Genomic analysis reveals hypoxia adaption of Tibetan Mastiff by introgression of Grey Wolf from Tibetan Plateau

Benpeng Miao1,2, Zhen Wang#1, Yixue Li#1,3

1 Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P. R. China

2 University of Chinese Academy of Sciences, Beijing, P. R. China

3 Shanghai Center for Bioinformation Technology, Shanghai Industrial Technology Institute, Shanghai, P. R. China

Correspondence should be addressed to Zhen Wang (zwang01@sibs.ac.cn) and Yixue Li (yxli@sibs.ac.cn).

Abstract

The Tibetan Mastiff (TM) is a native of the Tibetan Plateau, which has acquired fast adaption to the extreme highland environment. Recently, the impact of positive selection on the genome of Tibetan Mastiffs has been studied and potential hypoxia-adaptive genes have been identified. However, the origin of the adaptive variants remains unknown. In this study, we investigated the signature of genetic introgression on the adaption of TMs for the highland environment with the genomic data of dogs and wolves from the high and low altitudes. Using the method of 'ABBA/BABA' test, we identified genomic regions of TM possibly introgressed from grey wolves living in the Tibetan Plateau. Several of the regions including the EPAS1 and HBB loci also showed dominant signature of selective sweeps in the TM genome. We validated the introgression of the two loci by excluding the possibility of convergent evolution and ancestral polymorphisms, and examined the haplotypes on all available canid genomes. The estimated time of introgression based on a non-coding region of the EPAS1 locus was consistent with the history of the initial colonization of the Tibetan Plateau by modern humans. Our results demonstrated that adaptive introgression of wolves from the highland plays an important role for the TMs living in the hypoxic environment, which indicated that domestic animals could acquire local adaption quickly by secondary contact with their wild relatives.

Key words: Tibetan Mastiff, grey wolf, introgression, adaption

Introduction

The history of dog domestication is often depicted as that dogs were domesticated from their wild ancestors and then were further selected to form many modern breeds[1-5]. However, the origin and evolution of the domestic dog remains a controversial question. For a conclusion of recent studies, the origin time of domestic dog is about 11,000-16,000 or 27,000–40,000 years ago with four mainly potential origin areas: southern East Asia, central Asia, Middle East and Europe[2,6-9]. In spite of the precise origin areas, the domestic dog should origin from the low altitude areas according to these studies. Since the domestication by humans, dogs have dispersed into new environments during the expansion of human civilization[10]. Among these was the Tibetan Mastiff (TM), a native of the Tibetan Plateau, which has been reported to origin from the domesticated Chinese native dog that resides at lower altitudes[11,12]. According to the previous studies, the human settlement history on Tibetan Plateau dated from about 25,000 years ago[13,14]. Thus the TM have adapted to high altitude over a relative shorter time, compared with highland wild animals, such as yak, Tibetan antelope, snow leopard and wild boar, of which the whole genome have been sequenced[15-18]. The Tibetan Plateau is an extreme environment with low atmospheric oxygen pressure, cold climate and limited resources[19], and the TM has acquired various adaptive traits to the highland such as a thicker coat and a lower hemoglobin level than the low-altitude dogs[12,20]. With the whole-genome sequencing technology, many recent studies focused on the genetic basis of the adaptive traits[12,21-24]. Intriguingly, some genes with signatures of positive selection in the TM genome have been reported including EPAS1, which played a central role in the hypoxic responses and contributed to the adaption of Tibetans[21,23,25-28]. However, how the TM quickly acquired the adaptive variations on these genes remains unknown.

Introgression is one of the potential sources of genetic variation, which has been discovered in a variety of animal genomes[29-31]. For example, Fan et al. found extensive admixture between domestic dogs and wolves, with up to 25% of Eurasian wolf genomes showing signs of dog ancestry[32]. Ai et al. found a large region on the X chromosome of domestic pigs might be introgressed from an extinct Sus species[33]. Introgression could also result in adaptive evolution[34,35]. One example of such adaptive introgression was observed in butterfly mimicry that closely related Heliconius species exchange protective color-pattern genes promiscuously[36,37].

In addition to the TM, we noted a recent study which proposed that EPAS1 was also one of the genes that underwent positive selection in the Tibetan grey wolf (TW)[38]. Since grey wolves could have a much more ancient history of occupying habitats on the Tibetan Plateau than dogs that migrated with humans[11,12,38-42], it was likely that the TM acquired the adaptive traits for the highland over a relative shorter time by genetic introgression of the TW. Interestingly, the rationality of the assumption was reported in the Tibetan people, where the EPAS1 haplotype specific to Tibetans was caused by introgression of Denisovan-like DNA[19]. In this study, we integrated the genomic data of dogs and wolves from previous analyses to test the

hypothesis.

Result

SNP data and genetic relationship

The raw SNP data of 29 canids that represented 7 populations were gathered from Dog Genome SNP Database (DoGSD, http://dogsd.big.ac.cn/snp/index.jsp)[43]. Among this collection, three populations lived at the high altitudes, including 2 Tibet grey wolves (TW), 2 Qinghai grey wolves (QW) and 10 Tibetan Mastiffs (TM); three populations came from the low altitudes, containing 2 Xinjiang grey wolves (XW), 2 Inner Mongolia grey wolves (IW) and 10 Yingjiang indigenous dogs (YJ); and a golden jackal (GLJ) was taken as an outgroup[1,21,38,44]. By the genotyping pipeline of the Genome Analysis Toolkit[45], a total of 18,771,565 SNPs and 5,347,179 indels were identified across the 29 individual genomes (Supplementary table 1). The levels of genome-wide genetic diversity were measured by Watterson's θ and pairwise nucleotide diversity π. The low altitude wolves had higher genetic diversity than high altitude wolves, suggesting the small population of highland wolves due to limited sphere of activity. Meanwhile the genetic diversity of low altitude wolves was higher than the diversity of dogs (Figure 1A, Supplementary table 1). These results were correspond with previous studies[1,38].

To explore the genetic relationships among the individuals, we conducted principal component analysis (PCA) based on the whole-genome SNPs. The PCA exhibited a clear picture of the groupings (Figure 1B). The first principal component (PC1) strongly separated GLJ from dogs and wolves. The dogs, highland wolves and lowland wolves were separated with each other significantly along the PC2. When more eigenvectors were incorporated, all four wolf populations were separated, while the two dog populations, YJ and TM, were still clustered together (Figure 1B). To figure out their phylogenetic relationships, we constructed a neighbor-joining (NJ) tree (Figure 1C). The tree showed that dogs and wolves were firstly split into two distinct branches, and within each branch, the high- and low-altitude populations were then separated. The results of PCA and tree analysis suggested that the genetic background of the TM was much closer to dogs than wolves, which was in agreement with the prior knowledge that it was derived from Chinese native dogs residing at lower altitudes[11,12].

