Estimation of Population Divergence Times from Non-Overlapping Genomic Sequences: Examples from Dogs and Wolves

Pontus Skoglund,* Anders Götherström, and Mattias Jakobsson

Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden

***Corresponding author:** E-mail: pontus.skoglund@ebc.uu.se.

Associate editor: Rasmus Nielsen

Abstract

Despite recent technological advances in DNA sequencing, incomplete coverage remains to be an issue in population genomics, in particular for studies that include ancient samples. Here, we describe an approach to estimate population divergence times for non-overlapping sequence data that is based on probabilities of different genealogical topologies under a structured coalescent model. We show that the approach can be adapted to accommodate common problems such as sequencing errors and postmortem nucleotide misincorporations, and we use simulations to investigate biases involved with estimating genealogical topologies from empirical data. The approach relies on three reference genomes and should be particularly useful for future analysis of genomic data that comprise of nonoverlapping sets of sequences, potentially from different points in time. We applied the method to shotgun sequence data from an ancient wolf together with extant dogs and wolves and found striking resemblance to previously described fine-scale population structure among dog breeds. When comparing modern dogs to four geographically distinct wolves, we find that the divergence time between dogs and an Indian wolf is smallest, followed by the divergence times to a Chinese wolf and a Spanish wolf, and a relatively long divergence time to an Alaskan wolf, suggesting that the origin of modern dogs is somewhere in Eurasia, potentially southern Asia. We find that less than two-thirds of all loci in the boxer and poodle genomes are more similar to each other than to a modern gray wolf and that—assuming complete isolation without gene flow—the divergence time between gray wolves and modern European dogs extends to 3,500 generations before the present, corresponding to approximately 10,000 years ago (95% confidence interval [CI]: 9,000–13,000). We explicitly study the effect of gene flow between dogs and wolves on our estimates and show that a low rate of gene flow is compatible with an even earlier domestication date ~30,000 years ago (95% CI: 15,000–90,000). This observation is in agreement with recent archaeological findings and indicates that human behavior necessary for domestication of wild animals could have appeared much earlier than the development of agriculture.

Key words: population divergence, demographic inference, domestic dogs, shotgun sequencing, ancient DNA.

Research article

Introduction

Methods for reconstructing the demographic history and divergence of relatively undifferentiated populations have attracted great interest in the last few decades, but these methods are presently facing a number of challenges ranging from sheer computational problems to issues with data quality and magnitude (Nielsen and Beaumont 2009). Although recent advances in sequencing technology have enabled retrieval of genomic sequences from several extinct mammals (e.g., Poinar et al. 2006; Noonan et al. 2006; Blow et al. 2008; Miller et al. 2008), rigid population genetic analyses of multiple loci have so far been restricted to Neanderthals (Noonan et al. 2006; Green et al. 2006; Wall and Kim 2007; Green et al. 2010) and a single modern human (Rasmussen et al. 2010), and direct genomic analysis of extinct populations has yet to fulfill its full promise to shed light on past population processes (Willerslev and Cooper 2005; Millar et al. 2008; Green et al. 2009).

A major problem for ancient DNA (aDNA) studies stems from the fact that the most powerful approaches are based on direct shotgun sequencing (Millar et al. 2008, but see also Burbano et al. 2010). In most applications of direct ancient genomic sequencing, endogenous sequences will be outnumbered by microbial DNA, making it easier to generate large amounts of data compared with targeted methods, but difficult to achieve high coverage assemblies (Millar et al. 2008), resulting in little overlap between loci obtained from different individuals (e.g., Green et al. 2010). Population genetic analyses of ancient genome data can use reference genomes (Green et al. 2010) and single-nucleotide polymorphism (SNP) information (Rasmussen et al. 2010), but multiple reference genomes and public SNP databases currently only exist for a few vertebrates. As more genomic reference data becomes available, large-scale studies will become available also for non-model organisms, and we need clear paradigmatical approaches for analyzing paleogenomic data that 1) allow statistical testing of explicit demographic models (e.g., Nielsen and Beaumont 2009), 2) are resilient to contamination problems (Gilbert et al. 2005; Wall and Kim 2007; Green et al. 2009), 3) can handle postmortem nucleotide misincorporations (Pääbo 1989; Briggs et al. 2007;

Brotherton et al. 2007; Axelsson et al. 2008) and sequencing errors (Johnson and Slatkin 2008; Jiang et al. 2009; Lynch 2009; Liu et al. 2010), and 4) allow comparison between data from different points in time (Depaulis et al. 2009).

We describe a method which utilizes reference genomes for comparing multiple loci that do not overlap among samples and allows maximum likelihood estimation of population divergence times. To illustrate the method, we analyzed available genomic shotgun sequences from modern dogs (*Canis lupus familiaris*) and gray wolves (*Canis lupus lupus*) together with an ancient wolf to investigate the origin of dog domestication. Although it is clear that the closest living relatives of domestic dogs are Eurasian gray wolves (Olsen 1985; Clutton-Brock 1987; 1995; Vilà et al. 1997, 1999, Leonard et al. 2002), the timing and process of domestication remain contentious (e.g., Morey 2006; Boyko et al. 2009; Pang et al. 2009; vonHoldt et al. 2010). Initial analyses of mtDNA yielded estimates of a first domestication more than 100 thousand years ago (Ka) (Vilà et al. 1997), but recent phylogeographic mtDNA studies claim a date less than 16,300 years ago (Savolainen et al. 2002, Pang et al. 2009; but see also Boyko et al. 2009). Population genetic analyses using multiple autosomal loci have also produced disparate estimates, corresponding to 10–27 Ka depending on assumptions about generation time (Lindblad-Toh et al. 2005; Gray et al. 2009) and has pinpointed the Middle East as the most likely region of origin (vonHoldt et al. 2010). The archaeological record of dogs has pointed to a more recent domestication time based on fossils and burials ~12,000–14,000 years old (Davis and Valla 1978; Nobis 1979; Musil 1984; Olsen 1985; Benecke 1987; Sablin and Khlopachev 2002; Morey 2006), but a recent study documents dog-like morphological features in European canid fossils as old as 31,000 years (Germonpré et al. 2009).

Using our approach, we estimate that dogs and wolves first diverged more than 10 Ka, and that postdivergence gene flow is compatible with an even more ancient divergence date. We investigate possible biases arising from sequencing error, nucleotide misincorporations, sequence alignment, short fragment length, time-structured sampling, and incomplete genealogy information and show that our approach could be particularly useful for analyzing multilocus aDNA sequence data.

Materials and Methods

Genealogical Inference of Divergence

To estimate population divergence times in a four-way alignment between a sample (canid) sequence, two modern reference genomes (boxer and poodle, see below) and an outgroup, we first infer the genealogical topology by recording each position where two of the canid sequences has a derived allele and the third retains the ancestral variant found in the outgroup using a simple parsimony method that assumes an infinite sites model of mutation (fig. 1). If two positions in the same alignment results in different topologies, for example due to recombination, the locus is discarded.

We estimate the internode divergence time T (measured in units of N_e generations, where N_e is the effective population size in chromosomes) between a sample (s) and a corresponding population ancestral to two differentiated populations ("p" for poodles and "b" for boxers) using probabilities of concordant or discordant topologies under a coalescent divergence model (Hudson 1983; Takahata 1989; Rosenberg 2002). In this model, there are three possible topologies describing the genealogy of the three lineages, of which we denote the $(s,(b,p))$ case concordant, and the remaining two, $(b,(p,s))$ and $(p,(b,s))$, discordant. The probability of each discordant topology equals the probability that p and b do not coalesce during the time spent in the internode population (e^{-T}), where they are unable to coalesce with s, multiplied by the probability that s and p coalesce in the ancestral population (one-third). The probability of the remaining concordant topology equals $1 - 2e^{-T}/3$. We used the log-likelihood function of the internode divergence time T to compute the maximum likelihood estimate (MLE) of T and to obtain confidence intervals (CIs, Wakeley 2008)

$$\log(L(T)) = G_c \times \log\left(1 - \frac{2}{3}e^{-T}\right) + G_d \times \log\left(\frac{2}{3}e^{-T}\right),$$

where G_c is the number of concordant topologies and G_d is the total number of discordant topologies. To plot log-likelihood functions on the same scale, we subtract the maximum value of $\log(L(T))$ from all values to obtain a relative log-likelihood function.

