Canine lineages reveal genetic drivers of dog behavioral diversification

Emily V. Dutrow[1], James A. Serpell[2], Elaine A. Ostrander[1,3]

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda MD, 20892

[2]Department of Clinical Sciences and Advanced Medicine, University of Pennsylvania School of Veterinary Medicine, Philadelphia PA, 19104

[3]Correspondence to: Elaine A. Ostrander, Ph.D., National Human Genome Research Institute, 50 South Drive, Building 50, Room 5351. Bethesda MD, 20892; Phone: 301 594 5184; eostrand@mail.nih.gov

Summary

Selective breeding of domestic dogs has generated diverse breeds often optimized for performing specialized tasks. Despite the heritability of breed-typical behavioral traits, identification of causal loci has proven challenging due to the complexity of canine population structure. We overcome longstanding difficulties in identifying genetic drivers of canine behavior by developing an innovative framework for understanding relationships between breeds and the behaviors that define them utilizing a dataset of over 4,000 domestic, semi-feral and wild canids and behavioral survey data for over 46,000 dogs. We identify ten major canine genetic lineages and their behavioral correlates, and show that breed diversification is predominantly driven by non-coding regulatory variation. We determine that lineage-associated genes converge in neurodevelopmental co-expression networks, identifying a sheepdog-associated enrichment for interrelated axon guidance functions. This work presents a scaffold for canine diversification that positions the domestic dog as an unparalleled system for revealing the genetic origins of behavioral diversity.

Introduction

Selective breeding of domestic dogs to perform specialized tasks may constitute humankind's most consequential behavioral genetics experiment. Humans have employed dogs for a range of purposes for thousands of years: performing farm tasks such as herding livestock or killing vermin, tracking and retrieving game, pulling loads, protecting people and property, and companionship (Serpell, 2017; Wilcox and Walkowicz, 1989). To produce dogs that will reliably display traits conducive to executing these functions, humans have selectively bred toward a variety of behavioral ideals. Therefore, genetic analyses of domestic dogs present a unique system for studying how behavioral diversity is biologically encoded, a consequential yet often intractable area of inquiry. Recent advances in three key domains provide new opportunities to address this question: First, citizen science initiatives related to canine behavior have yielded comprehensive behavioral trait datasets for tens of thousands of domestic dogs (Hsu and Serpell, 2003; Salonen et al., 2020; Serpell and Hsu, 2005). Second, the proliferation of publicly available canine whole genome sequencing (WGS) data provided the statistical power necessary to study complex traits, with thousands of genomes representing hundreds of breed types available for interrogation (Jagannathan et al., 2019; Leinonen et al., 2011; Pendleton et al., 2018; Plassais et al., 2019). Third, the development of methods for analyzing 'Big Data', both genomic and otherwise, yielded powerful approaches for high-dimensional data exploration that surpass prior capabilities (Amir et al., 2013; Haghverdi et al., 2016; Kuchroo et al., 2022; McInnes et al., 2018; Moon et al., 2019; Street et al., 2018).

Capitalizing on these developments, this study makes three significant advances in the study of canine history and behavior. First, we identify the major canine genetic lineages

corresponding to historical breed occupation using a dimensionality reduction approach aimed at capturing and quantifying the local and global complexity inherent in the relationships between individual dogs, breeds and larger canine populations, rather than requiring a hierarchical phylogenetic framework often inconsistent with the nature of dog breed development. Second, we identify behavioral correlates of canine lineage diversification, uncovering key shifts in breed-typical traits conducive to performing specific tasks. Third, we characterize the genetic variants principally driving canine phenotypic specificity, demonstrating the outsized role of non-coding regulatory variation in conferring breed-typical traits and the influence of selective breeding for behavioral versus morphological or aesthetic phenotypes on canine genetic diversification. This study uncovers precise neurodevelopmental pathways underlying canine behavioral specification, revealing how human beings shaped the world's most versatile domesticated animal and establishing canine genetics as a platform for studying the genetic underpinnings of behavioral diversity within species.

Results

Dimensionality reduction reveals canine lineages corresponding to divergent working roles

To identify genetic drivers of breed-typical canine behaviors, we compiled an expansive genomic dataset aimed at maximally representing canine behavioral diversity while accurately capturing global genetic relationships among purebred, mixed breed, free-breeding, and wild canids. The dataset includes a combination of WGS and SNP array data from 4,261 individuals, 2,823 of which are purebred dogs representing 226 breeds recognized by the Fédération Cynologique Internationale (FCI; Supplemental Text and fci.be/nomenclature; Figure 1A, Table S1). Owner-reported breed was used as a metric for dataset curation and to aid in visualization; however, the genetic analyses herein were intended to be independent of breed membership or recognition. Thus, we included 687 pet dogs of mixed, unknown, or non-recognized breed (Table S1). To capture the full spectrum of selective breeding pressures ranging from purebred dogs to theoretically free-breeding populations, we also included 658 semi-feral village dogs from 47 countries and 93 wild canids from four continents (Table S1).

We produced a dataset for visualizing and quantifying canine genetic relationships by generating a variance-standardized relationship matrix from a merged dataset of 102,764 SNPs from the aforementioned samples (Figure S1A). While principal component analysis (PCA) alone is often used to visualize genetic data, capturing global trends and outliers, such transformations have a limited ability to capture the sum of all local and global relationships in complex data within two dimensions (Figure S1A). UMAP, a non-linear alternative to PCA, largely clustered individuals into isolated groups without preserving global relationships, in

much the same way as a phylogenetic tree (Figure S1A) (McInnes *et al.*, 2018). To account for global relationships transcending individual breed clusters, we sought to impose no prior assumptions on the structure of our data and capture trajectories: linear or bifurcating paths through high-dimensional data that traverse a continuum of states. To do so, we used Potential of Heat Diffusion for Affinity-Based Transition Embedding (PHATE), a tool that encodes local and global data structure in high-dimensional data by transforming relationships among data points into diffusion probabilities that capture the likelihood of transition between states (Moon *et al.*, 2019). By applying PHATE, we generated a maximally informative two-dimensional embedding, i.e., a low-dimensional representation of high-dimensional data for exploratory analysis and visualization (Moon *et al.*, 2019). We observed that the PHATE-embedded data formed eight major linear trajectories emerging from the central portion of the embedding, each roughly corresponding to variance in 2-5 of the input PCs (Figure 1B, Figure S1B).

We broadly characterized the distribution of purebred dogs in the PHATE embedding using the ten FCI breed group classifications (i.e., kennel club groups broadly categorizing breeds by shared occupation and/or traits; see Supplemental text). We determined that breeds belonging to individual kennel club groups generally separated into distinct trajectories emerging from the center of the plot. One exception was the sighthound group, which split between a trajectory made up of European sighthounds and another largely comprised of breeds categorized by the FCI as “primitive-type” dogs but more aptly described as hunting breeds not originating in Europe or Asia (Figure 1B). Moreover, FCI groups that did not form one of the major linear trajectories in the two-dimensional visualization formed discrete trajectories observable in three dimensions (Figure S1C). Free-breeding village dogs were generally positioned in the central region of the embedding, and the wild canids, other than the dingo-related populations, clustered

tightly near the base of the trajectories comprised of spitz and “primitive-type” breeds consistent with previous analyses (Figure 1C) (Parker et al., 2017).

Canine lineages reflect shifts toward increasingly divergent genetic states

Because characterization based on kennel club-based categories indicated lineage-like trajectories from wild canids to divergent purebred types, we formally characterized the branching patterns observed in the PHATE embedding using pseudotemporal reconstruction, a method for inferring lineages among branching trajectories in high-dimensional data (Street *et al.*, 2018). We designated lineage origin with gray wolves and, in an otherwise unsupervised manner, identified 10 lineages, each of which were named based on breed composition: scent hound, pointer-spaniel, terrier, retriever, herder, sled dog, African and Middle Eastern, Asian spitz, dingo, and sighthound (Figure 2A, S2A; Table S2). We calculated pseudotime, a variable that quantifies positions of data points along lineages without requiring fixed underlying temporal variables such as evolutionary time or a bifurcating branch framework (Street *et al.*, 2018). Pseudotime values corresponding to at least one lineage were assigned to every individual in the dataset, with 74% of purebred and mixed breed dogs belonging to a single lineage. On average, each lineage included 663 individuals from 67 breeds (range=17-125; Table S2).

For each individual in the dataset, we calculated identity-by-state (IBS) with each gray wolf, finding that lineage structure did not appear to be singularly driven by decreased mean ancestral identity (Figure 2B). By examining mean pairwise IBS between each individual in the dataset and all dogs of each breed, we determined that lineage progression was broadly characterized by decreased inter-breed IBS, with terminal breeds (those with higher

pseudotemporal positions) displaying the most unique constellation of genotypes compared to all other dogs. For instance, we examined inter-breed IBS with a subset of scent hound lineage breeds and found that breeds with the highest pseudotemporal positions (i.e., otterhound and the bloodhound), had the least amount of shared identity with all other dogs (Figure S2B). In contrast, breeds with lower pseudotemporal positions (e.g., American coonhounds and other non-European scent hounds) generally showed increased shared identity with all other dogs both within and outside the scent hound lineage (Figure S2B).

We then characterized the degree to which genetic relationships among and across lineages were related to the historical geographic origin of breeds. A major axis of variation corresponded to separation between east and west, with breeds of non-European origin concentrated in the Asian spitz and African and Middle Eastern lineages (Figure 2C, 2D). Breeds with origins in Great Britain comprised the termini of all major western lineages, indicative of the intense selective pressure that defined Victorian-era breed development (Figure 2C) (Worboys et al., 2018). The predominance of this single geographic variable across the entire dataset illustrates the outsized role that historical breed origin has on contemporary canine population structure. We ascertained genomic correlates of breed history potentially driving this phenomenon by estimating the correlation between maximum pseudotime value on any lineage with homozygosity and with inbreeding (F coefficient), showing that pseudotime is not significantly correlated with increased homozygosity ($R = -0.11$, $p = 0.071l$; Figure S2C, top). We identified a significant but weak correlation with inbreeding (F coefficient; $R = 0.27$, $p = 4.5E-6$), a particularly noteworthy finding given the largely purebred composition of lineage termini compared to the predominately mixed breed or free-breeding basal groups (Figure S2C, bottom). Together, these findings suggest that each of the lineages identified in this study

represents groups of dogs possessing similar ancestry but which heavily diverged from wolves and ancestral canines through a combination of strong genetic drift or selection.

We then examined the distribution of modern free-breeding dogs in the PHATE embedding to determine the extent to which contemporary geographic isolation underlies trajectory structure among non-purebred populations. As expected, geographic location was the major driving variable (Figure 2D). While contemporary purebred dogs are generally isolated, both temporally and geographically, from their free-breeding counterparts (e.g., AKC-registered Chinese shar-peis versus contemporary free-breeding dogs in China), we observed overlap between free-breeding populations and purebred dogs historically linked to the same geographic region (Figure 2D). This finding may reflect both shared ancestry and historical admixture between purebred and free-breeding populations, though likely predating formal breed formation. To underscore the accuracy with which the PHATE embedding captures complex inter-breed relationships, we examined breeds with known histories involving hybridization of existing breeds. We observed an accurate reflection of relationships, even when breeds with minimal shared ancestry were hybridized (Figure S2D). For instance, the Nova Scotia duck tolling retriever is known to have been established through hybridization of sheepdog and retriever breed dogs (Parker *et al.*, 2017). While the contributing breeds belong to two different lineages, the Nova Scotia duck tolling retrievers are distributed along a trajectory at the nexus of both lineages (Figure S2D, left). Together, the aforementioned findings establish that the lineage framework accurately accounts for complex relationships across domestic dogs, establishing a means by which global features of dog breed diversification, such as breed-typical behavioral traits, can be interrogated both phenotypically and genotypically.

Behavioral traits are major phenotypic correlates of canine lineages

We used the lineage framework described above to characterize the genetic basis of breed-typical behavioral traits. To identify global behavioral correlates of the major canine lineages, we generated breed-average behavioral metrics based on data from over 46,000 purebred dogs using the Canine Behavioral Assessment and Research Questionnaire (C-BARQ) (Hsu and Serpell, 2003; Serpell and Hsu, 2005). C-BARQ is a comprehensive and well-validated questionnaire for behavioral assessment comprised of owner-reported numeric scores for 100 questions related to both breed-typical and atypical or undesirable canine behaviors corresponding to 14 distinct categories (listed in Figure 3A and Methods). Therefore, we calculated 14 numeric phenotype scores for each purebred dog in the C-BARQ dataset, each corresponding to a single behavioral factor (See Methods) (Hsu and Serpell, 2003; Serpell and Hsu, 2005).

Prior to integrating breed phenotypes into our genetic analyses, we verified that breeds with similar working roles broadly clustered based on calculated C-BARQ behavioral metrics, establishing the utility of the dataset for quantifying occupation-related breed traits. Using UMAP to dimensionally reduce the dataset of 14 C-BARQ scores per purebred dog, we visualized clustering of behaviorally similar individuals and identified the major axes of behavioral variation across the dataset, generally corresponding to fear, predatory drive and aggression, energy and excitability, and trainability (Figure 3A) (McInnes *et al.*, 2018). To test if these distributions corresponded to inherited, breed-related differences related to occupation, we examined the distribution of kennel club groups within the embedding and established that C-BARQ captures behavioral tendencies among breeds with shared purposes that are concordant

with historical working role requirements (Figure 3B). For example, we observed that terriers largely clustered in the bottom left of the plot, consistent with a tendency to exhibit predatory behavior and dog-directed aggression (Figure 3B). Companion and toy dogs clustered along the right side of the plot, consistent with increased social and non-social fear (Figure 3B). Consistent with a tendency toward increased trainability and decreased predatory drive and aggression, sheepdogs and retrievers clustered at the top of the plot (Figure 3B). We calculated mean C-BARQ scores for every breed in the behavioral dataset in order to assign behavioral metrics to breeds in our genetic dataset. By applying breed-average phenotypes, we reduced phenotypic noise caused by intra-breed variation and ensure that genetic associations with behavior were not driven by a preponderance of behavioral outliers in breeds genetically predisposed to behavioral problems (Dodman et al., 2016; Tang et al., 2014).

To broadly characterize breed-typical behavioral phenotypes across our purebred genetic dataset, we visualized breed-average C-BARQ scores on the PHATE embedding and observed distinct suites of behavioral traits across lineages (Figure S3A). We also observed significant overlap in behavioral tendencies across multiple lineages. For example, trainability was generally increased among breeds positioned on the herder, pointer-spaniel, and retriever trajectories (Figure S3A). Moreover, we determined that increased trainability was broadly a feature of working and sporting breeds by comparing trainability scores for working trial breeds versus non-working trial breeds (ANOVA p adj. = 3.06E-4) and non-recognized breeds (p adj. = 8.27E-3; Figure S3B).

In order to ascertain the degree to which genetic divergence along lineages corresponds to increasing behavioral distinctiveness, we measured correlation between breed-average C-BARQ scores and breed-average lineage pseudotime values. We identified significant

correlations for all eight of the lineages analyzed (dingo and African and Middle Eastern lineages excluded; Figure 3C, Table S3). The most significant positive correlations identified were between the herder lineage and non-social fear (p=1.86e-5, r=0.64), and between the terrier lineage and predatory chasing (p=1.96e-5, r=0.69), the latter consistent with working roles involving catching and killing prey (Figure 3C, Figure S3C, Table S3). The most significant negative correlation was between the scent hound lineage and trainability, consistent with selective breeding for traits advantageous for an independently driven and focused working style versus those conducive to human-directed responsiveness (p=4.50e-4, r=-0.68; Figure 3C, Figure S3C, Table S3).

These results revealed a unique repertoire of behavioral correlations for each lineage. Moreover, hierarchical clustering of correlation values for all 14 behavioral factor scores identified subsets of parallel behavior tendencies across lineages, such as decreased trainability and increased touch sensitivity, familiar dog aggression, and non-social fear among terriers and scent hounds (Figure 3C, Table S3). The observed overlap in certain behavioral tendencies across genetically distinct lineages underlines the utility of identifying phenotypic and genotypic correlates within lineages, as opposed to across all breeds, as a way of accurately identifying genetic drivers of breed behavioral diversification.

Lineage-associated variants are largely non-coding regions implicated in neurodevelopment

Having established significant behavioral correlates of the major canine lineages, we then identified genetic drivers of these global phenotypic shifts by performing a genome wide

association study (GWAS) on the WGS subset of the aforementioned dataset (1,171 samples). Using pseudotime values as quantitative phenotypes, we utilized a generalized linear model association test for the seven lineages sufficiently represented in the WGS dataset (Figure 4A, top left). We identified the top 100 associated loci of linked variants per lineage, ranked by LD-based index variant significance, showing that greater than 85% were unique to a single lineage (Figure 4A, Figure S4A). We also observed that the top lineage-associated loci were not evenly distributed throughout the genome, indicating apparent hotspots of associated loci in adjacent but independent genomic intervals (Figure 4A).

To test if the most significant associations were relevant to each lineage as a whole, as opposed to being driven by rare variants specific to a few terminal breeds, we examined the number of breeds with non-zero lineage-associated allele frequencies, finding that such variants were generally present in at least five breeds per lineage (Figure S4B). While allele frequency distributions varied, associated variants were not generally lineage-specific but rather present at moderate allele frequencies across the entire dataset, suggesting that breed diversification was principally driven by selection on standing variation in dogs that existed prior to modern breed formation (Figure S4C). To assess the degree to which this variation predates domestication, we examined allele frequencies of the top lineage-associated variants in gray wolves. For the terrier and herder lineages, the majority of associated variants were present in modern gray wolves, in contrast to the other five lineages tested which showed largely dog-specific and therefore likely derived variant associations (Figure 4A, bottom left). We determined that 25% of the top herder lineage-associated loci were present at >95% allele frequency in gray wolves, indicating that breed diversification along this lineage primarily involved selection on ancestral variation,

suggestive of ancient origins for variants driving herding dog-typical behavioral tendencies (Figure 4A, bottom left).

