Corrected: Author Correction     OPEN

# Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours

Anna V. Kukekova [1]*, Jennifer L. Johnson[1], Xueyan Xiang[2], Shaohong Feng[2], Shiping Liu[2], Halie M. Rando [1], Anastasiya V. Kharlamova[3], Yury Herbeck[3], Natalya A. Serdyukova[4], Zijun Xiong [2,5], Violetta Beklemischeva[4], Klaus-Peter Koepfli[6,7], Rimma G. Gulevich[3], Anastasiya V. Vladimirova[3], Jessica P. Hekman[1,13], Polina L. Perelman[4,8], Aleksander S. Graphodatsky[4,8], Stephen J. O'Brien[7,9], Xu Wang [10,14], Andrew G. Clark[10], Gregory M. Acland[11], Lyudmila N. Trut[3] and Guojie Zhang [2,5,12]*

**Strains of red fox (*Vulpes vulpes*) with markedly different behavioural phenotypes have been developed in the famous long-term selective breeding programme known as the Russian farm-fox experiment. Here we sequenced and assembled the red fox genome and re-sequenced a subset of foxes from the tame, aggressive and conventional farm-bred populations to identify genomic regions associated with the response to selection for behaviour. Analysis of the re-sequenced genomes identified 103 regions with either significantly decreased heterozygosity in one of the three populations or increased divergence between the populations. A strong positional candidate gene for tame behaviour was highlighted: *SorCS1*, which encodes the main trafficking protein for AMPA glutamate receptors and neurexins and suggests a role for synaptic plasticity in fox domestication. Other regions identified as likely to have been under selection in foxes include genes implicated in human neurological disorders, mouse behaviour and dog domestication. The fox represents a powerful model for the genetic analysis of affiliative and aggressive behaviours that can benefit genetic studies of behaviour in dogs and other mammals, including humans.**

The red fox (*Vulpes vulpes*) and the domestic dog (*Canis familiaris*) are closely related species that only diverged about 10 million years ago within the family Canidae[1]. However, these two species occupy very different ecological niches. The red fox has a geographic range wider than that of any other wild species in the order Carnivora[2] and has even become a common resident of many major cities[3–6]. The dog, on the other hand, has become widespread for a different reason: it was domesticated from the grey wolf at least 15,000 years ago[7,8] and became 'man's best friend'.

There is no evidence that the fox was domesticated historically, although a red fox was found co-buried with humans in a Natufian grave from 14.5–11.6 thousand years ago at a southern Levant site in northern Jordan[9], the same geographic region where the oldest co-burials of humans and dogs are found[10]. The first strong evidence of fox domestication comes instead from the late nineteenth century, when the farm breeding of red foxes for fur began in Prince Edward Island, Canada[11]. Though many animal species are not well-suited to breeding in captivity[12], fox breeding has continued successfully for more than a century[11,13–17]. Conventional farm-bred foxes have adapted to the farm environment, yet their behaviour still clearly differentiates them from dogs because they generally exhibit fear or aggression toward humans.

In 1959, the experimental domestication of farm-bred foxes began at the Institute of Cytology and Genetics of the Russian Academy of Sciences[18–23]. For over 50 generations, foxes were selected for positive responses toward humans, leading to the establishment of a tame strain of foxes that are eager to interact with humans from a very young age[21,24]. Beginning in the late 1960s, a complementary strain of foxes selected for aggressive behavior toward humans was also developed and has proceeded for more than 40 generations[22,23]. A conventional population comparable to the farm-bred founder population of both selected strains has also been maintained but was not subjected to deliberate selection for behaviour. The fox strains have remained outbred during the entire course of the breeding programme, and a strong genetic contribution to the behavioural differences between the tame and

[1]Animal Sciences Department, College of ACES, University of Illinois at Urbana, Champaign, IL, USA. [2]China National Genebank, BGI -Shenzhen, Shenzhen, China. [3]Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. [4]Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. [5]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [6]Smithsonian Conservation Biology Institute, National Zoological Park, Washington DC, USA. [7]Theodosius Dobzhansky Center for Genome Bioinformatics, Saint Petersburg State University, Saint Petersburg, Russia. [8]Novosibirsk State University, Novosibirsk, Russia. [9]Guy Harvey Oceanographic Center, Halmos College of Natural Sciences and Oceanography, Nova Southeastern University, Fort Lauderdale, FL, USA. [10]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. [11]Baker Institute for Animal Health, Cornell University, College of Veterinary Medicine, Ithaca, NY, USA. [12]Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [13]Present address: The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [14]Present address: Department of Pathobiology, Auburn University, Auburn, AL, USA. *e-mail: avk@illinois.edu; guojie.zhang@bio.ku.dk

aggressive strains has been confirmed[20,23,25,26]. Unlike modern dogs, which have been selected for a wide variety of traits, these fox strains were selected solely for behaviour, and the shifts in their behaviour were recent and well documented.

Maximizing the scientific value of these experimental fox populations requires the development of genomic tools for the fox. In contrast to the dog, whose karyotype consists of 38 pairs of acrocentric autosomes in addition to the sex chromosomes, the red fox karyotype comprises 16 pairs of metacentric autosomes, the sex chromosomes and 0–8 supernumerary B chromosomes[27,28]. Synteny between the dog and fox chromosomes has been established but at a low resolution[29–33], hindering identification of the regions in the dog genome that correspond to genomic regions of interest in the fox.

Here, we present the sequence assembly of the red fox genome and a population genetic analysis of whole re-sequenced genomes of foxes from the tame, aggressive and conventional farm-bred populations. Selection on the tame and aggressive strains is likely to have influenced genetic diversity and the fixation of variants across the genome, yielding a robust model for understanding the genetic basis of variation in social behaviour, which is a long-standing problem in evolutionary biology.

## Results

**The red fox genome assembly and annotation.** A male red fox with a standard karyotype (Supplementary Fig. 1) was sequenced to 93.9× coverage using Illumina HiSeq and assembled with SOAPdenovo v.2.04.4[34]. The genome comprises 676,878 scaffolds (scaffold N50 is 11,799,617 bp) and includes 21,418 annotated fox protein coding genes (Supplementary Tables 1,2).

Alignment of the largest 500 scaffolds against the dog genome revealed that 84% of the scaffolds mapped to one dog chromosome, 15% mapped to two or more dog chromosomes and 1% could not be assigned to a position in the dog genome (Supplementary Table 3; Supplementary Fig. 2). Among the scaffolds that mapped to more than one dog chromosome, five mapped to two dog chromosomes that are known to be syntenic to a single fox chromosome[29–32,35].

**Genetic structure of fox populations.** The genomes of 10 foxes from each of the three populations (tame, aggressive and conventional farm-bred) were sequenced with a coverage of ~2.5× , yielding ~75× total genome coverage across all 30 animals (Supplementary Table 4). The 96% of the reads were aligned to the fox scaffolds and the 8,458,133 identified SNPs were retained for subsequent analyses (Supplementary Table 5). The assessment of the relationship among 30 individuals using principal component analysis (PCA), neighbour-joining analysis[36] and STRUCTURE 2.3.4[37–40] indicated the presence of three populations in the data set and less divergence between the conventional and aggressive populations than between the tame and either the conventional or aggressive population (Fig. 1).

**Genomic regions differentiating fox populations.** Simulations were performed in order to support the identification of genomic regions targeted by selection rather than genetic drift (Supplementary Note 1; Supplementary Figs. 3–6). To identify regions of complete or nearly complete fixation within each of the three populations, pooled heterozygosity ($H_p$) was estimated. $H_p$ was calculated for 9,151 windows of 500 kb that were moved along the genome in steps of 250 kb. Population-specific cut-offs corresponding to $P < 0.0001$ revealed 96 low-$H_p$ windows in the tame ($H_p^T$), 60 windows in the aggressive ($H_p^A$) and 14 windows in the conventional population ($H_p^C$) (Fig. 2; Supplementary Tables 6, 7). None of the identified $H_p^T$ windows overlapped with the $H_p^A$ and $H_p^C$ windows, but two $H_p$ windows were significant in both the aggressive and conventional populations. In total, 138 annotated genes were found in $H_p^T$

windows, 159 in $H_p^A$ windows and 51 in $H_p^C$ windows (Supplementary Tables 7,8).

Fixation index ($F_{ST}$) was calculated for the same 9,151 windows used in the $H_p$ analyses to identify regions of extreme differentiation between the fox populations. Only 3% of windows in the analysis of the tame and aggressive populations had $F_{ST}$ values of 0.458 or higher. Using an $F_{ST}$ value of 0.458 as a cut-off for significance (Supplementary Note 2), we identified 275 windows in the analysis of the tame and aggressive populations ($F_{ST}^{TA}$), 106 windows in the analysis of the tame and conventional populations ($F_{ST}^{TC}$) and 1 window in the analysis of the aggressive and conventional populations ($F_{ST}^{AC}$) (Supplementary Table 7; Fig. 2). In total, 650 annotated genes are located in the identified $F_{ST}^{TA}$ windows, 234 in $F_{ST}^{TC}$ windows and three in $F_{ST}^{AC}$ windows. Among the identified $F_{ST}$ windows, 18.7% were also significant in the $H_p$ analysis and 35.7% of significant $H_p$ windows were significant in the $F_{ST}$ analysis (Supplementary Tables 7,8).

PANTHER over-representation analysis[41] (Supplementary Table 9) identified significant enrichment for the GO term "carbohydrate binding" in the $H_p^A$ and $F_{ST}^{TA}$ windows as well as terms related to "clathrin-coated vesicle" and immune response, specifically "cytokine activity" ($H_p^A$) and "interleukin-1 receptor binding" ($F_{ST}^{TA}$). The analysis of the $H_p^T$ windows identified enrichment for "single guanine insertion binding" and "damaged DNA binding". Other terms identified in the $F_{ST}^{TA}$, $F_{ST}^{TC}$ and $F_{ST}^{AC}$ windows are presented in Supplementary Table 9.

More than 80% of genes located in the 9,151 windows were found to be brain-expressed and no over-representation for brain-expressed genes in the significant windows was observed (Supplementary Table 10). Several receptor-coding genes for glutamatergic (*GRIN2B*, *GRM6*), GABAergic (*GABBR1*, *GABRA3*, *GABRQ*) and cholinergic (*CHRM3*, *CHRNA7*) synapses have been identified among the genes located in significant windows (Supplementary Table 11).

To avoid splitting a single sweep across multiple windows, significant $H_p$ windows located close to each other were merged, yielding 30, 19 and 10 combined $H_p$ windows in the tame, aggressive and conventional populations, respectively. Although most of the combined windows comprised one–five windows, two combined $H_p^T$ windows and two combined $H_p^A$ windows were longer than 5 Mb (Supplementary Tables 7,12; Supplementary Fig. 7).

The same rule was used to merge significant $F_{ST}$ windows and produced 57 combined $F_{ST}^{TA}$ windows, 42 combined $F_{ST}^{TC}$ windows and one combined $F_{ST}^{AC}$ window (Supplementary Table 7; Supplementary Fig. 7). Among the six combined $F_{ST}TA$ windows that were 5 Mb or larger, one overlaps completely with a large combined $H_p^T$ window (VVU14, region 86), and two overlap completely (VVU4, region 27) or partially (VVU8, region 46) with large combined $H_p^A$ windows.

The analysis of the positions of all significant windows revealed 103 regions in the fox genome (Supplementary Table 7). The comparison of these regions to the regions associated with domestication and positive selection in dogs[42–45] highlighted 45 fox regions. Three candidate domestication regions (CDR) identified in ref. [44], ten CDRs identified in ref. [42], 22 regions of positive selection in dogs identified in ref. [43] and 38 regions identified in ref. [45] overlap or are located near the genomic regions identified in foxes (Supplementary Table 13). A tentative enrichment of fox regions for CDRs and regions of positive selection in dogs was observed ($P = 0.06$).

