J Forensic Sci, May 2007, Vol. 52, No. 3
doi:10.1111/j.1556-4029.2007.00425.x
Available online at: www.blackwell-synergy.com

Rebekah L. Gundry,[1] Ph.D.; Marc W. Allard,[2] Ph.D.; Tamyra R. Moretti,[3] Ph.D.; Rodney L. Honeycutt,[4] Ph.D.; Mark R. Wilson,[5] Ph.D.; Keith L. Monson,[6] Ph.D.; and David R. Foran,[7] Ph.D.

Mitochondrial DNA Analysis of the Domestic Dog: Control Region Variation Within and Among Breeds

ABSTRACT: The mitochondrial DNA (mtDNA) control regions of 125 domestic dogs (*Canis familiaris*) encompassing 43 breeds, as well as one coyote and two wolves were sequenced and subsequently examined for sequence variation in an effort to construct a reference dog mtDNA data set for forensic analysis. Forty informative variable sites were identified that described 45 haplotypes, 29 of which were observed only once. Substantial variation was found both within and among breeds in the mtDNA derived from tissue, indicating that analysis of the mtDNA derived from dog hairs could be a valuable, discriminating piece of evidence in forensic investigations. The dog data set single nucleotide polymorphisms (SNPs) ranged from having one to six changes on a phylogenetic tree. On average, there were 1.9 character changes for each variable position on the tree. The most variable sites (with four or more changes each, listed from the most changes to the fewest) observed were 15,639 ($L = 6$), 16,672 ($L = 5$), 15,955 ($L = 4$), 15,627 ($L = 3$), 16,431 ($L = 3$), and 16,439 ($L = 3$). These sites were consistent with other reports on variable positions in the dog mtDNA genome. A total of 26 SNPs were chosen to best identify all major clusters in the domestic dog data set. The descriptive analyses revealed that this data set is similar to other published canine data sets and further demonstrates that this domestic dog data set is a useful resource for forensic applications. This reference data set has been compiled and validated against the published dog genetic literature with an aim to aid forensic investigations that seek to incorporate mtDNA sequences and SNPs from trace evidence such as dog hair.

KEYWORDS: forensic science, trace evidence, domestic dog, mitochondrial DNA, sequence variation, control region, interbreed and intrabreed studies, *Canis familiaris*

The first published mitochondrial DNA (mtDNA) genome of the domestic dog (*Canis familiaris*) contains 16,727 bp (1) with the control region (CR, see review in (2)) spanning positions 15,458–16,727 (1270 bp). While the dog mtDNA genome closely resembles the mtDNA genome of other mammals, the dog (and related canids) mtDNA CR differs due to the presence of a 10 bp repeat unit (5′-GTACACGT(A/G)C-3′) that begins at base 16,130 and varies in number and sequence both within and among individuals. The CR of dog mtDNA, like that of human mtDNA, has been the focus of a number of studies investigating the variation among individuals. Previous studies (1–9) have revealed more than 100 single nucleotide polymorphisms (SNPs) throughout the mtDNA CR of the domestic dog. Additional studies that have focused on the evolutionary genetics of dogs and the history of their domestication (5,6,10–14) have provided important genetic information pertinent to forensic applications. These studies have examined dog breeds from Europe, Asia, Africa, Siberia, India, America, and Japan (5,6,10–14) and have identified SNPs, haplotypes, and haplogroups that are defined by variable mtDNA sites observed among individuals (2,5,6,9,10,14). Furthermore, the utility of mtDNA sequence information to forensic casework has been demonstrated for dog hairs and saliva (3,8,9,15). For example a recent dog database examined 109 dogs for 573 bp of the 5′ end of the control region with animals from Germany, Sweden, and Europe, and compared those samples with others from Japan, China, and the United Kingdom. However, despite previous reports on the sequence and utility of dog mtDNA, there is still a need for a publicly available United States reference data set for forensic analyses. In addition, further investigation of the variation along the entire CR, a validation of SNP sites against known genetic data, and additional examination of both intrabreed and interbreed variation are needed.

The reference database presented here contains the complete dog mtDNA CR sequences, variable SNPs, and haplotypes of 125 U.S. domestic dogs and three wild canids. The addition of detailed phylogenetic analyses to the sequence comparisons allowed for the identification and confirmation of informative variable sites. The variation reported herein was compared with the published dog genetic data (1–9) in order to determine whether or not the genetic variation in this reference database was typical of other domestic dogs. While previous studies (1–9) rarely examined more than two individuals per breed, the current study includes broader sampling within two selected breeds (Golden Retrievers and Labrador Retrievers, with $n = 34$ and $n = 30$, respectively) in order to examine the additional variation uncovered with broader intrabreed sampling. This diverse data set, including full mtDNA CR sequences, detailed phylogenetic analyses, and validation against previously reported data, allows for further assessment

[1] Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205.
[2] Department of Biological Sciences, George Washington University, Washington, DC 20052.
[3] Federal Bureau of Investigation, DNA Unit 1, Quantico, VA 22135.
[4] Department of Wildlife and Fisheries Sciences, Texas A&M University, College Station, TX 77843.
[5] Federal Bureau of Investigation, Chem-Bio Sciences Unit, FBI Laboratory, Quantico, VA 22135.
[6] Counterterrorism and Forensic Science Research Unit, FBI Academy, Quantico, VA 22135.
[7] School of Criminal Justice and Department of Zoology, 560 Baker Hall, Michigan State University, East Lansing, MI 48824.

Received 1 Aug. 2006; and in revised form 11 Nov. 2006; accepted 11 Nov. 2006; published 13 April 2007.

of the genetic variation in the mtDNA CR of domestic dogs. This data set therefore provides a valuable forensic asset, additional data for a dog mtDNA CR reference database, and a developing tool for the examination of a common piece of trace evidence, namely shed dog hair.

