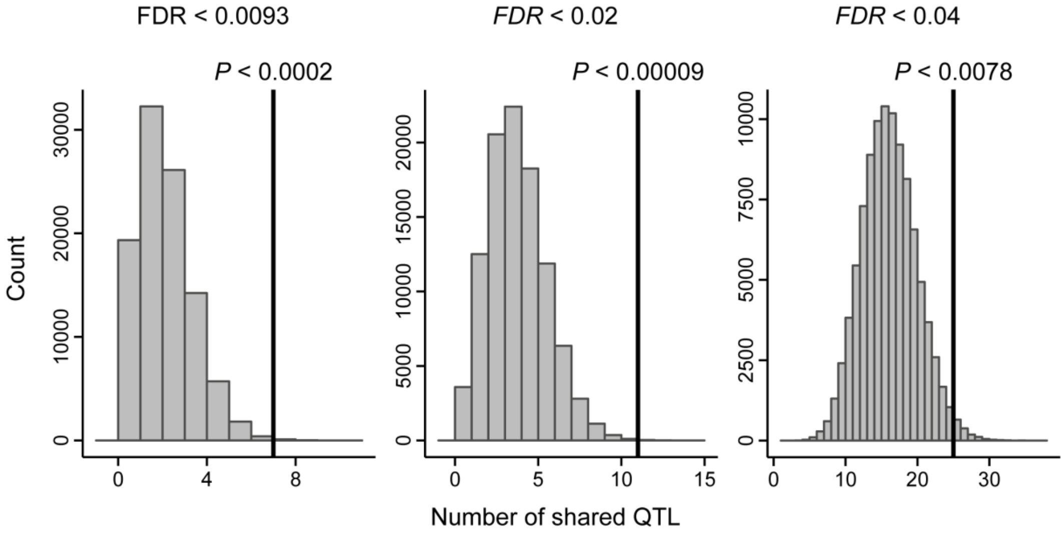

Figure 1-16 Likelihood of QTL sharing with 378 random draws per ortholog set. Vertical lines represent observed sharing with 241 draws.

An argument can be made that since a quarter to a third of all genes is estimated to have a potential impact on size, the pool of random human-dog orthologs should only include these potential genes. To address this criticism, the number of random orthologs can be increased from 378 genes. A reasonable approach is to multiply the number of QTL regions (n=241) by three, thus giving each random ortholog set 723 genes. The logic is that since the 241 shared QTL are only derived from genes that can impact size, the randomization test must take into account the fact that up to two-thirds of random draws have no chance of impacting size.

Such a test reveals that at *FDR* < 0.0093 and 0.02 the extent of sharing is still more than expected by chance (*P* < 0.01 and *P* < 0.02, respectively) (Figure 1-17). However, at *FDR* < 0.04 more shared height QTL are expected than observed (*P* < 0.79), indicating that true associations do taper off at this least conservative level of significance, as expected.

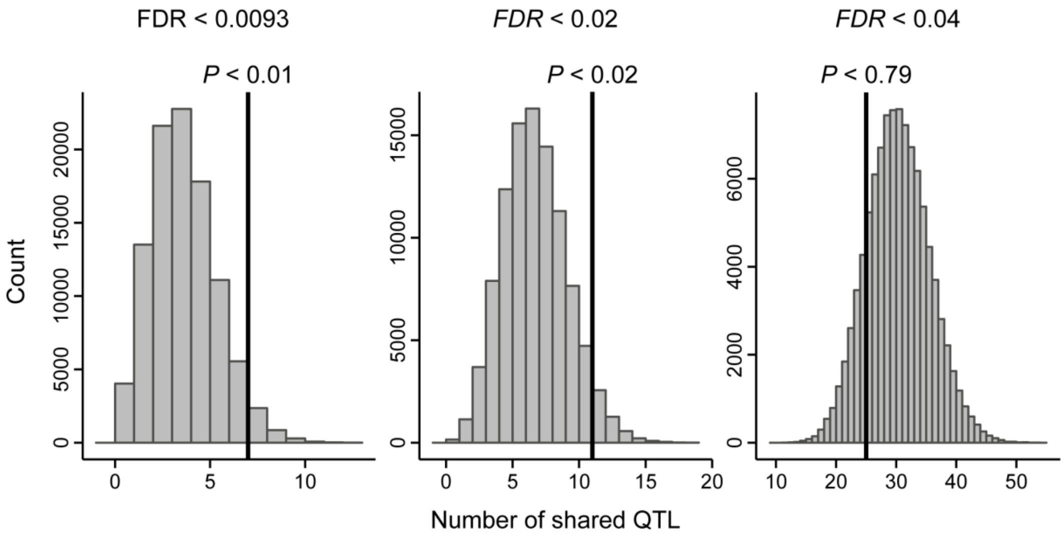

Figure 1-17 Liklihood of QTL sharing with 723 random draws per ortholog set. Vertical lines represent observed sharing with 241 draws.

Notable QTLs

From the NHGRI catalog, ten different studies report that *HMGA2* and *LCORL* influence human height variation. Interestingly, both of these genes have also been shown to influence size variation in the cow and horse (Makvandi-Nejad et al., 2012; Metzger et al., 2013; Pryce et al., 2011; Signer-Hasler et al., 2012). *HMGA2* is a known size QTL in dogs, but this is the first time *LCORL* has been discussed as an interbreed size variant in dogs. As previously mentioned, *LCORL* appears to impact body length in dogs as well as humans (Soranzo et al., 2009). This identification of *LCORL* as an interbreed variant that likely impacts variation in body length demonstrates the type of insights that can be gained from interspecies comparisons of intraspecific variation.

Another interesting observation from these shared QTLs is that *BMP3* impacts dog size. This gene has previously been associated with dog skull diversity (Schoenebeck et al., 2012). In this situation pleiotropy as well as allelic heterogeneity could explain this apparent double association.

Notable human height genes that don't appear to influence interbreed dog height variation include *HMGA1* and *ZBTB38*, genes that associated with height in 11 different human studies. *ZBTB38* is particularly interesting because it is associated with height variation within the Portuguese Water Dog breed (K. Chase and K.G. Lark). This finding underlines the reality that interbreed studies only provide a vignette of the genetic basis of dog size variation because substantial variation also exists within breeds. So while *ZBTB38* is not an interbreed shared height QTL with humans, it is still a shared QTL because it influences size within a breed.

FBN2 and *IHH* are other human height genes not identified as shared QTLs, although like *ZBTB38*, they influence height variation in the Portuguese Water Dog (K. Chase and K.G. Lark). Perhaps unsurprisingly, *FBN2* and *IHH* narrowly miss the *FDR* < 0.04 threshold of significance, demonstrating that meaningful SNPs exist even below the least conservative cutoff applied in this study. It is important to point out that the human height QTLs are more similar to those found in the Portuguese Water Dog because human studies have primarily focused on variation within populations.

Some genes appear to impact intrabreed variation as well as interbreed variation. For example, the gene *IGF1* appears as a shared QTL with the interbreed associations, and is also associated with height variation in the Portuguese Water Dog (Chase et al., 2005; Sutter et al., 2007). Had associations near *FBN2* and *IHH* been slightly more significant, they would have also fallen into this category of intrabreed and interbreed variants.

Despite the considerable overlap of height QTLs, obviously not all QTLs are shared between humans and dogs. A major QTL influencing height in dogs but not humans is the *FGF4* retrogene insertion on chromosome 18 that is responsible for chondrodysplasia in extremely short-limbed breeds like the dachshund and basset hound

(Parker et al., 2009). That this QTL is not shared in humans is unsurprising since the retrotransposition event recently occurred in dogs, and placed a mis-regulated, active copy of *FGF4* into a novel genomic location. Interestingly, a similar condition in humans is caused by mutations in other components of the FGF signaling pathway as mentioned previously in this study in the context of achondroplasia and QTNs (Velinov et al., 1994).

Why Some QTLs Are Missing

Without redefining QTL sharing altogether (for example by including paralogs), it is valuable to consider factors that might lead to an underestimate of the extent of QTL sharing in humans and dogs. One possibility to consider is how limitations of the dog associations for height might lead to missing shared QTLs. The probability that QTLs unique to dogs are due to poor sampling or array coverage in humans is less likely because human height studies have been performed with multiple populations using high density microarrays. This is not the case for unique human QTLs that are missing in dogs.

Closer examination of the distribution of dog SNPs can elucidate technical reasons for why some human QTLs might not appear to influence height in the dog. Of the 378 orthologous human height genes, 7 lack a non-zero SNP that is within the required 200 kb to act as genetic marker (*CYP20A1*, *CS*, *LRRC37B*, *STAT2*, *PASK*, *GFPT2*, *SCMH1*). This means that these genes may be dog height QTLs, but the microarray lacked the coverage to ascertain it. The average number of non-zero SNPs assigned per gene is 8.9, with 8 SNPs as the median. A total of 53 genes have three or fewer non-zero SNPs within 200 kb. These results indicate that the extent of QTL sharing reported in this study is likely underestimated due to limited informative SNP coverage on the dog microarray platform.

Conclusion

At least 263 regions of the genome are associated with variation in human height. Of these regions, 241 have reliable orthologs with the dog. This study compares how many of the same loci contribute to height variation within human and dogs. To make the analysis comparable, height at the withers, body length, and neck length are combined in the dog so as to anatomically capture human height. This composite measure of height yields 289 significant dog SNPs that are validated through a series of gene ontology and selective sweep analyses that demonstrate a substantial portion of associations are meaningful.

Although one might predict that relatively few loci will be shared between species as distantly related as the human and dog, this study finds at least 25 orthologous regions contribute to intraspecific height variation in humans and dogs. This represents over 10% of all human height QTLs (25 of 241) and 44% of dog height QTLs (7 of the top 16 loci).

Relative to the 55% estimate of QTL sharing for simple traits based on candidate gene studies of parallelism and convergence (Conte et al., 2012), the rate of QTL sharing for the highly polygenic trait height is lower (10 - 44%). Interestingly, for more distantly related taxa (like the human and dog), the rate of QTL sharing for candidate gene studies of parallelism and convergence is between 0.1 and 0.4 (Conte et al., 2012), almost exactly the range of QTL sharing observed in this study (0.1 to 0.44). As is the case for traits influenced by a few large effect size loci, highly polygenic traits reflect constraints on genetic variability and variation, likely leading to what we observe as extensive and comparable levels of QTL sharing. These similar rates of QTL sharing for simple and highly polygenic traits suggests that there is nothing fundamentally different between what we call simple and complex traits, other than perhaps the resolution of our understanding.

Despite this extent of QTL sharing, highly polygenic traits like height are still not ideal for candidate gene studies, except perhaps for a few loci that are beginning to emerge across mammalian species like *HMGA2* and *LCORL*. With enough studies like this work, in time perhaps this will change.

Another outcome of this study comes from the decomposition of dog height using genome-wide associations to identify loci that are associated with limb, torso, and neck variation. In addition to confirming the action of *IGF1* and the *FGF4* retrogene, *LCORL* has been associated with body length variation between dog breeds, a finding bolstered by a human study of *LCORL* variation (Soranzo et al., 2009).

Chapter 2

Effect Size Distributions for Height in Humans and Dogs

Introduction

Chapter 1 examined the extent of height QTL sharing in humans and dogs. After identifying QTLs that influence how a phenotypic character varies, the next step is often to estimate the relative effect sizes of contributing QTLs. The effect sizes of QTLs determine how many loci are necessary to explain a meaningful proportion of trait variation. Effect sizes depend on the substitution effect of replacing one allele for another as well as allele and genotype frequencies.

The range of effect sizes can span from Mendelian, where one or two loci explain the entirety of trait variation, to infinitesimal where an infinite number of loci contribute equally (Figure 2-1) (Fisher, 1918; Mather, 1943; Robertson, 1967). Despite the desire by some researchers for phenotypic characters to be influenced by a few large effect size loci, most traits are influenced by many loci that resemble the exponential distribution of effect sizes (Flint and Mackay, 2009; Plomin et al., 2009).

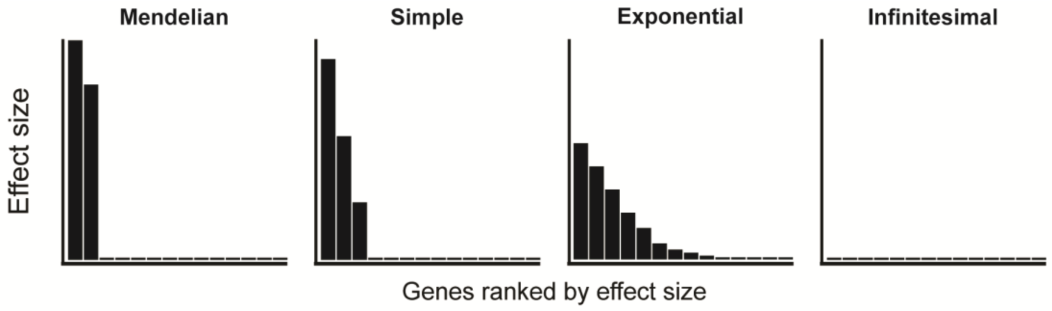

Figure 2-1 Distributions of QTL effect sizes

This exponential distribution of effect sizes was the prevailing assumption going into human genome-wide association studies. The thought was that some of the largest effect size QTLs could individually explain up to 10% of trait variation in humans (Flint and Mackay, 2009). As mentioned in Chapter 1, results were not as expected: hundreds of loci are apparently necessary to explain 10% of trait variation.

This observation has prompted widespread discussion about the missing heritability for complex traits in humans. In some cases this has led to the conclusion that rather than an exponential model, perhaps the effect size distribution of QTLs in humans more closely resembles the infinitesimal model (Makowsky et al., 2011). On the other extreme, reports in dogs and horses have suggested that quantitative traits have been simplified in these domesticates and only require a few QTL to explain the majority of trait variation (Boyko et al., 2010; Makvandi-Nejad et al., 2012). Chapter 2 discusses QTL effect sizes in humans and formally tests the idea of a simple genetic architecture in dogs.

The Effect Size Distribution of Human Height

The Problem of Missing Heritability

As the genomics era began near the turn of the century 20th, Eric Lander suggested that common variants might explain a substantial portion of common diseases (Lander, 1996). What became known as the 'common disease, common variant' hypothesis contributed to the optimism going into human genome-wide association studies (Knight, 2009). When genome-wide studies started rolling out, many novel variants were identified, but a relatively small amount of the heritable portion of trait variation could be explained. This observation led many researchers to speculate where this missing heritability might be for complex traits like height, type 2 diabetes, and autism (Donnelly, 2008; Maher, 2008).

Ideas Concerning Missing Heritability

Much literature has been published debating and speculating the topic of missing heritability; several of the major themes will be mentioned here. One explanation is that rare variants contribute substantially to the phenotypes of complex traits (Brachi et al., 2011; Eichler et al., 2010; Manolio et al., 2009). This possibility can be explored with forthcoming affordable high-throughput sequencing technology that can capture many rare variants as well as classes of variants beyond SNPs (Luo et al., 2011). Another explanation for missing heritability is allelic heterogeneity where there are multiple functional alleles with different phenotypic consequences at the same locus (Bergelson and Roux, 2010; Wood et al., 2011; Zhang et al., 2012). Although, by definition, epistasis does not explain missing narrow-sense heritability, it has been suggested as a source of creating overestimates of heritability (Zuk et al., 2012). Yet another source of proposed missing heritability is transgenerational epigenetic variation (Johannes et al., 2009).

Another perspective offered by Greg Gibson is that the debate over missing heritability is overblown (Eichler et al., 2010). He argues that hidden environmental structure is not being accounted for and is causing estimates of narrow-sense heritability to be inflated. He goes on to clarify his position as follows:

> More fundamentally then, there is a missing genetic variance problem, which really relates to misplaced preconceptions. It would have been nice if GWA studies typically uncovered a dozen associations each explaining 5–10% of the variance, but the fact that they do not suggests only that the allelic effects are smaller or the causal alleles are too rare.

The reality is that for most complex traits, the missing heritability is probably in a lot of places, including the possibility that it has been overestimated to begin with.

For human height, one of the most extensive genome-wide studies explains ~10% of the phenotypic variation in height with 180 loci (Lango Allen et al., 2010). Another study takes the approach of considering almost 300,000 SNPs simultaneously and explains 45% of variance in height (Yang et al., 2010). The authors argue that the heritability is not missing, but is in QTLs with effect sizes too small to pass current significance tests. They go on to point out that low allele frequencies and incomplete linkage disequilibrium between causal variants and SNPs are the reason why many QTLs have such low effect sizes. Clearly many loci contribute to human height variation, but there is no reason to assume that they have equal effect sizes, or are all small.

Effect Size Distributions

Perhaps the most useful construct to use when considering the loci that contribute to complex traits is the distribution of QTL effect sizes (Figure 2-1). Even if the largest loci that contribute to trait variation only explain 1% of variation, the distribution of QTL effects could still more closely resemble the exponential rather than the infinitesimal model (Flint and Mackay, 2009). From sample composition to technical considerations such as whether SNPs are effectively tagging causal variants, multiple factors likely contribute to small effect sizes coming out of genome-wide association studies.

For example, human height associations within populations don't typically include individuals with extreme height differences like those affected by achondroplasia. As mentioned in Chapter 1, one form of short-limbed dwarfism is primarily the result of a single nucleotide polymorphism (G380R) in *FGFR3* (Bellus et al., 1995). Even if a few individuals with achondroplasia were included in height genome-wide associations, the *FGFR3* QTL would have a relatively small effect size because effect sizes also depend on allele and genotype frequencies. Thus in order for a variant to have a large effect size in a genome-wide association it must have a large substitution effect as well as be common. This excludes rare variants like G380R in *FGFR3* that have very large effect sizes.

The take-home message here is that the absence of large effect size QTLs really reflects the frequency-distribution of substitution effect sizes. The effect size distribution of height in humans is exponentially distributed, although depending on the study population, this distribution will be shifted towards larger or smaller effect sizes.

Effect Sizes in Model Organisms

Jonathan Flint and Trudy Mackay provide an insightful and relevant perspective on the effect sizes of quantitative traits in humans as well as in flies and mice (Flint and Mackay, 2009). As argued in this dissertation, Flint and Mackay also assert that effect sizes in humans are most consistent with an exponential distribution.

They describe how initial studies in mice and flies revealed QTLs with large effects that could account for large portions of phenotypic variation. These studies in mice and flies influenced expectations for human studies and likely contributed to the missing heritability conundrum. Interestingly, as the sample sizes of studies in flies and mice increased, the number of QTLs also increased, and effect sizes decreased, a pattern similar to human studies (Lai et al., 2007; Turri et al., 2001a, 2001b). Since allelic effect sizes depend on substitution effects as well as allele frequencies, it makes sense that effect sizes can be inflated from estimates based on relatively few samples.

Flint and Mackay also point out that inbreeding can drastically alter perceptions of genetic architecture, making traits appear simplified in inbred populations. After considering studies from humans, mice, and flies, Flint and Mackay conclude that the genetic architecture of quantitative traits is fairly consistent across both traits and species and fits the exponential distribution of gene or QTL effects (Flint and Mackay, 2009).

The Effect Size Distribution of Dog Height

A Simple Genetic Architecture?

In contrast to studies of quantitative traits in humans, mice and flies, a fairly recent study in dogs argues that only a few QTLs are necessary to explain the majority of variation for morphological traits (Boyko et al., 2010). The authors conclude that domestication and breed formation have simplified the genetic architecture of morphological traits in dogs relative to humans (Boyko et al., 2010). This study consisted of genome-wide associations of over 60,000 SNPs with 57 breed measures for 915 dogs from 80 domestic breeds. Ironically, some of the data from this study was leveraged to identify at least 25 shared height QTLs in humans and dogs in Chapter 1 of this dissertation.

The conclusion of a simple genetic architecture is based on the finding that for most traits, >70% of trait variation can be explained by the action of three or fewer loci. Consistent with this idea of a simple genetic architecture, variants of the *IGF1* gene have been associated with dog size differences between many breeds, explaining an enormous 50% of phenotypic variation (Boyko et al., 2010; Chase et al., 2002; Sutter et al., 2007). Curiously, a notable exception that does not align with an *IGF1* effect size of near Mendelian proportions is the large Rottweiler breed that has the small dog *IGF1* allele (Sutter et al., 2007).

The data invoked by Boyko et al. 2010 to support the model of a simple genetic architecture in dogs is that estimated QTL effects explain a large proportion of variation. Since the whole hypothesis rests on estimated QTL effect sizes, accuracy of these estimates is essential. Instead of highlighting QTL effect sizes for height at the withers or body length, Boyko et al. 2010 focus on the more environmentally influenced metric of body weight. They find that *IGF1* explains R^2 = 50% of variance in purebred dogs and R^2

= 17% in a sample of 50 outbred village dogs from Puerto Rico and Africa (Boyko et al., 2009). While a QTL effect size of R^2 = 17% is certainly larger than what has been observed for quantitative traits studied in humans, it is more consistent with an exponential QTL effect size distribution, not a simple genetic architecture as proposed by Boyko et al. 2010.

Rather than letting the results of the village dogs temper the idea of a simple genetic architecture, the authors of Boyko et al. 2010 marginalize these data, citing reduced linkage disequilibrium and non-genetic factors in village dogs. The National Geographic popularized this idea in February of 2012 with the article titled *Mix*, *Match*, *Morph*: "Flip a few switches, and your dachshund becomes a Doberman.... Flip again, and your Doberman is a Dalmatian."

In contrast to the conclusions of Boyko et al. 2010, the breeding history and resulting phenotypic distributions of dog height within and between breeds suggests the action of many variants rather than just a few of large effect (Figure 1-6). This observation and recognition of height as highly polygenic is supported by the practical experience of dog breeders (Greyhound, 2014). In addition, laboratory animals that have gone through somewhat similar breeding conditions as dogs still retain a complex genetic architecture for quantitative traits (Flint and Mackay, 2009). To clarify this issue I test the hypothesis of a simple genetic architecture in various ways by leveraging data from Boyko et al. 2010 as well as by studying an independent sample of mixed-breed dogs.

Genome-wide Association Data

The first conclusion from Boyko et al. 2010 to consider is the number of loci that contribute to morphological trait variation. How many associations are significant for each morphological trait considered? The Boyko et al. 2010 study employs two different levels of significance, one for absolute traits (P < 5.0e-5) and one for skeletal and proportional traits (*P* < 1.0e-4). Based on these definitions of significance, the 57 traits have an average of 18 significant SNPs from an average of 7 different chromosomes (Figure 2-2). Since many traits are heavily influenced by body size, the authors also provide allometric associations that have been normalized by the logarithm of body weight. For the allometric associations, an average of 18 SNPs are significant from an average of 9 different chromosomes. After accounting for the fact that multiple associations sometimes tag the same locus, I find that on average between 7 and 9 different loci contribute to trait variation. Since *P* < 5.0e-5 and *P* < 1.0e-4 are fairly conservative cutoffs, the number of loci contributing to interbreed trait variation could be much higher.

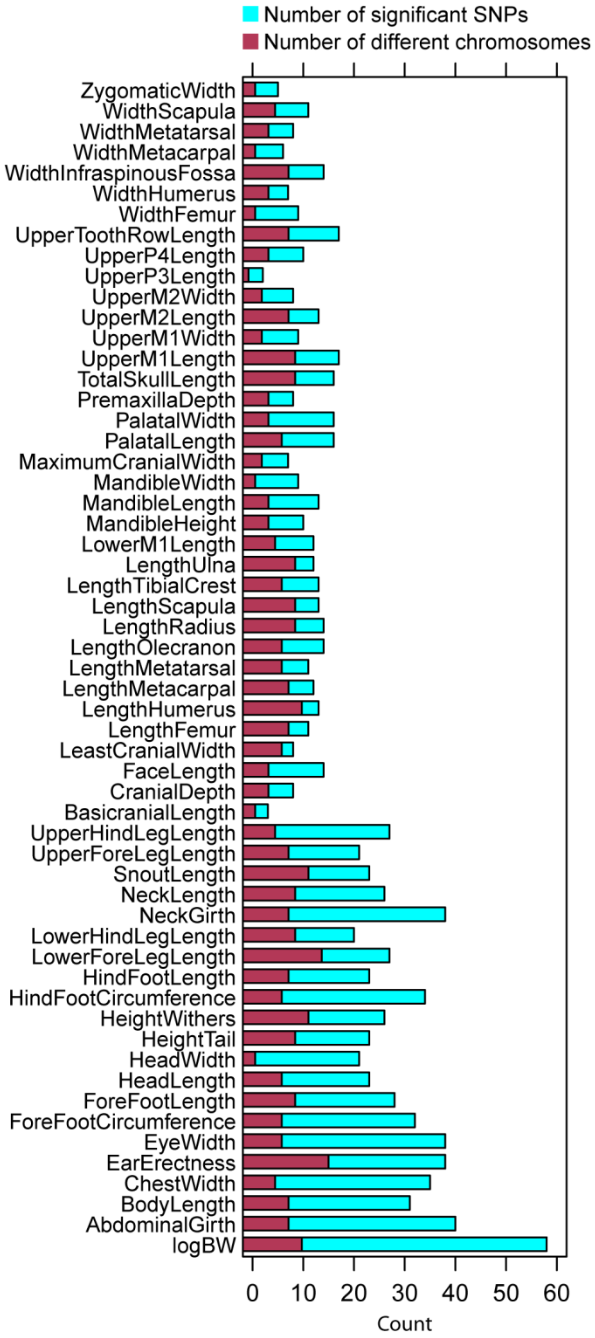

Figure 2-2 Number of different SNPs and chromosomes associated with morphological differences between dog breeds.

While it appears as though more than two or three loci contribute to each trait in the Boyko et al. 2010 data, the most relevant question in regards to teasing apart the genetic architecture of morphology is what are the effect sizes of these variants? If a few QTLs are indeed responsible for >70% of trait variation, there is an argument for a simplified genetic architecture. The next part of this analysis examines SNPs associated with height at the withers to explore QTL effect sizes.

Rather than examine effect sizes with the composite measure of height used in Chapter 1, I will examine height at the withers individually here. Height at the withers has 26 significant associations from 11 different chromosomes at the Bonferroni level of significance (Figure 2-2). The Boyko et al. 2010 study reports p-values for associated SNPs as well as allele frequencies by breed, but phenotypic breed information is unfortunately not provided. This is part of why height at the withers was chosen because between previous studies of height and reported breed standards, a fairly reliable dataset of breed heights can be assembled (Appendix B) (Alderton, 2008; Sutter et al., 2008). This is not an ideal approach, but it is at least reasonable because the Boyko et al. 2010 study also used breed averages for phenotypes rather than individual dog measures. Without the raw data it is impossible to reconstruct the exact associations from the Boyko et al. 2010 study; nonetheless, calculated associations for height at the withers do mirror those from Boyko et al. 2010 fairly well and will be used as a proxy for the original p-values (Figure 2-3). Reconstructed height at the withers associations are based on 70 breeds that have at least 9 dogs per breed. Associations are calculated in R using the base linear model function (R Core Team, 2013).

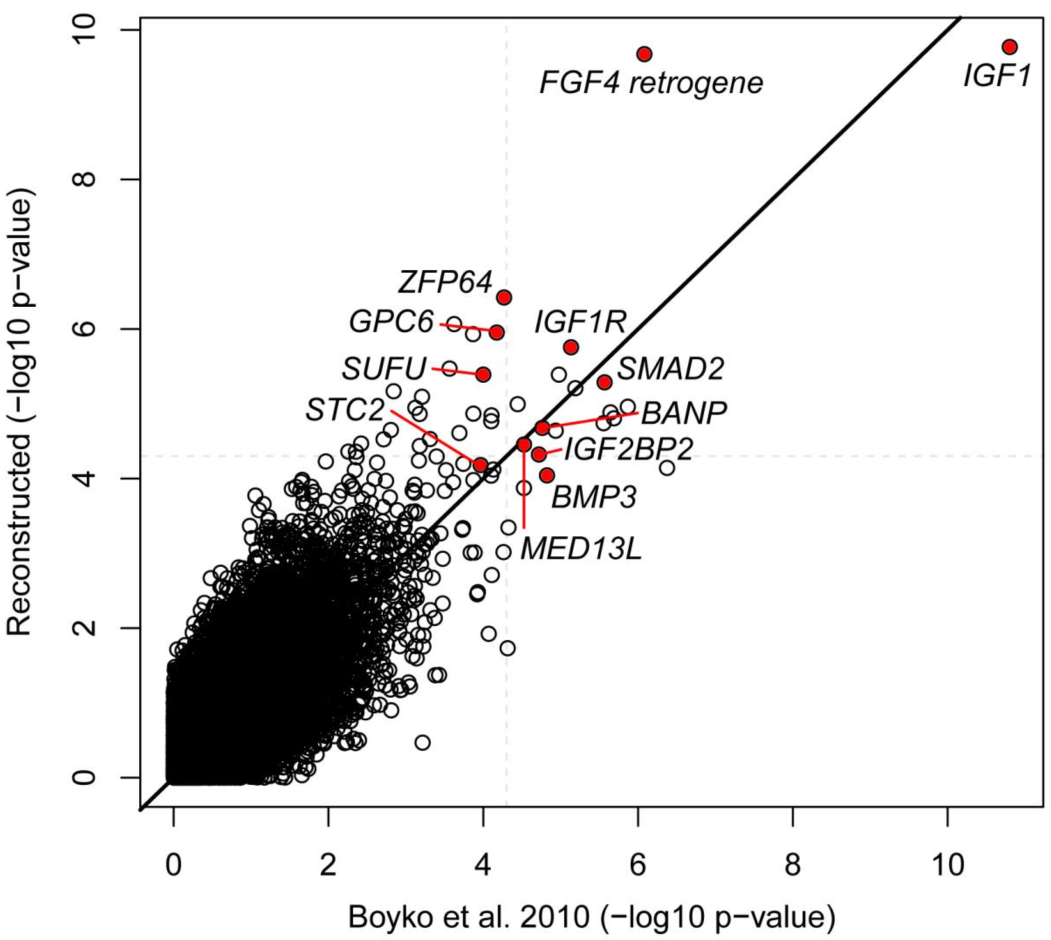

Figure 2-3 Comparison of Boyko et al. 2010 p-values for height at the withers with reconstructed p-values from this study. Twelve loci chosen for additional analysis are highlighted in red. Dotted line represents Boyko et al. 2010 Bonferroni significance.

Consistent with the report of Boyko et al. 2010, the SNP associated with *IGF1* explains 45% of height at the withers and the *FGF4* retrogene also explains 45% of variation. Note that these associations are based on reconstructed p-values, and that the reconstructed p-value for *FGF4* retrogene is substantially more significant than the Boyko et al. 2010 p-value, suggesting that *FGF4* retrogene likely explains less than 45% (Figure 2-3). Nonetheless, in concert these two loci explain 70% of variation in height at the withers between breeds (adjusted R^2), consistent with a simple genetic architecture. Since *IGF1* and *FGF4* retrogene do not collectively explain 90% (45+45) of interbreed variation, they must be explaining somewhat overlapping portions of height variation. Curiously, including the next most significant SNP near *ZFP64* only increases the percent of variation explained from 70% to 71%. By including all 12 loci highlighted in red (Figure 2-3), the percent of variation explained rises to 82%. Interestingly, if *IGF1* is removed from this 12 locus model, the percent of variation explained only drops by 5 percent. From a statistical perspective, this suggests multicollinearity where multiple predictor variables are correlated with each other.

Examining some of the other significant loci can provide further hints if multicollinearity is undermining the 12 locus model. Surprisingly, the SNPs near the next three most significant loci (*IGF1R*, *GPC6*, *SUFU*) explain a whopping 54% of variation. The next three most significant SNPs near *SMAD2*, *BANP*, and *MED13L* explain 47% of variation, and the last three loci highlighted in red (*IGF2BP2, STC2, BMP3*) explain 40% of variation. Clearly these 12 variants are explaining overlapping portions of interbreed height variation.

When variants explain overlapping portions of phenotypic variation this makes the linear model unstable, complicating estimates of individual QTL effect sizes. Can the variants near *IGF1* and *FGF4* retrogene explain 70% of variation? Yes, this has just been observed. Do the variants near *IGF1* and *FGF4* retrogene cause 70% of interbreed height variation? If all of the other variants like *SMAD2* don't have a meaningful contribution, then yes, that would suggest that they not only explain, but cause 70% of interbreed height variation. However, recent work has further validated some of these other height loci by identifying the causative mutations behind the *SMAD2, STC2,* and *IGF1R* associations, so clearly *IGF1* and *FGF4* retrogene do not cause 70% of interbreed height variation (Hoopes et al., 2012; Rimbault et al., 2013). This presents a bit of a conundrum: how can variants like *IGF1* and *FGF4* retrogene explain variation that they do not exclusively cause?

Assigning effect sizes according to variation explained from an interbreed study implicitly assumes that all other loci are random with respect to the locus of interest. If multiple alleles contribute to differences between breeds, inasmuch as the effects of those alleles are correlated, estimated individual effect sizes will represent the action of all correlated alleles, which can lead to vastly overestimated individual effect sizes. In this way correlated allelic effects can give the caricature of a simple genetic architecture as variants have high predictive power, but conceal the action of other meaningful loci. This scenario is most likely to occur for polygenic traits that undergo intense directional selection, like the selection dogs have experienced for reduced body size the last several hundred years. Thus, *IGF1* and *FGF4* retrogene can be highly predictive of interbreed height differences, but are not the whole reason or cause of those differences. Perhaps the best way to describe this phenomenon is that it is correlation with partial causation.

To test this hypothesis that correlated allelic effects lead to an overly simplistic view of genetic architecture, allelic correlations, or linkage disequilibrium, can be examined between SNPs. Unfortunately standard measures of linkage disequilibrium (LD) cannot be calculated because individual genotypes are not given in the Boyko et al. 2010 data. Instead, allele frequencies are given for each SNP according to breed. Thus for this analysis, allele frequencies from 70 breeds, each containing at least 9 dogs, are compared between loci using the Pearson product-moment correlation coefficient (*R*). Correlations are given relative to the *IGF1* SNP because it is the most significant association and the reconstructed p-value is quite similar to the Boyko et al. 2010 p-value.

After a Bonferroni correction for multiple testing, the allele frequencies for the *IGF1* SNP are significantly correlated with the allele frequencies of 6 of the other top 11 SNPs in Figure 2-3 (*STC2*, *BANP*, *SMAD2*, *GPC6*, *ZFP64*, *IGF2BP2*). This high degree of correlation between allele frequencies suggests that correlated allelic effects are inflating effect size estimates of some of the top loci like *IGF1*.

To examine the extent of correlated allelic effects, 12 loci on chromosomes other than those that contain the top 12 loci are randomly chosen with -log10 p-values between 2 and 2.5. Other chromosomes are chosen so that there is no linkage between these loci and the top 12 loci, thus avoiding signal due to the top loci. After a Bonferroni correction, the allele frequencies of 2 of these 12 mid-range loci are significantly correlated with the *IGF1* SNP. These results suggest that correlated allelic effects well beyond the most significant loci contribute to an overly simplistic view of genetic architecture by inflating effect size estimates of top loci like *IGF1*.

This analysis doesn't mean that *IGF1* and *FGF4* retrogene have small effect sizes, but it does suggest that they do not cause 70% of variation between breeds. Based on an intrabreed study of *IGF1* in the Portuguese Water Dog, this variant is reported to explain ~15% of skeletal size variation (Chase et al., 2005; Sutter et al., 2007). Because this estimate is within a single breed it is more insulated from correlated allelic effects and thus represents a more accurate approximation of the substitution effect size of the *IGF1* variant. Relative to humans and other species, 15% is certainly a large effect, but it does not approach the 45% estimated earlier in this study. It is therefore reasonable to predict that *IGF1* causes no more than ~15% of interbreed height variation, but can explain, through the action of correlated alleles, ~45% of variation.

