UCLA

UCLA Electronic Theses and Dissertations

Title

Elucidating evolutionary processes in North American gray wolves: genetic subdivision, local adaptation, and coat coloration

Permalink

https://escholarship.org/uc/item/1kp8b0vf

Author

Schweizer, Rena Madeleine

Publication Date

2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Elucidating evolutionary processes in North American gray wolves: genetic subdivision, local adaptation, and coat coloration

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Biology

by

Rena Madeleine Schweizer

2015

© Copyright by

Rena Madeleine Schweizer

2015

ABSTRACT OF THE DISSERTATION

Elucidating evolutionary processes in North
American gray wolves: genetic subdivision, local
adaptation, and coat coloration

by

Rena Madeleine Schweizer
Doctor of Philosophy in Biology
University of California, Los Angeles, 2015
Professor Robert Wayne, Chair

A fundamental question in evolutionary biology concerns how organisms adapt to challenges in their environment and how genetic variation is acted upon by natural selection. Thus, the gray wolf (*Canis lupus*) is an excellent study species in this regard because coat color and morphological variation exists throughout its range, and a variety of genetic resources are available. In this doctoral dissertation, we explored three facets of evolution in North American gray wolves. First, we determined environmentally-related genetic subdivision and evidence for local adaptation through the use of 42K single nucleotide polymorphisms (SNPs) genotyped on a SNP array in 111 wolves from six ecotypes, and identified consistent signals of selection on genes related to morphology, coat coloration, metabolism, vision and hearing. Second, we designed a targeted capture of 1040 genes, including all exons and flanking regions, as well as

5000 1kb non-genic neutral regions and resequenced these regions in 107 wolves. Using selection tests, we identified potentially functional variants related to local adaptation. Finally, we focused on understanding positive selection at the K locus, a gene responsible for black coat color in wolves and domestic dogs. A previous study suggested that the melanistic K^B allele was introduced into the genome of North American wolves from the domestic dog via interbreeding, and then underwent positive selection. We designed a custom capture array to resequence five megabases surrounding the K locus core deletion in a larger sample of North American wolves from multiple areas to assess patterns of nucleotide and haplotype diversity, population-specific decay in linkage disequilibrium, and hierarchical patterns of genetic divergence among populations. From these data we infer that adaptive introgression most likely occurred first in the Northwest Territories or Yukon area of Canada, when native dogs and humans were co-existing in the Arctic. Furthermore, we find evidence for a strong, ongoing selective sweep in Yellowstone wolves that may be related to immunity and disease prevalence. These analyses set an important precedent for the use of cutting-edge genetic techniques to solve long-standing evolutionary questions about wild populations.

The dissertation of Rena Madeleine Schweizer is approved.

Eleazar Eskin

Kirk Lohmueller

Robert Wayne, Committee Chair

University of California, Los Angeles

2015

This dissertation is dedicated to my mom and dad,
who have inspired me to question, discover, explore,
joke, dance, and be passionate to my “level best.”

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge the tremendous support and guidance of my committee members, Eleazar Eskin, Kirk Lohmueller, Victoria Sork, and Robert Wayne. Bob, as my advisor, has been an extraordinary collaborator and mentor, and has continuously sustained my dissertation studies and challenged me to think harder and perform better. He also makes a great margarita! I look forward to future collaborations with Bob, Eleazar, Kirk, and Victoria, hopefully for the rest of my academic career.

Chapter One is a version of Schweizer, Rena M., vonHoldt, Bridgett M., Harrigan, Ryan, Knowles, James C., Musiani, Marco, Coltman, David, Novembre, John, Wayne, Robert K. (Submitted) Genetic subdivision and candidate genes under selection in North American gray wolves. *Molecular Ecology*. Author contributions were as follows: 1) Designed research: RMS, BVH, RH, JCK, MM, DC, JN, RKW; 2) Performed research: RMS, BVH, RH, JCK; 3) Contributed samples: MM, DC; 4) Analyzed data: RMS, BVH, RH, JN, RKW; 5) Wrote the paper: RMS, JCK, RKW. The authors are grateful to Torsten Günther and Graham Coop for advice on `Bayenv`, to Eunjung Han and Kirk Lohmueller for advice on Fst/XP-EHH, to Chris Darimont for insight into British Columbia wolves, and to John Pollinger for lab management and mentoring. We wish to thank H. Dean Cluff and Paul Paquet for critical sample collection and contribution, and Eleazar Eskin, Kirk Lohmueller, René Malenfant, and Victoria Sork, for reading earlier versions of this manuscript.

