Journal of Heredity 2009:100(Supplement 1):S28–S36
doi:10.1093/jhered/esp012
Advance Access publication April 1, 2009

Intrabreed Stratification Related to Divergent Selection Regimes in Purebred Dogs May Affect the Interpretation of Genetic Association Studies

Melanie L. Chang, Jennifer S. Yokoyama, Nick Branson, Donna J. Dyer, Christophe Hitte, Karen L. Overall, and Steven P. Hamilton

From the Department of Psychiatry and Institute for Human Genetics, University of California, San Francisco (Chang, Yokoyama, Hamilton); the Center for Neurobiology and Behavior, Department of Psychiatry, University of Pennsylvania, Philadelphia (Branson, Dyer, Overall); and the Institut de Génétique et Développement, CNRS UMR6061, Université de Rennes, 2 Av du Pr. Léon Bernard, 35043 Rennes, France (Hitte).

Address correspondence to Steven P. Hamilton at the address above, or e-mail: steveh@lppi.ucsf.edu.

Abstract

Until recently, canine genetic research has not focused on population structure within breeds, which may confound the results of case–control studies by introducing spurious correlations between phenotype and genotype that reflect population history. Intrabreed structure may exist when geographical origin or divergent selection regimes influence the choices of potential mates for breeding dogs. We present evidence for intrabreed stratification from a genome-wide marker survey in a sample of unrelated dogs. We genotyped 76 Border Collies, 49 Australian Shepherds, 17 German Shepherd Dogs, and 17 Portuguese Water Dogs for our primary analyses using Affymetrix Canine v2.0 single-nucleotide polymorphism (SNP) arrays. Subsets of autosomal markers were examined using clustering algorithms to facilitate assignment of individuals to populations and estimation of the number of populations represented in the sample. SNPs passing stringent quality control filters were employed for explicitly phylogenetic analyses reconstructing relationships between individuals using maximum parsimony and Bayesian methods. We used simulation studies to explore the possible effects of intrabreed stratification on genome-wide association studies. These analyses demonstrate significant stratification in at least one of our primary breeds of interest, the Border Collie. Demographic and pedigree data suggest that this population substructure may result from geographic isolation or divergent selection regimes practiced by breeders with different breeding program goals. Simulation studies indicate that such stratification could result in false discovery rates significant enough to confound genome-wide association analyses. Intrabreed stratification should be accounted for when designing and interpreting the results of case–control association studies using purebred dogs.

Key words: *Bayesian analysis, canine genetics, maximum parsimony, phylogenetics, population stratification, purebred dogs*

The purebred dog has become a popular and useful model organism for genetic studies of disease and, more recently, behavior (Sutter and Ostrander 2004). Its popularity stems from its familiarity and from the history of breeding practices that have produced and maintain the phenotypic diversity of contemporary dog breeds, combined with relative intrabreed genetic homogeneity. These characteristics have been used for successful linkage-based approaches for gene mapping and, lately, in association-based approaches (Salmon Hillbertz et al. 2007).

Several assumptions are made in most genetic association studies using purebred dogs. First, it is assumed that purebred dogs constitute separate, closed, inbred populations exhibiting intense founder effect. We therefore expect that there will be limited phenotypic and genetic variation within breeds but broad variation between breeds. These

assumptions are probably safe, given registration practices for purebred dogs that dictate a dog cannot be considered purebred unless both of its parents are registered purebreds. However, the common corollary, which is that pure breeds make up homogeneous populations, may be problematic.

Most modern purebred dog breeders select breeding animals based on conformation, judged via success at dog shows. Animals are evaluated at these shows according to written standards that are specific on numerous points of the dog's appearance. The desired qualities of appearance (and to some extent carriage and attitude) are referred to as "type," and winning dogs are described as exhibiting "excellent breed type." Top winning dogs produce a disproportionate share of offspring and are often inbred so as to be "prepotent" (high homozygosity), meaning they will reliably "stamp" their offspring with their successful phenotypes. Variation is discouraged and outcrossing is forbidden, as they decrease consistency of type. These breeding practices can result in remarkable phenotypic homogeneity and further decrease the effective population sizes of breeds that may already descend from only a handful of founders. Such patterns of suppressed gene flow between dog breeds and intense inbreeding within breeds produces apparent population isolates that are in many ways ideal for genetic analyses (Sutter and Ostrander 2004).

However, dog breeds often fragment into "lines" or "types" based on geography, divergent selection regimes, or both. For example, differences in preferred appearance and strong founder effects in each population led to marked divergence between American and Japanese Akitas following World War II. The Federation Cynologique Internationale formalized a split in 1999, designating the American dogs as "American Akita" and the Japanese dogs as "Akita Inu"—separate breeds. However, the American Kennel Club (AKC) still considers both populations to be a single breed and registers both in a single studbook. This is only one example suggesting that the assumption of homogeneity within breeds may be faulty.