An Independent Sample of Mixed-breed Dogs

Another way to test the hypothesis that correlated allelic effects lead to an overly simplistic view of genetic architecture is to estimate individual effect sizes in a panel of mixed-breed dogs with a similar distribution of height variation as the purebreds used in Boyko et al. 2010 (Figure 2-4) (Appendix B). The admixture in mixed-breed dogs, like the village dogs studied in Boyko et al. 2010, can break up correlated alleles to some extent and allow for a more accurate estimate of individual QTL effect sizes. If the effect sizes of large effect loci are smaller in mixed-breeds than previously reported in purebreds, this would suggest an underlying genetic architecture more similar to the exponential QTL effect size distribution than the simple model (Figure 2-1).

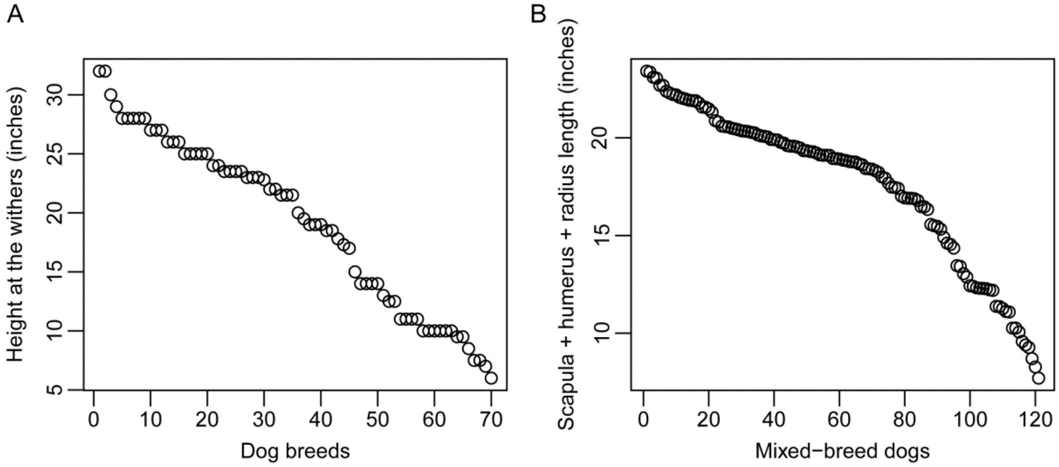

Figure 2-4 Comparison of (A) height at the withers by breed with (B) scapula + humerus + radius length in a panel of mixed-breed dogs.

The 121 mixed-breed dogs used in this study are former pets obtained postmortem through Skulls Unlimited International Inc. (Oklahoma City, OK, USA). Because these mixed-breed dogs are former pets, it can be assumed that the quality of care that might impact a trait like height is comparable to that received by purebred pets. In addition to the full skeleton for each mixed-breed dog, tissue samples were obtained for molecular analysis. To control for age and sex, all dogs in the mixed-breed panel are adult males.

The sum of the lengths of the scapula, humerus, and radius serves as a reasonably comparable measure for height at the withers in live dogs (Figure 2-5A). Maximum bone lengths were measured to the nearest 0.0001 inch with digital calipers for both the left and right (Figure 2-5B). Left and right measurements were then averaged to obtain a single value for each bone. To estimate measurement error, 11 random dogs were measured again and had an average correlation (*R*) of 0.9998 with previous measurements, indicating that these measurements reliably capture bone lengths.

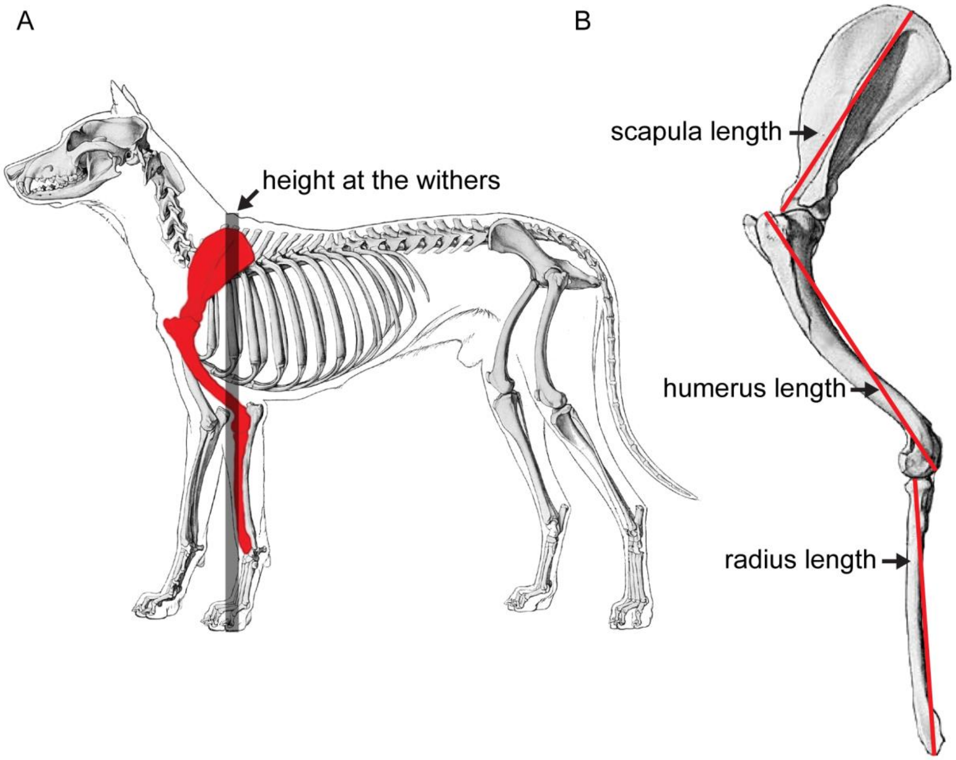

Figure 2-5 (A) Scapula, humerus, and radius (red) as a proxy for height at the withers (black line) in live dogs. (B) Specific measurements for scapula, humerus, and radius.

For each mixed-breed dog whole genomic DNA was extracted from 10 mg of tissue. Cells were lysed in 500 µL of PK buffer (0.025 µL 1M Tris (pH8), 0.001 µL 0.5M EDTA (pH8), 0.01 µL 5M NaCl, 0.025 µL 10% SDS, 0.439 µL H_2O) with 10 µL proteinase K overnight at 55°C. Samples were transferred to Phase Lock Gel Heavy 2.0 mL tubes (5 Prime) and 0.5 mL phenol:chloroform:isoamyl alcohol (PCI 25:24:1) was added and mixed by inversion. Following centrifugation for 30 minutes, 0.5 mL chloroform was added and mixed by inversion. After 15 minutes of centrifugation the aqueous phase was transferred to a new 1.5 mL tube and 1 mL cold 95% ethanol was added and mixed by inversion. Following 15 minutes of centrifugation alcohol was decanted and 1 mL of 70% ethanol was added and mixed by inversion. After 5 minutes of centrifugation ethanol was decanted once more and samples were dried for at least 6 hours before being resuspended in 100 µL sterile H_2O. A 1:50 dilution of extracted genomic DNA was used for PCR-based assays.

The strength of using mixed-breed dogs or village dogs is that linkage disequilibrium between influential loci is reduced, allowing for a more accurate estimate of individual QTL effect sizes. The challenge with this type of panel is that linkage disequilibrium often breaks down between causative mutations and the genomic markers as well. This is one of the primary reasons why the Boyko et al. 2010 study trusted purebred effect sizes rather than the substantially smaller village dog estimates. This study circumvents the problem by primarily genotyping causative variants. Variants from five different genes will be examined here: *IGF1*, *FGF4* retrogene, *SMAD2*, *STC2*, and *BANP* (Table 2-1).

The *IGF1* QTL on chromosome 15 contains three variants that could individually or in concert contribute to a reduction in size. There is a diagnostic SNP in the second intron that has an allele unique to dogs relative to the ancestral gray wolf (Gray et al., 2010). In complete linkage disequilibrium with this SNP is a ~200 bp bimorphic short-interspersed transposable element (SINE) also in the second intron of *IGF1* (Gray et al., 2010). In addition to these variants, a microsatellite in the promoter of *IGF1* is also strongly associated with a reduction in size. The SINE is a strong candidate to be the causative mutation, and because all three variants are in strong linkage disequilibrium with each other, the SINE is genotyped in the mixed-breed dogs.

Primers were designed (Untergasser et al., 2007) to amplify an ~850 bp product with the diagnostic short-interspersed element (SINE) and a 643 bp product without it (Table 2-1). Polymerase chain reactions (PCRs) were run as 10 µL reactions with 5 µL Epicentre PreMix D master mix, 0.3 µL forward primer, 0.3 µL reverse primer, 0.1 µL New

England Biolabs *Taq* DNA polymerase, 3.3 µL H_2O, and 1 µL genomic DNA. The following PCR cycle conditions were run on the MJ Research PTC-200 DNA Engine: 96°C for 3 minutes, 40 cycles of [96°C for 30 seconds, 65°C for 60 seconds, 72°C for 2 minutes], 72°C for 10 minutes. PCRs were run on a 1% agarose/TBE gel with a 100 bp New England Biolabs DNA ladder (Figure 2-6A).

Table 2-1 Primers and associated amplicon lengths for height variants

Gene	Forward	Reverse	Amplicon length (bp)
IGF1 (PCR)	GGGCCTGGTCTTCTGCACTG	GGGACTGGCCAAGTCTCAGC	~850/643
IGF1 (qPCR)	GGGCCTGGTCTTCTGCACTGATATT	TGCCCCCAGCTGCCCTAAGA	~623/416
FGF4 retrogene (set 1)	GTCCTGCTGGCGGTGCTG	GGGGAGGAAGTGGGTGACCT	~1700/570
FGF4 retrogene (set 2)	TGTGACACACAGATGGACCATGA	CTCTCCCCCTTTCCCTCTGG	164
SMAD2 (set 1)*	GGAAGCCTTAGGGGATTTTG	CTCCACCACCCACAGAAACT	683
SMAD2 (set 2)*	GGCATGGGAGAGTGACCTAA	GAGCAGCCTGTGAAGGAAAC	440
STC2	CCGTTCCAGAGCCTCTACAC	GAGCTCCCTATGGTTCCAGC	296
BANP	TTTCCTCAGCTGCCACCTTC	GCTGCAGAAGCCTAGCTACA	254

* Primers reported in Rimbault et al. 2013

To verify traditional PCR results, the *IGF1* PCR assay was redesigned for quantitative PCR (qPCR) melt-curve analysis (Table 2-1). qPCRs were run as 20 µL reactions with 10 µL Promega GoTaq qPCR master mix, 1 µL forward primer, 1 µL reverse primer, 0.2 µL ROX, 5.8 µL H_2O, and 2 µL genomic DNA. The following qPCR cycle conditions were run on the Applied Biosystems 7300 Real-Time PCR System: 50°C for 2 minutes, 95°C for 10 minutes, 40 cycles of [95°C for 15 seconds, 60°C for 60 seconds, 72°C for 2 minutes], dissociation step (95°C for 15 seconds, 60°C for 30 seconds, 95°C for 15 seconds). Genotype calls are based on the combination of melt-curve plots as well as the ratio of P2/P1 (Figure 2-6B).

Multiple SNPs are significant near *STC2* and *BOD1* on chromosome 4 for the association with height at the withers in Boyko et al. 2010. This locus was also investigated in the more recent Rimbault et al. 2013 study. At this point no specific variant has been identified as the most likely causative mutation, so primers were designed to amplify a 296 bp product that contains a SNP reported in Boyko et al. 2010 with a diagnostic restriction enzyme site (HpyCH4V) to form a restriction fragment length polymorphism (RFLP) (Table 2-1). PCRs were run as 10 µL reactions with 5 µL Epicentre PreMix G master mix, 0.5 µL forward primer, 0.5 µL reverse primer, 0.1 µL New England Biolabs *Taq* DNA polymerase, 2.9 µL H_2O, and 1 µL genomic DNA. The following PCR cycle conditions were run on the Applied Biosystems Veriti Thermal Cycler: 95°C for 3 minutes, 40 cycles of [95°C for 30 seconds, 63°C for 30 seconds, 72°C for 60 seconds], 72°C for 5 minutes. Restriction digestions were run for 1.5 hours at 37°C with 10 µL PCR product, 0.5 µL HpyCH4V, and 1.0 µL NEB buffer 4 from New England Biolabs. RFLPs were run on a 2.5% agarose/TBE gel with a 100 bp New England Biolabs DNA ladder. The HpyCH4V enzyme also cuts another site within the PCR amplicon besides the diagnostic SNP such that the C allele results in two products: one that is 26 bp and another that is 270 bp. The T allele results in three products: one that is 26 bp, one that is 96 bp, and one that is 174 bp. The 26 bp product is indistinguishable from excess primers, but the other RFLP producs allow genotypes to be determined (Figure 2-6C).

The *FGF4* retrotransposition on chromosome 18 is tagged by a SNP in the Boyko et al. 2010 association for height at the withers; in this study the retrotranspostion is directly tested in mixed-breed dogs. To capture all possible genotypes two sets of primers were designed (Table 2-1). Since the retrotransposition lacks introns, one set of primers (primer set 1) was designed within *FGF4* that yields a 570 bp product for the retrotransposition allele and a ~1700 bp product for the endogenous *FGF4*. The ~1700 bp product pushes the technical capacity for reliable PCR amplification and so another set of primers (primer set 2) was designed at the retrotransposition insertion site that yields a 164 bp product for alleles that do not have the retrotransposition. PCRs were run as 10 µL reactions with 5 µL Epicentre PreMix G master mix (primer set 1) and PreMix D (primer set 2), 0.5 µL forward primer, 0.5 µL reverse primer, 0.1 µL New England Biolabs *Taq* DNA polymerase, 2.9 µL H_2O, and 1 µL genomic DNA. The following PCR cycle conditions were run on the MJ Research PTC-200 DNA Engine: 96°C for 3 minutes, 40 cycles of [96°C for 30 seconds, 67°C (primer set 1) and 66°C (primer set 2) for 30 seconds, 72°C for 60 seconds], 72°C for 10 minutes. PCRs were run on a 2% agarose/TBE gel with a 100 bp New England Biolabs DNA ladder (Figure 2-6D).

Like *IGF1* and *FGF4* retrogene, the Boyko et al. 2010 association for height at the withers identifies significant SNPs near *SMAD2* on chromosome 7. A recent study proposes that a pair of deletions ~15 kb from *SMAD2*, and in complete linkage disequilibrium with each other, could be cis-regulatory causative mutations (Rimbault et al., 2013). The 9.9 kb deletion, the larger of the two, is genotyped by Rimbault et al. 2013 with two sets of primers (Table 2-1). This study follows the same protocol as outlined in Rimbault et al. 2013 where the first primer set detects the allele without the deletion and the second primer set detects the 9.9 kb deletion allele (Figure 2-6E).

On chromosome 5 a SNP near the gene *BANP* was identified in the Boyko et al. 2010 study, although it was not specifically discussed. Primers were designed to amplify a 254 bp product containing a diagnostic restriction enzyme site (NlaIII) (Table 2-1). PCRs were run as 10 µL reactions with 5 µL Epicentre PreMix G master mix, 0.5 µL forward primer, 0.5 µL reverse primer, 0.1 µL New England Biolabs *Taq* DNA polymerase, 2.9 µL H_2O, and 1 µL genomic DNA. The following PCR cycle conditions were run on the Applied Biosystems Veriti Thermal Cycler: 95°C for 3 minutes, 40 cycles of [95°C for 30 seconds, 63°C for 30 seconds, 72°C for 60 seconds], 72°C for 5 minutes. Restriction digestions were run for 3 hours at 37°C with 10 µL PCR product, 0.05 µL NlaIII, and 1.525 µL NEB buffer 4, and 0.425 µL (1:10 dilution of 100X) BSA from New

England Biolabs. RFLPs were run on a 2.5% agarose/TBE gel with a 100 bp New England Biolabs DNA ladder. Like with the *STC2* RFLP, the *BANP* PCR product contains two restriction sites, one of which is for the diagnostic SNP. The T allele results in three products: one that is 52 bp, one that is 79 bp, and one that is 115 bp. The C allele results in two products: one that is 79 bp and one that is 171 bp (Figure 2-6F).

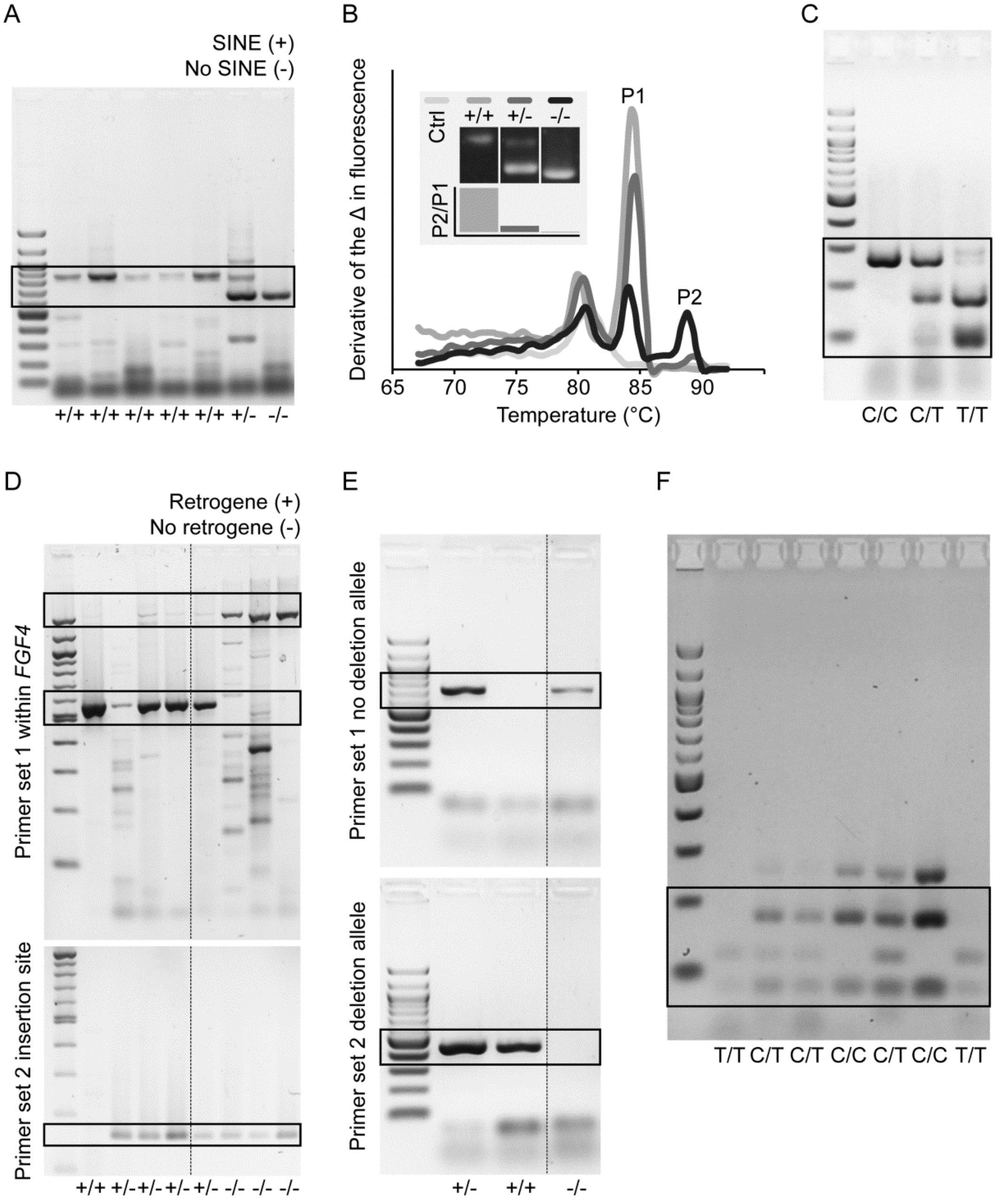

Figure 2-6 Representative genotyping results for mixed-breed dogs. (A) *IGF1* short-interspersed element (SINE). (B) Melt-curves for *IGF1* SINE (inset comparison of PCR and qPCR results). (C) *STC2* RFLP. (D) *FGF4* retrogene insertion. (E) *SMAD2* deletion. (F) *BANP* RFLP.

Inspection of the correlations between genotyped variants and the sum of scapula, humerus, and radius reveals that the variants for *IGF1*, *FGF4* retrogene, and *SMAD2* are significantly correlated with height; the correlations for *STC2* and *BANP* are less clear (Figure 2-7).

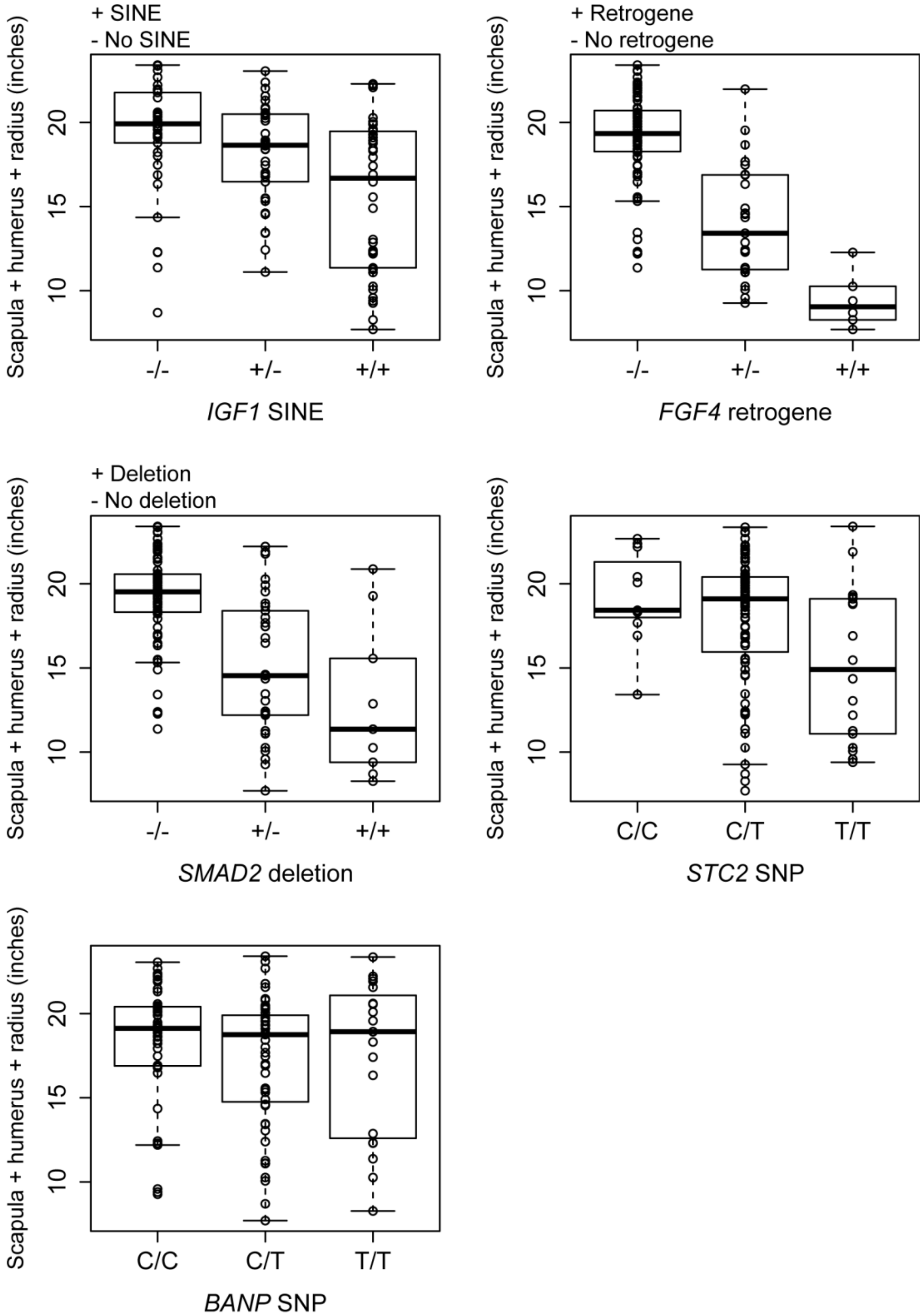

Figure 2-7 Correlation of genetic variants and measure of height in mixed-breed dogs.

With mixed-breed dogs interrelatedness is expected to play a less important role than with the purebred panel in Boyko et al. 2010, but to capture broad trends of population structure and interelatedness within the panel, six microsatellites were analyzed with *SPAGeDi* to generate a kinship matrix (Figure 2-8) (Fondon and Garner, 2004; Hardy and Vekemans, 2002).

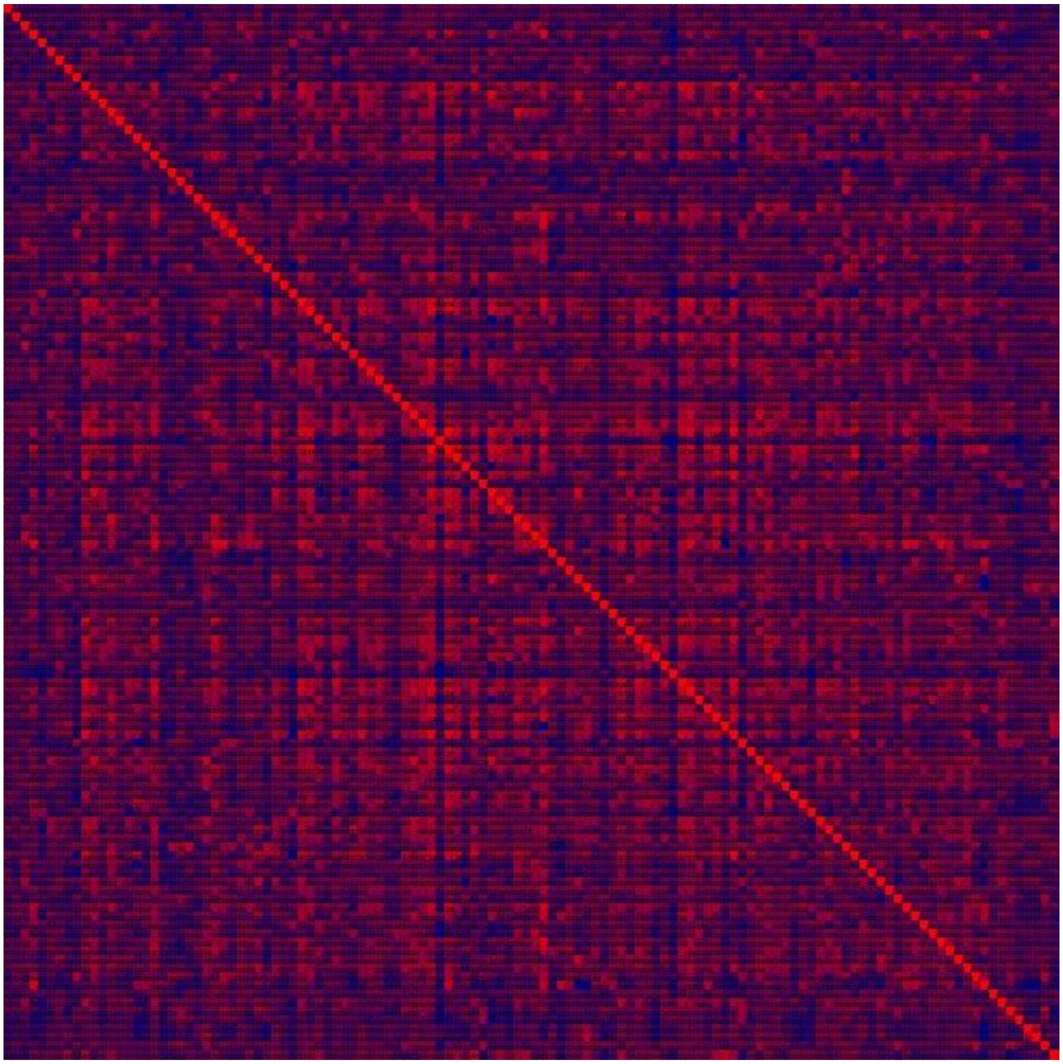

Figure 2-8 Kinship matrix of mixed-breed dogs where red is more related and blue is less related.

The software package Tassel was used to estimate the effect sizes of height variants, with (MLM) and without (GLM) applying the aforementioned kinship matrix (Bradbury et al., 2007). Results for both models are summarized in Table 2-2. Although the kinship matrix does impact effect size estimates, the effect is not severe enough to preclude performing multi-locus analysis without correcting for relatedness similar to as was done earlier in this study.

Table 2-2 Correlation between height variants and height in mixed-breed dogs

	GLM	MLM	GLM	MLM	GLM	MLM
Gene	*F*-value	*F*-value	p-value	p-value	R^2	R^2
IGF1	11.6	11.3	2.5e-5	3.2e-5	0.165	0.161
FGF4 retrogene	57.7	60.5	6.5e-18	1.7e-18	0.510	0.521
SMAD2	27.9	28.0	1.2e-10	1.2e-10	0.323	0.323
STC2	4.6	4.9	0.012	0.009	0.072	0.076
BANP	0.7	0.6	0.475	0.557	0.013	0.010

Consistent with Figure 2-7, *IGF1*, *FGF4* retrogene, and *SMAD2* are significantly associated with height variation in the mixed-breed dogs. At $P < 0.05$ *STC2* is also significant; with a Bonferroni correction for multiple testing *STC2* is only significant for the MLM that includes the kinship matrix ($P < 0.01$). The *BANP* SNP is not significant for either GLM or MLM models. It is likely that linkage disequilibrium has been broken up between the causal mutations and SNPs genotyped near *STC2* and *BANP*, thus leading to weak to no signal in the mixed-breed dogs. In contrast, where likely causative mutations have been genotyped for *IGF1*, *FGF4* retrogene, and *SMAD2*, correlations are much more significant.

Due to missing genotypes for some of the dog samples, 112 dogs are used for the multi-locus model because they have complete genotype data. In the multi-locus model with *IGF1*, *FGF4* retrogene, *SMAD2*, and *STC2,* 67% of variation in height is captured in the panel of mixed-breed dogs. Removing *IGF1* from this model drops the percentage explained to 62% (5% drop). Dropping *FGF4* retrogene from the full model lowers the percentage to 41% (26% drop). Removing *SMAD2* results in a 3% drop and *STC2* in a 1% decrease in the percentage explained.

The picture that emerges from this analysis is that *FGF4* retrogene is very much a large effect size variant, drastically altering the height of dogs. *IGF1* and *SMAD2* on the other hand correlate with a reduction in height, but have more moderate effect sizes. Although this mixed-breed panel does break up correlated allelic effects to some extent, correlated allelic effects still likely overestimate the percent of variation (67%) actually caused by these four variants. While these four loci can explain 67% of variation, they likely cause about 35% of variation (*FGF4* retrogene 26% + *IGF1* 5% + *SMAD2* 3% + *STC2* 1%). The effect size distribution that is most consistent with these large and moderate effect sizes is the exponential distribution (Figure 2-1), albeit likely shifted towards larger effect sizes relative to humans.

Conclusion

Chapter 1 examined how many of the same QTLs influence variation in height in humans and dogs. This work found that at least 25 height QTLs are shared between the species. Chapter 2 explored the distribution of effect sizes for variants that contribute to height differences in humans and dogs. Methodological considerations have led to interpretations of genetic architecture that are misleading and unproductive in both humans and dogs.

Despite reports of an infinitesimal model in humans and a simple model in dogs, this work suggests an exponential distribution of QTL effect sizes best describes the genetic architecture of height in both humans and dogs. Sample composition and the marker-based approach has likely biased the types of variants and perceived effect sizes in humans, leading to the ill-conceived notion to reject the exponential model when modest to large effect size variants absolutely exist. Correlated allelic effects contribute to inflating QTL effect size estimates in dogs and result in a simplified view of genetic architecture that ignores the rich genetic complexity that exists below large and modest effect size variants. This does not mean that the exponential distributions for humans and dogs are identical, rather, depending on the population, humans are likely shifted towards smaller effect sizes.

In conclusion, it is probably more productive to view gene effect size distributions as a continuum, thus keeping the focus on understanding the genetic architecture of variation rather than seeking to categorize it.

Chapter 3

The Influence of PRDM9 on Genetic Architecture

Introduction

Chapters 1 and 2 studied aspects of the genetic architecture of variation in humans and dogs after variants have already sufficiently increased in frequency to be detected by genome-wide scans. Chapter 3 takes a different angle on the genetic architecture of variation and explores the basis of mutation and recombination patterns that can influence the generation and fate of QTLs.

During prophase I of meiosis, crossing over of homologous chromosomes leads to increased genetic diversity and is essential for chromosomal segregation during anaphase (Cohen and Pollard, 2001; Hartl and Clark, 1997). Before crossovers between homologous chromosomes can form, double-strand breaks are introduced to chromosomes by a homolog of the yeast protein Spo11 (Keeney, 2001; Keeney et al., 1997; Szostak et al., 1983). Double-strand breaks are non-randomly distributed along chromosomes where some regions are considered hotspots for meiotic recombination (Gerton et al., 2000; Petes, 2001). While not every double-strand break results in a crossover, every double-strand break needs to be repaired (Haber, 2000; Szostak et al., 1983). Since the repair of double-strand breaks is inherently biased, error-prone, and can lead to genomic instability, understanding the process that influences the localization of double-strand break hotspots is of broad interest (Ferguson and Alt, 2001; Khanna and Jackson, 2001; O'Driscoll and Jeggo, 2006).

Work in humans and mice has identified PRDM9 as a protein central to directing the localization of double-strand breaks, and specifically directing them away from the promoter regions of genes (Baudat et al., 2010; Brick et al., 2012; Myers et al., 2010; Parvanov et al., 2010; Ségurel et al., 2011). Mechanistically, PRDM9 binds triplet DNA targets with a zinc finger domain and directs the initiation of double-strand breaks by trimethylating histone H3 at lysine 4 (Buard et al., 2009; Smagulova et al., 2011). An important feature of PRDM9 is that the zinc finger binding site is rapidly evolving, thus double-strand break localization should vary across time and among different alleles (Groeneveld et al., 2012; Oliver et al., 2009; Ponting, 2011; Thomas et al., 2009). Interestingly, before *PRDM9* was connected with meiotic recombination, it was labeled as a speciation gene for its role in causing infertility in mice with different alleles (Flachs et al., 2012; Mihola et al., 2009). The evolutionary impact of incompatible *PRDM9* alleles will be discussed in more detail later.

Recently, a functional copy of PRDM9 was shown to be missing in the canid lineage, a lineage that includes dogs, wolves, foxes, jackals, and coyotes. Relative to humans and mice, meiotic recombination hotspots in dogs are more stable and less pronounced (Axelsson et al., 2012). They are also often localized near repetitive GC-rich regions of the genome as well as the promoters of genes (Auton et al., 2013; Axelsson et al., 2012). Since the loss of PRDM9 occurred in the common ancestor of the canid lineage, it is challenging to know for certain if observed patterns of recombination in dogs are exclusively due to the loss of PRDM9, or are the consequence of subsequent evolutionary events. Teasing apart cause and consequence as it relates to the specific DNA motifs being targeted for double-strand breaks is particularly challenging because up to ~49 million of years have passed since PRDM9 was lost in the common ancestor of the canid lineage (Auton et al., 2013).