Chapter Two is a version of Schweizer, Rena M., Robinson, Jaqueline, Harrigan, Ryan, Silva, Pedro, Galaverni, Marco, Musiani, Marco, Green, Richard E., Novembre, John, Wayne, Robert K. (Submitted) Targeted capture and resequencing of 1040 genes reveal environmentally driven functional variation in gray wolves. *Molecular Ecology*. Author contributions were as

follows: 1) Designed research: RMS, JAR, PS, MG, RH, REG, JN, RKW; 2) Performed research: RMS, JAR, PS, MG, RH; 3) Contributed new reagents or analytical tools: REG, MM; 4) Analyzed data: RMS, RH, REG, JN, RKW; 5) Wrote the paper: RMS, RKW. The authors would like to thank Drs. Nikos Alachiotis, Chris Darimont, Brant Faircloth, Zhenxin Fan, Adam Freedman, Paul Gugger, Torsten Günther, Klaus-Peter Koepfli, Pavlos Pavlidis, and Clare Marsden for helpful discussions concerning methods implementation and data interpretation within this study, and Rachel Treadwell for assistance with lab work. Hamid Razzaghi kindly provided the use of a figure of *LIPG*. Comments from Eleazar Eskin, Kirk Lohmueller, and Victoria Sork on earlier versions of this manuscript were valuable.

My colleague, Sam Vohr, from UC Santa Cruz, was enormously helpful for my third chapter, and contributed custom scripts to calculate and plot EHH and haplotype structure. His dissertation advisor, Ed Green, also helped with data processing by providing feedback and some custom scripts. Olaf Thalmann, from University of Turku, provided initial array design and ideas for data analysis, from which my third chapter was developed.

Funding for this dissertation was provided as follows: National Science Foundation grants (DEB-1021397 and OPP-0733033) to JN and RKW, a NSF Graduate Research Fellowship to RMS, a UCLA Dissertation Year Fellowship to RMS, and a UCLA Edwin E. Pauley Fellowship to RMS.

I would like to acknowledge multiple family members, friends, and colleagues who made this dissertation possible: Francois and Linda, my parents, encouraged me from a young age to explore the world around me, and provided immense support as I pursued both my academic and non-academic passions; my three sisters, Briana, Maia, and Teia, are my dearest friends and provided me with plenty of encouragement, wisdom, and laughs; my partner, Andrew Wood,

who I am so grateful for, for making me laugh, pushing me to do better, convincing me there was a bright side when I had dissertation blues, and staying up late to help me with Python scripts when he could have been sleeping instead; Bridgett vonHoldt, for teaching and guiding me through grad school as both a colleague and a friend; my HLSP's Betsy Shimer Repogle and Romina Kim, with whom I shared many tables, coffees, and giggles at Kerhkoff; and, the multiple generations of Wayne labmates and other EEBers with whom I've shared many laughs, stimulating conversations, and late night study/dance sessions.

VITA

RENA MADELINE SCHWEIZER

EDUCATION

- B.S. in Biology, University of California, Los Angeles, June 2007
    - Departmental Honors in Ecology and Evolutionary Biology Department
    - Latin Honors conferred by College Honors Program

GRANTS, FELLOWSHIPS, AND AWARDS

- 2014-2015: UCLA Graduate Division Dissertation Year Fellowship ($20,000, plus tuition)
- 2014: NSF Research Experience for Undergraduates Supplement (Grant: "The genomic and ecological context of a major gene under selection in natural populations", PI: Robert Wayne, written by: Rena Schweizer) ($6,250)
- 2010-2013: NSF Graduate Research Fellowship ($128,000, plus tuition and fees)
- 2009 and 2010: UCLA Chancellor's Award ($10,000)
- 2008-9 and 2013-4: Edwin W. Pauley Fellowship ($40,000, plus tuition and fees)
- 2007: Departmental Honors in Department of Ecology and Evolutionary Biology
- 2007: Latin Honors, UCLA Honors Program, College of Letters and Science

PUBLICATIONS

- Lonsinger, RC, **Schweizer, RM**, Pollinger, JP, Wayne, RK, and Roemer, GW. (2015). Fine-scale genetic structure of the ringtail (*Bassariscus astutus*) in a sky island mountain range. *J of Mammalogy*. 96:257–268.
- Freedman, AH, Gronau, I, **Schweizer, RM**, *et al.* (2014). Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genetics*, 10:1-12.

- Thalmann, O, Shapiro, B, Cui, P, Schuenemann, VJ, Sawyer, SK, Greenfield, DL, Germonpré, MB, Sablin, MV, López-Giráldez, F, Napierala, H, Uerpmann, H-P, Loponte, DM, Acosta, AA, Giemsch, L, Schmitz, RW, Worthington, B, Buikstra, JE, Druzhkova, A, Graphodatsky, AS, Ovodov, ND, Wahlberg, N, Freedman, AH, **Schweizer, RM**, *et al*. (2013) Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. *Science,* 342:871-874.
- Fuller, T, Thomassen, HA, Peralvo, M, Buermann,W, Mila, B, Kieswetter, CM, Jarrın-V, P, Cameron, SE, Mason, E, **Schweizer, RM**, *et al*. (2013) Intraspecific morphological and genetic variation of common species predicts ranges of threatened ones. *Proceedings B*, 280:1-10.
- Thomassen, HA, Fuller, T, Buermann, W, Mila, B, Kieswetter, CM, Jarrın-V, P, Cameron, SE, Mason, E, **Schweizer, RM**, *et al*. (2011) Mapping evolutionary process: a multi-taxa approach to conservation prioritization. *Evol Applications*, 4:397–413.
- Slater, GJ, Thalmann, O, Leonard, JA, **Schweizer**, **RM**, *et al*. (2009) Evolutionary history of the Falklands wolf. *Current Bio*, 19: R937-8.
- **Schweizer**, **RM**, Roemer, G, Pollinger, J, Wayne, RK (2008) Characterization of 15 tetra-nucleotide microsatellite markers in the ringtail (*Bassariscus astutus*). *Mol Ecol Res*, 9:210-2.