Previous genetic mapping studies using cross-bred fox pedigrees identified nine fox behavioural quantitative trait loci (QTL)[26,46]. Comparison of the QTL intervals with the positions of the 103 genomic regions from Supplementary Table 7 revealed 30 regions that overlap with five of the QTL (Supplementary Table 14). The identified overlap is significantly higher ($P < 0.0001$) than expected by chance.
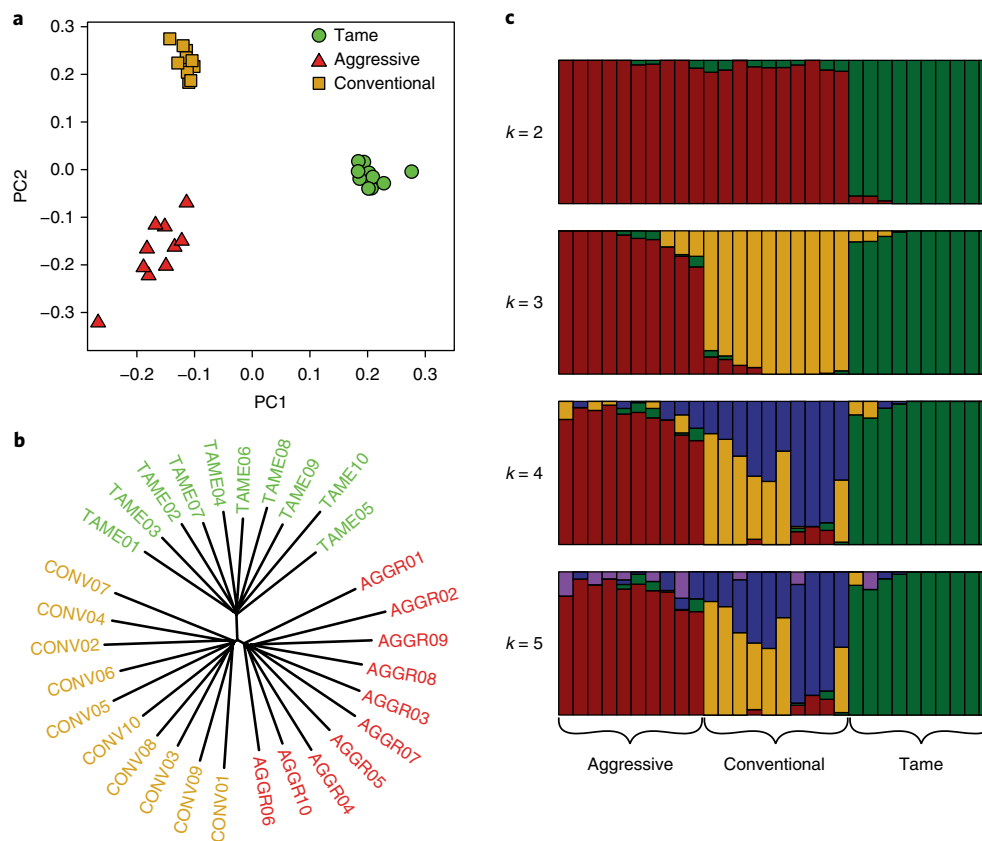
**Fig. 1 | Analyses of the relationship among aggressive, tame and conventional red fox populations. a–c**, PCA (**a**), neighbour-joining tree analysis (**b**) and STRUCTURE analysis (**c**) of the fox populations. All analyses were performed using SNP data for 30 foxes (10 each from the tame, aggressive and conventional populations) whose genomes were re-sequenced. In **a** and **b**, each individual is represented by a data point. In **c**, each individual is represented by a bar that is segmented into colours based on its assignment into inferred clusters given the assumption of *k* populations. The length of the coloured segment is the estimated proportion of the individual's genome belonging to that cluster. Each level of *k* was run four times, but the data shown are from the run giving the highest estimated probability of observing the data. The assumed number of clusters (*k*) is indicated on the *y*-axis. The population origin of individuals is indicated on the *x*-axis.

**Behaviour-related genes.** Identification of genes involved in aggression, sociability and anxiety in foxes is of particular interest because these behaviours are hallmarks of several human behavioural disorders. Analysis of the 971 annotated genes located within significant windows detected 13 genes associated with autism spectrum disorder[47], 13 genes associated with bipolar disorder[48] and three genes located at the border of the Williams–Beuren syndrome deletion in humans[49] (Supplementary Tables 15,16). Six genes from the fox regions have been previously associated with aggressive behaviour in mice[50,51] (Supplementary Table 15). The analysis of significant windows also highlighted fox genes that are not direct orthologues of human genes associated with behavioural disorders or of mouse genes for aggression but that belong to the same gene families and may have similar functions.

Several behaviour-associated genes in significant regions contained alleles corresponding to missense mutations with differences in frequency among the populations (Supplementary Tables 17,18). Two missense mutations in the autism-associated *CACNA1C* gene, CACNA1C-SNP1 (Ile937Thr) and CACNA1C-SNP2 (Thr1875Ile), are located at evolutionarily conserved sites and the CACNA1C-SNP1 was predicted by PolyPhen-2 v.2.2.2r398[52] to be 'possibly damaging' (score: 0.614; sensitivity: 0.87; specificity: 0.91). The derived fox-specific allele for CACNA1C-SNP1 was observed only in the tame population. By contrast, for CACNA1C-SNP2, the derived allele was observed in both the aggressive and conventional populations but not in the tame population (Supplementary Fig. 8).

***SorCS1* is a positional candidate for the QTL on fox chromosome 15.** From the 103 regions of interest identified in the fox genome, the 30 regions that overlapping the behavioural QTL mapped in fox pedigrees[26,46] should represent the most likely targets of selection for behaviour in the tame and aggressive populations (Supplementary Table 14). To test this assumption, we analysed an identified genomic region (region 94 on scaffold 1) that is located on VVU15 within the fox QTL interval (Supplementary Table 14; Fig. 3). Region 94 incudes a single significant $F_{ST}^{TA}$ window that corresponds to part of the *SorCS1* gene (Supplementary Tables 7,18). Although this window did not reach the significance thresholds for $H_p$ in the tame ($H_p^T = 0.20$) and aggressive ($H_p^A = 0.23$) populations, the likelihood of observing such extreme $H_p$ values is low (tame $P < 0.005$; aggressive $P < 0.001$).

The QTL on VVU15 was identified for the behavioural phenotype D.PC1 (a phenotype defined using PCA) that differentiates foxes that continue to solicit an observer's attention after an interaction (higher D.PC1) versus foxes that avoid the observer in the same context (lower D.PC1)[46]. The QTL on VVU15 explains 2.85% of D.PC1 variance in the $F_2$ population[46].

To test whether inheritance of certain *SorCS1* haplotypes predicts variation in D.PC1, we developed 25 short insertion/deletion markers distributed relatively equally across a 5 Mb interval that includes region 94 in the middle (Supplementary Table 19). The markers were genotyped in an additional sample of tame and aggressive foxes and in the $F_2$ pedigrees, whose offspring demonstrate a wide
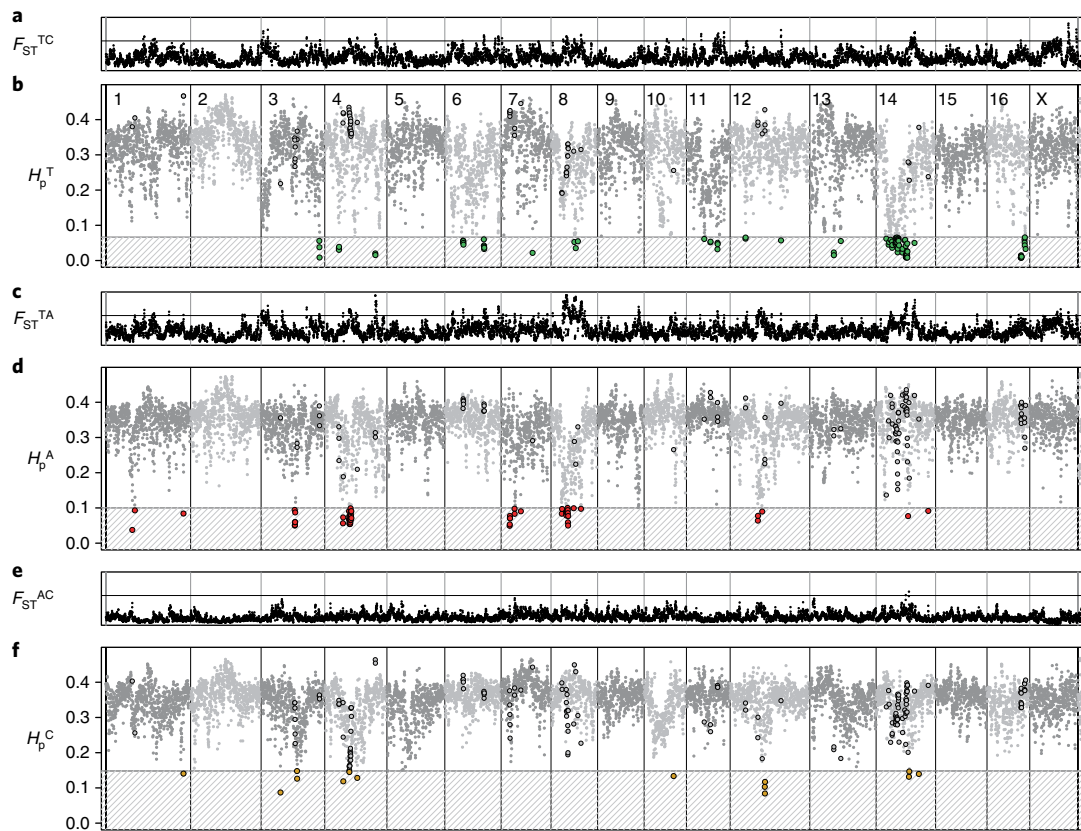
**Fig. 2 | Genome-wide fixation index and pooled heterozygosity analyses across the fox genome. a–f**, Dots on all panels represent windows of 500 kb analysed in both the $F_{ST}$ and $H_p$ analyses. The order of the windows is based on LASTZ mapping of individual windows to the dog genome and a previously published fox–dog synteny map[31] (chimeric scaffolds are split to reflect the most likely location of each window in the fox genome). Vertical lines separate fox chromosomes. Chromosome numbers are indicated in **b**. **a,c,e**, $F_{ST}$ analysis across the fox genome. The horizontal line is $F_{ST} = 0.458$. **b,e,f**, $H_p$ analysis across the fox genome. The grey patterned box and coloured dots indicate the windows that reached significance in that population (Supplementary Table 6; Supplementary Fig. 11). The dots that are outlined in the non-significant zones are windows that reached significance in a different population. **a**, $F_{ST}$ between the tame and conventional populations. **b**, $H_p$ in the tame population. **c**, $F_{ST}$ between tame and aggressive populations. **d**, $H_p$ in the aggressive population. **e**, $F_{ST}$ between aggressive and conventional populations. **f**, $H_p$ in the conventional population.

spectrum of behaviours. We analysed the genotypes of the tame and aggressive foxes to identify the most common haplotypes in the two populations and then tested the effect of the identified haplotypes on behavior in the F2 population.

Haplotype analysis of the tame population identified eight markers located within or in close proximity to the *SorCS1* gene (scaffold 1: 41,647,754–42,312,608 bp) as a single linkage disequilibrium (LD) block located in the middle of the genotyped 5 Mb interval (Supplementary Fig. 9). Within this LD block, Haploview[53] identified one haplotype (*olv*) with a frequency of 60.6% in the tame population that was not observed in the aggressive population, two haplotypes (*trq* and *lav*) that were rare in tame but frequent in the aggressive population, and a fourth haplotype (*pch*) that was found in both populations (Table 1; Fig. 4a; Supplementary Table 20). There were four additional uncommon haplotypes that did not reach 10% frequency in either population. Differences in the behaviour of F2 individuals homozygous for any of the three main haplotypes (*olv*, *trq* and *lav*) were statistically significant (Kruskal–Wallis, $P = 0.03$). F2 individuals that inherited two copies of the tame haplotype (*olv*) had the highest values for D.PC1 (mean: 0.068), while individuals that inherited two copies of one of the common aggressive haplotypes (*lav*) had the lowest values (mean: −0.546) (Table 1; Fig. 4b; Supplementary Fig. 10). A post-hoc Dunn's test with Benjamini–Hochberg[54] correction achieved $P = 0.0142$ for the comparison of the *lav* and *olv* homozygotes (Fig. 4b), while other pair-wise comparisons of homozygotes for the main haplotypes were not signifi-

cant ($P > 0.2$). Analysis of haplotypes for markers located on the left (5′) and right (3′) ends of the genotyped 5 Mb interval did not identify haplotypes with a significant effect on D.PC1 values in the F2 population (Supplementary Note 3). Significant allele frequency differences for *SorCS1* SNPs were also identified in the genotyping-by-sequencing experiment[55] that used a different sample of the tame and aggressive foxes. Taken together, these data strongly suggest that *SorCS1* is a positional candidate for the behavioural QTL on VVU15.

