Final Technical Report, 2004-DN-BX-K004 M. W. Allard

Report title, award number, author(s)
2004-2008 National Institute of Justice, Forensic DNA, $243,000. PI M. W. Allard,
Building a Genetic reference database for dog mtDNA sequences and SNPs. 2004-DN-BX-K004

Abstract: The mitochondrial control (mtCR) region was analyzed for 552 unrelated domestic dogs. One hundred and four haplotypes were identified, 36 of which had not been previously reported. Additionally, 24 new polymorphisms were identified when compared to previously published datasets. The random match probability of the current dataset was found to be 0.041. It was determined that there is no genetic basis, based on the mtCR sequences, for grouping dogs by purebred or mixed or geographic location within the continental United States. Dogs of the same breed share similarities in their mtCR sequence, even if the sequences are not identical. This shows that it is necessary to collect multiple individuals from the same breed to build a thorough database of mtCR polymorphisms. Also purebred and mixed breed individuals can be combined into one database. The same is true for dogs from separate geographic locations.

While the mtCR has proven successful for human forensic evaluations by indicating ethnic origin, domestic dogs (*Canis lupus familiaris*) of seemingly unrelated breeds often form large haplotype groups (haplogroups?) based on identical control region sequences. In an attempt to break up these large haplogroups, we have sequenced the remaining ~15,484 base pairs of the canine mitochondrial genome for 64 individuals, adding 15 previously published dog mitochondrial genomes to those collected. Phylogenetic and population genetic methods were used to search for additional variability in the form of single nucleotide polymorphisms (SNPs). We have identified 356 SNPs and 65 haplotypes using the mitochondrial genome excluding the control region. The exclusion capacity was found to be 0.018. The mtCR was also evaluated for the same 79 dogs (both for the CR and mtgenome). The genetic signals from the different pieces of the mitochondrial genome do not conflict, but instead support and provide additional resolution for common mitochondrial control region haplogroups.

Table of Contents: **Page**

Executive Summary:

A 2005-2006 survey found that there were approximately 73 million domestic dogs (*Canis lupus familiaris*) in the United States (www.appma.org) or 1 dog for every 4 people in the country. As demonstrated by several cases, not only is dog hair collected as evidence when the dog is directly involved in a crime (Schneider, 1999 #90), dog hair and other types of canine evidence are frequently found at crime scenes as secondary transfer from the criminal or victims based on high interactions between humans and dogs ({Savolainen, 1999 #2}, State of California vs. David Westerfield, 2002 and State of Iowa vs Andrew Rich, 2002). Mitochondrial DNA (mtDNA) from human hair evidence has been used in the United States courts since the case of Tennessee vs Paul William Ware in 1996. The procedures for isolating, analyzing and presenting human mtDNA data that satisfy the admissibility requirements for scientific or technical evidence are in place and have been accepted by the legal and forensic communities (1). Thus, we plan to use similar methods for the dog DNA work.

Microscopic analyses of hair rarely tells more than species type, as hair can vary both between individuals of the same species as well as within an individual (2). There is little or no nuclear DNA in the hair shaft, often leaving mtDNA as the only source of DNA that can be recovered from the hair shafts of telogen hairs {Graham, 2007 #86; Takayanagi, 2003 #87; Roberts, 2007 #88}.

Mitochondria are organelles that play a role in the body's energy production and are found in numbers as high as 1000 per cell with as many as 10 genome copies per mitochondria (3,4). The high copy number of mitochondrial genomes per cell is useful for forensic analyses, particularly where the amounts of DNA are small or degraded. Typing based on nuclear DNA markers may be more problematic due to lower copy numbers {Budowle, 2003 #89}. Also, the mitochondrial genome is maternally inherited and does not undergo recombination. Different regions of the mammalian mtGenome accumulate mutations more readily than others. In humans, as well as other mammals, the control region (also known as D-loop or hypervariable region) has the highest mutation rate (6,7), making it a popular region of analysis to search for DNA variation. Relative to humans, dogs have an additional region, a 10bp tandem repeat that is repeated up to 30 times within the control region. The number of repeats is known to vary within an individual (8).

It is well know that other forensic studies have investigated the potential uses of canine mtDNA as evidence and that private databases of canine mtDNA variation exist {Angleby, 2005 #8; Gundry, 2007 #5; Himmelberger, 2008 #50; Savolainen, 1999 #2; Savolainen, 1997 #1; Wetton, 2003 #13}. We plan to use DNA sequencing and analysis to further categorize canine mtDNA haplotypes and develop the first public reference database of canine mtDNA single nucleotide polymorphisms (SNPs) from the control region of the canine mitochondrial genome.

Discussion

This project was intended to survey the largest known sample set of mtCRs isolated from domestic dogs across the United States. While sequencing 427 new mtCRs, we searched for new SNPs and haplotypes and added this data to 128 published samples. We evaluated the need to distinguish between purebred and mixed breed dogs and dogs from different geographic regions across the continental United States. We also looked at the necessity of sequencing multiple individuals of the same breed for a thorough database.

When collecting samples, discrepancies were found in breed definition. For example, some samples were received labeled by the donor as "Spitz". While there are Finnish Spitz's

and German Spitz's, Spitz is not a true breed designation but another name for an American Eskimo dog. It is unknown whether the donor meant one of the specific Spitz breeds or if the dog was in fact an American Eskimo dog. Also, two samples were received with the breed listed as "unknown", but one was described as purebred and one described as mixed. Descriptions could not be clarified or changed as this could be error prone without seeing the dog. As a result of each of the above mentioned problems, the number of distinct breeds collected for this study may be inflated. Population analyses were done to assess the severity of this issue including an AMOVA.

During sequencing, the tandem repeat was excluded due to the known possibility of variation within an individual (8). While excluding the tandem repeat region from control region studies has come to be common practice (10,11, 16-18), it appears that the studies conducted by our lab are the first to have problems obtaining the sequence for the region following the repeat (9). The sequencing problems seem to result from either individuals having a different number of repeats in the tandem repeat region (16130-16430 bp), individuals having a different number of C's and/or T's at the C/T stretch (16663 – 16676 bp), or a combination of both. This resulted in multiple sequence runs from the same individual being slightly different across these regions. Because there are multiple mitochondria per cell and multiple mitochondrial genomes per mitochondria, the differences between the genomes per mitochondria and per cell caused the DNA sequence reads to be shifted by one or a few bases due to the insertion or deletion of bases in problematic regions. This resulted in ambiguous bases being coded with the corresponding IUB code and the region between 16663 (nucleotide position) np and 16676 np being excluded when using a multiple alignment to search for informative SNPs.

The phylogenetic analysis showed that all dogs in our current dataset grouped within previously defined haplogroups A, B, C and D (Table 4). The proportions of samples within each group are very similar to the portions of unique haplotypes previously identified for each group. This is particularly interesting because the samples used in previous studies came from all over the world, while the samples in the current study are from the United States alone. It appears that regardless of local origin, more domestic dogs have an A haplotype than any of the other types described. Next is haplotype B followed by C and then D. Additional local studies are needed to confirm this observation. The lack of individuals from groups E and F is most likely due to the fact that the individuals in previous studies that formed groups E and F were collected from Asian and/or Siberian localities {Kim, 1998 #10; Okumura, 1996 #14; Savolainen, 2002 #7; Tsuda, 1997 #11}. Individuals with D, E, and F haplotypes have been found in much lower frequencies compared to individuals with types A, B, and C in world-wide samplings {Savolainen, 2002 #7}, which demonstrates that these haplotypes are more rare in the dog population.

Seventy-four sub-haplogroups were found in the current dataset with 63% of the dogs grouping in 1 of 8 sub-haplogroups containing between 10 and 70 individuals (Figure 2). The distribution of sub-haplogroup sizes shows that while there are many canine mitochondrial control region haplotypes, the majority of dogs share a few common haplotypes while the minority had unique or fairly unique haplotypes. These results demonstrate a recurring problem with canine mitochondrial control region sequence data: most dogs share identical types. This also indicates a need for the evaluation of the remainder of the canine mitochondrial genome to look for additional SNPs that may further break up these large haplogroups (Webb and Allard, submitted).

All of the variable sites identified in the current dataset are listed in Table 2 with the informative and highly informative sites shown in Table 4. Identification of informative SNPs is important when trying to recognize the most useful SNPs for assessing population variation. How informative a SNP is said to be is relative to the size and variation present in the dataset. Knowing where these informative SNPs occur in the mtCR allows for the potential development of SNP panels. Rather than sequencing the entire domestic dog mtCR, one could target the specific sites that distinguish between haplogroups, cutting down on resources and DNA necessary for the analysis. Our identification of 24 new SNPs, 6 of which were found to be informative and 3 highly informative, shows that previous studies have not resulted in a complete sampling of dog mtCRs, especially the region downstream of the repeat. All of the newly identified informative and highly informative SNPs were found in this less commonly sequenced region. While this contradicts the other findings of more informative SNPs upstream of the repeat region, the lack of sequencing and analysis of the region downstream of the repeat most likely explains this finding. As more sequences are added to the dataset, new sites may become phylogenetically informative due to the discovery of shared SNPs. Sites already identified as informative may gain a higher ranking due to their presence in more individuals. Also, the requirement of defining 1% of the total individuals in the dataset as criteria for the third ranking of SNP is subjective and changing this requirement may lead to changes in the ranking of SNPs.

As forensic samples are often subjected to conditions that may degrade DNA, the presence of the 60bp hotspot within the mtCR is particularly useful. While the number of unique haplotypes gleaned from only 60 bases is not going to be as large as those from the entire mtCR, this provides a region of high variability to target when the entire mtCR cannot be sequenced.

Conversely, specific SNPs such as those occurring at position 16439 bp seem to show higher levels of heteroplasmy relative to the remainder of the dataset, which is represented in our dataset as ambiguous base calls. As such, we recommend that future researchers pay close attention to base calls at these sites when editing their raw sequence data, and if possible, clone this region to further investigate these ambiguous sites.

The exclusion capacity and random match probabilities calculated for the dataset are slightly more powerful but similar to those previously reported (20, 21). The additional power comes from a larger sampling of dogs leading to more genetic variation in the dataset. This statistic varies depending on the dataset, and ideally, all existing and future control region sequences should be stored in the same database. As a result, a single statistic calculated for all control regions would be collected.

The nucleotide diversity and fixation index (Fst) both identify the lack of genetic structure within dogs when grouped as purebred and mixed. This shows that the decision as to how to classify certain breeds (i.e. Labradoodles) is trivial as purebred dogs and mixed breed dogs are not distinct populations based on mtCR sequence (Table 5). The AMOVA analysis also resulted in a low Fst value when dogs were grouped by state of sample origin and the distribution of dogs within each major haplogroup was consistent across the different geographic regions. This finding, along with the consistent distribution of haplogroups across states (Figure 3), supports previous studies that there is no need for local canine mtCR SNP databases within the continental United States (20). The significant Fst value when dogs are grouped by breed is most likely due to the strong amount of inbreeding that occurs in purebred dog lineages. While dogs of the same breed do not always share identical mtCR sequences, there is a higher within breed similarity than among the breeds as a whole. This demonstrates why multiple individuals

from a single breed and, more importantly, individuals from a variety of different breed types need to be collected to establish a thorough database of domestic dog mtCR SNPs.

Conclusions

As a result of combining 427 newly sequenced domestic dog mtCRs with a previous study of 125 domestic dog mtCRs (9), we have identified both new haplotypes and new informative SNPs. The results of the current study were consistent with previous studies. They found that domestic dogs were grouped into one of four previously identified major groups when using mtCR DNA. The dogs in this study were grouped within 37 of the previously defined 179 sub-haplogroups or formed 1 of the 36 new sub-haplogroups defined by a previously unrecorded mtDNA haplotype. The majority of the 552 dogs, 63%, were grouped into 1 of the 8 large sub-haplogroups with between 10 and 70 individuals per group. This indicates the need for the sequencing and analysis of the remainder of the domestic dog mtGenome (mtGenome) in hopes of identifying additional discriminatory SNPs to break up these large haplogroups and sub-haplogroups (Webb and Allard, submitted). Additionally, 94 SNPs were identified in the current dataset. Of the 94, 54 SNPs were informative, and 33 SNPS were highly informative with 24, 6 and 3 SNP sites, respectively, being previously unrecognized in the published literature. In general, population analyses show that domestic dogs are one large population. Smaller populations such as "purebred" and "mixed" or geographic populations cannot be distinguished based on mtCR sequences. However, when dogs are grouped by breed, they have less genetic variation than the population as a whole. These population analyses demonstrate the need to sample across a variety of breeds, including multiple individuals of the same breed, and that local mtCR SNP databases are not needed within the United States.

Second part of project: the mtDNA genome of canids

Hair, both human and animal, is often found as evidence in criminal investigations. Because hair is a composite of dead cells, the DNA contained in even fresh hair samples can be degraded (1). Each cell contains only two copies of the nuclear genome, but a second genome is also present in much higher copy numbers, the mtGenome. Mitochondria are organelles responsible for many metabolic tasks within and between cells. When mtDNA is sequenced, focus tends to be on a region of the genome known as the mtCR (also knows as the D-loop or hypervariable region) (5-11)Webb and Allard, 08-027). In canines, the mtCR is approximately 1,272 base pairs (bps) in size, is non-coding, and thus accumulates substitutions faster than any other comparably sized region of the mtGenome (12). This high rate of substitution is useful when looking for variability to help identify samples. In human investigations, the mtCR can indicate the ethnicity of a person (6). Knowing how valuable human mtDNA can be, attempts have been made to analyze mtDNA from the domestic dog (*Canis lupus familiaris*) for instances when dog hair is found as evidence at a crime scene (5, 7, 8, 11, 13-15) Webb and Allard, 08-027). A 2005-2006 survey found that there were approximately 73 million domestic dogs in the United States (www.appma.org). Because of this, it is not unexpected that dog hair is often found in criminal investigations either when a dog is directly involved in a crime or as secondary transfer from either the victim or suspect. It has been shown that while highly variable, the control region does not distinguish between dog breeds or any of the main groupings of dogs. In a previous study, we found that out of 552 domestic dogs, there were groups containing as many as 59 dogs of varying breeds with identical control region sequences (Webb and Allard, 08-027). In fact, the random match probability of the mtCR for the domestic dog was found to be 4.3% as

compared to between 2.5% and 0.52% for the human mtCR ((4), Webb and Allard, 08-027). Knowing that the domestic dog mtCR does not have the discriminatory power of the human mtCR, and also knowing that there are an additional ~15,458 bp of mtGenome outside of the control region, we have sequenced the remainder of the genome for 64 domestic dogs from our mtCR study. We combined our sequences with 15 complete mtGenome sequences downloaded from Genbank (16, 17). We use phylogenetic and population genetic methods to analyze the 79 genomes and report relationships and variable sites in the remainder of the genome that will aid in further discriminating between dogs with common mtCR sequences.

Discussion

The aim of this study was to sequence multiple mtGenomes of domestic dog to search for informative SNPs that would break up the large haplotype groups formed by using the mtCR sequence alone and to assess the utility of the mtGenome for forensic analyses. Individuals were chosen for mtGenome sequencing because either they belonged to one of the large mtCR haplotype groups or the breed was of interest. The 64 newly sequenced mtGenomes combined with the 15 mtGenomes downloaded from Genbank form the largest domestic dog mtGenome dataset to be published to date and the first to be used to identify domestic dog mtGenome haplotypes.

During sample collection, donors were asked to determine breed and breed type (either purebred or mixed). As the authors never saw the actual dog, breed and type were never changed, even when the declarations were questionable. For example, 2 samples were received with one being labeled “West Highland White Terrier” and the other “West Highland Terrier.” While these two dogs could very well be of the same breed, they were distinguished as different breeds in the current dataset based on the differing donor descriptions. Individuals with unknown breed type were considered mixed unless otherwise listed by the donor.

The presence of the 2 bp insertion in sequences from the Bjornerfeld et al. (16) study and the fact that our sequencing strategy included designing PCR primers based on amplicon size and not flanking a particular gene or region allows us to conclude that this sequence is not from pseudogene. Upon presenting the results of this study at the 2007 Society of Molecular Biology and Evolution meeting, it was suggested that this might be an error in the DNA sequence of the domestic dog that is corrected by the translational machinery upon translation from DNA to amino acid.

When comparing the mtGenome excluding the mtCR to the mtCR, we first notice that while the mtGenome has more haplotypes, the mtCR has a higher overall percentage of SNPs. Also, the percentage of SNPs unique to an individual is about the same for the two datasets. While it may seem counter-intuitive that such a comparatively small region would have a higher percentage of SNPs, it must be remembered that the mtCR is non-coding, meaning it does not translate into an RNA or protein and therefore lacks strong biological constraints to prevent nucleotides from mutating. The majority of the mtGenome excluding the mtCR codes for RNA or proteins with important biological functions, making the probability of a SNP occurring in one of those regions much lower (12). When SNPs do occur in a coding region, it is more likely that they are unique or possessed by only a small number of individuals, leading to more haplotypes with unique SNPs or unique combinations of SNPs within the mtGenome. This is seen in our dataset. Collectively, our results show that while there is more variability in the mtCR, the percentage of unique SNPs is relatively constant throughout the genome. Incorporation of SNPs

outside of the mtCR increases the number of informative SNPs for forensic use to 57% of the total SNPs found.

Collectively, the 79 dogs in our dataset formed 10 groups and 47 unique haplotypes with 8 ambiguous sequences. The ambiguous base calls were due to true polymorphisms within the individual dog samples due to the multiple genomes per cell (2, 3). While the number of individuals with unique haplotypes may seem high, it is important to keep in mind that this is the first study of its kind, and the number will likely decrease as more dog mtGenomes are evaluated. Relative to the mtCR, this number will likely always to be higher due to larger region and higher constraints against mutation on the coding portions of the mtGenome.

As mentioned above, the number of individuals that share identical mtGenome sequences is smaller than the number of individuals that share mtCRs for the same dogs (Figures 2 and 3). This illustrates how the additional sequence variation of the mtGenome can be used to break up the large groups that often result from mtCR sequencing. Figure 2 shows how the dogs are situated relative to their haplotype. Figure 3 demonstrates the phenomenon that was seen in our larger mtCR study. While there are many canine mitochondrial control region haplotypes, most dogs share the common types while the minority of dogs have unique or fairly unique types. The distribution of the dogs within the mtGenome haplotype groups shows that the additional variation found in the remainder mtGenome breaks-up the large groups formed by mtCR sequences alone.

The distributions of dogs within each haplogroup were consistent with the mtCR groupings. As previously reported, when using only the mtCR sequence group A contained the most individuals followed by groups B, C and D (Webb and Allard, 08-027). When evaluating the mtGenome groups in the same manner, the same trend persists. Group A had the most individuals followed by B, C and D. When viewing the relationships of the dogs in the trees shown in Figure 1, it can be seen that not only do the sizes of the groups correspond between datasets, but also the members of each group. Dogs that grouped together based upon their mtCR also grouped together based upon their mtGenome excluding the mtCR sequences. This result indicates that the phylogenetic signal present in the mtCR is also present in the remainder of the mtGenome. This result is expected as the mitochondrial genome does not undergo recombination and as such acts as a single locus. This is promising for forensic use of canine mitochondrial DNA. It shows that the entire mitochondrial genome can be used to identify samples because the results from different regions of the genome do not conflict.

The importance of the mutational “hot spots” within the mtGenome is that forensic samples are often degraded, making it difficult to obtain complete sequence through large areas. Also, the mtGenome is 92% larger than the mtCR. As a result, it is much more expensive to sequence. By identifying the most variable regions, we have provided coordinates where future groups can focus sequencing efforts; conversely, the regions where no SNPs were found could be avoided. These SNP free sections are all coding regions of the mtGenome; therefore, it is not surprising that the nucleotide composition of this region is conserved among the dogs in our dataset. All regions of increased or decreased SNP occurrence were identified via haplotype pairwise alignment to the Kim et al (17) reference sequence.

The random match probability results show that when considering the remainder of the mtGenome, there is a lower chance of a random match compared to using the mtCR alone. This is significant since it provides extra confidence that a match between a suspect dog and the sample found at a crime scene are truly the same individual and not just the result of the two randomly sharing mtGenome haplotype.

These results of the pairwise difference and nucleotide diversity assessments are consistent with the findings of the mtCR study. Though not statistically significant, they indicate that mixed breed dogs come from a more variable gene pool and, as expected, have more diversity in their sequence than purebred dogs. The ancestral lines of purebreds should contain only the DNA of individuals from the same breed or the founding breeds resulting in more constrained physical as well as genetic characteristics.

