Variants That Differentiate Wolf and Dog Populations Are Enriched in Regulatory Elements

Pelin Sahlén[1,*], Liu Yanhu[2], Jinrui Xu[3], Eniko Kubinyi[4], Guo-Dong Wang[2,5], and Peter Savolainen[1]

[1]KTH Royal Institute of Technology, School of Chemistry, Biotechnology and Health, Science for Life Laboratory, Stockholm, Sweden
[2]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China
[3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA
[4]Department of Ethology, ELTE Eötvös Loránd University, Budapest, Hungary
[5]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

*Corresponding author: E-mail: pelin.akan@scilifelab.se.

Accepted: 6 April 2021

Abstract

Research on the genetics of domestication most often focuses on the protein-coding exons. However, exons cover only a minor part (1–2%) of the canine genome, whereas functional mutations may be located also in regions beyond the exome, in regulatory regions. Therefore, a large proportion of phenotypical differences between dogs and wolves may remain genetically unexplained. In this study, we identified variants that have high allelic frequency differences (i.e., highly differentiated variants) between wolves and dogs across the canine genome and investigated the potential functionality. We found that the enrichment of highly differentiated variants was substantially higher in promoters than in exons and that such variants were enriched also in enhancers. Several enriched pathways were identified including oxytocin signaling, carbohydrate digestion and absorption, cancer risk, and facial and body features, many of which reflect phenotypes of potential importance during domestication, including phenotypes of the domestication syndrome. The results highlight the importance of regulatory mutations during dog domestication and motivate the functional annotation of the noncoding part of the canine genome.

Key words: domestication, cis-regulatory regions, epigenetics.

Introduction

The dog was the first domesticated animal and is uniquely integrated into human society. Through domestication, dogs have evolved distinct morphological and behavioral traits which underly their adaptation to the human social environment (Miklósi 2015). Studies of the genetic components behind the evolution of the dog have, so far, focused on the coding part of the genome using variant genotyping and genome sequencing (Vaysse et al. 2011; Plassais et al. 2017, 2019). However, there are many variants in the noncoding part of the wolf and dog genomes, and it is unclear to what extent these variants contribute to phenotypic adaptations. In this study, we set out to answer this question, focusing on enhancers and promoters that are annotated by functional genomic data to increase our detection power.

Promoters and enhancers are the noncoding cis-regulatory elements orchestrating gene expression (Andersson and Sandelin 2020). Several experimental techniques are available for detecting active enhancer regions. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Barski et al. 2007) can profile the enhancer signatures, for example, H3K27Ac (Creyghton et al. 2010), H3K4me1 and transcription factor binding sites (Tian et al. 2011). ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is another method to map regulatory elements by locating open chromatin regions (Buenrostro et al. 2013). Although there is a plethora of information regarding the noncoding parts of the genome of humans and some model organisms (Davis et al. 2018), only a few studies are available for mapping enhancer elements in canine genomes. Two major comparative studies

Significance

Studies for finding genetic variants that mediated the evolution of the dogs from their wolf ancestors have been on the coding part of the canine genome. The role of noncoding variants in cis-regulatory elements is not well studied. We isolated variants that are highly differentiated (HD) between gray wolves and Southeast Asian village dogs and analyzed their distribution in the genome. We found that HD variants are enriched in cis-regulatory elements and this enrichment is larger than that of the protein-coding sequences. We also found that the elements containing HD variants regulate genes that are involved in oxytocin signaling, longevity, and digestion. We hope that our results will motivate a comprehensive annotation of the noncoding canine genome.

(Schmidt et al. 2010; Villar et al. 2015) reported transcription factor binding (*CEBPA* and *HNF4A*) and putative enhancer regions using ChIP-seq in dog livers. A large study, Barkbase, generated open chromatin maps of multiple tissues using ATAC-seq assays (Megquier et al. 2019).

