nature COMMUNICATIONS

ARTICLE

Received 24 Jul 2012 | Accepted 27 Mar 2013 | Published 14 May 2013

DOI: 10.1038/ncomms2814

The genomics of selection in dogs and the parallel evolution between dogs and humans

Guo-dong Wang[1,*], Weiwei Zhai[2,*], He-chuan Yang[1,3], Ruo-xi Fan[4], Xue Cao[1], Li Zhong[4], Lu Wang[4], Fei Liu[1], Hong Wu[4], Lu-guang Cheng[5], Andrei D. Poyarkov[6], Nikolai A. Poyarkov JR[7], Shu-sheng Tang[5], Wen-ming Zhao[2], Yun Gao[1], Xue-mei Lv[2], David M. Irwin[1,8], Peter Savolainen[9], Chung-I Wu[2,10] & Ya-ping Zhang[1,3,4]

The genetic bases of demographic changes and artificial selection underlying domestication are of great interest in evolutionary biology. Here we perform whole-genome sequencing of multiple grey wolves, Chinese indigenous dogs and dogs of diverse breeds. Demographic analysis show that the split between wolves and Chinese indigenous dogs occurred 32,000 years ago and that the subsequent bottlenecks were mild. Therefore, dogs may have been under human selection over a much longer time than previously concluded, based on molecular data, perhaps by initially scavenging with humans. Population genetic analysis identifies a list of genes under positive selection during domestication, which overlaps extensively with the corresponding list of positively selected genes in humans. Parallel evolution is most apparent in genes for digestion and metabolism, neurological process and cancer. Our study, for the first time, draws together humans and dogs in their recent genomic evolution.

[1] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, No. 32 Jiaochang Donglu, Kunming 650223, China. [2] Center for Computational Biology and Laboratory of Disease Genomics and Individualized Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, No. 1 Beichen West Road, Chaoyang District, Beijing 100101, China. [3] Department of Molecular and Cell Biology, School of Life Sciences, University of Science and Technology of China, No. 96 JinZhai Road, Hefei 230026, China. [4] Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, No. 2 Cuihu Beilu, Kunming 650091, China. [5] Kunming Police Dog Base, Ministry of Public Security, Heilongtan, Kunming 650204, China. [6] Severtsov Institute of Ecology and Evolution, Russian Academy of Science, Leninskiy prospect, 33, Moscow 119071, Russia. [7] Department of Vertebrate Zoology, Faculty of Biology, M.V. Lomonosov Moscow State University, 1-12 Leninskie Gory, Moscow 119991, Russia. [8] Department of Laboratory Medicine and Pathobiology, University of Toronto, 1 King's College Circle, Toronto, Ontario, Canada M5S 1A8. [9] School of Biotechnology, KTH Royal Institute of Technology, Science for Life Laboratory, Tomtebodavägen 23A, SE-171 21 Solna, Sweden. [10] Department of Ecology and Evolution, University of Chicago, 1101 E 57th Street, Chicago, Illinois 60637, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Y. -p. Z. (email: zhangyp@mail.kiz.ac.cn) or to C. I. W. (email: ciwu@uchicago.edu).

The genetic basis of animal and plant domestication is an interesting question that is also of practical value[1]. The remarkable diversity in the physical and behavioural traits in dogs is one of the most interesting examples of domestication[2–5]. The evolution of dogs is often depicted as a two-stage process. In the first stage, dogs were domesticated from their wild relatives, possibly the grey wolves of Southeast Asia[6–11]. Ever since then, dogs and humans lived commensally sharing the same living environments and food resources[3]. In the second stage spanning the last few hundred years, intensive breeding programs have created many modern breeds and selected for an assortment of human favourable characters[12]. Many studies have focused on the genetic basis of phenotypic variation in modern breeds[13,14]. In contrast, the genetic changes associated with the transition from wolves to ancestral dogs have received far less attention.

Previous studies using mtDNA and Y chromosome data found that the indigenous dogs from China, together with several dog breeds that originated from Southeast Asia/China (often designated as ancient breeds), have the highest genetic diversity and are the basal lineages connecting to the wild grey wolves[6,9,10]. Whole-genome analysis using single-nucleotide polymorphism (SNP) chips among a large number of canids also revealed a closer relationship between these ancient breeds and the wild wolves[4,5]. Thus, the native dogs of South China are likely the most primitive form of dogs and may represent the product of the first stage of domestication[6,9,10]. Coupled with the availability of the dog genome and the rapid advances in sequencing technology, the study of the native dog populations in China may shed considerable light on the early history of dog domestication.

In this study, we perform whole-genome sequencing of four grey wolves, three Chinese indigenous dogs and three modern breeds, and identify 13.92 million SNPs and 3.02 million small indels. Genome-wide analysis shows a general trend of decreasing diversity from wolves to Chinese indigenous dogs to dog breeds. Demographic analysis reveals a population split between wolves and Chinese indigenous dogs that is as old as 32,000 years ago and that subsequent bottlenecks are rather mild, suggesting that dogs may have been domesticated initially through their scavenging with humans. Population genetic analysis identifies 311 genes under positive selection with strong enrichment in the sexual reproduction, digestion and metabolism, and neurological processes. Interestingly, this list of genes is found to overlap extensively with those that have been selected in humans. The overlap in sets is most apparent for genes involved in digestion and metabolism, neurological process and cancer. Our study, for the first time, reveals striking parallelism in the recent evolution of dogs and humans.

Results

Sample collection and sequencing. Four grey wolves from locations across Eurasia and three Chinese indigenous dogs from Southwest China were collected for this work (Fig. 1a). In addition, we also sequenced dogs from three breeds, one German Shepherd, one Belgium Malinois and one Tibetan Mastiff (Table 1). Of the four grey wolves and six dogs we sequenced, the effective throughput for each individual ranges from 8.92X to 13.56X (Supplementary Table S1). Sanger sequence data for the