Characterization of lineage associations at the variant rather than locus level required assessing the degree to which variant significance with an associated locus predicts functionality; therefore, we applied two methods for assessing coding and non-coding functionality and determined that both basewise conservation scores (phyloP) and predicted non-coding functional significance (as measured using DeepSEA, see Methods) were skewed toward the top ranked variant at each locus (Figure S4D) (Pollard et al., 2010; Zhou and Troyanskaya, 2015). Based on these findings, we expanded our characterization of individual variants to the top five significant variants per locus and found that 74% of the top 100 loci identified in each association test contained only single nucleotide variants, and 96% contained no insertions or deletions larger than four base pairs (Figure S4E). We also determined that only 1% of these loci contained a non-synonymous coding variant and that all missense, splice, and UTR variants were predicted to be of either low or moderate impact (Figure S4E; See Methods). Strikingly, of the 16,250 significantly associated variants in all lineages analyzed, regardless of locus or variant ranking, only 76 were coding variants predicted to have moderate or high impact by either prediction tool, of which only two were predicted to have a high impact based on both prediction tools (Table S4).

We next examined overlap between lineage-associated variants and loci with known enhancer and promoter functions in other species using the ENCODE database of candidate cis-regulatory elements (cCREs). We determined that within 95 lineage-associated loci, at least one of the top five variants overlapped a human cCRE or mouse cCRE in the orthologous region of the human or mouse genome after filtering for replicating cCREs within tissue groups (Figure

S4F; Table S4) (Moore et al., 2020). The adult brain and the embryonic brain were the top human and mouse tissues, respectively, for lineage-associated variant overlaps overall. Of the overlapping tissue-specific elements identified, the greatest percentage were human embryonic brain-specific (31% of elements) followed by human adult brain-specific (26%). The intersection between lineage-associated variation and regulatory elements across species indicates that a subset of lineage-associated variants may affect evolutionarily conserved regulatory elements, particularly those with tissue-specific neurodevelopmental functions.

We next mapped predicted gene interactions for non-coding lineage-associated loci to perform gene-based tissue expression enrichment (See Methods). Using the set of genes predicted to interact with each set of lineage-associated non-coding loci (lineage-associated genes), we tested 52 human tissue types for enrichment of expression (See Methods) (Consortium et al., 2020). We found that of the 16 significantly enriched tissues for the scent hound lineage, 10 were brain tissues, as were 8 of 11 from the herder lineage (Figure 4B, Table S5). Notably, the most significant scent hound and herder lineage enrichments, respectively, were the frontal cortex Brodmann area 9 and the anterior cingulate cortex Brodmann area 24, both broadly implicated in socio-cognitive functions including learned fear response (Figure 4B, Table S5) (Apps et al., 2016; Jhang et al., 2018; Snow, 2016).

While identification of canine brain regions homologous to higher-order human cortical regions remains difficult, previous studies identifying significant variation in brain morphology across domestic dog breeds suggest that a subset of lineage-associated loci identified herein may contribute to gross neuroanatomical phenotypes relevant to behavior and with neurodevelopmental origins (Hecht et al., 2019). Therefore, we also tested if expression of lineage-associated genes was enriched during specific neurodevelopmental time periods by

performing gene set enrichment analysis using published gene expression data for 29 ages of human brain samples (Kang et al., 2011). We determined that the top two gene set enrichments for any lineage (corresponding to the scent hound and pointer-setter lineages) were both at the 26 post-conception week (pcw) stage, a period during human late fetal neurodevelopment characterized by thalamocortical innervation (p adj.=6.09E-5 and p. adj=2.82E-6, respectively; See Table S5 for all lineage enrichment results) (Vasung et al., 2010). Together, the spatial and temporal enrichment data suggest that scent hound-associated variation may principally affect innervation of the frontal cortex, evoking the hypothesis that human-directed selective pressure shaped the neurodevelopment of scent hound breeds toward a unique mode of sensory-associative learning (Alcaraz et al., 2018; Turner and Parkes, 2020).

Substantial temporal enrichment of lineage-associated genes suggested a complex mode of behavioral trait specification in domestic dogs potentially involving iterative modulation of gene co-expression networks. Therefore, we examined expression of lineage-associated genes across human neurodevelopmental stages (Kang *et al.*, 2011). For each lineage, we detected multiple clusters of co-expressed lineage-associated genes, including the scent hound lineage-associated late fetal co-expression timepoint (Figure 5). We performed a gene ontology enrichment test on all lineage-associated genes regardless of lineage or brain expression, and found that the most significantly enriched biological process categories were “positive regulation of developmental process” and “positive regulation of cell differentiation” (both p adj.=7.64E-7), and that 48% of genes in either category were also implicated in the overlapping category of “neurogenesis” (all genes enrichment p adj.=1.16E-3), indicating a broad role for fetal neuron formation in conferring canine behavioral phenotypes across lineages (Table S5).

Given that all lineages showed associations with genes implicated in fetal brain development, and the relevance of such pathways in common human neuropsychiatric conditions, we examined the overlap between lineage-associated genes and risk genes identified in a meta-analysis of eight DSM-defined psychiatric disorders and a database of Autism Spectrum disorder (ASD) risk genes (Figure 5; See Methods for risk gene curation methodology) (Banerjee-Basu and Packer, 2010; Consortium, 2019; Willsey et al., 2013). Forty-three of the top lineage-associated brain-expressed genes are also identified as risk-associated in one or both of the cross-disorder and ASD datasets (Figure 5, Table S5). Seventeen of these genes were associated with at least two of the eight disorders analyzed in the cross-disorder dataset and three were associated with >5 of the 8 disorders (*DCC, SORCS3, KCNQ5*), indicating an outsized role for pleiotropic human neuropsychiatric risk genes in the modulation of canine behavior (Consortium, 2019).

We also identified significant enrichments for non-disordered human behavioral associations among lineage-associated genes based on a comprehensive collection of 5,330 human GWAS publications (MacArthur et al., 2017; Watanabe et al., 2017). Notably, three lineages showed significant enrichment of genes associated with "general risk tolerance" in humans (pointer-setter lineage p adj. = 1.78E-4; terrier lineage p adj. = 1.33E-4; retriever lineage p adj. = 6.94E-5) (Table S5) (Karlsson Linnér et al., 2019). The significant overlap between lineage-associated genes and those implicated in human behavior, as well as pleiotropic mechanisms across human psychiatric disorders, suggests shared cellular mechanisms underlying neurodiversity among humans and dogs.

Sheepdog-associated loci are enriched near axon guidance genes

Livestock herding dogs display one of the most easily defined breed-typical behaviors, characterized by an instinctive herding drive coupled with unique motor patterns that move herds in complex ways (Coppinger and Coppinger, 2014; Serpell, 2017; Wilcox and Walkowicz, 1989). Because the terminus of the herder lineage formed a distinct cluster of sheepdog breeds, we sought to identify loci underlying the unique traits of this behaviorally distinct cluster versus the rest of the herder lineage (Figure 6A, Figure S5A). We performed a WGS-based GWAS followed by gene set enrichment analysis, and identified a single significantly enriched KEGG pathway, "axon guidance" (hsa04360), which included 14 sheepdog cluster-associated genes (p adj.=4.25e-2; Figure 6B, Table S5, Table S6) (Liberzon et al., 2011; Watanabe *et al.*, 2017). Eight of these sheepdog cluster-associated genes were axon guidance receptors, four were guidance cues, and two were GTPase activating proteins (Figure 6C). Several of these genes converge spatiotemporally in a variety of developmental contexts, particularly in midline patterning (Figure 6C) (Ducuing et al., 2019; Prieur and Rebsam, 2017).

While the associated axon guidance genes included those with roles in netrin, ephrin, slit, and semaphorin signaling pathways, the greatest number were part of the ephrin signaling pathway, including three receptor genes (Figure 6B, Figure 6C). For example, we identified an 11.3kb sheepdog-associated locus within the same topological domain as the axon guidance gene *EPHA5*, which encodes an ephrin receptor broadly expressed in the developing brain and implicated in dopaminergic neuronal projection guidance (Figure 6B, Figure S5A) (Cooper et al., 2009; Deschamps et al., 2009). The associated haplotype of 18 variants (index SNP adj. p = 2.577E-8) was present at an allele frequency of 77.5% in border collies, 3.1% in non-sheepdog cluster herder lineage dogs, and 7.6% in all non-border collie purebred dogs (Figure S5).

Examination of allele frequencies among “working” border collies versus a genetically distinct population of “show” border collies, established in the last few decades following AKC breed recognition (Figure S5B), revealed a haplotype frequency of 60% in show lines versus 83.3% in working lines, suggestive of relaxation of behavioral selective pressure among individuals bred for aesthetic conformation (Figure S5C). Moreover, prevalence of a core subset of variants within the border collie haplotype among wild canids indicates that human-directed pressure to produce dogs that herd sheep may have involved co-opting an ancient regulatory haplotype (Figure S5C). Together, these results indicate a relatively complex mode of inheritance for sheepdog-specific phenotypes involving a set of variants that iteratively modulate neuronal circuity, evoking an intriguing hypothesis regarding the ancient underpinnings of herding in mammalian behavior and the origins of genetic variants conferring such behaviors among early dogs.

Discussion

Humans selectively bred dogs for specific occupations, producing a powerful system for understanding how behavior is encoded in genomes. Previous canine genetic mapping and association studies have largely focused on loci underlying behavioral shifts related to domestication or the genetic causes of canine behavioral problems (Dodman *et al.*, 2016; Duffy et al., 2008; Pendleton *et al.*, 2018; Tang *et al.*, 2014; vonHoldt et al., 2017; Zapata et al., 2016). However, perhaps the most consequential application of the canine genetic system is its utility in understanding the genetics of breed-defining behaviors. Significant heritability for breed-typical canine behaviors has previously been noted (MacLean et al., 2019; Svartberg, 2006). However,

identification of contributing loci has historically been challenging due to the inherent complexity of canine population dynamics, featuring varying degrees of selective pressure for aesthetic and morphological traits, some of which may be linked to behavioral phenotypes, together with the frequent occurrence of population bottlenecks (van Rooy et al., 2014).

Embracing the intrinsically high-dimensional nature of canine genetic data while accurately distilling to its most fundamental axes of variation enabled our global approach to studying canine behaviors which focused on breed group-defining behavioral tendencies. Using genetic data from over 4,000 canids, we identified 10 major canine lineages characterized by behavioral features advantageous for distinct roles in human society. An exhaustive representation of every level of selective pressure applied to domestic dogs was critical for this study: from purebred dogs historically subjected to human-driven selective pressure, to those breeding under more permissive breed standards and, finally, to semi-feral dogs that experience little or no human-imposed selective pressure. By assembling a complete dataset and using methods capable of denoising complex genetic relationships without requiring adherence to a hierarchical phylogenetic structure, we ensured that the global trends identified herein were robust to the diverse genetic histories of breeds. We also employed the novel approach of implementing pseudotemporal reconstruction for analyzing genetic relationships, assigning a robust, quantifiable metric to individual dogs representing their positioning in the global context of domestic canid genomes. While our genetic dataset represents a temporally restricted snapshot of global canine diversity, our lineage-based methodology enabled identification of the most fundamental genetic drivers of canine diversification into behaviorally unique varieties based on their unique genotypes and traits relative to a lineage-specific ancestral domestic dog state. This lineage-based approach enables the identification of potentially divergent origins of similar

behaviors across lineages. Moreover, by establishing correlations between pseudotime and breed-average behavioral scores, we identify genetic associations robust to the noisiness of genotype-phenotype associations based on individual dogs and without the necessity for inferring breed-typical behaviors based on anecdote or kennel club classifications.

In addition to identifying loci underlying genetic diversification along lineages, we demonstrate the power of using a lineage framework for identification of alleles that differentiate genetically and phenotypically distinct groups of breeds from their lineage counterparts. For example, we identify loci associated with a genetically distinct cluster of herding breeds primarily comprised of sheepdogs. We determined that these sheepdog-associated loci are significantly overrepresented near axon guidance genes and genes significantly enriched for late mid-prenatal human brain expression (p adj. = 4.28E-2; Table S5). The striking frequency of loci associated with synaptic components, including both structural components of the synapse and various neurotransmitter receptors, may be pertinent to numerous developmental and functional processes relevant to behavior. Should genetic variation modulating expression of these genes converge spatially, they evoke an intriguing hypothesis regarding binocular vision in sheepdogs and potential implications for motor behavior, given their functional convergence in axon sorting at the optic chiasm (Larsson, 2015; Prieur and Rebsam, 2017). Our identification of a sheepdog-associated haplotype upstream of *EPHA5* is intriguing given its implication in multiple human neuropsychiatric conditions, most notably ADHD (Liu et al., 2021; Matoso et al., 2013; Noor et al., 2014; Weiss et al., 2009). Additionally, *EPHA5* and its ligand *EFNA5*, both sheepdog-associated genes, have been implicated in anxiety and maternal pup-gathering behavior in knockout mouse models, suggesting that herding drive among sheepdogs may involve augmentation of the same anxiety-associated biological mechanisms that drive spontaneous

maternal protective behaviors in other species (Mamiya et al., 2008; Sheleg et al., 2017; Weber and Olsson, 2008).

Whereas canine aesthetic traits or extreme morphological phenotypes often involve high-impact *de novo* mutations, we find that canine diversification along lineages is largely driven by non-coding variants often present at moderate frequencies in non-lineage populations, suggesting a substantially more complex mode of inheritance for canine behavioral versus non-behavioral traits (Boyko et al., 2010; Hayward et al., 2016; Plassais *et al.*, 2019). Therefore, this study characterized lineage-associated loci in a unified manner, identifying brain regions and developmental time periods most likely to be implicated in canine behavioral diversification, as opposed to interpreting the significance and biological functions of individual variants. By doing so, we address an outstanding question in the field of canine genetics: the degree to which breed diversification was driven by selective breeding for behavioral traits versus morphological (e.g., body size) or aesthetic (e.g., coat type) traits. While lineages identified herein include subsets of aesthetically or morphologically distinct breeds, this study establishes two lines of evidence indicating the outsized effect of behavior-based selective breeding on breed diversification: the correlation between genetic lineages and behavioral shifts and the overrepresentation of brain and behavior-related genes in lineage-associated loci. Moreover, by identifying human GWAS gene sets significantly enriched for lineage-associated genes, we approximate the types of traits humans principally targeted when molding domestic dogs into varieties required for human survival and sport.

In addition, our findings indicate that historical behavioral selective pressure was not uniform across all lineages. For example, the herder, scent hound, and terrier lineages showed the most robust behavioral correlations ($p < 1E-3$) as well as significant lineage-associated gene

enrichment in one or more prefrontal areas of the cortex associated with human social cognition, decision-making, and emotional intelligence (Duncan and Owen, 2000; Fleck et al., 2006; Lane et al., 1997). In contrast, the sighthound lineage showed no robust behavioral correlations ($p > 0.01$) nor forebrain or midbrain tissue enrichment for lineage-associated genes. Instead, the most significant enrichment for sighthound lineage-associated genes was for subcutaneous adipose (p adj. = 3.68E-6), consistent with selective breeding predominately aimed at morphological features conducive to athleticism.

Together, our findings indicate that the primary outcome of human-directed selective breeding on domestic dogs is a spectrum of diverse canine genomes that principally differ in how they regulate gene expression in the brain. This study provides a critical advance in the identification of genomic regions driving canine behavioral diversity, revealing lineage-associated brain co-expression networks and providing an entry point for targeted follow-up studies aimed at deepening our understanding of how complex mammalian behaviors are biologically encoded.

Limitations of the Study

This study establishes that canine breed diversification is principally driven by non-coding variation. However, the present study did not perform structural variant analysis and therefore does not identify potentially consequential copy number variants within associated loci (Serres-Armero et al., 2021). In addition to the above, this study relied on human annotations as part of a broad approach to characterizing loci. Identification of canine-specific regulatory elements within associated loci and topological associations will be done in future studies using newly available long-read canine reference genomes with improved annotations (Halo et al.,

2021; Jagannathan et al., 2021; Wang et al., 2021). Finally, our identification of behavioral shifts among domestic dog breeds, though robust, are based on a questionnaire designed to measure broad aspects of temperament in dogs rather than expressly targeting breed-typical behavioral functions (e.g., pointing, herding, scent-trailing, etc.). Therefore, future studies of behavioral phenotypes in dogs should aim to include measures of these specialized variants.

The identification of canine lineage-associated loci previously implicated in human behavioral traits and neuropsychiatric disorders is promising, as we show here. However, it remains difficult to ascribe finite canine behavioral traits to individual loci in much the same way assigning loci to risk for human psychiatric disorders is complicated by diagnostic overlap between disorders, locus heterogeneity, pleiotropy, and the effects of common variation on risk (Consortium, 2019). Nonetheless, this study uncovers shifts in canine behavior along lineages and the genomic loci associated with them, providing an essential foundation for future studies examining relationships between individual variants and behavior. Moreover, the enrichment among lineage-associated genes for pleiotropic human neuropsychiatric risk genes indicates a role for the canine genetic model in deconvoluting such complex genetic mechanisms underlying human psychopathology within a simpler genetic and behavioral architecture.

