Published in final edited form as:

Mamm Genome. 2016 December ; 27(11-12): 599–609. doi:10.1007/s00335-016-9660-9.

Evaluation of Artificial Selection in Standard Poodles Using Whole-Genome Sequencing

Steven G. Friedenberg^{1,3}, Kathryn M. Meurs^{1,3}, and Trudy F.C. Mackay^{2,3}

¹Department of Clinical Sciences, College of Veterinary Medicine, North Carolina State University, 1060 William Moore Drive, Raleigh, NC 27607 USA

²Department of Biological Sciences, College of Sciences, North Carolina State University, 3510 Thomas Hall, Raleigh, NC 27695 USA

³Comparative Medicine Institute, North Carolina State University, Raleigh, NC 27607 USA

Abstract

Identifying regions of artificial selection within dog breeds may provide insights into genetic variation that underlies breed-specific traits or diseases - particularly if these traits or disease predispositions are fixed within a breed. In this study, we searched for runs of homozygosity (ROH) and calculated the d_i statistic (which is based upon F_{ST}) to identify regions of artificial selection in Standard Poodles using high-coverage, whole genome sequencing data of 15 Standard Poodles and 49 dogs across seven other breeds. We identified consensus ROH regions 1 Mb in length and common to at least 10 Standard Poodles covering 0.6% of the genome, and d_i regions that most distinguish Standard Poodles from other breeds covering 3.7% of the genome. Within these regions, we identified enriched gene pathways related to olfaction, digestion, and taste, as well as pathways related to adrenal hormone biosynthesis, T cell function, and protein ubiquitination that could contribute to the pathogenesis of some Poodle-prevalent autoimmune diseases. We also validated variants related to hair coat and skull morphology that have previously been identified as being under selective pressure in Poodles, and flagged additional polymorphisms in genes such as ITGA2B, CBX4, and TNXB that may represent strong candidates for other common Poodle disorders.

Keywords

dog; genetics; next-generation sequencing

Address for correspondence: S.G. Friedenberg, Department Clinical Sciences, College of Veterinary Medicine, North Carolina State University, 1060 William Moore Drive, Raleigh, NC 27607, USA, steven_friedenberg@ncsu.edu, +1 919 515 3277.

Author contributions

SGF collected samples, designed the study, analyzed the data, and wrote the manuscript. KMM collected samples and supervised the study. TFCM also supervised and provided guidance to the study. All authors have read and edited the manuscript.

Conflict of interest

The authors declare no conflicts of interest.

Introduction

Recent population bottlenecks and selective breeding have largely been responsible for the development of modern dog breeds (Boyko 2011; Marsden et al. 2015). Within a dog breed, genetic diversity is severely reduced relative to ancestral canine populations and modern mixed-breed dogs (Karlsson and Lindblad-Toh 2008). Selective breeding has led to the development of breed-specific traits and characteristics such as size and coat color, many of which have been well described in the literature (Sutter et al. 2007; Cadieu et al. 2009; Boyko et al. 2010). In addition to these morphologic traits, many dog breeds have a relatively high prevalence of certain diseases, ranging from dilated cardiomyopathy in the Doberman Pinscher to atopic dermatitis in West Highland White Terriers (Sousa and Marsella 2001; Wess et al. 2010). Over the past 10 years, significant progress has been made in identifying disease-causing genetic polymorphisms that distinguish affected from unaffected dogs within a breed, primarily through genome-wide association studies (GWAS) followed by fine-mapping and/or additional sequencing (Parker et al. 2009; Goldstein et al. 2010; Meurs et al. 2010; Seppälä et al. 2011; Kyöstiä et al. 2012).

For some canine diseases, however, causative genetic polymorphisms may be fixed within a breed, making case/control GWAS impossible. Myxomatous mitral valve degeneration in Cavalier King Charles Spaniels could represent such a disease, as most dogs within the breed develop the characteristic nodular endocardial valvular lesions (Häggström et al. 1992; Egenvall et al. 2006). Additionally, for diseases that require an environmental trigger prior to the development of clinical signs (such as autoimmune disorders), some predisposing genetic variants may be at high frequency within a breed, which could make implementing an adequately powered GWAS challenging. For these reasons, identifying signatures of artificial selection could be helpful in identifying disease-causing or disease-predisposing genes for disorders that tend to cluster within a particular breed.

Several studies have been performed in dogs aimed at identifying such signatures of artificial selection. Akey *et al.* used data from an array of 21,000+ single nucleotide polymorphisms (SNPs) to identify regions of artificial selection in 10 phenotypically diverse dog breeds using the F_{ST} -based metric d_i (Akey et al. 2010). Vaysse *et al.* applied this metric to data generated from the Illumina 174,000+ SNP array in 46 dog breeds (Vaysse et al. 2011). Studies in dogs, cows, and horses have also used runs of homozygosity (ROH) as a method of detecting signatures of artificial selection within specific breeds (Freedman et al. 2014; Kim et al. 2015; Metzger et al. 2015).