Methods

One hundred and twenty-five dogs of 43 breeds (Tables 1 and 2) were sampled, including Alaskan Husky ($n = 1$), American Eskimo Dog ($n = 3$), American Spitz ($n = 1$), Anatolian Shepherd Dog ($n = 2$), Basset Hound ($n = 1$), Beagle ($n = 1$), Belgian Sheepdog ($n = 1$), Border Collie ($n = 2$), Boxer ($n = 2$), Brittany ($n = 1$), Cairn Terrier ($n = 1$), Chesapeake Bay Retriever ($n = 2$), Chihuahua ($n = 1$), Chow Chow ($n = 2$), Cocker Spaniel ($n = 2$), Dachshund ($n = 1$), Dalmatian ($n = 1$), Doberman Pinscher ($n = 2$), English Bulldog ($n = 1$), English Springer Spaniel ($n = 1$), English Terrier ($n = 1$), German Shorthaired Pointer ($n = 1$), Golden Retriever ($n = 34$), Great Dane ($n = 2$), Greyhound ($n = 1$), Husky ($n = 2$), Kerry Blue Terrier ($n = 1$), Labrador Retriever ($n = 30$), Lhasa Apso ($n = 3$), Maremma Sheepdog ($n = 2$), Miniature Schnauzer ($n = 2$), Old English Sheepdog ($n = 1$), Pug ($n = 1$), Rottweiler ($n = 2$), Shar Planinetz ($n = 2$), Siberian Husky ($n = 1$), Soft Coated Wheaten Terrier ($n = 1$), Staffordshire Bull Terrier ($n = 2$), Standard Poodle ($n = 2$), Toy Poodle ($n = 1$), West Highland White Terrier ($n = 1$), Whippet ($n = 1$), and Yorkshire Terrier ($n = 2$), as well as two gray wolves (*C. lupus*), and one coyote (*C. latrans*). Thirty-seven of the dog breeds are recognized by the American Kennel Club and thus are well established in the United States. The interbreed analysis was comprised of 61 dogs covering 41 breeds ($n = 1$–3 per breed). This includes all of the convenience samples collected except for those in the intrabreed analysis. The intrabreed analysis was based on a larger sampling of unrelated individuals from two select common breeds (Golden Retrievers [$n = 34$] and Labrador Retrievers [$n = 30$]).

Tissue samples (blood, heart, liver, testis, and uterus) were collected from dog breeders and veterinary clinics at various locations (Unknown $n = 3$, Texas $n = 62$, Massachusetts $n = 51$, Michigan $n = 7$, Italy $n = 2$, wolves from Minnesota and North West Territories Canada, coyote from Texas) and stored at -70°C. The unextracted tissues were rinsed with ethanol and ddH_2O, and *c.* $0.5\,cm^3$ was digested at 56°C for 2 h in 300 µL extraction buffer (10 mM Tris, 100 mM NaCl, 39 mM DTT, 10 mM EDTA, 2.0% SDS) and 1.2 U proteinase K (Amresco, Solon, OH). Residual tissues were transferred into a Spin-X® extraction tube (Costar®, Corning, NY) and centrifuged for 5 min. A solution of 300 µL of phenol:chloroform:isoamyl alcohol (25:24:1) was added to the filtrate. The samples were then vortexed and centrifuged. The aqueous phase was removed and then purified using a Microcon™ 100 concentrator (Millipore Corp., Bedford, MA). DNA was resuspended in sterile ddH_2O and quantified by spectrophotometry.

TABLE 1—*Interbreed analysis.*

#	Breed	(*n*) Per Breed	Total (*n*)	%
1	Alaskan Husky	1	1	1.64
2	American Eskimo Dog	1	2	3.28
	Belgian Sheepdog	1		
3	Doberman Pinscher	1	1	1.64
4	Great Dane	2	2	3.28
5	Dalmatian	1	2	3.28
	West Highland White Terrier	1		
6	Anatolian Shepherd Dog	1	3	4.92
	Shar Planinetz	2		
7	Border Collie	2	3	4.92
	Cocker Spaniel	1		
8	Doberman Pinscher	1	1	1.64
9	Siberian Husky	1	1	1.64
11	Chesapeake Bay Retriever	1	1	1.64
13	Cairn Terrier	1	1	1.64
16	Bassett Hound	1	8	13.11
	Dachshund	1		
	English Bulldog	1		
	German Shorthaired Pointer	1		
	Kerry Blue Terrier	1		
	Lhasa Apso	1		
	Standard Poodle	1		
	Toy Poodle	1		
17	Lhasa Apso	1	1	1.64
20	American Eskimo Dog	1	2	3.28
	American Spitz	1		
21	Yorkshire Terrier	1	1	1.64
23	Boxer	2	3	4.92
	Staffordshire Bull Terrier	1		
28	Chow Chow	1	1	1.64
30	American Eskimo Dog	1	1	1.64
31	Miniature Schnauzer	2	2	3.28
32	Brittany	1	1	1.64
33	Husky	1	1	1.64
34	Beagle	1	1	1.64
35	Soft Coated Wheaten Terrier	1	1	1.64
36	Yorkshire Terrier	1	1	1.64
37	Greyhound	1	1	1.64
38	Rottweiler	2	2	3.28
39	Anatolian Shepherd Dog	1	6	9.84
	Chihuahua	1		
	Chow Chow	1		
	English Springer Spaniel	1		
	Husky	1		
	Staffordshire Bull Terrier	1		
40	Maremma Sheepdog	1	1	1.64
41	Maremma Sheepdog	1	1	1.64
42	Cocker Spaniel	1	6	9.84
	Lhaso Apso	1		
	Old English Sheepdog	1		
	Pug	1		
	Standard Poodle	1		
	Whippet	1		
43	English Terrier	1	1	1.64
44	Chesapeake Bay Retriever	1	1	1.64

The haplotype distribution among 61 individuals in the interbreed analysis (41 breeds) is listed. Breed, haplotype #, number of individuals per breed, number of individuals per haplotype, and frequency (%) that the haplotype was observed are provided. Haplotype number refers to Table 5.