Methods

Data Ascertainment and Availability

We compiled a genomic dataset of 1,171 canid whole genome sequences obtained from the Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) or generated by the NIH Intramural Sequencing Center, 1,000 of which were previously published. Accession and

biosample numbers for previously published and unpublished WGS samples are listed in Table S1. In addition, the complete WGS dataset used in this study is available under accession number X. We also compiled a dataset of 3,090 Illumina Canine HD SNP chip (Illumina, San Diego, CA) arrays, a subset of which were previously published (Table S1). The full SNP genotype array dataset used in this study is available at GEO accession X and all SNP array sample information is available in Table S1. The combined WGS and SNP dataset (n = 4,261) is comprised of 2,823 purebred dogs (FCI-recognized, owner-reported or registered), 687 non-recognized or mixed breed dogs (including three wolf-dog hybrid breeds, one of which is FCI-recognized), 658 village dogs, and 93 wild canids. Table S1 contains sample metadata including FCI breed standard-based information and collection location for village dogs and wild canids.

WGS and SNP Array Quality Control

The 3,090 SNP array datasets used in this study were filtered based on sample and variant missingness and to include only biallelic autosomal SNVs. We extracted the SNP array variants from the filtered WGS dataset (WGS filtering described below) to generate a combined SNP dataset for 4,261 individuals. The maximum sample missingness for the final dataset of 102,764 SNPs was 0.20%. Duplicate samples from both datasets combined were removed using kinship analysis with KING by calculating a cutoff value based on kinship values between known duplicates in the dataset (Manichaikul et al., 2010). Duplicate samples were pruned by keeping the sample with the greater sequencing depth for WGS data or lower missingness for SNP array data. After filtering, the maximum kinship value for the combined dataset of WGS and SNP array samples was 0.35.

Alignment and variant calling for WGS data was performed as previously described (Plassais *et al.*, 2019). WGS data was subsequently filtered to remove low quality samples and genotypes. Hard filtering was performed based on GATK guidelines using bcftools: bcftools filter -e 'FS>60.0 || SOR>3 || MQ<40 || MQRankSum<-12.5 || MQRankSum>12.5 || QD<2.0 || ReadPosRankSum<-8.0' -Ob | bcftools filter -S . -e 'FMT/DP<3 | FMT/GQ<10' | bcftools filter -e 'F_MISSING > 0.05 || MAF <= 0.0017' -Ob | bcftools view -e 'AC==0 || AC==AN' -m2 -M2 (Danecek et al., 2021; Van der Auwera et al., 2013). The allele frequency threshold was determined based on a requirement for at least four minor alleles in the final dataset. The filtered WGS dataset of 1,171 samples contained 20,667,447 variants with a maximum sample missingness of 6.95% and minimum sequencing depth of 10x. The average WGS sample sequencing depth and missingness after filtering were 23x and 1.48%, respectively. Inbreeding (F coefficient) and homozygosity were calculated using PLINK v1.9 (Purcell et al., 2007).

Dimensionality reduction and pseudotemporal reconstruction

Principal component analysis (PCA) was performed using a variance-standardized relationship matrix based on the combined dataset of 102,764 SNPs (SNP array positions) for 4,261 samples using PLINK v1.9 (Purcell *et al.*, 2007). For subsequent analyses, 20 PCs (accounting for 47% of the cumulative variance) were used, with subsequent PCs each accounting for less than 1.15% of variance (Figure S1B). UMAP was run on a matrix containing the first 20 PCs using the umap package in R (R v4.1.0, umap 0.2.7.0) (McInnes *et al.*, 2018). UMAP plots were generated with multiple gamma, minimum distance, and spread values as indicated on plots in Figure S1A. The two-dimensional PHATE embedding was generated using

the first 20 PCs as input with phate(pca_matrix, ndim=2, knn=80, gamma=0, decay=100) (R v4.1.0, phateR v1.0.7, python 3.7) (Moon *et al.*, 2019). Three-dimensional PHATE was run with the same settings but with ndim=3 (Moon *et al.*, 2019).

Slingshot (R 3.6.2, Slingshot v1.4.0) was used for identifying lineages using PHATE coordinates as input and clusters defined using kmeans(centers=40) with the wild canid-containing cluster (other than Dingo-related wild canids) set as the starting cluster (Street *et al.*, 2018). Pseudotime values were obtained for each of the 10 Slingshot-identified lineages using slingPseudotime (Street *et al.*, 2018) (Table S2). For phylogenetic analysis, linkage pruning was performed on the merged SNP dataset with PLINK v1.9 using --indep 50 5 2 (Purcell *et al.*, 2007). Pruned SNPs were used for distance matrix generation with PLINK v1.9 using --distance 1-ibs square for neighbor-joining tree generation with Phylip v3.697 (100 matrices generated, neighbor and consense functions) (Felsenstein, 2005; Purcell *et al.*, 2007).

Analysis of C-BARQ behavioral data

C-BARQ survey data was compiled as of July 2021 and contained 67,970 responses. Scores for the 14 previously defined behavioral factors assessed in C-BARQ (Trainability, Stranger-Directed Aggression, Owner-Directed Aggression, Dog-Directed Aggression, Familiar Dog Aggression, Dog-Directed Fear, Stranger-Directed Fear, Nonsocial Fear, Touch Sensitivity, Separation-Related Problems, Excitability, Attachment/Attention-Seeking, Predatory Chasing, Energy Level) were calculated as averages of the raw response values for the corresponding questions (Duffy and Serpell, 2012; Hsu and Serpell, 2003; Serpell and Hsu, 2005). We filtered the C-BARQ dataset to eliminate multiple entries per individual as well as individuals missing

greater than 20% of responses for any one factor, resulting in a dataset of factor scores for 46,973 unique individuals. The filtered C-BARQ dataset contained 237 domestic dog breeds (as reported by owner) with an average of 136 individuals per breed. We generated a UMAP plot for all individuals based on a matrix of C-BARQ factor scores for each individual using uwot (v0.1.11) in R (v3.6.2) (McInnes *et al.*, 2018). Density of points within UMAP plot by group was calculated as two-dimensional kernel estimations evaluated using 100 grid points in each direction (MASS v7.3-51.5).

For correlation analyses, we generated breed-average C-BARQ factor scores for the 132 breeds having at least 30 individuals in the C-BARQ dataset as well as representation in the genetic dataset. The dingo and African and Middle Eastern lineages were eliminated from correlation analyses due to lack of sufficient representation in the C-BARQ dataset. For each lineage, we calculated breed-average pseudotime values, requiring that a majority of individuals in the breed belonged to the lineage in order to exclude breed averages based on single outlier individuals. We calculated the two-sided Spearman's rank correlation rho and p values between breed-average pseudotime value and breed-average C-BARQ factor score for each lineage independently for each of 14 behavioral factors using cor(method="spearman") (R 3.6.2 Stats package). We used the same method to calculate correlations with breed-average gray wolf mean IBS (as described above) and breed-average maximum lineage pseudotime. Working trial requirements for FCI-recognized breeds were obtained from FCI breed standards (fci.be/nomenclature/). ANOVA and Tukey tests for behavioral comparisons between working trial groups were performed with all breed averages using aov and TukeyHSD (R 3.6.2 Stats package).

Lineage Association Analysis

Lineage pseudotime genome-wide association analyses (GWAS) were performed using WGS data from 1,167 individuals. GWAS were performed separately for each lineage (other than the sled dog, African and Middle Eastern, and dingo lineages which were eliminated due to lack of representation in the WGS dataset) and also with maximum lineage pseudotime per individual regardless of lineage. GWAS were performed using the PLINK v2.3 association analysis command with --mac 10 --glm --adjust gc and --lamba where the lambda value was set as the LD score regression intercept calculated using LDSC (Bulik-Sullivan et al., 2015; Purcell *et al.*, 2007). The LD score regression intercepts for each lineage association were calculated with ldsc.py –l2 –ld-wind-kb 1000 –maf 0.004 followed by munge_sumstats.py and ldsc.py –h2 (Bulik-Sullivan *et al.*, 2015). Bonferroni-adjusted p values from each GWAS were used for the a variant clumping procedure implemented with PLINK (v1.9) using –maf 0.004 –clump-p2 0.01 –clump-p1 0.001 –clump-r2 0.50 –clump-kb 1000 –clump-field BONF in order to define start and end points of associated loci (Purcell *et al.*, 2007). The minor allele frequency (MAF) cutoff of 0.004 corresponds to restriction to variants with at least 10 minor alleles in the WGS dataset to match the cutoff (--mac 10) implemented in the association analysis command. Coordinates of loci were calculated based on the first and last positions of clumped variants.

For analyses based on the top 100 loci per lineage, locus ranks were calculated based on the index variant Bonferroni-adjusted p value. Of the top 100 loci for each of the seven individual lineages and combined lineage GWAS, 42% of loci were less than 10 kbp in length and 676 out of 800 loci had more than one clumped variant with GWAS Bonferroni-adjusted p value < 0.01. Circular Manhattan plot was generated using CMplot 3.7.0.

Annotation of lineage-associated variants and enrichment tests

Variant coordinates were lifted from CanFam3.l to hg19 (GRCh37), mm10 (GRCm38), and hg38 (GRCh38) genome builds with liftOver and default settings. We required that lifted variants reciprocally mapped to the correct CanFam3.1 position (Rhead et al., 2010). DeepSEA was performed on reciprocally-lifting hg19 coordinates using rundeepsea.py (Zhou and Troyanskaya, 2015). DeepSEA functional significance scores were calculated for each canine variant that had at least one allele matching the reference human allele by running DeepSEA with both canine alleles set as the human reference allele to account for either allele matching the human reference allele (approximately 57% of reciprocally lifting variants) (Zhou and Troyanskaya, 2015). To calculate phyloP scores for each variant mapped to hg38, we used bigWigAverageOverBed and hg38.phyloP100way.bw obtained from the UCSC Genome Browser at http://hgdownload.cse.ucsc.edu (Pollard *et al.*, 2010; Rhead *et al.*, 2010). We used both SnpEff (v5.0 with CanFam3.1.99) and VEP (v104) to identify lineage-associated coding variants (Cingolani et al., 2012; McLaren et al., 2016).

Lifted coordinates were intersected with all candidate cis-regulatory element (cCRE) regions for human (n=926,535) and mouse (n=339,815) obtained from ENCODE (Moore *et al.*, 2020). Regulatory element annotations for tissues were obtained using these ENCODE cCRE IDs and merged for all human and mouse tissues that had replicate samples available in the SCREEN database (Table S4). All cCREs were treated as active if annotated as anything other than unclassified or low-DNase (any combination of H3K27ac, H3K4me3, CTCF, DNase). cCREs were filtered for those that replicate at least once within the replication groups as

described in Table S4, resulting in 261,643 human cCREs and 169,594 mouse cCREs. Mouse adult brain, mouse adult liver, and mouse adult spleen were removed due to lack of available replicates. All cCREs with intersecting lineage-associated variants were also lifted to CanFam3.1 with liftOver using -minMatch=0.5 to require reciprocal mapping back to the correct CanFam3.1 position (Rhead *et al.*, 2010).

Predicted target genes for all loci were identified using GREAT (McLean et al., 2010). First, associated locus coordinates were defined as described above. The median locus size for input to GREAT was 42.7 kb. All clumps were lifted to hg38 using liftOver with minMatch=0.5 (Rhead *et al.*, 2010). Next, the rGREAT client for GREAT analysis R package (v1.20.0) was used to predict target genes using submitGreatJob (includeCuratedRegDoms=FALSE) and getEnrichmentTables with the default GREAT gene regulatory domain definition (Basal plus extension) (McLean *et al.*, 2010). Genes predicted for the top 100 loci per lineage or combined lineages (172-188 genes per lineage) were used for input to FUMA (v1.3.7) for gene set enrichment with the background gene set specified as protein coding genes (ensemble v92). Human gene expression values and enrichment significance values were obtained with FUMA based on the GTEx v8 (54 sample types, of which 52 are tissues versus cell lines; 30 general tissue types) and BrainSpan (29 different ages of brain samples and 11 general developmental stages of brain samples) datasets (Consortium *et al.*, 2020; Kang *et al.*, 2011; Watanabe *et al.*, 2017). The 11 developmental stages of BrainSpan human brain tissue are as follows: early prenatal (8-12 pcw), early mid-prenatal (13-18 pcw), late mid-prenatal (19-24 pcw), late prenatal (25-38 pcw), early infancy (0-5m), late infancy (1-18m), early childhood, (19m-5y) late childhood (6-11y), adolescence, (12-19y) young adulthood and middle adulthood (Kang *et al.*, 2011). All optional parameters were set to defaults (Benjamini-Hochberg multiple testing

correction for gene-set enrichment testing, maximum adjusted P-value for gene set association < 0.05, and minimum overlapping genes >= 2) (Benjamini and Hochberg, 1995; Watanabe *et al.*, 2017). Human behavioral trait GWAS gene set enrichment was performed using FUMA and was based on the NHGRI GWAS catalog version e104 (2021-09-15) (MacArthur *et al.*, 2017; Watanabe *et al.*, 2017). Hierarchical clustering in all heatmaps was performed using pheatmap (Kolde, 2015)

Sheepdog cluster GWAS, variant clumping, and locus annotation were performed as described above except with a case/control (case n=63 sheepdog cluster individuals; control n=321 non-sheepdog cluster herder lineage individuals; Table S1) rather than quantitative phenotype with PLINK v1.9 (Purcell *et al.*, 2007). KEGG pathway gene set enrichment analysis was performed using FUMA v1.3.7 as described above. Owner-reported data was used to identify working border collies versus those bred for aesthetic competitions (show border collies) and designations were verified based on pairwise IBS data as shown in Figure S5B. Previously published WGS data were obtained for the golden jackal (PRJNA274504, SAMN03366713), dhole (PRJNA266585, SAMN03168405), and Andean fox (PRJNA232497, SAMN02487034) for the purposes of identifying allele conservation in Figure S5C, but were not used in any other analyses. UCSC genome browser images show CanFam3.1 protein-coding genes (Ensembl v104) and human neural progenitor cell (hNPC) topologically associating domain (TAD) regions lifted to CanFam3.1 as described above (Aken et al., 2016; Dixon et al., 2015; Rhead *et al.*, 2010).

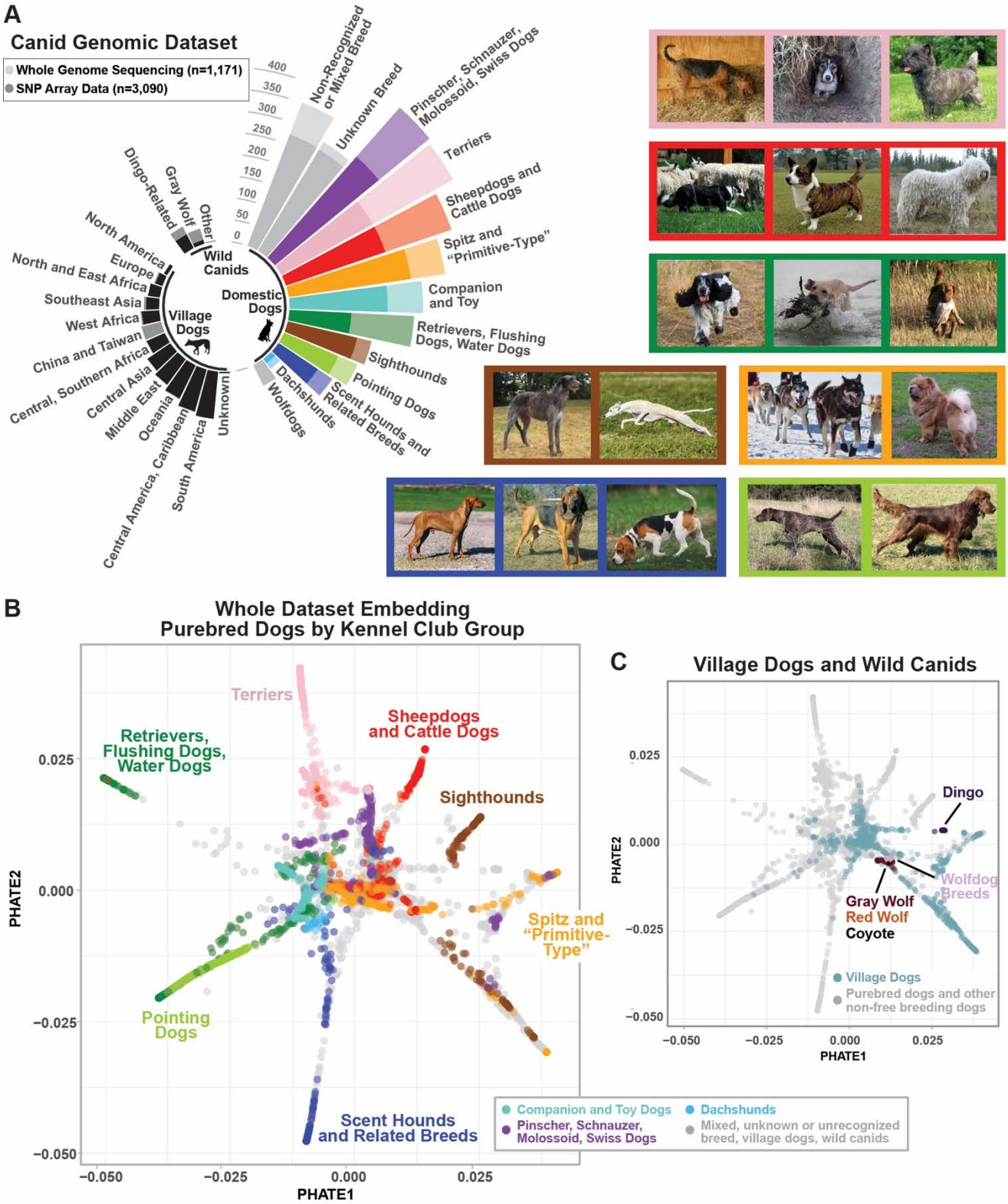

A
Canid Genomic Dataset
Whole Genome Sequencing (n=1,171)
SNP Array Data (n=3,090)
400
350
300
250
200
150
100
50
0
Non-Recognized or Mixed Breed
Unknown Breed
Pinscher, Schnauzer, Molossoid, Swiss Dogs
Terriers
Sheepdogs and Cattle Dogs
Spitz and "Primitive-Type"
Companion and Toy
Retrievers, Flushing Dogs, Water Dogs
Sighthounds
Pointing Dogs
Scent Hounds and Related Breeds
Dachshunds
Wolfdogs
Unknown
Dingo-Related
Gray Wolf
Other
Wild Canids
Domestic Dogs
Village Dogs
North America
Europe
North and East Africa
Southeast Asia
West Africa
China and Taiwan
Central, Southern Africa
Central Asia
Middle East
Oceania
Central America, Caribbean
South America
B
Whole Dataset Embedding
Purebred Dogs by Kennel Club Group
Terriers
Sheepdogs and Cattle Dogs
Retrievers, Flushing Dogs, Water Dogs
Sighthounds
Spitz and "Primitive-Type"
Pointing Dogs
Scent Hounds and Related Breeds
PHATE2
PHATE1
0.025
0.000
−0.025
−0.050
Companion and Toy Dogs
Dachshunds
Pinscher, Schnauzer, Molossoid, Swiss Dogs
Mixed, unknown or unrecognized breed, village dogs, wild canids
C
Village Dogs and Wild Canids
Dingo
Wolfdog Breeds
Gray Wolf
Red Wolf
Coyote
Village Dogs
Purebred dogs and other non-free breeding dogs

Figure 1. Canid dataset and dimensionality reduction (A) Left: Dataset frequencies by sample and data type grouped by kennel club group designation (Supplemental Text), non-recognized breed, unknown breed, or wolfdog hybrid breeds. Village dogs are grouped by sampling location. Three categories of wild canid are indicated: gray wolves, dingo-related canids and other (red wolf, coyote). Darker color of each stacked bar indicates number of SNP array samples; upper portion (light color) indicates number of WGS samples. See also Table S1. Right: Images of breeds within seven kennel club groups colored according to bar plot. **(B)** PHATE plot of all canid samples colored by group as in panel A. Wild canids and dogs with no kennel club group designation are colored in gray. **(C)** PHATE plot with wild canids and village dogs colored by sample type. All remaining domestic dog samples colored in gray.