In this study, we assumed that the variants that played significant roles during the evolution of domestic dogs from wolves should show a significant difference in allele frequencies in dogs and wolves. Combining information about the genomic position of regulatory regions and fixation index measure (F_{ST}) for genome variants from published studies, we identified variants that have high allelic frequency difference between wolves and dogs and that map within enhancers and promoters. We then investigated the potential functionality of the variants that mapped within enhancers and promoters. Our results show that the majority of variants with high F_{ST} value are located within promoter and enhancer sequences, many of which are linked to phenotypes of potential importance during domestication, suggesting the importance of changes in gene expression during dog domestication.

Results

We used publicly available data sets to annotate the enhancers and promoters in the canine genome. Putative enhancer regions covered approximately 5.4% of the genome (supplementary fig. 1 and table 1*a*–*c*, Supplementary Material online). To annotate promoter regions, we used the NCBI RefSeq annotation which includes both curated and predicted genes (supplementary table 1*d*, Supplementary Material online). We analyzed only the promoters of protein coding genes, which resulted in the selection of 24,471 promoters covering 1.7% of the genome (supplementary fig. 1, Supplementary Material online). We also included the exonic sequences in our analyses as a reference, since variants in protein-coding regions were already shown to explain a portion of the phenotypic divergence between dogs and wolves (Axelsson et al. 2013; Cagan and Blass 2016). The exons spanned 1.4% of the genome.

We used the fixation index measure (F_{ST}) to identify regions with high allelic frequency differences between dogs and wolves, which we call highly differentiated (HD) variants. To capture the general difference between dogs and wolves, and avoid signals from the recent intense selection for extreme morphologic types that formed modern dog breeds, we studied Southeast Asian village dogs, which have a noncontrolled reproduction and nonstandardized morphology and among the highest genetic diversity for dogs around the world, indicating limited population bottlenecks (Boyko et al. 2010; Wang et al. 2016). Therefore, we selected 38 Southeast Asian village dogs and 41 Eurasian and American wolves from a whole genome data set of 722 canids (Plassais et al. 2019), and calculated F_{ST} for 19.25 M autosomal SNPs. We focused on the variants with the top 1% F_{ST} values for our analyses, the distribution of which is shown in supplementary figure 2*a*, Supplementary Material online (F_{ST} values range: [0.54–1]). We then overlapped the positions of these top variants with the promoters and enhancer, finding significantly larger number of SNPs with high F_{ST} values in promoters and enhancers than in the genome as a whole (supplementary fig. 2*b*, Supplementary Material online).

Variants with High Differentiation between Dogs and Wolves Were Enriched in Exon Sequences

We observed enrichment for variants with high ($\geq$0.9) F_{ST} values in exons (fig. 1*a*). We looked at the consequences for all exonic variants with the top 1% F_{ST} values which ranges from 0.54 to 1, since F_{ST} values as low as 0.3 indicate significant population differentiation (Roux et al. 2016). In order to estimate the functional effect of these variants, we searched for the genes that contained a deleterious HD variant for its function using the SIFT software tool (supplementary table 3, Supplementary Material online). We identified 46 such genes, and although the list was not enriched with significance for any gene ontology category due to the small sample size, there were several marginally enriched phenotypes relating to dental and facial features (fig. 1*b*).

Regulatory Regions Were Enriched for Variants with High Differentiation between Dogs and Wolves

We then investigated the distribution of the variants in regulatory sequences. We grouped regulatory regions into two

