Unravelling selection signatures in a single dog breed suggests recent selection for morphological and behavioural traits

Unravelling selection signatures in a single dog breed suggests recent selection for morphological and behavioural traits

Juliane Friedrich[1], Andrea Talenti[1], Per Arvelius[2], Erling Strandberg[3], Marie J. Haskell[4], Pamela Wiener[1*]

[1]Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, UK

[2]Swedish Armed Forces Dog Training Center, PO Box 194, SE-195 24 MÄRSTA, Sweden

[3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, PO Box 7023, S-750 07 Uppsala, Sweden

[4]Scotland's Rural College (SRUC), Edinburgh, EH9 3JG, UK

*Corresponding author

Pamela Wiener: Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, UK; Telephone: +44 (0)131 651 9100; Fax: +44 (0) 131 651 9105; pam.wiener@roslin.ed.ac.uk

Acknowledgements

The authors want to thank all owners of German Shepherd dogs participating in this study for their time and effort to answer the questionnaires and send saliva samples for genotyping. Thanks are also extended to the Kennel Club, the British Association for German Shepherd Dogs, and the German Shepherd Dog Breed Council of Great Britain for assistance in participant recruitment for the UK cohort. Thanks to Zita Polgar, Carol-Anne Duthie and Joanna Warner for assistance in contacting dog owners. We would also like to thank the SAF Dog Training Centre, in particular Lisa Rutström, for recruiting participants for the Swedish cohort, and Susanne Gustafsson and Gabriela Bottani Claros (Swedish University of Agricultural Sciences) for extracting the DNA. We are very thankful to Lee Murphy and his colleagues at the Edinburgh Clinical Research Facility (University of Edinburgh) for DNA extraction and genotyping of the UK dogs. Further thanks to Vanessa Kapsona for conducting preliminary population structure analyses. Primary funding was provided by the Dogs Trust (UK); further funding was provided by BBSRC Institute Strategic Programme Grants (to the Roslin Institute; BBS/E/D/30002276, BBS/E/D/10002070) and RESAS, Scottish Government (to SRUC).

The authors declare no conflict of interest.

Abstract

Strong selection has resulted in substantial morphological and behavioural diversity across modern dog breeds, which makes dogs interesting model animals to study the underlying genetic architecture of these traits. However, results from between-breed analyses may confound selection signatures for behaviour and morphological features that were co-selected during breed development. In this study, we assess population genetic differences in a unique resource of dogs of the same breed but with systematic behavioural selection in only one population. We exploit these different breeding backgrounds to identify signatures of recent selection. Selection signatures within populations were found on chromosomes 4 and 19, with the strongest signals in behaviour-related genes. Regions showing strong signals of divergent selection were located on chromosomes 1, 24 and 32, and include candidate genes for both physical features and behaviour. Some of the selection signatures appear to be driven by loci associated with coat colour (Chr 24; *ASIP*) and length (Chr 32; *FGF5*), while others showed evidence of association with behaviour. Our findings suggest that signatures of selection within dog breeds have been driven by selection for morphology and behaviour. Furthermore, we demonstrate that combining selection scans with association analyses is effective for dissecting the traits under selection.

Introduction

The development of current dog breeds can be viewed as a unique long-term selection experiment to study the process of domestication[1] as well as short-term evolutionary change as a consequence of intensive breeding[2]. While the domestication of the modern dog (*Canis lupus familiaris*) from wolves took place at least 15,000 years ago[3], with some estimates considerably earlier (e.g. 20,000 to 40,000 years ago[4]), the popularity of dogs has led to ongoing strict selection according to breeding schemes and standards imposed by breed associations and national kennel clubs. The establishment of genetically and phenotypically distinctive breeds by this intense artificial selection pressure has resulted in high intra-species variation for physical and physiological features, disease susceptibility and behaviour traits[5–7], which makes dogs powerful models to investigate the underlying genetic architecture and signatures of selection for various traits.

Genetic manifestation of the development of dog breeds can be seen as selection signatures, genomic regions targeted by natural or artificial selection that exhibit various characteristics, including population differentiation, extreme linkage disequilibrium (LD) and patterns of the haplotype structure (e.g. long-range haplotypes) or mutations in coding region[8]. Accordingly, selection signatures between dog breeds have been reported for physical traits, domestication-related traits and some specific behaviours and have led to the identification of candidate genes, e.g. *IGF1* for body size, *FGF5* for coat length and *HAS2* for skin wrinkling[2], *AMY2B*, *MGAM* and *SGLT1* for adaptation to a starch-rich diet[9] and *TRPM3* and *ROBO1* for athletic success in sport-hunting[10]. In a recent whole-genome sequence study of 144 modern dog breeds, positive human-imposed selection was implicated in the fixation or high prevalence within breeds of a range of morphological characteristics (e.g. ear shape, height, weight)[11]. These recent studies for selection signatures in dogs have focused on between-breed or dog-wolf comparisons and while such studies have allowed detection of signatures related to notable physical features, signatures for more subtle traits like behaviour characteristics may be confounded with or masked by signals for the physical features, which might complicate the interpretation of these signatures as appears to be the case for association signals[12].

In this study, we analysed a single dog breed, the German Shepherd dog (GSD), to detect signals of selection. The breed was established in the late 19th century by crossing multiple breeds, with the initial purpose of creating a sheep herding dog[13] and later use as a general working dog within the military or police. GSDs used in this study originated from two populations, the UK and Sweden; while the UK population represented a random sample of pet, show and working dogs, the Swedish dogs were bred within a breeding program of the Swedish Armed Forces (SAF) and only dogs that pass a behaviour test can become working dogs or be used for breeding. Accordingly, in a previous study[14] we showed that there were significant differences between the two GSD populations for various behaviour traits as measured in a questionnaire, e.g. aggression against strangers or dogs, chasing and playfulness. In contrast, morphological differences between populations were reduced compared to between-breed studies. We hypothesise that by comparing populations of the same breed but with different behaviour-related selection strategies, we may be able to identify selection signatures for behaviour as well as those for physical traits. Furthermore, by applying multiple statistical tests for the detection of selection signatures, we have increased the power to detect true signals of selection. Nonetheless, despite the within-breed approach, one of the main difficulties that remains is the identification of the actual trait(s) under selection. We addressed this issue by characterising the relationship between selection signatures and statistical associations between genotype and phenotype (behaviour and morphological traits) from the same populations. We suggest that this approach, combining population genetics and quantitative genetics methods, may also be applicable in other contexts.

Results and discussion

Genomic structure of populations

Characterising the genetic relationships between individual dogs is a valuable tool to evaluate the genetic structure of GSDs in this study. The underlying population structure in the two GSD populations (250 dogs in total) was explored by applying a principal component analysis (PCA) and ancestry estimation on a pruned SNP data set. The PCA indicated a separation between the UK and Swedish populations based on the first two principal components (PCs), which explained 2.8% and 1.9% of the genetic variance, respectively (Figure 1). With respect to PC1 and PC2, the UK dogs had a broader distribution than the Swedish GSDs, suggesting a stronger founder effect in the Swedish cohort. However, some of the UK GSDs clustered with the Swedish GSDs. The overall separation of the two populations is likely due to the geographical separation and thus primarily independent pedigrees but may also reflect the more recent origins of the Swedish population, with the SAF as the only breeder and the primary goal to breed good working dogs. The partial overlap between the two populations is likely due to the use of external dogs in the SAF breeding program, leading to some shared ancestry. A visual assessment of the ancestry estimation based on the ADMIXTURE program[15] (Figure 2) also revealed a clear discrimination between the UK and Swedish populations. The lowest cross-validation error of 0.55 was identified for three clusters (K=3), with the blue cluster primarily associated with the Swedish population and the red and green clusters primarily associated with the UK population.

The average inbreeding coefficient calculated based on runs of homozygosity (F_{ROH}) was 0.29 ± 0.02 (standard deviation; SD) for Swedish GSDs and 0.31 ± 0.05 for UK GSDs. The significantly lower inbreeding estimate ($P < 0.05$) in the Swedish population might be a consequence of a strategic breeding scheme by the Swedish Armed Forces (SAF). The average nucleotide diversity (μ) was 0.30 ± 0.16 for both populations.

Selection signatures within populations

Selection signatures can be detected within populations by identifying distinctive patterns of linkage disequilibrium (LD). In the event of selective sweeps, favourable genetic variants increase in frequency and form extended haplotypes with neighbouring genomic regions due to LD, as reviewed in Ref. 16. We computed the integrated haplotype score (iHS), which is a variation of the extended haplotype homozygosity (EHH) statistic that aims to detect recent and incomplete selective sweeps within populations[17]. In total, 197 and 142 regions with extreme EHH were detected within the UK and Swedish GSD population, respectively. A list of SNPs belonging to the top 0.5% of the iHS statistic in the UK and Swedish populations is given in Table A2. The iHS statistic identified similar selection signatures in both populations, but the most extreme values differed between populations, as shown by the ten regions with the highest iHS statistics (Figure 3, Table 1). Regions with the highest iHS for the UK population were located on Chr 19 at 36.0 – 36.5 Mb and 37.5 – 37.7 Mb. A single marker on Chr 4 at 52.5 Mb showed the highest iHS in the Swedish population, followed by a region on Chr 18 at 54.9 – 55.3 Mb. The SNPs identified by iHS were further tested for their association with different traits (coat colour, coat length and behaviour) separately for each population to identify the putative trait under selection.

The genes located within or closest to the ten most extreme values of iHS (positional candidate genes) identified within populations (Table 1) have been previously associated with behaviour. Regarding those on Chr 19, variants in *TMEM163* (transmembrane protein 163) were associated with active behaviour in an open-field test involving cattle[18]. However, *TMEM163* is also a functional candidate for physical features, e.g. for eye width and depth[19] and hair colour[20] in humans. *NCKAP5* (NCK associated protein 5) was also identified as candidate gene for temperament in cattle[21] and has been associated with numerous neurological conditions in humans[22–24].

The iHS peak on Chr 4 in the Swedish population points to the *CLINT1* (Clathrin Interactor 1) gene. This gene is reported to be among the top risk genes for the susceptibility to schizophrenia in humans[25]

and markers near *CLINT1* were suggestive peaks associated with barking tendency in a genome-wide association study of behaviour traits in Labrador retrievers[26].

We conducted a gene list enrichment analysis with Enrichr[27,28] of the 256 and 338 genes that were located in and close to (within 40 kb of) the regions of the top 0.5% iHS in the UK and Swedish populations, respectively. No pathways were significantly enriched after accounting for multiple testing, however, Panther pathway analyses indicated nominally significant ($P < 0.05$) functional enrichment of several pathways for the UK population: "heterotrimeric G-protein signalling -Gi alpha and Gs alpha mediated" ($P = 0.01$; genes: *GRK4, GRK7, RGS12, ADCY2, ADRA2C, DRD2*), "Alzheimer disease-presenilin" ($P = 0.02$; *TRPC6, MMP7, MMP27, RBPJ, MMP20*), "heterotrimeric G-protein signalling -Gq alpha and Go alpha mediated" ($P = 0.02$; *GRK4, GRK7, CACNA1A, RGS12, DRD2*), "ionotropic glutamate receptor" ($P = 0.03$; *CACNA1A, SLC17A8, GRIA4*) and "axon guidance mediated by semaphorins" ($P = 0.03$*; CRMP1, FYN*). All of these functions have been shown to be relevant for behaviour among other functions, e.g. heterotrimeric G proteins in mood disorders, as reviewed in Ref. 29, ionotropic glutamate receptors for long term synaptic plasticity, as reviewed in Ref. 30, 31 and semaphorins in neuronal structure, as reviewed in Ref. 32. Nominally significant pathways for the Swedish population were "5-Hydroxytryptamine degradation" ($P = 0.003$; *ALDH3A2*, *ALDH3A1*), "apoptosis signaling" ($P = 0.01$; *MAP2K3, CASP9, DAXX, BAK1, BIRC2, BIRC3*) and "Thyrotropin-releasing hormone receptor signaling" ($P = 0.03$; *PLCE1, STX3, TRHR*). 5-hydroxytryptamine (serotonin) is an important neurotransmitter and plays a key role in numerous behavioural disorders and characteristics, e.g. depression[33] and aggressiveness[34].

Selection signatures between populations

Another approach to identify signatures of selection is the comparison of genetic variation (e.g. allele frequencies or haplotype structure) between different populations. Accordingly, signatures of differential selection between the two GSD populations were analysed employing three different tests: the fixation index (F_{ST}), the cross-population extended haplotype homozygosity (XP-EHH) and differences between ROH (ΔROH_{Prop}). F_{ST} was calculated to determine genetic differentiation between UK and Swedish GSD populations. Low genome-wide genetic differentiation was detected for the single SNP-based statistic ($F_{ST} = 0.021 \pm 0.029$) and for the SNP window-based statistic ($F_{ST} = 0.021 \pm 0.016$), consistent with previous within-dog-breed estimates [35].

We scanned the genome for regions of genetic differentiation within overlapping 1 Mb windows and found 17 distinctive peaks that comprise the top 1% window-based F_{ST} values on Chr 1, 9, 20, 22, 24, 29, 30 and 32, with values ranging from 0.07 to 0.16 (Table A3). The highest F_{ST} value (0.16) was found for a region on Chr 24 (22.0 – 24.5 Mb), which contains 46 genes. Among these genes are several with functions in physical characteristics and behaviour, e.g. *SPAG4* and *SUN5* involved in cytoskeletal anchoring, *NCOA6* involved in glucocorticoid and corticosteroid receptor signalling and *ASIP* and *RALY* associated with skin and fur pigmentation. Furthermore, seven members of the bactericidal/permeability-increasing (BPI) fold-containing (BPIF) superfamily of genes are located in this region (*BPIFB2, BPIFB6, BPIFB3, BPIFB4, BPIFA2, BPIFA3, BPIFA1* and *BPIFB1*). It was shown that these genes play a role in the innate immune system and lipoprotein metabolism, but also in the brain's response to oxidative stress (ageing), relevant for neuropsychiatric diseases[36]. Interestingly, high F_{ST} for Labrador retriever populations differentiated based on their coat colour and function (gundog and showdog) was also detected in the same region on Chr 24 (22.4 – 22.8 Mb) in a previous study[37].

While the F_{ST} statistic detects differences in allele frequencies between populations, the XP-EHH test, an approach based on linkage disequilibrium, is designed to detect regions that are fixed (or nearly fixed) in one population but remain segregating in the other population. Extreme high (positive) and

low (negative) scores are indicators of a region under strong positive selection in the UK and Swedish population, respectively. The region including the SNP with the highest score (3.4) for the UK population was located on Chr 35 (11.0 - 11.5 Mb) and contains three genes (*NEDD9, ADTRP*, and *TMEM170B)* (Table A3). The *NEDD9* (Neural Precursor Cell Expressed, Developmentally Down-Regulated 9) gene has been shown to be associated to cognitive impairment in mice[38], *ADTRP* is important for vascular development and function in mouse and zebrafish[39] and *TMEM170B* has been reported to be downregulated in TCGA human breast cancer data[40]. The region with the highest absolute score (3.8) for the Swedish population was located on Chr 12 (3.6-7.5 Mb). This region contains 59 genes; *RNF8* and *TBC1D22B* are closest to the SNP with the most extreme score. The ubiquitin gene *RNF8* (ring finger protein 8) plays a role in the immune system and has also been linked to autism; a recent study in *RNF8* knockout mice indicated a role of this gene in synapse formation and cerebellar-dependent learning abilities[41]. The function of *TBC1D22B* is largely unknown but it may encode a GTPase-activating protein.

As a third approach to identifying differential selection between the populations, we identified the regions showing differences in extended homozygosity. To identify these selection signatures, we calculated the between-population differences in runs of homozygosity (ΔROH_{Prop}), which describes the difference in the proportion of dogs with an ROH of a specified length at a given SNP. The average ΔROH_{Prop} value across the genome was low (0.07 ± 0.06), indicating considerable overlap of ROH between the UK and Swedish populations. However, some regions with ROH were predominantly present in only one population (Table A3). The highest absolute ΔROH_{Prop} indicating selection signatures in the UK population were found on Chr 17 and 32: the ROH mapped to Chr 17 (8.3 - 8.4 Mb) and Chr 32 (13.3 - 13.4 Mb) were present in over 70% of the UK dogs but less than 40% of the Swedish dogs. The genes located in these regions are *GREB1, NTSR2,* and *LPIN1* on Chr 17, with no characterised genes in the Chr 32 region. The neurotensin gene *NTSR2* is involved in dopamine modulation and a SNP in this gene has been tested in a polygenic model of highly sensitive personality in humans[42]. *LPIN1* plays a prominent role in lipid metabolism regulating adipocyte differentiation and co-regulating other genes involved in lipid metabolism. The highest absolute ΔROH_{Prop} indicating

selection signatures in the Swedish population was found on Chr 1: a ROH mapped to Chr 1 (24.7 to 25.5 Mb) was present in 90% of the Swedish dogs but only in 42% of the UK dogs and contains the genes *LDLRAD4*, *MOXD1* and *CTGF* (see below).

Target regions for divergent selection signatures between populations

In the detection of selection signatures, the application of multiple approaches is recommended to reduce the rate of false positive signals[16]. To identify target regions under differential selection in the two GSD populations, we selected regions from the 99th percentile (top 1%) of each score distribution (SNP window-based F_{ST}, ΔROH_{Prop}, and XP-EHH) and searched for intersecting signals between two or three of the approaches. Using this criterion, we identified 433 SNPs (Table A3), with the greatest overlap between the SNP window-based F_{ST} and ΔROH_{Prop} statistics (374 SNPs). No SNPs were detected by all three approaches. The 433 SNPs were located in 16 candidate selected regions on Chr 1, 9, 12, 22, 24, 32 and 34, which harbour 114 genes in total (Table 2; Figure 4). One Panther pathway was nominally significantly ($P < 0.05$) enriched by these 114 genes: "p53 pathway feedback loops" ($P = 0.03$; *CDKN1A, RBL1*). The SNPs identified as under divergent selection by these analyses were further tested for their association with different traits (coat colour, coat length and behaviour) separately for each population to identify the putative trait under selection.

A visual inspection of the Circos plot (Figure 4), which illustrates the results for the three approaches, indicates regions on Chr 1, 24 and 32 where peaks can be seen based on all three methods, although not belonging to the top 1% for XP-EHH. Linear plots for these three regions illustrate the results from association analyses for traits with SNPs located in that region that have adjusted $P < 0.1$ ("Regional association") and the selection signature test statistics ("Selection signatures") (Figure A2). The specific population showing evidence of selection can be determined by the ΔROH_{Prop} or XP-EHH score. Three regions showing evidence of selection in the Swedish population are located on Chr 1 (24.0 – 24.1, 24.4 – 25.1 and 25.3 – 25.9 Mb; 17 genes), each harbouring several interesting candidate genes. The *LDLRAD4 (*low density lipoprotein receptor class A domain containing 4) gene inhibits transforming growth factor-β signalling[43] and is a putative schizophrenia-related gene[44]. Another growth factor-

related gene in this region is *CTGF* (connective tissue growth factor). Other candidates for genes under selection in this region are the G-protein-associated melanocortin receptor genes *MC2R* and *MC5R*. *MC2R* (also known as the adrenocorticotropic hormone receptor gene, *ACTHR*) is a major modulator of glucocorticoid secretion regulation. *MC5R* has been associated with a range of phenotypes, including shedding and fur length in dogs[45], fatness in pigs, reviewed by Ref. 46, and psychiatric disorders in humans[47]. It was also differentially expressed in the brains of aggressive and tame foxes[48]. These reported associations with different traits highlight one of the difficulties in identifying phenotypic targets of selection. In our analysis, we found no significant associations (FDR-adjusted $P < 0.05$) between any of the selection signatures on Chr 1 with behaviour traits, coat colour or coat length, but there was a suggestive association (FDR-adjusted $P < 0.1$) with chasing behaviour in the UK population (Table 2). Regarding fur shedding, GSDs as a breed are considered to be shedders, making it unlikely that there are large differences between the two populations for this trait.

Regions showing evidence of selection in the UK population are located on Chr 24 and 32. The Chr 24 candidate region under selection (22.9 – 23.8 Mb; 18 genes) in the UK population comprises well-known genes associated with black-and-tan and saddle-tan coat colour in dogs (*ASIP, RALY*)[49,50]. We found highly significant associations in between coat colour and SNPs in this region showing evidence of selection (Table 2, Figure A2). The saddle and tan/ black and tan coat colour was the dominant coat colour in the UK GSDs while sable was predominant in the Swedish population (Table A1). The region on Chr 32 (5.4 – 5.7 Mb; 3 genes) encompasses two behaviour- and growth-related candidate genes: *PRKG2* and *RASGEF1B*. *RASGEF1B* (RasGEF domain family member 1B) has been identified as a positional candidate gene for dog rivalry in a genome-wide association study across multiple dog breeds[51]. Several case studies have been carried out in humans on chromosomal diseases related to a microdeletion of loci homologous to the region on Chr 4 comprising the *PRKG2* and *RASGEF1B* genes[52–54]. The loss of these genes leads to growth restriction, aggression, self-injurious behaviours and mental retardation in affected individuals. The association analysis revealed a significant association between SNPs in this region and aggressive behaviour towards strangers in the Swedish GSD population and *PRKG2* has previously been reported as a top candidate gene for anxiety in mice[55].

281 However, the region on Chr 32 is in close proximity to the *BMP3* gene associated with skull
282 morphology[56] and the *FGF5*[2] gene associated with coat length in dogs. Regarding *BMP3*, differences
283 in skull morphology have not previously been identified in GSDs nor have they been shown to carry a
284 derived allele in this gene previously associated with brachycephaly[56], thus selection on skull
285 morphology seems unlikely. However, we also found a highly significant association with coat length
286 in both populations (Table 2, Figure A2), suggesting that this trait drives the selection signature on Chr
287 32 (via *FGF5*).

288 **Which traits are under selection?**

289 One of the main difficulties in interpreting genomic selection signatures is the identification of the
290 actual trait(s) under selection. In dogs, the traits under selection are assumed to be primarily related to
291 physical traits (e.g. skull shape, coat colour, body size) and/or behaviour[57]. While between-breed studies
292 have greatly contributed to the understanding of the genetic control of physical traits[11,58], addressing
293 behaviour genetics by performing across-breed selection signature analyses is likely to be challenging
294 because breeds differ in multiple characteristics, including both behaviour and these physical traits,
295 many of which show Mendelian inheritance and thus tend to show very strong signals.

296 We employed several approaches to characterise the relationship between the detected selection
297 signatures and phenotypic traits that were recorded for these populations. First we repeated the
298 ADMIXTURE analysis using only genotypes from SNPs identified as selection signatures (Figure A1)
299 and fitted the ancestry assignment probabilities to the three individual clusters that were detected as
300 factors in linear models for the phenotypes. We observed significant associations between UK
301 (primarily associated with cluster 1) and Swedish (cluster 3) ancestries and some behaviour traits
302 (Stranger-directed interest, Dog-directed fear) (Table A4). Furthermore, highly significant associations
303 were identified between the ancestries and other dog characteristics, including the function of the dog
304 (working, pet or show dog), coat length and coat colour (Table A4). These results demonstrate a
305 statistical association between these phenotypes and the dog's genotypes in the selection signature
306 regions.

We then performed association analyses for behaviour traits, coat length and coat colour within each population only for markers within selection signature regions. We identified 87 SNPs with FDR-adjusted $P < 0.05$ associated with coat length, coat colour, human-directed playfulness, stranger-directed aggression, stranger directed fear and dog-directed fear (Table A5) in at least one of the populations. The striking significant associations for coat colour (lowest FDR-adjusted $P = 3.37x10^{-14}$) and coat length (lowest FDR-adjusted $P = 1.13x10^{-25}$), comprising regions on Chr 24 and 32, respectively, have previously been identified for these traits[49,59–61] (Table 2).

As discussed above, previous studies on selection signatures in dogs have generally focused on inter-breed or dog-wolf comparisons and primarily detected selection signatures (and thus candidate genes) for physical features, e.g. body size, coat characteristics and skeletal morphology[2,11,58]. Some studies, however, also identified signatures for neural crest development[1] or brain function and nervous system development[9], which might be relevant for behaviour especially in regard to domestication. We compiled a list of candidate genes reported in previous genomic analyses of phenotype associations and selection signatures in canids (dogs, wolves, foxes) focused on morphology and behaviour and compared them to genes located in regions showing evidence of selection in our study (Table A6, note that the number of overlapping genes is not informative for identifying the trait under selection because the number of reported candidate genes differs substantially between studies). The biological functions of genes in common between the two lists are diverse and include a number of genes that have been associated with behaviour. Major candidate genes for physical features in dogs, e.g. *IGF1*, *SMAD2*, *FGF5* and *BMP3*, as reviewed in Ref. 7, were not detected within selection signatures in our study. However, *FGF5*, which has previously been associated with coat length, is located in close proximity to the selection signature on Chr 32 and we detected a highly significant association with coat length for this region (*BMP3*, associated with skull morphology, is also located near this region, but as discussed above, our data does not support a signature of selection associated with this trait). We also detected well-described genes associated with coat colour (Chr 24: *ASIP, RALY*). Together these results suggest that selection for morphological traits (coat length and coat colour) has driven differences between the two populations in the genomic regions on Chr 24 and 32. In contrast, the region we

detected on Chr 1 showed an association with Chasing in the UK population and comprises candidate genes with functions in behaviour, but was not associated with morphological traits that we measured. Moreover, some of the selection signature regions showed associations with both morphological and behaviour traits, e.g. the region on Chr 32 was associated with both Stranger-directed aggression and coat length in the Swedish population (Table 2). Furthermore, genes associated with physical appearance like *ASIP* have previously been associated with behaviour traits, e.g. social behaviour in mice[62]. Thus, it is possible that some of the selection signatures we detected are also associated with multiple traits.

Limitations of the study

By comparing UK and Swedish GSDs, we hypothesised that we would be able to detect selection signatures for behaviour because behaviour was the main selection target in the Swedish population. However, we found that the geographical origin of the dogs was confounded with other attributes, e.g. coat colour and length. We addressed the issue of which trait(s) were under selection by characterising the relationship between selection signatures and associations with phenotypic attributes (behaviour, coat length, coat colour), recognizing that the sample size for the association analyses within populations was small and therefore these results should be interpreted with caution. In addition, measurements on other morphological traits (e.g. body size and weight) were not available, but these might also be under selection and should be considered in future studies. We conclude that our study of German Shepherd dogs has identified selection signatures probably driven by selection for coat colour and length (e.g. at the *ASIP* and *FGF5* genes) as well as other signatures that may be related to differential selection for behaviour between the Swedish and UK populations. Functional analyses are needed to test whether the identified candidate genes within regions showing evidence of selection do influence dog behaviour characteristics.