CONFERENCES AND PRESENTATIONS

- Poster: Targeted capture and re-sequencing of candidate genes in North American gray wolves reveal environmentally driven functional variation, *Evolution*, June 2015
- Poster: "Genome-wide selection in North American wolves" *American Genetics Association*, "Genomics and Biodiversity", July 2011

Genetic subdivision and candidate genes under selection in North American gray wolves

ABSTRACT

Previous genetic studies of the highly mobile gray wolf (*Canis lupus*) found population structure that coincides with habitat and phenotype differences. We hypothesized that these ecologically distinct populations (ecotypes) should exhibit signatures of selection in genes related to morphology, coat color, and metabolism. To test these predictions, we quantified population structure related to habitat using a genotyping array to assess variation in 42,036 SNPs in 111 North American gray wolves. Using these SNP data and individual-level measurements of 12 environmental variables, we identified six ecotypes: West Forest, Boreal Forest, Arctic, High Arctic, British Columbia, and Atlantic Forest. Next, we explored signals of selection across these wolf ecotypes through the use of three complementary methods to detect selection: F_{ST}/XP-EHH bivariate percentile, `BayeScan`, and environmentally correlated directional selection with `Bayenv`. Across all methods, we found consistent signals of selection on genes related to morphology, coat coloration, metabolism, as predicted, as well as vision and hearing. In several high-ranking candidate genes, including *LEPR*, *TYR*, and *SLC14A2*, we found marked clines in allele frequencies that follow environmental changes in temperature and precipitation, a result that is consistent with local adaptation rather than genetic drift. Our findings show that local adaptation can occur despite gene flow in a highly mobile species and can be detected through a moderately dense genomic scan. These patterns of local adaptation revealed by SNP genotyping likely reflect high fidelity to natal habitats of dispersing wolves, strong ecological divergence

among habitats, and moderate levels of linkage in the wolf genome. This chapter includes a supplemental appendix, plus supplemental figures and tables.

Introduction

By targeting genomic regions distinctly marked by positive selection, genes that are functionally important to individual fitness in natural populations can potentially be identified (Nielsen *et al.* 2007). Of particular interest are genomic regions having markers whose allele frequency variation is related to ecological differences among populations (Dobzhansky 1948; (Hancock *et al.* 2008; Novembre & Rienzo 2009; Coop *et al.* 2010; Jones *et al.* 2012). However, allele frequencies are typically correlated between closely related populations due to shared population histories and gene flow, which potentially leads to elevated false positive rates (Coop *et al.* 2010). This problem can be circumvented in part by comparing multiple unlinked loci between populations since the effects of demography are genome-wide while selection is generally locus-specific (Nielsen 2005). Specific outlier loci can then be statistically identified and presumed to be in linkage disequilibrium (LD) with (aka "tag" loci) genes or other genomic features under selection. Further, the broader categories and patterns of genes under selection can be determined through gene ontology (GO) enrichment methods, in which the frequency of certain categories of genes are measured relative to a background expectation (Primmer *et al.* 2013). Measurements of genome-wide patterns of variation using large scale SNP genotyping arrays is a crucial first step towards establishing evidence of local adaptation and illuminating the specific, functional variants under selection in natural populations (e.g. Akey *et al.* 2002; Jones *et al.* 2012; Staubach *et al.* 2012; Pyhäjärvi *et al.* 2013). Consequently, we used a SNP genotyping array to explore evidence of local adaptation and identify genes under selection in a highly mobile carnivore, the gray wolf (*Canis lupus*).