Oligonucleotide primers used to amplify and sequence the canid mtDNA CR were designed based on a dog reference sequence ((1); GenBank accession no. U96639). A primer naming convention was used where the primer name indicates the position of the 5′ base. Forward primers were defined as F15412 (5′-CCACTATCAGCACCCAAAG-3′), F15719 (5′-GTAATGTCCC TCTTCTCGCT-3′), F16072 (5′-CTCACGCATAAAATCAAG GTG-3′), and F16431 (5′-CACGCGCGTAAGACATTAAG-3′). Reverse primers were defined as R15803 (5′-TGAAGTAAGAA CCAGATGCCA-3′), R16114 (5′-CCTGAAACCATTGACTGAATAG-3′), R16527 (5′-GGGTTTGGCGGGACATAA-3′), and R42 (5′-GGCATTTTCAGTGCCTTGCTT-3′). Lyophilized primers (Operon Technologies Inc., Alameda, CA) were resuspended to a concentration of 10 mM in TE (10 mM Tris-HCL, 0.1 mM EDTA, pH 8.0). Primers F15412 and R42 are positioned outside of the CR and were used for amplification. All primers were used in sequencing reactions to generate overlapping bidirectional sequences covering both strands of the entire CR (Fig. 1).

TABLE 2—*Intrabreed analysis.*

Haplotype #	Individuals (*n*)	%
Labrador Retrievers (*n* = 30)		
19	1	3.33
22	1	3.33
23	1	3.33
26	1	3.33
39	1	3.33
29	3	10.00
16	4	13.33
24	4	13.33
27	14	46.67
Golden Retrievers (*n* = 34)		
14	1	2.94
15	1	2.94
18	1	2.94
25	1	2.94
27	4	11.76
16	9	26.47
26	17	50.00

The haplotype distribution among 34 Golden Retrievers and 30 Labrador Retrievers in the intrabreed analysis is listed. Breed, haplotype #, number of individuals per breed, and frequency (%) that the haplotype was observed within that breed are provided. Haplotype number refers to Table 5.

One nanogram of total genomic dog DNA was used to amplify the entire CR in a 25 μL reaction volume containing the following reagents: 200 μM of each dNTP (Applied Biosystems, Foster City, CA), 0.6 μM each of primers F15412 and R42, 5 U Ampli*Taq* Gold DNA polymerase (Applied Biosystems), 0.16 μM BSA, and 1 × GeneAmp^{TM} PCR Buffer containing $MgCl_2$ (Applied Biosystems). Amplifications were conducted in a GeneAmp$^{®}$ 9700 PCR System thermal cycler (Applied Biosystems), and consisted of denaturation for 11 min at 95°C, followed by 30 cycles of 1 min at 94°C, 1 min at 60°C, and 2 min at 72°C, plus a final incubation for 60 min at 60°C. Genomic dog DNA (Novagen Inc., Madison, WI) was used as a positive control and sterile ddH_2O as a negative control. ExoSAP-IT^{TM} (USB, Cleveland, OH) was used according to the manufacturer's instructions for inactivation and removal of residual dNTPs and primers from the reaction before sequencing. Cycle sequencing was performed using an ABI PRISM^{TM} dRhodamine Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems) according to the manufacturer's instructions. The sequencing primers were the same as those listed above. The sequencing reaction (containing 7 μL of the amplified DNA, 3.5 μL primer [at 1 μM] and 9.5 μL Ready Reaction mix) was carried out in a GeneAmp$^{®}$ 9700 PCR System thermal cycler and consisted of denaturation for 1 min at 96°C, followed by 25 cycles of 15 sec at 96°C, 1 sec at 50°C, and 1 min at 60°C, followed by a hold at 4°C until the next step. Following the sequencing reaction, the samples were filtered through a CentriSep 96-well filter plate (Princeton Separations, Adelphia, NJ) according to the manufacturer's instructions. The samples were then dried in a speedvac for 15 min and resuspended in 10 μL Hi Deionized Formamide (Applied Biosystems). The sample was then heated for 2 min at 95°C and then chilled on ice. An ABI PRISM^{TM} 310 Genetic Analyzer (Applied Biosystems) was used for sequencing. Instrument parameters included a 1 mL syringe, a POP-6^{TM} (Applied Biosystems) polymer, a 47 cm capillary, a 1 × Genetic Analyzer buffer containing EDTA (Applied Biosystems), a POP6 (1 mL) Rapid E run module, a DTPOP6 (dR set-any primer) mobility file, and a CE1 base caller. The running parameters included a 10-sec injection, a 2.0 kV injection, a 15 kV run, a 50°C run temperature, and a 35-min run time. Sequencing Analysis v. 3.4.1 (Applied Biosystems) was used to analyze the raw data and Sequencher v.4.1.2 (Gene Codes, Ann Arbor, MI) was used to make final base decisions and to edit and assemble the final CR sequences.