The present study addresses the phenomenon of "breed splits" and its possible consequences for genetic association studies. Although it is clear from previous evidence, as well as a point of common knowledge among purebred dog owners, trainers, and handlers, that population structure exists within breeds, this structure has not been systematically characterized. Previous studies of population structure in dogs have focused primarily on the relationships between breeds, incorporating relatively small samples of a large number of breeds, and using clustering methods to compare overall degrees of similarity between samples characterized either by microsatellite markers or small numbers of single-nucleotide polymorphisms (SNPs) localized to a limited sampling of the genome (e.g., Parker et al. 2004). Existing assessments of within-breed population structure are characterized by restricted genomic coverage (Quignon et al. 2007; Björnerfeldt et al. 2008) or were accomplished via pedigree analysis (Calboli et al. 2008).

We sought to determine if stratification may be predicted by knowledge of sample origin, geography, or selection regime. We incorporated autosomal SNP genotype data with broad genomic coverage, taking advantage of sizable, well-characterized samples in 4 breeds of interest. We used phylogenetic methods developed by systematic biologists to examine the evolutionary relationships of biological groups (recency of common ancestry) and to investigate population substructure within breeds. We interpreted our findings in the context of owner-reported demographic and pedigree information, in an effort to understand how we may identify probable stratification within samples for future genetic analyses. Finally, we conducted simulations to explore the effects of such stratification on genome-wide association studies (GWASs) and explored strategies for minimizing the risks of false-positive results in GWAS.

Methods

Sample Recruitment, Collection, and Data Generation

We recruited and collected samples of 4 pure dog breeds for investigation of within-breed stratification and smaller samples of 23 pure dog breeds for a comparative assessment of overall canine diversity. Owners of participating dogs were recruited at dog shows and working competitions (sheepdog trials) and through direct mail, e-mail lists, breed clubs, and training organizations. We drew samples from dogs on site or asked owners to send blood samples to our laboratory using a standardized protocol. We also collected pedigrees, demographic data, and a detailed behavioral questionnaire (Overall et al. 2006) for each dog.

Our sample included 3 herding breeds of interest for a project exploring the genetic background of canine noise phobia, a discrete behavioral phenotype with a probable genetic component: Border Collie (BOC), Australian Shepherd (AUS), and German Shepherd Dog (GSD). We also included the Portuguese Water Dog (PWD), characterized by a breed community that is enthusiastically supportive of canine genetic studies, and 24 additional dogs of 23 diverse breeds, for a comparative assessment of structure within our breeds of interest in the context of overall canine diversity. This latter group includes a number of breeds that have never been included in previous studies of canine population structure, such as the Otterhound (the rarest AKC registered breed) and the Thai Bangkaew Dog (TBD, an indigenous spitz-type breed of Thailand). This gave us a total of 183 dogs (Table 1). Our sample included unrelated dogs selected for GWAS, extended pedigrees segregating noise phobia in 2 breeds (BOC and AUS), and 5 small family groups (BOC and AUS, 1 trio and 4 quartets) included for assessment of Mendelization errors.

We collected whole blood samples of approximately 5 ml from each dog and extracted genomic DNA from each sample using the Gentra Puregene Blood Kit (Qiagen, Valencia, CA). We surveyed approximately 127 000 markers per dog using Affymetrix's Canine v2.0 SNP array and called genotypes using the BRLMM-P algorithm. We dropped X chromosome markers and filtered the remaining markers for

Table 1. Total dogs genotyped, by breed

Breed	Number genotyped
Border Collies	76
Australian Shepherds	49
German Shepherd Dogs	17
Portuguese Water Dogs	17
Thai Bangkaew Dogs	2
Other breeds[a]	22
Total	183

[a] American Staffordshire Terrier, Australian Cattle Dog, Australian Kelpie, Belgian Sheepdog, Briard, Bulldog, Rough Collie, Miniature Dachshund, Dalmatian, Doberman, English Springer Spaniel, Golden Retriever, Keeshond, Labrador Retriever, Newfoundland, Otterhound, Papillon, Pointer, Standard Schnauzer, Soft-Coated Wheaten Terrier, Thai Bangkaew Dog (2), Weimaraner, and Whippet.

call rate, concordance for a single dog between multiple (4) genotyping runs, significant deviations from Hardy–Weinberg equilibrium, Mendelization errors, and minor allele frequencies <0.02. We generated marker subsets constituting 700, 2100, 4200, and 21 000 SNPs that were spaced evenly across the genome for use in phylogenetic analyses. Multiple marker sets were used to address computational limitations associated with some analyses and to test the consistency of different-sized marker sets.

Cluster and Genetic Distance Analyses

Preliminary cluster analyses were conducted because their use in previous studies (Parker et al. 2004; Quignon et al. 2007; Björnerfeldt et al. 2008) would afford us comparable assessment of our results. We conducted a preliminary clustering analysis of 2100 high-quality SNPs (100% call rate, median inter-SNP distance 850 kb) from all 38 autosomes using "structure" (Pritchard et al. 2000; Falush et al. 2003, 2007). This data set included 48 BOC, 27 AUS, 17 PWD, and 16 GSD for a total of 108 unrelated dogs and was subjected to 30 iterations of $K = 1$ through $K = 8$, where the user-assigned value for K is the number of putative population groups predicted to be present in the given sample. We then used methods outlined by Evanno et al. (2005) to determine the "best fit" or number of population groups predicted given our data set.