Analysis of Pleistocene Siberian Wolf Sequences

Blow et al. (2008) sequenced (Illumina and 454 GS20) aDNA from an ancient wolf from the Altai region in Asia (Derevianko et al. 2003), dated by thermoluminescence to between 40 and 50 Ka). Both processed and raw data sets from Blow et al. (2008) were kindly provided by the authors. We initially aligned the processed Altai wolf sequences to the boxer (Canfam2.0; Lindblad-Toh et al. 2005), poodle (Kirkness et al. 2003), and cat (catChrV17e; Pontius et al. 2007) genome assemblies. However, a bias in shared allelic states with the boxer over the poodle was identified (the Altai wolf shared a boxer-specific allele at 181 positions but shared the poodle allele at only 1 position). Because this effect was likely due to conservatively stringent alignment criteria to the boxer genome in the pipeline for identification of authentic canid sequences implemented by Blow et al. (2008), we performed an independent analysis of the 2,745,862 raw sequence reads. We used megablast (Zhang et al. 2000) to identify canid aDNA sequences in the raw library by aligning sequences to the three genomes using word size 16. Sequences with bit score >35 and expect value <0.001 were extracted, in total 97,319 reads. These reads were then aligned to the entire NCBI *ref_seq* and *other_genomic* databases, leaving 93.7% (91,178 reads, mean length 40 bp, range 37–118 bp) of the originally identified reads with the best hit to a carnivore genome. This amounts to 3.45% canid reads, compared with the 2.16% (57,028 reads) identified by Blow et al. (2008). Comparison

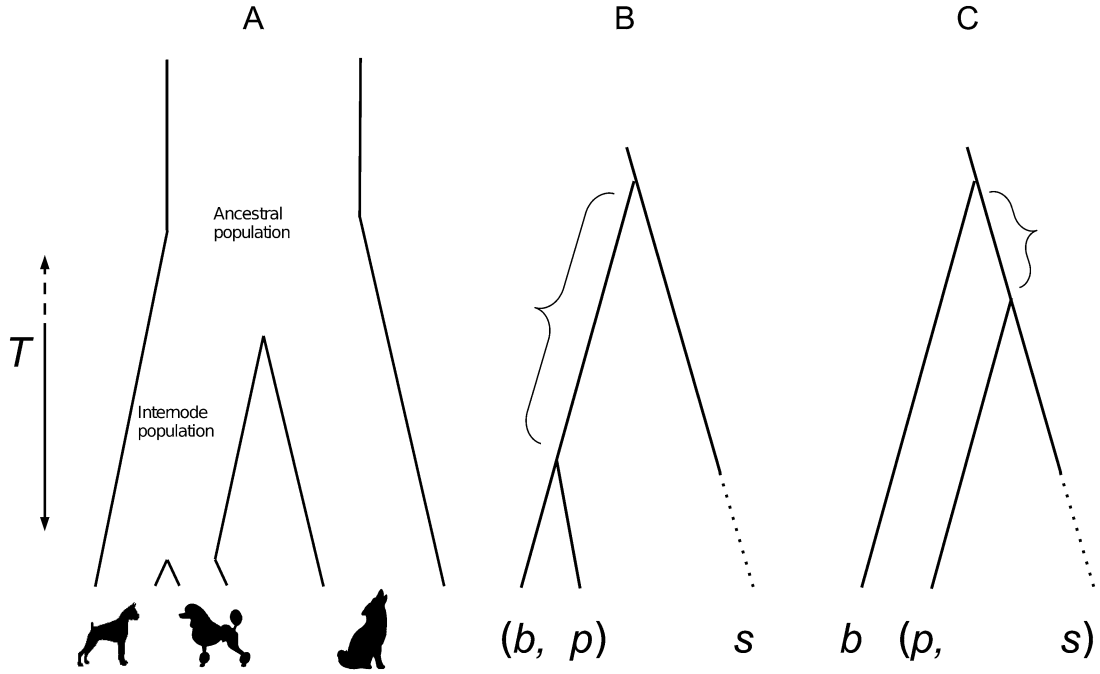

FIG. 1. Illustration of the model used for estimating divergence times from counts of different gene genealogy topologies. (*A*) A model of dog demographic history with the internode population and the ancestral population indicated. (*B*) The concordant genealogical topology with the internal branch indicated. (*C*) One of two possible discordant topologies with the internal branch indicated.

of these sequences with the boxer reference genome revealed patterns of molecular degradation characteristic for aDNA (fig. 2; Briggs et al. 2007). Additionally, the bias of sharing alleles with the boxer more often than the poodle was now similar to that in modern wolf sequences of similar fragment length (133 and 35, respectively, see below).

Alignment of Modern and Ancient Data to Carnivore Genomes

To identify homologous regions, identified endogenous sequence reads from the Altai wolf were realigned to the boxer, cat, and poodle genome assemblies. Megablast searches used the option "word size" set to 12, "culling limit" set to 20 and "e-value" = 0.001. In addition, trace reads from an Alaskan gray wolf (n = 21,687), Chinese gray wolf (n = 23,410), Indian gray wolf (n = 22,539), Spanish gray wolf (n = 22,116), German shepherd (n = 99,981), English shepherd (n = 99,373), Labrador retriever (n = 99,698), Beagle (n = 99,648), Alaskan malamute (n = 99,829), Rottweiler (n = 99,983), Bedlington terrier (n = 98,208), Portuguese water dog (n = 98,112), and Italian greyhound (n = 98,320), originally published by Lindblad-Toh et al. (2005), were downloaded from the NCBI trace archive and aligned in a similar way as the Altai wolf. Because the majority of these sequences were >800 bp, word size was set to 40 and 24 for searches of wolf sequences against the dog and cat genomes, respectively. Matches were required to align over at least 30 bases and hits to the two dog genomes had to cover >50% of the read length to be retained. The best alignment to each genome was identified as the hit which had the highest bit score and in case of several hits with equal score, the longest alignment was considered the best. Multiple alignment of a canid sequence

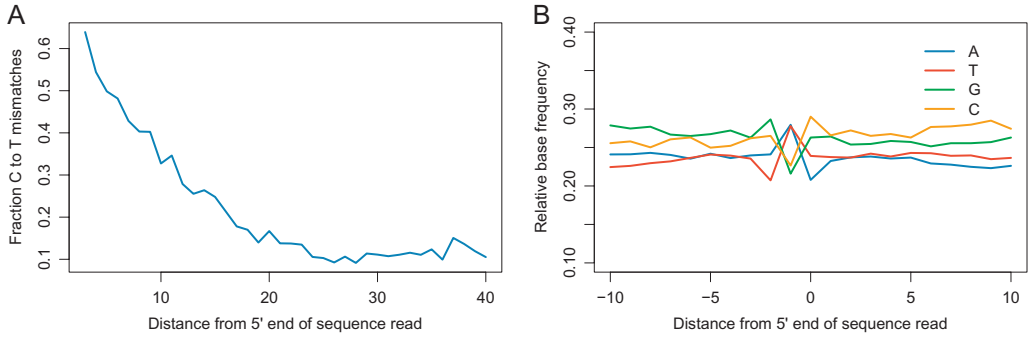

FIG. 2. Signs of molecular degradation in the Altai wolf sequences. (*A*) Increased frequency of C → T mismatches compared with other mismatches toward the 5′ end of the sequence read (*B*) biased base composition in the reference sequence at the 5′ end of the aligned sequence read.

and its best aligned sequence in each target genome were created using MUSCLE v3.7 with default parameters (Edgar 2004). In all downstream evolutionary analyses, we excluded indels due to uncertainties about the mutation rate and positions where one of the sequences had a low-quality base (Lindblad-Toh et al. 2005). In the topological analysis and estimates of pairwise mismatches involving the ancient wolf, we excluded C → T and G → A mismatches between sample sequence and reference genome because these mismatches are most likely due to postmortem nucleotide misincorporations.

Estimation of Sequencing and Alignment Error Rates

Using a rather distantly related outgroup genome, such as the cat, to infer the ancestral allelic state introduces a potential source of bias in that sequencing or alignment errors in the shotgun sequences not only inflates the length of the external branch but may also occur on a site that differs between canids and cats, changing the state in the canid to the ancestral variant found in the cat genome. These types of errors could result in the topology of a locus to be incorrectly inferred as "concordant" or "discordant" using our parsimony method. We therefore estimated the rate of sequencing and alignment errors present in the modern shotgun sequences using an approach that is based on assuming a constant mutation rate and implemented a correction to our inference method. Specifically, we used the observed differences between the boxer and cat genomes, both relatively high-quality genomes, to set a baseline for expected canid-felid pairwise differences to which the modern shotgun sequences can be compared (e.g., Burgess and Yang 2008). We considered the exact same positions as those used for the genealogical inference procedure. For example, in the alignments constructed as described above between Chinese wolf, cat, boxer, and poodle sequences, the boxer and cat sequences differ at 14.86% ($D_{c-b} = 0.1486$; 95% CI from 10,000 bootstrap replicates: 0.1475–0.1497) of the positions. Due to the high quality of the genomes, we consider this level of differences to be the "true" level of differences between canids and cat, unaffected by errors. However, for the same positions, the Chinese wolf (the sample) differs from the cat at 15.70% ($D_{c-w} = 0.1570$; CI: 0.1560–0.1582), and the additional fraction (0.84%) differing sites between the cat and the sample (B) can be caused by alignment or sequencing errors due to the lower quality of the single-pass sequences of the Chinese wolf. Errors that occur in positions where the sample (the Chinese wolf in this example) and the cat have the same true variant (fraction $1 - D_{c-b}$) will always be visible and increase the observed difference. Those errors that occur on true polymorphic sites (between sample and cat) can change the sample variant to another variant (not the cat variant; happens in two-thirds of the cases), which does not change the number of polymorphisms, or the error can change the sample variant to the same variant as the cat (happens in one-third of the cases), decreasing the observed difference between the cat and the sample. Assuming that the number of errors in the boxer and the cat reference genomes are negligible and that errors only hit a site once, an expression for the contribution by the error rate E to the additional fraction of differing sites between the cat and the sample B is

$$B = E \times (1 - D_{c-b}) - \frac{E \times D_{c-b}}{3}$$

rearranging and solving for E gives

$$E = \frac{3B}{3 - 4D_{c-b}}.$$

Correcting Genealogical Topologies for Sequencing and Alignment Error

Using the estimates of the error rate, we computed the fraction of sites f in the sample that erroneously display an ancestral (cat) allele, due to sequencing or alignment errors, in three-way alignments of the sample, the boxer and the cat. These events only occur at sites where the boxer and cat differ (D_{c-b}) and must confer a change from the boxer variant to the particular variant present in cat (happens in one-third of the cases), in total

$$f = E \times D_{c-b} \times \frac{1}{3}.$$

We used the high-quality boxer sequence to compute f but note that the lower-quality poodle data could also have been used for this analysis. Next, we used the average length of alignments in base pairs (L) to compute the expected occurrence $F = (f \times L)$ of such sites in an alignment. It is possible to have $F \geq 1$ for long alignments, but because our empirical data only consisted of relatively short alignments, which typically had one or very few variable sites, we can assume that $F < 1$ for each alignment and treat F as a probability that a locus displays a concordant site due to an error.