Here, we used these methods (ROH, d_i) to identify regions of artificial selection with a much denser set of SNPs derived from whole genome sequencing (WGS) of 64 dogs across eight distinct breeds. We chose these two metrics in order to identify those regions of the genome that are mostly fixed within a single breed (ROH) as well as those regions of the genome that most substantially contribute to cross-breed divergence (d_i). We focused specifically on the Standard Poodle as this breed has a relatively high prevalence of several autoimmune diseases (Tevell et al. 2008; Pedersen et al. 2015; Hanson et al. 2015) and also represents one of the most common breeds in our dataset of whole genome sequences. Additionally, because our data were derived from whole genome sequencing rather than SNP arrays, we

used our findings to pinpoint specific polymorphisms within regions of artificial selection that may be responsible for some Poodle-specific traits.

Our findings suggest that searching for signatures of artificial selection within a breed using WGS data is both feasible and accurate, and could help identify additional genetic variation that contributes to normal canine morphologic characteristics. Additionally, these methods may complement existing GWAS approaches by identifying some variants that contribute to the development of diseases that are at high frequency within a particular breed.

Methods

Samples

We selected 20 Boxers, 15 Standard Poodles, 6 Great Danes, 6 Scottish Terriers, 5 Scottish Deerhounds, 4 Collies, 4 Doberman Pinschers, and 4 West Highland White Terriers for inclusion in this study. Samples were collected as part of ongoing disease-related research in our laboratory. Based upon available pedigree data, there were no known familial relationships among any of the included dogs going back at least three generations. DNA was extracted from EDTA blood samples obtained from each dog using the standard protocol of the DNeasy Blood and Tissue Kit (Qiagen).

Next-generation sequencing

Approximately 3 µg of DNA was submitted for library preparation and whole genome sequencing at the University of North Carolina Chapel Hill High Throughput Sequencing Facility (46), the Medical University of South Carolina Proteogenomics Facility (8), the University of Missouri DNA Core (6), or the Genomics Sciences Laboratory at North Carolina State University (4) (numbers in parenthesis represent the numbers of samples sequenced at each institution). All sequencing experiments were designed as 100- or 125-bp paired-end reads and each sample was run on either 1 or 2 lanes of an Illumina HiSeq 2000 or 2500 high-throughput sequencing system.

Analysis of next-generation sequencing data was performed using standardized bioinformatics pipeline for all samples as described previously (Friedenberg and Meurs 2016). Briefly, sequence reads were trimmed using Trimmomatic 0.32 (Bolger et al. 2014) to a minimum phred-scaled base quality score of 30 at the start and end of each read with a minimum read length of 70 bp, and aligned to the canFam3 reference sequence (Lindblad-Toh et al. 2005) using BWA 0.7.10 (Li and Durbin 2009). Aligned reads were prepared for analysis using Picard Tools 1.115 (<http://broadinstitute.github.io/picard>) and GATK 3.4 (McKenna et al. 2010) following best practices for base quality score recalibration and indel realignment specified by the Broad Institute, Cambridge, MA (DePristo et al. 2011; Van der Auwera et al. 2013). Variant calls were made using GATK's HaplotypeCaller walker, and variant quality score recalibration (VQSR) was performed using sites from dbSNP 139 and the Illumina CanineHD BeadChip as training resources.

Variant filtering

In order to select variants for downstream analysis, we first applied a VQSR tranche sensitivity cutoff of 99.9% to SNPs and 99% to indels. Using both GATK and VCFtools (Danecek et al. 2011), we then set any genotype call with a phred-scaled quality score < 20 to missing, and further filtered variant sites to include only biallelic SNPs with a minimum call rate across all samples of 95%, minimum minor allele frequency (MAF) of 0.02, and Hardy-Weinberg equilibrium (HWE) p-value 1×10^{-7} . The resulting set of SNPs was used to determine ROH and d_i . A similarly filtered set of variants including both SNPs and indels was used to evaluate the variant effects within the ROH and d_i regions.

Runs of homozygosity

The filtered SNPs were subset to include only Standard Poodles, and autosomal runs of homozygosity ≥ 1 Mb in length and common to at least 10/15 dogs (to account for genotyping errors) were identified using PLINK 1.9 (Chang et al. 2015). Parameters were similar to those described elsewhere (Marsden et al. 2015), but with modifications to identify consensus regions in at least 10 of the dogs. Input flags were as follows: `--chr 1-3 8 --homozyg-snp 200 --homozyg-kb 1000 --homozyg-window-missing 100 --homozyg-window-het 1 --allow-no-sex --dog --homozyg-match 0.95 --homozyg group --pool-size 10`. The consensus regions were converted to a BED file using custom scripting in R 3.2.3 (Team 2015), and the combined SNP/indel file was subset using GATK to include only those variants within the consensus regions. In R, we converted GATK-calculated alternate allele frequencies into minor allele frequencies (relative to Standard Poodles) for downstream analyses. The effects of the resulting variants were evaluated using Variant Effect Predictor (VEP) 83 (McLaren et al. 2010) with both gene ontology and GERP++ plugins (Davydov et al. 2010); variant filtering was performed in R.

