This is to certify that the
dissertation entitled

IDENTIFICATION OF GENE-SPECIFIC SINGLE NUCLEOTIDE POLYMORPHISMS WITHIN THE CANINE GENOME AND THEIR USE TO DETERMINE NUCLEOTIDE DIVERSITY AND INBREEDING COEFFICIENTS WITHIN THE CANINE GENOME

presented by

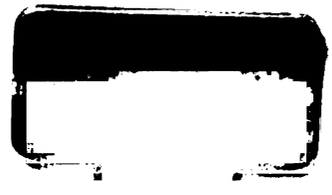
JAMES A. BROUILLETTE, MD

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Genetics

Patrick J. Venta
Major Professor's Signature

1/21/10
Date



IDENTIFICATION OF GENE-SPECIFIC SINGLE NUCLEOTIDE POLYMORPHISMS WITHIN THE CANINE GENOME AND THEIR USE TO DETERMINE NUCLEOTIDE DIVERSITY AND INBREEDING COEFFICIENTS WITHIN THE CANINE GENOME

By

James A. Brouillette, MD

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Genetics

2010

ABSTRACT

IDENTIFICATION OF GENE-SPECIFIC SINGLE NUCLEOTIDE POLYMORPHISMS WITHIN THE CANINE GENOME AND THEIR USE TO DETERMINE NUCLEOTIDE DIVERSITY AND INBREEDING COEFFICIENTS WITHIN THE CANINE GENOME

By

James A. Brouillette, MD

The domestic dog, *Canis familiaris,* has lived in close relationship with man for thousands of years, working as a hunter, herder, and loyal companion. Through selective breeding, the various dog breeds have a high prevalence of various genetic diseases. Genetic studies are ongoing to elucidate the nature of disease-causing mutations within the various dog breeds. In the work presented here, I elucidate a method to identify single nucleotide polymorphisms (SNPs) within canine genes of interest by pooling and sequencing DNA from across ten breeds of dog. The SNP markers generated are used to estimate heterozygosity within the canine genome, to demonstrate that the markers generated by across breed SNP identification will be heterozygous within breeds, and to estimate the coefficient of inbreeding within three breeds of dog.

DEDICATION

This manuscript is dedicated to my sons Jacob and Andrew, who made as many sacrifices as I did in order to complete this degree. I love you and appreciate your patience during the long course of my graduate education.

I also dedicate this work to my wife Tammie, without whom this work would not have been completed. Thank you for your tireless support.

ACKNOWLEDGEMENTS

I would like to thank my advisors Patrick Venta and Vilma Yuzbasiyan-Gurkan whose enthusiasm and support made this work possible. Your knowledge, inquisitiveness, and energy were greatly appreciated.

I would also like to thank Donna Housley, David Entz, Tracy Hammer, Sarah Colombini, Margo Machen, Tiffy Zachos, Susan Ewart, Martha Mulks, Betty Werner, Cheryl Swensen, John Kruger, and Simon Peterson-Jones for your many discussions and problem solving sessions along the way.

Finally, I would like to thank Kathy Lovell for shepherding this manuscript to completion. This project would not have been completed without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS[1]

Term	Abbreviation
Base(s) pairs	bp
centiMorgans	cM
Deoxyribonucleic acid	DNA
Kilobase(s) (pairs)	kb
Mega base(s) pairs	Mb
Micrograms	µg
Microliters	µl
Micromolar	µM
Milliliters	ml
Millimolar	mM
Nanograms	ng
Picograms	pg
Picomoles	pmol
Polymerase Chain Reaction	PCR
Ribonucleic acid	RNA
Units	U

1. Abbreviations for specific gene names are found in the text.

Chapter 1

Introduction

Introduction

Domestic dogs and humans have lived in close association for centuries. During that time, man has selectively bred dogs to the point where today there are over three hundred different dog breeds in existence worldwide (Ostrander et al., 2000).

Man's understanding of the genetic diseases that occur in domestic dogs has steadily grown over the last few decades. As recently as 1979, there were only 13 canine disorders that were established as congenital or inherited in dogs (Pearson, 1979). That number has rapidly increased. In 1988, there were more than 200 genetic disorders that had been identified in domestic dogs, more than 70% of which were inherited as autosomal recessive disorders (Patterson et al., 1988).

With the widespread use of antibiotics, antihelmintics, vaccination against viral diseases, and improved and standardized diets, genetic disease has grown in clinical importance among veterinarians, dog breeders, and dog owners (Patterson, 2000). At latest count, there are 370 different canine genetic disorders, with 50% of these having breed-specific aggregations (Patterson, 2000; Ostrander et al., 2000). Of these 370 diseases, 215 (58%) have clinical and laboratory abnormalities that resemble a human genetic disease and more than 70% of these are inherited as autosomal recessive traits, X-linked recessive traits, or have a complex pattern of inheritance (Ostrander et al., 2000). An additional 5-10 diseases are added to the growing list each year. The price of this genetic disease manifests itself as pain and suffering for the affected dogs, emotional suffering to owners

and breeders, and an estimated $500 million per year in costs to diagnose and treat affected animals (Padgett, 1998). Currently, 50 of the 370 canine genetic diseases have been defined at the molecular level (Giger et al., 2006).

Finding Canine Disease Genes

The disease genes that have been identified to date fall into three different categories. The first category is those diseases that have an identifiable protein that is absent or not functioning. The search for the mutation in these cases becomes a matter of searching the gene encoding the faulty protein for mutations in the coding or control regions. Factor IX deficiency causing hemophilia B in Cairn Terriers (Evans et al., 1989), von Willebrand factor deficiency causing von Willebrand's disease in Scottish Terriers (Venta et al., 2000), and C3 deficiency causing Complement third component deficiency in Brittany Spaniels (Ameratunga et al., 1998) are examples of diseases in this category.

A second category of disease genes are those that have been identified by examining candidate genes for the presence of a mutation. A candidate gene is any gene that is suspected to harbor the mutation causing the disease phenotype by virtue of its role in a similar disease in another species or based on some prior knowledge of its biochemical properties. Mutations in the dystrophin gene causing Duchenne's muscular dystrophy in Golden Retrievers (Sharp et al., 1992), the phosphodiesterase 6 beta gene causing rod-cone dysplasia in Irish Setters (Suber et al., 1993), and the phosphodiesterase alpha gene

causing rod-cone dysplasia in Cardigan Welsh Corgis (Petersen-Jones et al., 1999) fall into this category.

The final category of genetic diseases contains those diseases that have no obvious defective protein or candidate genes to test as causative agents. In these cases, purely genetic analysis can be performed using linkage analysis in families in which the disease is present or by association analysis in affected individuals. Yuzbasiyan-Gurkan et al. (1997) were able to identify a marker closely linked to the mutation causing copper toxicosis in Bedlington Terriers using linkage analysis. The COMMD1 gene was subsequentlyl cloned, leading to new insights in copper metabolism in mammals and the discovery of a whole new family of related proteins (van de Sluis et al., 2002; Burstein et al., 2005). Similarly, Lin et al. (1999) were able to identify a mutation in the hypocretin receptor which causes narcolepsy in Doberman Pinschers and Labrador Retrievers using linkage analysis, opening a whole new field of investigation for sleep and sleep-related disorders (Zeiter et al., 2006). Ostrander and Kruglyak have demonstrated the feasibility of whole genome association analysis in dogs using computer modeling (Ostrander and Kruglyak, 2000). More recent work has moved association mapping in dogs from a theoretical possibility to a future certainty, as all the necessary groundwork has been done to enable geneticists to undertake association studies in dogs (Kirkness et al., 2003; Lindblad-Toh et al., 2005; Sutter et al., 2004; Ostrander and Kruglyak, 2000; Clark et al., 2004). A recent use of association analysis in dogs resulted in the discovery that the merle coat color is due to a mutation in the SILV gene (Clarke et al., 2006).

Evolution of Canine Genetic Markers

The first genetic markers in canines were identified as protein polymorphisms observed as variants in electrophoretic mobility. The vast majority of these were structural proteins or enzymes of the various components of the blood such as isocitrate dehydrogenase, albumin, and hemoglobin (Meera Khan et al., 1973; Weiden et al., 1974; Simonsen, 1976). By today's standards, these markers exhibited little polymorphism and were limited in scope. These markers were soon replaced by RFLP markers.

During the 1980s, several RFLP markers were discovered in genes such as the DLA-D and DLA-A (Sarmiento and Storb, 1988, 1989). These markers, while more abundant and polymorphic than the protein polymorphisms, were labor and time intensive and required relatively large quantities of DNA.

A major leap forward in canine genetics occurred in the 1990s. Microsatellite (simple sequence length polymorphisms or SSLPs) markers, simple sequence motifs of two to six nucleotides repeated in arrays of varying lengths, were found to be highly abundant throughout mammalian genomes, highly polymorphic, and rapidly typable using PCR. They have the advantage of being present at about 20 kb intervals throughout the canine genome (Yuzbasiyan-Gurkan and Venta, Pers. Comm.). The one drawback is that they are only occasionally found associated with coding regions of the canine genome. Our collaborator, Dr. Vilma Yuzbasiyan-Gurkan, previously developed several hundred anonymous SSLP markers throughout the canine genome and used this resource to

identify a marker that is tightly linked to the mutation causing copper toxicosis in Bedlington Terriers (Yuzbasiyan-Gurkan et al., 1997). Many other SSLP type markers have been discovered and developed in other labs (Francisco et al., 1996; Ostrander et al., 1993, 1995).

This work centers on the identification and development of single nucleotide polymorphisms as genetic markers. At the time this work was undertaken, nothing was known about the frequency of occurrence of SNPs in the canine genome (Chapter 3). Based on data from the human genome, it was likely that the SNPs would occur more frequently in the canine genome than SSLP markers (Collins et al., 1997, Nickerson et al., 1998). In addition, the introduction of DNA arrays made high throughput genotyping of SNP markers feasible (Wang et al., 1998; Chee et al., 1996; Landegren et al., 1998, see below,), which would allow for more rapid whole-genome scans of SNP markers.

The Canine Genetic Map

The first maps of the canine genome were published in 1997 (Lingaas et al., 1997; Mellersh et al., 1997; Langston et al., 1997). Lingaas et al. established 16 linkage groups and assigned a total of 43 markers to those 16 groups.

Using 17 three generation families, Mellersh et al. (1997) assigned 139 microsatellite markers to 30 linkage groups containing at least one other linked marker (with a lod score of 3 or greater). The linkage groups ranged in size from 2.3 to 106.1 cM. This map

covered an estimated 884.2 cM of the canine genome with an average marker spacing of 14.03 cM. An additional 11 polymorphic markers were not linked to any other marker. Of the 150 markers, 47 were dinucleotide repeats, 102 were tetranucleotide repeats, and one was a hexanucleotide repeat.

In a companion paper, Langston et al. (1997) developed the first canine-rodent somatic cell hybrid panel. This panel contains a total of 43 microcell hybrid clones that each display unique canine chromosome retention patterns and three whole cell hybrids that contained the X chromosome. They assigned 181 microsatellite markers and 27 canine genes to 31 syntenic groups consisting of two or more markers and/or genes. Many of these markers were also used by Mellersh et al. (1997). Each of the syntenic groups had between 2 and 11 markers. Since the canine karyotype consists of 38 pairs of autosomes plus the X and Y chromosomes (Langston et al., 1997), this does not represent full coverage of the canine genome. It does, however, represent a substantial portion of the canine genome.

Priat et al. (1998) followed up these mapping projects with a radiation hybrid map of the canine genome. This map contains a total of 400 markers consisting of 218 gene markers and 182 microsatellite markers. The map contains 347 markers assigned to 57 groups with an additional 53 markers being unlinked in the current map. The groups contain between 2 and 11 linked markers. The radiation hybrid panel consists of 126 cell lines, and the map is thought to cover about 80% of the canine genome.

The work by Priat et al. began the process of integrating the linkage maps of Lingaas et al. (1997) and Mellersh et al. (1997) into a radiation hybrid (RH) map. It also indicates areas of synteny shared by the dog, human, and pig genomes.

The next generation of linkage map was produced by Neff et al. (1999). This map extends the original map of Mellersh et al. (1997) from 150 microsatellite markers to 276 markers divided into 40 linkage groups. Average marker spacing in this map dropped from 14 cM to 9.3 cM. This map is estimated to cover 90% of the canine genome (Neff et al., 1999).

The canine genetic map was integrated into a single map by the assignment of the linkage groups from the RH map (Priat et al., 1998) and linkage map (Lingaas et al., 1997; Mellersh et al., 1997; Neff et al., 1999) to specific canine chromosomes using chromosome painting (Yang et al., 1999). Of the 44 published RH groups and 40 published linkage groups, 39 and 33 groups were assigned to specific chromosomes, respectively. In addition, Yang et al. were able to align chromosomal regions of the canine karyotype with syntenic regions of both the human and red fox karyotypes. The syntenic alignments should enable researchers to identify genes in the human comparative map that will be candidates for canine genetic diseases by cross-species comparison, once linkage is established to a canine marker (Yang et al., 1999). A combined 3MB resolution RH map of CFA1 that incorporated SNP markers was published in 2004 (Housley et al., 2004).

Several refinements have been made in the canine genome map since the above reports were published (Werner et al., 1999; Sargan et al., 2000; Mellersh et al., 2000; Richman et al., 2001; Lingaas et al., 2001). At the time, the expanded and integrated RH and linkage map consisted of approximately 800 markers, more than 300 of which were genes (Mellersh et al., 2000; Lingaas et al., 2001; Sargan et al., 2000). The average spacing between markers was 9 cM. In addition, each of the synteny groups had been assigned to specific canine chromosomes (Sargan et al., 2001). One additional refinement was the characterization of a set of 172 markers for genome-wide screens of the canine genome (Minimal screening set-1 [MSS-1] Richman et al., 2001). This set of markers, all of which were microsatellites, were chosen because they provided as complete coverage of the canine genome as was possible, they were highly informative, and they had been ordered in linkage groups with high statistical certainty (Richman et al., 2001). It had been estimated that 42% of the canine genome is within 5 cM of at least one of these markers, and 77% of the genome was within 10 cM (Richman et al., 2001). While there were some gaps within the canine genetic map, the map taken in total was thought to be sufficient for the whole-genome linkage analysis (Richman et al., 2001; Ostrander and Kruglyak, 2000). An extended and improved marker set (MSS-2) was published in 2004 (Clark et al., 2004).

By 2004, Breen et al. created an integrated FISH and Radiation hybrid map of the canine genome (Breen et al., 2004). That map contained a total of 4250 markers, 4100 of which were assigned to linkage groups and to canine chromosomes. The genes were assigned to

60 different linkage groups that could be assigned to the 38 canine autosomes and two sex chromosomes (Breen et al., 2004).

In 2005, the entire canine genome was sequenced from a female Boxer (Lindblad-Toh et al, 2005). This map represents 7.5-fold redundancy and is thought to cover 99% of the canine genome. In addition, the sequence data revealed 2.5 million SNP markers within the canine genome. Commercially available SNP arrays (Lindblad-Toh et al., 2006) have been developed.

The SNP Markers

The canine genetic markers that have been developed in this series of studies are single nucleotide polymorphisms, or SNPs. These markers are even more abundant than SSLP markers and are also amenable to typing by PCR amplification. In addition, it has been possible to develop SNPs as type I, or gene-associated, markers (Werner et al., 1999; Sargan et al., 2000). This has the effect of anchoring them in the canine genome and furthers comparative genetics across mammalian species. Early reports of SNP markers in the canine genome began appearing in the mid-1990s in such genes as erythroid aminolevulinate synthase, γ-D-crystallin and opsin (Boyer et al., 1995; Shibuya et al., 1995; Ray et al., 1996).

SNPs are defined as single base pair substitutions in genomic DNA at which different sequence alternatives exist in normal individuals in some population. The allele frequency of the most common allele must also be 99% or less (Brookes, 1999).

The frequencies of transitions, transversions, and indels are not equal. Two thirds of all SNPs are C → T transitions while the other third is made up of all the other possible changes (Wang et al., 1998; Brookes, 1999). It has been speculated (Halliday and Grigg, 1993) that the reason for the propensity of C → T changes is that 3-5% of cytosine residues in mammalian genomes are presumed to be methylated. These residues can undergo spontaneous deamination to yield thymine (Halliday and Grigg, 1993). Thus, a methylated cytosine residue gives rise to a thymidine residue. The result is the conversion from a C-G base pair to a T-A base pair. An unmethylated C residue that undergoes deamination will be recognized as a uracil residue, and readily repaired back to the original C residue.

Before this work, the frequency of occurrence of SNPs in the canine genome was unknown. It had been established that in humans, if one randomly analyzes two chromosomes, a SNP will typically be observed to occur once per 1000 bp of DNA (Brookes, 1999). This means that there is a 0.1% chance that any base will be heterozygous in a given individual. Within gene coding regions, the frequency of occurrence of a SNP drops to around 1 in 4000 bp, with half of these changes resulting in non-synonymous changes (Brookes, 1999). These numbers indicated that there would be several million nucleotide differences between any two individuals and around 100,000

differences in their proteomes (Brookes, 1999). This estimate would later be borne out by sequence analysis of the canine genome (Lindblad-Toh et al., 2005).

Before this work, there had not been any publications on a systematic search for SNPs in the canine genome. The human genome project had resulted in a huge amount of human DNA sequence data being available to the scientific community. Several groups have taken advantage of this resource to locate SNPs in the human genome. Two groups (Buetow et al., 1999, Picoult-Newberg et al. 1999) searched the expressed sequence tag (EST) database for SNPs. They scanned the database for multiply sequenced ESTs and examined them for sequence differences. They then went back to the DNA and confirmed that the SNPs did indeed exist in the DNA among the population.

Taillon-Miller et al. (1998) used a similar strategy on genomic DNA. They analyzed overlapping clones of genomic DNA and looked for sequence differences. They then analyzed sequence differences to see if they represented sequencing errors or whether they indeed represented SNPs.

Searching for SNPs in the Human Genome

Several large-scale SNP identification projects have been undertaken to establish estimates of the frequency of occurrence of SNPs and to develop methods of identification and genotyping of SNPs once they had been located. Wang et al. (1998) examined 2.3 Mb of DNA from three individuals and a pool of 10 individuals using gel-

based sequencing and high-density variation-detection DNA chips. For this study, they selected 1,139 STS sequences for analysis by both sequencing and DNA chip hybridization. They found a total of 279 candidate SNPs distributed across 239 of the STS sequences, yielding a SNP in roughly every 1000 bp of DNA screened. Among the SNPs identified, the ratio of transitions to transversions was 2:1. In addition, 25% of changes occurred in CpG dinucleotides even though they made up only 2% of the sequence surveyed. Almost all of the changes were C → T transitions.

This project also involved using DNA arrays to survey STS sequences for SNPs. This was done by establishing 25-bp oligomers in groups of 4 with position 13 of each oligomer representing one of the four bases. By knowing the nucleotide occurring at this position based on the known reference sequence, variability could be detected as a change in the expected hybridization pattern. They identified 2748 SNPs in this manner, with a SNP occuring once every 721 nucleotides. Among these SNPs, the mean heterozygosity was 33% and the mean frequency of the minor allele was 25%.

In addition to the identification of SNPs, Wang et al. used chip hybridization to genotype individuals for the collection of SNPs they had previously identified. They were easily able to simultaneously test for 558 SNPs on one chip. They established two tiles for each SNP, one for each allele. The oligonucleotide arrays again consisted of 25-mers that were complementary to one of the two alleles at position 13. The individual's DNA to be hybridized was synthesized using specific PCR primers with uniform sequence on each end to allow batch labeling of all PCR products. They were able to perform multiplex

PCR on all 558 loci in a single PCR reaction and make allele determinations for each of 3 individuals tested at 50% of the loci tested. When dividing the loci into 24 sets of 23 primer pairs each, they were able to make allele determinations for all three individuals tested at 92% of the loci tested. They have thus demonstrated the feasibility of using chip hybridization to perform large-scale genotyping of hundreds of SNPs simultaneously. Another group (Lai et al., 1998), examined the region around the human APO E gene for the presence of SNPs with results similar to those of the other studies listed. However, they analyzed a contiguous stretch of DNA and confirmed that the development of a high-density SNP map (with SNP markers spaced every 30 kb) was feasible given current technology.

Similarly, Cargill et al. (1999) used DNA chip hybridization along with denaturing HPLC to identify SNPs occuring in the coding regions and adjacent sequences of 106 candidate genes for caridovascular disease, endocrine disease, and neuropsychiatric disease. They searched a total of 196.2 kb of DNA and identified 392 cSNPs and an additional 168 SNPs in the adjacent noncoding sequence. They found a SNP at a frequency of one SNP per 346 bp in the coding region and one SNP per 354 bp in the noncoding region. They calculated nucleotide diversity to be 0.0005 in coding regions and 0.00052 in noncoding regions.

In addition, they were able to examine the cSNPs for occurrence of synonymous vs. non-synonymous nucleotide changes. They found that roughly half of the cSNPs were of each type with 207 cSNPs being synonymous and 185 cSNPs being non-synonymous.

Since roughly two thirds of all random nucleotide mutations would be expected to alter the amino acid sequence of the encoded protein, they argue that there is strong selection against non-conservative DNA mutations. In fact, they calculate that non-synonymous nucleotide changes survive at only 38% of the rate of synonymous nucleotide changes (Cargill et al., 1999). Based on their data, they conclude that the average gene contains approximately 4 SNPs in their coding regions, each of which occur at a frequency of at least a few percent in the human population. By extrapolating these data, they would estimate the number of cSNPs in the human genome to be between 240,000 and 400,000. More recent estimates of the number of genes in the human genome would push this number down to between 120,000 and 160,000 (Venter et al., 2001).

In a companion study, Halushka et al. (1999) examined the coding sequences and adjacent sequences for SNPs in 75 candidate genes for essential hypertension by chip hybridization and gel-based sequencing. They surveyed a total of 28 Mb of DNA, 190 kb in 148 alleles. They identified a total of 874 SNPs, of which 387 were cSNPs. The nucleotide diversity from the data of Halushka et al. are very close to those for Cargill et al. (1999), with the nucleotide diversities reported by Halushka et al. being 0.00045 for coding regions and 0.00054 for noncoding regions.

In another series of experiments, an area of either 9.7 kb or 24 kb of contiguous DNA was sequenced around the lipoprotein lipase or angiotensin converting enzyme genes, respectively (Nickerson et al., 1998; Clark et al., 1998; Rieder et al., 1999). In the first set of experiments (Nickerson et al., 1998 and Clark et al., 1998), researchers sequenced

9.7 kb of DNA within the lipoprotein lipase gene in a total of 71 individuals. The individuals were African-American (24 individuals) Eurpoean (24 individuals) and European-American (23 individuals). They found a total of 79 SNPs, of which, 47 were transversions. They also found 9 insertion/deletion variations. There were 7 variable sites in the coding region, a stretch of 998 bp of DNA, with the remaining 81 variable sites in the 8,736 bp of noncoding DNA. This gave a nucleotide diversity of 0.002 in the entire sample and 0.0005 in the coding region.

In the second study (Rieder et al., 1999), the investigators sequenced 24 kb of DNA around the DCP1 gene, which encodes angiotensin converting enzyme. They did this in six individuals of European descent and 5 individuals of African descent. They identified a total of 78 varying sites on 22 chromosomes. They found the nucleotide diversity to be 0.00093 overall. Using a combination of techniques, they were able to determine that there were 13 distinct haplotypes among the individuals tested.

Taken together, these studies support one another and likely provide a reasonable estimate for the nucleotide diversity across the human genome. They also validate the chip hybridization approach as both a method of SNP screening and genotyping.

The work below follows these projects in several respects. A method was developed to systematically scan coding and noncoding regions of various canine genes for the presence of SNPs in a pool of ten dogs of different breeds. From these results, an estimate for nucleotide diversity was calculated for the canine genome (Chapter 3). At

the time of this work, a limited amount of canine nucleotide sequence data was available, and the sequencing data that resulted from the SNP search also represented newly cataloged sequence data for the canine genome (Genbank accession numbers in Chapter 3). Since that time, the complete nucleotide sequence of the canine genome has become available (Kirkness et al., 2003, Lindblad-Toh et al., 2005). The sequencing of the canine genome led to the identification of 2.5 million SNPs within the canine genome (Lindblad-Toh et al., 2005).

SNPs and Gene Mapping

SNPs are the most abundant form of polymorphism known to exist in the genome. Like any type of genetic marker, family-based linkage studies can be performed using SNP-based markers. One disadvantage of SNP markers compared to the more commonly used SSLP markers, is that the informativeness of the markers is less than that of SSLP markers, due to the fact that SNP markers are biallelic, whereas SSLPs generally have several alleles. With only two alleles, the maximum heterozygosity is 0.50. In contrast, SSLP markers have a heterozygosity that typically ranges from 0.65-0.80 (Kruglylak, 1997). However, the greater abundance of SNP markers easily makes up for this shortfall in heterozygosity because several can be combined to increase informativeness.

Kruglyak (1997) set out to test the feasibility of performing whole-genome linkage searches using SNP markers. In his computer modeling, he reached several key conclusions. First, a map of biallelic markers with a density of 2.25-2.5 times that of a

microsatellite map provides comparable information content. Thus, a 4 cM map of biallelic markers is comparable to a 10 cM map of microsatellites.

Next, the frequencies of the two alleles do not have a great effect on the information content of the map of biallelics as long as the frequency of the rare allele is 0.2 or greater. Thus, perfect "50/50" alleles are not required for an informative map.

Finally, the abundance of the SNPs in the genome makes development of large numbers of markers to create a very dense map of the genome (1 cM or less) theoretically and technically feasible. In fact, if current estimates hold, there should be on the order of 10 million SNPs in the human genome. While testing such large numbers of markers in family based linkage studies is technically daunting, methods are being developed to increase throughput to make such genotypings feasible (Wang et al., 1998; see above).

It has been suggested that one of the true breakthroughs in genetics that SNPs will allow to come to pass is the mapping of genes conferring risk for complex diseases (Risch and Merikangas, 1996; Collins et al., 1997; Kruglyak, 1999). Risch and Merikangas examined the possibility of detecting genes conferring a genome relative risk (GRR) between 1.5 and 4. (GRR is defined as the increased chance that an individual with a particular genotype has the disease.) They conclude that disease susceptibility alleles with moderate frequency in the population (p is 0.1 to 0.5) that confer a GRR of 4 or greater will be detectable by family-based linkage analysis. However, for disease

susceptibility loci with GRR of 2 or less, the number of families needed to detect linkage would exceed 2500 and thus be practically unachievable.

They suggest that association analysis is a much better approach in this case. Instead of family-based linkage analysis, association analysis would be performed using affected sib-pairs or single affected individuals and their parents. Association analysis would then be performed based on inheritance of a given allele or associated marker in affected individuals as compared to appropriately selected controls. A significant deviation from random inheritance based on allele frequencies would be suggestive of association between the marker under consideration and the disease susceptibility allele. Similar calculations could be performed based on inheritance of a given allele or marker from unaffected parents to affected offspring. An inheritance of a given allele or marker that was significantly greater than 50% would be suggestive of association between the marker and the disease susceptibility allele.

Two approaches to whole-genome association analysis have been suggested (Collins et al., 1997). The direct method involves characterizing the approximately 25,000 genes in the human genome to identify SNPs in the coding regions (cSNPs) of these genes. It is assumed that the SNPs resulting in an amino acid change in the encoded protein will be directly responsible for disease susceptibility. The tests would directly examine these coding changes for association with disease susceptibility (Collins et al., 1997). In fact, many investigators have begun identifiying these cSNPs within the human genome (Picoult-Newburg et al., 1999; Cargill et al., 1999; Halushka et al., 1999).