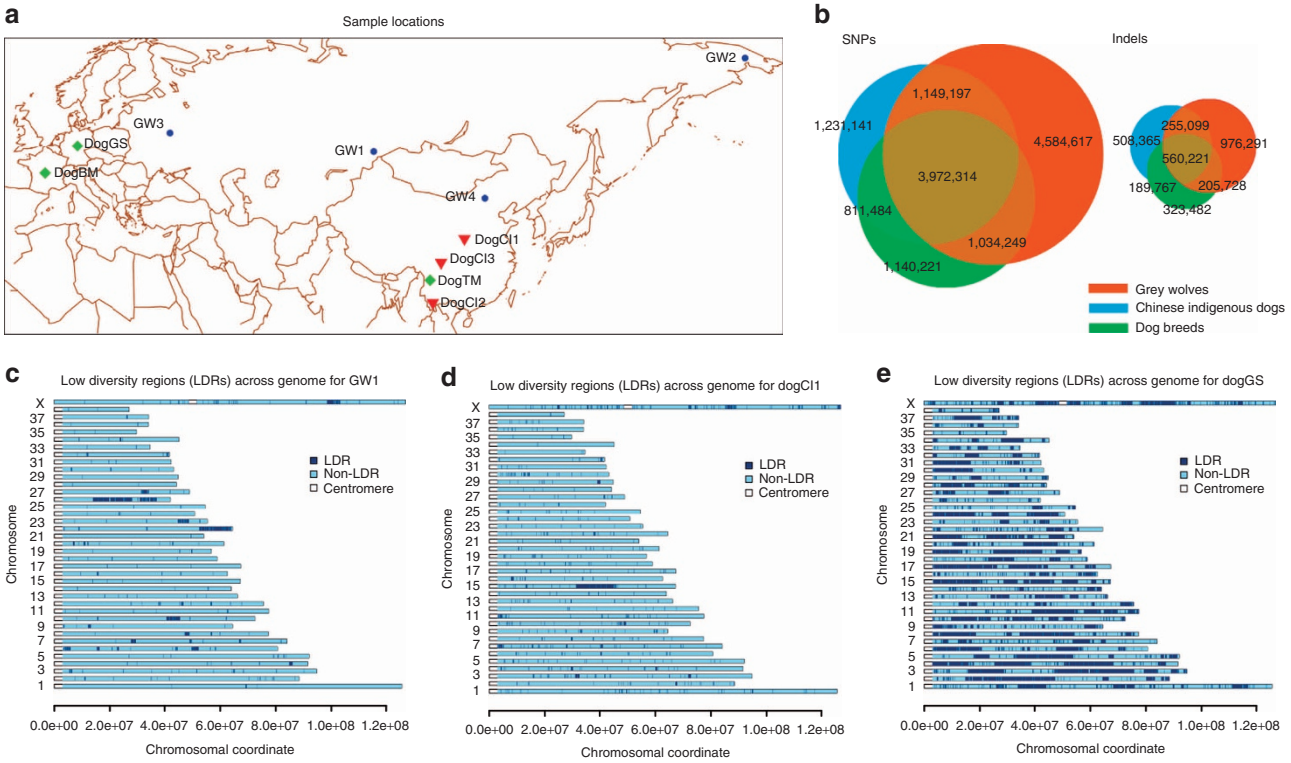

Figure 1 | Sampling and diversity information of the dog and wolf individuals. (**a**) The geographic locations for the four grey wolves (GW1–4), three Chinese indigenous dogs (dogCI1–3), two European dog breeds (dogGS: Germany Shepherd, and dogBM: Belgium Malinois), and one Tibetan Mastiff (dogTM) used in this study are indicated. (**b**) SNP and small indels overlapping between the three different populations, respectively, (wolves, Chinese indigenous dogs and dog breeds). (**c**) Low-diversity regions (LDRs) plotted across the genome for the grey wolf 1. The cutoff value for LDRs is 0.00005. (**d**) LDRs plotted across the genome for the Chinese indigenous dog 1. (**e**) LDRs plotted across the genome for the German shepherd. LDR plots for the other individuals are shown in the Supplementary Fig. S3.

Table 1 | Sample and sequencing throughput for all 11 individuals.

Individual	Sample	Location	Sequence depth (X)	SNP number	Diversity θ (4 Nμ) per kb	Indel number
Wolves						
GW1	Grey wolf	Altai, Russia	11.24	5,564,300	1.52	811,429
GW2	Grey wolf	Chukotka, Russia	8.92	5,276,100	1.32	764,838
GW3	Grey wolf	Bryansk, Russia	11.10	5,472,254	1.44	902,521
GW4	Grey wolf	Inner Mongolia, China	9.61	5,420,479	1.37	906,620
				Mean = 5,433,283	Mean = 1.41	Mean = 846,352
Chinese indigenous dogs						
DogCI1	Chinese indigenous dog	Xi'an, China	13.56	4,361,559	1.29	849,298
DogCI2	Chinese indigenous dog	Simao, China	9.83	3,571,772	0.81	570,177
DogCI3	Chinese indigenous dog	Ya'an, China	10.25	4,225,853	1.07	795,828
				Mean = 4,053,061	Mean = 1.06	Mean = 738,434
Ancient breed						
DogTM	Tibetan Mastiff	Lijiang, China	10.37	4,221,547	1.12	706,664
Modern breeds						
DogsGS	German Shepherd	Germany	9.56	3,448,915	0.67	528,265
DogBM	Belgium Malinois	France	10.11	3,664,565	0.93	687,172
DogREF	Boxer(reference)	na	12.18	1,212,888	0.59	—
				Mean = 2,775,456	Mean = 0.73	Mean = 607,719

Abbreviation: SNP, single-nucleotide polymorphism.
Indels were not called in the reference genome because of the difference in sequencing strategy (that is, Sanger sequencing).

reference Boxer genome was also downloaded from the NCBI trace archive for subsequent analysis[7].

After aligning the short reads to the reference genome, we identified single-nucleotide polymorphisms and small insertions and deletions (length <50) for all the individuals (Details of the data flow are presented in the Supplementary Fig. S1). Across the 11 individual genomes, a total of 13,923,223 SNPs were identified, of which 10,740,377 were found within the 4 wolves, 7,164,136 within the 3 Chinese indigenous dogs and 6,958,268 within the 4 breed dogs (Fig. 1b). A parallel analysis was also conducted for small indels, which yielded a similar pattern with the greatest number found in wolves and least within the breed dogs (Fig. 1b). Through experimental verification, we found current scheme in identifying variants maintains high levels of sensitivity with very limited amount of false positives. For example, we found that the overall false positive rate is less than 5% and for non-singleton polymorphism, genome-wide false negative is less than 10% (Supplementary Note 1).

Genetic diversity and population structure. Using the heterozygous sites called within a diploid organism, we performed a sliding window analysis of the genetic diversity θ (4 Nμ) along the genome for each individual. Interestingly, the genetic diversity shows a decreasing order from wild wolves, to Chinese indigenous dogs and then modern breeds (Table 1). This trend is most evident when we partition the genome into segments of very low diversity and plot this pattern across the genome (Fig. 1c–e). This decreasing order matches with the expectation from a two-stage history where Chinese indigenous dogs represent the groups following the first domestication event.