Calculation of d_i

Using the SNP dataset described above, we calculated the pairwise Weir and Cockerham F_{ST} (Weir and Cockerham 1984) for each SNP between Standard Poodles and each of the other 7 breeds in our dataset using VCFtools. We then imported this data into R, and using custom scripting, calculated the d_i statistic (Akey et al. 2010) all autosomal SNPs using the formula

$$d_i = \sum_{j \neq i} \frac{F_{ST}^{ij} - E[F_{ST}^{ij}]}{sd[F_{ST}^{ij}]}$$

where $E[F_{ST}^{ij}]$ is the expected value and $sd[F_{ST}^{ij}]$ is the standard deviation of the F_{ST} between Standard Poodles (i) and each other breed (j). To visualize trends across the genome, we used locally weighted scatterplot smoothing (LOESS) of the resulting d_i values for each chromosome with a very low α (0.02) in order to maximize peak resolution. Following methods published elsewhere (Akey et al. 2010; Vaysse et al. 2011), we identified peaks crossing the top 1% threshold of LOESS-smoothed values, and using the zoo package in R (Zeileis and Grothendieck 2005) identified local minima surrounding these peaks. Each peak – from local minimum to local maximum (crossing the 1% threshold) to local

minimum – was identified as a genomic region under selection for downstream analyses. These regions were converted to a BED file, and Standard Poodle SNPs and indels within these regions were selected and evaluated using GATK and R. Variant effects were determined using VEP 83 as described above.

Pathway analysis

We evaluated genes in both the ROH and d_i regions using Enrichr (Chen et al. 2013). Gene symbols were mapped to the human Entrez Gene database using biomaRt (Smedley et al. 2015) in order to permit recognition by pathway analysis software, and genes were grouped into ontologies/pathways using databases from Gene Ontology (GO) 2015 and KEGG 2016 (Kanehisa and Goto 2000; Ashburner et al. 2000; Kanehisa et al. 2013). Pathways were scored for enrichment using Fisher's exact test. We analyzed the resulting data to (1) identify gene clusters within the ROH and d_i regions and (2) identify overrepresented genes on a genome-wide level. For the first analysis, we filtered for pathways with a p-value ≤ 0.05 containing at least 5 genes per pathway; for the second analysis, we filtered for pathways with Benjamini-Hochberg corrected p-value ≤ 0.05 using the set of known human protein-coding genes as a reference.

Results

Variant callset

Average depth for whole genome sequences ranged from 19–37 $\times$; 96.5–98.1% of bases were covered to a minimum depth of 15 $\times$. After variant calling and filtering, 7,450,704 biallelic SNPs and 987,062 biallelic indels remained across all eight breeds evaluated.

ROH analysis

We identified a total of 2,164 individual ROH > 1 Mb in all 15 Standard Poodles (Table 1); these ROH are located on all 38 autosomes. We identified only 17 consensus ROH regions with overlapping segments in at least 10/15 Standard Poodles (Supplemental Table 1); these consensus regions are located on 5 chromosomes and span 13.6 Mb (approximately 0.6% of the canine genome).

Within the ROH consensus regions, we found 42,665 variants in 117 genes (Supplemental Table 2). Using the impact classification scheme defined by VEP (http://useast.ensembl.org/info/genome/variation/predicted_data.html#consequences), the vast majority of these variants (42,518 or 99.7%) are predicted to have a modifier effect on protein function, and well over half of these variants (23,861 or 55.9%) are predicted to be located in intergenic regions. A total of 1,072 variants have a GERP++ conservation score of 4 or greater (highly evolutionarily constrained), and an additional 1,636 variants have a GERP++ conservation score between 2–4 (moderately evolutionarily constrained) (Goode et al. 2010; Marsden et al. 2015).

Of the 42,665 variants in the ROH consensus regions, 19,833 (46%) had a MAF ≤ 0.05 , indicating a high level of fixation in our population of Standard Poodles. These variants are located in 115/117 of the previously described genes (all except ENSCAFG00000003812

and ENSCAFG00000029696). A similar percentage of these variants (99.6%) are predicted by VEP to have modifier effects on gene function and are located in intergenic regions (54%). Only 624 of these relatively fixed variants were predicted to have a strong effect on gene function, including 569 with a conservation score ≥ 4 , 12 with a “high” VEP consequence, and 46 with a “moderate” VEP consequence. These variants are located in 60 genes, which are flagged as “Contains fixed strong effect variants” in Supplemental Table 2.

d_i analysis

Using the d_i statistic, we identified a total of 52 regions that most strongly contribute to differentiation between Standard Poodles and the other breeds in our study (Supplemental Table 3); these regions are located on 24 chromosomes and span 90.1 Mb (approximately 3.7% of the canine genome). A plot of d_i values across the genome in Standard Poodles, with overlaid ROH consensus regions, is shown in Figure 1 (with a more expanded version Supplemental Figure 1).

We validated our calculation of the d_i statistic by testing whether genes known to be associated with coat features in Standard Poodles (and hence highly likely to be under positive selection) are present in the d_i regions of the genome we identified. We found three genes associated with coat style such as fur length, texture, and curl (RSPO2, KRT71, and FGF5 (Cadieu et al. 2009)) and two genes associated with coat color (TYRP1, CBD103 (Schmutz et al. 2002; Candille et al. 2007)) overlapping genomic regions in the 99th percentile of our calculated d_i values (Figure 2). Additionally, we identified two genes (BMP3 and IGFBP4) known to contribute to canine skull morphology (Jones et al. 2008; Schoenebeck et al. 2012) within the peak d_i regions.

Within the d_i regions, we found 344,389 variants in a total of 951 genes (Supplemental Table 4). Similar to the ROH consensus regions, the vast majority of these variants (342,608) are predicted to have a modifier effect on gene function and just over 60% (210,121) are predicted to be located in intergenic regions. A total of 8,502 variants have a GERP++ conservation score ≥ 4 and 13,684 variants have a score between 2–4.