Kruglyak (1999) has done computer modeling to assess the feasibility of whole-genome association analysis using an indirect approach. The indirect approach would rely on linkage disequilibrium (LD) between the variable site which confers the disease susceptibility and tighly linked markers. However, it has not been established what levels of linkage disequilibrium can be generally expected across the human genome. Based on his modeling, Kruglyak suggests that useful levels of LD are only on the order of a few kilobases in the outbred human population. This implies that it would take 500,000 SNPs to undertake whole-genome association studies in outbred human populations. He also suggests that similar numbers of SNPs would be required in isolated populations unless the founding population is very small (effective size of 10-100 unrelated individuals). The assertions of Kruglyak have been controversial. Collins et al. (1999) examined linkage disequilibria between 1000 pairs of loci and found that LD was on the order of 300 kb throughhout the human genome. They assert that unlike the computer models of Kruglyak which simulated the human population as steadily expanding to its current size, the human population has gone through a series of expansions and contractions over its existence. The contractions, due to events such as epidemics, famines, massacres and pressure from technologically more advanced or more aggressive neighbors would result in greater LD than the model suggested by Kruglyak. Collins et al. conclude that as few as 30,000 SNP markers, 1 per 100 kb of DNA, may be sufficient to perform whole-genome association analysis in the human genome.

More recent work by The International HapMap Consortium (2005) indicates that there is much more linkage disequilibrium in the human genome than simple modeling studies would indicate. The HapMap Consortium obtained complete DNA sequences from 269 individuals froom four different human populations, including ten 500kb regions in which essentially all common DNA variation was determined. This study, in addition to identifying more than 1 million SNPs, found that

Ostrander and Kruglyak (2000) performed computer modeling to evaluate the feasibility of association analysis in the various dog breeds. They concluded that LD mapping is practical given the current state of the canine linkage map, with microsatellite markers spaced an average of 8.86 cM apart (Ostrander and Kruglyak, 2000; Werner et al., 1999). Indeed, they herald some characteristics of purebred dogs that make them intriguing for LD mapping. First, gene flow between breeds is limited by the pedigree structure. (Registration of a dog as a member of a given breed requires that both his parents be registered members of the same breed.) The modern dog breeds are relatively young, with most being developed in the last 300 years (Wilcox and Walkowicz, 1995;Wayne and Ostrander, 1999; Ostrander and Kruglyak, 2000). Many breeds have a small founding population. Popular sires have decreased the effective population size of the breeds. Finally, for many breeds, the breed's natural history has been such that severe population bottlenecks have occurred in the recent past (Ostrander and Kruglyak, 2000). All of these factors combine to increase the area of linkage disequilibrium in the various dog breeds. For example, Ostrander and Kruglyak (2000) performed computer modeling on the Rottweiler breed. Based on pedigree data provided by the American Kennel Club

and breed history (Wilcox and Walkowicz, 1995), they estimated that there will be high levels of LD extending 5-10 cM around a disease mutation (Ostrander and Kruglyak, 2000). They further propose that screening a sample of 40 affected dogs for identity by descent will be sufficient for gene localization. While the above analysis is specific to Rottweilers, further modeling indicates that similar areas of LD will exist even in breeds that haven't suffered the types of severe population bottlenecks as those of the Rottweiler (Ostrander and Kruglyak, 2000). Similar results have been demonstrated by Lindblad-Toh et al. (2005) for the Boxer and in five different breeds by Sutter et al. (2004).

Methods of SNP Identification

Since the development of RFLP markers (Botstein et al., 1980) it has been known that there was nucleotide variation within mammalian genomes. When the search for markers was first undertaken, the only method available was to isolate a cloned gene fragment for use as a probe and perform restriction digestion with as many different restriction enzymes as were necessary to locate an RFLP marker. One of the drawbacks of this method is that even performing endonuclease digestion with all the restriction enzymes available today, only about 50% of the SNPs would be identified as RFLPs. In fact, Nickerson et al. (1998) report that if they had performed restriction digestion on their target DNA with all of the restriction enzymes with either five- or six-base specificities (Roberts and Macelis, 1997), only 34 of their 88 variable sites would have been discovered.

With the advent of high throughput DNA sequencing techniques, methods of SNP identification have been developed using DNA sequencing. Direct sequencing has the advantage of examining all nucleotides in a sequencing run for the presence of SNPs. Its other advantage is that only DNA sequencing will precisely define both the location and the exact nature of the DNA variation detected (Kwok et al., 1994). The major disadvantage has been the high cost associated with sequencing the DNA of several individuals within a population under study in order to locate variable nucleotide sequences.

We and others, most notably Kwok's research group at Washington University, have developed a method of identifying SNPs by pooling DNA for sequencing (Chapter 3, Taillon-Miller et al., 1999). This has the advantage of simultaneously surveying several copies of DNA sequence for the presence of nucleotide variability while reducing the cost to that of just two sequencing reactions.

One early effort to identify SNPs in the human genome by Kwok et al. sought to utilize the large overlapping clones already available from the human genome project and inspect these sequences for nucleotide variability (Taillon-Miller et al., 1998). Where no nucleotide sequence information is available, one must develop STSs and then sequence the DNA from several individuals in order to identify SNPs found in that area of the genome. Kwok et al. (1994, 1996) were performing automated sequencing of the DNA from 4 individuals plus a pooled DNA sample for allele frequency estimates. This

strategy enabled them to identify with > 85% probability all the SNPs that occurred in the regions sequenced at a frequency of greater than 20% (Kwok et al., 1994).

Kwok et al. (1996) then applied this technique on a larger scale by scanning a series of STS markers for the presence of SNPs. They obtained primers for 194 STSs from the Whitehead Institute's collection of 838 STSs (as of July, 1994). They were able to amplify DNAs from 154 of the primer sets in four individuals and a pool of 80 individuals, and examine the amplified DNA for SNPs as given above (Kwok et al., 1994, 1996). They found 39 SNPs among the 154 STSs tested and estimated that a polymorphism occurred at a frequency of once per 791 bp, similar to the SNP frequencies reported above.

Taillon-Miller et al., (1997, 1999) further refined this method. First, they used a complete hydatiform mole (CHM) to serve as a sequencing control (Taillon-Miller et al., 1997). A CHM is the product of an abnormal conception. It is generally the product of the union of an enucleated ovum with a single sperm cell that later duplicates its genome to give a diploid tumor (Taillon-Miller et al., 1997; Grimes, 1984; Kajii and Ohama, 1977). Since the genome of the mole is from a single haploid sperm cell that has undergone a duplication event, every nucleotide position should be homozygous in the CHM. This serves as a control reaction in that it allows false positive SNPs resulting from amplification of duplicated sequences in the genome to be distinguished from true SNPs. It is estimated that the worldwide incidence of hydatiform moles in humans is one

per one thousand pregnancies (Taillon-Miller et al., 1999; Grimes, 1984). Thus, they argue that sample material should be available for all populations of interest.

With the improvement in dye-labeled dideoxy chain terminators, Taillon-Miller et al. (1999) now recommend sequencing only two DNA samples in parallel. These are the CHM DNA as a control and a pool of 80 individuals. They found that they could cut the number of sequencing reactions by 60%, from five parallel sequencing reactions to just two and still identify SNPs with the same sensitivity as separately sequencing the four individuals' DNA as was done previously.

At the time of this work, several other methods of SNP identification had been developed. These methods have been reviewed by Kwok and Chen (1998) and are briefly outlined below. From the time of completion of this research to present time, the availability of automated sequencing has virtually eliminated the use of these techniques to identify SNPs. They are included for the purpose of placing the work completed here in the context of the time it was completed.

SSCP: SSCP is single strand conformational polymorphism. The technique is based on the fact that single stranded DNA will form a unique tertiary structure based on its DNA sequence (Kwok and Chen, 1998). Any changes in nucleotide sequence will change the tertiary structure of the molecule. When these single stranded molecules are electrophoresed on a native gel, molecules with sufficient differences in conformation will migrate at different rates and can be distinguished on the gel. The advantage of this

technique is its technical simplicity. The disadvantages are that target molecules in which a polymorphism are to be identified must be smaller than 300 bp for differences in single nucleotides to sufficiently influence conformation so as to be resolvable on the gel and the need for multiple buffer conditions to achieve 90% sensitivity.

DGGE: DGGE is denaturing gradient gel electrophoresis. It is based on the fact that denaturation of double stranded DNA is sequence dependent. A difference in a single nucleotide between two DNA molecules often causes a great enough difference in their denaturation temperatures to distinguish between the two molecules. When a partially-open DNA molecule is migrating through a gel, it is for all intents and purposes immobilzed at the site where one end first denatures. Thus, DNA molecules with different low-melting domains will have different final positions in the gel. When heteroduplex DNA is run on a denaturing gradient gel, the heteroduplex DNA will denature at a concentration of denaturant that is much lower than its homoduplex counterpart. This forms the basis of the detection procedure. The advantage of this procedure is its ability, with some modification of the basic procedure above, to locate polymorphisms in DNA fragments as large as 1000 bp. Its disadvantage is the need to use specialized equipment to perform the analysis.

Methods of SNP Detection

Once a SNP has been identified, the next step is to develop a means to genotype individuals at the marker. The "gold standard" of SNP detection is to use allele-specific

restriction digestion and gel electrophoresis to genotype individuals for a given SNP. This method has the advantage of being highly accurate, technically reliable, and inexpensive. The disadvantages are that the method is labor intensive and has a low throughput. Several other methods to speed throughput have been developed (Landegren et al., 1998). The current methods all use amplification by PCR followed by allele determination by allele-specific hybridization or allele-specific restriction digestion, determination of mismatched DNA substrates by polymerases or ligases, or by template specific incorporation of nucleotides by polymerases.

There is a great deal of overlap between methods of SNP identification and detection. Certainly, given unlimited budgets, DNA sequencing could be used for SNP detection. Other methods, such as SSCP, DDGE, and heteroduplex analysis could be used to determine if polymorphism existed in a given individual. They may even be used to determine which alleles were present in an individual, with the inclusion of appropriate control reactions.

DNA chip hybridization is best suited for high throughput SNP detection. It has the advantage of being able to genotype an individual at thousands of polymorphic sites simultaneously (Wang et al., 1998; Chee et al., 1996; Landegren et al., 1998). Its main disadvantage at the time this work was undertaken was its high cost. Since the completion of this work, the cost and availability of DNA chip hybridization technology had decreased and become more reliable, putting it within the budget of most laboratories.

Canine Breed History

In the series of experiments detailed below, DNA from ten different dog breeds were used to form a working pool of DNA. The breeds making up this pool were chosen because they differ in size, behavior, and temperment. Presumably, genetic variation among this pool of DNA will reflect such variation. A summary of breed characteristics is given in Appendix 1-1 following this chapter. Breed histories are also included in Appendix 1-2 at the end of this chapter.

I have published the following papers during the course of this work. Chapter 2 was previously published as Venta, et al., "Gene-specific Universal Mammalian Sequence Tagged Sites: Application to the Canine Genome", *Biochemical Genetics* **34:** 321-341 (1996). In this work, I designed approximately 20% of the primer pairs, performed all of the DNA amplifications, and performed all of the sequencing reactions within the paper.

Chapter 3 was previously published as Brouillette et al., "Estimate of Nucleotide Diversity in Dogs with a Pool-and-Sequence Method", *Mammalian Genome* **11:** 1079-1086 (2000). In this work, I performed approximately 90% of the experiments.

Chapter 4 was previously published as Brouillette and Venta, "Within-breed Heterozygosity of Canine Single Nucleotide Polymorphisms Identified by Across-Breed

Comparison", *Animal Genetics* **33:** 464-467 (2002). In this work, I performed all of the experiments.

In addition, I have coauthored four other papers. They have been previously published as follows: 1. Brouillette et al., "*Bsl I* PCR/RFLP Marker in the Canine Connexin 40 Gene", *Animal Genetics* **30:** 229 (1999), in which I performed about 75% of the experiments; 2. Brouillette and Venta, "*Tth I* PCR/ RFLP Marker in the Canine Rod Transducin Alpha Gene", *Animal Genetics* **31:** 68 (2000), in which I performed all of the experiments; 3. Lingaas et al., "A Canine Linkage Map: 39 Linkage Groups", *J. Animal Breeding and Genetics* **118:** 3-19 (2001), in which I performed segregation analysis for 7 of the 222 markers, as part of the DogMap consortium; and 4. Ernst et al., "Mapping of FES and FURIN Genes to Porcine Chromosome 7", *Animal Genetics* **35:** 142-167 (2004), for which I provided PCR primers and amplification conditions for the mapping of the FES gene.

Appendix 1-1

Summary of physical characteristics of dog breeds[a]

Breed	Height in inches	Weight in pounds	Coat color	Fur Style	Class[b]
Am. Cocker Spaniel	15	24-28	Black, tan, chocolate, cream, tricolor	Silky, long	Sporting
Greyhound	26-28	65-70	Cinnamon, chestnut, red, black, brindle	Short, smooth	Hound
Doberman Pinscher	26-28	66-88	Black, red, blue, fawn	Short, smooth	Working
Siberian Husky	21-23	45-60	Gray, black, red	Thick, dense	Working
Labrador Retriever	21-24	55-75	Black, chocolate, yellow	Moderately short	Sporting
Collie	24-26	60-75	Sable and white, tricolor, blue merle	Short, smooth, double	Herding
Scottish Terrier	10-11	19-23	Black, brindle, wheat, gray	Wiry	Terrier
German Shepherd	22-26	75-95	Black and tan, black, sable	Short, dense	Herding
Beagle	13-15	55-75	Any color	Short, dense, smooth	Hound
Pointer	25-28	55-75	Liver, lemon, orange, white	Short, dense smooth	Sporting

a. Information is from *The Complete Dog Book*, 1997 and Wilcox and Walkowicz, 1995. Height and weight are for male dogs of each breed. Where differences existed between the references cited, data was from the first reference above. In all cases, the female dog was slightly shorter and lighter than the male dog.

b. "Class" refers to the grouping used by the American Kennel Club in *The Complete Dog Book*, 1997.

Appendix 1-2

Breed History
The history of each of the breeds used in this study is outlined briefly below.

American Cocker Spaniel: The American Cocker Spaniel can trace its roots back as far as the 14th century. In 1368, the Spanyell was first mentioned in the literature. Through the years the spaniel family was divided into two groups, the land spaniels and the water spaniels. As time passed, the land spaniels were divided into the smaller cocker spaniels and the larger varieties. Later, the toy spaniels were divided from the cocker spaniels.

The first registry of the Cocker Spaniel breed was in England in 1892. It was brought to the United States in the 1880s and went through a change in breed standard such that by the 1930s it came to be considered a separate breed from the English Cocker Spaniel from which it originated. It is considered to be a sporting dog, and is reputed to be an excellent hunter. The breed is known for being handsome, happy, eager to please, trusting, and intelligent. These traits have made it one of the most popular dog breeds in the United States (American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

Greyhound: The greyhound can trace its lineage back to ancient times. The first known record of the greyhound dates back to the hieroglyphs of ancient Egypt, around 3000 B.C. The greyhound has long been a favorite of the aristocracy. Documents from 9th century England indicate that it was a favorite hunting dog of the Duke of Mercia.

The earliest accounts of the greyhound in America date back to Spanish explorers in the 1500s. Known as hunters, there are reports of greyhounds running down deer, stags, and foxes. Yet, it is probably best known for its hunting ability for rabbits and hares. Known today for being gentle, well-behaved, and graceful pets, greyhounds are elegant show dogs and thrilling competitors (American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

Doberman Pinscher: The origin of this breed is well established. The breed began in Thueringen, Germany in 1890 by Louis Dobermann. Dobermann, a tax collector by trade, needed a dog to protect him from bandits. The breed mixed the hardiness and intelligence of the German Shepherd, the reaction and fire of the German Pinscher, and the hunting ability of the pointer. Further outbreeding added the Rottewiler's strength, courage, and guarding instinct and the Greyhound's foot speed. In only ten years, the breed standard had been established.

The breed is known today for its intelligence, its ability to absorb and retain training, and its loyalty. It is these qualities that put the breed in demand as a police and military dog American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

Siberian Husky: The Siberian Husky traces its roots to the dogs of the ancient Chukchi people of northeastern Asia. The dog was bred to be a sled dog. Its primary mission was as a dog that would travel great distances at moderate speed while carrying a light load and wouldn't flinch at the subzero temperatures of the Arctic region.

The reputation of this breed of sled dog was made in the United States in 1925. A diphtheria epidemic was sweeping through Nome, Alaska and dogsleds were used to take antiserum from Anchorage to Nome. This serum run was the forerunner of the famous Iditarod dogsled race, and it focused the spotlight on the Siberian Husky.

The Siberian Husky is naturally friendly and gentle. He is an exceptional family dog, and is still the favorite of dog mushers across the United States (American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

Labrador Retriever: The Labrador Retriever was originally seen in the early 1800s in Canada as a hunting dog and was particularly useful at retrieving water fowl. The breed was transported to England on fishing boats and soon became a popular sport dog there as well. Later in that century, the breed was all but eliminated in Canada, due to a heavy dog tax. The breed's development into its current form occurred largely in England.

The breed was first recognized in England in 1903 and in the United States in 1917. This breed is known for its eagerness to please its master and still possesses its sensibility, even-temper, intelligence, and strong marking and retrieving skills. The breed is renowned as a bird flusher, companion, drug-detector, and as a guide dog for the blind. These traits consistently put the breed in the top five in popularity in both England and the United States (American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

Collie: The breed known as the collie is thought to have its origin in the dogs that were brought to Scotland with the Roman invaders of 50 BC. These ancient dogs interbred with other Scottish herding dogs to yield the breed known today. This breed of dogs has been used to herd sheep for centuries in Scotland.

In 1860, Queen Victoria became a fancier of the breed after a trip to Scotland. With her blessing, the breed became a favorite of the aristocracy and affluent, as well as maintaining its traditional role as a herding dog. The two types, rough and smooth (referring to the length of the coat), were fixed enough in characteristics by 1886 that little has changed with regard to the breed standard since that time. By 1877, the breed had become established in the United States, though the breed was first introduced in this country with the early settlers over a century earlier.

The breed is consistently popular today as a family pet. It is known for its loyalty and affection and as a self-appointed guardian of the entire family, but particularly of small children. In recent years, the dog has maintained its popularity, due in part, to the Lad stories of writer Albert Payson Terhune and the "Lassie" movies and television series (American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

Scottish Terrier: The Scottish Terrier has been in existence for centuries. There are those that will argue that descriptions of the Skye Terrier written in the 1570s are not the Skye Terrier that is known today but the Scottish Terrier of antiquity. At the very least, the modern breed can trace its lineage to Scotland in the 1860s. The first standard was established in England in 1880. It has remained the standard, with only minor changes, up to the present day.

The Scotty was first introduced into the United States in 1883. Since this time, there have been thousands of Scotties imported. The terrier temperament is taken to the extreme in Scotties. He is alert, quick, and feisty. These qualities make the breed well suited to being a watchdog and varmint killer. The breed of dog requires discipline to prevent him from becoming a bully (American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

German Shepherd Dog: This breed was founded in 1899 by Max von Stephanitz. It has always been a working dog, originally as a herder, and in a variety of roles today. This breed grew steadily in popularity around the world up to World War I, but the popularity of the breed suffered due to the anti-German backlash in Europe and America following World War I.

The breed is known today for its loyalty, courage, and ability to assimilate and retain training for a number of specific purposes. German Shepherds are often used as guide dogs for the blind, and as police dogs, military dogs, and as a key component of search-and-rescue units. Considered by some to be aloof, the German Shepherd Dog doesn't give affection freely. However, once the dog warms to a person, he is loyal and dedicated, even to the point of giving his life for his master (American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

Beagle: The history of the Beagle breed is cloudy. Some reports indicate that the origin of the Beagle dates as far back as ancient Rome. Other accounts note that the Beagle has been used to hunt hares in Wales for centuries. Modern records of the Beagle date at least to the middle 1700s. Their keen sense of smell and compact size has made them a favorite to hunt rabbits, hares, and foxes, either individually or in packs.

In the United States, the Beagle has been in existence since colonial times. However, these dogs had the look of a Basset Hound rather than that of the Beagle of today. Imports of Beagles from the kennels of Great Britain in the 1880s and 1890s gave rise to the Beagle that is recognized today.

The Beagle is known as a capable hunting dog as well as a playmate for adults and children alike. The breed's inquisitiveness and happy-go-lucky nature have made it a consistent member of the top ten dog breeds in the United States (American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

Pointer: The Pointer breed got its start in England around 1650. These are excellent hunting dogs and were considered to be the first true pointing dogs. They were originally used to hunt hares. The Pointer was sent out to locate a hare, at which time, greyhounds were brought in to chase the hare. During the early 1700s, the pointer's hunting ability was more thoroughly exploited due to the increased popularity of wing-hunting. Legends abound about this breed's pointing ability. One example is the story of a sportsman who lost his dog in the moors of England. He returned a year later to find the skeleton of the dog still pointing at the skeleton of a bird.

The pointer of today is a hunting specialist. He is muscular, courageous, speedy and has great endurance. His ability to concentrate on his job and ability to work with people other than his master keep him as a favorite among hunting dogs (American Kennel Club, 1997; Wilcox and Walkowicz, 1995).

Chapter 2

Gene-Specific Universal Mammalian Sequence-Tagged Sites:

Application to the Canine Genome

Gene-Specific Universal Mammalian Sequence-Tagged Sites: Application to the Canine Genome

Patrick J. Venta[1,2,3], James A. Brouillette[1,3], Vilma Yuzbasiyan-Gurkan[2], and George J. Brewer[4]

Departments of Microbiology[1] and Small Animal Clinical Sciences[2], College of Veterinary Medicine, and the Genetics Program[3], Michigan State University, East Lansing, MI 48824-1314 and the Department of Human Genetics[4], The University of Michigan Medical School, Ann Arbor, MI 48109-0618

Key Words: Genome mapping; evolution; homology; polymerase chain reaction

Corresponding Author:

Patrick J. Venta, Ph.D.

Phone: 517-432-2515

FAX: 517-432-2514

e-mail: venta@cvm.msu.edu

Abstract

We are developing a genetic map of the dog based partly upon markers contained within known genes. In order to facilitate the development of these markers, we have used PCR primers designed to conserved regions of genes that have been sequenced in at least two species. We have refined the method for designing primers to maximize the number that produce successful amplifications across as many mammalian species as possible. We report the development of primer sets for eleven loci in detail: CFTR, COL10A1, CSFIR, CYP1A1, DCN1, FES, GHR, GLB1, PKLR, PVALB, and RB1. We also report an additional 75 primer sets in the appendices. The PCR products were sequenced to show that the primers amplify the expected canine genes. These primer sets thus define a class of gene-specific sequence-tagged sites (STSs). There are a number of uses for these STSs, including the rapid development of various linkage tools and the rapid testing of genomic and cDNA libraries for the presence of their corresponding genes. Six of the eleven gene targets reported in detail have been proposed to serve as "anchored reference loci" for the development of mammalian genetic maps [O'Brien et al., Nat. Genet. 3:103-112, 1993]. The primer sets should cover a significant portion of the canine genome for the development of a linkage map. In order to determine how useful these primer sets would be for other genome projects, we tested the eleven primer sets on the DNA from species representing five mammalian orders. Eighty-four percent of the gene-species combinations amplified successfully. We have named these primer sets "universal mammalian sequence-tagged sites" (UM-STSs) because they should be useful for many mammalian genome projects.

Introduction

Efforts have intensified in recent years to develop comprehensive genomic maps for many eukaryotic species using molecular techniques. Many of these efforts have focused on mammalian species, including human, mouse, rat, ox, sheep, pig, horse, cat, and dog (e.g., Buchanan et al., 1994; Dietrich et al., 1992; Ellegren et al., 1992; O'Brien, 1986; Serikawa et al., 1992; Weissbach et al., 1992; WinterØ et al., 1991; Barendse et al., 1994; and the present report). For the non-human species, these projects should lead to more successful breeding strategies, both for selecting desirable characteristics and for removing genes that lead to various genetic diseases. Comparisons made between these genome maps should also lead to new insights on the mechanisms of chromosomal evolution (e.g., see O'Brien et al., 1993).

We are developing a comprehensive map of the canine genome, with our ultimate aim being to reduce the incidence of canine genetic diseases. In addition to developing random, highly polymorphic genetic markers (Type 2 markers; Yuzbasiyan-Gurkan et al., submitted), we are also developing markers for specific genes (Type 1 markers). An appropriate mix of these two types of markers should maximize our ability to map disease genes.

The traditional method for developing gene-specific markers, Southern blotting and cross-species hybridization, is very time consuming, labor intensive and limited in flexibility. This method has been the mainstay for developing gene-specific markers in

most animal genome projects. There is a need to develop more efficient methods. This is particularly important for animal genome projects where scientific resources are more limited. One method that has excellent potential is cross-species polymerase chain reaction (PCR). This method has been used successfully for the study of a number of individual genes but has not been applied on a genome-wide basis for the purpose of map development. To study a single gene, the cost associated with the failure of a few primers sets to amplify the correct target is negligible and new primer sets can be easily redesigned and synthesized. However, when primer sets are being designed for many genes, the cost for failed primers can become substantial, both in terms of time and other resources, so we have refined the design method to minimize this problem.

We describe here, in detail, eleven primer sets that can amplify gene-specific targets of dogs and other mammalian species. Seventy-five additional primer sets are listed in the Appendices. Because markers based on PCR primers are called sequence-tagged sites (STSs; Olsen et al., 1989), we call these primer sets universal mammalian STSs (UM-STSs) because they should be useful for many mammalian genome projects.

Materials and methods

DNA Isolation

DNA from dog, human, pigtail macaque, horse, pig, rat and mouse were isolated from various tissues by standard phenol-chloroform extraction methods (Sambrook et al., 1989). Goat DNA was kindly supplied by Dr. Karen Friderici, Michigan State University. DNA was purified by standard methods from a canine liver cDNA library (Clontech) and from a canine genomic DNA library (Clontech) after growing 1 x 10^6 phage in *E. coli* strain LE392 (Murray et al., 1977) in liquid culture (Sambrook et al., 1989).

Design of PCR Primers

The method of primer design detailed here was used throughout this series of experiments for primer design, unless the primer was designed based on available sequence data for the species being studied. Primers were designed to genes where the intron-exon structure was known in at least one species and where the nucleotide sequence was known in at least two species (the "index species") that are not closely related. Tandemly duplicated genes known to have undergone gene conversion in any species were avoided. Primers were generally designed so that the amplified product contained an intron. Since the canine sequence was unknown in most cases, the sizes of the canine introns were not known prior to amplification. We have since determined that

the vast majority of canine introns will be between 50% and 150% of the size of the corresponding human intron (Venta, unpublished observations). We have followed the human gene nomenclature system (ISGN, 1987) for naming the canine genes. The eleven loci described in detail in this chapter, and their protein products, are: CFTR, cystic fibrosis transmembrane regulator; COL10A1, type X collagen, alpha 1 chain; CSFIR, colony stimulating factor 1 receptor; CYP1A1, cytochrome P-450 1, alpha 1; DCN1, decorin; FES, c-fes (feline sarcoma) proto-oncogene; GHR, growth hormone receptor; GLB1, beta galactosidase; PKLR, pyruvate kinase - liver, RBC form; PVALB, parvalbumin; and RB1, retinoblastoma protein. The Genbank Accession numbers or reference for the sequence of the two index species for each locus are as follows: CFTR, M55129, M60493; COL10A1, X65120, X65121; CSFIR, X14720, K01643; CYP1A1, Uchida et al., 1990, X04300; DCN1, L01125, Z12298; FES, X06292, J02088; GHR, Z11802, J04811; GLB1, S59584, M57734; PVALB, X63578, M15452; PKLR, S59798, M17088; and RB1, L11910, M26391.