In North American gray wolves, genetically distinct populations are observed which correspond to differences in ecological factors such as prey type and habitat; consequently, these populations have been considered “ecotypes” (Muñoz-Fuentes *et al.* 2009; Koblmüller *et al.* 2009; vonHoldt *et al.* 2011). Suggested reasons for this pattern included dispersal by individuals to habitats similar to their natal environment (natal homing) and the presence of discrete habitat and prey relationships (Geffen *et al.* 2004, Musiani *et al.* 2007). In coastal British Columbia, for example, wolves specialize on fish and small deer in near shore environments, tend to be smaller and more gracile than wolves elsewhere, and live in an extremely wet temperate rainforest (Darimont *et al.* 2003). Previous studies have demonstrated that these wolves are genetically and ecologically distinct, even from inland British Columbia wolves (Muñoz-Fuentes *et al.* 2009). In addition, genetically distinct Arctic wolves are migratory, rather than territorial like most wolves, and follow barren-ground caribou during migratory movements of >1000 kilometers across cold, relatively dry, open terrain (Mech & Boitani 2003; Musiani *et al.* 2007). Similarly, genetically distinct wolves of the western forests of North America take larger prey, such as elk and moose, in heavily forested and mountainous terrain (Mech & Boitani 2003). These findings suggest the potential for divergent natural selection and resulting patterns of local adaptation (Hancock *et al.* 2008; Mullen & Hoekstra 2008; de Jong *et al.* 2012; Pujolar *et al.* 2014). Specifically, genes influencing morphologic features related to diet such as dentition, skull robustness and shape, vision (e.g. for open vs. closed terrain, conditional upon latitude), locomotion (e.g. for migratory vs. territorial behavior), metabolism and thermal regulation would be predicted to diverge among ecotypes. Variation in morphology has been found among North American wolves (Jolicoeur

1959; Musiani *et al.* 2007; Muñoz-Fuentes *et al.* 2009; O'Keefe *et al.* 2013) and diversification of cranial form corresponds to differences in prey size (Slater *et al.* 2009). Coat color and pattern likewise varies with paler pelage more common in Northern regions (Gipson *et al.* 2002; Musiani *et al.* 2007; Anderson *et al.* 2009). These phenotypic differences suggest functional categories of candidate genes that may underlie local adaptation in ecologically distinct populations of wolves.

In this study, we genotyped 111 wolves from across Canada and Alaska for variation in 42 587 SNPs using Affymetrix v2 Canine SNP arrays. Our intent was to uncover population structure, and to identify genomic signals of selection and local adaptation in North American gray wolves. As the first step, we defined genetic units by quantifying population structure, isolation by distance and differentiation between subpopulations. To validate ecotype designations, we used a random forest model on high-resolution data collected on 12 environmental variables relating to temperature, precipitation, and vegetation. Next, we applied three approaches to identify SNPs showing signal of selection. First, we identified SNPs having outlier allele and haplotype frequencies between ecotypes using a composite statistic of F_{ST} and cross population extended haplotype homozygosity (XP-EHH) (Sabeti *et al.* 2007; vonHoldt *et al.* 2010). Second, we applied a model-based method (`BayeScan`) to identify SNPs that are significantly differentiated among populations, further suggesting diversifying selection (Foll & Gaggiotti 2008). `BayeScan` tests whether subpopulation-specific allele frequencies, measured by an F_{ST} coefficient, are significantly different from the allele frequency within the common gene pool. Third, we applied a Bayesian approach (*`Bayenv`*; Coop *et al.* 2010) to identify significant correlations between SNPs and environmental variables. We took advantage of moderate levels of linkage disequilibrium in gray wolves (Gray *et al.* 2009) to identify candidate genes as those that are within 10 kilobases (kb) of an outlier SNP. Using GO enrichment analysis and published functional data of specific candidate genes, we showed that selection may have acted on genes with morphological, phenotypic, and metabolic functions in relation to specific environmental variables. We also found significant genic SNPs and GO categories related to vision and hearing. Altogether, we demonstrated local adaptation in a highly mobile carnivore, and provided a set of >500 candidate genes for verification in a comprehensive resequencing study using a gene capture array (Schweizer *et al.*, submitted).

Methods

Sample selection and genotyping

The samples that we genotyped were selected from a set of gray wolves used in previous studies (Carmichael *et al.* 2007; Musiani *et al.* 2007) with additional tissue samples obtained from the University of Alaska Museum (Fairbanks, AK) and from P. Paquet (University of Victoria, Canada) to maximize geographic representation in northern Canada and Alaska. All samples were collected under permit grants to researchers at these institutions (Carmichael *et al.* 2007; Musiani *et al.* 2007; Muñoz-Fuentes *et al.* 2009). Forty-five samples were previously genotyped on the genome-wide Affymetrix v2 Canine SNP array (vonHoldt *et al.* 2010). We genotyped an additional 87 samples on the same SNP arrays following the manufacturer's protocol (Appendix 1-I, Supporting Information).

After array hybridization and scanning, genotypes were called using the MAGIC algorithm (Boyko *et al.* 2010) in reference to the dog genome (CanFam2; Lindblad-Toh *et al.*

2005). After filtering (Appendix 1-I, Supporting Information), a total of 42,587 SNPs were retained for analysis (henceforth referred to as 42K SNPs). Fourteen closely related individuals were identified and removed from further analyses, following methods described previously (vonHoldt *et al.* 2011), and we used the remaining 111 individuals for analysis. Because of the potential for LD biasing our results, we also generated a reduced data set of 22,084 SNPs that were not in high LD due to physical proximity (henceforth referred to as 22K LD-pruned; see Appendix 1-I, Supporting Information). This 22K LD-pruned dataset was used for population genetic analyses where indicated below, but the 42K set was used for selection tests. It is possible that the use of a dog SNP array in wolves may impose SNP ascertainment bias, especially in studies comparing dogs to wolves and other canine species (vonHoldt *et al.* 2010; vonHoldt *et al.* 2011). However, this bias is expected to be consistent within wolves, and the large number of varying SNPs within our samples, of which >98% were ascertained in the domestic dog (vonHoldt *et al.* 2010; vonHoldt *et al.* 2011), supports the use of this array for an intraspecific study.