We computed the number of loci that erroneously display a concordant topology assuming that we can estimate this proportion from (1) the observed number of concordant loci N_{CO}, which is composed of both true concordant loci (N_{CT}) and loci without true informative sites that have been assigned as concordant due to the concordant sites arising from errors ($N_{NT} \times F$), $N_{CO} = N_{CT} + N_{NT} \times F$, and (2) the observed number of noninformative loci N_{NO}, from which a fraction has been erroneously moved to the concordant category, $N_{NO} = N_{NT} \times (1 - F)$. Using these two relationships, we obtained a corrected number of concordant loci N_{CT} as

$$N_{CT} = N_{CO} - \frac{N_{NO}}{1 - F} \times F.$$

We also applied a correction to the number of observed discordant topologies; because if a sequencing error converts a derived variant to an ancestral variant, it might also cause the appearance of a concordant site in loci that are discordant. If the locus contains sites supporting the true topology (i.e., discordant), this type of error would cause the locus to be discarded due to violating the assumption that sequence regions are free from recombination. The number of observed

discordant loci (N_{DO}) should therefore equal the number of true discordant loci (N_{DT}) that do not contain sites that erroneously display a concordant topology, $N_{DO} = N_{DT} \times (1 - F)$. The true number of discordant loci would then be computed as

$$N_{DT} = \frac{N_{DO}}{1 - F}.$$

In the above analysis, we focused on the observation of erroneous ancestral alleles and not the observation of erroneous derived alleles because erroneous derived alleles only influence our analysis if they appear at sites where the two canids already have different alleles. Because the fraction of polymorphic sites between two canids is approximately 0.1% (Lindblad-Toh et al. 2005) and the fraction of polymorphic sites between a cat and a canid is approximately 15% (see below), the potential error due to erroneous derived alleles is negligible compared with the effect of erroneous ancestral alleles.

Average Sequence TMRCA between Canids

To compare our method for population divergence time estimation with a different approach, we computed average sequence coalescence time (average time to most recent common ancestor [TMRCA]) between the boxer and our sample sequences using the same positions as above. We used a method where the error-prone polymorphisms specific to low-coverage shotgun sequences can be excluded by computing the average number of mutations on the fraction of the branch between the boxer and cat that postdate the TMRCA of the boxer and the sample sequence (Noonan et al. 2006; Green et al. 2006, 2010; Prüfer et al. 2010). We computed a statistic S_s which we define as the number of sites where the sample sequence "s" shares an (ancestral) allele with the cat but the boxer has a different (derived) allele, divided by the total number of investigated positions. In a standard neutral coalescent model without population structure, this value would correspond to the population mutation rate $\theta = 4N_e\mu$, such that $S_s = \theta/2$, because only mutations on one of the branches in the genealogy of the boxer and the sample are considered. In the case of population divergence, S_s is equal to $(\theta + T)/2$, where T is the divergence time between populations.

To correct for the effect of sequencing and alignment errors when using an outgroup as distant as the cat (see above and Prüfer et al. 2010), we used the relatively high-quality (but lower coverage than the boxer) poodle sequences. We based the error correction on the assumption that the poodle sequences and sequences from other modern European dog breeds should have approximately the same value of S_s and that any remainder can be attributed to errors (note that this procedure is different from the error correction applied to the genealogical approach above). First, we computed the fraction of differences between the boxer and the cat in which the poodle carries the cat allele, $S_p = 0.000668$. The average for other modern dog breed sequences was $S_d = 0.001270$ (excluding the Alaskan Malamute), and assuming that all low-coverage Sanger shotgun sequences from Lindblad-Toh et al. (2005) had identical inflation in sequencing error, we quantified the inflation due to errors as $I_D = S_d - S_p = 0.000601$. We then estimated the average value for wolves ($S_w = 0.001724$) and subtracted the estimated inflation due to sequencing errors (I_D), resulting in a corrected $S_w = 0.001122$. By normalizing this estimate of S_w with the corrected average for European dogs (S_w/S_d), we obtain an estimate of the relative average TMRCA between a modern dog chromosome and a modern wolf chromosome that is independent of assumptions about the mutation rate, other than the mutation rate being equal for all lineages. Because the Altai wolf sequences is from a different historical time and sequenced with a different technology (short sequences), we computed S_{Altai}/S_d without error correction.

Bootstrap Analysis

We obtained 95% bootstrap CIs for the fraction of concordant topologies, S_s, S_w/S_d and E from 10,000 pseudoreplicates over loci. These CIs reflect the amount of sampling error involved in each estimate. Note that the uncertainty for the estimates of the error rate E (the CIs are narrow, see below) is not carried through to the other estimates which involve E and therefore represents a potential additional source of uncertainty.

Simulations

We generated gene genealogies using Serial Sim-Coal (Excoffier et al. 2000; Anderson et al. 2005) and placed mutations on the genealogies with probability proportional to branch length using a custom program (available upon request). For 100,000 simulated genealogies, we sampled the number of segregating sites from a Poisson distribution with mean equal to the total genealogy length (Hudson 1990), assuming a 500-bp locus with mutation rate of 10^{-8} per bp and generation (reasonable for dogs, see, e.g., Lindblad-Toh et al. 2005). For each particular demographic model, we calculated the proportions of concordant and discordant genealogies from informative segregating sites and relative average TMRCA (S_w/S_d) between dogs and wolves with analogous procedures as for the empirical data. For all models, we assume breed creation 70 generations ago with no migration between breeds (Wayne and Ostrander 2007), in which case N_e of breeds has no affect on our analysis. Based on previous estimates, we set the ancestral dog effective population size to 13,000 chromosomes (Lindblad-Toh et al. 2005), the wolf effective population size to 45,000 chromosomes (Gray et al. 2009), and generation time to 3 years (Lindblad-Toh et al. 2005; Gray et al. 2009). We investigated the effect of very low ($N_em = 0.25$) and low ($N_em = 0.5$) symmetric migration rates between dogs and wolves. For testing scenarios involving the Altai wolf, we used 500,000 replicates and a locus mutation rate of 3.5×10^{-7} (corresponding to a 35-bp sequence with a mutation rate of 10^{-8} per bp and generation), an age of 13,000 generations for the Altai wolf, a dog-wolf divergence time of 5,000 generations ago without gene flow, and assumptions about N_e in dogs and wolves as above.

Results

Investigating the Effect of Incomplete Genealogy Information

To estimate the divergence time T between the ancestral population of the boxer and poodle and the population leading to a particular canid sample of interest, we used a coalescent-based approach based on the number of gene genealogies that show a discordant topology and the number of gene genealogies that show a concordant topology, when compared with the population topology (Takahata 1989; Rosenberg 2002; Wakeley 2008). The number of concordant and the number of discordant topologies allow us to compute the likelihood for different divergence times (scaled by generation time and effective population size N_e) assuming no migration between populations and assuming that loci are independent of each other (see Materials and Methods).

This framework has the inherent assumption that the true genealogy is always known for any given locus (Wakeley 2008). However, in practice, the method requires that the genealogical topology is inferred using genetic markers, a step which can potentially give rise to biased results (Nielsen 1998; Yang 2002). In our implementation, we assume an infinite sites mutation model and reconstruct a genealogical topology only when sufficient information is available from SNPs in the sequence alignments. For a mutation to be informative about the genealogy in this parsimony framework, it must occur on an internal branch of the genealogy (excluding the internal branch leading to the outgroup). This confers a possible bias if one category of topology is more likely to contain sufficient information for inferring the genealogy. Specifically, if one category of topology has, on average, genealogies with longer internal branches, the genealogies in this category would more often contain informative sites than a topology category where genealogies have short internal branches. In our divergence model, the concordant topology can arise when lineages sampled from the two more closely related populations coalesce either in the ancestral population of these two populations (the "internode" population) or in the ancestral population of all three populations (fig. 1). We can think of the genealogy as consisting of two parts: the part before the final ancestral population and the part where all remaining lineages are in the final ancestral population. The portion of genealogies that have a concordant topology that have arisen through a coalescent event prior to reaching the final ancestral population have a longer internal branch on which informative genealogies can arise than concordant or discordant topologies that are formed when all three lineages reach the ancestral population. This results in a higher probability of detecting concordant topologies. The magnitude of this bias will depend on the effective sizes of the internode and ancestral populations because a reduction in effective size in the ancestral population leads to a higher probability of rapid coalescence, resulting in a smaller portion of the total branch length due to lineages in the ancestral population.