Genetic distance analyses were then performed for comparison to previous analyses of the same type that used limited-coverage SNP data (Quignon et al. 2007). Average genome-wide proportions of alleles sharing identity by state (IBS) were calculated pairwise for 108 dogs from 4 breeds (48 BOC, 27 AUS, 17 PWD, 16 GSD) across 21 000 uncorrelated SNPs covering all 38 canine autosomes. These were used to create a distance matrix (1-IBS) of 108 × 108 individuals with pLINK v1.02 (Purcell et al. 2007). The distance matrix was visualized in *R*, and the number of optimal clusters "K" was calculated with a hierarchical agglomerative clustering method ("agnes" as implemented in *R*). The optimal value of clusters was calculated to be $K = 8$. Cluster stability was then assessed for $K = 8$, as well as for $K = 4$ (which corresponds to the number of breeds evaluated) via bootstrapping including outliers.

Phylogenetic Analysis

We conducted a series of phylogenetic analyses using SNP data sets of 700, 4200, and 21 000 autosomal markers in analyses that included the 4 breeds of primary interest and some that included 183 dogs of 27 breeds. As mentioned above, marker numbers were varied to compare the performance of varying marker set sizes and to address the computational constraints of analysis software. Unlike conventional measures of stratification, which produce clusters based on overall genetic similarity, phylogenetic methods incorporate models of evolution into the analysis and yield explicit hypotheses of nested relationships (i.e., recency of common ancestry) between taxa (in this case, breed samples and individuals) included in a study (Hennig 1966).

We performed parsimony analyses using PAUP* 4.0bv10 (Swofford 2003) and Nixon's parsimony ratchet, implemented in PAUPRat 1.0 (Sikes and Lewis 2001), to render large data sets tractable for analysis. Trees were constructed using random stepwise addition and TBR branch swapping and evaluated via bootstrapping analyses in PAUP (1000 reps, TBR-M) and Bremer support indices (decay analysis) using TREEROT 3.0 (Sorenson and Franzosa 2007). We performed Bayesian analyses using MrBayes 3.1.2 (Huelsenbeck et al. 2001; Ronquist and Huelsenbeck 2003) with runs of 500 000–2 000 000 generations and the following settings: temperature of heated chains = 0.05−0.20; Nst = 6; rates = invgamma. Bayesian trees were evaluated via posterior probabilities generated during the analysis.

The resulting trees were combined in a supertree analysis via Matrix Representation with Parsimony (MRP) that summarized the clades that were found across all analyses, implemented in Rainbow 1.2 (Chen et al. 2004) and PAUP 4.0. Trees were visualized using FigTree 1.1.2 (Rambaut 2008).

Pedigree Analysis

As part of data collection for our larger study examining the genetic basis of canine noise phobia, we collected pedigrees and questionnaire information from owners (Overall et al. 2006). We used this information to characterize the dogs in our sample and interpret the population structure we discovered through phylogenetic analysis. Data that proved particularly informative for this purpose were pedigree information about show ring performance of ancestors (such as show championships) and geographical origin of dogs, type of event at which the sample was collected (i.e., working trial vs. dog show), organization with which the dog was registered, and owner-reported information about titles achieved or activities regularly engaged in with their dogs, as well as the type of breeder from which the dog was obtained (show dog breeder or working sheepdog breeder). Using this information allowed us to describe the individuals of one breed in particular, BOC, as "show dogs" or "working dogs." These types were correlated with different

geographical origins, with show dogs tracing back to Australia/New Zealand and working dogs tracing back to the United Kingdom.

Simulation Studies

We conducted a simulated case–control association study (100 iterations) using observed genotypes and information about population substructure in BOC revealed by the cluster, distance, phylogenetic, and pedigree analyses. Our sample of unrelated BOC split into 2 groups across all analyses: a larger clade of 43 dogs and a small clade of 5 that were consistently differentiated (see Results). We randomly assigned case–control status to the large group of 43 BOC. We then assigned case–control status to the smaller group of 5 BOC as follows: "split sample" (2 randomly assigned as cases and 3 randomly assigned as controls), "all cases" (with the balance of the 43 randomly assigned case or control status), or "all controls" (with the balance of the 43 randomly assigned case or control status). We performed a genome-wide allelic association analysis on approximately 53 000 SNPs using all 48 unrelated BOCs, using the adjusted P value calculation to obtain the average chi-squared value and genomic inflation factor based on median chi-squared (pLINK v1.04), and evaluated our simulated results for significantly inflated false-positive association rates. Principal components were calculated using Eigenstrat (Price et al. 2006), and logistic regression with covariates was implemented in pLINK v1.04.

Results

To summarize our results, individual dogs were correctly assigned to their respective breeds using all methods. Related dogs that were included in some analyses consistently grouped together, supporting the credibility of the results of the analyses. A group of unrelated BOC formed a separate, well-supported clade across analyses. These 5 dogs are distinguished by the type of purpose for which they were bred (show vs. working) and by geographical origin either of themselves or close ancestors. In simulated GWAS, this stratification led to significantly inflated false-positive association rates.