Primers were designed to highly conserved nucleotide sequences contained within coding regions. It is presumed that, in the absence of parallel evolution, regions that are conserved among distantly related mammalian species represent the nucleotide sequence of the most recent common ancestor of the two modern mammals. Thus, any mismatches to the primers should result only from evolution of the canine nucleotide sequence since the divergence from the most recent common ancestor.

Within the areas that were conserved among the two index species, an attempt was made to place the primers in the nucleotide sequence corresponding to the least mutable amino acids (Collins and Jukes, 1994; Jones et al., 1992). While there are slight differences among studies, there is a consensus that Gly, Phe, Tyr, Trp, and Cys are among the least mutable amino acids (Collins and Jukes, 1994; Jones et al., 1992; Dayhoff et al., 1978).

In addition, an attempt was made to choose primers overlying codons with the fewest degenerate codon positions (Li and Grauer, 1987). With only a single codon, Met and Trp are excellent amino acids over which to design a primer. Others that are both rarely mutable and have few codons are Phe, Cys, Lys, and Gln.

The final step was to attempt to place the 3′ end of the primer in the second position of the codon for all codons except glycine, which has greater conservation at the first position rather than the second position (Venta, unpublished observations). An attempt was made to use general principles such as the avoidance of primer-dimer formation as well.

Conservation of amino acids within multigene families was also taken into account, when possible. Where unavoidable nucleotide mismatches occurred between the two index species, the primer sequence was designed to exactly match one of the two which we then call the "primary" index species. GC-rich genes were generally avoided due to the amplification difficulties that can occur, even with exactly matching primers. Primers were twenty bp in length on average. Each primer in a pair was adjusted to be of

approximately the same annealing temperature (Breslauer et al., 1986). All sets of primer pairs were designed to have approximately the same annealing temperature as well, in anticipation of performing multiplex amplifications. It was not always possible to follow every rule for every gene, given the actual circumstances; however, the majority of the rules were generally applicable. Primers were synthesized by either the Michigan State University Macromolecular Structure Facility or the University of Michigan DNA Synthesis Facility.

PCR Amplifications

Correct design and syntheses of the primers were examined by amplifying the DNA from the primary index species. Standard buffer, nucleotide, and primer concentrations were 50 mM Tris-HCl (pH 8.3 at room temperature), 50 mM KCl, 1.5 mM $MgCl_2$, 200 μM dNTPs, 0.1 μg of each primer, and 0.5 - 1.0 μg of target DNA in a 25 μl reaction. Reactions were routinely boiled for three min prior to the addition of 2.0 U of Taq DNA polymerase. Optimal cycling conditions for the amplification of canine genomic DNA were usually found by testing one of several sets of conditions in general use in the lab (see Appendix 2-5). Occasionally it was necessary to use "hot-start" conditions (Bassam and Caetano-Anolles, 1993) in order to get stronger, cleaner amplifications. The presence of an amplification product was determined by electrophoresis of a portion of the reaction on a 1% agarose TBE gel (TBE = 90 mM Tris, pH 8.3, 90 mM sodium borate, 2.5 mM EDTA) followed by staining with ethidium bromide.

DNA Sequence Analysis

The identity of each amplified canine gene was confirmed by "single pass" direct sequencing of PCR products using Sequenase or Taq cycle sequencing kits (United States Biochemical Corp., Cleveland). The PCR products were gel purified with Qiaex (Qiagen Corp., Chatsworth, CA) or by elution from polyacrylamide gel slices (Bergenhem et al., 1992) prior to their use in the sequencing reactions. The canine sequences were visually aligned with the sequences of the other species used to design the PCR primers in order to verify the degree of sequence identity.

Results

The primer sets for the various UM-STSs reported here are given in Appendix 2-7 and efficient amplification conditions for the canine genes are given in Appendix 2-5. It is probable that these conditions could be optimized further (e.g., reduction in the time in each cycle). However, the conditions reported here were found to work effectively while minimizing the number of conditions that had to be examined. A representative gel showing amplification of the canine target DNA along with the human target DNA is shown in Appendix 2-1. The human target serves as a positive control for the amplification system because these primers were designed to exactly match the human sequence. The ability to quickly screen genomic and cDNA libraries for the presence of sequences is also demonstrated in Appendix 2-1. The genomic clones for GHR, COL10A1, and DCN1 (a very faint signal, stronger on other gels [data not shown]) are present in this particular canine genomic library. The presence of a decorin cDNA clone (encoded by the DCN1 locus) in the canine liver cDNA library is shown by the presence of the 122 bp band; cDNA clones for GHR and COL10A1 are not present. The DCN1 PCR product from the cDNA library was sequenced and its identity confirmed (see Appendix 2-2). The human and canine genomic bands have different sizes for GHR and DCN1 because of the intron size differences. The size for the COL10A1 PCR product is the same between the species because an intron was not spanned for this is the UM-STS. Although the PCR product bands in Appendix 2-1 are unique, a few UM-STS-species combinations sometimes contained one to several non-specific amplification products. This is a minor problem with unique sequence primers, because it is almost always

possible to deduce the correct band based upon staining intensity, and on the similarity in size compared to the band of the primary index species.

The amplified products for all of the canine loci were sequenced to confirm their identity and the results are shown in Appendix 2-2. The degree of identity between the canine and index species sequences for each locus is within the range generally accepted (roughly 70 to 100%) as demonstrating homology between the genes of mammalian species (Li and Grauer, 1987). These results support the hypotheses that the canine PCR products are homologous to the respective index species' genes. The canine COL10A1 sequence matched the human and mouse sequences to a similar extent (data not shown). The sequences for PKLR and CYP1A1 exactly matched previously published canine coding sequences (Whitney et al., 1994; Uchida et al., 1990); the sequence for canine FES is given in Appendix 2-3. Although the majority of the canine sequence for PVALB is from an intron, we believe the degree of sequence identity from this region is sufficient evidence to confirm that the PCR product is from the correct canine locus. As expected, the canine sequences tend to show greater identity with the human sequences than with the rodent sequences because of the faster evolutionary rate of the rodent genome (Gu and Li, 1993). A microsatellite repeat was found within the amplified product itself for RB1. Preliminary results show that the RB1 repeat, $(GA)_{12(avg)}$, has moderate genetic variability within several canine breeds.

We hypothesized that each primer set should work for many different mammals, given the evolutionary rate at which nucleotide substitutions occur (Li and Graur, 1987) and the

number of primer nucleotide mismatches that can be tolerated by PCR. We tested the 'universal' utility of these primers on the DNAs from mammals representing several different orders. We used the same reaction conditions that were found to amplify the canine sequences. We have termed these reactions "Zoo PCRs." Appendix 2-4 shows a representative experiment. The FES proto-oncogene was amplified from all of the DNAs examined. These PCR products were purified and sequenced directly without subcloning (see Methods and Materials). The sequences are tabulated in Appendix 2-3. The degree of sequence identity makes it highly likely that the canine PCR products are all homologous to the corresponding index species' genes. The pattern of nucleotide interchange is also what would be expected for homologous genes; members of the same mammalian order share more sequence similarity with one another than with those of other orders.

The data for the Zoo PCRs for the other UM-STS primer sets reported in this paper are given in Appendix 2-6. Greater than eighty-four percent of the targets, excluding the index and canine species, amplified under the single condition used to amplify the canine sequence. These species represent five different mammalian orders; primates (human and macaque), carnivores (dog), artiodactyls (goat and pig), perissodactyls (horse), and rodents (mouse and rat). Limited experiments on other members of these orders (e.g., cat and ox) produced similar results (data not shown). Lack of amplification for DCN1 for one of the artiodactyls (goat) would be predicted because there are four mismatches between the UM-STS primers and the sequence of the closely related bovine DCN1 (Day et al., 1987). We have found it difficult (although not impossible) to amplify DNA using

primers that contain more than two mismatches with the target, when using 20-mers (P.V., unpublished results). It is likely that the homologous gene from at least some of the non-amplifying species would appear using these primer sets if other PCR conditions were examined.

Discussion

This study has shown the feasibility of generating a series of UM-STSs, useful for studies of many genomes, and addressed methodological considerations for their development. UM-STSs should serve as useful tools both for amplifying regions of interest from genomes as well as for isolation of clones from genomic and cDNA libraries and for cross-species comparisons. The data reported in this paper indicate that approximately 85% of all carefully designed UM-STSs will be useful for any given mammalian species. We believe that this method is far more efficient, less costly and considerably less labor intensive than traditional hybridization and Southern blotting-based methods. An additional important benefit is that the information for the necessary reagents (i.e., the primer sequences) is transmitted much more easily and quickly than the clones that are necessary for Southern blotting.

UM-STSs will also be useful for developing genetic markers within various genomes. We have found a microsatellite within one of the eleven loci reported here (RB1) and have found other microsatellite repeats associated with genomic clones isolated through the use of UM-STSs (unpublished results). Single site variability should also be found directly in at least some of the amplified products by using one of a number of techniques developed for scanning for variability, such as the single-strand conformation polymorphism technique. For example, this method has been used to find two polymorphic sites in a study of the canine ALAS*2* gene in a PCR product of a size similar to those reported here (Boyer et al., 1995). If the frequency of single site polymorphic

variability for other mammals is as high as that estimated for humans (roughly one in 200 to 400 nucleotides), then a significant portion of UM-STSs will have these sites. We are currently screening for this variability in the canine genome to estimate the frequency of such variation in the dog. It will be necessary to screen each species individually for genetic variability. However, the availability of previously designed UM-STS primer sets, such as those reported here, should make this work proceed more rapidly compared to the traditional method.

An example of the utility of cross-species comparisons is given by the case of Waardenburg syndrome. The clue to the location of one of the human Waardenburg syndrome genes--well known for causing a syndromic hearing loss--was first gleaned from comparative mapping with the mouse (Asher and Friedman, 1990). The map locations in the mouse suggested possible locations of the human disease gene, one of which eventually was proven correct (e.g., Morell et al., 1992). Because the identity of the gene in the mouse was not known at the time, this approach might more properly be called a 'positional candidate' approach. UM-STSs will be useful for rapidly producing mammalian genetic maps so that the positional candidate approach can be applied to more species.

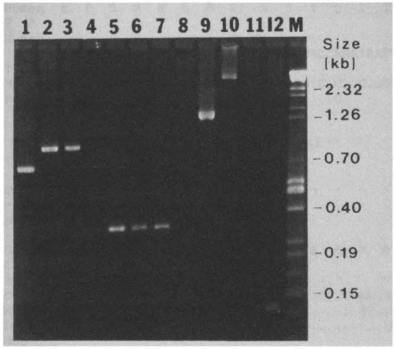
Very little is known about location of genes within the canine genome. Indeed, except for genes located on the X-chromosome (Meera-Khan, 1984; Deschenes et al., 1994) and a few small unassigned linkage groups (Meera-Khan, 1984), the rest of the genome has remained unexplored. New linkage groups are being developed by us and others

(Holmes et al., 1992; Ostrander et al., 1993; Rothuizen et al., 1994; Yuzbasiyan-Gurkan et al., submitted) based primarily on simple sequence repeats. The development of UM-STSs should help to rapidly identify the location of linkage groups on specific canine chromosomes. The identification of conserved syntenies will allow candidate linkages to be tested in the canine genome. The assignment of the proposed anchor loci (O'Brien et al., 1993) as defined by UM-STSs to specific chromosomes can be accomplished by the somatic cell hybrid, flow sorted chromosome, and fluorescent in situ hybridization (FISH) methodologies. Other methods, such as assignment by use of linkage to previously mapped loci, are also possible. We have already assigned several genes by FISH to canine chromosomes using cosmids isolated with UM-STSs (Fujita et al., in press). Using the methods described in this paper, we have developed a much greater number of UM-STSs that should cover, for linkage mapping purposes, a substantial portion of the canine and other mammalian genomes (see Appendices 2-8 and 2-9).

Acknowledgements

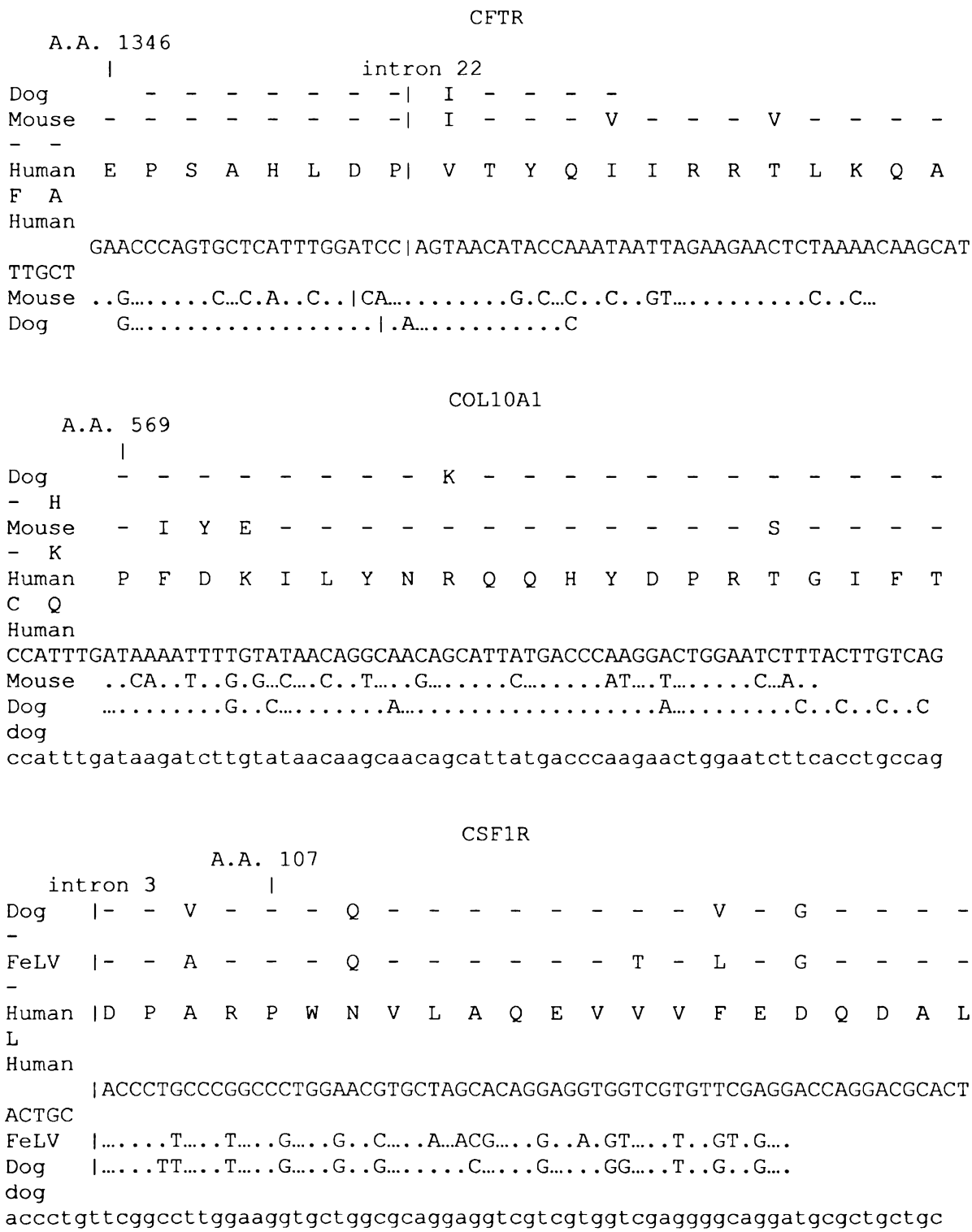
We thank Ya Shiou Yu and Murat Gurkan for their valuable technical assistance. We also thank Tracy Hammer, Neal Dittmer, Jessica Nadler, Elizabeth Tullett, Marc Crotteau, and Kristen Penner for their contributions to the development of a number of the primer sets. This work was supported by the American Kennel Club, the Orthopedic Foundation for Animals, and the Morris Animal Foundation. We also thank the Washington Regional Primate Facility for supplying the pigtail macaque tissues to Dr. Richard E. Tashian, with whom one of us (P.V.) originally isolated the DNA.

Appendix 2-1



Amplification of several canine gene segments using UM-STSs. The following lanes were amplified with the gene-specific primer sets (see table 1): lanes 1-4, GHR; lanes 5-8, COL10A1; and lanes 9-12, DCN1. Lane 13 contains a mixture of DNA size markers; λ bacteriophage cut with the restriction endonuclease BstEII and the plasmid pSK- (Stratagene) cut with MspI. Lanes 1, 5, and 9 contain PCR products amplified from human genomic DNA. Lanes 2, 6, and 10 contain PCR products amplified from canine genomic DNA. Lanes 3, 7, and 11 contain PCR product amplified from DNA purified from a canine genomic library contained in a λ bacteriophage vector. Lanes 4, 8, and 12 contain PCR products amplified from a canine liver cDNA library.

Appendix 2-2


CFTR
A.A. 1346
intron 22
Dog - - - - - - - -| I - - - -
Mouse - - - - - - - - -| I - - - V - - - V - - - -
- -
Human E P S A H L D P| V T Y Q I I R R T L K Q A
F A
Human
GAACCCAGTGCTCATTTGGATCC|AGTAACATACCAAATAATTAGAAGAACTCTAAAACAAGCAT
TTGCT
Mouse ..G….....C…C.A..C..|CA…….....G.C…C..C..GT…….....C..C…
Dog G…..................|.A…….....C
COL10A1
A.A. 569
Dog - - - - - - - - - K - - - - - - - - - - - - - -
- H
Mouse - I Y E - - - - - - - - - - - - - - S - - - -
- K
Human P F D K I L Y N R Q Q H Y D P R T G I F T
C Q
Human
CCATTTGATAAAATTTTGTATAACAGGCAACAGCATTATGACCCAAGGACTGGAATCTTTACTTGTCAG
Mouse ..CA..T..G.G…C….C..T….G…….....C…….....AT….T…….....C…A..
Dog ….....G..C…….....A…..................A…….....C..C..C..C
dog
ccatttgataagatcttgtataacaagcaacagcattatgacccaagaactggaatcttcacctgccag
CSF1R
A.A. 107
intron 3
Dog |- - V - - - Q - - - - - - - - - - V - G - - - -
-
FeLV |- - A - - - Q - - - - - - - T - L - G - - - -
-
Human |D P A R P W N V L A Q E V V V F E D Q D A L
L
Human
|ACCCTGCCCGGCCCTGGAACGTGCTAGCACAGGAGGTGGTCGTGTTCGAGGACCAGGACGCACT
ACTGC
FeLV |…..T…..T…..G…..G..C…..A…ACG…..G..A.GT…..T..GT.G….
Dog |…..TT…..T…..G…..G..G…..C…..G…..GG…..T..G..G….
dog
accctgttcggccttggaaggtgctggcgcaggaggtcgtcgtggtcgaggggcaggatgcgctgctgc

Appendix 2-2 (cont'd).

DCN1

```
    A.A. 231
       |                                        intron 6
Dog                          - - - - - - - |- - - - - - - - - -
- -
Rat    - - - - - - - - - - N S - - S |- - - - - - - - - -
- -
Human  V D A A S L K G L N N L A |K L G L S F N S
I S
Human
       GTTGATGCAGCTAGCCTGAAAGGACTGAATAATTTGGCTA|AGTTGGGATTGAGTTTCAACAGCA
TCTCT
Rat    ….....C….....A..TC…..T…|..C….T…..C…..T…...A.C
Dog                          ….............|…....C…..T……....N
dog                          ggactgaataatttggcta
agttgggactgagttttaacagcatctc
```

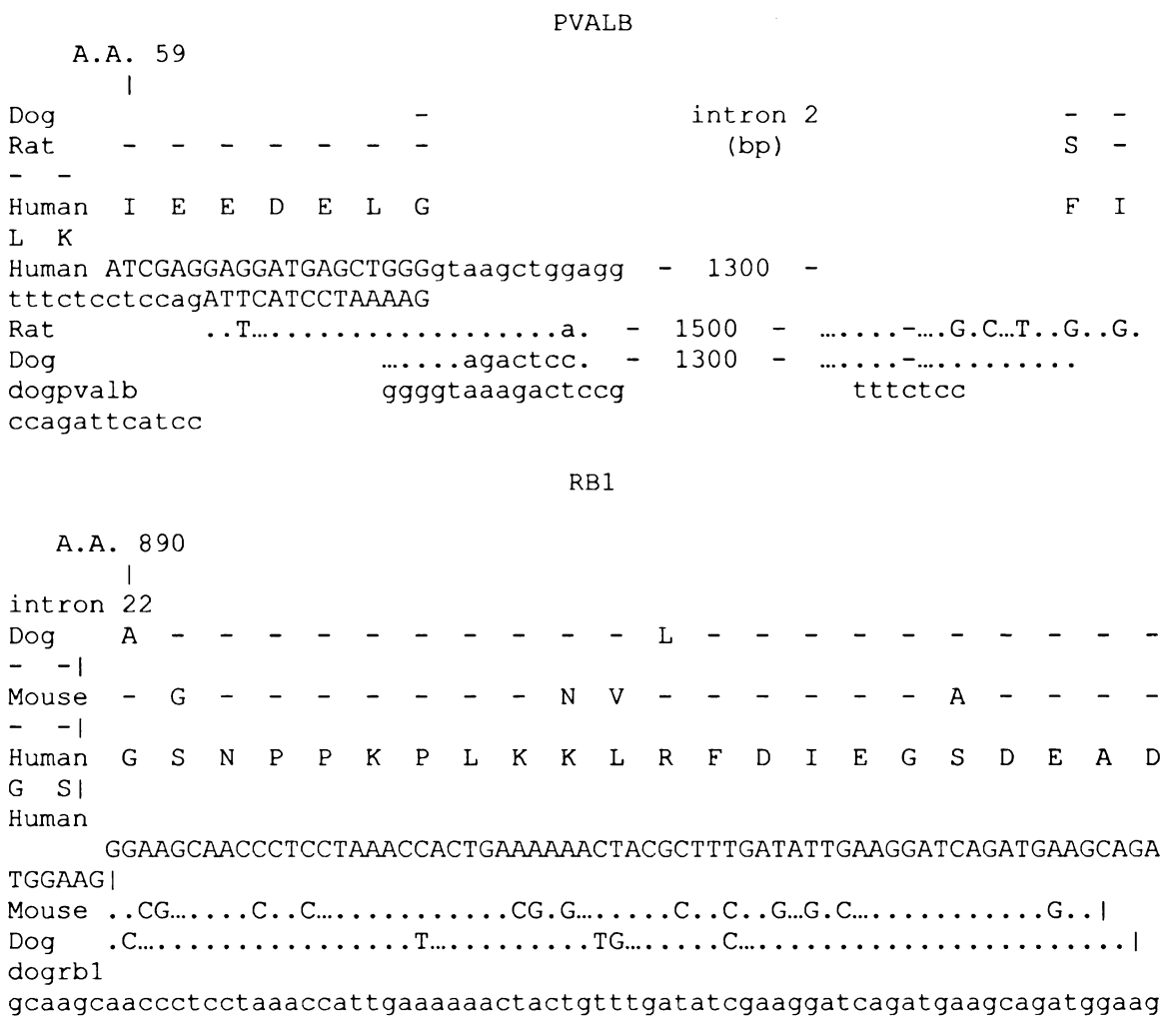
GHR

```
    A.A. 333
        |
Dog      - D L - - - - - - G - - - - - - - - - N - - -
- -
Rat      - D A - - - - - - - - - - - - - - - - D - Q -
- -
Human    D E P D E K T E E S D T D R L L S S D H E
K S
Human
       TGATGAGCCAGATGAAAAGACTGAGGAATCAGACACAGACAGACTTCTAAGCAGTGACCATGAGA
AATCA
Rat    ….…TG.G…..G…....A..G….....C…..........GA…...G….....
Dog    ….…C.T…..........C..A.G…..........................AC….............
dog
tgatgacctagatgaaaagaccgaaggatcagacacagacagacttctaagcaacgaccatgagaaatca
```

GLB1

```
   A.A. 268
        |
Dog                          - - - - - - - - - - V - - - V -
- -
Mouse  - - - - - - - - - - - - - K - - - - V - - K T L
- T
Human  E F Y T G W L D H W G Q P H S T I K T E A V
A S
Human
       GAATTCTATACTGGCTGGCTAGATCACTGGGGCCAACCTCACTCCACAATCAAGACCGAAGCAGT
GGCTTCC
Mouse ..G…..................C…....TA….C..T…..GG.G..A..TA..A..C…..A..
Dog                          ….T…....G..A…..A…G.G….T….TC…....
dogglb1
gatcattggggccagccacactcaacagtgaagactgaagtcgtggcttcc
```

Appendix 2-2 (cont'd).



Lineups of several canine gene sequences with homologous mammalian genes. The nucleotide and amino acid sequences are compared for each of several anchor loci between dog and two other species. The locations of PCR primers are underlined, although not all PCR primer sites are shown. Some of the lineups show intron sequence whereas others simply identify the location of the introns. Genbank accession numbers for canine sequences are as follows: CFTR, L77683 and L77689; COL10A1, L77672; CSF1R, L77670; DCN1, L77648; GHR, L77673; GLB1, L77671; PVALB, L77685 and L77686; and RB1, L77669.

