Supplementary information, Data S16 Supplementary materials and methods

1 Sample collection

Blood or tissue samples were obtained from 12 gray wolves (8 blood samples and 4 tissue samples), 27 village dogs (24 blood samples and 3 tissue samples), 19 breed dogs (16 with blood and 3 with tissue samples), and 1 dhole as an outgroup (*Cuon alpinus*). The geographic distribution of these individuals is shown in Figure 1A and Supplementary information, Table S1.

Of the wolves, 9 gray wolves are from China (Xinjiang, Shanxi, and Inner Mongolia), and 3 from Russia. 27 indigenous/village dogs are from East Asia and Africa. Among these 27 individuals, 11 are from Southern East Asia, 12 from northern East Asia, and 4 from Nigeria.

The 19 breeds are AFG, Afghan Hound; SLO, Sloughi; BEM, Belgian Malinois; CHI, Chihuahua; FIL, Finnish Lapphund; GAL, Galgo; GNE, Gray Norwegian Elkhound; GSD, German Shepherd Dog; JAM, Jamthund; LAH, Lapponian Herder; MEN, Mexican Naked (hairless); PEN. Peruvian Naked (hairless); SWL, Swedish Lapphund; SAM, Samoyed; ESL, East Siberian Laika; SIH, Siberian Husky; ALM, Alaska Malamute; GRD, Greenland dogs, and TIM, Tibetan Mastiff.

Of the 58 individuals, 10 individuals were sequenced in a previous study (Supplementary information, Table S1). Sample locations for the dogs and wolves are shown in Figure 1A, and Supplementary information, Table S1.

2 DNA library construction and sequencing.

Total genomic DNA was extracted from blood or tissue samples using the phenol/chloroform method. For each individual, 1-3 μg of DNA was sheared into fragments of 200–800 bp with the Covaris system. DNA fragments were then processed according to the Illumina DNA sample preparation protocol: fragments were end-repaired, A-tailed, ligated to paired-end adaptors and PCR-amplified with ~500 bp inserts for library construction. Sequencing was performed on the Illumina HiSeq 2000 platform, and 100-bp paired-end reads were generated.

3 Sequence data pre-processing and variant calling

Raw sequence reads (fastq format) were mapped to the dog reference genome (Canfam3, downloaded from the UCSC genome browser) with the Burrows-Wheeler Aligner (BWA, Version 0.6.2-r126)[25]. Reads with identical start/end points were dumped with PICARD (Version 1.87) and subsequently locally realigned and base-recalibrated using the Genome Analysis Tool Kit (GATK, Version 2.5-2-gf57256b)[3].

Individually generated BAM files were then conjugated to call variants using the UnifiedGenotypeCaller from the GATK package. Raw variants were then recalibrated using the Variant Quality Score Recalibration (VQSR module). During the base and variant recalibration, a list of known SNPs/indels were downloaded from the Ensembl database (ftp://ftp.ensembl.org/pub/release-73/variation/vcf/canis_familiaris/) and were used as the training set. Small indels were separately called using the SAMtools mpileup.

4 Genetic diversity, linkage disequilibrium and structure analysis

Beagle was used to impute the missing genotypes and phase the genotypes into associated haplotypes[26]. Genetic diversity for each individual as well as for several sub-groupings was calculated using a custom python script. Linkage disequilibrium for different populations was calculated using the haploview software (Version 1.2)[27]. Variants with minor allele frequencies less than 0.05 were filtered before the LD calculation. Population structure analysis was done using the EM algorithm implemented in the Frappe package (Version 1.1)[28]. Before the Frappe analysis, variants were first thinned to be at least 50kb in distance from nearby SNPs. Principle component analysis was carried out using the smartPCA program in the Eigensoft package (Version 4.2)[19].