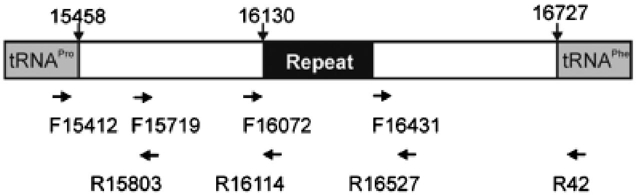

FIG. 1—*Schematic diagram of the dog mitochondrial DNA control region and flanking tRNA genes. The repeat region is shaded in black. The direction (arrows) and relative position of amplification (F15412 and R42) and sequencing primers (all) are indicated. Overlapping sequences were determined between the following primer sets: F15412 and R15803; F15719 and R16114; F16072 and R16527; and F16431 and R42.*

All sequences are available at the National Center for Biotechnology Information website in the GenBank database (http://www.ncbi.nlm.nih.gov/Genbank/index.html) under accession numbers AY240030-AY240157.

Sequence variants, haplotypes, and haplogroups are reported relative to a reference sequence (1) and in a manner similar to that used for human mtDNA (16–21). Variable positions were identified using the Winclada and NONA software (22,23) available at www.cladistics.com and named relative to the dog reference sequence just as human mtDNA studies have utilized the Cambridge reference sequence (24) to facilitate communication and nomenclature.

Phylogenetic analyses were performed using WinClada and Nona software. The CR sequences were aligned according to standardized rules for human mtDNA sequence alignments (16,17) and the aligned sequences were imported into the phylogenetic software. The repeat region was excluded from phylogenetic analyses and was not used in counts of the number of haplotypes or overall diversity due to the large amount of variation and the likely presence of heteroplasmy. Parsimony ratchet analysis of 2000 iterations was performed on the alignment and the most parsimonious tree(s) was used for all subsequent analyses. The coyote and wolf sequences were used as outgroups. The variation observed for canids was compared with a U.S. Caucasian data set (18) to assess the relative discriminating ability of mtDNA CR in dogs versus humans.

The most informative and variable sites for the dog CR sequences were determined by analysis and inspection of the phylogenetic tree. Estimates of the character length and retention index (Ri) were used in determining whether the variable sites were informative. Character length is the number of times a character is observed to change across the tree. Retention index is a measure of character congruence; hence, if a character arose once and defines all members of a clade then that character will have an Ri of 100. If there are any reversals or independent gains then the Ri is <100, and has a score of 0 if all character changes independently arose. Sites that were variable in two or more data samples were listed as phylogenetically informative, while sites that distinguished clusters of four or more samples and showed a low number of independent gains and/or reversals were considered to be highly informative.

Results

Complete mtDNA CR sequences were generated for all 128 canid samples (125 dogs, two gray wolves, and one coyote). Con-

TABLE 3—*Summary of the lengths of the CR in 128 canid samples.*

Length (bp)	Repeats	*n*
1225	25	1
1235	26	1
1242	27	1
1243	27	1
1252	28	2
1253	28	12
1254	28	2
1255	28	11
1256	28	2
1262	29	3
1263	29	7
1265	29	5
1266	29	1
1268	29	1
1272	30	5
1273	30	12
1274	30	1
1275	30	6
1277	30	1
1282	31	3
1283	31	5
1284	31	1
1285	31	8
1287	31	1
1292	32	1
1293	32	7
1294	32	1
1295	32	9
1303	33	2
1305	33	1
1313	34	5
1314	34	1
1315	34	2
1316	34	1
1323	35	1
1325	35	3
1422	38	1

Length of the CR, number of 10 bp repeat units, and number of individuals are listed.

CR, control region.

trol region sizes ranged from 1225 to 1422 bp (Table 3), with the largest sequence found in a Chesapeake Bay Retriever sample that contained a 67 bp insertion at position 15,597 (5′-CCCCTAT GTACGTCGTGCATTAATGGTTTGCCCCATGCATATAAGC ATGTACATAATATTATATCCT-3′). A total of 97 variable sites were found along the length of the CR excluding the repeat region and not counting the 67 bp insert. Of these variable positions, 40 were phylogenetically informative in two or more dogs (Table 4). Thirty-three of the variable positions were also reported in a study of Japanese native dog breeds, which identified a total of 42 informative variable sites (4,5). After independent phylogenetic analysis of data in GenBank, 23 of the most informative sites in the Okumura et al. (5) study were also the most informative in the current data set. In the complete sequence comparisons of all 128 individuals, 45 haplotypes were observed, and of these, 29 were observed once (Tables 1, 2 and 5). The coyote and wolf haplotypes were both unique as well in this data set. No discernable relationship between breed and haplotype (or repeat number) was discovered with this data set, a finding similar to that reported in all other published dog mtDNA data sets (3–6,9,25).

The intrabreed study included the sequences of 34 Golden Retrievers and 30 Labrador Retrievers. Twenty-one variable and informative sites were observed in this study, with 19 of the sites shared by both breeds. All of the shared sites were highly informative (15,526, 15,595, 15,612, 15,620, 15,627, 15,632, 15,639, 15,643, 15,652, 15,800, 15,815, 15,912, 15,955, 16,003, 16,025, 16,083, 16,128, 16,431, 16,439). The differences included one rare site (16,032) for Labrador Retrievers and a common dog SNP (16,672) observed in Golden Retrievers only. The 34 Golden Retrievers had a genetic diversity value [$h = n(1 - \sum X_i^2)/(n-1)$] of 0.683 and a random match probability [$P = \sum X_i^2$] of 33.7%. The 30 Labrador Retrievers had a genetic diversity of 0.756 and a random match probability of 26.9%. The most common haplotype (Table 2) for Golden Retrievers (haplotype #26) was observed in 17 individuals and the second most common (haplotype #16) was observed in nine individuals. The most common haplotype for Labrador Retrievers (haplotype #27) was observed in 14 individuals. Two haplotypes (haplotypes #16, 24), each of which was observed in four individuals, were the next most common. It will take additional sampling to determine whether this pattern of genetic variation continues for other breeds.