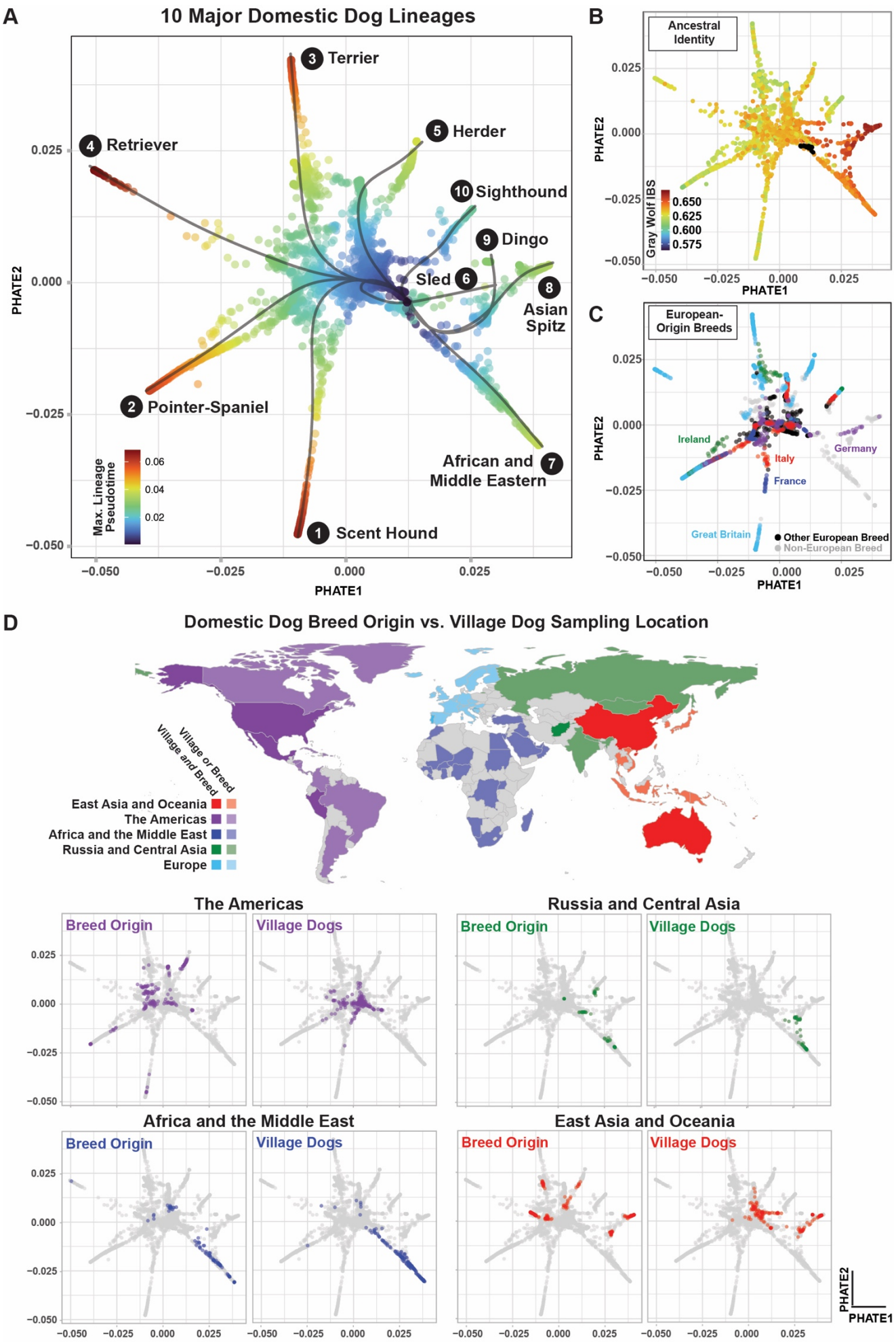
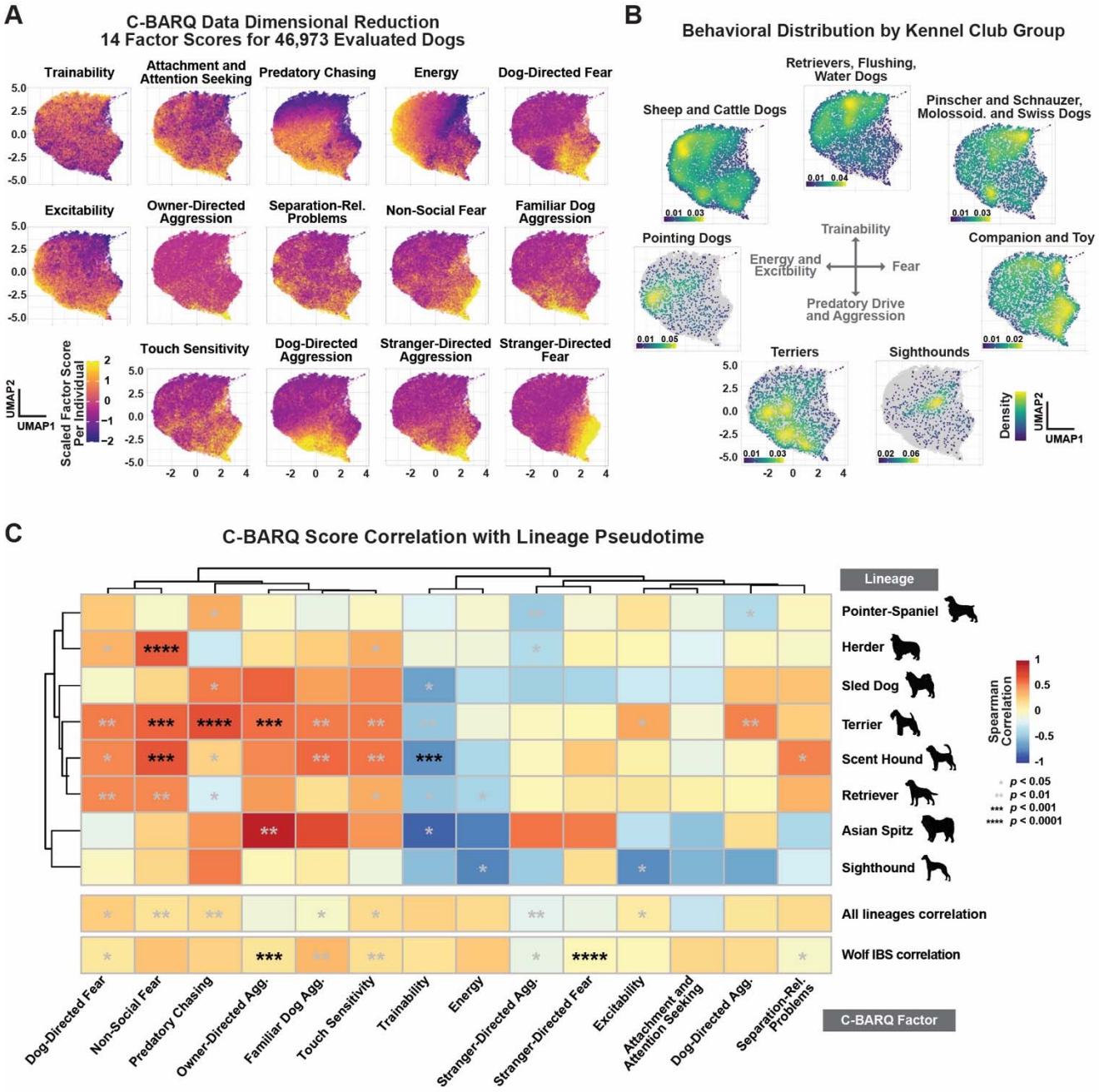

A
10 Major Domestic Dog Lineages
3 Terrier
4 Retriever
5 Herder
10 Sighthound
9 Dingo
Sled 6
8 Asian Spitz
2 Pointer-Spaniel
African and Middle Eastern 7
1 Scent Hound
Max. Lineage Pseudotime
0.06
0.04
0.02
PHATE1
PHATE2
B
Ancestral Identity
Gray Wolf IBS
0.650
0.625
0.600
0.575
C
European-Origin Breeds
Ireland
Italy
France
Germany
Great Britain
Other European Breed
Non-European Breed
D
Domestic Dog Breed Origin vs. Village Dog Sampling Location
Village or Breed
Village and Breed
East Asia and Oceania
The Americas
Africa and the Middle East
Russia and Central Asia
Europe
The Americas
Breed Origin
Village Dogs
Russia and Central Asia
Breed Origin
Village Dogs
Africa and the Middle East
Breed Origin
Village Dogs
East Asia and Oceania
Breed Origin
Village Dogs

Figure 2. Identification of canine lineages with pseudotemporal reconstruction. (A) PHATE plot delineating paths of 10 canine lineages originating with gray wolf. Points are colored by maximum pseudotime in any lineage and plotted lines are principal curves generated with Slingshot (Street *et al.*, 2018). **(B)** PHATE plot showing gray wolf IBS (average pairwise IBS between individual and each gray wolf). Wolf and coyote samples are colored black. **(C)**. PHATE plot indicating samples from breeds with European origins (non-European origin breeds in gray). **(D)** Top: Geographic origin of breeds (historical, FCI-based) and village dog sampling location. Darker shade indicates locations with breed origin and village dog representation in the dataset. Lighter shade indicates countries for which dataset includes either domestic or village dogs. Bottom: PHATE plots comparing breed origin (left in each pair) to village dogs of the same geographic location (right in each pair) for non-European geographic regions.


A
C-BARQ Data Dimensional Reduction
14 Factor Scores for 46,973 Evaluated Dogs
Trainability
Attachment and Attention Seeking
Predatory Chasing
Energy
Dog-Directed Fear
Excitability
Owner-Directed Aggression
Separation-Rel. Problems
Non-Social Fear
Familiar Dog Aggression
Touch Sensitivity
Dog-Directed Aggression
Stranger-Directed Aggression
Stranger-Directed Fear
Scaled Factor Score Per Individual
UMAP2
UMAP1
B
Behavioral Distribution by Kennel Club Group
Retrievers, Flushing, Water Dogs
Sheep and Cattle Dogs
Pinscher and Schnauzer, Molossoid. and Swiss Dogs
Pointing Dogs
Companion and Toy
Terriers
Sighthounds
Trainability
Energy and Excitbility
Fear
Predatory Drive and Aggression
Density
UMAP2
UMAP1

C
C-BARQ Score Correlation with Lineage Pseudotime
Lineage
Pointer-Spaniel
Herder
Sled Dog
Terrier
Scent Hound
Retriever
Asian Spitz
Sighthound
All lineages correlation
Wolf IBS correlation
Spearman Correlation
p < 0.05
p < 0.01
p < 0.001
p < 0.0001
Dog-Directed Fear
Non-Social Fear
Predatory Chasing
Owner-Directed Agg.
Familiar Dog Agg.
Touch Sensitivity
Trainability
Energy
Stranger-Directed Agg.
Stranger-Directed Fear
Excitability
Attachment and Attention Seeking
Dog-Directed Agg.
Separation-Rel. Problems
C-BARQ Factor

Figure 3. C-BARQ behavioral traits and lineage correlations. (A) UMAP plot generated from 14 C-BARQ factor scores for all 46,973 C-BARQ-evaluated dogs (see Methods). Points indicate individual dogs and are colored by one of 14 factor scores per plot. Scores were scaled and centered (z-score) across all individuals for each factor independently. **(B)** UMAP plot (as in A) showing density of individuals within kennel club breed groups (See Methods). **(C)** Heatmap showing the correlation (two-sided Spearman's rank correlation rho) between breed-average C-BARQ factor scores and breed-average pseudotime values, maximum lineage pseudotime value (all lineages correlation; Figure 2A), and average gray wolf IBS (see Figure 2B). Asterisks indicate corresponding significance values. See Figure S3C for scatterplots of selected factors and lineages.

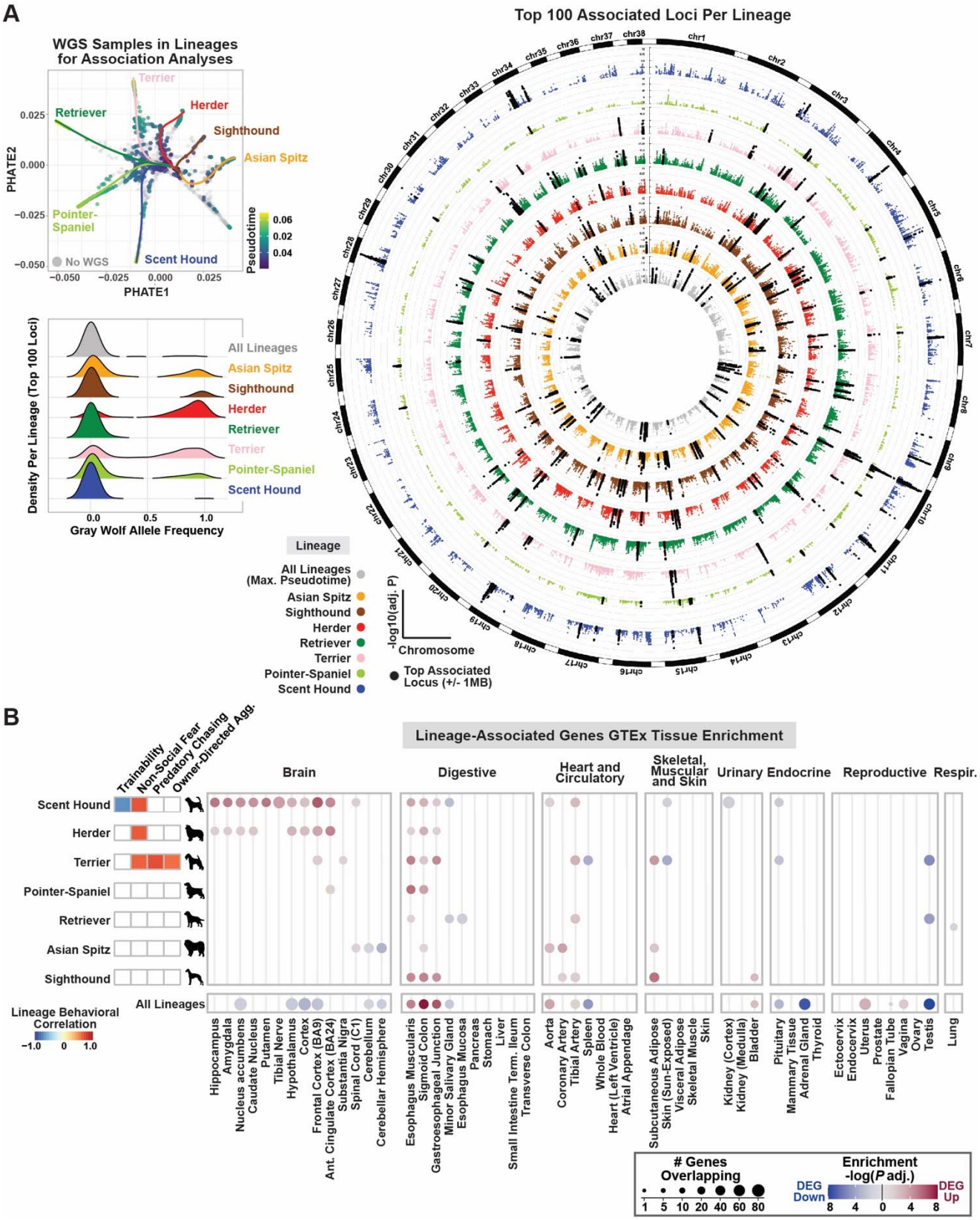

A
WGS Samples in Lineages for Association Analyses
Terrier
Herder
Retriever
Sighthound
Asian Spitz
Pointer-Spaniel
Scent Hound
No WGS
PHATE1
PHATE2
Pseudotime
Density Per Lineage (Top 100 Loci)
Gray Wolf Allele Frequency
All Lineages
Top 100 Associated Loci Per Lineage
Lineage
All Lineages (Max. Pseudotime)
Asian Spitz
Sighthound
Herder
Retriever
Terrier
Pointer-Spaniel
Scent Hound
-log10(adj. P)
Chromosome
Top Associated Locus (+/- 1MB)
B
Lineage-Associated Genes GTEx Tissue Enrichment
Trainability
Non-Social Fear
Predatory Chasing
Owner-Directed Agg.
Brain
Digestive
Heart and Circulatory
Skeletal, Muscular and Skin
Urinary
Endocrine
Reproductive
Respir.
Lineage Behavioral Correlation
All Lineages
Hippocampus
Amygdala
Nucleus accumbens
Caudate Nucleus
Putamen
Tibial Nerve
Hypothalamus
Cortex
Frontal Cortex (BA9)
Ant. Cingulate Cortex (BA24)
Substantia Nigra
Spinal Cord (C1)
Cerebellum
Cerebellar Hemisphere
Esophagus Muscularis
Sigmoid Colon
Gastroesophageal Junction
Minor Salivary Gland
Esophagus Mucosa
Pancreas
Stomach
Liver
Small Intestine Term. Ileum
Transverse Colon
Aorta
Coronary Artery
Tibial Artery
Spleen
Whole Blood
Heart (Left Ventricle)
Atrial Appendage
Subcutaneous Adipose
Skin (Sun-Exposed)
Visceral Adipose
Skeletal Muscle
Skin
Kidney (Cortex)
Kidney (Medulla)
Bladder
Pituitary
Mammary Tissue
Adrenal Gland
Thyroid
Ectocervix
Endocervix
Uterus
Prostate
Fallopian Tube
Vagina
Ovary
Testis
Lung
Genes Overlapping
Enrichment -log(P adj.)
DEG Down
DEG Up