Despite this limitation, it has been suggested that GC-biased gene conversion is a plausible mechanism to explain the enrichment of GC-rich sequences at recombination hotspots in dogs (Auton et al., 2013; Axelsson et al., 2012). GC-biased gene conversion is a phenomenon where G or C nucleotide bases are preferentially chosen as the template for single base-pair mismatches of heteroduplex DNA (Duret and Galtier, 2009; Galtier et al., 2001). However, the role of GC-biased gene conversion in this context must be viewed as speculative until it can be clarified whether recombination at repetitive GC-rich motifs is a consequence of GC-biased gene conversion at sites with recurrent recombination, or is directly due to where double-strand breaks tend to occur without PRDM9. Clarifying where double-strand breaks occurred when PRDM9 was lost in the common ancestor of the canids is also relevant because of how this story intersects with another seemingly unrelated observation: relative to other mammals, the canid lineage harbors an enrichment of tandem repeat mutations, particularly GC-rich repeats (Laidlaw et al., 2007).

Tandem repeats mutate 10 to 100,000 times more than other parts of the genome and can have functional consequences, and in some cases lead to rapid morphological evolution (Fondon and Garner, 2004; Gemayel et al., 2010; Usdin, 2008; Verstrepen et al., 2005). Although tandem repeat mutations are typically thought of as the consequence of slippage during replication, the contribution of recombination to tandem repeat mutations could be substantial (Gemayel et al., 2010). Interestingly, the destabilization of tandem repeats occurred in the canid lineage about the same time functionality of the recombination-directing gene *PRDM9* was lost (Axelsson et al., 2012). This coincidence prompts the question: could the loss of PRDM9 have caused the destabilization of tandem repeats in the canid lineage? In order to conclude that the death of PRDM9 gave rise to a repeat mutator in the canid lineage, two conditions must be met: 1) Meiotic double-strand breaks must lead to the destabilization of repetitive sequences and 2) Repetitive sequences must be a target for double-strand breaks in the absence of PRDM9. Data from a recent study in mice, where *PRDM9* is knocked out and double-strand breaks are directly interrogated, allows for a direct test that clarifies where double-strand breaks occur without PRDM9 and can reveal if the loss of PRDM9 gave rise to a repeat mutator (Brick et al., 2012).

Establishing the Loss of PRDM9 as a DNA Repeat Mutator

Recombination Destabilizes Tandem Repeats

The first condition that must be met to establish the loss of PRDM9 as a repeat mutator is that meiotic double-strand breaks must lead to the destabilization of repetitive sequences. While it was established years ago that meiotic recombination can lead to expansions and contractions of tandem repeats of ~375 bp (Pâques et al., 1998), what about simple sequence repeats?

It has been demonstrated that zinc-finger directed double-strand breaks destabilize triplet repeats in human cells (Mittelman et al., 2009). Mismatch repair of heteroduplex DNA, a consequence of meiotic double-strand breaks, can destabilize tandem repeats as well (Pearson et al., 2005). Based on these studies that identify double-strand breaks and mismatch repair as factors that destabilize tandem repeats, if repeats are targets of recurrent meiotic recombination they will become destabilized.

A recent study found that where heteroduplex DNA mismatches are insertions or deletions (indels), there is a bias towards using the insertion allele as a template for the other strand (Leushkin and Bazykin, 2013). This insertion bias is strongest for shorter insertions and tails off as insertion length increases. Unlike GC-biased gene conversion which is relatively weak (50.62% bias), or in some cases non-existent (in the fly), the insertion bias is particularly strong with an up to 5 fold excess in regions of high recombination (Duret and Galtier, 2009; Leushkin and Bazykin, 2013; Mancera et al., 2008; Robinson et al., 2013). This insertion-biased gene conversion could expand repeats that are destabilized through recurrent meiotic recombination.

A more stable recombination landscape in dogs means that the same genomic locations are subject to recurrent meiotic recombination (Axelsson et al., 2012), making the destabilization of tandem repeat DNA inevitable as long as tandem repeats have been a target for meiotic double-strand breaks since the common ancestor of the canid lineage lost PRDM9.

Hypothesis-based Approach

The first requirement to establish the loss of PRDM9 as a repeat mutator in the canid lineage is met: meiotic double-strand breaks lead to the destabilization of tandem repeats. Since up to ~49 million years of evolution has occurred since PRDM9 was lost in the common ancestor of the canid lineage, enriched motifs at meiotic recombination hotspots in dogs do not necessarily reflect motifs targeted when PRDM9 was first lost. However, if the most destabilized repeats in dogs are targets for double-strand breaks in the PRDM9 knockout mouse, this would suggest that the loss of PRDM9 led to the destabilization of repeats in the canid lineage.

The study of Brick et al. 2012 uses chromatin immunoprecipitation to pull down DNA that is bound to DMC1 after having been cut by SPO11 during meiotic recombination. These DNA fragments represent the locations of double-strand breaks and were subsequently sequenced and mapped to the mouse genome. In this way the locations of double-strand breaks were directly interrogated for mice with different *PRDM9* alleles and for a *PRMD9* knockout as well. By examining where double-strand breaks occur in a *PRDM9* knockout mouse, the millions of years of evolution in the canid lineage can be sidestepped and it becomes clear if the loss of PRDM9 led to the destabilization of tandem repeats in canids.

One of the observations that came out of the analysis of Laidlaw et al. 2007 is that GC-rich tandem repeats are among the most destabilized repeats in the dog genome. For example, the *BMP6* gene contains a massive expansion of CGG_n repeats in the canid lineage that is absent in other mammals (Laidlaw et al., 2007). Interestingly, the CGG_n motif also comes through as one of the motifs that is most enriched at recombination hotspots in dogs, suggesting that in the absence of PRDM9, CGG_n is a common target for meiotic double-strand breaks (Auton et al., 2013; Axelsson et al., 2012). Based on the work of Laidlaw et al. 2007, four particularly destabilized DNA motifs in the dog genome are: CGG_n, $CGGG_n$, $CGGGG_n$, and C_n. It can be concluded that the loss of PRDM9 caused a destabilization of tandem repeats in the canid lineage if these motifs match the genomic targets of double-strand breaks in the *PRDM9* knockout mouse.

One way to test these four GC-rich motifs is to examine the enrichment of word counts at double-strand break hotspots for different length k-mers in the *PRDM9* knockout mouse. Motifs are counted for the top quarter of mapped, non-overlapping, autosomal double-strand break hotspots as defined by Brick et al. 2012 using the EMBOSS wordcount tool (Rice et al., 2000). Null counts are the average of motif word counts from the same number of flanking base pairs as in the adjacent hotspot. This analysis reveals that the most destabilized repeats in the dog genome are overwhelmingly among the most enriched motifs targeted in the *PRDM9* knockout mouse (Table 3-1).

Table 3-1 Enriched motifs in the dog genome are targeted in *PRDM9* knockout mice

K-mer	Metric	CGG_n	$CGGG_n$	$CGGGG_n$	C_n
5 (n = 1,024)	Top rank	3	12	12	224
	Mean rank	9.8	19.5	42	226
	Percentile	99.0	98.1	95.9	78.0
6 (n = 4,096)	Top rank	2	50	74	1,150
	Mean rank	16.8	67.1	84.5	1,153
	Percentile	99.6	98.4	97.9	71.9
7 (n = 16,384)	Top rank	17	45	56	5,464
	Mean rank	21.5	207	306	5,485
	Percentile	99.9	98.7	98.1	66.5
8 (n = 65,536)	Top rank	97	214	264	23,460
	Mean rank	138	556	1,315	23,614
	Percentile	99.8	99.2	98.0	64.0
9 (n = 262,144)	Top rank	779	1,504	1,246	89,857
	Mean rank	1,117	2,033	4,166	90,667
	Percentile	99.6	99.2	98.4	65.4
10 (n = 1,048,576)	Top rank	1,035	2,320	2,621	235,408
	Mean rank	1,873	2,937	4,191	237,273
	Percentile	99.8	99.7	99.6	77.4
	Average percentile	99.6	98.9	98.0	70.5

Across 5-mers through 10-mers, the CGG_n motif is on average more enriched than 99.6% of all other possible motifs. With the 6-mer, CGG_n is the second most enriched motif out of 4,096 possible motifs. The top motif for 6-mers is $GCCGCG_n$, another GC-rich motif quite similar to CGG_n. The $CGGG_n$ and $CGGGG_n$ motifs are more enriched than 98.9% and 98% of all other motifs for 5-mers through 10-mers, respectively. The C_n motif is more modest in its enrichment at 70.5%.

Another approach to examine fold enrichment of the CGG_n, $CGGG_n$, $CGGGG_n$, and C_n motifs is with a 200 bp sliding window with 1 bp steps for hotspots in wild-type B6 mice and *PRDM9* knockout mice. To match motif length with k-mer lengths, the CGG_n motif was assessed as a 9-mer, $CGGG_n$ as an 8-mer, $CGGGG_n$ as a 10-mer, and C_n as a

10-mer. DNA complements and all possible reading frames are included for each motif considered. Double-strand break hotspot regions are 5 kb and centered on 2 kb autosomal hotspots as reported from the study of Brick et al. 2012. Five kb of null sequence is derived from 2.5 kb flanking each 5 kb hotspot; hotspots with overlapping null flanks are excluded, leaving 20,360 *PRDM9* knockout hotspots and 15,503 wild-type B6 hotspots. Enrichment is calculated as the mean of the motif frequency by position in 5 kb hotspots relative to the mean motif frequency of 5 kb of flanking null sequence. Sliding window analysis was performed in R using the BSGenome, Biostrings, plyr, stringr, and ggplot2 packages (Pages; Pages et al.; R Core Team, 2013; Wickham, 2009, 2011, 2012).

Sliding window analysis supports the word count analysis where all four motifs enriched in the dog genome are also targets for double-strand breaks in the *PRDM9* knockout mouse (Figure 3-1). These results indicate that the loss of PRDM9 directly causes double-strand breaks to occur at repetitive GC-rich regions of the genome. Also consistent with the word count analysis, the CGG_n motif has the highest degree of enrichment followed by $CGGG_n$, $CGGGG_n$, and C_n (Figure 3-1).

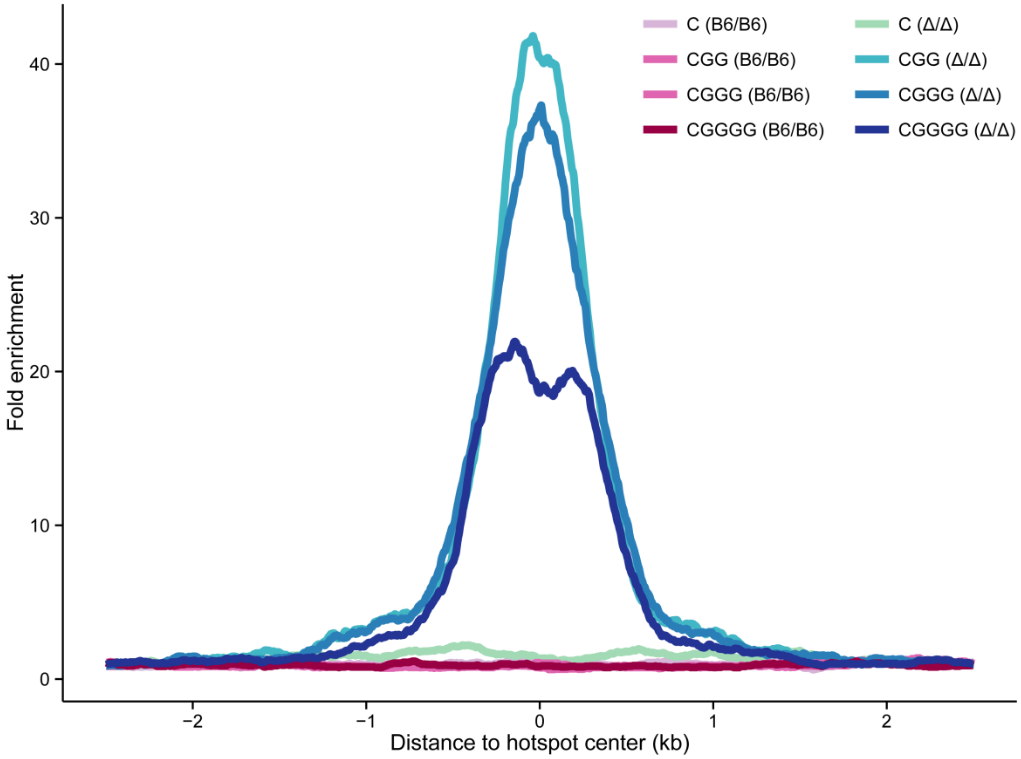

Figure 3-1 Fold enrichment at hotspots in wild-type (B6/B6) and *PRDM9* knockout (Δ/Δ) mice for repetitive DNA motifs enriched in the dog genome.

If repetitive GC content is a target for double-strand breaks in the absence of PRDM9, and there is bias for insertion alleles, this could certainly destabilize and in some cases expand existing repetitive GC content. Figure 3-1 only assesses perfect repeats; it does not capture repeats with mismatches or impurities. Are pure or impure GC-rich repeats better targets for double-strand breaks in the absence of PRDM9? To address this question, the perfect 9-mer CGG_n motif is compared to 9-mer CGG_n motifs with one mismatch in the 3 bp head or tail of the motif. Interestingly, the mismatch CGG_n motifs are actually more enriched (~2 fold) than the perfect CGG_n 9-mer motif (Figure 3-2). Note that in the wild-type B6 mouse there is a slight peak at the center of hotspots. Is it possible that even with a functional copy of *PRDM9*, some CGG_n motifs are targets of double-strand breaks? This idea will be revisited momentarily.

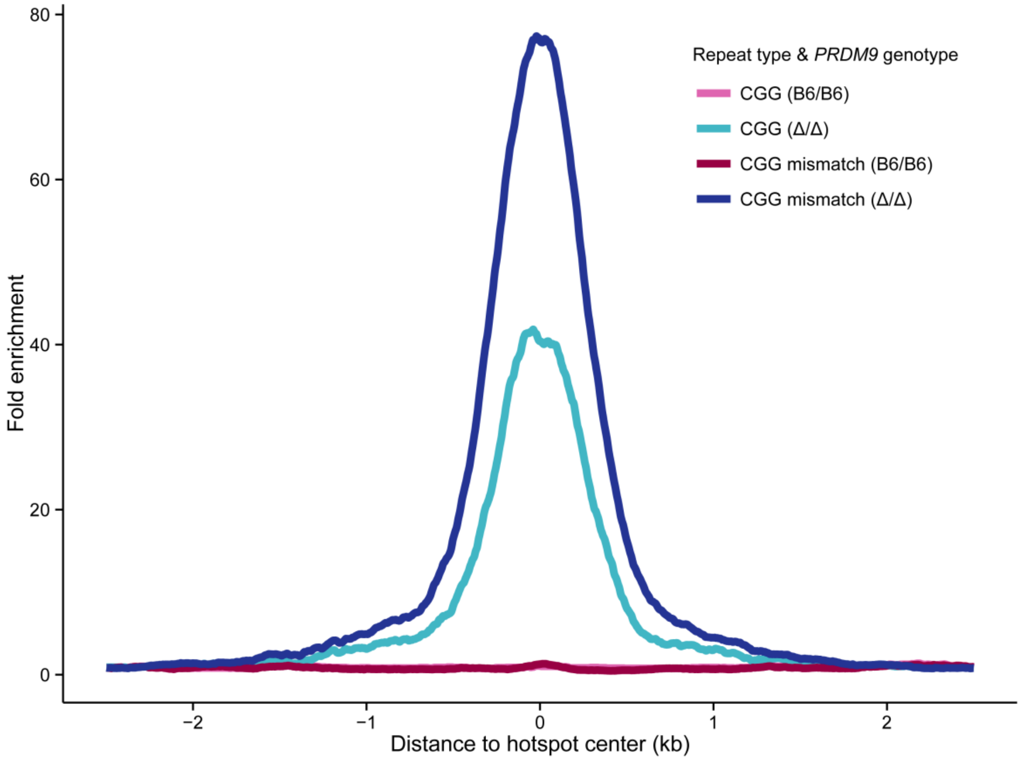

Figure 3-2 Fold enrichment at hotspots in wild-type (B6/B6) and *PRDM9* knockout (Δ/Δ) mice for perfect 9-mer CGG_n motifs and those with one mismatch in the 3 bp head or tail.

These results suggest that without PRDM9, regions of the genome that are only partially GC-rich and repetitive are also strong targets for double-strand breaks. Thus in the absence of PRDM9, existing GC-rich repeats will be destabilized, and new GC-rich repeats will be created. As hotspot strength increases, so does the proportion of hotspots that contain the CGG_n motif (Table 3-2). While the CGG_n motif is neither necessary nor sufficient for the formation of double-strand breaks in the absence of PRDM9, this motif is present in some of the hottest double-strand break hotspots, suggesting that the enrichment of this motif in dogs is not merely a consequence of recurrent double-strand breaks and GC-biased gene conversion. These sequences have been targeted ever since PRDM9 was lost.

Table 3-2 Percentage of hotspots with imperfect 9-mer CGG_n motif

Mapped autosomal hotspots	*PRDM9* knockout (n = 24,474)	Wild-type (B6) (n = 17,159)
All	43.5	5.1
Top half	56.2	4.9
Top quarter	66.3	4.7
Top eighth	73.7	4.5
Top 1000	77.4	5.1
Top 100	83	8

Collectively, these results indicate that the most destabilized repeat motifs in the dog genome are enriched targets for double-strand breaks in the PRDM9 knockout mouse. I conclude that the loss of PRDM9 is responsible for destabilizing tandem repeats in the common ancestor of the canid lineage.

Hypothesis-free Approach

With the hypothesis-based approach, the loss of PRDM9 was linked to some of the most destabilized tandem repeats in the dog genome. By taking a hypothesis-free approach, greater perspective can be gained on all of the targets of double-strand breaks when PRDM9 is lost. To do this, all 6-mer motifs are counted for the top 10,000 mapped, non-overlapping, autosomal double-strand break hotspots in wild-type B6 and *PRDM9* knockout mice using the EMBOSS wordcount tool (Rice et al., 2000). As before, null counts are the average of motif word counts from the same number of flanking base pairs as in the adjacent hotspot. This analysis reveals that high GC-content dominates enriched motifs at double-strand breaks in *PRDM9* knockout mice (Figure 3-3).

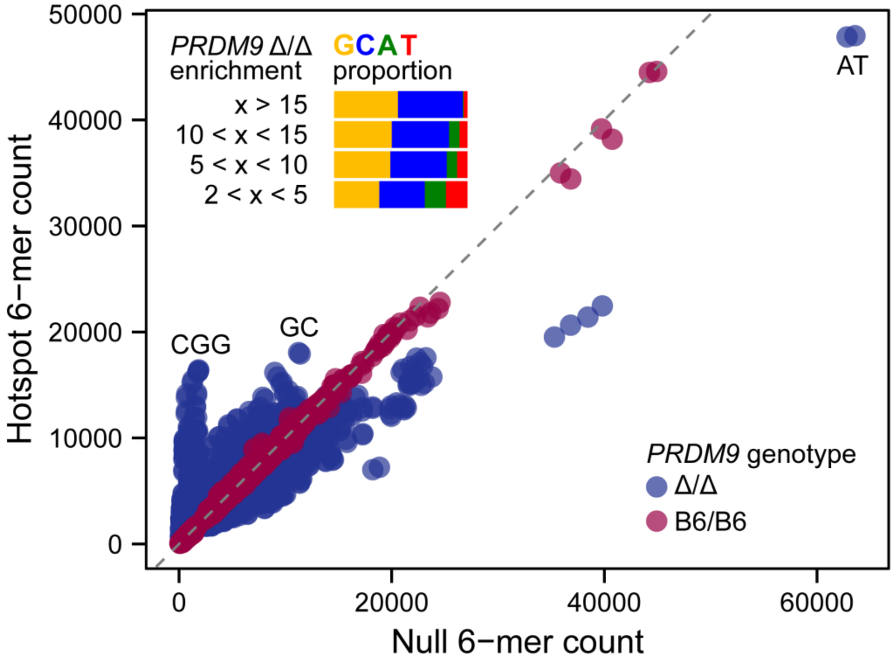

Figure 3-3 Wordcounts (6-mer) for wild-type (B6) and *PRDM9* knockout mice. GC-content is correlated with motif enrichment in *PRDM9* knockout mice (inset).

Consistent with the hypothesis-based approach that noted a slight enrichment of C_n, the hypothesis-free approach finds that mononucleotides of C and G are somewhat enriched (Figure 3-3). The plume of highly enriched motifs near zero on the x-axis in Figure 3-3 consists of GC-rich motifs, where a cluster of CGG_n motifs are the most enriched. Another observation from the comparison of wild-type and knockout *PRDM9* mice is that while many motifs are enriched in the knockout, only a few motifs show enrichment in the wild-type. This observation is consistent with a flatter distribution of hotspots in dogs where many sequences can be targets of double-strand breaks, but they do tend to be GC-rich (Auton et al., 2013; Axelsson et al., 2012). Reassuringly, some of the most enriched sequences in the B6 mouse match the predicted PRDM9 binding site in the study of Brick et al. 2012. An examination of word counts for 7-mers through 10-mers reinforces the finding that repetitive GC-rich content is a common target for double-strand breaks in the absence of PRDM9, although it is certainly not the only target (Table 3-3). In Table 3-3 enriched motifs, hot and coldspot counts, and fold enrichment are derived from the top quarter (n = 6065) of hotspots in the *PRDM9* knockout mouse. The percentage of hotspots with given motifs are calculated based on all reported hotspots in *PRDM9* knockout and wild-type mice. Since DNA complements and reading frame permutations were similarly enriched, reported motifs are representative.

Table 3-3 Top enriched hotspots for 7, 8, 9, and 10-mer motifs

Motif length	Motif	Repeat	Hotspot count	Coldspot count	Fold enrichment	Knockout hotspots with motif (%)	Wild-type hotspots with motif (%)
7	gcggcgg	CGG	5758	216.5	26.6	31.1	1.9
7	gcggggc	CGGGG	4095	258.5	15.8	34.4	3.2
7	cccgccc	–	5657	513.5	11.0	49.3	8.9
8	gcggcggc	CGG	3276	111	29.5	18.1	0.9
8	ggcggggc	CGGGG	2315	126.5	18.3	23.1	1.6
8	gggcgggg	–	2999	302	9.9	35.6	5.3
9	cggcggcgg	CGG	1820	59.5	30.6	12.7	0.5
9	gggcggggc	CGGGG	1481	79.5	18.6	16.0	1.0
9	ggggcgggg	–	1760	208.5	8.4	25.8	3.7
10	cggcggcggc	CGG	1302	40.5	32.1	8.0	0.3

The results of the hypothesis-free test are highly similar to studies in the dog where GC-rich simple repeats are extremely common targets for meiotic recombination (Auton et al., 2013; Axelsson et al., 2012). Thus high levels of repetitive GC-content at sites of recombination in dogs is a reflection of where double-strand breaks tend to occur in the absence of PRDM9, not the consequence of extensive GC-biased gene conversion. While GC-biased gene conversion may have played some role in increasing the GC-content in the dog genome, it is not required to fully explain why GC-rich repetitive DNA is a common target of meiotic recombination.

The Generation and Fate of QTLs

The loss of PRDM9 impacts the localization of meiotic double-strand breaks and has consequences for the generation and fate of QTLs. Studies in both mice and dogs find that in the absence of PRDM9, meiotic double-strand breaks are more prone to occur in the functional regions of genes (Auton et al., 2013; Brick et al., 2012). With recombination occurring more frequently in regions of the genome with a high gene density, the mutagenic aspect of recombination is predicted to have a larger role in generating QTLs in canids relative to other taxa with functional copies of *PRDM9*. In addition, those QTLs will have a greater likelihood of escaping their genetic contexts because they are in recombination hotspots. While the recombination landscape without PRDM9 tends to be flatter than with the protein, it is also more stable, leading to long-lasting recombination hotspots. A more stable recombination landscape promotes tandem repeat destabilization as well as the generation of recurrent QTLs.

Tandem repeats are highly mutagenic and could play a significant role in the seemingly never-ending supply of genetic variation in dogs, despite intense selective pressure and inbreeding. This is particularly relevant because of the role tandem repeats might be playing in rapid and continuous morphological evolution in dogs (Fondon and Garner, 2004). Perhaps these tandem repeats induced by the loss of PRDM9 are part of the explanation for why another member of the canid lineage, the silver fox, has been domesticated in only a few generations, with blue eyes, curly tails, and other dog-like attributes simultaneously appearing (Trut, 1999). The incredible evolvability and tendency to be domesticated in the canid lineage could be the consequence of how the loss of PRDM9 has altered the generation and fate of QTLs.

Destabilizing the repetitive GC-content of the genome can have more serious consequences than curly tails and blue eyes – it can lead to genomic instability. A recent review of repeat expansion diseases highlights the reality that GC-rich repeats are among the most common motifs that impact disorders associated with genomic instability like Fragile X disorders (Kumari et al., 2012). In addition, it has been proposed that the high GC-content at subtelomeres in the dog genome are the consequence of chromosomal fission at GC-rich regions of the genome (Webber and Ponting, 2005). A synteny map of humans and dogs reflects these extensive chromosomal fissions.

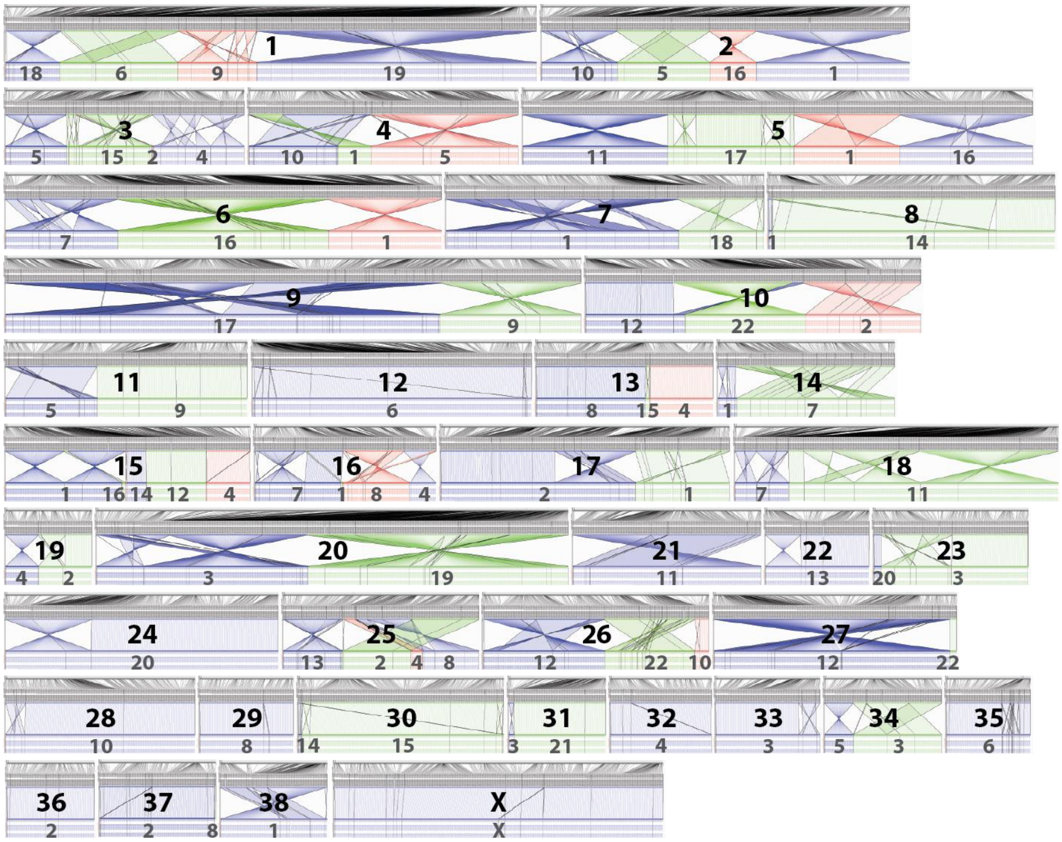

Figure 3-4 Synteny map of dog (larger numbers) and human (smaller numbers) chromosomes.

Is it possible that chromosomal fissions may be the eventual consequence of losing PRDM9 and the subsequent increase in GC-rich repeats? If the loss of PRDM9 has in fact led to genome instability, it is remarkable that the loss of a single protein has drastically altered genome architecture and the generation and fate of QTLs in the canid lineage. How could a protein as important as PRDM9 be lost in the first place? Shouldn't purifying selection keep PRDM9 around?

Model for the Death of PRDM9

It was previously mentioned how infertility can result from incompatibility of different *PRDM9* alleles in mice, hence the origin of its title as a speciation gene (Mihola et al., 2009). Since *PRDM9* knockout mice are infertile, how did the death of PRDM9 evolve in the canid lineage, an event that is predicted to be under considerable purifying selection?

Although mutators can offer a selective advantage in some contexts, because vital functions are unprotected from mutators, selection is predicted to act against them in the long run (Giraud et al., 2001). It has been proposed that mutators of simple sequence repeats can however be selectively advantageous because of their role as "tuning knobs" for genes and gene networks (Kashi and King, 2006a, 2006b; King, 2012). While this seems plausible, computer simulations suggest that repeat mutators can almost evolve neutrally, but are never advantageous (J.W. Fondon III). If different *PRDM9* alleles are compatible with each other and do not confer infertility, the loss of PRDM9 is unlikely to evolve because of its role as mutator (Figure 3-5A). If on the other hand, as is observed in mice, different *PRDM9* alleles are incompatible with each other and confer infertility, a heterozygous genotype with one null *PRDM9* allele could be compatible with all other *PRDM9* alleles, providing a selective advantage (Figure 3-5B). Once the null *PRDM9* allele increased sufficiently in frequency so that it started to occur in homozygosity, a compensating mutation would be necessary to sidestep the sterility that is observed in homozygous *PRDM9* knockout mice (Figure 3-5C).

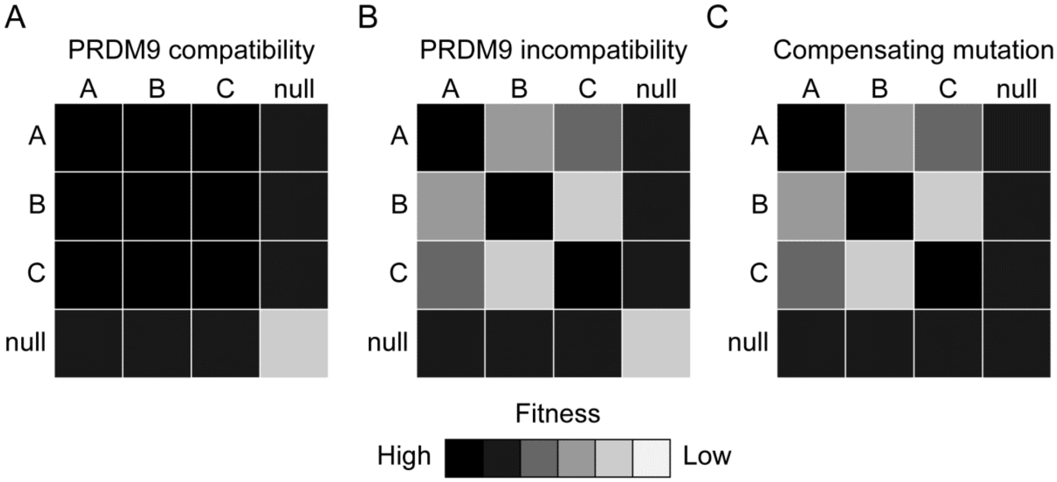

Figure 3-5 Model for how a null *PRDM9* allele could be selectively advantageous in heterozygosity and homozygosity.

Questioning the Human PRDM9 Binding Site

One of the observations from the sliding window analysis was that even in the wild-type B6 mouse, there appeared to be a slight enrichment of the CGG_n motif at the center of double-strand break hotspots (Figure 3-2). Why would the CGG_n motif be a target for double-strand breaks even with a functional copy of *PRDM9*?

A peculiar finding by the Brick et al. 2012 study is that in the pseudo-autosomal region of the X chromosome, PRDM9 does not direct where double-strand breaks occur. The pseudo-autosomal region of the X chromosome in humans and mice is characterized by extensive GC-content and is the only region of the genome where recombination always occurs, even with the Y chromosome, to guarantee proper segregation of sex chromosomes (Brick et al., 2012; Kent et al., 2002). Is it possible that if GC-content is enriched enough, the PRDM9-indepedent process, observed in the knockout mouse and in the canid lineage, supersedes PRDM9 to determine the location of double-strand breaks? This could explain the slight enrichment of CGG_n in the center of hotspots in mice with a functional copy of *PRDM9*: a subset of hotspots, like those in the pseudo-autosomal region of the X chromosome, is directed by a PRDM9-independent process that primarily targets GC-rich repetitive DNA.

If this hypothesis is true, perhaps the 13-mer predicted PRDM9 DNA binding site (CCNCCNTNNCCNC), derived from shared recombination hotspots in humans (Baudat et al., 2010), actually reflects the PRDM9-independent process. This is plausible since as was shown in mice, different *PRDM9* alleles alter the localization of double-strand breaks. Assuming that the populations from which shared recombination hotspots are derived have different *PRDM9* alleles, any shared hotspots would likely reflect the PRDM9-independent process. In addition, DNA binding site predictions based on bioinformatics, not biochemistry, should always be viewed with caution.

Since the human and mouse *PRDM9* alleles are different, there is no expectation that the human PRDM9 binding site should be a good predictor for the mouse PRDM9 binding site. However, if the human consensus PRDM9 binding motif is enriched in the *PRDM9* knockout mouse, this suggests that the human consensus binding site is actually tracking the PRDM9-independent process, not PRDM9. This is exactly what is observed where the human PRDM9 DNA binding site is predictive of hotspots in mice without a functional copy of *PRDM9* (Figure 3-6).

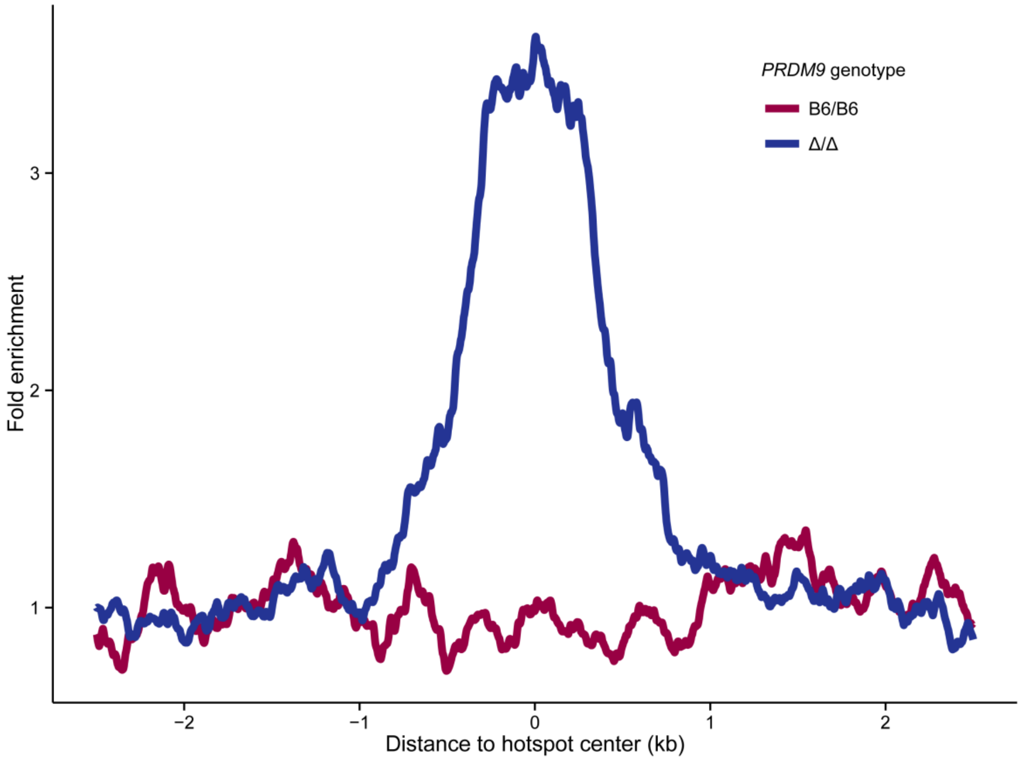

Figure 3-6 The predicted human PRDM9 DNA binding site is a motif enriched in *PRDM9* knockout mice.