Results of Cluster and Genetic Distance Analyses

Cluster analysis of unrelated individuals in 4 breeds (BOC, AUS, PWD, and GSD) identified 4 clusters corresponding to breed in the data (Supplemental Figure 1) and correctly assigned all dogs to the 4 reported breeds. These results suggest some degree of heterogeneity within BOC when $K = 4$. At user-assigned values of $K > 4$, 5 BOC become distinct from the rest of their breed across runs (Supplemental Figure 2). We also identified clustering that suggests some proportion of AUS ancestry is shared with the BOC, a result that would be predicted given the history of these breeds.

The hierarchical grouping via genetic distance analysis for $K = 4$ demonstrated perfect stability, with all dogs falling into their respective breed clusters (Figure 1). Hierarchical grouping for $K = 8$, calculated to be the "best fit" for this data set, demonstrated correct separation of dogs into 4 breeds and intrabreed stratification of AUS and BOC (Supplemental Figure 3). The AUS were broken into 4 separate clusters, though 2 of those were made up of singletons. The same 5 BOC that were differentiated in the cluster analyses grouped together and were distinct from the rest of the breed sample, forming a separate branch from the rest of the breed in the $K = 8$ dendrogram. A suggestive clustering of AUS was also detectable in the $K = 8$ dendrogram, though the clusters were composed of too few individuals to withstand rigorous stability testing.

Results of Phylogenetic Analyses

A series of parsimony and Bayesian analyses including the 4 breeds of primary interest agreed with the results of our cluster and genetic distance analyses, discovering 5 major clades, with 4 corresponding to the 4 breeds (BOC, AUS, PWD, and GSD) and a fifth clade made up of the same 5 distinctive BOC seen in cluster and distance analyses (Figure 2). In the results of Bayesian analyses, the clade containing these 5 BOC is located almost midway between the node leading to all the rest of the BOC and the node leading to the other 3 breeds. With branch length being proportional to distance, this suggests prominent divergence within the breed. The ability of Bayesian analysis to recover relationships at such low taxonomic levels, using these types of data, is confirmed by the behavior of the family trio and quartets included to check for Mendelization errors, which grouped together across all the analyses in which they were included (Supplemental Figure 4). The quality of the Bayesian phylogenetic hypotheses are indicated by posterior probabilities greater than 0.90 found throughout the trees supporting all major branches, including the branch leading to the 5 distinct BOCs separately.

We sought to determine the lower resolution performance of this approach when using single dogs from many breeds. Bayesian analyses including all unrelated dogs of 27 breeds were less well resolved. The relationships between breeds themselves were not well constructed in these analyses, but the major clades found in 4-breed analyses were still supported. Four out of five distinct BOCs still formed their own clade, albeit with a lower supporting posterior probability of 0.53 (Supplemental Figure 5). Parsimony analyses of the same SNP set were similarly less resolved, with only 4 clades showing bootstrap values greater than 50%, but the relationships between breeds agree with the results from Bayesian analyses (Supplemental Figure 6).

We constructed an MRP supertree to summarize the results of all Bayesian and parsimony analyses conducted that included either the 4 breeds of interest or all 27 breeds, using sets of 700, 4200, and 21 000 autosomal SNPs distributed evenly across the genome. The clades presented in this tree represent those clades found by every analysis in which they were included. Bootstrap values for all branches of this tree were more than 97%. Our supertree demonstrates separation of the 4 breeds of primary interest,