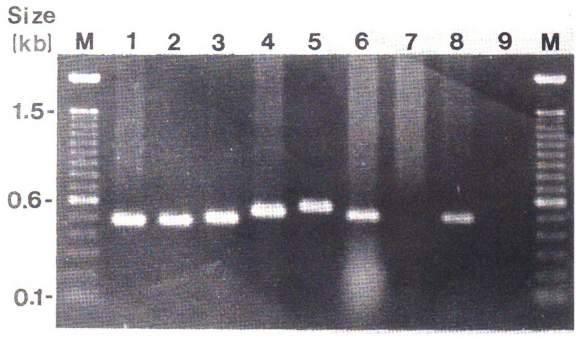
Appendix 2-3

Sequence of a portion of the FES proto-oncogene from several mammalian DNAs. Sequences are from exon 15 and intron 15. Notations for the sequence lineups are as follows: HUM, human; MAC, macaque; CAT, domestic cat; FES, feline sarcoma virus; DOG, dog; COW, ox; GOA, goat; HOR, horse; PIG, pig; RAT, rat; and MOU, mouse. The upper two lines for each block of text represent amino acid sequences and the lower lines represent nucleotide sequences. Dots indicate nucleotides in the various species that are identical to those of the human sequence. The human and cat sequences determined here exactly match the published sequences (Alcalay et al, 1990p; Roebroek et al, 1987). The feline sarcoma virus sequence was not determined in this study but is included for comparative purposes. Only a single amino acid interchange was found among these sequences; isoleucine (I) for macaque, cat and feline sarcoma virus and leucine (L) in all others. Sequence alignments for the intron were done visually and may not be optimal. Genbank accession numbers for these sequences are as follows: AMACFES, L77678; DOGFES, L77674, CATFES, L77675, COWFES, L77677; GOAFES, L77681; PIGFES, L77679; HORFES, L77676; RATFES, L77680, and MOUFES, L77682.

```
MAC,CAT,FES                                                        I
HUM   A  D  N  T  L  V  A  V  K  S  C  R  E  T  L  P  P  D  L  K
HUM  GCCGACAACACCCTGGTGGCGGTGAAGTCTTGTAGAGAGACGCTCCCACCTGACCTCAAG
MAC  ….........T…....A…..............................A…..
CAT  ….........T…....C….A….C.C….A…........A…..
FES  ….........T…....C….A….C.C….A….A…...A…..
DOG  ….....T..T…..........A….CC…....C….............
COW  ..A…..................A…...C.C….A…..............
GOA  ..A…..................A…...C…....A..G..C…........
PIG  ..A..T…................A….CC….A…................
HOR  ..T…..................A….CC…..........C..G…......
RAT  ..A…......C…...T…........C….............N…...
MOU     ….............T…........C….NNN…............


HUM   A  K  F  L  Q  E  A  R
HUM  GCCAAGTTTCTACAGGAAGCGAG GTGGGTGATAAACTAATGATCACCACGGGTCCCGCAT
MAC  ….................... …..................................
DOG  ….................... ….....C….G…C…--..CA.A.CT..A…
CAT  ….........T…....A.A ..A….AC…AG..C…--..CATAA.T…..C
FES  ….........T…....A.A
COW  ….........C…..... ….......G.AC.CCC…A.TGTA..C…CATA
GOA  ….........G…....A.. ….......G.AC.CC…A.TGTA..C…T.C.C
PIG  ….........G…..... ….........AG..CC….TGTGATAAAAGA.CC
HOR  ….............G..A.. ….....C…A…CC….TGGTAT.CTAA.G..
RAT  ….........G..NNNN…. ….…C….A.GGGA.CAGT..A..T…TTGTG
MOU  ….................A.. ….........A….AT
```

Appendix 2-4



Amplification of a portion of the FES protooncogene from several mammalian DNAs using UM-STS primers. Target DNAs for each lane are as follows: 1, human; 2, pigtailed macaque; 3, dog; 4, goat; 5, pig; 6, horse; 7, mouse; and 8, rat. The mouse DNA here was degraded; strong amplification was with another lot (sequence shown in Appendix 2-3). The DNA marker lane, (M) contains a 100-bp ladder.

Appendix 2-5

Amplification Conditions for Canine UM-STSs

Locus	Temperatures (C)	Times (min)	Size of PCR Product (bp) Human	Dog
CFTR	95, 57, 72	0.5, 1.5, 4	700	1000
COL10AI	94, 57, 72 (hs)[a]	1, 2, 3	384	384
CSF1R	94, 59, 72	1, 2, 3	730	730
CYP1A1	95, 57, 72	0.5, 1.5, 4	700	600
DCN1	94, 57, 72	1, 2, 3	1422	2000
FES	94, 57, 72	0.5, 1, 1.5	484	500
GHR	94, 57, 72	1, 2, 3	765	800
GLB1	94, 57, 72	1, 2, 3	238	240
PKLR	94, 59, 72	1, 2, 3	600	630
PVALB	94, 57, 72 (hs)	0.5, 1.5, 4	1400	1300
RB1	94, 59, 72	1, 2, 3	695	1300

a. hs indicates "hot start" used.

Appendix 2-6

Summary of Amplification Results for UM-STSs for Several Mammalian DNAs[a]

Locus	Human	Macaque	Dog	Goat	Pig	Horse	Mouse	Rat
CFTR	**+**[b]	+	+	-	+	+	+	+
COL10A1	+	+	+	-	+	+	+	+
CSF1R	+	+	+	+	-	-	+	+
CYP1A1	+	+	+	+	+	+	+	+
DCN1	+	+	+	-	+	+	+	+
FES	+	+	+	+	+	+	+	+
GHR	+	+	+	+	+	+	+	+
GLB1	+	+	+	+	+	+	+	+
PKLR	+	+	+	+	+	-	+	+
PVALB	+	+	+	+	+	+	-	-
RB1	+	+	+	-	+	+	+	+

a. +, Amplification; -, no amplification.
b. Boldface symbols indicate index species.

Appendix 2-7

Primer Sets for 11 Universal Mammalian Sequence-Tagged Sites.

Locus	Index Species[a]	Primer 1 (P1)[b]	Primer 2 (P2)	P1 Name[c]	P2 Name	P1 AA[d]	P2 AA	Human C'some
CFTR	Human	CTAAGCCATGGCCACAAGCA	CATTGCTTCTATCCTGTGTTC	HCFTR-EX22D	HCFTR-EX23U	1346	1407	7q31-q32
	Mouse	**********T*****	***C******************					
COL10A1	Human	ATTCTCTCCAAAGCTTACCC	GCCACTAGGAATCCTGAGAA	HCOL10A1-EX1D	HCOL10A1-EX2U	505	693	6q21-q22
	Mouse	********T**********	********************					
CSF1R	Human	TTCCAAAACACGGGGACCTA	CATGCCAGGGCGAGAAGGA	HCSF1R-EX3D	HCSF1R-EX4U	76	161	5q33-q35
	FeSV	********************	*G*****************					
CYP1A1	Dog	TTGGACCTCTTTGGAGCTGG	TGGTTGATCTGCCACTGGTT	DCYP1A1-EX3D	DCYP1A1-EX5U	319	417	7q31
	Human	********************	********************					
DCN1	Human	GTTGATGCAGCTAGCCTGAA	AAGTGAAGCTCCCTCAGATG	HDCN1EX6D	HDCN1EX7U	207	247	12q21-q23
	Rat	**********C********	*****G**********					
FES	Human	GGGGAACTTTGGCGAAGTGTT	TCCATGACGATGTAGATGGG	HFES-EX14D	HFES-EX15U	573	641	15q25-qter
	FeSV	**********A********	********************					
GHR	Human	CCAGTTCCAGTTCCAAAGAT	TGATTCTTCTGGTCAAGGCA	HGHR-EX9D	HGHR-EX10U	301	432	5p13.1-p12
	Rat	*****A**********	A******************					
GLB1	Human	GAATTCTATACTGGCTGGCT	CATTCCAATAGGCAAAATTGGT	HGLB-EX8D	HGLB-EX9U	268	319	3pter-p21
	Mouse	**G*****************	**********************					
PKLR	Human	CGCCTCAAGGAGATGATCAA	ATGAGCCCGTCGTCAATGTA	HPKLR-EX4D	HPKLR-EX6U	72	193	1q21
	Rat	********************	********************					

Appendix 2-7 (Cont'd).

PVALB1	Human	ATGTGAAGAAGGTGTTTCACAT	TCTTTGTCTCCAGCAGCCAT	HPVALB-EX3D	HPVALB-EX4U	43	93	22q12-q13.1
	Rat	**********C*****	**C**********					
RB1	Human	GTTCCAGAAAATAAATCAGATGGT	ACTCATTTCTGCCAGTTTCTG	HRB1-EX25D	HRB1-EX26U	844	905	13q14.2
	Mouse	**********C********	**********					

a. Primary index species listed first.
b. Stars (*) indicate identical nucleotides.
c. First letter, primary index species; next letters, locus; EX + number, exon number, D, down; U, up.
d. Amino acid (AA) over which 5' end of nucleotide primer lies.

Appendix 2-8

Eighty-six[a] universal mammalian sequence tagged sites - human chromosomal locations and names

Locus Name	Gene Product	Human Chromosome	Band	Primer 1 Name	Primer 2 Name
PND	Pronatriodilatin	1	p36	DPNDEX1D	DPNDEX2U
PKLR	Pyruvate Kinase – RBC	1	q21	HPKLREX4D	HPKLREX6U
AT3	Antithrombin III	1	q23-q25	HAT3EX3D	HAT3EX4U
REN	Renin	1	q32	HRENEX8D	HRENEX9U
SFTP3	Pulmonary Surfactant Protein 3	2	p11.2	DSFPT3EX4D	DSFTP3EX5U
SPTBN1	Beta Spectrin (Non-RBC)	2	p21	HSPTBN1EX13D	HSPTBN1EX14U
APOB	Apolipoprotein B	2	p24-p23	HAPOBEX26D	HAPOBEX26U
IL1A	Interleukin 1 Alpha	2	q13	HIL1AEX2D	SIL1AEX3U
COL3A1	Collagen III Alpha 1	2	q31-q32.3	HCOL3A1EX24D	HCOL3A1EX25U
ELN	Elastin	2	q31-qter	HELNEX32D	HELNEX33U
PAX3	Human Paired Domain 2	2	q34-q36	HHUP2EX2D	HHUP2EX3U
GCG	Glucagon	2	q36-q37	HGCGEX4D	HGCGEX5U
PIT1	Pituitary-Specific Transcription Factor 1	3	p11	HPIT1EX4D	HPIT1EX5U
GLB1	B-Galactosidase	3	pter-p21	HGLB1EX8D	HGLB1EX9U
GPX1	Glutathione Peroxidase 1	3	q11-q12	HGPX1EX1D	HGPX1EX2U
TF	Transferrin	3	q21	HTFEX7D	HTFEX8U
RHO1	Rhodopsin	3	q21-qter	HRHOEX3D	HRHOEX4U
GLUT2	Glucose Transport-like 2	3	q26.1-q26.3	HGLUT2EX9D	HGLUT2EX10U
SST	Somatostatin	3	q28	HSSTEX1D	HSSTEX2U
HOX7	Homeobox 7	4	p16.1	HHOX7EX2D	HHOX7EX2U
PDEB	cGMP Phosphodiesterase Beta	4	pter	HPDEBEX14D	HPDEBEX15U
ALB	Albumin	4	q11-q13	HALBEX4D	HALBEX5U
KIT	c-KIT Protooncogene	4	q12-q13	HKITEX18D	HKITEX20U
FGG	Fibrinogen Gamma	4	q28	HFGGEX8D	HFGGEX9U
GHR	Growth Hormone Receptor	5	p13.1-p12	HGHREX9D	HGHREX10U
HEXB	Beta Hexosaminidase	5	q13	HHEXBEX12D	HHEXBEX13U
IL4	Interleukin 4	5	q23-q31	HIL4EX1D	HIL4EX2U
ADRB2	Adrenergic Receptor Beta 2	5	q31-q32	HADRB2EX1D	HADRB2EX1U
CSF1R	CSF-1 Receptor	5	q33-q35	HCSF1REX3D	HCSF1EX4U
TNFA	Tumor Necrosis Factor Alpha	6	p21.3	HTNFAEX1D	HTNFAEX4U
EDN1	Endothelin 1	6	p24-p23	HEDN1EX3D	HEDN1EX4U
COL9A1	Collagen IX Alpha I	6	q12-q14	HCOL9A1EX3D	HCOL9A1EX4U

Appendix 2-8 (cont'd).

COL10A1	Collagen Type X Alpha I	6 q21-q22	HCOL10A1EX2D	HCOL10A1EX2U
PLG	Plasminogen	6 q25-q27	HPLGEX18D	HPLGEX19U
EPO	Erythropoeitin	7 q21	HEPOEX2D	HEPOEX3U
CFTR	Cystic Fibrosis Trans. Regulator	7 q31-q32	HCFEX22D	HCFEX23U
TCRB	T-Cell Receptor Beta	7 q35	DTCRBEX2D	DTCRBEX3U
SFTP2	Pulmonary Suractant Protein 2	8 p21	HSFTP2EX2D	HSFTP2EX4U
CA2	Carbonic Anhydrase II	8 q22	CAUNIVEX3D	HCAIIEX4U
TG	Thyroglobulin	8 q24	HTGEX9D	HTGEX10U
ALDOB	Aldolase B	9 q21.3-q22.2	HALDOBEX7D	HALDOBEX8U
C5	Complement Factor 5	9 q22-q34	HC5EX36D	HC5EX37U
ABL	ABL Proto-oncogene	9 q34	HABLEX10D	HABLEX11U
RET	RET Proto-Oncogene	10 q11.2	HRETEX19D	HRETEX20U
TDT	Terminal Transferase	10 q23-q24	HTDTEX9D	HTDTEX10U
OAT	Ornithine Aminotranferase	10 q26	HOATEX7D	HOATEX8U
WT1	Wilms Tumor 1	11 p13	HWT1EX8D	HWT1EX9U
LDHA	Lactate Dehydrogenase A	11 p14-15.5	HLDHAEX3D	HLDHAEX4U
INS	Insulin	11 p15.5	DINSEX2D	DINSEX3U
CD20	CD20	11 q12-q13.1	HCD20EX6D	HCD20EX7U
ROM1	Rod Outer Segment Protein-1	11 q13	HROM1EX1D	HROM1EX1U
APOC3	Apolipoprotein C3	11 q23-qter	DAPOC3EX2D	DAPOC3EX3U
VWF	von Willebrand's Factor	12 p	HVWFEX46D	HVWFEX47U
LDHB	Lactate Dehydrogenase B	12 p12.1-12.2	HLDHBEX3D	HLDHBEX4U
IL6	Interleukin 6	12 p12.2-p12	HIL6EX3D	DIL6EX4U
TPI	Triosphosphate Isomerase	12 p13	HTPIEX2D	HTPIEX5U
COL2A1	Collagen II Alpha 1	12 q14.3	HCOL2A1EX2D	HCOL2A1EX3U
DCN1	Decorin	12 q21-q23	HDCNEX6D	HDCNEX7U
IGF1	Insulin-Like Growth Factor 1	12 q22	HIGF1EX3D	HIGF1EX4U
PLA2	Phospholipase A2	12 q23-qter	DPLA2EX2D	DPLA2EX3U
RB1	Retinoblastoma 1	13 q14.2	HRB1EX25D	HRB1EX26U
F7	Clotting Factor VII	13 q34	HF7EX7D	HF7EX8U
CHY	Chymase (Mast Cell)	14 q11.2	DCHYEX4D	DCHYEX5U
CKBB	Creatine Kinase Brain	14 q32.3	DCKBEX6D	DCKBEX8U
TCRA	T-Cell Receptor Alpha	14 q34	DTCRAEX3D	DTCRAEX4U
B2M	Beta-2 Microglobulin	15 q21-q22.2	HB2MEX2D	HB2MEX3U
CYP1A1	Cytochrome P-450 (AHH)	15 q22-q24	DCYP1A1EX3D	DCYP1A1EX5U
PKM	Pyruvate Kinase – Muscle	15 q22-qter	HPKMEX2D	HPKMEX3U
FES	FES Proto-Oncogene	15 q25-qter	HFESEX14D	HFESEX15U
HGBA	Alpha Hemoglobin	16 p13.3	HHGBAEX2D	HHGBAEX3U

Appendix 2-8 (cont'd).

GOT2	Glutamate Oxaloacetate Transaminase 2	16 q21-q22	HGOT2EX5D	HGOT2EX7U
CTRB	Chymotrypsinogen	16 q22.3-q23.2	DCTRBEX5D	DCTRBEX6U
APRT	Adenosine PR Transferase	16 q24	HAPRTEX3D	HAPRTEX5U
TP53	Tumor Protein 53	17 p13.1	HTP53EX5D	HTP53EX7U
NF1	Neurofibromatosis 1	17 q11.2	HNF1EX6D	HNF1EX7U
SCN4A	Skeletal Muscle Sodium Channel	17 q23.1-q25.3	HSCN4AEX23D	HSCN4AEX24U
TS	Thymidylate Synthetase	18 pter-q12	HTSEX5D	HTSEX6U
APOC2	Apolipoprotein C2	19 q13.2	DAPOC2EX3D	DAPOC2EX4U
CKMM	Creatine Kinase Muscle	19 q13.2-q13.3	DCKMEX2D	DCKMEX3U
PVALB	Parvalbumin	22 q12-q13.1	HPVALBEX3D	HPVALBEX4U
DYS	Dystrophin	X p21	DDYSEX7D	DDYSEX7U
MNK	Menkes Protein	X q12-q13.3	HMNKEX	HMNKEX
HPRT	Hypoxanthine PR Transferase	X q26	HHPRTEX7D	HHPRTEX8U
F9	Clotting Factor IX	X q26.3-q27.1	DF9EX7D	DF9EX8U
F8	Clotting Factor VIII	X q28	HF8EX24D	HF8EX25U
SRY	Sex Determining Region – Y	Y p11.3	HSRYEX1D	HSRYEX1U

a- For convenience, the eleven loci described in detail are included in the Appendices

Appendix 2-9

Eighty-six universal mammalian sequence tagged sites - sequences and sizes

Locus Name	Primer 1 Sequence	Primer 2 Sequence	PCR Product Size Human Genomic	Dog Genomic[a]
PND	GCAGACCTGCTGGATTTCAAG	CAGTCCGCTCTGGGCTCCAAT	360	360
PKLR	CGCCTCAAGGAGATGATCAA	ATGAGCCCGTCGTCAATGTA	660	500
AT3	CTTCTTTGCCAAACTGAACTG	GGGCTGAACTTTGACTTCCA	658	660
REN	ACACTCCCCGACATCTCTTT	CGCCGATCAAACTCTGTGTA	137	137
SFTP3	GGAAGTTCCTGGAGCATGAG	CACAGGCCCAGGTGCTTACA	308	310
SPTBN1	TCTCAAGACTATGGCAAACA	CTGCCATCTCCCAGAAGAA	640	800
APOB	GTAAAAGCTCAGTATAAGAAAAAC	GTGCCCTCTAATTTGTACTG	460	460
IL1A	AGAAGTCAAGATGGCCAAAGT	TGATTCAGAGACAGATGGTC	1900	1900
COL3A1	GGACCAGGAAGTGATGGGAA	ACTTTCTCCTTGACTTCCCT	752	1400
ELN	GCTGCAGCCGCTAAAGCAG	AGGACACCTCCAAGGCCAG	600	1300
PAX3	GCCACAAGATCGTGGAGATG	GGTTCTCTCTTTTGTATTCCTC	1020	1170
GCG	TTCATTGCTTGGCTGGTGAA	GTGTTCATCTCATCAGAGAA	700	600
PIT1	TTCAGTCAAACAACAATCTG	GCTCCCACTTTTTCATTGTA	700	1000
GLB1	GAATTCTATACTGGCTGGCT	CATTCCAATAGGCAAAATTGGT	700	1000
GPX1	GACTACACCCAGATGAACGA	CAGGAACTTCTCAAAGTTCC	633	633
TF	GCTGACAGGGACCAGTATGA	AACAGCAGGTCCTTCCCATG	1700	585
RHO1	TACATGTTCGTGGTCCACTT	TGGTGGGTGAAGATGTAGAA	1479	553
GLUT2	TGGATGAGTTATGTGAGCAT	GACTTTCCTTTGGTTTCTGG	364	364
SST	GACTCCCGAGGCTTCCTCTTTGA	ATACTGCAGGAGAGAGAAGAA	1200	1200
HOX7	AAGTTCCGCCAGAAGCAGTA	ATCTTCAGCTTCTCCAGCTC	400	400
PDEB	CTGAAGAGCTACTACACGGA	TGACACTTGTTCATCCACCA	300	300
ALB	GGCTGACTGCTGTGCAAAACA	AAGTAAGGATGTCTTCTGGC	730	730
KIT	CCTGTGAAGTGGATGGCACC	GCATCCCAGCAAGTCTTCAT	1000	1000
FGG	CAATATAAAGAAGGATTTGGACA	TGACACTTGTTCATCCACCA	1422	3000
GHR	CCAGTTCCAGTTCCAAAGAT	TGATTCTTCTGGTCAAGGCA	238	200
HEXB	TTCATTGGTGGAGAAGCTTG	ATCTTTGGAACTCCAGAGTC	1400	1000
IL4	CTATTAATGGGTCTCACCTCCCAACT	TCAACTCGGTGCACAGAGTCTTGG	469	450
ADRB2	CCCATTCAGATGCACTGGTA	GCAGCCAGCAGAGGGTGAA	381	281
CSF1R	TTCCAAAACACGGGGACCTA	CATGCCAGGGCGAGAAGGA	1200	800
TNFA	CTCAGCCTCTTCTCCTTCCT	ATGGGCTCATACCAGGGCTT	1198	1200
EDN1	CCAAAAAGACAAGAAGTGCTG	TGGAACAGTCTTTTCCTTTCTT	1400	800
COL9A1	ATCAGGATTGGCCAAGATGA	GGAATCCTGAAGTCTACATT	484	500
COL10A1	ATTCTCTCCAAAGCTTACCC	GCCACTAGGAATCCTGAGAA	340	340
PLG	CAGCTCCCTGTGATTGAGAA	TAGACACCAGGCTTATTGGG	1100	1100

Appendix 2-9 (cont'd).

EPO	CTCCCTCTGGGCCTCCCAGT	CCATCCTCTTCCAGGCATAGAA	478	600
CFTR	CTAAGCCATGGCCACAAGCA	CATTGCTTCTATCCTGTGTT	765	800
TCRB	GACTGTGGCTTCACCTCGG	GATCTCATAGAGGATGGTGG	238	238
SFTP2	CAGAAACACACGGAGATGGT	GCCATCTTCATGATGTAGCA	500	600
CA2	CAGTTCCATTTTCAGTGGGG	GGCCAGTCCATCAGGTTGCT	350	1500
TG	TTCACCTCAGAGTGCTACTG	GCTTCTCTGTAGCTCATGATCTT	650	650
ALDOB	GTGACTGCTGGACATGCCTG	TTTGCAGCCTTGCCACCCC	463	450
C5	TGTGTACGATTCCGGATATTTGA	GCTCCTTCACAGACTTTCTG	300	300
ABL	TCAGACGAAGTGGAAAAGGA	AGAAGGCGCTCATCTTCATT	151	151
RET	CCCTTCCACATGGATTGAAA	CATCCAGTTAGCATATACAC	1800	1800
TDT	ACCTGGAAGGCCATCCGTGT	CGCCGGAGGTCTCTCTCAAA	1200	1250
OAT	CGTGCTCTTCAGGATCCAAA	GCCAGCCATCTACCAGTTCT	153	153
WT1	GAGAAACCATACCAGTGTGA	GTTTTACCTGTATGAGTCCT	820	850
LDHA	AACTCCAAGCTGGTCATTAT	GAATCCAGATTGCAACCACT	786	264
INS	GAGCGCGGCTTCTTCTACAC	GGTAGAGGGAGCAGATGCTGG	650	650
CD20	CTCTTTGCTGCCATTTCTGGAAT	TGGAAGAAGGCAAAGATCAGCAT	793	800
ROM1	CAGAGGACGGGCCACAGAA	GTTAAACACCACAGAGGCCTT	900	900
APOC3	CAGGAACAGAGGTGCCATGC	TGCGCCACCTGGGACTCCTG	1900	1800
VWF	CCAGAGCGCATGGAGGCCTG	CTGCACTCCAGCTTGAATCC	700	850
LDHB	TTCCTCAGATCGTCAAGTACA	CTGCTGGGATGAATGCCAAG	650	650
IL6	GCACTGGCAGAAAACAACCT	ATCTGAAACTCCACAAGACC	172	172
TPI	TATATCGACTTCGCCCGGCA	ATGGCCCACACAGGCTCATA	300	500
COL2A1	CTCTGCGACGACATAATCTG	TCTCCAGGTTCTCCTTTCTG	600	1100
DCN1	GTTGATGCAGCTAGCCTGAA	AAGTGAAGCTCCCTCAGATG	1200	1300
IGF1	GGCATCGTGGATGAGTGCTG	CTCCTTCTGTTCCCCTCCTG	950	900
PLA2	GACTACGGCTGCTACTGTG	TTACAGCTGGCCAGTTTCTT	420	300
RB1	GTTCCAGAAAATAAATCAGATGGT	ACTCATTTCTGCCAGTTTCTG	1600	1600
F7	AATGGAGCTCAGTTGTGTGG	CGATGTCGTGGTTGGTGGT	1600	700
CHY	GTCCCACCTGGGAGAATGTG	TGGGAGATTCGGGTGAAGAC	900	900
CKBB	TGGATCAACGAGGAAGACCA	TTCACACCATCCACCACCAT	200	950
TCRA	ACTGTCTGCCTGTTCACCGATTT	GTAACAACTTGGCATCACAGGAAT	900	900
B2M	TTCAGCAAGGACTGGTCTTT	CTGCTTACATGTCTCGATCT	1244	1100
CYP1A1	TTGGACCTCTTTGGAGCTGG	TGGTTGATCTGCCACTGGTT	800	860
PKM	GCCTTCATTCAGACCCAGCA	ATTCCAGACTTAATCATCTCCTT	1200	1200
FES	GGGGAACTTTGGCGAAGTGTT	TCCATGACGATGTAGATGGG	467	436
HGBA	CCCACCACCAAGACCTACTT	CGGTATTTGGAGGTCAGCAC	548	480
GOT2	TTTAAGTTCAGCCGAGATGT	CTTGGTAGGCCATGTCAAA	1400	1300
CTRB	AACGACATCACCCTGCTGTT	TGCAGGAGGAGACGCCACT	592	600
APRT	GACTCCCGAGGCTTCCTCTTTG	ATACTGCAGGAGAGAGAAGAA	695	1300
TP53	TACAAGCAGTCACAGCACAT	TCTTCCAGTGTGATGATGGT	1600	1600

Appendix 2-9 (cont'd).

NF1	ATTCACTCTCTGTGTACTTG	CAAAGCTTCTGTGACTGTTT	440	350
SCN4A	CTCAAGGTGGACATCCTGTACA	AGCAGCGTCCGGATGCCCTT A	1177	1100
TS	TGCCAGTTCTATGTGGTGAA	AGGTAAATATGTGCATCTCC	915	750
APOC2	GAATCACTCTACAGTTACTGG	AGCTGCTGTGCTTTTGCTGTA	201	201
CKMM	AAGAAGCTGCGGGACAAGGA	CAGCCCACGGTCATGATGAAA	1119	1100
PVALB	ATGTGAAGAAGGTGTTTCACAT	TCTTTGTCTCCAGCAGCCAT	420	830
DYS	GTTTCAGGCCAGACCTCTTT	TACCGACCTTCAGGATCAAG	500	500
MNK	GGCATGACTTGTAATTCCTG	CATCAAATCCCATGTCTTCTA T	669	750
HPRT	AGCTTGCTGGTGAAAAGGAC	TTATAGTCAAGGGCATATCC	766	650
F9	TGGGTGGTAACTGCAGCCCACT	CTACGCACACTCTTCACCCCA	650	650
F8	GATGCACAGATTACTGCTTC	GTAAGCAGAGATTTTACTCCC TG	779	800
SRY	AAGCGACCCATGAACGCATT	TTCGGGTATTTCTCTCTGTG	2500	2500

a - All PCR products have been sequenced or, in a few cases, are derived from primers made to the published canine sequence.