It is interesting to speculate if perhaps humans, with our various alleles of *PRDM9* (Berg et al., 2011), are vulnerable to the death of PRDM9 in our own lineage. As proposed earlier, if enough alleles of *PRDM9* are incompatible, this could provide the ideal environment for a null *PRDM9* allele to increase in frequency (Figure 3-5). With a compensating mutation to offset any negative consequences of a null *PRDM9* allele in homozygosity, humans could join the evolutionary trajectory of canids that possess an incredible capacity for adaptation and diversification.

Conclusion

This study finds that in the absence of PRDM9, double-strand break hotspots often localize to GC-rich repetitive regions of the genome. Recurrent double-strand break localization and mismatch repair has led to a massive destabilization of tandem repetitive DNA in the canid lineage, some of which has already been shown to have functional consequences. The death of PRDM9 not only changes the recombination landscape in canids, but affects genome stability and the generation and fate of QTLs.

In addition, a model is proposed for how PRDM9 could have been lost in the canids lineage, where incompatible *PRDM9* alleles make the a null mutation selectively favored, followed by a subsequent compensating mutation to prevent sterility in homozygosity.

Finally, this study questions the predicted PRDM9 DNA binding site in humans and concludes that the existing consensus sequence likely predicts for a PRDM9-independent process. This PRDM9-independent process directs the localization of double-strand breaks in a subset of chromosomal regions, including the pseudo-autosomal region, and is the primary mechanism responsible for the localization of double-strand breaks in the canid lineage.

Appendix A

Code and Associated Files for Chapter 1

```
#################################################################################
# R code to:
# Create human height plots by country, population, and for individuals
#################################################################################

#Plot world height by population
#File derived from supplementary table given in Appendix I
#of (Gustafsson and Lindenfors, 2004)
ht<-read.table("world_height_pops.txt",sep="\t", header=TRUE)
#Columns are: Region Population Sex Height
ht$Height <- as.numeric(ht$Height)*0.393701
ht$Population<-with(ht,reorder(ht$Population,ht$Height))
library(ggplot2)
palette1<- c("#980043","#253494")
cbbPalette <- c("#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2",
                "#D55E00", "#CC79A7")
ggplot(ht, aes(Population,as.numeric(Height),colour=Region)) +
  geom_point(size=2) + theme_bw() + ylab("Average height (inches)") +
  coord_flip() + scale_color_manual(values=cbbPalette) +
  theme(axis.text.y = element_blank(), axis.ticks.y = element_blank())

#Plot world height by country
ht<-read.table("world_height.txt",sep="\t",header=TRUE)
ht$Country<-with(ht,reorder(ht$Country,ht$Height))
ht$Height<-ht$Height*39.3701
#library(ggplot2)
palette1<- c("#980043","#253494")
ggplot(ht, aes(Country,Height,colour=Sex)) + geom_point(size=3) + theme_bw() +
  scale_color_manual(values=palette1) +
  ylab("Average height (inches)") + coord_flip()

#Plot individual height from Galton's data
library(UsingR)
data(galton)
data(father.son)
ggplot(father.son, aes(father.son$sheight)) + geom_bar() + theme_bw() +
  xlab("Height (inches)") + ylab("Count")
```

```
world_height.txt file to plot height by country

Country Height  Sex     Order
Argentina       1.7348  Male    130
Argentina       1.6076  Female  43
Australia       1.784   Male    159
Australia       1.748   Male    134
Australia       1.645   Female  73
Australia       1.634   Female  62
Austria 1.792   Male    164
Austria 1.676   Female  98
Bahrain 1.651   Male    83
Bahrain 1.542   Female  12
Belgium 1.786   Male    161
Belgium 1.681   Female  102
Bolivia 1.6     Male    39
Bolivia 1.422   Female  1
Brazil  1.74    Male    133
Brazil  1.731   Male    127
Brazil  1.611   Female  45
Brazil  1.601   Female  40
Cameroon        1.706   Male    119
Cameroon        1.613   Female  48
Canada  1.76    Male    141
Canada  1.751   Male    136
Canada  1.633   Female  60
Canada  1.623   Female  54
Chile   1.712   Male    123
Chile   1.71    Male    122
Chile   1.696   Male    108
Chile   1.591   Female  34
Chile   1.572   Female  20
Chile   1.561   Female  16
China   1.702   Male    112
China   1.663   Male    91
China   1.586   Female  26
China   1.57    Female  18
Colombia        1.706   Male    118
Colombia        1.587   Female  29
Croatia 1.805   Male    167
Croatia 1.663   Female  90
Czech Republic 1.8031 Male    166
Czech Republic 1.6722 Female  95
Denmark 1.826   Male    173
Denmark 1.687   Female  106
Egypt   1.703   Male    115
Egypt   1.589   Female  31
El Salvador     1.656   Male    85
El Salvador     1.603   Female  41
England 1.771   Male    149
England 1.768   Male    145
England 1.766   Male    144
England 1.639   Female  70
England 1.637   Female  67
England 1.632   Female  59
Finland 1.79    Male    163
Finland 1.77    Male    148
Finland 1.65    Female  81
Finland 1.63    Female  56
France  1.77    Male    147
France  1.756   Male    139
France  1.646   Female  75
France  1.625   Female  55
```

```
Germany 1.81   Male    168
Germany 1.78   Male    156
Germany 1.68   Female  101
Germany 1.65   Female  80
Ghana   1.695  Male    107
Ghana   1.585  Female  25
Greece  1.783  Male    158
Greece  1.666  Female  92
Hong Kong      1.717  Male    125
Hong Kong      1.587  Female  28
India   1.663  Male    89
India   1.647  Male    77
India   1.612  Male    46
India   1.526  Female  10
India   1.521  Female  8
India   1.519  Female  7
Indonesia      1.58   Male    24
Indonesia      1.47   Female  2
Iran    1.734  Male    129
Iran    1.703  Male    114
Iran    1.598  Female  35
Iran    1.572  Female  19
Iraq Baghdad   1.654  Male    84
Iraq Baghdad   1.558  Female  15
Ireland 1.775  Male    151
Ireland 1.635  Female  64
Italy   1.772  Male    150
Italy   1.76   Male    140
Italy   1.678  Female  99
Italy   1.65   Female  79
Ivory Coast    1.701  Male    111
Ivory Coast    1.591  Female  33
Jamaica 1.718  Male    126
Jamaica 1.608  Female  44
Japan   1.707  Male    121
Japan   1.58   Female  23
Lithuania      1.813  Male    169
Lithuania      1.675  Female  97
Malawi  1.66   Male    87
Malawi  1.55   Female  13
Malaysia       1.647  Male    76
Malaysia       1.533  Female  11
Mali    1.713  Male    124
Mali    1.604  Female  42
Malta   1.752  Male    137
Malta   1.699  Male    110
Malta   1.638  Female  69
Malta   1.599  Female  36
Mexico  1.67   Male    94
Mexico  1.6    Female  38
Mongolia       1.684  Male    105
Mongolia       1.577  Female  21
Montenegro     1.832  Male    175
Montenegro     1.684  Female  104
Netherlands    1.838  Male    176
Netherlands    1.832  Male    174
Netherlands    1.707  Female  120
Netherlands    1.699  Female  109
Nigeria 1.638  Male    68
Nigeria 1.578  Female  22
Norway  1.824  Male    172
Norway  1.816  Male    170
Norway  1.682  Female  103
```

Norway	1.68	Female	100
Peru	1.64	Male	72
Peru	1.51	Female	4
Philippines	1.634	Male	61
Philippines	1.619	Male	49
Philippines	1.517	Female	6
Philippines	1.502	Female	3
Poland	1.785	Male	160
Poland	1.651	Female	82
Portugal	1.737	Male	131
Portugal	1.637	Female	66
Scotland	1.782	Male	157
Scotland	1.75	Male	135
Scotland	1.635	Female	63
Scotland	1.613	Female	47
Serbia	1.82	Male	171
Serbia	1.668	Female	93
Singapore	1.706	Male	117
Singapore	1.6	Female	37
Slovenia Ljubljana	1.803	Male	165
Slovenia Ljubljana	1.674	Female	96
Spain	1.78	Male	155
Spain	1.662	Female	88
Sri Lanka	1.636	Male	65
Sri Lanka	1.514	Female	5
Sweden	1.779	Male	153
Sweden	1.646	Female	74
Switzerland	1.754	Male	138
Switzerland	1.64	Female	71
Thailand	1.703	Male	113
Thailand	1.59	Female	32
Turkey Ankara	1.761	Male	142
Turkey Ankara	1.74	Male	132
Turkey Ankara	1.62	Female	51
Turkey Ankara	1.589	Female	30
United Arab Emirates	1.734	Male	128
United Arab Emirates	1.564	Female	17
United States	1.789	Male	162
United States	1.78	Male	154
United States	1.776	Male	152
United States	1.763	Male	143
United States	1.706	Male	116
United States	1.648	Female	78
United States	1.632	Female	58
United States	1.632	Female	57
United States	1.622	Female	53
United States	1.587	Female	27
Vietnam	1.657	Male	86
Vietnam	1.621	Male	52
Vietnam	1.552	Female	14
Vietnam	1.522	Female	9
Wales	1.77	Male	146
Wales	1.62	Female	50

```
################################################################################
# R code to:
# Plot dog height distribution. Data compiled from (Alderton, 2008)
################################################################################

doght<-read.table("dog_height.txt",sep="\t",header=TRUE)
doght$Breed<-reorder(doght$Breed, doght$Range, mean)
ggplot(aes(Breed,Range),data=doght) + geom_violin(fill="black",colour="black") +
  theme_bw() + coord_flip() + geom_hline(aes(yintercept=31.5),lty=2) +
  geom_hline(aes(yintercept=33.5),lty=2) +
  ylab("Height at the withers (inches)") + xlab("Breeds") +
  theme(axis.text.y=element_blank()) + theme(axis.ticks=element_blank()) +
  theme(panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_line(size=.1, color="grey")) +
  scale_y_continuous(breaks=seq(5,35, by=5))
```

```
################################################################################
# R code to:
# Identify shared height QTLs in humans and dogs from genome-wide associations
################################################################################

#Get dog GWAS data:
#Read in CanMapAssociation from Boyko et al. 2010 (http://tinyurl.com/mn6kt22)
cma <- read.table("CanMapAssociation",header=T)
#Select only relevant rows and columns
cma <- subset(cma,cma$HeightWithers>0 | cma$BodyLength>0 | cma$NeckLength>0)
cma <- cma[,c(1:7,13,34,76)]
cma <- subset(cma,!cma$chrom %in% "chrX") #exclude the X chromosome

#Number of non-zero SNPs to be considered
length(cma[,1])

#Get maximum p-value between HeightWithers, BodyLength, and NeckLength
GetHeight <- function(x) {
  bl <- as.numeric(x[8])
  hw <- as.numeric(x[9])
  nl <- as.numeric(x[10])
  vec <- c(bl,hw,nl)
  ht.max <- max(vec)
  ht.min <- min(vec)
  ht.mid <- which(vec!=ht.max & vec!=ht.min)
  ht.mid <- vec[ht.mid]
  if(ht.max>ht.mid+1){ # ht.mid+1 is the default, remove for making Venns
    id<-which(vec==ht.max)
  }
  else {id<-0}
  return(c(ht.max,id))
}
ht <- apply(cma,1,GetHeight)
ht <- data.frame(t(ht))
names(ht) <- c("ht","id")
summary(factor(ht$id))
cma.merge <- data.frame(cma,ht)

#Calculate q-values
library(qvalue)
dht.qval <- qvalue(10^-ht$ht,robust=TRUE)

#Plot q-values and p-values with q-value cutoffs 0.025, and 0.045
library(ggplot2)
qp.df <- data.frame(dht.qval$qvalues,dht.qval$pvalues)
names(qp.df) <-c ("qvalue","pvalue")
qp <- ggplot(qp.df, aes(qvalue, pvalue))
qp + geom_point() + xlim(0,.075) + ylim(0,.1) + theme_classic() +
  geom_vline(xintercept=.025,lty=2) +
  geom_vline(xintercept=.045,lty=2) +
  xlab("q-value") +
  ylab("p-value")

#Summarizing impact of combining traits
# 0.01=3.96 | 0.02=3.27 | 0.025=3.03 | 0.03=2.865 | 0.04=2.457 | 0.045=2.251 |
# 0.05=1.955 | Bonferroni 5e-5=4.30103
library(VennDiagram)
SummarySNPs <- function(x, cutoff) {
  bl <- subset(x, x$BodyLength > cutoff)
  hw <- subset(x, x$HeightWithers > cutoff)
  nl <- subset(x, x$NeckLength > cutoff)
  bl <- paste0(bl$chrom, bl$chromEnd)
  hw <- paste0(hw$chrom, hw$chromEnd)
  nl <- paste0(nl$chrom, nl$chromEnd)
  blhw <- subset(bl, bl %in% hw)
  blnl <- subset(bl, bl %in% nl)
  hwnl <- subset(hw, hw %in% nl)
  i.111 <- length(subset(blhw, blhw %in% blnl & blhw %in% hwnl))
  i.011 <- length(blhw)
  i.110 <- length(blnl)
  i.101 <- length(hwnl)
  i.bl <- length(subset(bl, bl %in% blhw | bl %in% blnl))
  i.010 <- length(bl)
  i.hw <- length(subset(hw, hw %in% blhw | hw %in% hwnl))
  i.001 <- length(hw)
  i.nl <- length(subset(nl, nl %in% blnl | nl %in% hwnl))
  i.100 <- length(nl)
  i.all <- c(i.001, i.010, i.011, i.100, i.101, i.110, i.111)
  plot.new()
  overrideTriple = 1
  draw.triple.venn(i.001, i.010, i.100, i.011, i.110, i.101, i.111,
                   c("Height withers", "Body length", "Neck length"),
                   sep.dist = 0, fill = c("red", "blue", "green"),
                   lty = "blank")
}
#All plots were exported as PDFs with dimensions 3X3 and edited in Illustrator
SummarySNPs(cma.merge, 3.03) # 0.025
SummarySNPs(cma.merge, 2.251) # 0.045
SummarySNPs(cma.merge, 4.30103) # Bonferroni

#Only needed for manhattan plotting purposes for chromosome ordering
cma2 <- read.table("CanMapAssociation_chr_fixed",header=T)
cma2 <- subset(cma2,cma2$HeightWithers>0 | cma2$BodyLength>0 | cma2$NeckLength>0)
cma2 <- cma2[,c(1:7,13,34,76)]
ht2 <- apply(cma2,1,GetHeight)
ht2 <- data.frame(t(ht2))
names(ht2)<-c("ht","id")
summary(factor(ht2$id))
cma.merge2 <- data.frame(cma2,ht2)
library(gap)
hwdata <- with(cma2,cbind(chrom,chromEnd,HeightWithers))
bldata <- with(cma2,cbind(chrom,chromEnd,BodyLength))
nldata <- with(cma2,cbind(chrom,chromEnd,NeckLength))
htdata <- with(cma.merge2,cbind(chrom,chromEnd,ht))
hwdata <- as.data.frame(cbind(hwdata,rep(NA,length(hwdata[,1]))))
bldata <- as.data.frame(cbind(bldata,rep(NA,length(bldata[,1]))))
nldata <- as.data.frame(cbind(nldata,rep(NA,length(nldata[,1]))))
htdata <- as.data.frame(cbind(htdata,rep(NA,length(htdata[,1]))))
colorscheme <- rep(c(rgb(0,0,0,.15),rgb(.1,.1,.1,.15)),38)
ops <- mht.control(logscale=FALSE,colors=colorscheme,usepos=TRUE,srt=0,cex=2)
mhtplot2(hwdata,ops,pch=20,xlab="",ylab="")
mhtplot2(bldata,ops,pch=20,xlab="",ylab="")
mhtplot2(nldata,ops,pch=20,xlab="",ylab="")
mhtplot2(htdata,ops,pch=20,xlab="",ylab="")
abline(h=3.03,lty=2)
abline(h=2.251,lty=2)
abline(h=4.30103,lty=2)

#Read in NHGRI GWAS catalog for height (http://www.genome.gov/gwastudies/)
library(gdata)
hgwas <- read.xls("MyGWASSearch_1_22_14.xls")

#Get column of reported genes and clean it up
repG <- hgwas[,14]
CleanGenes <- function(x) {
  output <- character(0)
  for(i in 1:length(x)) {
    line <- as.character(x[i])
    lines <- unlist(sapply(line,strsplit,","))
    lines <- unlist(sapply(lines,strsplit,"/"))
    names(lines) <- c()
    lines<-gsub("\n","",lines)
    lines<-gsub(" ","",lines)
    output <- c(output,lines)
  }
  return(unique(output))
  return(output)
}
humanHeightGenes <- as.factor(CleanGenes(repG))
allhumanheights <- as.factor(CleanGenes(repG)) #comment out unique line for this

#Select SNPs according to defined cutoffs
# 0.01=3.96 | 0.02=3.27 | 0.025=3.03 | 0.03=2.865 | 0.04=2.457 | 0.045=2.251 |
# 0.05=1.955
library(GenomicRanges)
dht.bon <- subset(cma.merge,cma.merge$ht>4.30103) # Bonferroni
dht.02 <- subset(cma.merge,cma.merge$ht>3.03) # 0.02
dht.04 <- subset(cma.merge,cma.merge$ht>2.251) # 0.04
dhtG.bon <- with(dht.bon, GRanges(chrom,
                                IRanges(chromEnd-200000, chromEnd+199999), "+",
                                id=paste0(chrom,":",chromEnd,"_",ht,"_",id)))
dhtG.02 <- with(dht.02, GRanges(chrom,
                                IRanges(chromEnd-200000, chromEnd+199999), "+",
                                id=paste0(chrom,":",chromEnd,"_",ht,"_",id)))
dhtG.04 <- with(dht.04, GRanges(chrom,
                                IRanges(chromEnd-200000, chromEnd+199999), "+",
                                id=paste0(chrom,":",chromEnd,"_",ht,"_",id)))

#Get dog genes from Ensembl
library(biomaRt)
mart <- useMart("ensembl")
datasets <- listDatasets(mart)
Cmart <- useDataset("cfamiliaris_gene_ensembl",mart)
dog.genes <- getBM(attributes=c("chromosome_name","start_position",
                                "end_position","ensembl_gene_id","hgnc_symbol",
                                "strand"),
                   filters="with_hgnc", values=TRUE, mart=Cmart)

#Liftover coordinates from canFam3 to canFam2
library(rtracklayer)
dog.genes.G3 <- with(dog.genes, GRanges(paste0("chr",chromosome_name),
                                      IRanges(start_position, end_position),
                                      strand, id=ensembl_gene_id))
chain <- import.chain("canFam3ToCanFam2.over.chain")
dog.genes.G2.list <- liftOver(dog.genes.G3,chain)
dog.genes.G2 <- unlist(dog.genes.G2.list)

#Find SNP overlaps with dog genes
dog.bon<-findOverlaps(dhtG.bon,dog.genes.G2)
dog.02 <- findOverlaps(dhtG.02,dog.genes.G2)
dog.04 <- findOverlaps(dhtG.04,dog.genes.G2)

#Report genes with overlapping dog size SNPs
ensembl.ids <- data.frame(as.factor(dog.genes.G2$values.gr..queryHits.ol....))
names(ensembl.ids) <- "ensembl_gene_id"
ensembl.df <- merge(ensembl.ids, dog.genes, by="ensembl_gene_id", sort=FALSE)
ensembl.df <- ensembl.df[,c(1,5)]

ReportGeneHits <- function(x) {
  genes <- data.frame(ensembl.df, as.table(t(x)))
  hit.genes <- subset(genes,genes[,3]>0)
  hit.genes <- hit.genes[order(hit.genes[,1]),]
  hit.genes <- hit.genes[,2]
  hit.genes <- unique(hit.genes)
  print(length(hit.genes))
  return(hit.genes)
}
hit.dog.bon <- ReportGeneHits(dog.bon)
write.table(hit.dog.bon, "dog.bon.GO.txt", quote=FALSE, row.names=FALSE,
            col.names=FALSE)
hit.dog.02 <- ReportGeneHits(dog.02)
write.table(hit.dog.02, "dog.02.GO.txt", quote=FALSE, row.names=FALSE,
            col.names=FALSE)
hit.dog.04 <- ReportGeneHits(dog.04)
write.table(hit.dog.04, "dog.04.GO.txt", quote=FALSE, row.names=FALSE,
            col.names=FALSE)

#Do QTL sharing analysis
Hmart <- useDataset("hsapiens_gene_ensembl",mart)
humanDogOrtholog <- getLDS(attributes=c("chromosome_name","start_position",
                                      "end_position","ensembl_gene_id",
                                      "hgnc_symbol"),
                         filters="with_homolog_cfam", values=TRUE,mart=Hmart,
                         attributesL=c("chromosome_name","start_position",
                                       "end_position"),martL=Cmart)
names(humanDogOrtholog) <- c("chr","start","end","eID","HGNC","Dchr","Dstart",
                           "Dend")
humanDogOrtholog <- subset(humanDogOrtholog,humanDogOrtholog$HGNC!="")

#Remove duplicate entries, retaining the top entry which is the best hit,
#and thus likely ortholog
hdH <- subset(humanDogOrtholog,!duplicated(humanDogOrtholog$HGNC))

#Liftover ortholog coordinates from canFam3 to canFam2
library(rtracklayer)
hdHG <- with(hdH,GRanges(paste0("chr",Dchr),IRanges(Dstart, Dend),"+",
                       HGNC,id=eID))
chain <- import.chain("canFam3ToCanFam2.over.chain")
hdHG2.list <- liftOver(hdHG,chain)
hdHG2 <- unlist(hdHG2.list)

#Read in manually annotated orthologs that are human height genes to supplement
#ensembl. These coordinates are already in CanFam2
man.hdH <- read.table("manual.orthologs.txt",header=TRUE,sep="\t",
                    stringsAsFactors=FALSE)
man.hdHG <- with(man.hdH,GRanges(paste0("chr",chr),IRanges(start, end),"+",HGNC,
                               id=eID))

#Combine ensembl and manually annotated orthologs
all.HG <- c(man.hdHG,hdHG2)
all.HG <- subset(all.HG,!duplicated(all.HG$HGNC))

#Find overlaps of dog height SNPs and orthologous human genes
hd.bon <- findOverlaps(dhtG.bon,all.HG)
hd.02 <- findOverlaps(dhtG.02,all.HG)
hd.04 <- findOverlaps(dhtG.04,all.HG)

#Report trait decomposition
hd.bon.ht <- findOverlaps(dhtG.bon,subset(all.HG,all.HG$HGNC %in%
humanHeightGenes))
hd.02.ht <- findOverlaps(dhtG.02,subset(all.HG,all.HG$HGNC %in% humanHeightGenes))
hd.04.ht <- findOverlaps(dhtG.04,subset(all.HG,all.HG$HGNC %in% humanHeightGenes))
```

```
report.bon <- data.frame(dht.bon,as.table(hd.bon.ht))
subset(report.bon,report.bon$as.table.hd.bon.ht.>0 & report.bon$id>0)
report.02 <- data.frame(dht.02,as.table(hd.02.ht))
subset(report.02,report.02$as.table.hd.02.ht.>0 & report.02$id>0)
report.04 <- data.frame(dht.04,as.table(hd.04.ht))
subset(report.04,report.04$as.table.hd.04.ht.>0 & report.04$id>0)

#Report shared hits
report.hits <- function(x) {
  genes <- data.frame(all.HG$HGNC,as.table(t(x)))
  hit.genes <- subset(genes,genes[,2]>0)
  hit.genes <- hit.genes[order(hit.genes[,1]),]
  hit.genes <- hit.genes[,1]
  hit.genes <- unique(hit.genes)
  ht.genes <- subset(hit.genes,hit.genes %in% humanHeightGenes)
  print(length(hit.genes))
  print(length(ht.genes))
  print((length(ht.genes)/length(hit.genes))*100)
  return(as.character(ht.genes))
}
hit.bon <- report.hits(hd.bon)
hit.02 <- report.hits(hd.02)
hit.04 <- report.hits(hd.04)

#Report regions
read.regions <- read.table("regions.txt",stringsAsFactors=FALSE)
regions <- apply(read.regions,1,strsplit,"_")
get.regions <- function(x, regions) {
  region <- character()
  if(length(grep(paste0("\\<",x,"\\>"),regions))>0) {
    for(i in 1:length(regions)) {
      if(length(grep(paste0("\\<",x,"\\>"),regions[[i]]))>0){
        region<-regions[[i]][[1]][[1]]
      }
    }
  }
  else {region<-x}
  return(region)
}
hit.hum <- toString(subset(humanHeightGenes, humanHeightGenes %in% all.HG$HGNC))
hit.hum <- strsplit(hit.hum,", ")
hit.hum <- unlist(hit.hum)
rhit.hum <- sapply(hit.hum,get.regions,regions)
rhit.hum <- unlist(unique(rhit.hum))
hum.reg.num <- length(rhit.hum)

#There are 241 orthologous regions
rhit.bon <- sapply(hit.bon,get.regions,regions)
rhit.bon <- unlist(unique(rhit.bon))
rhit.02 <- sapply(hit.02,get.regions,regions)
rhit.02 <- unlist(unique(rhit.02))
rhit.04 <- sapply(hit.04,get.regions,regions)
rhit.04 <- unlist(unique(rhit.04))
rhit.bon
rhit.02
rhit.04

#Report all gene hits
get.hits <- function(x) {
  genes <- data.frame(all.HG$HGNC,as.table(t(x)))
  hit.genes <- subset(genes,genes[,2]>0)
  hit.genes <- hit.genes[order(hit.genes[,1]),]
  hit.genes <- hit.genes[,1]
  hit.genes<-unique(hit.genes)
  return(hit.genes)
}
ghit.bon <-get.hits(hd.bon)
ghit.02 <-get.hits(hd.02)
ghit.04 <-get.hits(hd.04)

#All plots were exported as PDFs with dimensions 3X3 and edited in Illustrator
library(VennDiagram)
human <- length(subset(humanHeightGenes, humanHeightGenes %in% all.HG$HGNC))
dog.bon <- length(ghit.bon)
dog.02 <- length(ghit.02)
dog.04 <- length(ghit.04)
hd.share.bon <- length(hit.bon)
hd.share.02 <- length(hit.02)
hd.share.04 <- length(hit.04)
plot.new()
draw.pairwise.venn(human, dog.bon, hd.share.bon, c("Human", "Dog"),
                   fill = c("orange2", "maroon"), lty = "blank")
plot.new()
draw.pairwise.venn(human, dog.02, hd.share.02, c("Human", "Dog"),
                   fill = c("orange2", "maroon"), lty = "blank")
plot.new()
draw.pairwise.venn(human, dog.04, hd.share.04, c("Human", "Dog"),
                   fill = c("orange2", "maroon"), lty = "blank")

perms <- 100000
random.hits <- function(x) {
  r.hits <- vector(mode="numeric",length=perms)
  for(i in 1:perms) {
    rG <- sample(1:length(all.HG[,1]), hum.reg.num, replace=FALSE)
    rsub.G <- all.HG[rG,]
    r.genes <- subset(x,x %in% rsub.G$HGNC)
    r.hits[i] <- length(r.genes)
  }
  return(r.hits)
}
ran.hit.bon <- random.hits(ghit.bon)
ran.bon.df <- as.data.frame(ran.hit.bon)
plot.bon <- ggplot(data.frame(ran.bon.df), aes(ran.hit.bon))
plot.bon + geom_histogram(binwidth=1,fill="grey",colour="grey30") +
  theme_classic() + geom_vline(xintercept=length(rhit.bon),lwd=1) +
  xlab("Number of shared QTL") +
  ylab("Count")
length(subset(ran.hit.bon,ran.hit.bon>length(rhit.bon)))

ran.hit.02 <- random.hits(ghit.02)
ran.02.df <- as.data.frame(ran.hit.02)
plot.02 <- ggplot(data.frame(ran.02.df), aes(ran.hit.02))
plot.02 + geom_histogram(binwidth=1,fill="grey",colour="grey30") +
  theme_classic() + geom_vline(xintercept=length(rhit.02),lwd=1) +
  xlab("Number of shared QTL") +
  ylab("Count")
length(subset(ran.hit.02,ran.hit.02>length(rhit.02)))

ran.hit.04 <- random.hits(ghit.04)
ran.04.df <- as.data.frame(ran.hit.04)
plot.04 <- ggplot(data.frame(ran.04.df), aes(ran.hit.04))
plot.04 + geom_histogram(binwidth=1,fill="grey",colour="grey30") +
  theme_classic() + geom_vline(xintercept=length(rhit.04),lwd=1) +
  xlab("Number of shared QTL") +
  ylab("Count")
length(subset(ran.hit.04,ran.hit.04>length(rhit.04)))
```

manual.orthologs.txt file with manually annotated human and dog orthologs (CanFam 2 coordinates)

chromosome_name	start_position	end_position	ensembl_gene_id	hgnc_symbol	strand
17	22244523	22268843	AL137731	RBJ	1
9	14688492	14719256	BC001264	WDR68	1
15	44029102	44046127	BC098313	C12orf48	1
29	27942792	27960274	NM_000318.2	PXMP3	1
9	14811403	14819705	NM_001003788	LYK5	1
1	68234777	68234906	NM_001012507.2	C6orf173	1
25	46560794	46565632	NM_001252198.1	PTMA	1
10	11335085	11480772	NM_003483.4	HMGA2	1
12	3936729	3989633	NM_005514	HLA-B	1
2	62483245	62483519	NM_005949.3	MT1F	1
37	28798656	28800541	NM_006000	TUBA1	1
1	100690099	100695836	NM_012119	CCRK	1
20	36681254	36710360	NM_015224	C3orf63	1
11	53876460	53907919	NM_015397	WDR40A	1
37	28547486	28552016	NM_017521.2	FEV	1
9	44141698	44166437	NM_018404	CENTA2	1
12	4459857	4503164	NM_019105.6	HLAclassIII	1
26	31858383	31869521	NM_020070.2	IGLL1	1
7	20319481	20350372	NM_052965.2	C1orf19	1
29	10668871	10679388	NM_138969	RDHE2	1
12	6541267	6543870	NM_178508.3	C6orf1	1
22	4006786	4126740	NM_198989.2	DLEU7	1
30	29776230	29778899	NM_207322.2	C2CD4A	1
18	20988215	20988901	NR_003680.1	RPL13AP17	1
15	36713462	36746645	NR_038159.1	MRPL42	1
15	36661327	36698391	NM_003348.3	UBE2N	1
23	31241760	31246185	NM_001242375.1	ANAPC13	1

regions.txt file that defines human QTL regions

```
IGF1R_ADAMTS17
ACAN_POLG
NUP37_PMCH_C12orf48
ANAPC13_CEP63_PCCB
LCORL_NCAPG
BMP3_PRKG2_RASGEF1B
BOD1_STC2_FBXW11
IGF1_CCDC53_NUP37_C12orf48_PMCH_GNPTAB
GH1_CSH1_WDR68_LYK5_MAP3K3_MT1F
NPPC_PDE6D_COPS7B_DIS3L2_ALPP_PTMA
NOG_DGKE_TRIM25_COIL_RISK
TLE3_UACA_LARP6_LRRC49
ADAMTS10_OR2Z1_MYO1F_PRAM1
CS_STAT2
TMEM100_PCTP
PIGF_CRIPT_C2orf34_SOCS5
ATAD5_RNF135
FGFR4_NSD1
PTPRJ_SLC39A13
RUNX2_SUPT3H
BRCA2_PDS5B
FGFR3_SLBP
KCNJ16_KCNJ2
CTU2_GALNS
PTCH1_FANCC
ADAMTSL3_SH3GL3
LHX3_QSOX2
SIN3A_PTPN9
MFAP2_ATP13A2_SDHB
ATF7_ATP5G2
GDF5_UQCC_CEP250_EIF6_MMP24
PDXDC1_NTAN1
ZBTB38_ACPL2
HMGA1_C6orf1_NUDT3_C6orf106_LBH
PPARD_FANCE_ANKS1A_TCP11_ZNF76_DEF6_SCUBE3
KRT23_KRT20
TBX4_TBX2_C17orf82_BCAS3_NACA2
TNRC6B_ADSL
NPR3_C5orf23
TMEM126B_TMEM126A
SGSM3_MKL1
MTMR11_SV2A_SF3B4_Histoneclass2A
GLT25D2_C1orf19
EFEMP1_PNPT1_CCDC88A
HIST1H1D_Histonecluster_Histoneclass1_Butyrophilingenes
CDK6_PEX1_GATAD1_ERVWE1
PLAG1_MOS_CHCHD7_RDHE2_RPS20_LYN_TGS1_PENK
PXMP3_ZFHX4
CRLF3_ATAD5_CENTA2_RNF135
CABLES1_RBBP8_C18orf45
ADCY3_RBJ_POMC_DNMT3A_DTNB_DNAJC27
IHH_CRYBA2_FEV_SLC23A3_TUBA1_TNS1
GOLIM4_SERPINI1
PITX1_PCBD2_CATSPER3_TXNDC15_DDX46_CAMLG
NUP153_CAP2_KIF13A
LIN28B_HACE1_BVES_POPDC3
PPIL6_CD164_SMPD2_MNICAL1_ZBTB24
L3MBTL3_SAMD3
SPIN1_CCRK
PDE3A_SLCO1C1_SLCO1B3
LYZ_YEATS4_FRS2_CPSF6_CCT2_LRRC10
```

```
SOCS2_MRPL42_CRADD_UBE2N
ZDHHC7_CRISPLD2_USP10
BCR_GNAZ_RTDR1_IGLL1
TRIP11_FBLN5_ATXN3_CPSF2
CHCHD7_RDHE2
HTR1D_CLIC4_CATSPER4_LIN28
PKN2_RPL5
TGFB2_LYPLAL1
ZNF678_JMJD4
ANTXR1_ZNF638
FASTKD2_CYP20A1
DOCK3_RTF1
MICA_OR2J3_OR2I1P_HLA-B_HLAclassIII_HLAlocus
C6orf173_LOC387103
```

Complete results from gene ontology analysis of dog height associations:

FDR < 0.0093

			Fold
Term	Count	P-Value	Enrichment
Genome-wide association analysis identifies 20 loci that influence adult height	3	0.005	26.2
regulation of growth	6	0.012	4.3
regulation of small GTPase mediated signal transduction	5	0.019	4.8
cell division	5	0.032	4.1
small GTPase mediated signal transduction	5	0.035	4.0
phosphorylation	8	0.042	2.4
regulation of cell growth	4	0.044	5.0
protein amino acid phosphorylation	7	0.053	2.5
regulation of Ras protein signal transduction	4	0.054	4.6
Ras protein signal transduction	3	0.068	6.9
lipid storage	2	0.071	26.8
amine transport	3	0.083	6.1
organic cation transport	2	0.090	21.0
cell cycle	7	0.094	2.2
regulation of cellular component size	4	0.098	3.6
phosphorus metabolic process	8	0.098	2.0
phosphate metabolic process	8	0.098	2.0
hsa04110:Cell cycle	3	0.071	6.4