Of the 344,389 variants in the d_i regions, 108,980 (32%) had a MAF ≥ 0.05 , which as might be expected, is lower than those in the ROH consensus regions. These variants are located in 907/951 of the previously described genes. A similar percentage of these variants (99.4%) are predicted by VEP to have modifier effects on gene function and are located in intergenic regions (62.9%). Only 3,146 of these fixed variants were predicted to have a strong effect on gene function, including 2,667 with a conservation score ≥ 4 , 111 with a “high” VEP consequence, and 422 with a “moderate” VEP consequence. These variants are located in 521 genes, which are similarly flagged as “Contains fixed strong effect variants” in Supplemental Table 4

Pathway analysis

Of the 117 genes in the ROH consensus regions, 108 map to the human Entrez Gene database and were recognized by Enrichr (all unmapped genes are pseudogenes not in the Entrez gene database). Ten pathways (five GO Biological Process, four GO Molecular Function, one KEGG) have clusters with ≥ 5 genes (Figure 3a). Five of these pathways

are related to olfaction, three are related to digestion, and two are related to cell cycle regulation.

We also noted that eight genes in the ROH consensus regions are located in the T-cell receptor (TCR) β chain locus of the canine genome (chromosome 16, ~6.8–7.0 Mb). These genes include TRBV30, TRBV3-1, TRBV15, TRBV16, ENSCAFG00000003811, ENSCAFG00000003812, ENSCAFG00000014478, and ENSCAFG00000024810). However, these genes are very broadly defined in by GO as “protein binding” (and are not annotated by KEGG) and therefore did not meet our initial criteria for enrichment.

Of the 951 genes in the d_i regions, 830 map to the human Entrez Gene database and were recognized by Enrichr (57 genes have no human ortholog and 64 genes are not in the Entrez Gene database as they are mostly RNA genes or pseudogenes). Nineteen pathways (12 GO Biological Process, two GO Cellular Component, five GO Molecular Function) have clusters with 5 genes (Figure 3b). Overrepresented pathways include those related to taste, digestion, keratin proteins, ubiquitination, gene expression (DNA methylation, DNA binding), and hormone biosynthesis.

Within the d_i regions, we noted two apparent groupings of genes that were not identified using Enrichr: the eight genes in the TCR β chain locus described previously, as well as 33 zinc finger proteins located throughout the genome. Like the TCR genes, some of these zinc finger genes are not consistently annotated and hence did not meet our criteria for enrichment.

We also evaluated whether any pathways are significantly enriched for genes on a genome-wide level in either the ROH consensus regions or d_i regions. Only two pathways met this criterion (Table 2): one for taste and smell perception containing 14 genes in the ROH consensus regions (B-H $p = 0.036$), and one for protein ubiquitination containing 21 genes across the d_i regions (B-H $p = 0.021$).

Variant analysis

We subset the fixed, strong-effect variants within the ROH and d_i regions to identify specific polymorphisms more common in Standard Poodles than other breeds and whose consequences might be readily interpretable. We first filtered for variants with a MAF ≥ 0.2 in the other seven breeds combined (345 remaining within the ROH consensus regions; 1,076 remaining within the d_i regions). We then filtered to include only those variants with a “high” or “moderate” VEP impact (10 in ROH regions, 77 in d_i regions), followed by an additional filter to include only those variants with a GERP++ conservation score ≥ 2 , or a SIFT (Kumar et al. 2009) prediction of deleterious, deleterious low confidence, or tolerated low confidence. After filtering, one variant remained in the ROH consensus regions and 33 variants remained in the d_i regions (Supplemental Table 5); the one variant in the ROH consensus regions was also present in the d_i regions.

This list of variants includes polymorphisms in FGF5 and KRT71 that have been previously reported (Cadieu et al. 2009), as well as polymorphisms in CBX4 and TNXB which contribute to skin morphology and development (Mao et al. 2002; Mardaryev et al. 2016).

Many of the other variants have no specifically identifiable role at present, but are broadly responsible for gene (SETX (Skourti-Stathaki et al. 2011), ZNF658 (Ogo et al. 2015), RPL34 (Kenmochi et al. 1998), ENSCAFG00000005986 (Ben Yehuda et al. 1998), ENSCAFG00000003493 (Mao et al. 2016)) and protein (TRAPPC12 (Scrivens et al. 2011), LRRC3C (Kobe and Kajava 2001) regulation.

Discussion

In this study, we used SNP data derived from WGS to identify regions of selection in Standard Poodles using both runs of homozygosity and the d_i statistic. By grouping genes within these regions and examining sequence-level data, we identified specific pathways and polymorphisms under selection in this breed. While our study focused specifically on Standard Poodles, the methods we have documented here should be easily applicable to any dog breed or within other species.

Runs of homozygosity identify regions of reduced genomic diversity and have been shown to develop as a result of strong selective pressures (Ku et al. 2011; Purfield et al. 2012). Within the ROH consensus regions in Standard Poodles, we identified gene pathways that are mostly related to taste, olfaction, and digestion (Figure 3a). Dogs are well known to have an exquisite sense of smell, and over 1,000 genes have been identified related to canine olfaction (Quignon et al. 2005); hence it is unsurprising that many polymorphisms in these genes have become fixed within the genome and are overrepresented in our analysis. Additionally, Axelsson *et al.* documented that dog domestication has involved adaptation to a starch-rich diet (Axelsson et al. 2013), and along with these adaptations it is also unsurprising that other genes related to taste and digestion may be under selective pressure as well.