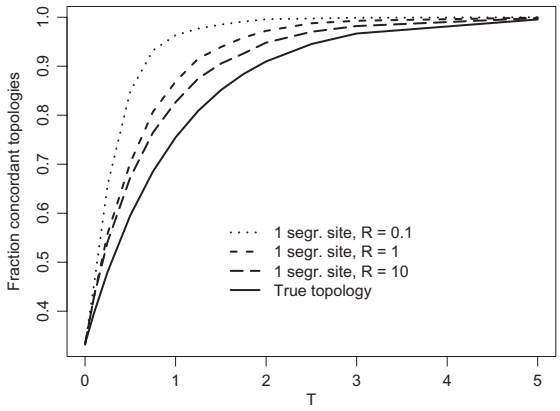

FIG. 3. The bias in the observed fraction of concordant genealogies inferred from SNPs depends on the effective sizes of ancestral and internode populations. The fraction of concordant topologies as a function of divergence time (T) given by directly counting the topologies from simulations is shown as a solid line and the fraction of concordant topologies obtained when inferring the topology from a single segregating site are shown as broken lines. Results from three models assuming different ratios ($R = N_a/N_i$) between the ancestral effective population size (N_a) and the internode effective population size (N_i) are shown ($R = 0.1$, 1, and 10). Divergence time (T) is shown in units of N_i generations.

We investigated this bias using coalescent simulations for different ratios (0.1, 1, and 10) of ancestral effective population size N_a and internode effective population size N_i. We placed a single segregating site on the genealogy and compared the proportion of observed concordant genealogies based on the information provided by SNPs and the true proportion of concordant topologies. The simulations show that a proportionally larger effective size of the ancestral population does indeed result in a lower bias (fig. 3). Because sequences in empirical data do not always contain exactly one SNP, we investigated the bias when using 500-bp sequences (similar to the average alignment length in our data) and our model of dog demographic history (see Materials and Methods). By comparing the true fraction of concordant topologies and the fraction of concordant topologies computed from sequence data, we estimate the bias to less than 3% (fig. 4).

Divergence between Dogs and Wolves

Using 4,085 alignments (599–1,475 from each of four wolves) of autosomal wolf and dog sequences that had informative sites which could be used to determine the topology of the genealogy, we found that 49–58% of all loci display the concordant topology for the poodle and boxer compared with the wolf (table 1). This fraction was similar for all three Eurasian gray wolves from India (49.5%, 95% bootstrap CI: 47.0–52.0%), China (50.8%, CI 48.1–53.5%), and Spain (51.2%, CI: 47.1–55.0%) but higher for the individual from Alaska (57.5%, CI: 53.5–61.4%). We computed the likelihood of divergence times given the sequence data and find MLEs of T from 0.28 to 0.45 coalescent time units (table 1 and fig. 5A), between modern European dogs (poodle and boxer) and each of the wolves. Assuming that all

Table 1. Summary of Data and Divergence Time Estimates (corrected for potential sequencing and alignment errors).

	Number of Alignments	Average base pair /Alignment	Inferred Topologies	Fraction Concordant Topologies (95% Bootstrap CI)	Divergence Time *T* MLE (95% maximum likelihood CI)	Reference
Altai wolf	7,655	35	841	80.0% (77.2–82.7%)	1.21 (1.07 to 1.35)	Blow et al. (2008)
Alaska wolf	2,287	563	637	57.5% (53.5–61.4%)	0.45 (0.36 to 0.54)	Lindblad-Toh et al. (2005)
Spanish wolf	1,992	627	599	51.2% (47.1–55.0%)	0.31 (0.23 to 0.40)	Lindblad-Toh et al. (2005)
Chinese wolf	5,067	501	1,311	50.8% (48.1–53.5%)	0.30 (0.25 to 0.36)	Lindblad-Toh et al. (2005)
India wolf	5,606	495	1,475	49.5% (47.0–52.0%)	0.28 (0.23 to 0.33)	Lindblad-Toh et al. (2005)
Alaskan malamute	19,302	539	4,564	34.1% (32.7–35.5%)	0.011 (−0.009 to 0.032)	Lindblad-Toh et al. (2005)
Italian greyhound	19,436	540	4,480	33.8% (32.5–35.2%)	0.007 (−0.013 to 0.029)	Lindblad-Toh et al. (2005)
Beagle	20,414	538	4,876	33.7% (32.4–35.0%)	0.005 (−0.014 to 0.025)	Lindblad-Toh et al. (2005)
German shepherd	26,388	495	6,093	32.6% (31.5–33.8%)	−0.01 (−0.027 to 0.008)	Lindblad-Toh et al. (2005)
Rottweiler	20,940	540	4,933	32.3% (31.0–33.6%)	−0.015 (−0.034 to 0.005)	Lindblad-Toh et al. (2005)
Portuguese water dog	18,747	524	4,212	32.6% (30.8–33.7%)	−0.016 (−0.036 to 0.005)	Lindblad-Toh et al. (2005)
Bedlington terrier	20,117	542	4,720	31.8% (30.5–33.1%)	−0.023 (−0.042 to −0.003)	Lindblad-Toh et al. (2005)
English shepherd	20,014	539	4,325	28.5% (27.2–30.0%)	−0.07 (−0.09 to −0.05)	Lindblad-Toh et al. (2005)
Labrador retriever	27,739	512	5,637	21.8% (20.7–22.9%)	−0.16 (−0.17 to −0.15)	Lindblad-Toh et al. (2005)

sequences represent independent samples from one population, these sequences can be analyzed jointly using our approach. Combining the sequence data for all Eurasian gray wolves (excluding the Alaskan wolf) resulted in an estimated divergence time of 0.29 (maximum likelihood CI: 0.26–0.33) (fig. 5*B*).

For comparison, we also used a different method that estimates the TMRCA between sequences rather than population divergence times (Green et al. 2006; Noonan et al. 2006). This method proceeds by computing the fraction of pairwise differences that have occurred on the boxer lineage since the TMRCA, thus ignoring polymorphisms unique to the shotgun sequences. Setting the cat–dog ancestor to 55 million years ago (Pontius et al. 2007), the mean TMRCA between the boxer and the poodle is ~36,700 years ago and between the boxer and wolves ~61,800 years ago. Because this value is heavily dependent on the assumed canid-felid divergence and a clock-like mutation rate, we normalized the S_w statistic with the value for dogs S_d (see Materials and Methods) and obtained $S_w/S_d = 1.70$ (CI: 1.64–1.77). This statistic reflects the relative TMRCA between a dog and a wolf compared with the TMRCA between two dogs and is independent of mutation rate. For comparison, Lindblad-Toh et al. (2005) estimated average pairwise differences to ~1/900 between European dogs and ~1/578 between a boxer and a wolf, corresponding to $S_w/S_d = 1.56$.

Because backcrossing between dogs and gray wolves still occurs in sympatric regions and may have been even more frequent in the past (Vilá et al. 2005; Wayne and Ostrander 2007), making inferences based on a model with complete isolation of dogs and wolves might underestimate divergence times. To investigate the impact of migration on our estimates of *T*, we simulated genetic data using a population divergence model with migration (see Materials and Methods). We qualitatively assessed the ability of different assumptions on migration rate and divergence time to reproduce the genealogical concordance and S_w/S_d in our empirical data. Barring migration between the dog and wolf populations, genealogical concordance (estimated from sequences) corresponding to 49.5% (observed value for the Indian wolf, CI: 47.0–52.0%) can readily be explained by a domestication ~10,000 years ago (3,500

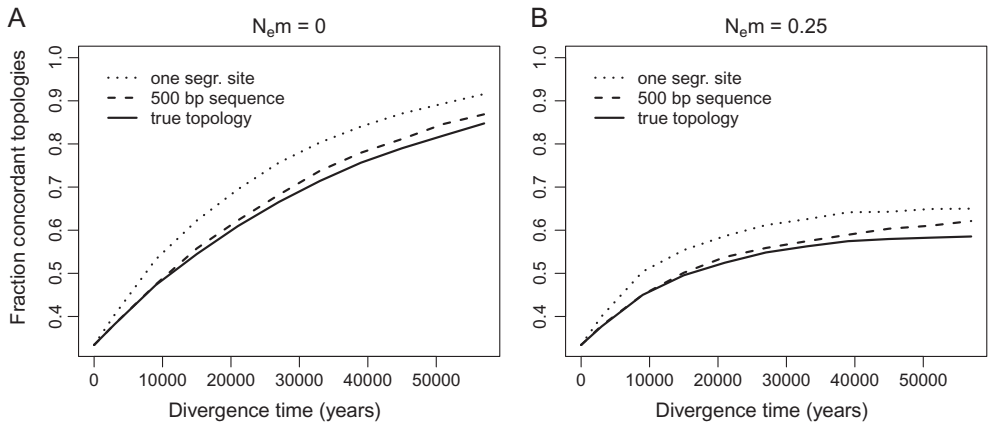

Fig. 4. The bias for observing concordant topologies due to incomplete genealogy information for (*A*) a model of dog demographic history without migration and (*B*) a model with migration. The fraction of concordant topologies estimated from a single segregating site is shown as a dotted line, the fraction estimated from a 500-bp locus with a mutation rate of 10^{-8} per bp and generation is shown as a dashed line and the true fraction of concordant genealogies is shown in as a solid line.