Material and methods

SNP genotyping and quality control

DNA was extracted from saliva samples collected with Performagene PG-100 swabs (UK population) or blood samples (Swedish population). The genotyping was performed using the CanineHD Whole-Genome Genotyping BeadChip[63] featuring 172,115 SNPs. The data was filtered for sample call rate of > 90%, SNP call rate > 98%, reproducibility (GTS) > 0.6 and low or confounded signal characterised by AB R mean (mean normalized intensity of the AB cluster) > 0.3 in GenomeStudio version 2.0. Minor allele frequency filtering of > 0.01 was used to include rare but informative variants, leaving a final dataset of 108,817 SNPs for analyses. Genotype information was available for 741 GSDs. Following further sample-based quality control, closely related dogs were removed following the procedure described in Chen et al.[64]. Briefly, a pruned genotype data set to remove closely related dogs was created for SNPs with MAF > 0.05 using PLINK version 1.9[65]: based on the variance inflation factor, a function of the multiple correlation coefficient of a given SNP regressed on all other SNPs within a window (using default parameters: window size = 50 SNPs, overlapping SNPs for shifting windows = 5, the variance inflation factor threshold = 2). Then, GCTA version 1.24.7[66] was used to compute the genetic relationship matrix and to remove one dog per pair with a genetic relationship higher than 0.2 (equivalent to 2nd degree or closer relatives) leaving a final set of 182 UK and 68 Swedish GSDs for subsequent analyses.

Samples and phenotypes

The GSDs used in this analysis originated from the UK and Sweden. For the UK population, GSDs that were at least two years old and registered with the UK Kennel Club were recruited via email to participate in a study on behaviour genetics[14,67]. GSDs from the UK population were bred by multiple breeders and primarily were pet dogs. All GSDs from the Swedish population were bred within the breeding program of the Swedish Armed Forces (SAF) starting in 2004 with the purpose of becoming working dogs. The strongest systematic selection pressure in the SAF breeding program is for behaviour

383 traits. Briefly, puppies were raised at the SAF, weaned at the age of 8 weeks and then fostered by
384 members of the Swedish public[68]. After a behaviour test at the age of 15-18 months, some dogs started
385 working with the SAF, Swedish Police or other authorities and companies, and/or were selected as
386 breeding animals, whereas others were kept as pet dogs. For the Swedish population, owners, trainers
387 or handlers of GSDs bred within the breeding program of the SAF were invited via email or letter to
388 participate in the study. Several phenotypes were analysed. Data on GSD behaviour was assessed using
389 the Canine Behaviour and Research Questionnaire (C-BARQ)[69]. The C-BARQ consists of questions
390 related to training and obedience, aggression, fear and anxiety, separation-related behaviour,
391 excitability, attachment and attention seeking, and miscellaneous behaviours. To calculate the
392 behaviour traits, a principal component analysis (PCA) was applied to the data to condense the questions
393 to a smaller number of 13 components, as described in Ref. 14. The dogs' scores for the 13 components,
394 adjusted for fixed effects (excluding cohort) as described in Ref. 67, were considered as adjusted
395 behaviour traits in the subsequent analyses. Other dog characteristics (e.g. sex, coat colour, coat length,
396 role) were assessed using a lifestyle survey[14]. Summary statistics for behaviour traits and other
397 characteristics within the two GSD populations are given in supplementary material (Table A1).

398

399 **Genomic structure of populations**

400 To characterise the genomic structure of the GSD populations, a principal component analysis (PCA)
401 and a cluster analysis were performed. PLINK version 1.9[65] with default parameters was used to create
402 a pruned SNP dataset with reduced linkage disequilibrium (LD) between SNPs, leaving a pruned dataset
403 of 9,180 SNPs. This dataset was employed only to characterise the genomic structure of populations,
404 via PCA and ADMIXTURE analyses. The PCA was performed in PLINK version 1.9 [65] and ancestry
405 estimation was performed using ADMIXTURE version 1.3.0[15]**.** The best number of clusters (K) was
406 determined by comparing 5-fold cross-validation (CV) errors.

407 Inbreeding, heterozygosity and nucleotide diversity were calculated within both GSD populations on
408 the final dataset of 108,817 SNPs. To determine inbreeding coefficients based on runs of homozygosity

(F_{ROH}), runs of homozygosity (ROH) were computed in PLINK version 1.9[65] using the default settings of a ROH length of 1000 kb and a window size of 65 SNPs, as in Pfahler and Distl[70]. The inbreeding was then estimated as the individual's total ROH length divided by the total genome length. ROH-based methods have been shown to perform best in relation to the true inbreeding[71]. Finally, nucleotide diversity (Nei's μ) was calculated per SNP using the --pi specifier in VCFtools[72].

Identification of selection signatures

Within populations

Signatures of selection within the two GSD populations were identified using the integrated haplotype score (iHS) statistic, which measures the extended haplotype homozygosity (EHH) in the genome as an indicator of selective sweeps. The iHS statistic is based on the integrated EHH (iHH_i), which is the integral of the observed decay of EHH away from a specified core allele *i* until the EHH reaches a specified cut-off. Phased genotypes of the final SNP dataset generated by Beagle version 4.1[73] (the phasing in Beagle was performed without specifying a reference population) were used to compute the SNP-wise iHS statistic using hapbin[74], specifying that the iHH should be calculated up to the point at which EHH drops below 0.05 (--cutoff 0.05). As in Voight et al.[17], the standardized iHS (iHS) for a SNP was calculated as

$$iHS = \frac{unstandardized\ iHS - \mu_{unstandardized\ iHS}}{\sigma_{unstandardized\ iHS}}$$

where the $unstandardized\ iHS$ is $\ln(iHH_i/iHH_j)$ for alleles *i* and *j*, and μ and σ are the mean and the standard deviation of the unstandardized iHS estimated from the empirical distribution of SNPs for which the derived allele frequency matches the frequency at the core SNP.

Between populations

To detect divergent signatures of selection between populations, three different approaches were used: the fixation index (F_{ST}), cross-population extended haplotype homozygosity (XP-EHH) and differences between runs of homozygosity (ROH).

First, the F_{ST} analysis was performed using the script described in Talenti et al.[75]. The F_{ST} between UK and Swedish dogs was calculated for each SNP according to the formula reported by Karlsson et al.[76], which is a comparison of the allele frequencies between populations:

$$F_{ST} = \frac{f_1^{UK}\left(f_2^S - f_2^{UK}\right) + f_1^S\left(f_2^{UK} - f_2^S\right)}{\left(f_1^{UK} * f_2^S\right) + \left(f_2^{UK} * f_1^S\right)}$$

where f_1^{UK} and f_2^{UK} are frequencies in the UK population for the two alleles and f_1^S and f_2^S are allele frequencies in the Swedish population. Next, the mean F_{ST} was calculated in 1 Mb sliding windows (window-based F_{ST}) with an overlap between windows of 500 kb, resulting in each SNP being located in exactly one or two windows. To derive a SNP-based value (to select the top 1% for calculating the intersection with other methods as described below), we averaged the window-based F_{ST} for the one or two windows in which the SNP was found.

Second, the XP-EHH statistic[77] was calculated to compare the EHH between populations, i.e. whether alleles are homozygous in one population and polymorphic in the other population. The XP-EHH statistic was calculated for the UK and Swedish populations using phased haplotypes generated by Beagle version 4.1[73] in hapbin[74], as described above.

For the third approach, ROH were computed in PLINK version 1.9[65]. We ran the analysis with the default settings of a ROH length of 1000 kb and a window size of 65 SNPs, as described above[70]. For every SNP, a homozygosity score (ROH_{Prop}) was calculated by dividing the number of dogs with a ROH at a specific SNP by the total number of dogs, such that ROH_{Prop} ranges from 0 to 1, as described in Bertolini et al.[78]. The absolute difference between ROH_{Prop} between populations (ΔROH_{Prop}) was used as statistic to determine which ROH are highly represented in one population but underrepresented in

the other population. Therefore, for every SNP, ΔROH_{Prop} values were calculated to identify ROH that are present in the majority of dogs in one population but not in the other.

Gene identification and Gene ontology (GO) analysis

To detect putative genomic regions showing evidence of selection, the most extreme values from the test statistics were selected for both the within- and between-population analyses to define selection signatures. For iHS, SNPs belonging to the top 0.5% of the distribution were selected. For F_{ST}, XP-EHH and ΔROH_{Prop}, the top 1% of each test distribution were selected and the overlap between these top SNPs was determined to identify SNPs that had most extreme values for at least two of the three methods, to reduce the chance of false positive signals. We chose a less stringent threshold for top SNPs for between-population statistics to allow for greater overlap since the three approaches differ in their methodologies and thus the ranking of top SNPs will vary. For a visual representation of target regions under selection between populations, the visualisation tool Circos[79] was used. For every SNP, the ΔROH_{Prop} and XP-EHH scores were plotted. Since the F_{ST} was calculated as a window-based average and Circos required a SNP-based value, we averaged the window-based F_{ST} for the one or two window in which the SNP was found, as described above.

The pairwise distances between the top SNPs were calculated and SNPs located within 200 kb were merged into a region. The distance of 200 kb was determined based on the linkage disequilibrium in the genome. First, the squared correlation (r^2) between all pairs of SNPs within 10Mb was calculated in PLINK version 1.9[65]. The average r^2 was then calculated for bins of increasing distance between SNPs to identify the distance around SNPs at which average r^2 drops below 0.5. The longest bin for which average $r^2 \geq 0.5$ was 200 kb.

To characterise functional relevance of regions showing evidence of selection, the top SNPs or regions (if multiple SNPs were found within 200 kb) were annotated for genes based on the CanFam3.1 genome assembly[80], using BEDtools 2.27 software[81]. SNPs were annotated considering a flanking region of ± 40kb, chosen based on the average between-marker distance of the array (~20kb), which was doubled to account for non-evenly spaced SNPs and SNPs lost through quality-control filtering. The genes

479 detected for these selection signatures were then submitted to Enrichr[27,28] to perform gene set
480 enrichment analyses. Enrichr is an integrative web-based application that compares submitted gene lists
481 to various gene-set libraries; the standard Fisher exact test option was used to calculate P-values for this
482 study.

483 **Characterising trait(s) under selection**

484 We employed two approaches to gain insights into the trait(s) under selection, as detected as genomic
485 selection signatures: (I) we modelled behaviour traits and other dog characteristics as a function of the
486 dog's ancestry based on selection signature regions and (II) we analysed the association within each
487 population between these traits and SNP markers in these regions. For both approaches, we compiled a
488 genotype data set of SNPs within the regions showing evidence of selection; this included SNPs
489 belonging to the top 0.5% of the iHS distribution in UK and Swedish populations and SNPs belonging
490 to the top 1% of F_{ST}, XP-EHH and ΔROH_{Prop} distributions that overlapped between at least two methods.

491 For (I), we repeated the ADMIXTURE analysis as described above, but only used genotypes of SNPs
492 from putatively selected regions to estimate the ancestry. Then, a linear regression was performed, as
493 described in Ref. 82, to model the relationship between the traits and ancestry assignment probabilities.

494 For (II), we analysed the association between the traits and SNP markers within the regions showing
495 evidence of selection, separately for each population. Behaviour traits were adjusted based on other
496 fixed effects as defined in the previous study[67] and treated as quantitative traits, while coat colour
497 ("saddle tan", "sable", "black", "other") and coat length ("long", "short") were treated as categorical
498 traits and not corrected for environmental factors. The association analysis was performed using
499 GEMMA[83], fitting the genomic relationship matrix (based on 108,817 genome-wide SNPs) as a random
500 effect to account for population stratification. To correct for multiple testing, P-values were adjusted
501 using the false discovery rate (FDR).

Data availability

Genotype and phenotype data for the UK dogs is available under CC-BY license from the Dryad Digital Repository[84]. The data for the Swedish dogs is restricted by the Swedish Armed Forces for reasons of national security.

References

1. Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, et al. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. BMC Biology. 2018 Jun 28;16:64.

2. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, et al. Tracking footprints of artificial selection in the dog genome. PNAS. 2010 Jan 19;107(3):1160–5.

3. Larson G, Karlsson EK, Perri A, Webster MT, Ho SYW, Peters J, et al. Rethinking dog domestication by integrating genetics, archeology, and biogeography. PNAS. 2012 Jun 5;109(23):8878–83.

4. Botigué LR, Song S, Scheu A, Gopalan S, Pendleton AL, Oetjens M, et al. Ancient European dog genomes reveal continuity since the Early Neolithic. Nat Commun. 2017 18;8:16082.

5. Mehrkam LR, Wynne C. Behavioral differences among breeds of domestic dogs (Canis lupus familiaris): Current status of the science. Applied Animal Behaviour Science. 2014;155:12–27.

6. Lewis TW, Wiles BM, Llewellyn-Zaidi AM, Evans KM, O'Neill DG. Longevity and mortality in Kennel Club registered dog breeds in the UK in 2014. Canine Genetics and Epidemiology. 2018 Oct 17;5(1):10.

7. Schoenebeck JJ, Ostrander EA. Insights into Morphology and Disease from the Dog Genome Project. Annual Review of Cell and Developmental Biology. 2014;30(1):535–60.

8. Nielsen R. Molecular Signatures of Natural Selection. Annual Review of Genetics. 2005;39(1):197–218.

9. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013 Mar;495(7441):360–4.

10. Kim J, Williams FJ, Dreger DL, Plassais J, Davis BW, Parker HG, et al. Genetic selection of athletic success in sport-hunting dogs. PNAS. 2018 Jul 24;115(30):E7212–21.

11. Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, et al. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nature Communications. 2019 Apr 2;10(1):1489.

12. Ostrander EA, Wayne RK, Freedman AH, Davis BW. Demographic history, selection and functional diversity of the canine genome. Nature Reviews Genetics. 2017 Dec;18(12):705–20.

13. Lord K, Schneider RA, Coppinger R. Evolution of working dogs [Internet]. The Domestic Dog: Its Evolution, Behavior and Interactions with People. 2016 [cited 2019 Oct 8]. Available from: /core/books/domestic-dog/evolution-of-working-dogs/CC5083D37F741470DDFA69AFBB238AB1

14. Friedrich J, Arvelius P, Strandberg E, Polgar Z, Wiener P, Haskell MJ. The interaction between behavioural traits and demographic and management factors in German Shepherd dogs. Applied Animal Behaviour Science [Internet]. 2018 Dec 5 [cited 2018 Dec 12]; Available from: http://www.sciencedirect.com/science/article/pii/S0168159118303265

15. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Jan 9;19(9):1655–64.

16. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97–120.

17. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. PLoS Biol [Internet]. 2006 Mar [cited 2018 Nov 9];4(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1382018/

18. Friedrich J, Brand B, Ponsuksili S, Graunke KL, Langbein J, Knaust J, et al. Detection of genetic variants affecting cattle behaviour and their impact on milk production: a genome-wide association study. Anim Genet. 2016 Feb 1;47(1):12–8.

19. Crouch DJM, Winney B, Koppen WP, Christmas WJ, Hutnik K, Day T, et al. Genetics of the human face: Identification of large-effect single gene variants. PNAS. 2018 Jan 23;115(4):E676–85.

20. Morgan MD, Pairo-Castineira E, Rawlik K, Canela-Xandri O, Rees J, Sims D, et al. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. Nature Communications. 2018 Dec 10;9(1):5271.

21. Valente TS, Baldi F, Sant'Anna AC, Albuquerque LG, Costa MJRP da. Genome-Wide Association Study between Single Nucleotide Polymorphisms and Flight Speed in Nellore Cattle. PLOS ONE. 2016 Jun 14;11(6):e0156956.

22. Luciano M, Huffman JE, Arias-Vásquez A, Vinkhuyzen AA, Middeldorp CM, Giegling I, et al. Genome-wide association uncovers shared genetic effects among personality traits and mood states. Am J Med Genet B Neuropsychiatr Genet. 2012 Sep;0(6):684–95.

23. Smith EN, Bloss CS, Badner JA, Barrett T, Belmonte PL, Berrettini W, et al. Genome-wide association study of bipolar disorder in European American and African American individuals. Mol Psychiatry. 2009 Aug;14(8):755–63.

24. Wang K-S, Liu X-F, Aragam N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. Schizophrenia Research. 2010 Dec 1;124(1):192–9.

25. Sun J, Kuo P-H, Riley BP, Kendler KS, Zhao Z. Candidate genes for schizophrenia: A survey of association studies and gene ranking. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2008;147B(7):1173–81.

26. Ilska J, Haskell MJ, Blott SC, Sánchez-Molano E, Polgar Z, Lofgren SE, et al. Genetic Characterisation of Dog Personality Traits. Genetics. 2017 Jan 1;genetics.116.192674.

27. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013 Apr 15;14:128.

28. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 08;44(W1):W90-97.

29. González-Maeso J, Meana JJ. Heterotrimeric G Proteins: Insights into the Neurobiology of Mood Disorders. Curr Neuropharmacol. 2006 Apr;4(2):127–38.

30. Lipsky RH, Marini AM. Brain-Derived Neurotrophic Factor in Neuronal Survival and Behavior-Related Plasticity. Annals of the New York Academy of Sciences. 2007;1122(1):130–43.

31. Lüscher C, Malenka RC. NMDA Receptor-Dependent Long-Term Potentiation and Long-Term Depression (LTP/LTD). Cold Spring Harb Perspect Biol [Internet]. 2012 Jun [cited 2019 Jun 18];4(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367554/

32. Pasterkamp RJ, Giger RJ. Semaphorin function in neural plasticity and disease. Current Opinion in Neurobiology. 2009 Jun 1;19(3):263–74.

33. Jacobsen JPR, Medvedev IO, Caron MG. The 5-HT deficiency theory of depression: perspectives from a naturalistic 5-HT deficiency model, the tryptophan hydroxylase 2Arg439His knockin mouse. Philos Trans R Soc Lond B Biol Sci. 2012 Sep 5;367(1601):2444–59.

34. de Almeida RMM, Ferrari PF, Parmigiani S, Miczek KA. Escalated aggressive behavior: Dopamine, serotonin and GABA. European Journal of Pharmacology. 2005 Dec 5;526(1):51–64.

35. Quignon P, Herbin L, Cadieu E, Kirkness EF, Hédan B, Mosher DS, et al. Canine Population Structure: Assessment and Impact of Intra-Breed Stratification on SNP-Based Association Studies. PLoS ONE [Internet]. 2007 Dec 19 [cited 2016 Mar 22];2(12). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2129117/

36. Moriya S, Soga T, Wong DW, Parhar IS. Transcriptome composition of the preoptic area in mid-age and escitalopram treatment in male mice. Neuroscience Letters. 2016 May 27;622:67–71.

37. Wiener P, Sánchez-Molano E, Clements DN, Woolliams JA, Haskell MJ, Blott SC. Genomic data illuminates demography, genetic structure and selection of a popular dog breed. BMC Genomics. 2017 Aug 14;18:609.

38. Knutson DC, Mitzey AM, Talton LE, Clagett-Dame M. Mice null for NEDD9 (HEF1) display extensive hippocampal dendritic spine loss and cognitive impairment. Brain Research. 2016 Feb 1;1632:141–55.

39. Patel MM, Silasi-Mansat R, Keshari RS, Sansam CL, Jones DA, Lupu C, et al. Role of Androgen Dependent TFPI-Regulating Protein (ADTRP) in Vascular Development and Function. Blood. 2016 Dec 2;128(22):556–556.

40. Li M, Han Y, Zhou H, Li X, Lin C, Zhang E, et al. Transmembrane protein 170B is a novel breast tumorigenesis suppressor gene that inhibits the Wnt/β-catenin pathway. Cell Death Dis [Internet]. 2018 Jan 24 [cited 2019 Jul 16];9(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5833782/

41. Valnegri P, Huang J, Yamada T, Yang Y, Mejia LA, Cho HY, et al. RNF8/UBC13 ubiquitin signaling suppresses synapse formation in the mammalian brain. Nature Communications. 2017 Nov 2;8(1):1271.

42. Chen C, Chen C, Moyzis R, Stern H, He Q, Li H, et al. Contributions of Dopamine-Related Genes and Environmental Factors to Highly Sensitive Personality: A Multi-Step Neuronal System-Level Approach. PLOS ONE. 2011 Jul 13;6(7):e21636.

43. Nakano N, Maeyama K, Sakata N, Itoh F, Akatsu R, Nakata M, et al. C18 ORF1, a Novel Negative Regulator of Transforming Growth Factor-β Signaling. J Biol Chem. 2014 Feb 5;289(18):12680–92.

44. Meerabux JMA, Ohba H, Iwayama Y, Maekawa M, Detera-Wadleigh SD, DeLisi LE, et al. Analysis of a t(18;21)(p11.1;p11.1) translocation in a family with schizophrenia. Journal of Human Genetics. 2009 Jul;54(7):386–91.

45. Hayward JJ, Castelhano MG, Oliveira KC, Corey E, Balkman C, Baxter TL, et al. Complex disease and phenotype mapping in the domestic dog. Nat Commun. 2016 Jan 22;7:10460.

46. Switonski M, Mankowska M. Dog obesity – The need for identifying predisposing genetic markers. Research in Veterinary Science. 2013 Dec;95(3):831–6.

47. Miller CL, Murakami P, Ruczinski I, Ross RG, Sinkus M, Sullivan B, et al. Two complex genotypes relevant to the kynurenine pathway and melanotropin function show association with schizophrenia and bipolar disorder. Schizophrenia Research. 2009 Sep 1;113(2):259–67.

48. Wang X, Pipes L, Trut LN, Herbeck Y, Vladimirova AV, Gulevich RG, et al. Genomic responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes). PNAS. 2018 Oct 9;115(41):10398–403.

49. Dreger DL, Schmutz SM. A SINE Insertion Causes the Black-and-Tan and Saddle Tan Phenotypes in Domestic Dogs. J Hered. 2011 Sep 1;102(Suppl_1):S11–8.

50. Dreger DL, Parker HG, Ostrander EA, Schmutz SM. Identification of a Mutation that Is Associated with the Saddle Tan and Black-and-Tan Phenotypes in Basset Hounds and Pembroke Welsh Corgis. J Hered. 2013 May 1;104(3):399–406.

51. Zapata I, Serpell JA, Alvarez CE. Genetic mapping of canine fear and aggression. BMC Genomics. 2016;17:572.

52. Bonnet C, Andrieux J, Béri-Dexheimer M, Leheup B, Boute O, Manouvrier S, et al. Microdeletion at chromosome 4q21 defines a new emerging syndrome with marked growth restriction, mental retardation and absent or severely delayed speech. Journal of Medical Genetics. 2010 Jun 1;47(6):377–84.

53. Bhoj E, Halbach S, McDonald-McGinn D, Tan C, Lande R, Waggoner D, et al. Expanding the spectrum of microdeletion 4q21 syndrome: a partial phenotype with incomplete deletion of the minimal critical region and a new association with cleft palate and Pierre Robin sequence. Am J Med Genet A. 2013 Sep;161A(9):2327–33.

54. Fee A, Noble N, Valdovinos MG. Functional Analysis of Phenotypic Behaviors of a 5-Year-Old Male with Novel 4q21 Microdeletion. J Pediatr Neuropsychol. 2015 Dec 1;1(1):36–41.

55. Le-Niculescu H, Balaraman Y, Patel SD, Ayalew M, Gupta J, Kuczenski R, et al. Convergent functional genomics of anxiety disorders: translational identification of genes, biomarkers, pathways and mechanisms. Transl Psychiatry. 2011 May;1(5):e9.

56. Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, et al. Variation of BMP3 Contributes to Dog Breed Skull Diversity. PLOS Genetics. 2012 Aug 2;8(8):e1002849.

57. Rimbault M, Ostrander EA. So many doggone traits: mapping genetics of multiple phenotypes in the domestic dog. Hum Mol Genet. 2012 Oct 15;21(R1):R52-57.

58. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Pielberg GR, Sigurdsson S, et al. Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. PLOS Genet. 2011 Oct 13;7(10):e1002316.

665 59. Legrand R, Tiret L, Abitbol M. Two recessive mutations in FGF5 are associated with the long-
666 hair phenotype in donkeys. Genet Sel Evol [Internet]. 2014 Sep 25 [cited 2019 Feb 20];46(1).
667 Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4175617/

668 60. Housley DJE, Venta PJ. The long and the short of it: evidence that FGF5 is a major determinant
669 of canine 'hair'-itability. Animal Genetics. 2006;37(4):309–15.

670 61. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, Parker HG, et al. Coat Variation in the
671 Domestic Dog Is Governed by Variants in Three Genes. Science. 2009 Oct 2;326(5949):150–3.

672 62. Carola V, Perlas E, Zonfrillo F, Soini HA, Novotny MV, Gross CT. Modulation of social behavior
673 by the agouti pigmentation gene. Front Behav Neurosci [Internet]. 2014 Aug 1 [cited 2020 Jan
674 27];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4117936/

675 63. Illumina I: Canine HD BeadChip. In Data Sheet: DNA Genotyping; 2010.
676 https://www.illumina.com/documents/products/datasheets/datasheet_caninehd.pdf

677 64. Chen M, Wang J, Wang Y, Wu Y, Fu J, Liu J. Genome-wide detection of selection signatures in
678 Chinese indigenous Laiwu pigs revealed candidate genes regulating fat deposition in muscle.
679 BMC Genet [Internet]. 2018 May 18 [cited 2019 May 30];19. Available from:
680 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5960162/

681 65. Purcell SM, Chang CC. PLINK 1.9 [Internet]. Available from: www.cog-genomics.org/plink/1.9/

682 66. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait
683 analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

684 67. Friedrich J, Strandberg E, Arvelius P, Sánchez-Molano E, Pong-Wong R, Hickey JM, et al.
685 Genetic dissection of complex behaviour traits in German Shepherd dogs. Heredity. 2019 Oct
686 14;1–13.

687 68. Wilsson E, Sinn DL. Are there differences between behavioral measurement methods? A
688 comparison of the predictive validity of two ratings methods in a working dog program. Applied
689 Animal Behaviour Science. 2012 Nov;141(3–4):158–72.

690 69. Hsu Y, Serpell JA. Development and validation of a questionnaire for measuring behavior and
691 temperament traits in pet dogs. Journal of the American Veterinary Medical Association. 2003
692 Nov 1;223(9):1293–300.

693 70. Pfahler S, Distl O. Effective Population Size, Extended Linkage Disequilibrium and Signatures
694 of Selection in the Rare Dog Breed Lundehund. PLoS One [Internet]. 2015 Apr 10 [cited 2016
695 Aug 17];10(4). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4393028/

696 71. Forutan M, Ansari Mahyari S, Baes C, Melzer N, Schenkel FS, Sargolzaei M. Inbreeding and runs
697 of homozygosity before and after genomic selection in North American Holstein cattle. BMC
698 Genomics. 2018 Jan 27;19(1):98.

699 72. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call
700 format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

701 73. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference
702 for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum
703 Genet. 2007 Nov;81(5):1084–97.

74. Maclean CA, Chue Hong NP, Prendergast JGD. hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets. Mol Biol Evol. 2015 Nov;32(11):3027–9.

75. Talenti A, Bertolini F, Pagnacco G, Pilla F, Ajmone-Marsan P, Rothschild MF, et al. The Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep regions. Mamm Genome. 2017 Apr 1;28(3):114–28.

76. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, et al. Efficient mapping of mendelian traits in dogs through genome-wide association. Nature Genetics. 2007 Nov;39(11):1321–8.

77. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007 Oct 18;449(7164):913–8.

78. Bertolini F, Gandolfi B, Kim ES, Haase B, Lyons LA, Rothschild MF. Evidence of selection signatures that shape the Persian cat breed. Mamm Genome. 2016 Apr 1;27(3):144–55.

79. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res [Internet]. 2009 Jun 18 [cited 2019 Jul 17]; Available from: http://genome.cshlp.org/content/early/2009/06/15/gr.092759.109

80. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D754–61.

81. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

82. Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, et al. Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies. PLOS Genetics. 2012 Apr 26;8(4):e1002641.

83. Zhou X, Stephens M. Genome-wide Efficient Mixed Model Analysis for Association Studies. Nat Genet. 2012 Jun 17;44(7):821–4.

84. Friedrich, J. et al. (2020), Data from: Unravelling selection signatures in a single dog breed suggests recent selection for morphological and behavioural traits, [Dataset], Dryad, https://doi:10.5061/dryad.g4f4qrfmr

85. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, et al. A Simple Genetic Architecture Underlies Morphological Variation in Dogs. PLOS Biol. 2010 Aug 10;8(8):e1000451.

86. MacLean EL, Snyder-Mackler N, vonHoldt BM, Serpell JA. Highly heritable and functionally relevant breed differences in dog behaviour. Proceedings of the Royal Society B: Biological Sciences. 2019 Oct 9;286(1912):20190716.

87. Freedman AH, Schweizer RM, Vecchyo DO-D, Han E, Davis BW, Gronau I, et al. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. PLOS Genetics. 2016 Mar 4;12(3):e1005851.

88. Kukekova AV, Johnson JL, Xiang X, Feng S, Liu S, Rando HM, et al. Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. Nature Ecology & Evolution. 2018 Sep;2(9):1479–91.

745 89. Schlamp F, van der Made J, Stambler R, Chesebrough L, Boyko AR, Messer PW. Evaluating
746 the performance of selection scans to detect selective sweeps in domestic dogs. Mol Ecol. 2016
747 Jan;25(1):342–56.

748 90. Saxena R, Voight BF, Lyssenko V, Burtt NP, Bakker PIW de, Chen H, et al. Genome-Wide
749 Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. Science. 2007
750 Jun 1;316(5829):1331–6.

751

752

Tables

Table 1. Top selection signatures within the UK and Swedish GSD populations, showing the ten highest integrated haplotype score (iHS) statistics. SNPs within 200 kb were summarised into selection signature regions.

Chr	Start (Mb)	Stop (Mb)	Distance (Mb)	N_{SNPs}†	iHS peak‡	iHS mean§	Gene(s)□	Phenotypic association††
UK population								
5	29.2	29.8	0.62	16	3.18	2.84	*ENSCAFG00000015899; MMP20; MMP27; MMP7;* ***ENSCAFG00000030873****;* ***BIRC2****; BIRC3;* ***YAP1****;* ***C11orf70****; CEP126; ANGPTL5*	-
12	68.1	68.2	0.06	2	3.22	2.96	***TRAF3IP2***	-
19	33.0	33.1	0.04	4	3.26	2.84	n.a.	-
19	36.0	36.5	0.51	10	3.46	2.93	***NCKAP5***	-
19	36.8	37.0	0.19	5	3.18	2.90	n.a.	-
19	37.5	37.7	0.20	6	3.48	3.19	***TMEM163***	-
19	38.3	38.6	0.31	9	3.19	2.79	***ZRANB3****;* *ENSCAFG00000005064;* ***R3HDM1****; UBXN4*	-
19	39.5	39.5	0.03	2	3.23	2.91	n.a.	-
20	57.6	57.7	0.07	3	3.18	3.10	*ENSCAFG00000031730; ENSCAFG00000023991;* ***ARHGAP45****; ATP5F1D; CIRBP; MIDN;* ***STK11****;* ***SBNO2****; POLR2E*	-
35	7.9	8.1	0.14	4	3.26	3.09	*BMP6;* ***TXNDC5****;* ***BLOC1S5****;* *ENSCAFG00000009583; ENSCAFG00000024482*	-
Swedish population								
4	44.3	n.a.	n.a.	1	3.09	n.a.	***ENSCAFG00000017171***	-
4	46.9	n.a.	n.a.	1	3.27	n.a.	*ENSCAFG00000028841*	-
4	50.0	50.2	0.15	4	3.09	2.90	***ATP10B***	-
4	52.5	n.a.	n.a.	1	3.47	n.a.	***CLINT1***	-
12	66.7	67.2	0.47	10	3.36	3.13	*GPR6;* ***WASF1****;* ***CDC40****;* ***METTL24****;* ***DDO****; SLC22A16; CDK19*	-
12	67.7	n.a.	n.a.	1	3.13	n.a.	***SLC16A10***	-
18	54.9	55.3	0.36	7	3.45	2.99	*LRRC10B; PPP1R32;* ***SYT7****; PGA; DDB1;* ***VWCE****;* ***ENSCAFG00000016314****;* ***SLC15A3****; CD5; VPS37C; CD6*	-

19	50.6	n.a.	n.a.	1	3.12	n.a.	***KIF5C***	-
24	42.4	42.5	0.05	3	3.33	3.05	*RBM38;* ***CTCFL***	-
36	30.1	30.6	0.05	6	3.11	2.82	*GULP1; COL3A1;* ***COL5A2***	-

757 †Number of top SNPs in region
758 ‡Standardised absolute iHS of the peak SNP (in that region)
759 §Average standardised absolute iHS across the SNPs of a region
760 Genes located within and +/- 40 kb around selection signatures. Genes highlighted in bold include a
761 SNP that belongs to the top 0.5% of the test statistic; all others are located within the region or +/- 40
762 kb around selection signatures
763 ††There were no phenotypic associations (behaviour, coat colour or coat length) with FDR-adjusted P-
764 value<0.1 for markers located within the top ten selection signatures within populations.

Table 2. Selection signatures that belonged to the top 1% of the distribution of at least two methods used to detect signatures of different selection between the GSD populations. SNPs within 200 kb were summarised into selection signature regions.

Chr	Start	Stop	N_{SNPs}†	Population	F_{ST}‡	ΔROH_{Prop}§	XP-EHH□	Gene(s)	Phenotypic association††
1	24024856	25483783	61	Sweden	0.12	0.46	NA	***ME2; MRO***; *MC2R; MC5R;* ***ENSCAFG00000000172***; *ENSCAFG00000029562; ENSCAFG00000029833; FAM210A;* ***LDLRAD4***; *ENSCAFG00000023012;* ***MOXD1***; *ENSCAFG00000031561;* ***CTGF***	Chasing*(UK)
9	16472361	16493753	4	UK	0.09	NA	2.81	*KCNJ16; KCNJ2*	-
12	5349354	6130868	44	Sweden	NA	0.27	3.44	*BRPF3;* ***PNPLA1***; ***C12H6orf222***; ***ETV7***; ***PXT1***; *ENSCAFG00000001396;* ***KCTD20***; *STK38; SRSF3; CDKN1A;* ***ENSCAFG00000001418***; *ENSCAFG00000001419;* ***CPNE5***; ***PPIL1***; ***C12H6orf89***; *MTCH1;* ***PI16***; ***FGD2***	Stranger-directed fear**(UK)
12	6466863	6554339	7	Sweden	NA	0.27	3.46	*FGD2;* ***CMTR1***; ***ENSCAFG00000030835***	Separation anxiety* (Sweden)
22	1027334	1140100	6	UK	0.08	0.26	NA	*RNASEH2B*	-
22	1683950	2496568	46	UK	0.12	0.26	NA	***KCNRG***; ***TRIM13***; ***SPRYD7***; ***KPNA3***; *ENSCAFG00000031710;* ***EBPL***; *ENSCAFG00000010362;* ***RCBTB1***; ***PHF11***; ***SETDB2***; ***CAB39L***; ***CDADC1***; *ENSCAFG00000028525; MLNR; FNDC3A*	-
24	22002778	22463326	24	UK	0.07	0.29	NA	*COMMD7;* ***DNMT3B***; ***MAPRE1***; ***EFCAB8***; ***SUN5***; ***BPIFB2***; ***BPIFB6***; ***BPIFB3***; ***BPIFB4***; ***ENSCAFG00000032553***; ***BPIFA2***; *ENSCAFG00000007369;* ***BPIFA3***; *BPIFA1*	Coat colour**(UK)
24	22908179	23816844	37	UK	0.14	0.28	NA	*ENSCAFG00000029918; ENSCAFG00000007430; ENSCAFG00000007435; ENSCAFG00000029879; NECAB3; PXMP4;* ***ZNF341***; ***CHMP4B***; *EIF2S2;* ***RALY***; ***ASIP***; *ENSCAFG00000007508; AHCY;* ***ITCH***; *DYNLRB1;* ***PIGU***; *MAP1LC3A;* ***NCOA6***; *TP53INP2*	Coat colour**(UK)

24	24867975	25952679	64	UK	0.13	0.28	NA	***CNBD2***; ***EPB41L1***; ***AAR2***; ***DLGAP4***; ***MYL9***; ***TGIF2***; ***SLA2***; *TGIF2-C20orf24*; ***NDRG3***; ***DSN1***; ***SOGA1***; ***TLDC2***; ***SAMHD1***; ***RBL1***; ***MROH8***; ***RPN2***; *GHRH*; *MANBAL*; *SRC*	Coat colour**(UK)
32	4172082	4455360	7	UK	0.09	0.27	NA	***ANTXR2***; ***PRDM8***	Coat length**(UK)
32	5350389	5399877	4	UK	0.13	0.26	NA	***PRKG2***	Coat length**(UK) and * (Sweden) Stranger-directed aggression** (Sweden)
32	5609507	5667788	4	UK	0.12	0.26	NA	*ENSCAFG00000008928*; *RASGEF1B*	Coat length** (UK and Sweden)
32	13000437	14125551	44	UK	0.11	0.37	NA	***SNCA***; ***MMRN1***; ***CCSER1***	Coat colour* (UK) Separation anxiety*(UK) Stranger-directed aggression* (Sweden)
32	14527559	14597957	4	UK	0.11	0.38	NA	*ENSCAFG00000009954*	-
32	14952127	15194499	4	UK	0.10	0.28	NA	*ENSCAFG00000009965*	-
34	33480270		1	UK	NA	0.27	2.80		-

†Number of top SNPs in region
‡Fixation index
§Differences between runs of homozygosity
Cross-population extended haplotype homozygosity.
NA indicates that this selection signature was not present in the top 1% of the test distribution
Genes highlighted in bold include a SNP that belongs to the top 1% of the test distribution; all others are located within the region or +/- 40 kb around selection signatures
††Significant phenotypic associations (behaviour, coat colour, coat length) for the UK and Swedish population within selection signature region. P-values were adjusted using False Discovery Rate (FDR), with significant associations determined as adjusted P-values <0.05 (**) and suggestive associations as adjusted P-values <0.1 (*). The population for which the phenotypic association was identified is specified in parentheses.

Figure legends

Figure 1. Principal Component Analysis of the pruned genomic data. Eigenvectors for the first two principal components are plotted and individuals are coloured according to the population of origin. The variances explained by the principal components are given in parentheses.

Figure 2. Ancestry proportions of studied GSDs based on the pruned genomic data assuming three underlying ancestries (K = 3 clusters) as revealed by ADMIXTURE. Each cluster is represented by a colour and the length of the specific coloured segment indicates the dog's proportion of membership in that cluster.

Figure 3. Distribution of integrated haplotype score (iHS) in the UK (upper plot) and Swedish population (lower plot). The red line indicates the threshold for the top 0.5% iHS.

Figure 4. Circos plot for signatures of selection between GSD populations. The plot shows the three statistics used to identify regions under differential selection: differences between runs of homozygosity (ΔROH_{Prop}, outer circle, blue track), cross-population extended haplotype homozygosity (XP-EHH, middle circle, green track) and the fixation index ($F_{ST,}$ inner circle, purple track). The plot indicates concordant evidence in regions on Chr 1, 24 and 32, where peaks can be seen based on all three methods (although not within the top 1% of SNPs for XP-EHH, shown in red for the three methods).

Appendices

Table A1. Description of German Shepherd dog populations. Summary statistics for behaviour traits and other dog attributes within the UK and the Swedish GSD populations.

Table A2. List of SNPs belonging to the top 0.5% of the iHS statistic in the UK and Swedish populations.

Table A3. Lists of SNPs belonging to the top 1% of the F_{ST}, XP-EHH and ΔROH_{Prop} statistics and the SNPs that belonged to the top 1% for at least two methods.

Table A4. Significance of associations between population attributes and genetic ancestries. The proportion of ancestries estimated by ADMIXTURE (cluster 1, cluster 2, cluster 3) based on markers located within selection signature regions were fitted as fixed effects in separate linear models to test their association with different response variables (population attributes: behaviour traits, role of the dog, coat colour and coat length). The P-values for the respective models are shown in the table.

Table A5. Markers located in selection signature regions and showing significant associations (FDR-adjusted P<0.1) with phenotypic traits (behaviour, coat colour, coat length).

Table A6. Overlaps between genes located in selection signature regions and candidate genes for morphological traits and behaviour reported in other studies. A list of candidate genes in canids was compiled using the following references[1, 2, 9, 10, 11, 26, 37, 45, 50, 51, 58, 61, 67, 76, 85-89] and was compared to genes located in regions detected as selection signatures in this study.

Figure A1. Ancestry proportions of GSDs based on genotypes of SNPs from putatively selected regions assuming three underlying ancestries (K = 3 clusters) as revealed by ADMIXTURE. Each cluster is represented by a colour and the length of the specific coloured segment indicates the dog's proportion of membership in that cluster. The labels indicate the origin of the dog (Sweden or UK) and the coat colour (1 = saddle tan, 0 = sable, black or others).

Figure A2. Fine-mapping of target regions under divergent selection between German Shepherd dog populations. Particularly compelling regions that showed evidence of divergent selection in all three selection signature test statistics (SNP window-based F_{ST}, ΔROH_{Prop}, and XP-EHH) are located on Chr 1, 24 and 32. The plots illustrate the FDR-adjusted P-values from association analyses for phenotypic traits (behaviour, coat colour, coat length) (above, "Regional association") and the selection signature test statistics (below, "Selection signatures") for all SNPs in these regions. The plots were created using a modified R code from that of Saxena et al. 2007 [90].

Unravelling selection signatures in a single dog breed suggests recent selection for morphological and behavioural traits

Juliane Friedrich[1], Andrea Talenti[1], Per Arvelius[2], Erling Strandberg[3], Marie J. Haskell[4], Pamela Wiener[1*]

[1]Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, UK

[2]Swedish Armed Forces Dog Training Center, PO Box 194, SE-195 24 MÄRSTA, Sweden

[3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, PO Box 7023, S-750 07 Uppsala, Sweden

[4]Scotland's Rural College (SRUC), Edinburgh, EH9 3JG, UK

*Corresponding author

Pamela Wiener: Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, UK; Telephone: +44 (0)131 651 9100; Fax: +44 (0) 131 651 9105; pam.wiener@roslin.ed.ac.uk

Acknowledgements

The authors want to thank all owners of German Shepherd dogs participating in this study for their time and effort to answer the questionnaires and send saliva samples for genotyping. Thanks are also extended to the Kennel Club, the British Association for German Shepherd Dogs, and the German Shepherd Dog Breed Council of Great Britain for assistance in participant recruitment for the UK cohort. Thanks to Zita Polgar, Carol-Anne Duthie and Joanna Warner for assistance in contacting dog owners. We would also like to thank the SAF Dog Training Centre, in particular Lisa Rutström, for recruiting participants for the Swedish cohort, and Susanne Gustafsson and Gabriela Bottani Claros (Swedish University of Agricultural Sciences) for extracting the DNA. We are very thankful to Lee Murphy and his colleagues at the Edinburgh Clinical Research Facility (University of Edinburgh) for DNA extraction and genotyping of the UK dogs. Further thanks to Vanessa Kapsona for conducting preliminary population structure analyses. Primary funding was provided by the Dogs Trust (UK); further funding was provided by BBSRC Institute Strategic Programme Grants (to the Roslin Institute; BBS/E/D/30002276, BBS/E/D/10002070) and RESAS, Scottish Government (to SRUC).

The authors declare no conflict of interest.

Abstract

Strong selection has resulted in substantial morphological and behavioural diversity across modern dog breeds, which makes dogs interesting model animals to study the underlying genetic architecture of these traits. However, results from between-breed analyses may confound selection signatures for behaviour and morphological features that were co-selected during breed development. In this study, we assess population genetic differences in a unique resource of dogs of the same breed but with systematic behavioural selection in only one population. We exploit these different breeding backgrounds to identify signatures of recent selection. Selection signatures within populations were found on chromosomes 4 and 19, with the strongest signals in behaviour-related genes. Regions showing strong signals of divergent selection were located on chromosomes 1, 24 and 32, and include candidate genes for both physical features and behaviour. Some of the selection signatures appear to be driven by loci associated with coat colour (Chr 24; *ASIP*) and length (Chr 32; *FGF5*), while others showed evidence of association with behaviour. Our findings suggest that signatures of selection within dog breeds have been driven by selection for morphology and behaviour. Furthermore, we demonstrate that combining selection scans with association analyses is effective for dissecting the traits under selection.

Introduction

The development of current dog breeds can be viewed as a unique long-term selection experiment to study the process of domestication[1] as well as short-term evolutionary change as a consequence of intensive breeding[2]. While the domestication of the modern dog (*Canis lupus familiaris*) from wolves took place at least 15,000 years ago[3], with some estimates considerably earlier (e.g. 20,000 to 40,000 years ago[4]), the popularity of dogs has led to ongoing strict selection according to breeding schemes and standards imposed by breed associations and national kennel clubs. The establishment of genetically and phenotypically distinctive breeds by this intense artificial selection pressure has resulted in high intra-species variation for physical and physiological features, disease susceptibility and behaviour traits[5–7], which makes dogs powerful models to investigate the underlying genetic architecture and signatures of selection for various traits.

Genetic manifestation of the development of dog breeds can be seen as selection signatures, genomic regions targeted by natural or artificial selection that exhibit various characteristics, including population differentiation, extreme linkage disequilibrium (LD) and patterns of the haplotype structure (e.g. long-range haplotypes) or mutations in coding region[8]. Accordingly, selection signatures between dog breeds have been reported for physical traits, domestication-related traits and some specific behaviours and have led to the identification of candidate genes, e.g. *IGF1* for body size, *FGF5* for coat length and *HAS2* for skin wrinkling[2], *AMY2B*, *MGAM* and *SGLT1* for adaptation to a starch-rich diet[9] and *TRPM3* and *ROBO1* for athletic success in sport-hunting[10]. In a recent whole-genome sequence study of 144 modern dog breeds, positive human-imposed selection was implicated in the fixation or high prevalence within breeds of a range of morphological characteristics (e.g. ear shape, height, weight)[11]. These recent studies for selection signatures in dogs have focused on between-breed or dog-wolf comparisons and while such studies have allowed detection of signatures related to notable physical features, signatures for more subtle traits like behaviour characteristics may be confounded with or masked by signals for the physical features, which might complicate the interpretation of these signatures as appears to be the case for association signals[12].

In this study, we analysed a single dog breed, the German Shepherd dog (GSD), to detect signals of selection. The breed was established in the late 19th century by crossing multiple breeds, with the initial purpose of creating a sheep herding dog[13] and later use as a general working dog within the military or police. GSDs used in this study originated from two populations, the UK and Sweden; while the UK population represented a random sample of pet, show and working dogs, the Swedish dogs were bred within a breeding program of the Swedish Armed Forces (SAF) and only dogs that pass a behaviour test can become working dogs or be used for breeding. Accordingly, in a previous study[14] we showed that there were significant differences between the two GSD populations for various behaviour traits as measured in a questionnaire, e.g. aggression against strangers or dogs, chasing and playfulness. In contrast, morphological differences between populations were reduced compared to between-breed studies. We hypothesise that by comparing populations of the same breed but with different behaviour-related selection strategies, we may be able to identify selection signatures for behaviour as well as those for physical traits. Furthermore, by applying multiple statistical tests for the detection of selection signatures, we have increased the power to detect true signals of selection. Nonetheless, despite the within-breed approach, one of the main difficulties that remains is the identification of the actual trait(s) under selection. We addressed this issue by characterising the relationship between selection signatures and statistical associations between genotype and phenotype (behaviour and morphological traits) from the same populations. We suggest that this approach, combining population genetics and quantitative genetics methods, may also be applicable in other contexts.

Results and discussion

Genomic structure of populations

Characterising the genetic relationships between individual dogs is a valuable tool to evaluate the genetic structure of GSDs in this study. The underlying population structure in the two GSD populations (250 dogs in total) was explored by applying a principal component analysis (PCA) and ancestry estimation on a pruned SNP data set. The PCA indicated a separation between the UK and Swedish populations based on the first two principal components (PCs), which explained 2.8% and 1.9% of the genetic variance, respectively (Figure 1). With respect to PC1 and PC2, the UK dogs had a broader distribution than the Swedish GSDs, suggesting a stronger founder effect in the Swedish cohort. However, some of the UK GSDs clustered with the Swedish GSDs. The overall separation of the two populations is likely due to the geographical separation and thus primarily independent pedigrees but may also reflect the more recent origins of the Swedish population, with the SAF as the only breeder and the primary goal to breed good working dogs. The partial overlap between the two populations is likely due to the use of external dogs in the SAF breeding program, leading to some shared ancestry. A visual assessment of the ancestry estimation based on the ADMIXTURE program[15] (Figure 2) also revealed a clear discrimination between the UK and Swedish populations. The lowest cross-validation error of 0.55 was identified for three clusters (K=3), with the blue cluster primarily associated with the Swedish population and the red and green clusters primarily associated with the UK population.

The average inbreeding coefficient calculated based on runs of homozygosity (F_{ROH}) was 0.29 ± 0.02 (standard deviation; SD) for Swedish GSDs and 0.31 ± 0.05 for UK GSDs. The significantly lower inbreeding estimate ($P < 0.05$) in the Swedish population might be a consequence of a strategic breeding scheme by the Swedish Armed Forces (SAF). The average nucleotide diversity (μ) was 0.30 ± 0.16 for both populations.

Selection signatures within populations

Selection signatures can be detected within populations by identifying distinctive patterns of linkage disequilibrium (LD). In the event of selective sweeps, favourable genetic variants increase in frequency and form extended haplotypes with neighbouring genomic regions due to LD, as reviewed in Ref. 16. We computed the integrated haplotype score (iHS), which is a variation of the extended haplotype homozygosity (EHH) statistic that aims to detect recent and incomplete selective sweeps within populations[17]. In total, 197 and 142 regions with extreme EHH were detected within the UK and Swedish GSD population, respectively. A list of SNPs belonging to the top 0.5% of the iHS statistic in the UK and Swedish populations is given in Table A2. The iHS statistic identified similar selection signatures in both populations, but the most extreme values differed between populations, as shown by the ten regions with the highest iHS statistics (Figure 3, Table 1). Regions with the highest iHS for the UK population were located on Chr 19 at 36.0 – 36.5 Mb and 37.5 – 37.7 Mb. A single marker on Chr 4 at 52.5 Mb showed the highest iHS in the Swedish population, followed by a region on Chr 18 at 54.9 – 55.3 Mb. The SNPs identified by iHS were further tested for their association with different traits (coat colour, coat length and behaviour) separately for each population to identify the putative trait under selection.

The genes located within or closest to the ten most extreme values of iHS (positional candidate genes) identified within populations (Table 1) have been previously associated with behaviour. Regarding those on Chr 19, variants in *TMEM163* (transmembrane protein 163) were associated with active behaviour in an open-field test involving cattle[18]. However, *TMEM163* is also a functional candidate for physical features, e.g. for eye width and depth[19] and hair colour[20] in humans. *NCKAP5* (NCK associated protein 5) was also identified as candidate gene for temperament in cattle[21] and has been associated with numerous neurological conditions in humans[22–24].

The iHS peak on Chr 4 in the Swedish population points to the *CLINT1* (Clathrin Interactor 1) gene. This gene is reported to be among the top risk genes for the susceptibility to schizophrenia in humans[25]

and markers near *CLINT1* were suggestive peaks associated with barking tendency in a genome-wide association study of behaviour traits in Labrador retrievers[26].

We conducted a gene list enrichment analysis with Enrichr[27,28] of the 256 and 338 genes that were located in and close to (within 40 kb of) the regions of the top 0.5% iHS in the UK and Swedish populations, respectively. No pathways were significantly enriched after accounting for multiple testing, however, Panther pathway analyses indicated nominally significant ($P < 0.05$) functional enrichment of several pathways for the UK population: "heterotrimeric G-protein signalling -Gi alpha and Gs alpha mediated" ($P = 0.01$; genes: *GRK4, GRK7, RGS12, ADCY2, ADRA2C, DRD2*), "Alzheimer disease-presenilin" ($P = 0.02$; *TRPC6, MMP7, MMP27, RBPJ, MMP20*), "heterotrimeric G-protein signalling -Gq alpha and Go alpha mediated" ($P = 0.02$; *GRK4, GRK7, CACNA1A, RGS12, DRD2*), "ionotropic glutamate receptor" ($P = 0.03$; *CACNA1A, SLC17A8, GRIA4*) and "axon guidance mediated by semaphorins" ($P = 0.03$*; CRMP1, FYN*). All of these functions have been shown to be relevant for behaviour among other functions, e.g. heterotrimeric G proteins in mood disorders, as reviewed in Ref. 29, ionotropic glutamate receptors for long term synaptic plasticity, as reviewed in Ref. 30, 31 and semaphorins in neuronal structure, as reviewed in Ref. 32. Nominally significant pathways for the Swedish population were "5-Hydroxytryptamine degradation" ($P = 0.003$; *ALDH3A2*, *ALDH3A1*), "apoptosis signaling" ($P = 0.01$; *MAP2K3, CASP9, DAXX, BAK1, BIRC2, BIRC3*) and "Thyrotropin-releasing hormone receptor signaling" ($P = 0.03$; *PLCE1, STX3, TRHR*). 5-hydroxytryptamine (serotonin) is an important neurotransmitter and plays a key role in numerous behavioural disorders and characteristics, e.g. depression[33] and aggressiveness[34].

Selection signatures between populations

Another approach to identify signatures of selection is the comparison of genetic variation (e.g. allele frequencies or haplotype structure) between different populations. Accordingly, signatures of differential selection between the two GSD populations were analysed employing three different tests: the fixation index (F_{ST}), the cross-population extended haplotype homozygosity (XP-EHH) and differences between ROH (ΔROH_{Prop}). F_{ST} was calculated to determine genetic differentiation between UK and Swedish GSD populations. Low genome-wide genetic differentiation was detected for the single SNP-based statistic ($F_{ST} = 0.021 \pm 0.029$) and for the SNP window-based statistic ($F_{ST} = 0.021 \pm 0.016$), consistent with previous within-dog-breed estimates [35].