Introgression between TM and TW

To examine admixture among these populations, we performed population structure analysis with frappe[46], which estimates individual ancestry and admixture proportions assuming k ancestral populations (Figure 2A). When k=3, the GLJ, dogs and wolves are separated with each other, and admixture between dogs and wolves was observed. When k=4, the low- and high-altitude dogs were split, and admixture mainly occurred between low-altitude dogs and wolves. When k=5, the highland wolves were separated from lowland wolves, with limited admixture between TM and

YJ. However, in any of these cases, we did not found admixture between high-altitude dogs and wolves.

As TW could have a longer history of living in the Tibetan Plateau than TM, we assumed that the introgression of TW may be a genetic source of TM to adapt to the environment. To examine the possibility, we used the method of 'ABBA/BABA' test, which have been successfully performed in the introgression studies of butterflies and Tibetans[19,36,47]. The test was based on phylogenetic tree among the YJ, TM, TW and GLJ (Supplementary figure 1), where the YJ and TM coalesced firstly, and then they coalesced with the TW. The GLJ was treated as outgroup in the tree. Given a site that is bi-allelic in the four populations where the TW is in state B and the outgroup is in state A, there are two possible allelic configurations of YJ-TM-TW-O with respect to possible introgression between TW and either TM or YJ: ABBA and BABA (Figure 2B). In the cases of random sorting of ancestral polymorphisms, the frequency of the two configurations should be approximately equal[37,44,47]. In our study, the null hypothesis was that there was no introgression between the TW and TM or YJ after the divergence of TW from the ancestor of YJ and TM (Patterson's D=0; D>0 for introgression between TM and TW; D<0 for introgression between YJ and TW). Nonetheless, we did not find an excess of ABBA sites and got a negative Patterson's D of -0.0041 (one-tailed Z-test for D=0, p-value=0.0265), indicating no genome-wide introgression between the TM and TW, which was consistent with the frappe result.

Although no significant introgression at the genome-wide level, it was possible that most of the introgressed DNA were lost during the breeding of TMs, and only those that contributed to the adaptive traits could be preserved. To identify the local genomic regions with introgression signatures, we divided the genome into 216 segments of 10Mb and performed the 'ABBA/BABA' test for each segment. We found 14 outliers (D > 0.046) of these segments on different chromosomes, all of which were significantly positive (one-tailed Z-test for D=0, P-value $< 4 \times 10^{-5}$). These results indicated that the introgression between TM and TW could occurred at local genomic regions (Figure 2C, Supplementary figure 2 and Supplementary table 2).

To further explore the possibility of adaptive introgression, we collected the genomic regions with signals of selective sweeps in the TM or TW based on previous studies [21,38] (see Methods). There were 5 and 6 'ABBA/BABA' outliers containing signals of selective sweeps in the TM (Supplementary table 3) and TW (Supplementary table 4), respectively, and 3 of them were overlapped (Figure 2D). Among the overlapped regions, a 200k region (chr10: 48,500,000-48,700,000) and a 300k region (chr21: 28,000,000-28,300,000) showed the most excessive ABBA sites in the TM, which strongly indicated adaptive introgression (one-tailed Z-test for D=0; D=0.185±0.002, p-value<0.001 and D=0.127±0.001, p-value<0.001, respectively). Meanwhile, the TW selective candidates also contained the 200k region, but not the 300k region (Supplementary table 4).

Introgression of EPAS1 locus

To examine the potential introgression at the 200k region in more details, we

analyzed the distribution of almost fixed ABBA/BABA sites and found a dominant peak of shared and fixed ABBA sites combined with a complete lack of BABA sites (Figure 3A). Meanwhile, this region had the most significant Fst values between the TM and YJ (Figure 3A), indicating a strong selective sweep occurred at the region. All those 496 almost fixed ABBA sites were distributed in a 124.3kb locus (chr10:48,564,800-48,689,200), covering the EPAS1 gene as we expected. To validate this result on more populations, we built a haplotype network for the locus including all available canid genomes in DoGSD (see Methods). We observed a clear pattern in which all the haplotypes of dogs and wolves living in the highland were clustered together, which was distinct from the haplotypes from low altitudes (Figure 3B). The long range of ABBA patterns and the broad haplotype network would be very hard to explain in terms of convergent molecular evolution between the highland dogs and wolves[36].

It was pointed out that the 'ABBA/BABA' patterns on local regions could be confounded with ancestral polymorphism interacted with selection[37,48]. Thus we tested the probability that the shared pattern in the EPAS1 locus resulted from ancestral variation instead of introgression. Firstly, we estimated the chance of a 124.3kb haplotype to be maintained without recombination in both TM and TW lineages since the time of divergence. The shorter divergence time would result in larger probability. The probability of maintaining the EPAS1 haplotype was significantly unlikely even for the shortest divergence time (p=0.007) assuming a constant recombination rate of 0.78×10^{-8} per base pair per generation[1,8,44,49] (Figure 3C, see Methods). Secondly, introgression should result in recent splitting of alleles, and lower sequence divergence, compared with the genomic background (Supplementary figure 3A). In contrast, ancestral polymorphism should result in more ancient splitting of alleles and greater sequence divergence (Supplementary figure 3B)[37]. So we compared the average sequence divergence between the TM and TW across the EPAS1 locus versus the genomic background (see Methods). We found significantly reduced sequence divergence at the EPAS1 locus than the background (p-value < 2.2e-16, Figure 3D). All the results supported that the shared pattern in the locus was due to the introgression instead of ancestral polymorphism.

To date the introgression event, we chose a non-coding region (48,639,440-48,689,200) from the EPAS1 locus and constructed a neighbor-joining tree for it (Supplementary figure 4). Differing from the phylogenetic tree on the genome scale, the tree showed that the TM firstly coalesced with the TW, which could be used to represent the time of introgression. We then used the bpp3.2a[50] software to estimate the divergence time based on the non-coding region (see Methods). The estimated introgression time was about 24,000 years ago, with Bayesian 90% credible interval of 8,700-36,000 years ago assuming an average mutation rate per generation of $\mu = 1 \times 10^{-8}$ [1,44]. The time of introgression coincided with previous studies that the initial colonization of the Tibetan Plateau by modern humans occurred during the Upper Paleolithic (10,000-40,000 years ago) rather than Neolithic (4,000-10,000 years ago)[14,51].

Introgression of HBB locus

Next, we examined the potential introgression of the 300k region on the chromosome 21 with the similar analyses to the EPAS1 locus. This region contained a 179kb locus (chr21:28,060,000-28,239,000) with 182 almost fixed ABBA sites, which constituted two close peaks with the length of 74kb and 75kb, respectively. The locus covered the HBB gene cluster and some olfactory receptors, and it also showed significant selective sweeps in the TM compared with YJ (Figure 4A). The roles of HBB in hypoxia adaption were widely reported. The variations of HBB could alter the Hb-O_2 affinity to response the hypoxia, which has been studied in highland animals[22,52-55]. Through the haplotype network of the HBB locus based on all available canid genomes in DoGSD, we confirmed that the haplotypes of highland dogs were much closer to the highland wolves than those from low altitudes (Figure 4B). The shorter the haplotype is, the more likely of maintaining it. The probability of maintaining the two peaks of 74kb and the 75kb throughout the shortest divergence time was also significantly unlikely (0.026 and 0.024, respectively, Figure 4C). The sequence divergence of the 74kb and 75kb peaks in HBB locus between TM and TW was respectively significantly reduced compared with the genomic background (p-value = 4.501e-05 and p-value = 2.064e-08, separately, Figure 4D). These evidences indicated that the two peaks at the HBB locus of the TM was also acquired by adaptive introgression of the TW.