Figure 4. Identification and characterization of lineage-associated loci and genes. (A) Top Left: PHATE plot showing maximum lineage pseudotime for WGS samples. Colored lines indicate principal curves for the seven lineages with sufficient WGS representation for association analyses (see Methods). Right: Circular Manhattan plots showing log-transformed Bonferroni-adjusted GWAS p-values. Loci highlighted in black are the 100 most significantly associated per lineage. Highlighted loci were expanded to 1MB for clarity. Bottom left: Ridge plot showing the distribution of gray wolf minor allele frequencies for each of the top 100 index variants per lineage as shown in Manhattan plot. Y-axis indicates density for each lineage. **(B)** Left: Correlation values for associations identified in the C-BARQ analysis with $p<0.001$ (See Figure 3C). Right: FDR-corrected significance values for tissue enrichment analysis with the GTEx v8 dataset using predicted target genes for the top 100 loci per lineage or combined lineages (see Methods). Cell lines (2 of 54 GTEx samples) not shown. Circle size denotes number of differentially expressed genes (continuous scale) and circles are colored by significance (log-transformed FDR-adjusted p values), with up and down-regulated differentially expressed genes (DEG) indicated by red and blue color scale, respectively.

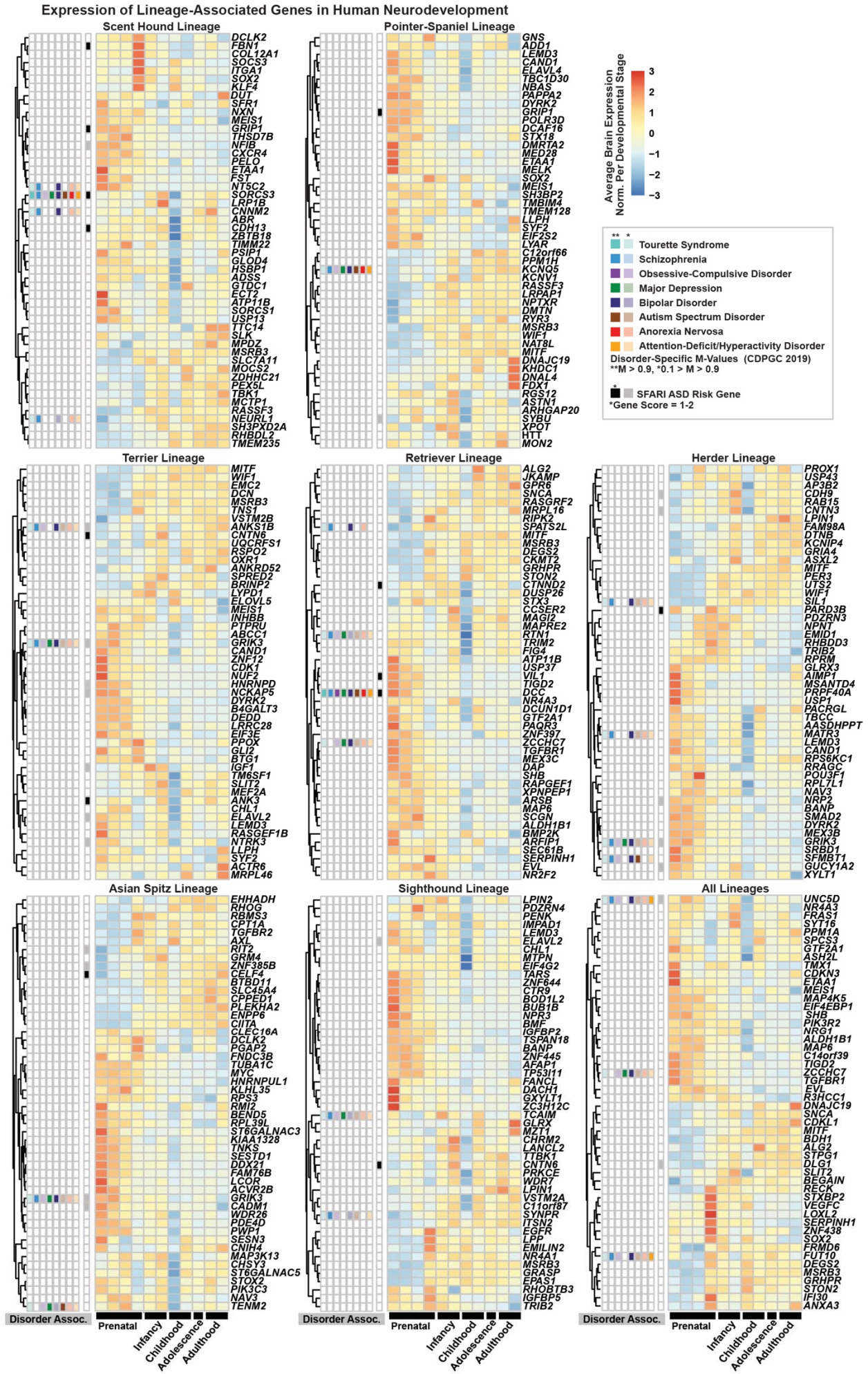

Expression of Lineage-Associated Genes in Human Neurodevelopment
Scent Hound Lineage
DCLK2
FBN1
COL12A1
SOCS3
ITGA1
SOX2
KLF4
DUT
SFR1
NXN
MEIS1
GRIP1
THSD7B
NFIB
CXCR4
PELO
ETAA1
FST
NT5C2
SORCS3
LRP1B
CNNM2
ABR
CDH13
ZBTB18
TIMM22
PSIP1
GLOD4
HSBP1
ADSS
GTDC1
ECT2
ATP11B
SORCS1
USP13
TTC14
SLK
MPDZ
MSRB3
SLC7A11
MOCS2
ZDHHC21
PEX5L
TBK1
MCTP1
RASSF3
NEURL1
SH3PXD2A
RHBDL2
TMEM235
Pointer-Spaniel Lineage
GNS
ADD1
LEMD3
CAND1
ELAVL4
TBC1D30
NBAS
PAPPA2
DYRK2
GRIP1
POLR3D
DCAF16
STX18
DMRTA2
MED28
ETAA1
MELK
SOX2
MEIS1
SH3BP2
TMBIM4
TMEM128
LLPH
SYF2
EIF2S2
LYAR
C12orf66
PPM1H
KCNQ5
KCNV1
RASSF3
LRPAP1
NPTXR
DMTN
RYR3
MSRB3
WIF1
NAT8L
MITF
DNAJC19
KHDC1
DNAL4
FDX1
RGS12
ASTN1
ARHGAP20
SYBU
XPOT
HTT
MON2
Average Brain Expression
Norm. Per Developmental Stage
3
2
1
0
−1
−2
−3
** *
Tourette Syndrome
Schizophrenia
Obsessive-Compulsive Disorder
Major Depression
Bipolar Disorder
Autism Spectrum Disorder
Anorexia Nervosa
Attention-Deficit/Hyperactivity Disorder
Disorder-Specific M-Values (CDPGC 2019)
**M > 0.9, *0.1 > M > 0.9
SFARI ASD Risk Gene
*Gene Score = 1-2
Terrier Lineage
MITF
WIF1
EMC2
DCN
MSRB3
TNS1
VSTM2B
ANKS1B
CNTN6
UQCRFS1
RSPO2
OXR1
ANKRD52
SPRED2
BRINP2
LYPD1
ELOVL5
MEIS1
INHBB
PTPRU
ABCC1
GRIK3
CAND1
ZNF12
CDK1
NUF2
HNRNPD
NCKAP5
DYRK2
B4GALT3
DEDD
LRRC28
EIF3E
PPOX
GLI2
BTG1
IGF1
TM6SF1
SLIT2
MEF2A
ANK3
CHL1
ELAVL2
LEMD3
RASGEF1B
NTRK3
LLPH
SYF2
ACTR6
MRPL46
Retriever Lineage
ALG2
JKAMP
GPR6
SNCA
RASGRF2
MRPL16
RIPK2
SPATS2L
MITF
MSRB3
DEGS2
CKMT2
GRHPR
STON2
CTNND2
DUSP26
STX3
CCSER2
MAGI2
MAPRE2
RTN1
TRIM2
FIG4
ATP11B
USP37
VIL1
TGD2
DCC
NR4A3
DCUN1D1
GTF2A1
PAQR3
ZNF397
ZCCHC7
TGFBR1
MEX3C
DAP
SHB
RAPGEF1
XPNPEP1
ARSB
MAP6
SCGN
ALDH1B1
BMP2K
ARFIP1
SEC61B
SERPINH1
EVL
NR2F2
Herder Lineage
PROX1
USP43
AP3B2
CDH9
RAB15
CNTN3
LPIN1
FAM98A
DTNB
KCNIP4
GRIA4
ASXL2
MITF
PER3
UTS2
WIF1
SIL1
PARD3B
PDZRN3
NPNT
EMID1
RHBDD3
TRIB2
RPRM
GLRX3
AIMP1
MSANTD4
PRPF40A
USP1
PACRGL
TBCC
AASDHPPT
MATR3
LEMD3
CAND1
RPS6KC1
RRAGC
POU3F1
RPL7L1
NAV3
NRP2
BANP
SMAD2
DYRK2
MEX3B
GRIK3
SRBD1
SFMBT1
GUCY1A2
XYLT1
Asian Spitz Lineage
EHHADH
RHOG
RBMS3
CPT1A
TGFBR2
AXL
RIT2
GRM4
ZNF385B
CELF4
BTBD11
SLC45A4
CPPED1
PLEKHA2
ENPP6
CIITA
CLEC16A
DCLK2
PGAP2
FNDC3B
TUBA1C
MYC
HNRNPUL1
KLHL35
RPS3
RMI2
BEND5
RPL39L
ST6GALNAC3
KIAA1328
TNKS
SESTD1
DDX21
FAM76B
LCOR
ACVR2B
GRIK3
CADM1
WDR26
PDE4D
PWP1
SESN3
CNIH4
MAP3K13
CHSY3
ST6GALNAC5
STOX2
PIK3C3
NAV3
TENM2
Sighthound Lineage
LPIN2
PDZRN4
PENK
IMPAD1
LEMD3
ELAVL2
CHL1
MTPN
EIF4G2
TARS
ZNF644
CTR9
BOD1L2
BUB1B
NPR3
BMF
IGFBP2
TSPAN18
BANP
ZNF445
AFAP1
TP53I11
FANCL
DACH1
GXYLT1
ZC3H12C
TCAIM
GLRX
MZT1
CHRM2
LANCL2
TTBK1
CNTN6
PRKCE
WDR7
LPIN1
VSTM2A
C11orf87
SYNPR
ITSN2
EGFR
LPP
EMILIN2
NR4A1
MSRB3
GRASP
EPAS1
RHOBTB3
IGFBP5
TRIB2
All Lineages
UNC5D
NR4A3
FRAS1
SYT16
PPM1A
SPCS3
GTF2A1
ASH2L
TMX1
CDKN3
ETAA1
MEIS1
MAP4K5
EIF4EBP1
SHB
PIK3R2
NRG1
ALDH1B1
MAP6
C14orf39
TIGD2
ZCCHC7
TGFBR1
EVL
R3HCC1
DNAJC19
SNCA
CDKL1
MITF
BDH1
ALG2
STPG1
DLG1
SLIT2
BEGAIN
RECK
STXBP2
VEGFC
LOXL2
SERPINH1
ZNF438
SOX2
FRMD6
FUT10
DEGS2
MSRB3
GRHPR
STON2
IFI30
ANXA3
Disorder Assoc.
Prenatal
Infancy
Childhood
Adolescence
Adulthood

Figure 5. Neurodevelopmental clustering of lineage-associated genes. Heatmaps showing the average expression of the 50 most significantly lineage-associated brain-expressed genes for each of 11 human neurodevelopmental stages (see Methods) (Kang *et al.*, 2011). Heatmap values are scaled in rows. Gene order is determined by hierarchical clustering. Genes included have log2-transformed expression values >1 for at least one developmental stage (prior to row-wise scaling). Human psychiatric disorder annotations for each gene are plotted to the left. The first eight columns include annotations for eight psychiatric disorders (See Methods). Colors correspond to each disorder with darker shades indicating a disorder-specific M-value (posterior probability of association) > 0.9 (Consortium, 2019). Lighter shades indicate 0.1 < M-value < 0.9. For SFARI ASD risk genes (black and gray squares), color coding indicates gene score according to SFARI database (Banerjee-Basu and Packer, 2010).

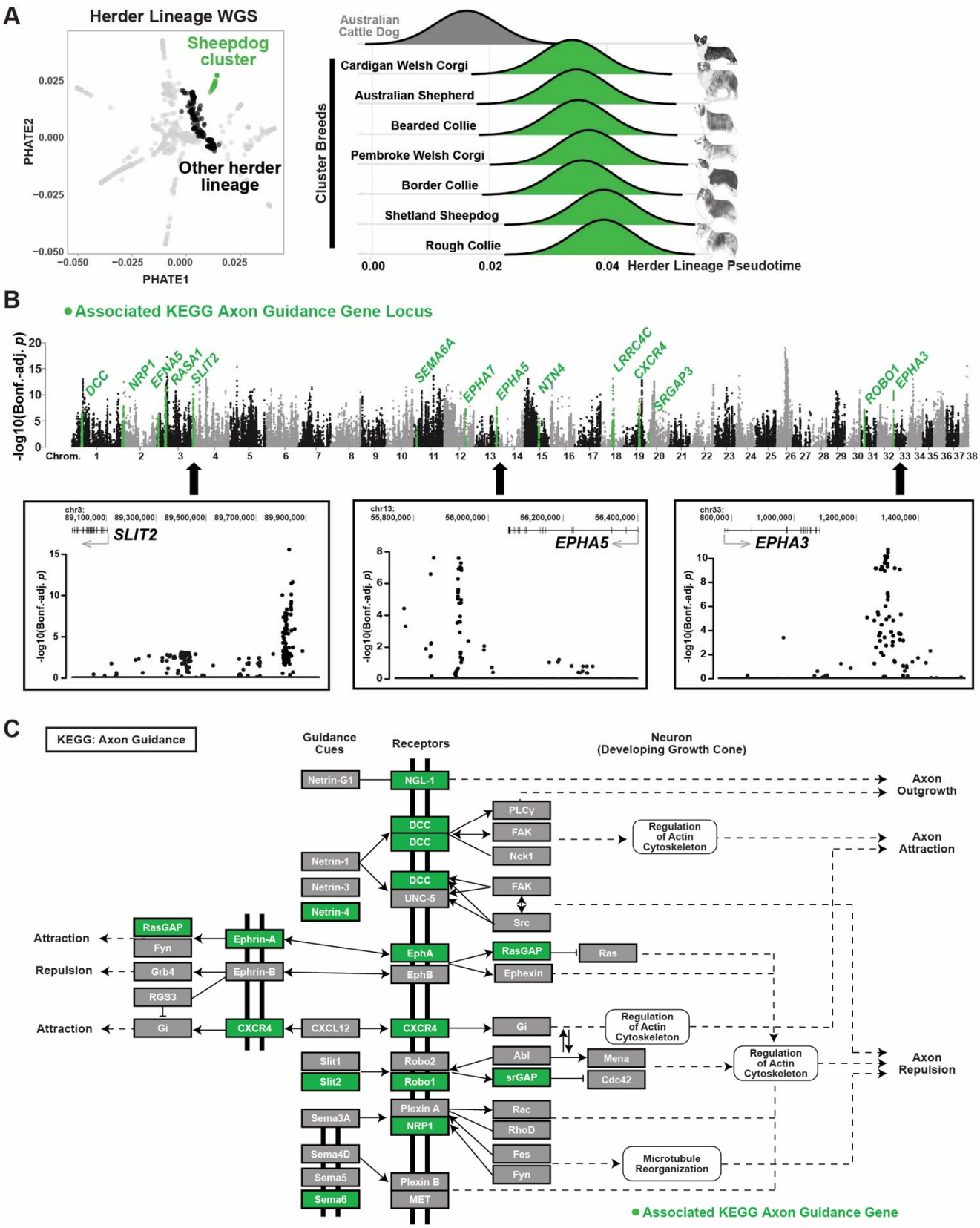

A
Herder Lineage WGS
Sheepdog cluster
Other herder lineage
PHATE2
PHATE1
Australian Cattle Dog
Cluster Breeds
Cardigan Welsh Corgi
Australian Shepherd
Bearded Collie
Pembroke Welsh Corgi
Border Collie
Shetland Sheepdog
Rough Collie
Herder Lineage Pseudotime
B
Associated KEGG Axon Guidance Gene Locus
-log10(Bonf.-adj. p)
Chrom.
DCC
NRP1
EFNA5
RASA1
SLIT2
SEMA6A
EPHA7
EPHA5
NTN4
LRRC4C
CXCR4
SRGAP3
ROBO1
EPHA3
SLIT2
EPHA5
EPHA3
C
KEGG: Axon Guidance
Guidance Cues
Receptors
Neuron (Developing Growth Cone)
Netrin-G1
NGL-1
Axon Outgrowth
DCC
PLCγ
FAK
Nck1
Regulation of Actin Cytoskeleton
Axon Attraction
Netrin-1
Netrin-3
Netrin-4
UNC-5
Src
Attraction
Repulsion
RasGAP
Fyn
Grb4
RGS3
Ephrin-A
Ephrin-B
EphA
EphB
Ras
Ephexin
Gi
CXCR4
CXCL12
Slit1
Slit2
Robo2
Robo1
Abl
srGAP
Mena
Cdc42
Axon Repulsion
Sema3A
Plexin A
NRP1
Rac
RhoD
Sema4D
Sema5
Sema6
Plexin B
MET
Fes
Microtubule Reorganization
Associated KEGG Axon Guidance Gene