5 Mutation rate estimation using comparative genomic data

In order to determine the long-term mutation rate, we extracted whole genome alignment data for 15 eutherian mammals from the Ensembl database (http://asia.ensembl.org/info/genome/compara/analyses.html#ancestral). Human was used as the outgroup and sister species (cat, horse or cattle) were chosen for dogs. For each possible sister species, we did a three species comparison (Human, (dog, sister_species)) by extracting information from the multiple species alignment. The branch length along the dog lineage was estimated using the baseml package in the PAML package (Version 4.7)[29]. The long term evolutionary of rate along the dog lineage is then calculated using the branch length divided by species divergence time between the sister species and dog. Between species divergence time is extracted from Hedges 2002[30].

6 Population demography

Population level admixture analysis was first carried out using the TreeMix program (Version 1.12)[31]. Allele count data for each SNP was first extracted and subsequently ported to the TreeMix program. Admixture analysis was conducted by allowing a number of migration tracks. An alternative way of inferring the admixture histories for multiple populations is the F3/F4 test[6]. The threepop/fourpop module from the TreeMix package is used to carry out the F3/F4 test.

Pairwise Sequentially Markovian Coalescent (PSMC) model was used to estimate the population histories from individual genomes using the PSMC package[32]. Because sequence coverage is an important factor in determining the inferred population sizes, a correction factor is invoked to correct for the false negatives in SNP calling. This correction factor is inferred by creating datasets with a series of different coverages and finds the values to recover the true curve (high coverage result).

The joint site frequency spectrum between wolves and the South Chinese Indigenous dogs was used to infer the population history with the dadi package (Version 1.6.3)[33]. The joint site frequency spectrum for indigenous dogs and wolves was extracted following these steps: a) first, genotypes with qualities less than 18 (corresponding to roughly 30 in samtools, i.e. 0.1% error rates) were masked. b) The remaining site frequencies were then projected down to low dimensions. For example, from (wolf_samplesize=24, dog_samplesize=22) to (16, 16) using the dadi package. A wide range of projections were tried in parallel. c) Lineage specific substitution matrix was estimated using the ambiore package (Version 1.0)[34] with the whole genome sequence alignments between the dhole (Supplementary information, Table S1) and dog genome. This information is then used to estimate the lineage specific substitution matrix. d) A corrected site-frequency spectra (SFS) was then used to perform the demographic inference.

Because the ancestral population of wolves might not be at equilibrium, we allowed the wolf population to start changing continuously from an equilibrium population at some time in the past (T1). During the continuous change (i.e. from T1 to now), at some more recent time T2, dog population splits off and start to change continuous from size one (S1) to the end size (S2) (Figure 2C).

Bayesian analysis on species evolutionary history was conducted using both the BPP (Version 2.2) and G-PhoCS package (Version 1.2.2) independently[12, 13]. First, noncoding regions of 500bp length were first extracted from the genome-wide data (at least xkb distance away from nearest coding regions. We tried a variety of distances including 10-100kb). Because of the computational burden, this large set was subselected into smaller subsets (500 loci -3000loci) before the Markov Chain Monte Carlo method was carried out. The difference between BPP and G-PhoCS is that, G-PhoCS can allow migrations to happen between populations, while BPP doesn't allow for any gene flow between populations.

Population admixture time was conducted using the HAPMIX program (Version

1.2)[35]. We used the Southern Chinese indigenous dogs and the breeds as the two source populations. The program will give maximum likelihood estimates on the admixture time for each individual. The genetic distances between SNPs were extracted from a previous published genetic map[36]. The overall admixture time is inferred by maximizing the likelihood combing the likelihood values from all the individuals.

7 Targets of positive selection

The SweepFinder algorithm was used to extract regions of the genome that show the strongest signal of positive selection[24]. The genome-wide site frequency spectrum is used as the background site frequency distribution before fitting a sweep model to the data. In order to leverage all the samples (not just the Southern Chinese indigenous group), the genetic diversity and Fst values were also used to extract top candidate signals. We first calculated the mean Fst value and diversity ratio (wolf/dog) for each window (50kb stepping at 20kb). We then filtered target-selected regions by requiring both measurements to be in the top 2% of each distribution. Gene Ontology (GO) analysis was carried using the DAVID[37] and phenotypic enrichment was conducted using GeneDecks programs available at the genecards website (www.genecards.org).