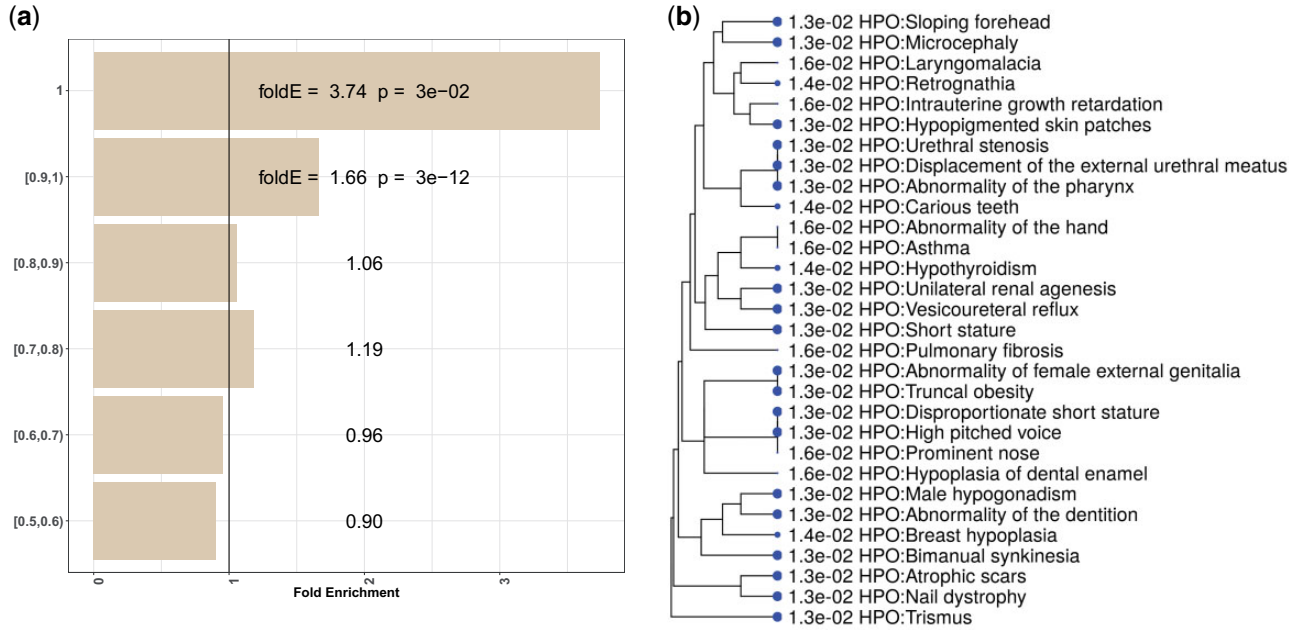

Fig. 1.—(*a*) Fold enrichment of variants for coding exonic sequences in different F_{ST} bins (*y* axis). (*b*) Human phenotypes that are enriched for genes carrying at least one missense exonic variant with an F_{ST} value greater than or equal to 0.5. All *P*-values are FDR-corrected and FDR threshold 0.05 is used. The size of the blue dots is proportional to the FDR. The human phenotype ontology database (https://hpo.jax.org/app/) was used for enrichment analyses due to the lack of such functional annotation for the canine genes. The tree is constructed using the distance between two gene sets based on the number of genes in the intersection and the union of two sets. The distance matrix is then used to construct a hierarchical clustering tree based on the number of shared and unique genes between the different sets.

groups: Promoters (defined as 1,000 bases upstream and 500 bases downstream of all transcription start sites of all coding transcripts), and enhancers (the nonpromoter regions that were situated within open chromatin regions and/or carried an H3K27Ac mark) (supplementary fig. 1 and methods, Supplementary Material online).

We then overlapped the promoter and enhancer regions with the variants with the top 1% F_{ST} values. Out of the variants with F_{ST} equal to 1, 20% (169/834) were situated within promoter sequences (15.4-fold enrichment, $P = 4.5e{-}258$), compared with 9.8% (82/834) for enhancer sequences (3.6-fold enrichment, $P = 1.2e{-}23$) and 2% (17/834) for exonic sequences (3.74-fold enrichment, $P = 3e{-}02$) (figs. 1 and 2). Thus, the enrichment in promoters and enhancers was higher than the enrichment in exons and, particularly, the enrichment was substantially higher in promoters than in exons. The variants with F_{ST} values greater or equal to 0.9 but less than 1 were also enriched for promoters and enhancers as well as exons, but at similar enrichment levels for all three classes.