Discussion

As TMs were derived from native Chinese dogs and migrated to the Tibetan Plateau with humans, how they acquired the adaptive traits in a relatively short time remains elusive. The recent study that the Tibetans could achieve this goal by gene flow from the Denisovans highlighted the role of introgression on the issue[19]. In this study, we tested the possibility that the TWs, the wild relatives of dogs but with more ancient history of living in the highland, could provide genetic sources for the adaption of TMs. As a result, we did not found any significant introgression on the genome-wide level with the population structure analysis and the ‘ABBA/BABA’ test. However, when applied the ‘ABBA/BABA’ test to the introgression on the local genomic level, we found that the EPAS1 and HBB loci of the TM, which have been identified to play very important roles in the hypoxia adaption at the highland in previous studies[21-23,52-55], came from the TW. Both the loci also showed signature of selective sweeps in the TM, suggesting an adaptive role of the introgression. Thus, it was plausible that after the admixture of the ancestral breed of TMs and TWs, most of the DNA from TWs were lost during the breeding of TMs, and only those that benefited their livings on the highland were preserved. These results indicated that by secondary contact with wild relatives, the adaptive introgression may be an effective and rapid way for domestication animals to adapt to the new environment, and it was necessary to examine local genomic regions to identify the events.

In the EPAS1 and HBB loci, the long range of almost fixed ABBA sites between

the highland dogs and wolves was hard to explain by convergent evolution. A possible concern may be that the results of ‘ABBA/BABA’ test on local regions should be cautious to interpret for ancestral polymorphism. According to our results, the chance of maintaining so long a haplotype without recombination in both the TM and TW lineages due to ancestral polymorphism was highly unlikely. In addition, the sequence divergences between the TM and TW of the two loci were also significantly reduced compared with the genomic background. Both of the evidences were more consistent with introgression rather than ancestral polymorphism. The introgression pattern could also be observed in other dog breeds living at the highland, as we showed in the haplotype networks based on all available canid genomes. Taking together, introgression was the most plausible explanation for the two loci in the highland dogs, which enabled them to adapt to the hypoxia environment in a relatively short time period.

Methods

SNP calling

The raw per-sample SNP data (.g.vcf files) of 29 canids were downloaded from Dog Genome SNP Database (DoGSD, http://dogsd.big.ac.cn/snp/index.jsp)[43]. SNPs and small indels were called using the GATK (v3.5) pipelines[45]. All per-sample .g.vcf files were combined together by the joint genotyping tool, GenotypeGVCFs, which produced a raw set of joint SNPs and indels. Then the raw joint SNPs and indels were separately filtered by VariantFiltration with the following options:
SNPs: DP >846.0||QD < 2.38 || FS > 29.195 || SOR > 4.449 || MQ < 3.76 || MQRankSum < -9.340e-01 || ReadPosRankSum < -9.210e-01
Indels: DP >707.0 || QD < 2.86 || FS > 28.542 || SOR > 4.406 ||InbreedingCoeff < -0.2884 || ReadPosRankSum < -9.340e-01"

Population genetics analysis

The GCTA (v1.25)[56] was used to perform the principle component analysis on the individuals (Figure 1B). We constructed a neighbor-joining tree (Figure 1A) with MEGA (v5.1)[57] based on the IBS distance matrix between all individuals calculated by PLINK v1.07[58]. The population structure was inferred by frappe (v1.1)[46] with the maximum likelihood method (Figure 2A). The ancestral clusters k was ranged from 3 to 5, with 10000 iterations for each run. The genetic diversity and Fst values were calculated with 100k siding window by the PopGenome package[59] (v2.1.6) in R (v3.1.3).

‘ABBA/BABA’ test

The genome-wide Patterson’s D statistics was calculated by the qpDstat program in the ADMIXTOOLS package (v1.1) [60]. On the local genomic level, the reference genome was divided into 216 segments of 10Mb, and the D statistics was calculated for each segment by the PopGenome package[59] (v2.1.6) in R (v3.1.3). Following Durand et al.[47] , the standard error of the statistics was calculated using a jackknife procedure[61]. A Z-score was then obtained by dividing the value of the D statistic by its standard error. The number of ABBA/BABA sites was counted for each segment with a strict criteria. The ‘A’ sites of GLJ should equal to 2, the ‘B’ sites of TW should be more than 2 and the ‘A/B’ sites of TM or YJ should be more than 16. To count the almost fixed ABBA/BABA sites, the fixed ‘A’ sites of GLJ should equal to 2, the fixed ‘B’ sites of TW should be more than 2 and the fixed ‘A/B’ sites of TM or YJ should be more than 18.

Meanwhile, we collected 132 potential selective regions of highland dogs based on the supplementary table 6 of X. et al. paper and 84 potential hypoxia-related genes (potential selective regions) of highland wolves according to the table_S5 of W. et al. paper[21,38]. If the selective region overlapped with one of 216 segments of 10Mb above, we considered the 10Mb region as a candidate in the dog or wolf group for selective sweeps and thus several selective regions may belong to one same candidate. We did the overlap analyses among introgression outliers, dog and wolf groups.

Haplotype network

We constructed haplotype networks for the EPAS1 (chr10:48,564,800-48,689,200) and HBB loci (chr21:28,060,000-28,239,000), respectively. Besides the 29 canids, SNP data of all available dog and wolf populations for the two regions were collected from the DoGSD (see Supplementary table 5). For the EPAS1 locus, there were a total of 29 SNPs that passed the filter with minor allele count greater than 4 and the r^2 of linkage disequilibrium (LD) greater than 0.3. To limit the number of haplotypes to display, we used the top 50 common haplotypes. We used Beagle[62] (v3.3.2) to phase the SNPs and PopART[63] (v1.7) to build the haplotype network with the method of minimum spanning[64]. For the HBB locus, there were a total of 35 SNPs that passed the filter with minor allele count greater than 6 and the LD greater than 0.3. We built a network with the top 42 common haplotypes.

DNA sequence divergence

The average sequence divergences of the EPAS1 and HBB loci were measured by the diversity of 2 kb sliding window divided by the length of the window with the PopGenome package[59] (v2.1.6). Sliding windows (2kb for chr10 and chr21) across corresponding chromosomes excluding the introgression locus were treated as genomic background. We used the Welch's t-test to compare the divergence between the potential introgression region and background intervals. Based on the two close 'ABBA/BABA' peaks of HBB locus, we divided the locus to a 74kb part (28,060,000~28,134,000) and a 75 kb part (28,164,000~28,239,000) to compare the divergence.