Figure 6. Sheepdog-associated genes are enriched for axon guidance functions. (A) Left: PHATE plot showing only samples with WGS data. Individuals with herder lineage pseudotime values > 0.03 (sheepdog cluster) are green and all others with non-zero herder lineage pseudotime values (comparison group for GWAS) are black. WGS samples not part of sheepdog GWAS (non-herder lineage) are gray. Right: Ridge plot showing distribution of herder lineage pseudotime values for WGS samples within each breed in the sheepdog cluster (green) with the y-axis indicating density. The next closest breed by breed-average pseudotime values (Australian cattle dog) is also shown (gray). **(B)** Top: Manhattan plot showing sheepdog cluster GWAS results with log-transformed Bonferroni-adjusted p-values. KEGG axon guidance gene loci are indicated in green. Highlighted loci were expanded to 1MB for clarity. Bottom: genome browser images of three significantly associated loci. **(C)** Axon guidance pathway generated by KEGG (hsa04360) modified for clarity with highlighted factors indicating significant gene associations.

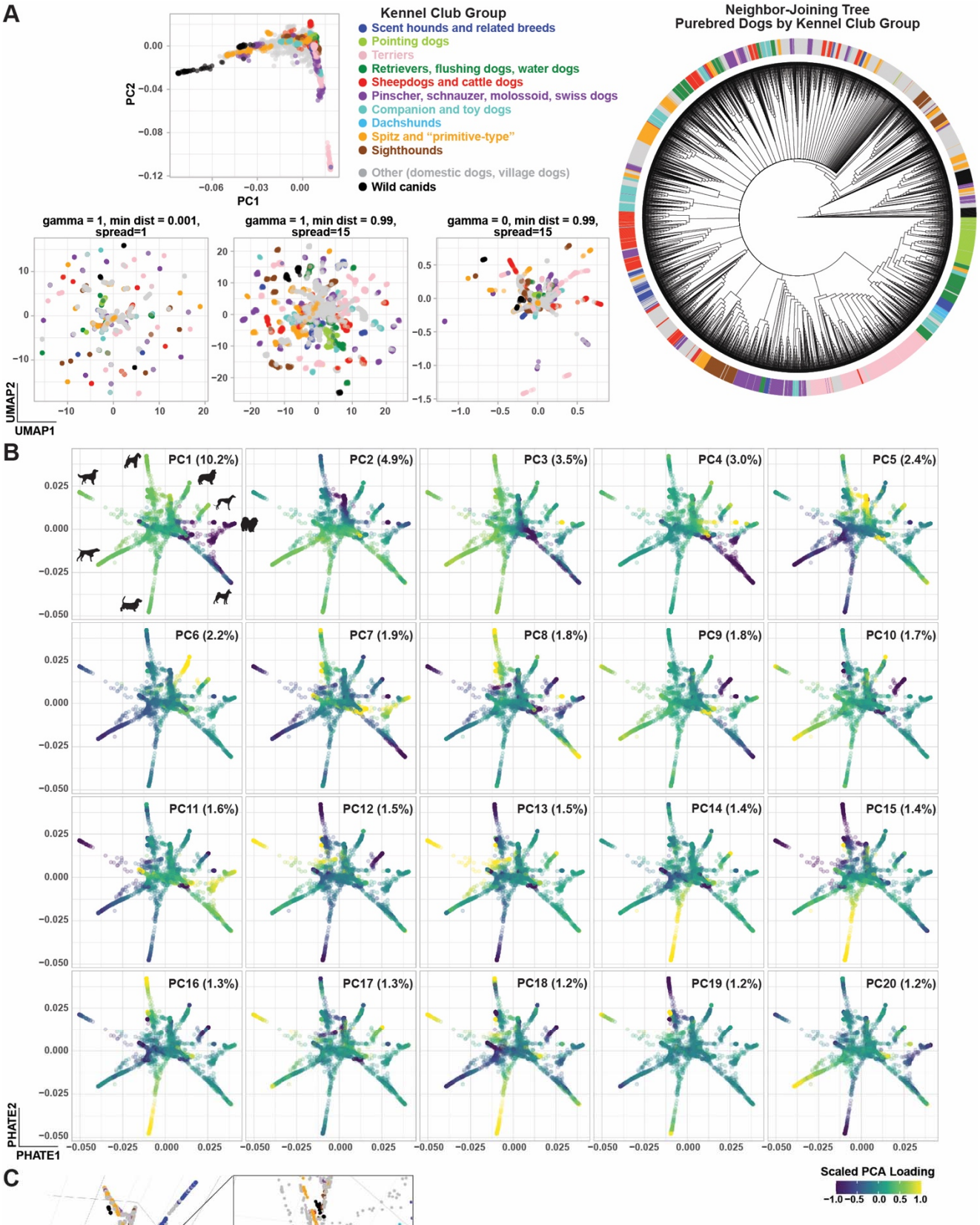

A
Kennel Club Group
Scent hounds and related breeds
Pointing dogs
Terriers
Retrievers, flushing dogs, water dogs
Sheepdogs and cattle dogs
Pinscher, schnauzer, molossoid, swiss dogs
Companion and toy dogs
Dachshunds
Spitz and "primitive-type"
Sighthounds
Other (domestic dogs, village dogs)
Wild canids
PC1
PC2
Neighbor-Joining Tree
Purebred Dogs by Kennel Club Group
gamma = 1, min dist = 0.001, spread=1
gamma = 1, min dist = 0.99, spread=15
gamma = 0, min dist = 0.99, spread=15
UMAP1
UMAP2
B
PC1 (10.2%)
PC2 (4.9%)
PC3 (3.5%)
PC4 (3.0%)
PC5 (2.4%)
PC6 (2.2%)
PC7 (1.9%)
PC8 (1.8%)
PC9 (1.8%)
PC10 (1.7%)
PC11 (1.6%)
PC12 (1.5%)
PC13 (1.5%)
PC14 (1.4%)
PC15 (1.4%)
PC16 (1.3%)
PC17 (1.3%)
PC18 (1.2%)
PC19 (1.2%)
PC20 (1.2%)
PHATE1
PHATE2
Scaled PCA Loading
C
PHATE1
PHATE2
PHATE3
Companion and Toy
Pinscher, Schnauzer, Molossoid, Swiss Dogs

Figure S1. Characterization of dimensionality-reduction methodologies and classification of breeds, related to Figure 1. (A) PCA plot (top left), UMAP plots (bottom left), and neighbor-joining tree (right) colored by kennel club group (Supplemental text). Each UMAP plot was generated with varying gamma, minimum distance, and spread values as indicated on each plot. **(B)** PHATE plots colored by PCA eigenvector. PC values were scaled and centered for visualization. Variance explained by each PC is shown at the top of each plot. **(C)** Three-dimensional PHATE plot (left) colored as in A with zoomed view of the central portion of the plot (right).

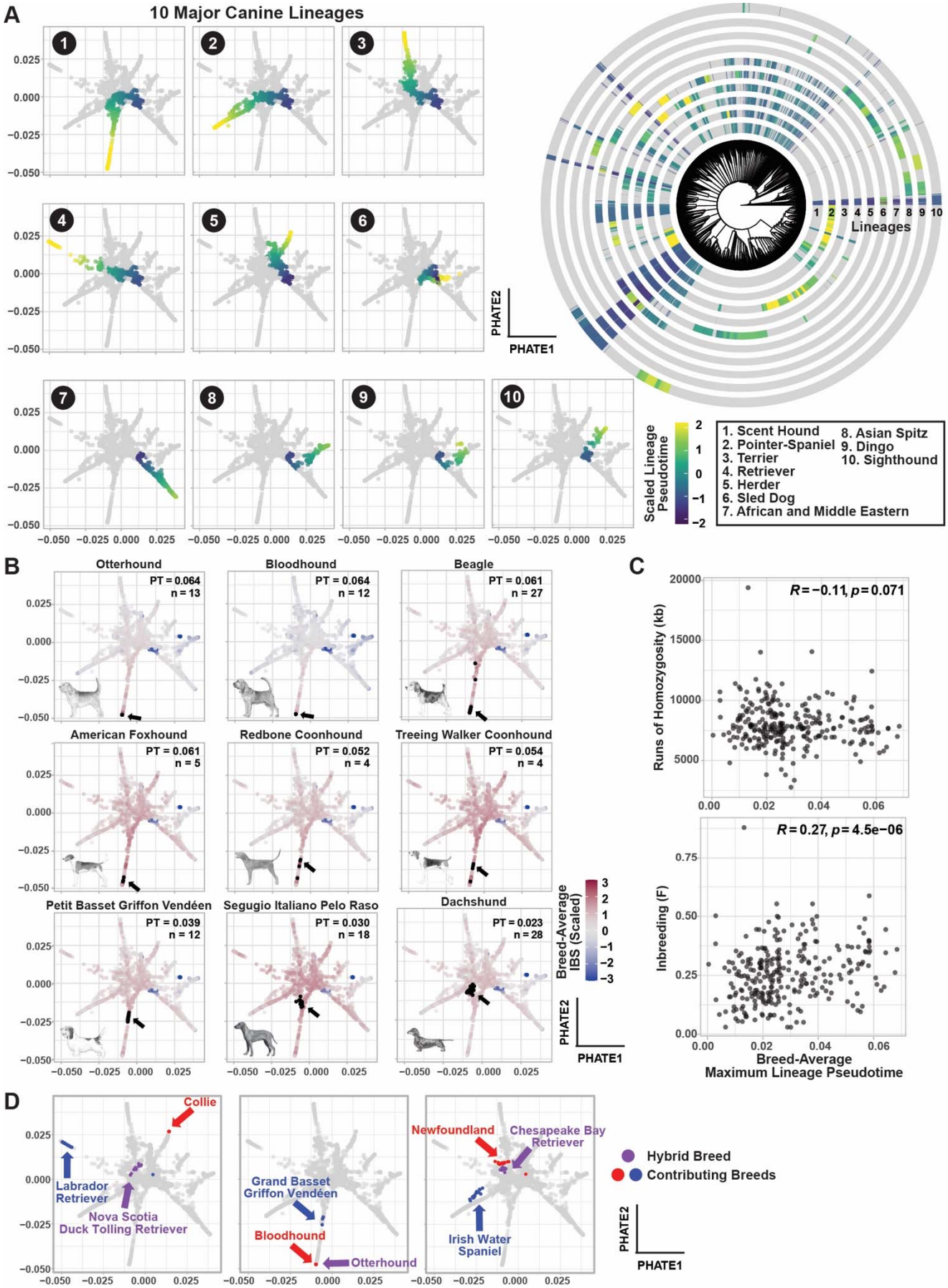

A
10 Major Canine Lineages
1
2
3
4
5
6
7
8
9
10
PHATE2
PHATE1
Lineages
Scaled Lineage Pseudotime
1. Scent Hound
2. Pointer-Spaniel
3. Terrier
4. Retriever
5. Herder
6. Sled Dog
7. African and Middle Eastern
8. Asian Spitz
9. Dingo
10. Sighthound
B
Otterhound
PT = 0.064
n = 13
Bloodhound
PT = 0.064
n = 12
Beagle
PT = 0.061
n = 27
American Foxhound
PT = 0.061
n = 5
Redbone Coonhound
PT = 0.052
n = 4
Treeing Walker Coonhound
PT = 0.054
n = 4
Petit Basset Griffon Vendéen
PT = 0.039
n = 12
Segugio Italiano Pelo Raso
PT = 0.030
n = 18
Dachshund
PT = 0.023
n = 28
Breed-Average IBS (Scaled)
C
R = −0.11, p = 0.071
Runs of Homozygosity (kb)
R = 0.27, p = 4.5e−06
Inbreeding (F)
Breed-Average Maximum Lineage Pseudotime
D
Collie
Labrador Retriever
Nova Scotia Duck Tolling Retriever
Grand Basset Griffon Vendéen
Bloodhound
Otterhound
Newfoundland
Chesapeake Bay Retriever
Irish Water Spaniel
Hybrid Breed
Contributing Breeds

Figure S2. Canine lineage identification and characterization, related to Figure 2. (A) Pseudotime values for each lineage plotted individually in PHATE plots (left) and in concentric circular heatmaps (right) which annotate tips of the neighbor-joining tree (see also annotation of tree in Figure S1A). Pseudotime values for all plots are scaled and centered (z-score) for visualization. **(B)** PHATE plots showing mean pairwise IBS values between each individual in the plot and individuals belonging to each of nine scent hound breeds. Scaling and centering (z-score) of mean pairwise IBS values was performed on the entire dataset simultaneously. All individuals belonging to the comparison breed are shown in black and their locations are highlighted with arrows corresponding to the breed centroid. Number of individuals in breed and breed-average scent hound lineage pseudotime (PT) are also shown for each breed. **(C)** Breed-average runs of homozygosity length in kb (top) and method-of-moments F coefficient estimate (bottom) versus breed-average maximum lineage pseudotime values with associated spearman correlation rho and p values. **(D)** PHATE plots indicating individuals belonging to selected breeds created by hybridization of existing breeds (purple), with contributing breeds (red and blue) indicated on each plot. Contributing breeds are approximated based on anecdotal breed history and/or prior genetic analyses (Parker *et al.*, 2017; Wilcox and Walkowicz, 1989).

