Inference of the Demographic History of the Domestic Dog
(*Canis lupus familiaris*)

Honors Thesis
Presented to the College of Agriculture and Life Sciences, Physical Sciences
of Cornell University
in Partial Fulfillment of the Requirements for the
Research Honors Program

by
Julie Marie Granka
January 2008
Dr. Carlos Bustamante

Table of Contents

List of Figures

List of Tables

Abstract

The domestic dog (*Canis lupus familiaris*), the oldest domesticated species, has a unique demographic history through its domestication from the gray wolf (*Canis lupus*) and in the formation of behaviorally and morphologically diverse dog breeds. Using information contained in the site frequency spectrum of purebred dogs and the Poisson Random Field framework, we infer the demography of the dog at domestication, in the formation of individual dog breeds, and of several wild canid populations. First, we find evidence for a slight contraction in population size approximately 15,000 years ago during the domestication of the dog. As these results may be an artifact of using breed dogs to infer a pre-breed dog population, it is likely that continued introgression between dogs and wolves or multiple domestication events have maintained high levels of dog diversity. Demography in the formation of several dog breeds is also examined, where the relatively rare breeds of the Bernese Mountain Dog and Pekingese appear to have gone through the most severe population contractions. In contrast, less severe contractions are found for the Golden and Labrador Retrievers, both popular breeds, and the Akita, which has likely introgressed with wolves. Finally, we examine data from several wild canid populations, finding evidence for population contractions in the gray wolf populations of Spain and Israel, but none in North American populations or coyote. We have developed a more comprehensive picture of the domestic dog's demographic history, which can prove useful in its application to other studies of the domestic dog currently underway.

I. Introduction

The domestic dog (*Canis lupus familiaris*) has recently become a model organism of great interest, so much so that it has been called the "geneticists' best friend" (Pennisi 2007). From the toy poodle to the Saint Bernard, domestic dogs differ drastically in size, shape, color, musculature, and other features. The existence of extreme differences among dog breeds, a result of intense selective breeding among purebred dogs, makes domestic dogs particularly useful in mapping complex traits related to morphology, behavior, and disease. The dog's demographic history has also had a profound effect on the canine genome in levels of linkage disequilibrium among breeds (Sutter et al. 2004), making the dog an ideal model organism. Although the history of the domestic dog has been extensively studied, much remains to be discovered about dog domestication and the formation of individual dog breeds. In researching the dog's demographic history in detail, we can obtain insight into the effects of domestication on the dog genome and aid studies currently underway to map complex traits using *Canis familiaris* as a model system. Here, we provide an overview of past canine research and introduce the demographic history of the dog.

Recent Developments in the Domestic Dog

The domestic dog, *Canis lupus familiaris*, is an ideal model organism with continually improving genetic resources. In 2003, a radiation hybrid map of the dog was published (Guyon et al. 2003), as well as a 1.5x genome sequence of the dog obtained from a male standard poodle (Kirkness et al. 2003). In 2005, a 7.5x coverage sequence of a Boxer was published (Lindblad-Toh et al.), increasing our knowledge of the dog genome as well as the number of tools currently available for research of the dog.

Current research demonstrates several salient features that make the domestic dog a particularly useful model organism.

The dog is a model system well suited to mapping human disease genes, as along with sharing our environment, many dog breeds are at high risk for the same diseases seen in humans. These diseases include cancer, epilepsy, thyroid disorders, allergies, heart disease, and many others (Sutter and Ostrander 2004). Several such diseases have already been extensively studied in the dog, such as hip dysplasia and Addison's disease (Chase et al. 2004; Chase et al. 2006).

In addition, the fact that many dog breeds share the same morphological and behavioral characteristics such as retrieving abilities, achondroplasia, tail wagging, and other traits can be harnessed in genetic studies. Dog breeds genetically cluster given their roles in human activities, geographic location, or morphological characteristics; main clusters that have been found are ancient breeds such as the Akita and Shiba-Inu, mastiff breeds such as the Mastiff, Bullmastiff, and Boxer, and herding dogs such as the Belgian Sheepdog and Collie (Parker et al. 2004). Several recent studies highlight the use of similarities between breeds to map complex traits. Sutter et al. (2007) identified a gene, IGF1 (encoding insulin-like growth factor 1), which appears to play a major role in body size in all small dogs. In addition, Mosher et al. (2007) used the whippet to link athletic performance to a genetic basis, where heterozygotes of a mutation in the myostatin gene are seen to have an increased racing speed.

One of the most useful features of the domestic dog genome is the extent of linkage disequilibrium (LD), or non-random association of alleles, among dog breeds. As a result of selective breeding and small founding populations of most breeds, LD is

approximately 20-50 times more extensive within dog breeds than in humans (Ostrander and Wayne 2005). Long-range LD extends furthest in rare breeds such as the Akita and Bernese Mountain Dog, with the least extensive LD in more common breeds such as the Labrador and Golden Retrievers (Sutter et al. 2004). This makes association mapping in dogs less costly than in humans, as using dogs can decrease the number of genetic markers needed by nearly two orders of magnitude (Sutter et al. 2004). Harnessing the extent of LD for use in discovering genes associated with diseases and other morphological traits is a very exciting area for future research.

From this brief overview of recent research, it is clear that the domestic dog is a very promising model organism. In its tractability for gene mapping, canine research has the potential to be immensely powerful in discovering the genetic basis for complex traits, many of which are also seen in humans. Of additional interest are the genetic bases of breed-specific behaviors and genes associated with domestication. However, we have only limited knowledge of the history of individual dog breeds and of the domestic dog as a whole (Sutter and Ostrander 2004). Studies of dog demography can be very useful in identifying particular breeds to study, in researching genes associated with domestication, and in discerning the effects of demography and other factors, such as selection, in the dog genome. We describe past research of the history of the domestic dog, highlighting the focus of this research study.

Dog Demographic History

The domestic dog is classified in the order Carnivora in the family Canidae along with its closest relative, the gray wolf (*Canis lupus*). Mitochondrial DNA sequence analysis appears to unambiguously support the classification of the gray wolf as the dog's

closest relative, with mtDNA sequence differing less between the wolf and dog than between the wolf and the coyote, the wolf's closest wild canid relative (Wayne 1993). Although there is little debate regarding the dog's closest relative, the exact details of the domestication of the dog remain uncertain.

Currently, there exist many plausible estimates of the timing of dog domestication. Archaeological evidence points to an origin roughly 12-15,000 years ago (Olsen 1985). Even among archaeologists there exists debate, however, as insufficient amounts of canid archaeological material often make distinctions between a domesticated gray wolf and domestic dog unclear (Olsen 1985). Genetic evidence may support a more ancient origin of domestic dogs. Through the examination of linkage disequilibrium among various dog breeds, Lindblad-Toh et al. (2005) suggest domestication occurred approximately 27,000 years ago. In another study of mitochondrial DNA control region sequences, high divergence between dog and gray wolf sequences indicates a timing of domestication of as early as 135,000 years ago (Vilá et al. 1997). The authors attribute the difference in the fossil record and their estimate to the fact that domesticated dogs may not have been morphologically distinct from the gray wolf until the transition to hunter-gatherer societies 10-15,000 years ago, possibly causing morphological changes in the dog. As there are limitations to studies performed using mitochondrial DNA, such as strictly maternal inheritance, this motivates the need for analyses of nuclear DNA.

Other than the issue of timing, there is the issue of the location and number of founding events of the domestic dog. It is believed that both New and Old World domestic dogs originated from the Old World, without an independent domestication

event in the New World (Leonard et al. 2002). East Asia has been proposed as the location in the Old World from which dogs have originated (Savolainen et al. 2002).

Examination of diversity among dogs can also provide insight into other questions of domestic dog origins. If the dog were domesticated from only a small number of gray wolves, one would see very little diversity among today's dogs. In contrast, high levels of diversity in the dog could be maintained through continued interbreeding between dogs and wolves or multiple domestication events. From examination of MHC genes, there is evidence that introgression often occurs between domesticated species, such as cattle and pigs, and their wild ancestors (Vilá et al. 2005). This trend also appears to apply to the domestic dog. Most mitochondrial DNA analyses suggest origins of dogs in multiple locations or continued admixture between dogs and wolves (Vilá et al. 1999a). Tsuda et al. (1997) find evidence for admixture between dogs and wolves in the matriarchal origins of dogs, and Randi and Lucchini (2001) detect introgression and admixture of rare domestic dogs genes in the wild gray wolf. Continued breeding with wolves likely acted to maintain diversity in the domesticated dog population, though whether this was done by humans intentionally is still debatable (Vilá et al. 2005).

Humans have in fact played an extremely large role in the creation of today's diverse dog breeds. Currently, over 400 domestic dog breeds exist, most of which are less than 400 years old. In 2003, the American Kennel Club (AKC) had roughly 916,000 dog registrations, with the two most popular breeds (the Labrador Retriever and the Golden Retriever) making up 16% and 6% of all breeds respectively (Sutter and Ostrander 2004). It is believed that today's dog breeds were formed not from a highly inbred, but rather from a genetically diverse, ancestral dog population (Vilá et al. 1999a).

While this explains the diversity seen among dog breeds, the exact history of individual breeds is unclear. No kennel club, and therefore few systematic records of dog breeding, existed prior to 1873 (Dangerfield and Howell, 1971). As a result, study of the formation of individual dog breeds, as well as dog domestication, are areas of interest.

While much is known about the demographic history of the domestic dog, much remains to be discovered that could potentially aid the mapping of complex traits in the dog or provide insights into human history during dog domestication. We hope to contribute to the developing field of dog genetics in a more thorough study of dog demography, making future research developments in dogs and humans more promising. In this study, we draw independent conclusions regarding questions of the severity of a population contraction at dog domestication and in the formation of several dog breeds. We also examine the history of several wild canid populations, linking the demographic history of the dog to that of its closest ancestors.

II. Demographic Models

We are interested not only in the details of the domestication event when the dog diverged from the gray wolf but also in the more recent formation of individual dog breeds. In order to model this history, we considered a two-stage bottleneck model similar to Lindblad-Toh et al. (2005). A graphical representation of the model is shown in Figure 1.

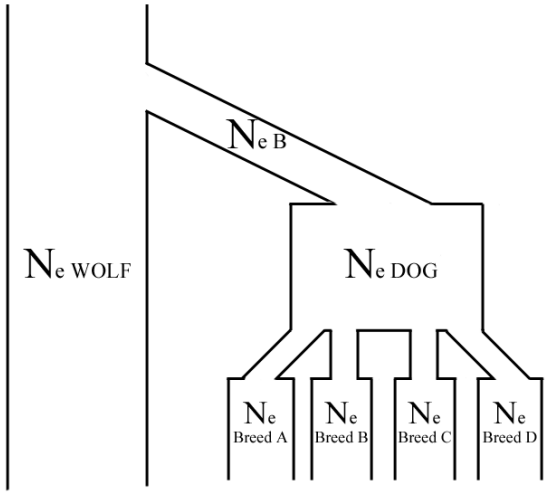

Figure 1. Demographic model of domestic dog origins, from past to present. N_{eWOLF}, N_{eB}, N_{eDOG}, and N_{eBreed} are the effective population sizes of wolf, during the domestication bottleneck, of dogs after the bottleneck, and of individual breeds, respectively.

The model of Figure 1 assumes that gray wolves have maintained a constant population size (N_{eWOLF}) throughout time. The founding event of dogs from the gray wolf is characterized by a bottleneck of size N_{eB}, lasting until the dog population expands to a size N_{eDOG}. Individual dog breeds are then formed, each characterized by their own unique founding events and bottlenecks. Current breed effective populations sizes are denoted by $N_{eBreed\ A,\ B,\ C,\ and\ D.}$ In our study, we research these demographic models in two parts – one, for domestication, and second, for breed formations.

The model we propose is rather simplistic, not accounting for the possibility of continued interbreeding between dogs and wolves, multiple domestication events, gene

flow between dog breeds, or subdivision among wolf populations. We also assume a constant wolf effective population size (N_{eWOLF}), although several wolf populations are known to have undergone severe population size changes (Blanco et al. 1992; Wayne et. al 1992). Remarks on the validity of these and other assumptions will be discussed in the analyses to follow. [For analysis of wolf population structure and demography, see VII. Demographic Analysis of Wild Canids].

Domestication Model

For inference of the domestication event, we analyze the demographic model shown in Figure 2. The wolf population is assumed constant throughout time, and τ, the time of the domestication event, is assumed to be 15,000 years from the present (Olsen 1985). Three unknown parameters are to be estimated. The first is τ_B, the length of the domestication event. The second is ω_B, the bottleneck population size scaled by N_{eWOLF}, or N_{eB}/N_{eWOLF}. The third parameter is ω, the scaled domesticated dog population size after the bottleneck, or N_{eDOG}/N_{eWOLF}.

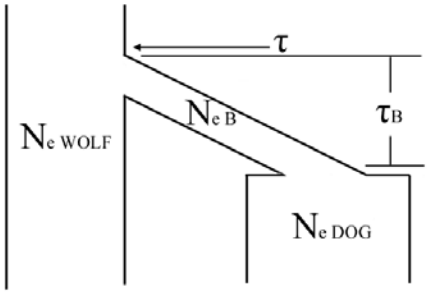

Figure 2. Demographic model of dog domestication event, from past to present. N_{eWOLF}, N_{eB}, and N_{eDOG} are the effective population sizes of wolf, of the population during the domestication bottleneck, and of dogs after the bottleneck, respectively. τ is the time of the domestication event from the present, and τ_B is the bottleneck duration.

Breed Formation Model

A similar model describes the formation of an individual dog breed (Figure 3). The breed is formed at time τ from the present from an ancestral "pre-breed" dog population of size N_{eDOG}. The founding population of the breed (N_{eB}) lasts for time τ_B until the population size expands to its current effective size, N_{eBreed}. All parameters (τ, τ_B, ω (N_{eDOG}/N_{eBreed}), and ω_B (N_{eB}/N_{eBreed})) are to be estimated, modeling the intense selective breeding involved in the formation of a breed.

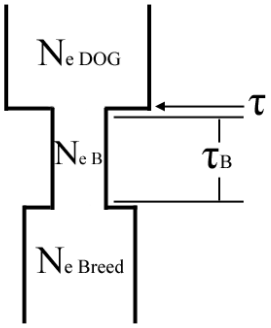

Figure 3. Demographic model of dog breed formation, from past to present. N_{eDOG}, N_{eB}, and N_{eBreed} are the effective population sizes of pre-breed dogs, of the population during the breed formation bottleneck, and of the current breed, respectively. τ is the time of the breed formation event from the present, and τ_B is the bottleneck duration.

III. Materials

Available Data

Data analyzed is obtained primarily from Sutter et al. (2004) in the analysis of the extent of linkage disequilibrium in 17 Akita and 20 each of the Bernese Mountain Dog, Golden Retriever, Labrador Retriever, and Pekingese. These five dog breeds were chosen to encompass a wide range of breed histories: while the Labrador and Golden Retrievers are both more common breeds, the Akita, Bernese Mountain Dog, and Pekingese are all rarer breeds with more severe population declines in their histories. Segments of ordered synteny found by comparing the 1.5x standard poodle sequence and the human genome were sequenced on canine chromosomes 1, 2, 3, 34, and 37 (Sutter et al. 2004). Single nucleotide polymorphisms (SNPs) were discovered by resequencing in 95 dogs, including all of the aforementioned dogs except two Akitas, to result in a total of 200 SNPs. The total length sequenced on these five chromosomes is 52,018 bp, determined as the summed length of all amplicons sequenced. Additional details of the SNP discovery can be found in Sutter et al. (2004). We refer to this original data from the five dog breeds as the "sequence" data.

A subset of 106 out of the total 200 SNPs ascertained by Sutter et al. (2004) were genotyped by Gray et al. (in prep) in an additional 17 dog breeds (listed in Appendix Table 1). This results in a total of 22 dog breeds (577 dogs) available for analysis across the 106 SNPs. We refer to this data as the "genotype" data. In addition, SNPs are genotyped in the Golden Jackal (*Canis aureus*), whose genotype in each position is assumed to be the ancestral base. This information is used to root all ascertained SNPs.

Regions on chromosome 1 were resequenced in the original five dog breed samples of Sutter et al. (2004) as well as in four gray wolf populations, a coyote (*Canis latrans*) population, and two Golden Jackals (Gray et al., in prep). The gray wolf populations are from four geographic locations: Alaska (n = 19), Israel (n = 14), Spain (n = 20), and Yellowstone National Park (n = 20). These four wolf populations, as well as the coyote population, are analyzed using this sequence data on chromosome 1. In total, 11,279 bp were sequenced on chromosome 1, again determined as the summed lengths of amplicons.

Sequence data obtained from the five initial breeds (Akita, Bernese Mountain Dog, Golden Retriever, Labrador Retriever, and Pekingese) was phased by the program PHASE (Gray et al., in prep; Stephens & Donnelly 2003; Stephens et al. 2001). Uninformative SNPs, sites segregating in the Golden Jackal or for which the Golden Jackal had an unknown genotype, were excluded (Appendix Table 2). In the genotype data for the total of 22 breeds, 24 sites uninformative in rooting the SNPs were excluded, reducing the genotyped SNP count to 82 (Appendix Table 3).

Preliminary Statistics

The effective population size of wolves was estimated from the phased sequence data on chromosome 1. All wolf populations from Alaska, Israel, Spain, and Yellowstone National Park were pooled (2n = 144). Using the number of segregating sites among wolves (S = 54), Watterson's (1975) estimate of $\theta = 4N\mu$ for the entire region is 9.74105, while the per-bp θ is 0.00086. This value is rather similar to the value seen in both dogs and humans (Parker et al. 2004). Using a mutation rate μ of 1×10^{-8}

per generation (Lindblad-Toh et al. 2005), the estimated current effective population size of wolves is approximately 21,591.

Additional summary statistics for this data are shown in Table 1, calculated by programs written in Python for more flexibility in the analysis. Statistics are obtained pooling all wolf populations, and for each wolf population individually. Statistics are also obtained for the sequence data for all five chromosomes in the five original dog breeds. Results from combining all chromosomes are shown in Table 2, both for all dogs pooled and for each dog breed separately.

Diversity levels indicated by π and θ in wolves and dogs are rather comparable. Values for Tajima's D, which compares values of π and θ, appear to be rather positive for all dog breeds and the Spanish wolf population. Under a population decline, there will be fewer recent mutations contributing to the number of segregating sites, making the value of Tajima's D positive (Tajima 1989). Though we do not assess the significance of these values, a positive Tajima's D could be indicative of a population decline in these breeds and populations. In addition, θ, indicating levels of diversity, is lowest for the Pekingese and Bernese Mountain Dog and highest for the Akita. There are slight differences in values of π between breeds, with the Bernese Mountain Dog having the lowest nucleotide diversity and Akita the greatest. We will explore these statistics in further detail in later sections.

Table 1. Summary statistics of sequence data for wolves. Data is obtained from chromosome 1 only, with a summed length of amplicons of 11,279 bp.

Wolf Population	**2n**	**Segregating Sites**	**θ (Watterson)**	**θ (per site)**	**Number of Singletons**	**π (per site)**	**Tajima's D**	**Average Heterozygosity**
All Wolves	144	54	9.74105	0.00086	4	0.00118	1.11374	0.24390
Alaska Wolf	38	41	9.75822	0.00087	4	0.00103	0.67976	0.27613
Israel Wolf	28	18	4.62552	0.00041	3	0.00044	0.24078	0.26488
Spain Wolf	40	34	7.99333	0.00071	1	0.00112	2.02833	0.36298
Yellowstone Wolf	38	49	11.6623	0.00103	5	0.00120	0.56350	0.26816
Coyote	36	48	11.5752	0.00103	10	0.00095	-0.25822	0.21769

Table 2. Summary statistics of sequence data for dogs. Data is pooled from chromosomes 1, 2, 3, 34, and 37, with a summed length of amplicons of 52,018 bp.

Dog Breed	**2n**	**Segregating Sites**	**θ (Watterson)**	**θ (per site)**	**Number of Singletons**	**π (per site)**	**Tajima's D**	**Average Heterozygosity**
All Dogs	194	188	32.17804	0.00062	4	0.00095	1.70429	0.26192
Akita	34	138	33.75075	0.00065	10	0.00088	1.33231	0.32115
Bernese Mountain Dog	40	102	23.98001	0.00046	5	0.00066	1.58661	0.32850
Golden Retriever	40	128	30.09256	0.00058	3	0.00075	1.09292	0.29723
Labrador Retriever	40	124	29.15217	0.00056	11	0.00075	1.21510	0.30494
Pekingese	40	104	24.45021	0.00047	5	0.00076	2.26648	0.37097

IV. Methods

We use the program PRFREQ (Williamson et al. 2005) to estimate the demographic parameters of our given models (Figure 2 and Figure 3) in a composite likelihood framework. First, we introduce the basics of relevant population genetic theory including coalescent theory, which describes the history of a sample of DNA sequences. The implications of demographic history on observed sequence data will also be discussed, most importantly in how it relates to the demographic modeling to follow. We then discuss the analysis program PRFREQ, along with a method involving coalescent simulations to assess the significance of our demographic models.

Theory Background

Coalescent theory is a very powerful theory describing the genealogical history of a sample of DNA sequences. The coalescent involves tracing a sample of genes backwards throughout time until all genes in the sample "coalesce," or share a common ancestor. In other words, this is an implementation of "identity by descent" for a sample of genes (Kingman 2000). Under the Wright-Fisher model of random mating and constant population size, the probability that two genes in a sample will coalesce in the previous generation is 1/N, where N is the number of genes in the sample. An example of a coalescent tree is pictured in Figure 4.

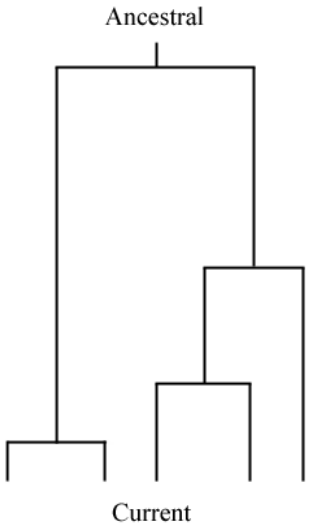

Figure 4. Example of a coalescent tree. Lower, external branches represent the current sample of sequences (n = 5), and the upper node represents the common ancestor of all the sequences. Lengths of branches represent the time between coalescent events.

The coalescent can be placed in a statistical framework. Coalescent times, T_i, denote the time it takes for a sample having *i* ancestors to have *i-1* ancestors. These T_i are distributed exponentially with expected value 2/[(i)(i-1)], scaled in units of N generations (Kingman 1982). As can be seen by this formula and in Figure 4, coalescent times increase as the number of ancestors decreases (i.e., T_2, the time until the last coalescent event, is the longest). In order to model the segregating sites on a given coalescent tree, mutations are distributed according to the Poisson distribution with rate $\theta/2$ per lineage (Kingman 1982), where $\theta = 4N\mu$ and μ is the mutation rate. Therefore, longer branches in the coalescent tree will accumulate more mutations. In modeling coalescent times and mutations, coalescent theory can explain the distribution and number of segregating sites in an observed sample of sequences. In the statistical framework of coalescent theory, we can later incorporate population size changes, selection, and other factors. Using the coalescent is extremely helpful when dealing with

DNA sequence data and when generating random samples under particular demographic models.

Demographic inferences can be made by examining the site frequency spectrum (SFS), a method of summarizing single nucleotide polymorphism (SNP) data that provides information about the history of a sample of DNA sequences. The "unfolded" SFS is a vector, $x = (x_1, x_2, x_3, \ldots, x_{n-1})$, obtained from a sample of *n* sequences. Each entry x_i denotes the number of SNPs with derived allele at frequency *i* out of *n* in the sample. Generally, the ancestral state is inferred from an outgroup species, where the outgroup genotype is assumed to be the "ancestral" allele and the other the "derived" allele. If the ancestral state of each SNP is unknown, we must construct the "folded" site frequency spectrum, where each entry $\zeta_i = x_i + x_{n-i}$. The sum of the entries in the site frequency spectrum is the number of segregating sites in the population, or S.

According to coalescent theory, under neutrality the expected entry x_i of the site frequency spectrum is θ/i (i.e., the expected number of singletons, x_1, is equal to θ). An example of a SFS under neutrality is shown in Figure 5. Under coalescent theory, this expectation can be violated by a number of deterministic and stochastic factors, such as substructure, natural selection at linked sites, population size changes, or a combination of these. Because of this, examining the site frequency spectrum and its deviations from neutrality will be extremely informative when inferring the demography of the domestic dog.

