The impact of identity by descent on fitness and disease in dogs.

Permalink
https://escholarship.org/uc/item/6jx667vq

Journal
Proceedings of the National Academy of Sciences of the United States of America, 118(16)

ISSN
0027-8424

Authors
Mooney, Jazlyn A
Yohannes, Abigail
Lohmueller, Kirk E

Publication Date
2021-04-01

DOI
10.1073/pnas.2019116118

Peer reviewed
The impact of identity by descent on fitness and disease in dogs

Jazlyn A. Mooneya, Abigail Yohannesb, and Kirk E. Lohmuellerb,dc,e

aDepartment of Human Genetics, University of California, Los Angeles, CA 90095; bDepartment of Biology, Stanford University, Stanford, CA 94305; cDepartment of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095; and dDepartment of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095

Edited by Elaine A. Ostrander, National Human Genome Research Institute, Bethesda, MD, and approved March 11, 2021 (received for review September 17, 2020)

Domestic dogs have experienced population bottlenecks, recent inbreeding, and strong artificial selection. These processes have simplified the genetic architecture of complex traits, allowed deleterious variation to persist, and increased both identity-by-descent (IBD) segments and runs of homozygosity (ROH). As such, dogs provide an excellent model for examining how these evolutionary processes influence disease. We assembled a dataset containing 4,414 breed dogs, 327 village dogs, and 280 wolves genotyped at 117,208 markers and data for clinical and morphological phenotypes. Breed dogs have an enrichment of IBD and ROH, relative to both village dogs and wolves, and we use these patterns to show that breed dogs have experienced differing severities of bottlenecks in their recent past. We then found that ROH burden is associated with phenotypes in breed dogs, such as lymphoma. We next test the prediction that breeds with greater ROH have more disease alleles reported in the Online Mendelian Inheritance in Animals (OMIA). Surprisingly, the number of causal variants identified correlates with the popularity of that breed rather than the ROH or IBD burden, suggesting an ascertainment bias in OMIA. Lastly, we use the distribution of ROH across the genome to identify genes with depletions of ROH as potential hotspots for inbreeding depression and find multiple exons where ROH are never observed. Our results suggest that inbreeding has played a large role in shaping genetic and phenotypic variation in dogs and that future work on understudied breeds may reveal new disease-causing variation.

Significance

Dogs and humans have coexisted together for thousands of years, but it was not until the Victorian Era that humans practiced selective breeding to produce the modern standards we see today. Strong artificial selection during the breed formation period has simplified the genetic architecture of complex traits and caused an enrichment of identity-by-descent (IBD) segments in the dog genome. This study demonstrates the value of IBD segments and utilizes them to infer the recent demography of canids, predict case-control status for complex traits, locate regions of the genome potentially linked to inbreeding depression, and to identify understudied breeds where there is potential to discover new disease-associated variants.

Author contributions: J.A.M. and K.E.L. designed research; J.A.M. and A.Y. performed research; J.A.M. and K.E.L. contributed new reagents/analytic tools; J.A.M. and A.Y. analyzed data; and J.A.M. and K.E.L. wrote the paper.

The authors declare no competing interest.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2019116118/-/DCSupplemental.

Published April 14, 2021.
Disease Traits Are Associated with ROH Burden. We hypothesize that the prevalence of ROH and IBD segments could be associated with recessive genetic disease in each breed (Fig. 1 and SI Appendix, Fig. S2). ROH form when an individual inherits the same genome within an ROH [odds-ratio (OR) $\beta = 0.238$] and Golden Retrievers ($\beta = 0.286$ and $P < 0.027$). For lymphoma in Golden Retrievers, case status is positively associated with the amount of the genome within an ROH [odds-ratio (OR) = 2.491, SI Appendix, Table S1], and on average cases carried more ROH than controls (SI Appendix, Fig. S6). Conversely, ROH appeared to show a no trait associations (6 observed associations versus 1.45 associations expected under the null, $P = 0.0027$, binomial test; Fig. 2).