Table 2. Pairwise Differences, Error Estimates, and Di Divergence Time Corrections.

	S	S_s/S_p	E	$N_{CO}-N_{CT}$ (%)
India wolf	0.00181 (0.00174–0.00189)	1.81 (1.71–1.92)	0.01110 (0.01060–0.01163)	27.70
Spanish wolf	0.00177 (0.00167–0.00188)	1.74 (1.60–1.91)	0.01078 (0.01008–0.01152)	27.90
Chinese wolf	0.00172 (0.00165–0.00180)	1.67 (1.56–1.79)	0.01054 (0.01004–0.01105)	26.40
Alaska wolf	0.00159 (0.00150–0.00169)	1.48 (1.35–1.63)	0.01003 (0.00949–0.01059)	20.80
Alaskan malamute	0.00140 (0.00137–0.00144)	1.20 (1.15–1.25)	0.01011 (0.00991–0.01031)	35.80
English shepherd	0.00133 (0.00130–0.00136)	1.08 (1.04–1.14)	0.01024 (0.01005-0.01044)	40.70
German shepherd	0.00130 (0.00127–0.00133)	1.04 (1.00–1.09)	0.00832 (0.00814–0.00849)	32.60
Rottweiler	0.00130 (0.00126–0.00139)	1.04 (0.99–1.09)	0.00907 (0.00891–0.00923)	34.30
Italian greyhound	0.00129 (0.00125–0.00132)	1.02 (0.98–1.07)	0.00920 (0.00901–0.00940)	34.40
Labrador retriever	0.00126 (0.00123–0.00129)	0.98 (0.94–1.02)	0.00946 (0.00927–0.00965)	44.00
Bedlington terrier	0.00125 (0.00122–0.00128)	0.97 (0.93–1.02)	0.00858 (0.00841–0.00876)	33.70
Beagle	0.00124 (0.00121–0.00127)	0.96 (0.91–1.00)	0.00833 (0.00815–0.00850)	31.70
Portuguese water dog	0.00122 (0.00119–0.00125)	0.92 (0.87–0.97)	0.00836 (0.00818–0.00853)	33.00
Standard Poodle	0.00067 (0.00066–0.00078)	1.00	0.00014 (0.00008–0.00019)	NA

S is the fraction of differences per site between the boxer and the cat in which the sample sequence carries the cat allele, S_s / S_p is the corrected estimate of the relative TMRCA between the sample and the boxer genome normalized by the estimate for poodle sequences, E is the estimated frequency of sequencing/alignment errors per bace pair, N_{CO} – N_{CT} is the difference between the fraction of concordant topologies before and after error correction. Bootstrap CIs (95%) obtained from 10,000 pseudo-replicates over loci are shown in parentheses.

generations, CI: 9,000–13,000 years). However, assuming a low migration rate (N_em = 0.25) between dogs and wolves from T to the creation of breeds pushes the domestication time to ~30,000 years ago (CI: 15,000–90,000) (fig. 6A). The two models with migration affect the S_w/S_d statistic in a similar way, but we found that the assumption of a three to four times larger wolf effective population size compared with dogs (see Materials and Methods) is not enough to reproduce the above estimated divergence time and S_w/S_d values for the Indian wolf under our model (fig. 6B), but note that S_w/S_d for the three other wolves was slightly lower (table 2). A greater difference between the effective population sizes of wolves (N_w) and dogs (N_d) would make the simulation results more similar to our observations because TMRCAs between populations are increased by both divergence time and ancestral effective size.

Divergence between the Ancient Altai Wolf and Modern Canids

Sequences from the ancient Altai wolf displayed a fraction of 80.0% concordant genealogies with respect to the modern dog genomes (CI: 77.2–82.7%). This corresponds to an MLE of T = 1.21 (CI: 1.07–1.35), which is, compared with the estimate of the modern gray wolf–dog divergence, approximately four times as far back in time (fig. 5B). We also investigated possible divergence times using simulations. Our results indicate a divergence between the Altai wolf and the wolf population ancestral to dogs approximately 90,000 years ago (CI: 75,000–110,000, fig. 7). These simulations also illustrate that if the ancient sample is from a population that is directly ancestral to the modern populations, the fraction of concordant topologies corresponds directly to the age of the ancient specimen (fig. 7A). In addition, we obtained a S_{Altai}/S_d statistic of 3.81 (CI: 3.52–4.10), which together with the high divergence estimate indicates that the Altai wolf has a significantly deeper divergence to modern dogs than the modern gray wolves in our data set.

It has previously been shown that sequence read length can bias sequence comparisons between genomes (Green et al. 2009). To investigate the sensitivity of our estimates to fragment length, with special regard to the short read length produced by the Illumina technology used in the Altai wolf study (Blow et al. 2008), we created artificially fragmented data sets from the modern wolf sequences with sizes corresponding to the typical range of aDNA (35–250 bp). We found that data sets with short fragment length tend to have a lower uncorrected fraction of concordant topologies, resulting in underestimation of divergence times, but note that the error correction alleviates this bias (fig. 8A). In contrast, short fragment length has a tendency to increase the estimate of relative TMRCA (fig. 8B and Green et al. 2009). The bias for the uncorrected fraction of concordant topologies could be due to short fragments more frequently being mapped to an incorrect location in one of the two dog genomes but not the other, causing an incorrectly scored discordant topology. Indeed, a bias in the uncorrected analysis where the discordant topology that has the poodle lacking a canid-specific allele (found in wolf and boxer) over the discordant topology where the boxer lacks the canid-specific allele increased for short fragment lengths (Fig. 8C). A possible explanation for this is that short reads are more prone to align to a nonhomologous region in the lower-quality 1.5× poodle genome. Regardless, this bias would imply a possible underestimation of the divergence time for very short reads, supporting the conclusion that the Altai wolf is from a population distantly related to modern dogs.

Divergence between Dog Breeds

For the nine dog breeds, a range of 4,212–5,637 alignments showed informative sites that could be used to determine the topology of the gene genealogies (table 1). Using our approach based on discordant/concordant gene genealogies, we estimated the population divergence between the ancestral population of boxers and poodles and the other breeds as follows: Labrador retriever, −0.16; English

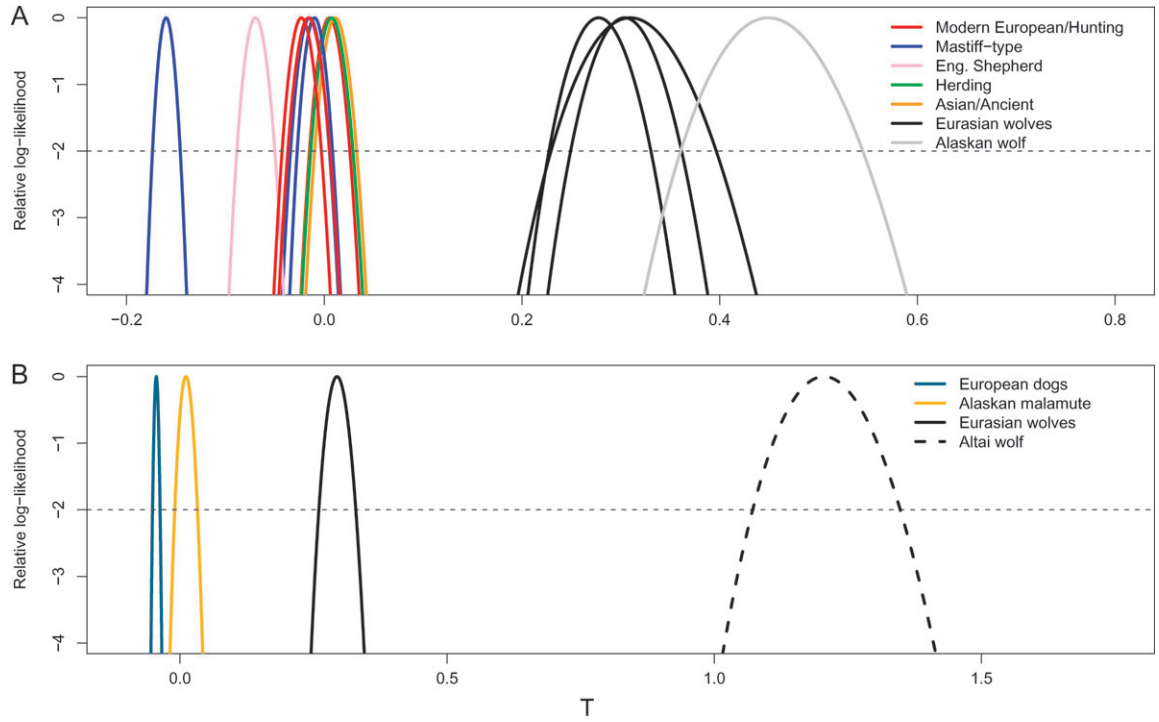

Fig. 5. Relative log-likelihood estimates of divergence time T between different dog breeds and wolves to the boxer/poodle ancestral population. (*A*) Divergence between individual dog breeds (color indicates classification by Parker et al. [2004]) and wolf populations and (*B*) joint divergence between canid populations and the ancient wolf. Dashed line represents the 95% CI cutoff.