## Discussion

The sequencing and assembly of the red fox genome facilitated the analysis of tame and aggressive populations developed through five decades of selection for behaviour. The population structure analysis clearly differentiated three populations and showed more divergence between the tame and conventional than between the aggressive and conventional populations (Fig. 1). These findings are consistent with the fact that foxes from the conventional farm-bred population were ancestors to both the tame and aggressive strains, but the tame population has been under selection for a decade longer than the aggressive. Secondary introduction of conventional foxes into the aggressive population in the 1990s also led to the reduced divergence observed between these two populations.

Because the tame and aggressive populations were selected solely for their specific behaviours and efforts were made to minimize
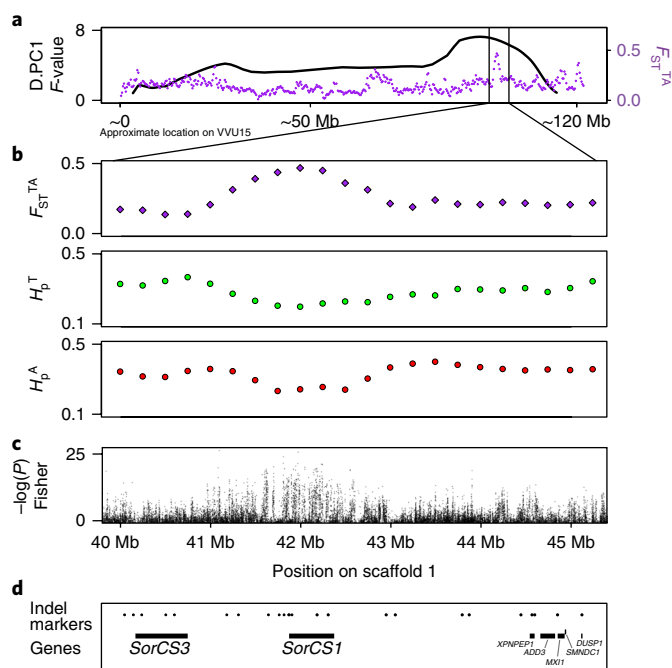
**Fig. 3 | *SorCS1* region on VVU15. a**, The QTL plot for D.PC1 (black line) and the distribution of $F_{ST}$ values (purple) between the tame and aggressive populations. The position on VVU15 is approximated by adding the lengths of the portions of the scaffolds that mapped to the syntenic dog chromosomes (CFA31, CFA28). **b**, $F_{ST}$ and $H_p$ values on fox scaffold 1: 40–45 Mb, the segment that maps to the region on VVU15 (CFA28 segment) that has the QTL for D.PC1 and a significant hit for $F_{ST}$ between the tame and aggressive populations. **c**, Results of the Fisher exact tests for allele frequency differences between the tame and aggressive populations conducted using individual SNPs. The y-axis is the $-\log_{10}$ transformed unadjusted P-values. **d**, Positions of the genes in the region (black lines) and all genotyped markers (black dots).

inbreeding, these populations are well suited to the identification of genomic targets of selection[22,23]. The 103 highlighted regions (Supplementary Table 7) include 30 intervals identified in the tame population and 19 intervals identified in the aggressive population as showing a lower level of heterozygosity than would be expected due to genetic drift (Supplementary Note 1). The longest regions were found on fox chromosomes 4, 8 and 14. Region 27 on VVU4 and region 46 on VVU8 had the lowest heterozygosity in the aggressive population, while regions 79–87 on VVU14 had the lowest heterozygosity in the tame population. The extended length of these selective sweeps is most likely associated with their locations in pericentromeric regions of fox chromosomes where the recombination rate is dramatically reduced[31], but it is also possible that each of these regions harbour several genetic variants associated with selection for behaviour.

Among the 56 regions that contain $F_{ST}^{TA}$ windows, only 18 regions include windows identified in the $H_p^T$ or $H_p^A$ analyses (Supplementary Table 7). The remaining $F_{ST}^{TA}$ windows did not approach fixation in either of the two populations. Similarly, the analyses of allele frequencies in lines of Virginia chickens selected for body weight and in strains of rats selected for behaviour found that many identified loci did not reach fixation in these selected populations[56,57] suggesting that even after 50 generations of selective breeding for complex phenotypes, many loci targeted by selection are retained in a heterozygous state. Mechanisms that could prevent their fixation include non-additive effects, a small effect of a locus on a phenotype and epistasis, all of which were observed in QTL mapping of fox pedigrees[46].

Changes in physiology, morphology and reproduction have also been observed over the course of fox domestication[22,23,58–60]. These by-products of selection for behaviour could be caused by several mechanisms[61,62] including pleiotropy, hitchhiking, random fixation, trade-offs between different biological systems and targeting of genes that have a broad effect on the genome, for example DNA methylation. The GO terms overrepresented in the $H_p^T$ windows (Supplementary Table 9) raise a question of whether selection for tame behaviour was associated with mechanisms involved in regulation of DNA stability. The $H_p^A$ and $F_{ST}^{TA}$ windows showed enrichment for genes associated with the immune response suggesting that immune genes may play an important role in selection of foxes for aggressive behaviour. Previously, it was demonstrated that rats from a strain selected for aggressive behavior showed a higher immune response than rats selected for tameness[63–65]. A link between aggressive behavior and immunological responsiveness was indicated in multiple studies[66–70]. Interestingly, the same set of interleukin genes and receptors that was identified in fox region 52 on VVU8 was also identified on dog chromosome 17 in a region that differentiates dogs from wolves[44] (Supplementary Table 13), suggesting a role of immune genes in both dog and fox domestication.

Comparison of the identified regions to the genomic intervals comprising behavioural QTL[26,46] revealed significant enrichment for QTL-associated regions (Supplementary Table 14). We focused on region 94 and identified *SorCS1* as a strong candidate for a behavioural QTL on VVU15[46] (Supplementary Note 3). *SorCS1* is a member of the Vps10p-domain receptor family, which mediates intracellular protein trafficking and sorting[71]. The major proteins sorted by SorCS1 are neurexin and AMPA glutamate receptors (AMPARs)[72]. Mutations in *SorCS1* and in genes coding neurexins and AMPAR subunits have been found to be associated with several human behavioural disorders[73–81]. The function of SorCS1 as a global regulator of synaptic receptor trafficking supports the role of *SorCS1* in the regulation of behavioural differences between tame and aggressive foxes. These results also demonstrate the advantage of applying a combination of approaches, namely genomic analysis in fox populations and QTL mapping of cross-bred fox pedigrees, to the identification of positional candidate genes for behaviour.

Comparing genes from the fox regions to genes related to autism and bipolar disorder identified 22 shared genes (Supplementary Table 15), including the gene *CACNA1C*, in which we identified non-synonymous mutations at evolutionarily conserved sites (Supplementary Fig. 8). *CACNA1C* plays an important role in dendritic development, neuronal survival, synaptic plasticity, memory and learning[82]. Although no significant enrichment for genes associated with any neurotransmitter system (Supplementary Table 11) was observed, the identification of genes involved in glutamatergic signalling in foxes supports previous reports that genes coding for different types of glutamate receptors are associated with domestication in dogs, cats, and rabbits[83–85]. The identification of genes involved in synapse formation and functioning further supports a role for synaptic plasticity in fox domestication and highlights the fox strains as a model for human behavioural disorders.

There are significant similarities between the behaviour of tame foxes and domestic dogs, and the identified fox regions overlap with canine candidate domestication regions (Supplementary Table 13). In addition to CDRs, previous studies reported an SNP[44] and several transposon indels[86] located in the region syntenic to the Williams–Beuren syndrome in humans as differentiating dogs from wolves. The *POM121* gene reported in the latter study[86] was also identified in the fox region 18 which is approaching fixation in the aggressive population (Supplementary Tables 7,16). Differently sized deletions and inversions in the Williams–Beuren syndrome region can lead to different behavioural phenotypes in humans[87]. Identification of signatures of selection in this region in both dogs and foxes underscores the importance of this region for behaviour in a variety of

**Table 1 | Major *SorCS1* haplotypes**

| Haplotype | Frequency in tame | Frequency in aggressive | Number of homozygotes in $F_2$ | Mean D.PC1 in $F_2$ | Variance D.PC1 in $F_2$ |
|---|---|---|---|---|---|
| *SorCS1* region | | | | | |
| *trq* | 7% | 47% | 15 | −0.29 | 0.64 |
| *lav* | 2% | 37% | 40 | −0.55 | 1.06 |
| *olv* | 61% | 0% | 50 | 0.07 | 0.95 |
| *pch* | 16% | 10% | 7 | −0.53 | 1.23 |
| Left of *SorCS1* | | | | | |
| *re* | 19% | 33% | 35 | −0.14 | 1.17 |
| *gr* | 20% | 31% | 37 | −0.03 | 0.95 |
| *yl* | 40% | 0% | 26 | −0.39 | 1.14 |
| Right of *SorCS1* | | | | | |
| *p* | 17% | 29% | 61 | −0.50 | 1.14 |
| *s* | 51% | 10% | 17 | −0.09 | 0.77 |

The major haplotypes that were found in each of the three regions within the genotyped 5.Mb interval surrounding *SorSC1* (left of *SorCS1*, at *SorCS1* (middle) and right of *SorCS1*). Only haplotypes that reached 10% frequency in at least one of the two populations are shown. The haplotype names in column one are the same names listed in Supplementary Table 20, which contains extended information about the haplotypes found. The frequencies in the tame and aggressive populations are based on the data from Haploview. The number of homozygotes in the $F_2$ population is listed, along with the mean value and variance for D.PC1 in the homozygous individuals. The cumulative distribution of D.PC1 values in $F_2$ homozygotes for the haplotypes in the three regions is shown in Supplementary Fig. 10.
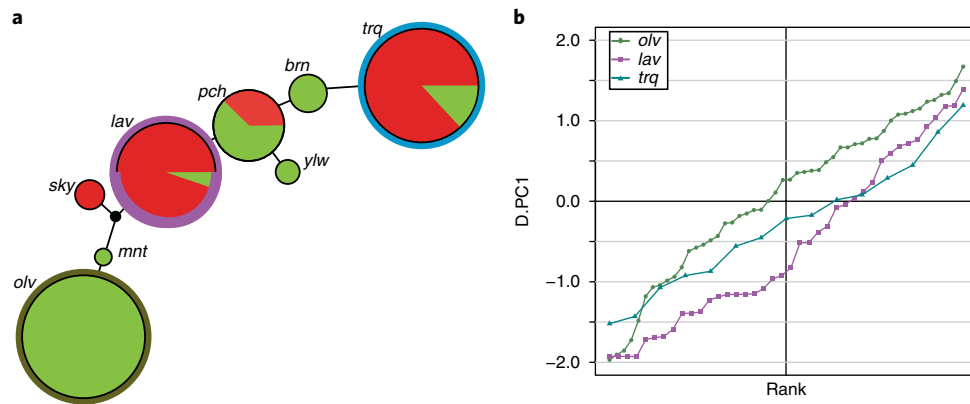


**Fig. 4 | The *SorCS1*-associated haplotypes and their effect on behaviour in the $F_2$ population. a**, A network of the *SorCS1*-associated haplotypes in the tame and aggressive populations. The sizes of the circles are scaled relative to the frequency of the haplotype in both populations combined, and the centre of the circle is coloured to show the relative frequencies of the haplotype in either population (tame being green and aggressive being red). The outer circle of the three major haplotypes (*olv*, *trq* and *lav*) follows the colouring in **b**. The length of the lines between the haplotypes is scaled relative to the number of genotypes for individual markers that differ between the haplotypes, ranging from 1 to 3. The black node is the calculated median vector from the Network 5 run. **b**, Cumulative distributions of the scores for the behavioural phenotype D.PC1 among $F_2$ individuals homozygous for the three main haplotypes: *olv*, *lav* and *trq*. The primary tame haplotype, *olv*, is shown in olive green, and the most common aggressive haplotypes, *lav* and *trq*, are shown in purple and blue, respectively. The points on the lines are individual $F_2$ foxes.

mammalian species. The fact that synergistic analysis of dogs and foxes here implicated shared loci highlights the value of investigating whether comparable behaviours in closely related species are regulated through shared molecular mechanisms and gene networks.