We observed a significant association between the burden of ROH and case-control status for five traits: portosystemic vascular anomalies (PSVA) in Yorkshire Terriers ($\beta = -0.394$ and $P = 0.027$), lymphoma within both Labrador ($\beta = -0.604$ and $P < 0.0340$) and Golden Retrievers ($\beta = 0.913$ and $P < 0.001$), cranial cruciate ligament disease (CCLD) in Labrador Retrievers ($\beta = -0.403$ and $P < 0.003$), elbow dysplasia across all breeds ($\beta = 0.238$ and $P < 0.047$), and mast cell tumors across all breeds ($\beta = 0.286$ and $P < 0.027$). For lymphoma in Golden Retrievers, case status is positively associated with the amount of the genome within an ROH [odds-ratio (OR) = 2.491, SI Appendix, Table S1], and on average cases carried more ROH than controls (SI Appendix, Fig. S6). Conversely, ROH appeared to show a no trait associations (6 observed associations versus 1.45 associations expected under the null, $P = 0.0027$, binomial test; Fig. 2).

We observed a significant association between the burden of ROH and case-control status for five traits: portosystemic vascular anomalies (PSVA) in Yorkshire Terriers ($\beta = -0.394$ and $P = 0.027$), lymphoma within both Labrador ($\beta = -0.604$ and $P < 0.0340$) and Golden Retrievers ($\beta = 0.913$ and $P < 0.001$), cranial cruciate ligament disease (CCLD) in Labrador Retrievers ($\beta = -0.403$ and $P < 0.003$), elbow dysplasia across all breeds ($\beta = 0.238$ and $P < 0.047$), and mast cell tumors across all breeds ($\beta = 0.286$ and $P < 0.027$). For lymphoma in Golden Retrievers, case status is positively associated with the amount of the genome within an ROH [odds-ratio (OR) = 2.491, SI Appendix, Table S1], and on average cases carried more ROH than controls (SI Appendix, Fig. S6). Conversely, ROH appeared to show a
The impact of identity by descent on fitness and disease in dogs

Mooney et al.
PNAS | 3 of 7
The impact of identity by descent on fitness and disease in dogs

https://doi.org/10.1073/pnas.2019116118
IBD and ROH have had no causal variants identified through 2020. Thus, even if there are false-positive associations in OMIA, these false positives cannot explain the lack of OMIA variants in breeds that are less popular, have more ROH, and regions of the genome in IBD segments. Researchers should consider shifting their focus to some of these understudied breeds, as there may be more potential to discover new disease-associated variants. Such breeds include the Bearded Collie, Belgian Sheepdog, Bedlington Terrier, or Dogue de Bordeaux, all of which are prone to serious health conditions according to the AKC (akc.org/dog-breeds).

Ascertainment bias is not unique to OMIA and has been observed in human databases like Online Mendelian Inheritance in Man (OMIM) (37). In the case of human data, authors found that OMIM contains an enrichment of diseases caused by high-frequency recessive alleles because of the method through which these variants have been identified. Many variants were identified in isolated human populations, where there may be elevated levels of relatedness, which increases the probability of mapping higher-frequency deleterious variants (37). This ascertainment bias was not, and could not be discovered, in previous work on ROH in dogs (7), which focused on the more limited, but important, question of whether already identified disease variants were located in ROH.

**ROH Reveal Genes with Recessive Lethal Mutations.** Given the relatively high values of FROH observed for breed dogs, much of the genome should be in ROH in at least one of the 4,342 individuals in our study. We hypothesized that the genes not contained within an ROH in any individual or showing a deficit of ROH compared to the rest of the genome contain recessive lethal variants because individuals homozygous for these mutations are not viable (lethal path, Fig. 1). Across 4,342 dogs, we observed 27 genes (coordinates available on GitHub) where at least one exon does not overlap an ROH in any individual. To test whether this is unusual, we permuted the locations of the ROH within each individual and recounted the number of genes with an exon not containing an ROH. We found that if ROH were randomly distributed across the genome, we would expect to see ROH in all exons across genes (SI Appendix, Fig. S10). Thus, there are more genes not overlapping ROH than expected by chance ($P < 0.0001$), suggesting the presence of segregating recessive lethal mutations across breed dogs. This result is similar to what has been shown in inbred Scandinavian wolves (16), where ROH distribution fluctuates nonrandomly across the genome. Thus, these genomic regions could be lacking ROH because strongly deleterious recessive lethal mutations lurk as heterozygotes in the founders of the breed. If offspring become homozygous for these regions, they are no longer viable and are not sampled in our study (Fig. 1).