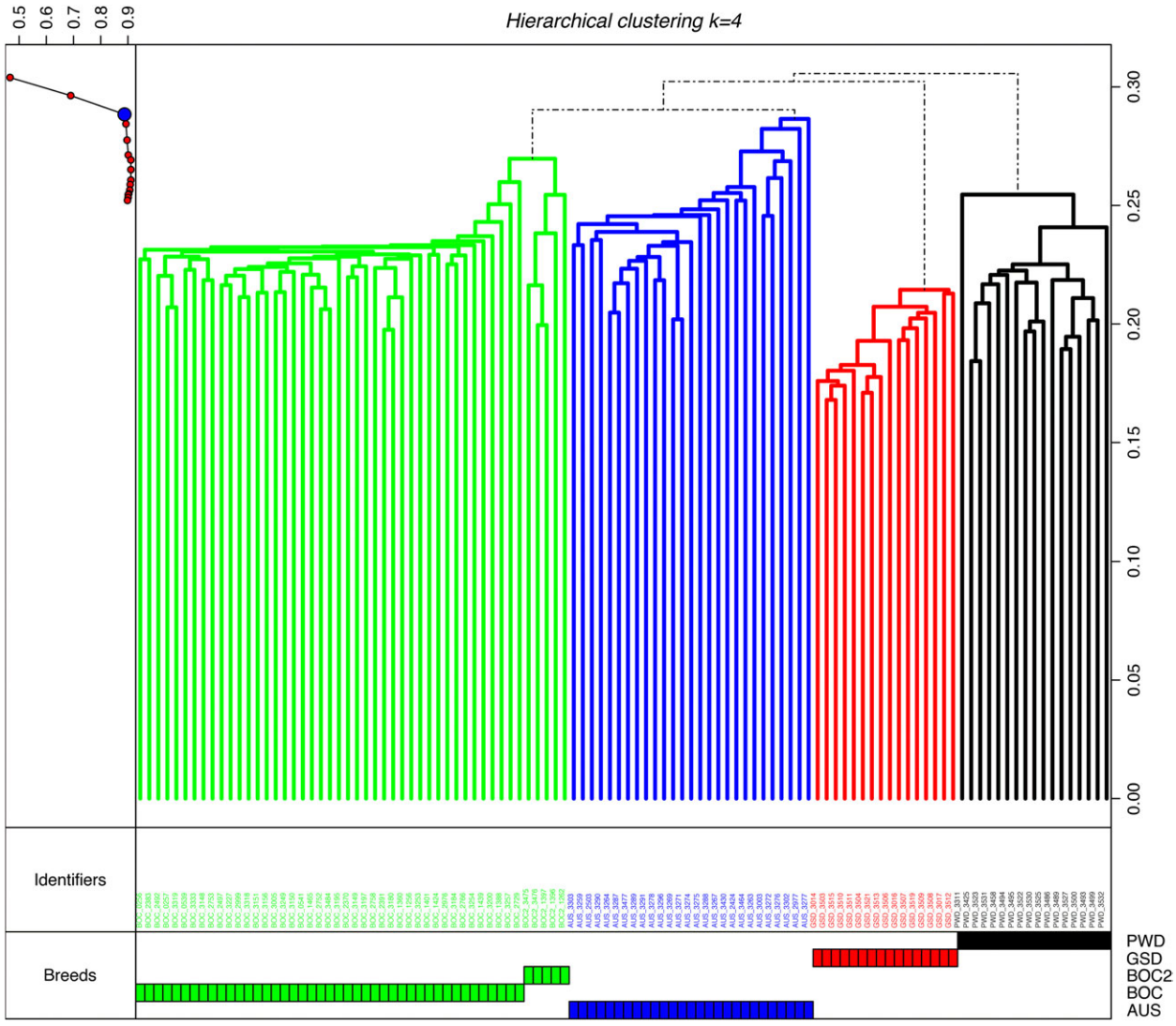

Figure 1. Dendrogram of 108 dogs of 4 breeds constructed by pairwise genetic distance analysis for $K = 4$. Each dog is plotted on the x axis with the distance of IBS given on the y axis. Clusters are represented by different colors, with breeds indicated in the bottom-most panel of the x axis. Breeds included: BOC (green), AUS (blue), PWD (black), and GSD (red).

with PWD and GSD shown as relatively closely related and the 5 distinct BOC together on their own separate branch (Figure 3).

Closer examination of our supertree reflects relationships that can be explained in terms of known pedigree relationships, breed history, phenotype, and geography. The 2 extended pedigrees (BOC and AUS) included for our noise phobia analysis, and the small family groups included to check for Mendelization errors cluster together across all analyses in which they were included. Most of the groups found in this tree agree broadly with the results of an earlier study that used cluster analysis to examine the relationships of breeds with minor exceptions (Parker et al. 2004).

Neither the GSD nor PWD demonstrated significant population structure in any analyses. The homogeneity of these 2 breeds can be explained by sampling (the GSD were sampled from among a close-knit community of European working police and military dog breeders) and breed history (all of today's PWD descend from a limited number of founding dogs).

Simulations

We sought to determine if intrabreed stratification would confound GWAS by carrying out simulations using our observed genotypes. As described above, we identified a group of BOC divergent from the larger group of BOC samples. We randomly assigned the 43 unrelated dogs of this latter group to case or control status. When the 5 distinct but unrelated BOC were split between case–control status, we obtained a near-null distribution with an average chi-squared statistic of 1.005 and a genomic inflation factor