Since we never actually saw the dogs from which our samples were obtained, we were able to test the significance of the purebred versus mixed labels. Our results agree with the nucleotide diversity results which showed that there is not significant genetic variation between the group of dogs labeled "purebred" and those dogs labeled "mixed." This illustrates that not knowing whether a dog is purebred or mixed has very little consequence on the dataset in terms of mtDNA. Additionally, we show that geographic location of sample collection is not relevant when evaluating dogs from the United States via mtGenome haplotypes. Conversely, the fixation index becomes larger when dogs are grouped based on breed, which demonstrates that dogs of the same breed, while perhaps not possessing identical mtGenome sequences, have similar sequence composition than expected at random. These results support our previous mtCR dataset findings, which allows us to draw the same conclusions. First, classifying breeds by breed type (purebred or mixed) is trivial when it comes to mtDNA. Second, there is no need for local canine mitochondrial SNP databases. Finally, there is population substructure when dogs are grouped by breed. This is most likely due to the higher amounts of inbreeding of purebred dogs, which exemplifies that the need to collect multiple individuals of the same breed is necessary for a thorough mitochondrial SNP database.

Conclusions

Consistent with the mtCR results, analysis of the SNPs in the remainder of the mtGenome does not group dogs by breed or any other common domestic dog grouping. However, the SNPs found in the remainder of the mtGenome are useful since they provide additional discriminatory sites that break up common mtCR haplotype groups. Within our dataset of 79 domestic dog mtGenomes excluding the mtCR, 2.3% of the nucleotides were found to be variable. Fifty-seven percent of the variable sites were informative by supporting groups of two or more dogs, and 26% of the informative sites were highly informative by supporting groups of eight or more dogs. When comparing haplotype groups formed from the mtCR sequences alone and the mtGenome sequences without the mtCR for the same set of 79 dogs, it becomes obvious that the SNPs found in the remainder of the mtGenome have a higher discriminatory power. When looking at the mtCR alone, there are 18 individuals (25.7%) with unique mtCR sequences and 52 dogs (74.3%) forming 14 groups with up to 7 dogs per group. Comparatively, when looking at the same 79 dogs using mtGenome sequences without the mtCR, the distribution shifts with 24 dogs (33.8%) forming 10 groups containing at most 3 dogs and the remaining 67.6% (n = 48) of the dogs having unique haplotypes. Using AMOVA, the current dataset shows that there is little need to be concerned with whether a dog is classified as purebred or mixed or knowing the geographic location within the United States from which a sample was obtained. We do see evidence that it is necessary to collect multiple individuals of the same breed for a comprehensive mitochondrial SNP database. This is the first study to report SNP variation outside of the mtCR for the domestic dog. Our data demonstrate the usefulness of the entire mtGenome for forensic use in identifying domestic dog samples.

Main Body of the Final Technical Report:

Identification of Forensically Informative SNPs in the Domestic Dog Mitochondrial Control Region*

Kristen M. Webb[1], B.S.; Marc W. Allard [1], Ph.D.

[1]Department of Biological Sciences, George Washington University, Washington, DC 20052.

* This work was supported by the National Institute of Justice through grant 2004-DN-BX-K004to M. W. Allard. This work has been presented at The NIJ Conference 2007 and at the GW Research and Discovery Day, 2007. Both instances were in poster form.

Ia. Introduction:

A 2005-2006 survey found that there were approximately 73 million domestic dogs (*Canis lupus familiaris*) in the United States (www.appma.org) or 1 dog for every 4 people in the country. As demonstrated by several cases, not only is dog hair collected as evidence when the dog is directly involved in a crime (Schneider, 1999 #90), dog hair and other types of canine evidence are frequently found at crime scenes as secondary transfer from the criminal or victims based on high interactions between humans and dogs ({Savolainen, 1999 #2}, State of California vs. David Westerfield, 2002 and State of Iowa vs Andrew Rich, 2002). Mitochondrial DNA (mtDNA) from human hair evidence has been used in the United States courts since the case of Tennessee vs Paul William Ware in 1996. The procedures for isolating, analyzing and presenting human mtDNA data that satisfy the admissibility requirements for scientific or technical evidence are in place and have been accepted by the legal and forensic communities (1). Thus, we plan to use similar methods for the dog DNA work.

Microscopic analyses of hair rarely tells more than species type, as hair can vary both between individuals of the same species as well as within an individual (2). There is little or no nuclear DNA in the hair shaft, often leaving mtDNA as the only source of DNA that can be recovered from the hair shafts of telogen hairs {Graham, 2007 #86; Takayanagi, 2003 #87; Roberts, 2007 #88}.

Mitochondria are organelles that play a role in the body's energy production and are found in numbers as high as 1000 per cell with as many as 10 genome copies per mitochondria (3,4). The high copy number of mitochondrial genomes per cell is useful for forensic analyses, particularly where the amounts of DNA are small or degraded. Typing based on nuclear DNA markers may be more problematic due to lower copy numbers {Budowle, 2003 #89}. Also, the mitochondrial genome is maternally inherited and does not undergo recombination. Different regions of the mammalian mtGenome accumulate mutations more readily than others. In humans, as well as other mammals, the control region (also known as D-loop or hypervariable region) has the highest mutation rate (6,7), making it a popular region of analysis to search for DNA variation. Relative to humans, dogs have an additional region, a 10bp tandem repeat that is repeated up to 30 times within the control region. The number of repeats is known to vary within an individual (8).

It is well know that other forensic studies have investigated the potential uses of canine mtDNA as evidence and that private databases of canine mtDNA variation exist {Angleby, 2005 #8; Gundry, 2007 #5; Himmelberger, 2008 #50; Savolainen, 1999 #2; Savolainen, 1997 #1;

Wetton, 2003 #13}. We plan to use DNA sequencing and analysis to further categorize canine mtDNA haplotypes and develop the first public reference database of canine mtDNA single nucleotide polymorphisms (SNPs) from the control region of the canine mitochondrial genome.

IIa Methods

Domestic dog blood, tissue, and buccal swab samples were collected as donations from veterinary practices and private donors across the United States. Blood and tissue samples were not collected by the practices solely for this study. The collected samples were those that otherwise would have been disposed of. The donor of the sample made the determination of breed type and whether a sample was purebred or mixed. The donor was also asked to indicate any known relationship of a particular sample to other samples donated to this study. As this study focused only on mitochondrial DNA and is inherited maternally, siblings would have identical mitochondrial DNA sequences. Unrecognized familial relationships could lead to a misinterpretation of individuals of the same breed being thought to have the same mitochondrial DNA and affect estimates of nucleotide diversity. A subset of the blood and tissue samples collected was used for sequencing and analysis.

All blood and tissue samples were stored at –20°C until needed. Approximately 1 gram of tissue was isolated and placed in a culture tube with 0.1X TAE (Tris-Acetate-EDTA (Ethylenediamine Tetraacetic Acid)) for preservation. Each tissue was ground into one cell slurry using a Janke and Kunkel Ultra Turrax T25 tissue grinder (Janke and Kunkel, Staufen, Germany). Total genomic DNA was extracted from the blood and tissue samples using the Invitrogen DNA Easy kit following the protocols for "Small Blood Samples and Hair Follicles" or "Small Amounts of Cells, Tissues or Plant Leaves" (Invitrogen Corporation, Carlsbad, CA). Following extraction, DNA samples were stored in 0.1X TE (Tris-EDTA). DNA was quantified using the Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE).

The oligonucleotide primers used in this study were taken from the previous study of Dog mtDNA CR genetic variation (9). PCR primers were redesigned relative to the previous study because a number of samples in the previous study yielded double-banded products. The new primers flanked the entire mtCR and sat outside of the mtCR than the original primers. The new primers were defined as R51 (5'-TATGTTTATGGAGTCGTGCGA-3') and F15406 (5'-TTTGCTCCACCATCAGCACC-3' Figure 1). The previously designed sequencing primers from Gundry et al. (9) were used for DNA sequencing in this study. The use of both sets of mtCR primers resulted in bidirectional, overlapping, high quality, 4 – 6x sequence coverage across the mtCR but excluding the tandem repeat region (Figure 1). This repeat region is found in both dogs and wolves and is known to vary within and among individuals; thus, it was not sequenced for the current study (8).

All primers were received lyophilized from Operon Biotechnologies, Inc (Huntsville, AL) and were resuspended to a concentration of 160mM in 0.1X TE. The mtCR was amplified with one primer pair designed to span the entire region. PCR amplifications were performed in 50µl reactions using 100ng total DNA, 1x Buffer (Fisher BioReagents, Fisher Scientific, Pittsburgh, PA), 5mM $MgCl_2$ (Fisher BioReagents, Fisher Scientific, Pittsburgh, PA), 0.4mM dNTP mix (Invitrogen Corporation, Carlsbad, CA), 0.1µM of each primer and 2.5units of Taq polymerase (Fisher BioReagents, Fisher Scientific, Pittsburgh, PA). The PCR amplification profile on the thermal cycler (MJ Research, DNA Engine) comprised of an initial denaturing step of 96°C for 10 minutes. The next step was denaturing at 94°C for 15 seconds, annealing at 56°C for 30 seconds, extension at 72°C for 1 minute for 39 amplification cycles and, lastly, a final

extension at 72°C for 7 minutes. PCR products were run on a 1% agarose gel at 70 volts for 1 hour. A 1 kilobase ladder was used to determine size of the amplified product, and a low-mass ladder was included in order to determine concentration of each product. Samples were diluted to 10 µl reactions with a concentration of 40-60ng/µl of DNA and cleaned using 2µl of ExoSAP-IT (USB Corporation, Cleveland, OH) according to the procedure recommended by the vendor. The ExoSAP-IT procedure includes a 37°C incubation step for 15 minutes followed by an inactivation step at 80°C for 15 minutes. Samples were then shipped on dry ice overnight to SeqWright DNA Technology Services in Houston, Texas. SeqWright (www.seqwright.com) completed all DNA sequencing according to their protocols using ABI technology.

Representative sequences of the haplotypes previously described (10,11) were downloaded from Genbank (Accession #'s AF531654-AF531741 and AY656703-AY656710). Additionally 125 domestic dog sequences collected from the previous study (9) were also included in the current dataset (Genbank Accession #'s AY240030-AY240072, AY240074-AY240093, AY240095-AY240154 and AY240156-AY240157).

The forensic version of Sequencher 4.1.4FB19 (Gene Codes Corporation, Ann Arbor, MI) was used to edit and align all mtCR sequences. This version of the software builds alignments according to the previously defined criteria for gap placement and priority for preference of sequence differences in forensic evaluations (12). All alignments were confirmed by eye. Standard IUB codes were used for polymorphic sites and N's were inserted for positions in which the base could not be determined. As with human forensic studies, a reference control region sequence was used as a comparison. Utilizing a reference sequence allows base coordinates to be compared across different studies (13), thus all coordinates mentioned in this research are in terms of the reference sequence. This has been previously recommended in an effort to standardize canine mitochondrial nucleotide nomenclature (13). Per Peirera's recommendations that the reference control sequence should be the first canine mitochondrial genome to be published by Kim et al. (5).

The region spanning from 16663 bp – 16676 bp was removed from the multiple alignment due to sequencing and alignment issues stemming from a polymorphic C/T stretch. These bases were considered when defining unique haplotypes. The tandem repeat, which comprises of a varying number of 10 bp fragments and located at 16130 np – 16430 np, was not sequenced due to variation within an individual (8). To account for the missing region, all mtCR sequences were divided relative to whether they are from the region 5' of the repeat (np 15458 – 16129) or 3' of the repeat (np 16430 – 16727) with respect to the published right strand of the mtCR (Figure 1). Two multiple alignments of all downloaded and newly sequenced mtCRs were created, one for each region on either side of the repeat.

In Winclada (14), the "new matrix merge" command was used to combine the separate alignments based on matching identical taxon names. Arlequin was then used to search within the dataset for groups of dogs with identical control region sequences or haplotypes and to calculate the frequency of these haplotypes. The sequences that were identical to at least one other sequence in the dataset were also identified in Winclada and removed by using the "mark identical taxa" and "delete selected terms" commands. The entire mtCR sequence, excluding the tandem repeat, of samples representing unique haplotypes were aligned to the reference sequence using Sequencher. The coordinates and base calls of the SNPs were recorded in an Excel spreadsheet. Because the 13 bases between 16663 and 16676 were not used in the multiple alignment, manual checks were done to ensure the uniqueness of all haplotypes. As the majority of the previously published haplotype definitions lacked sequence in the region after the tandem

repeat (on the 3'side), our new sequences allowed for more descriptive haplotype definitions. If the region downstream from the repeat was not sequenced in an earlier study, the region was represented by the question mark symbol (?) to signify missing data. New sequences are grouped with previously defined haplotypes based on the SNPs present in the 5' region, which was previously sequenced. If two or more individuals are identical to a previously defined haplotype in the 5' region and new SNPs are found in the 3' region not previously sequenced, then the additional SNPs become the definition of a subtype within the previously defined haplogroup. A completely new haplotype is defined as at least one individual possessing a unique set of SNPs relative to the reference sequence and do not match the 5' region or complete mtCR sequence of any of the previously published sequences.

In order to determine the relationship relative to previously defined haplogroups of the new mtCR sequence haplotypes, the matrix was transposed from DNA to numeric characters (A=0, C=1, G=2, T=3). The number "4" was inserted manually to replace any missing data that was truly a gap between the query sequence and the Kim reference sequence. This way gaps would be considered as potential informative sites by Winclada as opposed to "missing data" or regions without base calls due to unobtainable sequence. Winclada was then used to assess the relationships of the different dogs by constructing a phylogenetic tree using a parsimony ratchet search method on the entire dataset. Recommended search strategies for using the parsimony ratchet for large data matrices were followed (15). If multiple equally likely trees were obtained, they were combined to make a consensus tree and the placements of individuals with new mtCR haplotypes were assessed relative to the previously published haplotypes.

Winclada was also used to identify informative SNPs defined as a nucleotide that supports a group of two or more individuals. This was done by using the "mop informative characters/delete selected characters" function and then using the character diagnoser to trace each character, or informative SNP, on the tree. The length and retention index (ri) statistics were recorded for each informative SNP. The length is the number of times the nucleotide state at a given position changes on the tree. The ri is a measure of a nucleotide position having the same base in two individuals being the result of shared common ancestry and not convergence. The ri score can range from 100 to 0. A score of 100 denote that the character change arose only once and defines all members of a group. The scores get progressively lower until a score of 0 is reached. This indicats that all character changes arose independently.

SNPs were classified into three ranks: the first rank was simply the presence of a SNP at a nucleotide position, the second rank was assigned to characters found to be phylogenetically informative by Winclada based on character length and ri, and the third level of ranking contains informative SNPs that define groups of six or more individuals, or 1% of the total dogs in the dataset.

All population statistics were either calculated in Arlequin or by hand. The dataset was analyzed as a whole with each individual defined as a unique haplotype (ignoring identical taxa). The dataset was also analyzed by separating dogs by their purebred or mixed description. This analysis was used for suspected evidence of inbreeding in purebred individuals and for evaluating if the "purebred" and "mixed" characterizations actually represent two unique populations. The samples were also separated by large regional groupings to look for local substructure and by those breed groups with a high number of purebred individuals (n>6) to look for within breed structure. The mean number of pairwise differences, nucleotide diversity, and assessment of variation within and between each grouping were calculated through Arlequin. Additional statistics such as exclusion capacity ($1 - \Sigma X_i^2$) and random match probability (ΣX_i^2),

where X_i is the frequency of the ^{ith} haplotype, were calculated by hand following the grouping of individuals with identical sequences into haplotypes.

IIIa Results

Six hundred and ninety-eight domestic dog blood, tissue and buccal swab samples were collected from various veterinary practices and private donors across the United States. Of the 698 samples collected, 427 blood and tissue samples were used for sequencing and analysis and are available on Genbank (Accession #'s EU223385 – EU223811). The distribution of these samples across the United States was as follows: California = 189, Maryland = 1, Mississippi = 8, New York = 1, Pennsylvania = 100, Nevada = 52, Texas = 14, Vermont = 1, Virginia = 61. Three hundred and ten of these samples came from purebred individuals and the remaining 116 were mixed breed. Samples with of unknown breed type were considered mixed. The 426 newly collected samples were combined with the standard reference sequence {Kim, 1998 #10} and 125 purebred dogs from a previous study (9) for a final dataset of 552 domestic dogs. A complete list of the different breeds and the number of each breed included in this study can be found in Table 1.

The complete mtCR excluding the tandem repeat was sequenced for 417 of the 427 newly collected individuals. The 10 individuals that did not have complete sequence were missing bases immediately after the repeat. The heteroplasmy of the repeat region caused the resultant sequence after this area to be unreadable due to the varying number of repeat units within the same dog. The missing bases in these sequences were coded as missing data and were not considered when looking for unique SNPs or haplotypes.

Previously defined haplotypes (n=123) were downloaded from Genbank as we planned to continue using the established nomenclature (10,11). Haplotype A15 could not be downloaded from Genbank as it was not found with the other sequences from the publication. Haplotypes labeled A37, A74, A75, A76, A77, A78, and A79 do not appear to exist in previously published datasets.

The sizes of the newly sequenced complete mtCR ranged from 965 bp to 975 bp, excluding the tandem repeat. The final dataset of the newly sequenced mtCR and those from the three previous studies (9-11) consisted of 733 taxa, including the reference sequence. Following the separate alignments of each unique haplotype to the reference sequence, the size of the total matrix was 985 characters. Sixteen gaps were inserted into the alignment when haplotypes were aligned to the Kim et al. (5) reference sequence: 15464.1, 15539.1, 15546.1, 16129.1, 16507.1, 16542.1, 16562.1, 16663.1, 16663.2, 16671.1, 16671.2, 16671.3, 16673.1, 16674.1, 16711.1, 16711.2.

The search for individuals with identical mtCR sequences resulted in 311 unique haplotypes from the starting dataset of 731 domestic dog control region sequences. Tree searches of the unique sequences only resulted in 508 equally parsimonious trees. This means that there were 508 equally likely resolutions of the relationships of the 311 dogs using the control region data and the parsimony ratchet method of grouping. A single consensus tree was made from all resultant trees as a way to summarize non-conflicting groupings. These groupings, as well as the spreadsheet of all individuals and the variable SNPs they possessed, were used to identify haplotypes in the current dataset. Excluding those sequences/haplotypes from previously published studies (10,11) resulted in the identification of 104 unique haplotypes in our dataset of 552 sequences. These 104 haplotypes do not include those individuals that did not match any other sequence due to the presence of ambiguous base calls.

Canine mitochondrial control region sequences had previously been grouped into six main types, A, B, C, D, E and F (10,11). Analysis of the parsimony ratchet trees showed that all of the newly sequenced control regions fell within four of these main types, namely A, B, C and D.

Hereafter the number and percentage of individuals reported for a haplogroup is based only on the 552 dog dataset and does not include previously published individuals. The majority of the newly sequenced samples either fell within one of the previously defined haplogroups or a sub-haplogroup of a previously defined haplogroup. Additionally, 36 newly defined haplotypes were identified and 60 sequences had ambiguous base calls and could only be classified in terms of major haplogroup. A complete list of all haplotypes found in this study and a list of dogs belonging to each haplogroup are given in Tables 2 and 3.

Haplogroup A was the largest haplogroup in the previously published data and also the haplogroup in which more of the newly sequenced samples clustered relative to the other groups. Previously published studies reported 76 haplotypes within group A, which make up 61.8% of all previously published haplotypes (10,11). Three hundred and seventy of the 552 individuals, or 67%, from the current study fall within haplogroup A. Most of these individuals possessed one of the 24 previously identified haplotypes, namely A1, A2, A5, A11, A16, A17, A18, A19, A20, A22, A24, A26, A27, A28, A29, A31, A33, A40, A66, A68, A70, A71, A80 and A82. Most of these haplotypes were further divided into sub-haplotypes as a result of newly obtained complete control region sequence (excluding the repeat) as opposed to only sequences upstream of the repeat being collected in previous studies. Additionally, 24 new A haplotypes were described from the current dataset. In keeping with the previous naming scheme, the new haplotypes are A84 – A107. Counting sub-haplotypes and excluding the ambiguous individuals that clustered into haplogroup A, 70 unique A haplotypes were found in the current dataset.

Haplogroup B was the second largest set both in terms of previously defined haplotypes and where newly sequenced individuals grouped. Previous studies reported 20 haplotypes within the B haplogroup. These 20 types make up 16.3% of all previously defined groups (10,11). In the current dataset, it was found that 139 or 25.2% of all individuals possessed B haplotypes. New individuals were found to contain 8 of the 20 previously defined haplotypes: B1, B3, B6, B8, B10, B11, B12 and B20. Nine new haplotypes were defined, B21 – B29. Counting sub-haplotypes and excluding the ambiguous individuals that clustered into haplogroup B, 24 new B haplotypes were found in the current dataset. The largest single grouping of individuals (n=70) with the same haplotype occurred in B1, which was further sub-divided (Table 2).