Probability of maintaining a long haplotype from shared ancestral lineage.

We followed a method used by Emilia Huerta-Sa´nchez et al. [19] to calculate the probability of a long haplotype shared by TM and TW due to ancestral polymorphism:

$$p = 1 - pgamma(m,\ shape = 2,\ rate = \frac{1}{L})$$

L is the expected length of a shared ancestral sequence and $\mathrm{L} = 1/(\mathrm{r} \times \mathrm{t})$, where r is the recombination rate per generation per bp, and t is the length of the TM and TW branches since divergence. Assuming an exponential distribution of admixture tracts, the probability of seeing a shared fragment of length$\geq$ m is $\exp(-\mathrm{m}/\mathrm{L})$. However, conditional on observing the TW nucleotide at position j, the expected length is the sum of two exponential random variables with expected lengths L. Therefore it follows a Gamma distribution with shape parameter 2, and rate parameter 1/L.

There were two different ideas about the divergence time between dog and wolf in recent papers. One is about 27,000–40,000 years ago, and the other is about 11,000-16,000 years ago[1,8,44]. The recombination rate of chr10 is 0.78×10^{-8} per generation per bp and the recombination rate of chr21 is 1.02×10^{-8} per generation per bp with a generation time of 3 years[49]. As a shorter divergence time will result in a

larger probability, we used the shortest time, about 11,000 years ago, to calculate the p value (Figure 3C and Figure 4C).

Time of introgression

The divergence time of TM and TW based on a non-coding region (chr10:48,639,440-48,689,200) of the EPAS1 locus (Supplementary figure 4) was used to represent the time of introgression. We used the bpp3.2a[50] software to estimate the divergence time under the multispecies coalescent model on a fixed species phylogeny with the gamma priors:$\theta \sim G(2,\ 2000)$; $\tau \sim G(2{,}2000)$, and the nsample is 3,200,000. The estimated mean divergence time (τ_{MT}) is 8×10^{-5}, and the 90% Bayesian confidence interval ranged from 0.000029 to 0.00012. The real divergence time was inferred by assuming an average mutation rate per generation of $\mu = 1 \times 10^{-8}$and a generation time of three years[1,44].

Figure legends

Figure 1. Genetic relationships based on all autosomal SNPs. (A) Distributions of nucleotide diversity π and Watterson's θ across the genome. Both statistics were calculated for a 100 kb sliding window along the genome. (B) Principle component plots for the 7 populations. The insert figure are the third (PC3) and forth component (PC4) plots. (C) Unrooted neighbor-joining (NJ) tree of individuals. Different colors represent different populations. The scale bar represents the identity-by-state (IBS) score between individuals. The PCA plot and NJ tree showed that TM was closer to YJ than wolves. GLJ, Golden Jackal; YJ, Yingjiang indigenous dogs; TM, Tibetan Mastiff; XW, Xinjiang grey wolf; IW, Inner Mongolia grey wolf; TW, Tibet grey wolf; QW, Qinghai grey wolf.

Figure 2. The analysis of the introgression between TM and TW. (A) Population structures with the number of ancestral clusters k from 3 to 5. Each color represents one ancestral cluster and each vertical bar represents an individual. Although there was admixture between dogs and wolves, there was not admixture between TM and TW. (B) The pattern of ABBA/BABA nucleotide sites in the analysis. GLJ is as outgroup in state A, TW is in state B, and TM and YJ, having the recent common ancestor that diverged from the ancestral population of TW, are in different state either A or B. (C) Manhattan plot of the Patterson's D-statistic for the 216 segments of 10 Mb across all autosomal chromosomes. Each dot represents a segment and the segments from one chromosome have the same color. Several significant outliers with positive Patterson's D values indicated the introgression between TM and TW. (D) Venn plot showing the overlapped regions among 'ABBA/BABA' outliers and selective sweeps in the TM and TW. The regions were all measured by windows of 10 Mb.

Figure 3. The introgression evidences of the EPAS1 locus. (A) The distribution of fixed ABBA/BABA sites and Fst in this locus. This locus has significantly excessive ABBA sites (one-tailed Z-test for D=0; D=0.185±0.002, p-value<0.001) and significant Fst between TM and YJ. (B) A haplotype network based on the top 50 common haplotypes in the locus. The haplotypes were defined from 29 representative SNPs across all available canid genomes in DoGSD. Each circle represents a haplotype and the size is proportional to the number of individuals belonging to that haplotype. The colors represent different populations. Lines connect each haplotype to its most similar relative. Bars represent mutational steps between haplotypes. RSW, Russia grey wolves; ISW, wolf from Israel; CRW, wolf from Croatia; DQ, Diqing indigenous dog; GS, German Shepherd; KM, Kumming indigenous dog; LJ, Lijiang indigenous dog; DG, Dingo from Australia; BJ, Basenji from USA; BM, Belgian Malinois. (C) The probability of maintaining different length of haplotypes assuming the recombination rate of 0.78×10^{-8} per base pair per generation. The shorter the divergence time is, the larger the probability is. Even if the shortest time, 11,000 years, was considered, the probability of maintaining the EPAS1 haplotype was significantly

unlikely (p=0.007). (D) Sequence divergence is reduced between TM and TW in the locus compared with the genomic background (Welch's t-test, p-value < 2.2e-16).

Figure 4. The introgression evidences of the HBB locus. (A) The distribution of fixed ABBA/BABA sites and Fst in the locus. This locus has significantly excessive ABBA sites (one-tailed Z-test for D=0; D=0.127±0.001, p-value<0.001) accompanied by elevated Fst between TM and YJ. (B) A haplotype network among the top 42 common haplotypes in the locus. The haplotypes were defined from 35 representative SNPs across all available canid genomes in DoGSD. RSW, Russia grey wolves; ISW, wolf from Israel; CRW, wolf from Croatia; DQ, Diqing indigenous dog; GS, German Shepherd; KM, Kumming indigenous dog; LJ, Lijiang indigenous dog; DG, Dingo from Australia; BJ, Basenji from USA; BM, Belgian Malinois. (C) The probability of different length of haplotypes to be maintained assuming the recombination rate of 1.02×10^{-8} per base pair per generation. The shorter divergence time would result in larger probability. The probability of maintaining the two peaks of the 74kb and 75kb throughout 11,000 years was significantly unlikely (0.026 and 0.024, separately) (D) Sequence divergence is respectively reduced between TM and TW for the 74kb and 75kb peaks (Welch's t-test, p-value = 4.501e-05 and p-value = 2.064e-08, separately).

Reference

1 Wang, G. D. *et al.* The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nature communications* **4**, 1860, doi:10.1038/ncomms2814 (2013).

2 Wang, G. D. *et al.* Out of southern East Asia: the natural history of domestic dogs across the world. *Cell research* **26**, 21-33, doi:10.1038/cr.2015.147 (2016).

3 Parker, H. G. *et al.* Genetic structure of the purebred domestic dog. *Science* **304**, 1160-1164 (2004).

4 Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).