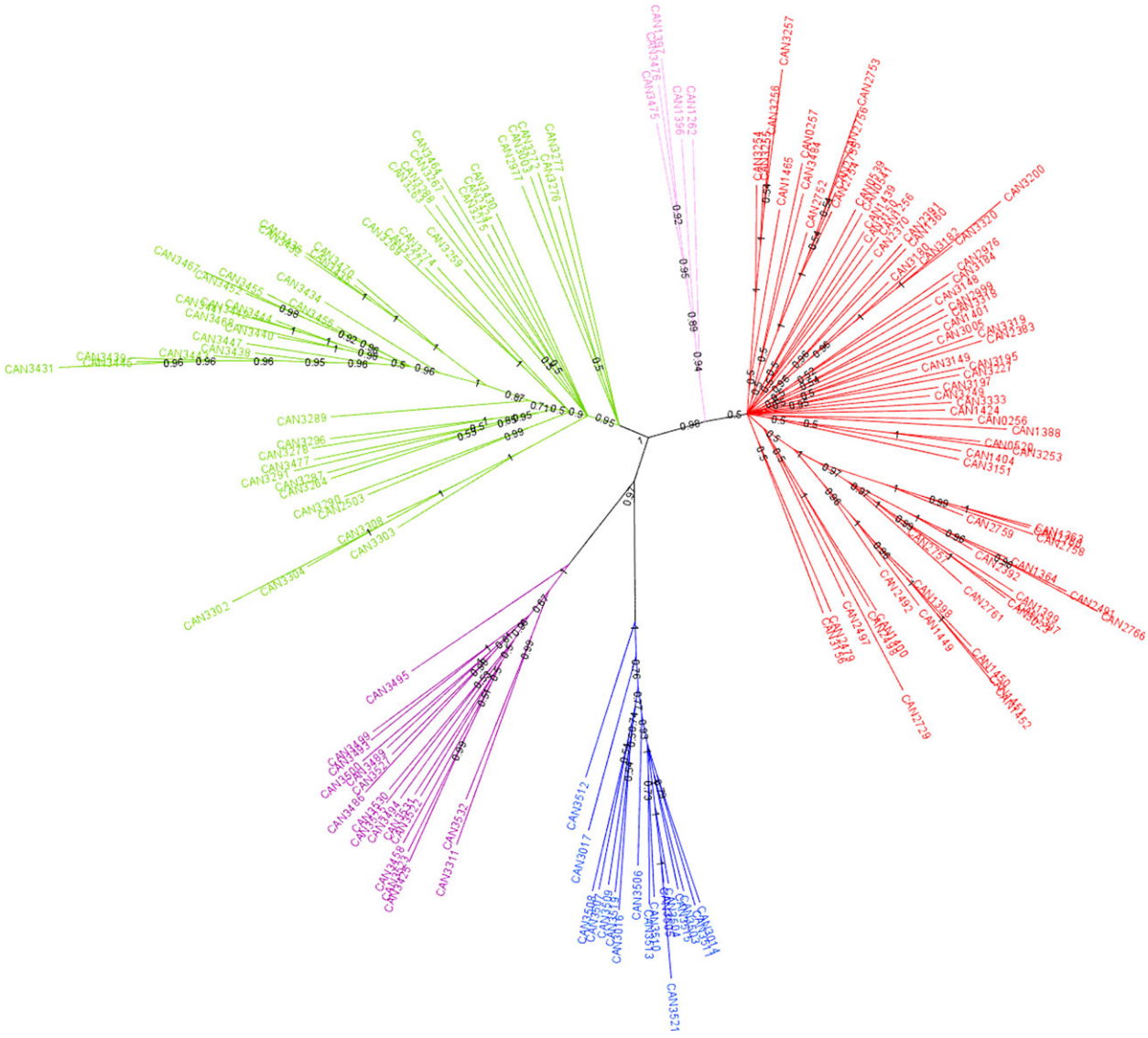

Figure 2. Fifty percent majority rule consensus tree of Bayesian analyses including 4 breeds and 4200 SNPs. Counterclockwise from left: AUS (green), PWD (purple), GSD (blue), and BOC (red). A fifth clade (pink) is made up of the 5 show BOC included in our sample. This is a representative unrooted network with branch lengths proportional to distance. PWD and GSD are sister taxa in this network. The clade containing the 5 show BOC is located almost midway between the node leading to all the rest of the BOC and the node leading to the other 3 breeds. Posterior probabilities are shown for each branch. The posterior probability of the clade containing show dogs in this particular tree is 0.94, which is nearly as high as those supporting the clades representing the 3 other pure breeds, and almost twice as high as the branch supporting all the rest of the BOC (3).

of 1.112 (Table 2). However, when all 5 of these BOCs were assigned to either a case or control group, average chi-squared statistics were 1.180 or 1.168, respectively, with genomic inflation factors of 1.358 and 1.343, respectively, demonstrating significantly increased false-positive rates secondary to stratification artifact (Table 2).

Principal components analyses (PCA) were carried out using the uncorrelated set of 21 100 genome-wide markers, and the positions on each of the first 3 eigenvectors were used as covariates in the simulated GWAS in the BOC samples in a logistic regression framework. This led to genomic inflation factors of 1.0 regardless of whether the 5 divergent BOCs were all assigned to either case or control status, effectively correcting for the observed stratification (Table 2).