Functional Profile of Enhancers and Promoters Enriched for Variants with High Differentiation between Dogs and Wolves

Our results showed significant enrichment of HD variants within both promoter and enhancer sequences. We, therefore, looked at the functional annotations of genes regulated by the enriched regions. We first performed target gene assignment for the enhancers with HD variants since enhancers are often located far away from the genes they regulate (Akerborg et al. 2019). We used the GREAT software (McLean et al. 2010) which takes gene expression and curated enhancer data sets into account, which should increase the accuracy of the target gene assignment compared with assigning the gene nearest to the enhancer (supplementary material, Supplementary Material online). In our analysis, we only included the enhancers that contained at least one variant in every 500 bases. This resulted in assignment of 2,923 genes to the 5,294 enhancer elements (supplementary fig. 3 and table 4*a* and *b*, Supplementary Material online). Additionally, there were 1,618 genes with promoters containing at least one HD variant which we included. We then performed a functional pathway enrichment analysis using all the genes assigned to enhancers or promoters. Several pathways of potential relevance for dog domestication were enriched, such as oxytocin signaling (Nagasawa et al. 2015) (FDR = 7.7e–5), carbohydrate digestion, absorption (Axelsson et al. 2013) (FDR = 7.3e–5), and longevity regulating pathway (FDR = 8e–4) (fig. 3*a* and supplementary table 4*c*, Supplementary Material online). The term "Pathways in cancer" was one of the most enriched (FDR = 1.2e–10), 30% (160/528) of the genes belonging to this term. Out of the 1,618 genes with promoters containing HD variants, 363