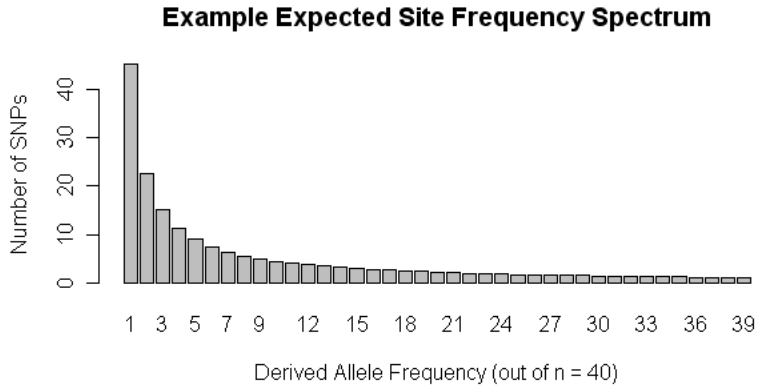

Figure 5. Expected unfolded site frequency spectrum under neutrality for a sample of 40 sequences. x-axis is the derived allele frequency out of 40, and the y-axis is the number of SNPs with derived allele at that frequency.

Both population genetic and coalescent theory describe the effect of deviations from neutrality on the site frequency spectrum (for more information, see Wakeley 2007, in press). To picture these scenarios, we use the coalescent and look backwards in time. While long external (current) branches translate to an increase in rare alleles that are not shared by many sequences in the sample, long internal branches translate to an increase in middle to high frequency derived alleles. Under a situation of population growth, the external (current) branches of the coalescent tree must coalesce before the population becomes smaller in the past. This results in a "star-shaped" genealogy, a coalescent tree with very long external branches and an increase in rare alleles or singletons. For a population that has declined in size, there are shorter external branches and longer internal branches when the population was larger. More mutations will accumulate on these internal branches, resulting in an excess of middle to high frequency derived alleles. Substructure and isolation also affect the site frequency spectrum, where isolation results in a long time before the subpopulations are joined by a coalescent event. As a result of these long internal branches, we see an excess of middle frequency derived alleles.

As described above, since population size changes can have large effects in the site frequency spectrum, we use the SFS to infer demographic parameters governing both domestication and breed formation.

Analysis program PRFREQ

We use the program PRFREQ (Williamson et al. 2005) for inference of demography. The program was initially developed to jointly infer selection and demography for putatively neutral and selected site frequency spectra. Since selection does not play a role in our demographic inference, as the noncoding sites we observe are assumed to be neutral, we ignore the selection aspect of the program and work only with its inference of population size changes. The program does so in a maximum likelihood framework, finding the predicted site frequency spectrum under given demographic models.

The framework of the program is the Poisson Random Field (PRF) approach (Sawyer and Hartl 1992), which uses single-locus diffusion theory to predict the distribution of allele frequency across sites. Diffusion theory describes the random motion of particles in a set (Sawyer 1976) and can directly be applied to the "diffusion" of alleles in a population. The model assumes the two-allele Wright-Fisher model of mutation, with non-overlapping generations and random mating. The approach also assumes that all sites examined are in linkage equilibrium; i.e., that all sites are unlinked and independent.

According to theory, the expected number of SNPs x_i where *i* sites have the derived allele and *n-i* have the ancestral allele (and $i = 1, 2, 3, \ldots, n-1$) are distributed according to a Poisson distribution (Hartl 1994), the mean of which follows from the

equilibrium densities under the Wright-Fisher model (Sawyer and Hartl 1992). Sawyer and Hartl (1992) derive this result using the stationary solution to the derived diffusion equations, assuming no changes in population size. With changing population sizes, such as a contraction of severity ω at time τ in the past, the transient solution to the diffusion equation is used (Williamson et al. 2005). Classifying mutations as occurring either before the population size change or after the population size change, one obtains an equation for the distribution of allele frequency across sites given the parameters ω and τ. The expected value of each entry in the SFS is $E(x_i|\tau,\omega) = \theta F(i)$, where F(i) is found as described in Williamson et al. (2005).

In the PRFREQ program, there are two calculations that can be performed given a particular demographic history. The first is the multinomial calculation, which does not require an *a priori* estimate of θ. The multinomial calculation calculates the probability that a given SNP is segregating at derived allele frequency *i* out of *n*, where *i* = 1, 2, 3, …, *n-1* (Williamson et al. 2005). A cancellation of terms involving θ makes this probability independent of the mutation rate, as the denominator of the probability sums over all possible frequency classes. We find the likelihood of the observed SFS given the demographic history described by τ and ω by multiplying over all frequency classes in the SFS (Equation 1). In this equation, *n* is the sample size, x_i is the number of alleles with derived frequency *i* out of *n*, and $F(i|\tau,\omega)$ is found using Williamson et al. (2005).

Equation 1.

$$L(x \mid \tau, \omega) = \prod_{i=1}^{n-1} \left(\frac{F(i \mid \tau, \omega)}{\sum_{j=1}^{n-1} F(j \mid \tau, \omega)} \right)^{x_i}$$

Because this calculation does not depend on the mutation rate and is based only on the shape of an observed site frequency spectrum, if an estimate of θ is not available, the multinomial calculation should be used.

PRFREQ can also perform a calculation using the fact that the number of SNPs in each frequency class is distributed according to the Poisson distribution with mean $E(x_i|\tau,\omega) = \theta F(i|\tau,\omega)$ (Bustamante et al. 2001). Here, a known value of θ is required for the calculation. As in the multinomial calculation, we calculate the likelihood of the observed data by multiplying over all classes in the site frequency spectrum (Equation 2). Here, *n-1* is the number of classes in the SFS, and x_i is the number of SNPs with derived frequency *i* out of *n*.

Equation 2.

$$L(x \mid \tau, \omega) = \prod_{i=1}^{n-1} \exp[\theta F(i \mid \tau, \omega)] \frac{[\theta F(i \mid \tau, \omega)]^{x_i}}{x_i!}$$

Given an observed SFS, we can find the maximum likelihood estimates (MLEs) of the demographic parameters using both the Poisson and multinomial calculations. PRFREQ calculates the likelihood of the data (using either Equation 1 or Equation 2) for given ranges of the demographic parameters of interest and returns the parameter combination with the highest likelihood. After examining the results of PRFREQ, we manually adjust the parameter ranges to find the MLEs over the entire likelihood surface. In addition to obtaining the MLEs and likelihood of the data under particular demographic models, we can also easily obtain the likelihood of the data under the neutral model. In this case, τ, the time of the population size change, occurs effectively at a time ∞ from the present, whereas ω, the ratio of the current and effective population sizes, is 1.

Given a manageable likelihood framework, we can use a likelihood ratio test to assess the significance of incorporating additional demographic parameters into our demographic models. Under a null hypothesis of constant population size, the likelihood ratio test statistic is equal to $2\log[\mathbf{L}(\tau, \omega)/\mathbf{L}(\infty, 1)]$, which, when maximum likelihood estimates of both τ and ω are calculated, has approximately a χ^2 distribution with 2 degrees of freedom (Williamson et al. 2005).

The likelihoods of the multinomial and Poisson calculations cannot be directly compared, given that the multinomial is dealing with proportions of SNPs and the Poisson is dealing with actual numbers of SNPs given a value of θ. To make the likelihoods comparable between the two calculations, we calculate the maximum likelihood estimate of θ used in the multinomial calculation (Equation 3), where $F(i|\tau,\omega)$ is calculated given the final estimates of τ and ω obtained from the multinomial calculation, and S is the number of segregating sites observed in the data.

Equation 3.

$$\hat{\theta} = \frac{S}{\sum_{i=1}^{n-1} F(i \mid \tau, \omega)}$$

Substituting this value of θ into the Poisson likelihood equation (Equation 2) results in a "rescaled" value of the multinomial likelihood, allowing the likelihoods from the multinomial and Poisson calculations to be compared using the likelihood ratio test statistic ($2*(L_{Multinomial}-L_{Poisson})$). Since θ is effectively maximized in the new multinomial likelihood, whereas the Poisson calculation requires a given value of θ, the multinomial likelihood has one more degree of freedom. A p-value can be calculated using the χ^2

approximation with 1 df, which can indicate whether allowing θ to vary from the given value greatly increases the likelihood.

An important caveat is that the preceding discussion of methods assumes that observed SNPs are unlinked. However, this assumption does not hold for our data set. SNPs that we observe are tightly linked within amplicons, the regions of DNA amplified for sequencing, which range from roughly 500 to 700 bp in length. In contrast, SNPs are nearly independent between amplicons, some of which lie nearly 1 Mbp apart. Because we do not incorporate this linkage among sites, the calculations we make are based on the composite-likelihood function, which should be interpreted as an approximation of the true likelihood function (Caicedo et al. 2007). The true likelihood function, in contrast to the composite-likelihood function, would explicitly take into account linkage among observed SNPs.

The program PRFREQ was adjusted to infer the specific demographic models of dog domestication and breed formation (Figure 2 and Figure 3), estimating τ_B and ω_B, the length and severity of a bottleneck, as well as τ and ω. Scaling of time and size change parameters can be done either in terms of the ancestral or the current effective population size. Details on the particular demographic models tested as well as the specific likelihood ratio tests conducted are to be described in later sections.

Coalescent Simulations

As previously mentioned, the analysis of PRFREQ assumes that all observed sites are unlinked and independent. In our data set, we are dealing with closely linked sites within amplicons. Using msHOT (Hellenthal and Stevens 2007), we simulate data to account for increased recombination between amplicons but tight linkage within

amplicons. msHOT is a modification of ms (Hudson 2002), a program popularly used to generate samples under the coalescent model. While ms assumes a constant recombination rate across an entire region, msHOT allows for recombination "hotspots," or areas of increased recombination, along a chromosome. We simulate data separately for each chromosome given the unique lengths between amplicons, modeling the spaces between amplicons as recombination "hotspots." Using msHOT we can more efficiently simulate only the amplicons, rather than the entire chromosome, while accounting for recombination. In doing so, we can observe how linkage affects the analysis of PRFREQ, comparing the results of the coalescent simulations to our observed data.

Input for msHOT requires the number of hot spots (in our case, equal to the number of amplicons on the chromosome minus 1), the start and end site of each amplicon (denoting the length of the amplicon), and the intensity of each hotspot in comparison to the background recombination rate (the distance in base pairs between two amplicons). Under the neutral model, we perform 2000 coalescent simulations and obtain the SFS from each simulation. We use these simulations as another method of calculating model significance aside from the χ^2 approximation of the likelihood ratio test statistic described above. Further details of these simulations will be described in later sections.

V. Demographic Analysis of Domestication Event

Analysis with PRFREQ

Inference of demographic parameters of the dog domestication event (Figure 2) was made using the program PRFREQ (Williamson et al. 2005). In order to perform a statistically rigorous comparison between demographic models, a nested likelihood ratio approach was taken. The nested models tested are shown in Table 3, where parameters are explained in II. Demographic Models, Domestication Model.

Table 3. Nested likelihood models used in inference of the domestication event. Parameters of each model, as well as their associated degrees of freedom, are given.

Model	Parameter	df
A0 (Stationary demography)	None	0
A1 (Size change at domestication)	τ = fixed (15,000 years) ω = vary	1
A2 (Size change at any time in past)	τ = vary ω = vary	2
A3 (2 size changes – bottleneck at domestication and after)	τ = fixed (15,000 years) τ_B = vary ω_B = 0.1 ω = vary	2
A4 (2 size changes – at domestication and after domestication)	τ = fixed (15,000 years) τ_B = vary ω_B = vary ω = vary	3

First, we assume for the A1, A3, and A4 models that domestication occurred approximately 15,000 years ago, or 5,000 generations ago assuming a generation time of 3 years (Mech and Seal 1987; Vilá et al. 1999b). Because we assume wolves have maintained a constant population size throughout time equal to 21,591 (calculated in III. Materials, Preliminary Statistics), we scale all values by the ancestral wolf effective population size. ω, the parameter indicating the severity of the population size change, is

equal to N_{eDOG}/N_{eWOLF}, and ω_B, indicating the severity of the bottleneck, is equal to N_{eB}/N_{eWOLF}. An ω greater than 1 indicates a population expansion.

Significance of the improvements between models is assessed by the likelihood ratio test statistic, with p-values estimated from the χ^2 distribution with degrees of freedom equal to the difference in the degrees of freedom of the models in question. A significant difference in the likelihoods of the A0 and A1 models is evidence of a size change at dog domestication. A significant difference between models A2 and A1 is evidence of a population size change that occurred at some time other than 15,000 years ago. If the maximum likelihood of A3 is significantly greater than that of A1, there is evidence of a 10-fold contraction during a population bottleneck rather than a simple population contraction. Finally, if the A4 model has a significantly higher likelihood than the A1 model, we have significant evidence of a bottleneck with ω_B taking on a value other than 0.1. Although there are additional model selection criteria aside from the likelihood ratio test that could be used, in this analysis we primarily use the likelihood ratio test.

In order to perform coalescent simulations with msHOT as described, we use a background recombination rate equal to the per-bp θ of wolves (0.00086) and multiply this by the number of base pairs in all amplicons of the chromosome. For an estimate of $\rho = 4Nr$, we use an r equal to 1×10^{-8}, which makes ρ effectively equal to θ. Input used for msHOT for the domestication is shown in Appendix Table 5.

We simulate 2000 samples under the neutral model with no change in population size, where the current effective population size is the same as in wolves (~ 21,600). We obtain the SFS from each sample and obtain the multinomial likelihood of each under the

neutral A0 model with PRFREQ. We analyze only multinomial likelihoods, as using the Poisson calculations may be biased by an improper value of θ used in the coalescent simulations. We also optimize demographic parameters with PRFREQ under the A1 model for the 2000 neutral samples, keeping τ constant at 5000 generations but allowing ω, the severity of the contraction, to vary between 0.1 and 3.1. We examine the difference in likelihood under the contraction model and the neutral model to supplement the p-values we obtain from the approximation to the χ^2 distribution.

Data Manipulation

In order to infer demography of the initial domestication event, we want to use as input to PRFREQ a site frequency spectrum that represents the ancestral "pre-breed" dog population that existed after dogs diverged from wolves. We do this in a number of ways, using both the sequence data from the five breeds initially sequenced as well as the genotype data from the additional 17 breeds.

Sequence Data

In order to infer the domestication event, all dog breeds are pooled together into one population in an attempt to represent the "ancestral" dog population. The site frequency spectrum of the observed sequence data is examined pooling data across all five chromosomes (Figure 6), using the Golden Jackal as the outgroup to root each given SNP as either ancestral or derived. The observed SFS is compared to the expected SFS under neutrality, obtained using Watterson's estimate of θ (32.178) from the number of segregating sites (S = 188). The observed deviations from the expected SFS has several causes, most notably population subdivision among breeds and the strong recent bottlenecks of individual breeds. The effect of subdivision in the SFS is clear in the

deficiency of rare alleles and a perceived excess of intermediate frequency variants. Recent population contractions from breed formation also show their effect by decreasing the expected number of rare alleles.

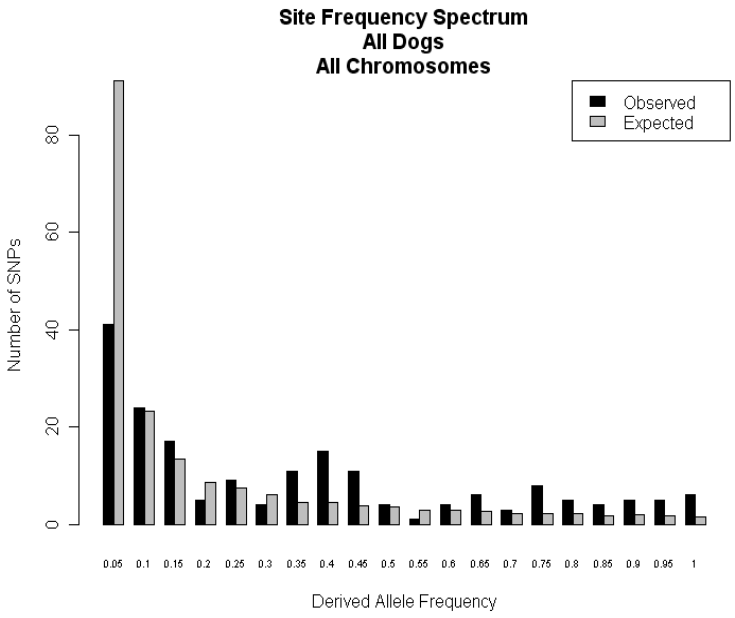

Figure 6. Site frequency spectrum for all chromosomes and all dog breeds pooled. Black bars indicate observed data, and gray bars indicate the expectation under neutrality. x-axis is the derived allele frequency, and the y-axis is the number of SNPs with derived allele frequency less than or equal to the value on the x-axis.

In order to reduce the signatures of individual breed bottlenecks and population subdivision to have a more accurate inference of the domestication bottleneck, a sampling method was used. For every SNP, one allele was sampled from each of the five dog breeds, and the number of derived alleles in the sample of five is counted. For each SNP, this is done 2000 times, and an average of the number of derived alleles for each SNP is obtained over the 2000 iterations. These average counts are rounded and the SFS is created from these averages, counting the number of SNPs at frequency 1/5, 2/5, 3/5 and 4/5 (Figure 7). Expected values under neutrality are obtained from the same method as above, with $\theta = 54.24$ from the number of segregating sites.

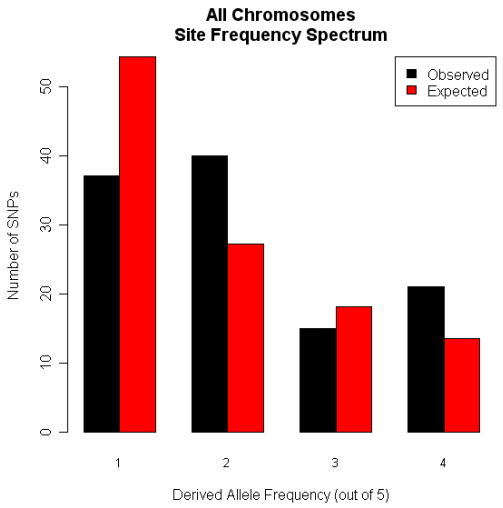

Figure 7. Site frequency spectrum of data sampled by each SNP as described in text, where black bars indicate observed data, and red bars indicate the expectation under neutrality. x-axis is the derived allele frequency out of 5, and the y-axis is the number of SNPs with derived allele at that frequency.

To ensure that this sampling method did not drastically change our results, we also performed another sampling method where we sample each chromosome, rather than each SNP, individually. For each chromosome, we sample one chromosome from each of the five breeds and pool all chromosomes to construct the SFS for that particular sample. We perform this 2000 times, and average the site frequency spectra from each run. The results of this are shown in Figure 8, where expected values are obtained as outlined above with $\theta = 59.216$.

In addition, general summary statistics of each sequence data set were obtained (Table 4). Values of the statistics are relatively comparable between the two sampling methods. Interestingly, however, while Tajima's D is negative for the data set sampled by SNP, Tajima's D is positive for the data set sampled by chromosome (Tajima 1989). However, these differences may not be significant. Nucleotide diversity, or π, is nearly identical between the two sampling methods.

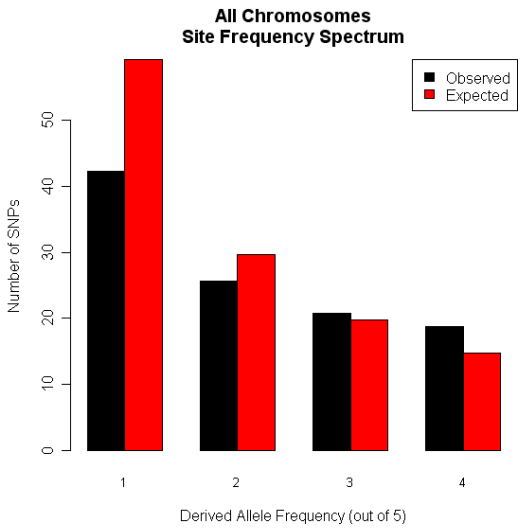

Figure 8. Site frequency spectrum of data sampled by each chromosome as described in text, where black bars indicate observed data, and red bars indicate the expectation under neutrality. x-axis is the derived allele frequency out of 5, and the y-axis is the number of SNPs with derived allele at that frequency.

Table 4. Summary statistics obtained for sequence data sets, sampled by SNP and by chromosome as described in text. S is the number of segregating sites, and θ_W is Watterson's estimate of θ.

Data Set	**S**	**θ_W (per site)**	**π (per site)**	**Tajima's D**	**Average Heterozygosity**
Sampled by SNP	113	0.00104	0.001006	-0.26542	0.397876
Sampled by Chromosome	107.577	0.00099	0.001006	0.101698	0.375276

It is important to note that these sampling methods likely do not entirely reduce the strong effects of breed subdivision and breed formation in the SFS. The sampling methods performed assume that domesticated dogs originated by randomly breeding selected breed dogs, which is not entirely accurate. However, given our data, this approach was the most plausible to minimize subdivision between breeds and reduce the effects of demography within a breed. With additional data and more breeds from which to sample, as in the genotype data, these sampling methods would become more effective.

Genotype Data

In addition to using the sequence data from the five breeds alone, we also worked with inference of the domestication event using the genotype data from the additional 17 breeds. This dataset was extremely important to analyze, as in contrast to the sequence data with only five breeds, the genotype data contains data from 22 breeds. These additional breeds will add considerably more information to the site frequency spectrum representative of the ancestral dog population, and will likely increase the power of the demographic analyses conducted.

Since this genotype data was not phased, there are unresolved missing genotypes for the 82 SNPs remaining after removing uninformative sites in the Golden Jackal (Appendix Table 3). Out of a maximum sample size of 2n = 1154 for each SNP, the number of known genotypes (i.e. the sample size) for the 82 SNPs ranged from the lowest value of 170 to the highest value of 1022. An arbitrary sample size cutoff of 577 (half of the total value of 1154) was chosen, and three sites with a sample size less than this value were excluded (Appendix Table 4).

From this data, one cannot directly produce a site frequency spectrum with the typical categories of singletons, doubletons, and (*n-1*)-tons, as each SNP has a different sample size. In order to create a SFS with one sample size from SNPs with different sample sizes, we use the hypergeometric distribution to "project" a given site frequency spectrum to a particular sample size *n* (as in Clark et al. 2005). For each SNP with original sample size *N*, the probability P(x = i) of the SNP having *i* derived alleles out of *n* is calculated, taking the form of the hypergeometric distribution (Equation 4).

Equation 4.

$$P(x=i)=\frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

The two "classes" of elements are derived and ancestral alleles, where *m* and *N-m* are the original number of derived and ancestral alleles, respectively, and *i* and *n-i* are the "projected" number of derived alleles and ancestral alleles out of the new sample size, respectively. Note that *n* must be less than or equal to *N*, meaning that the projected SFS must have a sample size less than or equal to the sample size of each SNP. The probability P(x = i) is summed over all SNPs for a given *i*, generating the i^{th} entry of the site frequency spectrum. Again, this projection makes no assumptions regarding missing data, as each SNP is projected to a lower sample size than the original.

After pooling all individuals and SNPs with sample sizes above the cutoff value of 577, the lowest sample size was n = 628. The site frequency spectrum of the genotype data is projected to n = 628 using the hypergeometric projection, creating a SFS as if the sample size were only 628 individuals (Figure 9).

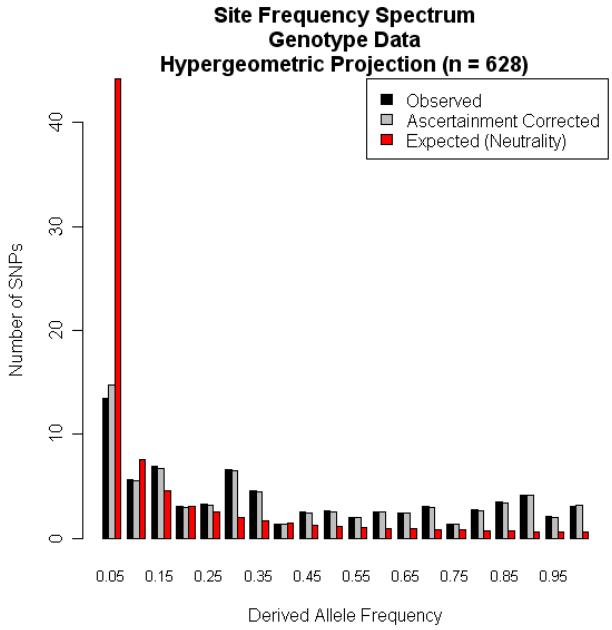

Figure 9. Site frequency spectrum of genotype data, using a hypergeometric projection to n = 628 as described in text. Black bars indicate observed data, gray bars indicate data corrected for SNP ascertainment as described in text, and red bars are the expectation under neutrality. x-axis is the derived allele frequency, and the y-axis is the number of SNPs with derived allele frequency less than or equal to the value on the x-axis.

Ascertainment Bias

Due to the manner in which the genotyped SNPs were ascertained in the additional dog breeds, there is a bias in the site frequency spectrum of the genotype data. The 82 SNPs we examine were not actually discovered in all dogs, but rather in the initial panel of 97 dogs in the five breeds of the Akita, Bernese Mountain Dog, Golden Retriever, Labrador Retriever, and Pekingese. When discovering SNPs in a small subset of the entire population, rare SNPs segregating in the larger population will likely not be discovered. In the site frequency spectrum, this translates to seemingly fewer low frequency derived alleles and a skew towards higher frequency alleles (Nielsen et. al. 2004).

In order to correct for this bias, we apply methods outlined in Nielsen et al. (2004) to correct the observed SFS for the bias of ascertaining SNPs in a small discovery panel.