Using the phased genotypes, linkage disequilibrium, in terms of the correlation coefficient (r^2), was calculated for wolves and the Chinese indigenous dog populations. As seen in Fig. 2a, linkage disequilibrium decreases rapidly for both wolves and the Chinese indigenous dogs. Within distances as short as 5 kb, levels of correlation decrease very rapidly to below 0.2, with this trend being slightly stronger in the wolves than in the Chinese indigenous dogs. The similarity in linkage disequilibrium observed here suggests that a relative weak population bottleneck might have occurred during dog domestication.

Given the genotypes across the genomes, we did Bayesian clustering inferences by partitioning the individuals into $K=2$ and $K=3$ groups. As seen from Fig. 2b, when we try to cluster the individuals into two groups, the first cluster separates all of the grey wolves from the dogs. Interestingly, the Chinese indigenous dogs and the Tibetan Mastiff showed a closer relationship with the wolves. When we tried to partition the sample into three clusters, the analysis started to split the wolves into further groups, likely due to the higher distances within the wolves (Supplementary Fig. S4).

In order to further explore the relative relationships between these individuals, a principle component analysis with all the individuals were carried out. When plotting the first two principle components, dogs and wolves were separated as two distinct groups (Fig. 2c). Interestingly, all of the dogs clustered quite tightly together and distantly from the wolves, however, the Chinese dogs, including the Tibetan Mastiff, were located slightly closer to the wolves (Fig. 2c inset).

Previous studies, using SNP genotyping arrays, have surveyed the global distribution of genetic diversity across a large number of dogs and wolf-like canids. When we combined the sequenced individuals with the 1,191 canids surveyed previously[5], we found that the Chinese native dogs, together with several dog breeds that originated from China/Southeast Asia, are among the first tier of individuals that is closest to the grey wolves (Fig. 2d). In addition, when we compared the Chinese indigenous groups with native dogs from other geographic regions (for example, African village dogs[15]), Chinese indigenous dogs are also found to be much closer to wolves than native dogs from other places surveyed to date (Supplementary Note 2). The close proximity of the Chinese indigenous dogs and breeds originated from Southeast Asia to grey wolves, together with the high genetic diversity observed in the Chinese native dogs, support a Southeast Asia origin for dogs[9,10].

Demographic history. Using joint site frequency spectra generated after polarizing the polymorphisms with an outgroup species

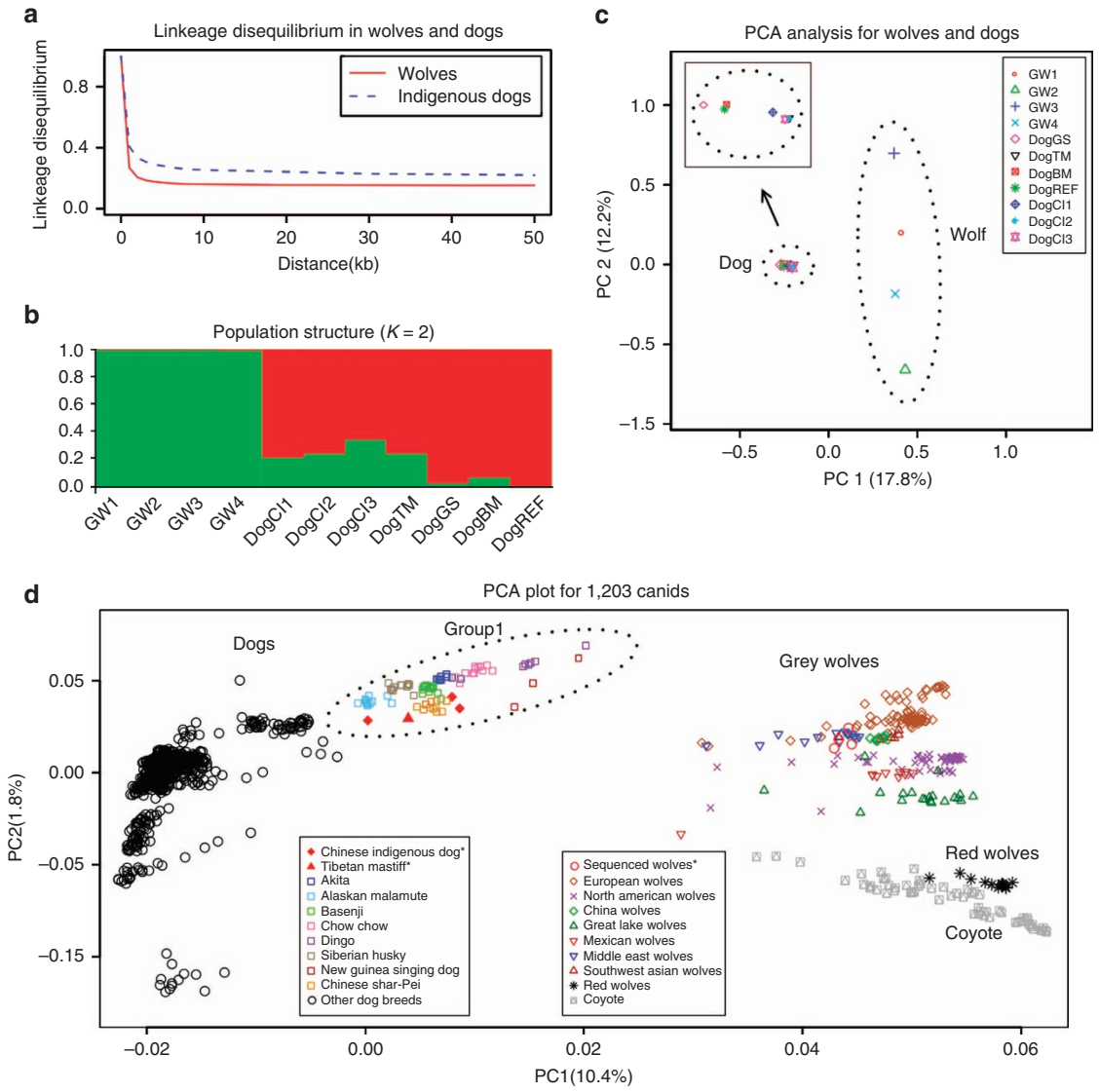

Figure 2 | Population structure and principle component analysis. (**a**) Correlation coefficients (r^2) were calculated for the wolf/dog populations over 50 kb windows. (**b**) Structure analysis on all the individuals with $K = 2$. (**c**) Principle component plots for the first two PCs for all 11 individuals. Inset figure is a zoomed-in version of the dog group. (**d**) Principle component plot for 1203 canids including our data and individuals from a previous study[4]. The group 1 is the cluster of dogs that are closest to grey wolves.