We scanned the genome for regions of genetic differentiation within overlapping 1 Mb windows and found 17 distinctive peaks that comprise the top 1% window-based F_{ST} values on Chr 1, 9, 20, 22, 24, 29, 30 and 32, with values ranging from 0.07 to 0.16 (Table A3). The highest F_{ST} value (0.16) was found for a region on Chr 24 (22.0 – 24.5 Mb), which contains 46 genes. Among these genes are several with functions in physical characteristics and behaviour, e.g. *SPAG4* and *SUN5* involved in cytoskeletal anchoring, *NCOA6* involved in glucocorticoid and corticosteroid receptor signalling and *ASIP* and *RALY* associated with skin and fur pigmentation. Furthermore, seven members of the bactericidal/permeability-increasing (BPI) fold-containing (BPIF) superfamily of genes are located in this region (*BPIFB2, BPIFB6, BPIFB3, BPIFB4, BPIFA2, BPIFA3, BPIFA1* and *BPIFB1*). It was shown that these genes play a role in the innate immune system and lipoprotein metabolism, but also in the brain's response to oxidative stress (ageing), relevant for neuropsychiatric diseases[36]. Interestingly, high F_{ST} for Labrador retriever populations differentiated based on their coat colour and function (gundog and showdog) was also detected in the same region on Chr 24 (22.4 – 22.8 Mb) in a previous study[37].

While the F_{ST} statistic detects differences in allele frequencies between populations, the XP-EHH test, an approach based on linkage disequilibrium, is designed to detect regions that are fixed (or nearly fixed) in one population but remain segregating in the other population. Extreme high (positive) and

low (negative) scores are indicators of a region under strong positive selection in the UK and Swedish population, respectively. The region including the SNP with the highest score (3.4) for the UK population was located on Chr 35 (11.0 - 11.5 Mb) and contains three genes (*NEDD9, ADTRP*, and *TMEM170B)* (Table A3). The *NEDD9* (Neural Precursor Cell Expressed, Developmentally Down-Regulated 9) gene has been shown to be associated to cognitive impairment in mice[38], *ADTRP* is important for vascular development and function in mouse and zebrafish[39] and *TMEM170B* has been reported to be downregulated in TCGA human breast cancer data[40]. The region with the highest absolute score (3.8) for the Swedish population was located on Chr 12 (3.6-7.5 Mb). This region contains 59 genes; *RNF8* and *TBC1D22B* are closest to the SNP with the most extreme score. The ubiquitin gene *RNF8* (ring finger protein 8) plays a role in the immune system and has also been linked to autism; a recent study in *RNF8* knockout mice indicated a role of this gene in synapse formation and cerebellar-dependent learning abilities[41]. The function of *TBC1D22B* is largely unknown but it may encode a GTPase-activating protein.

As a third approach to identifying differential selection between the populations, we identified the regions showing differences in extended homozygosity. To identify these selection signatures, we calculated the between-population differences in runs of homozygosity (ΔROH_{Prop}), which describes the difference in the proportion of dogs with an ROH of a specified length at a given SNP. The average ΔROH_{Prop} value across the genome was low (0.07 ± 0.06), indicating considerable overlap of ROH between the UK and Swedish populations. However, some regions with ROH were predominantly present in only one population (Table A3). The highest absolute ΔROH_{Prop} indicating selection signatures in the UK population were found on Chr 17 and 32: the ROH mapped to Chr 17 (8.3 - 8.4 Mb) and Chr 32 (13.3 - 13.4 Mb) were present in over 70% of the UK dogs but less than 40% of the Swedish dogs. The genes located in these regions are *GREB1, NTSR2,* and *LPIN1* on Chr 17, with no characterised genes in the Chr 32 region. The neurotensin gene *NTSR2* is involved in dopamine modulation and a SNP in this gene has been tested in a polygenic model of highly sensitive personality in humans[42]. *LPIN1* plays a prominent role in lipid metabolism regulating adipocyte differentiation and co-regulating other genes involved in lipid metabolism. The highest absolute ΔROH_{Prop} indicating

selection signatures in the Swedish population was found on Chr 1: a ROH mapped to Chr 1 (24.7 to 25.5 Mb) was present in 90% of the Swedish dogs but only in 42% of the UK dogs and contains the genes *LDLRAD4*, *MOXD1* and *CTGF* (see below).

Target regions for divergent selection signatures between populations

In the detection of selection signatures, the application of multiple approaches is recommended to reduce the rate of false positive signals[16]. To identify target regions under differential selection in the two GSD populations, we selected regions from the 99th percentile (top 1%) of each score distribution (SNP window-based F_{ST}, ΔROH_{Prop}, and XP-EHH) and searched for intersecting signals between two or three of the approaches. Using this criterion, we identified 433 SNPs (Table A3), with the greatest overlap between the SNP window-based F_{ST} and ΔROH_{Prop} statistics (374 SNPs). No SNPs were detected by all three approaches. The 433 SNPs were located in 16 candidate selected regions on Chr 1, 9, 12, 22, 24, 32 and 34, which harbour 114 genes in total (Table 2; Figure 4). One Panther pathway was nominally significantly ($P < 0.05$) enriched by these 114 genes: "p53 pathway feedback loops" ($P = 0.03$; *CDKN1A, RBL1*). The SNPs identified as under divergent selection by these analyses were further tested for their association with different traits (coat colour, coat length and behaviour) separately for each population to identify the putative trait under selection.

A visual inspection of the Circos plot (Figure 4), which illustrates the results for the three approaches, indicates regions on Chr 1, 24 and 32 where peaks can be seen based on all three methods, although not belonging to the top 1% for XP-EHH. Linear plots for these three regions illustrate the results from association analyses for traits with SNPs located in that region that have adjusted $P < 0.1$ ("Regional association") and the selection signature test statistics ("Selection signatures") (Figure A2). The specific population showing evidence of selection can be determined by the ΔROH_{Prop} or XP-EHH score. Three regions showing evidence of selection in the Swedish population are located on Chr 1 (24.0 – 24.1, 24.4 – 25.1 and 25.3 – 25.9 Mb; 17 genes), each harbouring several interesting candidate genes. The *LDLRAD4 (*low density lipoprotein receptor class A domain containing 4) gene inhibits transforming growth factor-β signalling[43] and is a putative schizophrenia-related gene[44]. Another growth factor-

related gene in this region is *CTGF* (connective tissue growth factor). Other candidates for genes under selection in this region are the G-protein-associated melanocortin receptor genes *MC2R* and *MC5R*. *MC2R* (also known as the adrenocorticotropic hormone receptor gene, *ACTHR*) is a major modulator of glucocorticoid secretion regulation. *MC5R* has been associated with a range of phenotypes, including shedding and fur length in dogs[45], fatness in pigs, reviewed by Ref. 46, and psychiatric disorders in humans[47]. It was also differentially expressed in the brains of aggressive and tame foxes[48]. These reported associations with different traits highlight one of the difficulties in identifying phenotypic targets of selection. In our analysis, we found no significant associations (FDR-adjusted $P < 0.05$) between any of the selection signatures on Chr 1 with behaviour traits, coat colour or coat length, but there was a suggestive association (FDR-adjusted $P < 0.1$) with chasing behaviour in the UK population (Table 2). Regarding fur shedding, GSDs as a breed are considered to be shedders, making it unlikely that there are large differences between the two populations for this trait.

Regions showing evidence of selection in the UK population are located on Chr 24 and 32. The Chr 24 candidate region under selection (22.9 – 23.8 Mb; 18 genes) in the UK population comprises well-known genes associated with black-and-tan and saddle-tan coat colour in dogs (*ASIP, RALY*)[49,50]. We found highly significant associations in between coat colour and SNPs in this region showing evidence of selection (Table 2, Figure A2). The saddle and tan/ black and tan coat colour was the dominant coat colour in the UK GSDs while sable was predominant in the Swedish population (Table A1). The region on Chr 32 (5.4 – 5.7 Mb; 3 genes) encompasses two behaviour- and growth-related candidate genes: *PRKG2* and *RASGEF1B*. *RASGEF1B* (RasGEF domain family member 1B) has been identified as a positional candidate gene for dog rivalry in a genome-wide association study across multiple dog breeds[51]. Several case studies have been carried out in humans on chromosomal diseases related to a microdeletion of loci homologous to the region on Chr 4 comprising the *PRKG2* and *RASGEF1B* genes[52–54]. The loss of these genes leads to growth restriction, aggression, self-injurious behaviours and mental retardation in affected individuals. The association analysis revealed a significant association between SNPs in this region and aggressive behaviour towards strangers in the Swedish GSD population and *PRKG2* has previously been reported as a top candidate gene for anxiety in mice[55].

281 However, the region on Chr 32 is in close proximity to the *BMP3* gene associated with skull
282 morphology[56] and the *FGF5*[2] gene associated with coat length in dogs. Regarding *BMP3*, differences
283 in skull morphology have not previously been identified in GSDs nor have they been shown to carry a
284 derived allele in this gene previously associated with brachycephaly[56], thus selection on skull
285 morphology seems unlikely. However, we also found a highly significant association with coat length
286 in both populations (Table 2, Figure A2), suggesting that this trait drives the selection signature on Chr
287 32 (via *FGF5*).

288 **Which traits are under selection?**

289 One of the main difficulties in interpreting genomic selection signatures is the identification of the
290 actual trait(s) under selection. In dogs, the traits under selection are assumed to be primarily related to
291 physical traits (e.g. skull shape, coat colour, body size) and/or behaviour[57]. While between-breed studies
292 have greatly contributed to the understanding of the genetic control of physical traits[11,58], addressing
293 behaviour genetics by performing across-breed selection signature analyses is likely to be challenging
294 because breeds differ in multiple characteristics, including both behaviour and these physical traits,
295 many of which show Mendelian inheritance and thus tend to show very strong signals.

296 We employed several approaches to characterise the relationship between the detected selection
297 signatures and phenotypic traits that were recorded for these populations. First we repeated the
298 ADMIXTURE analysis using only genotypes from SNPs identified as selection signatures (Figure A1)
299 and fitted the ancestry assignment probabilities to the three individual clusters that were detected as
300 factors in linear models for the phenotypes. We observed significant associations between UK
301 (primarily associated with cluster 1) and Swedish (cluster 3) ancestries and some behaviour traits
302 (Stranger-directed interest, Dog-directed fear) (Table A4). Furthermore, highly significant associations
303 were identified between the ancestries and other dog characteristics, including the function of the dog
304 (working, pet or show dog), coat length and coat colour (Table A4). These results demonstrate a
305 statistical association between these phenotypes and the dog's genotypes in the selection signature
306 regions.

We then performed association analyses for behaviour traits, coat length and coat colour within each population only for markers within selection signature regions. We identified 87 SNPs with FDR-adjusted $P < 0.05$ associated with coat length, coat colour, human-directed playfulness, stranger-directed aggression, stranger directed fear and dog-directed fear (Table A5) in at least one of the populations. The striking significant associations for coat colour (lowest FDR-adjusted $P = 3.37 \times 10^{-14}$) and coat length (lowest FDR-adjusted $P = 1.13 \times 10^{-25}$), comprising regions on Chr 24 and 32, respectively, have previously been identified for these traits[49,59–61] (Table 2).

As discussed above, previous studies on selection signatures in dogs have generally focused on inter-breed or dog-wolf comparisons and primarily detected selection signatures (and thus candidate genes) for physical features, e.g. body size, coat characteristics and skeletal morphology[2,11,58]. Some studies, however, also identified signatures for neural crest development[1] or brain function and nervous system development[9], which might be relevant for behaviour especially in regard to domestication. We compiled a list of candidate genes reported in previous genomic analyses of phenotype associations and selection signatures in canids (dogs, wolves, foxes) focused on morphology and behaviour and compared them to genes located in regions showing evidence of selection in our study (Table A6, note that the number of overlapping genes is not informative for identifying the trait under selection because the number of reported candidate genes differs substantially between studies). The biological functions of genes in common between the two lists are diverse and include a number of genes that have been associated with behaviour. Major candidate genes for physical features in dogs, e.g. *IGF1*, *SMAD2*, *FGF5* and *BMP3*, as reviewed in Ref. 7, were not detected within selection signatures in our study. However, *FGF5*, which has previously been associated with coat length, is located in close proximity to the selection signature on Chr 32 and we detected a highly significant association with coat length for this region (*BMP3*, associated with skull morphology, is also located near this region, but as discussed above, our data does not support a signature of selection associated with this trait). We also detected well-described genes associated with coat colour (Chr 24: *ASIP, RALY*). Together these results suggest that selection for morphological traits (coat length and coat colour) has driven differences between the two populations in the genomic regions on Chr 24 and 32. In contrast, the region we

334 detected on Chr 1 showed an association with Chasing in the UK population and comprises candidate
335 genes with functions in behaviour, but was not associated with morphological traits that we measured.
336 Moreover, some of the selection signature regions showed associations with both morphological and
337 behaviour traits, e.g. the region on Chr 32 was associated with both Stranger-directed aggression and
338 coat length in the Swedish population (Table 2). Furthermore, genes associated with physical
339 appearance like *ASIP* have previously been associated with behaviour traits, e.g. social behaviour in
340 mice[62]. Thus, it is possible that some of the selection signatures we detected are also associated with
341 multiple traits.

342

343 **Limitations of the study**

344 By comparing UK and Swedish GSDs, we hypothesised that we would be able to detect selection
345 signatures for behaviour because behaviour was the main selection target in the Swedish population.
346 However, we found that the geographical origin of the dogs was confounded with other attributes, e.g.
347 coat colour and length. We addressed the issue of which trait(s) were under selection by characterising
348 the relationship between selection signatures and associations with phenotypic attributes (behaviour,
349 coat length, coat colour), recognizing that the sample size for the association analyses within
350 populations was small and therefore these results should be interpreted with caution. In addition,
351 measurements on other morphological traits (e.g. body size and weight) were not available, but these
352 might also be under selection and should be considered in future studies. We conclude that our study of
353 German Shepherd dogs has identified selection signatures probably driven by selection for coat colour
354 and length (e.g. at the *ASIP* and *FGF5* genes) as well as other signatures that may be related to
355 differential selection for behaviour between the Swedish and UK populations. Functional analyses are
356 needed to test whether the identified candidate genes within regions showing evidence of selection do
357 influence dog behaviour characteristics.

Material and methods

SNP genotyping and quality control

DNA was extracted from saliva samples collected with Performagene PG-100 swabs (UK population) or blood samples (Swedish population). The genotyping was performed using the CanineHD Whole-Genome Genotyping BeadChip[63] featuring 172,115 SNPs. The data was filtered for sample call rate of > 90%, SNP call rate > 98%, reproducibility (GTS) > 0.6 and low or confounded signal characterised by AB R mean (mean normalized intensity of the AB cluster) > 0.3 in GenomeStudio version 2.0. Minor allele frequency filtering of > 0.01 was used to include rare but informative variants, leaving a final dataset of 108,817 SNPs for analyses. Genotype information was available for 741 GSDs. Following further sample-based quality control, closely related dogs were removed following the procedure described in Chen et al.[64]. Briefly, a pruned genotype data set to remove closely related dogs was created for SNPs with MAF > 0.05 using PLINK version 1.9[65]: based on the variance inflation factor, a function of the multiple correlation coefficient of a given SNP regressed on all other SNPs within a window (using default parameters: window size = 50 SNPs, overlapping SNPs for shifting windows = 5, the variance inflation factor threshold = 2). Then, GCTA version 1.24.7[66] was used to compute the genetic relationship matrix and to remove one dog per pair with a genetic relationship higher than 0.2 (equivalent to 2nd degree or closer relatives) leaving a final set of 182 UK and 68 Swedish GSDs for subsequent analyses.

Samples and phenotypes

The GSDs used in this analysis originated from the UK and Sweden. For the UK population, GSDs that were at least two years old and registered with the UK Kennel Club were recruited via email to participate in a study on behaviour genetics[14,67]. GSDs from the UK population were bred by multiple breeders and primarily were pet dogs. All GSDs from the Swedish population were bred within the breeding program of the Swedish Armed Forces (SAF) starting in 2004 with the purpose of becoming working dogs. The strongest systematic selection pressure in the SAF breeding program is for behaviour

383 traits. Briefly, puppies were raised at the SAF, weaned at the age of 8 weeks and then fostered by
384 members of the Swedish public[68]. After a behaviour test at the age of 15-18 months, some dogs started
385 working with the SAF, Swedish Police or other authorities and companies, and/or were selected as
386 breeding animals, whereas others were kept as pet dogs. For the Swedish population, owners, trainers
387 or handlers of GSDs bred within the breeding program of the SAF were invited via email or letter to
388 participate in the study. Several phenotypes were analysed. Data on GSD behaviour was assessed using
389 the Canine Behaviour and Research Questionnaire (C-BARQ)[69]. The C-BARQ consists of questions
390 related to training and obedience, aggression, fear and anxiety, separation-related behaviour,
391 excitability, attachment and attention seeking, and miscellaneous behaviours. To calculate the
392 behaviour traits, a principal component analysis (PCA) was applied to the data to condense the questions
393 to a smaller number of 13 components, as described in Ref. 14. The dogs' scores for the 13 components,
394 adjusted for fixed effects (excluding cohort) as described in Ref. 67, were considered as adjusted
395 behaviour traits in the subsequent analyses. Other dog characteristics (e.g. sex, coat colour, coat length,
396 role) were assessed using a lifestyle survey[14]. Summary statistics for behaviour traits and other
397 characteristics within the two GSD populations are given in supplementary material (Table A1).

398

399 **Genomic structure of populations**

400 To characterise the genomic structure of the GSD populations, a principal component analysis (PCA)
401 and a cluster analysis were performed. PLINK version 1.9[65] with default parameters was used to create
402 a pruned SNP dataset with reduced linkage disequilibrium (LD) between SNPs, leaving a pruned dataset
403 of 9,180 SNPs. This dataset was employed only to characterise the genomic structure of populations,
404 via PCA and ADMIXTURE analyses. The PCA was performed in PLINK version 1.9 [65] and ancestry
405 estimation was performed using ADMIXTURE version 1.3.0[15]**.** The best number of clusters (K) was
406 determined by comparing 5-fold cross-validation (CV) errors.

407 Inbreeding, heterozygosity and nucleotide diversity were calculated within both GSD populations on
408 the final dataset of 108,817 SNPs. To determine inbreeding coefficients based on runs of homozygosity

(F_{ROH}), runs of homozygosity (ROH) were computed in PLINK version 1.9[65] using the default settings of a ROH length of 1000 kb and a window size of 65 SNPs, as in Pfahler and Distl[70]. The inbreeding was then estimated as the individual's total ROH length divided by the total genome length. ROH-based methods have been shown to perform best in relation to the true inbreeding[71]. Finally, nucleotide diversity (Nei's μ) was calculated per SNP using the --pi specifier in VCFtools[72].

Identification of selection signatures

Within populations

Signatures of selection within the two GSD populations were identified using the integrated haplotype score (iHS) statistic, which measures the extended haplotype homozygosity (EHH) in the genome as an indicator of selective sweeps. The iHS statistic is based on the integrated EHH (iHH_i), which is the integral of the observed decay of EHH away from a specified core allele *i* until the EHH reaches a specified cut-off. Phased genotypes of the final SNP dataset generated by Beagle version 4.1[73] (the phasing in Beagle was performed without specifying a reference population) were used to compute the SNP-wise iHS statistic using hapbin[74], specifying that the iHH should be calculated up to the point at which EHH drops below 0.05 (--cutoff 0.05). As in Voight et al.[17], the standardized iHS (iHS) for a SNP was calculated as

$$iHS = \frac{unstandardized\ iHS - \mu_{unstandardized\ iHS}}{\sigma_{unstandardized\ iHS}}$$

where the $unstandardized\ iHS$ is $\ln(iHH_i/iHH_j)$ for alleles *i* and *j*, and μ and σ are the mean and the standard deviation of the unstandardized iHS estimated from the empirical distribution of SNPs for which the derived allele frequency matches the frequency at the core SNP.

Between populations

To detect divergent signatures of selection between populations, three different approaches were used: the fixation index (F_{ST}), cross-population extended haplotype homozygosity (XP-EHH) and differences between runs of homozygosity (ROH).

First, the F_{ST} analysis was performed using the script described in Talenti et al.[75]. The F_{ST} between UK and Swedish dogs was calculated for each SNP according to the formula reported by Karlsson et al.[76], which is a comparison of the allele frequencies between populations:

$$F_{ST} = \frac{f_1^{UK}\left(f_2^S - f_2^{UK}\right) + f_1^S\left(f_2^{UK} - f_2^S\right)}{\left(f_1^{UK} * f_2^S\right) + \left(f_2^{UK} * f_1^S\right)}$$

where f_1^{UK} and f_2^{UK} are frequencies in the UK population for the two alleles and f_1^S and f_2^S are allele frequencies in the Swedish population. Next, the mean F_{ST} was calculated in 1 Mb sliding windows (window-based F_{ST}) with an overlap between windows of 500 kb, resulting in each SNP being located in exactly one or two windows. To derive a SNP-based value (to select the top 1% for calculating the intersection with other methods as described below), we averaged the window-based F_{ST} for the one or two windows in which the SNP was found.

Second, the XP-EHH statistic[77] was calculated to compare the EHH between populations, i.e. whether alleles are homozygous in one population and polymorphic in the other population. The XP-EHH statistic was calculated for the UK and Swedish populations using phased haplotypes generated by Beagle version 4.1[73] in hapbin[74], as described above.

For the third approach, ROH were computed in PLINK version 1.9[65]. We ran the analysis with the default settings of a ROH length of 1000 kb and a window size of 65 SNPs, as described above[70]. For every SNP, a homozygosity score (ROH_{Prop}) was calculated by dividing the number of dogs with a ROH at a specific SNP by the total number of dogs, such that ROH_{Prop} ranges from 0 to 1, as described in Bertolini et al.[78]. The absolute difference between ROH_{Prop} between populations (ΔROH_{Prop}) was used as statistic to determine which ROH are highly represented in one population but underrepresented in

453 the other population. Therefore, for every SNP, ΔROH_{Prop} values were calculated to identify ROH that
454 are present in the majority of dogs in one population but not in the other.

455 *Gene identification and Gene ontology (GO) analysis*

456 To detect putative genomic regions showing evidence of selection, the most extreme values from the
457 test statistics were selected for both the within- and between-population analyses to define selection
458 signatures. For iHS, SNPs belonging to the top 0.5% of the distribution were selected. For F_{ST}, XP-
459 EHH and ΔROH_{Prop}, the top 1% of each test distribution were selected and the overlap between these
460 top SNPs was determined to identify SNPs that had most extreme values for at least two of the three
461 methods, to reduce the chance of false positive signals. We chose a less stringent threshold for top SNPs
462 for between-population statistics to allow for greater overlap since the three approaches differ in their
463 methodologies and thus the ranking of top SNPs will vary. For a visual representation of target regions
464 under selection between populations, the visualisation tool Circos[79] was used. For every SNP, the
465 ΔROH_{Prop} and XP-EHH scores were plotted. Since the F_{ST} was calculated as a window-based average
466 and Circos required a SNP-based value, we averaged the window-based F_{ST} for the one or two window
467 in which the SNP was found, as described above.

468 The pairwise distances between the top SNPs were calculated and SNPs located within 200 kb were
469 merged into a region. The distance of 200 kb was determined based on the linkage disequilibrium in
470 the genome. First, the squared correlation (r^2) between all pairs of SNPs within 10Mb was calculated
471 in PLINK version 1.9[65]. The average r^2 was then calculated for bins of increasing distance between
472 SNPs to identify the distance around SNPs at which average r^2 drops below 0.5. The longest bin for
473 which average $r^2 \geq 0.5$ was 200 kb.

474 To characterise functional relevance of regions showing evidence of selection, the top SNPs or regions
475 (if multiple SNPs were found within 200 kb) were annotated for genes based on the CanFam3.1 genome
476 assembly[80], using BEDtools 2.27 software[81]. SNPs were annotated considering a flanking region of ±
477 40kb, chosen based on the average between-marker distance of the array (~20kb), which was doubled
478 to account for non-evenly spaced SNPs and SNPs lost through quality-control filtering. The genes

479 detected for these selection signatures were then submitted to Enrichr[27,28] to perform gene set
480 enrichment analyses. Enrichr is an integrative web-based application that compares submitted gene lists
481 to various gene-set libraries; the standard Fisher exact test option was used to calculate P-values for this
482 study.

483 **Characterising trait(s) under selection**

484 We employed two approaches to gain insights into the trait(s) under selection, as detected as genomic
485 selection signatures: (I) we modelled behaviour traits and other dog characteristics as a function of the
486 dog's ancestry based on selection signature regions and (II) we analysed the association within each
487 population between these traits and SNP markers in these regions. For both approaches, we compiled a
488 genotype data set of SNPs within the regions showing evidence of selection; this included SNPs
489 belonging to the top 0.5% of the iHS distribution in UK and Swedish populations and SNPs belonging
490 to the top 1% of F_{ST}, XP-EHH and ΔROH_{Prop} distributions that overlapped between at least two methods.

491 For (I), we repeated the ADMIXTURE analysis as described above, but only used genotypes of SNPs
492 from putatively selected regions to estimate the ancestry. Then, a linear regression was performed, as
493 described in Ref. 82, to model the relationship between the traits and ancestry assignment probabilities.

494 For (II), we analysed the association between the traits and SNP markers within the regions showing
495 evidence of selection, separately for each population. Behaviour traits were adjusted based on other
496 fixed effects as defined in the previous study[67] and treated as quantitative traits, while coat colour
497 ("saddle tan", "sable", "black", "other") and coat length ("long", "short") were treated as categorical
498 traits and not corrected for environmental factors. The association analysis was performed using
499 GEMMA[83], fitting the genomic relationship matrix (based on 108,817 genome-wide SNPs) as a random
500 effect to account for population stratification. To correct for multiple testing, P-values were adjusted
501 using the false discovery rate (FDR).

Data availability

Genotype and phenotype data for the UK dogs is available under CC-BY license from the Dryad Digital Repository[84]. The data for the Swedish dogs is restricted by the Swedish Armed Forces for reasons of national security.

References

1. Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, et al. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. BMC Biology. 2018 Jun 28;16:64.

2. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, et al. Tracking footprints of artificial selection in the dog genome. PNAS. 2010 Jan 19;107(3):1160–5.

3. Larson G, Karlsson EK, Perri A, Webster MT, Ho SYW, Peters J, et al. Rethinking dog domestication by integrating genetics, archeology, and biogeography. PNAS. 2012 Jun 5;109(23):8878–83.

4. Botigué LR, Song S, Scheu A, Gopalan S, Pendleton AL, Oetjens M, et al. Ancient European dog genomes reveal continuity since the Early Neolithic. Nat Commun. 2017 18;8:16082.

5. Mehrkam LR, Wynne C. Behavioral differences among breeds of domestic dogs (Canis lupus familiaris): Current status of the science. Applied Animal Behaviour Science. 2014;155:12–27.

6. Lewis TW, Wiles BM, Llewellyn-Zaidi AM, Evans KM, O'Neill DG. Longevity and mortality in Kennel Club registered dog breeds in the UK in 2014. Canine Genetics and Epidemiology. 2018 Oct 17;5(1):10.

7. Schoenebeck JJ, Ostrander EA. Insights into Morphology and Disease from the Dog Genome Project. Annual Review of Cell and Developmental Biology. 2014;30(1):535–60.

8. Nielsen R. Molecular Signatures of Natural Selection. Annual Review of Genetics. 2005;39(1):197–218.

9. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013 Mar;495(7441):360–4.

10. Kim J, Williams FJ, Dreger DL, Plassais J, Davis BW, Parker HG, et al. Genetic selection of athletic success in sport-hunting dogs. PNAS. 2018 Jul 24;115(30):E7212–21.

11. Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, et al. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nature Communications. 2019 Apr 2;10(1):1489.