5 Wayne, R. K. & Ostrander, E. A. Lessons learned from the dog genome. *Trends in Genetics* **23**, 557-567 (2007).

6 Laura M. Shannona, R. H. B., Marta Castelhanoc, Elizabeth Coreyc, Jessica J. Haywarda, Corin McLeand,, Michelle E. Whitea, M. A. S., Baddley A. Anitaf, Nono Ikombe Bondjengog, Jorge Caleroh, Ana Galovi,, Marius Hedimbij, B. I., Rajashree Khalapl, Douglas Lallym, Andrew Mastan, Kyle C. Oliveiraa, Lucía Pérezo,, Julia Randallp, N. M. T., Francisco J. Trujillo-Cornejoo, Carlos Valerianoh, Nathan B. Sutterr, & Rory J. Todhunterc, C. D. B., and Adam R. Boykoa,1. Genetic structure in village dogs reveals a Central Asian domestication origin. *PNAS* **112**, 13639-13664, doi:10.1073/pnas.151621511210.5061/dryad.v9t5h (2015).

7 Pollinger, J. P. *et al.* Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**, 898-902 (2010).

8 Skoglund, P., Ersmark, E., Palkopoulou, E. & Dalen, L. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current biology : CB* **25**, 1515-1519, doi:10.1016/j.cub.2015.04.019 (2015).

9 Thalmann, O. *et al.* Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. *Science* **342**, 871-874 (2013).

10 Germonpré, M. *et al.* Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. *Journal of Archaeological Science* **36**, 473-490 (2009).

11 Li, Y. & Zhang, Y. High genetic diversity of Tibetan Mastiffs revealed by mtDNA sequences. *Chinese Science Bulletin* **57**, 1483-1487 (2012).

12 Li, Y. *et al.* Population variation revealed high-altitude adaptation of Tibetan mastiffs. *Molecular biology and evolution* **31**, 1200-1205, doi:10.1093/molbev/msu070 (2014).

13 Zhao, M. *et al.* Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. *Proceedings of the National Academy of Sciences* **106**, 21230-21235 (2009).

14 Qi, X. *et al.* Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau. *Molecular biology and evolution* **30**, 1761-1778, doi:10.1093/molbev/mst093 (2013).

15 Qiu, Q. *et al.* The yak genome and adaptation to life at high altitude. *Nature genetics* **44**, 946-949 (2012).

16 Ge, R.-L. *et al.* Draft genome sequence of the Tibetan antelope. *Nature communications* **4**, 1858 (2013).

17 Cho, Y. S. *et al.* The tiger genome and comparative analysis with lion and snow leopard genomes. *Nature communications* **4** (2013).

18 Li, M. *et al.* Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nature genetics* **45**, 1431-1438 (2013).

19 Huerta-Sanchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194-197, doi:10.1038/nature13408 (2014).

20 Wen, J. & Yuan, N. Analysis on the parameters of blood in healthy Tibetan mastiffes. *Journal of Qinghai Medical College* **20**, 7-9 (1998).

21 Gou, X. *et al.* Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome research* **24**, 1308-1315, doi:10.1101/gr.171876.113 (2014).

22 Wang, G. D. *et al.* Genetic convergence in the adaptation of dogs and humans to the high-altitude environment of the tibetan plateau. *Genome biology and evolution* **6**, 2122-2128, doi:10.1093/gbe/evu162 (2014).

23 van Patot, M. C. & Gassmann, M. Hypoxia: adapting to high altitude by mutating EPAS-1, the gene encoding HIF-2alpha. *High altitude medicine & biology* **12**, 157-167, doi:10.1089/ham.2010.1099 (2011).

24 Storz, J. F., Scott, G. R. & Cheviron, Z. A. Phenotypic plasticity and genetic adaptation to high-altitude hypoxia in vertebrates. *The Journal of experimental biology* **213**, 4125-4136 (2010).

25 Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78 (2010).

26 Bigham, A. *et al.* Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet* **6**, e1001116 (2010).

27 Beall, C. M. *et al.* Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences* **107**, 11459-11464 (2010).

28 Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72-75 (2010).

29 Anderson, E. Introgressive hybridization. *Introgressive hybridization.* (1949).

30 Dowling, T. E. & Secor, C. L. The role of hybridization and introgression in the diversification of animals. *Annual review of Ecology and Systematics*, 593-619 (1997).

31 Harrison, R. G. & Larson, E. L. Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity* **105**, 795-809 (2014).

32 Fan, Z. *et al.* Worldwide patterns of genomic variation and admixture in gray wolves. *Genome research* **26**, 163-173, doi:10.1101/gr.197517.115 (2016).

33 Ai, H. *et al.* Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature genetics* **47**, 217-225, doi:10.1038/ng.3199 (2015).

34 Hedrick, P. W. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular ecology* **22**, 4606-4618, doi:10.1111/mec.12415 (2013).

35 Castric, V., Bechsgaard, J., Schierup, M. H. & Vekemans, X. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet* **4**, e1000168 (2008).

36 Heliconius Genome, C. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94-98, doi:10.1038/nature11041 (2012).

37 Smith, J. & Kronforst, M. R. Do Heliconius butterfly species exchange mimicry alleles? *Biology*

letters **9**, 20130503, doi:10.1098/rsbl.2013.0503 (2013).

38 Zhang, W. *et al.* Hypoxia adaptations in the grey wolf (Canis lupus chanco) from Qinghai-Tibet Plateau. *PLoS Genet* **10**, e1004466 (2014).

39 Aggarwal, R. K., Ramadevi, J. & Singh, L. Ancient origin and evolution of the Indian wolf: evidence from mitochondrial DNA typing of wolves from Trans-Himalayan region and Pennisular India. *Genome Biology* **4**, P6 (2003).

40 Vilà, C. *et al.* Mitochondrial DNA phylogeography and population history of the grey wolf Canis lupus. *Molecular ecology* **8**, 2089-2103 (1999).

41 Sousa, V. & Hey, J. Understanding the origin of species with genome-scale data: modelling gene flow. *Nature reviews. Genetics* **14**, 404-414, doi:10.1038/nrg3446 (2013).

42 Mech, L. D. & Boitani, L. *Wolves: behavior, ecology, and conservation*. (University of Chicago Press, 2010).

43 Bai, B. *et al.* DoGSD: the dog and wolf genome SNP database. *Nucleic acids research* **43**, D777-D783 (2015).

44 Freedman, A. H. *et al.* Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet* **10**, e1004016, doi:10.1371/journal.pgen.1004016 (2014).

45 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).

46 Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology* **28**, 289-301 (2005).

47 Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular biology and evolution* **28**, 2239-2252 (2011).

48 Klein, J., Sato, A., Nagl, S. & O'hUigín, C. Molecular trans-species polymorphism. *Annual Review of Ecology and Systematics*, 1-C1 (1998).

49 Wong, A. K. *et al.* A comprehensive linkage map of the dog genome. *Genetics* **184**, 595-605, doi:10.1534/genetics.109.106831 (2010).