Population structure

In order to determine the population structure within our samples, we first used STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003) to identify genetic clusters of individuals. Using the 111-individual 22K LD-pruned data set, we ran STRUCTURE v2.3.4 with 20 000 burn-in iterations followed by 50 000 sampling iterations for $K = 1$ through 10, assuming correlated allele frequencies under the admixture ancestry model. Each run was performed 10 times, and the ΔK statistic of Evanno *et al.* (2005) was calculated to help determine the most appropriate number of genetic clusters using Structure Harvester v0.6.93 (Earl & vonHoldt 2012). We used the *greedy* algorithm within CLUMPP v1.1.2 (Jakobsson & Rosenberg 2007) to account for variation in cluster labels across the 10 random iterations of STRUCTURE. Individuals that had > 50% assignment to a single genetic cluster were considered part of a population, and individuals with a lower percentage assignment were characterized as "admixed".

We performed a principal components analysis (PCA) (Patterson *et al.* 2006) using SMARTPCA within EIGENSTRAT v3.0 (Price *et al.* 2006) for the 111-individual 22K LD-pruned SNP data set. To measure the degree of genetic differentiation between clusters identified by non-admixed individuals (n=94), we used custom scripts to calculate Weir and Cockerham's (1984) θ, an estimator of Wright's (1951) F_{ST} (and henceforth referred to as F_{ST}), across all clusters and between each pair of clusters. Finally, to further visualize patterns of population structure, majority-rule neighbor-joining trees based on allele-sharing distances, which were calculated using PLINK (Purcell *et al.* 2007), were constructed using the package ape 3.1-2 in R (Paradis *et al.* 2004; http://www.R-project.org). Trees were generated using the 22K LD-pruned SNP set, with 1000 bootstraps, then visualized within the ape package.

Isolation by distance and spatial autocorrelation

To assess isolation by distance (IBD), Mantel tests were performed to compare pairwise geographic distances with genetic distance, D_{IBS}, calculated within PLINK for the 42K SNPs. The Mantel analysis was performed with the *vegan* v.2.0-10 package (Oksanen et al 2013) in R using 1000 permutations to test the correlation between genetic and Log_{10}–transformed geographic distance. Using the same data within GENALEX v6.501 (Peakall & Smouse 2006), we measured spatial autocorrelation within 130 even distance classes of 100km each. Significance was assessed using 9999 permutations, and the 95% confidence interval around the correlation *r* was determined using 9999 bootstraps.

Environmental layers and habitat classification

We acquired environmental characteristics for each individual using georeferenced environmental layer datasets (Hijmans *et al.* 2005) consisting of variation in annual means, extremes, and seasonal variation in temperature and precipitation, measured at 1-km spatial resolution. We used a set of 12 environmental variables: eight WorldClim variables were previously determined to maximize the variation within North America while minimizing correlation (Harrigan *et al.* 2014), and four satellite/radar variables were added after minimizing the Pearson correlation among a larger set, as in Harrigan, *et al.* (2014). The 12 variables measure temperature (annual mean temperature, mean diurnal temperature range, temperature seasonality, maximum temperature of warmest month, minimum temperature of coldest month), precipitation (annual precipitation, precipitation seasonality, precipitation of coldest quarter), vegetation (percent tree cover, normalized difference vegetation index, and land cover category), and altitude. This set of environmental variables includes those such as precipitation that have been demonstrated to significantly affect wolf population structure (Geffen *et al.* 2004) and morphology (O'Keefe *et al.* 2013).

To test whether our population groupings, as determined through genetic tests alone, were ecologically different and could be justified as unique ecotypes for downstream methods, we used a tree classification method called random forest (Breiman 2001) by utilizing the `randomForest` package (Liaw & Wiener 2002) in `R`. This test uses environmental data for each individual, in conjunction with our suggested population assignment based on genetic data, to test how well each individuals can be assigned to a group based on environmental data alone. The software uses a subset of individuals to train the model, and then attempts to assign "test" individuals to a group. Accuracy is measured by how often the model correctly assigns test individuals based on environmental data to the group specified for them. Assignment of individuals to populations using the 12 environmental variables had an accuracy of 82.98% from 50 000 trees. Accuracy was highest for British Columbia and Atlantic Forest populations, and most errors occurred when assigning individuals to the West Forest or Boreal Forest populations and to the Arctic or High Arctic populations. This difficulty in assigning individuals from these populations was also observed in STRUCTURE assignment tests (see *Results*). Based on the close correspondence of populations with unique environments, we subsequently classified them as "ecotypes".