12. Ostrander EA, Wayne RK, Freedman AH, Davis BW. Demographic history, selection and functional diversity of the canine genome. Nature Reviews Genetics. 2017 Dec;18(12):705–20.

13. Lord K, Schneider RA, Coppinger R. Evolution of working dogs [Internet]. The Domestic Dog: Its Evolution, Behavior and Interactions with People. 2016 [cited 2019 Oct 8]. Available from: /core/books/domestic-dog/evolution-of-working-dogs/CC5083D37F741470DDFA69AFBB238AB1

14. Friedrich J, Arvelius P, Strandberg E, Polgar Z, Wiener P, Haskell MJ. The interaction between behavioural traits and demographic and management factors in German Shepherd dogs. Applied Animal Behaviour Science [Internet]. 2018 Dec 5 [cited 2018 Dec 12]; Available from: http://www.sciencedirect.com/science/article/pii/S0168159118303265

15. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Jan 9;19(9):1655–64.

16. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97–120.

17. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. PLoS Biol [Internet]. 2006 Mar [cited 2018 Nov 9];4(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1382018/

18. Friedrich J, Brand B, Ponsuksili S, Graunke KL, Langbein J, Knaust J, et al. Detection of genetic variants affecting cattle behaviour and their impact on milk production: a genome-wide association study. Anim Genet. 2016 Feb 1;47(1):12–8.

19. Crouch DJM, Winney B, Koppen WP, Christmas WJ, Hutnik K, Day T, et al. Genetics of the human face: Identification of large-effect single gene variants. PNAS. 2018 Jan 23;115(4):E676–85.

20. Morgan MD, Pairo-Castineira E, Rawlik K, Canela-Xandri O, Rees J, Sims D, et al. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. Nature Communications. 2018 Dec 10;9(1):5271.

21. Valente TS, Baldi F, Sant'Anna AC, Albuquerque LG, Costa MJRP da. Genome-Wide Association Study between Single Nucleotide Polymorphisms and Flight Speed in Nellore Cattle. PLOS ONE. 2016 Jun 14;11(6):e0156956.

22. Luciano M, Huffman JE, Arias-Vásquez A, Vinkhuyzen AA, Middeldorp CM, Giegling I, et al. Genome-wide association uncovers shared genetic effects among personality traits and mood states. Am J Med Genet B Neuropsychiatr Genet. 2012 Sep;0(6):684–95.

23. Smith EN, Bloss CS, Badner JA, Barrett T, Belmonte PL, Berrettini W, et al. Genome-wide association study of bipolar disorder in European American and African American individuals. Mol Psychiatry. 2009 Aug;14(8):755–63.

24. Wang K-S, Liu X-F, Aragam N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. Schizophrenia Research. 2010 Dec 1;124(1):192–9.

25. Sun J, Kuo P-H, Riley BP, Kendler KS, Zhao Z. Candidate genes for schizophrenia: A survey of association studies and gene ranking. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2008;147B(7):1173–81.

26. Ilska J, Haskell MJ, Blott SC, Sánchez-Molano E, Polgar Z, Lofgren SE, et al. Genetic Characterisation of Dog Personality Traits. Genetics. 2017 Jan 1;genetics.116.192674.

27. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013 Apr 15;14:128.

28. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 08;44(W1):W90-97.

29. González-Maeso J, Meana JJ. Heterotrimeric G Proteins: Insights into the Neurobiology of Mood Disorders. Curr Neuropharmacol. 2006 Apr;4(2):127–38.

30. Lipsky RH, Marini AM. Brain-Derived Neurotrophic Factor in Neuronal Survival and Behavior-Related Plasticity. Annals of the New York Academy of Sciences. 2007;1122(1):130–43.

31. Lüscher C, Malenka RC. NMDA Receptor-Dependent Long-Term Potentiation and Long-Term Depression (LTP/LTD). Cold Spring Harb Perspect Biol [Internet]. 2012 Jun [cited 2019 Jun 18];4(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367554/

32. Pasterkamp RJ, Giger RJ. Semaphorin function in neural plasticity and disease. Current Opinion in Neurobiology. 2009 Jun 1;19(3):263–74.

33. Jacobsen JPR, Medvedev IO, Caron MG. The 5-HT deficiency theory of depression: perspectives from a naturalistic 5-HT deficiency model, the tryptophan hydroxylase 2Arg439His knockin mouse. Philos Trans R Soc Lond B Biol Sci. 2012 Sep 5;367(1601):2444–59.

34. de Almeida RMM, Ferrari PF, Parmigiani S, Miczek KA. Escalated aggressive behavior: Dopamine, serotonin and GABA. European Journal of Pharmacology. 2005 Dec 5;526(1):51–64.

35. Quignon P, Herbin L, Cadieu E, Kirkness EF, Hédan B, Mosher DS, et al. Canine Population Structure: Assessment and Impact of Intra-Breed Stratification on SNP-Based Association Studies. PLoS ONE [Internet]. 2007 Dec 19 [cited 2016 Mar 22];2(12). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2129117/

36. Moriya S, Soga T, Wong DW, Parhar IS. Transcriptome composition of the preoptic area in mid-age and escitalopram treatment in male mice. Neuroscience Letters. 2016 May 27;622:67–71.

37. Wiener P, Sánchez-Molano E, Clements DN, Woolliams JA, Haskell MJ, Blott SC. Genomic data illuminates demography, genetic structure and selection of a popular dog breed. BMC Genomics. 2017 Aug 14;18:609.

38. Knutson DC, Mitzey AM, Talton LE, Clagett-Dame M. Mice null for NEDD9 (HEF1) display extensive hippocampal dendritic spine loss and cognitive impairment. Brain Research. 2016 Feb 1;1632:141–55.

39. Patel MM, Silasi-Mansat R, Keshari RS, Sansam CL, Jones DA, Lupu C, et al. Role of Androgen Dependent TFPI-Regulating Protein (ADTRP) in Vascular Development and Function. Blood. 2016 Dec 2;128(22):556–556.

40. Li M, Han Y, Zhou H, Li X, Lin C, Zhang E, et al. Transmembrane protein 170B is a novel breast tumorigenesis suppressor gene that inhibits the Wnt/β-catenin pathway. Cell Death Dis [Internet]. 2018 Jan 24 [cited 2019 Jul 16];9(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5833782/

41. Valnegri P, Huang J, Yamada T, Yang Y, Mejia LA, Cho HY, et al. RNF8/UBC13 ubiquitin signaling suppresses synapse formation in the mammalian brain. Nature Communications. 2017 Nov 2;8(1):1271.

42. Chen C, Chen C, Moyzis R, Stern H, He Q, Li H, et al. Contributions of Dopamine-Related Genes and Environmental Factors to Highly Sensitive Personality: A Multi-Step Neuronal System-Level Approach. PLOS ONE. 2011 Jul 13;6(7):e21636.

43. Nakano N, Maeyama K, Sakata N, Itoh F, Akatsu R, Nakata M, et al. C18 ORF1, a Novel Negative Regulator of Transforming Growth Factor-β Signaling. J Biol Chem. 2014 Feb 5;289(18):12680–92.

44. Meerabux JMA, Ohba H, Iwayama Y, Maekawa M, Detera-Wadleigh SD, DeLisi LE, et al. Analysis of a t(18;21)(p11.1;p11.1) translocation in a family with schizophrenia. Journal of Human Genetics. 2009 Jul;54(7):386–91.

45. Hayward JJ, Castelhano MG, Oliveira KC, Corey E, Balkman C, Baxter TL, et al. Complex disease and phenotype mapping in the domestic dog. Nat Commun. 2016 Jan 22;7:10460.

46. Switonski M, Mankowska M. Dog obesity – The need for identifying predisposing genetic markers. Research in Veterinary Science. 2013 Dec;95(3):831–6.

47. Miller CL, Murakami P, Ruczinski I, Ross RG, Sinkus M, Sullivan B, et al. Two complex genotypes relevant to the kynurenine pathway and melanotropin function show association with schizophrenia and bipolar disorder. Schizophrenia Research. 2009 Sep 1;113(2):259–67.

48. Wang X, Pipes L, Trut LN, Herbeck Y, Vladimirova AV, Gulevich RG, et al. Genomic responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes). PNAS. 2018 Oct 9;115(41):10398–403.

49. Dreger DL, Schmutz SM. A SINE Insertion Causes the Black-and-Tan and Saddle Tan Phenotypes in Domestic Dogs. J Hered. 2011 Sep 1;102(Suppl_1):S11–8.

50. Dreger DL, Parker HG, Ostrander EA, Schmutz SM. Identification of a Mutation that Is Associated with the Saddle Tan and Black-and-Tan Phenotypes in Basset Hounds and Pembroke Welsh Corgis. J Hered. 2013 May 1;104(3):399–406.

51. Zapata I, Serpell JA, Alvarez CE. Genetic mapping of canine fear and aggression. BMC Genomics. 2016;17:572.

52. Bonnet C, Andrieux J, Béri-Dexheimer M, Leheup B, Boute O, Manouvrier S, et al. Microdeletion at chromosome 4q21 defines a new emerging syndrome with marked growth restriction, mental retardation and absent or severely delayed speech. Journal of Medical Genetics. 2010 Jun 1;47(6):377–84.

53. Bhoj E, Halbach S, McDonald-McGinn D, Tan C, Lande R, Waggoner D, et al. Expanding the spectrum of microdeletion 4q21 syndrome: a partial phenotype with incomplete deletion of the minimal critical region and a new association with cleft palate and Pierre Robin sequence. Am J Med Genet A. 2013 Sep;161A(9):2327–33.

54. Fee A, Noble N, Valdovinos MG. Functional Analysis of Phenotypic Behaviors of a 5-Year-Old Male with Novel 4q21 Microdeletion. J Pediatr Neuropsychol. 2015 Dec 1;1(1):36–41.

55. Le-Niculescu H, Balaraman Y, Patel SD, Ayalew M, Gupta J, Kuczenski R, et al. Convergent functional genomics of anxiety disorders: translational identification of genes, biomarkers, pathways and mechanisms. Transl Psychiatry. 2011 May;1(5):e9.

56. Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, et al. Variation of BMP3 Contributes to Dog Breed Skull Diversity. PLOS Genetics. 2012 Aug 2;8(8):e1002849.

57. Rimbault M, Ostrander EA. So many doggone traits: mapping genetics of multiple phenotypes in the domestic dog. Hum Mol Genet. 2012 Oct 15;21(R1):R52-57.

58. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Pielberg GR, Sigurdsson S, et al. Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. PLOS Genet. 2011 Oct 13;7(10):e1002316.

59. Legrand R, Tiret L, Abitbol M. Two recessive mutations in FGF5 are associated with the long-hair phenotype in donkeys. Genet Sel Evol [Internet]. 2014 Sep 25 [cited 2019 Feb 20];46(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4175617/

60. Housley DJE, Venta PJ. The long and the short of it: evidence that FGF5 is a major determinant of canine 'hair'-itability. Animal Genetics. 2006;37(4):309–15.

61. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, Parker HG, et al. Coat Variation in the Domestic Dog Is Governed by Variants in Three Genes. Science. 2009 Oct 2;326(5949):150–3.

62. Carola V, Perlas E, Zonfrillo F, Soini HA, Novotny MV, Gross CT. Modulation of social behavior by the agouti pigmentation gene. Front Behav Neurosci [Internet]. 2014 Aug 1 [cited 2020 Jan 27];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4117936/

63. Illumina I: Canine HD BeadChip. In Data Sheet: DNA Genotyping; 2010. https://www.illumina.com/documents/products/datasheets/datasheet_caninehd.pdf

64. Chen M, Wang J, Wang Y, Wu Y, Fu J, Liu J. Genome-wide detection of selection signatures in Chinese indigenous Laiwu pigs revealed candidate genes regulating fat deposition in muscle. BMC Genet [Internet]. 2018 May 18 [cited 2019 May 30];19. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5960162/

65. Purcell SM, Chang CC. PLINK 1.9 [Internet]. Available from: www.cog-genomics.org/plink/1.9/

66. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

67. Friedrich J, Strandberg E, Arvelius P, Sánchez-Molano E, Pong-Wong R, Hickey JM, et al. Genetic dissection of complex behaviour traits in German Shepherd dogs. Heredity. 2019 Oct 14;1–13.

68. Wilsson E, Sinn DL. Are there differences between behavioral measurement methods? A comparison of the predictive validity of two ratings methods in a working dog program. Applied Animal Behaviour Science. 2012 Nov;141(3–4):158–72.

69. Hsu Y, Serpell JA. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. Journal of the American Veterinary Medical Association. 2003 Nov 1;223(9):1293–300.

70. Pfahler S, Distl O. Effective Population Size, Extended Linkage Disequilibrium and Signatures of Selection in the Rare Dog Breed Lundehund. PLoS One [Internet]. 2015 Apr 10 [cited 2016 Aug 17];10(4). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4393028/

71. Forutan M, Ansari Mahyari S, Baes C, Melzer N, Schenkel FS, Sargolzaei M. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. BMC Genomics. 2018 Jan 27;19(1):98.

72. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

73. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet. 2007 Nov;81(5):1084–97.

74. Maclean CA, Chue Hong NP, Prendergast JGD. hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets. Mol Biol Evol. 2015 Nov;32(11):3027–9.

75. Talenti A, Bertolini F, Pagnacco G, Pilla F, Ajmone-Marsan P, Rothschild MF, et al. The Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep regions. Mamm Genome. 2017 Apr 1;28(3):114–28.

76. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, et al. Efficient mapping of mendelian traits in dogs through genome-wide association. Nature Genetics. 2007 Nov;39(11):1321–8.

77. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007 Oct 18;449(7164):913–8.

78. Bertolini F, Gandolfi B, Kim ES, Haase B, Lyons LA, Rothschild MF. Evidence of selection signatures that shape the Persian cat breed. Mamm Genome. 2016 Apr 1;27(3):144–55.

79. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res [Internet]. 2009 Jun 18 [cited 2019 Jul 17]; Available from: http://genome.cshlp.org/content/early/2009/06/15/gr.092759.109

80. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D754–61.

81. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

82. Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, et al. Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies. PLOS Genetics. 2012 Apr 26;8(4):e1002641.

83. Zhou X, Stephens M. Genome-wide Efficient Mixed Model Analysis for Association Studies. Nat Genet. 2012 Jun 17;44(7):821–4.

84. Friedrich, J. et al. (2020), Data from: Unravelling selection signatures in a single dog breed suggests recent selection for morphological and behavioural traits, [Dataset], Dryad, https://doi:10.5061/dryad.g4f4qrfmr

85. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, et al. A Simple Genetic Architecture Underlies Morphological Variation in Dogs. PLOS Biol. 2010 Aug 10;8(8):e1000451.

86. MacLean EL, Snyder-Mackler N, vonHoldt BM, Serpell JA. Highly heritable and functionally relevant breed differences in dog behaviour. Proceedings of the Royal Society B: Biological Sciences. 2019 Oct 9;286(1912):20190716.

87. Freedman AH, Schweizer RM, Vecchyo DO-D, Han E, Davis BW, Gronau I, et al. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. PLOS Genetics. 2016 Mar 4;12(3):e1005851.

88. Kukekova AV, Johnson JL, Xiang X, Feng S, Liu S, Rando HM, et al. Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. Nature Ecology & Evolution. 2018 Sep;2(9):1479–91.

745 89. Schlamp F, van der Made J, Stambler R, Chesebrough L, Boyko AR, Messer PW. Evaluating
746 the performance of selection scans to detect selective sweeps in domestic dogs. Mol Ecol. 2016
747 Jan;25(1):342–56.

748 90. Saxena R, Voight BF, Lyssenko V, Burtt NP, Bakker PIW de, Chen H, et al. Genome-Wide
749 Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. Science. 2007
750 Jun 1;316(5829):1331–6.

751

752

Tables

Table 1. Top selection signatures within the UK and Swedish GSD populations, showing the ten highest integrated haplotype score (iHS) statistics. SNPs within 200 kb were summarised into selection signature regions.

Chr	Start (Mb)	Stop (Mb)	Distance (Mb)	N_{SNPs}†	iHS peak‡	iHS mean§	Gene(s)□	Phenotypic association††
UK population								
5	29.2	29.8	0.62	16	3.18	2.84	*ENSCAFG00000015899; MMP20; MMP27; MMP7;* ***ENSCAFG00000030873****;* ***BIRC2****; BIRC3;* ***YAP1****;* ***C11orf70****; CEP126; ANGPTL5*	-
12	68.1	68.2	0.06	2	3.22	2.96	***TRAF3IP2***	-
19	33.0	33.1	0.04	4	3.26	2.84	n.a.	-
19	36.0	36.5	0.51	10	3.46	2.93	***NCKAP5***	-
19	36.8	37.0	0.19	5	3.18	2.90	n.a.	-
19	37.5	37.7	0.20	6	3.48	3.19	***TMEM163***	-
19	38.3	38.6	0.31	9	3.19	2.79	***ZRANB3****;* *ENSCAFG00000005064;* ***R3HDM1****; UBXN4*	-
19	39.5	39.5	0.03	2	3.23	2.91	n.a.	-
20	57.6	57.7	0.07	3	3.18	3.10	*ENSCAFG00000031730; ENSCAFG00000023991;* ***ARHGAP45****; ATP5F1D; CIRBP; MIDN;* ***STK11****;* ***SBNO2****; POLR2E*	-
35	7.9	8.1	0.14	4	3.26	3.09	*BMP6;* ***TXNDC5****;* ***BLOC1S5****;* *ENSCAFG00000009583; ENSCAFG00000024482*	-
Swedish population								
4	44.3	n.a.	n.a.	1	3.09	n.a.	***ENSCAFG00000017171***	-
4	46.9	n.a.	n.a.	1	3.27	n.a.	*ENSCAFG00000028841*	-
4	50.0	50.2	0.15	4	3.09	2.90	***ATP10B***	-
4	52.5	n.a.	n.a.	1	3.47	n.a.	***CLINT1***	-
12	66.7	67.2	0.47	10	3.36	3.13	*GPR6;* ***WASF1****;* ***CDC40****;* ***METTL24****;* ***DDO****; SLC22A16; CDK19*	-
12	67.7	n.a.	n.a.	1	3.13	n.a.	***SLC16A10***	-
18	54.9	55.3	0.36	7	3.45	2.99	*LRRC10B; PPP1R32;* ***SYT7****; PGA; DDB1;* ***VWCE****;* ***ENSCAFG00000016314****;* ***SLC15A3****; CD5; VPS37C; CD6*	-

19	50.6	n.a.	n.a.	1	3.12	n.a.	***KIF5C***	-
24	42.4	42.5	0.05	3	3.33	3.05	*RBM38;* ***CTCFL***	-
36	30.1	30.6	0.05	6	3.11	2.82	*GULP1; COL3A1;* ***COL5A2***	-

[†]Number of top SNPs in region
[‡]Standardised absolute iHS of the peak SNP (in that region)
[§]Average standardised absolute iHS across the SNPs of a region
Genes located within and +/- 40 kb around selection signatures. Genes highlighted in bold include a SNP that belongs to the top 0.5% of the test statistic; all others are located within the region or +/- 40 kb around selection signatures
[††]There were no phenotypic associations (behaviour, coat colour or coat length) with FDR-adjusted P-value<0.1 for markers located within the top ten selection signatures within populations.

Table 2. Selection signatures that belonged to the top 1% of the distribution of at least two methods used to detect signatures of different selection between the GSD populations. SNPs within 200 kb were summarised into selection signature regions.

Chr	Start	Stop	N_{SNPs}†	Population	F_{ST}‡	ΔROH_{Prop}§	XP-EHH□	Gene(s)	Phenotypic association††
1	24024856	25483783	61	Sweden	0.12	0.46	NA	***ME2; MRO***; *MC2R; MC5R;* ***ENSCAFG00000000172***; *ENSCAFG00000029562; ENSCAFG00000029833; FAM210A;* ***LDLRAD4***; *ENSCAFG00000023012;* ***MOXD1***; *ENSCAFG00000031561;* ***CTGF***	Chasing*(UK)
9	16472361	16493753	4	UK	0.09	NA	2.81	*KCNJ16; KCNJ2*	-
12	5349354	6130868	44	Sweden	NA	0.27	3.44	*BRPF3;* ***PNPLA1***; ***C12H6orf222***; ***ETV7***; ***PXT1***; *ENSCAFG00000001396;* ***KCTD20***; *STK38; SRSF3; CDKN1A;* ***ENSCAFG00000001418***; *ENSCAFG00000001419;* ***CPNE5***; ***PPIL1***; ***C12H6orf89***; *MTCH1;* ***PI16***; ***FGD2***	Stranger-directed fear**(UK)
12	6466863	6554339	7	Sweden	NA	0.27	3.46	*FGD2;* ***CMTR1***; ***ENSCAFG00000030835***	Separation anxiety* (Sweden)
22	1027334	1140100	6	UK	0.08	0.26	NA	*RNASEH2B*	-
22	1683950	2496568	46	UK	0.12	0.26	NA	***KCNRG***; ***TRIM13***; ***SPRYD7***; ***KPNA3***; *ENSCAFG00000031710;* ***EBPL***; *ENSCAFG00000010362;* ***RCBTB1***; ***PHF11***; ***SETDB2***; ***CAB39L***; ***CDADC1***; *ENSCAFG00000028525; MLNR; FNDC3A*	-
24	22002778	22463326	24	UK	0.07	0.29	NA	*COMMD7;* ***DNMT3B***; ***MAPRE1***; ***EFCAB8***; ***SUN5***; ***BPIFB2***; ***BPIFB6***; ***BPIFB3***; ***BPIFB4***; ***ENSCAFG00000032553***; ***BPIFA2***; *ENSCAFG00000007369;* ***BPIFA3***; *BPIFA1*	Coat colour**(UK)
24	22908179	23816844	37	UK	0.14	0.28	NA	*ENSCAFG00000029918; ENSCAFG00000007430; ENSCAFG00000007435; ENSCAFG00000029879; NECAB3; PXMP4;* ***ZNF341***; ***CHMP4B***; *EIF2S2;* ***RALY***; ***ASIP***; *ENSCAFG00000007508; AHCY;* ***ITCH***; *DYNLRB1;* ***PIGU***; *MAP1LC3A;* ***NCOA6***; *TP53INP2*	Coat colour**(UK)

24	24867975	25952679	64	UK	0.13	0.28	NA	***CNBD2***; ***EPB41L1***; ***AAR2***; ***DLGAP4***; ***MYL9***; ***TGIF2***; ***SLA2***; *TGIF2-C20orf24*; ***NDRG3***; ***DSN1***; ***SOGA1***; ***TLDC2***; ***SAMHD1***; ***RBL1***; ***MROH8***; ***RPN2***; *GHRH*; *MANBAL*; *SRC*	Coat colour**(UK)
32	4172082	4455360	7	UK	0.09	0.27	NA	***ANTXR2***; ***PRDM8***	Coat length**(UK)
32	5350389	5399877	4	UK	0.13	0.26	NA	***PRKG2***	Coat length**(UK) and * (Sweden) Stranger-directed aggression** (Sweden)
32	5609507	5667788	4	UK	0.12	0.26	NA	*ENSCAFG00000008928*; *RASGEF1B*	Coat length** (UK and Sweden)
32	13000437	14125551	44	UK	0.11	0.37	NA	***SNCA***; ***MMRN1***; ***CCSER1***	Coat colour* (UK) Separation anxiety*(UK) Stranger-directed aggression* (Sweden)
32	14527559	14597957	4	UK	0.11	0.38	NA	*ENSCAFG00000009954*	-
32	14952127	15194499	4	UK	0.10	0.28	NA	*ENSCAFG00000009965*	-
34	33480270		1	UK	NA	0.27	2.80		-

†Number of top SNPs in region

‡Fixation index

§Differences between runs of homozygosity

Cross-population extended haplotype homozygosity.

NA indicates that this selection signature was not present in the top 1% of the test distribution

Genes highlighted in bold include a SNP that belongs to the top 1% of the test distribution; all others are located within the region or +/- 40 kb around selection signatures

††Significant phenotypic associations (behaviour, coat colour, coat length) for the UK and Swedish population within selection signature region. P-values were adjusted using False Discovery Rate (FDR), with significant associations determined as adjusted P-values <0.05 (**) and suggestive associations as adjusted P-values <0.1 (*). The population for which the phenotypic association was identified is specified in parentheses.

Figure legends

Figure 1. Principal Component Analysis of the pruned genomic data. Eigenvectors for the first two principal components are plotted and individuals are coloured according to the population of origin. The variances explained by the principal components are given in parentheses.

Figure 2. Ancestry proportions of studied GSDs based on the pruned genomic data assuming three underlying ancestries (K = 3 clusters) as revealed by ADMIXTURE. Each cluster is represented by a colour and the length of the specific coloured segment indicates the dog's proportion of membership in that cluster.

Figure 3. Distribution of integrated haplotype score (iHS) in the UK (upper plot) and Swedish population (lower plot). The red line indicates the threshold for the top 0.5% iHS.

Figure 4. Circos plot for signatures of selection between GSD populations. The plot shows the three statistics used to identify regions under differential selection: differences between runs of homozygosity (ΔROH_{Prop}, outer circle, blue track), cross-population extended haplotype homozygosity (XP-EHH, middle circle, green track) and the fixation index ($F_{ST,}$ inner circle, purple track). The plot indicates concordant evidence in regions on Chr 1, 24 and 32, where peaks can be seen based on all three methods (although not within the top 1% of SNPs for XP-EHH, shown in red for the three methods).

Appendices

Table A1. Description of German Shepherd dog populations. Summary statistics for behaviour traits and other dog attributes within the UK and the Swedish GSD populations.

Table A2. List of SNPs belonging to the top 0.5% of the iHS statistic in the UK and Swedish populations.

Table A3. Lists of SNPs belonging to the top 1% of the F_{ST}, XP-EHH and ΔROH_{Prop} statistics and the SNPs that belonged to the top 1% for at least two methods.

Table A4. Significance of associations between population attributes and genetic ancestries. The proportion of ancestries estimated by ADMIXTURE (cluster 1, cluster 2, cluster 3) based on markers located within selection signature regions were fitted as fixed effects in separate linear models to test their association with different response variables (population attributes: behaviour traits, role of the dog, coat colour and coat length). The P-values for the respective models are shown in the table.

Table A5. Markers located in selection signature regions and showing significant associations (FDR-adjusted P<0.1) with phenotypic traits (behaviour, coat colour, coat length).

Table A6. Overlaps between genes located in selection signature regions and candidate genes for morphological traits and behaviour reported in other studies. A list of candidate genes in canids was compiled using the following references[1, 2, 9, 10, 11, 26, 37, 45, 50, 51, 58, 61, 67, 76, 85-89] and was compared to genes located in regions detected as selection signatures in this study.

Figure A1. Ancestry proportions of GSDs based on genotypes of SNPs from putatively selected regions assuming three underlying ancestries (K = 3 clusters) as revealed by ADMIXTURE. Each cluster is represented by a colour and the length of the specific coloured segment indicates the dog's proportion of membership in that cluster. The labels indicate the origin of the dog (Sweden or UK) and the coat colour (1 = saddle tan, 0 = sable, black or others).