50 Yang, Z. The BPP program for species tree estimation and species delimitation. *Current Zoology* **61**, 854-865 (2015).

51 Chen, F. *et al.* Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 BP. *Science* **347**, 248-250 (2015).

52 Projecto-Garcia, J. *et al.* Repeated elevational transitions in hemoglobin function during the evolution of Andean hummingbirds. *Proceedings of the National Academy of Sciences* **110**, 20669-20674 (2013).

53 Natarajan, C. *et al.* Epistasis among adaptive mutations in deer mouse hemoglobin. *Science* **340**, 1324-1327 (2013).

54 Storz, J. F., Runck, A. M., Moriyama, H., Weber, R. E. & Fago, A. Genetic differences in hemoglobin function between highland and lowland deer mice. *The Journal of experimental biology* **213**, 2565-2574 (2010).

55 Storz, J. F. *et al.* Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse hemoglobin. *Proceedings of the National Academy of Sciences* **106**, 14450-14455 (2009).

56 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76-82 (2011).

57 Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood,

evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution* **28**, 2731-2739 (2011).

58 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-575 (2007).

59 Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Molecular biology and evolution*, msu136 (2014).

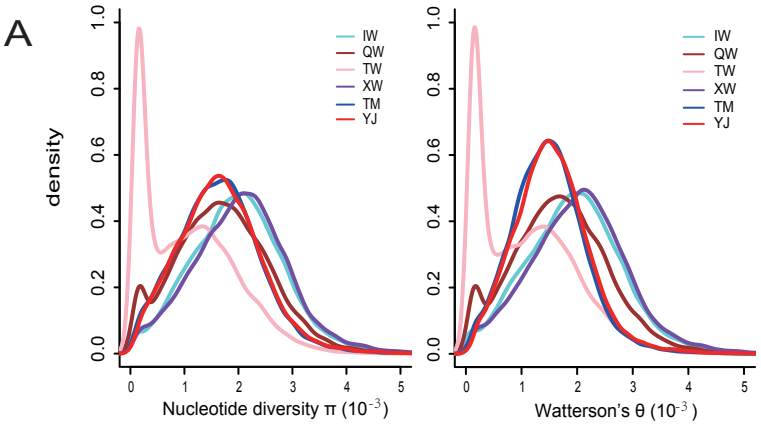
60 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).

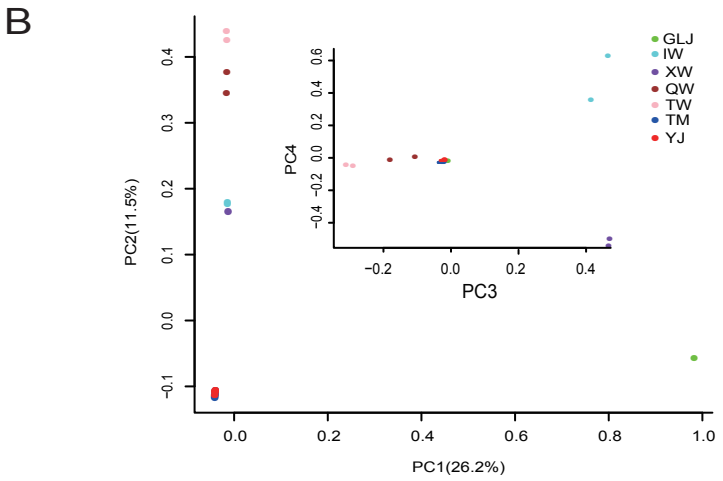
61 Efron, B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* **68**, 589-599 (1981).

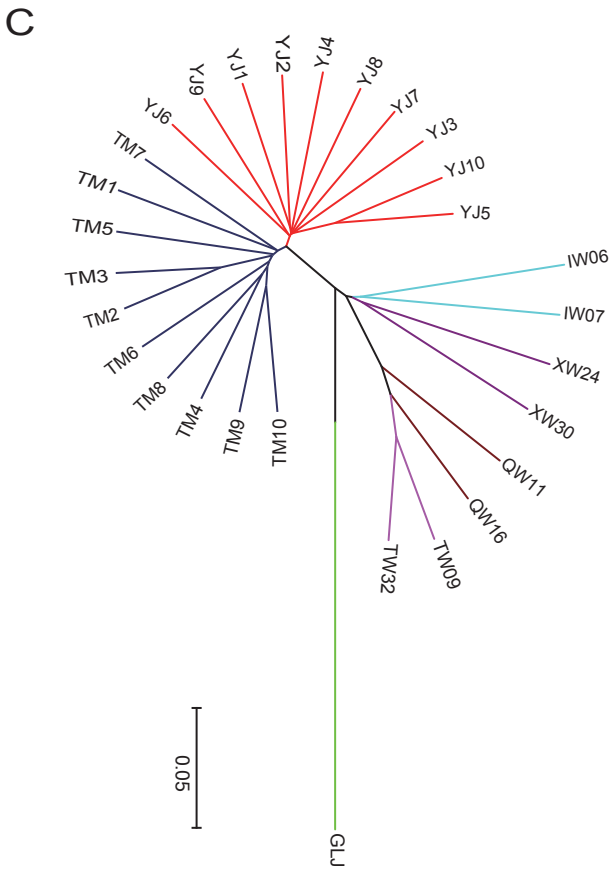
62 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**, 1084-1097 (2007).

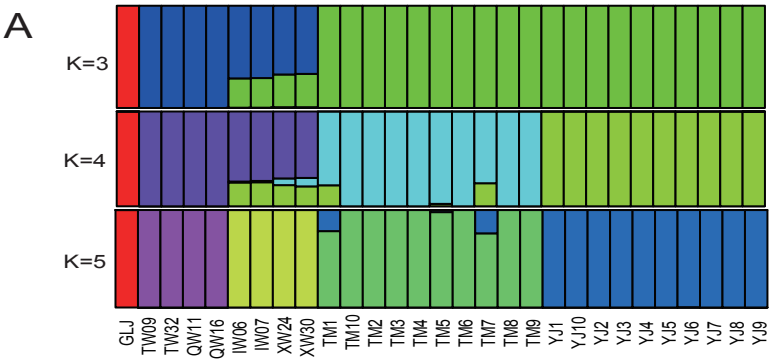
63 Leigh, J. W. & Bryant, D. popart: full‐feature software for haplotype network construction. *Methods in Ecology and Evolution* **6**, 1110-1116 (2015).

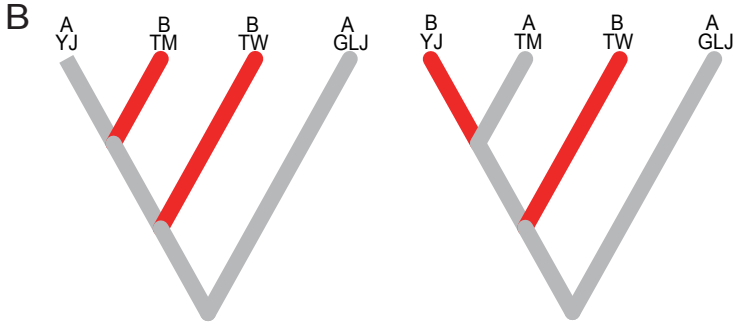
64 Bandelt, H.-J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution* **16**, 37-48 (1999).