The d_i statistic, which is based upon F_{ST} , provides an alternate method of identifying genomic regions under selection (Akey et al. 2010). Because F_{ST} is primarily a measure of population differentiation, d_i is likely to highlight those regions of the genome that help distinguish Standard Poodles from other breeds rather than regions that are fixed across all (or most) dogs. Within these d_i regions, we flagged seven genes that have been shown by others to be associated with morphologic characteristics common in Standard Poodles. Additionally, we did *not* find evidence of selection for certain genes associated with characteristics that are absent from Standard Poodles, such as inverted hair growth (Karlsson et al. 2007) or a corkscrew tails (Jones et al. 2008).

Within the d_i regions, we identified several overrepresented gene pathways in Standard Poodles (Figure 3b). For example, genes related to taste and digestion that we identified in the ROH regions are also present in the d_i regions, indicating that certain digestive adaptations may be breed-specific. We also identified a pathway related to adrenal hormone biosynthesis that includes the genes CYP21A2, CYP19A1, and HSD17B3. Autoantibodies against CYP21A2 have been well documented in the pathogenesis of human Addison's disease (Brønstad et al. 2014), and Standard Poodles are the most overrepresented dog breed for canine Addison's disease (Hanson et al. 2015). Additionally, we noted a locus under selection in Standard Poodles on chromosome 16 containing many of the TCR β chain

genes. The β chain comprises one of the hypervariable regions of the TCR that interacts with peptide-loaded MHC molecules, and polymorphisms in these genes have been associated with the development of autoimmune disease in humans (Ohashi 2002; Nicholson et al. 2005; Koehli et al. 2014). Further study of these genes – which are relatively unexplored in dogs – could be helpful in explaining the pathogenesis of some autoimmune diseases common in Standard Poodles, including Addison's disease, sebaceous adenitis, and immune-mediated thrombocytopenia (Grindem et al. 1991; Tevell et al. 2008; Hanson et al. 2015).

We also noted several overrepresented pathways containing genes whose specific functions are more difficult to discern. These pathways relate to protein ubiquitination, DNA methylation, and DNA binding. We might hypothesize that these genes help define some of the subtle secretory, transcriptional, and translational processes that differentiate dog breeds, however without further study, any specific consequences of selection within these pathways is unknown. Interestingly, genes associated with ubiquitination have been associated with autoimmune diseases because of their role in tagging proteins for degradation by the proteasome (Bhoj and Chen 2009). Hence, these genes could also be studied further to elucidate any role they may play in Standard Poodle autoimmune diseases.

In addition to identifying overrepresented gene pathways, we searched for particular genetic variants that may contribute to Poodle-specific traits (Supplemental Table 5). Some of the polymorphisms we flagged have reasonably well-defined functional roles. For example, we identified the missense variants in *FGF5* and *KRT71* that have been previously associated with canine coat phenotypes (Cadieu et al. 2009). Additionally, we identified a missense mutation in the gene *ITGA2B* which encodes for the platelet glycoprotein α_{IIb} and is a recognized component of the human platelet antigen system (Metcalfe et al. 2003). This polymorphism has been previously evaluated in a small number of dogs with immune-mediated thrombocytopenia (Callan et al. 2013), but may be especially relevant in Standard Poodles given the high within-breed prevalence of the disease (Grindem et al. 1991). Lastly, we identified variants in two genes (*CBX4* and *TNXB*) that are associated with skin morphology, and these could represent reasonable candidate genes for further study of sebaceous adenitis in the breed.

Similar to the pathway analysis, however, many of these individual variants are located in genes whose specific role difficult to characterize at present. For example, we identified a highly conserved missense variant in the gene *senataxin* (*SETX*) which is associated with the development of amyotrophic lateral sclerosis in humans (Hirano et al. 2011) - a disease not documented in dogs. However, as many genes have different functions across (and within) species, further analyses are required to identify how, if at all, these particular variants play a role in the development of some Poodle-specific traits.

There are several limitations to this study. First, for some dog breeds used in the estimation of the d_i statistic, we had as few as four dogs/breed (because of limited data availability), which may not provide a completely accurate representation of allele frequencies within those breeds; this may have slightly biased our d_i calculation. Second, we only used two methods to identify regions of artificial selection. Other researchers have used measures

such as S_i , which is based upon relative heterozygosity (Vaysse et al. 2011), Tajima's D, or nucleotide diversity (π) (Schlamp et al. 2016). Application of these techniques may have allowed us to identify other regions of artificial selection in Standard Poodles. Third, our pathway analysis was constrained by the current state of gene annotation, as we observed several genes (e.g. the TCR β chain genes) that were not automatically clustered into relevant pathways or ontologies. We attempted to mitigate improper gene classification by converting canine gene names to their human orthologues, however this approach still resulted in some incomplete or poorly classified annotations. Finally, our analysis of specific polymorphisms within the d_i and ROH regions was limited by our ability to understand the role of a particular variant. We focused solely on VEP high or moderate impact variants that are likely to cause changes in the protein-coding portion of the genome. However, it is likely that other highly conserved variants (which may be intronic, in UTRs, etc.) also play a role in the development of some breed-specific traits. Ongoing work aimed at improving canine genomic annotation may help shed light on the role of these particular polymorphisms.