The third haplogroup described, haplogroup C, was represented by eight haplotypes in the previously published literature (10,11). Again, this distribution of only 6.5% of the total types previously published closely agrees with the distribution of individuals from the current dataset, 7.6% (n=42), grouping within haplogroup C. Of the eight haplotypes, five were represented in the current dataset: C1, C2, C3, C5 and C8. Additionally, three new haplotypes were described, C9, C10 and C11. Counting sub-haplotypes and excluding the ambiguous individuals that clustered into haplogroup C, nine unique C haplotypes were found in the current dataset.

Haplogroup D was represented by six (4.8%) haplotypes in the literature (10,11) while only one individual (0.2%) from the current dataset fell within haplogroup D, specifically forming sub-haplotype D1a.

No individuals from the current dataset matched any types from haplogroups E or F from previously published studies. The number of total unique haplotypes identified in the current dataset is 104 with a haplotype distribution of A=70, B=24, C=9, D=1, E=0, F=0.

As seen in Figure 2, the distribution chart of the 74 major haplogroups and 3 ambiguous sequence groups, the majority of the haplogroups, 85.1% (n = 63), have less than 10 members. Variants, or sub-haplogroups, were not counted as unique but rather grouped together with the main haplogroup (i.e. B1 = B1a, B1b, B1c, B1d, B1e and B1f). These smaller haplogroups contain only 26.1% (n = 144) of the 552 total individuals in the dataset. Three of the remaining 11 haplogroups are comprised of those sequences that contain ambiguous base calls for each of the 3 major haplogroups. These individuals make up 10.8% (n=60) of the total dataset. While these haplogroups are large, the individuals are only grouped together due to the presence of an ambiguous base in their sequence and not because they share a common control region sequence; thus, they should not be considered when examining haplotype frequency. The remaining 8 haplogroups range in size from 10 individuals to 70 individuals with 63% (n = 348) of the individuals falling in one of these larger haplogroups.

Of the 987 characters in the total mtCR dataset, 9.5% (n=94) were found to be SNPs (Table 2). In the abridged 965 character dataset, excluding the problematic region between 16663 bp and 16676 bp, 5.6% (n=54) of the characters were found to be informative SNPs meaning the SNP was present in two or more individuals (Table 4). Thirty-three of the 54 informative SNPs were found to be highly informative by defining a group that contains 1% or more of the total dogs in the current dataset. Of the 94 SNPs identified in the current study, 24 had not been previously recognized as variable sites in the published literature (10, 11, 19) with 6 of these 24 sites found to be informative and 3 highly informative. Of the 54 informative SNPs, 44 were found in the region upstream of the tandem repeat and only 10 were found in the region downstream of repeat. For the highly informative SNPs, only 6 of the 33 were found in the region following the repeat. Of the informative SNPs found and had not been previously reported in the data, two were in the region upstream of the repeat and the four were in the region downstream of the repeat. Three of which were highly informative. Finally, SNP length varied from 1-85 with ri's between 0 – 100 (Table 4).

A mutational "hotspot" has been identified in the region between 15595 bp and 15653 bp (20). The previous study identified 30 SNPs in the region upstream from the repeat in the dog mtCR with 12 of these 30 SNPs occurring in this 60 bp region. In the current study, 22 SNPs were found in this "hotspot" region, 16 of which were informative. As with the previous study, more SNPs occurred in this 60 bp region than any other comparatively sized region of the mtCR.

Treating all newly collected sequences as a single population, the average pairwise nucleotide difference was 12.49 +/- 5.65 and the nucleotide diversity was 0.013 +/- 0.006. The exclusion capacity of the canine mitochondrial control region excluding the tandem repeat, or 1- ΣX_i^2 (where X_i is the frequency of the i^{th} haplotype), was 0.959 and the random match probability, ΣX_i^2, was 0.041 for all unique haplotypes in the current dataset. In other words, the probability of two dogs having the same control region sequence at random is 4.1 out of 100 relative to this dataset. When the population was split into purebred and mixed breed individuals, the uncorrected pairwise differences decreased slightly, though not significantly, to 12.36 +/- 5.59 for purebred and increased for mixed breed to 12.79 +/- 5.80. Rounded to the thousandths, the nucleotide diversities of the purebred and mixed separate datasets were identical to the combined dataset: 0.013 +/- 0.006. Accordingly, the AMOVA analysis on the dataset showed that there is not a significant difference in genetic variation between the purebred and

mixed populations (Table 5). Dogs were also divided based on the large amount of samples from California (n=189), Pennsylvania (n=100) Virginia (n=61) and Nevada (n=52). Again, AMOVA analysis showed that there is no significant difference in genetic variation in dogs sampled from the different geographic regions based on mtCR sequence (Table 5). The dogs from each state were also evaluated based on how they were distributed among the four major haplogroups. As can be seen from Figure 3, the distribution among haplogroups is consistent regardless of geographic location. The third AMOVA analysis of large purebred groups (n>6) consisted of Golden Retrievers (n=39), Labrador Retrievers (n=31), Basset Hounds (n=8), Dachshunds (n=8), Poodles (n=8), Border Collies (n=7), Boston Terriers (n=7), Cavalier King Charles Spaniels (n=7), Cocker Spaniels (n=7), Jack Russell Terriers (n=7), Miniature Schnauzers (n=7), Rottweilers (n=7) and West Highland Terriers (n=7). As can be seen from Table 4, all dogs from the same breed do not consistently share a haplotype. However, the AMOVA results do show evidence of genetic population substructure when dogs are grouped according to breed (Table 5).

Discussion

This project was intended to survey the largest known sample set of mtCRs isolated from domestic dogs across the United States. While sequencing 427 new mtCRs, we searched for new SNPs and haplotypes and added this data to 128 published samples. We evaluated the need to distinguish between purebred and mixed breed dogs and dogs from different geographic regions across the continental United States. We also looked at the necessity of sequencing multiple individuals of the same breed for a thorough database.

When collecting samples, discrepancies were found in breed definition. For example, some samples were received labeled by the donor as “Spitz”. While there are Finnish Spitz’s and German Spitz’s, Spitz is not a true breed designation but another name for an American Eskimo dog. It is unknown whether the donor meant one of the specific Spitz breeds or if the dog was in fact an American Eskimo dog. Also, two samples were received with the breed listed as “unknown”, but one was described as purebred and one described as mixed. Descriptions could not be clarified or changed as this could be error prone without seeing the dog. As a result of each of the above mentioned problems, the number of distinct breeds collected for this study may be inflated. Population analyses were done to assess the severity of this issue including an AMOVA.

During sequencing, the tandem repeat was excluded due to the known possibility of variation within an individual (8). While excluding the tandem repeat region from control region studies has come to be common practice (10,11, 16-18), it appears that the studies conducted by our lab are the first to have problems obtaining the sequence for the region following the repeat (9). The sequencing problems seem to result from either individuals having a different number of repeats in the tandem repeat region (16130-16430 bp), individuals having a different number of C’s and/or T’s at the C/T stretch (16663 – 16676 bp), or a combination of both. This resulted in multiple sequence runs from the same individual being slightly different across these regions. Because there are multiple mitochondria per cell and multiple mitochondrial genomes per mitochondria, the differences between the genomes per mitochondria and per cell caused the DNA sequence reads to be shifted by one or a few bases due to the insertion or deletion of bases in problematic regions. This resulted in ambiguous bases being coded with the corresponding IUB code and the region between 16663 (nucleotide position) np and 16676 np being excluded when using a multiple alignment to search for informative SNPs.

The phylogenetic analysis showed that all dogs in our current dataset grouped within previously defined haplogroups A, B, C and D (Table 4). The proportions of samples within each group are very similar to the portions of unique haplotypes previously identified for each group. This is particularly interesting because the samples used in previous studies came from all over the world, while the samples in the current study are from the United States alone. It appears that regardless of local origin, more domestic dogs have an A haplotype than any of the other types described. Next is haplotype B followed by C and then D. Additional local studies are needed to confirm this observation. The lack of individuals from groups E and F is most likely due to the fact that the individuals in previous studies that formed groups E and F were collected from Asian and/or Siberian localities {Kim, 1998 #10; Okumura, 1996 #14; Savolainen, 2002 #7; Tsuda, 1997 #11}. Individuals with D, E, and F haplotypes have been found in much lower frequencies compared to individuals with types A, B, and C in world-wide samplings {Savolainen, 2002 #7}, which demonstrates that these haplotypes are more rare in the dog population.

Seventy-four sub-haplogroups were found in the current dataset with 63% of the dogs grouping in 1 of 8 sub-haplogroups containing between 10 and 70 individuals (Figure 2). The distribution of sub-haplogroup sizes shows that while there are many canine mitochondrial control region haplotypes, the majority of dogs share a few common haplotypes while the minority had unique or fairly unique haplotypes. These results demonstrate a recurring problem with canine mitochondrial control region sequence data: most dogs share identical types. This also indicates a need for the evaluation of the remainder of the canine mitochondrial genome to look for additional SNPs that may further break up these large haplogroups (Webb and Allard, submitted).

All of the variable sites identified in the current dataset are listed in Table 2 with the informative and highly informative sites shown in Table 4. Identification of informative SNPs is important when trying to recognize the most useful SNPs for assessing population variation. How informative a SNP is said to be is relative to the size and variation present in the dataset. Knowing where these informative SNPs occur in the mtCR allows for the potential development of SNP panels. Rather than sequencing the entire domestic dog mtCR, one could target the specific sites that distinguish between haplogroups, cutting down on resources and DNA necessary for the analysis. Our identification of 24 new SNPs, 6 of which were found to be informative and 3 highly informative, shows that previous studies have not resulted in a complete sampling of dog mtCRs, especially the region downstream of the repeat. All of the newly identified informative and highly informative SNPs were found in this less commonly sequenced region. While this contradicts the other findings of more informative SNPs upstream of the repeat region, the lack of sequencing and analysis of the region downstream of the repeat most likely explains this finding. As more sequences are added to the dataset, new sites may become phylogenetically informative due to the discovery of shared SNPs. Sites already identified as informative may gain a higher ranking due to their presence in more individuals. Also, the requirement of defining 1% of the total individuals in the dataset as criteria for the third ranking of SNP is subjective and changing this requirement may lead to changes in the ranking of SNPs.

As forensic samples are often subjected to conditions that may degrade DNA, the presence of the 60bp hotspot within the mtCR is particularly useful. While the number of unique haplotypes gleaned from only 60 bases is not going to be as large as those from the entire mtCR, this provides a region of high variability to target when the entire mtCR cannot be sequenced.

Conversely, specific SNPs such as those occurring at position 16439 bp seem to show higher levels of heteroplasmy relative to the remainder of the dataset, which is represented in our dataset as ambiguous base calls. As such, we recommend that future researchers pay close attention to base calls at these sites when editing their raw sequence data, and if possible, clone this region to further investigate these ambiguous sites.

The exclusion capacity and random match probabilities calculated for the dataset are slightly more powerful but similar to those previously reported (20, 21). The additional power comes from a larger sampling of dogs leading to more genetic variation in the dataset. This statistic varies depending on the dataset, and ideally, all existing and future control region sequences should be stored in the same database. As a result, a single statistic calculated for all control regions would be collected.

The nucleotide diversity and fixation index (Fst) both identify the lack of genetic structure within dogs when grouped as purebred and mixed. This shows that the decision as to how to classify certain breeds (i.e. Labradoodles) is trivial as purebred dogs and mixed breed dogs are not distinct populations based on mtCR sequence (Table 5). The AMOVA analysis also resulted in a low Fst value when dogs were grouped by state of sample origin and the distribution of dogs within each major haplogroup was consistent across the different geographic regions. This finding, along with the consistent distribution of haplogroups across states (Figure 3), supports previous studies that there is no need for local canine mtCR SNP databases within the continental United States (20). The significant Fst value when dogs are grouped by breed is most likely due to the strong amount of inbreeding that occurs in purebred dog lineages. While dogs of the same breed do not always share identical mtCR sequences, there is a higher within breed similarity than among the breeds as a whole. This demonstrates why multiple individuals from a single breed and, more importantly, individuals from a variety of different breed types need to be collected to establish a thorough database of domestic dog mtCR SNPs.

Conclusions

As a result of combining 427 newly sequenced domestic dog mtCRs with a previous study of 125 domestic dog mtCRs (9), we have identified both new haplotypes and new informative SNPs. The results of the current study were consistent with previous studies. They found that domestic dogs were grouped into one of four previously identified major groups when using mtCR DNA. The dogs in this study were grouped within 37 of the previously defined 179 sub-haplogroups or formed 1 of the 36 new sub-haplogroups defined by a previously unrecorded mtDNA haplotype. The majority of the 552 dogs, 63%, were grouped into 1 of the 8 large sub-haplogroups with between 10 and 70 individuals per group. This indicates the need for the sequencing and analysis of the remainder of the domestic dog mtGenome (mtGenome) in hopes of identifying additional discriminatory SNPs to break up these large haplogroups and sub-haplogroups (Webb and Allard, submitted). Additionally, 94 SNPs were identified in the current dataset. Of the 94, 54 SNPs were informative, and 33 SNPS were highly informative with 24, 6 and 3 SNP sites, respectively, being previously unrecognized in the published literature. In general, population analyses show that domestic dogs are one large population. Smaller populations such as “purebred” and “mixed” or geographic populations cannot be distinguished based on mtCR sequences. However, when dogs are grouped by breed, they have less genetic variation than the population as a whole. These population analyses demonstrate the need to sample across a variety of breeds, including multiple individuals of the same breed, and that local mtCR SNP databases are not needed within the United States.

Acknowledgments

We thank Sheri Church and Mark Wilson and two anonymous reviewers for careful review of this manuscript. The research was conducted at The George Washington University. The 698 new domestic dog samples were collected through donations from veterinary practices and private donors. We would like to thank Adobe Animal Hospital, Austin Vet Hospital, Caring Hands Animal Hospital, Del Paso Vet Clinic, Little River Vet Clinic, The College of Veterinary Medicine at Mississippi State University, Pet Medical Center of Las Vegas, Seneca Hill Animal Hospital, The Animal Clinic of Clifton, West Flamingo Animal Hospital for blood and tissue samples and 80 private donors who provided buccal swabs to add to our collection. Aisling Kelley, Stephanie Carnation and Dunia Qutub helped to collect sequence data. This research was funded by the National Institute of Justice through grant 2004-DN-BX-K004 to M. W. Allard.

Va References

1. Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med. 1995;108:68-74.

2. Deedrick DW. Hairs, Fibers, Crime, and Evidence. Forensic Science Communications. 2000;2.

3. Bogenhagen D, Clayton D. The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial dexoyribonucleic acid. Journal of Biol Chem. 1974;249:7991-7995.

4. Nass M. Mitochondrial DNA. I. Intramitochondrial distribution and structural relations of single- and double-length circular DNA. Journal of Molecular Biology. 1969;42:521-528.

5. Kim KS, Lee SE, Jeong HW, Ha JH. The complete nucleotide sequence of the domestic dog (Canis familiaris) mitochondrial genome. Mol Phylogenet Evol. 1998;10:210-20.

6. Parsons TJ, Coble MD. Increasing the Forensic Discrimination of Mitochondrial DNA Testing through Analysis of the Entire Mitochondrial DNA Genome. Croatian Medical Journal. 2001;42(3):304 - 309.

7. Pesole G, Gissi C, De Chirico A, Saccone C. Nucleotide substitution rate of mammalian mitochondrial genomes. J Mol Evol. 1999;48:427-34.

8. Savolainen P, Arvestad L, Lundeberg J. A novel method for forensic DNA investigations: repeat-type sequence analysis of tandemly repeated mtDNA in domestic dogs. J Forensic Sci. 2000;45:990-9.

9. Gundry RL, Allard MW, Moretti TR, Honeycutt RL, Wilson MR, Monson KL, Foran DR. Mitochondrial DNA analysis of the domestic dog: control region variation within and among breeds. J Forensic Sci. 2007;52:562-72.

10. Angleby H, Savolainen P. Forensic informativity of domestic dog mtDNA control region sequences. Forensic Sci Int. 2005;154:99-110.

11. Savolainen P, Zhang YP, Luo J, Lundeberg J, Leitner T. Genetic evidence for an East Asian origin of domestic dogs. Science. 2002;298:1610-3.

12. Wilson MR, Allard MW, Monson K, Miller KW, Budowle B. Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region. Forensic Sci Int. 2002;129:35-42.

13. Pereira L, Van Asch B, Amorim A. Standardisation of nomenclature for dog mtDNA D-loop: a prerequisite for launching a Canis familiaris database. Forensic Sci Int. 2004;141:99-108.

14. Nixon K. WinClada ver 1.00.08, www.cladistics.com

15. Nixon KC, The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis. Cladistics. 1999;15:407-414.

16. Takahasi S, Miyahara K, Ishikawa H, Ishiguro N, Suzuki M. Lineage classification of canine inheritable disorders using mitochondrial DNA haplotypes. J Vet Med Sci. 2002;64:255-9.

17. Tsuda K, Kikkawa Y, Yonekawa H, Tanabe Y. Extensive interbreeding occurred among multiple matriarchal ancestors during the domestication of dogs: evidence from inter- and intraspecies polymorphisms in the D-loop region of mitochondrial DNA between dogs and wolves. Genes Genet Syst. 1997;72:229-38.

18. Wetton JH, Higgs JE, Spriggs AC, Roney CA, Tsang CS, Foster AP. Mitochondrial profiling of dog hairs. Forensic Sci Int. 2003;133:235-41.

19. Okumura N, Ishiguro N, Nakano M, Matsui A, Sahara M. Intra- and interbreed genetic variations of mitochondrial DNA major non-coding regions in Japanese native dog breeds (Canis familiaris). Anim Genet. 1996;27:397-405.

20. Himmelberger AL, Spear TF, Satkoski JA, George DA, Garnica WT, Malladi VS, et al. Forensic utility of the mitochondrial hypervariable region 1 of domestic dogs, in conjunction with breed and geographic information. J Forensic Sci. 2008 Jan;53(1):81-9.

21. Savolainen P, Rosen B, Holmberg A, Leitner T, Uhlen M, Lundeberg J. Sequence analysis of domestic dog mitochondrial DNA for forensic use. J Forensic Sci 1997; 42 (4): 593-600.