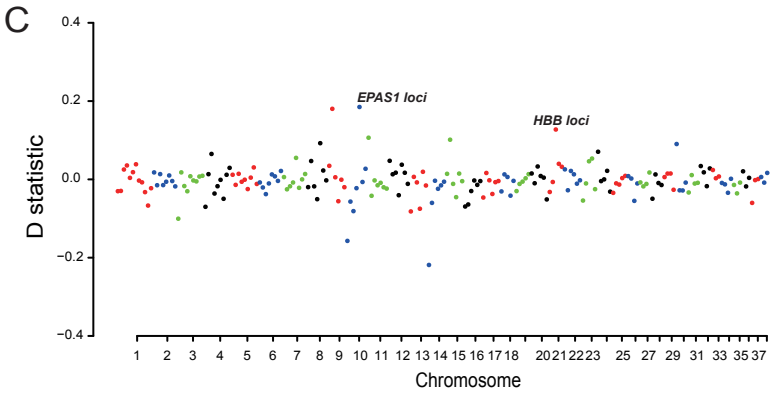

A
density
IW
QW
TW
XW
TM
YJ
Nucleotide diversity π (10^{-3})
Watterson's θ (10^{-3})

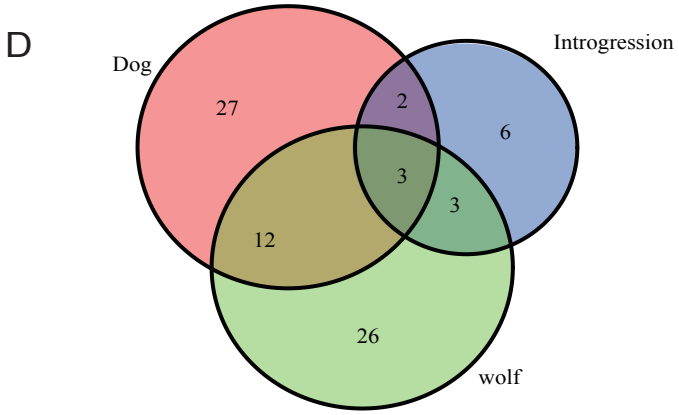

B
GLJ
IW
XW
QW
TW
TM
YJ
PC1(26.2%)
PC2(11.5%)
PC3
PC4

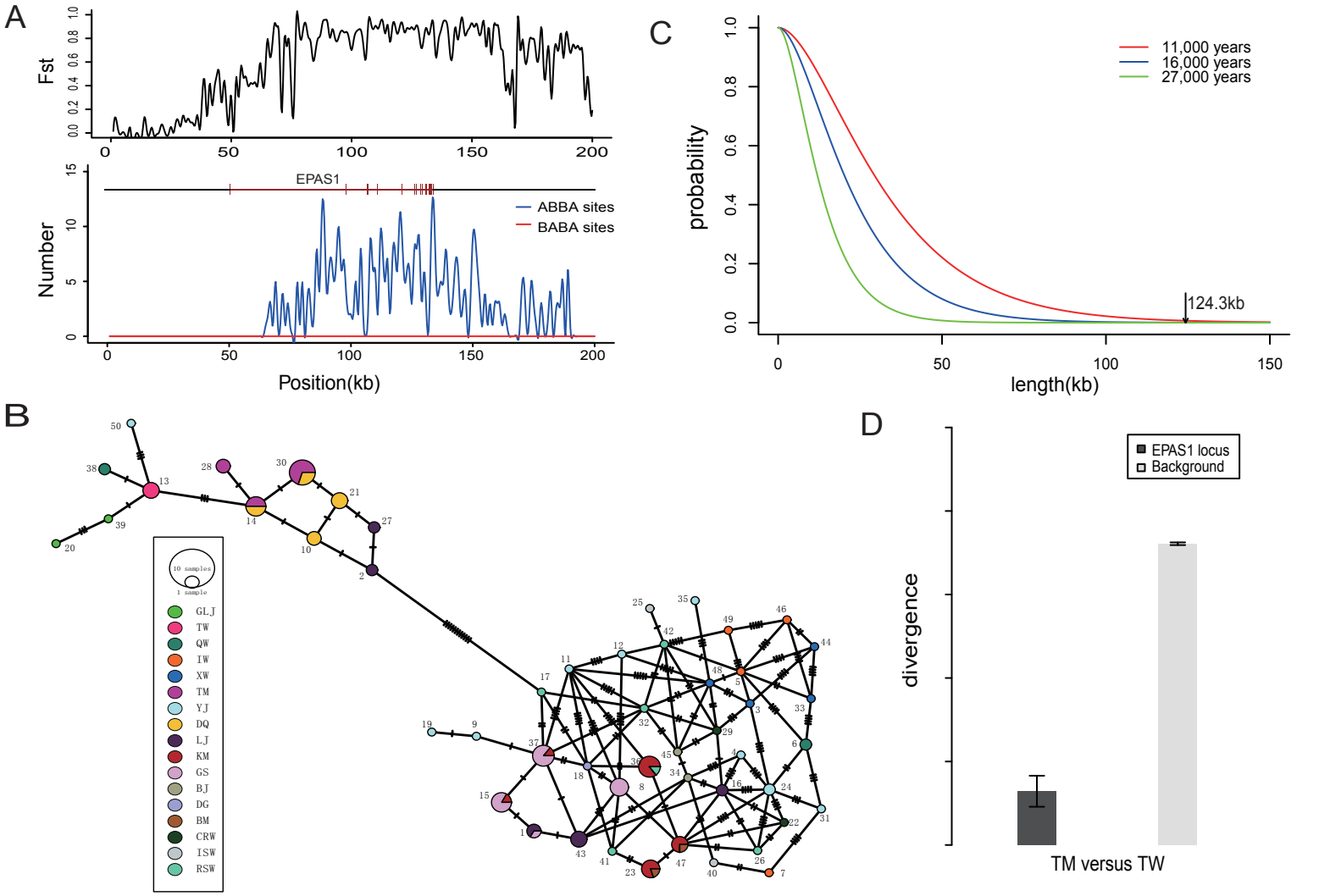

C
YJ9
YJ1
YJ2
YJ4
YJ8
YJ7
YJ3
YJ10
YJ5
YJ6
TM7
TM1
TM5
TM3
TM2
TM6
TM8
TM4
TM9
TM10
IW06
IW07
XW24
XW30
QW11
QW16
TW32
TW09
GLJ
0.05