The sequencing and assembly of the fox genome has revealed that a combination of genetic mapping and genome re-sequencing can be used to identify targets of selection for behaviour in the fox strains. Decades of documented selection that have resulted in dramatic differences in the behaviour of tame and aggressive foxes render these populations valuable to genomic studies of behaviour. The fox model expands the spectrum of behaviours that can be studied using animal models and provides insight into the evolution and regulation of mammalian social behaviors.

## Methods

**Fox samples and history of the fox experimental populations.** Samples were collected from adult foxes maintained at the experimental farm of the Institute of Cytology and Genetics (ICG) (Novosibirsk, Russia).

The samples from three populations maintained at the ICG farm were used in this study:

1. The conventional farm-bred population is a standard farm bred population that is outbred and has not been deliberately selected for behaviour. The conventional farm-bred population originated from foxes from eastern Canada[16] where fox farm breeding began in the second part of the nineteenth century.
2. The tame population was developed through selection of conventional farm-bred foxes for a tame response to humans beginning in 1959 at the ICG. The population began with 198 individuals that were selected from several fox farms across the former Soviet Union due to their less aggressive and fearful behaviour towards humans. A description of the selective breeding programme was published previously[20,22,23,88]. Pedigree records were carefully maintained and a significant effort was made to avoid inbreeding throughout the breeding programme. A representative video of tame fox behaviour is available online: https://www.youtube.com/watch?v=vrqOSgEh0fQ
3. The aggressive population was developed by selecting conventional farm-bred foxes for an aggressive response towards humans, beginning in the late 1960s at the ICG. The population started with approximately 150 initial founders, but an additional 70 conventional farm-bred foxes were introduced into

the aggressive population in 1990s. This introduction aimed to increase the population size, which had been reduced shortly after the dissolution of the Soviet Union (1993). A description of the selective breeding programme was published previously[20,22,23,88]. Pedigree records were carefully maintained and a strong effort was made to avoid inbreeding during the entire breeding programme. A representative video of behavior of aggressive foxes is available online: https://www.youtube.com/watch?v=GeAWbLLNesY

*Sample used for whole-genome sequencing.* A blood sample from an $F_1$ male produced by cross-breeding a female from the aggressive strain and a male from the tame strain was used for whole-genome sequencing. DNA from blood was extracted using the phenol–chloroform method[89].

*Samples used for re-sequencing.* Blood samples from 30 individuals, corresponding to 10 from each of the tame, aggressive and conventional farm-bred populations, were collected for re-sequencing. Samples were chosen so as not to share any parents or grandparents, and each population sample included an equal number of males and females (Supplementary Table 4). DNA was extracted using Qiagen Maxi Blood Kits, as per the manufacturer's instructions.

*Samples used for RNA-seq.* Brain samples were collected from 24 male foxes (12 from the tame and 12 from the aggressive populations) into RNAlater and then stored at −80 °C. RNA was extracted from three brain regions: the right basal forebrain, the right prefrontal cortex, and the right part of the hypothalamus. Sequencing was performed on an Illumina HiSeq2000. The basal forebrain and prefrontal cortex samples were sequenced using single-end 50 bp reads, and the hypothalamus samples were sequenced using single-end 100 bp reads. In total, 37.2, 41.3 and 72.6 Gb of data were produced for samples from the basal forebrain, right prefrontal cortex and hypothalamus, respectively. The RNA-seq reads were quality filtered and used for annotation of the fox assembly.

RNA-seq quality filtering included several steps. Data quality, GC content and distribution of sequence length were initially assessed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and then reads were processed with flexbar[90] in two passes: the first to trim adapters, remove low-quality reads and remove reads less than 35 bp in length, and the second to remove polyA tails. Third, reads that mapped to fox mitochondrial DNA sequences from NCBI (accession numbers JN711443.1, GQ374180.1, NC_008434.1 and AM181037.1) using Bowtie2[91,92] were discarded, and finally, any remaining reads that mapped to ribosomal DNA sequences were discarded.

*Samples used for genotyping.* Samples from 64 tame, 70 aggressive, 109 $F_1$ and 537 $F_2$ foxes were used for genotyping. Fox $F_2$ pedigrees were produced by cross-breeding tame and aggressive foxes to produce $F_1$ and then breeding $F_1$ foxes to each other to produce $F_2$ pedigrees. The same set of $F_2$ pedigrees was previously used for QTL mapping[46].

**Sequencing and assembly of the fox genome.** Fox paired-end and mate-pair DNA libraries with nine different insert size lengths (from 170 bp to 20 kb) were constructed (Supplementary Table 1). The libraries were sequenced on an Illumina HiSeq2000, with the short insert size libraries yielding read-lengths of 100 and 150 bp and the long insert size, mate-pair libraries yielding 49 bp ends (Supplementary Table 1). In total, 366 Gb of raw reads were produced. A series of strict filtering steps was performed to remove artificial duplications, adapter contamination and low-quality reads[93]. The program SOAPdenovo v.2.04.4[34] was used for de novo assembly (Supplementary Table 1). Briefly, reads from the short-insert libraries (<2,000 bp) were first assembled into contigs on the basis of $k$-mer overlap information. Then, reads from the long-insert libraries (≥2,000 bp) were aligned onto the contigs to construct scaffolds. Finally, we used the paired-end information to retrieve read-pairs and then performed a local assembly of the collected reads to fill gaps between the scaffolds. The program SSPACE v.2.0[94] was used to extend the pre-assembled scaffolds with reads from all long-insert (2–20 kb) libraries (9 libraries, in total). SSPACE v.2.0 was run with the following parameters: -x 0 -k 5 -n 20. Genome assembly quality was evaluated using GC content and the sequencing depth distribution by mapping all the reads back to reference genome using SOAP2[95].

The fox genome was assembled into 676,878 scaffolds with a total length of 2,495,544,672 bp, contig N50 of 20,012 bp and scaffold N50 of 11,799,617 bp (Supplementary Table 1). The raw reads and the longest 82,429 scaffolds, which are all scaffolds at least 200 bp in size, were deposited in NCBI (BioProject PRJNA378561).

**Annotation of the fox genome.** Fox RNA-seq data, de novo gene prediction and homology with canine and human proteins were used to annotate the protein-coding genes in the fox assembly (Supplementary Table 2).

*Homologue-based prediction.* Protein sequences available for the dog and human from Ensembl release-70 were mapped to the fox genome assembly using TBLASTN (BLASTall 2.2.23) with an $e$-value cutoff $1 \times 10^{-5}$. The aligned sequences were then analysed with GeneWise (v.2.2.0)[96] to search for accurate spliced alignments.

*De novo prediction.* Repetitive sequences were masked in the fox genome assembly using RepeatMasker (v.3.3.0) (http://www.repeatmasker.org/). De novo gene prediction was then performed with AUGUSTUS (v.2.5.5)[97]. The parameters were optimized using the gene models with high GeneWise scores from the homologue-based prediction.

*RNA-Seq prediction.* Filtered RNA-Seq reads from three tissues were aligned against the fox genome assembly using TopHat[98]. The candidate exon regions identified by TopHat were then used by Cufflinks[99] to construct transcripts. Finally, the Cufflinks assemblies for the three tissues were merged using the Cuffmerge option in Cufflinks.

The three gene sets obtained by each of the three approaches (homologue-based prediction, de novo prediction and RNA-Seq prediction) were integrated based on gene structures. Finally, all gene evidence was merged to form a comprehensive and non-redundant gene set. In total, 21,418 protein-coding genes were identified in the fox genome.

*Gene annotation.* In order to assign gene symbols to the fox genes with high confidence, a reciprocal blast method was applied. Fox protein sequences and the dog protein sequences that are located on the dog chromosomes, not chromosome fragments (as downloaded from Ensembl release-73), were analysed with BLASTP in both directions. The BLASTP-aligned results were filtered using an $e$-value cutoff of $1 \times 10^{-5}$, and reciprocal best hit (RBH) pairs were determined using the following condition: for two genes (for example A and B) from the fox gene set and the dog gene set, respectively, they would be accepted as an RBH pair if and only if they were reciprocally each other's top-BLASTP-score hits, meaning there was no gene in the fox gene set with a higher score than A to B, and there was no gene in the dog gene set with a higher score than B to A. This analysis of the fox predicted genes against the dog Ensembl database identified 16,620 dog Ensembl IDs, 14,419 with gene symbols available. Fox protein sequences and human protein sequences on human chromosomes, not chromosome fragments (downloaded from Ensembl release-73), were analysed with BLASTP using the same protocol. This analysis identified 15,826 human Ensembl ID's, all having an associated gene symbol. These 15,826 high confidence gene symbols were assigned to the associated fox genes and were used in downstream analysis.

The 21,418 predicted protein coding genes were compared against several databases to produce a preliminary annotation. Genes were aligned using BLASTP to the SwissProt and TrEMBL databases[100] and assigned to the best match of their alignments. Motifs and domains of genes were determined by InterProScan[101] against protein databases including ProDom, PRINTS, Pfam, SMART, PANTHER and PROSITE. Furthermore, all genes were aligned against the KEGG[102] proteins, and the pathway in which the gene might be involved was derived from the matched genes in KEGG. The fox genome annotaion statistics are presented in Supplementary Table 2.

**Alignment of the fox scaffolds against the dog genome.** The top 500 longest scaffolds (size range: 47,686 bp to 55,683,013 bp), which contain 94% of the fox genome by length, were aligned against the CanFam3.1 assembly (autosomes, mitochondrial DNA and X-chromosome) and the dog Y-chromosome assembly[103] using LAST[104]. Because each scaffold mapped to multiple locations in the dog genome, we sought to identify the dog chromosome(s) to which it was most likely syntenic. For each scaffold, the maximum LAST score corresponding to each dog chromosome was identified. These scores were Z-transformed using the formula $(x_i - \overline{x})/\sigma$, and the dog chromosome(s) with Z-scores significant at $P < 0.05$ to a particular scaffold were considered syntenic to that scaffold (Supplementary Table 3).

To confirm the accuracy of the assignment of each fox scaffold to one or more syntenic positions in the dog, the LAST mapping results were then scanned with a Python script to determine the best hit at each nucleotide along each scaffold. The LAST mapping data were imported into a MySQL database to identify which dog chromosome corresponded to the highest-scoring mapped segment overlapping each nucleotide along the fox scaffold. Regions mapping to an individual chromosome were plotted as lines using MatPlotLib, with the position on the scaffold as the $x$-axis and the position on the dog chromosome as the $y$-axis. Dog chromosomes to which the scaffold mapped robustly are identified in the legend on the plot. Robust mapping was defined as cases where the best mapping score for the scaffold against that chromosome was at least one standard deviation above the average highest score across all chromosomes. This strategy allowed for visualization of the relationship between each scaffold and the dog genome based on this high score alone, and the fact that it showed an overwhelming consensus with the Z-score data supported the assignment of dog syntenic fragments using the first approach (Supplementary Fig. 2).