Discussion

The BOC, our primary breed of interest for a study of noise phobia, has a long history of selection as a working

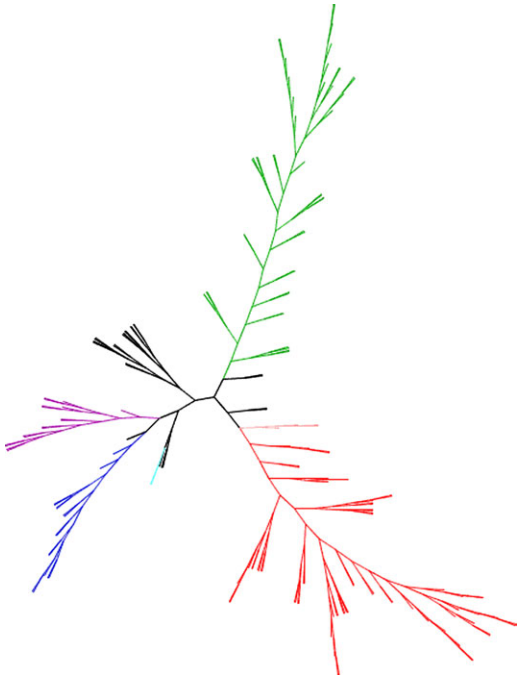

Figure 3. MRP supertree summarizing results of all analyses including 700, 4200, and 21 000 SNPs, with ID numbers removed to improve legibility. Clockwise from top: AUS (green), BOC (red), TBD (light blue), GSD (blue), and PWD (purple). The 5 show BOC included in our sample are shown in pink. Bootstrap values for all branches of this tree were more than 97%.

sheepdog. Breeders of working BOC have historically followed a selection regime that prioritizes behavioral traits considered desirable for herding. This has resulted in an extremely consistent behavioral phenotype, comprising a stereotyped habitus (incorporating working traits, such as "eye," "style," and "sheep sense," or an ability to anticipate the actions of livestock) accompanied by a wide variation in appearance or what conformation (show) breeders would refer to as "type" (Supplemental Figure 7).

But since the 1960s, the breed has also been developed, primarily in Australia and New Zealand, as a show dog. Conformation breeders select for appearance and evaluate the breed worthiness of their dogs on the basis of success in the show ring. The full breed standard published by the Australian National Kennel Council illustrates specifics ranging from idealized body length/height proportions to acceptable and unacceptable ear set (ANKC 2005). This selection regime has produced dogs of extreme homogeneity in appearance, both in Australasia and America, as highly ranked American show dogs are usually derived from Australasian ancestors. In general, these dogs exhibit few or none of the behavioral characteristics desired in working sheepdogs.

Using pedigree, registration, and other demographic information, we were able to determine that the 5 BOC that consistently formed a separate, well-supported clade were distinguished from the rest of our sample, because either they or their ancestors were successful show dogs, and all 5 traced back to show champions from Australasia, either directly or within less than 4 generations. These 5 samples were all collected at conformation shows or were sent to us by owners who participated in AKC-sponsored conformation events with their dogs. The majority of our BOC sample, by contrast, was collected at working sheepdog trials, traces back to British ancestors, and came from owners who use working farm dogs, or breed and train dogs for sheepdog trial competitions, or both. Some suggestive population structure was also found within our AUS sample, but the variation within this breed is not as straightforward to characterize. AUS are characterized by a long history of "dual purpose" breeding, and the heterogeneity we found within this sample probably reflects this fact.

It has previously been suggested that differences in geographic origin in case versus control samples may confound genome-wide association results (Quignon et al. 2007). We suggest that differing selection regimes may exacerbate the situation. Our results are consistent with the results of studies using pedigree analysis or smaller marker

Table 2. Results of 100 simulations of GWAS of approximately 53 000 autosomal SNPs in a total of 48 unrelated BOC

Subpopulation assignment	Test performed	Genomic inflation factor	Average chi-square
Australasian—split case/control[a]	Allelic association	1.112 ± 0.051	1.005 ± 0.030
Australasian—all cases[b]	Allelic association	1.358 ± 0.060	1.180 ± 0.042
Australasian—all cases[c]	Logistic regression w/3 covariates	1 ± 0	0.829 ± 0.023
Australasian—all controls[c]	Allelic association	1.343 ± 0.055	1.168 ± 0.042
Australasian—all controls[c]	Logistic regression w/3 covariates	1 ± 0	0.831 ± 0.024

Five show BOC of Australasian descent were split (2:3) between cases : controls or assigned all to cases or all to controls as noted above, with the balance of dogs randomly assigned either case or control status. Genomic inflation factors and average chi-square values were calculated for all simulations, and descriptive statistics of each are given for allelic association tests or logistic regression using 3 covariates to account for population structure. GWAS with all 5 divergent BOC assigned either case or control status demonstrated inflated false-positives. However, this inflation can be reduced to null by using covariates that account for population substructure.

[a] 22–26 Cases assigned randomly.

[b] 24–28 Cases assigned randomly.

[c] 25–28 Controls assigned randomly.

sets to identify population substructure within single dog breeds (Calboli et al. 2008; Björnerfeldt et al. 2008), further emphasizing the importance of understanding the geographic origin and functional context within which samples are collected for large-scale studies.

Results of previous analyses of smaller marker sets or those sampling only a portion of the genome are concurrent with our results using extensive genome-wide coverage. However, the samples and methods we used in our study, utilizing dense SNP data sampling of all 38 canine autosomes, and modern methods of phylogenetic analysis, allow us to assess relationships both between and within breeds with much finer resolution than previous studies. Awareness of sample origin helps explain the patterns of population substructure that were revealed through our analyses and should allow other researchers to avoid introducing stratification into future analyses by constructing study samples in ways that reduce this confounding effect.