Figure A2. Fine-mapping of target regions under divergent selection between German Shepherd dog populations. Particularly compelling regions that showed evidence of divergent selection in all three selection signature test statistics (SNP window-based F_{ST}, ΔROH_{Prop}, and XP-EHH) are located on Chr 1, 24 and 32. The plots illustrate the FDR-adjusted P-values from association analyses for phenotypic traits (behaviour, coat colour, coat length) (above, "Regional association") and the selection signature test statistics (below, "Selection signatures") for all SNPs in these regions. The plots were created using a modified R code from that of Saxena et al. 2007 [90].

Unravelling selection signatures in a single dog breed suggests recent selection for morphological and behavioural traits

Juliane Friedrich[1], Andrea Talenti[1], Per Arvelius[2], Erling Strandberg[3], Marie J. Haskell[4], Pamela Wiener[1*]

[1]Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, UK

[2]Swedish Armed Forces Dog Training Center, PO Box 194, SE-195 24 MÄRSTA, Sweden

[3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, PO Box 7023, S-750 07 Uppsala, Sweden

[4]Scotland's Rural College (SRUC), Edinburgh, EH9 3JG, UK

*Corresponding author

Pamela Wiener: Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, UK; Telephone: +44 (0)131 651 9100; Fax: +44 (0) 131 651 9105; pam.wiener@roslin.ed.ac.uk

Author contributions

Juliane Friedrich: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing-original draft, Writing-review & editing

Andrea Talenti: Methodology, Software, Visualization, Writing-review & editing

Per Arvelius: Data curation, Project administration, Resources, Supervision, Writing-review & editing

Erling Strandberg: Data curation, Project administration, Resources, Supervision, Writing-review & editing

Marie J. Haskell: Conceptualization, Funding acquisition, Project administration, Supervision, Writing-review & editing

Pamela Wiener: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing-review & editing

Acknowledgements

The authors want to thank all owners of German Shepherd dogs participating in this study for their time and effort to answer the questionnaires and send saliva samples for genotyping. Thanks are also extended to the Kennel Club, the British Association for German Shepherd Dogs, and the German Shepherd Dog Breed Council of Great Britain for assistance in participant recruitment for the UK cohort. Thanks to Zita Polgar, Carol-Anne Duthie and Joanna Warner for assistance in contacting dog owners. We would also like to thank the SAF Dog Training Centre, in particular Lisa Rutström, for recruiting participants for the Swedish cohort, and Susanne Gustafsson and Gabriela Bottani Claros (Swedish University of Agricultural Sciences) for extracting the DNA. We are very thankful to Lee Murphy and his colleagues at the Edinburgh Clinical Research Facility (University of Edinburgh) for

DNA extraction and genotyping of the UK dogs. Further thanks to Vanessa Kapsona for conducting preliminary population structure analyses. Primary funding was provided by the Dogs Trust (UK); further funding was provided by BBSRC Institute Strategic Programme Grants (to the Roslin Institute) and RESAS, Scottish Government (to SRUC). [AU: do you wish to specify grant numbers?]

The authors declare no conflict of interest.

Abstract

Strong selection has resulted in substantial morphological and behavioural diversity across modern dog breeds, which makes dogs interesting model animals to study the underlying genetic architecture of these traits. However, results from between-breed analyses may confound selection signatures for behaviour and morphological features that were co-selected during breed development. In this study, we assess population genetic differences in a unique resource of dogs of the same breed but with systematic behavioural selection in only one population. We exploit these different breeding backgrounds to identify signatures of recent selection. Selection signatures within populations were found on chromosomes 4 and 19, with the strongest signals in behaviour-related genes. Regions showing strong signals of divergent selection were located on chromosomes 1, 24 and 32, and include candidate genes for both physical features and behaviour. Some of the selection signatures appear to be driven by loci associated with coat colour (Chr 24; *ASIP*) and length (Chr 32; *FGF5*), while others showed evidence of association with behaviour. Our findings suggest that signatures of selection within dog breeds have been driven by selection for morphology and behaviour. Furthermore, we demonstrate that combining selection scans with association analyses is effective for dissecting the traits under selection.

Introduction

The development of current dog breeds can be viewed as a unique long-term selection experiment to study the process of domestication[1] as well as short-term evolutionary change as a consequence of intensive breeding[2]. While the domestication of the modern dog (*Canis lupus familiaris*) from wolves took place at least 15,000 years ago[3], with some estimates considerably earlier (e.g. 20,000 to 40,000 years ago[4]), the popularity of dogs has led to ongoing strict selection according to breeding schemes and standards imposed by breed associations and national kennel clubs. The establishment of genetically and phenotypically distinctive breeds by this intense artificial selection pressure has resulted in high intra-species variation for physical and physiological features, disease susceptibility and behaviour traits[5–7], which makes dogs powerful models to investigate the underlying genetic architecture and signatures of selection for various traits.

Genetic manifestation of the development of dog breeds can be seen as selection signatures, genomic regions targeted by natural or artificial selection that exhibit various characteristics, including population differentiation, extreme linkage disequilibrium (LD) and patterns of the haplotype structure (e.g. long-range haplotypes) or mutations in coding region[8]. Accordingly, selection signatures between dog breeds have been reported for physical traits, domestication-related traits and some specific behaviours and have led to the identification of candidate genes, e.g. *IGF1* for body size, *FGF5* for coat length and *HAS2* for skin wrinkling[2], *AMY2B*, *MGAM* and *SGLT1* for adaptation to a starch-rich diet[9] and *TRPM3* and *ROBO1* for athletic success in sport-hunting[10]. In a recent whole-genome sequence study of 144 modern dog breeds, positive human-imposed selection was implicated in the fixation or high prevalence within breeds of a range of morphological characteristics (e.g. ear shape, height, weight)[11]. These recent studies for selection signatures in dogs have focused on between-breed or dog-wolf comparisons and while such studies have allowed detection of signatures related to notable physical features, signatures for more subtle traits like behaviour characteristics may be confounded with or masked by signals for the physical features, which might complicate the interpretation of these signatures as appears to be the case for association signals[12].

In this study, we analysed a single dog breed, the German Shepherd dog (GSD), to detect signals of selection. The breed was established in the late 19th century by crossing multiple breeds, with the initial purpose of creating a sheep herding dog[13] and later use as a general working dog within the military or police. GSDs used in this study originated from two populations, the UK and Sweden; while the UK population represented a random sample of pet, show and working dogs, the Swedish dogs were bred within a breeding program of the Swedish Armed Forces (SAF) and only dogs that pass a behaviour test can become working dogs or be used for breeding. Accordingly, in a previous study[14] we showed that there were significant differences between the two GSD populations for various behaviour traits as measured in a questionnaire, e.g. aggression against strangers or dogs, chasing and playfulness. In contrast, morphological differences between populations were reduced compared to between-breed studies. We hypothesise that by comparing populations of the same breed but with different behaviour-related selection strategies, we may be able to identify selection signatures for behaviour as well as those for physical traits. Furthermore, by applying multiple statistical tests for the detection of selection signatures, we have increased the power to detect true signals of selection. Nonetheless, despite the within-breed approach, one of the main difficulties that remains is the identification of the actual trait(s) under selection. We addressed this issue by characterising the relationship between selection signatures and statistical associations between genotype and phenotype (behaviour and morphological traits) from the same populations. We suggest that this approach, combining population genetics and quantitative genetics methods, may also be applicable in other contexts.

Results and discussion

Genomic structure of populations

Characterising the genetic relationships between individual dogs is a valuable tool to evaluate the genetic structure of GSDs in this study. The underlying population structure in the two GSD populations (250 dogs in total) was explored by applying a principal component analysis (PCA) and ancestry estimation on a pruned SNP data set. The PCA indicated a separation between the UK and Swedish populations based on the first two principal components (PCs), which explained 2.8% and 1.9% of the genetic variance, respectively (Figure 1). With respect to PC1 and PC2, the UK dogs had a broader distribution than the Swedish GSDs, suggesting a stronger founder effect in the Swedish cohort. However, some of the UK GSDs clustered with the Swedish GSDs. The overall separation of the two populations is likely due to the geographical separation and thus primarily independent pedigrees but may also reflect the more recent origins of the Swedish population, with the SAF as the only breeder and the primary goal to breed good working dogs. The partial overlap between the two populations is likely due to the use of external dogs in the SAF breeding program, leading to some shared ancestry. A visual assessment of the ancestry estimation based on the ADMIXTURE program[15] (Figure 2) also revealed a clear discrimination between the UK and Swedish populations. The lowest cross-validation error of 0.55 was identified for three clusters (K=3), with the blue cluster primarily associated with the Swedish population and the red and green clusters primarily associated with the UK population.

The average inbreeding coefficient calculated based on runs of homozygosity (F_{ROH}) was 0.29 ± 0.02 (standard deviation; SD) for Swedish GSDs and 0.31 ± 0.05 for UK GSDs. The significantly lower inbreeding estimate ($P < 0.05$) in the Swedish population might be a consequence of a strategic breeding scheme by the Swedish Armed Forces (SAF). The average nucleotide diversity (μ) was 0.30 ± 0.16 for both populations.

Selection signatures within populations

Selection signatures can be detected within populations by identifying distinctive patterns of linkage disequilibrium (LD). In the event of selective sweeps, favourable genetic variants increase in frequency and form extended haplotypes with neighbouring genomic regions due to LD, as reviewed in Ref. 16. We computed the integrated haplotype score (iHS), which is a variation of the extended haplotype homozygosity (EHH) statistic that aims to detect recent and incomplete selective sweeps within populations[17]. In total, 197 and 142 regions with extreme EHH were detected within the UK and Swedish GSD population, respectively. A list of SNPs belonging to the top 0.5% of the iHS statistic in the UK and Swedish populations is given in Table A2. The iHS statistic identified similar selection signatures in both populations, but the most extreme values differed between populations, as shown by the ten regions with the highest iHS statistics (Figure 3, Table 1). Regions with the highest iHS for the UK population were located on Chr 19 at 36.0 – 36.5 Mb and 37.5 – 37.7 Mb. A single marker on Chr 4 at 52.5 Mb showed the highest iHS in the Swedish population, followed by a region on Chr 18 at 54.9 – 55.3 Mb. The SNPs identified by iHS were further tested for their association with different traits (coat colour, coat length and behaviour) separately for each population to identify the putative trait under selection.

The genes located within or closest to the ten most extreme values of iHS (positional candidate genes) identified within populations (Table 1) have been previously associated with behaviour. Regarding those on Chr 19, variants in *TMEM163* (transmembrane protein 163) were associated with active behaviour in an open-field test involving cattle[18]. However, *TMEM163* is also a functional candidate for physical features, e.g. for eye width and depth[19] and hair colour[20] in humans. *NCKAP5* (NCK associated protein 5) was also identified as candidate gene for temperament in cattle[21] and has been associated with numerous neurological conditions in humans[22–24].

The iHS peak on Chr 4 in the Swedish population points to the *CLINT1* (Clathrin Interactor 1) gene. This gene is reported to be among the top risk genes for the susceptibility to schizophrenia in humans[25]

and markers near *CLINT1* were suggestive peaks associated with barking tendency in a genome-wide association study of behaviour traits in Labrador retrievers[26].

We conducted a gene list enrichment analysis with Enrichr[27,28] of the 256 and 338 genes that were located in and close to (within 40 kb of) the regions of the top 0.5% iHS in the UK and Swedish populations, respectively. No pathways were significantly enriched after accounting for multiple testing, however, Panther pathway analyses indicated nominally significant ($P < 0.05$) functional enrichment of several pathways for the UK population: "heterotrimeric G-protein signalling -Gi alpha and Gs alpha mediated" ($P = 0.01$; genes: *GRK4, GRK7, RGS12, ADCY2, ADRA2C, DRD2*), "Alzheimer disease-presenilin" ($P = 0.02$; *TRPC6, MMP7, MMP27, RBPJ, MMP20*), "heterotrimeric G-protein signalling -Gq alpha and Go alpha mediated" ($P = 0.02$; *GRK4, GRK7, CACNA1A, RGS12, DRD2*), "ionotropic glutamate receptor" ($P = 0.03$; *CACNA1A, SLC17A8, GRIA4*) and "axon guidance mediated by semaphorins" ($P = 0.03$*; CRMP1, FYN*). All of these functions have been shown to be relevant for behaviour among other functions, e.g. heterotrimeric G proteins in mood disorders, as reviewed in Ref. 29, ionotropic glutamate receptors for long term synaptic plasticity, as reviewed in Ref. 30, 31 and semaphorins in neuronal structure, as reviewed in Ref. 32. Nominally significant pathways for the Swedish population were "5-Hydroxytryptamine degradation" ($P = 0.003$; *ALDH3A2*, *ALDH3A1*), "apoptosis signaling" ($P = 0.01$; *MAP2K3, CASP9, DAXX, BAK1, BIRC2, BIRC3*) and "Thyrotropin-releasing hormone receptor signaling" ($P = 0.03$; *PLCE1, STX3, TRHR*). 5-hydroxytryptamine (serotonin) is an important neurotransmitter and plays a key role in numerous behavioural disorders and characteristics, e.g. depression[33] and aggressiveness[34].

Selection signatures between populations

Another approach to identify signatures of selection is the comparison of genetic variation (e.g. allele frequencies or haplotype structure) between different populations. Accordingly, signatures of differential selection between the two GSD populations were analysed employing three different tests: the fixation index (F_{ST}), the cross-population extended haplotype homozygosity (XP-EHH) and differences between ROH (ΔROH_{Prop}). F_{ST} was calculated to determine genetic differentiation between UK and Swedish GSD populations. Low genome-wide genetic differentiation was detected for the single SNP-based statistic ($F_{ST} = 0.021 \pm 0.029$) and for the SNP window-based statistic ($F_{ST} = 0.021 \pm 0.016$), consistent with previous within-dog-breed estimates [35].

We scanned the genome for regions of genetic differentiation within overlapping 1 Mb windows and found 17 distinctive peaks that comprise the top 1% window-based F_{ST} values on Chr 1, 9, 20, 22, 24, 29, 30 and 32, with values ranging from 0.07 to 0.16 (Table A3). The highest F_{ST} value (0.16) was found for a region on Chr 24 (22.0 – 24.5 Mb), which contains 46 genes. Among these genes are several with functions in physical characteristics and behaviour, e.g. *SPAG4* and *SUN5* involved in cytoskeletal anchoring, *NCOA6* involved in glucocorticoid and corticosteroid receptor signalling and *ASIP* and *RALY* associated with skin and fur pigmentation. Furthermore, seven members of the bactericidal/permeability-increasing (BPI) fold-containing (BPIF) superfamily of genes are located in this region (*BPIFB2, BPIFB6, BPIFB3, BPIFB4, BPIFA2, BPIFA3, BPIFA1* and *BPIFB1*). It was shown that these genes play a role in the innate immune system and lipoprotein metabolism, but also in the brain's response to oxidative stress (ageing), relevant for neuropsychiatric diseases[36]. Interestingly, high F_{ST} for Labrador retriever populations differentiated based on their coat colour and function (gundog and showdog) was also detected in the same region on Chr 24 (22.4 – 22.8 Mb) in a previous study[37].

While the F_{ST} statistic detects differences in allele frequencies between populations, the XP-EHH test, an approach based on linkage disequilibrium, is designed to detect regions that are fixed (or nearly fixed) in one population but remain segregating in the other population. Extreme high (positive) and

low (negative) scores are indicators of a region under strong positive selection in the UK and Swedish population, respectively. The region including the SNP with the highest score (3.4) for the UK population was located on Chr 35 (11.0 - 11.5 Mb) and contains three genes (*NEDD9, ADTRP*, and *TMEM170B)* (Table A3). The *NEDD9* (Neural Precursor Cell Expressed, Developmentally Down-Regulated 9) gene has been shown to be associated to cognitive impairment in mice[38], *ADTRP* is important for vascular development and function in mouse and zebrafish[39] and *TMEM170B* has been reported to be downregulated in TCGA human breast cancer data[40]. The region with the highest absolute score (3.8) for the Swedish population was located on Chr 12 (3.6-7.5 Mb). This region contains 59 genes; *RNF8* and *TBC1D22B* are closest to the SNP with the most extreme score. The ubiquitin gene *RNF8* (ring finger protein 8) plays a role in the immune system and has also been linked to autism; a recent study in *RNF8* knockout mice indicated a role of this gene in synapse formation and cerebellar-dependent learning abilities[41]. The function of *TBC1D22B* is largely unknown but it may encode a GTPase-activating protein.

As a third approach to identifying differential selection between the populations, we identified the regions showing differences in extended homozygosity. To identify these selection signatures, [EA: added a comma] we calculated the between-population differences in runs of homozygosity (ΔROH_{Prop}), which describes the difference in the proportion of dogs with an ROH of a specified length at a given SNP. The average ΔROH_{Prop} value across the genome was low (0.07 ± 0.06), indicating considerable overlap of ROH between the UK and Swedish populations. However, some regions with ROH were predominantly present in only one population (Table A3). The highest absolute ΔROH_{Prop} indicating selection signatures in the UK population were found on Chr 17 and 32: the ROH mapped to Chr 17 (8.3 - 8.4 Mb) and Chr 32 (13.3 - 13.4 Mb) were present in over 70% of the UK dogs but less than 40% of the Swedish dogs. The genes located in these regions are *GREB1, NTSR2,* and *LPIN1* on Chr 17, with no characterised genes in the Chr 32 region. The neurotensin gene *NTSR2* is involved in dopamine modulation and a SNP in this gene has been tested in a polygenic model of highly sensitive personality in humans[42]. *LPIN1* plays a prominent role in lipid metabolism regulating adipocyte differentiation and co-regulating other genes involved in lipid metabolism. The highest absolute

241 ΔROH_{Prop} indicating selection signatures in the Swedish population was found on Chr 1: a ROH mapped
242 to Chr 1 (24.7 to 25.5 Mb) was present in 90% of the Swedish dogs but only in 42% of the UK dogs
243 and contains the genes *LDLRAD4*, *MOXD1* and *CTGF* (see below).

244 *Target regions for divergent selection signatures between populations*
245 In the detection of selection signatures, the application of multiple approaches is recommended to
246 reduce the rate of false positive signals[16]. To identify target regions under differential selection in the
247 two GSD populations, we selected regions from the 99th percentile (top 1%) of each score distribution
248 (SNP window-based F_{ST}, ΔROH_{Prop}, and XP-EHH) and searched for intersecting signals between two
249 or three of the approaches. Using this criterion, we identified 433 SNPs (Table A3), with the greatest
250 overlap between the SNP window-based F_{ST} and ΔROH_{Prop} statistics (374 SNPs). No SNPs were
251 detected by all three approaches. The 433 SNPs were located in 16 candidate selected regions on Chr
252 1, 9, 12, 22, 24, 32 and 34, which harbour 114 genes in total (Table 2; Figure 4). One Panther pathway
253 was nominally significantly ($P < 0.05$) enriched by these 114 genes: "p53 pathway feedback loops" (P
254 $= 0.03$; *CDKN1A, RBL1*). The SNPs identified as under divergent selection by these analyses were
255 further tested for their association with different traits (coat colour, coat length and behaviour)
256 separately for each population to identify the putative trait under selection.

257 A visual inspection of the Circos plot (Figure 4), which illustrates the results for the three approaches,
258 indicates regions on Chr 1, 24 and 32 where peaks can be seen based on all three methods, although not
259 belonging to the top 1% for XP-EHH. Linear plots for these three regions illustrate the results from
260 association analyses for traits with SNPs located in that region that have adjusted $P < 0.1$ ("Regional
261 association") and the selection signature test statistics ("Selection signatures") (Figure A2). The specific
262 population showing evidence of selection can be determined by the ΔROH_{Prop} or XP-EHH score. Three
263 regions showing evidence of selection in the Swedish population are located on Chr 1 (24.0 – 24.1, 24.4
264 – 25.1 and 25.3 – 25.9 Mb; 17 genes), each harbouring several interesting candidate genes. The
265 *LDLRAD4 (*low density lipoprotein receptor class A domain containing 4) gene inhibits transforming
266 growth factor-β signalling[43] and is a putative schizophrenia-related gene[44]. Another growth factor-

related gene in this region is *CTGF* (connective tissue growth factor). Other candidates for genes under selection in this region are the G-protein-associated melanocortin receptor genes *MC2R* and *MC5R*. *MC2R* (also known as the adrenocorticotropic hormone receptor gene, *ACTHR*) is a major modulator of glucocorticoid secretion regulation. *MC5R* has been associated with a range of phenotypes, including shedding and fur length in dogs[45], fatness in pigs, reviewed by Ref. 46, and psychiatric disorders in humans[47]. It was also differentially expressed in the brains of aggressive and tame foxes[48]. These reported associations with different traits highlight one of the difficulties in identifying phenotypic targets of selection. In our analysis, we found no significant associations (FDR-adjusted $P < 0.05$) between any of the selection signatures on Chr 1 with behaviour traits, coat colour or coat length, but there was a suggestive association (FDR-adjusted $P < 0.1$) with chasing behaviour in the UK population (Table 2). Regarding fur shedding, GSDs as a breed are considered to be shedders, making it unlikely that there are large differences between the two populations for this trait.

Regions showing evidence of selection in the UK population are located on Chr 24 and 32. The Chr 24 candidate region under selection (22.9 – 23.8 Mb; 18 genes) in the UK population comprises well-known genes associated with black-and-tan and saddle-tan coat colour in dogs (*ASIP, RALY*)[49,50]. We found highly significant associations in between coat colour and SNPs in this region showing evidence of selection (Table 2, Figure A2). The saddle and tan/ black and tan coat colour was the dominant coat colour in the UK GSDs while sable was predominant in the Swedish population (Table A1). The region on Chr 32 (5.4 – 5.7 Mb; 3 genes) encompasses two behaviour- and growth-related candidate genes: *PRKG2* and *RASGEF1B*. *RASGEF1B* (RasGEF domain family member 1B) has been identified as a positional candidate gene for dog rivalry in a genome-wide association study across multiple dog breeds[51]. Several case studies have been carried out in humans on chromosomal diseases related to a microdeletion of loci homologous to the region on Chr 4 comprising the *PRKG2* and *RASGEF1B* genes[52–54]. The loss of these genes leads to growth restriction, aggression, self-injurious behaviours and mental retardation in affected individuals. The association analysis revealed a significant association between SNPs in this region and aggressive behaviour towards strangers in the Swedish GSD population and *PRKG2* has previously been reported as a top candidate gene for anxiety in mice[55].

294 However, the region on Chr 32 is in close proximity to the *BMP3* gene associated with skull
295 morphology[56] and the *FGF5*[2] gene associated with coat length in dogs. Regarding *BMP3*, differences
296 in skull morphology have not previously been identified in GSDs nor have they been shown to carry a
297 derived allele in this gene previously associated with brachycephaly[56], thus selection on skull
298 morphology seems unlikely. However, we also found a highly significant association with coat length
299 in both populations (Table 2, Figure A2), suggesting that this trait drives the selection signature on Chr
300 32 (via *FGF5*).

301 **Which traits are under selection?**

302 One of the main difficulties in interpreting genomic selection signatures is the identification of the
303 actual trait(s) under selection. In dogs, the traits under selection are assumed to be primarily related to
304 physical traits (e.g. skull shape, coat colour, body size) and/or behaviour[57]. While between-breed studies
305 have greatly contributed to the understanding of the genetic control of physical traits[11,58], addressing
306 behaviour genetics by performing across-breed selection signature analyses is likely to be challenging
307 because breeds differ in multiple characteristics, including both behaviour and these physical traits,
308 many of which show Mendelian inheritance and thus tend to show very strong signals.

309 We employed several approaches to characterise the relationship between the detected selection
310 signatures and phenotypic traits that were recorded for these populations. First, we repeated the
311 ADMIXTURE analysis using only genotypes from SNPs identified as selection signatures (Figure A1)
312 and fitted the ancestry assignment probabilities to the three individual clusters that were detected as
313 factors in linear models for the phenotypes. We observed significant associations between UK
314 (primarily associated with cluster 1) and Swedish (cluster 3) ancestries and some behaviour traits
315 (Stranger-directed interest, Dog-directed fear) (Table A4). Furthermore, highly significant associations
316 were identified between the ancestries and other dog characteristics, including the function of the dog
317 (working, pet or show dog), coat length and coat colour (Table A4). These results demonstrate a
318 statistical association between these phenotypes and the dog's genotypes in the selection signature
319 regions.

We then performed association analyses for behaviour traits, coat length and coat colour within each population only for markers within selection signature regions. We identified 87 SNPs with FDR-adjusted $P < 0.05$ associated with coat length, coat colour, human-directed playfulness, stranger-directed aggression, stranger directed fear and dog-directed fear (Table A5) in at least one of the populations. The striking significant associations for coat colour (lowest FDR-adjusted $P = 3.37 \times 10^{-14}$) and coat length (lowest FDR-adjusted $P = 1.13 \times 10^{-25}$), comprising regions on Chr 24 and 32, respectively, have previously been identified for these traits[49,59–61] (Table 2).

As discussed above, previous studies on selection signatures in dogs have generally focused on inter-breed or dog-wolf comparisons and primarily detected selection signatures (and thus candidate genes) for physical features, e.g. body size, coat characteristics and skeletal morphology[2,11,58]. Some studies, however, also identified signatures for neural crest development[1] or brain function and nervous system development[9], which might be relevant for behaviour especially in regard to domestication. We compiled a list of candidate genes reported in previous genomic analyses of phenotype associations and selection signatures in canids (dogs, wolves, foxes) focused on morphology and behaviour and compared them to genes located in regions showing evidence of selection in our study (Table A6, note that the number of overlapping genes is not informative for identifying the trait under selection because the number of reported candidate genes differs substantially between studies). The biological functions of genes in common between the two lists are diverse and include a number of genes that have been associated with behaviour. Major candidate genes for physical features in dogs, e.g. *IGF1*, *SMAD2*, *FGF5* and *BMP3*, as reviewed in Ref. 7, were not detected within selection signatures in our study. However, *FGF5*, which has previously been associated with coat length, is located in close proximity to the selection signature on Chr 32 and we detected a highly significant association with coat length for this region (*BMP3*, associated with skull morphology, is also located near this region, but as discussed above, our data does not support a signature of selection associated with this trait). We also detected well-described genes associated with coat colour (Chr 24: *ASIP, RALY*). Together these results suggest that selection for morphological traits (coat length and coat colour) has driven differences between the two populations in the genomic regions on Chr 24 and 32. In contrast, the region we

detected on Chr 1 showed an association with Chasing in the UK population and comprises candidate genes with functions in behaviour, but was not associated with morphological traits that we measured. Moreover, some of the selection signature regions showed associations with both morphological and behaviour traits, e.g. the region on Chr 32 was associated with both Stranger-directed aggression and coat length in the Swedish population (Table 2). Furthermore, genes associated with physical appearance like *ASIP* have previously been associated with behaviour traits, e.g. social behaviour in mice[62]. Thus, [EA: added a comma] it is possible that some of the selection signatures we detected are also associated with multiple traits.