**Re-sequencing of fox samples from three populations.** DNA samples from 30 foxes (10 foxes from tame, 10 foxes from aggressive, and 10 foxes from conventional farm-bred population) were sequenced using individual libraries. The libraries were constructed using the Nextera DNA Sample Preparation kit V2 (Illumina) and included individual barcodes. The libraries were quantified by qPCR and pooled by combining five individuals from a single population (six pools in total). Each pool was sequenced on one lane of Illumina HiSeq2000 using

a TruSeq SBS sequencing kit version 3 (Illumina) for 100 cycles from each end of the fragments. Reads were analysed with Casava1.8.2 (Illumina). The genome of each individual was sequenced at approximately 2.5×. Nine samples, four from the tame and five from the conventional populations, that received lower sequence coverage were re-sequenced on a part of a lane to balance the total amount of sequencing data obtained for all individuals. In total, 75.9 Gb, 81.8 Gb and 67.5 Gb of sequencing were obtained for the tame, aggressive and conventional samples, respectively (Supplementary Table 4). The sequencing data was deposited to NCBI (BioProject PRJNA376561).

**Read alignment and SNP calling.** The reads obtained for each sample were mapped, for each individual, with Bowtie2[91,92] to the 676,878 scaffolds of the fox assembly. Reads that mapped to more than one location or that mapped with a quality lower than a Phred score of 20 were removed using SAMtools[105]. The MarkDuplicates tool of Picard (http://picard.sourceforge.net) was utilized to remove duplicated reads. The ten samples from each population were then combined into population pools, and GATK[106–108] was used to re-align indels. Fox SNPs were identified using two SNP-calling programs, UnifiedGenotyper and ANGSD (Supplementary Table 5):

1.  SNPs were called by GATK UnifiedGenotyper with the pooled data from each of the three populations (three pools of 10 individuals each). SNPs with more than 2 alleles and with extremely high or low read coverage (more than 3× the average depth across all samples, or less than 1/3 the average depth across all samples) were removed using vcftools (–min-alleles 2 –max-alleles 2 –min-meanDP 8.60543 –max-meanDP 77.44887).
2.  SNPs were also called using ANGSD[109] for individual samples from each of the three populations (30 individual samples). SNPs were called using the parameters: -doMajorMinor 1 -GL 2 -doMaf 2 -doGeno 7 -realSFS 1 -doSNP 1 -doPost 1 -doCounts 1 -dumpCounts 4 -doHWE 1 and then filtered with parameter –lrt 50.

SNPs called by both programs were identified using scaffold locations, and a total of 8,458,133 SNPs identified by both programs were retained for further analysis (Supplementary Table 5).

**Principal component analysis.** PCA was performed using the genotypes of all individuals across the set of 8,458,133 SNPs without providing any information about the populations of origin of the re-sequenced samples (tame, aggressive or conventional). A covariance matrix for the SNP data was calculated using the EIGENSOFT software[110]. The eigenvectors from the covariance matrix were generated with the R function 'eigen,' and significance was determined with a Tracy–Widom test to evaluate the statistical significance of each principal component ($P < 0.01$ for both the first and the second principal components). The results of PC analysis were visualized using R.

**Construction of the individual tree.** A tree of relationships among the sequenced individuals from the tame, aggressive and conventional farm-bred populations was constructed using the neighbour-joining method[36]. Individual genotypes for the 8,458,133 SNPs were used. The distances ($D_{ij}$) between each pair of individuals ($i$ and $j$), were calculated using the formula:

$$D_{ij} = \sum_{m=1}^{M} d_{ij}/L$$

Where $M$ is the number of segregating sites in $i$ and $j$; $L$ is the length of regions and $d_{ij}$ is the distance between individuals $i$ and $j$ at site $m$. We set $d_{ij}$ equal to 0 when individuals $i$ and $j$ were both homozygous for the same allele (AA/AA); 0.5, when at least one of the genotypes of an individual $i$ or $j$ was heterozygous (Aa/AA, AA/Aa or Aa/Aa); and 1, when individuals $i$ and $j$ were both homozygous but for different alleles (AA/aa or aa/AA). We used the distance matrix of $d_{ij}$ to construct a phylogenetic tree using the neighbour-joining method and the program fneighbor[111].

**STRUCTURE analysis.** Clustering analysis was performed using the Bayesian inference program STRUCTURE 2.3.4[37–40]. Individual genotypes for 680,000 SNPs randomly chosen from the 8,458,133-SNP set were used. Four independent runs were performed at each level of $k$ from 1 to 5 with a burn-in of 100,000 and 100,000 Markov-chain Monte Carlo replicates using the admixture model without prior information about populations. The values for estimated log probability of data, $L(K)$, were used to calculate delta $k$ for the levels of $k$ from 2 to 4 in order to find the optimal number of subpopulations following a typical procedure[112] (Supplementary Table 21). The value for both delta $k$ and the mean of the estimated log probability of the data were highest at $k = 3$ (Supplementary Table 21).

**Analysis of allele frequency differences.** *Pooled heterozygosity.* Pooled heterozygosity ($H_p$) is a measure of heterozygosity in a set of samples across a region containing multiple SNPs[113]. Re-sequenced samples from each population (10 samples per population) were combined, and $H_p$ was estimated for each of the three populations separately. Because each individual was sequenced with low

coverage (~2.5×) we used allelic read depth in pooled data (~25 coverage) for $H_p$ estimation in each population. The depth of each individual allele was counted using the SNP data from the GATK/UnifiedGenotyper run and used to determine the major and minor allele frequencies for each SNP in each population. $H_p$ was calculated using a sliding window approach. The selection of window size has considered several factors, including the estimated linkage disequilibrium (LD) length in tame and aggressive populations[55], simulations of the allelic fixation rate (Supplementary Note 1; Supplementary Fig. 5), and the results of a pilot analysis with smaller window sizes. The 500 kb windows were moved along the fox scaffolds in 250 kb steps. Only scaffolds of 500,000 bases and longer were included in this analysis, corresponding to the largest 309 scaffolds. Within the scaffolds, only windows containing 20 or more SNPs were considered. The average number of SNPs per window was 1,784 (median: 1,739; standard deviation: 1,084; max: 6,730). In total there were 9,151 windows in the analysis. The average read depth per window is presented in Supplementary Table 7. $H_p$ was calculated separately for each population using the formula: $H_p = 2\Sigma n_{MAJ}\Sigma n_{MIN}/(\Sigma n_{MAJ} + \Sigma n_{MIN})^2$, where $n_{MAJ}$ and $n_{MIN}$ are the number of reads for major and minor alleles for each SNP, respectively, $\Sigma n_{MAJ}$ is the sum of the reads of the major alleles for all SNPs in that window, and $\Sigma n_{MIN}$ is the same for the minor alleles[113]. Calculations were performed using in-house scripts written in R. Because the window $H_p$ values were not normally distributed (Supplementary Fig. 11), the significance threshold was established in each population by 10,000 permutations following a previous study[114]. The allele depth data were permutated using the complete set of 8,458,133 SNPs. SNP positions were held constant, and $H_p$ was calculated for all windows with over 20 SNPs in every permutation run. 10,000 permutations were conducted in R, and the minimum $H_p$ values and values at multiple percentile levels were recorded from each permutation.

For a threshold $P$-value of <0.0001, the 0.0001 percentile of the minimum values from the 10,000 permutations was calculated in R for each population. All windows in a population with $H_p$ values at that calculated value or lower were considered to be significant at $P < 0.0001$ (Supplementary Table 6). The $P$-value threshold of 0.0001 (1/10,000) was chosen because there were 9,151 (just under 10,000) windows analysed. This criterion represents a stringent threshold with an expected false positive rate of less than one window per population.

For the window corresponding to the *SorCS1* gene, region 94, we estimated the probability of observing the $H_p$ values in the tame and aggressive populations compared to a null distribution estimated using 10,000 permutations. We compared the tame and aggressive $H_p$ values in the region to the minimum $H_p$ value for various percentiles recorded while running the permutations, that is, if the lowest $H_p$ value at percentile 0.01 for all of the 10,000 permutations for that population was higher than the observed $H_p$ value, the $P$-value for the observed value is <0.01. The lowest possible percentile for which this is true was reported.

*Combined $H_p$ windows.* The significant $H_p$ windows that were identified on the same scaffold and in the same population when the gap between them was not larger than 1 Mb were merged into combined $H_p$ windows (Supplementary Table 7). Our reasons for combining these windows were twofold: (1) uneven distribution of reads among windows could impact our analysis; (2) evaluation of the $H_p$ values in gap windows (windows located in the 1 Mb interval between $H_p$ windows significant within a single population) showed low heterozygosity although these windows did not meet the population's significance cut-off.

*Fixation index.* The fixation index ($F_{ST}$) was calculated in R using the estimator formula reported previously[115], following an earlier publication[116], which allows for the use of pooled data in windows. $F_{ST}$ was calculated for the same 9,151 windows that were used in the pooled heterozygosity analysis. For each SNP the following estimators were calculated:

$$\widehat{N}^{[k]} = \left(\frac{a_1}{n_1} - \frac{a_2}{n_2}\right)^2 - \frac{h_1}{n_1} - \frac{h_2}{n_2}$$

$$\widehat{D}^{[k]} = \widehat{N}^{[k]} + h_1 + h_2$$

$$h_i = \frac{a_i(n_i - a_i)}{n_i(n_i - 1)}$$

Where $k$ *is* the individual SNP; $a_1$ is the number of reads for allele 1 in population 1; $n_1$ is the depth of reads for that SNP; $a_2$ is the number of reads for allele 1 in population 2; and $n_2$ is the depth of reads for that SNP.

For each window $F_{ST}$ was estimated using the formula:

$$\widehat{F} = \frac{\sum_{k=1}^{K} \widehat{N}^{[k]}}{\sum_{k=1}^{K} \widehat{D}^{[k]}}$$

*Combined $F_{ST}$ windows.* The significant $F_{ST}$ windows that were identified on the same scaffold and in the same type of analysis when the gap between them was not larger than 1 Mb were merged into combined $F_{ST}$ windows (Supplementary Table 7).

*Identification of 103 regions of interest.* The positions of all significant windows identified in the fox genome were analysed and used to establish regions where either a single significant window was identified, or any combination of classes of significant windows ($H_p$ or $F_{ST}$ in any population(s)) were located on a single scaffold within 1 Mb of each other (Supplementary Table 7).

*Simulations.* Simulations were conducted in forqs[117] (see Supplementary Note 1 for more details). Population parameters were selected for the simulation based on pedigree information and breeding records from 1959 (when the population was founded) through 2010, as the DNA samples used in the current study were collected no later than 2010. A base simulation with fifty generations of breeding and 240 animals was conducted and three parameters were varied:

1. To evaluate the effect of population size, the population was simulated with population sizes of 120, 480 and 960 individuals. Each of these scenarios assumed that every founder had two unique haplotypes and that the population was bred for 50 generations.

2. To evaluate the effect of the relatedness of the founding animals, two alternate levels of relatedness were simulated. The populations were set to have either 50 or 100 founding haplotypes distributed evenly in the first generation, in contrast to the 480 in the base simulation. In these scenarios, populations of 240 individuals were bred for 50 generations.

3. To evaluate the effect of the number of generations, breeding of the base population (240 unrelated individuals) was simulated over 100, 250 and 500 generations.

The simulations were run using fox chromosome 1 (VVU1) as a proxy for the fox genome. The chromosomal length (220 Mb) and recombination map (120 cM) were approximated using a meiotic linkage map of VVU1 aligned against the dog genome[35] (Supplementary Note 1, Supplementary Fig. 3).

Haplotype frequencies were calculated at 100,000 bp intervals in each simulation scenario. The distribution of the haplotype frequencies (Supplementary Fig. 4) included all non-zero haplotype frequencies across all 100 replications of each scenario. The length of haplotypes that were identical-by-descent with founder haplotypes was calculated for every haplotype in every individual in the final generation. The haplotype lengths were recorded for all 100 replicates of each simulation scenario. The proportion of the genome represented by haplotypes of a given size or shorter was calculated and is shown in Supplementary Fig. 5. The distribution of the average haplotype lengths along chromosome 1 was calculated by dividing the chromosome into one hundred 2.2 Mb windows and averaging the lengths of all haplotypes that have a midpoint falling in the window (Supplementary Fig. 6).