Table 1 – Breed List

Breed	Purebred	Mixed
Airedale	3	
AiredaleTerrier	1	
Akita	2	
AlaskanHusky	1	
AlaskanMalamute	1	
AmericanCocker	1	
AmericanEskimoDog	2	1
AmericanSpitz	1	
AmericanStaffordshire	1	
AnatolianShepherd	2	
AustralianShepherd	6	3
AustralianTerrier	1	
Basset	1	
BassetHound	8	
Beagle/Corgi		1
Beagle/Labrador		1
Beagle	5	4
BeardedCollie	1	
BelgianSheepdog	1	
BerneseMountainDog	4	
BichonFrise	5	4
BloodHound	1	
BlueHeeler	2	
Bolognese	1	
BorderCollie	7	4
BostonTerrier	7	
Boxer	5	1
BrittanySpaniel	2	1
Bulldog	3	
BullMastiff	4	
BullTerrier	2	
CairnTerrier	1	1
CardiganCorgi	2	
CarrinTerrier	1	
Catahoula		1
CavalierKingCharlesSpaniel	7	
ChesapeakBayRetriever	3	
Chihuahua	5	9
ChocolateLabradorRetriever	6	1
Chow	1	1

ChowChow	2	
Cockapoo		2
CockerSpaniel/Poodle		1
CockerSpaniel	7	1
Collie	2	1
Corgi	5	1
CotonDeTulear	3	
Cur	1	
Dachshund	8	
Dalmation	3	1
Doberman	2	
DobermanPinscher	5	1
DoguedeBordeaux	1	
EnglishBulldog	2	
EnglishMastiff	3	
EnglishShepherd		1
EnglishSpringerSpaniel	2	
EnglishTerrier	1	
EskimoDog	1	
FinnishSpitz	1	
FlatCoatedRetriever	3	
FoxTerrier	1	1
FrenchBulldog	1	
GermanShepherd	4	1
GermanShortHairedPointer	2	
GoldenRetriever/Poodle		1
GoldenRetriever	39	
GreatDane	6	
GreatPyrenees	1	
Greyhound	1	
Havanese	5	
HuntingDog	1	
Husky/Retriever		1
Husky/Shepherd		1
Husky	4	1
ItalianGreyhound	1	
JackRussell/Beagle		1
JackRussell	7	2
JapaneseChin/LhasaApso		1
Keeshond	3	
KerryBlueTerrier	1	
Labradoodle	3	4
Labrador/BorderCollie		1
Labrador/Dane		1

Labrador	2	
LabradorRetriever	31	4
Leonberger	1	
LhasaApso	4	2
Maltese/ShihTzu		1
Maltese	5	3
Maltipoo	1	
ManchesterTerrier	2	
Maremma	2	
Mastiff	2	
MiniatureDachshund	2	
MiniaturePinscher	2	1
MiniaturePoodle	4	
MiniatureSchnauzer	7	
Mix		3
MunsterlanderPointer	1	
NeapolitanMastiff	2	
Newfoundland	1	
NorwegianElkhound	1	
OldEnglishSheepdog	4	
Papillon/Sheltie		1
Papillon	1	
PembrokeCorgi	1	
PembrokeWelshCorgi	1	
PharaohHound	1	
PitBull	2	
PitBullTerrier	5	3
Pointer	1	
Pomeranian	5	3
Poodle	8	6
PortugueseWaterDog	3	1
Pug/JackRussell		1
Pug/Jug		1
Pug	6	
RatTerrier	1	
Ridgeback	1	
Rottweiler/St.Bernard		1
Rottweiler	7	
RoughCollie	1	
SaintBernard	1	
Samoyed	1	
Sapsaree*	1	
Schipperke	2	
Schnauzer/Poodle		1

Schnauzer	4	2
ScottishTerrier	2	
SharPei	3	
SharPlaninetz	2	
Sheltie	4	1
Shepherd/Chow		2
Shepherd/Labrador		1
Shepherd		8
ShetlandSheepdog	1	
ShibaInu	6	
ShihTzu/LhasaApso		1
ShihTzu	5	2
ShiloShepherd	1	
SiberianHusky	1	
Spitz		1
SpringerSpaniel	1	
StaffordBullTerrier	1	
StandardPoodle	1	
SwissMountainDog	1	
TeacupMaltese	1	
Terrier		3
TibetanMastiff	1	
TibetanSpaniel	1	
TibetanTerrier	1	
ToyChow	1	
ToyFoxTerrier	1	
ToyPoodle	6	
Unknown	2	1
Vizsla	3	
WalkerHound	1	
Weimaraner	3	
WelshCorgi	1	
WestHighlandTerrier	7	
WestHighlandWhiteTerrier	2	
WheatonTerrier	2	
Whippet	1	
WhiteSchnauzer		1
Wire-hairedDachshund	1	
Yorkie-Chihuahua		1
Yorkie-Poodle		2
YorkshireTerrier	6	1

Table 1. A complete list of all dogs used in the study. Each breed is listed as well as the number of purebred and mixed breed for each breed. All breed names and types were determined by the sample donor. *Sapsarsee is the dog used by Kim et al. (5) and is the reference sequence.

Table 2 – Haplotype Descriptions
(Note to editor – Table 2 can be found in the separate excel spreadsheet)

Table 2 - This table lists the haplotype name in the left most column, followed by the number of dogs that possess the haplotype and the SNPs defining each type. The row at the top contains the coordinates of each SNP relative to the Kim et al. (5) reference sequence, whose nucleotides are listed immediately below the coordinates at the varying sites. Asterisks (*) above a coordinate indicate a new SNP (not including ambiguous base calls) found in this study relative to previously published data. All SNPs are listed as the variable nucleotide at the corresponding position. Coordinates shaded in grey indicate informative SNPs in Table 5. A dot (.) indicates a match to the reference sequence and a blank cell indicates that when aligned to the reference sequence that position does not exist in the sample.

Table 3 - Distribution of Haplotypes

Haplotype	Breed	(n) per breed	Total (n)	%
A1	American Eskimo Dog	1	7	1.27
	Belgian Sheepdog	1		
	Border Collie	1		
	Catahoula	1		
	Doberman Pinscher	2		
	Rough Collie	1		
A2	French Bulldog	1	11	2.17
	Great Dane	5		
	Leonberger	1		
	Saint Bernard	1		
	Schnauzer	1		
	Scottish Terrier	2		
A2a	West Highland Terrier	1	2	0.36
	Pit Bull Terrier	1		
A5a	Labrador Retriever	3	3	0.54
A5b	Jack Russell	1	7	1.27

	Pug/Jack Russell	1		
	Pug/Jug	1		
	Sheltie	3		
	Shetland Sheepdog	1		
A5c	Labrador Retriever	1	1	0.18
A11	American Staffordshire	1	40	7.25
	Anatolian Shepherd	4		
	Australian Shepherd	1		
	Border Collie	1		
	Border Collie	2		
	Boston Terrier	1		
	Boxer	1		
	Bulldog	1		
	Chihuahua	1		
	Chihuahua	1		
	Chocolate Labrador Retriever	1		
	Chow Chow	1		
	Cocker Spaniel	1		
	Collie	1		
	English Bulldog	1		
	English Springer Spaniel	1		
	Husky/Shepherd	1		
	Husky	1		
	Jack Russell	2		
	Labrador Retriever	1		
	Labrador Retriever	1		
	Miniature Dachshund	1		
	Miniature Schnauzer	1		
	Old English Sheepdog	1		
	Pembroke Welsh Corgi	1		
	Pit Bull Terrier	1		
	Rottweiler	2		
	Schnauzer	1		
	Shepherd	3		
	Shih Tzu	1		
	Springer Spaniel	1		
	Yorkshire Terrier	1		
A11a	Labrador Retriever	1	8	1.45
	Manchester Terrier	2		

	Rottweiler	4		
	Rottweiler/St. Bernard	1		
A11b	Chihuahua	1	3	0.54
	Dachshund	1		
	Papillon	1		
A11c	Terrier	1	1	0.18
A11d	Greyhound	1	1	0.18
A11e	Airedale	1	1	0.18
A16	Brittany Spaniel	1	37	6.70
	Chesapeake Bay Retriever	2		
	Chocolate Labrador Retriever	1		
	Chow	1		
	English Mastiff	2		
	Golden Retriever	5		
	Italian Greyhound	1		
	Labradoodle	2		
	Labradoodle	2		
	Labrador/Border Collie	1		
	Labrador/Dane	1		
	Labrador	16		
	Labrador Retriever	1		
	Yorkshire Terrier	1		
A17a	Beagle	1	57	10.33
	Bichon Frise	2		
	Bichon Frise	1		
	Boston Terrier	3		
	Boxer	3		
	Bull Mastiff	3		
	Bull Terrier	1		
	Cavalier King Charles Spaniel	4		
	Chihuahua	2		
	Chocolate Labrador Retriever	3		
	Coton de Tulear	1		
	Dalmatian	1		
	Dalmatian	1		
	Dogue de Bordeaux	1		
	English Mastiff	1		
	Flat Coated Retriever	2		
	Great Dane	1		

	Jack Russell	1		
	Jack Russell	2		
	Labrador	2		
	Mastiff	2		
	Miniature Dachshund	1		
	Miniature Pinscher	1		
	Pit Bull	3		
	Pomeranian	1		
	Pug	3		
	Rottweiler	1		
	Samoyed	1		
	Shar Pei	1		
	Shepherd/Labrador	1		
	Shepherd	2		
	Shiba Inu	1		
	Stafford Bull Terrier	1		
	Toy Fox Terrier	1		
	Unknown	1		
A17b	Dalmatian	1	1	0.18
A17c	Yorkshire Terrier	1	1	0.18
A17d	PitBullTerrier	1	1	0.18
A18	Bearded Collie	1	44	7.97
	Chihuahua	3		
	Cockapoo	1		
	Cocker Spaniel	1		
	Dachshund	2		
	English Springer Spaniel	1		
	Fox Terrier	1		
	German Shepherd	1		
	Havanese	4		
	Husky	1		
	Jack Russell	2		
	Lhasa Apso	2		
	Lhasa Apso	1		
	Maltese	1		
	Maltese	2		
	Old English Sheepdog	2		
	Pomeranian	1		
	Poodle	1		

	Pug	3		
	Sheltie	1		
	Shepherd	1		
	Teacup Maltese	1		
	Toy Chow	1		
	Toy Poodle	3		
	Vizsla	3		
	Weimaraner	2		
	Whippet	1		
A18a	Miniature Schnauzer	1	4	0.72
	Schnauzer	2		
	White Schnauzer	1		
A18b	American Cocker	1	2	0.36
	Dachshund	1		
A18c	Sheltie	1	1	0.18
A19	Australian Shepherd	1	13	2.36
	Beagle/Corgi	1		
	Beagle/Labrador	1		
	Dachshund	1		
	English Terrier	1		
	German Shepherd	3		
	German Short Haired Pointer	1		
	Jack Russell/Beagle	1		
	Mix	1		
	Portuguese Water Dog	1		
	Shilo Shepherd	1		
A20	Chihuahua1	1	6	1.09
	Coton de Tulear	1		
	Maremma	1		
	Papillon/Sheltie	1		
	Pharaoh Hound	1		
	Pointer	1		
A20a	Miniature Poodle	1	3	0.54
	Poodle	2		
A20b	English Shepherd	1	1	0.18
A22	Bernese Mountain Dog	4	7	1.27
	Bull Mastiff	1		
	Neapolitan Mastiff	2		
A24	Brittany Spaniel	2	3	0.54

	Ridgeback	1		
A26	Cairn Terrier	1	8	1.45
	Cairn Terrier	1		
	Cavalier King Charles Spaniel	2		
	Newfoundland	1		
	West Highland Terrier	1		
	Wheaton Terrier	2		
A27	Bichon Frise	1	5	0.91
	Keeshond	3		
	Lhasa Apso	1		
A27b	Corgi	1	1	0.18
A27c	Pit Bull Terrier	1	1	0.18
A28	Cur	1	2	0.36
	Hunting Dog	1		
A29a	Husky/Retriever	1	4	0.72
	Husky	3		
A31a	EskimoDog	1	1	0.18
A33	Golden Retriever/Poodle	1	16	2.90
	Golden Retriever	14		
	Labrador Retriever	1		
A33a	Golden Retriever	1	1	0.18
A33b	Golden Retriever	1	1	0.18
A40a	Swiss Mountain Dog	1	1	0.18
A66	Cavalier King Charles Spaniel	1	1	0.18
A68	Shiba Inu	3	3	0.54
A70	Collie	1	1	0.18
A71	Cardigan Corgi	1	5	0.91
	Corgi	2		
	Miniature Pinscher	1		
	Pembroke Corgi	1		
A71a	Akita	1	1	0.18
A80a	Munsterlander Pointer	1	2	0.36
	Yorkshire Terrier	1		
A80b	Yorkshire Terrier	1	1	0.18
A82a	German Shepherd	1	2	0.36
	Terrier	1		
A84*	Poodle	2	2	0.36
A85*	Golden Retriever	1	5	0.91
	Labrador Retriever	4		

A86*	Bichon Frise	1	3	0.54
	Beagle	1		
	Boxer	1		
A87*	Miniature Schnauzer	5	5	0.91
A88*	Cocker Spaniel	1	2	0.36
	Shih Tzu	1		
A89*	Maremma	1	1	0.18
A90*	Alaskan Malamute	1	1	0.18
A91*	Miniature Pinscher	1	1	0.18
A92*	Bulldog	1	1	0.18
A93*	Golden Retriever	1	1	0.18
A94*	Chow Chow	1	1	0.18
A95*	Old English Sheepdog	1	1	0.18
A96*	Beagle	1	1	0.18
A97*	Tibetan Mastiff	1	1	0.18
A98*	Chihuahua	1	1	0.18
A99*	American Spitz	1	1	0.18
A100*	American Eskimo Dog	1	1	0.18
A101*	Mix	1	1	0.18
A102*	Shepherd/Chow	1	1	0.18
A103*	Shar Pei	1	1	0.18
A104*	Finnish Spitz	1	1	0.18
A105*	West Highland Terrier	1	1	0.18
A106*	Alaskan Husky	1	1	0.18
A107*	Doberman	1	1	0.18
A ambig 1	Akita	1	1	n/a
A ambig 2	Australian Shepherd	1	2	n/a
	Cocker Spaniel	1		
A ambig 3	Beagle	1	1	n/a
A ambig 4	Beagle	1	1	n/a
A ambig 5	Boxer	1	1	n/a
A ambig 6	Bull Terrier	1	1	n/a
A ambig 7	Chihuahua	1	1	n/a
A ambig 9	Pit Bull	1	1	n/a
A ambig 10	Pomeranian	1	1	n/a
A ambig 11	Shepherd	1	1	n/a
B1a	Airedale	2	59	10.69
	Australian Shepherd	1		
	Basset Hound	5		

	Beagle	1		
	Blue Heeler	1		
	Bolognese	1		
	Border Collie	1		
	Bulldog	1		
	Chihuahua	1		
	Corgi	1		
	Corgi	1		
	Dachshund	1		
	English Bulldog	1		
	Fox Terrier	1		
	German Short Haired Pointer	1		
	Golden Retriever	10		
	Great Pyrenees	1		
	Kerry Blue Terrier	1		
	Labradoodle	1		
	Labradoodle	1		
	Labrador Retriever	5		
	Lhasa Apso	1		
	Miniature Poodle	1		
	Poodle	3		
	Schnauzer/Poodle	1		
	Schnauzer	1		
	Shar Pei	1		
	Shih Tzu/Lhasa Apso	1		
	Shih Tzu	1		
	Shih Tzu	2		
	Standard Poodle	1		
	Terrier	1		
	Tibetan Spaniel	1		
	Tibetan Terrier	1		
	Weimaraner	1		
	Welsh Corgi	1		
	West Highland Terrier	2		
B1b	Beagle	1	7	1.27
	Maltese/Shih Tzu	1		
	Mix	1		
	Poodle	2		
	Rat Terrier	1		

	Shepherd/Chow	1		
B1c	Golden Retriever	1	1	0.18
B1d	Golden Retriever	1	1	0.18
B1e	Golden Retriever	1	1	0.18
B1f	Golden Retriever	1	1	0.18
B3a	Maltipoo	1	6	1.09
	Miniature Poodle	1		
	Poodle	1		
	Toy Poodle	2		
	West Highland White Terrier	1		
B6a	Schipperke	1	2	0.36
	Walker Hound	1		
B6b	Shepherd	1	1	0.18
B8a	Flat Coated Retriever	1	1	0.18
B10a	Cocker Spaniel	1	1	0.18
B10b	Maltese	1	1	0.18
B11a	Cocker Spaniel/Poodle	1	3	0.54
	Dachshund	1		
	Shih Tzu	1		
B12a	Bichon Frise	1	1	0.18
B20a	Portuguese Water Dog	2	2	0.36
B21*	Cocker Spaniel	1	3	0.54
	Labrador Retriever	1		
	Yorkshire Terrier	1		
B22*	Bichon Frise	1	2	0.36
	Maltese	1		
B23*	Maltese	2	3	0.54
	Spitz	1		
B24*	Carrin Terrier	1	1	0.18
B25*	Golden Retriever	1	1	0.18
B26*	Chesapeake Bay Retriever	1	1	0.18
B27*	Unknown	1	1	0.18
B28*	Cockapoo	1	1	0.18
B29*	Japanese Chin/Lhasa Apso	1	1	0.18
B Ambigs 1	Airedale Terrier	1	7	n/a
	Basset Hound	1		
	Cardigan Corgi	1		
	Chocolate Labrador Retriever	1		
	Labradoodle	1		

	Shih Tzu	1		
	Yorkie-Poodle	1		
B Ambigs 2	American Eskimo Dog	1	1	n/a
B Ambigs 3	Australian Shepherd	1	1	n/a
B Ambigs 4	Australian Terrier	1	1	n/a
B Ambigs 5	Basset Hound	1	1	n/a
B Ambigs 7	Basset Hound	1	1	n/a
B Ambigs 8	Basset Hound	1	1	n/a
B Ambigs 9	Beagle	1	2	n/a
	Boston Terrier	1		
B Ambigs 10	Bichon Frise	2	2	n/a
B Ambigs 12	Blood Hound	1	1	n/a
B Ambigs 15	Chihuahua	1	1	n/a
B Ambigs 17	Chocolate Labrador Retriever	1	3	n/a
	Corgi	1		n/a
	Coton de Tulear	1		n/a
B Ambigs 20	Dachshund	1	1	n/a
B Ambigs 21	Doberman Pinscher	1	1	n/a
B Ambigs 22	Doberman Pinscher	2	2	n/a
B Ambigs 24	Golden Retriever	1	1	n/a
B Ambigs 25	Jack Russell	1	1	n/a
B Ambigs 27	Maltese	1	1	n/a
B Ambigs 28	Poodle	2	2	n/a
B Ambigs 30	Portuguese Water Dog	1	1	n/a
B Ambigs 31	Schipperke	1	1	n/a
B Ambigs 33	Toy Poodle	1	1	n/a
B Ambigs 34	Unknown	1	1	n/a
B Ambigs 35	Wire-haired Dachshund	1	1	n/a
B Ambigs 36	Yorkie-Chihuahua	1	1	n/a
C1a	Siberian Husky	1	1	0.18
C2a	Dalmatian	1	3	0.54
	West Highland Terrier	2		
C2b	Boston Terrier	1	4	0.72
	Chihuahua	1		
	Lhasa Apso	1		
	Yorkshire Terrier	1		
C3a	Australian Shepherd	1	12	2.17
	Border Collie	4		
	Cocker Spaniel	1		

	Havanese	1		
	Pomeranian	2		
	Pomeranian	1		
	Poodle	1		
	Shiba Inu	1		
C5a	Anatolian Shepherd	1	3	0.54
	Shar Planinetz	2		
C8a	Border Collie	1	5	0.91
	Doberman	1		
	Doberman Pinscher	1		
	Pit Bull Terrier	1		
	Pomeranian	1		
C9*	Boston Terrier	1	1	0.18
C10*	Pomeranian	1	1	0.18
C11*	Border Collie	1	1	0.18
C Ambig 1	Beagle	1	1	n/a
C Ambig 2	Blue Heeler	1	1	n/a
C Ambig 3	Chow	1	1	n/a
C Ambig 4	Cocker Spaniel	1	3	n/a
	Miniature Poodle	1		
	Schnauzer	1		
C Ambig 5	Collie	1	1	n/a
C Ambig 7	Pit Bull Terrier	1	1	n/a
C Ambig 9	Shiba Inu	2	1	n/a
	Yorkie-Poodle		1	
C Ambig 10	West Highland White Terrier	1	1	n/a
D1a	Norwegian Elkhound	1	1	0.18

Table 3 - The haplotype distribution of 551 domestic dogs relative to the Kim et al. (5) reference sequence. Haplotype name, breed, number of individuals per breed, number of individuals per haplotype, and frequency (%) that haplotype observed are provided. Haplotype names refer to Table 2.

Table 4 – Informative Sequence Variants in the Domestic Dog mtCR

Coordinate	Reference	Observed	L	ri	Coordinate	Reference	Observed	L	ri
15464.1	-	C	6	37	15800	T	C	2	99
15475	T	C	1	100	15807	C	T	1	100
15483	C	T	1	100	15814	C	T	1	100
15508	C	T	1	100	15815	T	C	2	99
15513	G	A	1	100	15819	T	C	1	100
15526	C	T	2	99	15912	C	T	2	99
15553	A	G	13	20	15931	A	-	2	92
15595	C	T	7	95	15938	G	-	5	90
15611	T	C	1	100	15955	C	T	39	85
15612	T	C	1	100	15959	C	T	4	25
15620	T	C	68	48	16003	A	G	1	100
15621	C	T	3	75	16025	T	C	82	38
15622	T	C	1	100	16032	A	G	3	71
15625	T	C	5	55	16083	A	G	4	97
15627	A	G	85	56	16084*	A	G	1	100
15628	T	C	2	75	16128	G	A	2	99
15632	C	T	2	99	16129.1*	-	G	12	26
15635	A	G	2	66	16430*	G	T/-	12	94
15639	T	A/C/G	45	84	16431	C	-	10	94
15643	A	G	1	100	16432*	A	-	8	95
15650	T	C	2	97	16433*	C	-	9	95
15652	G	A	3	98	16439	T	C	4	97
15653	A	G	2	66	16501	T	C	1	100
15665	T	C	5	60	16507	T	A	1	100
15710	C	T	2	95	16576	A	G	12	21
15750	C	T	1	100	16617*	G	A	2	0
15781	C	T	1	100	16705	C	T	2	94

Table 4 - SNPs that have been found to be variable in 2 or more individuals. The nucleotide coordinate relative to the Kim et al. (5), the reference sequence base (5), the observed base, the character length (L) and character retention index (ri) are listed. See materials and methods for definitions of character length and retention index. Shaded boxes indicate highly informative sites in the current study. Asterisks (*) indicate unrecognized sites in previously published literature.