Limitations of the study

By comparing UK and Swedish GSDs, we hypothesised that we would be able to detect selection signatures for behaviour because behaviour was the main selection target in the Swedish population. However, we found that the geographical origin of the dogs was confounded with other attributes, e.g. coat colour and length. We addressed the issue of which trait(s) were under selection by characterising the relationship between selection signatures and associations with phenotypic attributes (behaviour, coat length, coat colour), recognizing that the sample size for the association analyses within populations was small and therefore these results should be interpreted with caution. In addition, measurements on other morphological traits (e.g. body size and weight) were not available, but these might also be under selection and should be considered in future studies. We conclude that our study of German Shepherd dogs has identified selection signatures probably driven by selection for coat colour and length (e.g. at the *ASIP* and *FGF5* genes) as well as other signatures that may be related to differential selection for behaviour between the Swedish and UK populations. Functional analyses are needed to test whether the identified candidate genes within regions showing evidence of selection do influence dog behaviour characteristics.

Material and methods

SNP genotyping and quality control

DNA was extracted from saliva samples collected with Performagene PG-100 swabs (UK population) or blood samples (Swedish population). The genotyping was performed using the CanineHD Whole-Genome Genotyping BeadChip[63] featuring 172,115 SNPs. The data was filtered for sample call rate of > 90%, SNP [EA: removed an extra space] call rate > 98%, reproducibility (GTS) > 0.6 and low or confounded signal characterised by AB R mean (mean normalized intensity of the AB cluster) > 0.3 in GenomeStudio version 2.0. Minor allele frequency filtering of > 0.01 was used to include rare but informative variants, leaving a final dataset of 108,817 SNPs for analyses. Genotype information was available for 741 GSDs. Following further sample-based quality control, closely related dogs were removed following the procedure described in Chen et al.[64]. Briefly, a pruned genotype data set to remove closely related dogs was created for SNPs with MAF > 0.05 using PLINK version 1.9[65]: based on the variance inflation factor, a function of the multiple correlation coefficient of a given SNP regressed on all other SNPs within a window (using default parameters: window size = 50 SNPs, overlapping SNPs for shifting windows = 5, the variance inflation factor threshold = 2). Then, GCTA version 1.24.7[66] was used to compute the genetic relationship matrix and to remove one dog per pair with a genetic relationship higher than 0.2 (equivalent to 2nd degree or closer relatives) leaving a final set of 182 UK and 68 Swedish GSDs for subsequent analyses.

Samples and phenotypes

The GSDs used in this analysis originated from the UK and Sweden. For the UK population, GSDs that were at least two years old and registered with the UK Kennel Club were recruited via email to participate in a study on behaviour genetics[14,67]. GSDs from the UK population were bred by multiple breeders and primarily were pet dogs. All GSDs from the Swedish population were bred within the breeding program of the Swedish Armed Forces (SAF) starting in 2004 with the purpose of becoming working dogs. The strongest systematic selection pressure in the SAF breeding program is for behavior

396 traits. Briefly, puppies were raised at the SAF, weaned at the age of 8 weeks and then fostered by
397 members of the Swedish public[68]. After a behaviour test at the age of 15-18 months, some dogs started
398 working with the SAF, Swedish Police or other authorities and companies, and/or were selected as
399 breeding animals, whereas others were kept as pet dogs. For the Swedish population, owners, trainers
400 or handlers of GSDs bred within the breeding program of the SAF were invited via email or letter to
401 participate in the study. Several phenotypes were analysed. Data on GSD behaviour was assessed using
402 the Canine Behaviour and Research Questionnaire (C-BARQ)[69]. The C-BARQ consists of questions
403 related to training and obedience, aggression, fear and anxiety, separation-related behaviour,
404 excitability, attachment and attention seeking, and miscellaneous behaviours. To calculate the
405 behaviour traits, a principal component analysis (PCA) was applied to the data to condense the questions
406 to a smaller number of 13 components, as described in Ref. 14. The dogs' scores for the 13 components,
407 adjusted for fixed effects (excluding cohort) as described in Ref. 67, were considered as adjusted
408 behaviour traits in the subsequent analyses. Other dog characteristics (e.g. sex, coat colour, coat length,
409 role) were assessed using a lifestyle survey[14]. Summary statistics for behaviour traits and other
410 characteristics within the two GSD populations are given in supplementary material (Table A1).

411

412 **Genomic structure of populations**

413 To characterise the genomic structure of the GSD populations, a principal component analysis (PCA)
414 and a cluster analysis were performed. PLINK version 1.9[65] with default parameters was used to create
415 a pruned SNP dataset with reduced linkage disequilibrium (LD) between SNPs, leaving a pruned dataset
416 of 9,180 SNPs. This dataset was employed only to characterise the genomic structure of populations,
417 via PCA and ADMIXTURE analyses. The PCA was performed in PLINK version 1.9 [65] and ancestry
418 estimation was performed using ADMIXTURE version 1.3.0[15]**.** The best number of clusters (K) was
419 determined by comparing 5-fold cross-validation (CV) errors.

420 Inbreeding, heterozygosity and nucleotide diversity were calculated within both GSD populations on
421 the final dataset of 108,817 SNPs. To determine inbreeding coefficients based on runs of homozygosity

(F_{ROH}), runs of homozygosity (ROH) were computed in PLINK version 1.9[65] using the default settings of a ROH length of 1000 kb and a window size of 65 SNPs, as in Pfahler and Distl[70]. The inbreeding was then estimated as the individual's total ROH length divided by the total genome length. ROH-based methods have been shown to perform best in relation to the true inbreeding[71]. Finally, nucleotide diversity (Nei's μ) was calculated per SNP using the --pi specifier in VCFtools[72].

Identification of selection signatures

Within populations

Signatures of selection within the two GSD populations were identified using the integrated haplotype score (iHS) statistic, which measures the extended haplotype homozygosity (EHH) in the genome as an indicator of selective sweeps. The iHS statistic is based on the integrated EHH (iHH_i), which is the integral of the observed decay of EHH away from a specified core allele *i* until the EHH reaches a specified cut-off. Phased genotypes of the final SNP dataset generated by Beagle version 4.1[73] (the phasing in Beagle was performed without specifying a reference population) were used to compute the SNP-wise iHS statistic using hapbin[74], specifying that the iHH should be calculated up to the point at which EHH drops below 0.05 (--cutoff 0.05). As in Voight et al.[17], the standardized iHS (iHS) for a SNP was calculated as

$$iHS = \frac{unstandardized\ iHS - \mu_{unstandardized\ iHS}}{\sigma_{unstandardized\ iHS}}$$

where the $unstandardized\ iHS$ is $\ln(iHH_i/iHH_j)$ for alleles *i* and *j*, and μ and σ are the mean and the standard deviation of the unstandardized iHS estimated from the empirical distribution of SNPs for which the derived allele frequency matches the frequency at the core SNP.

Between populations

To detect divergent signatures of selection between populations, three different approaches were used: the fixation index (F_{ST}), cross-population extended haplotype homozygosity (XP-EHH) and differences between runs of homozygosity (ROH).

First, the F_{ST} analysis was performed using the script described in Talenti et al.[75]. The F_{ST} between UK and Swedish dogs was calculated for each SNP according to the formula reported by Karlsson et al.[76], which is a comparison of the allele frequencies between populations:

$$F_{ST} = \frac{f_1^{UK}\left(f_2^S - f_2^{UK}\right) + f_1^S\left(f_2^{UK} - f_2^S\right)}{\left(f_1^{UK} * f_2^S\right) + \left(f_2^{UK} * f_1^S\right)}$$

where f_1^{UK} and f_2^{UK} are frequencies in the UK population for the two alleles and f_1^S and f_2^S are allele frequencies in the Swedish population. Next, the mean F_{ST} was calculated in 1 Mb sliding windows (window-based F_{ST}) with an overlap between windows of 500 kb, resulting in each SNP being located in exactly one or two windows. To derive a SNP-based value (to select the top 1% for calculating the intersection with other methods as described below), we averaged the window-based F_{ST} for the one or two windows in which the SNP was found.

Second, the XP-EHH statistic[77] was calculated to compare the EHH between populations, i.e. whether alleles are homozygous in one population and polymorphic in the other population. The XP-EHH statistic was calculated for the UK and Swedish populations using phased haplotypes generated by Beagle version 4.1[73] in hapbin[74], as described above.

For the third approach, ROH were computed in PLINK version 1.9[65]. We ran the analysis with the default settings of a ROH length of 1000 kb and a window size of 65 SNPs, as described above[70]. For every SNP, a homozygosity score (ROH_{Prop}) was calculated by dividing the number of dogs with a ROH at a specific SNP by the total number of dogs, such that ROH_{Prop} ranges from 0 to 1, as described in Bertolini et al.[78]. The absolute difference between ROH_{Prop} between populations (ΔROH_{Prop}) was used as statistic to determine which ROH are highly represented in one population but underrepresented in

the other population. Therefore, for every SNP, ΔROH_{Prop} values were calculated to identify ROH that are present in the majority of dogs in one population but not in the other.

Gene identification and Gene ontology (GO) analysis

To detect putative genomic regions showing evidence of selection, the most extreme values from the test statistics were selected for both the within- and between-population analyses to define selection signatures. For iHS, SNPs belonging to the top 0.5% of the distribution were selected. For F_{ST}, XP-EHH and ΔROH_{Prop}, the top 1% of each test distribution were selected and the overlap between these top SNPs was determined to identify SNPs that had most extreme values for at least two of the three methods, to reduce the chance of false positive signals. We chose a less stringent threshold for top SNPs for between-population statistics to allow for greater overlap since the three approaches differ in their methodologies and thus the ranking of top SNPs will vary. For a visual representation of target regions under selection between populations, the visualisation tool Circos[79] was used. For every SNP, the ΔROH_{Prop} and XP-EHH scores were plotted. Since the F_{ST} was calculated as a window-based average and Circos required a SNP-based value, we averaged the window-based F_{ST} for the one or two windows [EA: added an s] in which the SNP was found, as described above.

The pairwise distances between the top SNPs were calculated and SNPs located within 200 kb were merged into a region. The distance of 200 kb was determined based on the linkage disequilibrium in the genome. First, the squared correlation (r^2) between all pairs of SNPs within 10Mb was calculated in PLINK version 1.9[65]. The average r^2 was then calculated for bins of increasing distance between SNPs to identify the distance around SNPs at which average r^2 drops below 0.5. The longest bin for which average $r^2 \geq 0.5$ was 200 kb.

To characterise functional relevance of regions showing evidence of selection, the top SNPs or regions (if multiple SNPs were found within 200 kb) were annotated for genes based on the CanFam3.1 genome assembly[80], using BEDtools 2.27 software[81]. SNPs were annotated considering a flanking region of ± 40kb, chosen based on the average between-marker distance of the array (~20kb), which was doubled to account for non-evenly spaced SNPs and SNPs lost through quality-control filtering. The genes

492 detected for these selection signatures were then submitted to Enrichr[27,28] to perform gene set
493 enrichment analyses. Enrichr is an integrative web-based application that compares submitted gene lists
494 to various gene-set libraries; the standard Fisher exact test option was used to calculate P-values for this
495 study.

496 **Characterising trait(s) under selection**

497 We employed two approaches to gain insights into the trait(s) under selection, as detected as genomic
498 selection signatures: (I) we modelled behaviour traits and other dog characteristics as a function of the
499 dog's ancestry based on selection signature regions and (II) we analysed the association within each
500 population between these traits and SNP markers in these regions. For both approaches, we compiled a
501 genotype data set of SNPs within the regions showing evidence of selection; this included SNPs
502 belonging to the top 0.5% of the iHS distribution in UK and Swedish populations and SNPs belonging
503 to the top 1% of F_{ST}, XP-EHH and ΔROH_{Prop} distributions that overlapped between at least two methods.

504 For (I), we repeated the ADMIXTURE analysis as described above, but only used genotypes of SNPs
505 from putatively selected regions to estimate the ancestry. Then, a linear regression was performed, as
506 described in Ref. 82, to model the relationship between the traits and ancestry assignment probabilities.

507 For (II), we analysed the association between the traits and SNP markers within the regions showing
508 evidence of selection, separately for each population. Behaviour traits were adjusted based on other
509 fixed effects as defined in the previous study[67] and treated as quantitative traits, while coat colour
510 ("saddle tan", "sable", "black", "other") and coat length ("long", "short") were treated as categorical
511 traits and not corrected for environmental factors. The association analysis was performed using
512 GEMMA[83], fitting the genomic relationship matrix (based on 108,817 genome-wide SNPs) as a random
513 effect to account for population stratification. To correct for multiple testing, P-values were adjusted
514 using the false discovery rate (FDR).

Data availability

Genotype and phenotype data for the UK dogs is available under CC-BY license from the Dryad Digital Repository[84] [AU: please ensure this link is live prior to production – ED] .

The data for the Swedish dogs is restricted by the Swedish Armed Forces for reasons of national security.

References

1. Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, et al. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. BMC Biology. 2018 Jun 28;16:64.

2. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, et al. Tracking footprints of artificial selection in the dog genome. PNAS. 2010 Jan 19;107(3):1160–5.

3. Larson G, Karlsson EK, Perri A, Webster MT, Ho SYW, Peters J, et al. Rethinking dog domestication by integrating genetics, archaeology, and biogeography. PNAS. 2012 Jun 5;109(23):8878–83.

4. Botigué LR, Song S, Scheu A, Gopalan S, Pendleton AL, Oetjens M, et al. Ancient European dog genomes reveal continuity since the Early Neolithic. Nat Comms. 2017 18;8:16082.

5. Mehrkam LR, Wynne C. Behavioral differences among breeds of domestic dogs (Canis lupus familiaris): Current status of the science. Applied Animal Behaviour Science. 2014;155:12–27.

6. Lewis TW, Wiles BM, Llewellyn-Zaidi AM, Evans KM, O'Neill DG. Longevity and mortality in Kennel Club registered dog breeds in the UK in 2014. Canine Genetics and Epidemiology. 2018 Oct 17;5(1):10.

7. Schoenebeck JJ, Ostrander EA. Insights into Morphology and Disease from the Dog Genome Project. Annual Review of Cell and Developmental Biology. 2014;30(1):535–60.

8. Nielsen R. Molecular Signatures of Natural Selection. Annual Review of Genetics. 2005;39(1):197–218.

9. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013 Mar;495(7441):360–4.

10. Kim J, Williams FJ, Dreger DL, Plassais J, Davis BW, Parker HG, et al. Genetic selection of athletic success in sport-hunting dogs. PNAS. 2018 Jul 24;115(30):E7212–21.

11. Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, et al. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nature Communications. 2019 Apr 2;10(1):1489.

12. Ostrander EA, Wayne RK, Freedman AH, Davis BW. Demographic history, selection and functional diversity of the canine genome. Nature Reviews Genetics. 2017 Dec;18(12):705–20.

13. Lord K, Schneider RA, Coppinger R. Evolution of working dogs [Internet]. The Domestic Dog: Its Evolution, Behavior and Interactions with People. 2016 [cited 2019 Oct 8]. Available from: /core/books/domestic-dog/evolution-of-working-dogs/CC5083D37F741470DDFA69AFBB238AB1

14. Friedrich J, Arvelius P, Strandberg E, Polgar Z, Wiener P, Haskell MJ. The interaction between behavioural traits and demographic and management factors in German Shepherd dogs. Applied Animal Behaviour Science [Internet]. 2018 Dec 5 [cited 2018 Dec 12]; Available from: http://www.sciencedirect.com/science/article/pii/S0168159118303265

15. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Jan 9;19(9):1655–64.

16. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97–120.

17. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. PLoS Biol [Internet]. 2006 Mar [cited 2018 Nov 9];4(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1382018/

18. Friedrich J, Brand B, Ponsuksili S, Graunke KL, Langbein J, Knaust J, et al. Detection of genetic variants affecting cattle behaviour and their impact on milk production: a genome-wide association study. Anim Genet. 2016 Feb 1;47(1):12–8.

19. Crouch DJM, Winney B, Koppen WP, Christmas WJ, Hutnik K, Day T, et al. Genetics of the human face: Identification of large-effect single gene variants. PNAS. 2018 Jan 23;115(4):E676–85.

20. Morgan MD, Pairo-Castineira E, Rawlik K, Canela-Xandri O, Rees J, Sims D, et al. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. Nature Communications. 2018 Dec 10;9(1):5271.

21. Valente TS, Baldi F, Sant'Anna AC, Albuquerque LG, Costa MJRP da. Genome-Wide Association Study between Single Nucleotide Polymorphisms and Flight Speed in Nellore Cattle. PLOS ONE. 2016 Jun 14;11(6):e0156956.

22. Luciano M, Huffman JE, Arias-Vásquez A, Vinkhuyzen AA, Middeldorp CM, Giegling I, et al. Genome-wide association uncovers shared genetic effects among personality traits and mood states. Am J Med Genet B Neuropsychiatr Genet. 2012 Sep;0(6):684–95.

23. Smith EN, Bloss CS, Badner JA, Barrett T, Belmonte PL, Berrettini W, et al. Genome-wide association study of bipolar disorder in European American and African American individuals. Mol Psychiatry. 2009 Aug;14(8):755–63.

24. Wang K-S, Liu X-F, Aragam N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. Schizophrenia Research. 2010 Dec 1;124(1):192–9.

25. Sun J, Kuo P-H, Riley BP, Kendler KS, Zhao Z. Candidate genes for schizophrenia: A survey of association studies and gene ranking. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2008;147B(7):1173–81.

26. Ilska J, Haskell MJ, Blott SC, Sánchez-Molano E, Polgar Z, Lofgren SE, et al. Genetic Characterisation of Dog Personality Traits. Genetics. 2017 Jan 1;genetics.116.192674.

27. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013 Apr 15;14:128.

28. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 08;44(W1):W90-97.

29. González-Maeso J, Meana JJ. Heterotrimeric G Proteins: Insights into the Neurobiology of Mood Disorders. Curr Neuropharmacol. 2006 Apr;4(2):127–38.

30. Lipsky RH, Marini AM. Brain-Derived Neurotrophic Factor in Neuronal Survival and Behavior-Related Plasticity. Annals of the New York Academy of Sciences. 2007;1122(1):130–43.

31. Lüscher C, Malenka RC. NMDA Receptor-Dependent Long-Term Potentiation and Long-Term Depression (LTP/LTD). Cold Spring Harb Perspect Biol [Internet]. 2012 Jun [cited 2019 Jun 18];4(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367554/

32. Pasterkamp RJ, Giger RJ. Semaphorin function in neural plasticity and disease. Current Opinion in Neurobiology. 2009 Jun 1;19(3):263–74.

33. Jacobsen JPR, Medvedev IO, Caron MG. The 5-HT deficiency theory of depression: perspectives from a naturalistic 5-HT deficiency model, the tryptophan hydroxylase 2Arg439His knockin mouse. Philos Trans R Soc Lond B Biol Sci. 2012 Sep 5;367(1601):2444–59.

34. de Almeida RMM, Ferrari PF, Parmigiani S, Miczek KA. Escalated aggressive behavior: Dopamine, serotonin and GABA. European Journal of Pharmacology. 2005 Dec 5;526(1):51–64.

35. Quignon P, Herbin L, Cadieu E, Kirkness EF, Hédan B, Mosher DS, et al. Canine Population Structure: Assessment and Impact of Intra-Breed Stratification on SNP-Based Association Studies. PLoS ONE [Internet]. 2007 Dec 19 [cited 2016 Mar 22];2(12). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2129117/

36. Moriya S, Soga T, Wong DW, Parhar IS. Transcriptome composition of the preoptic area in mid-age and escitalopram treatment in male mice. Neuroscience Letters. 2016 May 27;622:67–71.

37. Wiener P, Sánchez-Molano E, Clements DN, Woolliams JA, Haskell MJ, Blott SC. Genomic data illuminates demography, genetic structure and selection of a popular dog breed. BMC Genomics. 2017 Aug 14;18:609.

38. Knutson DC, Mitzey AM, Talton LE, Clagett-Dame M. Mice null for NEDD9 (HEF1) display extensive hippocampal dendritic spine loss and cognitive impairment. Brain Research. 2016 Feb 1;1632:141–55.

39. Patel MM, Silasi-Mansat R, Keshari RS, Sansam CL, Jones DA, Lupu C, et al. Role of Androgen Dependent TFPI-Regulating Protein (ADTRP) in Vascular Development and Function. Blood. 2016 Dec 2;128(22):556–556.

637 40. Li M, Han Y, Zhou H, Li X, Lin C, Zhang E, et al. Transmembrane protein 170B is a
638 novel breast tumorigenesis suppressor gene that inhibits the Wnt/β-catenin pathway. Cell
639 Death Dis [Internet]. 2018 Jan 24 [cited 2019 Jul 16];9(2). Available from:
640 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5833782/

641 41. Valnegri P, Huang J, Yamada T, Yang Y, Mejia LA, Cho HY, et al. RNF8/UBC13
642 ubiquitin signaling suppresses synapse formation in the mammalian brain. Nature
643 Communications. 2017 Nov 2;8(1):1271.

644 42. Chen C, Chen C, Moyzis R, Stern H, He Q, Li H, et al. Contributions of Dopamine-
645 Related Genes and Environmental Factors to Highly Sensitive Personality: A Multi-Step
646 Neuronal System-Level Approach. PLOS ONE. 2011 Jul 13;6(7):e21636.

647 43. Nakano N, Maeyama K, Sakata N, Itoh F, Akatsu R, Nakata M, et al. C18 ORF1, a Novel
648 Negative Regulator of Transforming Growth Factor-β Signaling. J Biol Chem. 2014 Feb
649 5;289(18):12680–92.

650 44. Meerabux JMA, Ohba H, Iwayama Y, Maekawa M, Detera-Wadleigh SD, DeLisi LE, et
651 al. Analysis of a t(18;21)(p11.1;p11.1) translocation in a family with schizophrenia.
652 Journal of Human Genetics. 2009 Jul;54(7):386–91.

653 45. Hayward JJ, Castelhano MG, Oliveira KC, Corey E, Balkman C, Baxter TL, et al.
654 Complex disease and phenotype mapping in the domestic dog. Nat Commun. 2016 Jan
655 22;7:10460.

656 46. Switonski M, Mankowska M. Dog obesity – The need for identifying predisposing
657 genetic markers. Research in Veterinary Science. 2013 Dec;95(3):831–6.

658 47. Miller CL, Murakami P, Ruczinski I, Ross RG, Sinkus M, Sullivan B, et al. Two complex
659 genotypes relevant to the kynurenine pathway and melanotropin function show
660 association with schizophrenia and bipolar disorder. Schizophrenia Research. 2009 Sep
661 1;113(2):259–67.

662 48. Wang X, Pipes L, Trut LN, Herbeck Y, Vladimirova AV, Gulevich RG, et al. Genomic
663 responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes).
664 PNAS. 2018 Oct 9;115(41):10398–403.

665 49. Dreger DL, Schmutz SM. A SINE Insertion Causes the Black-and-Tan and Saddle Tan
666 Phenotypes in Domestic Dogs. J Hered. 2011 Sep 1;102(Suppl_1):S11–8.

667 50. Dreger DL, Parker HG, Ostrander EA, Schmutz SM. Identification of a Mutation that Is
668 Associated with the Saddle Tan and Black-and-Tan Phenotypes in Basset Hounds and
669 Pembroke Welsh Corgis. J Hered. 2013 May 1;104(3):399–406.

670 51. Zapata I, Serpell JA, Alvarez CE. Genetic mapping of canine fear and aggression. BMC
671 Genomics. 2016;17:572.

672 52. Bonnet C, Andrieux J, Béri-Dexheimer M, Leheup B, Boute O, Manouvrier S, et al.
673 Microdeletion at chromosome 4q21 defines a new emerging syndrome with marked
674 growth restriction, mental retardation and absent or severely delayed speech. Journal of
675 Medical Genetics. 2010 Jun 1;47(6):377–84.

676 53. Bhoj E, Halbach S, McDonald-McGinn D, Tan C, Lande R, Waggoner D, et al.
677 Expanding the spectrum of microdeletion 4q21 syndrome: a partial phenotype with
678 incomplete deletion of the minimal critical region and a new association with cleft palate
679 and Pierre Robin sequence. Am J Med Genet A. 2013 Sep;161A(9):2327–33.

680 54. Fee A, Noble N, Valdovinos MG. Functional Analysis of Phenotypic Behaviors of a 5-
681 Year-Old Male with Novel 4q21 Microdeletion. J Pediatr Neuropsychol. 2015 Dec
682 1;1(1):36–41.

683 55. Le-Niculescu H, Balaraman Y, Patel SD, Ayalew M, Gupta J, Kuczenski R, et al.
684 Convergent functional genomics of anxiety disorders: translational identification of
685 genes, biomarkers, pathways and mechanisms. Transl Psychiatry. 2011 May;1(5):e9.

686 56. Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, et al.
687 Variation of BMP3 Contributes to Dog Breed Skull Diversity. PLOS Genetics. 2012 Aug
688 2;8(8):e1002849.

689 57. Rimbault M, Ostrander EA. So many doggone traits: mapping genetics of multiple
690 phenotypes in the domestic dog. Hum Mol Genet. 2012 Oct 15;21(R1):R52-57.

691 58. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Pielberg GR, Sigurdsson S, et al.
692 Identification of Genomic Regions Associated with Phenotypic Variation between Dog
693 Breeds using Selection Mapping. PLOS Genet. 2011 Oct 13;7(10):e1002316.

694 59. Legrand R, Tiret L, Abitbol M. Two recessive mutations in FGF5 are associated with the
695 long-hair phenotype in donkeys. Genet Sel Evol [Internet]. 2014 Sep 25 [cited 2019 Feb
696 20];46(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4175617/

697 60. Housley DJE, Venta PJ. The long and the short of it: evidence that FGF5 is a major
698 determinant of canine 'hair'-itability. Animal Genetics. 2006;37(4):309–15.

699 61. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, Parker HG, et al. Coat Variation in
700 the Domestic Dog Is Governed by Variants in Three Genes. Science. 2009 Oct
701 2;326(5949):150–3.

702 62. Carola V, Perlas E, Zonfrillo F, Soini HA, Novotny MV, Gross CT. Modulation of social
703 behavior by the agouti pigmentation gene. Front Behav Neurosci [Internet]. 2014 Aug 1
704 [cited 2020 Jan 27];8. Available from:
705 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4117936/

706 63. Illumina I: Canine HD BeadChip. In Data Sheet: DNA Genotyping; 2010.
707 https://www.illumina.com/documents/products/datasheets/datasheet_caninehd.pdf

708 64. Chen M, Wang J, Wang Y, Wu Y, Fu J, Liu J. Genome-wide detection of selection
709 signatures in Chinese indigenous Laiwu pigs revealed candidate genes regulating fat
710 deposition in muscle. BMC Genet [Internet]. 2018 May 18 [cited 2019 May 30];19.
711 Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5960162/

712 65. Purcell SM, Chang CC. PLINK 1.9 [Internet]. Available from: www.cog-
713 genomics.org/plink/1.9/

714 66. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex
715 trait analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

716 67. Friedrich J, Strandberg E, Arvelius P, Sánchez-Molano E, Pong-Wong R, Hickey JM, et
717 al. Genetic dissection of complex behaviour traits in German Shepherd dogs. Heredity.
718 2019 Oct 14;1–13.

719 68. Wilsson E, Sinn DL. Are there differences between behavioral measurement methods? A
720 comparison of the predictive validity of two ratings methods in a working dog program.
721 Applied Animal Behaviour Science. 2012 Nov;141(3–4):158–72.