**Mapping the fox windows against the dog genome.** The 9,151 windows used in the $H_p$ and $F_{ST}$ analyses were mapped against the dog genome (CanFam3.1) using LASTZ (v.1.03.66)[118] to identify the window order on the fox chromosomes. The 'multiple' option of LASTZ was used to map to the entire dog genome in one run, and the alignments were then chained using the '–chain' option. All other parameters were set to default. LASTZ computed alignments separately for the forward and reverse sequence of each window and produced a separate list of alignments for each strand. To identify the best match and the secondary best match for each window, the LASTZ alignments were then filtered using the following protocol:

1. The mapped window segments were sorted by their starting nucleotide positions in the window. The alignments of the first two mapped segments in each window were compared, and if they overlapped by more than 50% of the length of either (after chaining by LASTZ, so this only happened when the same region mapped in different directions), the segment with the lower mapping score was removed, and the one with the higher mapping score was compared again to the next mapped segment in the window for overlap. All mapped window segments that did not overlap with other mapped segments were also retained.

2. Segments that mapped sequentially and in the same direction to the same dog chromosome were combined into a single segment if the ratio of the length of the combined dog segment to the length of the combined fox segment was between 0.8 and 1.2. This step allowed the identification of extended regions where fox segments were mapped to the same dog chromosome in the expected order and without large gaps. When segments were combined, the mapping score of the new, longer segment was calculated as the sum of the mapping scores of the two combined segments.

3. Short mapping segments (<1,000 bp) remaining after the joining of sequential segments were removed.

4. The second filtering step (combining segments mapped to the same dog chromosome, in the same orientation, and of similar length between dog and fox) was run again to combine any segments that were previously separated by a short segment.

5. Medium size mapping segments (<10,000 bp) were removed.

6. The second filtering step was run again to combine any segments that were previously separated by a medium segment.

When there was one filtered result for a window, this result was considered to be the main hit. When there were two hits for a window, the hit with the higher mapping score was reported as the main hit and the lower score was reported as the secondary hit. When there were three or more remaining hits, the window was examined manually and if two or more non-adjacent mapping segments were on the same dog chromosome, in the same direction, and were located close to each other, they were combined to a single extended segment. The top score is used as the primary mapping location and the second highest is reported as the secondary hit. All subsequent matches are not reported.

Out of 9,151 windows analysed, 8,715 (95.3%) mapped to one location in the dog genome, 402 to two locations, 18 to more than two locations and 6 did not receive a location after filtering. The order of windows in the fox genome (Fig. 2) was established using the alignment of the fox scaffolds against the dog genome and the known synteny between dog and fox chromosomes[29–31,35].

**Gene enrichment analysis.** The human gene symbols assigned by reciprocal blast in the course of the gene annotation of the fox genome were used in this analysis. Fox orthologues of human genes located inside of, or overlapping with, windows used in the pooled heterozygosity ($H_p$) and $F_{ST}$ analyses are listed in Supplementary Table 7. To determine the genes overlapping with each window, the intersect tool of bedtools was used with the options –wa and –wb with the windows as the 'a' file and the genes as the 'b' file.

*GO term over-representation analysis.* GO term over-representation analysis was performed for the significant windows identified in the $H_p$ and $F_{ST}$ analyses using the PANTHER (protein analysis through evolutionary relationships) classification system (PANTHER, v.13.0)[41]. The six data sets (genes identified in significant $H_p^T$, $H_p^A$, $H_p^C$, $F_{ST}^{TA}$, $F_{ST}^{TC}$ and $F_{ST}^{AC}$ windows) were analysed. The following over-representation tests were performed: "PANTHER GO-Slim Biological Process," "PANTHER GO-Slim Molecular Function," "PANTHER Protein Class," "GO biological process complete," "GO molecular function complete" and "GO cellular component complete". Annotations from the human (all genes in the database) were used as a reference list. Only results of the over-representation test with $P < 0.05$ after Bonferroni correction were reported (Supplementary Table 9).

*Brain-expressed genes.* The genes found in the windows were checked for enrichment of genes expressed in the brain. Version 17 of Human Protein Atlas[119] (http://www.proteinatlas.org/) was used and downloaded from http://v17.proteinatlas.org/download/normal_tissue.tsv.zip. Brain tissues were considered to be caudate, cerebellum, cerebral cortex, hippocampus, hypothalamus and pituitary gland. All genes that have any expression level in any brain tissue except 'none detected' were included in the list of brain-expressed genes. Of the 12,976 genes in the version of the protein atlas with relevant data, there were 10,424 genes that showed expression in the brain. Among 15,694 annotated genes in 9,151 fox windows (15,826 high-confidence annotated genes total, but not all are in the windows used in the analysis), 10,991 have data in the Human Protein Atlas and 9,058 are brain-expressed (82.4%). There are 971 annotated genes in our significant windows, among which 698 have data in the Human Protein Atlas and 571 show brain expression (81.8%) (Supplementary Table 10). A hypergeometric test was conducted at https://www.geneprof.org/GeneProf/tools/hypergeometric.jsp and did not find enrichment for brain-expressed genes in significant windows ($P = 0.69$).

Genes from significant windows were also compared to genes involved in glutamatergic, serotonergic, dopaminergic, GABAergic and cholinergic synapses as listed in the KEGG database (KEGG last updated: 7 December 2017). The enrichment for synapse-related genes from KEGG database (Supplementary Table 11) was tested using a hypergeometric test (https://www.geneprof.org/GeneProf/tools/hypergeometric.jsp) and adjusted for multiple testing with Benjamini–Hochberg correction. No significant enrichment for genes in glutamatergic (adjusted $P = 0.148$), serotonergic (0.241), dopaminergic (0.381), GABAergic (0.148) and cholinergic (0.148) synapses was observed.

**Comparison of fox significant windows with regions associated with domestication and positive selection in dogs.** The positions of the 103 fox regions from Supplementary Table 7 were compared with the dog regions associated with domestication and positive selection from four publications[42–45]. In three of these studies the dog regions were reported according their location in CanFam2[42,44,45], the positions of these regions were identified in CanFam3.1 using the liftOver tool from the UCSC browser. The syntenic regions were then identified using an alignment between the fox and dog genomes. Fox windows located within 2 Mb of the fox syntenic positions of the dog regions were considered to be regions that overlap between fox and dog. To test whether this overlap occurred at a rate higher than expected by chance, the extent to which these regions would be expected to overlap was computed by permutation. We combined the four sets of reported dog regions[42–45] into one set of regions for the permutation test. Our 103 fox regions were randomly permuted across the all 9,151 fox windows 10,000 times and the positions of the dog regions were held constant. The number of permuted fox regions that overlapped or were within 2 Mb of the dog regions was recorded for each permutation. The $P$-value for the actual number of overlap/close regions is the percentage of the 10,000 replications where the number of permuted regions marked as overlapping/close to the dog regions was at or higher than the actual number of overlapping/close regions.

**Comparison of 103 fox regions from Supplementary Table 7 with fox behavioural QTL.** The positions of fox regions from Supplementary Table 7 were compared with positions of nine fox behavioural QTL identified in previous studies[26,46]. Only QTL for behavioural phenotypes defined using PCA were included in this analysis. A QTL interval was defined as the genomic region extending 5–15 cM in both directions from the QTL peak, which is the cM position of the QTL with the most significant statistical support. The interval boundary on either side of the QTL peak was defined by the position of the mapped microsatellite marker[46] located within the 5–15 cM interval from the QTL peak that was farthest from the QTL peak. For example, if there were three markers on the fox meiotic linkage map[46] that fell on same side of the QTL peak at distances 7, 14 and 17 cM, respectively, the boundary of the QTL interval on this side would be placed at the position of the marker located 14 cM from the QTL peak. All microsatellite markers used for QTL mapping were dog-derived markers with known positions in the dog genome. Because the current QTL intervals are large and often correspond to several fox scaffolds, we used the locations of the microsatellite markers in the dog genome[46] to define the length and positions of the dog genomic regions syntenic to the fox QTL intervals. These regions were then compared to the dog genomic coordinates of the 103 fox regions from Supplementary Table 7. This analysis identified 30 fox regions (positive regions) that overlap with five out of the nine fox behavioural QTL (Supplementary Table 14).

To test whether the observed overlap between the fox regions and fox QTL intervals is statistically significant, we compared the proportion of the dog genome represented in QTL intervals (that is, the length of all nine QTL intervals relative to the total length of dog autosomes and the X chromosome in CanFam3.1) to the proportion of the windows in the 103 regions from Supplementary Table 7 that overlap with the QTL intervals (that is, the number of windows that are located in the 30 positive regions and that overlap with QTL intervals relative to the total number of windows in 103 regions). The null hypothesis was that the proportion of windows that overlap with the QTL intervals would be similar to the proportion of the dog genome that is represented in the QTL intervals. Based on dog–fox synteny[46], we estimated that the length of all nine QTL intervals corresponds to 474,130,369 bases in the dog genome; therefore, 20% of the dog genome is represented in QTL intervals (corresponding to 474,130,369 bases in the QTL intervals out of 2,327,633,984 bases in dog autosomes in CanFam3.1). Out of the 103 regions in Supplementary Table 7, 29 regions completely overlapped QTL intervals (that is, all windows in these regions overlap with QTL intervals) and one region (region 46) partly overlapped a QTL interval (61 out of 77 windows in that region overlapped the QTL interval). In total, the proportion of windows that overlapped QTL intervals was 40% (corresponding to 228 windows across the 30 positive regions overlapping the QTL intervals out of a total of 555 windows in the 103 regions). We performed a chi-square test (http://vassarstats.net/tab2x2.html) and found that that the proportion of the windows that overlap with the QTL intervals was significantly higher than would be expected by chance ($\chi^2 = 82.84$, d.f. = 1, $P$-value < 0.0001).

**Functional analysis of intergenic SNPs in significant windows.** We used the well-annotated dog genome for functional analysis of intergenic SNPs. As with variant calling in the fox de novo assembly, the reads obtained for the tame, aggressive and conventional populations were aligned to the dog genome (CanFam3.1) using Bowtie2[91,92], and SNPs were called using the UnifiedGenotyper tool from GATK[106–108]. Sequence variants that showed differences only between the dog and the fox (that is, positions where all foxes were identical and different from dog) were removed. The remaining SNPs were polymorphic in foxes and were filtered using VCFtools[120] to include only those that had two alleles, a mean depth from 30–180 reads, and a quality of 100 or greater. This filtering step used the parameters: "–min-meanDP 10 –max-meanDP 60 –min-alleles 2 –max-alleles 2 –minQ 100". The predicted effects of the SNPs that passed the filtering (Supplementary Table 22) were analysed with the program SNPeff[121] using the CanFam3.1.82 database from SNPeff. To find the SNPs located in significant $H_p$ and $F_{ST}$ windows, we utilized the results of mapping the windows to the dog genome to extract the variants that were located in dog regions that mapped to our significant windows.

**Fine mapping of the region on VVU15.** Twenty-five short polymorphic indels (1–7 nucleotides) were identified by analysing the sequences of the re-sequenced foxes aligned to fox scaffold 1. Primers were designed with AmplifX v.1.7.0 (http://crn2m.univ-mrs.fr/pub/amplifx-dist) using the sequence of fox scaffold 1. Forward primers were tagged with fluorescent tags and markers were arranged into five multiplexes (Supplementary Table 19). PCR was performed at a volume of 15 µl using 20 ng of DNA, 1 × Promega GoTaq Colorless Master Mix (Promega), and 0.3 pMol each of the tagged forward and untagged reverse primer. The following conditions were used: 96 °C 2 m; 30 cycles of 96 °C (20 s), 58 °C (20 s), 72 °C (20 s); final extension of 72 °C 1 h. The PCR products were combined post-PCR and analysed on ABI3730 Genetic Analyzer (PE Biosystems). PCR products were sized relative to an internal size standard using ABI GeneMapper 3.5 software package (PE Biosystems). In total, 70 aggressive, 64 tame, 109 $F_1$ and 537 $F_2$ individuals were genotyped.

Haploview[53] analysis of the tame and aggressive individuals was performed separately to determine the haplotypes in the two populations (Supplementary Fig. 9). Based on the Haploview data and the distances between the genotyped markers, three different sets of markers were chosen for haplotype analysis in the $F_2$ population. The three maker sets were: upstream (left) of *SorCS1* (i13, i16, i17, i19, i20), over *SorCS1* (i11, i10, i9, i7, i3, i4, i1, i12) and downstream (right) of *SorCS1* (i34, i37, i45, i47, i49, i52) (Supplementary Tables 19, 20). The frequency of the haplotypes for these three marker sets in the tame and aggressive populations were calculated by Haploview, and the $F_2$ individuals were examined manually using the pedigree information to determine their haplotypes for each marker set.