The interbreed comparisons excluded the Golden and Labrador Retrievers, and thus included 61 individuals consisting of 41 breeds (Table 1). These sequence comparisons identified an additional 19 informative positions, for a total of 40 informative sites (Table 4). Twenty-six informative variable sites (shaded in Table 4) were determined to be highly informative. Among breeds, the genetic diversity was 0.977 and the random match probability was 3.87%. The most common haplotype (haplotype #16) was observed in eight individuals and the second most common consisted of two haplotypes (haplotypes #39 and 42), each of which was observed in six individuals.

Length heteroplasmy was found in the region containing 10 bp tandem repeat units beginning at position 16,130. The number of repeat units ranged from 25 to 38 (Table 3), the most common being 28 repeats ($n = 29$) and average 30.3 repeats. All individuals were variable in the ninth position and two individuals (Beagle, Yorkshire Terrier) were variable in the 10th position of the repeat. The sequence reads after the repeat region for some samples were out of phase (data not shown), a pattern that is consistent with the presence of length heteroplasmy in humans (19,26). As length heteroplasmy typically is not utilized in forensic databases, this additional information is not included in the current dog reference data set (16,17,19).

Another known CR length variant, referred to as a T-stretch, is recorded where the reference sequence (1) has eight Ts beginning at position 16,664. In the current data set, seven different variants of this T-stretch region were observed as follows (Table 5): (1) 16671.1T (haplotypes #12, 14, 18, 38, following nomenclature rules for human mtDNA CR sequences (16,17,21) where the number after the decimal indicates the position for base insertion); (2) 16671.1T, 16674G (haplotype #15); (3) 16672T (haplotypes #1, 11, 13, 16, 17, 19, 34, 36); (4) 16,663.1C, 16663.2C, 16672T (haplotypes #22–29, 44); (5) 16663.1C, 16663.2C, 16664C, 16672T (haplotype #21); (6) 16671C (haplotypes #7, 20); and (7) 16671C, 16705T (haplotypes #5, 6, 8, 9). In some cases, the T-stretch contained nine Ts either as a result of an insertion at 16671.1T (haplotypes #12, 14, 15, 18) or by a transition at position 16672T (haplotypes #1, 11, 13, 16, 17, 19, 21–29, 34, 36, 44). The insertion of two C residues before the T-stretch was also observed (haplotypes #21–29, 44). This region sequenced smoothly regardless of the C and T-stretches.

Discussion

The goals of this study were to develop a reference mtDNA CR database, to identify informative variable sites along the entire

TABLE 4—*Informative sequence variants observed in the dog reference data set interbreed study.*

Position	Reference	Observed	L	Ri
15,483	C	T	1	100
15,508	**C**	**T**	**1**	**100**
15,526	**C**	**T**	**1**	**100**
15,553	A	G	2	0
15,595	**C**	**T**	**1**	**100**
15,611	**T**	**C**	**2**	**90**
15,612	**T**	**C**	**2**	**96**
15,620	T	C	2	97
15,621	C	T	2	50
15,622	T	C	2	0
15,625	T	C	1	100
15,627	**A**	**G**	**3**	**96**
15,632	**C**	**T**	**2**	**96**
15,639	**T**	**A/G**	**6**	**91**
15,643	**A**	**G**	**2**	**96**
15,650	**T**	**C**	**1**	**100**
15,652	**G**	**A**	**2**	**96**
15,665	T	C	1	100
15,710	C	T	1	100
15,781	C	T	1	100
15,800	**T**	**C**	**1**	**100**
15,807	C	T	1	100
15,814	C	T	1	100
15,815	**T**	**C**	**1**	**100**
15,819	T	C	1	100
15,912	**C**	**T**	**2**	**97**
15,955	**C**	**T**	**4**	**93**
16,003	**A**	**G**	**1**	**100**
16,025	**T**	**C**	**2**	**97**
16,032	A	G	2	75
16,083	**A**	**G**	**2**	**96**
16,128	**G**	**A**	**1**	**100**
16,431	**C**	**T**	**3**	**95**
16,439	**T**	**C**	**3**	**94**
16,480	G	A	2	0
16,501	T	C	1	100
16,576	A	G	2	50
16,671	**T**	**C**	**2**	**90**
16,672	**C**	**T**	**5**	**92**
16,705	C	T	1	100

Position, reference (1) base, observed base, character length (L) and retention index (Ri) are listed. Shaded boxes indicate highly informative characters in current study. Bold indicates highly informative in Okumura et al. (5) study.

CR, and to document and validate the important variable sites that discriminate the major haplogroups of the domestic dog.

An important aspect in the development of a reference database is to establish a common nomenclature. Previous reports containing dog CR sequence variation have not developed a common nomenclature for describing observed genetic variation. However, Periera et al. (2) recently recommended several rules for reporting dog sequence variation and adopted a nomenclature system based on that used by the forensic community for reporting human mtDNA CR sequence variation. The key element suggested by Periera et al. (2), and implemented in the current study, is that a reference sequence should be defined and any variation from that sequence is reported. By developing a common nomenclature, much of the human mtDNA forensics community has been able to speak with one voice and enhance communication among laboratories (16–21). Therefore, these nomenclatural practices should be extended to all species, including dogs (2). Importantly, as issues of sequence alignment are essential for a consistent nomenclature of genetic variation in dogs, the current data set follows those rules recommended for human CR sequences (16,17,19).