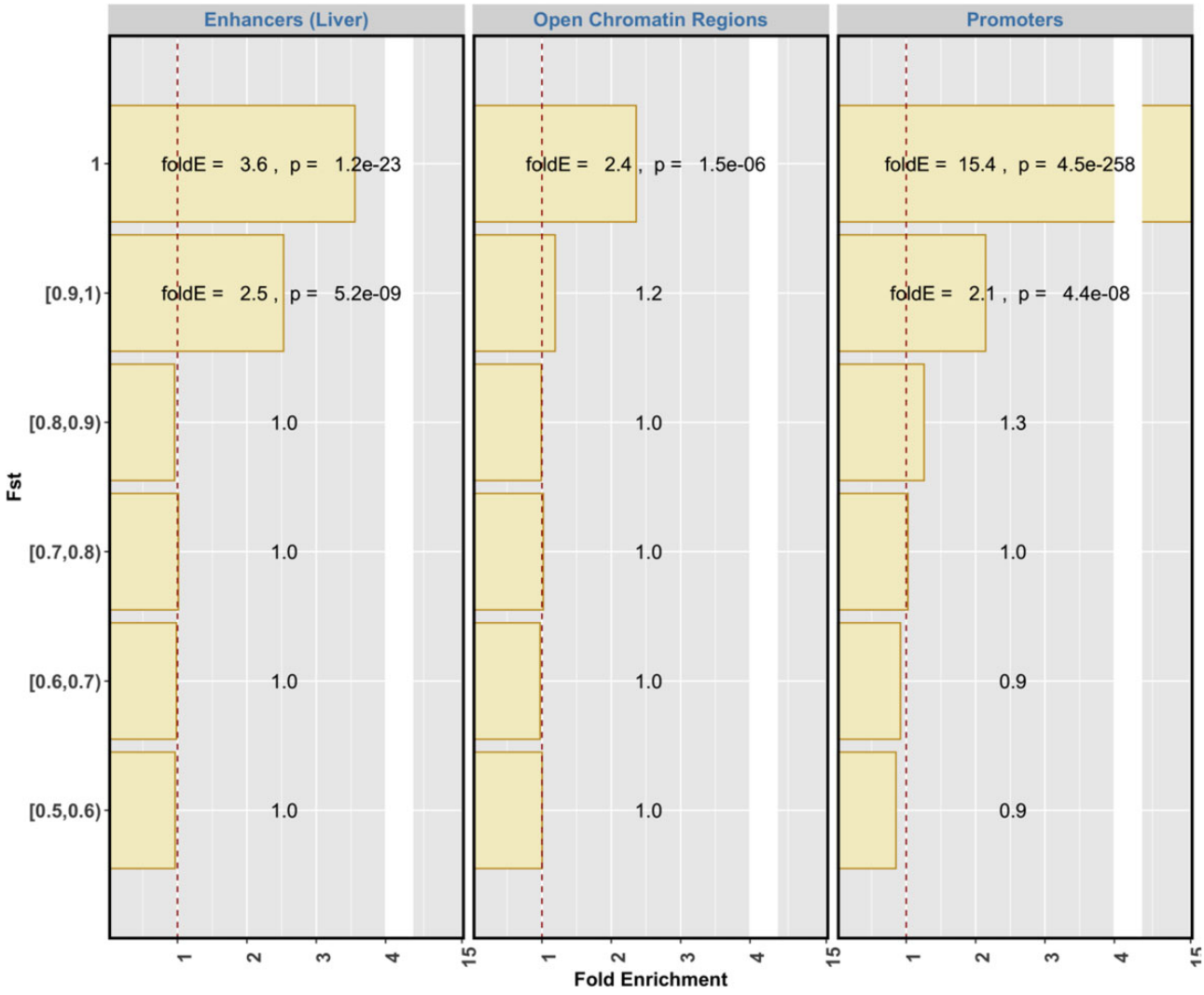

Fig. 2.—The enrichment of highly differentiated variants between wolves and dogs in regulatory elements; open chromatin regions assayed by ATAC-seq in multiple tissues, enhancers in livers assayed by ChIP-seq against H3K27Ac and promoters of protein-coding genes.

were associated with autosomal dominant diseases (FDR = 3.6e–14) and 534 were associated with autosomal recessive diseases (FDR = 1.1e–13). Many phenotypes related to facial and body features were also enriched, for example, micrognathia, hypertelorism, wide nasal bridge, anteverted nares, and cupped ear (fig. 3*b*).

Interestingly, when we used only the promoters containing the HD variants, only three pathways and four phenotypes were enriched and none of the pathways above were within the enriched terms (supplementary table 5*a*–*c*, Supplementary Material online). However, when only the genes assigned to enhancers were used, almost all of the above terms and phenotypes were enriched but at a lower degree (supplementary table 4*c*, Supplementary Material online). This indicates an important role for enhancers in phenotypic differentiation.

In addition, we selected all promoters that contain HD variants irrespective of whether they overlap with an ATAC-seq or ChIP-seq peak. There were 2,317 such promoters (supplementary table 6*a*, Supplementary Material online). We then investigated if there were particular binding motifs enriched for these promoters, finding that transcription factors, such as CGBP, TET1, DNMT1 were highly enriched (supplementary table 6*b*, Supplementary Material online). These transcription factors are either bound to methylated CpG dinucleotides or are required for DNA cytosine methylation (Tahiliani et al. 2009) and regulate the expression of multiple genes via suppression or activation through DNA methylation.

Discussion

In this study, we identified functional variants that have substantially different allele frequencies between dogs and wolves and potentially shape the different phenotypes of the dog and wolf populations. Consistent with previous