We correct under the basic model, assuming all SNPs are ascertained at the same depth, *d*. Though this may be an oversimplication, since SNPs may have missing data and unequal sample sizes, we ignore this in the context of our analysis. To correct for ascertainment, we effectively find the maximum likelihood estimates of the true probabilities of each entry in the site frequency spectrum, p_i, given our observed values of the entries, x_i, where $i = 1, 2, 3, \ldots n-1$ and n is the sample size of the entire sample.

The probability of ascertaining a SNP of frequency *i*, given the observed data, is one minus the probability of not ascertaining the SNP (Equation 5.). Not ascertaining the SNP, or finding that the site is not segregating in the sample, involves sampling exclusively *d* ancestral alleles or *d* derived alleles.

Equation 5.

$$P(Asc_i \mid X_i = x_i) = 1 - \frac{\binom{x_i}{d} + \binom{n - x_i}{d}}{\binom{n}{d}}$$

Given this equation for the probability of ascertainment, we can find the maximum likelihood estimate of each p_i using the formula in Equation 6., where the denominator is the sum over all classes in the site frequency spectrum.

Equation 6.

$$\hat{p}_i = \frac{n_i}{P(Asc \mid X = i)} \left[\sum_{j=1}^{n-1} \frac{n_j}{P(Asc \mid X = j)} \right]^{-1}$$

From each p_i, we calculate the expected entries in the reconstituted site frequency spectrum given no ascertainment bias. For additional details, see Nielsen et al. (2004).

The total number of individuals in the discovery panel (*d*) for this data is 97, the individuals from the five breeds where the SNPs were discovered. This ascertainment

correction is applied to the genotype data and shown with the uncorrected data in Figure 9. The difference between the corrected and uncorrected site frequency spectra is minimal, although we do see a slight increase in lower frequency derived alleles in the ascertainment-corrected SFS. This minimal difference is likely due to the relatively large discovery panel of individuals. Also, since we ignore that SNPs are ascertained in a substructured population, this may not be an entirely appropriate correction to use.

Genotype Data Sampling

As for the sequence data, the site frequency spectrum pooled for all dog breeds (Figure 9) is not appropriate for use in inference of dog domestication. Compared to the expectation under neutrality, the observed SFS has fewer low frequency derived alleles and an excess of intermediate and high frequency derived alleles. Again, these deviations are due to a number of factors, most especially population subdivision and recent breed bottlenecks.

In a manner similar to that performed for the sequence data, we sample from each breed to reduce these effects. Using all 82 informative SNPs (not excluding those in Appendix Table 4), we sample one chromosome from one individual from each of the 22 breeds. For each SNP, we keep track of the number of derived alleles and the sample size. The sample sizes and derived counts are averaged over 2000 iterations and rounded to the nearest integer.

Because each SNP has a different sample size, we use the hypergeometric distribution as described previously to project to a given sample size. Out of a possible sample size of 22 (as one chromosome is sampled from each of 22 breeds), the SNP with the lowest sample size is position 46808493 on chromosome 1, with an average sample

size of 4. There is a tradeoff between having more entries in the site frequency spectrum and excluding more SNPs with a low sample size. Because of this, we first project to a sample size of 14, removing all six SNPs in Table 5 ($n < 14$) to bring the total number of SNPs to 76. This generates a site frequency spectrum as if there were only 14 individuals, one individual sampled from each of a hypothetical 14 breeds. The resulting SFS is plotted in Figure 10. Note that if there were no missing data for any of the SNPs, the site frequency spectrum would have a sample size of 22.

In addition to projecting to 14, we project to a sample size of 11 (half the entire sample size of 22) by excluding only one SNP. We see if adding information from additional SNPs, while having fewer entries in the SFS, has any effect (Figure 11).

Table 5. Sites with low average sample size ($n < 14$) after sampling genotype data, as described in text. Amplicons, the positions within the amplicons, and the chromosome position according to CanFam1, are given.

Chromosome	**Amplicon/ Position within Amplicon**	**Position on Chromosome**	**Average Sample Size**
1	BLA11_742	46808493	4
	BLA51_378	51983668	12
2	None	---	---
3	BLD12_662	17105503	11
34	BLE41_434	26011335	13
37	BLB44_394	7496699	12
	BLB15_243	4045764	13

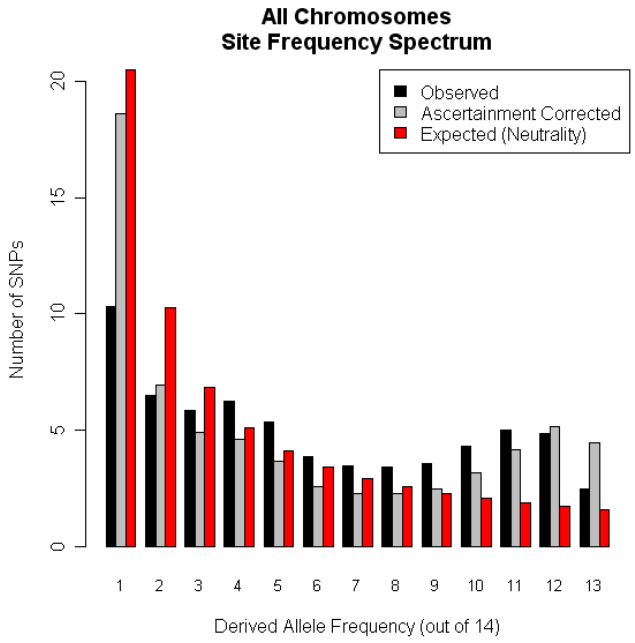

Figure 10. Site frequency spectrum for genotype data sampled by chromosome using a hypergeometric projection to n = 14 as described in text. Black bars indicate observed data, gray bars indicate data corrected for SNP ascertainment as described in text, and red bars are the expectation under neutrality. x-axis is the derived allele frequency out of 14, and the y-axis is the number of SNPs with derived allele at that frequency.

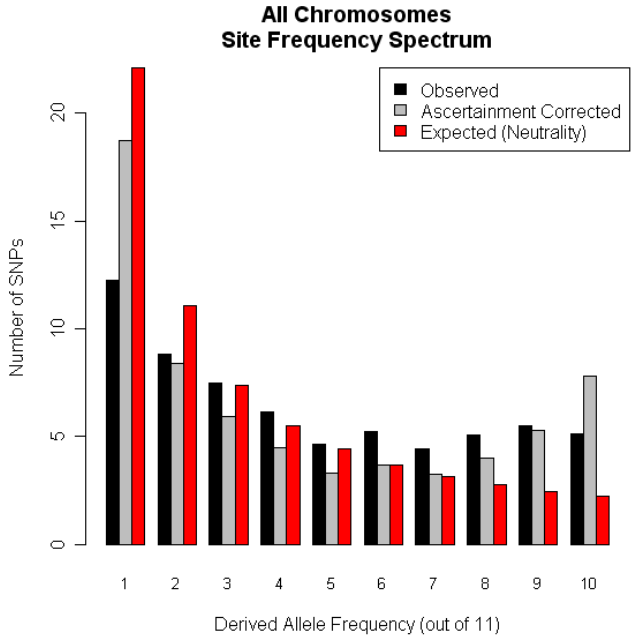

Figure 11. Site frequency spectrum for genotype data sampled by chromosome using a hypergeometric projection to n = 11 as described in text. Black bars indicate observed data, gray bars indicate data corrected for SNP ascertainment as described in text, and red bars are the expectation under neutrality. x-axis is the derived allele frequency out of 11, and the y-axis is the number of SNPs with derived allele at that frequency.

There again exists the issue of ascertainment bias for the sampled site frequency spectra. Since we sample only one chromosome from each breed, we assume that we have discovered the SNPs in a discovery panel of five dogs, one each from the Akita, Bernese Mountain Dog, Golden Retriever, Labrador Retriever, and Pekingese, the breeds in the discovery panel. Under this ascertainment scheme, we expect to observe an excess of high frequency derived alleles and fewer low frequency derived alleles compared to that actually present in the population. We apply the ascertainment bias correction outlined in Equation 5 and Equation 6 using a discovery panel depth (d) of 5 for both the sample size of n = 14 (Figure 10) and sample size of n = 11 (Figure 11).

In comparing the reconstituted SFS corrected for ascertainment bias with the observed SFS, it is clear that there is a correction for having observed fewer singletons. The corrected SFS also increases the number of SNPs at high frequency. Comparing the SFS of the n = 11 and n = 14 projected data, we see only minor differences. We observe a slight decrease in the number of SNPs in the highest frequency class for the larger sample size, whereas we do not see such a decrease for the lower sample size. There also appears to be a more pronounced hump of middle-to-high frequency derived alleles in the SFS of the larger sample size. The additional three entries of the SFS projected to a sample size of 14 may increase our power to infer demographic parameters of domestication.

Therefore, to infer the domestication event, we use several independent and different site frequency spectra. For the sequence data, we use two site frequency spectra, each sampled in a different manner. For the genotype data, we have site

frequency spectra projected to two different sample sizes, both correcting for and ignoring ascertainment bias.

Results

Using the site frequency spectra described above, the program PRFREQ was used to infer the composite maximum likelihood estimates for parameters of the models in Table 3 pictured in Figure 2. We use both the multinomial and Poisson calculations of PRFREQ, described in [IV. Methods, Analysis program PRFREQ]. Scaling is done in terms of the constant wolf population – all size change parameters are scaled by the wolf effective population size (i.e., $\omega = N_{eDOG}/N_{eWOLF}$).

Sequence Data

First, we analyzed the site frequency spectra from the sequence data of the original five breeds, both sampling by SNP (Figure 7) and by chromosome (Figure 8). The value of the ancestral θ used in the Poisson likelihood calculations is 44.925, the per-bp wolf θ (0.00086) multiplied by the total number of base pairs sequenced (52018 bp). While the choice of sampling method does not greatly change the conclusions of the analysis, the Poisson likelihood calculations do yield different results than the multinomial.

Results from sampling by SNP using both calculations are shown in Table 6. While the A1 (contraction 15,000 years ago) model is not significantly different than the A0 (neutral) model for the Poisson calculation, the multinomial calculation is significant. This is evidence for a significant contraction at domestication where the newly formed dog population was 0.21 (ω) times the size of the ancestral wolf population. Although allowing the time of the contraction to vary in the A2 model detects a more recent

contraction for both calculations, the improvement in likelihood is not significant. As a result, there is no evidence for a contraction at a time other than 15,000 years ago. Similarly, the other models (A3, A4) with two size changes are not significant. Interestingly, in the A4 model, parameter estimates indicate a prolonged expansion at the time of domestication followed by a severe contraction. However, from these methods, we do not have power to pick up signatures of anything other than an approximate four-fold contraction at domestication.

The results obtained after sampling by chromosome (Table 7) do not result in largely different conclusions. Again, only the multinomial calculation detects a significant difference between the neutral model and the contraction model, estimating a contraction of size 0.235 relative to the ancestral wolf effective population. This is of similar intensity to the estimate obtained when sampling by SNP, where ω was equal to 0.21. Again, higher models were not significant for either calculation.

Although the results of the two sampling methods are relatively comparable, we do observe differences when comparing the site frequency spectra of the two sampling methods (Figure 12, Figure 13). While the number of SNPs decreases as the derived frequency increases in the data sampled by chromosome, the data sampled by SNP has a more jagged appearance with an increase of derived alleles at frequency ¾. That we obtain similar results from the demographic modeling although the observed site frequency spectra are rather different may indicate that we may only have limited power when performing inference on a site frequency spectrum with only four entries. As can be seen in Figure 12 for the data sampled by SNP, the fit between the observed data and the contraction model is only slightly improved in comparison to the neutral model.

Similarly, only a slight improvement is seen in the data sampled by chromosome for the contraction model (Figure 13). Again, this may indicate our limited power to infer demography from a sample size of only five. Thus, examining the genotype data with a larger sample size may provide more information about dog domestication.

Table 6. Results of PRFREQ analysis for sequence data sampled by SNP as described in text, for both Poisson and multinomial calculations. All τ are given in number of generations from the present, and values of ω are given in terms of the ancestral population size (i.e., $\omega = N_{eDOG}/N_{eWOLF}$). p-values are given for the comparisons in parentheses using the χ^2 distribution.

Sequence data sampled by SNP					
Poisson					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	263.053	--	Constant size
A1	1	$\tau = 5000$ $\omega = 0.78$	263.195	0.595 (A1 vs. A0)	Contraction
A2	2	$\tau = 431.822$ $\omega = 0.228$	263.437	0.486 (A2 vs. A1)	Contraction
A3	2	$\tau = 5000$ $\tau_B = 218.07$ $\omega_B = 0.1$ $\omega = 100$	263.484	0.447 (A3 vs. A1)	Contraction, then expansion
A4	3	$\tau = 5000$ $\tau_B = 4331.84$ $\omega_B = 100$ $\omega = 0.109$	263.921	0.484 (A4 vs. A1)	Expansion, then contraction
Multinomial					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	377.939	--	Constant size
A1	1	$\tau = 5000$ $\omega = 0.21$	381.911	0.005 (A1 vs. A0)	Contraction
A2	2	$\tau = 431.822$ $\omega = 0.03$	382.261	0.527 (A2 vs. A1)	Contraction
A3	2	$\tau = 5000$ $\tau_B = 1468.19$ $\omega_B = 0.1$ $\omega = 100$	382.149	0.491 (A3 vs. A1)	Contraction, then expansion
A4	3	$\tau = 5000$ $\tau_B = 4318.22$ $\omega_B = 100$ $\omega = 0.001$	382.481	0.564 (A4 vs. A1)	Expansion, then contraction

Table 7. Results of PRFREQ analysis for sequence data sampled by chromosome as described in text, for both Poisson and multinomial calculations. All τ are given in number of generations from the present, and values of ω are given in terms of the ancestral population size (i.e., $\omega = N_{eDOG}/N_{eWOLF}$). p-values are given for the comparisons in parentheses using the χ^2 distribution.

Sequence data sampled by chromosome **Poisson**					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	249.094	--	Constant Size
A1	1	$\tau = 5000$ $\omega = 0.882$	249.121	0.815 (A1 vs. A0)	Contraction
A2	2	$\tau = 431.822$ $\omega = 0.337$	249.156	0.792 (A2 vs. A1)	Contraction
A3	2	$\tau = 5000$ $\tau_B = 43.1822$ $\omega_B = 0.1$ $\omega = 100$	249.136	0.865 (A3 vs. A1)	Contraction, then expansion
A4	3	$\tau = 5000$ $\tau_B = 3886.40$ $\omega_B = 100$ $\omega = 0.208$	249.296	0.840 (A4 vs. A1)	Expansion, then contraction
Multinomial					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	357.667	--	Constant size
A1	1	$\tau = 5000$ $\omega = 0.235$	359.887	0.035 (A1 vs. A0)	Contraction
A2	2	$\tau = 2159.11$ $\omega = 0.01$	360.009	0.622 (A2 vs. A1)	Contraction
A3	2	$\tau = 5000$ $\tau_B = 2893.207$ $\omega_B = 0.1$ $\omega = 3.007$	360.059	0.558 (A3 vs. A1)	Contraction, then expansion
A4	3	$\tau = 5000$ $\tau_B = 1295.466$ $\omega_B = 200$ $\omega = 0.0109$	360.043	0.856 (A4 vs. A1)	Expansion, then contraction

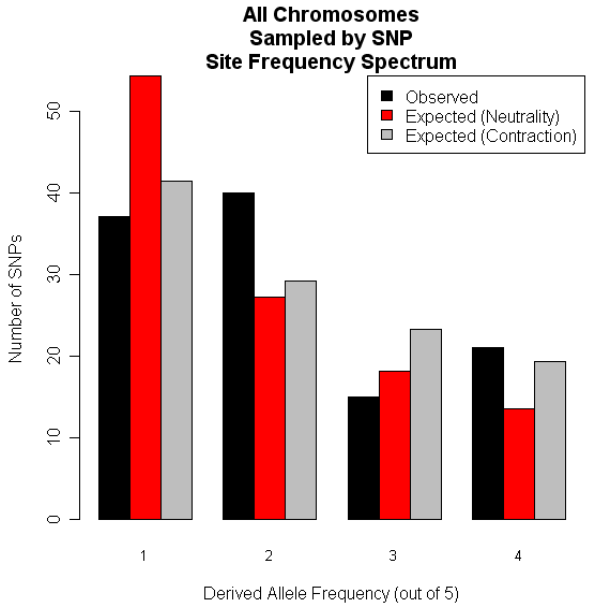

Figure 12. Site frequency spectra of data sampled by each SNP as described in text. Black bars are observed data, red bars are the expectation under neutrality, and gray bars are the expectation under the contraction (A1) model obtained from the multinomial calculation ($\omega = 0.21$). x-axis is the derived allele frequency out of 5, and the y-axis is the number of SNPs with derived allele at that frequency.

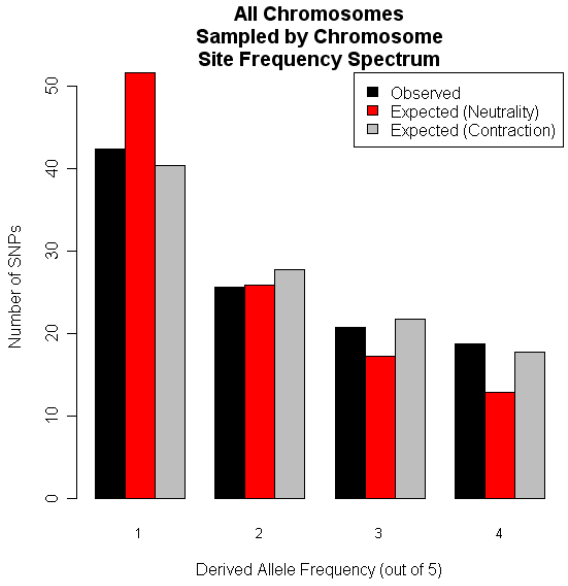

Figure 13. Site frequency spectra of data sampled by chromosome as described in text. Black bars are observed data, red bars are the expectation under neutrality, and gray bars are the expectation under the contraction (A1) model obtained from the multinomial calculation ($\omega = 0.235$). x-axis is the derived allele frequency out of 5, and the y-axis is the number of SNPs with derived allele at that frequency.

Genotype Data

While the analysis of the sequence data for the five breeds had a SFS with only four entries, we suspect that having more entries in the SFS will give us more power to detect demographic history. As a result, we analyze the genotype data collected from an additional 17 dog breeds using the four SFS pictured in Figure 10 and Figure 11, for sample sizes of 14 and 11 respectively, both with and without the ascertainment bias correction.

As for the sequence data, we perform both the multinomial and Poisson calculations. For the Poisson likelihood calculation, we use the same per-bp $\theta = 0.00086$ used for the sequence data. However, as only 105 out of the 200 original SNPs were genotyped in the additional dog breeds (see III. Materials), we sum only the lengths of those amplicons including the informative genotyped SNPs to obtain the per-region θ. From a total region length of 37,057 base pairs, a θ of 32.004 is used in the Poisson likelihood. Parameter estimates obtained from the Poisson and multinomial models are slightly different, as was seen for the sequence data analysis, while correcting for ascertainment or slightly changing the sample size does not appear to have a large effect.

We examine the data set projected to $n = 14$ for the data uncorrected for ascertainment bias (Table 8). For the Poisson inference of the A1 model (with τ constant at 15,000 years ago), the composite maximum likelihood estimate of ω is 0.225, indicating a dog ancestral population size 0.225 times the size of the wolf ancestral population size. Allowing the time of contraction, τ, to differ in the A2 model is not significantly different than the A1 model. In contrast, results for the multinomial calculation predict a more severe contraction. Under the A1 model, the maximum

likelihood estimate of $\omega = 0.064$, roughly four times the severity estimated from the Poisson. Allowing τ to vary does not significantly improve the fit of the multinomial model, although a more severe population decline is predicted than for the Poisson. Results of expected and predicted models are shown in the upper panel of Figure 14, where we can see that the predicted contraction models are not perfect at capturing the entire shape of the observed SFS.

Correcting for ascertainment bias does not appear to have a large affect on the demographic inference (Table 9). For the corrected data set, the maximum likelihood estimate of ω is 0.25 for the A1 model under the Poisson calculation and 0.031 under the multinomial calculation, only slightly different than the uncorrected SFS estimates. Examining the lower panel of Figure 14 shows that the predicted contraction models do not entirely match the observed SFS.

We also examine the significance of models beyond the A1 and A2 contraction models. For the uncorrected Poisson and multinomial inferences, no models beyond the A1 contraction model (with constant τ) were significant according to the χ^2 p-value approximation (Table 8). In the corrected data set, however, the A3 model is significant under the Poisson calculation (p-value = 0.033), but not the multinomial calculation (Table 9). This significant model detects a 10-fold contraction 15,000 years ago (at domestication) for a bottleneck lasting approximately 2600 generations. This is followed by an expansion to 2.5 times the size of the wolf effective population size. This significant p-value is suspicious, given that we see no other evidence of the significance of a model beyond the A1 model. Because PRFREQ assumes that sites are unlinked,

whereas the sites observed are in fact linked, this p-value obtained using the χ^2 approximation may not be appropriate.

Table 8. Results of PRFREQ analysis for genotype data with the hypergeometric projection to 14, uncorrected for ascertainment bias, for both Poisson and multinomial calculations. All τ are given in number of generations from the present, and values of ω are given in terms of the ancestral population size (i.e., $\omega = N_{eDOG}/N_{eWOLF}$). p-values are given for the comparisons in parentheses using the χ^2 distribution.

Genotype data (n=14) uncorrected for ascertainment					
Poisson					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	27.135	--	Constant Size
A1	1	$\tau = 5000$ $\omega = 0.225$	42.206	4.01×10^{-8} (A1 vs. A0)	Contraction
A2	2	$\tau = 2245.474$ $\omega = 0.12$	42.427	0.506 (A2 vs. A1)	Contraction
A3	2	$\tau = 5000$ $\tau_B = 1835.244$ $\omega_B = 0.01$ $\omega = 64.5$	42.425	0.508 (A3 vs. A1)	Contraction, then expansion
A4	3	$\tau = 5000$ $\tau_B = 3022.754$ $\omega_B = 0.9$ $\omega = 0.11$	42.426	0.803 (A4 vs. A1)	Expansion, then contraction
Multinomial					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	99.874	--	Constant size
A1	1	$\tau = 5000$ $\omega = 0.064$	108.532	3.17×10^{-5} (A1 vs. A0)	Contraction
A2	2	$\tau = 5397.775$ $\omega = 0.084$	108.533	0.957 (A2 vs. A1)	Contraction
A3	2	$\tau = 5000$ $\tau_B = 4.3182$ $\omega_B = 0.1$ $\omega = 0.0001$	108.532	1 (A3 vs. A1)	Contraction, then expansion
A4	3	$\tau = 5000$ $\tau_B = 5000$ $\omega_B = 0.064$ $\omega = 0.01$	108.532	1 (A4 vs. A1)	Expansion, then contraction

Table 9. Results of PRFREQ analysis for genotype data with the hypergeometric projection to 14, corrected for ascertainment bias as described in text, for both Poisson and multinomial calculations. All τ are given in number of generations from the present, and values of ω are given in terms of the ancestral population size (i.e., ω = NeDOG/NeWOLF). p-values are given for the comparisons in parentheses using the $\chi 2$ distribution.