The current study is the largest reported data set to date that covers the entire CR, with sequence data obtained for 128 canid samples, including 125 dogs from 43 breeds, and three wild canids. All SNP sites were categorized into two groups based on the level of discrimination they provided in separating canine haplotypes from one another. Sites that were variable in two or more data samples were considered as informative, while sites that distinguished clusters of four or more samples and showed a low number of independent gains and/or reversals were identified as highly informative phylogenetically. Of the 40 variable informative characters observed in this database, 26 were highly informative and important for distinguishing among dog haplotypes. Site 15,639 showed the greatest variability.

Importantly, the sequence data obtained here were compared with previous forensic and genetic studies. Of the 116 variable CR sites reported previously (3–6,8,9,27), 34 were informative and 26 were highly informative in the interbreed study (Table 4) presented here. Interestingly, 24 of the highly informative sites are found in both the dogs sampled here and in the Okumura et al. (5) study, indicating that the informative sequence variants found herein are useful for a wide variety of dogs from disparate geographic areas. The interbreed analyses here revealed four phylogenetically informative sites (15,819, 16,032, 16,501, 16,705) and 13 unique sites (15,640, 15,761, 15,925, 15,956, 15,959, 16,436, 16,480, 16,507.1, 16,598, 16,664, 16,674, 16,706.1, 16,706.2) that had not been reported previously in the literature. Additionally, the 67 bp insertion observed herein was previously reported in another Chesapeake-Bay Retriever (3).

Individuals within a particular breed are more likely to be genetically similar due to inbreeding, and thus of concern to forensic analysis and interpretation. As expected, the genetic diversity and number of differences were greater among breeds than they were within breeds. In intrabreed comparisons, the Labrador Retrievers had a higher genetic diversity than the Golden Retrievers examined. The intrabreed study revealed one site (16,032) that was unique to four of the Labrador Retrievers (haplotype #24). As this site was also observed in one American Eskimo Dog (haplotype #30) in the interbreed study, no additional informative SNPs were uncovered by the sequencing of extra dogs within a breed than had been discovered in interbreed comparisons which included one to three individuals per breed. One may not need to carry out extensive within-breed sampling if limited sampling shows few differences.

In addition to identified SNP sites, the sequence and length variability in the 10 bp repeats (beginning at position 16,130) is consistent with that reported previously (28,29). The variation in the number and sequences of these repeats suggests that these data may be useful for discriminating among individuals. However, other studies have found that the number of repeat units differs within individual hairs, along the shaft of a single hair, and among tissues (28,29). Owing to these reported levels of variation, the repeat region was not included in our analyses, although further study is warranted. The current data set also reveals sequence variation in the CR T-stretch, which has not been described previously in the literature, in part because analysis of the T-stretch necessitates sequencing the entire CR rather than only the 5′ portion as most investigators have done.

An interbreed comparison of 64 dogs and the same number of randomly selected human mtDNA CR sequences indicated that more variation exists in humans (data not shown). Dogs have fewer unique haplotypes and fewer informative variants than have been observed in humans, who have approximately three times as many informative sites and twice as many haplotypes as dog

TABLE 5—*Sequence differences relative to a dog mtDNA reference sequence (1) observed in 45 different canid haplotypes.*

Haplotype Number	*n*	15,459	15,460	15,464.1	15,483	15,499	15,508	15,513	15,514	15,515	15,519	15,526	15,530	15,532	15,533	15,534	15,553	15,557	15,594	15,595	15,597.1-.67
Reference Sequence:		**C**	**A**		**C**	**T**	**C**	**G**	**G**	**T**	**C**	**C**	**T**	**C**	**C**	**C**	**A**	**T**	**T**	**C**	
1	1				T																
2	2				T																
3	1				T																
4	2				T																
5	2						T					T									
6	3						T					T									
7	3						T					T									
8	1						T					T									
9	1						T					T									
10	1											T								T	
11	1											T								T	
12	1											T								T	
13	1											T								T	
14	1											T								T	
15	1											T								T	
16	21											T								T	
17	1											T								T	
18	1											T								T	
19	1											T								T	
20	2																				
21	1																				
22	1																				
23	4																				
24	4																				
25	1																				
26	18																				
27	18																				
28	1																				
29	3																				
30	1																				
31	2																				
32	1																				
33	1																				
34	1																G				
35	1																G				
36	1																				
37	1																				
38	2																				
39	7																				
40	1																				
41	1																				
42	6																				
43	1																				
44	1																				*
45	1	T	G	C		C		A	-	-	T		C	-	-	-		C	-		

Haplotype Number	*n*	15,598.1	15,611	15,612	15,613	15,617	15,620	15,621	15,622	15,623	15,625	15,627	15,628	15,632	15,639	15,640	15,643	15,647.1	15,648	15,649	15,650
Reference Sequence:			**T**	**T**	**A**	**C**	**T**	**C**	**T**	**C**	**T**	**A**	**T**	**C**	**T**	**T**	**A**		**A**	**A**	**T**
1																					
2	1											G			A						
3	2											G			A						
6	1											G			A						
4	2											G			A						
5	2		C												G						C
6	3		C												A						C
7	3		C												G						C
8	1		C					T							G						C
9	1		C												G						C
10	1				G				C		C			T	A						
11	1			C					C					T	G		G				
12	1			C							C			T	A						
13	1			C										T	G		G				
14	1			C										T	G		G				
15	1			C										T	G		G				
16	21			C										T	G		G				
17	1			C										T	G		G				
18	1			C										T	G		G				
19	1			C										T	G		G				
20	2							T							A						
21	1						C					G			A						
22	1											G			A						
23	4						C					G			A						
24	4						C					G			A						
25	1						C					G			A						
26	18						C					G			A						
27	18						C					G			A						
28	1						N					G			A						
29	3											G			A						
30	1											G			A	C					
31	2											G		T	A						
32	1											G			G						
33	1											G			A						
34	1														A						
35	1														A						
36	1														A						
37	1														A						
38	2														A						
39	7														A						
40	1														A						
41	1																				
42	6																				
43	1																				
44	1						C					G			A						
45	1	T	C			T	C			T	C	G	C	T		-		T	G	G	