722 69. Hsu Y, Serpell JA. Development and validation of a questionnaire for measuring behavior
723 and temperament traits in pet dogs. Journal of the American Veterinary Medical
724 Association. 2003 Nov 1;223(9):1293–300.

725 70. Pfahler S, Distl O. Effective Population Size, Extended Linkage Disequilibrium and
726 Signatures of Selection in the Rare Dog Breed Lundehund. PLoS One [Internet]. 2015
727 Apr 10 [cited 2016 Aug 17];10(4). Available from:
728 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4393028/

729 71. Forutan M, Ansari Mahyari S, Baes C, Melzer N, Schenkel FS, Sargolzaei M. Inbreeding
730 and runs of homozygosity before and after genomic selection in North American Holstein
731 cattle. BMC Genomics. 2018 Jan 27;19(1):98.

732 72. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant
733 call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

734 73. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data
735 Inference for Whole-Genome Association Studies By Use of Localized Haplotype
736 Clustering. Am J Hum Genet. 2007 Nov;81(5):1084–97.

737 74. Maclean CA, Chue Hong NP, Prendergast JGD. hapbin: An Efficient Program for
738 Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets.
739 Mol Biol Evol. 2015 Nov;32(11):3027–9.

740 75. Talenti A, Bertolini F, Pagnacco G, Pilla F, Ajmone-Marsan P, Rothschild MF, et al. The
741 Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep
742 regions. Mamm Genome. 2017 Apr 1;28(3):114–28.

743 76. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N,
744 et al. Efficient mapping of mendelian traits in dogs through genome-wide association.
745 Nature Genetics. 2007 Nov;39(11):1321–8.

746 77. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide
747 detection and characterization of positive selection in human populations. Nature. 2007
748 Oct 18;449(7164):913–8.

749 78. Bertolini F, Gandolfi B, Kim ES, Haase B, Lyons LA, Rothschild MF. Evidence of
750 selection signatures that shape the Persian cat breed. Mamm Genome. 2016 Apr
751 1;27(3):144–55.

79. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res [Internet]. 2009 Jun 18 [cited 2019 Jul 17]; Available from: http://genome.cshlp.org/content/early/2009/06/15/gr.092759.109

80. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D754–61.

81. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

82. Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, et al. Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies. PLOS Genetics. 2012 Apr 26;8(4):e1002641.

83. Zhou X, Stephens M. Genome-wide Efficient Mixed Model Analysis for Association Studies. Nat Genet. 2012 Jun 17;44(7):821–4.

84. Friedrich, J. et al. (2020), Data from: Unravelling selection signatures in a single dog breed suggests recent selection for morphological and behavioural traits, [Dataset], Dryad, https://doi:10.5061/dryad.g4f4qrfmr [AU: please ensure this link is live prior to production – ED]

85. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, et al. A Simple Genetic Architecture Underlies Morphological Variation in Dogs. PLOS Biol. 2010 Aug 10;8(8):e1000451.

86. MacLean EL, Snyder-Mackler N, vonHoldt BM, Serpell JA. Highly heritable and functionally relevant breed differences in dog behaviour. Proceedings of the Royal Society B: Biological Sciences. 2019 Oct 9;286(1912):20190716.

87. Freedman AH, Schweizer RM, Vecchyo DO-D, Han E, Davis BW, Gronau I, et al. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. PLOS Genetics. 2016 Mar 4;12(3):e1005851.

88. Kukekova AV, Johnson JL, Xiang X, Feng S, Liu S, Rando HM, et al. Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. Nature Ecology & Evolution. 2018 Sep;2(9):1479–91.

89. Schlamp F, van der Made J, Stambler R, Chesebrough L, Boyko AR, Messer PW. Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. Mol Ecol. 2016 Jan;25(1):342–56.

90. Saxena R, Voight BF, Lyssenko V, Burtt NP, Bakker PIW de, Chen H, et al. Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. Science. 2007 Jun 1;316(5829):1331–6.

Tables

Table 1. Top selection signatures within the UK and Swedish GSD populations, showing the ten highest integrated haplotype score (iHS) statistics. SNPs within 200 kb were summarised into selection signature regions.

Chr	Start (Mb)	Stop (Mb)	Distance (Mb)	N_{SNPs}†	iHS peak‡	iHS mean§	Gene(s)□	Phenotypic association††
UK population								
5	29.2	29.8	0.62	16	3.18	2.84	*ENSCAFG00000015899; MMP20; MMP27; MMP7;* ***ENSCAFG00000030873****;* ***BIRC2****; BIRC3;* ***YAP1****;* ***C11orf70****; CEP126; ANGPTL5*	-
12	68.1	68.2	0.06	2	3.22	2.96	***TRAF3IP2***	-
19	33.0	33.1	0.04	4	3.26	2.84	n.a.	-
19	36.0	36.5	0.51	10	3.46	2.93	***NCKAP5***	-
19	36.8	37.0	0.19	5	3.18	2.90	n.a.	-
19	37.5	37.7	0.20	6	3.48	3.19	***TMEM163***	-
19	38.3	38.6	0.31	9	3.19	2.79	***ZRANB3****; ENSCAFG00000005064;* ***R3HDM1****; UBXN4*	-
19	39.5	39.5	0.03	2	3.23	2.91	n.a.	-
20	57.6	57.7	0.07	3	3.18	3.10	*ENSCAFG00000031730; ENSCAFG00000023991;* ***ARHGAP45****; ATP5F1D; CIRBP; MIDN;* ***STK11****;* ***SBNO2****; POLR2E*	-
35	7.9	8.1	0.14	4	3.26	3.09	*BMP6;* ***TXNDC5****;* ***BLOC1S5****; ENSCAFG00000009583; ENSCAFG00000024482*	-
Swedish population								
4	44.3	n.a.	n.a.	1	3.09	n.a.	***ENSCAFG00000017171***	-
4	46.9	n.a.	n.a.	1	3.27	n.a.	*ENSCAFG00000028841*	-
4	50.0	50.2	0.15	4	3.09	2.90	***ATP10B***	-
4	52.5	n.a.	n.a.	1	3.47	n.a.	***CLINT1***	-
12	66.7	67.2	0.47	10	3.36	3.13	*GPR6;* ***WASF1****;* ***CDC40****;* ***METTL24****;* ***DDO****; SLC22A16; CDK19*	-
12	67.7	n.a.	n.a.	1	3.13	n.a.	***SLC16A10***	-
18	54.9	55.3	0.36	7	3.45	2.99	*LRRC10B; PPP1R32;* ***SYT7****; PGA; DDB1;* ***VWCE****;* ***ENSCAFG00000016314****;* ***SLC15A3****; CD5; VPS37C; CD6*	-

19	50.6	n.a.	n.a.	1	3.12	n.a.	***KIF5C***	-
24	42.4	42.5	0.05	3	3.33	3.05	*RBM38;* ***CTCFL***	-
36	30.1	30.6	0.05	6	3.11	2.82	*GULP1; COL3A1;* ***COL5A2***	-

[†]Number of top SNPs in region
[‡]Standardised absolute iHS of the peak SNP (in that region)
[§]Average standardised absolute iHS across the SNPs of a region
Genes located within and +/- 40 kb around selection signatures. Genes highlighted in bold include a SNP that belongs to the top 0.5% of the test statistic; all others are located within the region or +/- 40 kb around selection signatures
[††]There were no phenotypic associations (behaviour, coat colour or coat length) with FDR-adjusted P-value<0.1 for markers located within the top ten selection signatures within populations.

Table 2. Selection signatures that belonged to the top 1% of the distribution of at least two methods used to detect signatures of different selection between the GSD populations. SNPs within 200 kb were summarised into selection signature regions.

Chr	Start	Stop	N_{SNPs}†	Population	F_{ST}‡	ΔROH_{Prop}§	XP-EHH□	Gene(s)	Phenotypic association††
1	24024856	25483783	61	Sweden	0.12	0.46	NA	***ME2; MRO***; *MC2R; MC5R;* ***ENSCAFG00000000172***; *ENSCAFG00000029562; ENSCAFG00000029833; FAM210A;* ***LDLRAD4***; *ENSCAFG00000023012;* ***MOXD1***; *ENSCAFG00000031561;* ***CTGF***	Chasing*(UK)
9	16472361	16493753	4	UK	0.09	NA	2.81	*KCNJ16; KCNJ2*	-
12	5349354	6130868	44	Sweden	NA	0.27	3.44	*BRPF3;* ***PNPLA1***; ***C12H6orf222***; ***ETV7***; ***PXT1***; *ENSCAFG00000001396;* ***KCTD20***; *STK38; SRSF3; CDKN1A;* ***ENSCAFG00000001418***; *ENSCAFG00000001419;* ***CPNE5***; ***PPIL1***; ***C12H6orf89***; *MTCH1;* ***PI16***; ***FGD2***	Stranger-directed fear**(UK)
12	6466863	6554339	7	Sweden	NA	0.27	3.46	*FGD2;* ***CMTR1***; ***ENSCAFG00000030835***	Separation anxiety* (Sweden)
22	1027334	1140100	6	UK	0.08	0.26	NA	*RNASEH2B*	-
22	1683950	2496568	46	UK	0.12	0.26	NA	***KCNRG***; ***TRIM13***; ***SPRYD7***; ***KPNA3***; *ENSCAFG00000031710;* ***EBPL***; *ENSCAFG00000010362;* ***RCBTB1***; ***PHF11***; ***SETDB2***; ***CAB39L***; ***CDADC1***; *ENSCAFG00000028525; MLNR; FNDC3A*	-
24	22002778	22463326	24	UK	0.07	0.29	NA	*COMMD7;* ***DNMT3B***; ***MAPRE1***; ***EFCAB8***; ***SUN5***; ***BPIFB2***; ***BPIFB6***; ***BPIFB3***; ***BPIFB4***; ***ENSCAFG00000032553***; ***BPIFA2***; *ENSCAFG00000007369;* ***BPIFA3***; *BPIFA1*	Coat colour**(UK)
24	22908179	23816844	37	UK	0.14	0.28	NA	*ENSCAFG00000029918; ENSCAFG00000007430; ENSCAFG00000007435; ENSCAFG00000029879; NECAB3; PXMP4;* ***ZNF341***; ***CHMP4B***; *EIF2S2;* ***RALY***; ***ASIP***; *ENSCAFG00000007508; AHCY;* ***ITCH***; *DYNLRB1;* ***PIGU***; *MAP1LC3A;* ***NCOA6***; *TP53INP2*	Coat colour**(UK)

24	24867975	25952679	64	UK	0.13	0.28	NA	***CNBD2***; ***EPB41L1***; ***AAR2***; ***DLGAP4***; ***MYL9***; ***TGIF2***; ***SLA2***; *TGIF2-C20orf24*; ***NDRG3***; ***DSN1***; ***SOGA1***; ***TLDC2***; ***SAMHD1***; ***RBL1***; ***MROH8***; ***RPN2***; *GHRH*; *MANBAL*; *SRC*	Coat colour**(UK)
32	4172082	4455360	7	UK	0.09	0.27	NA	***ANTXR2***; ***PRDM8***	Coat length**(UK)
32	5350389	5399877	4	UK	0.13	0.26	NA	***PRKG2***	Coat length**(UK) and * (Sweden) Stranger-directed aggression** (Sweden)
32	5609507	5667788	4	UK	0.12	0.26	NA	*ENSCAFG00000008928*; *RASGEF1B*	Coat length** (UK and Sweden)
32	13000437	14125551	44	UK	0.11	0.37	NA	***SNCA***; ***MMRN1***; ***CCSER1***	Coat colour* (UK) Separation anxiety*(UK) Stranger-directed aggression* (Sweden)
32	14527559	14597957	4	UK	0.11	0.38	NA	*ENSCAFG00000009954*	-
32	14952127	15194499	4	UK	0.10	0.28	NA	*ENSCAFG00000009965*	-
34	33480270		1	UK	NA	0.27	2.80		-

†Number of top SNPs in region

‡Fixation index

§Differences between runs of homozygosity

Cross-population extended haplotype homozygosity.

NA indicates that this selection signature was not present in the top 1% of the test distribution

Genes highlighted in bold include a SNP that belongs to the top 1% of the test distribution; all others are located within the region or +/- 40 kb around selection signatures

††Significant phenotypic associations (behaviour, coat colour, coat length) for the UK and Swedish population within selection signature region. P-values were adjusted using False Discovery Rate (FDR), with significant associations determined as adjusted P-values <0.05 (**) and suggestive associations as adjusted P-values <0.1 (*). The population for which the phenotypic association was identified is specified in parentheses.

Figure legends

Figure 1. Principal Component Analysis of the pruned genomic data. Eigenvectors for the first two principal components are plotted and individuals are coloured according to the population of origin. The variances explained by the principal components are given in parentheses.

Figure 2. Ancestry proportions of studied GSDs based on the pruned genomic data assuming three underlying ancestries (K = 3 clusters) as revealed by ADMIXTURE. Each cluster is represented by a colour and the length of the specific coloured segment indicates the dog's proportion of membership in that cluster.

Figure 3. Distribution of integrated haplotype score (iHS) in the UK (upper plot) and Swedish population (lower plot). The red line indicates the threshold for the top 0.5% iHS.

Figure 4. Circos plot for signatures of selection between GSD populations. The plot shows the three statistics used to identify regions under differential selection: differences between runs of homozygosity (ΔROH_{Prop}, outer circle, blue track), cross-population extended haplotype homozygosity (XP-EHH, middle circle, green track) and the fixation index ($F_{ST,}$ inner circle, purple track). The plot indicates concordant evidence in regions on Chr 1, 24 and 32, where peaks can be seen based on all three methods (although not within the top 1% of SNPs for XP-EHH, shown in red for the three methods).

Appendices

Table A1. Description of German Shepherd dog populations. Summary statistics for behaviour traits and other dog attributes within the UK and the Swedish GSD populations.

Table A2. List of SNPs belonging to the top 0.5% of the iHS statistic in the UK and Swedish populations.

Table A3. Lists of SNPs belonging to the top 1% of the F_{ST}, XP-EHH and ΔROH_{Prop} statistics and the SNPs that belonged to the top 1% for at least two methods.

Table A4. Significance of associations between population attributes and genetic ancestries. The proportion of ancestries estimated by ADMIXTURE (cluster 1, cluster 2, cluster 3) based on markers located within selection signature regions were fitted as fixed effects in separate linear models to test their association with different response variables (population attributes: behaviour traits, role of the dog, coat colour and coat length). The P-values for the respective models are shown in the table.

Table A5. Markers located in selection signature regions and showing significant associations (FDR-adjusted P<0.1) with phenotypic traits (behaviour, coat colour, coat length).

Table A6. Overlaps between genes located in selection signature regions and candidate genes for morphological traits and behaviour reported in other studies. A list of candidate genes in canids was compiled using the following references[1, 2, 9, 10, 11, 26, 37, 45, 50, 51, 58, 61, 67, 76, 85-89] and was compared to genes located in regions detected as selection signatures in this study.

Figure A1. Ancestry proportions of GSDs based on genotypes of SNPs from putatively selected regions assuming three underlying ancestries (K = 3 clusters) as revealed by ADMIXTURE. Each cluster is represented by a colour and the length of the specific coloured segment indicates the dog's proportion of membership in that cluster. The labels indicate the origin of the dog (Sweden or UK) and the coat colour (1 = saddle tan, 0 = sable, black or others).

Figure A2. Fine-mapping of target regions under divergent selection between German Shepherd dog populations. Particularly compelling regions that showed evidence of divergent selection in all three selection signature test statistics (SNP window-based F_{ST}, ΔROH_{Prop}, and XP-EHH) are located on Chr 1, 24 and 32. The plots illustrate the FDR-adjusted P-values from association analyses for phenotypic traits (behaviour, coat colour, coat length) (above, "Regional association") and the selection signature test statistics (below, "Selection signatures") for all SNPs in these regions. The plots were created using a modified R code from that of Saxena et al. 2007 [90].

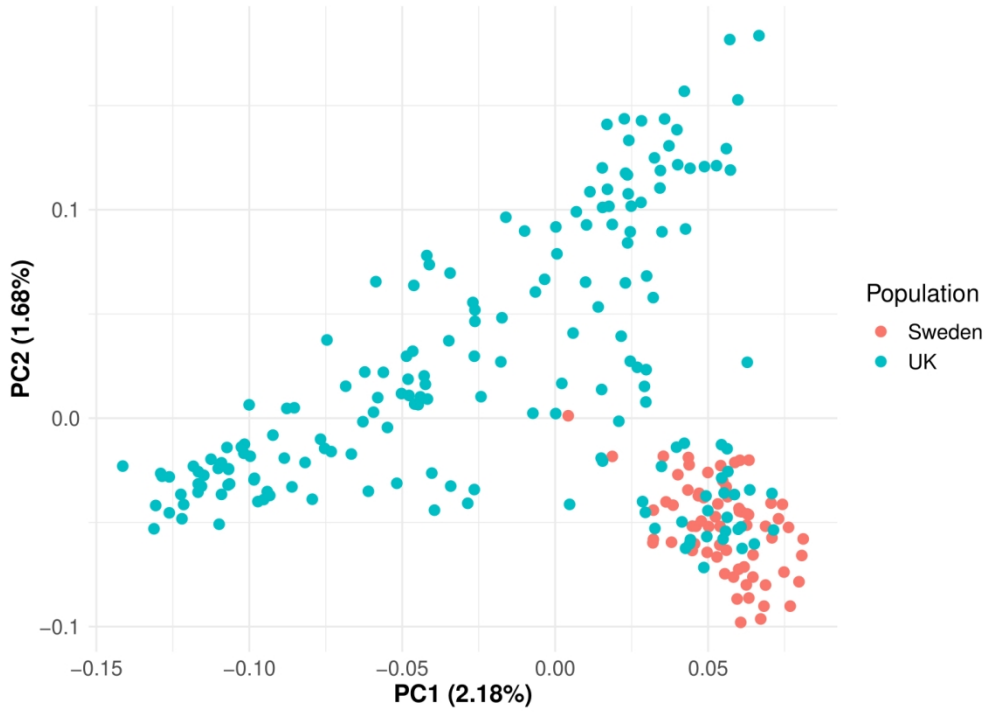

Figure 1. Principal Component Analysis of the pruned genomic data. Eigenvectors for the first two principal components are plotted and individuals are coloured according to the population of origin. The variances explained by the principal components are given in parentheses.

564x405mm (72 x 72 DPI)

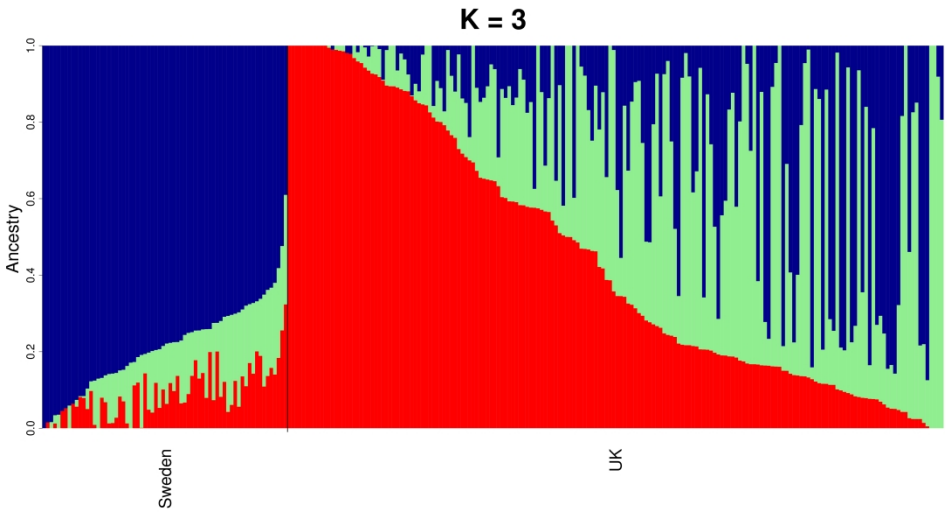

Figure 2. Ancestry proportions of studied GSDs based on the pruned genomic data assuming three underlying ancestries (K = 3 clusters) as revealed by ADMIXTURE. Each cluster is represented by a colour and the length of the specific coloured segment indicates the dog's proportion of membership in that cluster.

2116x1128mm (72 x 72 DPI)

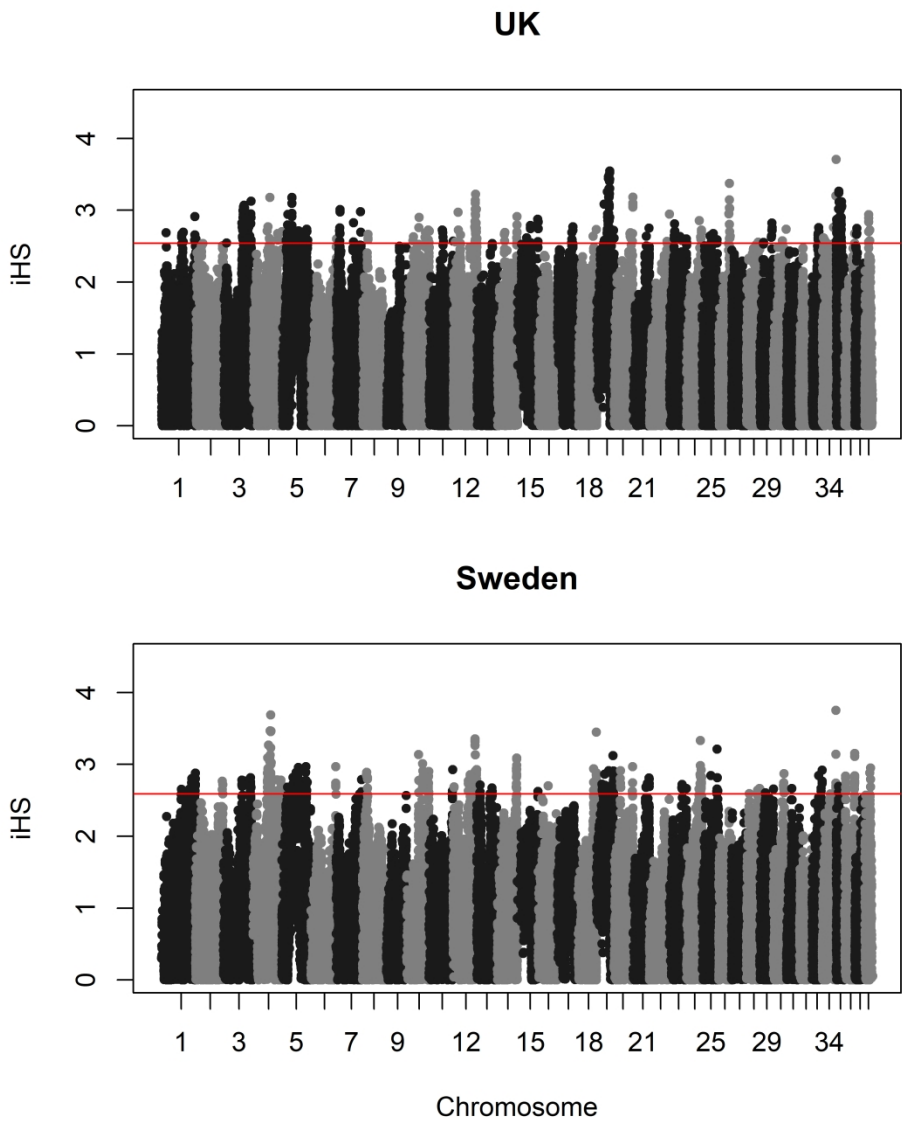

Figure 3. Distribution of integrated haplotype score (iHS) in the UK (upper plot) and Swedish population (lower plot). The red line indicates the threshold for the top 0.5% iHS.

152x188mm (600 x 600 DPI)

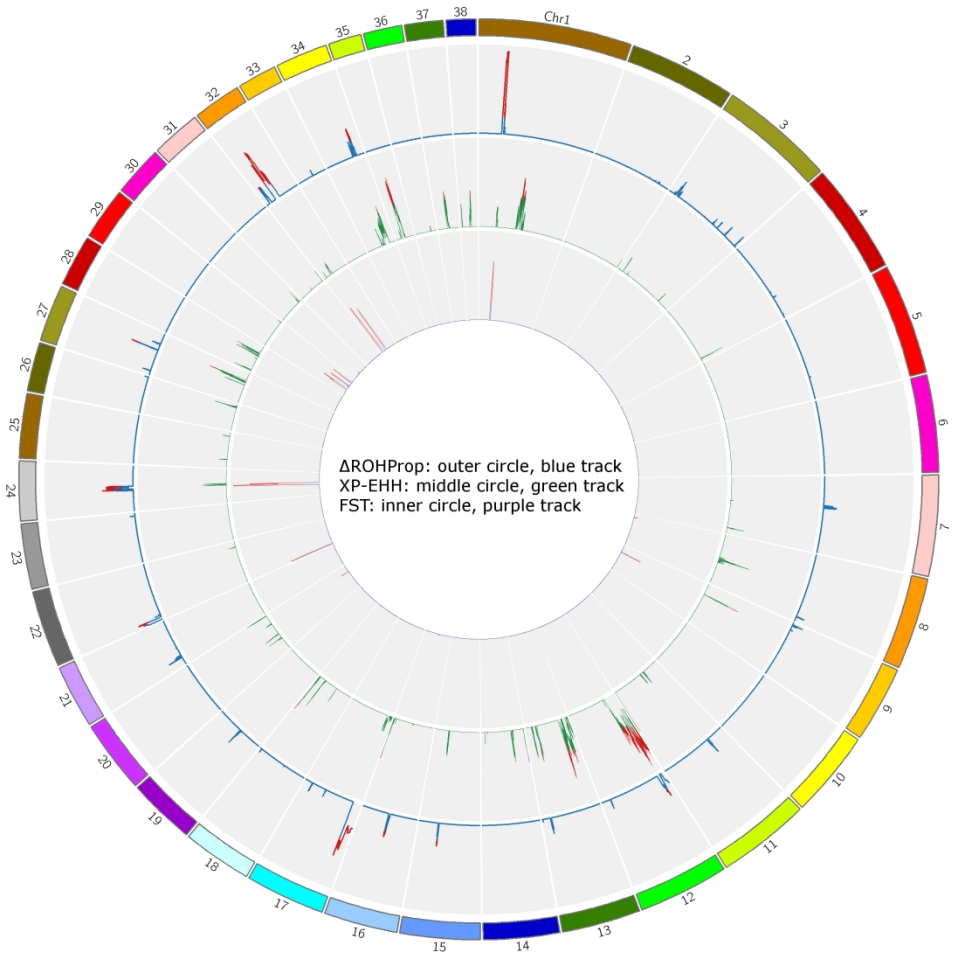

Figure 4. Circos plot for signatures of selection between GSD populations. The plot shows the three statistics used to identify regions under differential selection: differences between runs of homozygosity (ΔROHProp, outer circle, blue track), cross-population extended haplotype homozygosity (XP-EHH, middle circle, green track) and the fixation index (FST, inner circle, purple track). The plot indicates particularly compelling regions on Chr 1, 24 and 32, where peaks can be seen based on all three methods (although not within the top 1% of SNPs for XP-EHH, shown in red for the three methods).

793x793mm (96 x 96 DPI)