(a red wolf), we inferred the population demographic history under an isolation migration model[16]. As presented in Fig. 3, the effective population size for the wolf was found to have been relatively stable. The inferred effective population size for the extant wolf population is very similar to that inferred for the ancestral population, with the extant population being 94% of the size of the ancestral population. Interestingly, during domestication, the Chinese indigenous dog population experienced a mild bottleneck and the effective population size was reduced to 16% of the ancestral population size. Following the bottleneck, the population size has been steadily increasing to about 32% of that of the ancestral wolf population, which is largely consistent with the mild reduction in genetic diversity and the slight increase in linkage disequilibrium observed in the Chinese native dogs relative to the wolves.

With an assumed mutation rate of 2.2×10^{-9} per year[17] and a generation time of 3 years, the effective population size of dogs at the beginning of the bottleneck is found to be around 8,500 and the effective size of the extant Chinese indigenous dog population to be around 17,000. Compared with other domesticated species, which typically experienced a population shrinkage of several magnitudes[18,19], this level of population size reduction is rather weak.

The population divergence time is estimated to be around 32,000 years ago, which is much older than previous estimates using mtDNA data[9,10] (see discussion). The estimated migration rate is not very large either. The migration rate from wolves to dogs (M_{dw}) is slightly higher than that estimated for the other direction. The estimated migration rate is compatible with our observation that dogs and wolves exist as two rather disjoint clusters in the PCA and structure analysis, and is also in agreement with previous observations that introgressive hybridization between dogs and wild wolves is rare[20]. Behavioural or selective constraints imposed on these two groups might be the limiting factor contributing to the low level of gene flow[20,21].

In order to access the statistical confidence in the estimated parameter values, we performed a non-parametric bootstrap test

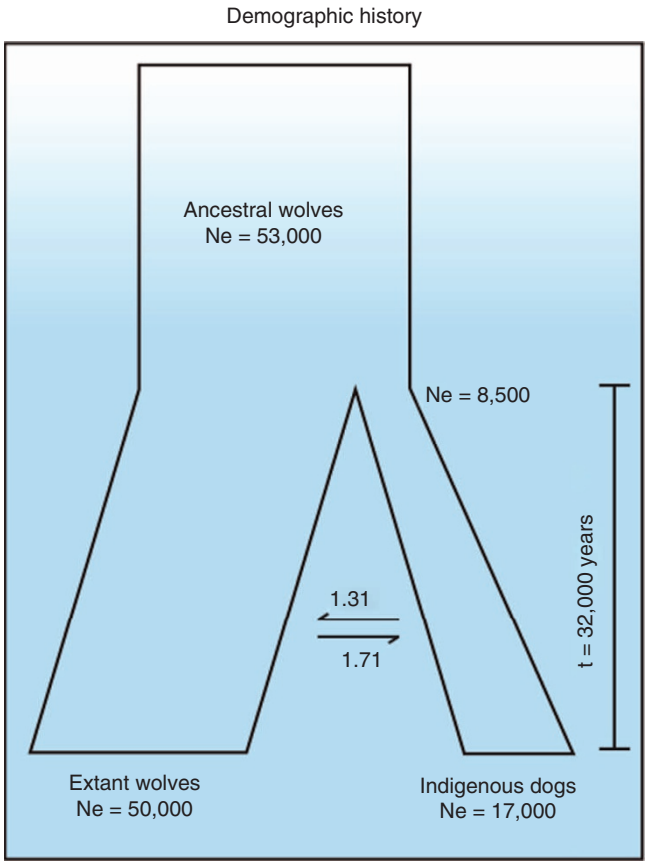

Figure 3 | Inferred demographic history for the wild wolves and the Chinese indigenous dogs. The extent and ancestral population sizes of two species are labelled. The migration rates between two populations are also labelled. As the current wolf's average diversity θ is equal to 0.00141 ($\theta = 4\,N_e\mu$) per kb and current wolves have an effective size that is 94% of the ancestral population, we estimated that the effective population size of the ancestral wolf to be around 53,000.

of the demographic history by resampling the SNPs to generate data sets of the same size with replacement. Under a variety of parameter settings, we found that the estimated values show a similar profile to that presented in Fig. 3 (see Methods as well as Supplementary Note 3), thus, the inferred demographic history shown here is supported with strong statistical confidence.

Putatively selected genes during dog domestication. As selection acting during the first stage of domestication should be shared among all dogs, we thus screened for candidate positively selected genes during dog domestication by looking for regions that show low diversity in all seven dogs and have high divergence between dogs and wolves. To avoid the possibility that a low-diversity segment was inherited from the wolf population, we filtered regions that showed relatively low diversity in wolves.

Using a set of stringent conditions for positive selection, we identified the top 1% of the genome that is expected to be enriched for genes bearing the signature of positive selection. This portion of the genome is distributed across 198 segments carrying a total of 311 genes (Supplementary Note 4, Table S6 and Fig. S11). It is worth pointing out that demographic factors also tend to generate genetic patterns that mimic traces of positive selection[22]. Thus, this candidate list is expected to be enriched for genes responsible for the domestication of the dog. When genes were analysed by their broad classification in the Gene Ontology, three major categories, namely reproduction, digestion and metabolism and neurological process stood out strongly (Table 2).

Genes related to digestion and metabolism are particularly interesting. Multiple GO terms ranging from nutrient transport (for example, lipid) to the regulation of the digestion process (for example, cholesterol) are over-represented. An example of a gene that shows evidence of positive selection is the *MGAM* gene, an important maltase-glucoamylase in the final steps of starch digestion[23]. Along with the recent shared history between dogs and humans, in particular adopting an agricultural based living condition, large changes in the food source for dogs, during the transition from being a carnivore to an omnivore, might have been the driving force for the positive selection for these types of genes[24].

The other interesting GO category is the neurological process. Genes associated with nerve cells themselves (for example, axon) and their connectivity (for example, neuron projection) are among the set of genes that are positively selected. Strong selection on behaviour (for example, reducing aggression) and neurological traits (for example, complex interactions with human beings) is often involved in the first steps of animal domestication[25]. Genes of this class thus might underlie the processes that led to the successful domestication of the dog (see later sections). In addition, quite a few genes involved in sensing local environmental stimuli, for example, sound (*MYO3A*) and smell (*NCAM2* and *OR2F1*), are also on the list of selected genes. Large changes in the environment for dogs during domestication might have driven positive selection in these genes, some of which might reflect relaxed selective constraints on these proteins where loss of the activities of these genes is often adaptive (for example, less is more[26]).