To test whether these 27 genes may have a functional effect, we intersected them with the 90th percentile constrained coding regions (CCRs) identified in human populations (38). CCRs were found to be enriched for disease-causing variants, especially in dominant Mendelian disorders (38). We expected that genes containing recessive lethal mutations would be conserved across species and asked whether the genes not overlapping ROH were enriched for CCRs. After resampling sets of 27 genes and intersecting them with the 90th percentile of CCRs 100,000 times, we would expect to see 18 of 27 genes falling above the 90th percentile of CCRs if exons lacking ROH in dogs were distributed randomly with respect to CCRs. In contrast, 23 out of 27 genes not overlapping ROH in dogs fall above the 90th percentile of CCRs if exons lacking ROH in dogs were distributed randomly with respect to CCRs. In contrast, 27 out of 27 genes not overlapping ROH in dogs fall above the 90th percentile of the CCR distribution ($P = 0.025$) (Fig. 4). Additionally, we observed a 2.94-fold enrichment of non-ROH genes relative to ROH genes in CCRs ($P = 0.041$, Fisher's exact test) (Fig. 4). Taken together, these results suggest that the genes with an exon not overlapping an ROH in dogs are enriched for exons devoid of variation in humans. Thus, these genes may be targets of strongly deleterious mutations affecting viability. Interestingly, there are some CCRs that do not have any known pathogenic or likely pathogenic variants, suggesting mutations in these exons could cause extreme developmental disorders or potentially be embryonically lethal (38). Studying mutations in these exons without ROH that overlap the CCR data could be fruitful both for identifying new disease phenotypes and for identifying variants with large fitness effects that could be linked to inbreeding depression or embryonic lethality.

We also tested whether our ROH analyses could be affected by low single nucleotide polymorphism (SNP) density, since the analyses thus far used only SNP genotype data. Because whole-genome sequence data would have an increased density of SNPs,

![Fig. 4. Histogram of the expected number of genes that fall into the top 10% CCRs over 100,000 randomly drawn sets of 27 genes. The empirical data are demarcated by the blue line ($P = 0.025$). The contingency table shows the count of genes classified as to whether all exons overlap an ROH ("ROH") and whether any exons overlap a CCR ("CCR"). There is a 2.94-fold enrichment of genes with at least one exon without an ROH ("non-ROH") genes in CCRs ($P = 0.041$) relative to genes where all exons overlap an ROH.](https://doi.org/10.1073/pnas.2019116118)
we repeated our analyses using two sets of sequence data. The first dataset represents samples from four different breeds of dog: Pug (n = 15) with ∼47× coverage (39), Labrador Retriever (n = 10) with ∼30× coverage (40), Tibetan Mastiff (n = 9) with ∼15× coverage (41), and Border Collie (n = 7) with ∼24× coverage (40). The second dataset was previously published (see ref. 16) and contains 220 samples from human populations. We find that of the 27 genes with at least one exon not overlapping the ROH in any dog, 3 genes ANKH, FYTTD1, and PRMT2 have exons not overlapping an ROH in any of the three data sets (SI Appendix, Table S3). One of these genes, ANKH, has known Mendelian phenotypes that have been reported in OMIM and is also a 95th percentile CCR (42). It should also be noted that these three genes ANKH, FYTTD1, and PRMT2 all reside toward the end of the chromosome in dogs (SI Appendix, Fig. S11). Nevertheless, the relative distribution of ROH and these three genes not containing ROH were concordant across both VCFTools and PLINK (SI Appendix, Fig. S11). We further examined the locations of exons devoid of ROH and observed that ROH tend not to occur at all within the genes, or ROH occur in the exons toward the end of the gene (SI Appendix, Fig. S12).