In summary, our findings demonstrate that identification of regions under artificial selection in dogs using whole-genome sequencing data is both accurate and feasible, and that the identified regions can be used to interrogate specific polymorphisms that may be responsible for breed-specific traits. These methods may therefore serve as a useful adjunct to genome-wide association studies for traits that are relatively fixed within a breed. Given the current limitations of genomic annotation, however, the effects of many of the identified variants are difficult to immediately decipher. Nevertheless, our findings also suggest some key pathways that could be interrogated for their contribution to Standard Poodle-specific diseases, and further studies are warranted that evaluate the effects of the genes within these pathways.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

SGF is supported by a National Institutes of Health T32 training award (5T32OD011130-07). Funding for whole genome sequencing was provided in part by the Poodle Club of America Foundation, the American Kennel Club Canine Health Foundation, the Morris Animal Foundation, and the NCSU Cardiac Genetics Laboratory. Some whole genome sequencing data were graciously contributed by Drs. Natasha J. Olby and Thierry Olivry (10 dogs), and Dr. Leigh Anne-Clark (8 dogs).

References

- Akey JM, Ruhe AL, Akey DT, et al. Tracking footprints of artificial selection in the dog genome. *Proc Natl Acad Sci.* 2010; 107:1160–1165. [PubMed: 20080661]
- Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25:25–29. [PubMed: 10802651]
- Axelsson E, Ratnakumar A, Arendt M-L, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature.* 2013; 495:360–364. [PubMed: 23354050]
- Ben Yehuda S, Dix I, Russell CS, et al. Identification and functional analysis of hPRP17, the human homologue of the PRP17/CDC40 yeast gene involved in splicing and cell cycle control. *RNA.* 1998; 4:1304–1312. [PubMed: 9769104]

- Bhoj VG, Chen ZJ. Ubiquitylation in innate and adaptive immunity. *Nature*. 2009; 458:430–437. [PubMed: 19325622]
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; 30:2114–2120. [PubMed: 24695404]
- Boyko AR. The domestic dog: man's best friend in the genomic era. 2011
- Boyko AR, Quignon P, Li L, et al. A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol*. 2010; 8:e1000451. [PubMed: 20711490]
- Brønstad I, Skinningsrud B, Bratland E, et al. CYP21A2 polymorphisms in patients with autoimmune Addison's disease, and linkage disequilibrium to HLA risk alleles. *Eur J Endocrinol*. 2014; 171:743–750. [PubMed: 25249698]
- Cadiou E, Neff MW, Quignon P, et al. Coat variation in the domestic dog is governed by variants in three genes. *Science*. 2009; 326:150–153. [PubMed: 19713490]
- Callan MB, Werner P, Mason NJ, et al. Polymorphisms in canine platelet glycoproteins identify potential platelet antigens. *Comp Med*. 2013; 63:348–354. [PubMed: 24209971]
- Candille SI, Kaelin CB, Cattanauch BM, et al. A beta-defensin mutation causes black coat color in domestic dogs. *Science*. 2007; 318:1418–1423. [PubMed: 17947548]
- Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015; 4:7–7. [PubMed: 25722852]
- Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013; 14:128. [PubMed: 23586463]
- Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011; 27:2156–2158. [PubMed: 21653522]
- Davydov EV, Goode DL, Sirota M, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++ *PLoS Comp Biol*. 2010; 6:e1001025.
- DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011; 43:491–498. [PubMed: 21478889]
- Egenvall A, Bonnett BN, Häggström J. Heart disease as a cause of death in insured Swedish dogs younger than 10 years of age. *J Vet Intern Med*. 2006; 20:894–903. [PubMed: 16955814]
- Freedman AH, Gronau I, Schweizer RM, et al. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet*. 2014; 10:e1004016–e1004016. [PubMed: 24453982]
- Friedenberg SG, Meurs KM. Genotype imputation in the domestic dog. *Mamm Genome*. 2016
- Goldstein O, Mezey JG, Boyko AR, et al. An ADAM9 mutation in canine cone-rod dystrophy 3 establishes homology with human cone-rod dystrophy 9. *Mol Vis*. 2010; 16:1549–1569. [PubMed: 20806078]
- Goode DL, Cooper GM, Schmutz J, et al. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res*. 2010; 20:301–310. [PubMed: 20067941]
- Grindem CB, Breitschwerdt EB, Corbett WT, Jans HE. Epidemiologic survey of thrombocytopenia in dogs: a report on 987 cases. *Vet Clin Pathol*. 1991; 20:38–43. [PubMed: 12673543]
- Hanson JM, Tengvall K, Bonnett BN, Hedhammar A. Naturally occurring adrenocortical insufficiency - An epidemiological study based on a Swedish-insured dog population of 525,028 Dogs. *J Vet Intern Med*. 2015
- Häggström J, Hansson K, Kvart C. Chronic valvular disease in the cavalier King Charles spaniel in Sweden. 1992
- Hirano M, Quinzii CM, Mitumoto H, et al. Senataxin mutations and amyotrophic lateral sclerosis. *Amyotroph Lateral Scler*. 2011; 12:223–227. [PubMed: 21190393]
- Jones P, Chase K, Martin A, et al. Single-nucleotide-polymorphism-based association mapping of dog stereotypes. *Genetics*. 2008; 179:1033–1044. [PubMed: 18505865]
- Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000; 28:27–30. [PubMed: 10592173]
- Kanehisa M, Goto S, Sato Y, et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2013; 42:D199–D205. [PubMed: 24214961]
- Karlsson EK, Baranowska I, Wade CM, et al. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet*. 2007; 39:1321–1328. [PubMed: 17906626]