Genotype data (n=14) corrected for ascertainment **Poisson**					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	41.089	--	Constant size
A1	1	τ = 5000 ω = 0.25	50.066	2.27×10^{-5} (A1 vs. A0)	Contraction
A2	2	τ = 16193.33 ω = 0.375	51.362	0.107 (A2 vs. A1)	Contraction
A3	2	τ = 5000 τ_B = 2590.932 ω_B = 0.1 ω = 2.5	52.328	0.033 (A3 vs. A1)	Contraction, then expansion
A4	3	τ = 5000 τ_B = 2159.11 ω_B = 0.0825 ω = 1.8	52.369	0.099 (A4 vs. A1)	Contraction, then expansion
Multinomial					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	113.829	--	Constant Size
A1	1	τ = 5000 ω = 0.031	117.190	0.009 (A1 vs. A0)	Contraction
A2	2	τ = 5937.553 ω = 0.03875	117.353	0.567 (A2 vs. A1)	Contraction
A3	2	τ = 5000 τ_B = 3778.443 ω_B = 0.1 ω = 22	118.001	0.444 (A3 vs. A1)	Contraction, then expansion
A4	3	τ = 5000 τ_B = 1295.466 ω_B = 100 ω = 0.02	117.005	1 (A4 vs. A1)	Expansion, then contraction

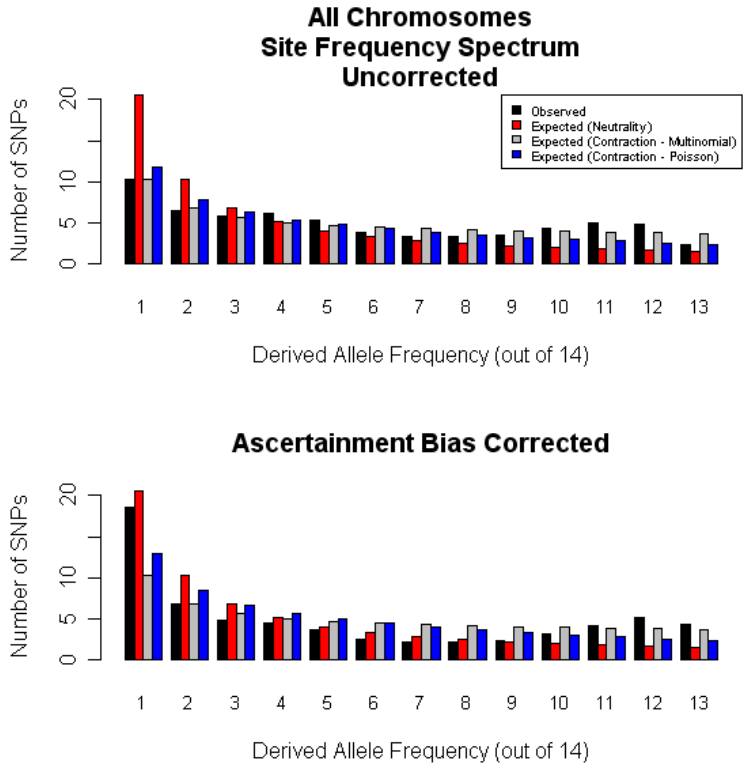

Figure 14. Site frequency spectrum for genotype data both uncorrected and corrected for SNP ascertainment as described in text, with a hypergeometric projection to 14. Black bars are the observed data, red bars are the expectation under neutrality, and gray and blue bars are the expectations under the contraction models from the PRFREQ multinomial and Poisson calculations, respectively, as indicated in Table 8 and Table 9. x-axis is the derived allele frequency out of 14, and the y-axis is the number of SNPs with derived allele at that frequency.

When using PRFREQ to analyze the SFS projected to a sample size of 11 rather than 14, we obtain similar results. For the uncorrected SFS under the Poisson calculation, an approximate four-fold contraction at the time of domestication is significantly different than a model of constant population size (Table 10). The multinomial calculation yields a more severe maximum likelihood estimate of $\omega = 0.0487$ for the A1 model. No models beyond the contraction (A1) model are significant for the uncorrected data set (Table 10). More complicated models with multiple size changes (A3, A4) are not significant and are not shown.

As with the data projected to n = 14, correcting for ascertainment bias does not drastically change the demographic modeling results (Table 11). The maximum

likelihood estimate of ω under the A1 model with the Poisson calculation is 0.2625, with a significant composite likelihood. Again, the multinomial calculation detects a more severe population contraction to 0.0325 times the ancestral wolf effective population size. In general, maximum likelihood estimates appear to be rather comparable between the uncorrected and corrected data sets between the n = 11 and n = 14 analyses. The observed and expected site frequency spectra both under neutrality and the contraction model for the data projected to a sample size of 11 are shown in Figure 15.

Ideally, further modifications on the parameters estimated using the ascertainment bias-corrected SFS should be made. The composite maximum likelihood estimates of the demographic parameters should theoretically be adjusted for the uncertainty created in using a SFS that was not actually observed (Nielsen et al. 2004). This has not been completed, but is an important area for further exploration.

Table 10. Results of PRFREQ analysis for genotype data with the hypergeometric projection to 11, uncorrected for ascertainment bias, for both Poisson and multinomial calculations. All τ are given in number of generations from the present, and values of ω are given in terms of the ancestral population size (i.e., $\omega = N_{eDOG}/N_{eWOLF}$). p-values are given for the comparisons in parentheses using the χ^2 distribution.

Genotype data (n=11) uncorrected for ascertainment **Poisson**					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	47.713	--	Constant size
A1	1	$\tau = 5000$ $\omega = 0.23$	58.401	3.77×10^{-6} (A1 vs. A0)	Contraction
A2	2	$\tau = 2418.203$ $\omega = 0.14$	58.503	0.651 (A2 vs. A1)	Contraction
Multinomial					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	117.453	--	Constant Size
A1	1	$\tau = 5000$ $\omega = 0.0486667$	124.169	0.0002 (A1 vs. A0)	Contraction
A2	2	$\tau = 4534.131$ $\omega = 0.0393333$	124.175	0.906 (A1 vs. A0)	Contraction

Table 11. Results of PRFREQ analysis for genotype data with the hypergeometric projection to 11, corrected for ascertainment bias as described in text, for both Poisson and multinomial calculations. All τ are given in number of generations from the present, and values of ω are given in terms of the ancestral population size (i.e., $\omega = N_{eDOG}/N_{eWOLF}$). p-values are given for the comparisons in parentheses using the χ^2 distribution.

Genotype data (n=11) corrected for ascertainment **Poisson**					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	56.023	--	Constant Size
A1	1	$\tau = 5000$ $\omega = 0.2625$	63.017	0.0002 (A1 vs. A0)	Contraction
A2	2	$\tau = 15113.77$ $\omega = 0.44$	63.766	0.221 (A2 vs. A1)	Contraction
Multinomial					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
A0	0	None	251.528	--	Constant size
A1	1	$\tau = 5000$ $\omega = 0.0325$	258.850	0.007 (A1 vs. A0)	Contraction
A2	2	$\tau = 6909.152$ $\omega = 0.05$	259.003	0.696 (A2 vs. A1)	Contraction

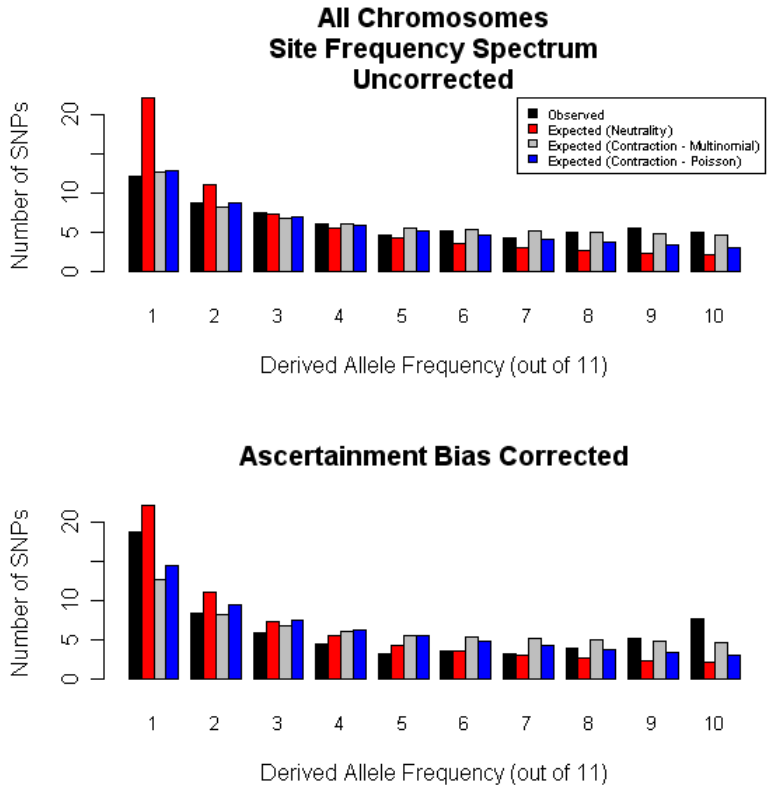

Figure 15. Site frequency spectrum for genotype data both uncorrected and corrected for SNP ascertainment as described in text, with a hypergeometric projection to 11. Black bars are the observed data, red bars are the expectation under neutrality, and gray and blue bars are the expectations under the contraction models from the PRFREQ multinomial and Poisson calculations, respectively, as indicated in Table 10 and Table 11. x-axis is the derived allele frequency out of 11, and the y-axis is the number of SNPs with derived allele at that frequency.

Finally, we compare the composite likelihoods of the multinomial and Poisson calculations as described in [IV. Methods, Analysis program PRFREQ]. Using Equation 3 we calculate the values of θ in the multinomial calculation using the optimized multinomial demographic parameters shown in Table 8 and Table 9. With this estimated value of θ and the optimized parameters, we perform the Poisson calculation of the composite likelihood as a "rescaling" of the original multinomial likelihood (Table 12). Note that the value of the ancestral θ used in the Poisson calculations is 32.004.

The Poisson likelihood with multinomial-estimated parameters has a significantly higher likelihood than the original Poisson likelihood under all neutral (A0) models,

where the values of θ estimated from the multinomial are approximately 20 (Table 12). The wolf θ used in the Poisson calculations (32.004) is in fact significantly different from these values. For the non-neutral models, the multinomial likelihood is significantly greater than the Poisson likelihood for only the A1 model of the n = 14 corrected data. Though maximizing θ does in fact increase the fit of the model, θ is estimated at approximately 414, more than 10 times the original θ used for the Poisson.

This rather unrealistic value of θ shows that the multinomial calculation has incorrectly estimated the mutation rate by an order of magnitude in relying only on the shape of the SFS. This explains why the multinomial calculation perceives a much stronger contraction than the Poisson; a very strong contraction given the θ we propose would decrease diversity too severely. Settling on unrealistic values of θ in the multinomial calculations may indicate overfitting of the data, possibly due to limited power in our dataset. The severe contraction results we obtain from the multinomial are likely a result of overfitting the data rather than a severe domestication contraction.

Table 12. Results of rescaling multinomial likelihoods for comparison between multinomial and Poisson calculations for the given models. p-value is obtained by taking twice the difference in log likelihood and using the χ^2 distribution with 1 df. A p-value < 0.05 indicates that the likelihood under the multinomial calculation is significantly greater than the likelihood under the Poisson calculation (indicated by asterisks).

Data Set	Model	θ (Estimated from Multinomial)	Poisson LL (Multinomial Parameters)	Poisson LL (Poisson Parameters)	p-value
n = 14 Uncorrected	A0	20.5	34.6800	27.1347	0.0001 *
	A1	108.265	43.3375	42.2064	0.1326
n = 14 Corrected	A0	20.499	48.6348	41.0894	0.0001 *
	A1	414.324	51.9956	50.0656	0.0495 *
n = 11 Uncorrected	A0	22.083	52.7715	47.7130	0.0015*
	A1	178.253	59.4863	58.4012	0.1407
n = 11 Corrected	A0	22.083	61.0818	56.0233	5.44×10^{-15} *
	A1	397.224	64.7431	63.0174	0.0632

Assessment of Model Significance with Coalescent Simulations

We simulate 2000 neutral coalescent samples using msHOT (Hellenthal and Stevens 2007) in order to account for the recombination between amplicons as well as between SNPs within amplicons. We sample 14 chromosomes in each sample, mirroring the estimation of the projected n = 14 genotype dataset. We input the SFS from each of these simulations into PRFREQ to obtain the multinomial likelihood of the neutral (A0) model. We also use the multinomial calculation of PRFREQ to optimize the contraction parameter, ω, while leaving τ fixed at 5000 generations for all simulations. We examine the distribution of the likelihood ratio test statistic between the neutral A0 and contraction A1 models (Figure 16), allowing us to obtain a verification of the p-value initially obtained using the χ^2 distribution from the likelihood ratio test statistic (Table 8, Table 9). Using the upper and lower 2.5% of values, we obtain a confidence interval of (-0.000068, 5.958706) for the LRT statistic. The fact that we observe negative values of the LRT statistic is due to the fact that we converge upon some of the boundary values of ω when performing the maximization.

For the uncorrected data, the value of the statistic is 17.315, and for the corrected data, the value of the statistic is 6.722, seen in Table 8 and Table 9. Each of these values lies outside of the 95% confidence interval of the simulated neutral data. Since the difference in likelihoods we observe is significantly greater than the difference observed for neutral data, the multinomial contractions we estimate are in fact significant even when accounting for linkage between sites.

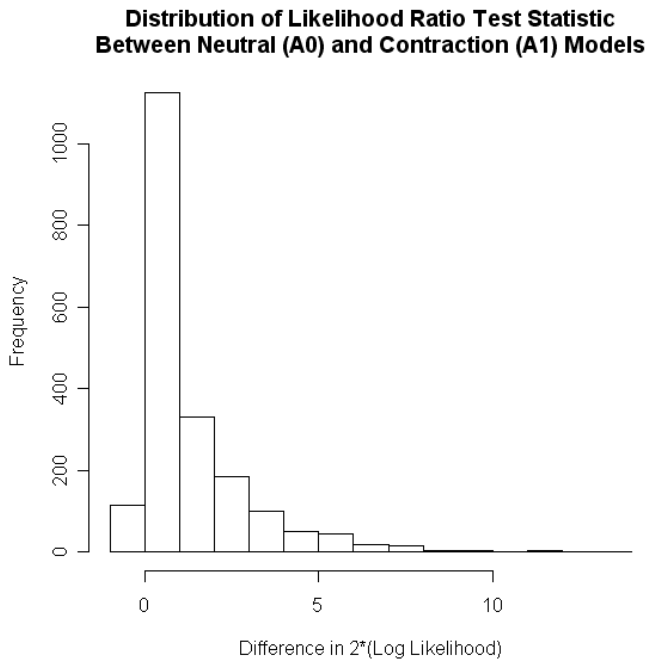

Figure 16. Distribution of the likelihood ratio test statistic between the optimized A1 (contraction) model and neutral (A0) model for 2000 neutral coalescent simulations of genotype data with a hypergeometric projection to 14. Multinomial calculations are used.

Interpretation and Domestication Conclusions

We have provided an analysis of the domestication event when dogs diverged from the gray wolf, assuming a simplistic demographic model of a one-time population size change without introgression between the dog and wolf. Using the site frequency spectra of one chromosome per dog breed to represent the dog population after dog domestication, we find evidence for a contraction at the time of domestication. We obtain this result for both the sequence and genotype data, although for the sequence data, the contraction model is only significant for the multinomial calculation. Here, we discuss the results of different methods used, as well as the implications of our results on dog demographic history.

First, we discuss the differences between the multinomial and Poisson calculations of PRFREQ. The Poisson calculation may not be entirely appropriate, as we

obtain an estimate of the ancestral wolf θ based on the current wolf population. This assumes that wolves are not subdivided and have maintained a constant population size throughout time, although these may not be entirely appropriate assumptions (for additional details see VII. Demographic Analysis of Wild Canids). There is also the possibility that wolf and dog have different mutation rates, skewing the results of the Poisson analysis. The Poisson likelihood calculations for both the sequence and genotype data should be interpreted in light of these caveats.

The multinomial inference takes into account only the shape of the observed site frequency spectra, not the observed number of segregating sites. For the sequence data, only the multinomial calculation provides evidence for a significant contraction at dog domestication. For the genotype data, the multinomial calculations estimate a much more severe contraction due to the fact that the calculation is largely affected by an excess of high frequency derived alleles in the SFS (as in Figure 14). This also appears to explain why the multinomial picks up a slightly stronger contraction for the data corrected for ascertainment bias in comparison to the uncorrected data (Table 8 and Table 9), as the corrected data has an increase of higher frequency derived alleles. As seen through the unrealistic values of θ estimated by the multinomial (Table 12), the multinomial calculation is likely estimating a more severe contraction by overfitting to the data. In contrast, the number of singletons largely affects the Poisson calculation. As a result, the Poisson calculation detects a more severe contraction for the uncorrected rather than the ascertainment bias-corrected SFS (Table 8, Table 9), as the ascertainment-corrected data has a greater number of singletons more similar to the value expected under neutrality.

We also discuss implications of the different results seen for the genotype and sequence data. We detect a more severe contraction in the genotype data than in the sequence data for both calculations. This is likely a result of the fact that we have more information about the true SFS of the ancestral "pre-breed" populations when we have a sample size of 11 or 14 as opposed to five. There is little difference between the results of the genotype data SFS projected to 11 as opposed to 14, indicating that these additional three entries and SNPs removed likely have little effect. Including more entries in the SFS certainly provides more power for demographic inference.

Overall, however, we have limited power to detect more complicated demographic models given that we are examining only 82 SNPs. In order to detect an expansion following a bottleneck at domestication, we would need to detect rare mutations that have arisen since the bottleneck event. Given very few SNPs and limited data, it is unlikely that we would be able to detect such mutations. In general, we do not detect any significant models beyond the A1 contraction models. We only detect slight evidence for a bottleneck with the Poisson calculation in the data corrected for ascertainment and n = 14, which is likely a result of the χ^2 approximation used (Table 9). We also address the issue of ascertainment bias for the genotype data and find that correcting for ascertainment for this particular data set does not appear to drastically affect our estimates of the demographic parameters. However, since methods of SNP ascertainment may have a large affect in other scenarios, accounting for ascertainment is essential in obtaining an accurate estimate of demography.

Another important issue regards the assumptions of PRFREQ, namely that all sites are unlinked. While PRFREQ allows us to find the composite likelihood of our data as an

approximation to the true likelihood, it may behave poorly when SNPs are tightly linked. We address this issue through coalescent simulations, which seem to suggest that the conclusions we draw from PRFREQ are not necessarily violated by SNP linkage.

In examination of the genotype data due to its increased power over the sequence data, it is thus possible that there was indeed a contraction at domestication approximately 15,000 years ago. The multinomial A1 estimates predict ω near 0.04, indicating a 25-fold contraction at domestication. Such a strong contraction would likely eliminate a much of the diversity we see in dog breeds; as mentioned, this estimate is likely due to overfitting to high frequency derived alleles in the SFS. The Poisson A1 calculation calculates perhaps a more realistic estimate of a four-fold contraction at domestication.

Though we detect evidence for a significant contraction at the time of domestication, it is possible that this is not actually due to an actual contraction at dog domestication. Specifically, deviations from neutrality in the SFS we observe are likely artifacts of using breed dogs, which show evidence of substructure as well as severe bottlenecks. It is likely that sampling one individual from each breed only slightly minimizes the effect of using breed dogs. This emphasizes the importance of using data from feral non-breed dogs to obtain a more accurate picture of dog domestication. Creation of an ancestral "pre-breed" dog SFS may be more plausible with such data, as there will be fewer spurious effects of recent breed bottlenecks and subdivision. Using breed dogs certainly imposes a limitation on the amount of information we can infer about dog domestication prior to breed formation.

In conclusion, we have evidence for a slight domestication bottleneck approximately 5,000 generations ago, or 15,000 years ago, when dogs diverged from the gray wolf. Given demographic signatures caused by strong recent breed bottlenecks in our limited data, however, it seems unlikely that there has been a severe contraction at dog domestication. If there were a very severe contraction during dog formation, we would expect diversity to be much lower in dogs than in wolves. However, the per-bp θ of the original five breeds on the sequence data of chromosome 1 is 0.00064, similar to the wolf θ of 0.00086. The minimal intensity of the domestication contraction can be demonstrated through these comparable levels of diversity.

It is likely that high levels of diversity in the dog were maintained through continued interbreeding between dogs and wolves or through multiple domestication events (Vilá et. al 1997; Randi and Lucchini 2002; Tsuda et al. 1997). Another possibility, proposed by Björnerfeldt, et. al (2006) is that while the initial dog population was small, relaxation of selective constraint allowed for accumulation of more diversity. Although we do not explore those models here, considering more complicated and realistic scenarios when modeling dog domestication is an important avenue for future research. In conclusion, the high level of diversity seen in today's domestic dog was not lost in one severe contraction at dog domestication, but was rather maintained through the domestic dog's gradual integration into human society.

VI. Demographic Analysis of Breed Formations

Analysis with PRFREQ

Next, we explore the demographic history of the five dog breeds of the Akita, Bernese Mountain Dog, Golden Retriever, Labrador Retriever, and Pekingese using the original sequence data and PRFREQ. Our goal is to compare the severity of the bottlenecks of the breeds and provide insight into the breeds' formations. As for the domestication event, we compare composite likelihoods between nested models, shown in Table 13 and governed by the parameters described in Figure 3. The nested models are similar to those examined for the domestication event.

Table 13. Nested likelihood models used in inference of breed bottleneck events. Parameters of each model, as well as their associated degrees of freedom, are given.

Model	Parameter	df
B0 (stationary demography)	None	0
B1a (1 size change, τ fixed)	τ = 100 generations ω = vary	1
B1b (1 size change)	τ = vary ω = vary	2
B2a (2 size changes (bottleneck) – population decline and expansion, with a fixed, short bottleneck length)	τ = vary τ_B = fixed ω_B = vary ω = vary	3
B2b (2 size changes (bottleneck) – population decline and expansion, with a varying bottleneck length)	τ = vary τ_B = vary ω_B = vary ω = vary	4

The B0 model is the neutral model. A significant comparison between the B0 and B1a model indicates a significant contraction at a fixed time τ, 100 generations from the present. Although this timing may not be historically valid for all breeds, it facilitates a better comparison of results between breeds. In the B1b model, we allow τ to vary; a significant comparison between B1b and B1a is evidence for a population size change at

a time other 100 generations ago. We also allow for the possibility of a classical bottleneck model, where a population contraction is followed by a subsequent expansion. If the B2a model, fixing the length of the bottleneck (τ_B) at a short period of time and optimizing all other parameters, performs significantly better than the B1 model, there is evidence for a bottleneck model. Finally, if the B2b model performs significantly better than the B2a model, we have evidence for a population bottleneck, with a different duration than that fixed in B2a.

As for the domestication event, coalescent simulations are performed to obtain verifications of the p-values obtained by the χ^2 approximation. We perform 2000 coalescent simulations under the neutral model with msHOT to model the Akita, Bernese Mountain Dog, Golden Retriever, Labrador Retriever, and Pekingese, using the same method of obtaining the background recombination rate and ρ as for the domestication. Input for each chromosome is shown in Appendix Table 6.

We obtain the SFS from each simulation and its multinomial likelihood under the neutral (B0) model. The neutral samples are also optimized under the B1a model with the multinomial PRFREQ calculation, keeping τ constant at 100 generations and allowing ω to vary between 0.1 and 3.1. We examine the distribution of the likelihood ratio test statistic between the neutral and contraction models to supplement the p-values obtained from the approximation to the χ^2 distribution.

Data Manipulation

The site frequency spectrum is constructed from the sequence data of the five dog breeds, pooling all chromosomes separately for each breed (Figure 17). As a result of intense selective breeding programs, however, breeds are highly inbred. Demographic

inference of breed formations could potentially be affected by the fact that under inbreeding, an individual's chromosomes are more similar than expected under random mating. We attempted to reduce the effects of inbreeding within breeds by sampling one chromosome per individual for each breed 2000 times. Using the Golden Jackal as the outgroup to root SNPs for each iteration, we average the SFS from each iterations. The results of this sampling are shown in Figure 18, and basic summary statistics are shown in Table 14.

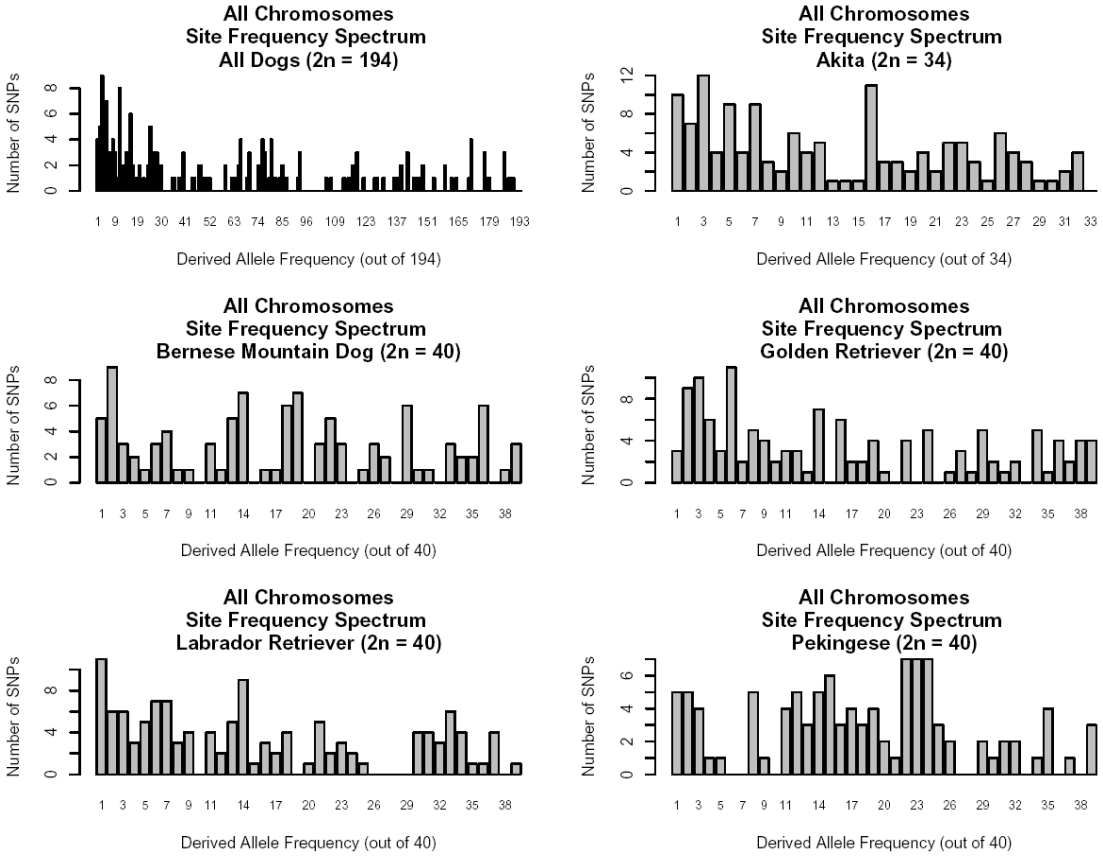

Figure 17. Observed site frequency spectra of sequence data, pooling all chromosomes, for each breed as indicated. Data is also shown pooling all dog breeds together as a comparison. x-axis is the derived allele frequency out of 2n as indicated, and the y-axis is the number of SNPs with derived allele at that frequency.