Parallel selection in both human and dog. Humans and dogs both experienced a suite of similar environments in the recent past. Natural selection, driven by convergent environmental pressures, might thus have worked on a similar set of genes in the two genomes. Genome-wide scans for positive selection in humans have been conducted using a wide variety of methods and data sets[27,28]. For example, Akey[22] compiled a collection of human genome regions that had been identified in at least two of nine different genome scans for positive selection[22]. To identify genes that may have been positively selected in parallel, we compared our list of positively selected genes in dogs with that from humans compiled in Akey[22].

Among the orthologous gene pairs between human and dog (a total of 17,661 gene pairs), 1,708 positively selected genes were identified for humans and 233 genes were found for dogs. Comparing these two data sets, 32 genes exist in the overlapping set between the two species (1.4 fold enrichment at a marginal significance of 0.03). Table 3 highlights genes of particular interests, with a full list summarized and presented in Supplementary Note 5 and Table S8.

A group of genes that appear to be under positive selection in both humans and dogs are those involved in digestion and metabolism. For example, two members of the ATP-binding cassette transporters superfamily, *ABCG5* and *ABCG8*, which have pivotal roles in the selective transport of dietary cholesterol[29], were found on both lists. As domestication has lead to drastic changes in the proportions of plant food, relative to animal food, natural selection on these genes in both species is expected due to this shared evolutionary history.

A second groups of genes selected in both species are those involved in neurological processes. An example of an interesting gene is *SLC6A4*, an integral membrane protein that transports the neurotransmitter serotonin[30] and is a target of

Table 2 | Gene ontology analysis of the candidate selected genes.

GO_term	*P*-value	Functional grouping*	Fold enrichment
Biological process			
Reproductive process in a multicellular organism	0.002	SR	2.68
Multicellular organism reproduction	0.002	SR	2.68
Regulation of digestive system process	0.006	DM	25.44
Gamete generation	0.010	SR	2.60
Sexual reproduction	0.010	SR	2.44
Macromolecule catabolic process	0.019	DM	1.91
Cell recognition	0.021	General	6.79
Negative regulation of digestive system process	0.021	DM	93.30
Negative regulation of intestinal phytosterol absorption	0.021	DM	93.30
Negative regulation of intestinal cholesterol absorption	0.021	DM	93.30
Cellular macromolecule catabolic process	0.022	DM	1.93
Cell–cell adhesion	0.028	General	2.70
Cellular amino-acid catabolic process	0.036	DM	5.49
Regulation of lipid transport	0.040	DM	9.33
Cell–cell adhesion mediated by integrin	0.042	General	46.65
Amine catabolic process	0.050	DM	4.78
Cellular component			
Cell projection	5.14E-04	General	2.43
Nuclear lumen	0.010	General	1.63
Axon	0.011	NP	3.73
Intracellular organelle lumen	0.014	General	1.52
Organelle lumen	0.019	General	1.49
Neuron projection	0.019	NP	2.48
Membrane-enclosed lumen	0.025	General	1.46
Neuromuscular junction	0.027	NP	11.54
Molecular function			
Endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 3′-phosphomonoesters	1.94E-07	General	26.15
Pancreatic ribonuclease activity	5.02E-07	General	35.03
Endoribonuclease activity, producing 3′-phosphomonoesters	1.30E-06	General	29.50
Ribonuclease activity	6.08E-06	General	11.32
Nuclease activity	4.69E-05	General	5.91
Endonuclease activity	9.26E-05	General	7.47
Endoribonuclease activity	1.35E-04	General	11.92

Abbreviations: DM, digestion and metabolism; GO, gene ontology; NP, neurological process; SR, sexual reproduction.
*There are three major enriched GO categories: SR, DM and NP. GO categories not belonging to the above three categories are classified as General. Only categories with *P*-values less than 5% are shown in this table. A more detailed table is presented in the Supplementary Table S7.

many psychomotor stimulants such as amphetamines and cocaine. Variation in this gene is responsible for a wide range of neurological pathogenic conditions such as aggressive behaviour[31], obsessive-compulsive disorder[32], depression and autism[33,34]. The most striking aspect is compulsive disorders, of which the two species share many similar phenotypes. Most interestingly, dogs respond similarly to the drugs that are used to treat humans (for example, clomipramine hydrochloride, a serotonin-reuptake inhibitor often also used as an anti-depressant drug), suggesting possible common genetic components for these behaviours in humans and dogs. Association studies have found that both the receptor and the downstream metabolite of *SLC6A4* are correlated with aggressive behaviour in dogs[35,36]. The protein coded by *SLC6A4* might underlie the genetic component of many neurological traits in both dogs and humans.

Aside from genes involved in metabolism and neurological processes, the other most prevalent class of genes that overlap between the two species is the cancer related genes. A good example is *MET*, the mesenchymal epithelial transition factor, which is an important proto-oncogene. Abnormal activation of the *MET* pathway leads to a variety of tumours. Many other cancer related genes, including those involved in the cell cycle and apoptotic pathways, are present in our shared list, and are further discussed in Supplementary Note 5.

Discussion

Chinese indigenous dogs might represent the missing link in dog domestication. The dense clustering of all dogs in the PCA plot, the closer distances between grey wolves and Chinese indigenous dogs together with the high genetic diversity within Chinese native dogs support a Southeast Asia origin for domesticated dogs. The whole-genome pattern also agrees with previous studies, based on mtDNA[9,10] and Y chromosome[6] data, as well as whole-genome SNP chip data[4,5], that the Chinese indigenous dogs, and several ancient dog breeds originated from Southeastern Asia, are the basal groups connected to their wild ancestors. The Chinese indigenous dogs are likely one of the early groups that resulted from the first stage of dog domestication and were subsequently the source from which dog breeds were further selected. Thus, the study of the Chinese indigenous dog might hold great promise for illuminating the origin of dogs.

The geographic location for dog domestication presented here, though quite strong, is not fully compatible with earlier studies that used wolves to identify the site of domestication. In

Table 3 | Positively selected genes found in both humans and dogs.