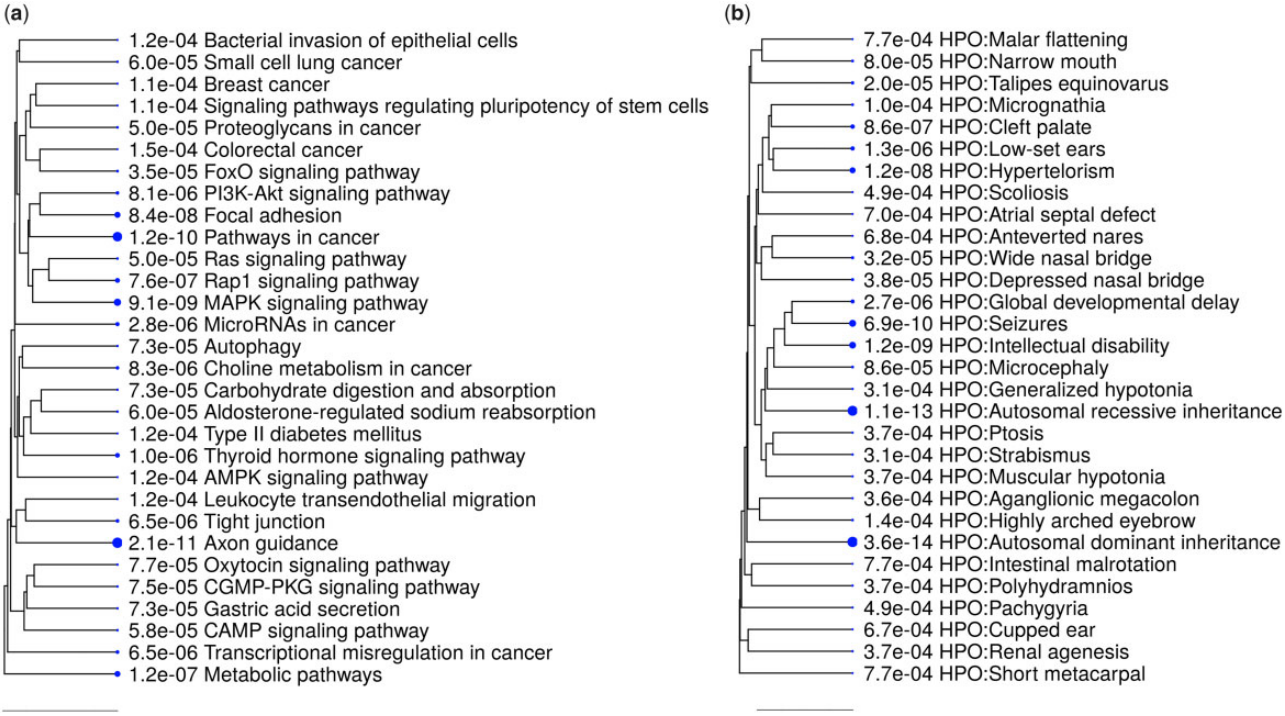

Fig. 3.—(*a*) The list of KEGG (the database of manually drawn pathway maps) pathways (*b*) human phenotypes (HPO) enriched for promoters regulated by regions with highly differentiated variants between wolves and dogs. FDR threshold of 0.01 is used and only the first 30 terms are shown. The size of the blue dots is proportional to the FDR value. The tree is constructed using the distance between two gene sets based on the number of genes in the intersection and the union of two sets. The distance matrix is then used to construct a hierarchical clustering tree based on the number of shared and unique genes between the different sets.

studies, we also found that such variants are enriched in the exons of coding genes. However, for the noncoding genome, promoters and enhancers showed stronger enrichment for the variants, supporting that many adaptive changes are mediated through changes in gene expression levels rather than protein structures (Wray 2007). These results strongly indicate that given the little divergence between wolf and dog proteins, many phenotypic differences can be due to regulatory mutations (King and Wilson 1975; Carroll 2005).

There are mainly two hypotheses that summarize the advantages of using cis-regulatory elements to change phenotypes, compared with using coding genes (Wray 2007). Both hypotheses are based on the flexibility of the cis-regulatory machinery. First, many mutations in cis-regulatory elements can fine-tune the target gene expression. In contrast, only a small portion of mutations are acceptable at protein-coding regions, whereas most mutations likely substantially change the protein stability, and thus drastically reduce the concentration of functional proteins. Consistent with this, the protein coding sequences are under strong purifying selection pressure (Lindblad-Toh et al. 2011; Wang et al. 2013). In addition, since mutations in regulatory elements often act according to additive rather than a recessive model, such mutations can be positively selected immediately (Ruvkun et al. 1991; Consortium 2013; Ponsuksili et al. 2015; Fallahsharoudi et al. 2017; Liu et al. 2020). Second, the mutations in cis-regulatory elements may be less pleiotropic than the mutations in protein-coding regions. For example, one of the cis-regulatory elements of a gene may be used only in a small number of tissues or developmental stages, and thus the mutation associated with the element can fine-tune the target gene expression for particular tissues and stages. By contrast, a nonsynonymous coding mutation permanently impacts the resulting protein (Stern 2000; Wray 2007).