Chapter 3

Estimate of Nucleotide Diversity in Dogs Using a Pool-and-Sequence Method

Estimate of Nucleotide Diversity in Dogs Using a Pool-and-Sequence Method

Running Head: Dog Nucleotide Diversity

James A. Brouillette,[1,3,4] Jennifer R. Andrew,[1] Patrick J. Venta[1,2,3]

[1]Department of Microbiology, College of Veterinary Medicine, Michigan State University, East Lansing, MI 48824, USA

[2]Department of Small Animal Clinical Sciences, College of Veterinary Medicine, Michigan State University, East Lansing, MI 48824-1314, USA

[3]Genetics Program, Michigan State University, East Lansing, MI 48824, USA

[4]College of Human Medicine, Michigan State University, East Lansing, MI 48824, USA

Corresponding Author:

Patrick J. Venta

Telephone: 517-432-2515

FAX: 517-432-2524

E-mail: venta@cvm.msu.edu

Abstract

Nucleotide diversity (π), the average number of base differences per site for two homologous sequences randomly selected from a population, is an important parameter used to understand the structure and history of populations. It is also important for determining the feasibility of developing a genetic map for a species based upon single nucleotide polymorphisms (SNPs). Nucleotide diversity has never been estimated for dogs. Segments of twelve canine genes from ten diverse dog breeds were examined for nucleotide variation by using a pool-and-sequence method. We identified three SNPs in the coding regions (2,501 bp) and eleven SNPs in the introns (2,953 bp). Each of these putative SNPs was tested by restriction enzyme analysis and all were verified. Six additional SNPs were identified in a single SINE contained in one gene. Using these data, canine sequence diversity across breeds was estimated to be 0.001 and 0.0004 in intronic and coding regions, respectively, with SNPs spaced every 400 bp on average. Discovery of useful SNPs in seven of the twelve genes suggests that construction of canine SNP-based map can be accomplished using current technology. Thirteen polymorphic SNPs were also found in 5,847 bp in the cat, horse, ox, and pig, using four of the same genes from which canine nucleotide diversity was estimated. These results suggest that these species may have similar amounts of nucleotide diversity.

Introduction

Although canine isozyme studies have been used to estimate genetic heterozygosity in dogs, nucleotide diversity has never been examined (Fisher et al., 1976; Simonsen, 1976; Weiden et al., 1974). Estimating the nucleotide diversity, π, of a species is important for answering questions about population structures and for understanding how genetic variation is maintained in populations (Nei and Li, 1979). Although most nucleotide diversity studies have been conducted using mitochondrial DNAs, several studies have been done in the past few years on human nuclear genes (e.g., Harding et al., 1997; Clark et al., 1998; Nickerson et al., 1998, Cargill et al., 1999, Halushka et al., 1999). These studies are already yielding information on the histories and relationships of populations around the world.

From the standpoint of mapping medically important genes, the studies of human nucleotide diversity have shown that it will be possible to build genetic maps with a very high density of single nucleotide polymorphisms (SNPs). It is anticipated that tens or hundreds of thousands of SNPs will be used in linkage- and disequilibrium-based mapping because of the potential to automate the typing of SNPs (Brookes, 1999; Collins et al., 1999; Kruglyak, 1999). In domestic animals, similar mapping strategies could be used to locate genes of interest. Linkage disequilibrium methods may be even easier to use, given the possible greater amount of linkage disequilibrium in domestic animal species (cf. Farnir et al., 2000). However, estimates of nucleotide diversity for most domesticated animals do not exist and these would be useful for determining the amount of effort needed to develop SNP-based maps.

Dogs are important work and companion animals in many human societies. We are interested in identifying the mutations that cause inherited diseases in dogs so that breeders can avoid producing affected animals (Padgett, 1998; Willis, 1989; Clark and Stainer, 1994). We would also like to gain insights on the genetic history and relatedness of the various breeds. Although many SNPs have been found in various genes in the dog, the amount of nucleotide diversity cannot be calculated from these studies because negative results for polymorphism searches are rarely reported.

This study provides an estimate of canine nucleotide diversity by ascertaining all variant and non-variant nucleotides in portions of twelve canine genes using a pool-and-sequence (PAS) method. Four of the genes studied in the dog were also examined for SNPs in four additional species (pigs, horses, cattle, and cats) and the results are reported here.

Materials and methods

Target DNA

Equimolar DNA pools were made from the following breeds for each species: dogs; Cocker spaniel, Greyhound, Doberman pinscher, Siberian husky, Labrador retriever, Collie, Scottish terrier, German shepherd dog, Beagle, and Pointer; cattle; Red-and-White Holstein, Holstein, Aryshire, Brown Swiss, Hereford, and Angus; horses; Appaloosa, Arabian, Belgian, Percheron, Miniature, Thoroughbred, Quarterhorse, and Rocky Mountain; and pigs; Chesterwhite, Duroc, Hampshire, Landrace, and Yorkshire. A DNA pool was also made from the DNA of 10 unrelated cats of unspecified breed. The unmixed purified DNA samples were kindly provided by each of the following individuals; dog, Dr. Vilma Yuzbasiyan-Gurkan; horse, Dr. Susan Ewart; pig, Dr. Cathy Ernst; and cat, Dr. John Kruger. Dr. Robert Holland kindly provided peripheral blood samples from cattle for DNA isolation. DNA was isolated from white blood cells using a previously published phenol-chloroform extraction method, except that the proeinase K digestion step was omitted (Sambrook et al., 1989). DNA was precipitated using ammonium chloride / ethanol precipitation. A DNA sample was used from an individual animal to serve as a sequencing control. Cats, cattle, horses and pigs were chosen as variation controls (see Results) because they are domesticated animal species representing three different mammalian orders. In addition, mapping projects are underway in each of these species as well as in dogs.

Target genes

Six genes were chosen for the study because they are candidate genes for other studies by our lab (*RHO1, GNAT1, NCAD, RDS, ACTC,* and *RYR2*). Six additional genes (*C5, CFTR, CYP1A1, FES, TS,* and *WT1*) were chosen because the primer pairs had already shown to amplify reliably, the product size was between 500 bp and 1000 bp, the primer pairs spanned an intron, and some of them had been shown to amplify target regions in other mammalian species of interest.

Primer Design

Primer pairs have been previously described for seven of the genes examined, *C5, CFTR, CYP1A1, FES, RHO1, TS,* and *WT1* (Venta et al., 1996). For the other five genes (*ACTC, GNAT1, NCAD, RDS,* and *RYR2*), primers were designed by comparing nucleotide sequences of two known mammalian species (usually human with mouse or rat), for which at least one exon-intron gene structure was available (Appendix 3-1).

New primers were designed as before, except that the 3' ends of the primers always overlie the most conserved codon position for a given amino acid (Venta et al., 1996; Fitch, 1976). Pairs of primers were designed to have similar annealing temperatures, and

the potential for primer-dimer formation was avoided. Primer sequences are given in Appendix 3-9.

PCR Amplification of Gene Segments

PCR was performed using 50 ng of genomic DNA from either the pooled sample or from the individual control sample, in 25 μl vols containing the following: 20 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 100 μM dNTPs, 10 pmol each primer, and 1.25 U of Taq polymerase. Cycling conditions for each primer pair are given in Appendix 3-9. PCR products were run on 1% agarose, 1 X TBE gels at 125 volts, stained in ethidium bromide, and photographed on a 360 nm ultraviolet transilluminator.

Purification of PCR Products

PCR products were run on agarose gels, gel slices containing PCR products were taken, and the DNA was extracted from the agarose using the QIAEX II Gel Extraction Kit (QIAGEN Corp., Valencia, CA) as recommended by the manufacturer. Aliquots of the extraction products were visualized after electrophoresis on a 1% agarose, 1 X TBE gel by staining with ethidium bromide.

Pool-and-Sequence Method

Approximately 100 ng of each purified PCR product was used as template for cycle sequencing. Cycle sequencing was performed using the Thermo Sequenase ^{33}P-radiolabeled terminator cycle sequencing kit (USB Corp., Cleveland, OH), using the dGTP termination mix, as directed by the manufacturer. Products of the sequencing reaction were run at 50 W for 1.5 to 4.5 hr on 6% acrylamide gels (19:1 acrylamide:bisacrylamide) containing 7M urea. Gels were dried and exposed to x-ray film (Kodak X-omat) at room temperature for 24-48 hr. The identity of the sequence was examined by performing a BLAST search (http://www.ncbi.nlm.nih.gov/BLAST) (Altschul et al., 1997). Possible polymorphic sites were identified by visual inspection of the sequencing ladders for bands in two lanes at identical base positions or by the band-doubling effect seen for insertion/deletion polymorphisms. Presumptive false positive bands were identified by comparison of the single animal and pooled animal sequencing ladders.

Diagnostic Tests for Identified SNPs

Once a putative single nucleotide polymorphism was located on the sequencing gel, a restriction site was created by PCR primer mutagenesis adjacent to the polymorphic site, unless a naturally occurring site was present (Venta et al., 1991). New primers were designed to complement the species-specific sequence as determined by the sequencing

reactions, except that one or two nucleotides were substituted to create a restriction site in one of the two alleles at the polymorphic site (Appendix 3-10). Nested amplifications were required to produce consistent results for cat WT01, cat WT02, dog CFTR01, horse WT01, and ox TS01 using the sequencing primers first, followed by the diagnostic primers.

Restriction Digestion and Estimation of Nucleotide Diversity

All restriction enzymes were purchased from one supplier (New England Biolabs, Berverly, MA) except for *Tai*I (Fermentas, Amherst, NJ) and *Mae*III (Boehringer Mannheim, Indianapolis, IN). Restriction digestion was performed directly in the PCR buffer with the addition of 50 mM $MgCl_2$ to reach a final concentration of approximately 10 mM. Restriction digestion reactions were carried out at the temperature recommended by the manufacturer. All of the enzymes used in this study worked well in PCR buffer supplemented with extra $MgCl_2$ except *Tth111*I and *Nla*III. For these enzymes, PCR products were precipitated with ethanol and ammonium acetate, resuspended in a suitable volume of TE (10 mM Tris, 1 mM EDTA, pH 7.4-8.0) and digested following the manufacturer's instructions.

Nucleotide diversity was calculated according to a previously published method (Nei and Li, 1979). Briefly, the average number of nucleotide differences across the total length of DNA sequenced in the study was calculated taking the allele frequency for each variant site into account.

Results

Identity of the Genes

In order to confirm that we were amplifying the homologue of the gene of interest, a BLAST search of the Genbank database (release 113) was performed using the nucleotide sequence data we obtained (Genbank accession numbers are listed in Appendix 3-1). The highest BLAST score always matched one or more mammalian homologs of the gene for which the primers had been designed (data not shown). Most of the nucleotide sequence matches were contained within the coding region of each gene. The intronic sequence generally had a relatively minor impact on the score.

Use of the Pool-and-Sequence Method for Identification of SNPs

Several SNPs can be seen as bands that occur at identical positions in the sequencing ladders of Appendix 3-6 and Appendix 3-7. The use of a sequencing ladder from a single animal served as an important control to recognize false positive SNPs for several of the genes examined (data not shown). The allele frequencies for any given SNP could be estimated from the relative intensities of the bands. For example, in Appendix 3-6A, SNP 1 (TS01) shows a very weak band in the C lane (Pool) and a much stronger one at the same position in the A lane. The alleles of this SNP can be distinguished by digestion of the PCR product with *Mwo*I (Appendix 3-2). The allele counts for the Pool are 2 (the C allele) and 18 (the A allele) (Appendix 3-3). The allele counts for SNP 4 (TS04) are 10

and 10 which are reflected in the banding intensities of the sequencing ladder (Appendix 3-6, Pool).

Indels (insertion/deletions) were recognized as the occurrence of doublet bands above the site of variation (Appendix 3-6A, SNP 2 [TS02] and Appendix 3-7B, SNP 3 [CFTR03]). Banding intensities for indels also give an indication of the relative allele frequencies as shown by CFTR03 (Appendix 3-7B, Pool and Appendix 3-3). A SNP caused by a transition was recognized as a single band among the doublets in the sequencing ladder above the indel site (Appendix 3-6A, SNP 3 [TS03], Pool).

A total of twenty SNPs were identified in 2,501 bp of coding and 2,953 bp of intronic sequence. Six of the SNPs were contained within one SINE element in the TS gene. This SINE is 88% identical to the previously reported consensus sequence (Minnick et al., 1992, Coltman and Wright, 1994, Bentolila et al., 1999). The fourteen non-SINE SNPs were confirmed by restriction enzyme digestion of the DNA from the individual pool members, either using naturally occurring restriction enzyme sites or ones that were created by PCR mutagenesis (Appendix 3-2). The minor allele frequencies for each SNP for the animals in the pool varied from 0.05 to 0.5 (Appendix 3-3). The rarest allele discovered (0.05) was found in the heterozygous state in the single mixed breed animal, and so it is not certain that it would have been detected by itself in the pool. We were able to easily detect a minor allele which occurred at a frequency of 0.15 in the pool that was not present in the single animal (*RYR2*, Appendix 3-3). We have also been able to detect a rare allele at a frequency of 0.062 in horses (*TS*, see below).

Four SNPs were found in the *TS* gene in addition to the six SNPs seen in the SINE (Appendix 3-2). Four SNPs were also found in the *CFTR* gene, three in the *ACTC* gene, and one each in the *GNAT1*, *NCAD*, and *RYR2* genes. Three of the fourteen SNPs were found in the coding regions of the *CFTR* and *GNAT1* genes. These three changes are silent. Three of the SNPs (ACTC01, CFTR02, and GNAT01) had a CpG dinucleotide overlying one of the alleles. CpGs are thought to be over-represented in SNPs due to their higher mutation rates compared to other dinucleotides (Krawczak et al., 1998). Three of the SNPs overlie indels. Of the remaining eleven non-indel SNPs, ten are transitions and one is a transversion (Appendix 3-2). One simple tandem repeat (STR) was also contained in one gene (*WT1*), although we knew of its existence before the SNP search was started (Shibuya et al., 1996; Venta et al., 1996).

Estimate of Canine Sequence Diversity

Canine sequence diversity was calculated with and without the six SNPs found in the TS SINE. The allele frequency of the TS SINE SNPs were estimated by the relative banding intensity from the sequencing film. The estimate for the rarer allele in each SNP in the SINE is 0.4. Sequence diversity was calculated from the number of bases sequenced (Appendix 3-1) and the amount of variability observed (Appendix 3-3). The estimate of sequence diversity for intronic sequences with the SINE SNPs included was 0.0013 and without the SINE SNPs was 0.0010. Coding region sequence diversity was 0.0004. If we had missed all of the alleles whose frequency is less that 0.15, the sequence diversity estimates would be biased downward about 16% from their true value, assuming that the

infinite site model would produce a reasonable approximation for the distribution of nucleotide variation in dog populations (cf. Clark et al., 1998).

Another useful parameter to estimate sequence diversity is the average frequency of segregating polymorphic sites. One SNP was found every 834 bp of coding region sequenced and every 268 bp of intronic region sequenced for an average of one SNP every 390 bp of the total sequence.

We have not determined the haplotypes for the genes in which more than one SNP was found. In many cases, however, it is possible to deduce that three or more haplotypes exist by inspection of the SNP pattern across breeds. For example, at least four haplotypes can be inferred for the SNPs found in the *ACTC* gene (Appendix 3-3). For SNPs ACTC01, ACTC02, and ACTC03, the Cocker spaniel must have a 2,1,2 haplotype, the Doberman pinscher must have a 2,1,1 haplotype, the Siberian husky must have a 2,2,2 haplotype, and the Scottish terrier must have a 1,1,2 haplotype. For those genes containing SNPs in the region sequenced (about 500 bp), the average heterozygosity is 0.33 (that is, one third of dogs examined are heterozygous for at least one SNP in a given polymorphic gene). This value is about half of that seen for the average STR (Weber, 1990).

Variability of SNPs Within Dog Breeds

The variability of the SNPs within breeds is of interest because most linkage studies in dogs will be conducted in single breeds. In order to examine the within-breed variability of SNPs discovered by pooling across breeds, four SNPs were examined in which the minor allele frequency in the initial pool were typed. The dogs were unrelated for at least three generations. The number of breeds in which heterozygotes are seen for each SNP is correlated with the allele frequencies seen in the original pool (Appendices 3-3 and 3-4). Only four gene/breed combinations (ACTC02, Scottish terriers; RYR201, Pointers; and TS04, Greyhounds and Collies) appear to show deviations from Hardy-Weinberg (HW) expectations (as might be expected from inbreeding), although the sample sizes are relatively small. For the data summed across breeds, there is a significant decrease in the observed number of heterozygotes relative to the number expected under HW equilibrium for ACTC02 (Chi-square = 24.49, $P < 0.001$), as would be predicted for subdivisions of a base population (Wahlund, 1928; Wright, 1951). However, no significant deviation was seen for TS04 (Chi-square = 1.70, $P = 0.43$). It is uncertain if the other two SNPs tested (RYR201 and TS02) show a significant deviation because of the low number of animals carrying the rarer allele. The important practical observation is that a reasonable number of heterozygotes are seen within breeds for the SNPs that were discovered by pooling breeds (Appendix 3-4).

Identification of SNPs in Other Mammalian Species

Examination of the same gene regions in other mammalian species would be expected to show a correlation in sequence diversity if certain genes had a greater-than-average tendency to maintain genetic variation. No correlation would suggest that our estimate of nucleotide diversity was not biased by genes that were more, or less, prone to mutation. Four genes (*CFTR*, *FES*, *TS*, and *WT1*) used in the canine SNP search were also used to search for SNPs in cats, horses, pigs, and cattle. Of the 16 gene-species combinations, all targets amplified except for two (cat FES and pig TS). For the 14 gene-species combinations that amplified, putative SNPs were found in seven of them, including four with two or more SNPs (Appendix 3-1).

Ten of the non-canine SNPs were tested by restriction enzyme analysis. All of the ten tested SNPs were confirmed. Allele frequencies for the ten SNPs examined were (species, SNP, allele 1, 2): horse; CFTR01, 13, 3; TS01, 1, 15; WT101, 11, 5: cattle, CFTR01, 6, 8; TS01, 1, 13: cat; WT01, 6, 14; WT02, 14, 6; WT03, 12, 6 (one sample failed to amplify): and pig, FES01, 4, 10. A simple tandem repeat was also discovered in the ox TS gene, although it was not examined for genetic variability (GenBank Accession No. AF203030). Each of these SNPs is unique to the species in which it was located. There is no obvious correlation across species for the presence of SNPs within a given gene (Appendix 3-5), suggesting that the nucleotide diversity estimate has not been biased by the inclusion of unusually polymorphic or non-polymorphic genes .

Discussion

The DNA sequence diversity estimate (Nei's π) of 0.001 suggests that the construction of a SNP based map of the canine genome is a feasible task with current technology. In the present study, seven of the twelve genes contained useful SNPs within an average of only 500 bp sequenced, suggesting that many of the available canine STSs will also contain SNPs (e.g., Venta et al., 1996; Lyons et al., 1997). Many of the universal mammalian STS primer sets tested on the canine genome will also be directly applicable to the search for SNPs in other mammalian genomes (Appendix 3-5).

The estimated frequency of SNPs in the human genome is about 1 SNP every 500-1000 bp of DNA (Collins et al., 1997; Nickerson et al., 1998; Wang et al., 1998). The unusual number of SNPs found in the canine TS SINE may represent a feature of some repetitive elements and would not seem to reflect the average nucleotide diversity in a species (Harumi et al., 1995). However, their inclusion in the diversity calculation had only a minor impact on the result. The rest of the data suggest that the SNP frequency in the canine genome is about one SNP every 400 bp. The frequency of SNPs (and the within breed π) for a given breed will be lower (Appendix 3-4), but still adequate for most mapping purposes. The data for the feline, bovine, porcine, and equine genomes suggest that the levels of sequence diversity in these genomes may also be similar to that of dog and human.

Pool-and-sequence allows a large number of alleles to be scanned simultaneously for

variability. The current detection limit for the minor allele is between five and ten percent, although work is underway to determine if this detection limit can be lowered. The manual method reported here is similar to recently reported automated PAS methods, but may be more sensitive because of the more uniform banding intensities seen with manual Thermo Sequenase with labeled terminators (Appendices 3-6 and 3-7). By pooling ten animals (20 alleles) at once, it is possible to be at least 95% certain to detect any SNP that occurs with an allele frequency of at least 20% in a given population (Lai et al., 1998).

The method of PCR primer mutagenesis followed by restriction digestion to make allele determinations has the advantage of being relatively inexpensive and simple to use. Many other techniques are also available (e.g., allele-specific PCR, ASO hybridization, or single-base extension) that can be used in combination with automated systems to increase throughput.

It has been suggested (Kruglyak, 1999, Collins et. al., 1999) that a whole genome search for disease genes could be performed utilizing linkage disequilibrium. Based on computer modeling, it had been suggested that the area of linkage disequilibrium in humans will be only a few kilobases (Kruglyak, 1999). Recent empirical data, however, suggest that the area of linkage disequilibrium may be closer to 300 kb (Collins et al., 1999). To our knowledge, no data exists on the range of linkage disequilibrium in dogs. Due to the similarity in breeding patterns between dogs and cows (that is, a significant influence of the founder effect), we suspect that the extent of linkage disequilibrium in

the canine populations will be similar to that found in cattle (Farnir et al., 2000). In cattle, D' values were around 50% at distances of 5 cM, falling off rapidly at distances greater than 5 cM. Further work will be needed to judge the extent of linkage disequilibrium in dogs.

A significant number of SNPs were found within a canine SINE element in the thymidylate synthetase (*TS*) gene. There are an estimated 400,000 SINE elements in the haploid canine genome, although the distribution of the SINES is not known (Bentolila et al., 1999). It may be necessary to use nested primers with at least one of the primers in unique flanking sequence to obtain usable data when a polymorphism is contained within a SINE (e.g., Venta et al., 1999).

We expected to see a decrease in heterozygosity within breeds because some inbreeding is typically used to fix desirable characteristics in lines. Surprisingly, there did not appear to be much deviation from Hardy-Weinberg expectations. However, the number of animals in each data set is small, and larger samples may show an effect of inbreeding like that seen for other breeds using STR markers (Zajc et al., 1997; Morera et al., 1999). A decrease in heterozygosity was also expected when the data was summed across breeds (Wahlund, 1928; Wright, 1951). For the two most polymorphic SNPs tested, one showed the expected decrease (ACTC02), but the other did not (TS04). The reason for this difference is presently unknown.

The SNP markers described in this report are all Type I (gene-specific) STS markers. These markers could be used in concert with high-density DNA probe arrays or other methods of automating SNP typing to make a rapid whole-genome search for linkage (e.g., Wang et al., 1998). It has been calculated that about 1000 evenly spaced SNPs with minor allele frequencies greater than or equal to 0.2 would produce the linkage power of a 300-marker STR map (Kruglyak, 1997). It should be feasible to produce this many Type I SNPs in the near future. SNPs and STRs have complementary advantages for linkage analysis (simpler automation vs. greater single marker genetic power [Brookes, 1999; Weber, 1990]) and therefore it seems likely that they will be used synergistically to locate genes of interest. Finally, because the mutation rate is lower for single nucleotides compared to tandem repeats, SNPs may help to produce a clearer picture of the history and genetic relationships of the many dog breeds.

Acknowledgments

We thank Dr. Jerry Dodgson, Dr. Vilma Yuzbasiyan-Gurkan, Dr. Susan Ewart, and the anonymous reviewers for a number of helpful comments on this manuscript. We also thank Dr. Jill Murtha, Ellen Bishop, Maika Symkowiak, Pam Edwards, and Michael Mienaltowksi for performing some of the typings done in this study. This work was supported by grants from the American Kennel Club Canine Health Foundation, The Morris Animal Foundation, and the Doberman Pinscher Foundation of America, and from private donations from individuals interested in canine health and welfare.

Appendix 3-1

Gene segments with amount of DNA sequenced.

Gene	Species	GenBank 1[a]	GenBank 2[a,b]	Coding Seq.	Intron Seq.
FES	Dog	**L77674**		157	239
	Horse	**L77676**		141	221
	Pig	**L77679**		99	118
	Ox	**L77677**		150	272
WT1	Dog	**U00687**		87	373
	Cat	AF201739	AF201740	89	345
	Horse	AF201736	AF201737	82	342
	Pig	AF202067	AF202068	81	377
	Ox	AF202074	AF201738	85	349
CFTR	Dog	**L77683**	**L77689**	139	248
	Cat	AF203024	AF203023	142	304
	Horse	AF202072	AF202071	146	305
	Pig	AF202075	AF203018	147	327
	Ox	AF201741	AF201742	144	294
TS	Dog	AF201743	AF202073	136	183
	Cat	AF203027	AF203028	134	264
	Horse	AF202065	AF202066	134	313
	Ox	AF203029	AF203030	135	307
C5	Dog	AF202069	AF202070	99	295
CYP1A1	Dog	AF203025	AF203026	147	238
ACTC	Dog	AF203020	AF203019	202	87
GNAT1	Dog	**AF153706**		186	208
NCAD	Dog		AF203017	284	107
RDS	Dog	**U27349**		416	0
RHO	Dog	**X71380**		319	307
RYR2	Dog	AF203022	AF203021	93	436

a. Bold entries in these columns indicate sequences previously submitted by us or by others.
b. Blank cells in this column indicate that the complete sequence is contained in Genbank entry in the column to the left.

Appendix 3-2

Location of SNPs, types of base changes, and diagnostic tests.

Species	SNP	Genbank Accession No.	Nucleotide Position[a]	Type of Change[b]	Restriction Enzyme[c]	Invariant Bands[d]	Diagnostic Bands[d]
Dog	ACTC01	5	289	A/T	Tai I	---	**331,311, 20**
	ACTC02	3	38	C/T	Cac8 I	---	**178,157, 18**
	ACTC03	3	166	C/T	**Tsp509 I**	410,(290), 230	450,310, **144**
	CFTR01	5	52	A/G[e]	**Ban II**	500	500,430, **78**
	CFTR02	5	73	A/G	Taq I	430,**72, 68**,50	400,370, **27**
	CFTR03	3	64	I/D (1)[f]	Bsl I	---	**161,141, 20**
	CFTR04	3	106	C/T	Hinf I	---	**120,100, 20**
	GNAT01	3	123	C/T	Tth111 I	---	400,380, **20**
	NCAD01	3	47	A/G	Bsl I	**95**	**145,128, 17**
	RYR201	5	174	A/G	**Mae III**	---	1010,800, **200**
	RYR201[i]	5	174	A/G	Hinf I	117	**101,75, 26**
	TS01	5	133	A/C	**Mwo I**	450,(182), **132,70, 22,14**	**75,66, 9**
	TS02	5	180	I/D (1)	**Mse I**	225,186, (182),**14**	**246,200, 46**
	TS03	5	215	C/T	**Pst I**	---	770,550, **219**
	TS04	3	221	A/G	**Mse I**	225,186, (182),**14**	**92,87, 5**
	TS05	3	42	I/D (2)	NT	---	---
	TS06	3	46	C/G	NT	---	---

Appendix 3-2 (cont'd).

	TS07	3	54	A/G	NT	---	---
	TS08	3	68	C/T	NT	---	---
	TS09	3	76	A/G	NT	---	---
	TS10	3	92	A/C	NT	---	---
Horse	CFTR01	5	136	C/T	Cac8 I	---	**161,137, 24**
	CFTR02	3	144	C/G	NT	---	---
	TS01	5	169	A/T	**Hinf I**	200,166	**229,192, 37**
	TS02	3	157	A/C	**Tsp509 I**[g]	219,161, **15**	**146,88, 58**
	WT101	5	70	C/T	**Mnl I**	---	**103,68, 35**
Ox	CFTR01	5	233	A/G	Mwo I	---	**274,258, 16**
	TS01	5	127	A/G	Bsl I		390,370, **15**

Appendix 3-2 (cont'd).