Gene	NS	Description
Genes involved in the digestion and metabolism		
ABCG5	4	ATP-binding cassette, sub-family G, member 5
ABCG8	4	ATP-binding cassette, sub-family G, member 8
PLA2G10	3	Phospholipase A2, group X
PRSS1	6	Protease, serine, 1 (trypsin 1)
Genes involved in the neurological process		
GRM8	2	Glutamate receptor, metabotropic 8
SLC6A4	4	Solute carrier family 6 (neurotransmitter transporter, serotonin), member 4
Genes involved in cancer (including apoptosis and cell cycle)		
BFAR	3	Bifunctional apoptosis regulator
BRE	2	Brain and reproductive organ-expressed (TNFRSF1A modulator)
ITGB1	2	Integrin, beta 1
MET	2	Met proto-oncogene (hepatocyte growth factor receptor)
STK17B	5	Serine/threonine kinase 17b
ZMYM2	6	Zinc finger, MYM-type 2

Abbreviation: NS, number of human studies, which found evidence of positive selection in this gene.
References are listed in the Supplementary Table S8. A full discussion can be found in the Supplementary Note 5.

particular, a previous study has argued for a Middle-Eastern origin of dogs based on the finding that Middle-Eastern wolves, as a group, seem to be closer to dogs than wolves from other places using the 48K SNP chip data[5]. However, the geographic distribution of wild wolves has been greatly affected by human activities in recent history. For example, the ancestral Chinese wolf, from which domesticated dogs may have originated, may already be extinct[9]. In addition, several wolves from Europe and Mexico are closer to dogs than the Middle-Eastern wolves (Fig. 2d), thus, it may be difficult to use patterns from extant wolves to infer domestication location. Nevertheless, it appears to be the case that the patterns revealed from wolves and dogs are not yet fully coherent. Further re-sequencing studies with more samples of wolves and indigenous dogs from around the world should bridge the two pictures drawn with dogs and wolves.

The divergence time between the dog and wolf that we estimated implies a more ancient age for domestication than suggested by previous studies[9,10]. Even though the genetic evidence and fossil records in many parts of the world are still very preliminary[37], archaeological remains of wolf-like canids, with some resemblance to the dog, as old as 30,000 years ago have recently been reported, although their status as dog is debated[38–41]. A deeper divergence and a mild population size reduction during domestication suggest an evolutionary trajectory for dogs that is often called self-domestication[42]. Early wolves might have been domesticated as scavengers that were attracted to live and hunt commensally with humans. With successive adaptive changes, these scavengers became progressively more prone to human custody. In light of this view, the domestication process might have been a continuous dynamic process, where dogs with extensive human contact were derived from these scavengers much latter when humans began to adopt an agricultural life style.

Our study on positive selection in humans and dogs found an extraordinary amount of parallel evolution, which was likely driven by their similar environments. Natural selection acting on genes involved in neurological processes in both species is of particular interest. As domestication is often associated with large increases in population density and crowded living conditions, these 'unfavourable' environments might be the selective pressure that drove the rewiring of both species. Positive selection in neurological pathways, in particular the serotonin system, could be associated with the constant need for reduced aggression stemming from the crowded living environment[43,44]. Moreover, the complex intimate interactions between dogs and humans might have also driven some of the striking parallelism seen in these two species.

Many genes that have undergone positive selection seem to be involved in similar diseases in both species. This could potentially be due to the pleiotropic effects of natural selection driven by the convergent environments (that is, antagonistic pleiotropy)[45]. Studying the genetic basis of these phenotypes among dog groups, in particular the disease associated traits including the many neurological diseases, might shed light on the genetic architecture of these disorders in humans. Parallel evolution happening in two species bestows on us an unprecedented opportunity to understand these traits by studying the evolution and the phenotypes in both species simultaneously. Interestingly, a companion study on hypoxic adaptation in Tibetan dogs also found strong evidence for parallel evolution between humans and dogs, implying that convergent evolution might be much more pervasive than observed here. Our best friend in the animal kingdom might provide us with one of the most enchanting systems for illuminating our understandings of human evolution and disease.

Methods

Sample collection for whole-genome sequencing. The genomes of four grey wolves and six domesticated dogs were sequenced for this study. The four grey wolves are from three different locations in Russia (Bryansk, Altai, and Chukotka), and one place in Inner Mongolia province of China. Of the six domesticated dogs sequenced, three are Chinese indigenous dogs. The Chinese indigenous dogs are the local dog populations that have lived in China for a long period of time and contain many ancestral polymorphisms retained since domestication from their wild ancestors[6,9,10]. The three indigenous dogs are sampled from the provinces of Shanxi, Yunnan and Sichuan. In addition to the three indigenous dogs, we also sequenced one individual each from three different modern dog breeds, the German shepherd, Belgian Malinois and Tibetan Mastiff. These breeds are selected from our sample collection and were chosen to broadly represent the breeds from Europe and Asia. The reference genomic sequence of a boxer was also extracted from the NCBI trace database for this study. Sample locations for the dogs and wolves are shown in Fig. 1a.

Genome sequencing and mapping. Total genomic DNA was extracted from blood samples using the phenol/chloroform method[9]. Whole-genome sequencing of each individual wolf and dog was performed on the Illumina GAIIx platform using a variety of fragment sizes (Supplementary Table S1) and read lengths

resulting in roughly 24.7–57.4 Gb of raw data for each individual. Details of the throughput and read lengths are summarized in Table 1. Paired-end reads were aligned to the dog reference genome assembly CanFam2[7] using the Burrows-Wheeler algorithm implemented in BWA-short[46] with default parameters. Trace data used for assembling the reference boxer genome was downloaded from NCBI and aligned to the reference genome with BWA-SW[47].

SNP calling and genotype estimation. After sequence reads were mapped to the reference genome, mpileup files against the dog reference genome were generated using samtools[48]. After removing duplicated reads with same start/end points, candidate SNP positions were extracted based on the following conditions: (1) SNP quality greater than 20 and (2) no indel in the surrounding $+/-$ 5 bp region[48]. After accumulating SNP positions, total coverage across all individuals was extracted. SNP positions with too low (total coverage <20) or two high coverage (total coverage >185) (possibly bad assembly or repetitive regions) were trimmed to ensure good quality in our final list. Given a SNP position, samtools was used to calculate the probability of each possible genotype conditioned on the observed reads from each individual. The genotype with maximal posterior probability was picked as the genotype for that locus.

Identification of insertions and deletions. The Pindel package[49] was used to curate a list of candidate indel positions together with the Dindel program. First, pair-end reads where one side could be uniquely mapped but not the other were collected. Unmapped reads were then spit and locally aligned according to library insert sizes. High quality candidate positions (single score $s_1 > 3$ and probability score $s_2 > 30$) were then extracted. Candidate positions in addition to information available in Dindel[50] were subsequently analysed where the local multiple sequence alignments were further refined and associated quality scores recomputed. High quality (filter: pass) candidates from the Dindel output were extracted as our final list of small insertions and deletions.