Table 5 – AMOVA analysis within and breed populations

Dataset	Source of Variation	Degrees of Freedom	Percentage of Variation
Purebred vs Mixed	Among populations	1	1.06
	Within populations	550	98.94

	Total	551	100
		Fst = 0.01057	
By States	Among populations	3	0 (-0.46)
	Within populations	398	100.46
	Total	401	100
		Fst = 0 (-0.00457)	
By Breed	Among populations	12	28.14
	Within populations	139	71.86
	Total	151	100
		Fst = 0.28137	

Table 5 - Grouping and results of AMOVA analysis as preformed in Arlequin to assess population structure between purebred and mixed breed dogs, dogs grouped by geographic state of origin and large breed groups of purebred dogs.

Figure 1 – Canine Mitochondrial Control Region Primers

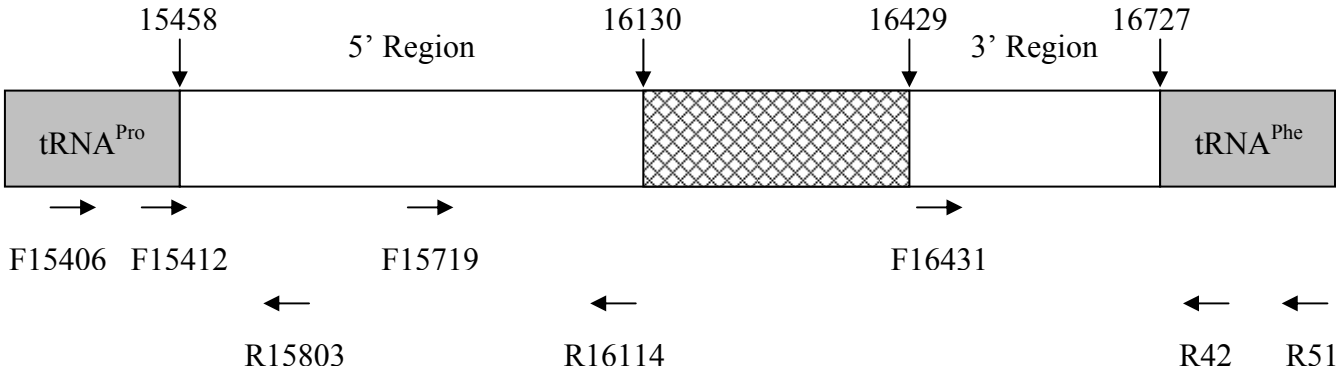
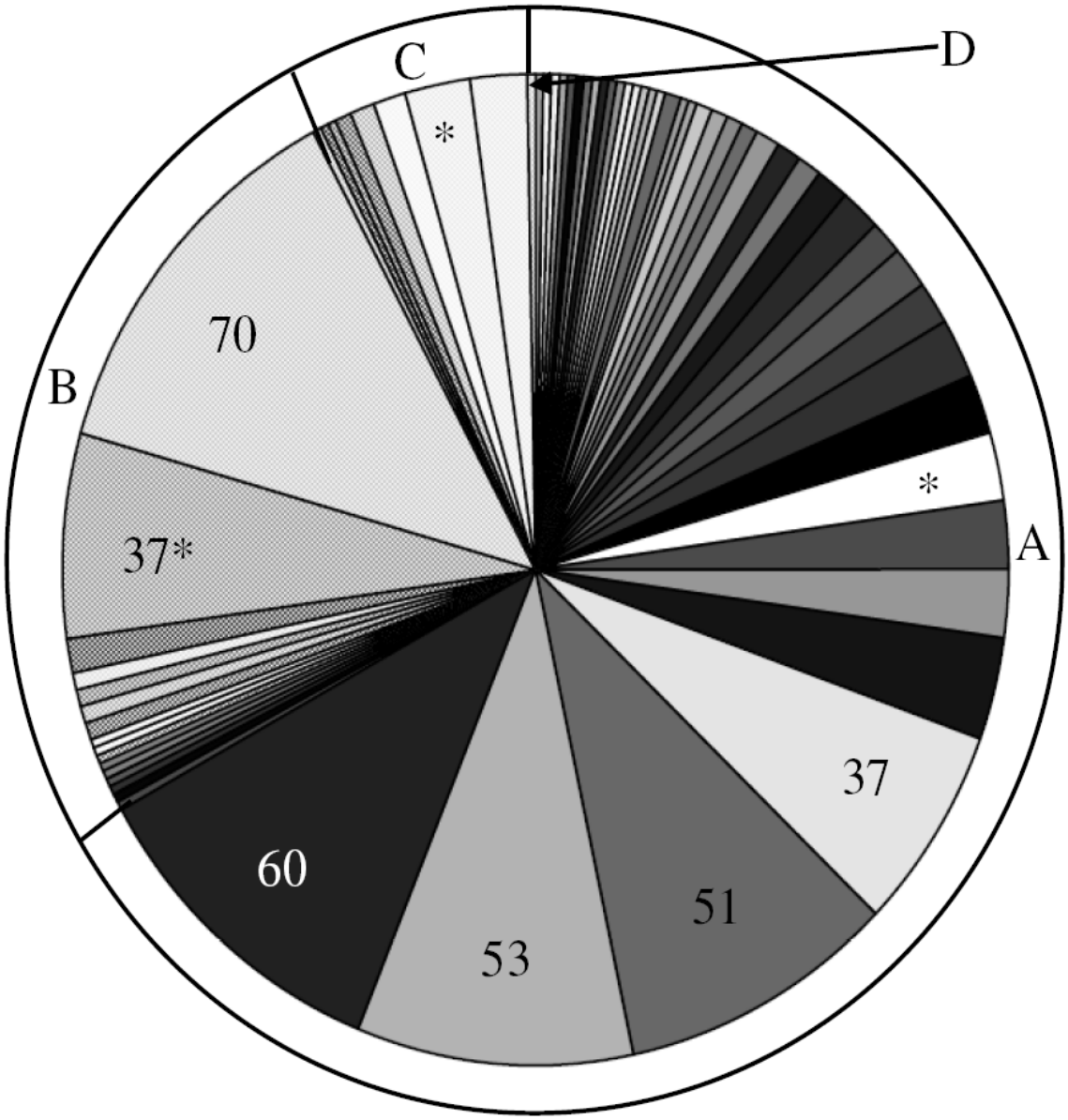


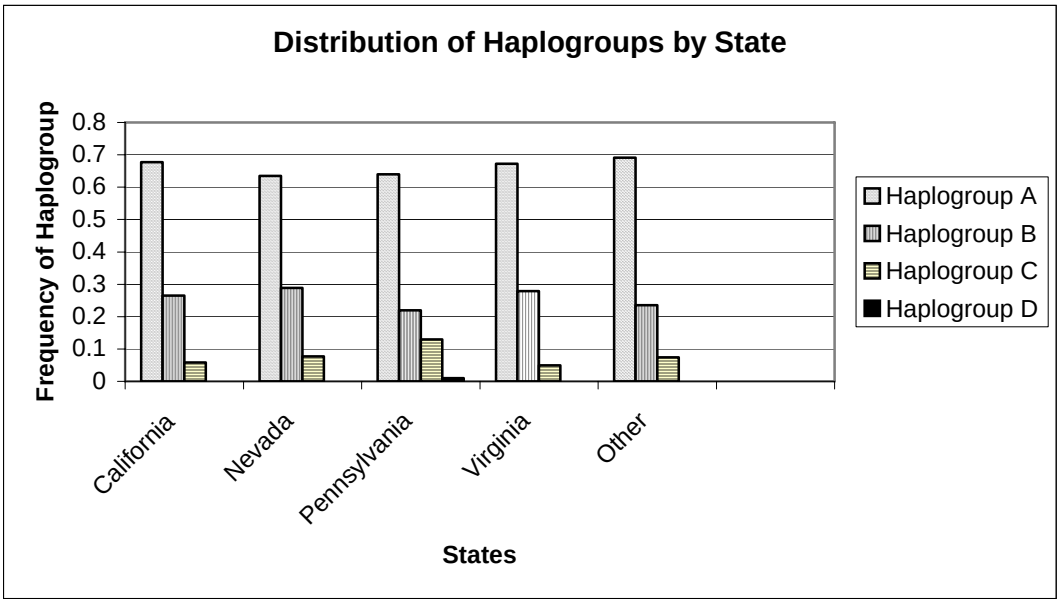
Figure 1. Coordinates and orientation of all canine mitochondrial control region primers. Beginning and end coordinates of control region are shown as well as coordinates of the unsequenced repeat region (indicated by the checkered box) relative to the Kim et al., (5). Primers F15406 and R51 were the primers used for PCR amplification of the control region. All 8 primers were used to obtain 4-6x sequence coverage. All primers except F15406 and R51 were designed by Gundry et al. (9)

Figure 2 – Distribution of Haplogroups



A pie chart showing distribution of haplogroup sizes. Haplogroup A is the largest containing 67% of all dogs surveyed followed by B with 25.2%, C with 7.6% and D with 0.2% of all dogs surveyed. The numbers inside of the slices represent the number of individuals in those large sub-haplogroups. The asterisks (*) identify the 3 haplogroups that are comprised entirely of sequences with ambiguous base calls. Haplogroup B has the largest single group with 70 individuals sharing a haplotype. It should be noted that over half of the pie is comprised of a few haplotypes with many individuals, indicating the need for identification of additional mitochondrial SNPs to break up these large haplotypes.

Figure 3 - Distribution of Haplogroups by State



Bar graph showing the distribution of domestic dogs based on geographic location according to haplogroup. The number of dogs from each location varied: California = 189, Nevada = 52, Pennsylvania = 100, Virginia = 61. The "Other" group was comprised of the remaining 150 from Maryland (n=1), Mississippi (n=8), New York (n=1), Texas (n=14), Vermont (n=1) and Unknown (n=125). The graph shows that there is no bias towards a specific haplogroup based on geographic region.

Main Body of the Final Technical Report:

Mitochondrial Genome DNA Analysis of the Domestic Dog: Identifying Informative SNPs Outside of the Control Region*.

Kristen M. Webb[1]*, B.S.; Marc W. Allard* [1]*, Ph.D.*
[1]Department of Biological Sciences, George Washington University, Washington, DC 20052.

* This work has been presented at The NIJ Conference 2007 and at the GW Research and Discovery Day, 2007. Both instances were in poster form. This work was supported by the National Institute of Justice through grant 2004-DN-BX-K004 to M. W. Allard

Ib. Introduction:

Hair, both human and animal, is often found as evidence in criminal investigations. Because hair is a composite of dead cells, the DNA contained in even fresh hair samples can be degraded (1). Each cell contains only two copies of the nuclear genome, but a second genome is also present in much higher copy numbers, the mtGenome. Mitochondria are organelles responsible for many metabolic tasks within and between cells. When mtDNA is sequenced, focus tends to be on a region of the genome known as the mtCR (also knows as the D-loop or hypervariable region) (5-11)Webb and Allard, 08-027). In canines, the mtCR is approximately 1,272 base pairs (bps) in size, is non-coding, and thus accumulates substitutions faster than any other comparably sized region of the mtGenome (12). This high rate of substitution is useful when looking for variability to help identify samples. In human investigations, the mtCR can indicate the ethnicity of a person (6). Knowing how valuable human mtDNA can be, attempts

have been made to analyze mtDNA from the domestic dog (*Canis lupus familiaris*) for instances when dog hair is found as evidence at a crime scene (5, 7, 8, 11, 13-15) Webb and Allard, 08-027). A 2005-2006 survey found that there were approximately 73 million domestic dogs in the United States (www.appma.org). Because of this, it is not unexpected that dog hair is often found in criminal investigations either when a dog is directly involved in a crime or as secondary transfer from either the victim or suspect. It has been shown that while highly variable, the control region does not distinguish between dog breeds or any of the main groupings of dogs. In a previous study, we found that out of 552 domestic dogs, there were groups containing as many as 59 dogs of varying breeds with identical control region sequences (Webb and Allard, 08-027). In fact, the random match probability of the mtCR for the domestic dog was found to be 4.3% as compared to between 2.5% and 0.52% for the human mtCR ((4), Webb and Allard, 08-027). Knowing that the domestic dog mtCR does not have the discriminatory power of the human mtCR, and also knowing that there are an additional ~15,458 bp of mtGenome outside of the control region, we have sequenced the remainder of the genome for 64 domestic dogs from our mtCR study. We combined our sequences with 15 complete mtGenome sequences downloaded from Genbank (16, 17). We use phylogenetic and population genetic methods to analyze the 79 genomes and report relationships and variable sites in the remainder of the genome that will aid in further discriminating between dogs with common mtCR sequences.

.

IIb Methods

Sample collection and DNA extraction methods were carried out as described in Webb and Allard (08-027).

Primers to amplify and sequence the mitochondrial genome were designed by hand. Eleven PCR primer pairs were designed to amplify products ranging in size from 835 bp to 1918 bp. The PCR primers were designed based on the predicted sizes of the resultant amplified regions rather than based on the coordinates of a specific gene or region. This design scheme lessened our chances of amplifying mitochondrial pseudogenes known to be present in canines (18). The PCR primers were also used as sequencing primers and an additional 69 sequencing primers were designed for a total of 92 primers (Table 1). Due to sequence variability, varying combinations of the 92 primers were used to sequence each dog sample. As a set, the complete genome primers resulted in bidirectional, overlapping, 3-4x high quality sequence coverage across the mitochondrial genome.

PCR and sequencing were carried out as described in Webb and Allard (08-027). Upon completion of sequencing, an additional check against pseudogenes was conducted by translating all genes into their corresponding amino acids and proteins and comparing the translations to the known translation of the representative domestic dog mtGenome published on Genbank (17).

A Genbank search revealed 15 additional complete mitochondrial genomes had been sequenced for the domestic dog. Of these previously published sequences, 14 came from a paper investigating the possibility of selection acting on the domestic dog mtGenome following domestication and the other was the first canine mitochondrial genome to be published (16, 17).

The forensic version of Sequencher 4.1.4FB19 (Gene Codes Corporation, Ann Arbor, MI) was used to edit and align all mtGenome sequences. Alignments were built according to the previously defined criteria for gap placement in forensic evaluations (19). Standard IUB codes were used for polymorphic sites. A recommendation has been made to follow human mtCR methods and compare domestic dog mtCR sequences to a standard reference sequence in an effort to standardize canine mitochondrial nucleotide nomenclature (14). We continued with this

recommendation by using the first canine mtGenome to be published as the reference mtGenome sequence (17). Using a reference sequence allows base coordinates to be compared across different studies (14), thus all coordinates mentioned in this research are in terms of the Kim et al. (17) reference sequence.

Arlequin 3.11 (20) was used to search for groups of dogs with identical mtGenome sequences, or haplotypes, and to calculate the frequency of these haplotypes. Individuals representing each unique haplotype were aligned to the reference sequence and the coordinates and base calls of the single nucleotide polymorphisms (SNPs) were recorded in an Excel spreadsheet.

Using Winclada (21), the alignment was transposed from DNA to numeric characters (A=0, C=1, G=2, T=3) using the view, numeric mode option. As with our previous control region study, Nona (22) and Winclada were used to build a phylogenetic tree to evaluate the relationships between the canines based on mtGenome sequences. A heuristic search was performed on the data following recommended search strategies (23). If the search resulted in multiple trees, a strict consensus tree was created. A strict consensus tree shows only those groups that exist in complete agreement among all fundamental trees. Upon obtaining a final tree, the relationships of the dogs were evaluated and dogs were assigned to a haplogroup based on mtGenome sequences and spatial relation on the tree with other dogs. Since this is the first study to identify and name haplotypes of the mtGenome, we built upon the previously established mtCR naming scheme with the intent of including the haplotype information of the entire genome, mtCR + mtGenome, in the new name. To convey the mtCR haplotype information, the mtCR haplotype name is used within the mtGenome haplotype name but modified by inserting the word “mtGenome” before the mtCR haplotype. A decimal is followed by a numerical distinction indicating different mtGenome types. For example, two individuals with the mtCR haplotype B1a but with different mtGenome haplotypes would now be called mtGenomeB1a.1 and mtGenomeB1a.2. (See results and Table 5 for further clarification). With the mtCR naming scheme, if an ambiguous base is present in the haplotype, the word “Ambig” is inserted into the haplotype name.

Winclada was also used to identify informative SNPs or those nucleotides that define a group of 2 or more individuals. Using the “mop informative characters/delete selected characters” function and then the character diagnoser to trace each character on the tree, informative SNPs were identified. The length and retention index (ri) statistics were recorded for each informative SNP. The length is the number of times the nucleotide state at a given position changes on the tree. The ri is a measure of a nucleotide position having the same base in two individuals being the result of shared common ancestry and not convergence. The ri scores can range from 100 to 0. A score of 100 being obtained when the character change arose only once in the evolution of the group and thus defines all members of a clade. The scores get progressively lower until a score of 0 is reached. This indicates all character changes arose independently.

SNPs were classified into three rankings based on the same criteria as Webb and Allard (08-027) except, due to the smaller dataset size, the third level of ranking contains informative SNPs that define groups of 8 or more individuals, or 10% of the total dogs in the dataset.

All statistics were either calculated in Arlequin or by hand. General population statistics, mean number of pairwise differences and nucleotide diversity, were calculated in Arlequin on the dataset as a whole with each individual defined as a unique haplotype (not removing identical taxa) as well as for purebred and mixed dogs to look for suspected evidence of inbreeding in

purebred individuals, and if individuals labeled "purebred" and "mixed" are distinguishable at the mitochondrial sequence level. The samples were also separated by regional groupings to look for local substructure. The samples were grouped by state: California = 31, Pennsylvania = 16, Nevada = 9, Virginia = 6, Mississippi = 1 and Texas = 1. Dogs were also separated into those breeds with more than one purebred individual to look for within breed structure: Australian Shepherd = 2, Dachshund = 2, German Shepherd = 2, Neapolitan Mastiff = 2, Poodle = 2, Jamthund = 2, Rottweiler = 2, Keeshond = 3, Cocker Spaniel = 3, Basset Hound = 3. The remaining 23 purebred dogs were included as a single group, Singles=23. Additional statistics such as exclusion capacity, $1 - \Sigma X_i^2$, and random match probability, ΣX_i^2, where X_i is the frequency of the i^{th} haplotype, were calculated by hand following the arrangement of individuals with identical sequences into the same group. A gamma value, which is used to account for multiple substations at the same nucleotide site, was calculated by Garli (24) and incorporated into Arlequin for population statistic estimations under the Tamura and Nei model of evolution (25) using AMOVA.

IIIb Results

Six hundred and ninety-eight domestic dog blood, tissue and buccal swab samples were collected from various veterinary practices and private donors across the United States. Of the 698 samples collected, 426 blood and tissue samples were used for control region sequencing and analysis (Webb and Allard, 08-027). Based on the results of the control region analysis, 64 individuals were chosen for complete genome sequencing and are available on Genbank (Table 2). These individuals were chosen based on their sharing of a mtCR haplotype with a large number of other dogs in the dataset (Allard and Webb, 08-027) and/or if the breed type was rare or interesting. Fifty-three of the samples came from purebred individuals and 11 were mixed breed. The 64 newly collected genomes were combined with the 15 purebred dogs downloaded from Genbank (16, 17) for a final dataset of 79 domestic dogs. Table 2 lists the different breeds of dog and number of each included in this study.

All new genomes were sequenced in their entirety and the genomes ranged in size from 15459 bp to 15461 bp excluding the control region. When translating the DNA sequence into corresponding amino acids to check for pseudogenes, a 2 bp "AG" insertion was found at positions 99141.1 and 99141.2 that disrupted the reading frame of the ND4L gene. This insertion was found in all of the newly sequenced dogs as well as those downloaded from the Bjornerfeldt et al. (16) study. The only sample not possessing the 2bp insertion was the Kim et al. (17) reference sequence.

Fourteen of the 79 dogs were identified as being identical to at least one other dog in the dataset based on mtGenome sequence. There was one instance of a purebred and a mixed breed dog sharing an identical mtGenome sequence and the remaining 13 instances of shared mtGenome sequences occurred within the purebred dogs. Of the 65 unique mtGenome haplotypes, 8 of those were due to individuals having ambiguous base calls in their sequence. Excluding these 8 sequences from the calculations, 72.2% of the mtGenomes sequenced were unique. This is much higher than the 18.3% unique canine mtCR haplotypes found in our previous study of 552 mtCRs. When considering only the mtCRs of the 79 dogs used in the current study excluding those dogs with ambiguous mtCR base calls (n=9), 52 dogs were identical to at least one other dog in the dataset or only 25.7% (n=18) of the mtCR sequences were unique.