Table 14. Summary statistics obtained for each breed after sampling one chromosome from each individual as described in text. S is the number of segregating sites, and θ is Watterson's estimate of θ.

Dog Breed	n	S	θ (per bp)	Number of Singletons	π (per bp)	Tajima's D	Average Heterozygosity
Akita	17	128.26	0.00073	16.649	0.00088	0.8822	0.3359
Bernese Mountain Dog	20	94.911	0.00051	9.627	0.00066	1.1822	0.3449
Golden Retriever	20	119.466	0.00065	12.674	0.00075	0.6661	0.3111
Labrador Retriever	20	115.587	0.00063	13.753	0.00075	0.7905	0.3192
Pekingese	20	98.540	0.00053	7.963	0.00076	1.7489	0.3818

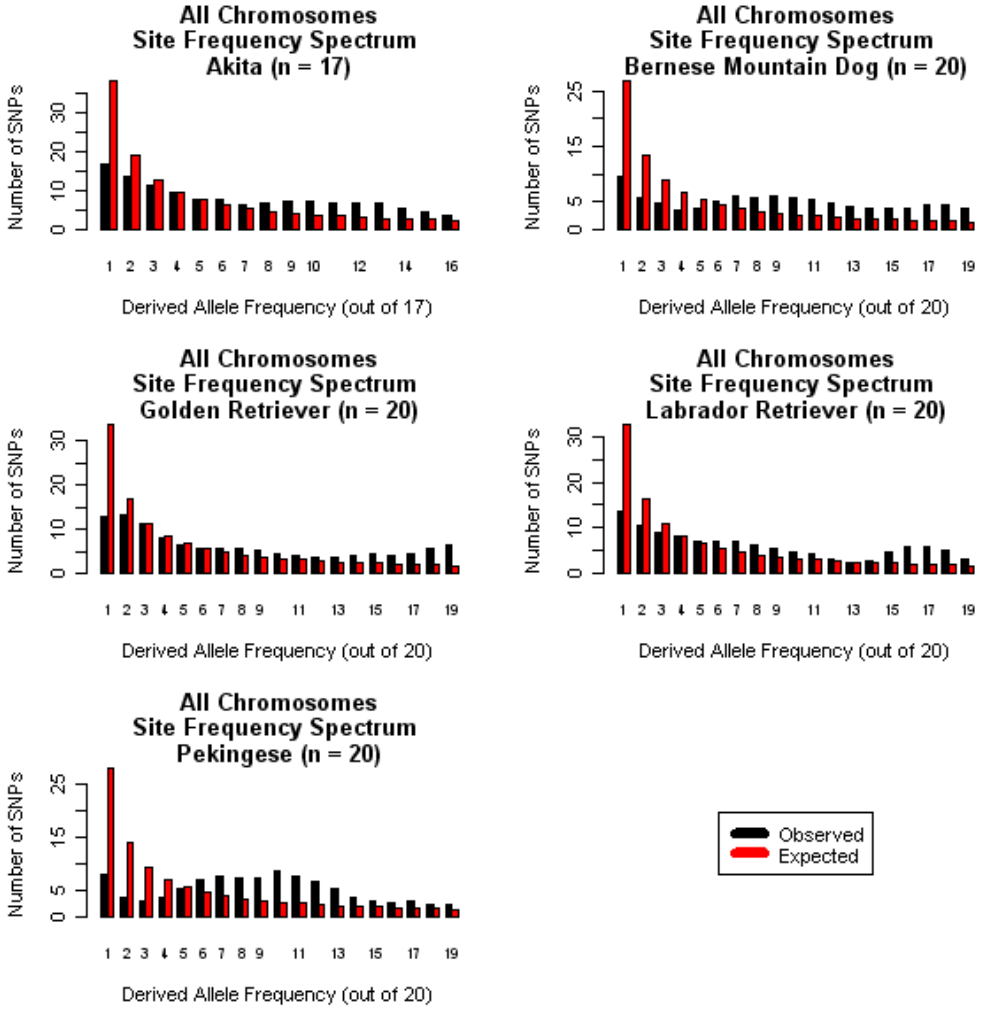

Figure 18. Site frequency spectrum for sequence data sampled one chromosome per individual in each breed as described in text. Black bars are the observed data, and red bars are the expectation under neutrality. x-axis is the derived allele frequency out of n as indicated, and the y-axis is the number of SNPs with derived allele at that frequency.

Comparisons can be made between the sampled (Figure 18) and unsampled (Figure 17) site frequency spectra, keeping in mind that the sampled allele frequencies are out of a total of n, rather than 2n, chromosomes. Even ignoring that the sampled data has a generally smoother site frequency spectrum as a result of averaging over many iterations, the sampled SFS lacks the large spikes seen in the unsampled site frequency spectrum. The sampling method implemented seems to be effective in accounting for at least some of the effects of inbreeding in the breed site frequency spectra.

A brief comparison of the site frequency spectra between breeds seems to be consistent with what is known about each breed (Figure 18). The Pekingese site frequency spectrum has the most notable excess of middle frequency variants, characteristic of a severe population decline. The site frequency spectra of the Bernese Mountain Dog, a rare breed, also has an excess of middle frequency variants, though to a lesser extent than that seen in the Pekingese. The Akita has only a slight excess of middle frequency SNPs. Finally, the Golden Retriever and the Labrador Retriever appear to have site frequency spectra most similar to that under neutrality, though noticeable deviations exist for SNPs of high frequency. Somewhat agreeing with these observations are values of Tajima's D (Table 14), where the Bernese Mountain Dog and Pekingese have greater positive values than seen in other breeds. Although we have not performed a rigorous test of significance of these values, it is possible that these deviations are due to a population decline. We explore these observations in further detail through analysis with PRFREQ.

Results

We obtain estimates of the demographic parameters for each breed using the sampled site frequency spectra (Figure 18) of each breed as the observed SFS in PRFREQ. As in the inference of the domestication event, both the multinomial and Poisson calculations are performed. We presume that a decline, if any, at dog domestication was not very severe and that the ancestral "pre-breed" dog population is the same size as the wolf effective population size. Due to this assumption of $N_{eDOG} = N_{eWOLF}$, the ancestral θ used in the Poisson calculation is 44.925, which is the per-bp wolf θ (0.00086) multiplied by the total number of base pairs sequenced in all five chromosomes (52018 bp). Another possible estimate for the pre-breed dog θ could be obtained from pooling all dog breeds; however, since values of θ among all dogs and all wolves do not vary greatly, this would likely have little effect.

First, we performed inference with the multinomial calculation. When performing our initial calculations, we scaled by the current effective population size because we were still unsure as to how to estimate the ancestral breed effective population size. In this scaling, $\omega = N_{eDOG}/N_{eBreed}$ and τ is in units of $2*N_{eBreed}$ generations (where N_{eBreed} is the current effective population size of the breed and N_{eDOG} is the ancestral population size). For B1b, B2a, and B2b models, we scaled by the current effective population size, where for the B2a model, τ is fixed at an arbitrarily short value of 0.02, in units of $2*N_{eBreed}$ generations.

Because we did not have any estimates available for the effective population sizes of each individual breed, we realized that this made the interpretation of the timing and contraction severity estimates rather difficult. To facilitate interpretation, we scale the

multinomial parameter estimates of the B2b models by the ancestral effective population size. For size parameters, we take the reciprocal of the values estimated, where $1/\omega$ is N_{eBreed}/N_{eDOG}. Assuming that the ancestral dog effective population is the same size as the wolf effective population size (21,591), we use the ω estimates to calculate the current effective breed population sizes. Then, we scale the estimated τ's (originally in units of $2*N_{eBreed}$ generations) to generations. When estimating the B1a multinomial model, however, we perform the inference in PRFREQ scaling by the ancestral population size, fixing τ at 100 generations and allowing ω to vary.

Results of the multinomial calculations are shown in Table 15 with consistent scaling across all models, despite the fact that different scaling was used in parameter estimation. We report values of both ω and $1/\omega$ for easier interpretation in the B1a and B1b models, where ω (N_{eDOG}/N_{eBreed}) is scaled by the current dog effective population size and $1/\omega$ (N_{eBreed}/N_{eDOG}) is scaled by the ancestral effective population size. For the B2a and B2b models, because we have more parameters, we only report values of ω and ω_B scaled by the current effective population size. τ is given in generations. Models beyond the B1a contraction model, fixing τ at 100 generations, are not significant.

In the Poisson calculation, because we use a value of θ equal to the θ of wolves, we perform scaling in the estimation based on the ancestral population. Results of the Poisson likelihood calculations are shown for each of the five breeds (Table 16), with parameters given as in the multinomial results of Table 15: timing in generations, ω in terms of the current effective population size (N_{eDOG}/N_{eBreed}), and $1/\omega$ as N_{eBreed}/N_{eDOG}. As in the multinomial calculations, models beyond the B1a contraction model are not significant. Although the B2a and B2b models were calculated, because they are not

significant, they are not shown.

Table 15. Results of PRFREQ analysis of sequence data for breed bottlenecks for the multinomial calculation. All τ are given in number of generations from the present, and values of ω are given in terms of the current breed size (i.e., $\omega = N_{eDOG}/N_{eBreed}$). $1/\omega$, the size change parameter in terms of the ancestral population size (i.e. N_{eBreed}/N_{eDOG}), is also reported. p-values are given for the comparisons in parentheses using the χ^2 distribution.

Multinomial Calculation					
Akita					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	257.659	--	Constant Size
B1a	1	$\tau = 100$ $\omega = 172.41$ $1/\omega = 0.0058$	274.355	7.53×10^{-9} (B1a vs. B0)	Contraction
B1b	2	$\tau = 3068$ $\omega = 9.5$ $1/\omega = 0.105$	274.952	0.274 (B1b vs. B1a)	Contraction
B2a	3	$\tau = 1919$ $\tau_B = 95.9$ $\omega_B = 0.08$ $\omega = 9$	274.954	0.486 (B2a vs. B1b)	Contraction, expansion (bottleneck)
B2b	4	$\tau = 1919$ $\tau_B = 143.9$ $\omega_B = 0.11$ $\omega = 9$	274.954	0.974 (B2b vs. B2a)	Contraction, expansion (bottleneck)
Bernese Mountain Dog					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	128.507	--	Constant Size
B1a	1	$\tau = 100$ $\omega = 370.37$ $1/\omega = 0.0027$	153.610	1.39×10^{-12} (B1a vs. B0)	Contraction
B1b	2	$\tau = 1679$ $\omega = 49.5$ $1/\omega = 0.0202$	154.283	0.246 (B1b vs. B1a)	Contraction
B2a	3	$\tau = 302$ $\tau_B = 86.3$ $\omega_B = 0.016$ $\omega = 10$	154.618	0.413 (B2a vs. B1b)	Contraction, expansion (bottleneck)
B2b	4	$\tau = 302$ $\tau_B = 91.8$ $\omega_B = 0.02125$ $\omega = 10$	154.618	0.972 (B2b vs. B2a)	Contraction, expansion (bottleneck)
Golden Retriever					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	211.800	--	Constant Size

B1a	1	$\tau = 100$ $\omega = 158.73$ $1/\omega = 0.0063$	227.499	2.10×10^{-8} (B1a vs. B0)	Contraction
B1b	2	$\tau = 5613$ $\omega = 10$ $1/\omega = 0.1$	228.574	0.143 (B1b vs. B1a)	Contraction
B2a	3	$\tau = 5613$ $\tau_B = 86.4$ $\omega_B = 0.7$ $\omega = 10$	228.574	0.988 (B2a vs. B1b)	Contraction, expansion (bottleneck)
B2b	4	$\tau = 5613$ $\tau_B = 38.9$ $\omega_B = 0.5$ $\omega = 10$	228.574	0.999 (B2b vs. B2a)	Contraction, expansion (bottleneck)
Labrador Retriever					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	200.332	--	Constant Size
B1a	1	$\tau = 100$ $\omega = 161.290$ $1/\omega = 0.0062$	215.957	2.27×10^{-8} (B1a vs. B0)	
B1b	2	$\tau = 3948$ $\omega = 8.75$ $1/\omega = 0.114$	216.727	0.215 (B1b vs. B1a)	Contraction
B2a	3	$\tau = 3948$ $\tau_B = 98.7$ $\omega_B = 1.75$ $\omega = 8.75$	216.727	0.987 (B2a vs. B1b)	Contraction, contraction
B2b	4	$\tau = 4145$ $\tau_B = 1554.6$ $\omega_B = 1$ $\omega = 8.75$	216.731	0.925 (B2b vs. B2a)	Contraction
Pekingese					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	132.055	--	Constant Size
B1a	1	$\tau = 100$ $\omega = 344.83$ $1/\omega = 0.0029$	163.497	2.22×10^{-15} (B1a vs. B0)	Contraction
B1b	2	$\tau = 69$ $\omega = 500$ $1/\omega = 0.002$	163.512	0.864 (B1b vs. B1a)	Contraction
B2a	3	$\tau = 86.4$ $\tau_B = 1.7$ $\omega_B = 0.8$ $\omega = 500$	163.512	1 (B2a vs. B1b)	Contraction, expansion (bottleneck)
B2b	4	$\tau = 34.5$ $\tau_B = 34.5$ $\omega_B = 0.5$ $\omega = 500$	163.522	0.884 (B2b vs. B2a)	Contraction, expansion (bottleneck)

Table 16. Results of PRFREQ analysis of sequence data for breed bottlenecks for the Poisson calculation. All τ are given in number of generations from the present, and values of ω are given in terms of the current breed size (i.e., $\omega = N_{eDOG}/N_{eBreed}$). $1/\omega$, the size change parameter in terms of the ancestral population size (i.e. N_{eBreed}/N_{eDOG}), is also reported. p-values are given for the comparisons in parentheses using the χ^2 distribution.

Poisson Calculation					
Akita					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	127.609	--	Constant Size
B1a	1	$\tau = 100$ $\omega = 81.967$ $1/\omega = 0.0122$	142.179	6.74E-08 (B1a vs. B0)	Contraction
B1b	2	$\tau = 91.762$ $\omega = 88.889$ $1/\omega = 0.01125$	142.179	0.960 (B1b vs. B1a)	Contraction
Bernese Mountain Dog					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	18.336	--	Constant Size
B1a	1	$\tau = 100$ $\omega = 181.818$ $1/\omega = 0.0055$	55.746	0 (B1a vs. B1a)	Contraction
B1b	2	$\tau = 755.689$ $\omega = 25.807$ $1/\omega = 0.03875$	55.787	0.773 (B1b vs. B1a)	Contraction
Golden Retriever					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	86.773	--	Constant Size
B1a	1	$\tau = 100$ $\omega = 91.743$ $1/\omega = 0.0109$	106.167	4.72E-10 (B1a vs. B0)	Contraction
B1b	2	$\tau = 91.762$ $\omega = 100$ $1/\omega = 0.01$	106.167	0.987 (B1b vs. B1a)	Contraction
Labrador Retriever					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	78.186	--	Constant Size
B1a	1	$\tau = 100$ $\omega = 105.263$ $1/\omega = 0.0095$	99.068	1.03E-10 (B1a vs. B0)	Contraction
B1b	2	$\tau = 367.049$ $\omega = 28.571$ $1/\omega = 0.035$	99.090	0.832 (B1b vs. B1a)	Contraction

Pekingese					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	20.096	--	Constant Size
B1a	1	$\tau = 100$ $\omega = 178.571$ $1/\omega = 0.0056$	61.434	0 (B1a vs. B0)	Contraction
B1b	2	$\tau = 64.773$ $\omega = 285.714$ $1/\omega = 0.0035$	61.447	0.874 (B1b vs. B1a)	Contraction

In order to make comparisons between breeds, we examine the B1a model under the Poisson and multinomial calculations (Table 15, Table 16). Because of the relationship between the two parameters, a severe population contraction could be represented by either an ancient τ and mild ω or a severe ω and recent τ. Fixing τ at 100 generations across all breeds, although it may not be historically accurate, allows the severity of a bottleneck to be reflected in estimates of ω that can be compared across breeds. For the multinomial calculation (Table 15), the breed with the strongest contraction is the Bernese Mountain Dog, with a current effective population size approximately 370 times smaller than that of the ancestral dog effective population size. The next severe contraction is for the Pekingese, whose value of $\omega = 345$ is similar to that of the Bernese Mountain Dog. Next is the Akita, with an estimated 179-fold contraction, the Labrador Retriever with a 161-fold contraction, and finally the Golden Retriever with a 159-fold contraction.

The order of the contraction severities in the B1a model among breeds for the Poisson calculation (Table 16) is rather comparable, though differences exist. The Bernese Mountain Dog is estimated to have had the strongest contraction of 182-fold, followed by the Pekingese with a similar ω of 179. Next is the Labrador Retriever, with

a contraction of 105-fold, the Golden Retriever, with a contraction of approximately 91-fold, and lastly the Akita, with a value of 82. Overall, it appears that multinomial calculated estimates are more severe.

We plot the results of the B1a models for both calculations in Figure 19. We observe that the predicted site frequency spectra do not account for every aspect of the shapes of the observed SFS, most notably for the Pekingese and Bernese Mountain Dog.

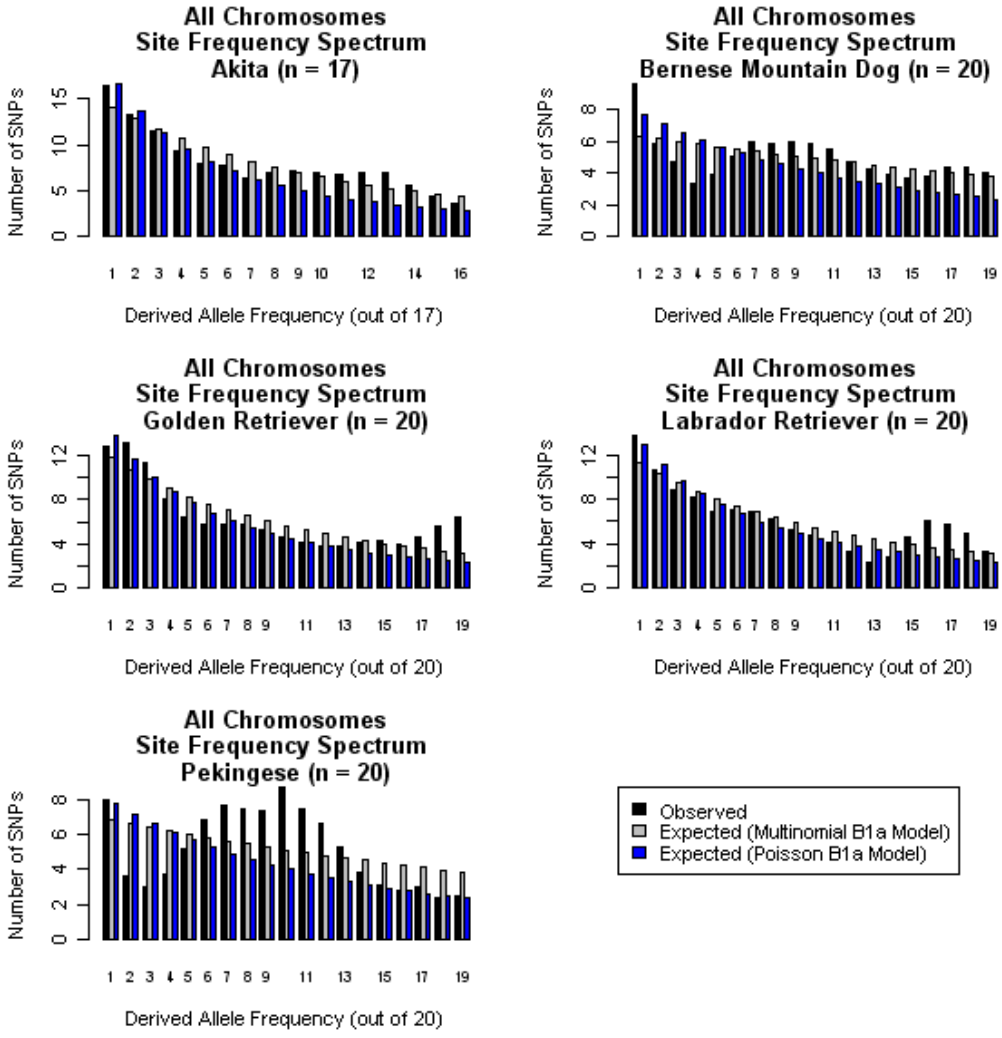

Figure 19. Site frequency spectra of data sampled one chromosome per individual as described in text for each breed. Black bars are the observed data, and gray and blue bars are the expectations under the contraction (B1a) models from the PRFREQ multinomial and Poisson calculations, respectively, as indicated in Table 15 and Table 16. x-axis is the derived allele frequency out of n as indicated, and the y-axis is the number of SNPs with derived allele at that frequency.

We also examine and compare the results of the B1b inference for both the Poisson and the multinomial calculations, where values of ω and τ are optimized (Table 15 and Table 16, plotted in Figure 20). Generally, it appears that the multinomial calculations estimate more ancient contractions while the Poisson estimates more recent ones; this causes the contraction parameters of the multinomial to be much smaller than in the Poisson. For example, whereas for the Akita the Poisson calculation estimates ω to equal approximately 44, the multinomial results in a value of 9.5. In another example, the Poisson calculation for the Golden Retriever picks up a strong contraction with ω = 100, whereas the multinomial estimate of ω is equal to 19. Interestingly, some of the calculated values of τ from the multinomial B1b model do not make logical sense given the timing of dog domestication. For instance, the Golden Retriever has a timing estimated at 5613 generations ago, larger than the assumed domestication timing of 5000 generations. The excess of high frequency derived alleles in the site frequency spectra may cause the estimate of a long, sustained population decline for many of the breeds in the multinomial calculation. Interestingly, Poisson and multinomial estimates for the Pekingese timing are very similar, approximately 60 generations ago.

We compare the ordering of the estimates of the timing of breed formation between the multinomial and Poisson calculations. In the multinomial calculation (Table 15), the breed with the earliest population decline is the Golden Retriever, followed by the Labrador Retriever, the Akita, the Bernese Mountain Dog, and finally the Pekingese, with the most recent population contraction. The order of timing is rather different for the Poisson calculation (Table 16): the breed with the earliest contraction is the Bernese Mountain Dog, followed by the Labrador Retriever, the Golden Retriever and Akita, and

the Pekingese. Implications of these timing and size change parameter estimates in relation to the formation of these five dog breeds are to be discussed.

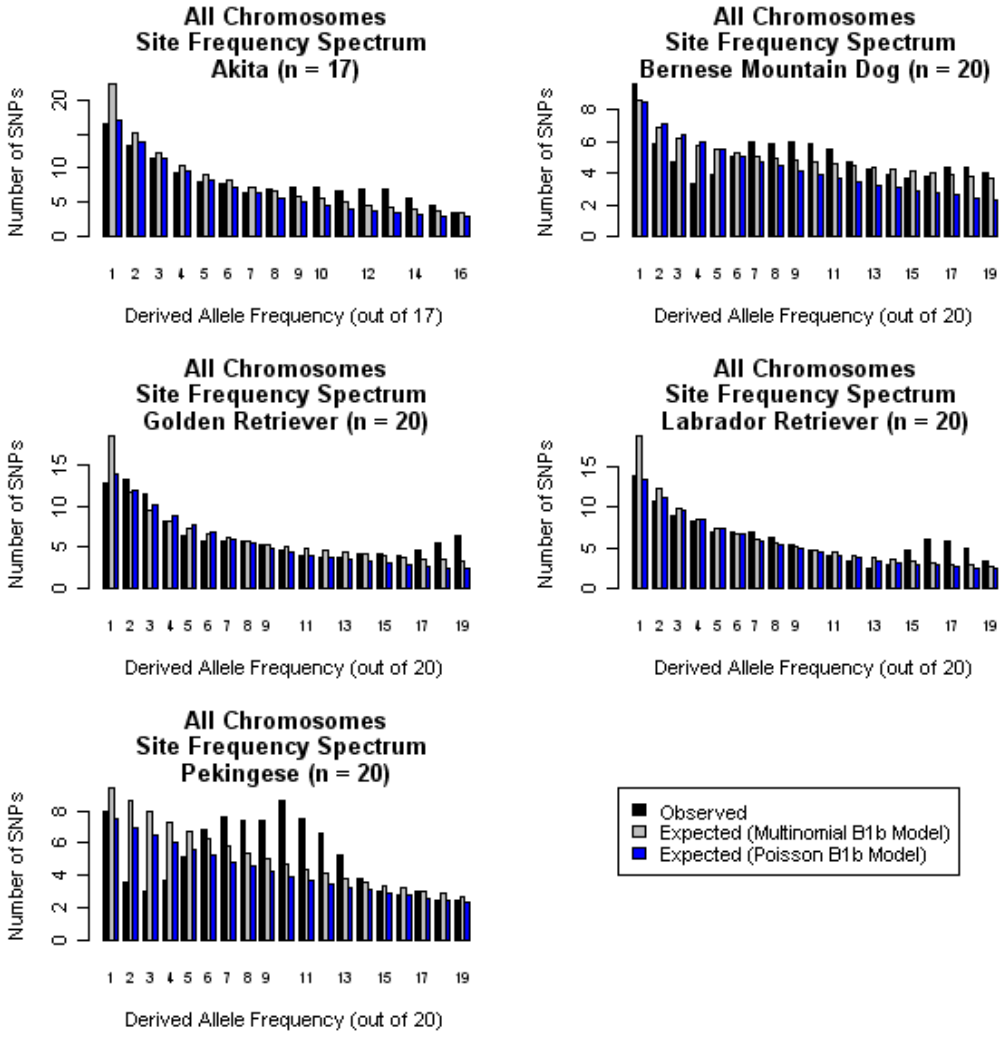

Figure 20. Site frequency spectra of data sampled one chromosome per individual as described in text for each breed. Black bars are the observed data, and gray and blue bars are the expectations under the contraction (B1b) models from the PRFREQ multinomial and Poisson calculations, respectively, as indicated in Table 15 and Table 16. x-axis is the derived allele frequency out of n as indicated, and the y-axis is the number of SNPs with derived allele at that frequency.