Conclusions

Here, we have shown how the population history of dogs has increased the number of regions of the genome carried in ROH and IBD segments, affecting phenotypes and fitness. Our genetic contributions to understanding the number of studies associating ROH burden with complex traits and directly shows this association in dogs (7, 43, 44). Our findings have implications for understanding the architecture of complex traits in other species, such as humans. Specifically, the fact that we find a relationship between ROH and certain phenotypes (Fig. 2), suggests that recessive mutations play a role in some traits. Much of the existing GWAS in humans has largely suggested that complex traits are highly polygenic with many additive effects (45–48). These differences across species likely reflect differences in genetic architecture driven by the demographic history of the populations combined with natural selection. Nevertheless, searching for recessive variants underlying complex traits in humans may be a fruitful avenue of research. Furthermore, variation in the amount of the genomes in ROH across human populations (8, 14, 17, 35) and the concordance of our results.

Computing IBD and ROH Scores. We computed each population’s IBD and ROH scores using an approach similar to that of Natafeka et al. (54). A population’s IBD score was calculated by computing the total length of all IBD segments between 4 and 20 cm and normalizing by the sample size. A population’s ROH score was computed using all ROH that passed quality control and normalizing by the sample size.

Association Test and Effect Size Estimates. For these analyses, we only used the subset of breed-dog data from Hayward et al. where we had phenotype information (10), allowing us to follow the same strict requirements for each clinical phenotype as originally done. For breed-specific analyses, we required there to be at least 10 cases and 10 controls per breed. For the “all breed” analyses, we include all individuals combined in the same analysis, rather than conducting a meta-analysis of all the results from the separate breed analyses. If there were more cases than controls, in either the breed-specific or all-breed analyses, the cases were down sampled to match the sample size of controls (reference SI Appendix, Table S1 for sample sizes).

We computed the association between ROH and each trait using a general linear mixed model, specifically a logistic mixed model, which is implemented in the R package GMMAT (53). Following the protocol from Hayward et al. (10), we did not include covariates in the association test and included the SNP-based or ROH-based kinship matrix as a random effect in the model to control for population stratification due to covariation of the amount of ROH per breed with the incidence of the phenotype in the breed. The P values were determined using a Wald test with a significance threshold of P = 0.05. Note, we use nominal P values in the manuscript and SI Appendix, these are not corrected for multiple testing (reference SI Appendix, SI Text for rationale). For more details on clinical-trait ascertainment see ref. 10. We generated the kinship matrix two different ways: 1) using the R package PRelate (56) on the SNP genotype matrix, and 2) by computing the total amount of the genome within an ROH that is shared between two individuals (SR).

\[
SR = \sum_{j=1}^{7} X_{Gj}\]

Here, \(X_{Gj} \in \{0, 1\}\) and equals 1 if the genotype (G) at the jth site falls within an ROH shared by both individuals and equals 0 otherwise. SR was computed for each pair of individuals and bounded between 0 (no sharing) and 1 (complete sharing) as follows:

Mooney et al.
The impact of identity by descent on fitness and disease in dogs

PNAS | 5 of 7

https://doi.org/10.1073/pnas.2019116118
To find the number of genes expected to lack ROH, we performed our permutation test 10,000 times to create a null distribution. The shuffled locations of ROH could occur at the ends of chromosomes. We repeated this randomization equivalent to the sum of all the autosome lengths. As such, the shuffled locations of ROH could occur at the ends of chromosomes. We repeated our permutation test 10,000 times to create a null distribution. The P value was computed as the proportion of permuted datasets with as many or more genes with an exon not overlapping an ROH relative to what was seen empirically (27 genes).