The haplotype network for the middle haplotypes was calculated using Network 5[122]. The median-joining method was used to calculate the network, leaving all options at the default settings. All haplotypes that were found by Haploview were used in the calculation (Fig. 4).

The effect of haplotypes on behaviour was analysed in the $F_2$ population (see Supplementary Note 3 for details). $F_2$ individuals that were homozygous for any haplotype in any of the three regions (left of *SorCS1*, at *SorCS1* (middle) and right of *SorCS1*) were identified. The haplotypes that were present in a homozygous state in more than 10 of the $F_2$ were selected for the analysis of their effect on DPC.1 phenotype[46]. The D.PC1 values of $F_2$ individuals from the groups homozygous for different haplotypes were compared using the Kruskal–Wallis test, and, for haplotypes found to be significant with Kruskal–Wallis, a post-hoc Dunn's test was used to compare individual haplotypes to each other. This analysis used the kruskal.test and dunn.test functions in R.

**Karyotype analysis.** *Chromosome preparation and banding techniques.* A fibroblast cell line was established from an ear skin biopsy using conventional techniques[123]. Metaphase preparations were obtained as previously described[29,124,125]. Standard G- and C-bandings were made using the methods described in Seabright[126] and Sumner[127]. Chromosomes were identified according to a previous study[128].

*Fluorescence in situ hybridization.* Metaphase chromosomes from the fox primary fibroblast cell line were GTG-stained and captured. Slides were then washed in methanol–acetic acid fixative following xylol treatment. In situ hybridization was performed with a digoxigenin-11-dUTP-labelled (TTAGGG)$_n$ telomere repeats probe and a biotin-11-dUTP labelled 18s RNA plus 28s RNA probe[29,129]. Hybridization signals were assigned to specific chromosomes or chromosome regions defined by G-banding patterns captured before hybridization.

*Image capture.* Digital images of the banded metaphase spreads and hybridization signals were captured as described[29,125,130] using the VideoTest system with a CCD camera (Jenoptic) mounted on a Zeiss microscope Axioscope 2 (Carl Zeiss). Metaphase spreads images were edited by Corel Paint Shop Pro Photo X2.

*Ethics statement.* All animal procedures complied with standards for humane care and use of laboratory animals by foreign institutions.

**Data availability.** The red fox genome assembly and raw reads that were used to generate it are under NCBI project number PRJNA378561. The sequencing data for the tame, aggressive and conventional fox populations are under NCBI project number PRJNA376561. The RNA-seq data is under NCBI/GEO project number GSE76517. Scripts used for all analyses are available upon request.

## References

1. Wayne, R. K. et al. Molecular systematics of the Canidae. *Syst. Biol.* **46**, 622–653 (1997).
2. Macdonald, D. W. & Reynolds, J. in *Canids: Foxes, Wolves, Jackals, and Dogs: Status Survey and Conservation Action Plan* (eds Sillero-Zubiri, C., Hoffmann, M. & Macdonald, D. W.) 129–136 (IUCN, Gland, 2004).
3. Baker, P. J., Funk, S. M., Harris, S. & White, P. C. Flexible spatial organization of urban foxes, *Vulpes vulpes*, before and during an outbreak of sarcoptic mange. *Anim. Behav.* **59**, 127–146 (2000).
4. Deplazes, P., Hegglin, D., Gloor, S. & Romig, T. Wilderness in the city: the urbanization of *Echinococcus multilocularis*. *Trends Parasitol.* **20**, 77–84 (2004).
5. Doncaster, C. P. & Macdonald, D. W. Drifting territoriality in the red fox *Vulpes Vulpes*. *J. Anim. Ecol.* **60**, 423–439 (1991).
6. Harris, S. & Smith, G. Demography of two urban fox (*Vulpes vulpes*) populations. *J. Appl. Ecol.* **24**, 75–86 (1987).
7. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
8. Wang, G. D. et al. Out of southern East Asia: the natural history of domestic dogs across the world. *Cell Res.* **26**, 21–33 (2016).

9.  Maher, L. A. et al. A unique human-fox burial from a pre-Natufian cemetery in the Levant (Jordan). *PLoS ONE* **6**, e15815 (2011).
10. Morey, D. *Dogs: Domestication and the Development of a Social Bond* (Cambridge Univ. Press, New York, 2010).
11. Westwood, R. Early fur-farming in Utah. *Utah Hist. Quart.* **57**, 320–339 (1989).
12. Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700–707 (2002).
13. Petersen, M. *The Fur Traders and Fur Bearing Animals* (Hammond, Buffalo, 1914).
14. Nes, N. N., Einarsson, E. J., Lohi, O. & Joergensen, G. *Beautiful Fur Animals: Their Colour Genetics* (Scientifur, Oslo, 1988).
15. Bespyatih, O. The consequences of amber acid feeding in different genotypes of farm-bred foxes. *VOGIS* **13**, 639–646 (2009).
16. Statham, M. J. et al. On the origin of a domesticated species: identifying the parent population of Russian silver foxes (*Vulpes vulpes*). *Biol. J. Linn. Soc.* **103**, 168–175 (2011).
17. Statham, M. J., Sacks, B. N., Aubry, K. B., Perrine, J. D. & Wisely, S. M. The origin of recently established red fox populations in the United States: translocations or natural range expansions? *J. Mammal.* **93**, 52–65 (2012).
18. Belyaev, D. K. Domestication of animals. *Sci. J.* **5**, 47–52 (1969).
19. Belyaev, D. K. Destabilizing selection as a factor in domestication. *J. Hered.* **70**, 301–308 (1979).
20. Trut, L. N. The genetics and phenogenetics of domestic behaviour. In *Proc. XIV Int. Cong. Genet.* Vol. 2 (ed. Belyaev, D. K.) 123–137 (Mir Publishers, Moscow, 1980).
21. Trut, L. N. Early canid domestication: the farm-fox experiment. *Am. Sci.* **87**, 160–169 (1999).
22. Trut, L. N., Plyusnina, I. Z. & Oskina, I. N. An experiment on fox domestication and debatable issues of evolution of the dog. *Genetika* **40**, 644–655 (2004).
23. Trut, L., Oskina, I. & Kharlamova, A. Animal evolution during domestication: the domesticated fox as a model. *Bioessays* **31**, 349–360 (2009).
24. Hare, B. et al. Social cognitive evolution in captive foxes is a correlated by-product of experimental domestication. *Curr. Biol.* **15**, 226–230 (2005).
25. Kukekova, A. V. et al. Measurement of segregating behaviors in experimental silver fox pedigrees. *Behav. Genet.* **38**, 185–194 (2008).
26. Kukekova, A. V. et al. Mapping loci for fox domestication: deconstruction/reconstruction of a behavioral phenotype. *Behav. Genet.* **41**, 593–606 (2011).
27. Wipf, L. & Shackelford, R. M. Chromosomes of the red fox. *Proc. Natl Acad. Sci. USA* **28**, 265–268 (1942).
28. Belyaev, D., Volobuev, V., Radzhabli, S. & Trut, L. Supernumary chromosome polymorphism and mosaicism in silver foxes. *Genet.* **10**, 58–67 (1974).
29. Yang, F. et al. A complete comparative chromosome map for the dog, red fox, and human and its integration with canine genetic maps. *Genomics* **62**, 189–202 (1999).
30. Yang, F. et al. Chromosome identification and assignment of DNA clones in the dog using a red fox and dog comparative map. *Chromosome Res.* **8**, 93–100 (2000).
31. Kukekova, A. V. et al. A meiotic linkage map of the silver fox, aligned and compared to the canine genome. *Genome Res.* **17**, 387–399 (2007).
32. Becker, S. E. et al. Anchoring the dog to its relatives reveals new evolutionary breakpoints across 11 species of the Canidae and provides new clues for the role of B chromosomes. *Chromosome Res.* **19**, 685–708 (2011).
33. Graphodatsky, A. S. et al. Phylogenomics of the dog and fox family (Canidae, Carnivora) revealed by chromosome painting. *Chromosome Res.* **16**, 129–143 (2008).
34. Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
35. Kukekova, A. V., Temnykh, S. V., Johnson, J. L., Trut, L. N. & Acland, G. M. Genetics of behavior in the silver fox. *Mamm. Genome* **23**, 164–177 (2012).
36. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
37. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
38. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
39. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* **7**, 574–578 (2007).
40. Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9**, 1322–1332 (2009).
41. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.* **44**, D336–D342 (2016).
42. Axelsson, E. et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**, 360–364 (2013).
43. Freedman, A. H. et al. Demographically-based evaluation of genomic regions under selection in domestic dogs. *PLoS Genet.* **12**, e1005851 (2016).
44. von Holdt, B. M. et al. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**, 898–902 (2010).
45. Wang, G. D. et al. The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat. Commun.* **4**, 1860 (2013).
46. Nelson, R. M. et al. Genetics of interactive behavior in silver foxes (*Vulpes vulpes*). *Behav. Genet.* **47**, 88–101 (2017).
47. Abrahams, B. S. et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4**, 36 (2013).
48. Douglas, L. N., McGuire, A. B., Manzardo, A. M. & Butler, M. G. High-resolution chromosome ideogram representation of recognized genes for bipolar disorder. *Gene.* **586**, 136–147 (2016).
49. Scherer, S. & Osborne, L. in *Genomic Disorders: The Genomic Basis of Disease* (eds Lupski, J. R. & Stankiewicz, P.) 221–236 (Humana, Totowa, 2006).
50. Freudenberg, F., Carreno Gutierrez, H., Post, A. M., Reif, A. & Norton, W. H. Aggression in non-human vertebrates: genetic mechanisms and molecular pathways. *Am. J. Med. Genet. B* **171**, 603–640 (2015).
51. Takahashi, A., Quadros, I. M., de Almeida, R. M. & Miczek, K. A. Behavioral and pharmacogenetics of aggressive behavior. *Curr. Top. Behav. Neurosci.* **12**, 73–138 (2012).
52. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
53. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
55. Johnson, J. L. et al. Genotyping-by-sequencing (GBS) detects genetic structure and confirms behavioral QTL in tame and aggressive foxes (*Vulpes vulpes*). *PLoS ONE* **10**, e0127013 (2015).
56. Sheng, Z., Pettersson, M. E., Honaker, C. F., Siegel, P. B. & Carlborg, O. Standing genetic variation as a major contributor to adaptation in the Virginia chicken lines selection experiment. *Genome Biol.* **16**, 219 (2015).
57. Heyne, H. O. et al. Genetic influences on brain gene expression in rats selected for tameness and aggression. *Genetics* **198**, 1277–1290 (2014).
58. Kharlamova, A. V., Chase, K., Lark, K. G. & Trut, L. N. Variation of skeletal parameters in silver fox (*Vulpes vulpes*), selected for behavior, and in domestic dog (*Canis familiaris*). *VOGIS* **12**, 32–38 (2008).
59. Trut, L. N., Dzerzhinskii, F. & Nikol'skii, V. S. Intracranial allometry and craniologic changes during domestication of silver foxes. *Genetika* **27**, 1605–1611 (1991).
60. Trut, L. N. et al. in *The Dog and its Genome* (eds Ostrander, E., Giger, U. & Lindblad-Toh, K.) 81–96 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2006).
61. Wilkins, A. S., Wrangham, R. W. & Fitch, W. T. The "domestication syndrome" in mammals: a unified explanation based on neural crest cell behavior and genetics. *Genetics* **197**, 795–808 (2014).
62. Wright, D. The genetic architecture of domestication in animals. *Bioinform. Biol. Insights* **9**, 11–20 (2015).
63. Oskina, I. N., Shikhevich, S. G. & Gulevich, R. G. Relationship between behavioral selection and primary and secondary immune response in wild gray rats. *Bull. Exp. Biol. Med.* **136**, 404–407 (2003).
64. Idova, G. et al. Immune reactivity in rats selected for the enhancement or elimination of aggressiveness towards humans. *Neurosci. Lett.* **609**, 103–108 (2015).
65. Oskina, I., Herbeck, Y., Shikhevich, S., Plyusnina, I. & Gulevich, R. Alterations in the hypothalamus-pituitary-adrenal and immune systems during selection of animals for tame behavior. *VOGIS* **12**, 39–49 (2008).
66. Laundeslager, M. L. & Kennedy, S. in *Psychoneuroimmunology* Vol. 1 (ed. Ader, R.) 475–496 (Elsevier Academic Press, San Diego, 2007).
67. Mommersteeg, P. M., Vermetten, E., Kavelaars, A., Geuze, E. & Heijnen, C. J. Hostility is related to clusters of T-cell cytokines and chemokines in healthy men. *Psychoneuroendocrinology* **33**, 1041–1050 (2008).
68. Miller, G. E. et al. Low early-life social class leaves a biological residue manifested by decreased glucocorticoid and increased proinflammatory signaling. *Proc. Natl Acad. Sci. USA* **106**, 14716–14721 (2009).
69. Patel, A., Siegel, A. & Zalcman, S. S. Lack of aggression and anxiolytic-like behavior in TNF receptor (TNF-R1 and TNF-R2) deficient mice. *Brain Behav. Immun.* **24**, 1276–1280 (2010).