We conducted GO functional analyses of the genes associated with the HD variants. As expected, there were no significantly enriched GO functions in the genes with the variants in their exons, presumably due to the small numbers of such genes. Most of the enriched phenotypes for exonic variants were related to facial and body features. However, we detected more genes whose promoters or enhancers carrying the HD variants. Analyzing these genes confirmed that the enrichment of functions associated with facial and body features was statistically significant, consistent with the domestication syndrome phenomenon (Pendleton et al. 2018). Shorter muzzles, floppy ears, reduced brain size are shared traits among domesticated mammals. Our findings support that these traits are linked, possibly through the mild deficit of the neural crest embryonic development, resulting in "neurocristopathies," such as micrognathia (reduced jaw

size), facial hypoplasia (smaller zygomatic bones), malformed external ear cartilages, and microcephaly (Wilkins et al. 2014).

Moreover, more GO functions, such as digestive functions and cancer-related functions, were also detected. The cancer-related functions are expected to play an important regulatory role in the cell cycle, suggesting substantial changes between dogs and wolves in terms of cell growth, proliferation, and differentiation. These cancer-related functions were enriched in both promoters and enhancers but had higher enrichment levels in enhancers than in promoters. This result supports the notion that enhancers tend to determine cell identities.

We observed that the enrichment for HD variants was higher in promoters than in enhancers, which might be due to the distinct functions between promoters and enhancers. It is widely observed that multiple enhancers are required to interact with one promoter to regulate the expression of its gene in a cell type (Karnuta and Scacheri 2018; Akerborg et al. 2019). This observation suggests that a genetic variant in the promoter may influence the gene expression more directly and effectively, compared with a variant in one of the individual enhancers. Therefore, the adaptive variants in promoters likely have larger effect sizes than those in enhancers, and thus are more likely to become HD variants during domestication.

We also observed that different pathways are associated with enhancers and promoters, respectively, which is likely due to the different regulatory functions of the enhancers and promoters. The enhancers are important to cell-type specific gene expression, and thus determine cell identity, whereas promoters tend to maintain basal gene expression. Due to the cell-type specific functions of the enhancers, the pathways associated with enhancer variants can be different from those associated with promoter variants.

The limitation of this study is that our dog sample (Southeast Asian village dogs) might not well represent the genomic changes that happened during the first step of domestication. Future studies should include a broader geographic sampling of village dogs to verify that the changes we described in this study are generalizable. However, a study (Shannon et al. 2015) including 549 village dogs from 38 countries found strong evidence that dogs were domesticated in Central Asia, in the proximity of Southeast Asia, therefore it is highly likely that our sample faithfully represents the first domesticated dogs.

In summary, this study highlights the importance of regulatory mutations for the study of dog evolution and domestication and will hopefully motivate the annotation of the noncoding canine genomes.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgements

The computations and data handling were enabled by resources in project [SNIC 2018011] provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This work was supported by a grant from Agria and SKK Forskningsfond.

Data Availability

All data are incorporated into the article and its supplementary material, Supplementary Material online.