shepherd, −0.07; Bedlington terrier, −0.02; Portuguese water dog, −0.02; Rottweiler, −0.01; German shepherd, −0.01; Beagle, 0.01; Italian greyhound, 0.01; and Alaskan Malamute, 0.01 coalescent time units (table 1). The observation of negative divergence times is not entirely unexpected considering that the boxer and poodle belonged to separate population clusters in previous studies (e.g., Parker et al. 2004, 2007; vonHoldt et al. 2010). A closer relationship of a sample dog breed to either the boxer or the poodle is a violation of the assumed population topology in our method. Although in principle the population topology can be chosen that maximizes the number of concordant topologies, our error correction method considers the two discordant topologies jointly. However, in our basic model, negative divergence times could be interpreted as relative to the divergence time of boxer and poodle populations, with time scaled in units appropriate to historical effective population size fluctuations. The rank order of inferred divergence times corresponds roughly to four previously suggested higher-order groups: "Modern European/Hunting," "Mastiff-type," "Herding," and "Asian/Ancient" (fig. 5*A*; Parker et al. 2004; Wayne and Ostrander 2007). For instance, we retrieve a separation of the Alaskan Malamute (Asian/Ancient) to other breeds, and the Italian greyhound—the

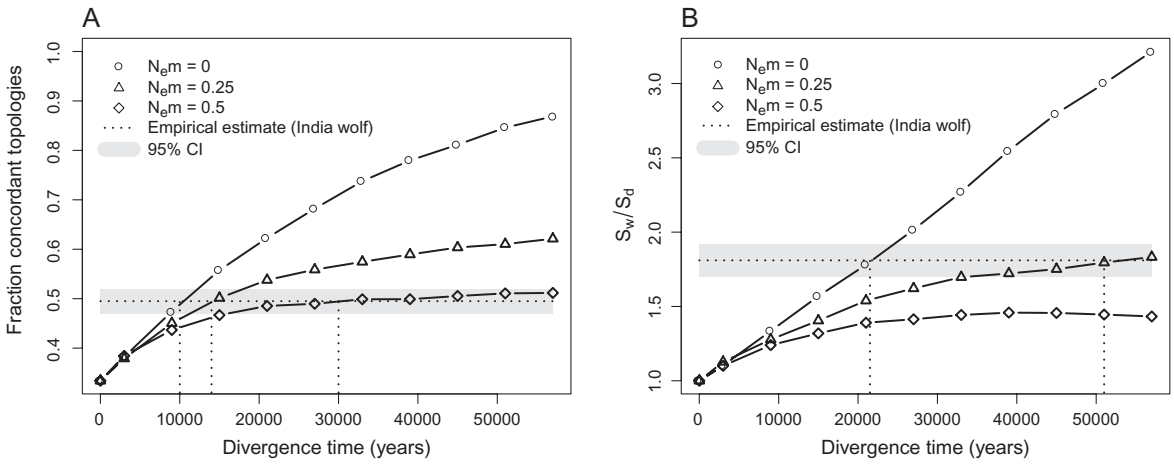

Fig. 6. The effect of assuming different migration rates (N_em) between dogs and wolves on the divergence time. The results are obtained from simulations of a model of dog demographic history (see Materials and Methods). (*A*) The effect on the fraction of concordant genealogies and (*B*) the effect on the relative TMRCA of dog and wolf sequences to the boxer genome (S_w/S_d).

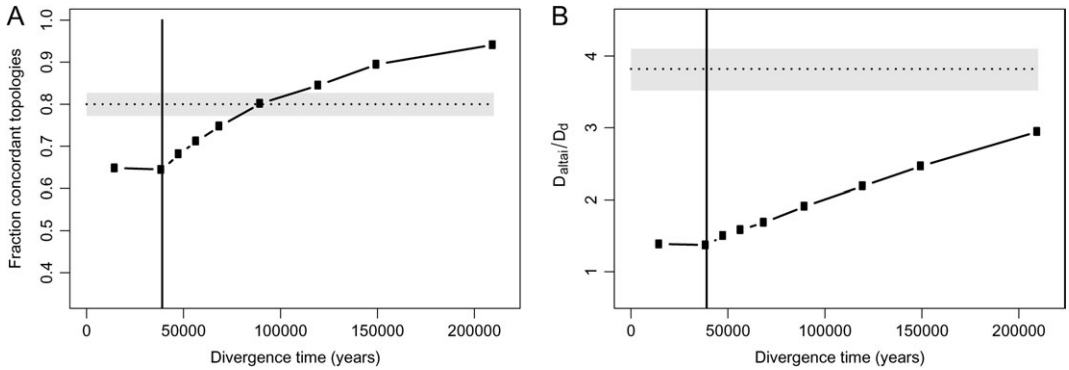

Fig. 7. Results of simulations of a potential divergence between the Altai wolf and the ancestral dog population. (*A*) Fraction of concordant topologies for the Altai wolf compared with modern dogs and (*B*) relative average TMRCA between the Altai wolf and modern dog genomes (S_{Altai}/S_d). The age of the Altai remains is shown by a solid vertical line, and the empirically estimated values with a horizontal dashed line, with the 95% CI indicated by gray shading.

only representative from the Herding breed cluster—displayed the second largest divergence time from the boxer–poodle ancestor, followed by most "Mastiff-type" breeds and finally most breeds from the "Modern European" cluster. Our analysis also included data from an English Shepherd, a breed that to our knowledge has not been included in previous multilocus studies on population structure. We found that its inferred divergence time is most similar to dogs from the Modern European/Hunting cluster.

Discussion

The probability of concordant or discordant genealogies has been treated in the literature (e.g., Hudson 1983; Takahata 1989; Rosenberg 2002; Wakeley 2008) but has mainly been used in different methods to estimate ancestral effective population sizes in primate evolution (e.g., Chen and Li 2001; Yang 2002; Hobolth et al. 2007). To our knowledge, the approach has previously not been used to infer divergence times from nonoverlapping sets of shotgun sequences. The use of this summary statistic is robust to a number of problems associated with evolutionary analysis of aDNA data. For instance, due to only utilizing information from mutations occurring on the internal branches of genealogies of three sequences (plus the outgroup), a postmortem mutation or sequencing error must confer a change from a derived variant to the allele present in the outgroup sequence to affect the outcome of the analysis. Although we have illustrated that utilizing mutations confers a bias if genealogies in the concordant and discordant classes have different average lengths of the internal branch and that this effect is stronger if the internode time T is much greater than N_e in the ancestral population, the bias can be accounted for with simulations (fig. 4). Only utilizing information on the internal branch also makes the approach appropriate for dealing with samples from different time points (Figs 1 and 7A). In practice, inferring the genealogy of a sample of sequences requires that some mutation event happened on the internal lineages, and we only need to assume that the mutation rate is equal for all lineages. In principle, this makes it possible to jointly consider markers with different mutation rates (e.g., SNPs and indels), but without a closely related outgroup sequence the states of indels are difficult to determine, a problem that might be smaller for other studies (e.g., chimpanzee-human). Indeed, the use of a closely related outgroup genome is likely to improve most population genetic estimates from ancient genomic data (Prüfer et al. 2010), but we have shown that appropriate corrections can alleviate biases introduced by sequencing errors (table 2) and short fragment length (fig. 8A), even for rather distant genome comparisons such as between cats and canids.

We obtained estimates of divergence time between wolf and dog populations using a methodological framework that is computationally flexible and robust to problems such as time-structured sampling. Assuming no gene flow between wolves and dogs after domestication, our results indicate a divergence between dogs and wolves at least 10,000 years ago (CI: 9,000–13,000), a conclusion that follows from assuming an ancestral effective population size (N_e) of ~13,000 chromosomes and a generation time of 3 years (Lindblad-Toh et al. 2005; Gray et al. 2009). This number is consistent with some previous estimates based on a divergence model with complete isolation between the populations (e.g., Savolainen et al. 2002; Gray et al. 2009). However, several studies have documented hybridization between gray wolves and dogs in sympatric regions (Randi 2008), and this interbreeding may have been even more extensive in the past (Vilá et al. 2005; vonHoldt et al. 2010). For instance, 5% of all samples from an encroached Italian wolf population had admixed ancestry despite showing lupine mtDNA and Y-haplotypes (Verardi et al. 2006).