Following the separate alignments of each unique genome sequence to the Kim et al. (17) reference sequence, 6 gaps were inserted into the matrix: 1493.1, 2679.1, 7015.1, 9865.1, 9914.1 and 9914.2. The final multiple alignment matrix size was 15463 bp by 79 dogs.
Within the roughly 15460 bases of the mtGenome excluding the mtCR, 356 SNPs were found (2.3%). Of the 356 SNPs, 57% (n=202) were found to be informative and 26% (n=94) were found to be highly informative by defining groups of 8 or more dogs (Table 3). In other words, 43% of the SNPs are variations unique to an individual. Comparatively, 9.5% of 987 mtCR bases were found to be variable with 42% being unique to an individual.

When assessing the same set of dogs for the two different mitochondrial regions the phylogenetic relationships were highly similar. In fact, when using mtGenome sequence excluding the mtCR all individuals formed groups with the same individuals as they did using mtCR sequence alone (Figure 1).

A complete list of haplotypes can be found in Table 4 and the frequency of each haplotype as well as each dog possessing a given haplotype can be found in Table 5. Haplogroup A was the largest group containing 60.75% (n=48) of the total individuals in the dataset. Within group A, there were 7 groups of individuals sharing one haplotype, 25 haplotypes unique to an individual, and 6 individuals with ambiguous base calls that could not be placed within a haplotype group. Haplogroup B was the second largest group of dogs containing 25.3% (n=20) of all individuals. Of the 20 individuals, only two groups were formed, 14 individuals had unique mtGenome sequences and one individual was ambiguous. Haplogroup C was the third largest group with 10.1% (n = 8) of all individuals. Seven of the 8 individuals had unique haplotypes and one individual was ambiguous. Haplogroup D was the smallest group containing only 3.8% (n = 3) of all individuals and contained one group of two dogs sharing a haplotype and one individual with a unique haplotype. Figure 2 shows the distribution of individuals relative to their haplotype.

Figure 3 illustrates the distribution of the haplotypes relative to group size. When looking at the mtCR alone there are 18 individuals with unique mtCR sequences and 14 groups with two or more dogs sharing a haplotype. Fifty-two individuals, 65.8%, fall within these 14 groups. When looking at the 79 dogs using mtGenome sequences without the mtCR, the distribution shifts with 10 groups containing a total of 24 (30.4%) dogs and the remaining 69.6% (n = 55) of the dogs having unique haplotypes.

A mutational “hotspot” has been reported in the canine mtCR (26) and confirmed by Webb and Allard (08-027). In the most recent study, this hotspot was defined by 22 mutations occurring in 60 bases, 1 mutation in every 2.7 bases, as opposed the calculated average rate of 1 mutation in every 15 bases for the mtCR. In the mtGenome the calculated average mutation frequency is 1 mutation in every 50 bases. From the distribution of mutation within the mtGenome shown in Figure 4, it can be seen that there are clusters of sequence variation and stretches of the genome where no SNPs are found. The regions with some of the highest frequency of SNPs were bases 10251-10354 with 9 SNPs in 103 bases, bases 11800-12006 with 16 SNPs in 206 bases, and bases 8661-9028 with 23 SNPs in 367 bases. The frequency of SNPs in these 3 regions is 1 in 11.5, 1 in 13, and 1 in 16, respectively. While this is not close to the 1 in 2.7 frequency of the mtCR hotspot, it is significantly greater than the 1 in 50 mutation rate that the mtGenome averages. Conversely, there were regions of 400 base pairs or larger that had very few SNPs. The regions spanning 1767 – 2645 (878 bp) and 9220 – 9824 (604 bp) have only 3 SNPs, and the region spanning 13792 – 14328 (536 bp) has only 2 SNPs. The largest region without any SNPs occurs between bases 9253 – 9707. This 454 bp region, as well as the

larger 604 bp region with only 3 SNPs in which it is contained, spans the coding region for the end of COIII gene, the tRNA-Gly and the beginning of the ND3 gene. Likewise, the other regions with only a few SNPs span the coding region for 16S rRNA, the ND6 gene, tRNA-Glu, and the CYTB gene.

Based upon the frequency of each haplotype, the random match probability for the mtGenome dataset as a whole was calculated to be 0.018, and the exclusion capacity was calculated to be 0.982. This implies that 98 individuals out of 100 can be excluded based on the mtGenome dataset, or 2 out of 100 individuals may have identical haplotypes by chance. Comparatively, the random match probability for the mtCR was calculated to be 0.041 with 96 out of 100 individuals excluded based on the mtCR dataset.

Using Garli, an alpha value for the gamma correction to recognize multiple substitutions at a single nucleotide site was calculated to be 0.0087, which was rounded to 0.01. Treating all newly collected sequences as a single population, the mean number of pairwise differences was 84.14 +/- 36.58 and the nucleotide diversity was 0.005441 +/- 0.002621. When the population was split into purebred and mixed breed individuals, the mean number of pairwise differences decreased slightly to 83.20 +/- 36.24 for purebred and increased for mixed breed to 90.12 +/- 42.05. The nucleotide diversity also decreased slightly to 0.005380 +/- 0.002598 for purebred and increased for mixed breed to 0.005829 +/- 0.003069.

The fixation index (Fst) values in Table 6, which represent the proportion of genetic variation within a subpopulation relative to the total population, are very low for the purebred vs mixed breed, and state comparison shows that grouping dogs by these factors has no genetic basis. However, when grouped by breed, the Fst value becomes significantly larger, which indicates that the presence of population structure within dogs of the same breed even when they do not have identical mtGenome sequences (Table 5). The 23 purebred dogs were grouped together because their were no other dogs of the same breed in the dataset. The population had a Fst of 0.19. Besides the group of Cocker Spaniels which had a population Fst of 0.18 and the German Shepherds which had a population Fst of 0.24, all other breed groups has scores of 0.43 or higher.

Discussion

The aim of this study was to sequence multiple mtGenomes of domestic dog to search for informative SNPs that would break up the large haplotype groups formed by using the mtCR sequence alone and to assess the utility of the mtGenome for forensic analyses. Individuals were chosen for mtGenome sequencing because either they belonged to one of the large mtCR haplotype groups or the breed was of interest. The 64 newly sequenced mtGenomes combined with the 15 mtGenomes downloaded from Genbank form the largest domestic dog mtGenome dataset to be published to date and the first to be used to identify domestic dog mtGenome haplotypes.

During sample collection, donors were asked to determine breed and breed type (either purebred or mixed). As the authors never saw the actual dog, breed and type were never changed, even when the declarations were questionable. For example, 2 samples were received with one being labeled “West Highland White Terrier” and the other “West Highland Terrier.” While these two dogs could very well be of the same breed, they were distinguished as different breeds in the current dataset based on the differing donor descriptions. Individuals with unknown breed type were considered mixed unless otherwise listed by the donor.

The presence of the 2 bp insertion in sequences from the Bjornerfeld et al. (16) study and the fact that our sequencing strategy included designing PCR primers based on amplicon size and not flanking a particular gene or region allows us to conclude that this sequence is not from pseudogene. Upon presenting the results of this study at the 2007 Society of Molecular Biology and Evolution meeting, it was suggested that this might be an error in the DNA sequence of the domestic dog that is corrected by the translational machinery upon translation from DNA to amino acid.

When comparing the mtGenome excluding the mtCR to the mtCR, we first notice that while the mtGenome has more haplotypes, the mtCR has a higher overall percentage of SNPs. Also, the percentage of SNPs unique to an individual is about the same for the two datasets. While it may seem counter-intuitive that such a comparatively small region would have a higher percentage of SNPs, it must be remembered that the mtCR is non-coding, meaning it does not translate into an RNA or protein and therefore lacks strong biological constraints to prevent nucleotides from mutating. The majority of the mtGenome excluding the mtCR codes for RNA or proteins with important biological functions, making the probability of a SNP occurring in one of those regions much lower (12). When SNPs do occur in a coding region, it is more likely that they are unique or possessed by only a small number of individuals, leading to more haplotypes with unique SNPs or unique combinations of SNPs within the mtGenome. This is seen in our dataset. Collectively, our results show that while there is more variability in the mtCR, the percentage of unique SNPs is relatively constant throughout the genome. Incorporation of SNPs outside of the mtCR increases the number of informative SNPs for forensic use to 57% of the total SNPs found.

Collectively, the 79 dogs in our dataset formed 10 groups and 47 unique haplotypes with 8 ambiguous sequences. The ambiguous base calls were due to true polymorphisms within the individual dog samples due to the multiple genomes per cell (2, 3). While the number of individuals with unique haplotypes may seem high, it is important to keep in mind that this is the first study of its kind, and the number will likely decrease as more dog mtGenomes are evaluated. Relative to the mtCR, this number will likely always to be higher due to larger region and higher constraints against mutation on the coding portions of the mtGenome.

As mentioned above, the number of individuals that share identical mtGenome sequences is smaller than the number of individuals that share mtCRs for the same dogs (Figures 2 and 3). This illustrates how the additional sequence variation of the mtGenome can be used to break up the large groups that often result from mtCR sequencing. Figure 2 shows how the dogs are situated relative to their haplotype. Figure 3 demonstrates the phenomenon that was seen in our larger mtCR study. While there are many canine mitochondrial control region haplotypes, most dogs share the common types while the minority of dogs have unique or fairly unique types. The distribution of the dogs within the mtGenome haplotype groups shows that the additional variation found in the remainder mtGenome breaks-up the large groups formed by mtCR sequences alone.

The distributions of dogs within each haplogroup were consistent with the mtCR groupings. As previously reported, when using only the mtCR sequence group A contained the most individuals followed by groups B, C and D (Webb and Allard, 08-027). When evaluating the mtGenome groups in the same manner, the same trend persists. Group A had the most individuals followed by B, C and D. When viewing the relationships of the dogs in the trees shown in Figure 1, it can be seen that not only do the sizes of the groups correspond between datasets, but also the members of each group. Dogs that grouped together based upon their

mtCR also grouped together based upon their mtGenome excluding the mtCR sequences. This result indicates that the phylogenetic signal present in the mtCR is also present in the remainder of the mtGenome. This result is expected as the mitochondrial genome does not undergo recombination and as such acts as a single locus. This is promising for forensic use of canine mitochondrial DNA. It shows that the entire mitochondrial genome can be used to identify samples because the results from different regions of the genome do not conflict.

The importance of the mutational "hot spots" within the mtGenome is that forensic samples are often degraded, making it difficult to obtain complete sequence through large areas. Also, the mtGenome is 92% larger than the mtCR. As a result, it is much more expensive to sequence. By identifying the most variable regions, we have provided coordinates where future groups can focus sequencing efforts; conversely, the regions where no SNPs were found could be avoided. These SNP free sections are all coding regions of the mtGenome; therefore, it is not surprising that the nucleotide composition of this region is conserved among the dogs in our dataset. All regions of increased or decreased SNP occurrence were identified via haplotype pairwise alignment to the Kim et al (17) reference sequence.

The random match probability results show that when considering the remainder of the mtGenome, there is a lower chance of a random match compared to using the mtCR alone. This is significant since it provides extra confidence that a match between a suspect dog and the sample found at a crime scene are truly the same individual and not just the result of the two randomly sharing mtGenome haplotype.

These results of the pairwise difference and nucleotide diversity assessments are consistent with the findings of the mtCR study. Though not statistically significant, they indicate that mixed breed dogs come from a more variable gene pool and, as expected, have more diversity in their sequence than purebred dogs. The ancestral lines of purebreds should contain only the DNA of individuals from the same breed or the founding breeds resulting in more constrained physical as well as genetic characteristics.

Since we never actually saw the dogs from which our samples were obtained, we were able to test the significance of the purebred versus mixed labels. Our results agree with the nucleotide diversity results which showed that there is not significant genetic variation between the group of dogs labeled "purebred" and those dogs labeled "mixed." This illustrates that not knowing whether a dog is purebred or mixed has very little consequence on the dataset in terms of mtDNA. Additionally, we show that geographic location of sample collection is not relevant when evaluating dogs from the United States via mtGenome haplotypes. Conversely, the fixation index becomes larger when dogs are grouped based on breed, which demonstrates that dogs of the same breed, while perhaps not possessing identical mtGenome sequences, have similar sequence composition than expected at random. These results support our previous mtCR dataset findings, which allows us to draw the same conclusions. First, classifying breeds by breed type (purebred or mixed) is trivial when it comes to mtDNA. Second, there is no need for local canine mitochondrial SNP databases. Finally, there is population substructure when dogs are grouped by breed. This is most likely due to the higher amounts of inbreeding of purebred dogs, which exemplifies that the need to collect multiple individuals of the same breed is necessary for a thorough mitochondrial SNP database.

Conclusions

Consistent with the mtCR results, analysis of the SNPs in the remainder of the mtGenome does not group dogs by breed or any other common domestic dog grouping. However, the SNPs

found in the remainder of the mtGenome are useful since they provide additional discriminatory sites that break up common mtCR haplotype groups. Within our dataset of 79 domestic dog mtGenomes excluding the mtCR, 2.3% of the nucleotides were found to be variable. Fifty-seven percent of the variable sites were informative by supporting groups of two or more dogs, and 26% of the informative sites were highly informative by supporting groups of eight or more dogs. When comparing haplotype groups formed from the mtCR sequences alone and the mtGenome sequences without the mtCR for the same set of 79 dogs, it becomes obvious that the SNPs found in the remainder of the mtGenome have a higher discriminatory power. When looking at the mtCR alone, there are 18 individuals (25.7%) with unique mtCR sequences and 52 dogs (74.3%) forming 14 groups with up to 7 dogs per group. Comparatively, when looking at the same 79 dogs using mtGenome sequences without the mtCR, the distribution shifts with 24 dogs (33.8%) forming 10 groups containing at most 3 dogs and the remaining 67.6% (n = 48) of the dogs having unique haplotypes. Using AMOVA, the current dataset shows that there is little need to be concerned with whether a dog is classified as purebred or mixed or knowing the geographic location within the United States from which a sample was obtained. We do see evidence that it is necessary to collect multiple individuals of the same breed for a comprehensive mitochondrial SNP database. This is the first study to report SNP variation outside of the mtCR for the domestic dog. Our data demonstrate the usefulness of the entire mtGenome for forensic use in identifying domestic dog samples.

Acknowledgements

We thank Sheri Church and Mark Wilson and two anonymous reviewers for careful review of this manuscript. The research was conducted at The George Washington University. The 698 new domestic dog samples were collected through donations from veterinary practices and private donors. We would like to thank Adobe Animal Hospital, Austin Vet Hospital, Caring Hands Animal Hospital, Del Paso Vet Clinic, Little River Vet Clinic, The College of Veterinary Medicine at Mississippi State University, Pet Medical Center of Las Vegas, Seneca Hill Animal Hospital, The Animal Clinic of Clifton, West Flamingo Animal Hospital for blood and tissue samples and 80 private donors who provided buccal swabs to add to our collection. This research was funded by the National Institute of Justice through grant 2004-DN-BX-K004 to M. W. Allard.

Table 1 – Genome Primers

Primer Name	Primer 5'-3'	5’ Coordinate	3’ Coordinate
1620F (PCR1)	TGTTGAGCTGGAACGCTTTC	1639	1620
549F	GCTAGTAGTCCTCTGGCGAA	574	549
84F	GGTTTGCTGAAGATGGCG	701	684
1191F	GGTACTATCTCTATCGCTCC	1210	1191
16625R (PCR1)	CGCATTTGGTCTCGTAGTCT	16625	16644

171R	GGAGCAGGTATCAAGCACAC	171	190
556R	GAGGACTACTAGCAATAGCT	556	575
997R	CATACCGGAAGGTGTGCTT	997	1015
2978F (PCR2)	GTTAGGGCTAGTGATAGAGC	2997	2978
1770F	GTGGTCTATCCGTTCCTGAT	1789	1770
2400F	GGTCGTAAACCCTATTGTCG	2419	2400
1418R (PCR2)	AAGCCTAACGAGCCTGGTG	1418	1436
1999R	CGGTATCCTGACCGTGCAA	1999	2017
2512R	GGAGTAATCCAGGTCGGTTT	2512	2531
2556R	GTACGAAAGGACAAGGGATG	2556	2575
4411F (PCR3)	GTTTGATTTAGTCCGCCTCAG	4431	4411
3220F	GCGTGGATAGTGTAAATGAC	3239	3220
3804F	GGTAGCACGAAGATCTTTGA	3823	3804
3945F	GGTTCCTGTCATGATAGTTG	3964	3945
2881R (PCR3)	CCTTCAACCAATCGCAGACG	2881	2900
3479R	GCATTCCACAACCCATTCAT	3479	3498
3645R	TATGCATATGACATGTTGCC	3645	3664
4188R	CCATCGCATCCATCATGATA	4188	4207
5949F (PCR4)	GTAATTCCAGCAGCCAGTAC	5968	5949
4939F	CCTAGTCCAAGACTGATAGT	4958	4939
5407F	GGCTCATGCTCCAAATAGTA	5426	5407
5583F	GGAAACTGACTAGTGCCGTT	5602	5583
6118F	CCTGAGTAGTAAGTGACAA	6136	6118
4241R (PCR4)	CCATTCCACTTCTGAGTTCC	4241	4260
4188R	CCATCGCATCCATCATGATA	4188	4207
4274R	GGAATTACGCTCATATCAGG	4274	4293
4792R	CCTGCGACTCACATATAGCA	4792	4811
4793R	CTGCGACTCACATATAGCAC	4793	4812
5481R	GGTACTTTACTAGGTGACGA	5481	5500
7642F (PCR5)	CAATGGGTATAAAGCTGTGG	7661	7642
6352F	AAGCTCATAGCATAGCTGG	6372	6352
6415F	GGACGAATTAGCTAGGACAA	6434	6415
Primer Name	Primer 5'-3'	5’Coordinate	3’Coordinate
7035F	GAGTTGAAATGGGTACGCCA	7054	7035
5871R (PCR5)	GCAATATCCCAGTATCAAACT	5871	5891
6044R	ACACCTATTCTGATTCTTCG	6044	6063
6212R	AGCTCACCATATGTTTACCG	6212	6231
6352R	CTCCAGCTATGCTATGAGCT	6352	6371
7032R	CTATGGCGTACCCATTTCAA	7032	7054

9264F (PCR6)	GAATGTAGAGCCAATAATTACG	9285	9264
8015F	CGATCAGTACCACAATAGG	8033	8015
8152F	GAGCTCAGGTTCGTCCCTTT	8171	8152
8825F	GAATGTGCCTTCTCGGATCA	8844	8825
7512R (PCR6)	TGCATTCATGAGCCGTTCC	7512	7530
7804R	TGCCACAGCTAGATACATCC	7804	7823
8084R	CGGTTAATCTCCATTCAGCA	8084	8103
8681R	CAAGCCCATGACCGCTGACA	8681	8700
11021F (PCR7)	CTGTTTGACGGAGACAGATAG	11041	11021
9722F	TTGGTTTGTGACGCTCAGG	9740	9722
9994F	CCTCTAAGCATAGTAGCGAT	10013	9994
10625F	GTAGAGTCCTGCGTTTAGTC	10644	10625
9190R (PCR7)	GAGACATCTTTTACAATCTCCG	9190	9211
9628R	GGATCTGCTCGCCTACCTT	9628	9646
9785R	TCCTAGCTGCGAGCCTAG	9785	9802
10278R	CACGACAACATATGGTTTGC	10278	10297
10565R	TTGAAGCAACACTGATTCCG	10565	10584
12543F (PCR8)	GCGGATAAGAAGAAATACTCC	12563	12543
11508F	GCAGTAGGTGCAAGGTCATT	11527	11508
12062F	CTATGATAGACCACGTGACA	12081	12062
10844R (PCR8)	GACTACCAAAAGCACACGTAG	10844	10864
10886R	TAGTACTTGCCGCTGTACTCC	10886	10906
11270R	CCTGATGACTATTAGCAAGC	11270	11289
11892R	GCTACTTCTTACGCGTTCAT	11892	11911
11945R	CTCAGGACAGGAAACAATCA	11945	11964
13799F (PCR9)	GTTGTCTGAATTGTTGACTGC	13819	13799
12723F	GGCTGGTTAATGCCAATTGT	12742	12723
12730F	TAAGTAGGGCTGGTTAATGC	12749	12730
13268F	GTTCTAGTGCCAGGATGAAA	13287	13268
13565F	TAAGGATTAGTAGACTGAGG	13584	13565
12234R (PCR9)	CTACTTATTGGATGATGGTACG	12234	12255
12415R	TACTTGGCCTACTACTAGC	12415	12433
Primer Name	Primer 5'-3'	5’ Coordinate	3’ Coordinate
12525R	AGCACAATAGTTGTAGCAGG	12525	12544
12759R	CACATCTGCACTCACGCATT	12759	12778
13206R	ATCCCACAGATAACTATGCC	13206	13225
13352R	CCTTGGCTACTATCCAACCA	13352	13371
14810F (PCR10)	GTCTGAGTCTGATGTGATTCC	14830	14810
14030F	GCCACTAAACCATCTCCTAT	14049	14030

14253F	TCAAGCAGAGATGTTAGACG	14272	14253
14390F	CGTAGTTAACGTCTCGGCA	14408	14390
13622R (PCR10)	ATTAATAATGATCAGCCTGTAAC	13622	13644
13973R	TTCAGAACAATCGCACAACC	13973	13992
14267R	GCTTGATGGAACTTCGGATC	14267	14286
15513F (PCR11)	GAGGGGAGAAGGGTTTACC	15531	15513
14933F	TGTAGTTATCTGGGTCTCC	14951	14933
15012F	GGATCGTAGGATAGCATAGG	15031	15012
14696R (PCR11)	AAAGCAACCCTAACACGATTC	14696	14716
14933R	GGAGACCCAGATAACTACT	14933	14951
15233R	GGACAAGTCGCTTCAATCTT	15233	15252

Table 1 – List of all primers (PCR and sequencing) used to sequence the canine mtGenome excluding the mtCR. The primer name, sequence (5'– 3' orientation), start coordinate and stop coordinate are listed.