Literature Cited

Akerborg O, et al. 2019. High-resolution regulatory maps connect vascular risk variants to disease-related pathways. Circ Genom Precis Med. 12(3):e002353.
Andersson R, Sandelin A. 2020. Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet. 21(2):71–87.
Axelsson E, et al. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495(7441):360–364.
Barski A, et al. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129(4):823–837.
Boyko AR, et al. 2010. A simple genetic architecture underlies morphological variation in dogs. PLoS Biol. 8(8):e1000451.
Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 10(12):1213–1218.
Cagan A, Blass T. 2016. Identification of genomic variants putatively targeted by selection during dog domestication. BMC Evol Biol. 16(1):10.
Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biol. 3(7):e245.
Consortium GT. 2013. The genotype-tissue expression (GTEx) project. Nat Genet. 45(6):580–585.
Creyghton MP, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA. 107(50):21931–21936.
Davis CA, et al. 2018. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 46(D1):D794–D801.
Fallahsharoudi A, et al. 2017. Genetic and targeted eQTL mapping reveals strong candidate genes modulating the stress response during chicken domestication. G3 (Bethesda) 7(2):497–504.
Karnuta JM, Scacheri PC. 2018. Enhancers: bridging the gap between gene control and human disease. Hum Mol Genet. 27(R2):R219–R227.
King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188(4184):107–116.
Lindblad-Toh K, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478(7370):476–482.
Liu Y, et al. 2020. Genome-wide analysis of expression QTL (eQTL) and allele-specific expression (ASE) in pig muscle identifies candidate genes for meat quality traits. Genet Sel Evol. 52(1):59.
McLean CY, et al. 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 28(5):495–501.
Megquier K, et al. 2019. BarkBase: epigenomic annotation of canine genomes. Genes (Basel). 10(6):433.
Miklósi A. 2015. Dog behaviour, evolution, and cognition. Oxford: Oxford University Press.

Nagasawa M, et al. 2015. Social evolution. Oxytocin-gaze positive loop and the coevolution of human–dog bonds. Science 348(6232):333–336.
Pendleton AL, et al. 2018. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. BMC Biol. 16(1):64.
Plassais J, et al. 2017. Analysis of large versus small dogs reveals three genes on the canine X chromosome associated with body weight, muscling and back fat thickness. PLoS Genet. 13(3):e1006661.
Plassais J, et al. 2019. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nat Commun. 10(1):1489.
Ponsuksili S, et al. 2015. Integrated Genome-wide association and hypothalamus eQTL studies indicate a link between the circadian rhythm-related gene PER1 and coping behavior. Sci Rep. 5(1):16264.
Roux C, et al. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS Biol. 14(12):e2000234.
Ruvkun G, Wightman B, Burglin T, Arasu P. 1991. Dominant gain-of-function mutations that lead to misregulation of the *C. elegans* heterochronic gene lin-14, and the evolutionary implications of dominant mutations in pattern-formation genes. Dev Suppl. 1:47–54.
Schmidt D, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328(5981):1036–1040.
Shannon LM, et al. 2015. Genetic structure in village dogs reveals a Central Asian domestication origin. Proc Natl Acad Sci USA. 112(44):13639–13644.
Stern DL. 2000. Evolutionary developmental biology and the problem of variation. Evolution 54(4):1079–1091.
Tahiliani M, et al. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324(5929):930–935.
Tian Y, et al. 2011. Global mapping of H3K4me1 and H3K4me3 reveals the chromatin state-based cell type-specific gene regulation in human Treg cells. PLoS One 6(11):e27770.
Vaysse A, et al. 2011. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 7(10):e1002316.
Villar D, et al. 2015. Enhancer evolution across 20 mammalian species. Cell 160(3):554–566.
Wang GD, et al. 2013. The genomics of selection in dogs and the parallel evolution between dogs and humans. Nat Commun. 4:1860.
Wang GD, et al. 2016. Out of southern East Asia: the natural history of domestic dogs across the world. Cell Res. 26(1):21–33.
Wilkins AS, Wrangham RW, Fitch WT. 2014. The "domestication syndrome" in mammals: a unified explanation based on neural crest cell behavior and genetics. Genetics 197(3):795–808.
Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 8(3):206–216.

Associate editor: Selene Fernández Valverde