References for all Supplementary information

1 Boyko AR, Boyko RH, Boyko CM *et al.* Complex population structure in African village dogs and its implications for inferring dog domestication history. *Proc Natl Acad Sci U S A* 2009; **106**:13903-13908.

2 Crapon de Caprona M-D, Savolainen P. Extensive Phenotypic Diversity among South Chinese Dogs. *ISRN Evol Biol* 2013; **2013**:1-8.

3 DePristo MA, Banks E, Poplin R *et al.* A framework for variation discovery and genotyping

using next-generation DNA sequencing data. *Nat Genet* 2011; **43**:491-498.

4 The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**:1061-1073.

5 vonHoldt BM, Pollinger JP, Earl DA *et al.* A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Res* 2011; **21**:1294-1305.

6 Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. *Nature* 2009; **461**:489-494.

7 Kumar S, Subramanian S. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* 2002; **99**:803-808.

8 Liu GE, Matukumalli LK, Sonstegard TS, Shade LL, Van Tassell CP. Genomic divergences among cattle, dog and human estimated from large-scale alignments of genomic sequences. *BMC Genomics* 2006; **7**.

9 Freedman AH, Gronau I, Schweizer RM *et al.* Genome sequencing highlights the dynamic early history of dogs. *Plos Genet* 2014; **10**:e1004016.

10 Skoglund P, Götherström A, Jakobsson M. Estimation of population divergence times from non-overlapping genomic sequences: examples from dogs and wolves. *Mol Biol Evol* 2011; **28**:1505-1517.

11 Lindblad-Toh K, Wade CM, Mikkelsen TS *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 2005; **438**:803-819.

12 Yang Z, Rannala B. Bayesian species delimitation using multilocus sequence data. *Proc Natl Acad Sci U S A* 2010; **107**:9264-9269.

13 Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human

demography from individual genome sequences. *Nat Genet* 2011; **43**:1031-1034.

14 Riehl S, Zeidi M, Conard NJ. Emergence of Agriculture in the Foothills of the Zagros Mountains of Iran. *Science* 2013; **341**:65-67.

15 Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 2002; **18**:337-338.

16 Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences : CABIOS* 1997; **13**:235-238.

17 Pang JF, Kluetsch C, Zou XJ *et al.* mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. *Mol Biol Evol* 2009; **26**:2849-2864.

18 Patterson N, Moorjani P, Luo Y *et al.* Ancient admixture in human history. *Genetics* 2012; **192**:1065-1093.

19 Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *Plos Genet* 2006; **2**:e190.

20 Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet* 2014; **30**:377-389.

21 Giuffra E, Kijas JMH, Amarger V, Carlborg O, Jeon JT, Andersson L. The origin of the domestic pig: Independent domestication and subsequent introgression. *Genetics* 2000; **154**:1785-1791.

22 Huang X, Kurata N, Wei X *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* 2012.

23 Durand EY, Patterson N, Reich D, Slatkin M. Testing for Ancient Admixture between Closely Related Populations. *Mol Biol Evol* 2011; **28**:2239-2252.

24 Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. *Genome Res* 2005; **15**:1566-1575.

25 Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010; **26**:589-595.

26 Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81**:1084-1097.

27 Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**:263-265.

28 Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiol* 2005; **28**:289-301.

29 Yang ZH. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007; **24**:1586-1591.

30 Hedges SB. The origin and evolution of model organisms. *Nat Rev Genet* 2002; **3**:838-849.

31 Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *Plos Genet* 2012; **8**:e1002967.

32 Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature* 2011; **475**:493-496.

33 Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint

demographic history of multiple populations from multidimensional SNP frequency data. *Plos Genet* 2009; **5**:e1000695.

34 Hwang DG, Green P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A* 2004; **101**:13994-14001.

35 Price AL, Tandon A, Patterson N *et al.* Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *Plos Genet* 2009; **5**:e1000519.

36 Auton A, Rui Li Y, Kidd J *et al.* Genetic recombination is targeted towards gene promoter regions in dogs. *Plos Genet* 2013; **9**:e1003984.

37 Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**:44-57.

38 Wang GD, Zhai W, Yang HC, *et al.* The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat Commun* 2013; **4**:1860.