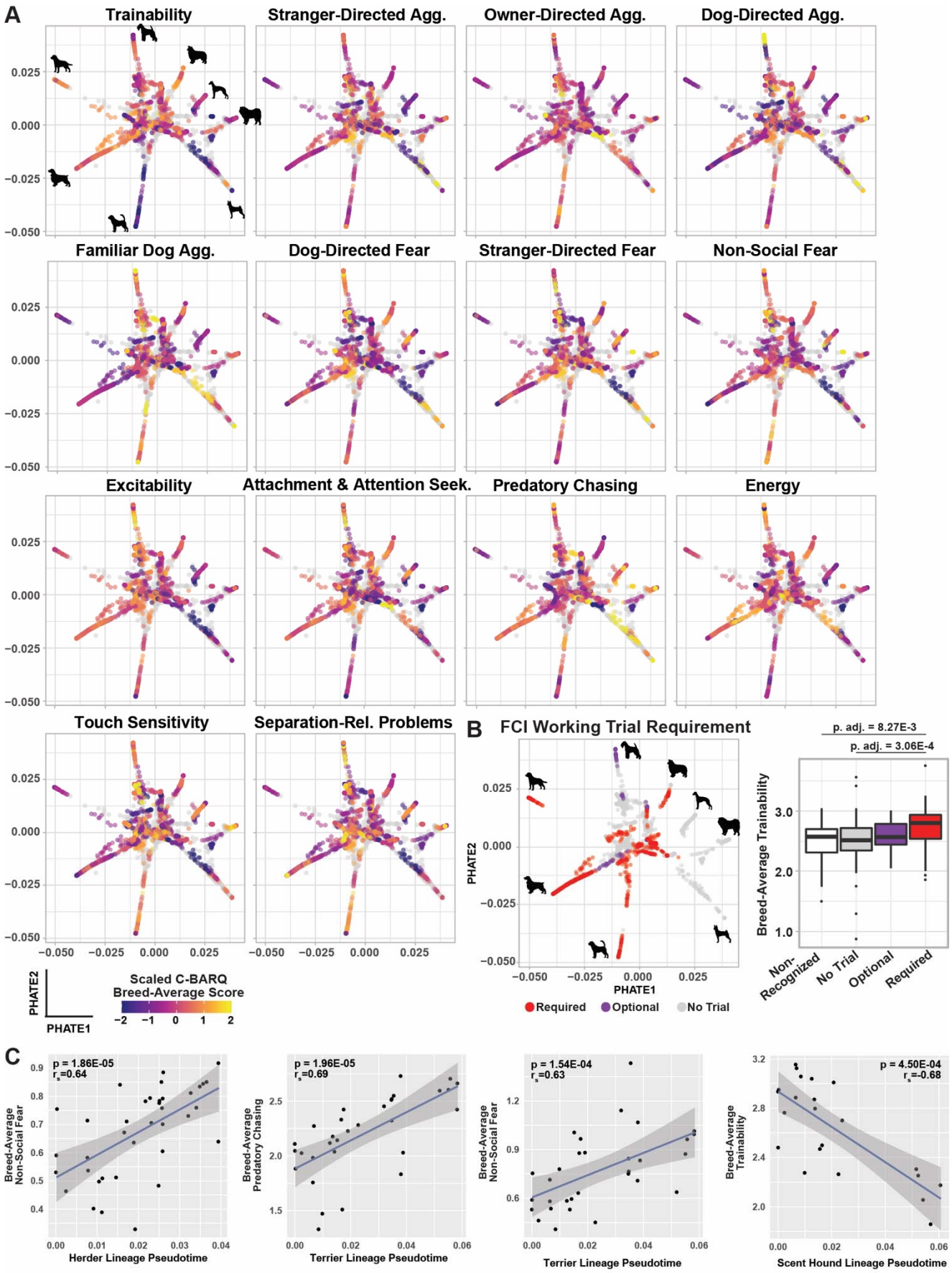

A
Trainability
Stranger-Directed Agg.
Owner-Directed Agg.
Dog-Directed Agg.
Familiar Dog Agg.
Dog-Directed Fear
Stranger-Directed Fear
Non-Social Fear
Excitability
Attachment & Attention Seek.
Predatory Chasing
Energy
Touch Sensitivity
Separation-Rel. Problems
PHATE2
PHATE1
Scaled C-BARQ Breed-Average Score
B
FCI Working Trial Requirement
Required
Optional
No Trial
p. adj. = 8.27E-3
p. adj. = 3.06E-4
Breed-Average Trainability
Non-Recognized
No Trial
Optional
Required
C
p = 1.86E-05
rs=0.64
Breed-Average Non-Social Fear
Herder Lineage Pseudotime
p = 1.96E-05
rs=0.69
Breed-Average Predatory Chasing
Terrier Lineage Pseudotime
p = 1.54E-04
rs=0.63
Breed-Average Non-Social Fear
Terrier Lineage Pseudotime
p = 4.50E-04
rs=-0.68
Breed-Average Trainability
Scent Hound Lineage Pseudotime

Figure S3. Characterization of C-BARQ scores mapped to genetic dataset, related to Figure 3. (A) PHATE plots showing purebred dogs colored by their respective breed-average C-BARQ score. All scores were scaled and centered for visualization. Non-purebred dogs or purebred dogs not represented in C-BARQ are colored in gray. **(B)** Left: FCI-recognized purebred dogs in PHATE plot colored by FCI working trial requirement (required in red, optional in purple, no trial in gray). Right: Boxplot showing breed-average trainability scores for non-FCI-recognized breeds, FCI-recognized breeds with no working trial requirement, an optional trial requirement, and a required trial. Significance calculations were obtained using a one-way ANOVA and Tukey test for all comparisons (See Methods). **(D)** Breed-average C-BARQ score versus breed-average lineage pseudotime for the top significantly correlated factor for four lineages. Significance and correlation values correspond to Spearman correlations as shown in Figure 3C. Linear regression line with 95% confidence interval is shown.

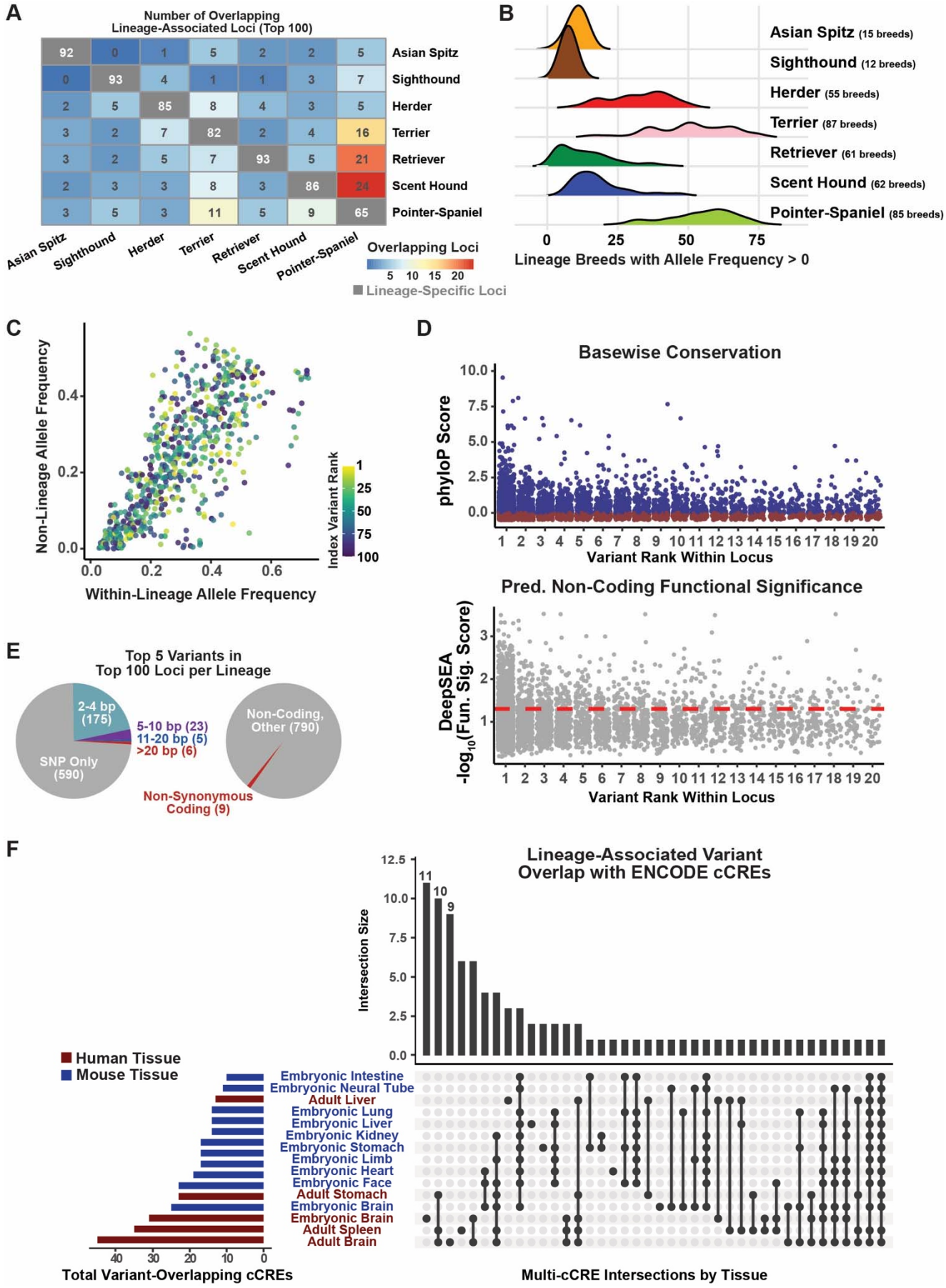

A
Number of Overlapping Lineage-Associated Loci (Top 100)
Asian Spitz
Sighthound
Herder
Terrier
Retriever
Scent Hound
Pointer-Spaniel
Overlapping Loci
5 10 15 20
Lineage-Specific Loci
B
Asian Spitz (15 breeds)
Sighthound (12 breeds)
Herder (55 breeds)
Terrier (87 breeds)
Retriever (61 breeds)
Scent Hound (62 breeds)
Pointer-Spaniel (85 breeds)
Lineage Breeds with Allele Frequency > 0
C
Non-Lineage Allele Frequency
Within-Lineage Allele Frequency
Index Variant Rank
D
Basewise Conservation
phyloP Score
Variant Rank Within Locus
Pred. Non-Coding Functional Significance
DeepSEA
-log10(Fun. Sig. Score)
Variant Rank Within Locus
E
Top 5 Variants in Top 100 Loci per Lineage
2-4 bp (175)
5-10 bp (23)
11-20 bp (5)
>20 bp (6)
SNP Only (590)
Non-Coding, Other (790)
Non-Synonymous Coding (9)
F
Lineage-Associated Variant Overlap with ENCODE cCREs
Intersection Size
Human Tissue
Mouse Tissue
Embryonic Intestine
Embryonic Neural Tube
Adult Liver
Embryonic Lung
Embryonic Liver
Embryonic Kidney
Embryonic Stomach
Embryonic Limb
Embryonic Heart
Embryonic Face
Adult Stomach
Embryonic Brain
Embryonic Brain
Adult Spleen
Adult Brain
Total Variant-Overlapping cCREs
Multi-cCRE Intersections by Tissue

Figure S4. Characterization of lineage-associated variants, related to Figure 4. (A) Heatmap showing overlap between the top 100 loci for each lineage. Number of overlaps was calculated using pairwise comparisons between each lineage to identify overlap between any portion of the locus. Numbers indicate the number of unique loci for each lineage designated in each column that overlap with lineage designated in the row name. Numbers in gray squares on the diagonal show the number of loci in the corresponding lineage that have no overlap with any other lineage. **(B)** Ridge plot showing the number of breeds (within the corresponding lineage only) with non-zero minor allele frequencies for the top 100 most significantly associated index variants per lineage. **(C)** Non-lineage versus within-lineage allele frequencies for index variants for each of the top 100 loci per lineage. Color indicates locus rank (see Methods). **(D)** Top: Basewise phyloP conservation scores for the top 20 variants per top 100 loci for each of seven lineages (as in panels A-C) and combined lineages by significance-based variant rank within locus. Positive scores are plotted in blue and indicate conservation. Bottom: Predicted functional significance of ranked variants. Plotted values are log-transformed predicted functional significance scores (see Methods). Horizontal dashed line indicates the significance cutoff (0.05). **(E)** Left: The top 100 loci per lineage or combined lineages annotated based on the maximum insertion or deletion size within the top five most significantly associated variants per locus. Right: The number of top 100 loci per lineage or combined lineages with one of the top 5 most significantly associated variants per locus annotated as non-synonymous coding based on SnpEff (see Methods). **(F)** Upset plot showing the number of lineage-associated loci (one or more of top 5 variants within the top 100 loci for any lineage) intersecting one or more human or mouse ENCODE candidate cis-regulatory elements (cCRE) by tissue after lifting the variant position to the respective genome (see Methods).

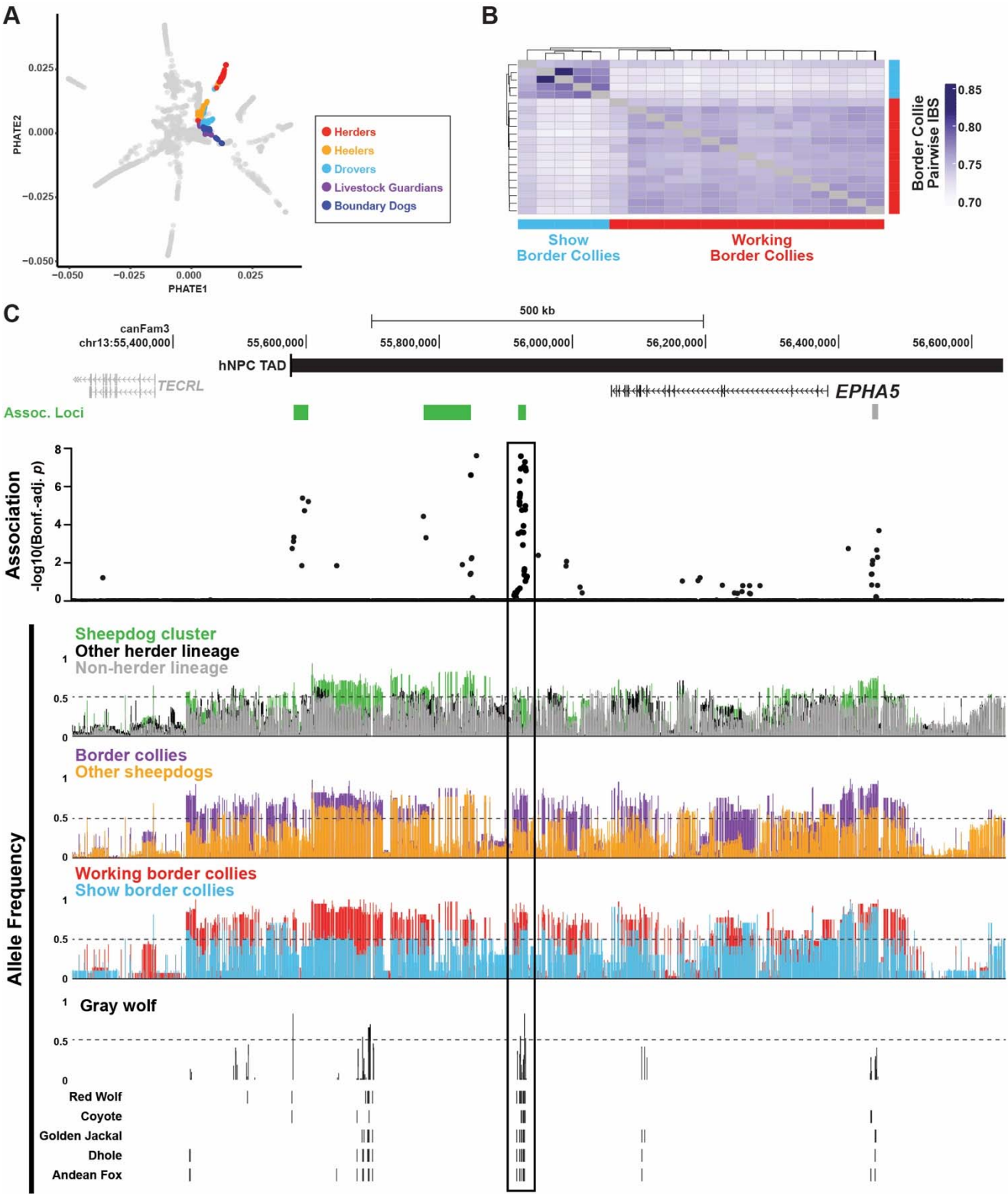

A
PHATE2
PHATE1
Herders
Heelers
Drovers
Livestock Guardians
Boundary Dogs
B
Border Collie Pairwise IBS
Show Border Collies
Working Border Collies
C
canFam3
chr13:55,400,000
500 kb
hNPC TAD
TECRL
EPHA5
Assoc. Loci
Association
-log10(Bonf.-adj. p)
Allele Frequency
Sheepdog cluster
Other herder lineage
Non-herder lineage
Border collies
Other sheepdogs
Working border collies
Show border collies
Gray wolf
Red Wolf
Coyote
Golden Jackal
Dhole
Andean Fox

Figure S5. Characterization of the *EPHA5* sheepdog-associated locus, related to Figure 6. **(A)** PHATE plot showing all purebred dogs with selected breeds within the herder lineage highlighted by anecdotal working group (See Table S1). **(B)** Heatmap showing pairwise IBS values for border collies in the WGS dataset including owner report-based annotations of individuals belonging to working versus show lines**.** **(C)** UCSC genome browser image of a 1.4MB locus (CanFam3 chr13:55,248,790-56,647,365) that includes sheepdog cluster-associated loci (green bars), an overlapping human neural progenitor cell (hNPC) topologically associating domain (TAD; see Methods), and *EPHA5* (CanFam3 protein-coding Ensembl v104). Bonferroni-adjusted p values for all variants are shown below bars indicating associations. Allele frequency tracks are overlayed for multiple comparisons: sheepdog cluster individuals (green; n=63) versus other non-cluster individuals in the herder lineage (black; n=321) and non-herder lineage (gray; n=787), border collies (purple; n=20) versus other individuals in the sheepdog cluster (orange; n=43), working border collies (red; n=15) versus show border collies (blue; n=5), and gray wolf (black; n=27). For the gray wolf, only allele frequencies for variants present at > 75% frequency in working border collies and < 25% in the entire dataset are shown for clarity. Frequency tracks for the red wolf, coyote, golden jackal, dhole, and Andean fox (black) indicate positions at which the individual is homozygous for the aforementioned subset of variants.

Supplemental Tables

Table S1. Genetic dataset metadata and accession information, related to Figure 1 and Methods

Table S2. PHATE embedding coordinates and lineage pseudotime values, related to Figure 1 and Figure 2

Table S3. C-BARQ correlation data, related to Figure 3

Table S4. Lineage association GWAS and locus annotation, related to Figures 4 and 5

Table S5. Gene set enrichment results, related to Figures 5 and 6

Table S6. Sheepdog cluster GWAS and locus annotation, related to Figure 6

Acknowledgements

We thank Ed Giniger, Heidi Parker, Alex Harris, Dayna Dreger, and the creators of PHATE for their insightful comments regarding this study. E.V.D and E.A.O. were funded by the Intramural Program of the National Human Genome Research Institute. This work was supported by funding from the Intramural Program of the National Human Genome Research Institute (E.V.D., E.A.O.).