FDR < 0.02

			Fold
Term	Count	P-Value	Enrichment
Many sequence variants affecting diversity of adult human height	8	0.001	5.0
Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels	4	0.001	18.8
Genome-wide association analysis identifies 20 loci that influence adult height	3	0.021	12.8
gland development	6	0.005	5.5
regulation of growth	9	0.006	3.3
spleen development	3	0.010	19.6
mammary gland development	4	0.014	7.9
transmembrane transport	11	0.015	2.4
regulation of small GTPase mediated signal transduction	7	0.016	3.4
sodium ion transport	5	0.020	4.8
post-embryonic development	4	0.021	6.8
regulation of Ras protein signal transduction	6	0.026	3.5
fat-soluble vitamin metabolic process	3	0.028	11.3
transmembrane receptor protein tyrosine kinase signaling pathway	6	0.034	3.3
negative regulation of muscle cell apoptosis	2	0.039	49.6
skeletal system development	7	0.043	2.7
positive regulation of cellular biosynthetic process	11	0.047	2.0
regulation of ARF protein signal transduction	3	0.048	8.5
positive regulation of biosynthetic process	11	0.051	2.0
circadian rhythm	3	0.052	8.1
enzyme linked receptor protein signaling pathway	7	0.056	2.5
regulation of cyclic nucleotide biosynthetic process	4	0.058	4.5
regulation of nucleotide biosynthetic process	4	0.058	4.5
regulation of cyclic nucleotide metabolic process	4	0.062	4.4
regulation of muscle cell apoptosis	2	0.062	31.0
regulation of transcription from RNA polymerase II	11	0.065	1.9

promoter			
regulation of nucleotide metabolic process	4	0.066	4.3
insulin-like growth factor receptor signaling pathway	2	0.070	27.6
intracellular signaling cascade	16	0.074	1.6
cation transport	9	0.075	2.0
bone development	4	0.075	4.0
positive regulation of transcription from RNA polymerase II promoter	7	0.077	2.3
proteolysis	14	0.077	1.6
positive regulation of tyrosine phosphorylation of Stat5 protein	2	0.077	24.8
positive regulation of macromolecule biosynthetic process	10	0.078	1.9
regulation of multicellular organism growth	3	0.085	6.1
phosphorus metabolic process	13	0.087	1.7
phosphate metabolic process	13	0.087	1.7
positive regulation of transcription, DNA-dependent	8	0.087	2.1
cell division	6	0.087	2.5
modification-dependent macromolecule catabolic process	9	0.088	1.9
modification-dependent protein catabolic process	9	0.088	1.9
positive regulation of RNA metabolic process	8	0.090	2.1
regulation of gene-specific transcription	4	0.092	3.7
positive regulation of molecular function	9	0.096	1.9
small GTPase mediated signal transduction	6	0.097	2.4
regulation of cell proliferation	11	0.097	1.7
regulation of tyrosine phosphorylation of Stat5 protein	2	0.099	19.1
tube development	5	0.099	2.8
hsa05200:Pathways in cancer	7	0.046	2.6

FDR < 0.04

Term	Count	P-Value	Fold Enrichment
Many sequence variants affecting diversity of adult human height	13	0.010	2.3
Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity	4	0.032	5.5
Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels	4	0.037	5.3
Severe combined immunodeficiency, B cell-negative	2	0.078	25.0
Combined cellular and humoral immune defects with granulomas	2	0.078	25.0
Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX	3	0.080	6.2
positive regulation of macromolecule biosynthetic process	41	0.001	1.7
positive regulation of cellular biosynthetic process	42	0.002	1.7
skeletal system development	24	0.002	2.0
positive regulation of RNA metabolic process	32	0.002	1.8
positive regulation of biosynthetic process	42	0.002	1.6
skeletal system morphogenesis	12	0.003	2.9
positive regulation of nitrogen compound metabolic process	39	0.003	1.6
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	38	0.003	1.6
positive regulation of transcription, DNA-dependent	31	0.003	1.8

transmembrane receptor protein tyrosine kinase signaling pathway	18	0.004	2.2
enzyme linked receptor protein signaling pathway	24	0.004	1.9
positive regulation of macromolecule metabolic process	48	0.004	1.5
gland development	13	0.004	2.6
mRNA transport	10	0.005	3.1
positive regulation of cell migration	10	0.005	3.0
positive regulation of cell proliferation	27	0.006	1.8
regulation of DNA replication	8	0.008	3.5
induction of apoptosis	22	0.008	1.9
regulation of growth	23	0.008	1.8
induction of programmed cell death	22	0.008	1.9
mammary gland development	8	0.008	3.4
response to endogenous stimulus	26	0.008	1.7
sodium ion transport	12	0.009	2.5
regulation of fibroblast proliferation	6	0.009	4.6
positive regulation of gene expression	34	0.009	1.6
response to hormone stimulus	24	0.009	1.8
nucleic acid transport	10	0.009	2.8
RNA transport	10	0.009	2.8
establishment of RNA localization	10	0.009	2.8
regulation of cell migration	14	0.010	2.2
regulation of cell proliferation	43	0.010	1.5
positive regulation of cell motion	10	0.010	2.8
positive regulation of locomotion	10	0.010	2.8
cellular response to hormone stimulus	12	0.010	2.4
intracellular signaling cascade	63	0.011	1.4
positive regulation of transcription from RNA polymerase II promoter	24	0.011	1.8
limb morphogenesis	10	0.011	2.7
appendage morphogenesis	10	0.011	2.7
RNA localization	10	0.011	2.7
transmembrane transport	33	0.012	1.6
regulation of cellular component size	19	0.012	1.9
cellular response to insulin stimulus	8	0.012	3.2
appendage development	10	0.014	2.6
limb development	10	0.014	2.6
cell fate commitment	12	0.014	2.3
response to alkaloid	7	0.014	3.5
positive regulation of fibroblast proliferation	5	0.014	5.2
hemopoietic or lymphoid organ development	18	0.016	1.9
positive regulation of apoptosis	26	0.017	1.6
cell motion	28	0.017	1.6
positive regulation of transcription	32	0.018	1.5
growth	14	0.018	2.1
small GTPase mediated signal transduction	20	0.018	1.8
positive regulation of programmed cell death	26	0.019	1.6
positive regulation of DNA replication	5	0.019	4.8
positive regulation of cell death	26	0.019	1.6
regulation of cell development	15	0.020	2.0
regulation of cell size	15	0.021	2.0
chordate embryonic development	21	0.021	1.7
phosphate metabolic process	49	0.023	1.4
phosphorus metabolic process	49	0.023	1.4
response to nicotine	4	0.023	6.4
embryonic development ending in birth or egg hatching	21	0.023	1.7
nucleobase, nucleoside, nucleotide and nucleic acid transport	10	0.024	2.4
regulation of locomotion	14	0.026	2.0
response to organic substance	38	0.026	1.4
regulation of cell motion	14	0.027	2.0

monovalent inorganic cation transport	20	0.027	1.7
immune system development	18	0.027	1.8
response to peptide hormone stimulus	12	0.027	2.1
epithelial cell differentiation	11	0.030	2.2
phosphorylation	41	0.031	1.4
spleen development	4	0.031	5.7
response to insulin stimulus	9	0.031	2.4
regulation of apoptosis	41	0.033	1.4
tube development	15	0.034	1.8
response to vitamin	7	0.034	2.9
embryonic morphogenesis	19	0.036	1.7
regulation of neurotransmitter levels	7	0.037	2.8
steroid metabolic process	14	0.037	1.9
regulation of programmed cell death	41	0.037	1.4
bone development	10	0.038	2.2
cell fate specification	6	0.039	3.2
regulation of cell death	41	0.039	1.4
insulin-like growth factor receptor signaling pathway	3	0.041	9.0
embryonic appendage morphogenesis	8	0.041	2.5
embryonic limb morphogenesis	8	0.041	2.5
epithelium development	15	0.042	1.8
regulation of transcription from RNA polymerase II promoter	37	0.044	1.4
axon guidance	9	0.044	2.3
thymus development	4	0.046	4.9
leukocyte chemotaxis	5	0.046	3.7
Golgi transport vesicle coating	3	0.050	8.1
COPI coating of Golgi vesicle	3	0.050	8.1
Golgi vesicle budding	3	0.050	8.1
regulation of epidermal growth factor receptor signaling pathway	4	0.051	4.7
transcription from RNA polymerase II promoter	15	0.052	1.7
regulation of nervous system development	13	0.052	1.8
regulation of cell morphogenesis	10	0.053	2.1
axonogenesis	13	0.054	1.8
cell chemotaxis	5	0.054	3.5
positive regulation of DNA metabolic process	6	0.055	2.9
induction of apoptosis by extracellular signals	9	0.055	2.2
neuron differentiation	24	0.057	1.5
regulation of binding	11	0.057	1.9
regulation of neuron differentiation	10	0.058	2.0
energy coupled proton transport, down electrochemical gradient	5	0.059	3.4
ATP synthesis coupled proton transport	5	0.059	3.4
regulation of organelle organization	14	0.059	1.7
regulation of transferase activity	21	0.060	1.5
regulation of DNA metabolic process	9	0.060	2.1
keratinocyte proliferation	3	0.060	7.4
positive regulation of transferase activity	15	0.061	1.7
response to nutrient levels	13	0.061	1.8
embryonic digit morphogenesis	4	0.063	4.3
regulation of ossification	7	0.068	2.4
response to vitamin A	5	0.068	3.2
cell morphogenesis involved in differentiation	15	0.068	1.7
tissue morphogenesis	12	0.070	1.8
dorsal/ventral pattern formation	6	0.070	2.7
central nervous system projection neuron axonogenesis	3	0.070	6.8
vesicle targeting, to, from or within Golgi	3	0.070	6.8
negative regulation of molecular function	19	0.070	1.5
myoblast migration	2	0.072	27.1
response to low density lipoprotein stimulus	2	0.072	27.1

regulation of cellular catabolic process	6	0.074	2.7
cell growth	6	0.074	2.7
protein amino acid phosphorylation	33	0.075	1.3
regulation of DNA binding	9	0.079	2.0
negative regulation of organelle organization	7	0.082	2.3
cation transport	28	0.083	1.4
regulation of small GTPase mediated signal transduction	15	0.084	1.6
cell morphogenesis involved in neuron differentiation	13	0.087	1.7
positive regulation of kinase activity	14	0.087	1.6
regulation of neurogenesis	11	0.087	1.8
regulation of smooth muscle cell proliferation	5	0.089	2.9
regulation of behavior	5	0.089	2.9
skin development	4	0.090	3.7
positive regulation of developmental process	16	0.091	1.6
lipid biosynthetic process	18	0.091	1.5
response to estrogen stimulus	8	0.093	2.1
Ras protein signal transduction	8	0.093	2.1
steroid biosynthetic process	7	0.094	2.2
skeletal muscle organ development	6	0.096	2.5
skeletal muscle tissue development	6	0.096	2.5
neuron projection morphogenesis	13	0.096	1.7
intracellular transport	32	0.096	1.3
regulation of chromosome organization	4	0.097	3.6
tube morphogenesis	9	0.098	1.9
response to DNA damage stimulus	20	0.098	1.5
hemopoiesis	14	0.098	1.6
hsa05214:Glioma	10	0.001	4.2
hsa04360:Axon guidance	13	0.003	2.7
hsa05218:Melanoma	8	0.017	3.0
hsa04960:Aldosterone-regulated sodium reabsorption	6	0.018	3.9
hsa04144:Endocytosis	14	0.021	2.0
hsa04910:Insulin signaling pathway	11	0.031	2.1
hsa00564:Glycerophospholipid metabolism	7	0.042	2.7
hsa04722:Neurotrophin signaling pathway	10	0.044	2.1
hsa04730:Long-term depression	7	0.045	2.7
hsa04350:TGF-beta signaling pathway	8	0.045	2.4
hsa05200:Pathways in cancer	19	0.068	1.5
hsa05219:Bladder cancer	5	0.072	3.1
hsa00920:Sulfur metabolism	3	0.073	6.6
hsa04010:MAPK signaling pathway	16	0.078	1.6
hsa05216:Thyroid cancer	4	0.094	3.6

```
################################################################################
# R code to:
# Calculate the proportion of selective sweep loci that are also composite dog
# height SNPs. Data from (Akey et al., 2010; Vaysse et al., 2011)
################################################################################

#Read in CanMapAssociation from Boyko et al. 2010 (http://tinyurl.com/mn6kt22)
cma <- read.table("CanMapAssociation",header=T)
#Select only relevant rows and columns
cma <- subset(cma,cma$HeightWithers>0 | cma$BodyLength>0 | cma$NeckLength>0)
cma <- cma[,c(1:7,13,34,76)]
cma <- subset(cma, cma$chrom!="chrX")

#Get maximum p-value between HeightWithers, BodyLength, and NeckLength
GetHeight<-function(x) {
  bl <- as.numeric(x[8])
  hw <- as.numeric(x[9])
  nl <- as.numeric(x[10])
  vec <- c(bl,hw,nl)
  ht.max <- max(vec)
  ht.min <- min(vec)
  ht.mid <- which(vec!=ht.max & vec!=ht.min)
  ht.mid <- vec[ht.mid]
  if(ht.max>ht.mid){ # ht.mid+1 is the default
    id <- which(vec==ht.max)
  }
  else {id <- 0}
  return(c(ht.max,id))
}
ht <- apply(cma,1,GetHeight)
ht <- data.frame(t(ht))
names(ht) <- c("ht","id")
summary(factor(ht$id))
cma.merge <- data.frame(cma,ht)

# Without the X chromosome
# 0.02=3.537602 | 0.04=2.647817 | Bonferroni 0.0093=4.30103
dht.bon <- subset(cma.merge,cma.merge$ht>4.30103) # Bonferroni
dht.02 <- subset(cma.merge,cma.merge$ht>3.537602) # 0.02
dht.04 <- subset(cma.merge,cma.merge$ht>2.647817) # 0.04
library(GenomicRanges)
dhtG.bon <- with(dht.bon, GRanges(chrom,
                                  IRanges(chromEnd-1, chromEnd), "+",
                                  id=paste0(chrom,":",chromEnd,"_",ht,"_",id)))
dhtG.02 <- with(dht.02, GRanges(chrom,
                                  IRanges(chromEnd-1, chromEnd), "+",
                                  id=paste0(chrom,":",chromEnd,"_",ht,"_",id)))
dhtG.04 <- with(dht.04, GRanges(chrom,
                                  IRanges(chromEnd-1, chromEnd), "+",
                                  id=paste0(chrom,":",chromEnd,"_",ht,"_",id)))

akey <- read.table("akey.data")
names(akey) <- c("chrom","pos")
akey$chrom <- paste0("chr", akey$chr)
akey.G <- with(akey, GRanges(chrom,
                                  IRanges(pos-200000, pos+199999), "+",
                                  id=paste0(chrom,":",pos)))
akey.bon.o <- findOverlaps(dhtG.bon,akey.G)
table(as.table(t(akey.bon.o)))
table(as.table(akey.bon.o))
akey.02.o <- findOverlaps(dhtG.02,akey.G)
table(as.table(t(akey.02.o)))
table(as.table(akey.02.o))
```

```
akey.04.o <- findOverlaps(dhtG.04,akey.G)
table(as.table(t(akey.04.o)))
table(as.table(akey.04.o))

s4 <- read.table("S4.data", header=TRUE, sep='\t')
s4.cut <- s4[,c(1:3)]
s5.di <- read.table("S5.di.data", header=TRUE, sep='\t')
s5.di.FDR <- subset(s5.di, s5.di$X5..FDR.p.value < 0.1)
s5.di.cut <- s5.di.FDR[,c(3:5)]
names(s5.di.cut) <- c("chrom", "start", "end")
s5.si <- read.table("S5.si.data", header=TRUE, sep='\t')
s5.si.FDR <- subset(s5.si, s5.si$X5..FDR.p.value < 0.05)
s5.si.cut <- s5.si.FDR[,c(3:5)]
names(s5.si.cut) <- c("chrom", "start", "end")
vay <- rbind(s4.cut,s5.di.cut,s5.si.cut)
vay.G <- with(vay, GRanges(paste0("chr",chrom),
                           IRanges(start, end), "+",
                           id=paste0(chrom,":",start)))
vay.bon.o <- findOverlaps(dhtG.bon,vay.G)
table(as.table(t(vay.bon.o)))
table(as.table(vay.bon.o))
vay.02.o <- findOverlaps(dhtG.02,vay.G)
table(as.table(t(vay.02.o)))
table(as.table(vay.02.o))
vay.04.o <- findOverlaps(dhtG.04,vay.G)
table(as.table(t(vay.04.o)))
table(as.table(vay.04.o))
```

```
################################################################################
# R code to:
# Evaluate distribution of Boyko et al. 2010 SNPs used in QTL Sharing analysis
################################################################################

cma <- read.table("CanMapAssociation",header=T)
cma <- subset(cma, cma$HeightWithers>0 | cma$BodyLength>0 | cma$NeckLength>0)
cma <- cma[,c(1:7,13,34,76)]
cma <- subset(cma, cma$chrom!="chrX")

library(GenomicRanges)
library(biomaRt)
mart <- useMart("ensembl")
Cmart <- useDataset("cfamiliaris_gene_ensembl",mart)

#Do QTL sharing analysis
Hmart <- useDataset("hsapiens_gene_ensembl",mart)
humanDogOrtholog <- getLDS(attributes=c("chromosome_name","start_position",
                                      "end_position","ensembl_gene_id",
                                      "hgnc_symbol"),
                         filters="with_homolog_cfam", values=TRUE,mart=Hmart,
                         attributesL=c("chromosome_name","start_position",
                                       "end_position"),martL=Cmart)
names(humanDogOrtholog) <-
c("chr","start","end","eID","HGNC","Dchr","Dstart","Dend")
humanDogOrtholog <- subset(humanDogOrtholog, humanDogOrtholog$HGNC!="")
humanDogOrtholog <- subset(humanDogOrtholog, humanDogOrtholog$chr!="X")
humanDogOrtholog <- subset(humanDogOrtholog, humanDogOrtholog$Dchr!="X")
humanDogOrtholog <- subset(humanDogOrtholog, humanDogOrtholog$chr!="Y")
humanDogOrtholog <- subset(humanDogOrtholog, humanDogOrtholog$Dchr!="Y")

#Remove duplicate entries, retaining the top entry which is the best hit, and thus
likely ortholog
hdH <- subset(humanDogOrtholog,!duplicated(humanDogOrtholog$HGNC))

#Liftover ortholog coordinates from canFam3 to canFam2
library(rtracklayer)
hdHG <- with(hdH,GRanges(paste0("chr",Dchr),IRanges(Dstart,
Dend),"+",HGNC,id=eID))
chain <- import.chain("canFam3ToCanFam2.over.chain")
hdHG2.list <- liftOver(hdHG,chain)
hdHG2 <- unlist(hdHG2.list)

#Read in manually annotated orthologs that are human height genes to supplement
ensembl
#These coordinates are already in CanFam2
man.hdH <-
read.table("manual.orthologs.txt",header=TRUE,sep="\t",stringsAsFactors=FALSE)
man.hdHG <- with(man.hdH,GRanges(paste0("chr",chr),IRanges(start,
end),"+",HGNC,id=eID))

#Combine ensembl and manually annotated orthologs
all.HG <- c(man.hdHG,hdHG2)
all.HG <- subset(all.HG,!duplicated(all.HG$HGNC))

#Read in NHGRI GWAS catalog for height (http://www.genome.gov/gwastudies/)
library(gdata)
hgwas <- read.xls("MyGWASSearch_1_22_14.xls")

#Get column of reported genes and clean it up
repG<-hgwas[,14]
CleanGenes<-function(x) {
  output <- character(0)
  for(i in 1:length(x)) {
    line <- as.character(x[i])
    lines <- unlist(sapply(line,strsplit,","))
    lines <- unlist(sapply(lines,strsplit,"/"))
    names(lines)<-c()
    lines <- gsub("\n","",lines)
    lines <- gsub(" ","",lines)
    output <- c(output,lines)
  }
  return(unique(output))
  return(output)
}
humanHeightGenes <- as.factor(CleanGenes(repG))

hit.hum <- toString(subset(humanHeightGenes, humanHeightGenes %in% all.HG$HGNC))
hit.hum <- strsplit(hit.hum,", ")
hit.hum <- unlist(hit.hum)

hum.ht.cf2 <- subset(all.HG, all.HG$HGNC %in% hit.hum)
hum.ht.cf2.200K <- hum.ht.cf2 + 200000

dog.G <- with(cma, GRanges(chrom,
                                IRanges(chromEnd-1, chromEnd), "+",
                                id=paste0(chrom,":",chromEnd)))

dog.hum.ht <- findOverlaps(hum.ht.cf2.200K, dog.G)
snps.all <- as.table(t(dog.hum.ht))
hum.report <- as.table(dog.hum.ht)
hum.report <- data.frame(hum.ht.cf2$HGNC,hum.report)
names(hum.report) <- c("Gene", "SNPs")
summary(factor(hum.report$SNPs))
subset(hum.report, hum.report$SNPs == 0)
snps.report <- data.frame(cma,snps.all)
```