Haplotype Number	*n*	15,651	15,652	15,653	15,665	15,710	15,750	15,751	15,761	15,764	15,769	15,773	15,781	15,800	15,807	15,811	15,813	15,814	15,815	15,819	15,912
Reference Sequence:		**C**	**G**	**A**	**T**	**C**	**C**	**T**	**G**	**A**	**C**	**A**	**C**	**T**	**C**	**A**	**C**	**C**	**T**	**T**	**C**
1	1			G														T			
2	2																	T			T
3	1																	T			T
4	2																	T			
5	2													C				T			T
6	3													C				T			T
7	3					T								C				T			T
8	1													C				T			T
9	1													C				T			T
10	1		A											C				T	C		T
11	1		A											C				T	C		T
12	1		A											C				T	C		T
13	1		A											C				T	C		T
14	1		A											C				T	C		T
15	1		A											C				T	C		T
16	21		A											C				T	C		T
17	1		A											C				T	C		T
18	1		A											C				T	C		T
19	1		A											C				T	C		T
20	2				C								T		T			T			
21	1																	T			
22	1																	T			
23	4																	T			
24	4																	T			
25	1																	T			
26	18																	T			
27	18																	T			
28	1																	T			
29	3																	T			
30	1																	T			
31	2																	T		C	
32	1																	T			
33	1		A															T			
34	1								-									T			
35	1																	T			
36	1																	T			
37	1																	T			
38	2																	T			
39	7																	T			
40	1																	T			
41	1																	T			
42	6																	T			
43	1																				
44	1																	T			
45	1	T	A		C	T	T	C		T	T	G		C		G	T				T

Haplotype Number	*n*	15,925	15,930	15,931	15,938	15,955	15,956	15,959	16,003	16,025	16,032	16,083	16,122	16,125	16,128	16,130	16,131	16,431	16,436	16,439	16,480
Reference Sequence:		**A**	**T**	**A**	**G**	**C**	**C**	**C**	**A**	**T**	**A**	**A**	**G**	**T**	**G**	**G**	**T**	**C**	**G**	**T**	**G**
1	1																				
2	2																				
3	1				-																
4	2																				
5	2				-	T			G									T		C	
6	3				-	T			G									T		C	
7	3				-	T			G									T		C	
8	1				-	T			G									T		C	
9	1					T			G									T		C	
10	1				-	T			G			G						T		C	
11	1				-	T			G			G			A						
12	1					T			G			G						T		C	
13	1					T	-		G						A			T		C	
14	1					T			G			G			A			T			
15	1					T			G			G			A			T		C	
16	21					T			G			G			A			T		C	
17	1			-		T			G			G			A			T		C	
18	1					T			G			G			A			T		C	
19	1					T			G			G						T		C	
20	2																	T			
21	1					T															
22	1					T														C	
23	4					T															
24	4									C	G										
25	1									C											
26	18									C											
27	18																				
28	1							T		C											
29	3					T															
30	1			-						C	G							T			
31	2									C											
32	1									C											
33	1																	T			
34	1																				
35	1									C											
36	1					T													A		
37	1									C											A
38	2									C											
39	7									C											
40	1	T								C											
41	1									C											
42	6																				
43	1																				
44	1																				
45	1		-	C		T							A	C		A	C	T		C	A

Haplotype Number	*n*	16,501	16,507	16,507.1	16,576	16,598	16,633	16,663.1	16,663.2	16,664	16,670	16,671	16,671.1	16,672	16,674	16,702	16,703	16,705	16,706.1	16,706.2	16,714	16,716	16,727
Reference Sequence:		**T**	**T**		**A**	**T**	**G**			**T**	**T**	**T**		**C**	**C**	**T**	**T**	**C**			**C**	**A**	**A**
1	1													T									
2	2																						
3	1																						
4	2																						
5	2	C										C						T					
6	3											C						T					
7	3											C											
8	1											C						T					
9	1											C						T					
10	1																						
11	1													T									
12	1												T										
13	1													T									
14	1			T									T										
15	1												T		G								
16	21													T									
17	1													T									
18	1												T										
19	1													T									
20	2											C											
21	1							C	C	C				T									
22	1							C	C					T									
23	4							C	C					T									
24	4							C	C					T									
25	1							C	C					T					C	A			
26	18							C	C					T									
27	18							C	C					T									
28	1							C	C					T									
29	3							C	C					T									
30	1																						
31	2																						
32	1																						
33	1																						
34	1				G									T									
35	1																						
36	1				G									T									
37	1																						
38	2												T										
39	7					N																	
40	1																						
41	1																						
42	6																						
43	1																						
44	1							C	C					T									
45	1		C		G		A				C	C											

The number of sequences (*n*) with each haplotype is listed. Breeds for which the haplotype was observed is listed in Tables 1–2. Letters (A, C, G, T) represent base substitutions that are different from the dog reference sequence. A dash (–) indicates a deletion. Blanks in the data set represent the same base as that observed in the dog reference sequence. A *indicates a 67 bp insertion at position 15,597 that was observed in a Chesapeake Bay Retriever (5′-CCCCTATGTACGTCGTGCATTAATGGTTTGCCCCATGCATATAAGCATGTACATAATATTATATCCT-3′). Haplotypes 10 and 12 are wolves and haplotype 45 is a coyote.

samples of comparable size. The presence of more variation in humans could be due to several factors. The dogs examined in this study were all purebred animals, artificially selected for similar phenotypic traits, and thus are likely to share a similar genetic background that reduces their genetic variation. Also, the domestic dog is thought to be a younger species than humans (6,25). The current study purposely selected purebred animals to determine whether the mtDNA CR could be used to associate individuals to a particular breed. Future work should sample more dogs including mixed-breed animals as they also could become the focus of a forensic investigation. However, preliminary evidence in the literature indicates that it is unlikely that Mongrel Dogs or Mix-Breed dogs will be significantly different from purebred animals, at least over the CR region of mtDNA (9).