Cat	WT01	5	112	A/G	Bsl I	---	**92,76, 16**
	WT02	5	162	A/G	Sau3A I	---	**92,73, 19**
	WT03	3	119	A/G	Nla III	530	**139,104, 19**
Pig	FES01	5	196	C/T	**Hinf I**	---	[340][h],**249**, [190],**215**, [140],**34**
	FES02	5	233	C/T	NT	---	---
	FES03	5	290	C/T	NT	---	---

a. Position of base change in Genbank entry.
b. Base change seen on coding strand.
c. NT = not tested. Bold entries in this column show naturally occurring sites.
d. Band sizes in parentheses indicate confirmed or potential processed pseudogene bands that can occasionally appear if the annealing temperature is not high enough. Bold numbers indicate sizes determined by actual sequence analysis. Other sizes are estimated by the local Southern method (Southern, 1979).
e. Underlined nucleotides are cSNPs.
f. Number represents the number of bases involved in the indel.
g. Hind III was also needed to separate two co-migrating bands that interfered with allele calling.
h. A second segregating *Hinf* I site was also observed in a non-sequenced region that is in complete linkage disequilibrium with the first site for all samples tested so far (this report and Ernst and Raney, personal communication). The band sizes for this site are shown in brackets.
i. Alternative test for RYR201 that uses a more powerful restriction enzyme and a site created by PCR mutagenesis.

Appendix 3-3

Allele counts and heterozygosity for SNPs found among dogs used in the sequencing pool.

Breed	ACTC 01	ACTC 02	ACTC 03	CFTR 01	CFTR 02	CFTR 03	CFTR 04	GNAT1	NCAD	RYR2	TS 01	TS 02	TS 03	TS 04	Across Locus H[c]
Single animal[a]	2 2	1 2[b]	1 2	1 2	1 2	1 2	1 1	1 1	1 2	1 1	1 2	1 2	1 1	2 2	66.7
Cocker spaniel	2 2	1 1	2 2	1 1	1 1	2 2	1 1	2 2	2 2	2 2	2 2	2 2	1 1	1 2	16.7
Greyhound	2 2	1 1	2 2	1 2	1 2	2 2	1 2	1 2	1 2	2 2	2 2	2 2	1 1	2 2	50.0
Doberman pinscher	2 2	1 1	1 2	1 1	1 1	2 2	1 1	1 1	1 1	2 2	1 2	2 2	2 2	1 1	33.3
Siberian husky	2 2	2 2	1 2	1 1	1 1	2 2	1 1	2 2	2 2	1 2	2 2	2 2	1 1	2 2	33.3
Labrador retriever	2 2	1 1	1 2	1 1	1 1	2 2	1 1	2 2	2 2	2 2	2 2	2 2	1 2	1 2	33.3
Collie	2 2	2 2	1 1	1 1	1 2	1 2	1 2	2 2	1 1	1 1	1 2	2 2	2 2	1 1	33.3
Scottish terrier	1 1	1 1	2 2	2 2	2 2	2 2	2 2	1 1	2 2	2 2	2 2	1 2	2 2	2 2	16.7
German shepherd	2 2	1 1	2 2	2 2	1 2	1 2	2 2	1 1	1 1	2 2	2 2	2 2	1 2	1 2	33.3
Beagle	2 2	1 1	2 2	1 2	1 1	1 2	1 2	1 2	1 2	2 2	2 2	2 2	1 2	1 2	66.7
Pointer	1 1	1 1	2 2	2 2	2 2	1 2	2 2	1 1	1 1	2 2	2 2	2 2	2 2	1 1	16.7
Allele 1	4	16	5	12	13	4	11	10	10	17	2	1	9	10	
Allele 2	16	4	15	8	7	16	9	10	10	3	18	19	11	10	
Site heterozygosity	0	0	30	20	30	40	30	20	20	10	20	10	30	40	
Locus heterozyosity[d]			30				50	20	20	10				70	33.3

a. The single animal is presumed to be a mixed breed; the results for this animal are not included in calculations used in the rest of the table for the purebred animals.
b. Heterozygous sites are shaded.
c. Across locus heterozygosity is calculated for each animal (example; for the Cocker spaniel, one gene out of six is heterozygous (16.7%).
d. Locus heterozygosity is calculated on the basis of the number of animals in the pool that are heterozygous for at least one of the sites within the locus.

Appendix 3-4

Genotypes for four SNPs within Canine Breeds.

Breed	No. of Dogs	ACTC02 Genotype 1/1, 1/2, 2/2	RYR201 Genotype 1/1, 1/2, 2/2	TS02 Genotype 1/1, 1/2, 2/2	TS04 Genotype 1/1, 1/2, 2/2
Cocker Spaniel	7	7, 0, 0	0, 0, 6[a]	0, 0, 6[a]	1, 3, 2[a]
Greyhound	8	8, 0, 0	0, 0, 8	0, 0, 8	2, 2, 4
Doberman Pinscher	25	3, 13, 9	0, 0, 25	0, 0, 5	4, 1, 0
Siberian Husky	9	3, 5, 1	0, 1, 7[a]	0, 0, 9	0, 1, 8
Labrador Retriever	11	2, 5, 4	0, 2, 9	0, 0, 11	1, 9, 1
Collie	9	8, 1, 0	4, 5, 0	0, 3, 6	2, 3, 4
Scottish Terrier	11	2, 0, 9	0, 0, 11	2, 5, 4	1, 2, 8
German Shepherd	9	0, 3, 6	0, 0, 9	0, 0, 9	1, 6, 2
Beagle	11	11, 0, 0	0, 1, 10	0, 0, 11	0, 6, 5
Pointer	8	8, 0, 0	1, 0, 7	0, 0, 8	5, 3, 0
Total	108	52, 27, 29	5, 9, 92	2, 8, 77	17, 36, 34

a. One DNA sample failed to amplify in this group.

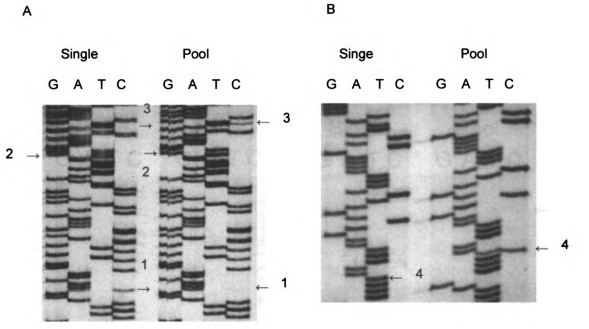
Appendix 3-5

Summary of SNPs identified vs. nucleotide and amino acid conservation in loci studied[a].

Gene	% Identity Amino acid	% Identity nucleotide[b]	SNPs Identified				
			Dog	Horse	Cat	Ox	Pig
ACTC	100	95	3	-	-	-	-
RHO	95	92	0	-	-	-	-
RYR2	99	92	1	-	-	-	-
WT1	96	91	0[c]	1	3	0	0
GNAT1	99	89	1	-	-	-	-
TS	89	89	10	2	0	1	-
NCAD	88	88	1	-	-	-	-
FES	80	87	0	0	-	0	3
CYP1A1	80	83	0	-	-	-	-
C5	79	80	0	-	-	-	-
CFTR	81	80	4	2	0	1[c]	0
RDS	91	76	0	-	-	-	-

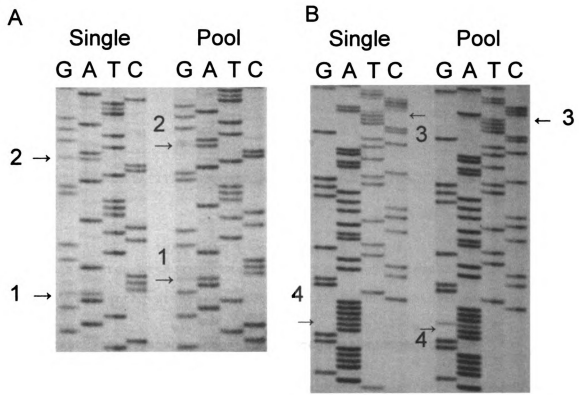
a. All SNPs were confirmed by restriction digestion except for six contained in a SINE in the canine TS gene, one horse CFTR SNP, and two pig FES SNPs.
b. The table is sorted by descending percentage of nucleotide sequence identity for human compared to rat or mouse.
c. Contains a microsatellite.

Appendix 3-6



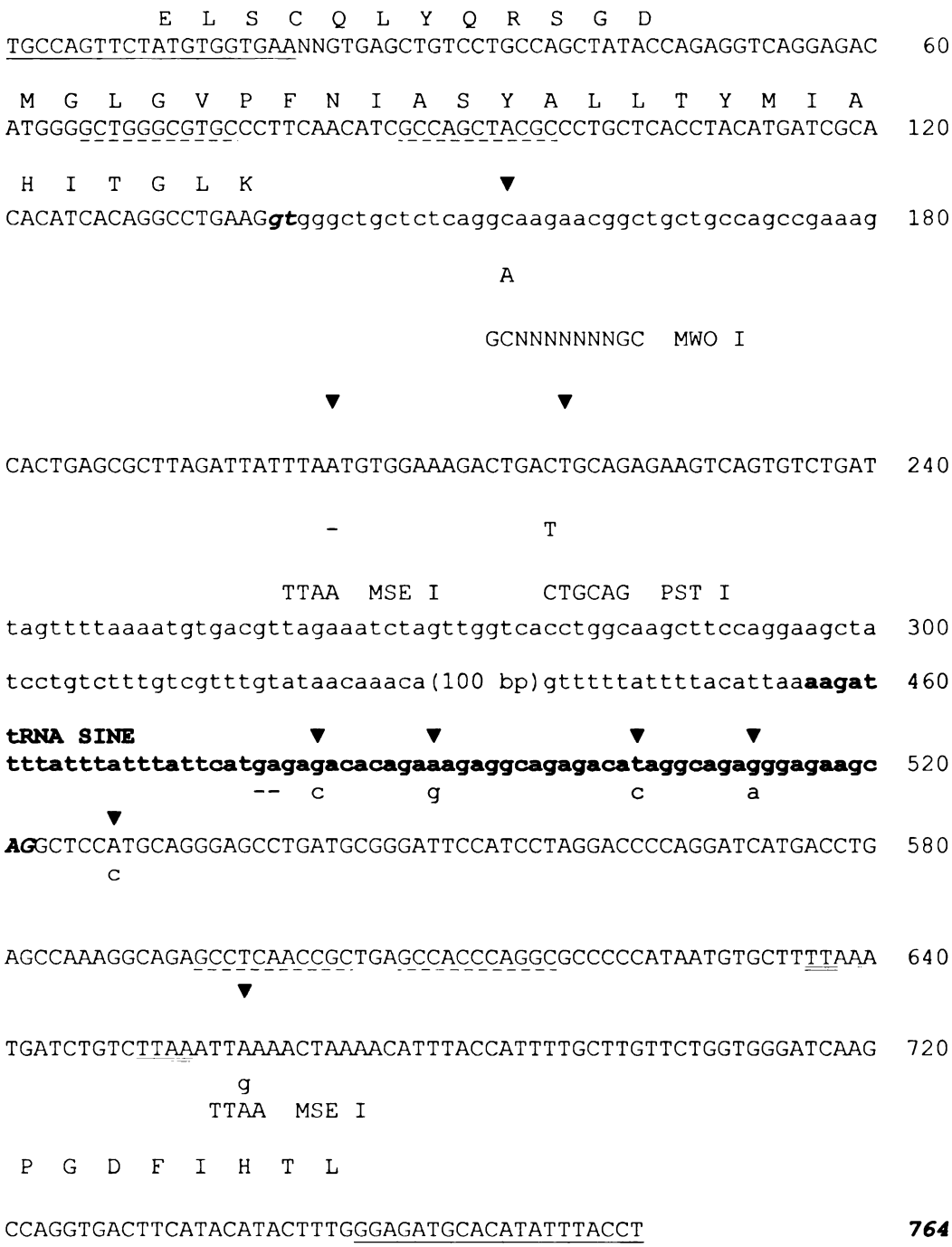
Identification of four SNPs in the canine *TS* gene by pool-and-sequence. Arrowheads indicate the identical location of four SNPs in a single animal and the pool of ten animals from different breeds as seen in sequencing ladders using the downstream (panel A) and upstream amplification primers (1 = TS01, 2 = TS02, 3 = TS03, 4 = TS04). Approximate allele frequencies can be visually estimated from the relative intensities of the two bands seen in the pool. Arrowhead 1 shows the position of A and C bands in the heterozygous single animal. The C overlies a naturally occurring *Mwo*I restriction enzyme site. The C band also appears weakly in the pool. Arrowheads labeled as 2 show the position of an insertion or deletion (indel) of a single base in the pool and heterozygous single animal. Note the doubling effect of bands above this point. Shadow bands could be faintly seen in the pool on the original film. One indel allele overlies a naturally occurring *Mse*I site. Arrowhead 3 shows two alleles in the pool (C and T) which is homozygous (T) in the single animal. The C allele overlies a naturally occurring *Pst*I site. Arrowhead 4 shows the location of a SNP found using the downstream primer. The relative location of these SNPs is shown in Fig. 3.

Appendix 3-7



Identification of four SNPs in the canine CFTR gene. The arrowheads in panel A indicate SNPs CFTR01 (arrowhead 1) and CFTR02 (2). These are two of the three coding region SNPs found in this study. The arrowheads in panel B indicate SNPs CFTR03 (3) and CFTR04 (4). For CFTR04, note the characteristic fall off in intensity in the A lane at the position of the SNP relative to the other bands in the A lane. CFTR03 is a second example of an indel SNP.

Appendix 3-8



Location of four SNPs identified in the canine *TS* gene. The PCR primers are underlined. The deduced amino acid sequence is shown above the nucleotide sequence. The invariant GT and AG splice signals are shown in bold italics. SNP alleles are identified with a triangle (▼) above the SNP, with the diagnostic restriction site shown underneath each SNP. A SINE is shown in bold. Double-underlined sequences are invariant *Mse*I sites and dash-underlined sequences are invariant *Mwo*I sites.

Appendix 3-9

UM-STSs primers used for PCR amplification and sequencing

Locus[a]	Primer 1	Primer 2	Size[b] (bp)	Time[c] (min)	Temperature (°C)	Reference
ACTC	5'GCCCTGGATTTTGAGAATGAGAT	5'ACGATCAGCAATACCAGGGTACA	1100	1, 2, 3	95, 57, 72	This report
C5	5'TGTGTACGATTCCGGATATTTGA	5'GCTCCTTCACAGACTTTCTG	1170	0.5 ,1.5, 4	95, 57, 72	Venta et al., 1996
CFTR	5'CTAAGCCATGGCCACAAGCA	5'CATTGCTTCTATCCTGTGTT	1020	0.5, 1.5, 4	95, 59, 72	Venta et al., 1996
CYP1A1	5'TTGGACCTCTTTGGAGCTGG	5'TGGTTGATCTGCCACTGGTT	600	0.5 ,1.5, 4	95, 57, 72	Venta et al., 1996
FES	5'GGGGAACTTTGGCGAAGTGTT	5'TCCATGACGATGTAGATGGG	450	0.5 ,1.5, 4	95, 57, 72	Venta et al., 1996
GNAT1	5'AGCACCATCGTCAAGCAGA	5'CTGGATACCCGAGTCCTTC	495	1, 1, 1	95, 61, 72	Brouillette and Venta, 2000
NCAD	5'CCATTAGCCAAGGGAATTCA	5'CCATTGTCAGAAGCAAGGA	1900	1, 1, 3	95, 57, 72	This report
RDS	5'TTTGACCAGAAGAAGCGGGT	5'TTGCTGATCCACTGAATCTC	550	1, 2, 3	95, 57, 72	This report
RHO	5'TACATGTTCGTGGTCCACTT	5'TGGTGGGTGAAGATGTAGAA	350	0.5, 1.5, 4	95, 57, 72	This report
RYR2	5'ATGCTTCATCTGTGGGATAG	5'TGTGTGTTCTGTTTCATCTT	1010	1, 1, 1	95, 57, 72	This report
TS	5'TGCCAGTTCTATGTGGTGAA	5'AGGTAAATATGTGCATCTCC	790	1, 2, 3	95, 57, 72	Venta et al., 1996
WT1	5'GAGAAACCATACCAGTGTGA	5'GTTTTACCTGTATGAGTCCT	800	1, 2, 3	95, 55, 72	Venta et al., 1996

a. Proteins encoded at the loci are: *ACTC*, alpha cardiac actin, *C5*; complement protein 5; *CFTR*, cystic fibrosis transmembrane regulator; *CYP1A1*, cytochrome P_1 450 (drug inducible); *FES*, cellular homologue of the feline sarcoma virus proto-oncogene; *GNAT1*, transducin alpha 1; *NCAD*, neural cadherin; *RDS*, retinal degeneration slow protein; *RHO*, rhodopsin; *RYR2*, cardiac ryanodine receptor; *TS*, thymidylate synthetase; and *WT1*, Wilm's tumor protein 1.

b. Sizes are given for the canine product. Sizes varied less than two-fold for other species.

c. All PCR reactions were run for 35 cycles.

Appendix 3-10

Diagnostic primers for testing SNPs[a]

Species	Locus	Primer 1	Primer 2	Size (bp)	Time (min)	Temp. (°C)
Dog	ACTC01	5'GCCCTGGATTTTGAGAATGAGAT	5'GTCAGTGTGTCATCTTTGAC	331	1,1,1	95,57,72
Dog	ACTC02	5'ACCTGTTCCACAAACGGCTT	5'ACGATCAGCAATACCAGGGTACA	178	1,1,1	95,57,72
Dog	CFTR02	5'CCCAGTGCTCATTTGGATTC	5'CATTGCTTCTATCCTGTGTT	930	1,2,3	95,57,72
Dog	CFTR03	5'AGGTGTGTCATCACCGGTA	5'CATTGCTTCTATCCTGTGTT	161	1,1,1	95,55,72
Dog	CFTR04	5'GTGATGTTACATGTGCCAGATT	5'CATTGCTTCTATCCTGTGTT	120	1,1,1	95,61,72
Dog	GNAT101	5'AGAGCACCATCGTCAAGCA	5'TGCATCAGCTTCCGGACATC	400	1,1,1	95,61,72
Dog	NCAD01	5'AATTAGTGAGCTGCCGACAT	5'CCATTGTCAGAAGCAAGGA	240	1,1,1,1	95,53,57,72
Dog	RYR201[b]	5'ATGCTTCATCTGTGGGATAG	5'ACGGTATAAAGAAACGGGAG	215	1,1,1	95,57,72
Dog	TS01,02	5'TACCAGAGGTCAGGAGACAT	5'CAGACACTGACTTCTCTGC	195	1,1,1	95,57,72
Ox	CFTR01	5'CTAAGCCATGGCCACAAGCA	5'GTACCTGTACGCATGTTAAG	274	1,1,1	95,57,72
Ox	TS01	5'CTGAAGGTGGCCTGCTCTGG	5'GAGTGTGCACGAAGTCACCTG	390	1,1,1	95,59,72
Horse	CFTR01	5'TACTGTGCAATCAGCAAAAG	5'GTAATTTAAAGAACTAAATTGCTA	161	1,2,3	95,57,72
Horse	WT101	5'GAGAAACCATACCAGTGTGA	5'CTAACAATGTGGGAGCATCA	103	1,1,2	95,57,72
Cat	WT101,02	5'CTTTGTTGTTGCCTATGGATG	5'ACAGGAAAATCCCCAGAGAT	92	1,2,3	95,57,72
Cat	WT103	5'GAGAAACCATACCAGTGTGA	5'CTAACAATGTGGGAGCATCA	670	1,1,1	95,57,72

a. Table shows only those diagnostic primer sets that are different from those used to perform the original pool-and-sequence experiments (Table 2). Nucleotides that were altered to create restriction sites for one allele are underlined.

b. Alternative test for RYR201.

Chapter 4

Within-Breed Heterozygosity of Canine Single Nucleotide Polymorphisms Identified By Across-Breed Comparison

Within-Breed Heterozygosity of Canine Single Nucleotide Polymorphisms Identified By Across-Breed Comparison

J A Brouillette[1,2,4], P J Venta[1,2,3,5]

Microbiology and Molecular Genetics[1], Genetics Program[2], Small Animal Clinical Sciences[3], College of Human Medicine[4], and College of Veterinary Medicine[5], Michigan State University, East Lansing, MI 48824-1314

Correspondence:

Patrick J. Venta, Ph.D.

Small Animal Clinical Sciences

College of Veterinary Medicine

Michigan State University

East Lansing, MI 48824-1314

Phone: 1-517-432-2515

Email: venta@cvm.msu.edu

Summary

Identification of single nucleotide polymorphisms (SNPs) by DNA sequence comparison across breeds is a strategy for developing genetic markers that are useful for many breeds. However, the heterozygosity of SNPs identified in this way might be severely reduced within breeds by inbreeding or population subdivision. The effect of inbreeding and population subdivision on heterozygosity of SNPs in dog breeds has never been investigated in a systematic way. We determined the genotypes of dogs from three divergent breeds for SNPs in four canine genes (ACTC, LMNA, SCGB, and TYMS) identified by across-breed DNA sequence comparison and compared the genotype frequencies to those expected under Hardy-Weinberg equilibrium. Although population subdivision significantly skewed allele frequencies across breeds for two of the SNPs, the deviations of observed heterozygosities compared to those expected within breeds were minimal. These results indicate that across-breed DNA sequence comparison is a reasonable strategy for identifying SNPs that are useful within many canine breeds.

Keywords: SNPs, Beagle, Doberman Pinscher, Scottish terrier, DNA pooling, inbreeding coefficient

Introduction

Purebred dog populations are affected by hundreds of different genetic diseases, many of which are homologous with human genetic diseases (Patterson 2000). The canine diseases are the bane of breeders and owners alike, while at the same time they present useful models of human genetic diseases. The same methods used to find human disease genes and mutations can be used to find canine disease genes and mutations (e.g., Yuzbasiyan-Gurkan et al, 1996). The most comprehensive method is linkage analysis, which requires the availability of genetic markers spread throughout the genome.

Single nucleotide polymorphisms (SNPs) have become more readily accessible for use in human mapping projects, and are thought to have good potential for very high throughput automation (Brookes 1999; Shi, 2001). A nucleotide diversity estimate in dogs suggests that the number of SNPs present in the dog genome is similar to the number found in the human genome and could be used for automated canine genotying (Brouillette et al, 2000). Unfortunately, few canine SNPs have been identified to date.

We recently developed a simple method to identify SNPs in any mammalian genome that can help to increase the number of SNPs for use in canine genetic studies (Brouillette et al. 2000). A pool of DNAs from different dogs is manually sequenced, after which SNPs are identified as bands occurring at identical positions in the sequencing ladder. In order to find SNPs that are useful for as many breeds as possible, we pool DNA from single representatives of ten diverse breeds. However, there is concern that the SNPs found

using a diverse pool will be of little use within dog breeds because variability will have been greatly reduced or lost due to inbreeding or population subdivision.

In order to test the hypothesis that these factors prevent useful SNPs from being found by pooling DNA samples across breeds, we determined over 500 genotypes using four of these SNPs in 42 to 49 individual dogs from three different breeds, and tested for significant departures from Hardy-Weinberg equilibrium predictions. We also calculated the inbreeding coefficients (a measure of the reduction of heterozygosity) for each breed-SNP combination.

Materials and Methods

The dogs genotyped were Beagles, Doberman pinschers, and Scottish terriers. They were unrelated for at least three generations, and should be a reasonably random sample of each breed. Some of the Beagles used were part of the DoGMap REfference Panel. Other dog DNA samples came from other projects in our lab.

DNA was isolated after collection of whole blood from dogs or from buccal cells (Brouillette et al. 2000; Richards et al. 1992). SNPs were selected to give a broad spectrum of allele frequencies as identified from a pool of single dog DNAs (Brouillette et al. 2000). Samples were genotyped for SNPs in the following genes: TYMS, SGCB, ACTC, and LMNA. Primers, amplification conditions, and restriction enzymes used for genotyping are given in Appendix 4-1. Two of the four SNPs were previously reported, including small genotype surveys (TYMS SNP TS04, and ACTC SNP ACTC02 in Brouillette et al. 2000). Allele frequencies in the ten-breed pool for these two SNPs were 0.50 and 0.20 (allele 2), respectively.

The cross-species primers for LMNA were 5'-ATCGCATCGACCTCCTCTC and 5'-AGGTCCTGGGACATGGCTGG. The LMNA SMP is a G/A transition located in intron 5 (position 174 in Genbank accession no. AF427092) and , in a survey of ten single dogs of different breeds (the same dogs used in the pool to identify the SNPs), allele 1 occurred at a frequency of 0.10. The cross-species primers for SGCB were 5'-ATTGGACCAAATGGCTGTG and 5'-GTCCTCGGGTCAAAAAACT. The SGCB

SNP is a T/C transition located in intron 4 (position 265 in Genbank accession number AF427093) and, in a survey of the same ten single dogs, allele 1 occurred with a frequency of 0.45.

Results and Discussion

The inbreeding coefficient was calculated using the equation $F = (H_e - H_o)/H_o$, where H_e is the expected number of heterozygotes based on Hardy-Weinberg equilibrium and H_o is the observed number of heterozygotes (Juneja et al. 1981). Among breed comparisons was calculated in the same manner combining data for all three breeds for each SNP. Chi-square values were used to determine significance.

Genotypes are given in Appendix 4-2. A few amplifications failed for some of the breed-SNP combinations, so that the number of dogs typed per breed is not always the same for the different SNPs. There were no cases in which there was a statistically significant difference between observed and expected genotypes based on Hardy-Weinberg equilibrium for any SNP within a given breed (Appendix 4-2).

Under the assumption that inbreeding is actually present but not statistically detectable, we calculated inbreeding coefficients. They ranged from a high of 0.143 for the SGCB SNP in Beagles to a low of -0.190 for the same SNP in Doberman pinschers. The average inbreeding coefficient for Beagles was 0.047 for the two markers for which inbreeding coefficients could be calculated, for Doberman pinschers was -0.044, for the four SNPs studied, and for Scottish terriers was 0.019 for the three SNPs that were variable in this breed.

The LMNA SNP showed variation only in the Doberman pinscher (allele 1 at 0.10). The ACTC SNP had allele 2 frequencies of 1.00, 0.38, and 0.67, and the SGCB SNP had allele 2 frequencies of 0.56, 0.84, and 0.44 in Beagles, Doberman pinschers, and Scottish terriers, respectively. The TYMS SNP had allele 1 frequencies of 0.39, 0.81, and 0.30 for the above breeds respectively. For the three breeds combined into one population, significant deficiencies of heterozygotes were seen for two SNPs (ACTC F = 0.298, P = 0.002; TYMS F = 0.223, P= 0.032), but not for the other two (SGCB F = 0.099, P = 0.51; LMNA F = -0.018, P = 0.98).

If the average inbreeding rates were high within breeds, correspondingly greater numbers of markers or dogs would be needed to obtain statistically significant linkages in genome scans. The inbreeding coefficients for the breeds found in this study suggest that inbreeding has only a very minor impact on the observed heterozygosity for SNPs. Population subdivision appears to have a greater impact on the variability of SNPs across breeds, but reasonable variability was still found within breeds in many cases (Appendix 4-2). This loss in heterozygosity across breeds is not unique to SNPs, but is shared with other commonly used marker types as well (Koskinen and Bredbacka, 2000; Juneja et al. 1981). Although further studies should be performed, we conclude from the present data that population subdivision and inbreeding will probably not preclude the use of pooling across breeds to find SNPs that are useful for studies within breeds.