Detection of ecotype-specific selective sweeps

In order to detect markers under selection within each ecotype, we grouped wolves based on STRUCTURE and Random Forest results. Only non-admixed wolves were analyzed (n = 94) so as to focus on detecting molecular evidence for local adaptation to specific habitats. We employed the joint F_{ST} and cross-population extended haplotype homozygosity test (XP-EHH; (Sabeti *et al.* 2007; vonHoldt *et al.* 2010), which has been used previously to identify selective sweep regions in multiple species (e.g. Sabeti *et al.* 2007; vonHoldt *et al.* 2010). The XP-EHH test uses the difference in haplotype length between two populations to identify regions that have undergone a hard selective sweep in one population if they show an extended haplotype in that population but not the other. The XP-EHH test requires a reference or ancestral population for the population being assessed. However, there is no straightforward ancestral population for each of the ecotypes identified here. Therefore, we compared each ecotype to its most closely related population, as determined by pairwise F_{ST}, and additionally to a pseudo-population consisting of all other ecotypes combined. Similar approaches have been previously applied (Yi *et al.* 2010; Carneiro *et al.* 2014). Consequently, we identified regions that diverged in each ecotype since splitting from the most recent ancestor, or regions that were specific to that ecotype in comparison to all other populations.

XP-EHH was calculated between each comparison pair mentioned above. This analysis requires data with known haplotype phase, so data were phased using `fastPHASE` software (Scheet & Stephens 2006) with subpopulations labeled according to their genetic population group (Appendix 1-I, Supporting Information). Using custom `R` scripts and previously developed methods (vonHoldt *et al.* 2010), we computed the empirical percentile for normalized F_{ST} and XP-EHH values associated with each SNP. A bivariate percentile score was calculated from the product of the F_{ST} and XP-EHH percentiles to obtain a single summary of the strength of the two signatures. If two or more SNPs were in the 95th percentile of the bivariate percentile score and were spaced <300kb apart, they were joined into a single cluster (vonHoldt *et al.* 2010). We ranked clusters by the number of SNPs they contained then by the bivariate percentile score of the central SNP. We selected the top 5% of empirical outlier clusters from each pairwise population, and then took the union of the two approaches (comparison to the population with the smallest F_{ST} and comparison to all other populations). Given that this might increase our rate of false positives, we also examined candidate sweep regions with a bivariate percentile score above 99.5th percentile as a more stringent test.

Model-based directional and balancing selection

To assess directional selection, we used the Bayesian method implemented in `BayeScan v2.1` (Foll & Gaggiotti 2008). `BayeScan` estimates selection by assigning a posterior probability (alpha) to a model in which selection explains a difference in allele frequencies better than a null model. A positive alpha indicates population-specific directional selection while a negative alpha suggests balancing or purifying selection. Given that `BayeScan` may suffer from elevated false positive rates under IBD and range expansion (Lotterhos & Whitlock 2014), and that balancing or purifying selection is especially prone to such issues (Lotterhos & Whitlock 2014), we focused on directional selection. Additionally, `BayeScan` was run using prior odds of 10, 1000, or 10 000 (Lotterhos & Whitlock 2014). Higher prior odds may reduce the false positive rate at the expense of identifying true loci under selection (Foll & Gaggiotti 2008). A false discovery rate (FDR) of 0.05 was used, with the caveat that although this reduces the number of false positives, true signals of selection may be missed (Foll & Gaggiotti 2008; Pilot *et al.* 2014).

Environmentally correlated selection

We used a Bayesian method (`Bayenv`) to identify allele frequencies that correlate with environmental variables (Coop et al 2010). In this approach, the empirical covariance in allele frequencies between geographically varying populations is initially estimated from a set of random markers (Hancock *et al.* 2008; 2010; Coop *et al.* 2010; Gunther & Coop 2013). Next, a Bayes Factor (BF) is assigned to each SNP of interest as a measure of how well the allele frequency of that SNP co-varies linearly with an environmental variable above the null model based on population structure alone.

In order to build a covariance matrix for the joint distribution of allele frequencies across populations, 10 000 SNPs were randomly chosen out of the full 42K SNP set after excluding SNPs that were out of Hardy-Weinberg equilibrium ($p<0.01$, exact test) using `PLINK`. These filters were applied to SNPs for the background covariance matrix as recommended by the authors (Coop *et al.* 2010), but all 42K SNPs were tested in the selection mode. Covariance matrices output by Bayenv after every 20 000th iteration were averaged over a total of 500 000 iterations. Following author recommendations (Coop *et al.* 2010), we compared the average correlation matrix (generated from the average covariance matrix with the *cov2cor* function in R) to our pairwise F_{ST} matrix for unusually high or low correlations, which might mean the MCMC model had not stabilized.