Appendix B

Code and Associated Data for Chapter 2

```
################################################################################
# R code to:
# Plot the number of significant SNPs and chromsomes for traits taken from
# Boyko et al. 2010.
################################################################################

cma <- read.table("CanMapAssociation_chr_fixed",header=T)
id.traits <- cma[,1:6]
reg.traits <- cma[,c(7,9,11,13,15,17,18,20,22,24,26,28,30,32,34,36,38,40,42,44,
                     46,48,50,52,54,56,58,60,62,64,66,68,70,72,74,76,78,80,82,
                     84,86,88,90,92,94,96,98,100,102,104,106,108,110,112,114,
                     116,118)]
reg.abs.traits <- reg.traits[,c(1:2,4:5,7:8,10:17,28:29,35:36,40,42:43)]
reg.skl.traits <- reg.traits[,c(3,6,9,18:27,30:34,37:39,41,44:57)]

allo.traits <- cma[,c(8,10,12,14,16,19,21,23,25,27,29,31,33,35,37,39,41,
                             43,45,47,49,51,53,55,57,59,61,63,65,67,69,71,73,75,
                             77,79,81,83,85,87,89,91,93,95,97,99,101,103,105,
                             107,109,111,113,115,117)]
allo.abs.traits <- allo.traits[,c(1,3:4,6,8:15,26:27,33:34,38,40:41)]
allo.skl.traits <- allo.traits[,c(2,5,7,16:25,28:32,35:37,39,42:55)]

SigSnps <- function(x, sig) {
  df <- data.frame(id.traits,x)
  df.sig <- subset(df, df$x > sig)
  return(length(df.sig[,1]))
}

SigChrs <- function(x, sig) {
  df <- data.frame(id.traits,x)
  df.sig <- subset(df, df$x > sig)
  return(length(summary(factor(df.sig$chrom))))
}
summary(c(apply(reg.abs.traits,2,SigSnps,4.30103),apply(reg.skl.traits,2,SigSnps,4
.0)))
summary(c(apply(reg.abs.traits,2,SigChrs,4.30103),apply(reg.skl.traits,2,SigChrs,4
.0)))
summary(c(apply(allo.abs.traits,2,SigSnps,4.30103),apply(allo.skl.traits,2,SigSnps
,4.0)))
summary(c(apply(allo.abs.traits,2,SigChrs,4.30103),apply(allo.skl.traits,2,SigChrs
,4.0)))

plot(c(apply(reg.abs.traits,2,SigSnps,4.30103),apply(reg.skl.traits,2,SigSnps,4.0)
))
plot(c(apply(reg.abs.traits,2,SigChrs,4.30103),apply(reg.skl.traits,2,SigChrs,4.0)
))

#exported as 4X8 pdfs and modified in Illustrator
barchart(c(apply(reg.abs.traits,2,SigSnps,4.30103),apply(reg.skl.traits,2,SigSnps,
4.0)))
barchart(c(apply(reg.abs.traits,2,SigChrs,4.30103),apply(reg.skl.traits,2,SigChrs,
4.0)))
```

```
################################################################################
# R code to:
# Reconstruct p-values for height at the withers, examine QTL effect sizes, and
# create relevant plots
################################################################################

dog <- read.table("CanMapAssociation_chr_fixed",header=T)
dog <- subset(dog,!dog$chrom %in% "chrX") #exclude the X chromosome
top12 <-
dog[c(5861,8597,11711,15801,29736,34459,40858,44692,47154,49762,54230,56381),1:3]
top12.names <- paste0(top12$chrom, top12$chromEnd)
dog <- subset(dog, dog$HeightWithers>0)
source("allele_freq.R")
size <- read.table("breed_height.txt",header=T)
size <- size[-c(3,20,34,40,50,51,54,58,77,79),] #At least 9 dogs in breed
attach(size)

ht1 <- data.frame(GreatDane1,IrishWolfhound1,Mastiff1,Greyhound1,SaintBernard1,
                BullMastiff1,Newfoundland1,Borzoi1,Kuvasz1,Akita1,
                DobermanPinscher1,AfghanHound1,Bloodhound1,BerneseMountainDog1,
                GiantSchnauzer1,Rottweiler1,Briard1,GermanShepherdDog1,Saluki1,
                IbizanHound1,AlaskanMalamute1,StandardPoodle1,GoldenRetriever1,
                Boxer1,FlatCoatedRetriever1,Bulldog1,LabradorRetriever1,
                IrishWaterSpaniel1,GermanShorthairedPointer1,BorderCollie1,
                OldEnglishSheepdog1,SiberianHusky1,AustralianShepherd1,
                PortugueseWaterDog1,ItalianGreyhound1,SpringerSpaniel1,
                MiniatureBullTerrier1,ChineseSharPei1,StandardSchnauzer1,
                Whippet1,ChowChow1,Brittany1,FrenchBulldog1,
                StaffordshireBullTerrier1,Basenji1,CockerSpaniel1,BassetHound1,
                GlenofImaalTerrier1,Beagle1,ShetlandSheepdog1,Pug1,
                JackRussellTerrier1,CavalierKingCharlesSpaniel1,
                PembrokeWelshCorgi1,CardiganWelshCorgi1,
                WestHighlandWhiteTerrier1,MiniaturePinscher1,ScottishTerrier1,
                CairnTerrier1,ShihTzu1,AustralianTerrier1,NorwichTerrier1,
                ToyPoodle1,Havanese1,Papillon1,Pomeranian1,
                PetitBassetGriffonVendeen1,Pekingese1,Dachshund1,Chihuahua1)
lm.ht <- apply(ht1,1,function(x) summary(lm(Height~x)))
lm.ht.out <- lapply(lm.ht,function(x) x$coefficients[8])
ht.pvals <- as.numeric(lm.ht.out)
ht.lp <- -log10(ht.pvals)
doggy.names <- dog[,1:3]
dog.names <- paste0(doggy.names$chrom,doggy.names$chromEnd)
d.tops <- which(dog.names %in% top12.names)
ht.tops<-ht1[d.tops,]
ht.tops[2,] <- 1-ht.tops[2,]
ht.tops[8,] <- 1-ht.tops[8,]
ht.tops[9,] <- 1-ht.tops[9,]
ht.tops[11,] <- 1-ht.tops[11,]
summary(lm(Height~as.numeric(ht.tops[5,]))) #IGF1 45%
summary(lm(Height~as.numeric(ht.tops[6,]))) #FGF4 retrogene 44%
summary(lm(Height~as.numeric(ht.tops[6,])+as.numeric(ht.tops[5,]))) #IGF1 and FGF4
70%
dht1 <- data.frame(dog,ht.lp)
top.12 <- dht1[d.tops,c(34,289)]

#82%, 77% without IGF1
tops <-
lm(Height~as.numeric(ht.tops[1,])+as.numeric(ht.tops[2,])+as.numeric(ht.tops[3,])+
          as.numeric(ht.tops[4,])+as.numeric(ht.tops[5,])+as.numeric(ht.tops[6,])+
          as.numeric(ht.tops[7,])+as.numeric(ht.tops[8,])+as.numeric(ht.tops[9,])+
       as.numeric(ht.tops[10,])+as.numeric(ht.tops[11,])+as.numeric(ht.tops[12,]))
summary(tops)
```

```
tops1 <-
lm(Height~as.numeric(ht.tops[5,])+as.numeric(ht.tops[6,])+as.numeric(ht.tops[8,]))
#71% IGF1, FGF4 retrogene, ZFP64

tops2 <-
lm(Height~as.numeric(ht.tops[1,])+as.numeric(ht.tops[7,])+as.numeric(ht.tops[10,])
) #54% IGF1R, GPC6, SUFU

tops3 <-
lm(Height~as.numeric(ht.tops[4,])+as.numeric(ht.tops[3,])+as.numeric(ht.tops[9,]))
#47% SMAD2, BANP, MED13L

tops4 <-
lm(Height~as.numeric(ht.tops[2,])+as.numeric(ht.tops[11,])+as.numeric(ht.tops[12,]
)) #40% IGF2BP2, STC2, BMP3

#Correlates alleles with IGF1
GetCors <- function(x) {
  test <- cor.test(as.numeric(x),as.numeric(ht.tops[5,]))
  return(c(test$p.value, test$estimate))
}
apply(ht.tops, 1, GetCors)

#Plot Boyko et al. 2010 p-values for height at the withers with reconstructed
#p-values
plot(dog$HeightWithers,ht.lp, ylab="Reconstructed (-log10 p-value)",
     xlab="Boyko et al. 2010 (-log10 p-value)",
     main="Height at the withers")
abline(0,1,lwd=2)
abline(v=4.3,lty=2)
abline(h=4.3,lty=2)
points(top.12$HeightWithers,top.12$ht.lp,col="red",pch=20)

#Plot height distributions for purebred and mixed-breed dogs
par(mfrow=c(1,2))
plot(size$Height,ylab="Height at the withers", xlab="Dog breeds")
plot(sort(red.t$Height, decreasing=TRUE), ylab="Scapula + humerus + radius
length", xlab="Mixed-breed dogs")
```

```
###############################################################################
# R code to:
# Calculate allele frequencies for each allele for every breed
###############################################################################

attach(dog)
Jackal1<-JackalAllele1/(JackalAllele1+JackalAllele2)
Jackal2<-JackalAllele2/(JackalAllele1+JackalAllele2)
RedWolf1<-RedWolfAllele1/(RedWolfAllele1+RedWolfAllele2)
RedWolf2<-RedWolfAllele2/(RedWolfAllele1+RedWolfAllele2)
BostonTerrier1<-BostonTerrierAllele1/(BostonTerrierAllele1+BostonTerrierAllele2)
BostonTerrier2<-BostonTerrierAllele2/(BostonTerrierAllele1+BostonTerrierAllele2)
Coyote1<-CoyoteAllele1/(CoyoteAllele1+CoyoteAllele2)
Coyote2<-CoyoteAllele2/(CoyoteAllele1+CoyoteAllele2)
Wolf1<-WolfAllele1/(WolfAllele1+WolfAllele2)
Wolf2<-WolfAllele2/(WolfAllele1+WolfAllele2)
AfghanHound1<-AfghanHoundAllele1/(AfghanHoundAllele1+AfghanHoundAllele2)
AfghanHound2<-AfghanHoundAllele2/(AfghanHoundAllele1+AfghanHoundAllele2)
Akita1<-AkitaAllele1/(AkitaAllele1+AkitaAllele2)
Akita2<-AkitaAllele2/(AkitaAllele1+AkitaAllele2)
AlaskanMalamute1<-
AlaskanMalamuteAllele1/(AlaskanMalamuteAllele1+AlaskanMalamuteAllele2)
AlaskanMalamute2<-
AlaskanMalamuteAllele2/(AlaskanMalamuteAllele1+AlaskanMalamuteAllele2)
AmericanEskimoDog1<-
AmericanEskimoDogAllele1/(AmericanEskimoDogAllele1+AmericanEskimoDogAllele2)
AmericanEskimoDog2<-
AmericanEskimoDogAllele2/(AmericanEskimoDogAllele1+AmericanEskimoDogAllele2)
AustralianShepherd1<-
AustralianShepherdAllele1/(AustralianShepherdAllele1+AustralianShepherdAllele2)
AustralianShepherd2<-
AustralianShepherdAllele2/(AustralianShepherdAllele1+AustralianShepherdAllele2)
AustralianTerrier1<-
AustralianTerrierAllele1/(AustralianTerrierAllele1+AustralianTerrierAllele2)
AustralianTerrier2<-
AustralianTerrierAllele2/(AustralianTerrierAllele1+AustralianTerrierAllele2)
Basenji1<-BasenjiAllele1/(BasenjiAllele1+BasenjiAllele2)
Basenji2<-BasenjiAllele2/(BasenjiAllele1+BasenjiAllele2)
BassetHound1<-BassetHoundAllele1/(BassetHoundAllele1+BassetHoundAllele2)
BassetHound2<-BassetHoundAllele2/(BassetHoundAllele1+BassetHoundAllele2)
Beagle1<-BeagleAllele1/(BeagleAllele1+BeagleAllele2)
Beagle2<-BeagleAllele2/(BeagleAllele1+BeagleAllele2)
BerneseMountainDog1<-
BerneseMountainDogAllele1/(BerneseMountainDogAllele1+BerneseMountainDogAllele2)
BerneseMountainDog2<-
BerneseMountainDogAllele2/(BerneseMountainDogAllele1+BerneseMountainDogAllele2)
Bloodhound1<-BloodhoundAllele1/(BloodhoundAllele1+BloodhoundAllele2)
Bloodhound2<-BloodhoundAllele2/(BloodhoundAllele1+BloodhoundAllele2)
BorderCollie1<-BorderCollieAllele1/(BorderCollieAllele1+BorderCollieAllele2)
BorderCollie2<-BorderCollieAllele2/(BorderCollieAllele1+BorderCollieAllele2)
Borzoi1<-BorzoiAllele1/(BorzoiAllele1+BorzoiAllele2)
Borzoi2<-BorzoiAllele2/(BorzoiAllele1+BorzoiAllele2)
Boxer1<-BoxerAllele1/(BoxerAllele1+BoxerAllele2)
Boxer2<-BoxerAllele2/(BoxerAllele1+BoxerAllele2)
Briard1<-BriardAllele1/(BriardAllele1+BriardAllele2)
Briard2<-BriardAllele2/(BriardAllele1+BriardAllele2)
Brittany1<-BrittanyAllele1/(BrittanyAllele1+BrittanyAllele2)
Brittany2<-BrittanyAllele2/(BrittanyAllele1+BrittanyAllele2)
BrusselsGriffon1<-
BrusselsGriffonAllele1/(BrusselsGriffonAllele1+BrusselsGriffonAllele2)
BrusselsGriffon2<-
BrusselsGriffonAllele2/(BrusselsGriffonAllele1+BrusselsGriffonAllele2)
BullMastiff1<-BullMastiffAllele1/(BullMastiffAllele1+BullMastiffAllele2)
```

```
BullMastiff2<-BullMastiffAllele2/(BullMastiffAllele1+BullMastiffAllele2)
BullTerrier1<-BullTerrierAllele1/(BullTerrierAllele1+BullTerrierAllele2)
BullTerrier2<-BullTerrierAllele2/(BullTerrierAllele1+BullTerrierAllele2)
Bulldog1<-BulldogAllele1/(BulldogAllele1+BulldogAllele2)
Bulldog2<-BulldogAllele2/(BulldogAllele1+BulldogAllele2)
CairnTerrier1<-CairnTerrierAllele1/(CairnTerrierAllele1+CairnTerrierAllele2)
CairnTerrier2<-CairnTerrierAllele2/(CairnTerrierAllele1+CairnTerrierAllele2)
CardiganWelshCorgi1<-
CardiganWelshCorgiAllele1/(CardiganWelshCorgiAllele1+CardiganWelshCorgiAllele2)
CardiganWelshCorgi2<-
CardiganWelshCorgiAllele2/(CardiganWelshCorgiAllele1+CardiganWelshCorgiAllele2)
CavalierKingCharlesSpaniel1<-
CavalierKingCharlesSpanielAllele1/(CavalierKingCharlesSpanielAllele1+CavalierKingC
harlesSpanielAllele2)
CavalierKingCharlesSpaniel2<-
CavalierKingCharlesSpanielAllele2/(CavalierKingCharlesSpanielAllele1+CavalierKingC
harlesSpanielAllele2)
Chihuahua1<-ChihuahuaAllele1/(ChihuahuaAllele1+ChihuahuaAllele2)
Chihuahua2<-ChihuahuaAllele2/(ChihuahuaAllele1+ChihuahuaAllele2)
ChineseSharPei1<-
ChineseSharPeiAllele1/(ChineseSharPeiAllele1+ChineseSharPeiAllele2)
ChineseSharPei2<-
ChineseSharPeiAllele2/(ChineseSharPeiAllele1+ChineseSharPeiAllele2)
ChowChow1<-ChowChowAllele1/(ChowChowAllele1+ChowChowAllele2)
ChowChow2<-ChowChowAllele2/(ChowChowAllele1+ChowChowAllele2)
CockerSpaniel1<-CockerSpanielAllele1/(CockerSpanielAllele1+CockerSpanielAllele2)
CockerSpaniel2<-CockerSpanielAllele2/(CockerSpanielAllele1+CockerSpanielAllele2)
Collie1<-CollieAllele1/(CollieAllele1+CollieAllele2)
Collie2<-CollieAllele2/(CollieAllele1+CollieAllele2)
Dachshund1<-DachshundAllele1/(DachshundAllele1+DachshundAllele2)
Dachshund2<-DachshundAllele2/(DachshundAllele1+DachshundAllele2)
DobermanPinscher1<-
DobermanPinscherAllele1/(DobermanPinscherAllele1+DobermanPinscherAllele2)
DobermanPinscher2<-
DobermanPinscherAllele2/(DobermanPinscherAllele1+DobermanPinscherAllele2)
EnglishCockerSpaniel1<-
EnglishCockerSpanielAllele1/(EnglishCockerSpanielAllele1+EnglishCockerSpanielAllel
e2)
EnglishCockerSpaniel2<-
EnglishCockerSpanielAllele2/(EnglishCockerSpanielAllele1+EnglishCockerSpanielAllel
e2)
FlatCoatedRetriever1<-
FlatCoatedRetrieverAllele1/(FlatCoatedRetrieverAllele1+FlatCoatedRetrieverAllele2)
FlatCoatedRetriever2<-
FlatCoatedRetrieverAllele2/(FlatCoatedRetrieverAllele1+FlatCoatedRetrieverAllele2)
FrenchBulldog1<-FrenchBulldogAllele1/(FrenchBulldogAllele1+FrenchBulldogAllele2)
FrenchBulldog2<-FrenchBulldogAllele2/(FrenchBulldogAllele1+FrenchBulldogAllele2)
GermanShepherdDog1<-
GermanShepherdDogAllele1/(GermanShepherdDogAllele1+GermanShepherdDogAllele2)
GermanShepherdDog2<-
GermanShepherdDogAllele2/(GermanShepherdDogAllele1+GermanShepherdDogAllele2)
GermanShorthairedPointer1<-
GermanShorthairedPointerAllele1/(GermanShorthairedPointerAllele1+GermanShorthaired
PointerAllele2)
GermanShorthairedPointer2<-
GermanShorthairedPointerAllele2/(GermanShorthairedPointerAllele1+GermanShorthaired
PointerAllele2)
GiantSchnauzer1<-
GiantSchnauzerAllele1/(GiantSchnauzerAllele1+GiantSchnauzerAllele2)
GiantSchnauzer2<-
GiantSchnauzerAllele2/(GiantSchnauzerAllele1+GiantSchnauzerAllele2)
GlenofImaalTerrier1<-
GlenofImaalTerrierAllele1/(GlenofImaalTerrierAllele1+GlenofImaalTerrierAllele2)
```

```
GlenofImaalTerrier2<-
GlenofImaalTerrierAllele2/(GlenofImaalTerrierAllele1+GlenofImaalTerrierAllele2)
GoldenRetriever1<-
GoldenRetrieverAllele1/(GoldenRetrieverAllele1+GoldenRetrieverAllele2)
GoldenRetriever2<-
GoldenRetrieverAllele2/(GoldenRetrieverAllele1+GoldenRetrieverAllele2)
GreatDane1<-GreatDaneAllele1/(GreatDaneAllele1+GreatDaneAllele2)
GreatDane2<-GreatDaneAllele2/(GreatDaneAllele1+GreatDaneAllele2)
Greyhound1<-GreyhoundAllele1/(GreyhoundAllele1+GreyhoundAllele2)
Greyhound2<-GreyhoundAllele2/(GreyhoundAllele1+GreyhoundAllele2)
Havanese1<-HavaneseAllele1/(HavaneseAllele1+HavaneseAllele2)
Havanese2<-HavaneseAllele2/(HavaneseAllele1+HavaneseAllele2)
IbizanHound1<-IbizanHoundAllele1/(IbizanHoundAllele1+IbizanHoundAllele2)
IbizanHound2<-IbizanHoundAllele2/(IbizanHoundAllele1+IbizanHoundAllele2)
IrishWaterSpaniel1<-
IrishWaterSpanielAllele1/(IrishWaterSpanielAllele1+IrishWaterSpanielAllele2)
IrishWaterSpaniel2<-
IrishWaterSpanielAllele2/(IrishWaterSpanielAllele1+IrishWaterSpanielAllele2)
IrishWolfhound1<-
IrishWolfhoundAllele1/(IrishWolfhoundAllele1+IrishWolfhoundAllele2)
IrishWolfhound2<-
IrishWolfhoundAllele2/(IrishWolfhoundAllele1+IrishWolfhoundAllele2)
ItalianGreyhound1<-
ItalianGreyhoundAllele1/(ItalianGreyhoundAllele1+ItalianGreyhoundAllele2)
ItalianGreyhound2<-
ItalianGreyhoundAllele2/(ItalianGreyhoundAllele1+ItalianGreyhoundAllele2)
JackRussellTerrier1<-
JackRussellTerrierAllele1/(JackRussellTerrierAllele1+JackRussellTerrierAllele2)
JackRussellTerrier2<-
JackRussellTerrierAllele2/(JackRussellTerrierAllele1+JackRussellTerrierAllele2)
Kuvasz1<-KuvaszAllele1/(KuvaszAllele1+KuvaszAllele2)
Kuvasz2<-KuvaszAllele2/(KuvaszAllele1+KuvaszAllele2)
LabradorRetriever1<-
LabradorRetrieverAllele1/(LabradorRetrieverAllele1+LabradorRetrieverAllele2)
LabradorRetriever2<-
LabradorRetrieverAllele2/(LabradorRetrieverAllele1+LabradorRetrieverAllele2)
Mastiff1<-MastiffAllele1/(MastiffAllele1+MastiffAllele2)
Mastiff2<-MastiffAllele2/(MastiffAllele1+MastiffAllele2)
MiniatureBullTerrier1<-
MiniatureBullTerrierAllele1/(MiniatureBullTerrierAllele1+MiniatureBullTerrierAllel
e2)
MiniatureBullTerrier2<-
MiniatureBullTerrierAllele2/(MiniatureBullTerrierAllele1+MiniatureBullTerrierAllel
e2)
MiniaturePinscher1<-
MiniaturePinscherAllele1/(MiniaturePinscherAllele1+MiniaturePinscherAllele2)
MiniaturePinscher2<-
MiniaturePinscherAllele2/(MiniaturePinscherAllele1+MiniaturePinscherAllele2)
Newfoundland1<-NewfoundlandAllele1/(NewfoundlandAllele1+NewfoundlandAllele2)
Newfoundland2<-NewfoundlandAllele2/(NewfoundlandAllele1+NewfoundlandAllele2)
NorwichTerrier1<-
NorwichTerrierAllele1/(NorwichTerrierAllele1+NorwichTerrierAllele2)
NorwichTerrier2<-
NorwichTerrierAllele2/(NorwichTerrierAllele1+NorwichTerrierAllele2)
OldEnglishSheepdog1<-
OldEnglishSheepdogAllele1/(OldEnglishSheepdogAllele1+OldEnglishSheepdogAllele2)
OldEnglishSheepdog2<-
OldEnglishSheepdogAllele2/(OldEnglishSheepdogAllele1+OldEnglishSheepdogAllele2)
Papillon1<-PapillonAllele1/(PapillonAllele1+PapillonAllele2)
Papillon2<-PapillonAllele2/(PapillonAllele1+PapillonAllele2)
Pekingese1<-PekingeseAllele1/(PekingeseAllele1+PekingeseAllele2)
Pekingese2<-PekingeseAllele2/(PekingeseAllele1+PekingeseAllele2)
```

```
PembrokeWelshCorgi1<-
PembrokeWelshCorgiAllele1/(PembrokeWelshCorgiAllele1+PembrokeWelshCorgiAllele2)
PembrokeWelshCorgi2<-
PembrokeWelshCorgiAllele2/(PembrokeWelshCorgiAllele1+PembrokeWelshCorgiAllele2)
PetitBassetGriffonVendeen1<-
PetitBassetGriffonVendeenAllele1/(PetitBassetGriffonVendeenAllele1+PetitBassetGrif
fonVendeenAllele2)
PetitBassetGriffonVendeen2<-
PetitBassetGriffonVendeenAllele2/(PetitBassetGriffonVendeenAllele1+PetitBassetGrif
fonVendeenAllele2)
Pomeranian1<-PomeranianAllele1/(PomeranianAllele1+PomeranianAllele2)
Pomeranian2<-PomeranianAllele2/(PomeranianAllele1+PomeranianAllele2)
PortugueseWaterDog1<-
PortugueseWaterDogAllele1/(PortugueseWaterDogAllele1+PortugueseWaterDogAllele2)
PortugueseWaterDog2<-
PortugueseWaterDogAllele2/(PortugueseWaterDogAllele1+PortugueseWaterDogAllele2)
Pug1<-PugAllele1/(PugAllele1+PugAllele2)
Pug2<-PugAllele2/(PugAllele1+PugAllele2)
Rottweiler1<-RottweilerAllele1/(RottweilerAllele1+RottweilerAllele2)
Rottweiler2<-RottweilerAllele2/(RottweilerAllele1+RottweilerAllele2)
SaintBernard1<-SaintBernardAllele1/(SaintBernardAllele1+SaintBernardAllele2)
SaintBernard2<-SaintBernardAllele2/(SaintBernardAllele1+SaintBernardAllele2)
Saluki1<-SalukiAllele1/(SalukiAllele1+SalukiAllele2)
Saluki2<-SalukiAllele2/(SalukiAllele1+SalukiAllele2)
Samoyed1<-SamoyedAllele1/(SamoyedAllele1+SamoyedAllele2)
Samoyed2<-SamoyedAllele2/(SamoyedAllele1+SamoyedAllele2)
ScottishDeerhound1<-
ScottishDeerhoundAllele1/(ScottishDeerhoundAllele1+ScottishDeerhoundAllele2)
ScottishDeerhound2<-
ScottishDeerhoundAllele2/(ScottishDeerhoundAllele1+ScottishDeerhoundAllele2)
ScottishTerrier1<-
ScottishTerrierAllele1/(ScottishTerrierAllele1+ScottishTerrierAllele2)
ScottishTerrier2<-
ScottishTerrierAllele2/(ScottishTerrierAllele1+ScottishTerrierAllele2)
ShetlandSheepdog1<-
ShetlandSheepdogAllele1/(ShetlandSheepdogAllele1+ShetlandSheepdogAllele2)
ShetlandSheepdog2<-
ShetlandSheepdogAllele2/(ShetlandSheepdogAllele1+ShetlandSheepdogAllele2)
ShihTzu1<-ShihTzuAllele1/(ShihTzuAllele1+ShihTzuAllele2)
ShihTzu2<-ShihTzuAllele2/(ShihTzuAllele1+ShihTzuAllele2)
SiberianHusky1<-SiberianHuskyAllele1/(SiberianHuskyAllele1+SiberianHuskyAllele2)
SiberianHusky2<-SiberianHuskyAllele2/(SiberianHuskyAllele1+SiberianHuskyAllele2)
SpringerSpaniel1<-
SpringerSpanielAllele1/(SpringerSpanielAllele1+SpringerSpanielAllele2)
SpringerSpaniel2<-
SpringerSpanielAllele2/(SpringerSpanielAllele1+SpringerSpanielAllele2)
StaffordshireBullTerrier1<-
StaffordshireBullTerrierAllele1/(StaffordshireBullTerrierAllele1+StaffordshireBull
TerrierAllele2)
StaffordshireBullTerrier2<-
StaffordshireBullTerrierAllele2/(StaffordshireBullTerrierAllele1+StaffordshireBull
TerrierAllele2)
StandardPoodle1<-
StandardPoodleAllele1/(StandardPoodleAllele1+StandardPoodleAllele2)
StandardPoodle2<-
StandardPoodleAllele2/(StandardPoodleAllele1+StandardPoodleAllele2)
StandardSchnauzer1<-
StandardSchnauzerAllele1/(StandardSchnauzerAllele1+StandardSchnauzerAllele2)
StandardSchnauzer2<-
StandardSchnauzerAllele2/(StandardSchnauzerAllele1+StandardSchnauzerAllele2)
SussexSpaniel1<-SussexSpanielAllele1/(SussexSpanielAllele1+SussexSpanielAllele2)
SussexSpaniel2<-SussexSpanielAllele2/(SussexSpanielAllele1+SussexSpanielAllele2)
ToyPoodle1<-ToyPoodleAllele1/(ToyPoodleAllele1+ToyPoodleAllele2)
```

```
ToyPoodle2<-ToyPoodleAllele2/(ToyPoodleAllele1+ToyPoodleAllele2)
WestHighlandWhiteTerrier1<-
WestHighlandWhiteTerrierAllele1/(WestHighlandWhiteTerrierAllele1+WestHighlandWhite
TerrierAllele2)
WestHighlandWhiteTerrier2<-
WestHighlandWhiteTerrierAllele2/(WestHighlandWhiteTerrierAllele1+WestHighlandWhite
TerrierAllele2)
Whippet1<-WhippetAllele1/(WhippetAllele1+WhippetAllele2)
Whippet2<-WhippetAllele2/(WhippetAllele1+WhippetAllele2)
YorkshireTerrier1<-
YorkshireTerrierAllele1/(YorkshireTerrierAllele1+YorkshireTerrierAllele2)
YorkshireTerrier2<-
YorkshireTerrierAllele2/(YorkshireTerrierAllele1+YorkshireTerrierAllele2)
villagedog1<-villagedogAllele1/(villagedogAllele1+villagedogAllele2)
villagedog2<-villagedogAllele2/(villagedogAllele1+villagedogAllele2)
```

breed_height.txt file that lists height by breed

```
Breed   Height
GreatDane1     32
IrishWolfhound1        32
ScottishDeerhound1     31
Mastiff1       30
Greyhound1     29
SaintBernard1  28
BullMastiff1   28
Newfoundland1  28
Borzoi1 28
Kuvasz1 28
Akita1  27
DobermanPinscher1      27
AfghanHound1   27
Bloodhound1    26
BerneseMountainDog1    26
GiantSchnauzer1        26
Rottweiler1    25
Briard1 25
GermanShepherdDog1     25
Collie1 25
Saluki1 25
IbizanHound1   25
AlaskanMalamute1       24
StandardPoodle1        24
GoldenRetriever1       23.5
Boxer1  23.5
FlatCoatedRetriever1   23.5
Bulldog1       23.5
LabradorRetriever1     23
IrishWaterSpaniel1     23
GermanShorthairedPointer1     23
BorderCollie1  22.8
OldEnglishSheepdog1    22
Samoyed1       22
SiberianHusky1 22
AustralianShepherd1    21.5
PortugueseWaterDog1    21.5
ItalianGreyhound1      21.5
SpringerSpaniel1       20
BullTerrier1   19.5
MiniatureBullTerrier1  19.5
ChineseSharPei1        19
StandardSchnauzer1     19
Whippet1       19
ChowChow1      18.5
Brittany1      18.5
FrenchBulldog1 17.8
StaffordshireBullTerrier1     17.3
Basenji1       17
EnglishCockerSpaniel1  16.5
BostonTerrier1 16
CockerSpaniel1 15
BassetHound1   14
SussexSpaniel1 14
GlenofImaalTerrier1    14
Beagle1 14
ShetlandSheepdog1      14
AmericanEskimoDog1     13.5
Pug1    13
JackRussellTerrier1    12.5
CavalierKingCharlesSpaniel1   12.5
PembrokeWelshCorgi1    11
CardiganWelshCorgi1    11
WestHighlandWhiteTerrier1     11
MiniaturePinscher1     11
ScottishTerrier1       10
CairnTerrier1  10
ShihTzu1       10
AustralianTerrier1     10
NorwichTerrier1        10
ToyPoodle1     10
Havanese1      9.5
Papillon1      9.5
Pomeranian1    8.5
PetitBassetGriffonVendeen1    7.5
Pekingese1     7.5
BrusselsGriffon1       7.5
Dachshund1     7
YorkshireTerrier1      6
Chihuahua1     6
```

Genotype and Phenotype Data for Mixed-breed Dogs

ID	BANP	STC2	SMAD2	FGF4	IGF1	Scapula (L)	Scapula (R)	Humerus (L)	Humerus (R)	Radius (L)	Radius (R)
35125	2	0	0	0	2	6.6845	6.7235	7.564	7.6135	7.9005	7.9025
35126	1	1	1	0	0	6.2955	6.3465	7.715	7.732	7.756	7.7055
35138	2	2	0	0	0	6.0285	6.01	6.5025	6.5285	6.6965	6.666
35154	NA	1	NA	1	0	6.5585	6.568	7.617	7.636	7.7985	7.774
35155	1	1	0	0	0	5.9565	6.0055	6.8665	6.8895	6.899	6.9125
35156	1	1	0	0	0	5.8195	5.768	6.437	6.4665	6.851	6.8695
35157	2	1	1	0	1	4.88	4.9175	5.736	5.746	5.834	5.8365
35172	1	1	0	1	2	5.337	5.347	5.182	5.1835	4.4035	4.348
35176	2	0	0	0	1	6.9445	6.8815	7.747	7.72	7.69	7.778
35177	1	1	0	0	1	6.328	6.3045	7.478	7.507	7.7395	7.782
35179	0	1	1	0	2	3.8585	3.902	4.205	4.3015	4.1895	4.185
36275	1	1	2	NA	2	5.8155	5.7935	6.877	6.855	6.626	6.5955
36276	1	1	2	2	0	3.391	3.3455	2.936	2.96	2.4065	2.3575
36277	2	1	1	1	2	4.5395	4.5955	4.035	4.059	3.663	3.698
36278	1	1	1	2	2	2.5035	2.4965	2.7425	2.745	2.4845	2.4305
36279	2	1	1	1	2	3.4285	3.4465	3.0495	3.0755	2.7435	2.7915
36550	0	1	0	0	1	6.5005	6.4935	7.5035	7.46	7.5425	7.628
36551	1	1	1	1	1	4.922	4.931	4.8815	4.9055	4.748	4.8255
36710	2	1	0	0	2	5.2155	5.2415	6.3265	6.2865	6.4235	6.368
36711	1	1	0	0	1	4.6495	4.671	5.329	5.352	5.3285	5.316
36712	NA	1	2	0	2	3.503	3.5065	3.9335	3.948	3.909	3.924
36761	2	2	1	0	0	6.6225	6.679	7.6425	7.6155	7.5975	7.635
36762	0	1	2	2	2	2.831	2.823	2.8905	2.905	2.54	2.5535
36763	0	1	2	2	2	3.589	3.6485	3.529	3.5615	3.137	3.0565
36994	0	1	1	0	0	6.6945	6.745	7.6975	7.7	7.789	7.805
36995	1	0	0	0	1	6.156	6.151	7.074	7.0515	7.202	7.1915
36996	1	1	0	0	1	6.5605	6.5605	6.7605	6.83	6.885	6.905
36997	1	0	1	1	1	5.4335	5.4445	6.229	6.259	5.987	6.0065
36998	1	1	0	0	1	6.3885	6.4315	7.2155	7.1945	7.226	7.156
37191	0	1	0	NA	0	4.2115	4.205	4.144	4.1355	3.9415	3.9635
37212	0	1	0	0	0	5.9505	5.9485	6.9415	6.9575	6.678	6.667
37213	0	1	0	NA	0	6.4805	6.6155	8.323	8.395	8.387	8.5165
37215	2	1	0	0	2	6.651	6.66	7.836	7.843	7.823	7.7695
37216	2	1	0	0	2	6.064	6.1335	6.691	6.6475	6.706	6.705
37247	0	1	2	1	2	3.9955	3.9955	4.454	4.4825	4.4	4.4145
37377	2	1	0	0	0	6.1395	6.134	7.1185	7.137	7.176	7.2055
37378	2	1	0	0	1	5.8695	5.8475	6.3665	6.3545	6.736	6.702
37379	2	0	0	0	0	6.1175	6.1155	7.096	7.0875	6.8455	6.887
37399	1	2	0	0	0	5.7165	5.745	6.5845	6.622	6.76	6.7805
37411	2	1	1	0	2	3.901	3.907	4.1185	4.1085	4.212	4.1925
37487	1	1	0	0	0	6.0435	6.0495	7.072	7.0705	7.2075	7.1935
37489	2	0	1	0	1	5.6265	5.666	6.3775	6.3465	6.389	6.38
37605	2	1	0	NA	0	5.913	5.971	7.273	7.255	7.124	7.177
37711	2	1	0	0	1	5.7415	5.77	6.436	6.3945	6.4265	6.461
37716	1	1	1	1	2	6.0305	6.0305	6.739	6.6405	6.856	6.765
37717	1	0	0	0	2	5.495	5.5055	6.4545	6.449	6.489	6.4625
37718	1	1	1	0	1	4.732	4.6155	4.602	4.555	4.2435	4.162
37815	1	2	0	0	0	6.964	6.91	8.1605	8.149	8.2945	8.3395
37836	1	1	0	0	0	6.088	6.148	6.752	6.7805	6.8435	6.836
37837	2	1	0	NA	0	6.153	6.111	7.0795	7.085	6.9315	6.94
37838	1	2	0	0	0	5.423	5.4665	6.7035	6.6665	6.6635	6.6475
37839	1	1	0	0	0	6.244	6.2305	6.998	6.985	7.162	7.064
37927	0	1	0	0	1	6.15	6.187	6.743	6.7165	7.1885	7.2025
37982	1	2	0	0	1	5.816	5.851	6.769	6.7685	6.7595	6.711
38028	1	2	0	0	1	5.659	5.667	6.4755	6.4455	6.724	6.75
38029	2	1	0	NA	2	6.252	6.287	6.7015	6.713	6.899	6.9175
38112	2	1	0	0	2	6.116	6.1515	6.904	6.8945	6.8445	6.8985
38113	2	0	0	0	0	6.8685	6.9055	7.89	7.8765	7.89	7.93
38125	1	1	0	0	2	5.9675	5.9715	7.008	6.9705	7.111	7.05
38200	2	1	0	0	1	6.6405	6.6365	8.2505	8.024	8.3165	8.2325
38201	0	1	0	0	2	5.718	5.6695	6.482	6.5005	6.76	6.706
38491	0	1	0	0	0	6.925	6.777	7.586	7.553	7.5315	7.47
38492	1	1	0	0	1	6.108	6.1625	7.1865	7.169	7.168	7.1845
38542	2	1	1	1	1	5.6095	5.6095	6.473	6.4455	6.569	6.6255
38572	1	1	1	1	1	4.808	4.825	4.974	5.013	4.7485	4.711
38574	1	1	2	0	1	6.2895	6.3035	7.2105	7.187	7.3585	7.4035
38578	1	1	0	0	2	5.8095	5.8175	6.512	6.521	6.78	6.754
38579	2	1	0	0	0	5.613	5.6145	6.3965	6.391	6.213	6.2195
39722	1	1	0	0	1	5.271	5.345	5.7155	5.7485	5.976	5.9685
39827	2	2	1	0	2	3.7545	3.7845	4.118	4.1405	4.2705	4.318
39828	1	1	0	0	0	7.0585	7.0585	8.0555	8.111	7.9465	7.962
39829	1	1	0	0	2	5.6155	5.667	6.59	6.546	6.5635	6.557
40101	2	1	0	0	1	6.4425	6.401	7.5285	7.515	7.36	7.3655
40262	0	1	0	0	0	6.351	6.3305	7.1595	7.0625	7.1235	7.0815
40439	2	1	0	0	1	6.6895	6.6545	7.701	7.7395	7.628	7.63
41059	2	1	0	0	0	6.6165	6.602	7.7715	7.7955	7.5165	7.499
41130	2	1	1	1	0	5.282	5.259	6.0845	6.132	6.052	6.1385
41604	2	1	0	0	0	5.9155	5.91	6.6785	6.66	6.741	6.768

41605	0	1	0	0	2	6.554	6.564	7.6415	7.626	7.8735	7.892
41620	0	1	0	0	2	5.173	5.187	5.9525	5.979	6.25	6.28
44404	2	2	2	2	2	3.311	3.3565	3.226	3.2625	2.8215	2.8165
44406	0	1	0	0	2	5.864	5.912	6.4125	6.423	6.5925	6.656
44407	1	1	0	0	2	5.9415	5.972	6.852	6.8225	6.812	6.7385
44430	0	1	0	1	0	5.271	5.263	5.59	5.5935	5.4685	5.4795
44590	1	1	0	0	1	5.848	5.9475	6.238	6.202	6.6105	6.661
47530	2	1	0	0	1	6.2645	6.213	7.0465	6.952	7.349	7.3225
47531	2	1	0	0	1	5.471	5.5285	6.5745	6.5545	6.34	6.364
47532	2	1	0	0	0	5.716	5.765	6.7785	6.8175	6.573	6.601
47533	1	1	0	0	0	6.845	6.84	7.8535	7.9075	7.939	7.985
47534	1	2	1	1	2	3.623	3.663	3.4565	3.4385	3.1805	3.1585
47535	1	2	1	1	2	3.199	3.195	3.4565	3.4705	3.3865	3.412
47536	2	1	1	0	0	6.116	6.0785	6.8035	6.8245	7.0025	7.005
47537	2	1	1	0	2	6.282	6.3195	7.1415	7.016	6.95	6.846
47538	2	0	0	0	1	5.2	5.211	5.8905	5.8825	5.8165	5.847
47580	1	1	1	0	0	5.372	5.413	6.176	6.135	6.4425	6.4605
47581	2	2	0	1	2	5.187	5.1365	5.953	5.911	5.832	5.7885
47582	1	2	1	1	2	3.745	3.799	3.8175	3.8365	3.6445	3.682
47583	1	1	2	0	2	4.9275	4.945	5.2755	5.375	5.309	5.2975
47584	NA	1	0	NA	2	5.8825	5.8865	6.771	6.774	6.976	6.913
47586	2	2	1	1	2	3.171	3.1845	3.315	3.403	3.098	2.9875
47587	2	1	0	0	0	6.5245	6.5455	7.4215	7.3435	7.558	7.5495
48050	0	1	0	0	0	6.216	6.2015	7.0415	7.028	7.367	7.349
48367	2	1	1	0	1	5.7375	5.7045	6.6015	6.5755	6.5575	6.498
48368	2	1	1	0	1	5.1555	5.1815	5.8895	5.8845	5.71	5.757
48548	0	0	0	0	2	5.6715	5.688	6.249	6.2815	6.363	6.3715
48549	1	1	0	0	2	5.1215	5.1415	5.639	5.6575	5.6475	5.726
48552	1	2	0	0	1	4.705	4.733	5.381	5.448	5.3165	5.346
48580	1	1	1	1	1	3.584	3.5455	3.7905	3.7835	3.741	3.776
49307	2	2	1	1	0	5.088	4.8535	5.0065	4.8205	4.504	4.4395
49308	0	1	0	1	0	4.2335	4.264	3.822	3.926	3.1925	3.309
49876	2	1	0	1	0	5.5805	5.593	5.651	5.6905	5.598	5.643
49877	1	1	1	0	1	5.4385	5.4715	5.9065	5.91	6.077	6.1125
50095	1	0	0	1	1	4.619	4.6095	4.468	4.4695	4.323	4.3455
50109	2	1	1	1	1	4.597	4.4015	4.128	4.228	3.755	3.7445
50115	1	1	0	0	1	5.2065	5.266	5.904	5.908	5.7445	5.7875
50116	2	1	0	0	0	5.61	5.669	6.9435	6.93	6.7095	6.7295
50167	1	1	0	0	1	4.836	4.83	5.261	5.276	5.387	5.4065
50168	2	1	0	2	0	4.712	4.6125	3.982	3.9525	3.5945	3.6945
50210	1	1	0	1	2	3.869	3.878	4.3345	4.3315	4.1825	4.1915
50211	1	2	1	0	2	4.4165	4.4205	4.227	4.2385	4.4095	4.3795
50222	1	2	1	1	2	3.472	3.468	3.8015	3.7975	3.812	3.8205

```
################################################################################
# R code to:
# Analyze and plot mixed-breed dog data
################################################################################

trait <- read.table("mixed.trait.txt", header=TRUE)
names(trait) <- c("ID", "Height", "Juvenile", "Sex")
gene <- read.table("mixed.geno.txt", header=TRUE)
red.t <- subset(trait, trait$Juvenile==0 & trait$Sex==1 & trait$Height < 25)
red.g <- subset(gene, gene$ID %in% red.t$ID)

lm.sex <- lm(trait$Height~trait$Sex)
lm.b.igf1 <- lm(red.t$Height ~ factor(red.t$Sex) + factor(red.g$igf1_sine))
summary(lm.b.igf1)
lm.b.fgf4 <- lm(red.t$Height ~ factor(red.t$Sex) + factor(red.g$fgf4_retro))
summary(lm.b.fgf4)
lm.b.smad2 <- lm(red.t$Height ~ factor(red.t$Sex) + factor(red.g$smad2_del))
summary(lm.b.smad2)
lm.b.stc2 <- lm(red.t$Height ~ factor(red.t$Sex) + factor(red.g$stc2_snp))
summary(lm.b.stc2)
lm.b.banp <- lm(red.t$Height ~ factor(red.t$Sex) + factor(red.g$banp_snp))
summary(lm.b.banp)

red.g <- na.omit(red.g)
red.t <- subset(red.t, red.t$ID %in% red.g$ID)

lm.all <- lm(red.t$Height ~ factor(red.g$igf1_sine) + factor(red.g$fgf4_retro) +
               factor(red.g$smad2_del) + factor(red.g$stc2_snp))
summary(lm.all)
lm.all <- lm(red.t$Height ~ factor(red.g$fgf4_retro) + factor(red.g$smad2_del) +
               factor(red.g$stc2_snp))
summary(lm.all)
lm.all <- lm(red.t$Height ~ factor(red.g$igf1_sine) + factor(red.g$smad2_del) +
               factor(red.g$stc2_snp))
summary(lm.all)
lm.all <- lm(red.t$Height ~ factor(red.g$igf1_sine) + factor(red.g$fgf4_retro) +
               factor(red.g$stc2_snp))
summary(lm.all)
lm.all <- lm(red.t$Height ~ factor(red.g$igf1_sine) + factor(red.g$fgf4_retro) +
               factor(red.g$smad2_del))
summary(lm.all)

igf1<-factor(red.g$igf1_sine)
fgf4<-factor(red.g$fgf4_retro)
smad2del<-factor(red.g$smad2_del)
stc2<-factor(red.g$stc2_snp)
banp<-factor(red.g$banp_snp)

par(mfrow=c(3,2))
plot(red.t$Height~igf1)
points(igf1,red.t$Height)
plot(red.t$Height~fgf4)
points(fgf4,red.t$Height)
plot(red.t$Height~smad2del)
points(smad2del,red.t$Height)
plot(red.t$Height~stc2)
points(stc2,red.t$Height)
plot(red.t$Height~banp)
points(banp,red.t$Height)
```