Unlike previous analyses (e.g., Vilà et al. 1997; Savolainen et al. 2002; Lindblad-Toh et al. 2005; Gray et al. 2009; Pang et al. 2009), we also consider the effect of gene flow after the domestication of dogs using an explicit population divergence model with migration. Although we do not estimate the extent of hybridization, we used relatively conservative population migration rates (N_em = 0.25–0.5), and found that gene flow between dogs and wolves is consistent with a divergence ~14,000 (CI: 11,000–18,000) to ~30,000 years ago (CI: 15,000–90,000), for the respective migration rates. Although the exact choices of

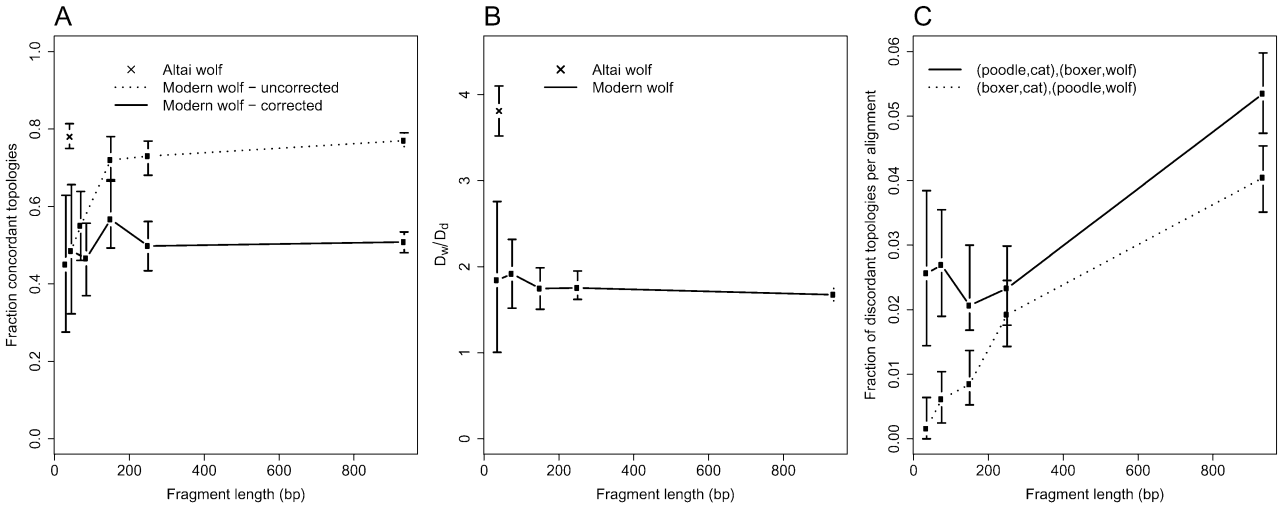

Fig. 8. The effect of fragment length on estimates of population divergence and relative TMRCA between dogs and wolves. (*A*) Uncorrected and corrected fraction of concordant topologies, (*B*) corrected S_w/S_d, and (*C*) uncorrected frequency of the two discordant topologies per alignment. In each plot is data shown for four randomly sampled data subsets from a modern (Chinese) wolf of 10,000 sequences with fragment length 35, 75, 150, and 250 bp, respectively, and the full Chinese wolf data set (mean fragment length 934 bp). In (*B*) and (*C*), observed values for the Altai wolf (mean fragment length 40 bp) is shown for comparison. Error bars show 95% bootstrap CIs.

migration rates in our simulations are merely examples, the accumulating evidence of gene flow between wolves and dogs in many parts of the world (e.g., vonHoldt et al. 2010) makes these observations hard to reconcile with the hypothesis of an origin of dogs during the Neolithic transition in East Asia suggested by Pang et al. (2009) on the basis of mtDNA variation. Instead, our results could be taken as support for the hypothesis that domestication was instigated earlier in human history, as indicated by recent archaeological finds (e.g., Sablin and Khlopachev 2002; Germonpré et al. 2009). More complex models incorporating bottlenecks and varying migration rates could possibly also explain the observed data. However, previous analyses have pointed to two important demographic events in the history of modern dog breeds, a first bottleneck during domestication and a second one during breed formation (Lindblad-Toh et al. 2005). A recent study on canid genetic variation concluded that domestication conferred a bottleneck in the form of a contraction in effective size without subsequent recovery (Gray et al. 2009), and we find support for this model in contrasting our divergence estimates with the average TMRCA between dogs and wolves (fig. 6).

In addition, genome-wide analyses of dogs and wolves recently showed that Middle Eastern wolves probably contributed the major part of the ancestry of modern dogs (vonHoldt et al. 2010). This is in agreement with our estimates that Indian wolves are most closely related to European dog genomes (followed by Chinese, Spanish, and Alaskan wolves) (table 1), but unfortunately no sequence data from Middle Eastern wolves were available for analysis. It is noteworthy that this rank order of divergence for wolves of different regions differs from what is obtained using a modified pairwise differences statistics (S_w/S_d in our notation), an approach that is commonly used for genomic aDNA (Green et al. 2006, 2009, 2010; Noonan et al. 2006; Prüfer et al. 2010). Here, the Alaskan wolf was closest (followed by Chinese, Spanish and Indian wolves, table 2), which indicates that the topological method is able to retrieve results more congruent with other sources of information.

Our analysis of genomic data from a Pleistocene wolf from Altai, Russia (dated to 40–50 Ka), could also provide information about the ancestral wolf population. However, we find that sequence and population divergence estimates between the Altai wolf and the modern dog genomes are difficult to explain by the age of the remains alone (fig. 7). Instead, our results indicate that this individual was part of a wolf population separated from the population that is ancestral to modern gray wolves by ~90,000 years (CI: 75,000–110,000 years). This observation is compatible with previous molecular and morphological evidence for several divergent gray wolf populations in the North American-Eurasian tundra during the Late Pleistocene (Leonard et al. 2007; Pilot et al. 2010) as well as extant endemic wolf populations (Sharma et al. 2004), and the divergence time for the Altai wolf estimated here highlights the possibility of regional discontinuity between extant and ancestral wolf populations also in Central Asia.

Using nonoverlapping single-individual sequence data from nine different dog breeds, we were also able to recover previously identified patterns of higher-order population clusters (Parker et al. 2004). We find estimates of divergence time that separates the Alaskan Malamute from European breeds, but the divergence time is surprisingly low compared with the observed number of pairwise differences in the same data (see table 2 and Lindblad-Toh et al. 2005), which could indicate that the divergence of the Alaskan malamute occurred at a time when the effective population size of dogs was large compared with the effective population size during the more recent period of breed

creation. Although the exact rank order of divergence between modern dog breeds is unresolved (vonHoldt et al. 2010), we found a high degree of correspondence between our inferred divergence times for European breeds and previously inferred breed clusters (compare fig. 5*A* and results in Parker et al. 2004, 2007, and vonHoldt et al. 2010). The observation of negative divergence times is not entirely unexpected considering that the boxer and poodle belonged to separate population clusters in previous studies (e.g., Parker et al. 2004, 2007; vonHoldt et al. 2010), and that the extremely low N_e in modern breeds (Sutter et al. 2004; Lindblad-Toh et al. 2005) causes the coalescent time scale to be up to 10 times faster in the last 200–300 years compared with the prehistoric population (Clutton-Brock 1987, 1995; Lindblad-Toh et al. 2005; Gray et al. 2009).

Conclusion

We show that recent population divergence times can be retrieved from nonoverlapping shotgun sequencing data from different points in time. Using this information, we have independently recapitulated previously described patterns of canid evolutionary history and obtained new estimates on the timing of dog domestication. In line with some previous estimates, we find that a model without migration is consistent with dogs being domesticated 10,000 years ago. However, a low rate of gene flow is compatible with a much earlier domestication date. We also demonstrate that a Central Asian Pleistocene wolf has a deeper divergence with modern dogs than would be expected by its mere age and that this individual is unlikely to represent a population directly ancestral to modern dogs. Our approach is tailored for analysis of large nonoverlapping sets of sequences and should be applicable to a number of species for which multiple reference genomes are becoming available.

Acknowledgments

We are grateful to Benoit Nabholz for helpful computational suggestions and to Tom Gilbert, Eske Willerslev, and two anonymous reviewers for valuable comments on a previous version of the manuscript. We also wish to thank Matthew Blow and Edward Rubin for kindly providing ancient wolf data. This work was supported by a Sven and Lilly Lawski Foundation for Natural Sciences scholarship to P.S. and a grant from the Swedish Research Council FORMAS to M.J. A.G. was supported by the Royal Swedish Academy of Science while participating in this study. Computations were performed on Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) resources under project p2009038.

References

Anderson CNK, Ramakrishnan U, Chan YL, Hadly EA. 2005. Serial SimCoal: a population genetic model for data from multiple populations and points in time. *Bioinformatics* 21:1733–1734.

Axelsson E, Willerslev E, Gilbert MTP, Nielsen R. 2008. The effect of ancient DNA damage on inferences of demographic histories. *Mol Biol Evol*. 25:2181–2187.

Benecke N. 1987. Studies on early dog remains from Northern Europe. *J Arch Sci*. 14:31–49.

Blow MJ, Zhang T, Woyke T, Speller CF, Krivoshapkin A, Yang DY, Derevianko A, Rubin EM. 2008. Identification of ancient remains through genomic sequencing. *Genome Res*. 18:1347–1353.

Boyko AR, Boyko RH, Boyko CM, et al. (15 co-authors). 2009. Complex population structure in African village dogs and its implication for inferring dog domestication history. *Proc Natl Acad Sci U S A*. 106:13903–13908.