Variant verification with the Sanger method. Randomly selected genome segments covering a total of 382 SNPs from the nuclear genome were validated by traditional Sanger sequence technology in order to evaluate the sensitivity and specificity of the SNV calling strategy. PCR primers were designed based on the coordinates of the SNV locations. After a total of 614 amplifications, the PCR products were purified and sequenced by traditional Sanger sequence technology.

Diversity estimation for each individual along the genome. Watterson's estimate of genetic diversity, which is based on the number of segregating sites were used to estimate the diversity across the genome[51]. For a single individual, the number of segregating sites is equivalent to the number of heterozygous sites in this individual within a segment of interest. The number of heterozygous sites was extracted for those candidate SNPs whose genotypes are most likely heterozygous.

When the number of reads covering a genomic position is not very high, there is a possibility that one of the alleles was missed during sequencing. Watterson's estimate of genetic diversity is modified to explicitly take into account this sampling effect[52]. Given the fact we have no less than 8X coverage of the genome, this correction was helpful, but not substantial.

Phasing and linkage disequilibrium. Given the genotype information across the genome for each individual, the program fastPHASE[53] was used to phase the genotypes into associated haplotypes with default parameters. Linkage disequilibrium was calculated using a custom written python script. We calculated the r^2 statistic, which is the correlation coefficient between two focal loci of interest.

Population structure analysis. SmartPCA program from the EIGENSOFT package (version 4.2)[54] was used to perform principle component analysis on the individuals that we sequenced. In addition, Structure (version 2.3.3)[55] was used to infer the population substructure among the samples. We varied the number for the population grouping parameter K to be 2 or 3 among different runs. SNP sets of different sizes after thinning the total number of SNPs with different distance conditions (that is, 100, 200 and 500 kb) between markers were implemented. The total length of the Markov Chain was set to be 1,100,000, of which 100,000 were burn-in steps.

Population demographic history. We inferred the population demographic history using methods implemented in the package ∂a∂i (version 1.60), which is based on the joint site frequency spectra between multiple populations[16]. Site frequency spectra is first extracted from our genotyping data and then polarized using a red wolf (an outgroup species) that we sequenced in a separate study. To avoid biases in the coding regions, only SNPs in the noncoding parts of the genome more than 5 kb from any coding region were extracted. Non-parametric bootstrapping was done by resampling (with replacement) the same number of SNPs from the total pool of SNPs.

We assumed that the mutation rate per year is 2.2×10^{-9} per year (ref. 17) and that the generation time is 3 years, thus the mutation rate per generation is 6.6×10^{-9} per generation. Using the genetic diversity θ ($4N_e\mu$) estimated across the genome and the mutation rate per generation, we can get a hold of the effective population size for the extant wolf population. Using the relative sizes of different populations (Fig. 3) inferred from the demographic inference, we can calculate the population sizes of the other populations. The divergence time is calculated by combining the information from ∂a∂i and the population size estimates. In particular, the divergence time (τ) from ∂a∂i is measured in $2N_e$ generations. The divergence time in years will be calculated as $2N_e\tau \times 3$.

In the demographic analysis, we were setting the possible range of time of domestication to be between 0 and 0.3 (equivalence of 100,000 years, that is, before modern human's migration out of Africa). In the bootstrap analysis, time spans of much larger range were also explored. In replicates where the estimated divergence time was far beyond the possible domestication time (that is, 250,000 years ago or further), those estimates were removed from the final results. This is equivalent to putting a hard bound on possible range of parameter estimates.

Orthologous gene pair and enrichment analysis. Gene orthologous relationship between human and dog was downloaded from Ensembl database (www.ensembl.org). In the enrichment analysis, the proportions of positively selected gene in two species were first computed (denoted as p_1 and p_2). The *P*-value was calculated as the proportion of simulated data sets that have equal or higher number of overlapped genes than the observed count. The simulation was done by randomly picking the same proportion of genes out of the total gene list, assuming independence among the two sets in human and dogs.

Fst calculation and potential hitchhiking regions. We used Weir and Cockerham[56] method to calculate Fst between wolf and dog populations using the inferred genotypes. After calculating the genome-wide diversity for each individual, the species specific mean diversities were calculated as the arithmetic mean across the seven individuals for the dog and the four individuals for wolves. Candidate hitchhiking regions were identified using three major criteria: (1) focal regions show reduced genetic diversity in the dog population (the bottom 5% quantile from the dog mean genome-wide distribution), (2) segments are not low-diversity regions in wolf (the bottom 20% quantile from the wolf mean genome-wide distribution), (3) there is a high divergence between the dog and wolf populations (we used top 95% quantile in the Fst distribution as the cutoff).

Gene ontology. Gene ontology enrichment test is performed using the Database for annotation, visualization and integrated discovery (DAVID)[57]. Associated transcript IDs were extracted from the Ensembl annotation.

References

1. Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature* **418,** 700–707 (2002).
2. Darwin, C. *The Variation of Animals and Plants under Domestication* (1868).
3. Serpell, J. *The Domestic Dog: Its Evolution, Behaviour, and Interactions with People.* 284 (Cambridge University Press, 1996).
4. vonHoldt, B. M. *et al.* A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Res.* **21,** 1294–1305 (2011).
5. Vonholdt, B. M. *et al.* Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464,** 898–902 (2010).
6. Ding, Z. L. *et al.* Origins of domestic dog in southern East Asia is supported by analysis of Y-chromosome DNA. *Heredity* **108,** 507–514 (2012).
7. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).
8. Ostrander, E. A. & Wayne, R. K. The canine genome. *Genome Res.* **15,** 1706–1716 (2005).
9. Pang, J. F. *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.* **26,** 2849–2864 (2009).
10. Savolainen, P., Zhang, Y. P., Luo, J., Lundeberg, J. & Leitner, T. Genetic evidence for an East Asian origin of domestic dogs. *Science* **298,** 1610–1613 (2002).
11. Wayne, R. K. & Ostrander, E. A. Lessons learned from the dog genome. *Trends Genet.* **23,** 557–567 (2007).
12. Parker, H. G. *et al.* Genetic structure of the purebred domestic dog. *Science* **304,** 1160–1164 (2004).
13. Candille, S. I. *et al.* A -defensin mutation causes black coat color in domestic dogs. *Science* **318,** 1418–1423 (2007).
14. Sutter, N. B. *et al.* A single IGF1 allele is a major determinant of small size in dogs. *Science* **316,** 112–115 (2007).
15. Boyko, A. R. *et al.* Complex population structure in African village dogs and its implications for inferring dog domestication history. *Proc. Natl Acad. Sci. USA* **106,** 13903–13908 (2009).
16. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5,** e1000695 (2009).

17. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99,** 803–808 (2002).
18. Caicedo, A. L. *et al.* Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3,** 1745–1756 (2007).
19. Lam, H. M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42,** 1053–1059 (2010).
20. Randi, E. & Lucchini, V. Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. *Conserv. Genet.* **3,** 31–45 (2002).
21. Vila, C. & Wayne, R. K. Hybridization between wolves and dogs. *Conserv. Biol.* **13,** 195–198 (1999).
22. Akey, J. M. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* **19,** 711–722 (2009).
23. Naim, H., Sterchi, E. & Lentze, M. Structure, biosynthesis, and glycosylation of human small intestinal maltase-glucoamylase. *J. Biol. Chem.* **263,** 19709–19717 (1988).
24. Larsen, C. S. Biological changes in human populations with agriculture. *Annu. Rev. Anthropol.* **24,** 185–213 (1995).
25. Belyaev, D. K. The Wilhelmine E. Key 1978 invitational lecture. Destabilizing selection as a factor in domestication. *J. Hered.* **70,** 301–308 (1979).
26. Olson, M. V. When less is more: Gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* **64,** 18–23 (1999).
27. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. G. Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* **8,** 857–868 (2007).
28. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312,** 1614–1620 (2006).
29. Lee, M. H. *et al.* Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. *Nat. Genet.* **27,** 79–83 (2001).
30. Gelernter, J., Pakstis, A. J. & Kidd, K. K. Linkage mapping of serotonin transporter protein gene SLC6A4 on chromosome 17. *Hum. Genet.* **95,** 677–680 (1995).
31. Patkar, A. A. *et al.* Serotonin transporter polymorphisms and measures of impulsivity, aggression, and sensation seeking among African-American cocaine-dependent individuals. *Psychiatry Res.* **110,** 103–115 (2002).
32. Di Bella, D., Erzegovesi, S., Cavallini, M. C. & Bellodi, L. Obsessive-compulsive disorder, 5-HTTLPR polymorphism and treatment response. *Pharmacogenomics J.* **2,** 176–181 (2002).
33. Brune, C. W. *et al.* 5-HTTLPR genotype-specific phenotype in children and adolescents with autism. *Am. J. Psychiatry* **163,** 2148–2156 (2006).
34. Bligh-Glover, W. *et al.* The serotonin transporter in the midbrain of suicide victims with major depression. *Biol. Psychiatry* **47,** 1015–1024 (2000).
35. Badino, P. *et al.* Modifications of serotonergic and adrenergic receptor concentrations in the brain of aggressive Canis familiaris. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **139,** 343–350 (2004).
36. Reisner, I. R., Mann, J. J., Stanley, M., Huang, Y. Y. & Houpt, K. A. Comparison of cerebrospinal fluid monoamine metabolite levels in dominant-aggressive and non-aggressive dogs. *Brain Res.* **714,** 57–64 (1996).
37. Larson, G. *et al.* Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proc. Natl Acad. Sci. USA* **109,** 8878–8883 (2012).
38. Germonpré, M. *et al.* Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. *J. Archaeol. Sci.* **36,** 473–490 (2009).
39. Napierala, H. & Uerpmann, H.-P. A 'new' palaeolithic dog from central Europe. *Int. J. Osteoarchaeol.* **22,** 127–137 (2012).
40. Pionnier-Capitan, M. *et al.* New evidence for upper Palaeolithic small domestic dogs in South-Western Europe. *J. Archaeol. Sci.* **38,** 2123–2140 (2011).
41. Wang, X. & Tedford, R. H. *Dogs: Their Fossil Relatives and Evolutionary History* (Columbia University Press, 2008).
42. Coppinger, R. & Coppinger, L. *Dogs: A Startling New Understanding of Canine Origin, Behavior and Evolution.* 1st edn (Scribner, 2001).
43. de Kloet, E. R., Joels, M. & Holsboer, F. Stress and the brain: from adaptation to disease. *Nat. Rev. Neurosci.* **6,** 463–475 (2005).
44. Popova, N. K. From genes to aggressive behavior: the role of serotonergic system. *Bioessays* **28,** 495–503 (2006).
45. Williams, G. C. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* **11,** 398–411 (1957).
46. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
47. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
48. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
49. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).
50. Albers, C. A. *et al.* Dindel: accurate indel calls from short-read data. *Genome Res.* **21,** 961–973 (2011).
51. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7,** 256–276 (1975).
52. Jiang, R., Tavare, S. & Marjoram, P. Population genetic inference from resequencing data. *Genetics* **181,** 187–197 (2009).
53. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78,** 629–644 (2006).
54. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
55. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
56. Weir, B. & Cockerham, C. Estimating F-statistics for the analysis of population structure. *Evolution* **38,** 1358–1370 (1984).
57. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).

Acknowledgements

We thank Thorfinn Korneliussen for his support using the program for estimating the site frequency (PMID:21663684). We thank Robert Wayne and Bridgett vonHoldt for helping with previously published canid genotype data. We also thank Rasmus Nielsen for helpful comments and suggestions. We thank V. Lukarevskiy, A. Gruzdev and T. Sipko for help with the sample collection. This work is supported by grants from National Natural Science Foundation of China (91231108), 973 program (2013CB835200), Bureau of Science and Technology of Yunnan Province, and Chinese Academy of Sciences. Peter Savolainen is a Royal Swedish Academy of Sciences Research Fellow supported by a grant from the Knut and Alice Wallenberg Foundation.

Author contributions

Y.P.Z. and C.I.W. supervised this work. G.D.W. and W.Z. designed the study and analysed data. H.C.Y, L.W. and W.M.Z. helped with the data analysis. R.X.F, X.C., L.Z., F.L., H.W., X.M.L. and Y.G. performed and helped with the experiments. L.G.C., A.D.P., N.A.P and S.S.T. provided samples. G.D.W., W.Z., D.M.I., P.S., C.I.W. and Y.P.Z wrote and revised the manuscript.

Additional information

Accession codes: All short read data have been deposited into the Short Read Archive under the accession number SRA068869.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Wang, G.-d. *et al.* The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat. Commun.* 4:1860 doi: 10.1038/ncomms2814 (2013).