An important implication of this study is that it should be possible to predict the number of heterozygotes within a purebred canine population with reasonable accuracy directly from allele frequencies under the assumption of Hardy-Weinberg equilibrium. Because

allele frequencies for SNPs can be determined within a few percent in a pooled DNA sample by various methods, the utility of a given SNP can be deduced by typing one pooled s\ample before a decision is made to type that marker in numerous individual dogs (Shi 2001). The identification of more canine gene-specific SNPs in the future by across-breed comparisons will make it possible to take advantage of the automated genotyping systems for mapping and ultimately identifying the genes that cause canine genetic diseases.

Acknowledgements

We thank Mr. Mark Stinson for help in obtaining some of the samples used in this study and Dr. Donna Housley for helpful suggestions. This work was supported by the Morris Animal Foundation, the Doberman Pinscher Foundation of America, Inc., and the American Kennel Club Canine Health Foundation with co-sponsorship by the Westie Foundation of America, the Health Trust Fund of the Scottish Terrier Club of America, and the Cairn Terrier Club of America.

Appendix 4-1

Diagnostic primers, restriction enzymes, and SNP locations

Locus	Primer sequences[1,2]	Restriction Enzyme[3]	Allele 1 (bp)	Allele 2 (bp)	Nucleotide change[4]	Genbank accession no.	Reference[5]
ACTC	5'-ACCTGTTCCACAAACG***G***CTT	Cac8I	178	157, 18	C/T, 38	AF203019	1
	5'-GCCAGCAGATTCCATACCTA						
LMNA	5'-ATCCGCATCGAC***C***GCCTCTC	Cac8I	127	107, 20	G/A, 174	AF42709	2
	5'-AGGTCCTGGGACATGGCTGG						
SCG8	5'-AACTGCATACCAATGT***G***ACT	HinfI	148	128, 20	G/A, 265	AF427093	2
	5'-ATGTTCTTGATGAATTTTGGC						
TYMS	5'-TGCCAGTTCTATGTGGTGAA	MseI	92[6]	97, 5	G/A, 221	AF202073	1
	5'-AGGTAAATATGTGCATCTCC						

1. Letters in bold and italics indicate altered bases used to create restriction sites for SNPs.
2. PCR conditions were 50 mM KCl, 10 mM Tris-Cl (pH8.3 at 20oC) 1.5 mM MgCl2, 100 µM dNTPs, 10 pmol each primer, 1-50 ng target DNA and 0.5 U Taq DNA polymerase per 25 µl reaction. Cycling conditions were 94o C, 1 min, 57o C, 1 min, 72o C, 5 min, for 35 cycles (blood DNA samples) or 30 cycles (buccal DNA samples), except the annealing temperature for LMNA was 65o C, and for TYMS the annealing time was 2 min and the extension time was 3 min.
3. Restriction enzyme digestions were carried out in PCR buffer with the addition of MgCl2 to a final concentration of 10 mM, and at the manufacturer's recommended temperature for 4-12 hr.
4. Number indicates nucleotide position in corresponding GenBank entry.
5. Number 1 is Brouillette et al. (2001) . Number 2 is this report.
6. Several invariant bands were also seen for this marker.

Appendix 4-2

Genotypes for dogs from three breeds using four SNPs.

Locus	Breed	Genotype	Expected[1]	Observed	p[2]
ACTC	Beagle	1/1	0	0	
		1/2	0	0	NA[3]
		2/2	42	42	
	Doberman Pinscher	1/1	19.0	19	
		1/2	23.6	24	0.99
		2/2	7.2	7	
	Scottish Terrier	1/1	5.3	6	
		1/2	21.3	20	0.86
		2/2	21.3	23	
LMNA	Beagle	1/1	0	0	
		1/2	0	0	NA
		2/2	42	42	
	Doberman Pinscher	1/1	0.1	0	
		1/2	4.8	5	0.95
		2/2	45.1	45	
	Scottish Terrier	1/1	0	0	
		1/2	0	0	NA
		2/2	47	47	
SGCB	Beagle	1/1	8.5	10	
		1/2	21.0	18	0.66
		2/2	12.7	14	
	Doberman Pinscher	1/1	1.2	0	
		1/2	12.6	15	0.41
		2/2	33.2	31	
	Scottish Terrier	1/1	14.7	14	
		1/2	23.1	25	0.85
		2/2	9.1	8	
TYMS	Beagle	1/1	6.4	6	
		1/2	20.0	21	0.95
		2/2	15.6	15	
	Doberman Pinscher	1/1	32.1	32	
		1/2	15.1	14	0.94
		2/2	1.8	2	
	Scottish Terrier	1/1	4.4	4	
		1/2	20.6	19	0.88
		2/2	24.0	25	

1. Under Hardy-Weinberg Equilibrium.
2. Chi-square probability.
3. NA = not applicable for chi-square analysis because expected cells are 0.

Chapter 5

Summary

Summary

In this thesis I have; (1) developed a method to produce genetic markers to test hypotheses regarding specific candidate genes as underlying traits and diseases in dogs and other domestic animals, and (2) shown that the markers developed across canine breeds are useful for linkage studies within breeds. Each of the methods will be briefly discussed in turn.

This process began with the design of canine gene-specific primer sets and establishing sequence tagged sites for several genes within the canine genome. For the work found in chapter two, cross-species PCR primer sets, referred to as UMSTSs, were developed for a total of 86 genes. Through additional work, UMSTSs have been developed for over 1100 genes (Chapter 4, Chapter 3, Venta and Vidal, 1999, Brouillette and Venta, 2000; Brouillette et al., 2000; Housley et al., 2006). The UMSTS were designed to optimize their usefulness across a wide spectrum of mammalian genomes. DNA was amplified from a set of ten different mammalian species with a subset of 11 primer pairs, and the amplified PCR products were sequenced to confirm that the PCR products were the species-specific orthologs of canine gene of interest. This verified the usefulness of these primer sets across mammalian species. Others have shown the utility of these primer pairs in the identification of genes in the equine and porcine genomes (Shubitowski et al, 2001; Farber et al., 2003).

Gene-specific sequence tagged sites could be used in a variety of ways. One way was to use the gene-specific amplification product to screen canine DNA libraries for the

presence of clones containing the gene of interest. The genes of interest could then be isolated from the library and a search made for SSLP markers within or near the gene of interest. The cloned gene could also be directly sequenced to yield nucleotide information about the particular mammalian gene.

With the development of a radiation hybrid map in the canine genome (Vignaux et al., 1999, Guyon et al, 2003, Housley et al, 2004, Breen et al., 2004), the primer pairs could be used to determine physical linkage groups and chromosome locations of specific genes within the canine genome.

At the time of this work, a limited amount of DNA sequence data was available for the canine genome. With the availability of the complete canine genome sequence (Lindblad-Toh et al., 2005), there is no need to design cross-species PCR primers for canine genes. However, the primers are still useful in other mammalian species when the species-specific nucleotide sequence of a gene of interest is unknown (eg. Brinkmeyer-Langford et al., 2005; Aitken et al., 2004).

The next step was to identify and isolate markers within specific genes within the canine genome. Our collaborator, Dr. Vilma Yuzbasiyan-Gurkan, has developed several hundred randomly occurring SSLP markers within the canine genome, which would be even more useful if they were linked to gene-specific markers, so that a comparative positional candidate gene approach could be realized. Because of the time and labor involved in isolating clones for the identification of SSLPs, it seemed that identification

of SNPs within the amplified intervals would be more efficient, provided that the canine SNPs were dense enough that many of the amplicons would contain a SNP.

The pool-and-sequence technique was developed to identify SNPs within genes of interest (Chapter 3). It was observed that individuals that were heterozygotes at a given nucleotide location in a gene being sequenced had two bands at a given position on the sequencing ladder that were about half of the intensity of the adjacent sites with single bands. I began experimenting with this observation to see if other ratios of nucleotide frequencies at a given position in the sequencing ladder would be indicated by corresponding changes in the intensities between bands in the sequencing ladder. Both a qualitative and a quantitative method to determine allele frequencies were developed, if indeed a polymorphism existed in a given DNA sample being sequenced.

This technique was used to locate polymorphisms within the canine genome that would be heterozygous across multiple breeds. A pool of DNA was created from a panel of ten dog breeds. These dogs were chosen based on the assumption that if the breeds were phenotypically different enough to be placed into different breeds by The American Kennel Club, they would also be different at the nucleotide level. Objective evidence by another lab has confirmed that this group of dog breeds is broadly representative of all other dog breeds (Parker et al., 2005). Taking the ability to detect alleles using the pool-and-sequence method and the ability to design primers for most canine genes based on the nucleotide sequence of known index species (Chapters 3 and 2), a search for SNPs was made in 12 canine genes. From this work, an estimate of nucleotide diversity was

made for the canine genome (Chapter 3). The nucleotide diversity was high enough that a SNP with a minor allele frequency of 0.2 was identified in all the genes studied. Similar frequencies were found in the other mammalian species investigated, suggesting that it is feasible to use this method to identify a SNP marker in virtually any candidate gene and any mammalian species of interest. It may, however, be necessary to sequence additional areas of a gene of interest if a SNP was not found within the first area amplified. These markers would then be used to test the hypothesis that a given candidate gene was associated with a given disease phenotype in a breed. This work also showed that the frequency of occurrence of SNPs in the canine genome is likely to be sufficient to perform genome-wide scans for linkage in the canine genome. This conclusion was borne out by the work of Lindblad- Toh et al. (2005). Finally, the work in chapter 3 provided the first estimate of nucleotide diversity and frequency of SNP occurrence in the canine genome.

Currently, a set of SNPs is being established throughout the canine genome (Lindblad-Toh et al., 2005), and a rapid scan of the entire canine genome will soon be possible for linkage or association studies using DNA arrays. It is estimated that approximately 10,000 evenly-spaced SNP markers would be necessary for association studies in the canine genome (Ostrander and Kruglyak, 2000; Sutter et al., 2004; Lindblad-Toh et al., 2005). The recent development of commercial arrays of 20,000 SNPs may make scans for association in the dog a practical reality (Lindblad-Toh, 2006).

The conserved synteny between canine, human, primate, and murine genomes is approximately 94% (Lindblad-Toh et al., 2005). This synteny immediately leads to comparative positional candidates for disease genes if a mutation causing a disease phenotype in humans or mice is mapped to a known location within the genome. The area of interest in the canine genome will correspond to the syntenic region of the human or murine genome where a candidate gene has already been mapped for a given disease phenotype of interest. Markers mapping to genes in that region could be immediately tested for association or linkage.

The pool-and sequence-technique allowed us to explore how much diversity would be found in the canine genome. Based on sequence data that incorporated fragments of 12 canine genes and 5.4 kbp of DNA, an estimate was made of how frequently a SNP would be found in the canine genome, approximately once every 400 nucleotides. It was estimated that the nucleotide diversity for exons and introns is 0.001 and 0.004, respectively in the canine genome (Chapter 3). These values are similar to what is found in the human genome (Collins et al., 1997; Wang et al., 1998), and are similar to the estimate of Lindblad-Toh et al. (2005). Lindblad-Toh et. al. compared nucleotide sequence data between a single Boxer and a single individual from various other dog breeds and found that the rate of SNP occurrence was 1 in 800 nucleotides (Lindblad-Toh et al., 2005). Preliminary estimates on a smaller data set in the feline, bovine, porcine, and equine genomes suggest similar nucleotide diversity in these species (data not shown). For the previously reported work (Chapter 3), the limit of detection for the minor allele is about 5% allele frequency. Other studies indicate that a detection limit of around 1% allele frequency for the minor allele is possible in our hands (Data not

shown). The manual pool-and-sequence method was used because it was considerably less expensive than automated methods at the time these experiments were performed. The relative cost of the two methods has changed and automated methods are now less expensive. Automated sequencing is recommended for SNP discovery in non-sequenced genomes.

A method of PCR mutagenesis was used to create restriction sites for endonuclease digestion that will preferentially cleave DNA with one or the other allele at the SNP site. This became a rapid and technically reliable means to type the SNPs that had been previously identified. Although primer mutagenesis remains an option for small scale studies, it would not be appropriate for large studies.

A final study was undertaken to determine if SNPs found across multiple breeds would be useful as markers for within-breed linkage or association analysis. This information would also be used to estimate coefficient of inbreeding by means of a reduction in heterozygosity within breeds.

It was known that population subdivision would decrease the heterozygosity of the individual population divisions within a population taken as a whole. This is known as the Wahlund Effect. If a population is divided into smaller units and breeding within those units are restricted, there will be a decrease of heterozygotes among the entire population due to the subdivisions, assuming that the alleles under study have different frequencies in the various subpopulations. An example of this effect is given in Appendix 5-1.

In this study, the genotypes were determined at four different SNP markers using 42-49 individuals from three different dog breeds. It was determined that population subdivision does significantly skew allele frequencies, at least in the breeds studied, and the coefficient of inbreeding ranged from 0.297 to -0.018. These values are in close agreement with other canine breeds. For example, Dorn and Schneider (1976) calculated inbreeding coefficients for 14 breeds of dog which varied from a low of 0.009 in Boxers to a high of 0.117 in German Shepherds. The median value across all 14 breeds tested was 0.18.

As a practical matter, there is sufficient heterozygosity of SNP markers within breeds for genetic studies to be undertaken without a significant increase in the number of animals or markers required for detection of linkage or association. The coefficient of inbreeding will be affected by a number of factors including popular sire effects, recent population bottlenecks, the age of the breed, and the population size of a given breed (Dorn and Schneider, 1976, Kathmann et al., 1999, Mandigers et al., 1993).

If these results can be generalized across breeds, the major implication is that SNP markers developed across breeds will be useful within breeds to a large extent, though the utility of breed-marker combinations will vary among the population of interest. This data supports the feasibility of the development of a panel of SNP markers across the canine genome to be used as part of a DNA array. The resulting array will have utility within many breeds for traditional linkage studies as well as candidate gene and allele association studies, with varying levels of usefulness depending on the breed under study and the markers used in the whole-genome analysis (Sutter et al., 2004; Parker et al., 2005).

The future of canine genomics and genetics is bright. The identification of SNPs within genes will allow for an increasingly dense canine linkage map (Mellersh et al., 2000), with more and more disease causing mutations being mapped in the coming years (Giger et al., 2006). Coupling this with the physical map (Mellersh et al., 2000; Guyon et al., 2003, Housley et al., 2004) will allow for increasingly detailed mapping of genes. DNA array technology will allow for higher throughput of linkage and association studies and further increase our knowledge of the canine genome as well as our understanding of mammalian evolution. The structure of the canine genome and pattern of breeding will make genome-wide linkage disequilibrium studies feasible in the near future, given the large number of SNP markers available today (Sutter et al., 2004; Ostrander and Kruglyak, 2000; Lindblad-Toh et al., 2005).

The first draft of the canine genome is now available with the identification of over 2.5 million SNP markers (Lindblad-Toh et al., 2005). The influx of data from this project will expand our knowledge of canine genetics as well as comparative genetics. There will likely be an explosion of information as disease-causing mutations are identified at an increasingly rapid rate. Breeders can use this information to improve overall breed health by selective breeding practices. This will lead to a better life for dogs and their owners alike.

Appendix 5-1

Brief Explanation of the Wahlund Effect

The Wahlund Effect is an observation that when a population is divided into subpopulations, there is an increase in the frequency of homozygotes relative to what would be expected by Hardy-Weinberg Equilibrium, even if the two subpopulations are each in Hardy-Weinberg Equilibrium[a,b]. This is a generalizable rule among populations which are stratified into subpopulations as long as the allele frequencies for a given gene are not equal to one another in the subpopulations.

As an example of this, consider a population of two hundred individuals divided into two subpopulations, each with 100 individuals. Assume that the subpopulations are each in Hardy-Equilibrium. For a gene, A, there are two alleles, "A" and "a". In subpopulation 1, the allele frequency of A is 0.7. In subpopulation 2, the allele frequency of A is 0.4.

By assuming the subpopulations are each in Hardy-Weinberg equilibrium, we can conclude that the allele frequency of a is 0.3 and 0.6, in subpopulations 1 and 2 respectively. We can use this information to calculate the occurrence of the various genotypes for gene A in each subpopulation.

Based on Hardy-Weinberg Equilibrium, the frequency of occurrence of the three genotypes in the two subpopulations is given below.

Subpopulation 1		Subpopulation 2		
Genotype	No. of Individuals	Genotype	No. of Individuals	Total
AA	49	AA	16	65
Aa	42	Aa	36	78
aa	9	aa	48	57

The allele frequency of allele A for the entire population is calculated as $2(65) + 78 / 2(65 + 78 + 57) = 0.52$

Given the frequency of allele A in the population is 0.52, the expected allele frequencies in the total population is:

(Expected): AA = 54.08 Aa = 99.84 aa = 46.08

(Observed): AA = 65 Aa = 78 aa = 57

This depression of heterozygotes is a mathematical phenomenon that can be generalized to any allele frequencies p1 and p2 in subpopulations 1 and 2, respectively, as long as the allele frequency in subpopulation 1 is not the same as the allele frequency of subpopulation 2, that is, $p1 \neq p2$.

If these subpopulations are in Hardy-Weinberg Equilibrium, the second allele in each subpopulation, q1 and q2, respectively, are related to one another by the following equation.

$p1 + q1 = 1$

$p2 + q2 = 1$

The frequency of the occurrence of heterozygotes in each subpopulation is:

2p1q1 and 2p2q2 respectively.

The Heterozygosity, H, for the entire population is simply the mean of the heterozygosities of each of the subpopulations. Stated mathematically, this is:

$$H = \frac{2p1q1 + 2p2q2}{2}$$

$$H = p1q1 + p2q2$$

$$H = p1(1 - p1) + p2(1 - p2)$$

This value is always smaller than

$p(1 - p)$, which equals $2pq$, unless $p = q$.

a. http://blackwellpublishing.com/ridley/a-z/Wahlund_effect.asp
b. http://statgen.dps.unipi.it/courses_file/documents/pdf/Holsinger-Wahlund.pdf

BIBLIOGRAPHY

Aitken, N., Smith, S., Schwarz, C. and Morin, P. "Single Nucleotide Polymorphism (SNP) Discovery in Mammals: A targeted-gene Approach" *Molecular Ecology* **13:** 1423-1431 (2004).

Alcalay, M., Antolini, F., Van de Ven, W.J., Lanfrancone, L., Grignani, F., and Pelicci, P.G. "Characterization of human and mouse c-fes cDNA clones and identification of the 5' end of the gene", *Oncogene* **5:** 267-275 (1990).

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Ahang, Z., Miller, W., and Lipman, D.J. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Research* **25:** 3389-3402 (1997).

Ameratunga, R., Winkelstein, J.A., Brody, L., Binns, M., Cork, L.C., Colombani, P., and Valle, D. "Molecular analysis of the Third Component of canine Complement (C3) and Identification of the Mutation responsible for Hereditary Canine C3 Deficiency", *Jounal of Immunology* **160(6):** 2824-2830 (1998).

American Kennel Club, "The Complete Dog Book", 19th ed., Howell Book House, New York, NY. 1997.

Asher, J.H.J. and Friedman, T.B. "Mouse and hamster mutants as models for Waardenburg syndromes in humans", *Journal of Medical Genetics* **27:** 618-626 (1990).

Barendse, W., Armitage, S.M., Kossarek, L.M., Shalom, A., Kirkpatrick, B.W., Ryan, A.M., Clayton, D., Li, L., Neibergs, H.L., Zhang, N., Grosse, W.M., Weiss, J., Creighton, P., McCarthy, F., Ron, M., Teale, A.J., Fries, R., McGraw, R.A., Moore, S.S., Georges, M., Soller, M., Womack, J.E., and Hetzel, D.J.S. " A genetic linkage map of the bovine genome", *Nature Genetics* **6:** 227-235 (1994).

Bassam, B.J., and Caetano-Anolles, G. "Automated "hot start" PCR using mineral oil and paraffin wax", *Biotechniques* **14:** 30-34 (1993).

Bentolila, S., Bach, J.M., Kessler, J.L., Bordelais, I., Cruaud, C., Weissenbach, J., and Panthier, J.J. "Analysis of major repetitive DNA sequences in the dog (*Canis familiaris*) genome", *Mammalian Genome* **10:** 699-705 (1999).

Bergenhem, N.C.H., Venta, P.J., Hopkins, P.J., and Tashian, R.E. "Variation in coding exons of two electrophoretic alleles at the pigtail macaque carbonic anhydrase I locus as determined by direct, double-stranded sequencing of polymerase chain reaction (PCR) products", *Biochemical Genetics* **30:** 279-287 (1992).

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. "Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms", *American Journal of Human Genetics* **32:** 314-331 (1980).

Boyer, G., Nonneman, D.J., Shibuya, H., Stoy, S.J., O'Brien, D., and Johnson, G.S. "A PCR-RFLP marker of the Erythroid Aminolevulinate Synthase Gene (ALAS2) on Canine Chromosome X", *Animal Genetics* **26:** 206-207 (1995).

Breen, M., Hitte, C., Lorentzen, T., Thomas, R., Cadieu, E., Sabacan, L., Scott, A., Evanno, G., Parker, H., Kirkness, E., Hudson, R., Guyon, R., Mahairas, G., Gelfenbeyn, B., Fraser, C., Andre, C., Galibert, F., and Ostrander, E. "An Integrated 4249 Marker FISH//RH Map of the Canine Genome", *BMC Genomics* **5:** 65 (2004).

Breslauer, K., Frank, R., Blocker, H., and Markey, L.A. "Predicting DNA duplex stability from the base sequence", *Proceedings of the National Academy of Science of the United States of America* **83:** 3746-3750 (1986).

Brinkmeyer-Langfore, C., Raudsepp, T., Lee, E-J., Goh, G., Schaffer, A., Agarwala, R., Wagner, M., Tozaki, T., Slow, L., Womack, J., Mickelson, J., and Chowdhary, B. "A High-Resolution Physical Map of the Equine Homologs of HSA19 Shows Divergent Evolution Compared to Other Mammals" *Mammalian Genome* **16:** 631-649 (2005).

Brookes, A. J. "The Essence of SNPs", *Gene* **234:** 177-186 (1999).

Brouillette, J.A. and Venta, P.J. "*Tth111*I PCR/RFLP marker in the canine rod transducin alpha (GNAT1) gene" *Animal Genetics* **31:** 68 (2000).

Brouillette, J.A., Andrew, J.R., and Venta, P.J. "Estimate of Nucleotide Diversity in Dogs using a Pool-and-Sequence Method", *Mammalian Genome* **11:** 1079-1086 (2000).

Buchanan, F.C., Adams, L.J., Littlejohn, R.P., Maddox, J.F., and Crawford, A.M. "Determination of evolutionary relationships among sheep breeds using microsatellites", *Genomics* **22:** 397-403 (1994).

Buetow, K.H., Edmonson, M.N., and Cassidy, A.B. "Reliable identification of large numbers of candidate SNPs from public EST data", *Nature Genetics* **21:** 323-325 (1999).

Burstein, E., Hogberg, J.E., Wilkinson, A.S., Rumble, J.M., Csomos, R.A., Komarck, C.M., Maine, G.N., Wilkinson, J.C., Mayo, M.W., and Duckett, C.S. "COMMD Proteins, A Novel Family of Structural and Functional Homologs of MURR1", *Journal of Biological Chemistry* **280 (23):** 22222-22232 (2005).

Cargill, M., Altculer, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemesh, J., Ziagura, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley G.Q., and Lander, E. S. "Characterization of single-nucleotide polymorphisms in coding regions of human genes", *Nature Genetics* **22:** 231-238 (1999).

Chee, M., Yang, R., Hubell, E., Berno, A., Huang, X.C., Stern, D., Winkler J., Lockhart, D.J., Morris, M.S., and Fodor, S.P.A. "Accessing Genetic Information with High-Density DNA Arrays", *Science* **274:** 610-614 (1996).

Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Buchanan, A., Stengard, J., Salomaa, V., Vartianien, E., Perola, M., Boerwinkle, E., and Sing, C.F. "Haplotype Structure and Population Genetic Inferences from Nucleotide-Sequence Variation in Human Lipoprotein Lipase", *American Journal of Human Genetics* **63:** 595-612 (1998).

Clark, L.A., Wahl, J.M., Rees, C.A., and Murphy, K.E. "Retrotransposon Insertion in SILV is Responsible for Merle Patterning of the Domestic Dog" *Proceedings of the National Academy of Sciences USA* **103(5):** 1376-1381 (2006).

Clark, L., Tsai, K., Steiner, J., Williams, D., Guerra, T., Ostrander, E., Galibert, F., and Murphy, K. "Chromosome-specific Microsatellite Multiplex Sets for Linkage Studies in the Domestic Dog", *Genomics* **85:** 550-554 (2004).

Clark, R.D. and Stainer, J.R. "Medical & Genetic Aspects of Purebred Dogs", Fairway, KS: Forum Publications, Inc (1994).

Collins, A., Lonjou, C., and Morton, N.E. "Genetic epidemiology of single-nucleotide polymorphisms", *Proceedings of the National Academy of Science USA* **96 (26):** 15173-15177 (1999).

Collins, D.W. and Jukes, T.H. "Rates of transition and transversion in coding sequences since the human-rodent divergence", *Genomics* **20:** 386-396 (1994).

Collins, F.S., Guyer, M.S., and Chakavarti, A. "Variations on a theme: cataloging human DNA sequence variation", *Science* **278:** 1580-1581 (1997).

Coltman, D.W. and Wright, J.M. "Can SINES: a family of tRNA-derived retroposons specific to the superfamily Canoidea", *Nucleic Acids Research* **22:** 2726-2730 (1994).

Day, A.A., McQuillan, C.I., Termine, J.D., and Young, M.R. "Molecular cloning and sequence analysis of the cDNA for small proteoglycan II of bovine bone" *Biochemical Journal* **248:** 801-805 (1987).

Deschenes, S.M., Puck, J.M., Dutras, A.S., Somberg, R.L., Felsburg, P.J., and Henthorn, P.S. "Comparative mapping of canine and human proximal Xq and genetic analysis of canine X-linked severe combined immunodeficiency", *Genomics* **23:** 62-68 (1994).

Dietrich, W., Katz, H., Lincoln, S.E., Shin, H.-P., Friedman, J., Dracopoli, N.C., and Lander, E.S. "A genetic map of the mouse suitable for typing intraspecific crosses", *Genetics* **131:** 423-447 (1992).

Dorn, C. and Schneider, R. "Inbreeding and canine mammary cancer: A retrospective study", *Journal of the National Cancer Institute* **57:**545-548 (1976).

Ellegren, H., Johansson, M., Sandberg, K., and Andersson, L. "Cloning of highly polymorphic microsatellites in the horse", *Animal Genetics* **23:** 133-142 (1992).

Evans, J.P., brinkhous, K.M., Brayer, G.D., Reisner, H.M., and High, K.A. "canine Hemophilia B Resulting from a Point Mutation with Unusual Consequences", *Proceedings of the National Academy of Sciences of the United States of America* **90(9):** 3968-3972 (1989).

Farber, C.R., Raney, N.E., Rilington, D.E., Venta, P.J., and Ernst, C.W. "Comparative Mapping of Genes Flanking the Human Chromosome 12 Evolutionary Break Points in the Pig", *Cytogenetics and Genome Research* **102(1-4):** 139-144 (2003).