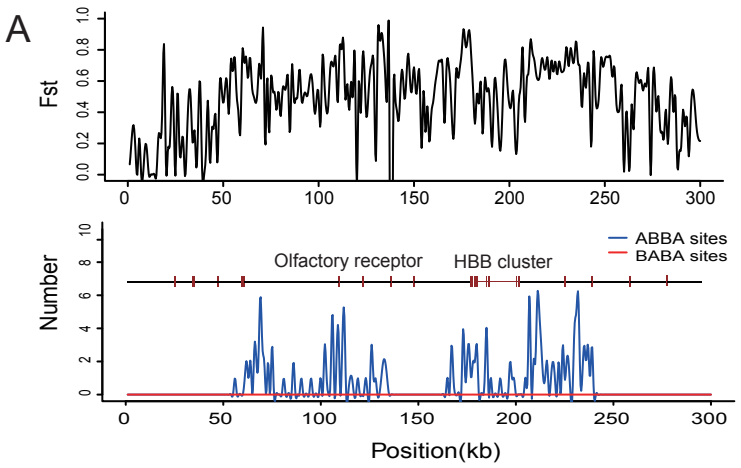

A
K=3
K=4
K=5
GLJ
TW09
TW32
QW11
QW16
IW06
IW07
XW24
XW30
TM1
TM10
TM2
TM3
TM4
TM5
TM6
TM7
TM8
TM9
YJ1
YJ10
YJ2
YJ3
YJ4
YJ5
YJ6
YJ7
YJ8
YJ9

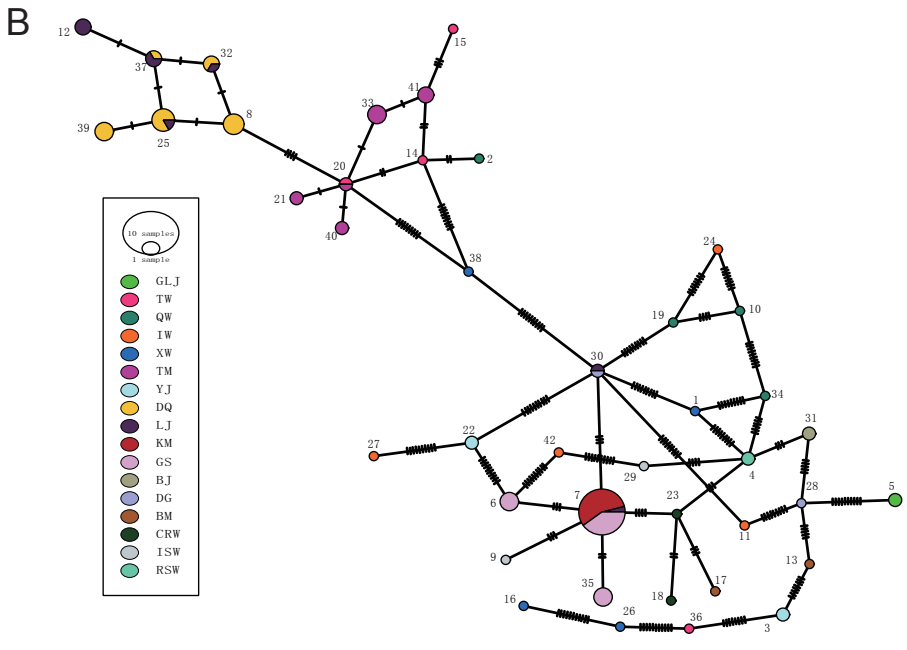

B
A
YJ
B
TM
B
TW
A
GLJ
B
YJ
A
TM
B
TW
A
GLJ

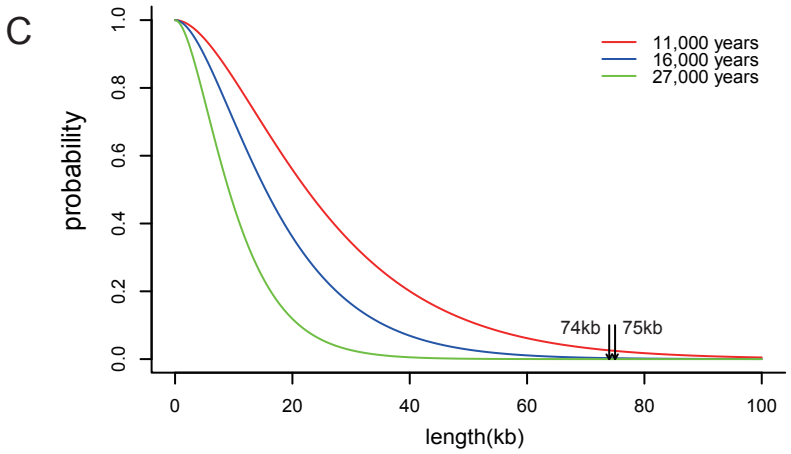

C
D statistic
0.4
0.2
0.0
−0.2
−0.4
EPAS1 loci
HBB loci
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 25 27 29 31 33 35 37
Chromosome

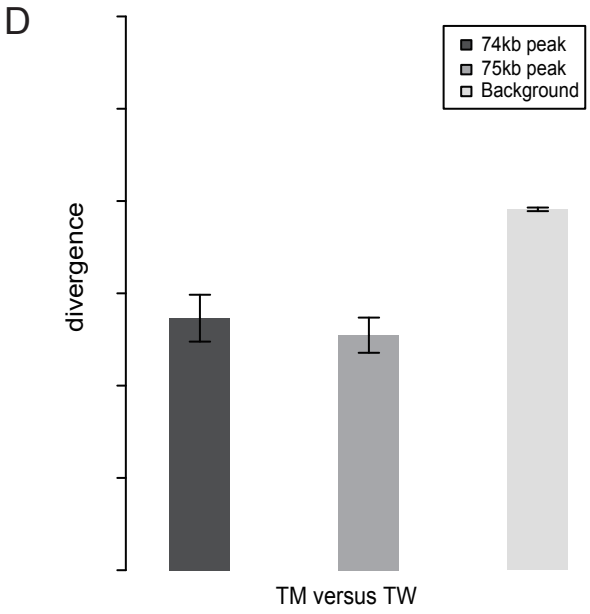

D
Dog
Introgression
wolf
27
2
6
3
3
12
26


A
Fst
0.0 0.2 0.4 0.6 0.8 1.0
0 50 100 150 200
EPAS1
Number
0 5 10 15
ABBA sites
BABA sites
0 50 100 150 200
Position(kb)
C
probability
0.0 0.2 0.4 0.6 0.8 1.0
11,000 years
16,000 years
27,000 years
124.3kb
0 50 100 150
length(kb)
B
10 samples
1 sample
GL.J
TW
QW
IW
XW
TM
YJ
DQ
LJ
KM
GS
BJ
DG
BM
CRW
ISW
RSW
D
EPAS1 locus
Background
divergence
TM versus TW


A
Fst
Number
Olfactory receptor
HBB cluster
ABBA sites
BABA sites
Position(kb)


B
10 samples
1 sample
GL_J
TW
QW
IW
XW
TM
YJ
DQ
LJ
KM
GS
BJ
DG
BM
CRW
ISW
RSW


C
probability
11,000 years
16,000 years
27,000 years
74kb
75kb
length(kb)


D
74kb peak
75kb peak
Background
divergence
TM versus TW