- Karlsson EK, Lindblad-Toh K. Leader of the pack: gene mapping in dogs and other model organisms. *Nat Rev Genet.* 2008; 9:713–725. [PubMed: 18714291]
- Kenmochi N, Kawaguchi T, Rozen S, et al. A map of 75 human ribosomal protein genes. *Genome Res.* 1998; 8:509–523. [PubMed: 9582194]
- Kim E-S, Sonstegard TS, Van Tassell CP, et al. The Relationship between Runs of Homozygosity and Inbreeding in Jersey Cattle under Selection. *PLoS One.* 2015; 10:e0129967–e0129967. [PubMed: 26154171]
- Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol.* 2001; 11:725–732. [PubMed: 11751054]
- Koehli S, Naeher D, Galati-Fournier V, et al. Optimal T-cell receptor affinity for inducing autoimmunity. *Proc Natl Acad Sci.* 2014; 111:17248–17253. [PubMed: 25411315]
- Ku CS, Naidoo N, Teo SM, Pawitan Y. Regions of homozygosity and their impact on complex diseases and traits. *Hum Genet.* 2011; 129:1–15. [PubMed: 21104274]
- Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 2009; 4:1073–1081. [PubMed: 19561590]
- Kyöstilä K, Cizinauskas S, Seppälä EH, et al. A SEL1L mutation links a canine progressive early-onset cerebellar ataxia to the endoplasmic reticulum-associated protein degradation (ERAD) machinery. *PLoS Genet.* 2012; 8:e1002759. [PubMed: 22719266]
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25:1754–1760. [PubMed: 19451168]
- Lindblad-Toh K, Wade CM, Mikkelsen TS, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature.* 2005; 438:803–819. [PubMed: 16341006]
- Mao JR, Taylor G, Dean WB, et al. Tenascin-X deficiency mimics Ehlers-Danlos syndrome in mice through alteration of collagen deposition. *Nat Genet.* 2002; 30:421–425. [PubMed: 11925569]
- Mao Y, Tamura T, Yuki Y, et al. The hnRNP-Htt axis regulates necrotic cell death induced by transcriptional repression through impaired RNA splicing. *Cell Death Dis.* 2016; 7:e2207–e2207. [PubMed: 27124581]
- Mardaryev AN, Liu B, Rapisarda V, et al. Cbx4 maintains the epithelial lineage identity and cell proliferation in the developing stratified epithelium. *J Cell Biol.* 2016; 212:77–89. [PubMed: 26711500]
- Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, et al. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences.* 2015
- McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297–1303. [PubMed: 20644199]
- McLaren W, Pritchard B, Rios D, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics.* 2010; 26:2069–2070. [PubMed: 20562413]
- Metcalfe P, Watkins NA, Ouwehand WH, et al. Nomenclature of human platelet antigens. *Vox Sang.* 2003; 85:240–245. [PubMed: 14516468]
- Metzger J, Karwath M, Tonda R, et al. Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. *BMC Genomics.* 2015; 16:764–764. [PubMed: 26452642]
- Meurs KM, Mauceli E, Lahmers S, et al. Genome-wide association identifies a deletion in the 3' untranslated region of striatin in a canine model of arrhythmogenic right ventricular cardiomyopathy. *Hum Genet.* 2010; 128:315–324. [PubMed: 20596727]
- Nicholson MJ, Hahn M, Wucherpfennig KW. Unusual features of self-peptide/MHC binding by autoimmune T cell receptors. *Immunity.* 2005; 23:351–360. [PubMed: 16226501]
- Ogo OA, Tyson J, Cockell SJ, et al. The zinc finger protein ZNF658 regulates the transcription of genes involved in zinc homeostasis and affects ribosome biogenesis through the zinc transcriptional regulatory element. *Mol Cell Biol.* 2015; 35:977–987. [PubMed: 25582195]
- Ohashi PS. T-cell signalling and autoimmunity: molecular mechanisms of disease. *Nat Rev Immunol.* 2002