In conclusion, this new database of information for determining domestic dog haplotypes will provide a useful baseline for forensic analyses of the dog mtDNA CR. Full-length CR sequences add additional variation as it does in humans and is therefore warranted. The interbreed data reveal that there are 26 informative SNPs within the CR that can be useful to determine the haplotype in forensic determinations of whether a particular dog can be included or excluded as a possible source of an evidentiary sample. The overall consistency of the dog data set with other published sequences supports the utility of this data set for forensic applications.

Acknowledgments

We thank Kerri Dugan, John Stewart, and Constance Fisher for their careful review of this manuscript. This research was conducted at the Counterterrorism and Forensic Science Research Unit (CFSRU) at the F.B.I. Academy and George Washington University. We thank Deborah Polanskey, Eric Pokorak, and Patricia Aagard (F.B.I.) and Eric Blair (Celera, Rockville, MD) for their time and use of sequencing instruments and James Robertson, (F.B.I., CFSRU) for the use of the fluorometer and spectrophotometer. All of the original tissue and DNA samples are housed in the Texas Cooperative Wildlife Collection at Texas A&M University. This research was supported in part by an appointment to the Visiting Scientists Program at the Federal Bureau of Investigation, Counterterrorism Forensic Science Research Unit administered by the Research Participation Program of the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the FBI-CFSRU. This is publication 04-10 of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only, and inclusion does not imply endorsement by the Federal Bureau of Investigation.

References

1. 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.
2. Periera 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.
3. 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;42:593–600.
4. 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.
5. 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.
6. Vila C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, et al. Multiple and ancient origins of the domestic dog. Science 1997;276:1687–9.
7. Randi E, Lucchini V, Christensen M, Mucci N, Funk S, Dolf G, et al. Mitochondrial DNA variability in Italian and East European wolves: detecting the consequences of small population size and hybridization. Conserv Biol 2001;14:464–73.
8. Schneider PM, Seo Y, Rittner C. Forensic mtDNA hair analysis excludes a dog from having caused a traffic accident. Int J Legal Med 1999;112: 315–6.
9. Wetton JH, Higgs JE, Spriggs AC, Roney CA, Tsang CS, Foster AP. Mitochondrial profiling of dog hairs. Forensic Sci Int 2003;133:235–41.
10. Vila C, Maldonado JE, Wayne RK. Phylogenetic relationships, evolution, and genetic diversity of the domestic dog. J Hered 1999;90:71–7.
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. Leonard JA, Wayne RK, Wheeler J, Valadez R, Guillen S, Vila C. Ancient DNA evidence for old world origin of new world dogs. Science 2002; 298:1613–6.
13. Wayne RK. Molecular evolution of the dog family. Trends Genet 1993; 9:218–24.
14. Wayne RK, Ostrander EA. Origin, genetic diversity, and genome structure of the domestic dog. Bioessays 1999;21:247–57.
15. Savolainen P, Lundeberg J. Forensic evidence based on mtDNA from dog and wolf hairs. J Forensic Sci 1999;44:77–81.
16. 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.
17. Wilson M, MW A, Monson K, Miller KW, Budowle B. Further discussions of the consistent treatment of length variants in the human mitochondrial DNA control region. Forensic Sci Commun 2002;4:1–8.
18. Allard MW, Miller K, Wilson M, Monson K, Budowle B. Characterization of the Caucasian haplogroups present in the SWGDAM forensic mtDNA dataset for 1771 human control region sequences. Scientific working group on DNA analysis methods. J Forensic Sci 2002;47:1215–23.
19. Budowle B, DiZinno J, Wilson M. Interpretation guidelines for mitochondrial DNA sequencing. Proceedings of the Tenth International Symposium on Human Identification. Madison, WI: Promega Corporation, http://www.promega.com/geneticidproc/ussymp10proc/default.htm, 1999.
20. Finnila S, Lehtonen MS, Majamaa K. Phylogenetic network for European mtDNA. Am J Hum Genet 2001;68:1475–84.
21. Helgason A, Hickey E, Goodacre S, Bosnes V, Stefansson K, Ward R, et al. mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. Am J Hum Genet 2001;68:723–37.
22. Goloboff P., NONA ver. 2, www.cladistics.com
23. Nixon K., WinClada ver 1.00.08, www.cladistics.com
24. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature 1981;290:457–65.
25. Angleby H, Savolainen P. Forensic informativity of domestic dog mtDNA control region sequences. Forensic Sci Int 2005;154:99–110.
26. Bendall KE, Sykes BC. Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. Am J Hum Genet 1995;57: 248–56.
27. Tsuchida S, Ikemoto S. Mitochondrial DNA polymorphism in dogs. J Vet Med Sci 1992;54:417–24.
28. 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.
29. Savolainen P, Arvestad L, Lundeberg J. mtDNA tandem repeats in domestic dogs and wolves: mutation mechanism studied by analysis of the sequence of imperfect repeats. Mol Biol Evol 2000;17:474–88.

Additional information and reprint requests:
Tamyra R. Moretti, Ph.D.
Federal Bureau of Investigation
DNA Unit 1
Quantico, VA 22135
E-mail: mallard@fbiacademy.edu