Briggs AW, Stenzel U, Johnson PL, et al. (11 co-authors). 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A*. 104:14616–14621.

Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, Cooper A. 2007. Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acid Res*. 35:5717–5728.

Burbano HA, Hodges E, Green RE, et al. (20 co-authors). 2010. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science*. 328:723–725.

Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol Biol Evol*. 25:1979–1994.

Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet*. 68:444–456.

Clutton-Brock J. 1987. A natural history of domesticated mammals. Cambridge (UK): Cambridge University Press, British Museum of Natural History.

Clutton-Brock J. 1995. Origins of the dog: domestication and early history. In: Serpell J, editor. The domestic dog: its evolution, behaviour and interactions with people. Cambridge (UK): Cambridge University Press. p. 7–20.

Davis SJM, Valla FR. 1978. Evidence for domestication of the dog 12,000 years ago in the Natufian of Israel. *Nature* 276:608–610.

Depaulis F, Orlando L, Hänni C. 2009. Using classical population genetics tools with heterochroneous data: time matters!. *PLoS ONE*. 4:e5541.

Derevianko AP, Shunkov MV, Agadjaniav AK, Baryshnikov GF, Malaeva EM, Ulianov VA, Kulik NA, Postnov AV, Anoikin AA. 2003. Paleoenvironment and paleolithic human occupation of Gorny Altai (subsistence and adaptation in the vicinity of Denisova Cave). Novosibirsk (Russia): Russian Academy of Sciences.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acid Res*. 32:1792–1797.

Excoffier L, Novembre J, Schneider S. 2000. SIMCOAL: a general coalescent program for simulation of molecular data in interconnected populations with arbitrary demography. *J Hered*. 91:506–509.

Germonpré M, Sablin MV, Stevens RE, Hedges REM, Hofreiter M, Stiller M, Despré VR. 2009. Fossil dogs and wolves from Paleolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. *J Arch Sci*. 36:473–490.

Gilbert MTP, Bandelt HJ, Hofreiter M, Barnes I. 2005. Assessing ancient DNA studies. *Trends Ecol Evol*. 20:541–544.

Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, Zhu L, Ostrander EA, Wayne RK. 2009. Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics* 181:1493–1505.

Green RE, Krause J, Ptak SE, et al. (11 co-authors). 2006. Analysis of one million basepairs of Neandertal DNA. *Nature* 444:330–336.

Green RE, Briggs AW, Krause J, Prüfer K, Burbano HA, Siebauer M, Lachmann M, Pääbo S. 2009. The Neandertal genome and ancient DNA authenticity. *EMBO J*. 28:2494–2502.

Green RE, Krause J, Briggs AW. (56 co-authors). 2010. A draft sequence of the Neandertal genome. *Science* 328:710–722.
Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet*. 3:e7.
Hudson RR. 1983. Testing the constant-rate neutral model with protein sequence data. *Evolution* 37:203–217.
Hudson RR. . Gene genealogies and the coalescent process. *Ox Surv Evol Biol*. 7:1–44.
Jiang R, Tavaré S, Marjoram P. 2009. Population genetic inference from resequencing data. *Genetics* 181:187–197.
Johnson PLF, Slatkin M. 2008. Accounting for bias from sequencing errors in population genetic estimates. *Mol Biol Evol*. 25:199–206.
Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL, Pop M, Wang W, Fraser CM, et al. 2003. The dog genome: survey sequencing and comparative analysis. *Science* 301:1898–1903.
Leonard JA, Vilà C, Fox-Dobbs K, Koch PL, Wayne RK, Van Valkenburgh B. 2007. Megafaunal extinctions and the disappearance of a specialized wolf morph. *Curr Biol*. 17: 1146–1150.
Leonard JA, Wayne RK, Wheeler J, Valadez R, Guillén S, Vilà C. 2002. Ancient DNA evidence for old world origin of new world dogs. *Science* 298:1613.
Lindblad-Toh K, Wade CM, Mikkelsen TS, et al. (11 co-authors). 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803–819.
Liu X, Fu Y-X, Maxwell TJ, Boerwinkle E. 2010. Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. *Genome Res*. 20:101–109.
Lynch M. 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* 182:295–301.
Millar CD, Huynen L, Subramanian S, Mohandesan E, Lambert DM. 2008. New developments in ancient genomics. *Trends Ecol Evol*. 23:386–393.
Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A, et al. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* 456:387–390.
Morey DF. 2006. Burying key evidence: the social bond between dogs and people. *Journal of Archaeological Science*. 33:158–175.
Musil R. 1984. The first known domestication of wolves in central Europe. Animals and archaeology. Vol. 4, Husbandry in Europe. Oxford, UK: British Archaeological. Reports International Series 227.
Nielsen R. 1998. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theor Popul Biol*. 53:143–151.
Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. *Mol Ecol*. 18:1034–1047.
Nobis G. 1979. Der älteste haushund lebte vor 14 000 Jahre. *Umschau* 19:610.
Noonan JP, Coop G, Kudaravalli S, et al. (11 co-authors). 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* 314:1113–1118.
Olsen SJ. 1985. Origins of the domestic dog: the fossil record. Tuscon (AZ): University of Arizona press.
Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, Malek TB, Johnson GS, DeFrance HB, Ostrander EA, Kruglyak L. 2004. Genetic structure of the purebred domestic dog. *Science* 304:1160–2116.
Parker HG, Kukekova AV, Akey DT, Goldstein O, Kirkness EF, Baysac KC, Mosher DS, Aguirre GD, Acland GM, Ostrander EA. 2007. Breed relationships facilitate fine-mapping studies: a 7.8-kb deletion cosegregates with Collie eye anomaly across multiple dog breeds. *Genome Res*. 17:1562–1571.
Pang J-F, Kluetsch C, Zou X-JZhang A, Luo L-Y, Angleby H, Ardalan A, Ekström C, Sköllermo A, Lundeberg J, et al. 2009. mtDNA data indicate a single origin for dogs south of Yangtze river, less than 16,300 years ago, from numerous wolves. *Mol Biol Evol*. 26:2849–2864.
Pääbo S. 1989. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc Natl Acad Sci U S A*. 86:1939–1943.
Pilot M, Branicki W, Jędrzejewski W, Goszczyński J, Jędrzejewska B, Dykyy I, Shkvyrya M, Tsingarska E. 2010. Phylogeographic history of grey wolves in Europe. *BMC Evol Biol*. 10:104.
Poinar HN, Schwarz C, Qi J, et al. (13 co-authors). 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 311:392–394.
Pontius JU, Mullikin JC, Smith DR, et al. (11 co-authors). 2007. Initial sequence and comparative analysis of the cat genome. *Genome Res*. 17:1675–1689.
Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE. 2010. Computational challenges in the analysis of ancient DNA. *Genome Biol*. 11:R47.
Randi E. 2008. Detecting hybridization between wild species and their domesticated relatives. *Mol Ecol*. 17:285–293.
Rasmussen M, Li Y, Lindgreen S, et al. (11 co-authors). 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463:757–762.
Rosenberg NA. 2002. The probability of topological concordance of gene trees and species trees. *Theor Popul Biol*. 61:225–247.
Sablin MV, Khlopachev GA. 2002. The earliest Ice Age dogs: evidence from Eliseevichi I. *Curr Anthr*. 43:795–799.
Savolainen P, Zhang J, Luo J, Lundeberg J, Leitner T. 2002. Genetic evidence for an East Asian origin of domestic dogs. *Science* 298:1610–1613.
Sharma DK, Maldonado JE, Jhala YV, Fleischer RC. 2004. Ancient wolf lineages in India. *Biol Lett*. 271:S1–S4.
Sutter NB, Eberle MA, Parker HG, Pullar BJ, Kirkness EF, Kruglyak L, Ostrander EA. 2004. Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Res*. 14:2388–2396.
Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122:957–966.
Wall JD, Kim SK. 2007. Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genetics*. 3:1862–1866.
Wayne RK, Ostrander EA. 2007. Lessons learned from the dog genome. *Trends Genet*. 23:557–567.
Wakeley J. 2008. Coalescent theory: an introduction. Greenwood Village (CO): Roberts & Company Publishers.
Verardi A, Lucchini V, Randi E. 2006. Detecting introgressive hybridisation between free-ranging domestic dogs and wild wolves (*Canis lupus*) by admixture linkage disequilibrium analysis. *Mol Ecol*. 15:2845–2855.
Vilà C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, Crandall KA, Lundeberg J, Wayne RK. 1997. Multiple and ancient origins of the domestic dog. *Science* 276:1687–1689.
Vilà C, Maldonado JE, Wayne RK. 1999. Phylogenetic relationships, evolution, and genetic diversity of the domestic dog. *J Hered*. 90:1.
Vilà C, Seddon J, Ellegren H. 2005. Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends Genet*. 21:214–218.
Willerslev E, Cooper A. 2005. Ancient DNA. *Proc R Soc Lond B Biol Sci*. 272:3–16.
vonHoldt BM, Pollinger JP, Lohmueller KE, et al. (36 co-authors). 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* 464:898–902.
Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 162:1811–1823.
Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comp Biol*. 7:203–221.