The selection mode of `Bayenv` was run separately for each SNP in the full 42K set with a total of 12 environmental variables and 100 000 iterations for each SNP. Each variable was normalized following author recommendations (Coop *et al.* 2010). `Bayenv` was run 10 times, since many MCMC sampling methods are sensitive to the initial conditions (Coop *et al.* 2010; Blair *et al.* 2014), and the final matrix of BFs was averaged over these 10 independent runs. For each of the 12 environmental variables, the empirical percentile of the $\log_{10}$ BF of each SNP was calculated. Both the top 5% and top 0.5% of outlier SNPs were candidates for further analysis (see below). For outlier SNPs, we plotted the mean value of the environmental variable within each population against the allele frequency for each population and calculated the Pearson correlation coefficient.

Candidate gene identification and gene ontology enrichment analysis

Using curated gene annotations from UCSC and Ensembl and accounting for different dog assembly versions (Freedman *et al.* 2014), we determined if there was any gene within 10 kb of each candidate SNP or sweep region. SNPs at this distance would likely be in LD with that gene (Gray *et al.* 2009). Across 42 587 SNPs on the array, we identified 26 108 SNPs (61%) that were within 10kb upstream or downstream of a gene.

Gene lists from each of the three selection tests were tested for significant enrichment of GO categories using `gProfiler` (Reimand *et al.* 2007; 2011). After correction for multiple testing using the Benjamini-Hochberg FDR, we examined significant categories ($p \leq 0.05$) with a minimum of two genes (Zhang *et al.* 2014). We also tested for an excess of genic SNPs among outliers using a one-sided conditional exact test (Agresti 2002) in `R`.

Phenotypic-genotypic association

For 33 of our wolf samples, we also had information on coat color phenotypes. Twenty-three of these wolves were sampled in a previous study on coat coloration and ecotype variation (Musiani *et al.* 2007) and were subsequently genotyped and included in this study. An additional 10 previously genotyped wolves from Yellowstone National Park (vonHoldt *et al.* 2011) that were not included in the analyses above because they represent translocated individuals, were included in the coat color analysis yielding 11 white, 11 black, and 11 gray (wild type) individuals.

In order to test for associations between SNPs near coat color genes and phenotypic variation within our samples, we performed a case/control association test using both the Fisher's exact test for allelic association (--fisher) and the full model testing for differences in any genotypes, with permutations for assigning significance (--model --cell 0 –perm) within PLINK. We implemented this for both white versus non-white coat color and black versus non-black coat color. For each significant SNP, we checked whether any known pigmentation gene was within 10 kb.

Results

Population structure and ecotypes

We observed notable population structure among our samples (Figure 1-1A-D). STRUCTURE runs showed the highest peak in Δ*K* values at K=3 and K=7 (Figure 1-S1). At K=3, there were distinct forest, arctic, and Atlantic groups (Figure 1-1C). Give the expansive geographic and subsequent environmental spread of these groupings (Figure 1-1A), we chose to examine higher values of K. Increasing values of up to K=6 appeared to separate different wolf ecotypes and confirm previous STRUCTURE groupings (Carmichael *et al.* 2007; vonHoldt *et al.* 2011). K=7 was not more informative with regard to geographic or habitat groupings and increasing K past 7 yielded no additional clusters to which more than three individuals were strongly assigned (i.e. ≥ 50%). We therefore used K=6 genetic clusters for subsequent analysis. The six clusters were geographically coherent (see Figure 1-1A), had high average assignment within each genetic cluster (84.5% ± 6.9%), and corresponded to specific habitats as found previously using microsatellite and SNP data: West Forest, Boreal Forest, Arctic, High Arctic, British Columbia, and Atlantic Forest (Carmichael *et al.* 2007; Muñoz-Fuentes *et al.* 2009; vonHoldt *et al.* 2011). After removal of two individuals whose STRUCTURE assignments showed they were migrants, we found that all six subpopulations, or ecotypes, were well circumscribed (Figure 1-1A).

Genetic differentiation of the 22K LD-pruned SNPs measured between all ecotypes was moderate, with global $F_{ST} = 0.09$. Pair-wise F_{ST} ranged from 0.0154 between Boreal Forest and West Forest ecotypes to 0.1128 between High Arctic and British Columbia ecotypes, with mean pair-wise $F_{ST} = 0.07$ (Figure 1-1B). The British Columbia ecotype appeared most distinct by this measure, showing high pairwise F_{ST} estimates with other ecotypes (Figure 1-1B).

There was high congruence between the STRUCTURE subpopulation assignments and their pattern of clustering by PCA. The same geographically coherent groups appeared clustered according to their scores on the first two axes, PC 1 and PC 2 (Figure 1-1D; Figure 1-S2). The first and second axes accounted for 4.2% and 3.8%, respectively, of the observed genetic variation. Within this PC space, admixed individuals were generally intermediate between the ecotypes for which their assignment was split in STRUCTURE analysis (Figure 1-1C-D). Results from neighbor-joining analyses generally supported structure and admixture population assignments with 100% support for all nodes in trees generated for the LD-pruned 22K SNP set (Figure 1-S3).