Table 2 – List of dogs used in current study

Accession Number	Source	Breed
DQ480493	Bjornerfeldt et al., 2006	Black Russian Terrier
DQ480495	Bjornerfeldt et al., 2006	Cocker Spaniel
DQ480490	Bjornerfeldt et al., 2006	Flat Coated Retriever
DQ480489	Bjornerfeldt et al., 2006	German Shepherd
DQ480491	Bjornerfeldt et al., 2006	Irish Setter
DQ480496	Bjornerfeldt et al., 2006	Irish Soft Coated Wheaten Terrier
DQ480492	Bjornerfeldt et al., 2006	Jamthund
DQ480502	Bjornerfeldt et al., 2006	Jamthund
DQ480498	Bjornerfeldt et al., 2006	Miniature Schnauzer
DQ480494	Bjornerfeldt et al., 2006	Poodle
DQ480500	Bjornerfeldt et al., 2006	Shetland Sheepdog
DQ480499	Bjornerfeldt et al., 2006	Siberian Husky
DQ480501	Bjornerfeldt et al., 2006	Swedish Elkhound
DQ480497	Bjornerfeldt et al., 2006	West Highland White Terrier
NC_002008	Kim et al., 1998	Sapsaree
EU408245	Webb and Allard, 2008	Akita1P
EU408246	Webb and Allard, 2008	American CockerSpaniel1P
EU408248	Webb and Allard, 2008	Australian Shepherd1P
EU408249	Webb and Allard, 2008	Australian Shepherd7P
EU408247	Webb and Allard, 2008	Australian Terrier1P
EU408254	Webb and Allard, 2008	Basset Hound2P
EU408255	Webb and Allard, 2008	Basset Hound3P
EU408256	Webb and Allard, 2008	Basset Hound4P
EU408250	Webb and Allard, 2008	Bichon Frise3P
EU408251	Webb and Allard, 2008	Blue Heeler1P
EU408252	Webb and Allard, 2008	Bolognese1P
EU408253	Webb and Allard, 2008	Boxer6P
EU408257	Webb and Allard, 2008	Brittany Spaniel1M
EU408264	Webb and Allard, 2008	Cairn Terrier4P
EU408260	Webb and Allard, 2008	Cardigan Corgi2P
EU408263	Webb and Allard, 2008	CavalierKingCharlesSpaniel9P
EU408262	Webb and Allard, 2008	Chihuahua5P
EU408261	Webb and Allard, 2008	Chihuahua11M
EU408258	Webb and Allard, 2008	Cockapoo1M
EU408259	Webb and Allard, 2008	Cockapoo3M
EU408266	Webb and Allard, 2008	Cocker Spaniel1P
EU408267	Webb and Allard, 2008	Cocker Spaniel3P
EU408268	Webb and Allard, 2008	Cocker Spaniel8P
EU408265	Webb and Allard, 2008	Corgi2P
EU408270	Webb and Allard, 2008	Dachshund4P

EU408272	Webb and Allard, 2008	Dachshund15P
EU408269	Webb and Allard, 2008	Doberman Pinscher5P
EU408271	Webb and Allard, 2008	Dogue de Bordeaux1P
EU408274	Webb and Allard, 2008	English Mastiff3P
EU408273	Webb and Allard, 2008	English Shepherd1M
EU408275	Webb and Allard, 2008	French Bulldog1P
EU408277	Webb and Allard, 2008	German Shepherd12P
EU408276	Webb and Allard, 2008	Great Dane2P
EU408278	Webb and Allard, 2008	Great Pyrenese1P
EU408279	Webb and Allard, 2008	Havanese3P
EU408280	Webb and Allard, 2008	Italian Greyhound
EU408281	Webb and Allard, 2008	Jack Russell6P
EU408282	Webb and Allard, 2008	Keeshond1P
EU408283	Webb and Allard, 2008	Keeshond2P
EU408284	Webb and Allard, 2008	Keeshond3P
EU408285	Webb and Allard, 2008	Labradoodle1P
EU408286	Webb and Allard, 2008	Miniature Dachshund2P
EU408289	Webb and Allard, 2008	Neapolitan Mastiff1P
EU408290	Webb and Allard, 2008	Neapolitan Mastiff2P
EU408287	Webb and Allard, 2008	Newfoundland1P
EU408288	Webb and Allard, 2008	Norwegian Elk Hound1P
EU408293	Webb and Allard, 2008	Pit Bull1M
EU408291	Webb and Allard, 2008	Pomerian2M
EU408292	Webb and Allard, 2008	Poodle7M
EU408294	Webb and Allard, 2008	Pug5P
EU408295	Webb and Allard, 2008	Rottweiler1P
EU408296	Webb and Allard, 2008	Rottweiler2P
EU408297	Webb and Allard, 2008	Schipperke1P
EU408299	Webb and Allard, 2008	Schnauzer4P
EU408298	Webb and Allard, 2008	Sheltie1M
EU408300	Webb and Allard, 2008	Tibetan Mastiff1P
EU408301	Webb and Allard, 2008	Tibetan Spaniel1P
EU408302	Webb and Allard, 2008	Toy Poodle3P
EU408304	Webb and Allard, 2008	Unknown1P
EU408303	Webb and Allard, 2008	Unknown1M
EU408305	Webb and Allard, 2008	Vizsla2P
EU408307	Webb and Allard, 2008	Walker Hound1P
EU408306	Webb and Allard, 2008	West Highland Terrier4P
EU408308	Webb and Allard, 2008	Yorkie/Chihuahua1M

Table 2 – List of Genbank accession number, reference and breed of each sequence used in the current study. All dogs from (16, 17) are purebred and all "Webb and Allard" dogs are either denoted "P" or "M" indicating purebred or mixed.

Table 3 - Informative Sites in the mtGenome excluding mtCR

Base	Reference	Sample	L	Ri	Base	Reference	Sample	L	Ri	Base	Reference	Sample	L	Ri	Base	Reference	Sample	L	Ri	Base	Reference	Sample	L	Ri
16	T	C	1	100	4303	A	G	2	85	8242	G	A	1	100	10776	T	C	2	83	13762	T	C	2	0
162	T	C	1	100	4360	T	C	1	100	8281	T	C	1	100	10785	A	G	2	83	13777	G	A	1	100
381	T	A	1	100	4390	T	C	1	100	8323	A	G	1	100	10863	A	G	1	100	13791	T	C	1	100
445	A	G	1	100	4466	G	A	2	66	8390	G	A	1	100	10917	G	A	1	100	14474	G	A	1	100
463	T	C	1	100	4484	G	A	1	100	8425	G	A	1	100	10992	G	A	1	100	14543	T	C	1	100
557	A	G	1	100	4503	A	G	1	100	8536	C	T	1	100	11172	A	G	1	100	14608	A	G	2	90
658	A	G	1	100	4517	G	A	1	100	8569	A	G	1	100	11176	C	T	1	100	14647	T	C	2	90
1046	G	A	1	100	4572	T	C	1	100	8670	C	T	1	100	11247	A	G	1	100	14671	G	A	1	100
1204	T	C	1	100	4591	G	A	1	100	8703	G	A	1	100	11250	T	C	1	100	14692	G	A	1	100
1351	A	G	1	100	4595	C	T	1	100	8736	T	C	1	100	11322	T	C	1	100	14800	C	T	1	100
1454	G	A	2	96	4646	T	C	1	100	8760	A	G	1	100	11400	T	C	1	100	14806	T	C	1	100
1522	G	A	2	0	4940	T	C	1	100	8764	G	T	1	100	11402	T	C	1	100	14930	T	C	1	100
1662	C	T	1	100	5009	C	T	1	100	8782	T	C	1	100	11572	A	C	1	100	14977	T	C	2	91
1689	C	T	1	100	5367	C	T	1	100	8817	A	G	1	100	11625	A	G	1	100	15185	T	C	1	100
1709	G	A	1	100	5519	C	T	1	100	8853	T	C	1	100	11657	C	A	1	100	15214	G	A	1	100
1748	T	C	1	100	5624	G	A	1	100	8877	A	G	1	100	11800	T	C	1	100	15287	G	A	1	100
1756	C	T	1	100	5855	C	T	1	100	8970	T	C	1	100	11813	A	G	1	100	15372	G	A	1	100
1766	T	C	1	100	5937	C	T	1	100	8991	A	G	1	100	11839	T	C	1	100	15435	G	A	1	100
1873	A	G	1	100	6053	C	T	1	100	9219	A	G	1	100	11897	T	C	1	100					
2185	T	C	1	100	6092	G	A	1	100	9222	C	T	1	100	11948	A	G	1	100					
2232	A	G	2	96	6257	G	A	1	100	9252	T	C	1	100	11959	C	T	1	100					
2656	G	A	1	100	6302	G	A	1	100	9708	C	T	2	83	11963	C	T	1	100					
2683	G	A	1	100	6401	C	T	1	100	9825	G	A	1	100	11984	A	G	1	100					
2812	C	T	1	100	6470	G	A	1	100	9835	A	G	1	100	12063	G	A	1	100					
2833	C	T	1	100	6518	G	A	1	100	9838	G	A	1	100	12122	C	T	1	100					
2854	A	G	1	100	6554	T	C	2	91	9865.1	-	A	3	71	12200	C	T	1	100					
2962	C	T	1	100	6629	T	C	1	100	9886	G	A	1	100	12260	A	G	1	100					
3028	A	C	1	100	6711	T	A	1	100	9896	T	C	1	100	12272	T	C	1	100					
3034	T	C	1	100	6740	G	A	1	100	10060	C	T	1	100	12330	A	G	1	100					
3196	T	C	1	100	6764	C	T	1	100	10159	C	T	1	100	12346	T	A	1	100					
3388	G	A	2	0	6767	G	A	1	100	10165	C	T	1	100	12401	T	C	1	100					
3406	C	T	2	96	6860	G	A	1	100	10195	T	C	1	100	12459	G	A	1	100					
3451	C	T	1	100	6863	C	T	1	100	10257	G	A	1	100	12636	T	C	1	100					
3465	T	C	1	100	6881	G	A	1	100	10311	C	T	1	100	12665	T	C	2	95					
3469	G	A	1	100	6967	A	G	1	100	10319	T	C	1	100	12788	T	C	1	100					
3494	T	C	1	100	7014	T	C	1	100	10346	C	T	2	75	12813	G	A	1	100					
3598	G	A	1	100	7058	T	C	1	100	10404	C	T	2	96	12818	C	T	1	100					
3628	A	G	1	100	7171	G	A	1	100	10440	T	C	1	100	12968	G	A	1	100					
3937	C	T	1	100	7186	C	A	1	100	10533	A	T	1	100	13102	T	C	1	100					
3940	C	T	1	100	7450	C	T	1	100	10542	A	G	1	100	13112	G	A	1	100					
3950	A	G	1	100	7593	T	C	1	100	10557	C	T	1	100	13261	C	T	1	100					
4135	C	T	1	100	7923	T	C	1	100	10611	A	T	2	91	13426	C	T	1	100					
4169	A	G	1	100	8101	G	A	1	100	10613	A	G	1	100	13594	G	A	2	95					
4204	G	A	1	100	8108	C	T	1	100	10680	C	T	1	100	13618	A	G	1	100					
4234	C	T	1	100	8221	A	C	1	100	10725	T	C	1	100	13660	C	T	1	100					
4277	A	G	1	100	8225	T	C	1	100	10773	T	C	1	100	13708	C	T	1	100					

Table 3 – Informative sites for the canine mtGenome excluding the mtCR. The nucleotide coordinate relative to the Kim et al. (17), the reference sequence base (17), the observed base, the character length (L) and character retention index (ri) are listed. Those coordinates shaded grey support groups of 8 or more dogs, making them the most informative SNPs found in the current dataset.

Table 5 – Distribution of Haplotypes

Distribution of Haplotypes					
Haplotype	mtCR Haplotype	Breed	(n) per breed	Total (n)	%
mtGenomeA2a.1	A2a	WestHighlandWhiteTerr(DQ480497)	1	1	1.27
mtGenomeA2b.1	A2b	GreatDane2P	1	2	2.53
	A2b	Schnauzer4P	1		
mtGenomeA2b.2	A2b	FrenchBulldog1P	1	1	1.27
mtGenomeA11e.1	A11e	Rottweiler1P	2	2	2.53
	A11e	Rottweiler2P			
mtGenomeA11e.2	A11e	MiniatureDachshund3P	1	1	1.27
mtGenomeA11e.3	A11e	AustralianShepherd7P	1	1	1.27
AmbigmtGenomeA11Ambig2.1	A11Ambig2	CockerSpaniel1P	1	1	1.27
mtGenomeA16a.1	A16a	BrittanySpaniel1M	1	1	1.27
mtGenomeA16a.2	A16a	ItalianGreyhound1P	1	1	1.27
mtGenomeA16a.3	A16a	EnglishMastiff3P	1	1	1.27
mtGenomeA17a.1	A17a	Boxer6P	1	3	3.80
	A17a	DoguedeBordeaux1P	1		
	A17a	MiniatureSchnauzer (D480498)	1		
mtGenomeA17a.2	A17a	Unknown1P	1	1	1.27
mtGenomeA17a.3	A17a	CavalierKingCharlesSpaniel9P	1	1	1.27
mtGenomeA17a.4	A17a	BichonFrise3P	1	1	1.27
AmbigmtGenomeA17a.1	A17a	Pug5P	1	1	1.27
mtGenomeA18b.1	A18b	AmericanCockerSpaniel1P	1	1	1.27
mtGenomeA18d.1	A18d	JackRussell6P	1	2	2.53
	A18d	Sheltie1M	1		
mtGenomeA18d.2	A18d	Dachshund15P	1	1	1.27
mtGenomeA18d.3	A18d	Vizsla2P	1	1	1.27
mtGenomeA18d.4	A18d	CockerSpaniel (DQ480495)	1	1	1.27
AmbigmtGenomeA18d.1	A18d	Cockapoo3M	1	1	1.27
AmbigmtGenomeA18d.2	A18d	ToyPoodle3P	1	1	1.27
mtGenomeA19a.1	A19a	Dachshund4P	1	2	2.53
	A19a	GermanShepherd12P	1		
mtGenomeA19a.2	A19a	Sapsaree(NC_002008)	1	1	1.27
mtGenomeA19a.2	A19a	AustralianShepherd1P	1	1	1.27
AmbigmtGenomeA20b.1	A20b	EnglishShepherd1M	1	1	1.27
mtGenomeA20c.1	A20c	Chihuahua11M	1	1	1.27
mtGenomeA22a.1	A22a	NeopolitanMastiff1P	1	1	1.27
mtGenomeA22a.2	A22a	NeopolitanMastiff2P	1	1	1.27
mtGenomeA26a.1	A26a	WestHighlandTerrier4P	1	3	3.80
	A26a	CairnTerrier4P	1		
	A26a	IrishSoftCoatedWT(DQ480496)	1		
mtGenomeA26a.2	A26a	NewFoundland1P	1	1	1.27
mtGenomeA27c.1	A27c	Keeshond1P	3	3	3.80
	A27c	Keeshond2P			
	A27c	Keeshond3P			
mtGenomeA29b.1	A29b*	SiberianHusky(DQ480499)	1	1	1.27

mtGenomeA71.1	A71	Corgi2P	1	1	1.27
mtGenomeA71.2	A71	Akita1P	1	1	1.27
AmbigmtGenomeA97.1	A97	TibetanMastiff1P	1	1	1.27
mtGenomeA98.1	A98	Chihuahua5P	1	1	1.27
mtGenomeA108.1	A108*	IrishSetter(DQ480491)	1	1	1.27
mtGenomeBAmbig4.1	BAmbig4	DobermanPinscher5P	1	1	1.27
mtGenomeBAmbig11.1	BAmbig11	Unknown1M	1	1	1.27
mtGenomeBAmbig12.1	BAmbig12	Yorkie/Chihuahua1M	1	1	1.27
mtGenomeB1Ambig1.1	B1Ambig1	AustralianTerrier1P	1	1	1.27
mtGenomeB1Ambig4.1	B1Ambig4	CardiganCorgi2P	1	1	1.27
mtGenomeB1a.1	B1a	Labradoodle1P	1	3	3.80
mtGenomeB1g.1	B1g*	ShetlandSheepdog(DQ480500)	1		
mtGenomeB1h.1	B1h*	Poodle(DQ480494)	1		
mtGenomeB1a.2	B1a	BassetHound4P	2	2	2.53
mtGenomeB1Ambig4.1	B1Ambig4	BassetHound2P			
mtGenomeB1a.3	B1a	TibetanSpaniel1P	1	1	1.27
mtGenomeB1a.4	B1a	Bolognese1P	1	1	1.27
mtGenomeB1a.5	B1a	Poodle7M	1	1	1.27
mtGenomeB1a.6	B1a	GreatPyrenese1P	1	1	1.27
AmbigmtGenomeB1a.1	B1a	BassetHound3P	1	1	1.27
mtGenomeB6a.1	B6a	WalkerHound1P	1	1	1.27
mtGenomeB6a.2	B6a	Schipperke1P	1	1	1.27
mtGenomeB10a.1	B10a	CockerSpaniel8P	1	1	1.27
mtGenomeB28.1	B28	Cockapoo1M	1	1	1.27
mtGenomeB30.1	B30*	FlatCoatedRet(DQ480490)	1	1	1.27
AmbigmtGenomeCAmbig1.1	CAmbig1	BlueHeeler1P	1	1	1.27
mtGenomeC3Ambig1.1	C3Ambig1	CockerSpaniel3P	1	1	1.27
mtGenomeC3a.1	C3a	Pomerian2M	1	1	1.27
mtGenomeC3a.2	C3a	Havanese3P	1	1	1.27
mtGenomeC3b.1	C3b*	BlackRussianTerrier(DQ480493)	1	1	1.27
mtGenomeC3b.2	C3b*	SwedishElkhound(DQ480501)	1	1	1.27
mtGenomeC8a.1	C8a	PitBull1M	1	1	1.27
mtGenomeC12.1	C12*	GermanShep(DQ480489)	1	1	1.27
mtGenomeD1a.1	D1a	NorweigianElkHound1P	1	1	1.27
mtGenomeD1b.1	D1b*	Jamthund(DQ480502)	2	2	2.53
mtGenomeD2.1	D2*	Jamthund(DQ480492)			

Table 5 – The distribution of all individuals in the dataset. Haplotype name, mtCR haplotype, sample id, number of individuals per breed sharing the haplotype, total number of individuals sharing the haplotype and frequency of haplotype are listed. Samples with mtCR haplotypes marked with an asterisk (*) are from Bjornerfeldt et al. (16) and had mtCR haplotypes not present in our previous study. The haplotype names correspond to the haplotypes listed in Table 4.

Table 6 – AMOVA analysis within and breed populations

Dataset	Source of Variation	Degrees of Freedom	Percentage of Variation
Purebred vs Mixed	Among populations	1	0.20
	Within populations	77	99.80
	Total	78	100
		Fst = 0.00198	
By State*	Among populations	5	0 (-4.72)
	Within populations	58	104.72
	Total	63*	100
		Fst = 0 (-0.04720)	
By Breed	Among populations	10	25.29
	Within populations	68	74.71
	Total	78	100
		Fst = 0.25288	

Grouping and results of AMOVA analysis as preformed in Arlequin to assess population structure between purebred and mixed breed dogs, dogs grouped by geographic state of origin and large breed groups of purebred dogs. *The "By State" calculations do not include the Kim et al (17) reference sequence or the 14 samples from Bjornerfeldt et al. (16).