Appendix C

Code for Chapter 3

```
##################################################################################
# R code to:
# Test enrichment of CGG, CGGG, CGGGG, and Cn motifs in PRDM9 knockout mice
# Data comes from Brick et al. 2012 and can be found at:
# http://www.nature.com/nature/journal/v485/n7400/extref/nature11089-s2.zip
##################################################################################

#The command line EMBOSS tool is required to call the wordcount function
#This script requires the BSgenome and stringr packages
install.packages("stringr", dependencies = TRUE)
source("http://bioconductor.org/biocLite.R")
biocLite("BSgenome")
library(BSgenome)
biocLite("BSgenome.Mmusculus.UCSC.mm9")
library(BSgenome.Mmusculus.UCSC.mm9)
library(stringr)

data <- read.table("2011-11-14340C-Supplementary_File_1.txt",header=T)
chr.include <-
c("chr1","chr2","chr3","chr4","chr5","chr6","chr7","chr8","chr9","chr10","chr11","
chr12","chr13","chr14","chr15","chr16","chr17","chr18","chr19")

get.DNA <- function(data, chr.include, pick.strain, top.pct) {
  output <- length(data[,1])
  print(paste(output,"rows in data file"))
  d <- subset(data,data$strain==pick.strain) #filter by mouse strain
  output <- length(d[,1])
  print(paste(output, pick.strain, "rows in data file"))
  d.hot <- subset(d,d$type=="Hotspot") #filter by Hotspot
  output <- length(d.hot[,1])
  print(paste(output, pick.strain, "hotspots in data file"))
  d.hot <- subset(d.hot, d.hot$chromosome %in% chr.include) #filter by chromosome
  output <- length(d.hot[,1])
  print(paste(output, pick.strain, "hotspots on mapped chromosomes in data file"))
  d.hot <- d.hot[,-(5:7)] #remove uninformative columns
  #split into lists of data.frames according to chromosome
  d.hot <- split(d.hot,d.hot$chromosome)
  #filter hotspots with overlapping ranges
  fix.overlaps <- function (x) {
    s <- x[,2] #starting positions
    s <- c(s,10000000000) #add big number to end of list of starts
    s <- s[-1] #remove first starting position
    #data.frame where fifth column is starting position of next hotspot
    df <- data.frame(x,s)
    #select only hotspots where end position is less than next start
    filter <- subset(df,df[,3]<df[,5])
    return(filter[,-5])
  }
  d.hot <- lapply(d.hot,fix.overlaps)
  library(plyr)
  d.hot <- ldply(d.hot, data.frame)
  d.hot <- d.hot[,-1]
  d.hot <- d.hot[order(-d.hot[,4]),]
  d.hot <- subset(d.hot,d.hot$start!=124074118)
  output <- length(d.hot[,1])
  print(paste(output, pick.strain, "non-overlapping hotspots on mapped chromosomes
in data file"))
  top <- (length(d.hot[,1]))%/%(100/top.pct)
  #top <- 10000 #Use for getting top 10,000 for 6-mers plot of PRDM9 and B6
  t.hot <- d.hot[1:top,]
  output <- length(t.hot[,1])
  print(paste(output, pick.strain, "in top", top.pct, "% of non-overlapping
hotspots on mapped chromosomes in data file"))
```

```
  b.null <- d.hot #b for before
  a.null <- d.hot #a for after
  width <- (b.null$end-b.null$start)+1
  b.null$start <- b.null$start-width
  b.null$end <- b.null$end-width
  a.null$start <- a.null$start+width
  a.null$end <- a.null$end+width
  t.b.null <- b.null[1:top,]
  t.a.null <- a.null[1:top,]
  seqs <-
getSeq(Mmusculus,t.hot[['chromosome']],start=as.integer(t.hot[['start']]),end=(as.
integer(t.hot[['end']])))
  hot.file <- paste(pick.strain,".",top.pct,".hot",sep="")
  writeXStringSet(seqs, file=hot.file, append=TRUE)
  print(paste("output file is",hot.file))
  seqs <-
getSeq(Mmusculus,t.b.null[['chromosome']],start=as.integer(t.b.null[['start']]),en
d=(as.integer(t.b.null[['end']])))
  null.b.file <- paste(pick.strain,".",top.pct,".b.null",sep="")
  writeXStringSet(seqs, file=null.b.file, append=TRUE)
  print(paste("output file is",null.b.file))
  seqs <-
getSeq(Mmusculus,t.a.null[['chromosome']],start=as.integer(t.a.null[['start']]),en
d=(as.integer(t.a.null[['end']])))
  null.a.file <- paste(pick.strain,".",top.pct,".a.null",sep="")
  writeXStringSet(seqs, file=null.a.file, append=TRUE)
  print(paste("output file is",null.a.file))
  return(list(hot.file,null.b.file,null.a.file))
}

#Modify previous code to get 10,000, not top quarter
PRDM9 <- get.DNA(data,chr.include,"PRDM9",1)
B6 <- get.DNA(data,chr.include,"B6",1)

#To get top quarter of hotspots in PRDM9 knockout and B6 mice
PRDM9 <- get.DNA(data,chr.include,"PRDM9",25)
B6 <- get.DNA(data,chr.include,"B6",25)

get.wordcounts <- function(x,wordsize,size.name) {
  #System commands using EMBOSS wordcount tool
  system2("wordcount",
paste(x[[1]],wordsize,paste(x[[1]],size.name,".wc",sep="")))
  system2("wordcount",
paste(x[[2]],wordsize,paste(x[[2]],size.name,".wc",sep="")))
  system2("wordcount",
paste(x[[3]],wordsize,paste(x[[3]],size.name,".wc",sep="")))
}

#For plotting PRDM9 and B6 word counts, take top 10,000 in each
get.wordcounts(PRDM9,"-wordsize 6 -outfile","6")
get.wordcounts(B6,"-wordsize 6 -outfile","6")

get.wordcounts(PRDM9,"-wordsize 5 -outfile","5")
get.wordcounts(PRDM9,"-wordsize 6 -outfile","6")
get.wordcounts(PRDM9,"-wordsize 7 -outfile","7")
get.wordcounts(PRDM9,"-wordsize 8 -outfile","8")
get.wordcounts(PRDM9,"-wordsize 9 -outfile","9")
get.wordcounts(PRDM9,"-wordsize 10 -outfile","10")

tabulate.wordcounts <- function(x,size.name) {
  b.null <- read.table(paste(x[[2]],size.name,".wc",sep=""))
  a.null <- read.table(paste(x[[3]],size.name,".wc",sep=""))
  names(b.null) <- c("motif","count")
  names(a.null) <- c("motif","count")
  null <- merge(b.null,a.null,by="motif")
  rm(b.null,a.null)
  null$count <- (null$count.x + null$count.y)/2
  null <- null[,-(2:3)]
  hot <- read.table(paste(x[[1]],size.name,".wc",sep=""))
  names(hot) <- c("motif","count")
  hotnull <- merge(hot,null,by="motif")
  names(hotnull) <- c("motif","hot","null")
  rm(hot)
  hotnull$ratio <- hotnull$hot/hotnull$null
  hotnull <- hotnull[order(-hotnull$ratio),]
  return(hotnull)
}

#For top 10,000 plot
PRDM9.out6 <- tabulate.wordcounts(PRDM9,"6")
B6.out6 <- tabulate.wordcounts(B6,"6")

#For top quarter analysis
PRDM9.out5 <- tabulate.wordcounts(PRDM9,"5")
PRDM9.out6 <- tabulate.wordcounts(PRDM9,"6")
PRDM9.out7 <- tabulate.wordcounts(PRDM9,"7")
PRDM9.out8 <- tabulate.wordcounts(PRDM9,"8")
PRDM9.out9 <- tabulate.wordcounts(PRDM9,"9")
PRDM9.out10 <- tabulate.wordcounts(PRDM9,"10")

m.cgg.5 <- c("CCGCC", "GCCGC", "CGCCG", "GGCGG", "CGGCG", "GCGGC")
m.cgg.6 <- c("CGGCGG", "GCGGCG", "GGCGGC", "GCCGCC", "CGCCGC", "CCGCCG")
m.cgg.7 <- c("CGGCGGC", "GCGGCGG", "GGCGGCG", "GCCGCCG", "CGCCGCC", "CCGCCGC")
m.cgg.8 <- c("CGGCGGCG", "GCGGCGGC", "GGCGGCGG", "GCCGCCGC", "CGCCGCCG",
"CCGCCGCC")
m.cgg.9 <- c("CGGCGGCGG", "GCGGCGGCG", "GGCGGCGGC", "GCCGCCGCC", "CGCCGCCGC",
"CCGCCGCCG")
m.cgg.10 <- c("CGGCGGCGGC", "GCGGCGGCGG", "GGCGGCGGCG", "GCCGCCGCCG",
"CGCCGCCGCC", "CCGCCGCCGC")

m.cggg.5 <- c("CGGGC", "GCGGG", "GGCGG", "GGGCG", "GCCCG", "CGCCC", "CCGCC",
"CCCGC")
m.cggg.6 <- c("CGGGCG", "GCGGGC", "GGCGGG", "GGGCGG", "GCCCGC", "CGCCCG",
"CCGCCC", "CCCGCC")
m.cggg.7 <- c("CGGGCGG", "GCGGGCG", "GGCGGGC", "GGGCGGG", "GCCCGCC", "CGCCCGC",
"CCGCCCG", "CCCGCCC")
m.cggg.8 <- c("CGGGCGGG", "GCGGGCGG", "GGCGGGCG", "GGGCGGGC", "GCCCGCCC",
"CGCCCGCC", "CCGCCCGC", "CCCGCCCG")
m.cggg.9 <- c("CGGGCGGGC", "GCGGGCGGG", "GGCGGGCGG", "GGGCGGGCG", "GCCCGCCCG",
"CGCCCGCCC", "CCGCCCGCC", "CCCGCCCGC")
m.cggg.10 <- c("CGGGCGGGCG", "GCGGGCGGGC", "GGCGGGCGGG", "GGGCGGGCGG",
"GCCCGCCCGC", "CGCCCGCCCG", "CCGCCCGCCC", "CCCGCCCGCC")

m.cgggg.5 <- c("CGGGG", "GCGGG", "GGCGG", "GGGCG", "GGGGC", "GCCCC", "CGCCC",
"CCGCC", "CCCGC", "CCCCG")
m.cgggg.6 <- c("CGGGGC", "GCGGGG", "GGCGGG", "GGGCGG", "GGGGCG", "GCCCCG",
"CGCCCC", "CCGCCC", "CCCGCC", "CCCCGC")
m.cgggg.7 <- c("CGGGGCG", "GCGGGGC", "GGCGGGG", "GGGCGGG", "GGGGCGG", "GCCCCGC",
"CGCCCCG", "CCGCCCC", "CCCGCCC", "CCCCGCC")
m.cgggg.8 <- c("CGGGGCGG", "GCGGGGCG", "GGCGGGGC", "GGGCGGGG", "GGGGCGGG",
"GCCCCGCC", "CGCCCCGC", "CCGCCCCG", "CCCGCCCC", "CCCCGCCC")
m.cgggg.9 <- c("CGGGGCGGG", "GCGGGGCGG", "GGCGGGGCG", "GGGCGGGGC", "GGGGCGGGG",
"GCCCCGCCC", "CGCCCCGCC", "CCGCCCCGC", "CCCGCCCCG", "CCCCGCCCC")
m.cgggg.10 <- c("CGGGGCGGGG", "GCGGGGCGGG", "GGCGGGGCGG", "GGGCGGGGCG",
"GGGGCGGGGC", "GCCCCGCCCC", "CGCCCCGCCC", "CCGCCCCGCC", "CCCGCCCCGC",
"CCCCGCCCCG")
```

```
m.c.5 <- c("CCCCC", "GGGGG")
m.c.6 <- c("CCCCCC", "GGGGGG")
m.c.7 <- c("CCCCCCC", "GGGGGGG")
m.c.8 <- c("CCCCCCCC", "GGGGGGGG")
m.c.9 <- c("CCCCCCCCC", "GGGGGGGGG")
m.c.10 <- c("CCCCCCCCCC", "GGGGGGGGGG")

summarize.motifs <- function(ranks, motifs) {
  get.rank <- function(mot, ran) {
    which(ran$mot==mot)
  }
  counts <- sapply(motifs, get.rank, ranks)
  return(c(min(counts), mean(counts)))
}
s.cgg.5 <- summarize.motifs(PRDM9.out5, m.cgg.5)
s.cgg.6 <- summarize.motifs(PRDM9.out6, m.cgg.6)
s.cgg.7 <- summarize.motifs(PRDM9.out7, m.cgg.7)
s.cgg.8 <- summarize.motifs(PRDM9.out8, m.cgg.8)
s.cgg.9 <- summarize.motifs(PRDM9.out9, m.cgg.9)
s.cgg.10 <- summarize.motifs(PRDM9.out10, m.cgg.10)

s.cggg.5 <- summarize.motifs(PRDM9.out5, m.cggg.5)
s.cggg.6 <- summarize.motifs(PRDM9.out6, m.cggg.6)
s.cggg.7 <- summarize.motifs(PRDM9.out7, m.cggg.7)
s.cggg.8 <- summarize.motifs(PRDM9.out8, m.cggg.8)
s.cggg.9 <- summarize.motifs(PRDM9.out9, m.cggg.9)
s.cggg.10 <- summarize.motifs(PRDM9.out10, m.cggg.10)

s.cgggg.5 <- summarize.motifs(PRDM9.out5, m.cgggg.5)
s.cgggg.6 <- summarize.motifs(PRDM9.out6, m.cgggg.6)
s.cgggg.7 <- summarize.motifs(PRDM9.out7, m.cgggg.7)
s.cgggg.8 <- summarize.motifs(PRDM9.out8, m.cgggg.8)
s.cgggg.9 <- summarize.motifs(PRDM9.out9, m.cgggg.9)
s.cgggg.10 <- summarize.motifs(PRDM9.out10, m.cgggg.10)

s.c.5 <- summarize.motifs(PRDM9.out5, m.c.5)
s.c.6 <- summarize.motifs(PRDM9.out6, m.c.6)
s.c.7 <- summarize.motifs(PRDM9.out7, m.c.7)
s.c.8 <- summarize.motifs(PRDM9.out8, m.c.8)
s.c.9 <- summarize.motifs(PRDM9.out9, m.c.9)
s.c.10 <- summarize.motifs(PRDM9.out10, m.c.10)

s.cgg.5
s.cgg.6
s.cgg.7
s.cgg.8
s.cgg.9
s.cgg.10

s.cggg.5
s.cggg.6
s.cggg.7
s.cggg.8
s.cggg.9
s.cggg.10

s.cgggg.5
s.cgggg.6
s.cgggg.7
s.cgggg.8
s.cgggg.9
s.cgggg.10
```

```
s.c.5
s.c.6
s.c.7
s.c.8
s.c.9
s.c.10

(1024-s.cgg.5[2])/1024
(4096-s.cgg.6[2])/4096
(16384-s.cgg.7[2])/16384
(65536-s.cgg.8[2])/65536
(262144-s.cgg.9[2])/262144
(1048576-s.cgg.10[2])/1048576

(1024-s.cggg.5[2])/1024
(4096-s.cggg.6[2])/4096
(16384-s.cggg.7[2])/16384
(65536-s.cggg.8[2])/65536
(262144-s.cggg.9[2])/262144
(1048576-s.cggg.10[2])/1048576

(1024-s.cgggg.5[2])/1024
(4096-s.cgggg.6[2])/4096
(16384-s.cgggg.7[2])/16384
(65536-s.cgggg.8[2])/65536
(262144-s.cgggg.9[2])/262144
(1048576-s.cgggg.10[2])/1048576

(1024-s.c.5[2])/1024
(4096-s.c.6[2])/4096
(16384-s.c.7[2])/16384
(65536-s.c.8[2])/65536
(262144-s.c.9[2])/262144
(1048576-s.c.10[2])/1048576

(0.9903971 + 0.9958903 + 0.9986877 + 0.9978994 + 0.995739 + 0.9982141) / 6 #CGG
(0.980957 + 0.9836121 + 0.9873962 + 0.9915237 + 0.9922447 + 0.9971991) / 6 #CGGG
(0.9589844 + 0.9793701 + 0.9813293 + 0.9799347 + 0.9841087 + 0.9960034) / 6 #CGGGG
(0.7797852 + 0.7185059 + 0.6652527 + 0.6396866 + .6541328 + 0.7737193) / 6 #Cn
```

```
################################################################################
# execute.R
# R code to:
# Perform sliding window analysis for CGG, CGGG, CGGGG, Cn, and predicted human
# PRDM9 motifs for PRDM9 knockout and B6 mice
# Data comes from Brick et al. 2012 and can be found at:
# http://www.nature.com/nature/journal/v485/n7400/extref/nature11089-s2.zip
################################################################################

source("quartile.window.R")

#Perfect CGG motif
dict.cgg.p <-
(c("CGGCGGCGG","GGCGGCGGC","GCGGCGGCG","GCCGCCGCC","CCGCCGCCG","CGCCGCCGC"))
pdict.cgg.p<-PDict(dict.cgg.p)

#Imperfect CGG motif with one mismatch
dict.cgg.i<-
c("NGGCGGCGG","CNGCGGCGG","CGNCGGCGG","CGGCGGNGG","CGGCGGCNG","CGGCGGCGN",
        "NGCGGCGGC","GNCGGCGGC","GGNGGCGGC","GGCGGCNGC","GGCGGCGNC","GGCGGCGGN",
        "NCGGCGGCG","GNGGCGGCG","GCNGCGGCG","GCGGCGNCG","GCGGCGGNG","GCGGCGGCN",
        "NCCGCCGCC","GNCGCCGCC","GCNGCCGCC","GCCGCCNCC","GCCGCCGNC","GCCGCCGCN",
        "NCGCCGCCG","CNGCCGCCG","CCNCCGCCG","CCGCCGNCG","CCGCCGCNG","CCGCCGCCN",
        "NGCCGCCGC","CNCCGCCGC","CGNCGCCGC","CGCCGCNGC","CGCCGCCNC","CGCCGCCGN")
pdict.cgg.i<-PDict(dict.cgg.i,tb.start=4,tb.width=3)

#Perfect CGGG motif
dict.cggg.p<-
(c("CGGGCGGG","GCGGGCGG","GGCGGGCG","GGGCGGGC","GCCCGCCC","CGCCCGCC","CCGCCCGC","C
CCGCCCG"))
pdict.cggg.p<-PDict(dict.cggg.p)

#Perfect CGGGG motif
dict.cgggg.p<-
(c("CGGGGCGGGG","GCGGGGCGGG","GGCGGGGCGG","GGGCGGGGCG","GGGGCGGGGC","GCCCCGCCCC","
CGCCCCGCCC","CCGCCCCGCC","CCCGCCCCGC","CCCCGCCCCG"))
pdict.cgggg.p<-PDict(dict.cgggg.p)

#Perfect C motif
dict.c.p<-(c("CCCCCCCCCC","GGGGGGGGGG"))
pdict.c.p<-PDict(dict.c.p)

#Human PRDM9 motif
dict.hum<-
c("ccnccatanccnc","ccnccattnccnc","ccnccatcnccnc","ccnccatgnccnc","ccnccttanccnc",
"ccnccttttnccnc","ccnccttcnccnc","ccnccttgnccnc","ccnccctanccnc","ccnccccttnccnc",
"ccnccctcnccnc","ccnccctgnccnc","ccnccgtanccnc","ccnccgttnccnc","ccnccgtcnccnc",
"ccnccgtgnccnc","ggnggtatnggng","ggnggtaanggng","ggnggtagnggng","ggnggtacnggng",
"ggnggaatnggng","ggnggaaanggng","ggnggaagnggng","ggnggaacnggng","ggngggatnggng",
"ggngggaanggng","ggngggagnggng","ggngggacnggng","ggnggcatnggng","ggnggcaanggng",
"ggnggcagnggng","ggnggcacnggng")
pdict.hum<-PDict(dict.hum,tb.start=6,tb.width=3)

#Predicted human PRDM9 DNA binding motif
PRDM9.hum <- quartile.window(data, "PRDM9", chr.include, pdict.hum)
B6.hum <- quartile.window(data, "B6", chr.include, pdict.hum)

prdm9Palette <- c("#253494", "#2C7FB8", "#41B6C4", "#A1DAB4")
ggplot(PRDM9.hum[[2]], aes(position,enrichment,colour=Quartile)) +
geom_line(lwd=2) + scale_color_manual(values=prdm9Palette)
ggplot(B6.hum[[2]], aes(position,enrichment,colour=Quartile)) + geom_line(lwd=2) +
  scale_color_manual(values=prdm9Palette)
```

```
ggplot(rbind(PRDM9.hum[[1]],R9.hum[[1]],R13.hum[[1]]),
aes(position,enrichment,colour=Strain)) + geom_line(lwd=2) +
  scale_color_manual(values=prdm9Palette)
#Perfect CGG motif
PRDM9.cgg.p <- quartile.window(data, "PRDM9", chr.include, pdict.cgg.p)
B6.cgg.p <- quartile.window(data, "B6", chr.include, pdict.cgg.p)

#Perfect CGGG motif
PRDM9.cggg.p <- quartile.window(data, "PRDM9", chr.include, pdict.cggg.p)
B6.cggg.p <- quartile.window(data, "B6", chr.include, pdict.cggg.p)

PRDM9.cgggg.p <- quartile.window(data, "PRDM9", chr.include, pdict.cgggg.p)
#perfect CGGGG
B6.cgggg.p <- quartile.window(data, "B6", chr.include, pdict.cgggg.p) #perfect
CGGGG
R9.cgggg.p <- quartile.window(data, "9R", chr.include, pdict.cgggg.p) #perfect
CGGGG
R13.cgggg.p <- quartile.window(data, "13R", chr.include, pdict.cgggg.p) #perfect
CGGGG

#Perfect Cn motif
PRDM9.c.p <- quartile.window(data, "PRDM9", chr.include, pdict.c.p)
B6.c.p <- quartile.window(data, "B6", chr.include, pdict.c.p)

#Imperfect CGG motif
PRDM9.cgg.i <- quartile.window(data, "PRDM9", chr.include, pdict.cgg.i)
B6.cgg.i <- quartile.window(data, "B6", chr.include, pdict.cgg.i)

PRDM9.c <- PRDM9.c.p[[1]]
PRDM9.c$Strain <- "C (knockout)"
PRDM9.cgg <- PRDM9.cgg.p[[1]]
PRDM9.cgg$Strain <- "CGG (knockout)"
PRDM9.cggg <- PRDM9.cggg.p[[1]]
PRDM9.cggg$Strain <- "CGGG (knockout)"
PRDM9.cgggg <- PRDM9.cgggg.p[[1]]
PRDM9.cgggg$Strain <- "CGGGG (knockout)"
B6.c <- B6.c.p[[1]]
B6.c$Strain <- "C (B6)"
B6.cgg <- B6.cgg.p[[1]]
B6.cgg$Strain <- "CGG (B6)"
B6.cggg <- B6.cggg.p[[1]]
B6.cggg$Strain <- "CGGG (B6)"
B6.cgggg <- B6.cgggg.p[[1]]
B6.cgggg$Strain <- "CGGGG (B6)"
comp1 <- rbind(PRDM9.c, PRDM9.cgg, PRDM9.cggg, PRDM9.cgggg, B6.c, B6.cgg, B6.cggg,
B6.cgggg)
names(comp1) <- c("position", "enrichment", "Motif")

prdm9Palette <- c("#D7B5D8", "#A1DAB4", "#DF65B0", "#41B6C4", "#DF65B0",
"#2C7FB8", "#980043", "#253494")
ggplot(comp1, aes(position, enrichment, colour= Motif)) + geom_line(lwd=2) +
  scale_color_manual(values=prdm9Palette) + theme_classic() +
  xlab("Distance to hotspot center (kb)") +
  ylab("Fold enrichment") +
  scale_x_continuous(breaks=c(-2000,-1000,0,1000,2000), labels=c(-2,-1,0,1,2)) +
  theme(legend.position = c(0.8,0.8))

PRDM9.cggi <- PRDM9.cgg.i[[1]]
PRDM9.cggi$Strain <- "CGG mismatch (knockout)"
B6.cggi <- B6.cgg.i[[1]]
B6.cggi$Strain <- "CGG mismatch (B6)"

prdm9Palette <- c("#DF65B0", "#41B6C4", "#980043", "#253494")
```

```
ggplot(rbind(PRDM9.cgg, PRDM9.cggi, B6.cgg, B6.cggi),
aes(position,enrichment,colour=Strain)) + geom_line(lwd=2) +
  scale_color_manual(values=prdm9Palette) + theme_classic() +
  xlab("Distance to hotspot center (kb)") +
  ylab("Fold enrichment") +
  scale_x_continuous(breaks=c(-2000,-1000,0,1000,2000), labels=c(-2,-1,0,1,2)) +
  theme(legend.position = c(0.8,0.8))
```

```
################################################################################
# quartile.window.R
# Generates sliding window figures of motifs specified in execute.R
################################################################################

quartile.window<-function(data,pick.strain,chr.include,pdict) {
  #This script requires the BSgenome and stringr packages
  #install.packages("stringr", dependencies = TRUE)
  #source("http://bioconductor.org/biocLite.R")
  #biocLite("BSgenome")
  library(BSgenome)
  #biocLite("BSgenome.Mmusculus.UCSC.mm9")
  library(BSgenome.Mmusculus.UCSC.mm9)
  library(stringr)
  source("get.DNA.R")
  print("Getting DNA sequences from genomic coordinates...")
  s.dna <- get.DNA(data,chr.include,pick.strain)
  s.length <- length(s.dna)
  quartile <- s.length%/%4
  source("get.matches.R")
  print("Matching pdict motifs to DNA sequences...")
  a.matches <- get.matches(s.dna,pdict)
  q1.matches <- get.matches(s.dna[1:quartile],pdict)
  q2.matches <- get.matches(s.dna[(quartile+1):(quartile*2)],pdict)
  q3.matches <- get.matches(s.dna[((quartile*2)+1):(quartile*3)],pdict)
  q4.matches <- get.matches(s.dna[((quartile*3)+1):(quartile*4)],pdict)
  source("get.coverage.R")
  print("Tabulating matched motifs...")
  a.cov <- get.coverage(a.matches)
  q1.cov <- get.coverage(q1.matches)
  q2.cov <- get.coverage(q2.matches)
  q3.cov <- get.coverage(q3.matches)
  q4.cov <- get.coverage(q4.matches)
  source("get.windows.R")
  print("Getting window counts...")
  a.win <- get.windows(q1.cov,"Strain",pick.strain)
  q1.win <- get.windows(q1.cov,"Quartile","1")
  q2.win <- get.windows(q2.cov,"Quartile","2")
  q3.win <- get.windows(q3.cov,"Quartile","3")
  q4.win <- get.windows(q4.cov,"Quartile","4")
  a.win$position <- a.win$position-2500
  q.plot <- rbind(q1.win,q2.win,q3.win,q4.win)
  q.plot$position <- q.plot$position-2500
  return(list(a.win,q.plot))
}
```

```
################################################################################
# get.DNA.R
# Captures DNA sequences associated with hotspots (MM9)
################################################################################

get.DNA <- function(data, chr.include, pick.strain) {
  output <- length(data[,1])
  print(paste(output,"rows in data file"))
  d <- subset(data,data$strain==pick.strain) #filter by mouse strain
  output <- length(d[,1])
  print(paste(output, pick.strain, "rows in data file"))
  d.hot <- subset(d,d$type=="Hotspot") #filter by Hotspot
  output <- length(d.hot[,1])
  print(paste(output, pick.strain, "hotspots in data file"))
  d.hot <- subset(d.hot, d.hot$chromosome %in% chr.include) #filter by chromosome
  output <- length(d.hot[,1])
  print(paste(output, pick.strain, "hotspots on chr.include chromosomes in data
file"))
  d.hot <- subset(d.hot,(d.hot$end-d.hot$start)==2000) #filter by hotspot size
  output <- length(d.hot[,1])
  print(paste(output, pick.strain, "2000 bp hotspots on chr.include chromosomes in
data file"))
  #set beginning of range for sliding window analysis
  d.hot$start <- d.hot$start-4100
  d.hot$end <- d.hot$end+4099 #set end
  d.hot <- d.hot[,-(5:7)] #remove uninformative columns
  #split into lists of data.frames according to chromosome
  d.hot <- split(d.hot,d.hot$chromosome)
  #filter out hotspots with overlapping ranges that would interfere with estimates
of hot and null
  fix.overlaps <- function (x) {
    s <- x[,2] #starting positions
    s <- c(s,10000000000) #add big number to end of list of starts
    s <- s[-1] #remove first starting position
    df <- data.frame(x,s) #data.frame where fifth column is starting position of
next hotspot
    #select only hotspots where end position is less than next start
    filter <- subset(df,df[,3]<df[,5])
    return(filter[,-5])
  }
  d.hot <- lapply(d.hot,fix.overlaps)
  library(plyr)
  d.hot <- ldply(d.hot, data.frame)
  d.hot <- d.hot[,-1]
  d.hot <- d.hot[order(-d.hot[,4]),]
  d.hot <- subset(d.hot,d.hot$end!=124080217)
  output <- length(d.hot[,1])
  print(paste(output, pick.strain, "non-overlapping hotspots on chr.include
chromosomes in data file"))
  d <- d.hot
  library(BSgenome)
  library(BSgenome.Mmusculus.UCSC.mm9)
  library(stringr)
  seqs <-
getSeq(Mmusculus,d[['chromosome']],start=as.integer(d[['start']]),end=(as.integer(
d[['end']])))
  return(seqs)
}
```

```
################################################################################
# get.matches.R
# Matches motifs specified in execute.R with those contained in DNA extracted
# with get.DNA.R
################################################################################

get.matches <- function(dna,pdict) {
  output <- vector(mode="list",length(dna))
  for(i in 1:(length(dna))) {
    output[i] <- matchPDict(pdict,dna[[i]],fixed="subject")
  }
  return(output)
}
```

```
################################################################################
# get.coverage.R
# Computes coverage for matched motifs from get.matches.R
################################################################################

get.coverage <- function(matches) {
  spots <- length(matches)
  cov.total <- seq(1:10200)
  for(i in 1:spots) {
    cov <- as.numeric(coverage(matches[[i]]))
    cov <- which(cov>0)
    cov.total <- c(cov.total,cov)
  }
  cov.y <- tabulate(cov.total)
  cov.x <- seq(1:length(cov.y))
  cov.xy <- data.frame(cov.x,cov.y)
  return(cov.xy)
}
```

```
################################################################################
# get.windows.R
# Performs sliding window analysis from coverage calculated in get.coverage.R
################################################################################

get.windows <- function(cov,category.type,category) {
  windows <- vector(mode="list",length=10000)
  for(i in 1:10000) {
    windows[[i]] <- sum(cov[i:(i+199),2])
  }
  win.x <- seq(1:5000)
  win.y <- ldply(windows)
  hot <- win.y[2501:7500,]
  null <- win.y[c(1:2500,7501:10000),]
  null.num <- (sum(null))/5000
  enrichment <- hot/null.num
  output <- data.frame(win.x,enrichment,category)
  names(output) <- c("position","enrichment",category.type)
  return(output)
}
```

References

Abdulrazzaq, Y.M., Moussa, M.A., and Nagelkerke, N. (2008). National growth charts for the United Arab Emirates. J. Epidemiol. Jpn. Epidemiol. Assoc. *18*, 295–303.

Aboyoun, P., Pages, H., and Lawrence, M. GenomicRanges: Representation and manipulation of genomic intervals. R Package Version 1106.

Akey, J.M., Ruhe, A.L., Akey, D.T., Wong, A.K., Connelly, C.F., Madeoy, J., Nicholas, T.J., and Neff, M.W. (2010). Tracking footprints of artificial selection in the dog genome. Proc. Natl. Acad. Sci. U. S. A. *107*, 1160–1165.

Alderton, D. (2008). Encyclopedia of Dogs (Parragon Books Ltd).

Allbrook, D. (1961). The estimation of stature in British and East African males. Based on tibial and ulnar bone lengths. J. Forensic Med. *8*, 15–28.

Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., and Kalchman, M.A. (1993). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington’s disease. Nat. Genet. *4*, 398–403.

Antón, S.C. (2003). Natural history of Homo erectus. Am. J. Phys. Anthropol. *Suppl 37*, 126–170.

Australian Bureau of Statistics (1998). How Australians Measure Up.

Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J.K., Hayward, J.J., Cohen, P.E., Greally, J.M., Wang, J., et al. (2013). Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLoS Genet *9*, e1003984.

Axelsson, E., Webster, M.T., Ratnakumar, A., Ponting, C.P., and Lindblad-Toh, K. (2012). Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. Genome Res. *22*, 51–63.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M.T., Perloski, M., Liberg, O., Arnemo, J.M., Hedhammar, Å., and Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature *495*, 360–364.

Barsh, G.S., Copenhaver, G.P., Gibson, G., and Williams, S.M. (2012). Guidelines for genome-wide association studies. PLoS Genet. *8*, e1002812.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and Massy, B. de (2010). PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science *327*, 836–840.

Bellus, G.A., Hefferon, T.W., Ortiz de Luna, R.I., Hecht, J.T., Horton, W.A., Machado, M., Kaitila, I., McIntosh, I., and Francomano, C.A. (1995). Achondroplasia is defined by recurrent G380R mutations of FGFR3. Am. J. Hum. Genet. *56*, 368–373.

Berg, I.L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N.J., and Jeffreys, A.J. (2011). Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. Proc. Natl. Acad. Sci. *108*, 12378–12383.

Bergelson, J., and Roux, F. (2010). Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. Nat. Rev. Genet. *11*, 867–879.

Blue, L. (2008). Why Are People Taller Today Than Yesterday? Time.

Bogin, B. (1999a). Patterns of Human Growth (Cambridge University Press).

Bogin, B. (1999b). Evolutionary Perspective on Human Growth. Annu. Rev. Anthropol. *28*, 109–153.

Boyko, A.R., Boyko, R.H., Boyko, C.M., Parker, H.G., Castelhano, M., Corey, L., Degenhardt, J.D., Auton, A., Hedimbi, M., Kityo, R., et al. (2009). Complex population structure in African village dogs and its implications for inferring dog domestication history. Proc. Natl. Acad. Sci. U. S. A. *106*, 13903–13908.

Boyko, A.R., Quignon, P., Li, L., Schoenebeck, J.J., Degenhardt, J.D., Lohmueller, K.E., Zhao, K., Brisbin, A., Parker, H.G., VonHoldt, B.M., et al. (2010). A simple genetic architecture underlies morphological variation in dogs. PLoS Biol. *8*, e1000451.

Brachi, B., Morris, G.P., and Borevitz, J.O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol. *12*, 232.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics *23*, 2633–2635.

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R.D., and Petukhova, G. V (2012). Genetic recombination is directed away from functional genomic elements in mice. Nature *485*, 642–645.

Buard, J., Barthès, P., Grey, C., and de Massy, B. (2009). Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. EMBO J. *28*, 2616–2624.

Callaway, E. (2014). Modern genomes reveal our inner Neanderthal. Nature.

Cavelaars, A.E., Kunst, A.E., Geurts, J.J., Crialesi, R., Grötvedt, L., Helmert, U., Lahelma, E., Lundberg, O., Mielck, A., Rasmussen, N.K., et al. (2000). Persistent variations in average height between countries and between socio-economic groups: an overview of 10 European countries. Ann. Hum. Biol. *27*, 407–421.

Charlesworth, B., and Charlesworth, D. (2010). Elements of evolutionary genetics (Greenwood Village, Colo.: Roberts and Co. Publishers).

Chase, K., Carrier, D.R., Adler, F.R., Jarvik, T., Ostrander, E.A., Lorentzen, T.D., and Lark, K.G. (2002). Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. Proc. Natl. Acad. Sci. U. S. A. *99*, 9930–9935.

Chase, K., Carrier, D.R., Adler, F.R., Ostrander, E.A., and Lark, K.G. (2005). Interaction between the X chromosome and an autosome regulates size sexual dimorphism in Portuguese Water Dogs. Genome Res. *15*, 1820–1824.

Chen, G. (2012). TGF-β and BMP Signaling in Osteoblast Differentiation and Bone Formation. Int. J. Biol. Sci. 272–288.

Chile (2010). Encuesta Nacional de Salud ENS Chile.

Cochran, G., and Harpending, H. (2009). The 10,000 Year Explosion: How Civilization Accelerated Human Evolution (Basic Books).

Cohen, P. (2011). Robert W. Fogel Investigates Human Evolution. N. Y. Times.

Cohen, P.E., and Pollard, J.W. (2001). Regulation of meiotic recombination and prophase I progression in mammals. BioEssays *23*, 996–1009.

Complex Trait Consortium (2003). The nature and identification of quantitative trait loci: a community's view. Nat. Rev. Genet. *4*, 911–916.

Connallon, T., and Clark, A.G. (2014). Evolutionary inevitability of sexual antagonism. Proc. R. Soc. B Biol. Sci. *281*, 20132123.

Connor Gorber, S., Shields, M., and Tremblay, M.S. (2008). Methodological issues in anthropometry: Self-reported versus measured height and weight.

Conte, G.L., Arnegard, M.E., Peichel, C.L., and Schluter, D. (2012). The probability of genetic parallelism and convergence in natural populations. Proc. R. Soc. B Biol. Sci. *279*, 5039–5047.

Corbett, J., Given, L., Gray, L., Leyland, A., MacGregor, A., Marryat, L., Miller, M., and Reid, S. (2009). The Scottish Health Survey 2008.

Delson, E., and Harvati, K. (2006). Palaeoanthropology: Return of the last Neanderthal. Nature *443*, 762–763.

Dettwyler, K.A. (1992). Nutritional status of adults in rural Mali. Am. J. Phys. Anthropol. *88*, 309–321.

Deurenberg, P., Bhaskaran, K., and Lian, P.L.K. (2003). Singaporean Chinese adolescents have more subcutaneous adipose tissue than Dutch Caucasians of the same age and body mass index. Asia Pac. J. Clin. Nutr. *12*, 261–265.

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. Nature *456*, 728–731.

Dubois, E. (1894). Pithecanthropus erectus: eine menschenaehnlich Uebergangsform aus Java (Batavia: Landsdrukerei).

Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. *10*, 285–311.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinforma. Oxf. Engl. *21*, 3439–3440.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Dusko Bjelica, S.P. (2012). Body Height and Its Estimation Utilizing Arm Span Measurements in Montenegrin Adults. Anthropol. Noteb. *18*, 69–83.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.

El-Zanaty, F., and Way, A. (2009). Egypt Demographic and Health Survey.

Eriksson, A., and Manica, A. (2012). Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. *109*, 13956–13960.

Fairbairn, D.J. (1997). Allometry for sexual size dimorphism: Pattern and Process in the Coevolution of Body Size in Males and Females. Annu. Rev. Ecol. Syst. *28*, 659–687.

Ferguson, D.O., and Alt, F.W. (2001). DNA double strand break repair and chromosomal translocation: Lessons from animal models. Oncogene *20*, 5572.

Fisher, R.A. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Trans. R. Soc. Edinb. *52*, 399–433.

Fitch, W.M. (1967). Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. J. Mol. Biol. *26*, 499–507.