Finally, we compare likelihoods between the multinomial and Poisson calculations as for the domestication event and described in [IV. Methods, Analysis program PRFREQ]. Values of θ calculated from the multinomial are entered in the calculation of the Poisson likelihood using the demographic parameters estimated by the multinomial in Table 15. These Poisson likelihoods with multinomial parameters, the original Poisson likelihoods, and the associated p-values comparing the two likelihoods are shown for each breed and model in Table 17. For all breeds, the multinomial neutral (B0) likelihoods are significantly higher than those of the Poisson; values of the multinomial θ tend to be less than the Poisson θ.

For the non-neutral B1a models, the values of θ estimated from the multinomial calculation are greater than the values of $\theta = 44.925$ used in the Poisson calculation and are even greater for the B1b models. However, the values of θ estimated do not seem as unrealistic as they did for the domestication event (Table 12). It is possible that the values of θ calculated with the multinomial calculation in breed formation may not be entirely inappropriate and that the effective population size of "pre-breed" dogs is indeed greater than the wolf effective population size. This causes the multinomial parameter estimates to indicate more severe population contractions than those of the Poisson. In addition, it is also possible that the multinomial calculation is overfitting to the given data and that we may have limited power to estimate both θ and a contraction.

Table 17. Results of rescaling multinomial likelihoods for comparison between multinomial and Poisson calculations for the given models and breeds. p-value is obtained by taking twice the difference in log likelihood and using the χ^2 distribution with 1 df. A p-value < 0.05 indicates that the likelihood under the multinomial calculation is significantly greater than the likelihood under the Poisson calculation (indicated by asterisks).

Breed	Model	θ (Estimated from Multinomial)	Poisson LL (Multinomial Parameters)	Poisson LL (Poisson Parameters)	p-value
Akita	B0	37.8837	129.5808	127.6094	0.0471*
	B1a	69.2242	146.2770	142.1789	0.0042 *
	B1b	81.0968	146.8743	142.1790	0.0022 *
Bernese Mountain Dog	B0	26.7434	33.6259	18.3361	3.20×10^{-8} *
	B1a	81.7190	58.7284	55.7457	0.0146 *
	B1b	206.4694	59.4018	55.7873	0.0072 *
Golden Retriever	B0	33.7051	92.2189	86.7733	9.66×10^{-4} *
	B1a	61.0294	107.9182	106.1674	0.0613
	B1b	117.165	84.8953	78.1865	0.0174 *
Labrador Retriever	B0	32.5372	100.5198	99.06760	2.49×10^{-4} *
	B1a	59.3297	101.2895	99.0901	0.088336
	B1b	77.4708	84.89532	78.1865	0.035963 *
Pekingese	B0	27.7515	33.5975	20.09616	2.03×10^{-7} *
	B1a	79.8652	65.03907	61.4344	0.00725 *
	B1b	80.1	65.05372	61.4469	0.00724 *

Assessment of Model Significance with Coalescent Simulations

First, we run 2000 coalescent simulations under a neutral model using msHOT in to account for recombination between SNPs and amplicons. To mirror the data set of the Akita, we set the number of chromosomes in the sample equal to 17; for the Bernese Mountain Dog, Golden Retriever, Labrador Retriever, and Pekingese, we set the sample size equal to 20. We input the SFS from each of these simulations into PRFREQ to obtain the multinomial composite likelihood of the neutral (B0) model. We also obtain the composite log likelihood from optimizing the neutral simulated data under the multinomial B1a model, fixing τ at 100 generations and allowing ω to vary from 0.1 to 3.1. We obtain the likelihood ratio test statistic for the two models (Figure 21) and the

95% confidence interval of the LRT statistic using the upper and lower 2.5% values. This yields confidence intervals of (0.000102, 2.949070) for the Akita and (0.00013, 3.33209) for all other breeds.

Examining the LRT statistic between the B1a and B0 models, the observed difference in likelihood for all breeds is much greater than what is expected under neutrality. The Akita (33.392), Bernese Mountain Dog (50.205), Golden Retriever (31.399), Labrador Retriever (31.249), and Pekingese (62.883), all have a value of the LRT statistic for the B1a and B0 models lying well outside of their respective confidence intervals. This indicates that the models that we estimate from the B1a multinomial contraction models are indeed significant, even when sites are in fact linked.

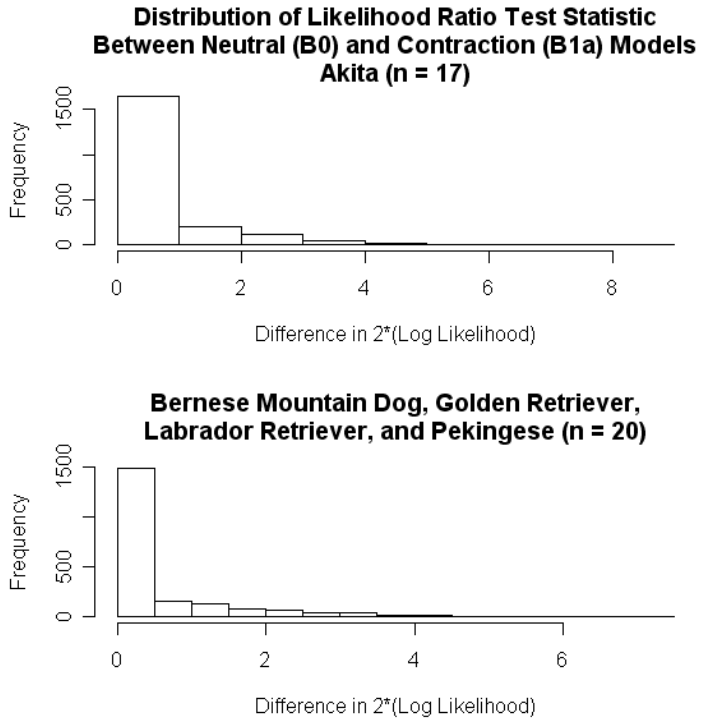

Figure 21. Distribution of likelihood ratio test statistic between the optimized contraction (B1a) model and neutral (B1) model for 2000 neutral coalescent simulations. Multinomial calculations are used. *Upper Panel:* Results for simulated data of the Akita (n = 17). *Lower Panel:* Results for the Bernese Mountain Dog, Golden Retriever, Labrador Retriever, and Pekingese (n = 20).

Interpretation

Before analyzing the results of the demographic inference in the context of each dog breed individually, we look at general results of the breed formation analyses and compare the results of the Poisson and multinomial calculations. The multinomial and Poisson calculations yield different estimates for the timing and severity of breed contractions for both the B1a and B1b models. As in both Figure 19 and Figure 20, it appears that the Poisson calculation fits more closely to the observed site frequency spectrum in lower frequency SNP categories, whereas the multinomial estimates appear to fit the observed site frequency spectrum more accurately for higher frequency SNPs. This again is a property of the calculations used; while the multinomial depends only on the shape of the site frequency spectrum, the Poisson calculation is heavily affected by the number of segregating sites and singletons.

The multinomial is influenced by an excess of middle to high frequency variants in the SFS. An excess of high frequency derived alleles is seen in the SFS of several breeds, most notably of the Golden and Labrador Retrievers (Figure 19). This excess could be due to a number of factors, such as ancestral misspecification, where the Golden Jackal allele is genotyped incorrectly. This would cause a SNP to appear to be at high frequency when in actuality it is at a lower frequency. Another possible cause of an excess of high frequency derived alleles is a breed founding event followed by the introduction of alleles from outside of the population. Either of these possibilities could play a role in the observed SFS, causing the multinomial calculation to possibly overfit a contraction model and perceive a more severe contraction.

The Poisson calculation, highly influenced by the number of singletons, picks up a less severe contraction. Although we use a θ for the Poisson calculation that assumes that the "pre-breed" dog population is the same size as the current wolf population, this may not be entirely accurate. A different value of θ would indeed change the Poisson composite likelihood and therefore the parameter estimates obtained. However, the scenario we propose does seem plausible, given that diversity levels are rather comparable between all dogs and all wolves and that there was likely not a very severe population decline at dog domestication.

For each breed and for both the Poisson and multinomial inferences, only the contraction models are significant. In order to detect a more complicated bottleneck model (a contraction followed by a subsequent expansion), we need to detect mutations that have arisen after the founding event. Since individual dog breeds are quite young, only a small number of SNPs have had time to arise during a recent expansion, if one has occurred. Given our limited data, we likely do not sample these rare SNPs, and therefore see no evidence of a bottleneck. It is also likely that while the census population sizes of the breeds studied have increased from the original number of founders, there has likely not been an expansion in the effective population size. With the use of popular sires and strong inbreeding within pure breeds (Ostrander and Kruglyak 2000), the effective population size of a breed likely has remained quite small. Evidence for only a contraction at breed formation is not surprising given our knowledge of breeding programs and the data we have available.

We also attempt to account for inbreeding in this study by sampling one chromosome from each individual of a breed. However, we have not performed an

identical analysis on the non-sampled data for each breed. In future studies, it may be interesting to do such an analysis to determine the effect of accounting for inbreeding in a demographic analysis. Also, we only examine the sequence data of the initial five breeds and have not explored the demographic histories of the additional 17 breeds genotyped. Though it is another interesting area for future research, it is possible that we may not have enough power to infer demography with the limited number of SNPs in each breed.

Finally, when examining the maximum likelihood estimates of ω from the multinomial and Poisson calculations, it appears that our results are rather consistent with what is currently known about individual dog breeds. The Pekingese and Bernese Mountain Dog appear to have undergone the most severe population declines in their histories, while common breeds such as the Golden Retriever and Labrador Retriever each appear to have undergone only mild population reductions. The Akita appears to have had a contraction similar to that of the Golden and Labrador Retrievers. Below, we examine our analyses of breed contraction in more detail for each dog breed; although we cannot place very much confidence in the various historical accounts available of breed origins (Dangerfield and Howell 1971), we interpret our results in light of this prior knowledge.

Akita

The Akita, or Akita-Inu, was traditionally a major breed in Japan, believed to be developed in the 17th century as a "large, versatile, intellectual hunting dog" (AKC 1997). Ownership of the Akita was once restricted to the Imperial family and aristocracy. Throughout the next 300 years, as the interest in selective breeding fluctuated in Japan, the breed suffered from several severe extinctions (AKC 1997). The first Akita was

registered with the American Kennel Club (AKC) in 1972 and is classified as a working breed, with 3200 registrations in 2003 (Sutter et al. 2004).

We find evidence in the Akita for a significant population contraction, though less severe than contractions found in most other breeds. When fixing τ at 100 generations, we find evidence for a 82-fold contraction, the least severe contraction estimated out of all five breeds for the Poisson calculation. When using the multinomial calculation, we detect that the Akita underwent a 172-fold population contraction, also the least severe. According to the Poisson calculation in the B1b model, the Akita underwent a 90-fold population contraction approximately 90 generations, or 270 years, ago (Table 16). Compared to the Labrador Retriever and the Bernese Mountain Dog, this contraction is rather recent. Although the contraction (both B1a and B1b) models do provide a better fit to the data than the neutral model, they do not account for the slight increase of higher frequency derived alleles in the Akita (Figure 19 and Figure 20). Our predicted model appears to not fully explain a more complicated demography of the Akita.

Although the Akita is known to be an ancient breed, which genetically clusters with the gray wolf (Parker, et al. 2004), we do not detect an ancient contraction in our analyses. It is possible that rather than finding evidence of a contraction at breed formation, we are detecting artifacts of the extinctions of the Akita known to have occurred between the 1600's and 1900's. That we do not see a population decline more severe than the Bernese Mountain Dog or Pekingese agrees with what is known about LD in the Akita, which does not extend as long as it does in either of these breeds (Sutter et al. 2004). Sutter et al. also find that the Akita has the lowest percentage of SNP pairs in

LD compared to the other four breeds, agreeing with the milder population decline we estimate.

For a breed that has neared extinction, we might expect the contraction severity to be greater than our analysis suggests. If fluctuations in population size did occur to the Akita due to varying interest in selective breeding (AKC 1997), subsequent increases and decreases in population size may decrease signatures of a severe contraction. Another explanation for this result is that United States Akita breeders have historically disagreed on which lines were the true "pure" lines of the Akita, creating substructure of multiple lines within the breed (Sutter, personal correspondence). As the samples used in this study are also likely composed of dogs from multiple lines, substructure within the Akita samples could produce the result of a less severe contraction. In addition, it is possible that continued hybridization between the gray wolf and Akita has minimized the effects of a severe contraction as well as the extent of LD in the Akita.

Bernese Mountain Dog

The Bernese Mountain Dog is classified by the AKC as a working breed, which clusters with other larger breeds such as the Mastiff (Parker et al. 2004). Invading Roman soldiers are believed to have introduced the breed's ancestors to Switzerland over 2000 years ago (AKC 1997). The breed almost disappeared in the 1800's; however, in the late 1800's, Herr Franz Schertenleib worked to find breeding stock and revive the breed (Fogle 2000). The number of individuals used to form the breed is believed to be quite small due to its severe contraction before its rehabilitation; according to the International Encyclopedia of the Dog (Dangerfield and Howell 1971), the Bernese Mountain Dog is one of the "most unfortunate" breeds. In the 1960's, there were still

less than 50 dogs registered annually, although the breed has expanded with 103 registered in 1970 (Dangerfield and Howell 1971), and 3100 in 2003 (Sutter et al. 2004).

In B1a analysis with τ set at 100 generations, we find evidence of a very severe contraction of 180-fold with the Poisson calculation (Table 16) and of 370-fold with the multinomial calculation (Table 15). According to the known breed history described above, the timing of 100 generations encompasses the time during which the breed had degenerated as well as during which the breed slightly increased in size during its rehabilitation. Compared to other breeds, this is the most severe population contraction, supporting the claim of an "unfortunate" and severe population decline in the past. When we allow the timing of contraction to vary for the Poisson (B1b) model, we see a less severe contraction that has occurred the furthest in the past, indicative of a severe and sustained population contraction.

For both the multinomial and Poisson calculations of in the B1a (Figure 19) and B1b (Figure 20) models, the predicted site frequency spectra are quite flattened compared to the neutral site frequency spectrum and seem to provide a much better fit to the observed data. However, the simple contraction demographic model does not account for the large peak of middle to high frequency SNPs observed. As with the Akita, this may reflect our limited power to detect more detailed events in population history.

This very strong contraction agrees with results of Sutter et al. (2004), where the breed was found to have long-range LD shorter than only the Pekingese by 0.3 Mb. In addition, Sutter et al. find haplotype sharing among the breed to be high, where for one particular region, one haplotype could account for 80% of the breed's chromosomes. Both the Bernese Mountain Dog's sustained population contraction and its founding

event appear to largely effect the site frequency spectrum we observe as well as the breed's properties of linkage disequilibrium.

Golden Retriever

The Golden Retriever, classified as a sporting breed, is believed to have ancestors used for game retrieving in the 1880's in England and Scotland (AKC 1997). In contrast to other breeds, the specific founder individuals of the Golden Retriever are known and recorded. In the late 1860's, the first yellow retriever (named "Nous") was crossed with a Tweed Water Spaniel (a now-extinct breed) named "Belle," who gave birth to four puppies. Along with the addition of other dogs such as Irish Setters, other Tweed Water Spaniels, and smaller Newfoundlands, these puppies appeared many times in the pedigree of the breed (AKC 1997). The Golden Retriever breed was very popular by the end of the 19th century and was registered by the AKC in 1925. In 1936, breed standards in England and Scotland changed to allow lighter and darker colors of the Golden Retriever, which were subsequently brought over to America (AKC 1997). Currently, the Golden Retriever is the 2nd most popular breed only after the Labrador Retriever (Sutter and Ostrander 2004).

Results from our analyses support this history. The Poisson B1a model estimates a population contraction 100 generations ago of approximately 92-fold, only greater than the value estimated for the Akita ($\omega = 82$, Table 16). Likewise, the multinomial calculation estimates that the Golden Retriever has had the least severe contraction of all breeds (Table 15). The B1b Poisson model yields similar estimates to the B1a model, with a τ of 92 generations (276 years) and ω of 100. Since this timing corresponds to a contraction in the 1700's, slightly predating the formation of the breed in the initial

"Nous x Belle" cross of the 1860's, this supports that there was indeed a contraction in the breed's formation. The rather mild contraction we estimate is supported by the fact that long-range LD decreases the quickest in the Golden Retriever (Sutter et al. 2004), with D' falling to half of its initial value in 370 kb as opposed to 3.3 Mb in the Pekingese.

The predicted SFS based on the contraction models fit the shape of the observed SFS slightly better than for the Akita and the Bernese Mountain Dog, with the exception of an observed increase in high frequency SNPs (Figure 19 and Figure 20). As mentioned previously, one possible cause of this increase is ancestral misspecification. Another possible explanation is substructure within the Golden Retriever population, possibly caused by the different coat colors (Light Golden, Golden, and Dark Golden) among the breed. Finally, another explanation is that the Golden Retriever was formed from a myriad of other breeds introduced into its breeding pool. Supporting this hypothesis, each of the four other breeds studied shares a high proportion of haplotypes with the Golden Retriever (Sutter et al. 2004). These explanations likely are the cause of the perceived increase in high frequency derived alleles in the Golden Retriever SFS.

The Golden Retriever has undergone a rapid population expansion after its formation, as the breed was rather popular by the end of the 19th century (AKC 1997). The breed's effective population size also likely increased due to the introduction of individuals with different coat colors or the introduction of other breeds into the breeding pool. We do not detect a subsequent expansion in the population size of the breed, as we likely do not have the power to do so. However, a subsequent population expansion likely minimizes the effects of the initial founding event, making our estimate of the founding contraction less severe. It appears that as a whole, the Golden Retriever has not

undergone an extremely severe bottleneck, most likely as a result of its strong popularity as a breed and thus larger current effective population size.

Labrador Retriever

The Labrador Retriever is believed to have been developed in Newfoundland from a small number of water dogs of local fishermen (AKC 1997). The breed is reported to have faded in popularity in Newfoundland as a result of a local heavy dog tax, while a quarantine law prevented importation into England and reduced the breeding of Labradors in Newfoundland (Dangerfield and Howell 1971). It is believed that two dogs in the 1870's contributed the most to produce the modern Labrador (AKC 1997). The Labrador was recognized in England in 1903 as a separate breed and is classified as a sporting breed. Currently, it is the most popular dog breed.

Examining results of the Poisson B1a model, we see evidence for an approximate 100-fold contraction approximately 100 generations ago (Table 16), a more severe contraction than both the Golden Retriever and Akita. For the multinomial calculation, we see a 161-fold contraction, only slightly more severe than that of the Golden Retriever (Table 15). When we allow τ to vary, we see evidence for a much earlier, but less severe, contraction approximately 367 generations ago (Table 16). This value of τ calculated by the Poisson approximation is the second-largest value of τ for all breeds. This could indicate that we are detecting evidence of a contraction at breed formation, rather than evidence of demographic events occurring afterwards. Again, our results agree with those of Sutter et al. (2004), who found the Labrador to have a range of LD slightly longer than that of the Golden Retriever.

The fact that two dogs are believed to have a large influence in the pedigrees of the Labrador Retrievers could result in the larger contraction predicted from the B1a model compared to that of the Golden Retriever. As with the Golden Retriever, there is a large peak in high frequency SNPs in the observed SFS (Figure 18), which could be due to substructure within the breed as well as the introduction of individuals into the breeding pool. Sutter et al. (2004) find the Labrador Retrievers to have the highest average number of haplotypes, supporting the hypothesis of introduction of alleles into the population. This also could reduce the severity perceived for the contraction in the Labrador Retriever. These results emphasize that breed histories are in fact very complex, and may not be fully explained by these two-epoch models and our limited data.

This analysis of the Labrador Retriever places its formation in context among the other breeds examined. While it appears to have had a more severe population contraction in its formation than the Golden Retriever, it has not undergone nearly as strong of a population contraction as the Bernese Mountain Dog or the Pekingese.

Pekingese

Dogs of the Pekingese breed were considered sacred in China in ancient times, with the earliest known record of the Pekingese in the Tang Dynasty of the 8th century (AKC 1997). The Pekingese clusters with ancient breeds such as the Akita, Shiba-Inu, and the Gray Wolf (Parker et al. 2004), and likely formed as a breed more than 1500 years ago (Dangerfield and Howell 1971). During the 1820's to 1850's, the Pekingese was not often found outside of the Chinese Imperial Palace. However, in 1860, the Imperial Palace was looted by Europeans, and between four and five Pekingese were

found and stolen. These dogs were brought to Europe, where three dogs, named "Ah Chum," "Mimosa," and "Boxer," were used primarily in the breed's development (AKC 1997). The breed was registered by the AKC in 1906 and is classified as a toy breed, with roughly 4700 registrations in 2003 (Sutter et al. 2004)

The Pekingese has undergone at least two bottlenecks; first, in the formation of the breed from the ancestral dog population, and second, in its introduction to the Western world. It is not surprising that our estimate of the intensity of the contraction for the breed is extremely large – nearly 180-fold for the Poisson B1a model and 344-fold for the multinomial calculation. This is supported by the fact that LD extends the longest in the Pekingese and that the breed has the lowest average number of haplotypes compared to the other four breeds (Sutter et al. 2004). When allowing τ to vary in the Poisson B1b model, we see a more recent and severe contraction of approximately 286-fold occurring 65 generations ago (Table 16). Interestingly, the multinomial detects a similar timing of 69 generations and a contraction of 500-fold (Table 15). This similarity may imply that the wolf population size does indeed reflect the ancestral dog effective population size of the Pekingese.

In examination of the site frequency spectrum, we see that this contraction again does not explain the entire shape of the observed site frequency spectrum. Much more so than in the other breeds, we see an extremely large excess of middle frequency derived alleles (Figure 19 and Figure 20), possibly indicative of multiple bottlenecks. Because the three-epoch B2a and B2b models are not significant, we do not find evidence for such a scenario. This again indicates the lack of power of our methods to detect multiple changes in demography.

The results of the Pekingese appear to be consistent with the small number of individuals brought into the Western world for the formation of the modern Pekingese breed. This extreme recent population reduction seems to overshadow the more ancient bottleneck that occurred in the formation of the Pekingese from the ancestral dog population. Our results for Pekingese demography agree with a strong recent bottleneck due to the importation of few individuals into the Western world.

Breed Conclusions

From our use of PRFREQ, we have attempted to find signatures of breed formation from the site frequency spectra of various breeds. Although we test more complicated models, we do not find a significant model other than a simple, one-time contraction model. This is somewhat expected, as continued inbreeding within a given breed may act to maintain a small effective population size, even if the census population size has in fact increased since breed formation.

In general, the results of the analyses agree with what is known about the particular breeds as well as with previous studies conducted on the same breeds. The Pekingese and Bernese Mountain Dog both were estimated to have very severe population contractions, agreeing with their histories of a small number of founder individuals and sustained small population size. The Golden Retriever and Labrador Retriever are more common breeds, where signatures of a contraction are lessened by subsequent population increases and introduction of individuals into the breeding pool. Finally, we perceive a population contraction in the Akita somewhat similar to that observed for the Golden and Labrador Retrievers, possibly due to the Akita's substructure or hybridization with the gray wolf. That our results agree with prior

estimates of linkage disequilibrium in the same breeds (Sutter et al. 2004) suggests that our methods could possibly be used to predict levels of LD in other breeds.