Farnir, F., Coppieters, W., Arranz, J.J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D., and Georges, M. "Extensive genome-wide linkage disequilibrium in cattle", *Genome Research* **10:** 220-227 (2000).

Fisher, R.A., Putt, W., and Hackel, E. "An investigation of the product of 53 gene loci in three species of wild canidae: canis lupus, canis latrans, and canis familiaris", *Biochemical Genetics* **14:** 963-974 (1976).

Fitch, W.M. (1976) "The molecular evolution of cytochrome c in eukaryotes", *Journal of Molecular Evolution* **8:** 13-40 (1976).

Francisco, L.V., Langston, A.A., Mellersh, C.S., Neal, C.L., and Ostrander, E.A. "A Class of Highly Polymorphic Tetranucleotide Repeats for Canine Genetic Mapping", *Mammalian Genome* **7:** 359-362 (1996).

Fujita, M., Loechel, R., McFarlin, K., Brewer, G.J., Venta, P., and Yuzbasiyan-Gurkan, V. "Assignment and localization of a set of genes to canine chromosomes using fluorescent in situ hybridization", *Mammalian Genome* (in press).

Giger, U., Sargan, D., and McNiel, E. "Breed-Specific Hereditary Diseases and Genetic Screening" in *The Dog and Its Genome,* Ostrander, E., Giger, U., and Lindblad-Toh, K. Eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY, 2006.

Grimes, D.A., "Epidemiology of Gestational Trophoblastic Disease", *American Journal of Obstetrics and Gynecology* **150:** 309-318 (1984).

Gu, X. and Li, W.-H. "Higher rates of amino acid substitution in rodents than in humans", *Molecular Phylogeny and evolution* **1:** 211-214 (1993).

Halliday, R. and Grigg, G. W. " DNA methylation and mutation", *Mutation Research* **285:** 61-67 (1993).

Halushka MK, Fan J-B, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakavarti, A. "Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis", *Nature Genetics* **22:** 239-247 (1999).

Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox, M.J., Schneider, J.A., Moulin, D.S., and Clegg, J.B. "Archaic African and Asian lineages in the genetic ancestry of modern humans", *American Journal of Human Genetics* **60:** 772-789 (1997).

Harumi, T., Kimura, M., and Yasue, H. "Survey on swine SINEs (PRE 1) as candidates for SSCP markers in genetic linkage analysis" *Animal Genetics* **26:** 403-406 (1995).

Holmes, N.G., Mellersh, C.S., Humphreys, S.J., Holliman, A., Curtis, R., and Sampson, J. "Isolation and characterization of microsatellites from the canine genome", *Animal Genetics* **24:** 289-292 (1992).

Holsinger, K.E. "Wahlund Effect, Wright's F Statistics", http://statgen.dps.unipi.it/courses_file/documents/pdf/Holsinger-Wahlund.pdf, (2001).

Housley, D., Zalewski, Z., Beckett, S., and Venta, P. " Design Factors that Influence PCR Amplification Success of Cross-species Primers among 1147 Mammalian Primer Pairs" *BMC Genomics* **7:** 253 (2006).

Housley, D.J.E., Ritzert, E., and Venta, P.J. "Comparative Radiation Hybrid Map of Canine Chromosome 1 Incorporating SNP and Indel Polymorphisms" *Genomics* **84(2):** 248-264 (2004).

ISGN. "Guidelines for human gene nomenclature: an international system for human gene nomenclature (ISGN, 1987)" *Cytogenetics and Cell Genetics* **46:** 11-28 (1987).

Juneja, R., Christensen, K., Andersen, E., and Gahne, B. "Frequencies of transferring types in various breeds of domestic dog", *Animal Blood Groups and Biochemical Genetics* **12:** 79-88 (1981).

Kajii, T. and Ohama, K. "Androgenetic origin of the hydatiform mole", *Nature* **268:** 633-634 (1977).

Kathmann, I., Jaggy, A., Busato,A., Bartschi, M., and Gaillard, C. "Clinical and genetic investigations of idiopathic epilepsy in the Bernese Mountain Dog", *Journal of Small Animal Practice* **40:** 319-325 (1999).

Kirkness, E., Bafna, V., Halpern, A., Levy, S., Remington, K., Rusch, D., Delcher, A., Pop, M., Wang, W., and Fraser, C. "The Dog Genome: Survey Sequencing and Comparative Analysis", *Science* **301:** 1898-1903 (2003).

Kirov, G., Nikolov, I., georgieva, L., Moskvina, V., Owen, M., and O'Donovan, M. "Pooled DNA Genotyping on Affymetrix SNP Genotyping Arrays", *BMC Genomics* **7:** 27 (2006).

Koskinen, M. and Bredbacka, P. "Assessment of the population structure of five Finnish dog breeds with microsatellites", *Animal Genetics* **31:** 310-317 (2000).

Krawczak, M., Ball, E.V., and Cooper, D.N. "Neighboring-nucleotide effects on the rates of germ- line single-base-pair substitution in human genes", *American Journal of Human Genetics* **63:** 474-488 (1998).

Kruglyak, L. "Prospects for whole-genome linkage disequilibrium mapping of common disease genes", *Nature Genetics* **22:** 139-144 (1999).

Kruglyak, L. "The use of a genetic map of biallelic markers in linkage studies", *Nature Genetics* **17:** 21-24 (1997).

Kwok, P-Y. and Chen, X. "Detection of Single Nucleotide Variations" *Genetic Engineering* **20:**125-134 (1998).

Kwok, P-Y., Deng, Q., Zakeri, H., and Nickerson D.A. "Increasing the Information Content of STS-Based Genome Maps: Identifying Polymorphisms in Mapped STSs", *Genomics* **31:** 123-126 (1996).

Kwok, P-Y., Carlson, C., Yager, T., Ankener, W., and Nickerson, D.A. "Comparative Analysis of human DNA Variations by Fluorescence-based Sequencing of PCR Products", *Genomics* **23:** 138-144 (1994).

Lai, E., Riley, J., Purvis, I., and Roses, A. "A 4-MB High-Density Single Nucleotide Polymorphism-Based Map around Human APOE", *Genomics* **54:** 31-38 (1998).

Landegren, U., Nilsson, M., and Kwok, P-Y. "Reading Bits of Genetic Information: Methods for Single-Nucleotide Polymorphism Analysis", *Genome Research* **8:** 769-776 (1998).

Langston, A.A., Mellersh, C.S., neal, C.L., Ray, K., Acland, G.M., Gibbs, M., Aguirre, G.D., Fournier, R.E.K., and Ostrander, E.A. "Construction of a Panel of Canine-Rodent Hybrid Cell Lines for Use in partitioning of the Canine Genome", *Genomics* **46:** 317-325 (1997).

Li, W.-H. and Graur, D. "Fundamentals of Molecular Evolution", Sunderland, MA: Sinauer Associates, Inc (1987).

Lin, L., Faraco, J., Li, RKadotani, H., Rogers, W., Lin, X., Qiu, X, deJong, P.J., nishino, S., and mignot, E. "The Sleep Disorder Canine Narcolepsy is Caused by a Mutation in the Hypocretin (Orexin) Receptor 2", *Cell* **98(3):** 365-376 (1999).

Lingaas, F., Aarskaug, T., Gerlach, J.A., Juneja, R.K., Fredolm, M., Sampson, J, Suter, N., Holmes, N.G., Binns, M.M., Ryder, E.J., Van Haeringen, W.A., Venta, P.J., Brouillette, J.A., Yuzbasiyan-Gurkan, V., Wilton, A.N., Bredbacka, P., Koskinen, M., Dunner, S., Parra, D., Schmutz, S., Schelling, C., Schlapfer, J., and Dolf, G. "A Canine Linkage Map: 39 Linkage Groups", *Journal of Animal Breeding and Genetics* **118:** 3-19 (2001).

Lindblad-Toh, K., Wade, C., Mikkelsen, T., Karlsson, E., Jaffe, D., Kamal, M., Clamp, M., Chang, J., Kulbokas III, E., Zody, M., Mauceli, E., Xie, X., Breen, M., Wayne, R., Ostrander, E., Ponting, C., Galibert, F., Smith, D., deJong, P., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C-W., Cook, A., Cuff, J., Daly,M., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardelben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K-P., Parker, H., Pollinger, J., Searle, S., Sutter, N., Thomas, R., Webber, C., Broad Institute Genome Sequencing Platform, and Lander, E. "Genome Sequence, Comparative Analysis, and Haplotype Structure of the Domestic Dog", *Nature* **438 (8):** 803-819 (2005).

Lindblad-Toh, K. "Trait Mapping Using a Canine SNP Array", *Third International Conference: Advances in Canine and Feline Genomics* University of California Davis, August 2-5, 2006.

Lingaas, F., Sorensen, A., Juneja, R.K., Johansson, S., Fredholm, M., Wintero, A.K., Sampson, J., Mellersh, C., Curzon, A., Holmes, N.G., Binns, M.M., Dickens, H.F., Ryder, E.J., Gerlach, J., Baumle, E., and Dolf, G. "Towards Construction of a Canine Linkage Map: Establishment of 16 Linkage Groups", *Mammalian Genome* **8:** 218-221 (1997).

Lyons, L.A., Laughlin, T.F., Copeland, N.G., Jenkins, N.A., Womack, J.E., and O'Brien, S.J. "Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes", *Nature Genetics* **15:** 47-56 (1997).

Mandiger, P., Knol, B., Ubbink, G., and Gruys, E. "Hereditary Necrotising Myelopathy in Kooiker Dogs", *Research in Veterinary Science* **54:** 305-316 (1993).

Meera Khan, P., Brahe, C., and Wijnen, L.M.M. "Gene map of the dog: six conserved and three disrupted syntenies", *Cytogenetics and Cell Genetics* **37:** 537-538 (1984).

Meera Khan, P., Los, W.R.T., van den Does, J.A., and Epstein, R.B. "Isoenzyme markers in Dog Blood Cells", *Transplantation* **15:** 624-628 (1973).

Mellersh, C.S., Hitte, C., Richman, M., Vignaux, F., Priat, C., Jouquant, S., Werner, P., Andre, C., DeRose, S., Patterson, D.F., Ostrander, E.A., and Galibert, F. "An integrated Linkage-Radiation Hybrid Map of the Canine Genome", *Mammalian Genome* **11:** 120-130 (2000).

Mellersh, C.S., Langston, A.A., Acland, G.M., Fleming, M.A., Ray, K., Weigand, N.A., Francisco, L.V., Gibbs, M., Aguirre, G.D., and Ostrander, E.A. "A Linkage Map of the Canine Genome", *Genomics* **46:** 326-336 (1997).

Minnick, M.F., Stillwell, L.C., Heineman, J.M., and Stiegler, G.L. "A highly repetitive DNA sequence possibly unique to canids", *Gene* **110:** 235-238 (1992).

Morell, R., Friedman, T.B., Moeljopawiro, S., Hartono, Soewito, and Asher, J.H., Jr. "A frameshift mutation in the HuP2 paired domain of the probable human homolog of murine Pax-3 is responsible for Waardenburg syndrome type 1 in an Indonesian family", *Human Molecular Genetics* **1:** 243-247 (1992).

Morera, L., Barba, C.J., Garrido, J.J., Barbancho, M., and de Andres, D.F. "Genetic variation detected by microsatellites in five Spanish breeds", *Journal of Heredity* **90:** 654-656 (1999).

Murray, N.E., Brammer, W.J., and Murray, K. "Lambdoid phages that simplify the recovery of in vitro recombinants", *Molecular and General Genetics* **156:** 53-59 (1977).

Neff, M.W., Broman, K.W., Mellersh, C.S., Ray, K., Acland, G.M., Aguirre, G.D., Ziegle, J.S., Ostrander, E.A., and Rine, J. "A Second-Generation Genetic Linkage Map of the Domestic Dog, *Canis Familiaris*", *Genetics* **151:** 803-820 (1999).

Nei, M. and Li W.H. "Mathematical model for studying genetic variation in terms of restriction endonucleases", *Proceedings of the National Academy of Science of the United States of America* **76:** 5269-5273 (1979).

Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, R.G., Stengard, J., Salomaa, V., Vartianen, E., Boerwinkle, E., and Sing, C.F. "DNA Sequence diversity in a 9.7 KB Region of the Human Lipoprotein Lipase Gene", *Nature Genetics* **19:** 233-240 (1998).

O'Brien, S.J. "Molecular genetics in the domestic cat and its relatives", *Trends in Genetics* **2:** 137-142 (1986).

O'Brien, S.J., Womack, J.E., Lyons, L.A., Moore, K.J., Jenkins, N.A., and Copeland, N.G. "Anchored reference loci for comparative genome mapping in mammals", *Nature Genetics* **3:** 103-112 (1993).

Olsen, M., Hood, L., Cantor, C., and Botstein, D. "A common language for physical mapping of the human genome", *Science* **245:** 1434-1435 (1989).

Ostrander, E.A. and Kruglyak, L. "Unleashing the Canine Genome", *Genome Research* **10:** 1271-1274 (2000).

Ostrander, E.A., Galibert, F., and Patterson, D.F. "Canine Genetics Comes of Age", *Trends in Genetics* **16(3):** 117-124 (2000).

Ostrander, E.A., Mapa, F.A., Yee, M., and Rine, J. "One Hundred and One Simple Sequence Repeat-Based Markers for the Canine Genome", *Mammalian Genome* **6:** 192-195 (1995).

Ostrander, E.A., Sprage, Jr., G.F., and Rine, J. "Identification and Characterization of Dinucleotide Repeat (CA)n markers for for Genetic Mapping in Dog", *Genomics* **16:** 207-213 (1993).

Padgett, G.A. *Control of canine genetic Diseases,* Howell Book House, New York, NY, (1998), pg. 4.

Parker, H., and Ostrander, E. "Canine genomics and genetics: running with the pack", *PLOS Genetics* **1(5):** 508-513 (2005).

Patterson, D.F. "Companion Animal Medicine in the Age of medical Genetics", *journal of Veterinary Internal medicine* **14:** 1-9 (2000).

Patterson, D.F., Haskins, M.E., Jezyk, P.F., Giger, U., meyers-Wallen, V.N., Aguirre, G., Fyfe, J.C., and Wolfe, J.H. "Research on Genetic Diseases: Reciprocal Benefits to Animals and Man", *Journal of the American Veterinary Medical Association* **193(9):** 1131-1144 (1988).

Pearson, H. "Changing Attitudes to Congenital and Inherited Diseases", *Veterinary Record* **105:** 318-323 (1979).

Petersen-Jones, S.M., Entz, D.D., and Sargan, D.R. "cGMP Phosphodiesterase-alpha Mutation Causes Progressive Retinal Atrophy in the Cardigan Welsh Corgi Dog", *investigative Ophthalmology and Visual Science* **40:** 1637-1644 (1999).

Picoult-Newberg, L, Ideker, T.E., Poh., M.G., Taylor, S.L., Donaldson, M.A., nickerson, D.A., and Boyce-Jacino, M. "Mining SNPs From EST Databases", *Genome Research* **9:** 167-174 (1999).

Priat, C., Hitte, C., Vignaux, F., Renier, C., Jiang, Z., jouquand, S., Cheron, A., Andre, C., and Galibert, F. "A Whole-Genome Radiation Hybrid Map of the Dog Genome", *Genomics* **54:** 361-378 (1998).

Ray, K., Trepanier, L.A., Acland, G.M., and Aguirre, G.D. "PCR/RFLP Marker in the Canine Opsin Gene", *Animal Genetics* **27:** 293-294 (1996).

Richards, B., Skoletsky, J., and Shuber, A. "Multiplex PCR amplification from the CFTR gene using DNA prepared from buccal brushes/swabs", *Human Molecular Genetics* **2:** 159-163 (1992).

Richman, M., Mellersh, C.S., Andre, C., Galibert, F., and Ostrander, E.A. "Characterization of a Minimal Screening Set of 172 microsatellite markers for Genome-Wide Screens of the Canine Genome", *Journal of Biochemical and Biophysical Methods"*, **47:** 137-149 (2001).

Ridley, M "Wahlund Effect" in *Evolution,* 3rd Edition, . http://blackwellpublishing.com/ridley/a-z/Wahlund_effect.asp, November, 2009.

Rieder, M.J., Taylor, S.L., Clark, A.G., and Nickerson, D.A. "Sequence Variation in the Human Angiotensin Converting Enzyme", *Nature Genetics* **22:** 59-62 (1999).

Risch, N. and Merikangas, K. "The Future of Genetic Studies of Complex Human Diseases", *Science* **273:** 1516-1517 (1996).

Roberts, R.J. and Macelis, D. "REBASE-Restriction Enzymes and Methylases", *Nucleic Acid Research* **25:** 248-262 (1997).

Robinson, R. "Genetic anomalies in dogs", *Canine Practice* **16:** 29-34 (1991).

Roebroek, A.J.M., Schalken, J.A., Onnekink, C., Bloemers, H.P.J., and Van de Ven, W.J. "Structure of the feline c-fes/fps proto-oncogene: Genesis of a retroviral oncogene", *Journal of Virology* **61:** 2009-2016 (1987).

Rothuizen, J., Wolfswinkel, J., Lenstra, J.A., and Frants, R.R. "The incidence of mini- and micro-satellite repetitive DNA in the canine genome", *Theoretical and Applied Genetics* **89:** 403-406 (1994).

Sambrook, J., Fritsch, E.F. and Maniatis, T. "Molecular Cloning. A Laboratory Manual", (2nd ed.) Cold Springs Harbor: Cold Springs Harbor Laboratory Press (1989).

Sargan, D.R., Yang, F., Squire, M., Milne, B.S., O'Brien, P.C.M., and Ferguson-Smith, M.A. "Use of Flow-Sorted Canine Chromosomes in the Assignment of Canine Linkage, Radiation Hybrid, and Syntenic Groups to Chromosomes: Refinement and Verification of the Comparative Chromosome Map for Dog and Human", *Genomics* **69:** 182-195 (2000).

Sarmiento, U.M. and Storb, R.F. "RFLP Analysis of DLA Class I Genes in the Dog", *Tissue Antigens* **34(3):** 158-163 (1989).

Sarmiento, U.M. and Storb, R.F. "Characterization of Class II Alpha Genes and DLA-D Region Allelic Asociations in the Dog", *Tissue Antigens* **32(4):** 224-234 (1988).

Serikawa, T., Kuramoto, T., Hilbert, P., Mori, M., Yamada, J., Dubay, C.J., Lindpainter, K., Ganten, D., Guenet, J.-L., Lathrop, G.M., and Beckmann, J.S. "Rat gene mapping using PCR-analyzed microsatellites", *Genetics* **131:** 701-721 (1992).

Sharp, N.J.H., Kornegay, J.N., VanCamp, S.D., Herbstreith, M.H., Secore, S.L., Kettle, S., Hung, W.Y., Dykestra, M.J., et al. "An Error in Dystrophin mRNA Processing in Golden Retriever Muscular Dystrophy, an Animal Homologue of Duchenne Muscular Dystrophy", *Genomics* **13(1):** 115-121 (1992).

Shi, M. "Enabling large-scale pharmacogenomic studies by high-throughput mutation detection and genotyping technologies", *Clinical Chemistry* **47:** 164-172 (2001).

Shibuya, H., Collins, B.K., Collier, L.L., Huang, T.H-M., Nonneman, D., and Johnson, G.S. "A polymorphic (GAAA)n microsatellite in a canine Wilms tumor 1 (WT1) gene intron", *Animal Genetics* **27:** 59-60 (1996).

Shibuya, H., Collins, B.K., Stoy, S.J., nonneman, D.J., and Johnson, G.S. "PCR/RFLP Markers in the Canine Gamma-D-Crystallin Gene", *Animal Genetics* **26:** 445-446 (1995).

Shubitowski, D. M., Venta, P.J., Douglass, C.L., Zhou, R.X., and Ewart, S.L. "Polymorphism Identification Within 50 Equine Gene-specific Sequence Tagged Sites", *Animal Genetics* **32(2):** 78-88 (2001).

Simonsen, V. "Electrophoretic Studies on the Blood proteins of Domestic Dogs and other Canidae", *Hereditas* **82:** 7-18 (1976).

Southern, E.M. "Measurement of DNA length by gel electrophoresis", *Analytical Biochemistry* **100:** 319-323 (1979).

Suber, M.L., Pittler, S.J., Qin, N., Wright, G.C., Holcombe, V., Lee, R.H., Craft, C.M., Lolley, R.N., Baehr, W., and Hurwitz, R.L. "Irish Settr Dogs Affected with Rod/Cone Dysplasia Contain a nonsense Mutation in the Rod cGMP Phosphodiesterase Beta-subunit Gene", *Proceedings of the National Academy of Sciences of the United States of America* **90(9):** 3968-3972 (1993).

Sutter, N., Eberle, M., Parker, H., Pullar, B., Kirkness, E., Kruglyak, L., and Ostrander, E. "Extensive and BAreed-Specific Linkage Disequilibrium in Canis familiaris", *Genome Research* **14:** 2388-2396 (2004).

Taillon-Miller, P., Piernot, E.E., and Kwok, P-Y., "Efficient Approach to Unique Single-Nucleotide Polymorphism Discovery", *Genome Research* **9:** 499-505 (1999).

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., and Kwok, P-Y. "Overlapping Genomic Sequences: A Treasure Trove of Single-Nucleotide Polymorphisms", *Genome Research* **8:** 748-754 (1998).

Taillon-Miller, P., Bauer-Sardina, I., Zakeri, H., Hillier, L., Mutch, D.G., and Kwok, P-Y. "The Homozygous Complete Hydatiform Mole: A unique Resource for Genome Studies", *Genomics* **46:** 307-310 (1997).

Uchida, T., Komori, M., Kitada, K., and Kamataki, T. "Isolation of cDNAs coding for three different forms of liver microsomal cytochrome P-450 form polychlorinated biphenyl-treated Beagle dogs", *Molecular Pharmacology* **38:** 644-651 (1990).

Van de Sluis, B., Rothuizen, J., Pearson, P.L., van Oost, B., and Wijmenga, C. "Identification of a New Copper Metabolism Gene by Positional Cloning in a Purebred Dog Population", *Human Molecular Genetics* **11(2):** 165-173 (2002).

Venta, P.J. and Vidal, M. "MspI RSP in Exon 42 of the Canine von Willebrand Factor (VWF) Gene", *Animal Genetics* **30:** 229 (1999).

Venta, P.J., Yuzbasiyan-Gurkan, V., Brewer, G.J., and Schall, W.D. "Mutation Causing vonWillebrand's Disease in Scottish Terriers", *Jounal of Veterinary Internal Medicine* **14:** 10-19 (2000).

Venta, P.J., Cao, Y., Alexander, L., and Yuzbasiyan-Gurkan, V. "A polymorphic microsatellite in the canine retinoblastoma (*RB1*) gene", *Animal Genetics* **30:** 470-472 (1999).

Venta, P.J., Brouillette, J.A., Yuzbasiyan-Gurkan, V., and Brewer, G.J. "Gene-specific universal mammalian sequence-tagged sites: application to the canine genome", *Biochemical Genetics* **34:** 321-341 (1996).

Venta, P.J., Hewett-Emmett, D., and Tashian, R.E. "Simple method to convert DNA sequence variation into sites cut by restriction endonucleases: utility shown by typing the human CA3 and mouse strain Car-2 polymorphisms", *American Journal of Human Genetics* **49:** a445 (1991).

Venter, J.C., Adams, m.D., myers, E.W., Li, P.W., Mural, R.J., et al. "The Sequence of the Human Genome", *Science* **291:** 1304-1351 (2001).

Vignaux, F., Hitte, C., Priat, C., Chuat, J. C., Andre, C., and Galibert,F. "Construction and Optimization of a Dog Whole-genome Radiation Hybrid Panel", *Mammalian Genome* **10:** 888-894 (1999).

Wahlund, S. "Zusammensetzung von populationen und korrelation-sercheinungen von standpunkt der vererbungslehre aus betrachtet", *Hereditas* **11:** 65-106 (1928).

Wang, D.G., Fan, J-B., Siao, C-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nubaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M., and Lander, E.S. "Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome", *Science* **280:** 1077-1082 (1998).

Weber, J. "Informativeness of human (dC-dA)n.(dG-dT)$_n$ polymorphisms", *Genomics* **7:** 524-30 (1990).

Weiden, P., Storb, R., Kolb, H.J., Graham, T., Anderson, J., and Giblett, E. "Genetic Variation of the Red Blood Cell Enzymes in the Dog", *Transplantation* **17(1):** 115-120 (1974).

Weissbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., and Lathrop, M. "A second-generation linkage map of the human genome" *Nature* **359:** 794-801 (1992).

Werner, P., Mellersh, C.S., Raducha, M.G., DeRose, S., Acland, G.M., prociuk, U., Wiegand, N., Aguirre, G.D., Henthorn, P.S., Patterson, D.F., and Ostrander, E.A. "Anchoring of Canine Linkage Groups with Chromosome-Specific Markers", *Mammalian Genome* **10:** 814-823 (1999).

Whitney, K.M., Goodman, S.A., Bailey, E.M., and Lothrop, C.D., Jr. "The molecular basis of canine pyruvate kinase deficiency", *Experimental Hematology* **22:** 866-874 (1994).

Wilcox, B. and Walkowicz, C. "The Atlas of Dog Breeds of the World", 5th ed., T.F.H. Publications, Neptune City, NJ. 1995.

Willis, M.B. "Genetics of the Dog", New York, NY: Howell Book House (1989).

WinterØ, A.K., Fredholm, M., and Thomsen, P.D. "Variable $(dG\text{-}dT)_n$-$(dC\text{-}dA)_n$ sequences in the porcine genome", *Genomics* **12:** 281-288 (1991).

Wright, S. "The genetical structure of populations", *Annals of Eugenics* **15:** 323-321 (1951).

Wu, D.Y., Ugozzoli, L., Pal, B.K., and Wallace, B.R. "Allele-specific enzymatic amplification of beta-globin genomic DNA for diagnosis of sickle cell anemia", *Proceedings of the National Academy of Science of the United States of America* **86:** 2757-2760 (1989).

Yang, F., O'Brien, P.C.M., Milne, B.S., Graphodatsky, A.S., Solanky, N., Trifonov, V., Rens, W., Sargan, D., and Ferguson-Smith, M.A. "A Complete Comparative Chromosome Map for the Dog, Red Fox, and Human and its Integration with Canine Genetic Maps", *Genomics* **62:** 189-202 (1999).

Yuzbasiyan-Gurkan, V., Blanton, S.H., Cao, Y., Ferguson, P., Li, J., Venta, P.J., and brewer, G.J. "Linkage of a Microsatellite Marker to the Canine Copper Toxicosis Locus in Bedlington Terriers", *American Journal of Veterinary Research* **58(1):** 23-27 (1997).

Yuzbasiyan-Gurkan, V., Wagnitz, S., Blanton, S.H., and Brewer, G.J. "Linkage studies of the esterase D and retinoblastoma genes to canine copper toxicosis: a model for Wilson disease", *Genomics* **15:** 86-90 (1993).

Zajc, I., Mellersh, C.S., and Sampson, J. "Variability of canine microsatellites within and between different dog breeds", *Mammalian Genome* **8:** 182-185 (1997).

Zeitzer, J.M., Nishino, S., and Mignot, E. "The Neurobiology of Hypocretins (Orexins), Narcolepsy and Related Therapeutic Interventions", *Trends in Pharmacological Sciences* **27(7):** 368-374 (2006).