70. Waltes, R., Chiocchetti, A. G. & Freitag, C. M. The neurobiological basis of human aggression: a review on genetic and epigenetic mechanisms. *Am. J. Med. Genet. B* **171**, 650–675 (2016).

71. Hermey, G. The Vps10p-domain receptor family. *Cell Mol. Life Sci.* **66**, 2677–2689 (2009).

72. Savas, J. N. et al. The sorting receptor SorCS1 regulates trafficking of neurexin and AMPA receptors. *Neuron* **87**, 764–780 (2015).

73. Ramanathan, S. et al. A case of autism with an interstitial deletion on 4q leading to hemizygosity for genes encoding for glutamine and glycine neurotransmitter receptor sub-units (*AMPA 2, GLRA3, GLRB*) and neuropeptide receptors NPY1R, NPY5R. *BMC Med. Genet.* **5**, 10 (2004).

74. Südhof, T. C. Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* **455**, 903–911 (2008).

75. Rujescu, D. et al. Disruption of the neurexin 1 gene is associated with schizophrenia. *Human. Mol. Genet.* **18**, 988–996 (2009).

76. Gauthier, J. et al. Truncating mutations in *NRXN2* and *NRXN1* in autism spectrum disorders and schizophrenia. *Human. Genet.* **130**, 563–573 (2011).

77. Gregor, A. et al. Expanding the clinical spectrum associated with defects in *CNTNAP2* and *NRXN1*. *BMC Med. Genet.* **12**, 106 (2011).

78. Tarabeux, J. et al. Rare mutations in *N*-methyl-D-aspartate glutamate receptors in autism spectrum disorders and schizophrenia. *Transl. Psychiatry* **1**, e55 (2011).

79. Reichelt, A. C., Rodgers, R. J. & Clapcote, S. J. The role of neurexins in schizophrenia and autistic spectrum disorder. *Neuropharmacology* **62**, 1519–1526 (2012).

80. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).

81. Soto, D., Altafaj, X., Sindreu, C. & Bayes, A. Glutamate receptor mutations in psychiatric and neurodevelopmental disorders. *Commun. Integr. Biol.* **7**, e27887 (2014).

82. Bhat, S. et al. *CACNA1C* (Cav1.2) in the pathophysiology of psychiatric disease. *Progress. Neurobiol.* **99**, 1–14 (2012).

83. Carneiro, M. et al. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* **345**, 1074–1079 (2014).

84. Li, Y. et al. Domestication of the dog from the wolf was promoted by enhanced excitatory synaptic plasticity: a hypothesis. *Genome Biol. Evol.* **6**, 3115–3121 (2014).

85. Montague, M. J. et al. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc. Natl Acad. Sci. USA* **111**, 17230–17235 (2014).

86. vonHoldt, B. M. et al. Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs. *Sci. Adv.* **3**, e1700398 (2017).

87. Pober, B. R. Williams-Beuren syndrome. *New Engl. J. Med.* **362**, 239–252 (2010).

88. Kukekova, A. V., Trut, L. N. & Acland, G. M. in *Genetics and the Behavior of Domestic Animals* (eds Grandin, T. & Deesing, M.) 361–396 (Elsevier, 2014).

89. Sambrook, J & Russell, D. W. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2001).

90. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR-flexible barcode and adapter processing for next-generation sequencing platforms. *Biology* **1**, 895–905 (2012).

91. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

92. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

93. Li, R. et al. The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).

94. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).

95. Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).

96. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).

97. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).

98. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

99. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

100. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).

101. Zdobnov, E. M. & Apweiler, R. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).

102. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

103. Li, G. et al. Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Res.* **23**, 1486–1495 (2013).

104. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinform.* **11**, 80 (2010).

105. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

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

107. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

108. Van der Auwera, G. et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **43**, 11.10.1–11.10.33 (2013).

109. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinform.* **15**, 356 (2014).

110. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).

111. Bovine Genome Sequencing and Analysis Consortium et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522–528 (2009).

112. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).

113. Rubin, C. J. et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587–591 (2010).

114. Qanbari, S. et al. A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. *PLoS ONE* **7**, e49525 (2012).

115. Karlsson, E. K. et al. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* **39**, 1321–1328 (2007).

116. Weir, B. S. & Hill, W. G. Estimating *F*-statistics. *Annu. Rev. Genet.* **36**, 721–750 (2002).

117. Kessner, D. & Novembre, J. forqs: forward-in-time simulation of recombination, quantitative traits and selection. *Bioinformatics* **30**, 576–577 (2014).

118. Harris, R. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, Pennsylvania State Univ. (2007).

119. Uhlen, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

120. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

121. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w[1118]; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

122. Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).

123. Stanyon, R. & Galleni, L. A rapid fibroblast culture technique for high resolution karyotypes. *Bolletino di Zool.* **58**, 81–83 (1990).

124. Graphodatsky, A. S. et al. Comparative cytogenetics of hamsters of the genus *Calomyscus*. *Cytogenet. Cell Genet.* **88**, 296–304 (2000).

125. Graphodatsky, A. S. et al. Phylogenetic implications of the 38 putative ancestral chromosome segments for four canid species. *Cytogenet. Cell Genet.* **92**, 243–247 (2001).

126. Seabright, M. A rapid banding technique for human chromosomes. *Lancet* **2**, 971–972 (1971).

127. Sumner, A. T. A simple technique for demonstrating centromeric heterochromatin. *Exp. Cell Res.* **75**, 304–306 (1972).

128. Makinen, A. The standard karyotype of the silver for (*Vulpes fulvus Desm*.). Committee for the standard karyotype of *Vulpes fulvus Desm*. *Hereditas* **103**, 171–176 (1985).

129. Trifonov, V. A. et al. Complex structure of B-chromosomes in two mammalian species: *Apodemus peninsulae* (Rodentia) and *Nyctereutes procyonoides* (Carnivora). *Chromosome Res.* **10**, 109–116 (2002).

130. Nie, W. et al. The genome phylogeny of domestic cat, red panda and five mustelid species revealed by comparative chromosome painting and G-banding. *Chromosome Res.* **10**, 209–222 (2002).

## Acknowledgements

## Author contributions

A.V.K., J.L.J., P.P., A.S.G., S.J.O.B., A.G.C., G.M.A., L.N.T. and G.Z. designed the study. A.V.K. and J.L.J. designed experiments. J.L.J., A.V.K., H.M.R., N.A.S., V.B., K.P.K., J.P.H., X.W. and A.V.V. performed experiments. A.V.K., J.L.J., X.X., S.F., S.L., H.M.R. and A.V.V. and performed analyses. X.X., S.F., S.L., Z.X. and G.Z. assembled the genome. A.V.K., A.V.Kh., R.G.G., A.V.V. and G.M.A. collected data. A.V.K., J.L.J., H.M.R., K.P.K., S.J.O.B., X.W., A.G.C., L.N.T. and G.Z. wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41559-018-0611-6.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Correspondence and requests for materials** should be addressed to A.V.K. or G.Z.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# nature research

Corresponding author(s):

*Double-blind peer review submissions: write DBPR and your manuscript number here instead of author names.*

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. $F$, $t$, $r$) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted *Give P values as exact values whenever suitable.* |
| ☐ | ☒ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |
| ☒ | ☐ | Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)* |

*Our web collection on statistics for biologists may be useful.*

## Software and code

Policy information about availability of computer code

| Data collection | Our experimental data was analyzed using such open data bases as UCSC browser, Ensemble, The Human Protein Atlas (version 17), KEGG, SwissProt, TrEMBL, ProDom, PRINTS, Pfam, SMART, and PANTHER. |
|---|---|
| Data analysis | The following programs were used: SOAPdenovo V2.04.4, SSPACE.v2.0, SOAP2, Casava1.8.2 (Illumina®, San Diego, CA), FastQC, flexbar, Bowtie2, TBLASTN, BLASTP, GeneWise (version 2.2.0), RepeatMasker (version 3.3.0), AUGUSTUS (version 2.5.5), TopHat, Cufflinks, LAST, LASTZ (version 1.03.66), SAMtools, Picard, GATK, ANGSD, EIGENSOFT, STRUCTURE 2.3.4, Forqs, PANTHER™ (version 13.0), liftOver tool from the UCSC browser, AmplifX (version 1.7.0), ABI3730 Genetic Analyzer (PE Biosystems, Foster City, CA), ABI GeneMapper 3.5 software package (PE Biosystems, Foster City, CA), Network 5, Haploview, Corel Paint Shop Pro Photo X2. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The red fox genome assembly and raw reads that were used to generate it are under NCBI project number PRJNA378561. The sequencing data for the tame, aggressive and conventional fox populations are under NCBI project number PRJNA376561. The RNA-seq data is under NCBI/GEO project number GSE76517.

# Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences     ☐ Behavioural & social sciences     ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](#)

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | The sample size was based on similar published studies (e.g. Rubin, C.J. et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 2010. 464:587-91) and our previous work (Nelson, R.M. et al. Genetics of Interactive Behavior in Silver Foxes (Vulpes vulpes). Behavior Genetics 2017. 47:88-101). |
| Data exclusions | The sequencing data was quality filtered and only SNPs identified by both UnifiedGenotyper and ANGSD were included in the analyses. Only windows containing at least 20 SNPs were included in the pooled heterozygosity (Hp) and fixation index (Fst) analyses. These were pre-established criteria. |
| Replication | The results of this study were compared with results from previous publications (Nelson, R.M. et al. Genetics of Interactive Behavior in Silver Foxes (Vulpes vulpes). Behavior Genetics 2017. 47:88-101; Johnson et al., Genotyping-By-Sequencing (GBS) Detects Genetic Structure and Confirms Behavioral QTL in Tame and Aggressive Foxes (Vulpes vulpes). PLoS One. 2015 10(6):e0127013). One identified genomic region was investigated in detail to assess reproducibility of the methods used in this study. |
| Randomization | No randomization was required. Study groups were based on populations. |
| Blinding | No blinding was required. |

# Reporting for specific materials, systems and methods

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ Unique biological materials |
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☐ | ☒ Animals and other organisms |
| ☒ | ☐ Human research participants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Unique biological materials

Policy information about [availability of materials](#)

| | |
|---|---|
| Obtaining unique materials | We sequenced DNA and RNA samples collected from three populations of foxes maintained at the experimental farm of the Institute of Cytology and Genetics of the Russian Academy of Sciences in Novosibirsk, Russia. The sequencing data was deposited and publicly available: NCBI BioProjects PRJNA378561 and PRJNA376561, NCBI/GEO Project GSE76517. |

## Animals and other organisms

| | |
|---|---|
| Laboratory animals | Three populations of farm-bred foxes (tame, aggressive, and conventional farm-bred) maintained at the experimental farm of the Institute of Cytology and Genetics of the Russian Academy of Sciences in Novosibirsk, Russia. |
| Wild animals | NA |
| Field-collected samples | NA |