We have only performed an analysis of the five initial dog breeds originally sequenced. In the future, it would be interesting to examine the demographic histories of the 17 additional dog breeds sequenced to make additional comparisons between breed histories. Furthermore, additional methods more fully accounting for inbreeding within breeds, as well as linkage between sites, may be necessary to increase the power we have to estimate demographic models with multiple epochs and size changes. Finally, another area for future exploration is the modeling of the entire demographic model shown in Figure 1, simultaneously inferring both the domestication and breed formation events. While we likely do not have the power to infer both demographic events with our given data, results of simultaneously estimating dog domestication as well as subsequent breed bottlenecks may be more accurate.

VII. Demographic Analysis of Wild Canids

Analysis with PRFREQ

We also examine the demography of individual wild canid populations to link the demography of the dog to that of its canine relatives. The populations available are four gray wolf populations, from Alaska, Israel, Spain, and Yellowstone National Park, and one coyote population. A major assumption that we have made in the demographic analysis of the domestication event is that wolves have maintained a constant population size throughout time. Using PRFREQ and the same nested likelihood models as we used for inference of dog breed formation (Table 13), we can assess the validity of this assumption.

We do not test the B1a model, which fixes the time of contraction, as we do in the inference of breed formation. Whereas we have a general hypothesis of the timing of breed formation, for individual wild canid populations, we do not know whether we should expect to see a population decline or increase or when such an event has occurred. In inferring the two-epoch contraction model, we optimize both the time and the severity of the contraction in the B1b model. As with the inference of breed domestication, we use both the multinomial and Poisson calculations. We refer to N_{ePOP} as the current effective population size of the population, and N_{eWOLF} as the effective population size of the ancestral wolf population.

Data Manipulation

For the wild canids, we have available sequence data only on chromosome 1, with a total length within amplicons of 11,279 bp. The site frequency spectra of this data for all four wolf populations and coyotes can be seen in Figure 22. Due to limited data,

the site frequency spectra are not very smooth; however, the overall shapes can be seen. At first glance, it appears that the Spain and Israel wolves have the most aberrant site frequency spectra, as they each have an excess of middle frequency variants.

It is known that since wolf packs are generally composed of closely related individuals, there may be slight inbreeding among wolf populations (Lehman et al. 1992). We perform the same correction done for inbreeding done when inferring breed demography (see VI. Demographic Analysis of Breed Formations) by sampling one chromosome from each individual in each population (Figure 23). In examination of the sampled data compared to the unsampled data it appears that the overall shapes are rather similar, perhaps indicating that inbreeding is not a very large influence in the wolf populations. We gather summary statistics for the sampled data, shown in Table 18.

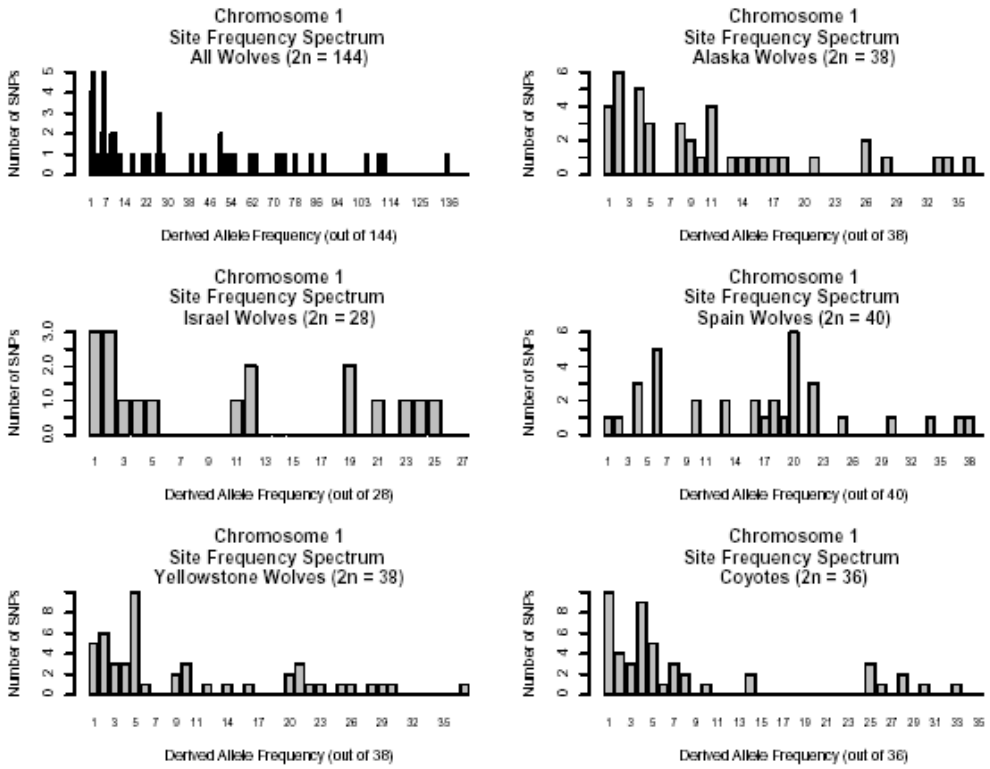

Figure 22. Observed site frequency spectra of sequence data, pooling all chromosomes, for each wild canid population of gray wolves and coyote as indicated. Data is also shown pooling all wolf populations together as a comparison. x-axis is the derived allele frequency out of 2n as indicated, and the y-axis is the number of SNPs with derived allele at that frequency.

Table 18. Summary statistics obtained for each wolf after sampling one chromosome from each individual as described in text. S is the number of segregating sites, and θ is Watterson's estimate of θ.

Canid Population	N	S	θ (per-site)	Number of Singletons	π (per-site)	Tajima's D	Average Heterozgosity
Alaska	19	36.959	0.00094	6.5525	0.00103	0.4229	0.29969
Israel	14	15.297	0.00043	3.7485	0.00044	0.0930	0.29867
Spain	20	33.361	0.00083	2.323	0.00112	1.3702	0.36077
Yellowstone	19	43.649	0.00111	9.0215	0.00119	0.3241	0.29280
Coyote	18	41.654	0.00107	12.0975	0.00096	-0.4410	0.24498

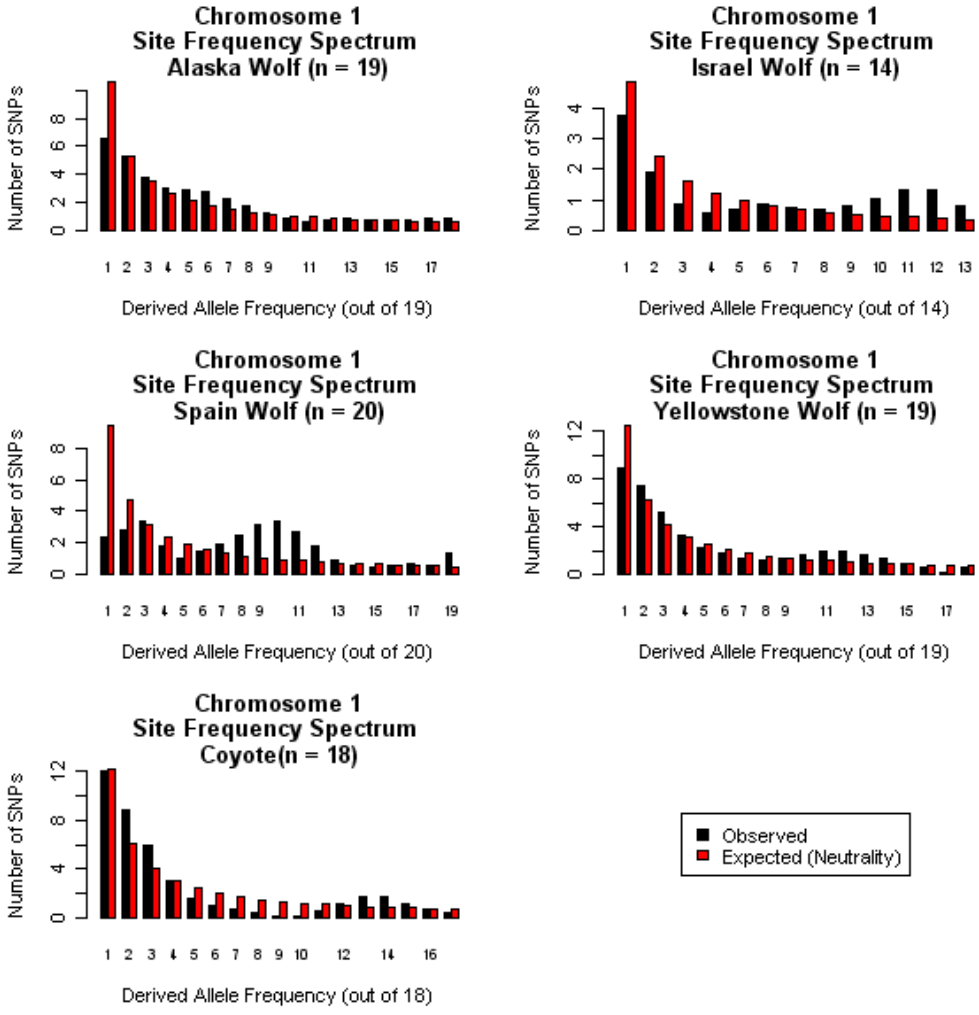

Figure 23. Site frequency spectrum for sequence data sampled one chromosome per individual in each wild canid population as described in text. Black bars are the observed data, and red bars are the expectation under neutrality. x-axis is the derived allele frequency out of n as indicated, and the y-axis is the number of SNPs with derived allele at that frequency.

In examination of the sampled site frequency spectra, we observe that the Spanish wolves have a large peak of middle frequency variants in their site frequency spectrum, indicative of a possible population decline. The Spanish wolves also appear to have a rather positive value of Tajima's D in comparison to the other wolf populations (Table 18), supporting this hypothesis. The Israel wolves also appear to deviate slightly from neutrality, but to a lesser extent than the Spain wolves. We use PRFREQ to explore these deviations and assess the demographic history of these gray wolf populations.

Unlike in the inference of breed formation and domestication, we do not perform coalescent simulations using msHOT to verify the p-values of the composite likelihoods of PRFREQ. This is because we do not infer parameters of the B1a contraction model but only for the B1b model. While it is straightforward to simulate data and obtain likelihoods under the neutral model, in order to obtain likelihoods under the B1b model we must optimize both τ and ω. Given 2000 samples and large possible ranges of τ, this may be too computationally intensive. We show no results of coalescent simulations for the wild canids, although such a verification of our PRFREQ results is an important area that should be explored in future research.

Results

As done for inference of dog domestication and breed formation, we use both the multinomial and Poisson calculations of PRFREQ to infer wild canid demography. When conducting the multinomial inference (Table 19), the sampled site frequency spectra of the Alaska, Israel, and Yellowstone wolf populations all provide no evidence of a significant population decline. Only the Spain wolf population has a B1b contraction model that is significant in relation to the B0 constant size model. The model indicates a

value of ω of 500 (the highest value examined), indicating a very severe population decline approximately 41 generations ago. Although it is not significant, for the coyote population, the maximum likelihood estimate of ω is 0.6, indicating a population expansion. B2a and B2b models are shown only for the Spain wolf population, where the B2a model sets τ_B at 0.02, in units of $2*N_{ePOP}$ generations.

Table 19. Results of PRFREQ analysis of sequence data for wild canid populations for the multinomial calculation. All τ are given in number of generations from the present, and values of ω are given in terms of the current population size (i.e., $\omega = N_{eWOLF}/N_{ePOP}$). 1/ω, the size change parameter in terms of the ancestral population size (i.e. N_{ePOP}/N_{eWOLF}), is also reported. p-values are given for the comparisons in parentheses using the χ^2 distribution. Results for models with increasing degrees of freedom (i.e., B2b), when ones with fewer df (i.e., B1b) are not significant, are not shown.

Multinomial Calculation					
Alaska Wolf					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	35.529569	--	Constant Size
B1b	1	τ = 7.8 ω = 500 1/ω = 0.002	36.756938	0.293063 (B1b vs. B0)	Contraction
Israel Wolf					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	2.8784	--	Constant Size
B1b	1	τ = 518.2 ω = 500 1/ω = 0.002	4.4196	0.2141 (B1b vs. B0)	Contraction
Spain Wolf					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	13.5209	--	Constant Size
B1b	1	τ = 41 ω = 500 1/ω = 0.002	20.7760	0.0007 (B1b vs. B0)	Contraction
B2a	2	τ = 26.3 τ_B = 1.7 ω_B = 0.109 ω = 500	20.7767	0.8731 (B2a vs. B1b)	Contraction, expansion (bottleneck)
B2b	3	τ = 17.3 τ_B = 14.7 ω_B = 0.4 ω = 500	20.7774	0.9690 (B2b vs. B2a)	Contraction, expansion (bottleneck)

Yellowstone Wolf					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	52.2687	--	Constant Size
B1b	1	$\tau = 5.1$ $\omega = 500$ $1/\omega = 0.002$	52.9196	0.5216 (B1b vs. B0)	Contraction
Coyote					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	59.3363	--	Constant Size
B1b	1	$\tau = 19{,}792$ $\omega = 0.6$ $1/\omega = 1.6667$	59.7582	0.6558 (B1b vs. B0)	Expansion

We also conduct inference using the Poisson calculation, scaling all values by the ancestral wolf effective population size. We use the same per-bp θ as used in the demographic inference of dog breeds, 0.00086, as estimated from the number of segregating sites in the wolf data. Multiplying by the number of base pairs in the amplicons of chromosome 1 (11,279), we obtain a value of θ of 9.741 for use in the Poisson likelihood of PRFREQ. The Poisson approximation for wolves may be more appropriate than for dog breeds, as the current wolf population possibly closely estimates the ancestral wolf population. Results for Poisson calculations are shown in Table 20.

Once again, there is a significant difference between the composite likelihoods of the B0 and B1b models for the Spain wolf population. However, we detect a contraction of approximately 36-fold, less severe than the 500-fold contraction estimated from the multinomial calculation. The time of this contraction is placed at 226 generations, or slightly less than 700 years, ago. As with the multinomial, the bottleneck B2b model is not significantly different than the contraction model.

With the Poisson calculation, we also find that the Israel wolf population has undergone a significant 4-fold population contraction over 10,000 generations, or 30,000 years, ago. This timing estimate may not be entirely realistic, though it may represent a tradeoff between a recent τ and a severe population decline and a more distant τ but less severe population decline. Interestingly, the B2b model detects evidence for two successive contractions of small size, although it is not significant. While an expansion is again detected from the coyote B1b model, it is not significant.

The observed and expected site frequency spectra under the various models are plotted in Figure 24, showing only the Israel and Spain populations because they are the only ones to show evidence of significant population declines. Note that although it is plotted, the multinomial contraction model for the Israel wolves is not significant. While the predicted models are an improvement to the neutral model, deviations still exist from the observed SFS.

Table 20. Results of PRFREQ analysis of sequence data for wild canid populations for the Poisson calculation. All τ are given in number of generations from the present, and values of ω are given in terms of the current population size (i.e., $\omega = N_{eWOLF}/N_{ePOP}$). $1/\omega$, the size change parameter in terms of the ancestral population size (i.e. N_{ePOP}/N_{eWOLF}), is also reported. p-values are given for the comparisons in parentheses using the χ^2 distribution. Results for models with increasing degrees of freedom (i.e., B2b), when ones with fewer df (i.e., B1b) are not significant, are not shown.

Poisson Calculation					
Alaska Wolf					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	-1.5376	--	Constant Size
B1b	1	τ = 97.15995 ω = 19.0476 $1/\omega$ = 0.0525	-1.0411	0.6087 (B1 vs. B0)	Contraction
Israel Wolf					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	-17.322	--	Constant Size

B1b	1	τ = 10795.55 ω = 4 1/ω = 0.25	-11.126	0.00204 (B1b vs. B0)	Contraction
B2b	3	τ = 3454.576 τ_B = 315.230 ω_B = 0.01 ω = 0.0073	-10.813299	0.7314 (B2b vs. B1b)	Contraction, contraction
Spain Wolf					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	-19.871767	--	Constant Size
B1b	1	τ = 226.7066 ω = 36.3636 1/ω = 0.0275	-15.586696	0.0138 (B1b vs. B0)	Contraction
B2b	2	τ = 863.644 τ_B = 781.5978 ω_B = 0.0181 ω = 23.5	-15.553876	0.9559 (B2a vs. B1b)	Contraction, expansion (bottleneck)
Yellowstone Wolf					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	7.410006	--	Constant Size
B1b	1	τ = 16.193325 ω = 22.2222 1/ω = 0.045	7.42498	0.9851 (B1b vs. B0)	Contraction
Coyote					
Model	**df**	**Parameter**	**Log Likelihood**	**p-value**	**Description**
B0	0	None	16.7743	--	Constant Size
B1b	1	τ = 20511.545 ω = 0.61538 1/ω = 1.625	18.1135	0.2621 (B1b vs. B0)	Expansion

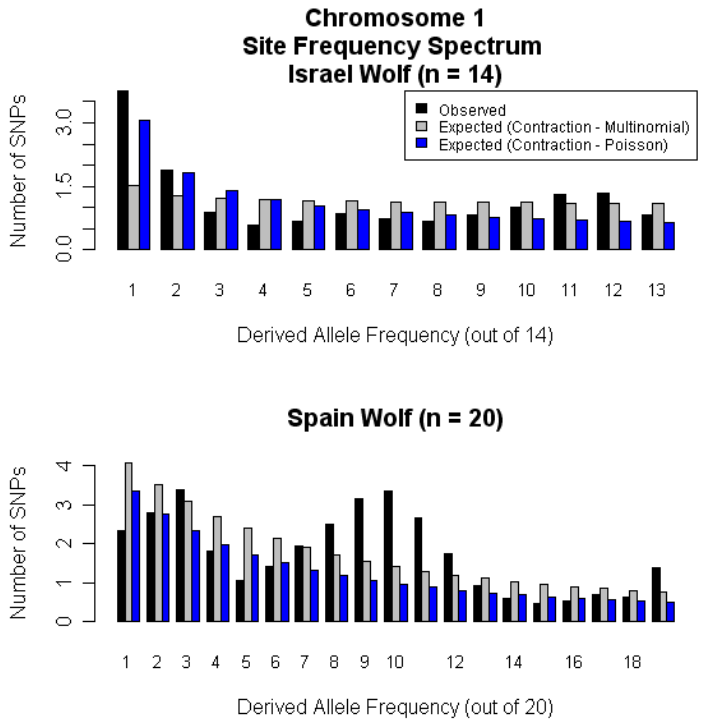

Figure 24. Site frequency spectra of data sampled one chromosome per individual as described in text for the Israel and Spain wolf populations. Black bars are the observed data, and gray and blue bars are the expectations under the contraction (B1b) models from the PRFREQ multinomial and Poisson calculations, respectively, as indicated in Table 19 and Table 20. x-axis is the derived allele frequency out of n as indicated, and the y-axis is the number of SNPs with derived allele at that frequency. Note only the Poisson contraction model is significant for the Israel wolf.

We also compare the composite likelihoods of the multinomial and Poisson calculations by rescaling the multinomial likelihood as for the domestication and breed inference. We use the multinomial parameter estimates in Table 19 to obtain the estimated θ under the multinomial; this θ is used in the Poisson likelihood calculation (Table 21). We do not see a significant difference in the composite likelihoods of the neutral (B0) models for the Spain population, as the multinomial estimate of θ (9.406) is very similar to the θ used in the Poisson calculations (9.741). However, for the Spain Blb calculation, the multinomial model is in fact significant, with an estimate of θ about twice as large as the Poisson θ. This may indicate that the ancestral wolf population is larger

than we assume with the Poisson. For the Israel wolf, there is a significant difference in likelihood for the neutral models, with the θ estimated to be approximately half of the value we estimate. Although the estimated value of θ is extremely large for the Israel B1b calculation, it is not significantly different from the Poisson calculation nor from the B0 model.

Table 21. Results of rescaling multinomial likelihoods for comparison between multinomial and Poisson calculations for the given models and wolf populations. p-value is obtained by taking twice the difference in log likelihood and using the χ^2 distribution with 1 df. A p-value < 0.05 indicates that the likelihood under the multinomial calculation is significantly greater than the likelihood under the Poisson calculation (indicated by asterisks). Note that the B1b Israel model is not shown because it is not significant in comparison to the B0 model.

Wolf Population	Model	θ (Estimated from Multinomial)	Poisson LL (Multinomial Parameters)	Poisson LL (Poisson Parameters)	p-value
Israel	B0	4.8272	-12.4731	-17.3222	0.0018 *
	B1b	1200.2	-10.9319	-11.1261	0.5331
Spain	B0	9.4063	-19.8511	-19.8718	0.8390
	B1b	19.3	-12.5960	-15.5867	0.0145 *

Interpretation

First, we discuss the differences between the multinomial and Poisson estimations, as values are not comparable between the two. For the Spain wolf, the multinomial calculation detects a 500-fold contraction 41 generations ago, compared to a 36-fold contraction 226 generations ago with the Poisson. For the Israel wolf, the Poisson detects an ancient population decline 10,000 generations ago, while a contraction is not significant in the multinomial.

As seen in Figure 24, the Poisson contraction estimates fit the observed SFS more appropriately for lower frequency SNPs, while the multinomial performs better for SNPs of high frequency. The fact that the multinomial calculation for the Spanish wolf maximizes the value of ω at 500 may indicate that it is overfitting to account for an increase in high frequency derived alleles. The discordance between the two estimates

may also indicate that the ancestral wolf effective population size of 21,600 may not be entirely accurate. Although we have limited data, it is still reasonable to conclude that there has been a significant contraction in the Spanish wolf population as well as possibly the Israel wolf population.

Interestingly, for the coyote population, maximum likelihood estimates of the multinomial and Poisson Blb calculations seem comparable, although they are not significant. Both calculations predict an expansion of approximately 1.6-fold, with a timing of 19,800 generations for the multinomial and 20,500 generations for the Poisson. The correspondence between the multinomial and Poisson calculations indicates that the value of θ used may be more appropriate for the coyote population than the individual wolf populations; a possible ancestral population size of the coyote is then approximately 21,000 individuals.

In our analyses, we detect a significant contraction in the Spain and possibly the Israel wolf populations. Since past research has been conducted regarding these wolf populations, we compare our results to the conclusions of these studies. In addition, we discuss the implications of a non-constant wolf population size on our previous demographic analyses.

Spain Wolf Population

As a result of the colonization of humans into their previous habitats and humans perceiving them as a threat to game and livestock, the gray wolf has undergone severe population declines over the past several hundred years (Wayne et al. 1992). From the 1940's to the 1970's, the population size of the Spanish wolf specifically has decreased (Ramírez et al. 2006). Although the use of poisoned baits for wolves was common in the

past, in 1970 the wolf was declared a legal game species and these poisoned baits became illegal in 1984 (Blanco et. al 1992). Since the 1990's, the Spanish wolf population size has slightly increased overall (Ramírez et al. 2006), and Blanco et al. (1992) conclude that the Spanish wolf population has increased in Northern Spain. Blanco et al. propose that the population has instead decreased in the South, where poisoned baits are likely still used illegally due to wolves' interference with red deer hunting. The authors extrapolate that roughly 500-750 wolves are killed each year in Spain (Blanco et al. 1992). According to our PRFREQ results of the Poisson B1b calculation, we perceive a significant population decline in the Spain wolf population of approximately 36-fold (Table 20). Although the timing of approximately 680 years ago may not be accurate, that we detect a severe population decline both in the multinomial and Poisson calculations agrees with these studies.

In examining the site frequency spectra of both the multinomial and Poisson models (Figure 24), we see that the SFS predicted from the contraction model does not seem to fit the observed data particularly well for SNPs of intermediate frequency. Given that Blanco et al. (1992) suspect different demographic forces to be acting on the northern and southern wolf populations, a likely cause of this deviation is population subdivision. Unfortunately, since the geographic area from which these wolves were obtained is unknown, this hypothesis cannot be directly assessed. In addition, we see a small increase of high frequency derived alleles that may be caused by multiple hits at given SNP sites or ancestral misspecification.

According to Blanco et al. (1992), the gray wolf population in Spain overall has been increasing due to protection efforts. The bottleneck model we predict is not

significant, although it does indicate a contraction followed by an expansion (Table 20). Since to detect a bottleneck one must to observe a sufficient number of new mutations that have arisen after the population contraction, we may not have enough power to detect this demography given limited data on only chromosome 1.

Israel Wolf Population

The Israel wolf population has also been previously studied. Again, as a result of human colonization, the wolf population has declined in Israel. However, harassing or killing wolves has become illegal in Israel, possibly minimizing the extent of the population contractions that would have occurred without such restrictions (Hefner and Geffen 1999). Regardless, in areas where humans coexist with wolves, humans are the largest cause of wolf mortality (Hefner and Geffen 1999). In a study of Arabian wolves in Israel, Hefner and Geffen conclude that these wolves have become habituated to living with humans. The Arabian wolf population has adjusted their foraging and pack habits, transitioning from feeding on livestock to scavenging, which has allowed them to live more harmoniously among humans (Hefner and Geffen 1999). Therefore, the gray wolf in Israel may not have endured a very drastic population decline due to these factors.