Supplemental Text

Description of FCI Breed Groups

The Fédération Cynologique Internationale (FCI) is the largest international federation of kennel clubs, recognizing registered breeds of its member kennel clubs (fci.be/nomenclature/). The FCI categorizes dog breeds into 10 breed groups. FCI group 1 includes all sheepdogs and cattledogs other than those classified as Swiss mountain dogs. These breeds assist in moving livestock, for example, by herding. FCI group 2 is named "pinscher and schnauzer, molossoid, and Swiss mountain and cattledogs". This group includes German ratters and guard dogs such as pinschers and schnauzers as well as mastiff-type dogs, which are generally large and may also be used as guard dogs. FCI group 3 encompasses terriers, of which many were developed for vermin control. While the ratters tend to be small, there are also subsets of large and medium sized breeds in this group, as well as the larger bull type terriers. FCI group 4 is comprised of only one breed, the dachshunds, which come in several size and coat varieties. While dachshunds exhibit scent hound-like qualities and are often categorized as such, they were bred to be burrowing dogs with a terrier-like hunting style. FCI group 5 is termed the "spitz and primitive types" group and is a relatively heterogeneous group consisting of European, Asian, and Nordic spitzes (the latter group consisting of diverse working breeds including sled dogs, hunting dogs, and guardians). Despite including working breeds, one of the defining characteristics of the spitz is a wolf-like appearance. Because Asian spitzes generally have increased wolf ancestry versus other breeds, the spitzes are grouped with "primitive-type" breeds, i.e. hunting breeds largely originating in Africa and the Middle East. FCI group 6 are the "scent hounds and related breeds." These sporting breeds are generally used by hunters to track game. Scent hound breeds represent a wide range of body sizes and coat types, as these traits are often suited to specific working environments or aesthetic tastes. FCI group 7 is the "pointing dogs" which, in addition to pointers, includes griffons, setters, and land spaniel types. FCI group 8, "retrievers, flushing

dogs, water dogs," comprise a group that includes not only retrievers and other water dogs, bred to flush and retrieve game from the water, but also several spaniel breeds bred specifically for flushing, as opposed to pointing or setting. FCI Group 9 is the "companion and toy dogs" group. While these breeds are generally small, they are relatively heterogeneous in terms of origin and aesthetic traits. Group 10, the sighthounds, are morphologically distinctive and were generally bred for speed.

References

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., et al. (2016). The Ensembl gene annotation system. Database *2016*, baw093. 10.1093/database/baw093.

Alcaraz, F., Fresno, V., Marchand, A.R., Kremer, E.J., Coutureau, E., and Wolff, M. (2018). Thalamocortical and corticothalamic pathways differentially contribute to goal-directed behaviors in the rat. Elife *7*, e32517. 10.7554/eLife.32517.

Amir, E.-a.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol *31*, 545-552. 10.1038/nbt.2594.

Apps, M.A.J., Rushworth, M.F.S., and Chang, S.W.C. (2016). The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. Neuron *90*, 692-707. 10.1016/j.neuron.2016.04.018.

Banerjee-Basu, S., and Packer, A. (2010). SFARI Gene: an evolving database for the autism research community. Dis Model Mech *3*, 133-135. 10.1242/dmm.005439.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) *57*, 289-300.

Boyko, A.R., Quignon, P., Li, L., Schoenebeck, J.J., Degenhardt, J.D., Lohmueller, K.E., Zhao, K., Brisbin, A., Parker, H.G., vonHoldt, B.M., et al. (2010). A simple genetic architecture underlies morphological variation in dogs. PLoS Biol *8*, e1000451-e1000451. 10.1371/journal.pbio.1000451.

Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M., and Schizophrenia Working Group of the Psychiatric Genomics, C. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics *47*, 291-295. 10.1038/ng.3211.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92. 10.4161/fly.19695.

Consortium, C.-D.G.o.t.P.G. (2019). Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. Cell *179*, 1469-1482.e1411. 10.1016/j.cell.2019.11.020.

Consortium, G., Aguet, F., Anand, S., Ardlie Kristin, G., Gabriel, S., Getz Gad, A., Graubert, A., Hadley, K., Handsaker Robert, E., Huang Katherine, H., et al. (2020). The GTEx Consortium

atlas of genetic regulatory effects across human tissues. Science *369*, 1318-1330. 10.1126/science.aaz1776.

Cooper, M.A., Kobayashi, K., and Zhou, R. (2009). Ephrin-A5 regulates the formation of the ascending midbrain dopaminergic pathways. Dev Neurobiol *69*, 36-46. 10.1002/dneu.20685.

Coppinger, L., and Coppinger, R. (2014). Dogs for herding and guarding livestock. Livestock Handling and Transport: Fourth Edition, 245-260.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. 10.1093/gigascience/giab008.

Deschamps, C., Faideau, M., Jaber, M., Gaillard, A., and Prestoz, L. (2009). Expression of ephrinA5 during development and potential involvement in the guidance of the mesostriatal pathway. Experimental Neurology *219*, 466-480. https://doi.org/10.1016/j.expneurol.2009.06.020.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331-336. 10.1038/nature14222.

Dodman, N.H., Ginns, E.I., Shuster, L., Moon-Fanelli, A.A., Galdzicka, M., Zheng, J., Ruhe, A.L., and Neff, M.W. (2016). Genomic Risk for Severe Canine Compulsive Disorder, a Dog Model of Human OCD. International Journal of Applied Research in Veterinary Medicine *14*.

Ducuing, H., Gardette, T., Pignata, A., Tauszig-Delamasure, S., and Castellani, V. (2019). Commissural axon navigation in the spinal cord: A repertoire of repulsive forces is in command. Seminars in Cell & Developmental Biology *85*, 3-12. https://doi.org/10.1016/j.semcdb.2017.12.010.

Duffy, D.L., Hsu, Y., and Serpell, J.A. (2008). Breed differences in canine aggression. Applied Animal Behaviour Science *114*, 441-460. https://doi.org/10.1016/j.applanim.2008.04.006.

Duffy, D.L., and Serpell, J.A. (2012). Predictive validity of a method for evaluating temperament in young guide and service dogs. Applied Animal Behaviour Science *138*, 99-109. https://doi.org/10.1016/j.applanim.2012.02.011.

Duncan, J., and Owen, A.M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends Neurosci *23*, 475-483. 10.1016/s0166-2236(00)01633-7.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package), version 3.6 (Joseph Felsenstein.).

Fleck, M.S., Daselaar, S.M., Dobbins, I.G., and Cabeza, R. (2006). Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. Cereb Cortex *16*, 1623-1630. 10.1093/cercor/bhj097.

Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods *13*, 845-848. 10.1038/nmeth.3971.

Halo, J.V., Pendleton, A.L., Shen, F., Doucet, A.J., Derrien, T., Hitte, C., Kirby, L.E., Myers, B., Sliwerska, E., Emery, S., et al. (2021). Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. Proceedings of the National Academy of Sciences *118*, e2016274118. 10.1073/pnas.2016274118.

Hayward, J.J., Castelhano, M.G., Oliveira, K.C., Corey, E., Balkman, C., Baxter, T.L., Casal, M.L., Center, S.A., Fang, M., Garrison, S.J., et al. (2016). Complex disease and phenotype mapping in the domestic dog. Nature Communications *7*, 10460. 10.1038/ncomms10460.

Hecht, E.E., Smaers, J.B., Dunn, W.D., Kent, M., Preuss, T.M., and Gutman, D.A. (2019). Significant Neuroanatomical Variation Among Domestic Dog Breeds. The Journal of Neuroscience *39*, 7748. 10.1523/JNEUROSCI.0303-19.2019.

Hsu, Y., and Serpell, J.A. (2003). Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. J Am Vet Med Assoc *223*, 1293-1300. 10.2460/javma.2003.223.1293.

Jagannathan, V., Drögemüller, C., Leeb, T., and Dog Biomedical Variant Database, C. (2019). A comprehensive biomedical variant catalogue based on whole genome sequences of 582 dogs and eight wolves. Animal Genetics *50*, 695-704. https://doi.org/10.1111/age.12834.

Jagannathan, V.A.-O., Hitte, C.A.-O.X., Kidd, J.A.-O., Masterson, P., Murphy, T.A.-O., Emery, S., Davis, B.A.-O.X., Buckley, R.A.-O., Liu, Y.H., Zhang, X.Q., et al. (2021). Dog10K_Boxer_Tasha_1.0: A Long-Read Assembly of the Dog Reference Genome. LID - 10.3390/genes12060847 [doi] LID - 847.

Jhang, J., Lee, H., Kang, M.S., Lee, H.-S., Park, H., and Han, J.-H. (2018). Anterior cingulate cortex and its input to the basolateral amygdala control innate fear response. Nature communications *9*, 2744-2744. 10.1038/s41467-018-05090-y.

Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M.M., Pletikos, M., Meyer, K.A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. Nature *478*, 483-489. 10.1038/nature10523.

Karlsson Linnér, R., Biroli, P., Kong, E., Meddens, S.F.W., Wedow, R., Fontana, M.A., Lebreton, M., Tino, S.P., Abdellaoui, A., Hammerschlag, A.R., et al. (2019). Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. Nature genetics *51*, 245-257. 10.1038/s41588-018-0309-3.

Kolde, R. (2015). pheatmap: Pretty Heatmaps. R package version 1.0.8.

Kuchroo, M., Huang, J., Wong, P., Grenier, J.-C., Shung, D., Tong, A., Lucas, C., Klein, J., Burkhardt, D.B., Gigante, S., et al. (2022). Multiscale PHATE identifies multimodal signatures of COVID-19. Nat Biotechnol. 10.1038/s41587-021-01186-x.

Lane, R.D., Reiman, E.M., Bradley, M.M., Lang, P.J., Ahern, G.L., Davidson, R.J., and Schwartz, G.E. (1997). Neuroanatomical correlates of pleasant and unpleasant emotion. Neuropsychologia *35*, 1437-1444. 10.1016/s0028-3932(97)00070-5.

Larsson, M. (2015). Binocular vision, the optic chiasm, and their associations with vertebrate motor behavior. Frontiers in Ecology and Evolution *3*.

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database, C. (2011). The sequence read archive. Nucleic acids research *39*, D19-D21. 10.1093/nar/gkq1019.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739-1740. 10.1093/bioinformatics/btr260.

Liu, L., Feng, X., Li, H., Cheng Li, S., Qian, Q., and Wang, Y. (2021). Deep learning model reveals potential risk genes for ADHD, especially Ephrin receptor gene EPHA5. Briefings in Bioinformatics *22*, bbab207. 10.1093/bib/bbab207.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic acids research *45*, D896-D901. 10.1093/nar/gkw1133.

MacLean, E.L., Snyder-Mackler, N., vonHoldt, B.M., and Serpell, J.A. (2019). Highly heritable and functionally relevant breed differences in dog behaviour. Proceedings of the Royal Society B: Biological Sciences *286*, 20190716. doi:10.1098/rspb.2019.0716.

Mamiya, P.C., Hennesy, Z., Zhou, R., and Wagner, G.C. (2008). Changes in attack behavior and activity in EphA5 knockout mice. Brain Res *1205*, 91-99. 10.1016/j.brainres.2008.02.047.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867-2873. 10.1093/bioinformatics/btq559.

Matoso, E., Melo, J.B., Ferreira, S.I., Jardim, A., Castelo, T.M., Weise, A., and Carreira, I.M. (2013). Insertional translocation leading to a 4q13 duplication including the EPHA5 gene in two siblings with attention-deficit hyperactivity disorder. American Journal of Medical Genetics Part A *161*, 1923-1928. https://doi.org/10.1002/ajmg.a.36032.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: uniform manifold approximation and projection, J. Open Source Softw *3*, 861.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biology *17*, 122. 10.1186/s13059-016-0974-4.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-501. 10.1038/nbt.1630.

Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., Elzen, A.v.d., Hirn, M.J., Coifman, R.R., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol *37*, 1482-1492. 10.1038/s41587-019-0336-3.

Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699-710. 10.1038/s41586-020-2493-4.

Noor, A., Lionel, A.C., Cohen-Woods, S., Moghimi, N., Rucker, J., Fennell, A., Thiruvahindrapuram, B., Kaufman, L., Degagne, B., Wei, J., et al. (2014). Copy number variant study of bipolar disorder in Canadian and UK populations implicates synaptic genes. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics *165*, 303-313. https://doi.org/10.1002/ajmg.b.32232.

Parker, H.G., Dreger, D.L., Rimbault, M., Davis, B.W., Mullen, A.B., Carpintero-Ramirez, G., and Ostrander, E.A. (2017). Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. Cell Rep *19*, 697-708. 10.1016/j.celrep.2017.03.079.

Pendleton, A.L., Shen, F., Taravella, A.M., Emery, S., Veeramah, K.R., Boyko, A.R., and Kidd, J.M. (2018). Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. BMC Biology *16*, 64. 10.1186/s12915-018-0535-2.

Plassais, J., Kim, J., Davis, B.W., Karyadi, D.M., Hogan, A.N., Harris, A.C., Decker, B., Parker, H.G., and Ostrander, E.A. (2019). Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nature Communications *10*, 1489. 10.1038/s41467-019-09373-w.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res *20*, 110-121. 10.1101/gr.097857.109.

Prieur, D.S., and Rebsam, A. (2017). Retinal axon guidance at the midline: Chiasmatic misrouting and consequences. Dev Neurobiol *77*, 844-860.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics *81*, 559-575. 10.1086/519795.

Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.A., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J., et al. (2010). The UCSC Genome Browser database: update 2010. Nucleic Acids Res *38*, D613-619. 10.1093/nar/gkp939.

Salonen, M., Sulkama, S., Mikkola, S., Puurunen, J., Hakanen, E., Tiira, K., Araujo, C., and Lohi, H. (2020). Prevalence, comorbidity, and breed differences in canine anxiety in 13,700 Finnish pet dogs. Scientific reports *10*, 1-11.

Serpell, J. (2017). The domestic dog (Cambridge University Press).

Serpell, J.A., and Hsu, Y.A. (2005). Effects of breed, sex, and neuter status on trainability in dogs. Anthrozoös *18*, 196-207. 10.2752/089279305785594135.

Serres-Armero, A., Davis, B.W., Povolotskaya, I.S., Morcillo-Suarez, C., Plassais, J., Juan, D., Ostrander, E.A., and Marques-Bonet, T. (2021). Copy number variation underlies complex phenotypes in domestic dog breeds and other canids. Genome research *31*, 762-774. 10.1101/gr.266049.120.

Sheleg, M., Yu, Q., Go, C., Wagner, G.C., Kusnecov, A.W., and Zhou, R. (2017). Decreased maternal behavior and anxiety in ephrin-A5(-/-) mice. Genes Brain Behav *16*, 271-284. 10.1111/gbb.12319.

Snow, P.J. (2016). The Structural and Functional Organization of Cognition. Front Hum Neurosci *10*, 501-501. 10.3389/fnhum.2016.00501.

Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics *19*, 477. 10.1186/s12864-018-4772-0.

Svartberg, K. (2006). Breed-typical behaviour in dogs—Historical remnants or recent constructs? Applied Animal Behaviour Science *96*, 293-313. https://doi.org/10.1016/j.applanim.2005.06.014.

Tang, R., Noh, H.J., Wang, D., Sigurdsson, S., Swofford, R., Perloski, M., Duxbury, M., Patterson, E.E., Albright, J., Castelhano, M., et al. (2014). Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. Genome Biology *15*, R25. 10.1186/gb-2014-15-3-r25.

Turner, K.M., and Parkes, S.L. (2020). Prefrontal regulation of behavioural control: Evidence from learning theory and translational approaches in rodents. Neuroscience & Biobehavioral Reviews *118*, 27-41. https://doi.org/10.1016/j.neubiorev.2020.07.010.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics *43*, 11.10.11-11.10.33. 10.1002/0471250953.bi1110s43.

van Rooy, D., Arnott, E.R., Early, J.B., McGreevy, P., and Wade, C.M. (2014). Holding back the genes: limitations of research into canine behavioural genetics. Canine Genet Epidemiol *1*, 7-7. 10.1186/2052-6687-1-7.

Vasung, L., Huang, H., Jovanov-Milošević, N., Pletikos, M., Mori, S., and Kostović, I. (2010). Development of axonal pathways in the human fetal fronto-limbic brain: histochemical characterization and diffusion tensor imaging. Journal of anatomy *217*, 400-417. 10.1111/j.1469-7580.2010.01260.x.

vonHoldt, B.M., Shuldiner, E., Koch Ilana, J., Kartzinel Rebecca, Y., Hogan, A., Brubaker, L., Wanser, S., Stahler, D., Wynne Clive, D.L., Ostrander Elaine, A., et al. (2017). Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs. Science Advances *3*, e1700398. 10.1126/sciadv.1700398.

Wang, C., Wallerman, O., Arendt, M.-L., Sundström, E., Karlsson, Å., Nordin, J., Mäkeläinen, S., Pielberg, G.R., Hanson, J., Ohlsson, Å., et al. (2021). A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. Communications Biology *4*, 185. 10.1038/s42003-021-01698-x.

Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nature Communications *8*, 1826. 10.1038/s41467-017-01261-5.

Weber, E.M., and Olsson, I.A.S. (2008). Maternal behaviour in Mus musculus sp.: An ethological review. Applied Animal Behaviour Science *114*, 1-22. https://doi.org/10.1016/j.applanim.2008.06.006.

Weiss, L.A., Arking, D.E., Gene Discovery Project of Johns, H., the Autism, C., Daly, M.J., and Chakravarti, A. (2009). A genome-wide linkage and association scan reveals novel loci for autism. Nature *461*, 802-808. 10.1038/nature08490.

Wilcox, B., and Walkowicz, C. (1989). Atlas of dog breeds of the world. New rev.

Willsey, A.J., Sanders, Stephan J., Li, M., Dong, S., Tebbenkamp, Andrew T., Muhle, Rebecca A., Reilly, Steven K., Lin, L., Fertuzinhos, S., Miller, Jeremy A., et al. (2013). Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. Cell *155*, 997-1007. https://doi.org/10.1016/j.cell.2013.10.020.

Worboys, M., Strange, J.-M., and Pemberton, N. (2018). The invention of the modern dog: breed and blood in Victorian Britain (Johns Hopkins University Press).

Zapata, I., Serpell, J.A., and Alvarez, C.E. (2016). Genetic mapping of canine fear and aggression. BMC Genomics *17*, 572. 10.1186/s12864-016-2936-3.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods *12*, 931-934. 10.1038/nmeth.3547.