Supplementary information, Data S7 BPP/G-PhoCS results

a) Dating three important nodes in the history of dog domestication

The site-frequency based method provides a good way of estimating population divergence, but due to the limited sample sizes of many clades, it is not possible to estimate all divergence times. In order to estimate divergence times between the different clades and also confirm the estimated divergence time from the dadi, we used BPP[12] and G-PhoCS[13] to date the different nodes in Figure 2B.

Both BPP and G-PhoCS are derived from an earlier program (MCMCcoal) developed by Ziheng Yang. The difference between the two programs is that G-PhoCS allows migrations between the different populations. As we show in the next section, under most conditions, the estimate of the population divergence time is relatively robust as long as migration rates are not too high ($4Nm \leq 10$). We thus used both programs to date several important nodes in the history of dog domestication.

Since the number of parameters increases quite quickly with the number of populations, we chose to stay with the three-population scheme (large number of populations can often run into difficulties associated with the convergence of the Markov Chain Monte Carlo, MCMC, method). For each internal node, we used an outgroup (wolves in this case) and two descendant populations. The two descendant populations are each chosen from the left and right clade underneath the internal node of interests. For example, if the left clade has X number of populations and right clade has Y number of populations, we try all possible X*Y combinations in dating the age of the internal node.

We dated three important internal nodes in the history of dog domestication. They are a) the divergence time between wolves/dogs (Tau1); b) the divergence time between southern Chinese dogs and other dog groups (Tau2), and c) the age of the most recent common ancestor (MRCA) for all European breeds (Tau3).

The first group analysis used the (wolves, (southern Chinese indigenous dogs, other dogs)) as the input. The other dogs are the dogs that do not show signal of admixture.

In this 58-sample dataset, aside from the groups showing signal of admixture (i.e., northern Chinese indigenous dogs, African and Middle Eastern dogs and Arctic groups), we did the analysis for all other dogs (15 breeds). For each population combination, we repeated the analysis twice by randomly picking 1000 loci from the genome data.

As presented in the Supplementary information, Figure S6, the estimated population sizes (Ne) for the contemporary and ancestral wolves are around 81,440 and 63,260, respectively (Supplementary information, Figure S6 and Table S5). The divergence time between the southern Chinese indigenous dogs and other dogs (Tau2) is about 14,545 years and the divergence between wolves and dogs (Tau1) is about 30,455 years using BPP and 37,273 with G-PhoCS program (Supplementary information, Table S5, see also section b for a discussion about this discrepancy).

The second group analysis is conducted to date the age for node 3 in Figure 2A. This node represents the most recent common ancestor of European breeds. As shown in Supplementary information, Table S5 and Figure S7, the population sizes estimated using this combination are very close to the results from the first group analysis. Interestingly, the time to the most common ancestor for the European breeds is found to be around 9,500 years, which was later than the origin of modern agriculture in the Middle East (around 14K years ago)[14].

b) Robustness of the divergence time estimates under a wide variety of demographic histories and the difference between the BPP and G-PhoCS estimates.

Given that the Bpp/G-PhoCS methods are built assuming very idealistic models (i.e., constant population sizes), we wanted to know how violations of the model assumptions affect our inferences on population divergence time. We thus simulated a wide variety of demographic scenarios and examined whether the inference on the divergence time can be strongly affected by the violation of the model assumptions.

We simulated a simple two-population divergence scenario using a standard model (case I) and a series of non-standard histories including migration (case 2-4), growth (case 5-7) and growth + migration (case 8-10) (Supplementary information, Table

S6). For each of these cases, we simulated data using the ms[15] and seqgen[16] to generate haploid data for both BPP and G-PhoCS to estimate the population parameters. In all the scenarios, we allowed migrations between the populations regardless whether migrations are allowed in the simulated data using G-PhoCS.

In Supplementary information, Figure S8, we can see that, in a wide variety of scenarios, the estimated species divergence time is very close to the simulated value. The divergence time estimation is quite robust to the perturbation of the model assumptions.