In our analysis, the multinomial inference does not detect a significant population decline, although the estimate is a 500-fold contraction (Table 19). For the Poisson calculation, a 4-fold population decline approximately 10,000 generations ago is significant. Since only the Poisson calculation finds significant evidence for a population decline, this indicates that a decline, if it has occurred, has had only mild effects on the shape of the site frequency spectrum in a deficit of rare alleles (Figure 23). The uncertainty of the value of θ used in the Poisson calculation should be taken into account

before concluding that the Israel wolf population has indeed gone through a severe population decline; however, a mild decline does seem to be plausible given the population's known recent history.

It is possible that an ancient contraction as estimated by the Poisson calculation has in fact occurred. When examining F_{ST}, a measure of population differentiation, between wild canid populations using the program DNAsp (Rozas et al. 2003), we find rather interesting results (Table 22). The Alaska, Yellowstone, and Spain populations have the lowest pairwise values of F_{ST} (between 0.02 and 0.19), indicating little population differentiation. However, pairwise F_{ST} of all wolf populations with the Israel population is slightly higher (between 0.3 and 0.39), indicating greater differentiation. Although we do not assess the significance of these values, that all other wolf populations have a higher F_{ST} with the Israel wolf population, even when compared to coyotes, indicates that the Israel wolf population could have diverged from an ancestral wolf population very long ago. Detecting a contraction approximately 10,000 years ago in our analysis may reflect this ancient divergence.

Table 22. Values of pairwise F_{ST} calculated between wolf and coyote populations.

F_{ST}	**Alaska Wolf**	**Coyote**	**Israel Wolf**	**Spain Wolf**	**Yellowstone Wolf**
Alaska Wolf	--	0.349227	0.385265	0.197512	0.026153
Coyote	--	--	0.602373	0.445591	0.344172
Israel Wolf	--	--	--	0.326050	0.305845
Spain Wolf	--	--	--	--	0.117090
Yellowstone Wolf	--	--	--	--	--

North American Wolf Populations

Old World wolf populations have historically been most threatened by humans; in contrast, wolf populations exist in very large numbers in the New World, such as in

Northern Canada and Alaska (Wayne et al. 1992). High diversity is also likely maintained in the North American wolf through interbreeding with coyote, as in Minnesota and Canada (Wayne et al. 1992). Compared to European populations, North American wolf populations exhibit much less subdivision, although genetic differentiation does exist (Wayne et al. 1992). However, there likely still have been slight declines in North American wolf populations, as Wayne et al. (1992) conclude through mitochondrial DNA analysis.

The North American gray wolf populations of Alaska and Yellowstone National Park both show no evidence of a significant population size change in either the multinomial or Poisson PRFREQ calculations. In addition, neither population has an observed SFS with a shape drastically different than under neutrality (Figure 23). This supports that wolves in North American have not undergone as severe population declines as those in Europe. It is not surprising that both the Alaska and Yellowstone wolves yield the same results, as the Yellowstone wolves are believed to be derived from Alaska. A very low pairwise F_{ST} for these two populations, indicating low genetic differentiation, supports this hypothesis (Table 22).

These wolf populations do not show evidence of a population history other than one of a constant population size. The effects of any declines that have indeed occurred in these gray wolf populations have likely been minimized by continued introgression with the coyote.

Coyote

In contrast to the wolf, human populations have not heavily persecuted the coyote, likely due to the coyote's smaller size and dependence on smaller prey (Wayne et al.

1992). As a result, the population size of the coyote has likely increased over most of North America (Wayne et al. 1992). Mitochondrial DNA analyses also suggest that the coyote has a population size an order of magnitude larger than that of gray wolves (Vilá 1999b). We do not detect a significant population expansion in the coyote, although our estimates suggest one. Given that we are only examining 11,279 bp of chromosome 1, we have limited power to detect significant population size changes other than those that have been very severe, as in the Spain wolf population. However, with additional data, we may be able to detect a significant expansion in the coyote with these methods.

As mentioned previously, the coyote parameter estimates of the Poisson calculation seem to agree with those of the multinomial calculation. This may indicate that an ancestral coyote population had an effective population size equal to the current effective population size of wolves. A population expansion of the coyote from this value may reflect that the coyote has a larger effective population size than wolves.

Implications on Demographic Inference

Lastly, we address the assumption made in our analysis of the domestication event of a constant wolf population size. In this analysis, all wolf populations were combined to estimate θ and the current wolf effective population size; these values were used to represent the ancestral wolf population. We have detected that wolves from Spain and possibly Israel have gone through population contractions, clearly violating the assumptions of our assumed demographic model. This calculation also ignored subdivision between the different wolf populations, seen through values of pairwise F_{ST} (Table 22). The incorrectness of these assumptions could affect the validity of the parameter estimates we obtain when analyzing dog domestication.

In order to minimize the effects of population size changes in wolves, as well as the effects of population subdivision, it is important to exclude the Israel and Spain wolf populations when calculating θ. The two remaining populations are the Alaska and Yellowstone wolves, which have not undergone population size changes and have a low pairwise F_{ST}. Excluding the Israel and Spain populations not only addresses the issue of changing population size but population subdivision as well.

We calculate a revised value of θ for wolves, including only the Alaska and Israel populations (Table 23). While the per-bp θ of all wolves, including all four populations, is 0.00086, the per-bp θ for only the Alaska and Yellowstone populations is 0.000904. This results in a new estimate of the wolf effective population size of approximately 22,600 compared to an original estimate of 21,600. The difference between these two values is not very large, indicating that having excluded the Israel and Spain populations from the domestication analysis would likely not have changed our conclusions. Nucleotide diversity (π) is also comparable between all wolves and the Alaska and Yellowstone populations, again indicating that using all four wolf populations, rather than only Alaska and Yellowstone, likely did not affect our results.

Table 23. Values of θ and π calculated for the indicated wolf populations from the sequence data on chromosome 1 (11,279 bp).

Wolf Population	**2n**	**Segregating Sites**	**θ (per site)**	**π (per site)**
All Wolves	144	54	0.000864	0.001176
Alaska	38	41	0.000865	0.001031
Israel	28	18	0.000410	0.000438
Spain	40	34	0.000709	0.001122
Yellowstone	38	49	0.001034	0.001196
Coyote	36	48	0.001026	0.000953
Alaska and Yellowstone	76	50	0.000904	0.001129

Wild Canid Conclusions

In conclusion, it appears that the Spain, and possibly Israel, wolf populations have undergone population declines. Neither the coyote nor the Alaska and Yellowstone gray wolf populations show significant evidence of any population size changes. The lack of significance for these three populations may in part be due to our limited power to detect demographic history when examining such a small portion of chromosome 1.

Having not excluded the Israel and Spain gray wolf populations in our calculations of the wolf effective population size does not appear to have had a large effect on our demographic analyses. This does not necessarily indicate that our assumption of a constant wolf population size is completely valid, however, as mitochondrial evidence does suggest that wolves have gone through a population decline (Vilá 1999b). In the future, incorporating a declining wolf population with subdivision into the demographic modeling could yield a more accurate representation of the domestication event as well as of individual breed formations.

VIII. Conclusions

We have used single nucleotide polymorphism data and information contained in the site frequency spectrum of the purebred dog to obtain a more complete picture of the demographic history of the domestic dog, *Canis lupus familiaris*. Using the Poisson Random Field approach in a composite likelihood framework, we attempt to infer the maximum likelihood estimates of parameters governing the domestication of the dog from the gray wolf, the formation of individual dog breeds, and the history of several wild canid populations.

In inference of the domestication event, we find evidence for a slight contraction during the domestication of dogs approximately 15,000 years ago, with no significant evidence of a different timing. The contraction perceived is likely an effect of analyzing breed dogs, which have gone through severe recent bottlenecks, rather than an effect of an actual contraction at domestication. More likely is that continued interbreeding between dogs and wolves or perhaps multiple domestication events have maintained the high levels of diversity seen in today's domestic dog. In future research of dog domestication, analysis of data from non-breed dogs will be essential to minimize the effects of breed demographic history and substructure. In addition, more complicated demographic models incorporating possibilities such as multiple domestication events, interbreeding between dogs and wolves, fluctuating population sizes in wolves, and subdivision among wolves, should be considered.

Results of inferring breed demographic histories appear to corroborate prior knowledge of the breeds studied. The strongest contraction is seen in the Bernese Mountain Dog, a relatively rare breed that is known to have maintained a small

population size throughout time. A very strong contraction is also seen for the Pekingese, a breed known to have originated from a very small number of founders. We see similar contraction severities for the Labrador and Golden Retrievers, the two most popular breeds of the American Kennel Club (AKC), which have likely been influenced by the introduction of other breeds into their breeding pools. Only a mild contraction is estimated for the Akita; although it is a relatively rare breed, the effects of a population contraction have likely been minimized through breed subdivision or its introgression with the gray wolf. An important area for future research is the joint inference of dog domestication and breed bottlenecks, rather than the independent inference of the two as performed in this study.

Finally, we examine the demographic history of several gray wolf populations. We find significant contractions in only the Spanish and possibly the Israel wolf populations, both of which are known to have undergone population declines. Our results also agree with studies indicating increased subdivision and smaller population sizes among European wolves compared to North American wolves (Wayne et al. 1992). Although our domestication analyses assume a constant wolf population size, we conclude that having excluded the Spain and Israel wolves in our calculation of the ancestral wolf effective population size would have had little effect. Nevertheless, in future studies of dog demography, incorporating changing wolf effective population sizes, as well as substructure within wolves, is an important addition.

While we have filled in several gaps in the history of the domestic dog through this study, future research on the same topic has the potential to be much more extensive. To yield more conclusive results, the methods and approaches presented in this study can

be implemented for more comprehensive data sets with greater numbers of dog breeds and larger numbers of SNPs. These results can be applied to the promising studies currently underway using the domestic dog as a model system to map complex traits. In obtaining more complete knowledge of its domestic history, what is already a powerful model organism has the potential to become an even more powerful one.

Acknowledgements

I would like to thank Dr. Carlos D. Bustamante for his ongoing support during my undergraduate research, as well as Dr. Adam Boyko for his endless guidance and assistance in this project. I also greatly appreciate the help of Dr. Nathan B. Sutter and Melissa M. Gray in providing the genetic data used in this study. I would also like to thank Dr. Robert K. Wayne and Dr. Elaine A. Ostrander. Finally, I would like to thank Dr. Andrew G. Clark, Dr. Jason G. Mezey, and Dr. Sutter for their time spent reviewing this thesis, as well as Dr. Arthur T. DeGaetano.

Literature Cited

American Kennel Club. 1997. The Complete Dog Book (19th Edition Revised). Howell Book House, New York.

Blanco, J.C., Reig, S., de la Cuesta, L. 1992. Distribution, status and conservation problems of the wolf *Canis lupus* in Spain. *Biological Conservation* 60, 73-80.

Björnerfeldt, S., Webster, M.T., Vilá, C. 2006. Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Research* 16: 990-994.

Bustamante, C.D., Wakeley, J., Sawyer, S., Hartl, D.L. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159:1779–1788.

Caicedo, A., Williamson, S., Hernandez, R.D., Boyko, A., Fledel-Alon, A., York, T., Polato, N., Olsen, K., Nielsen, R., McCouch, S.R., Bustamante, C.D., Purugganan M.D. 2007. Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. *PLoS Genetics* Sep 7;3(9):1745-56.

Chase, K., Lawler, D.F., Adler, F.R., Ostrander, E.A., and Lark, K.G. 2004. Bilaterally asymmetric effects of quantitative trait loci (QTLs): QTLs that affect laxity in the right versus left coxofemoral (hip) joints of the dog (*Canis familiaris*). *American Journal of Medical Genetics* 124**:** 239-247.

Chase, K., Sargan, D., Miller, K., Ostrander, E.A., Lark, K.G. 2006. Understanding the genetics of autoimmune disease: two loci that regulate late onset Addison's disease in Portuguese Water Dogs. *International Journal of Immunogenetics* 33:179–184.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., Nielsen, R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* 15: 1496-1502.

Dangerfield, S., and Howell, E. (eds).1971. The International Encyclopedia of Dogs. McGraw-Hill Book Company, New York.

Fogle, B. 2000. The New Encyclopedia of the Dog. Doring Kindersley, New York.

Gray, M.M., Bustamante, C.D., Granka, J.M., Sutter, N.B., Boyko, A., Zhu, L., Ostrander, E.A., Wayne, R.K. In prep. Linkage Disequilibrium and Demographic Modeling in Domestic and Wild Canid Populations.

Guyon, R., Lorentzen, T.D., Hitte, C., Kim, L., Cadieu, E., Parker, H.G., Quignon, P., Lowe, J.K., Renier, C., Gelfenbeyn, B., Vignaux, F., DeFrance, H.B., Gloux, S., Mahairas, G.G., André, C., Galibert, F., Ostrander, E.A. 2003. A 1 Mb resolution radiation hybrid map of the canine genome. *Proc. Natl. Acad. Sci. USA* 100, 5296–5301.

Hartl, D. L., Moriyama, E.M., Sawyer, S.A. 1994. Selection intensity for codon bias. *Genetics* 138**:**227-234

Hefner, R., and Geffen, E. 1999. Group size and home ranges of the Arabian Wolf (Canis lupus) in Southern Israel. *Journal of Mammalogy*, May; 80(2):611-619.

Hellenthal, G and M. Stephens. 2007. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics* Feb 15;23(4):520-1.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18:337-8.

Kingman, J.F.C. 1982. The coalescent. *Stochastic Process. Appl*. 13: 235–248.

Kingman, J.F.C. 2000. Origins of the coalescent 1974–1982. *Genetics* 156: 1461–1463.

Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M., Venter, J.C. 2003. The dog genome: survey sequencing and comparative analysis. *Science* 301, 1898–1903.

Lehman, N., Clarkson, P., Mech, L.D., Meier, T.J., Wayne, R.K. 2002. A study of the genetic relationships within and among wolf packs using DNA fingerprinting and mitochondrial DNA. *Behavioral Ecology and Sociobiology* 30:83-94.

Leonard, J.A., Wayne, R.K., Wheeler, J., Valadez, R., Guillen, S., and Vila, C. 2002. Ancient DNA evidence for Old World origin of New World dogs. *Science* 298**:** 1613-1616.

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

Mech, L.D., Seal, U.S. 1987. Premature reproductive activity in wild wolves. *Journal of Mammalogy*, 68, 871–873.

Mosher, D.S., Quignon, P., Bustamante, C.D., Sutter, N.B., Mellersh, C.S., Parker, H.G., Ostrander, E.A. 2007. A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs. *PLoS Genetics* 3(5):e79.

Nielsen, R., Hubisz, M.J., Clark, A.G. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168:2373–2382.

Nordborg, M. 2001. Coalescent theory. In: Balding, D., Bishop, M., Cannings, C. (eds) Handbook of statistical genetics. John Wiley & Sons, New York, 179-212.

Olsen, S.J. 1985. Origins of the Domestic Dog: The Fossil Record. University of Arizona Press, Tuscon, AZ.

Ostrander, E.A. and Wayne, R.K. 2005. The canine genome. *Genome Research* 15:1706-1716.

Ostrander, E. A., and Kruglyak, L. 2000. Unleashing the canine genome. *Genome Research* 10, 1271–1274.

Parker, H.G., Kim, L.V., Sutter, N.B., Carlson, S., Lorentzen, T.D., Malek, T.B., Johnson, G.S., DeFrance, H.B., Ostrander, E.A., and Kruglyak, L. 2004. Genetic structure of the purebred domestic dog. *Science* 304**:** 1160-1164.

Pennisi, E. 2007. The Geneticist's Best Friend. *Science* 317: 1668-1671.

Ramirez, O., Altet, L., Enseñat, C., Vilá, C., Sanchez, A., Ruiz, A. 2006. Genetic assessment of the Iberian wolf *Canis lupus signatus* captive breeding program. *Conservation Genetics* 7:861–878.

Randi, E. and Lucchini, V. 2002. Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. *Conservation Genetics* 3: 31–45.

Savolainen., P., Zhang, Y., Luo, J., Lundeberg, J., Leitner, T. 2002. Genetic Evidence for an East Asian Origin of Domestic Dogs, *Science* 298(5598):1610-3.

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X. and Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496-2497.

Sawyer, S. 1976. Branching diffusion processes in population genetics. *Advances in Applied Probability* Dec. 8(4): 659-689.

Sawyer, S. and Hartl, D.L. 1992. Population genetics of polymorphism and divergence. *Genetics* 132**:**1161-1176

Stephens, M., and P. Donnelly. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* 73:1162-1169.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* 68:978-989.

Sutter, N.B. and Ostrander, E.A. 2004. Dog star rising: The canine genetic system. *Nature Reviews Genetics* 5**:** 900-910.

Sutter, N.B., Eberle, M.A., Parker, H.G., Pullar, B.J., Kirkness, E.F., Kruglyak, L., Ostrander, E.A. 2004. Extensive and breed-specific linkage disequilibrium in Canis familiaris. *Genome Research* 14: 2388-2396.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-595.

Tsuda K., Kikkawa Y., Yonekawa H., Tanable Y. 1997. 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 and Genetic Systems*, 72, 229–238.

Vilá, C., Savolainen P., Maldonado J.E., Amorim I.R., Rice J.E., Honeycutt R.L., Crandall K.A., Lundeberg J., Wayne R.K. 1997. Multiple and ancient origins of the domestic dog. *Science* 276**:** 1687–1689.

Vilá, C., Maldonado J. E., Wayne R. K. 1999a. Phylogenetic relationships, evolutions, and genetic diversity of the domestic dog. *Journal of Heredity* 90: 71–77.

Vilá, C., Amorim, I.R., Leonard, J.A., Posada, D., Castroviejo, J., Petrucci-Fonseca, F., Crandall, K.A., Ellegren, H., Wayne, R.K. 1999b. Mitochondrial DNA phylogeography and population history of the grey wolf *Canis lupus. Molecular Ecology* 8: 2089–2103.

Vilá, C., Seddon, J., and Ellegren, H. 2005. Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends in Genetics* 21**:** 214-218.

Wakeley, J. 2007 (in press). *Coalescent Theory: An Introduction.* Roberts and Company Publishers, Greenwood Village, CO.

Watterson, G.A., 1975. On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology* 7**:**256-276.

Wayne, R.K.1993. Molecular evolution of the dog family. *Trends in Genetics* June; 9 (6): 218-224.

Wayne, R.K., Lehman, N., Allard, M.W., Honeycutt, R.L. 1992. Mitochondrial DNA variability of the gray wolf: genetic consequences of population decline and habitat fragmentation. *Conservation Biology* 6, 559–569.

Williamson, S.H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., Bustamante, C.D. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome, *Proc. Natl. Acad. Sci. USA* May 31; 102(22):7882-7.

Appendix

Appendix Table 1. Additional dog breeds genotyped, with sample size in number of individuals in parentheses.

Basset Hound (16)
Border Collie (25)
Briard (19)
Dachsund (65)
Doberman Pinscher (26)
Irish Wolfhound (20)
Mastiff (25)
Miniature Poodle (30)
Old English Sheepdog (15)
Pembroke Welsh Corgi (24)
Pomeranian (29)
Portugese Water Dog (31)
Saint Bernard (18)
Scottish Terrier (22)
Standard Poodle (60)
Toy Poodle (34)
Whippet (19)

Appendix Table 2. Sites of sequence data excluded in the analysis, giving the amplicon and position within the amplicon. The positions on the chromosome are those based on CanFam1.

Chromosome	**Amplicon_ Position within Amplicon**	**Position on Chromosome**
1	BLA31_734	50253721
2	None	None
3	BLD54_703	12277045
	BLD54_568	12277180
	BLD52_341	12327169
	BLD52_318	12327146
	BLD51_518	12348731
34	BLE24_403	24378033
37	BLB35_264	6774650

Appendix Table 3. Sites of genotype data excluded in the analysis, giving the amplicon and position within the amplicon. The positions on the chromosome are based on CanFam1.

Chromosome	Amplicon_ Position within Amplicon	Position on Chromosome
1	BLA11_499	48808250
	BLA12_291	46823601
	BLA23_315	48524388
	BLA23_527	48524600
	BLA32_524	50272391
	BLA33_184	50282636
	BLA51_603	51983449
2	BLC21_264	44844572
	BLC32_436	45765319
	BLC53_471	48390869
3	BLD53_30	12300813
	BLD43_121	13765384
	BLD11_429	17150625
34	BLE15_520	22632069
	BLE14_332	22695198
	BLE24_403	24378033
	BLE41_254	26011155
	BLE54_339	27681551
	BLE51_323	27712677
	BLE52_449	27726866
37	BLB21_720	5828538
	BLB21_822	5828640
	BLB22_440	5844650

Appendix Table 4. Sites of genotype data excluded due to low sample size, giving the amplicon and position within the amplicon. Positions on each chromosome are based on CanFam1.

Chromosome	Amplicon_ Position within Amplicon	Position on Chromosome	n
1	BLA11_742	46808493	170
3	BLD12_662	17105503	567
37	BLB44_394	7496699	570

Appendix Table 5. Command line arguments for each chromosome for msHOT for the domestication event is shown below, where nsam equals the number of chromosomes of the sample, and nreps = the number of iterations, or 2000. The –t flag indicates θ, the –r flag indicates ρ and the number of sites (the sum of the lengths of all amplicons), and the –v flag indicates the number and locations of the hotspots, along with the intensity of each hotspot.

Chromosome 1

```
msHOT  nsam nreps -t 5.300183 -r 5.300183 6137 -v 8 636 637 14923 1433
1434 1665549 2151 2152 71006 2852 2853 825936 3432 3433 31271 4088 4089
29553 4744 4745 2531831 5457 5458 37259
```

Chromosome 2

```
msHOT  nsam nreps -t 3.814716 -r 3.814716 4417 -v 5 750 751 33058 1475
1476 1563503 2259 2260 904809 2984 2985 13208 3704 3705 2624795
```

Chromosome 3

```
msHOT  nsam nreps -t 7.214882 -r 7.214882 8354 -v 10 765 766 23085 1478
1479 47308 2263 2264 1453105 2987 2988 771296 3690 3691 4784 4443 4444
84702 5274 5275 667708 6047 6048 1747490 6867 6868 22154 7653 7654
44569
```

Chromosome 34

```
msHOT  nsam nreps -t 7.201064 -r 7.201064 8338 -v 10 748 749 36356 1524
1525 25437 2365 2366 2531417 3130 3131 14558 3843 3844 4963 4556 4557
762065 5290 5291 9315 6012 6013 1628817 6809 6810 45131 7578 7579 7106
```

Chromosome 37

```
msHOT  nsam nreps -t 8.473211 -r 8.473211 9811 -v 12 784 785 4731 1494
1495 1781587 2240 2241 15646 3077 3078 899107 3795 3796 5571 4547 4548
23191 5258 5259 5894 5990 5991 643463 6698 6699 11234 7546 7547 58329
8332 8333 1540919 9163 9164 8591
```

Appendix Table 6. Command line arguments for each chromosome for msHOT for breed formation inference is shown below, where nsam equals the number of chromosomes of the sample, and nreps = the number of iterations, or 2000. The –t flag indicates θ, the –r flag indicates ρ and the number of sites (the sum of the lengths of all amplicons), and the –v flag indicates the number and locations of the hotspots, along with the intensity of each hotspot.

Chromosome 1

```
msHOT  nsam nreps -t 9.74104 -r 9.74104 11279 -v 17 636 637 33305 1349
1350 1681542 2049 2050 36733 2750 2751 688857 3310 3311 20897 3815 3816
56584 4425 4426 57923 5005 5006 31271 5661 5662 29553 6317 6318 17103
6949 6950 61777 7539 7540 3042 8131 8132 54151 8795 8796 842 9378 9379
661838 9886 9887 1729509 10599 10600 37259
```

Chromosome 2

```
msHOT  nsam nreps -t 5.126591 -r 5.126591 5936 -v 7 750 751 33058 1475
1476 1563503 2259 2260 904809 2984 2985 13208 3704 3705 882965 4478
4479 1699344 5223 5224 40967
```

Chromosome 3

```
msHOT  nsam nreps -t 9.706495 -r 9.706495 11239 -v 14 765 766 23085
1478 1479 25923 2282 2283 15123 2985 2986 4755 3770 3771 1416265 4404
4405 36206 5128 5129 771296 5831 5832 4784 6584 6585 82628 7328 7329
1330 8159 8160 667708 8932 8933 1747490 9752 9753 22154 10538 10539
44569
```

Chromosome 34

```
msHOT  nsam nreps -t 8.692577 -r 8.692577 10665 -v 13 748 749 36356
1524 1525 25437 2365 2366 1681923 3169 3170 848690 3934 3935 14558 4647
4648 4963 5360 5361 762065 6094 6095 39315 6816 6817 28433 7599 7600
1599601 8396 8397 31068 9136 9137 13323 9905 9906 7106
```

Chromosome 37

```
msHOT  nsam nreps -t 9.121808 -r 9.121808 10562 -v 13 751 752 42072
1535 1536 4731 2245 2246 1781587 2991 2992 15646 3828 3829 899107 4546
4547 5571 5298 5299 23191 6009 6010 5894 6741 6742 643463 7449 7450
11234 8297 8298 58329 9083 9084 1540919 9914 9915 8591
```