To examine the overlap between genes lacking ROH and CCRs from Havrilla et al. (38), we used BEDTools (57) to intersect non-ROH with the top 10% (90th percentile) of CCRs and exon ranges for CanFam3.1 (31), which came from Ensembl (58). Then, we tabulated the total number of genes where there was at least one exon where we never observed any overlap (including partial overlap) with an ROH (non-ROH genes) and the converse where there was at least one exon where we never observed any overlap with an ROH. For control edges across the chromosomes, we concatenated all 38 chromosomes into a single chromosome with a total length equivalent to the sum of all the autosome lengths. As such, the shuffled locations of ROH could occur at the ends of chromosomes. We repeated our permutation test 10,000 times to create a null distribution. The P value was computed as the proportion of permuted datasets with as many or more genes with an exon not overlapping an ROH relative to what was seen empirically (27 genes).

Identifying Depletions of ROH. To find the number of genes expected to lack at least one exon without an ROH, the ROH in each individual were permuted to a new location on the same chromosome using BEDTools shuffle (57). Next, we created a bed file containing the permuted ROH locations, intersected this file with the exon locations from CanFam3.1, and counted the number of genes with at least one exon where we did not observe any overlap with an ROH. To control for edge effects across the chromosomes, we concatenated all 38 chromosomes into a single chromosome with a total length equivalent to the sum of all the autosome lengths. As such, the shuffled locations of ROH could occur at the ends of chromosomes. We repeated our permutation test 10,000 times to create a null distribution. The P value was computed as the proportion of permuted datasets with as many or more genes with an exon not overlapping an ROH relative to what was seen empirically (27 genes).

\[ SR_{\text{bound}} = \frac{SR - \min(SR)}{\max(SR) - \min(SR)} \]

where the max amount of sharing in the equation above is the total length of the autosome and minimum is smallest SR value in base pairs. Results reported in the main text are for \( SR_{\text{bound}} \). We also compared these results to those when not using a kinship or ROH matrix (SI Appendix, Fig. S13).

Data Availability. SNP genotype data from the original projects (9, 10) are available on Dryad (https://datadryad.org/stash/dataset/doi:10.5061/dryad.g68k008 and https://datadryad.org/stash/dataset/doi:10.5061/dryad.266k44). Code used to process original data and generate intermediate files is available on GitHub (https://github.com/vaam92/DogProject_1ab) (39).

ACKNOWLEDGMENTS. We thank Eduardo Amorim, Arun Durvasula, Nelson Freimer, Malika Kumar-Freund, Jesse Garcia, Jacqueline Robinson, and Janet Sinheimer for helpful discussions about data analysis and curation as well as Bob Wayne and the reviewers for comments on the manuscript. This research was based upon work supported by the NSF Graduate Fellowship under Grant DGE-1650604 awarded to J.A.M. as well as NIH Grant R35GM119856 awarded to K.E.L.
46. T. A. Manolio et al., Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).
47. J. Yang et al., Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569 (2010).
48. E. A. Boyle, Y. I. Li, J. K. Pritchard, An expanded view of complex traits: From polygenic to omnigenic. Cell 160, 1177–1186 (2017).
49. C. C. Chang et al., Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015).
50. X. Zheng et al., A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328 (2012).
51. S. Ghirlanda, A. Acerbi, H. Herzog, J. A. Serpell, Fashion vs. function in cultural evolution: The case of dog breed popularity. PLoS One 8, e74770 (2013).
52. B. L. Browning, S. R. Browning, Detecting identity by descent and estimating genotype error rates in sequence data. Am. J. Hum. Genet. 93, 840–851 (2013).
53. A. Auton et al., Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res. 19, 795–803 (2009).
54. N. Nakatsuka et al., The promise of discovering population-specific disease-associated genes in South Asia. Nat. Genet. 49, 1403–1407 (2017).
55. H. Chen et al., Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am. J. Hum. Genet. 98, 653–666 (2016).
56. M. P. Conomos, A. P. Reiner, B. S. Weir, T. A. Thornton, Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. 98, 127–148 (2016).
57. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
58. F. Cunningham et al., Ensembl 2019. Nucleic Acids Res. 47, D745–D751 (2019).
59. J. Mooney, A. Yohannes, K. Lohmueller, The impact of identity-by-descent on fitness and disease in dogs. GitHub. https://github.com/jaam92/DogProject_Jaz. Deposited 6 January 2021.