Flachs, P., Mihola, O., Šimeček, P., Gregorová, S., Schimenti, J.C., Matsui, Y., Baudat, F., de Massy, B., Piálek, J., Forejt, J., et al. (2012). Interallelic and Intergenic Incompatibilities of the Prdm9 (Hst1) Gene in Mouse Hybrid Sterility. PLoS Genet *8*, e1003044.

Flint, J., and Mackay, T.F.C. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res. *19*, 723–733.

Fondon, J.W., and Garner, H.R. (2004). Molecular origins of rapid and continuous morphological evolution. Proc. Natl. Acad. Sci. *101*, 18058–18063.

Food and Nutrition Research Institute (2003). Philippine Facts and Figures 2003: Anthropometric.

France (2006). Campagne Nationale de Mensuration.

Frankenberg, E., and Jones, N.R. (2003). Self-Rated Health and Mortality: Does the Relationship Extend to a Low Income Setting?

Frey, U., Stormer, C., and Willfuhr, K. (2010). Homo Novus - A Human Without Illusions.

Furness, H. (2013). Sir David Attenborough: Humans have stopped evolving. Telegraph.co.uk.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics *159*, 907–911.

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. J. Anthropol. Inst. G. B. Irel. *15*, 246–263.

Garcia, J., and Quintana-Domeque, C. (2007). The evolution of adult height in Europe: A brief note. Econ. Hum. Biol. *5*, 340–349.

Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. Annu. Rev. Genet. *44*, 445–477.

Germonpré, M., Sablin, M.V., Stevens, R.E., Hedges, R.E.M., Hofreiter, M., Stiller, M., and Després, V.R. (2009). Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. J. Archaeol. Sci. *36*, 473–490.

Gerton, J.L., DeRisi, J., Shroff, R., Lichten, M., Brown, P.O., and Petes, T.D. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. *97*, 11383–11390.

Gibbons, A. (2010). Human Ancestor Caught in the Midst of a Makeover. Science *328*, 413–413.

Giraud, A., Radman, M., Matic, I., and Taddei, F. (2001). The rise and fall of mutator bacteria. Curr. Opin. Microbiol. *4*, 582–585.

Gray, M.M., Sutter, N.B., Ostrander, E.A., and Wayne, R.K. (2010). The IGF1 small dog haplotype is derived from Middle Eastern grey wolves. BMC Biol. *8*, 16.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A Draft Sequence of the Neandertal Genome. Science *328*, 710–722.

Greyhound, I. (2014). Breeding for proper size.

Groeneveld, L.F., Atencia, R., Garriga, R.M., and Vigilant, L. (2012). High Diversity at PRDM9 in Chimpanzees and Bonobos. PLoS ONE *7*, e39064.

Grueter, C.E., van Rooij, E., Johnson, B.A., DeLeon, S.M., Sutherland, L.B., Qi, X., Gautron, L., Elmquist, J.K., Bassel-Duby, R., and Olson, E.N. (2012). A Cardiac MicroRNA Governs Systemic Energy Homeostasis by Regulation of MED13. Cell *149*, 671–683.

Guernsey,, M.W., Ritscher,, L., Miller,, M.A., Smith,, D.A., Schöneberg,, T., and Shapiro,, M.D. (2013). A Val85Met Mutation in Melanocortin-1 Receptor Is Associated with Reductions in Eumelanic Pigmentation and Cell Surface Expression in Domestic Rock Pigeons (*Columba livia*). PLoS ONE *8*, e74475.

Gustafsson, A., and Lindenfors, P. (2004). Human size evolution: no evolutionary allometric relationship between male and female stature. J. Hum. Evol. *47*, 253–266.

Haber, J.E. (2000). Partners and pathways: repairing a double-strand break. Trends Genet. *16*, 259–264.

Haghdoost, A.A., Mirzazadeh, A., and Alikhani, S. (2008). Secular Trend of Height Variations in Iranian Population Born between 1940 and 1984. Iran. J. Public Health *37*, 1–7.

Hansen, T.F. (2006). The Evolution of Genetic Architecture. Annu. Rev. Ecol. Evol. Syst. *37*.

Hardy, O.J., and Vekemans, X. (2002). spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes *2*, 618–620.

Hare, B., and Woods, D. 20090-8199 U. (2013a). We Didn’t Domesticate Dogs. They Domesticated Us.

Hare, B., and Woods, V. (2013b). The Genius of Dogs: How Dogs Are Smarter than You Think (Dutton Adult).

Hartl, D.L., and Clark, A.G. (1997). Principles of population genetics (Sinauer associates Sunderland).

Hawks, J. (2005). Body mass in ancient humans and high latitude populations.

Hawks, J. (2010). Shrinking erectus.

Helsedirektoratet (2009). Fysisk aktivitet blant voksne og eldre i Norge Resultater fra en kartlegging i 2008 og 2009.

Herpin, N. (2003). La taille des hommes: son incidence sur la vie en couple et la carrière professionnelle. Économie Stat. *361*, 71–90.

Hindorff, L.A., MacArthur, J., Wise, A., Junkins, H.A., Hall, P.N., Klemm, A.K., and Manolio, T.A. (2013). A Catalog of Published Genome-Wide Association Studies.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.

Holliday, T.W. (1997). Body proportions in Late Pleistocene Europe and modern human origins. J. Hum. Evol. *32*, 423–447.

Hoopes, B.C., Rimbault, M., Liebers, D., Ostrander, E.A., and Sutter, N.B. (2012). The insulin-like growth factor 1 receptor (IGF1R) contributes to reduced size in dogs. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *23*, 780–790.

Huang, D.W., Sherman, B.T., and Lempicki, R. a (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huang, D.W., Sherman, B.T., and Lempicki, R. a (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Van Hung, M., and Sunyoung, P. (2008). The impact of environment on morphological and physical indexes of Vietnamese and South Korean students. VNU J. Sci. Nat. Sci. Technol. 50–55.

Instituto Brasileiro de Geografi a e Estatística (2010). Antropometria e Estado Nutricional de Criancas, Adolescentes e Adultos no Brasil.

Istat (2011). Italia in cifre.

Jaeger, U., Bruchhaus, H., Finke, L., Kromeyer-Hauschild, K., and Zellner, K. (1998). [Secular trend in body height since the Neolithic period]. Anthropol. Anz. Ber. Über Biol.-Anthropol. Lit. *56*, 117–130.

Japan (2011). Official Statistics by Ministry of Education, Culture, Sports, Science and Technology.

Johannes, F., Porcher, E., Teixeira, F.K., Saliba-Colombani, V., Simon, M., Agier, N., Bulski, A., Albuisson, J., Heredia, F., Audigier, P., et al. (2009). Assessing the Impact of Transgenerational Epigenetic Variation on Complex Traits. PLoS Genet *5*, e1000530.

Johnson, R.C., Nelson, G.W., Troyer, J.L., Lautenberger, J.A., Kessing, B.D., Winkler, C.A., and O'Brien, S.J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics *11*, 724.

Jones, P., Chase, K., Martin, A., Davern, P., Ostrander, E.A., and Lark, K.G. (2008). Single-Nucleotide-Polymorphism-Based Association Mapping of Dog Stereotypes. Genetics *179*, 1033–1044.

Jordan, S., Lim, L., Seubsman, S.-A., Bain, C., Sleigh, A., and Thai Cohort Study Team (2012). Secular changes and predictors of adult height for 86 105 male and female members of the Thai Cohort Study born between 1940 and 1990. J. Epidemiol. Community Health *66*, 75–80.

Jureša, V., Musil, V., and Kujundžić Tiljak, M. (2012). Growth Charts for Croatian School Children and Secular Trends in Past Twenty Years. Coll. Antropol. *36 supplement 1*, 47–57.

Kamadjeu, R.M., Edwards, R., Atanga, J.S., Kiawi, E.C., Unwin, N., and Mbanya, J.-C. (2006). Anthropometry measures and prevalence of obesity in the urban adult population of Cameroon: an update from the Cameroon Burden of Diabetes Baseline Survey. BMC Public Health *6*, 228.

Kashi, Y., and King, D.G. (2006a). Simple sequence repeats as advantageous mutators in evolution. Trends Genet. *22*, 253–259.

Kashi, Y., and King, D.G. (2006b). Has Simple Sequence Repeat Mutability Been Selected to Facilitate Evolution? Isr. J. Ecol. Evol. *52*, 331–342.

Kashi, Y., King, D., and Soller, M. (1997). Simple sequence repeats as a source of quantitative genetic variation. Trends Genet. *13*, 74–78.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. Database *2011*, bar049–bar049.

Katz, M., Amit, I., and Yarden, Y. (2007). Regulation of MAPKs by growth factors and receptor tyrosine kinases. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1773*, 1161–1176.

Katzmarzyk, P.T., and Leonard, W.R. (1998). Climatic influences on human body size and proportions: ecological adaptations and secular trends. Am. J. Phys. Anthropol. *106*, 483–503.

Keeney, S. (2001). Mechanism and control of meiotic recombination initiation. In Current Topics in Developmental Biology, (Academic Press), pp. 1–53.

Keeney, S., Giroux, C.N., and Kleckner, N. (1997). Meiosis-Specific DNA Double-Strand Breaks Are Catalyzed by Spo11, a Member of a Widely Conserved Protein Family. Cell *88*, 375–384.

Kemper, K.E., Visscher, P.M., and Goddard, M.E. (2012). Genetic architecture of body size in mammals. Genome Biol. *13*, 244.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, and D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

Khanna, K.K., and Jackson, S.P. (2001). DNA double-strand breaks: signaling, repair and the cancer connection. Nat. Genet. *27*, 247–254.

King, D.G. (2012). Evolution of Simple Sequence Repeats as Mutable Sites. In Tandem Repeat Polymorphisms, A.J.H. Phd, ed. (Springer New York), pp. 10–25.

Klitzman, R. (2013). The Failed Promise of 23andMe.

Knight, J.C. (2009). Human genetic diversity: functional consequences for health and disease (Oxford; New York: Oxford University Press).

Kułaga, Z., Litwin, M., Tkaczyk, M., Palczewska, I., Zajączkowska, M., Zwolińska, D., Krynicki, T., Wasilewska, A., Moczulska, A., Morawiec-Knysak, A., et al. (2011). Polish 2010 growth references for school-aged children and adolescents. Eur. J. Pediatr. *170*, 599–609.

Kumari, D., Lokanga, R., Yudkin, D., Zhao, X.-N., and Usdin, K. (2012). Chromatin changes in the development and pathology of the Fragile X-associated disorders and Friedreich ataxia. Biochim. Biophys. Acta BBA - Gene Regul. Mech. *1819*, 802–810.

Lai, C.-Q., Leips, J., Zou, W., Roberts, J.F., Wollenberg, K.R., Parnell, L.D., Zeng, Z.-B., Ordovas, J.M., and Mackay, T.F.C. (2007). Speed-mapping quantitative trait loci using microarrays. Nat. Methods *4*, 839–841.

Laidlaw, J., Gelfand, Y., Ng, K.-W., Garner, H.R., Ranganathan, R., Benson, G., Fondon, J.W., Enson, G.A.R.Y.B., and Iii, J.O.H.N.W.F.O. (2007). Elevated basal slippage mutation rates among the Canidae. J. Hered. *98*, 452–460.

Lander, E.S. (1996). The new genomics: global views of biology. Science *274*, 536–539.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

Larson, G., Karlsson, E.K., Perri, A., Webster, M.T., Ho, S.Y.W., Peters, J., Stahl, P.W., Piper, P.J., Lingaas, F., Fredholm, M., et al. (2012). Rethinking dog domestication by integrating genetics, archeology, and biogeography. Proc. Natl. Acad. Sci. 201203005.

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. Bioinforma. Oxf. Engl. *25*, 1841–1842.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol *9*, e1003118.

Leushkin, E.V., and Bazykin, G.A. (2013). Short Indels Are Subject to Insertion-Biased Gene Conversion. Evolution *67*, 2604–2613.

Li, W.-H., Wu, C.-I., and Luo, C.-C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. *21*, 58–71.

Lim, T.O., Ding, L.M., Zaki, M., Suleiman, A.B., Fatimah, S., Siti, S., Tahir, A., and Maimunah, A.H. (2000). Distribution of body weight, height and body mass index in a national sample of Malaysian adults. Med. J. Malaysia *55*, 108–128.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature *438*, 803–819.

Luo, L., Boerwinkle, E., and Xiong, M. (2011). Association studies for next-generation sequencing. Genome Res. *21*, 1099–1108.

Macgregor, S., Cornes, B.K., Martin, N.G., and Visscher, P.M. (2006). Bias, precision and heritability of self-reported and clinically measured height in Australian twins. Hum. Genet. *120*, 571–580.

Mackay, T.F. (2001). The genetic architecture of quantitative traits. Annu. Rev. Genet. *35*, 303–339.

Maher, B. (2008). Personal genomes: The case of the missing heritability. Nat. News *456*, 18–21.

Majerus, M.E.N., and Mundy, N.I. (2003). Mammalian melanism: natural selection in black and white. Trends Genet. *19*, 585 – 588.

Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., and de los Campos, G. (2011). Beyond Missing Heritability: Prediction of Complex Traits. PLoS Genet *7*, e1002051.

Makvandi-Nejad, S., Hoffman, G.E., Allen, J.J., Chu, E., Gu, E., Chandler, A.M., Loredo, A.I., Bellone, R.R., Mezey, J.G., Brooks, S.A., et al. (2012). Four Loci Explain 83% of Size Variation in the Horse. PLoS ONE *7*, e39929.

Mamidi, R.S., Kulkarni, B., and Singh, A. (2011). Secular trends in height in different states of India in relation to socioeconomic characteristics and dietary intakes. Food Nutr. Bull. *32*, 23–34.

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L.M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature *454*, 479–485.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Martincorena, I., Seshasayee, A.S.N., and Luscombe, N.M. (2012). Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature *485*, 95–98.

Mather, K. (1943). Polygenic inheritance and natural selection. Biol. Rev. *18*, 32–64.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.

Meisel, A., and Vega, M. (2004). A Tropical Success Story: A Century of Improvements in the Biological Standard of Living, Colombia 1910-2002.

Mendel, G.J. (1865). Experiments Concerning Plant Hybrids. Proc. Nat. Hist. Soc. Brünn *IV*, 3–47.

Metzger, J., Schrimpf, R., Philipp, U., and Distl, O. (2013). Expression levels of LCORL are associated with body size in horses. PloS One *8*, e56497.

Mexican Business Web (2012). ¿Cuánto miden los mexicanos?

Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J.C., and Forejt, J. (2009). A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science *323*, 373–375.

Mittelman, D., Moye, C., Morton, J., Sykoudis, K., Lin, Y., Carroll, D., and Wilson, J.H. (2009). Zinc-finger directed double-strand breaks within CAG repeat tracts promote repeat instability in human cells. Proc. Natl. Acad. Sci. *106*, 9607–9612.

Moosa, K. (2002). National Nutrition Survey for Adult Bahrainis Aged 19 Years and Above.

Morey, D.F. (1994). The Early Evolution of the Domestic Dog. Am. Sci. *82*, 336–347.

Mounier, A., Marchal, F., and Condemi, S. (2009). Is Homo heidelbergensis a distinct species? New insight on the Mauer mandible. J. Hum. Evol. *56*, 219–246.

Msamati, B.C., and Igbigbi, P.S. (2000). Anthropometric profile of urban adult black Malawians. East Afr. Med. J. *77*, 364–368.

Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G., and Donnelly, P. (2010). Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. Science *327*, 876–879.

National Center for Health Statistics (2008). Anthropometric Reference Data for Children and Adults: United States, 2003-2006.

National Statistics England (2011). Health Survey for England.

NSO Malta (2003). The Maltese Way of Life.

O'Driscoll, M., and Jeggo, P.A. (2006). The role of double-strand break repair — insights from human genetics. Nat. Rev. Genet. *7*, 45–54.

Ohman, J.C., Wood, C., Wood, B., Crompton, R.H., Günther, M.M., Yu, L., Savage, R., and Wang, W. (2002). Stature-at-death of KNM-WT 15000. Hum. Evol. *17*, 129–141.

Okosun, I.S., Cooper, R.S., Rotimi, C.N., Osotimehin, B., and Forrester, T. (1998). Association of waist circumference with risk of hypertension and type 2 diabetes in Nigerians, Jamaicans, and African-Americans. Diabetes Care *21*, 1836–1842.

Oliver, P.L., Goodstadt, L., Bayes, J.J., Birtle, Z., Roach, K.C., Phadnis, N., Beatson, S.A., Lunter, G., Malik, H.S., and Ponting, C.P. (2009). Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. PLoS Genet *5*, e1000753.

Ozer, B.K. (2008). Secular trend in body height and weight of Turkish adults. Anthropol. Sci. *116*, 191–199.

Özer, B.K., Sağır, M., and Özer, İ. (2011). Secular changes in the height of the inhabitants of Anatolia (Turkey) from the 10th millennium B.C. to the 20th century A.D. Econ. Hum. Biol. *9*, 211–219.

Pages, H. BSgenome: Infrastructure for Biostrings-based genome data packages.

Pages, H., Aboyoun, P., Gentleman, R., and DebRoy, S. Biostrings: String objects representing biological sequences, and matching algorithms.

Pâques, F., Leung, W.-Y., and Haber, J.E. (1998). Expansions and Contractions in a Tandem Repeat Induced by Double-Strand Break Repair. Mol. Cell. Biol. *18*, 2045–2054.

Parker, H.G., Kim, L.V., Sutter, N.B., Carlson, S., Lorentzen, T.D., Malek, T.B., Johnson, G.S., DeFrance, H.B., Ostrander, E.A., and Kruglyak, L. (2004). Genetic Structure of the Purebred Domestic Dog. Science *304*, 1160–1164.

Parker, H.G., VonHoldt, B.M., Quignon, P., Margulies, E.H., Shao, S., Mosher, D.S., Spady, T.C., Elkahloun, A., Cargill, M., Jones, P.G., et al. (2009). An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science *325*, 995–998.

Parvanov, E.D., Petkov, P.M., and Paigen, K. (2010). Prdm9 Controls Activation of Mammalian Recombination Hotspots. Science *327*, 835–835.

Pearson, K., and Lee, A. (1903). On the Laws of Inheritance in Man: I. Inheritance of Physical Characters. Biometrika *2*, 357–462.

Pearson, C.E., Edamura, K.N., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. *6*, 729–742.

Peltonen, M., Harald, K., Männistö, S., Saarikoski, L., Lund, L., Sundvall, J., Juolevi, A., Tiina, L., Aldén-Nieminen, H., Luoto, R., et al. (2008). Kansallinen FINRISKI 2007 - terveystutkimus : tutkimuksen toteutus ja tulokset. taulukkoliite.

Pennisi, E. (2013). Old Dogs Teach a New Lesson About Canine Origins. Science *342*, 785–786.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.

Petes, T.D. (2001). Meiotic recombination hot spots and cold spots. Nat. Rev. Genet. *2*, 360–369.

Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

Del Pino, M., Bay, L., Lejarraga, H., Kovalskys, I., Berner, E., and Rausch Herscovici, C. (2005). Peso y estatura de una muestra nacional de 1.971 adolescentes de 10 a 19 años: las referencias argentinas continúan vigentes. Arch. Argent. Pediatría *103*, 323–330.

Plomin, R., Haworth, C.M.A., and Davis, O.S.P. (2009). Common disorders are quantitative traits. Nat. Rev. Genet. *10*, 872–878.

Ponting, C.P. (2011). What are the genomic drivers of the rapid evolution of PRDM9? Trends Genet. *27*, 165–171.

Pryce, J.E., Hayes, B.J., Bolormaa, S., and Goddard, M.E. (2011). Polymorphic regions affecting human height also control stature in cattle. Genetics *187*, 981–984.

R Core Team (2013). R: A language and environment for statistical computing. (R Foundation for Statistical Computing).

Ranasinghe, P., Jayawardana, M.A.N.A.A.D., Constantine, G.R., Sheriff, M.H.R., Matthews, D.R., and Katulanda, P. (2011). Patterns and correlates of adult height in Sri Lanka. Econ. Hum. Biol. *9*, 23–29.

Reed, D.R., Lawler, M.P., and Tordoff, M.G. (2008). Reduced body weight is a common effect of gene knockout in mice. BMC Genet. *9*, 4.

Rensch, B. (1950). Die Abhängigkeit der relativen Sexualdifferenz von der Körpergrösse. Bonn. Zool. Beitr. *1*, 58–69.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Rightmire, G.P. (1998). Human evolution in the Middle Pleistocene: The role of Homo heidelbergensis. Evol. Anthropol. Issues News Rev. *6*, 218–227.

Rimbault, M., Beale, H.C., Schoenebeck, J.J., Hoopes, B.C., Allen, J.J., Kilroy-Glynn, P., Wayne, R.K., Sutter, N.B., and Ostrander, E.A. (2013). Derived variants at six genes explain nearly half of size reduction in dog breeds. Genome Res. *23*, 1985–1995.

Risch, N., Reich, E.W., Wishnick, M.M., and McCarthy, J.G. (1987). Spontaneous mutation and parental age in humans. Am. J. Hum. Genet. *41*, 218–248.

Robertson, A. (1967). The nature of quantitative genetic variation. Herit. Mendel 265–280.

Robinson, M.C., Stone, E.A., and Singh, N.D. (2013). Population genomic analysis reveals no evidence for GC-biased gene conversion in Drosophila melanogaster. Mol. Biol. Evol.

Rosindell, J., and Harmon, L.J. (2012). OneZoom: A Fractal Explorer for the Tree of Life. PLoS Biol *10*, e1001406.

Ruff, C. (2002). Variation in human body size and shape. Annu. Rev. Anthropol. 211–232.

Ruff, C.B., and Walker, A. (1993). Body size and body shape. Nariokotome Homo Erectus Skelet. 234–265.

Ruff, C., Niskanen, M., Junno, J.-A., and Jamison, P. (2005). Body mass prediction from stature and bi-iliac breadth in two high latitude populations, with application to earlier higher latitude humans. J. Hum. Evol. *48*, 381–392.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature.

Savolainen, P., Zhang, Y., Luo, J., Lundeberg, J., and Leitner, T. (2002). Genetic Evidence for an East Asian Origin of Domestic Dogs. Science *298*, 1610–1613.

Schilling, M.F., Watkins, A.E., and Watkins, W. (2002). Is Human Height Bimodal? Am. Stat. *56*, 223–229.

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. Chromosoma *109*, 365–371.

Schmitt, D. (2003). Insights into the evolution of human bipedalism from experimental studies of humans and other primates. J. Exp. Biol. *206*, 1437–1448.

Schoenebeck, J.J., Hutchinson, S.A., Byers, A., Beale, H.C., Carrington, B., Faden, D.L., Rimbault, M., Decker, B., Kidd, J.M., Sood, R., et al. (2012). Variation of BMP3 Contributes to Dog Breed Skull Diversity. PLoS Genet *8*, e1002849.

Schönbeck, Y., Talma, H., van Dommelen, P., Bakker, B., Buitendijk, S.E., HiraSing, R.A., and van Buuren, S. (2013). The world's tallest nation has stopped growing taller: the height of Dutch children from 1955 to 2009. Pediatr. Res. *73*, 371–377.

Schultz, T.P. (2005). Productive Benefits of Health: Evidence from Low-Income Countries (Rochester, NY: Social Science Research Network).

Schumacher, J., Kaneva, R., Jamra, R.A., Diaz, G.O., Ohlraun, S., Milanova, V., Lee, Y.-A., Rivas, F., Mayoral, F., Fuerst, R., et al. (2005). Genomewide Scan and Fine-Mapping Linkage Studies in Four European Samples with Bipolar Affective Disorder Suggest a New Susceptibility Locus on Chromosome 1p35-p36 and Provides Further Evidence of Loci on Chromosome 4q31 and 6q24. Am. J. Hum. Genet. *77*, 1102–1111.

Ségurel, L., Leffler, E.M., and Przeworski, M. (2011). The Case of the Fickle Fingers: How the PRDM9 Zinc Finger Protein Specifies Meiotic Recombination Hotspots in Humans. PLoS Biol *9*, e1001211.

Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., Hofreiter, M., Possnert, G., and Pääbo, S. (2004). No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol. *2*, E57.

Shea, B.T., and Bailey, R.C. (1996). Allometry and adaptation of body proportions and stature in African pygmies. Am. J. Phys. Anthropol. *100*, 311–340.

Shields, M., Connor Gorber, S., Janssen, I., and Tremblay, M.S. (2011). Bias in self-reported estimates of obesity in Canadian health surveys: an update on correction equations for adults. Health Rep. Stat. Can. Can. Cent. Health Inf. Rapp. Sur Santé Stat. Can. Cent. Can. Inf. Sur Santé *22*, 35–45.

Shinde, D.N., Elmer, D.P., Calabrese, P., Boulanger, J., Arnheim, N., and Tiemann-Boege, I. (2013). New evidence for positive selection helps explain the paternal age effect observed in achondroplasia. Hum. Mol. Genet. *22*, 4117–4126.

Siddle, K. (2011). Signalling by insulin and IGF receptors: supporting acts and new players. J. Mol. Endocrinol. *47*, R1–R10.

Signer-Hasler, H., Flury, C., Haase, B., Burger, D., Simianer, H., Leeb, T., and Rieder, S. (2012). A Genome-Wide Association Study Reveals Loci Influencing Height and Other Conformation Traits in Horses. PLoS ONE *7*, e37282.

Smagulova, F., Gregoretti, I. V, Brick, K., Khil, P., Camerini-Otero, R.D., and Petukhova, G. V (2011). Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature *472*, 375–378.

Smithsonian Institution (2010a). Homo erectus.

Smithsonian Institution (2010b). Homo heidelbergensis.

Smithsonian Institution (2010c). Homo neanderthalensis.

So, H.-K., Nelson, E.A.S., Li, A.M., Wong, E.M.C., Lau, J.T.F., Guldan, G.S., Mak, K.-H., Wang, Y., Fok, T.-F., and Sung, R.Y.T. (2008). Secular changes in height, weight and body mass index in Hong Kong Children. BMC Public Health *8*, 320.

Soranzo, N., Rivadeneira, F., Chinappen-Horsley, U., Malkina, I., Richards, J.B., Hammond, N., Stolk, L., Nica, A., Inouye, M., Hofman, A., et al. (2009). Meta-Analysis of Genome-Wide Scans for Human Adult Stature Identifies Novel Loci and Associations with Measures of Skeletal Frame Size. PLoS Genet *5*, e1000445.

Starc, G., and Strel, J. (2011). Is there a rationale for establishing Slovenian body mass index references of school-aged children and adolescents?

Statistics Netherlands, Haag, D., and Heerlen (2012). Lifestyle, preventive screening; sex, age.

Statistisches Bundesamt (2009). Mikrozensus - Fragen zur Gesundheit.

Steckel, R.H. (2004). New Light on the “Dark Ages”: The Remarkably Tall Stature of Northern European Men during the Medieval Era. Soc. Sci. Hist. *28*, 211–229.

Stevo Popovic, D.B. (2013). Body Height and Its Estimation Utilizing Arm Span Measurements in Serbian Adults. Int. J. Morphol. *31*, 271–279.

Stock, J.T. (2008). Are humans still evolving? Technological advances and unique biological characteristics allow us to adapt to environmental stress. Has this stopped genetic evolution? EMBO Rep. *9*, S51–S54.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. *100*, 9440–9445.

Subramanian, S.V., Özaltin, E., and Finlay, J.E. (2011). Height of Nations: A Socioeconomic Analysis of Cohort Differences and Patterns among Women in 54 Low- to Middle-Income Countries. PLoS ONE *6*, e18962.

Sutter, N.B., Bustamante, C.D., Chase, K., Gray, M.M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., Jones, P.G., et al. (2007). A single IGF1 allele is a major determinant of small size in dogs. Science *316*, 112–115.

Sutter, N.B., Mosher, D.S., Gray, M.M., and Ostrander, E.A. (2008). Morphometrics within dog breeds are highly reproducible and dispute Rensch's rule. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *19*, 713–723.

Syvänen, A.-C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat. Rev. Genet. *2*, 930–942.

Szostak, J.W., Orr-Weaver, T.L., Rothstein, R.J., and Stahl, F.W. (1983). The double-strand-break repair model for recombination. Cell *33*, 25–35.

Tawfeek, H. (2002). Relationship between waist circumference and blood pressure among the population in Baghdad, Iraq. Food Nutr. Bull. *23*, 402–406.

Thalmann, O., Shapiro, B., Cui, P., Schuenemann, V.J., Sawyer, S.K., Greenfield, D.L., Germonpré, M.B., Sablin, M.V., López-Giráldez, F., Domingo-Roura, X., et al. (2013). Complete Mitochondrial Genomes of Ancient Canids Suggest a European Origin of Domestic Dogs. Science *342*, 871–874.

Thomas, J.H., Emerson, R.O., and Shendure, J. (2009). Extraordinary Molecular Evolution in the PRDM9 Fertility Gene. PLoS ONE *4*, e8505.

Thorpe, S.K.S., Holder, R.L., and Crompton, R.H. (2007). Origin of Human Bipedalism As an Adaptation for Locomotion on Flexible Branches. Science *316*, 1328–1331.

Trut, L. (1999). Early Canid Domestication: The Farm-Fox Experiment. Am. Sci. *87*, 160.

Turchin, M.C., Chiang, C.W.K., Palmer, C.D., Sankararaman, S., Reich, D., and Hirschhorn, J.N. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat. Genet. *44*, 1015–1019.

Turri, M.G., Datta, S.R., DeFries, J., Henderson, N.D., and Flint, J. (2001a). QTL analysis identifies multiple behavioral dimensions in ethological tests of anxiety in laboratory mice. Curr. Biol. CB *11*, 725–734.

Turri, M.G., Henderson, N.D., DeFries, J.C., and Flint, J. (2001b). Quantitative trait locus mapping in laboratory mice derived from a replicated selection experiment for open-field activity. Genetics *158*, 1217–1226.

Tutkuviene, J. (2005). Sex and gender differences in secular trend of body size and frame indices of Lithuanians. Anthropol. Anz. Ber. Über Biol.-Anthropol. Lit. *63*, 29–44.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., and Leunissen, J.A.M. (2007). Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. *35*, W71–W74.

Usdin, K. (2008). The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. Genome Res. *18*, 1011–1019.

Vaysse, A., Ratnakumar, A., Derrien, T., Axelsson, E., Rosengren Pielberg, G., Sigurdsson, S., Fall, T., Seppälä, E.H., Hansen, M.S.T., Lawley, C.T., et al. (2011). Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. *7*, e1002316.

Velarde, C.N. (2006). Boletin Instituto Nacional de Salud.

Velinov, M., Slaugenhaupt, S.A., Stoilov, I., Scott, C.I., Jr, Gusella, J.F., and Tsipouras, P. (1994). The gene for achondroplasia maps to the telomeric region of chromosome 4p. Nat. Genet. *6*, 314–317.

Venkaiah, K., Damayanti, K., Nayak, M.U., and Vijayaraghavan, K. (2002). Diet and nutritional status of rural adolescents in India. Publ. Online 07 Novemb. 2002 Doi101038sjejcn1601457 *56*.

Vernot, B., and Akey, J.M. (2014). Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science 1245938.

Verstrepen, K.J., Jansen, A., Lewitter, F., and Fink, G.R. (2005). Intragenic tandem repeats generate functional variability. Nat. Genet. *37*, 986–990.

Vignerová, J., Brabec, M., and Bláha, P. (2006). Two centuries of growth among Czech children and youth. Econ. Hum. Biol. *4*, 237–252.

Vilà, C., Savolainen, P., Maldonado, J.E., Amorim, I.R., Rice, J.E., Honeycutt, R.L., Crandall, K.A., Lundeberg, J., and Wayne, R.K. (1997). Multiple and Ancient Origins of the Domestic Dog. Science *276*, 1687–1689.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Visscher, P.M. (2008). Sizing up human height variation. Nat. Genet. *40*, 489–490.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era — concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266.

Vonholdt, B.M., Pollinger, J.P., Lohmueller, K.E., Han, E., Parker, H.G., Quignon, P., Degenhardt, J.D., Boyko, A.R., Earl, D.A., Auton, A., et al. (2010). Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature *464*, 898–902.

Wagner, G.P., and Altenberg, L. (1996). Perspective: Complex Adaptations and the Evolution of Evolvability. Evolution *50*, 967–976.

Wang, S., Lachance, J., Tishkoff, S.A., Hey, J., and Xing, J. (2013). Apparent Variation in Neanderthal Admixture among African Populations is Consistent with Gene Flow from Non-African Populations. Genome Biol. Evol. *5*, 2075–2081.

Webber, C., and Ponting, C.P. (2005). Hotspots of mutation and breakage in dog and human chromosomes. Genome Res. *15*, 1787–1797.

Welsh Assembly Government (2010). Welsh Health Survey 2009.

WHO (2007). Mongolian Steps Survey on the Prevalence of Noncommunicable Disease Risk Factors 2006.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer N. Y.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. J. Stat. Softw. *40*, 1–29.

Wickham, H. (2012). stringr: Make it easier to work with strings.

Wood, A.R., Hernandez, D.G., Nalls, M.A., Yaghootkar, H., Gibbs, J.R., Harries, L.W., Chong, S., Moore, M., Weedon, M.N., Guralnik, J.M., et al. (2011). Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. Hum. Mol. Genet.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. *8*, 206–216.

Yamamoto, T., Kuboki, Y., Lin, S.Y., Sasaki, T., and Yano, M. (1998). Fine mapping of quantitative trait loci Hd-1, Hd-2 and Hd-3, controlling heading date of rice, as single Mendelian factors. Theor. Appl. Genet. *97*, 37–44.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

Yang, X., Li, Y., Ma, G.-S., Hu, X., Wang, J., Cui, Z., Wang, Z., Yu, W., Yang, Z., and Zhai, F. (2005). Study on weight and height of the Chinese people and the differences between 1992 and 2002. Zhonghua Liu Xing Bing Xue Za Zhi Zhonghua Liuxingbingxue Zazhi *26*, 489–493.

Zhang, G., Karns, R., Sun, G., Indugula, S.R., Cheng, H., Havas-Augustin, D., Novokmet, N., Durakovic, Z., Missoni, S., Chakraborty, R., et al. (2012). Finding Missing Heritability in Less Significant Loci and Allelic Heterogeneity: Genetic Variation in Human Height. PLoS ONE *7*, e51211.

Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. 201119675.

Biographical Information

Eldon Goodwin Prince attended Manor High School near Austin, Texas, graduating third in his class. During high school he served in many leadership roles including President of the National Honor Society, Band President, Class Vice-President, and editor-in-chief of the newspaper. He also started a lawn care business that eventually paid for much of his undergraduate education. Eldon enrolled at Brigham Young University in the fall of 2001 and was a member of the BYU Marching Band. After his freshman year he served a full-time religious mission for two years in central and southern California. Eldon returned to Brigham Young University in 2005 and graduated with a Bachelor of Science in molecular biology and a minor in statistics in 2008. In the laboratory of Dr. Richard Robison he received a research grant to lead a team of undergraduates in developing a more accurate clinical test for group A *Streptococcus*. As an undergraduate teaching assistant he also worked with Dr. Donald Breakwell on a study of students' learning preferences. In fall of 2008 Eldon joined the laboratory of Dr. Jeffery Demuth at University of Texas at Arlington to study genome-wide gene expression differences between male and female red flour beetles. In the summer of 2010 Eldon joined the laboratory of Dr. John Fondon III to study the genetic basis of rapid morphological evolution. His work in the lab of Dr. Fondon ranged from designing and constructing avian infrastructure to simulating DNA mutation rates and analyzing high-throughput genomic data in humans and dogs. After completing his PhD at the University of Texas at Arlington, Eldon will begin work as a Metrics and Reporting Sr. Analyst at Dell in Austin, Texas.