Figure 1 – mtCR and mtGenome Phylogenetic Trees

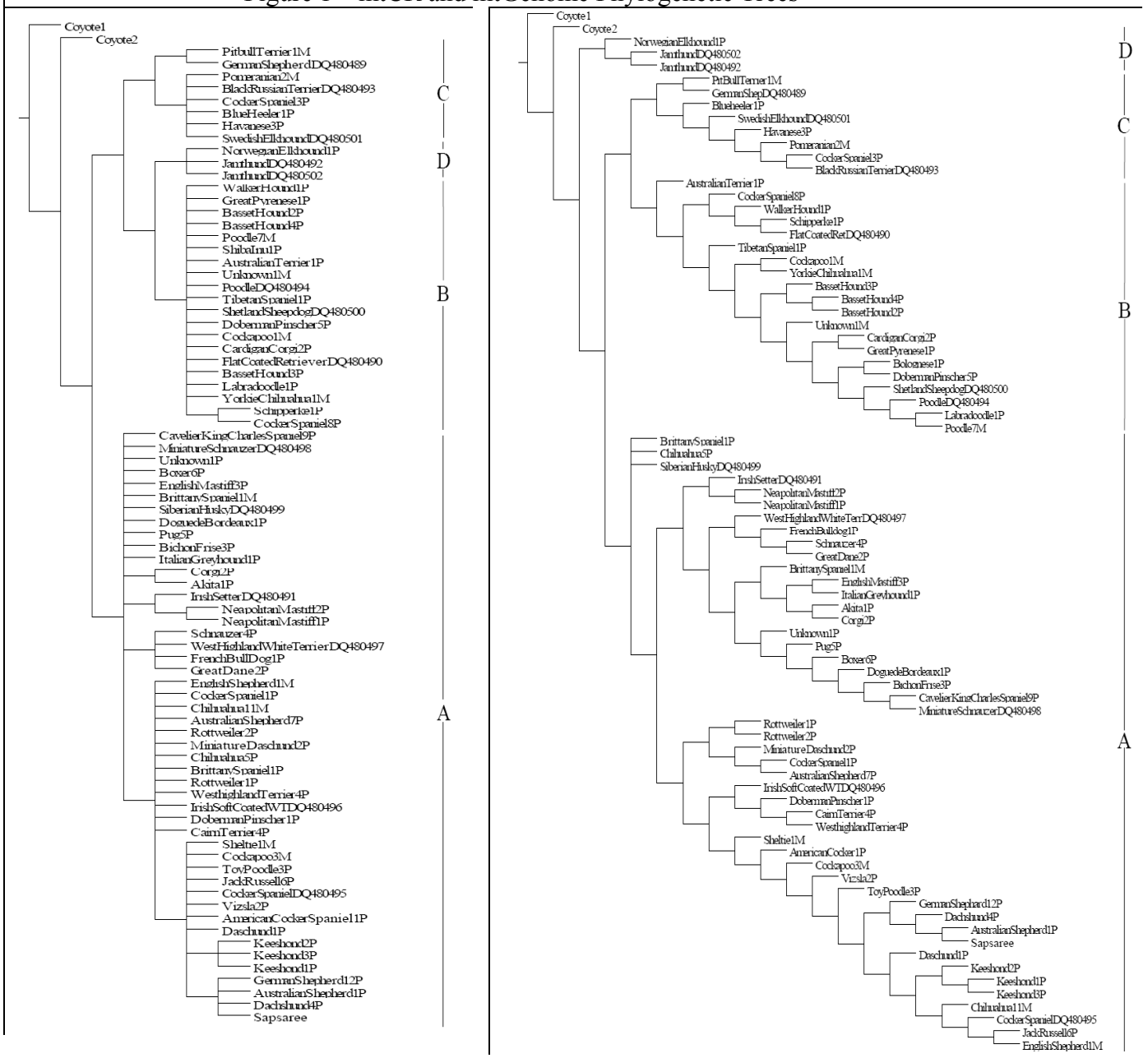


Figure 1 – Phylogenetic trees of the 79 dogs using only their mtCR sequences (left) and only their mtGenome sequences (right). The letters "A", "B", "C", and "D" represent the previously identified major haplogroup labels. While the relationships of the major haplogroups changes, and the order of the dogs within the groups changes, close inspection of each major group will show that the same dogs fall within the same groups regardless of the region of DNA sequence being used.

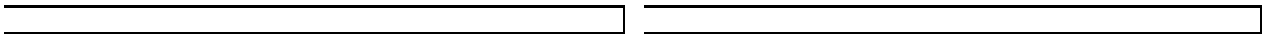


Figure 2 – Distribution of Haplotypes

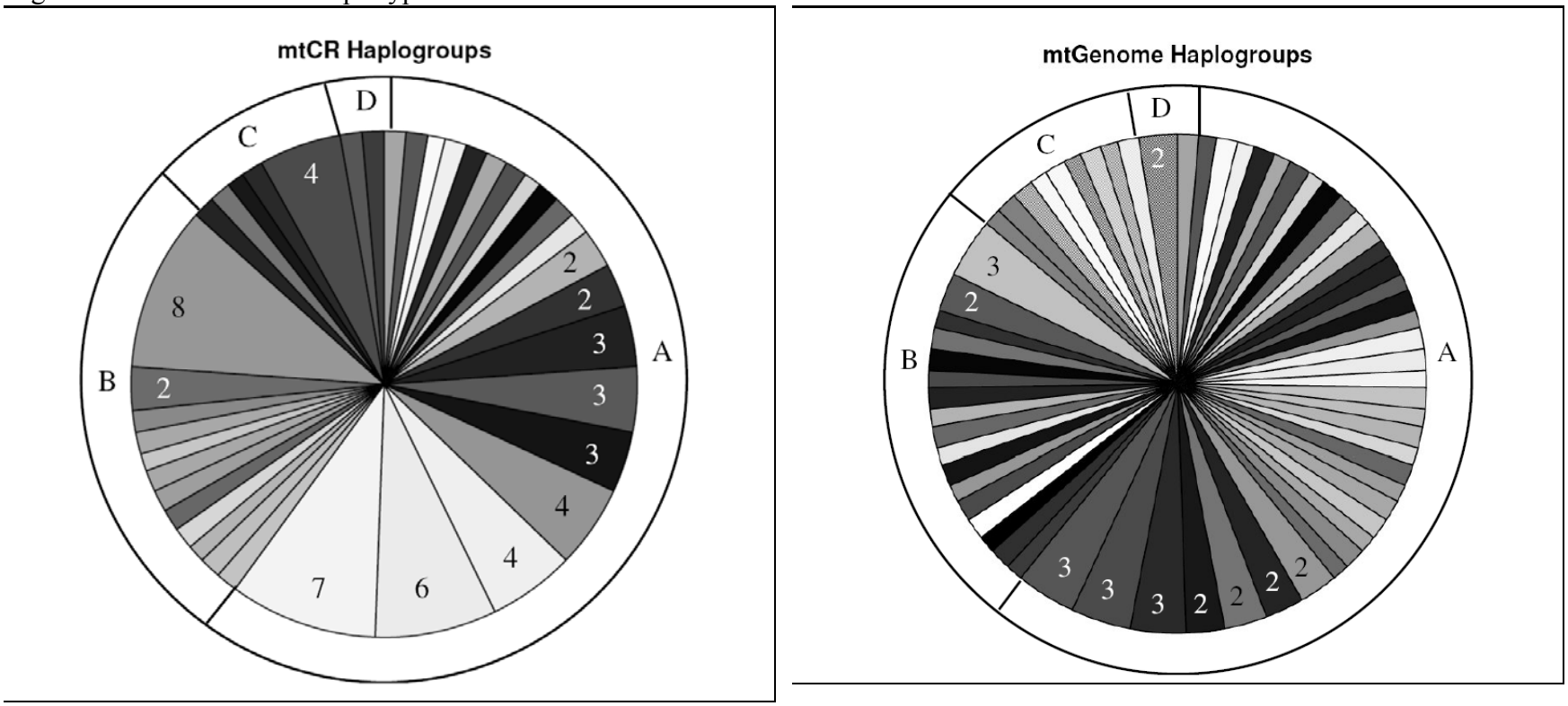


Figure 2 – Pie charts showing how haplotypes are distributed based on previously assigned haplotype names. Regardless of mtCR or mtGenome sequence, the trend of haplogroup A containing the most dogs followed by haplogroups B, C and then D persists.

Figure 3 - Distribution of Haplotypes Based on Group Size.

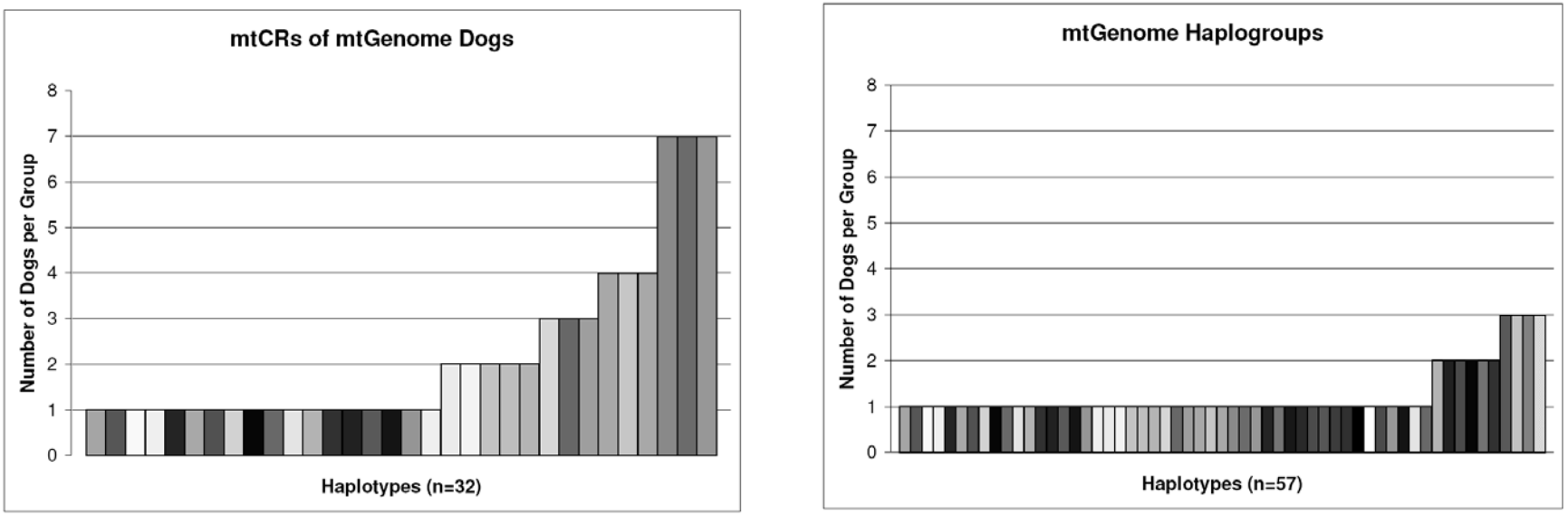


Figure 3 – A comparison of mtCR and mtGenome haplotype groups showing how the mtGenomes have more individuals with unique haplotypes and fewer groups or two or more identical samples for the same 79 dogs. Dogs from each dataset with ambiguous base calls were not included (mtCR, n = 9), (mtGenome, n = 8)

Figure 4 – Frequency of SNPs

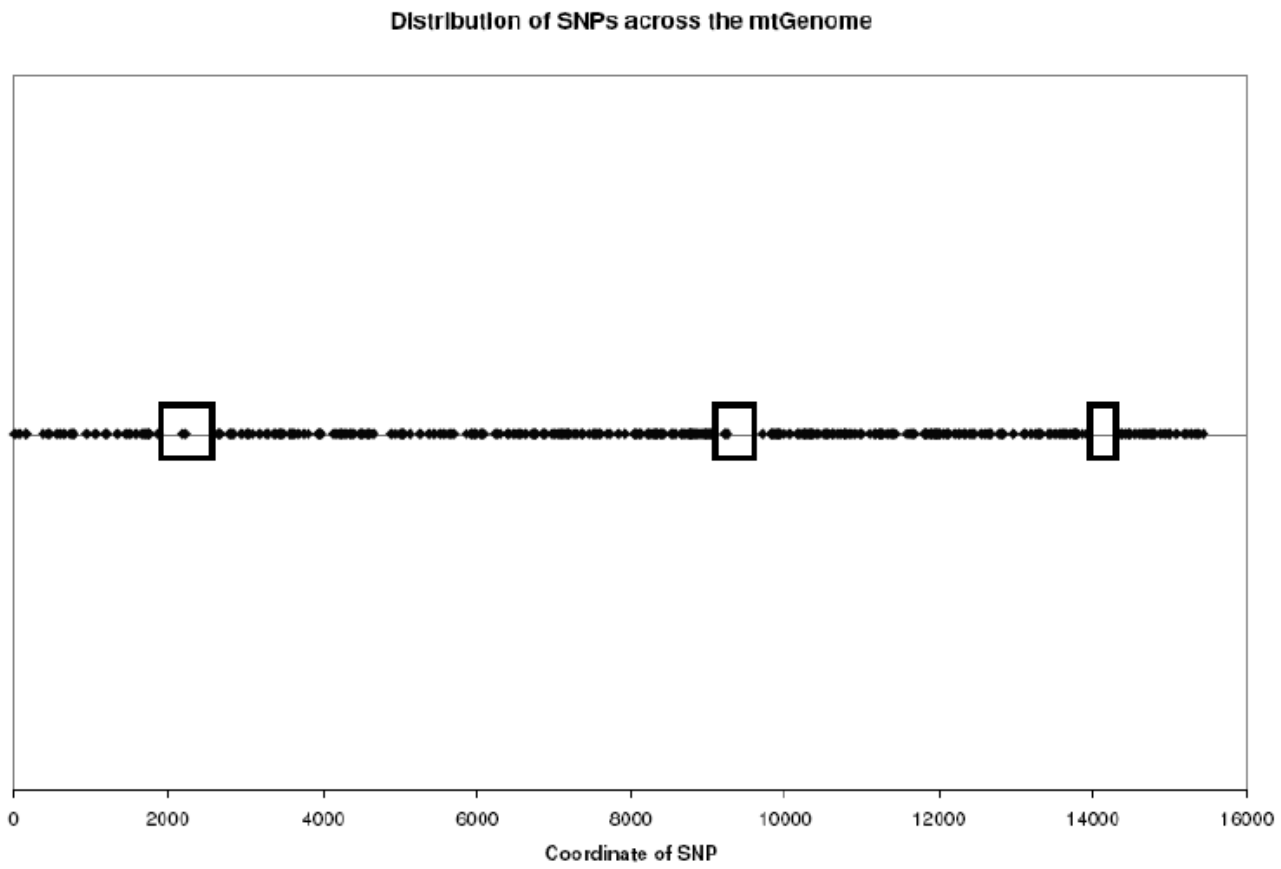


Figure 4 – Graph showing the distribution of SNPs across the mtGenome. The y-axis is labeled with coordinates relative to the Kim et al. (17) reference sequence. The boxes highlight regions with few or zero SNPs.

Vb REFERENCES

1. Vigilant L. An evaluation of techniques for the extraction and amplification of DNA from naturally shed hairs. Biol Chem. 1999 Nov;380(11):1329-31.

2. Bogenhagen D, Clayton D. The number of mitochondrial deoxyribonucleic acide genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial dexoyribonucleic acid. Journal of Biol Chem. 1974;249:7991-5.

3. Nass M. Mitochondrial DNA. I. Intramitochondrial distribution and structural relations of single- and double-length circular DNA. Journal of Molecular Biology. 1969;42:521 8.

4. Parsons TJ, Coble MD. Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. Croat Med J. 2001 Jun;42(3):304-9.

5. Angleby H, Savolainen P. Forensic informativity of domestic dog mtDNA control region sequences. Forensic Sci Int. 2005 Nov 25;154(2-3):99-110.

6. Budowle B, Wilson MR, DiZinno JA, Stauffer C, Fasano MA, Holland MM, et al. Mitochondrial DNA regions HVI and HVII population data. Forensic Sci Int. 1999 Jul 12;103(1):23-35.

7. Gundry RL, Allard MW, Moretti TR, Honeycutt RL, Wilson MR, Monson KL, et al. Mitochondrial DNA analysis of the domestic dog: control region variation within and among breeds. J Forensic Sci. 2007 May;52(3):562-72.

8. Okumura N, Ishiguro N, Nakano M, Matsui A, Sahara M. Intra- and interbreed genetic variations of mitochondrial DNA major non-coding regions in Japanese native dog breeds (Canis familiaris). Anim Genet. 1996 Dec;27(6):397-405.

9. Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, et al. A high observed substitution rate in the human mitochondrial DNA control region. Nat Genet. 1997 Apr;15(4):363-8.

10. Savolainen P, Lundeberg J. Forensic evidence based on mtDNA from dog and wolf hairs. J Forensic Sci. 1999 Jan;44(1):77-81.

11. Savolainen P, Rosen B, Holmberg A, Leitner T, Uhlen M, Lundeberg J. Sequence analysis of domestic dog mitochondrial DNA for forensic use. J Forensic Sci. 1997 Jul;42(4):593-600.

12. Pesole G, Gissi C, De Chirico A, Saccone C. Nucleotide substitution rate of mammalian mitochondrial genomes. J Mol Evol. 1999 Apr;48(4):427-34.

13. Halverson J, Basten C. A PCR multiplex and database for forensic DNA identification of dogs. J Forensic Sci. 2005 Mar;50(2):352-63.

14. Pereira L, Van Asch B, Amorim A. Standardisation of nomenclature for dog mtDNA D-loop: a prerequisite for launching a Canis familiaris database. Forensic Sci Int. 2004 May 10;141(2-3):99-108.

15. Wetton JH, Higgs JE, Spriggs AC, Roney CA, Tsang CS, Foster AP. Mitochondrial profiling of dog hairs. Forensic Sci Int. 2003 May 5;133(3):235-41.

16. Bjornerfeldt S, Webster MT, Vila C. Relaxation of selective constraint on dog mitochondrial DNA following domestication. Genome Res. 2006 Aug;16(8):990-4.

17. Kim KS, Lee SE, Jeong HW, Ha JH. The complete nucleotide sequence of the domestic dog (Canis familiaris) mitochondrial genome. Mol Phylogenet Evol. 1998 Oct;10(2):210 20.

18. Ishiguro N, Nakajima A, Horiuchi M, Shinagawa M. Multiple nuclear pseudogenes of mitochondrial DNA exist in the canine genome. Mamm Genome. 2002 Jul;13(7):365-72.

19. Wilson MR, Allard MW, Monson K, Miller KW, Budowle B. Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region. Forensic Sci Int. 2002 Sep 10;129(1):35-42.

20. Schneider S, Roessli D, Excoffier L. Arlequin: A software for population genetics data analysis. 2.0000 ed. University of Geneva: Genetics and Biometry Lab, Dept. of Anthropology; 2000.

21. Nixon KC. Winclada (BETA). 0.9.9 ed. Ithaca, NY: Published by author; 1999.

22. Goloboff PA. NONA (NO NAME). 2 ed. Tucuman, Argentina: Published by author; 1999.

23. Goloboff PA. Methods for Faster Parsimony Analysis. Cladistics. 1996 1996;12(3):199 220.

24. Zwickl DJ. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [Ph.D. dissertation]: The University of Texas at Austin; 2006.

25. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993 May;10(3):512-26.

26. Himmelberger AL, Spear TF, Satkoski JA, George DA, Garnica WT, Malladi VS, et al. Forensic utility of the mitochondrial hypervariable region 1 of domestic dogs, in conjunction with breed and geographic information. J Forensic Sci. 2008 Jan;53(1):81-9.

VI. Dissemination of Research Findings: In Press

Webb, K.M. and Allard, M.W. (2008). Announcement of the First Public Reference Database of dog Mitochondrial Single Nucleotide Polymorphisms (SNPs) for Forensic Use. In press the Journal of Forensic Science.

Webb, K.M. and Allard, M.W. (2008). Analysis of the Domestic Dog Mitochondrial Genome for Forensic Use. Use. In press, Journal of Forensic Science.

All of the sequences will be submitted to GenBank, see attached manuscripts for accession numbers. The data will also be placed on the web at The George Washington University, as well as offered to any other party who wishes to use or display it including the NIJ.

Several other dog evolutionary papers are also in preparation and will be submitted to the relevant journals.

One other forensic paper will be written on the remaining 100 dog mtgenome sequences that are being collected and analyzed at the University of California, Davis.