- Parker HG, VonHoldt BM, Quignon P, et al. An expressed Fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science*. 2009; 325:995–998. [PubMed: 19608863]
- Pedersen NC, Brucker L, Tessier NG, et al. The effect of genetic bottlenecks and inbreeding on the incidence of two major autoimmune diseases in standard poodles, sebaceous adenitis and Addison's disease. *Canine Genet Epidemiol*. 2015; 2:14. [PubMed: 26401342]
- Purfield DC, Berry DP, McParland S, Bradley DG. Runs of homozygosity and population history in cattle. *BMC Genet*. 2012; 13:70. [PubMed: 22888858]
- Quignon P, Giraud M, Rimbault M, et al. The dog and rat olfactory receptor repertoires. *Genome Biol*. 2005; 6:R83. [PubMed: 16207354]
- Schlamp F, van der Made J, Stambler R, et al. Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Mol Ecol*. 2016; 25:342–356. [PubMed: 26589239]
- Schmutz SM, Berryere TG, Goldfinch AD. TYRP1 and MC1R genotypes and their effects on coat color in dogs. *Mamm Genome*. 2002; 13:380–387. [PubMed: 12140685]
- Schoenebeck JJ, Hutchinson SA, Byers A, et al. Variation of BMP3 contributes to dog breed skull diversity. *PLoS Genet*. 2012; 8:e1002849–e1002849. [PubMed: 22876193]
- Scrivens PJ, Noueihed B, Shahrzad N, et al. C4orf41 and TTC-15 are mammalian TRAPP components with a role at an early stage in ER-to-Golgi trafficking. *Mol Biol Cell*. 2011; 22:2083–2093. [PubMed: 21525244]
- Seppälä EH, Jokinen TS, Fukata M, et al. LGI2 truncation causes a remitting focal epilepsy in dogs. *PLoS Genet*. 2011; 7:e1002194. [PubMed: 21829378]
- Skourti-Stathaki K, Proudfoot NJ, Gromak N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol Cell*. 2011; 42:794–805. [PubMed: 21700224]
- Smedley D, Haider S, Durinck S, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res*. 2015; 43:W589–W598. [PubMed: 25897122]
- Sousa CA, Marsella R. The ACVD task force on canine atopic dermatitis (II): genetic factors. *Vet Immunol Immunopathol*. 2001; 81:153–157. [PubMed: 11553376]
- Sutter NB, Bustamante CD, Chase K, et al. A single IGF1 allele is a major determinant of small size in dogs. *Science*. 2007; 316:112–115. [PubMed: 17412960]
- Team RC. R: A language and environment for statistical computing. 2015
- Tevell EH, Bergvall K, Egenvall A. Sebaceous adenitis in Swedish dogs, a retrospective study of 104 cases. *Acta Vet Scand*. 2008; 50:11–11. [PubMed: 18501018]
- Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013; 11:11.10.1–11.10.33.
- Vaysse A, Ratnakumar A, Derrien T, et al. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet*. 2011; 7:e1002316. [PubMed: 22022279]
- Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution*. 1984; 38:1358–1370.
- Wess G, Schulze A, Butz V, et al. Prevalence of Dilated Cardiomyopathy in Doberman Pinschers in Various Age Groups. *J Vet Intern Med*. 2010; 24:533–538. [PubMed: 20202106]
- Zeileis A, Grothendieck G. zoo: S3 infrastructure for regular and irregular time series. *J Stat Softw*. 2005; 14:1–27.

Figure 1.
LOESS-smoothed d_i values derived from whole-genome sequencing of 15 Standard Poodles compared to 7 other dog breeds. The dashed red line represents the 99th percentile cutoff value for the d_i statistic. Chromosome numbers are labeled at the top of the plot.

Figure 2.

LOESS-smoothed d_i values from portions of chromosomes 9, 11, 13, 16, 27, and 32 with overlaid positions of 3 genes known to be responsible for canine coat style (RSPO2, KRT71, FGF5), 2 genes known to be responsible for canine coat color (TYRP1, CBD103), and 2 genes known to contribute to canine skull morphology (BMP3, IGFBP4). The dashed red line represents the 99th percentile cutoff value for the d_i statistic; the dashed blue rectangles represent the boundaries of the d_i peaks. d_i values across the entire genome are shown in Supplemental Figure 1.

Figure 3. Heat map showing pathways with at least 5 genes per pathway in the (A) ROH consensus regions and (B) d_7 regions. Heat map colors represent Fisher's exact p-values for enrichment within each pathway; only pathways where $p \leq 0.05$ are shown. Within the ROH regions, overrepresented pathways include those related to olfaction, digestion, and cell cycle regulation. Within the d_7 regions, overrepresented pathways include those related to taste,

digestion, keratin proteins, ubiquitination, gene expression (DNA methylation, DNA binding), and hormone biosynthesis.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Summary of all runs of homozygosity (ROH) > 1 Mb in length identified from whole genome sequencing of 15 unrelated Standard Poodles.

ROH length (Mb)	% genome	Number of runs per dog	SNPs per run
> 1	20.5	144.2	10,820
> 2	16.9	85.6	14,869
> 4	11.3	38.5	21,442
> 8	4.8	10.2	34,199

Table 2

Pathways significantly enriched for genes on a genome-wide basis within either the ROH consensus regions or the d_f regions in 15 unrelated Standard Poodles.

Pathway	Gene count/ Genes in pathway	B-H p-value	Genes	Analysis
ubiquitinyl hydrolase activity (GO:0036459)	21/105	0.021	USP36; USP17L15; USP17L18; USP17L17; USP17L19; USP33; USP17L10; USP17L21; USP17L20; USP17L2; USP17L12; USP17L23; USP17L3; USP17L11; USP17L22; USP17L13; USP17L24; USP17L7; USP17L4; USP17L5; USP17L8	FST
detection of chemical stimulus involved in sensory perception (GO:0050907)	14/465	0.036	OR2A1; OR2F2; OR2F1; TAS2R38; OR9A4; OR2A12; OR2A7; OR2A5; TAS2R3; OR2A4; OR6B1; TAS2R5; OR2A25; OR2A2	ROH