For practical reasons, it may not always be the case that balanced study samples can be obtained. Rather than limit a study's sample size, it may be desirable to explore and implement other means to statistically account for population substructure. In addition to the methods outlined here, intrabreed stratification can also be detected by multidimensional scaling or PCA of an uncorrelated marker set of genome-wide SNP data. Covariates from either of these calculations can then be used in GWAS to statistically correct for substructure, a practice used in human studies to correct for population stratification (Price et al. 2006). For example, the elevated genomic inflation factors resulting from our simulation studies were reduced to null when the complete BOC sample including working and show dogs was instead analyzed by logistic regression using the first 3 principal component vectors as covariates. Another possible approach, when phylogenetic analyses result in well-supported hypotheses of relationships within a given sample of dogs, might incorporate methods such as phylogenetically independent contrasts that are effective for identifying associations between characters (marker and phenotype) reflecting shared ancestry rather than causative genetic factors (Felsenstein 1985; Midford et al. 2005).

These results have important implications for genetic association studies in dogs. Contrary to common assumptions, within-breed population structure can be significant in some breeds, and this stratification may be explained by geographical origin, by artificial selection criteria used by dog breeders, or both. Demographic and pedigree information should be used to guide the collection of study samples that are free of significant within-breed population structure. In addition, genetic data gathered during the performance of GWAS can be used to statistically adjust for substructure.

Supplementary Data

Supplementary Figures 1–7 can be found at http://www.jhered.oxfordjournals.org/.

Funding

McKnight Foundation (to S.P.H.), the Hellman Family Fund (to S.P.H.), a University of California, San Francisco, Innovations in Basic Science award (to S.P.H.), the AKC (ACORN award 850-A to S.P.H.), and the Defense Advanced Research Projects Agency (DARPA FY-06-0028/52731-LS-DRP to K.L.O.).

Acknowledgments

The authors would like to thank the numerous owners and breeders contributing genetic material for this work. Distribution Statement "A" Approved for Public Release, Distribution Unlimited.

References

Australian National Kennel Council (ANKC), 2005. The extended breed standard of the Border Collie. Victoria (Australia): The Border Collie Club of Victoria and the Australian National Kennel Council.

Björnerfeldt S, Hailer F, Nord M, Vilà C. 2008. Assortative mating and fragmentation within dog breeds. BMC Evol Biol. 8:28.

Calboli FC, Sampson J, Fretwell N, Balding DJ. 2008. Population structure and inbreeding from pedigree analysis of purebred dogs. Genetics. 179(1):593–601.

Chen D, Eulenstein O, Fernández-Baca D. 2004. Rainbow: a toolbox for phylogenetic supertree construction and analysis. Bioinformatics. 20(16): 2872–2873.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol. 14(8):2611–2620.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164:1567–1587.

Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes. 7(4):574–578.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat. 125:1–15.

Hennig W. 1966. Phylogenetic systematics. Urbana (IL): University of Illinois Press.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294:2310–2314.

Midford PE, Garland T, Jr, Maddison WP. 2005. PDAP package of mesquite. Version 1.07. Available from: URL http://mesquiteproject.org/pdap_mesquite/index.html.

Overall KL, Hamilton SP, Chang ML. 2006. Understanding the genetic basis of canine anxiety: phenotyping dogs for behavioral, neurochemical, and genetic assessment. J Vet Behav: Clin Appl Res. 1(3):124–141.

Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, Malek TB, Johnson GS, DeFrance HB, Ostrander EA, Kruglyak L. 2004. Genetic structure of the purebred domestic dog. Science. 304:1160–1164.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38(8):904–909.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. pLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 81(3):559–575.

Quignon P, Herbin L, Cadieu E, Kirkness EF, Hédan B, Mosher DS, Galibert F, André C, Ostrander EA, Hitte C. 2007. Canine population structure: assessment and impact of intra-breed stratification on SNP-based association studies. PLoS ONE. 2(12):e1324.

Rambaut A. 2008. FigTree v1.1.2. Edinburgh (UK): Institute of Evolutionary Biology, University of Edinburgh.

Ronquist F, Huelsenbeck JP. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Salmon Hillbertz NH, Isaksson M, Karlsson EK, Hellmén E, Pielberg GR, Savolainen P, Wade CM, von Euler H, Gustafson U, Hedhammar A, et al. 2007. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. Nat Genet. 39(11):1318–1320.

Sikes DS, Lewis PS. 2001. Beta software, version 1. PAUPRat: PAUP implementation of the parsimony ratchet. Distributed by the authors. Storrs (CT): Department of Ecology and Evolutionary Biology, University of Connecticut.

Sorenson MD, Franzosa EA. 2007. TreeRot, version 3. Boston (MA): Boston University.

Sutter NB, Ostrander EA. 2004. Dog Star Rising: the canine genetic system. Nature Genetics. 5:900–910.

Swofford DL. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Received November 23, 2008; Revised February 13, 2009; Accepted February 25, 2009

Corresponding Editor: Francis Galibert