As described in the methods, we used a Random Forest approach to confirm our ecotypes and to identify the environmental variables most significant in distinguishing them (Figure 1-2). The ecotypes demonstrated extensive variation in annual precipitation, mean diurnal temperature range, elevation, and maximum temperature (Figure 1-2A), and the relative importance of each of the 12 environmental variables to distinguishing ecotypes can be visualized (Figure 1-2B).

Isolation by distance and spatial autocorrelation

We found a significant correlation between geographic distance and genetic distance, D_{IBS}, across the 111 individual 22K LD-pruned data set, there was ($r = 0.431$; Mantel test $P =$

0.001; Figure 1-3). This correlation was slightly weaker among wolves located more than ~300km apart ($r = 0.385$; Mantel test $P = 0.001$; Figure 1-3) and stronger among those separated by shorter geographic distances ($r = 0.456$; Mantel test $P = 0.001$) (Figure 1-3). Spatial autocorrelation analysis showed a significant positive spatial autocorrelation in distance classes from 0km to 2500km (Figure 1-S5). Between 2500km to 4100km, there was a significant slightly negative autocorrelation, and beyond 4100km the trend showed negative spatial autocorrelation, but without significance (Figure 1-S5).

Selective sweeps within ecotypes (*F_{ST}/XP-EHH*)

The numbers of candidate genes from the F_{ST}/XP-EHH selection scan outlier regions are provided in Table 1-S1 and the coordinates of the top 60 clusters for each ecotype comparison are provided, along with their size, F_{ST}/XP-EHH percentile and genes in Table 1-S2. Only British Columbia wolves showed a significant increase in the proportion of genic SNPs in the top 5% of outlier regions compared to the full data set (one-sided exact conditional test, 1 degree freedom, $P < 0.05$).

GO enrichment tests performed on each of these sets of genes in `gProfiler` identified several enriched categories (Table 1-S1, Table 1-S3, Table 1-S4). GO categories relating to skeletal morphology, vision, organismal system, metabolism, immunity, response to environment, and dentition were enriched in all ecotypes, although the specific GO categories for each ecotype were usually different, implying that slightly different sets of genes were enriched (See examples in Table 1-1, Table 1-S3, Table 1-S4).

Several high-ranking joint F_{ST}/XP-EHH selective sweep regions contained notable candidate genes (details in Table 1-S2 and Appendix 1-II). A top candidate gene for morphology within the West Forest ecotype was *NOTCH2* (*Notch (Drosophila) Homolog 2*), which included GO categories such as “positive regulator of the BMP signaling pathway” and “limb morphogenesis”. Within the West Forest wolves, the cluster containing *NOTCH2* contained three SNPs above the 95th percentile, including one SNP with a joint percentile of 99.9%, and was an outlier using both reference populations. A top candidate gene within the Boreal Forest ecotype was *GDF5* (*Growth Differentiation Factor 5*), which encodes a protein that is a member of the bone morphogenetic protein (BMP) family (Figure 1-4A) (Nie *et al*. 2006). Two SNPs within the cluster containing *GDF5* together ranked at the 99.5th percentile. One of the top sweep regions within the Arctic ecotype contained *GALNT5* (*Polypeptide N-Acetylgalactosaminyltransferase 5*). *GALNT5* is a member of a large family of genes involved in protein glycosylation (Bennett *et al.* 2012). The other gene within this sweep region was *ERMN* (*Ermin*), which functions in cellular development within the central nervous system (Brockschnieder 2006). Together these two genes were located in a region with two SNPs ranking at the 99.7th percentile. A high-ranking candidate region within High Arctic wolves contained a single gene, *KIT* (v-*kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog*), which is an essential cell-surface receptor in the melanogenesis pathway (Wehrle-Haller 2003). This sweep region contained a single SNP with a joint percentile of 96.1% (Figure 1-S6). A high-ranking sweep cluster within British Columbia wolves contained *WNT5A* (*Wingless-Type MMTV Integration Site Family, Member 5A*), a gene which plays a critical role in determining size during murine tooth development (Lin *et al.* 2011; Cai *et al.* 2011). *WNT5A* was the most suitable candidate gene within the sweep region, with the only other annotated gene being related to nerve cells (Figure 1-S7). The sweep region contained five SNPs, including one with a maximum joint percentile of 98.7%. Two selective sweep regions within the Atlantic Forest wolves contained multiple candidate genes related to vision, oncogenesis, and lipid metabolism. The highest ranking cluster (4 SNPs, max. percentile product: 97.9%) contained *PLEKHB1* (*Pleckstrin Homology Domain Containing, Family B Member 1*), which is involved in retinal development in mice (Wan *et al.* 2011), and *MRPL48* (*Mitochondrial Ribosome Protein L48*), which showed evidence in Antarctic icefish of gene duplications to increase mitochondrial function (Coppe *et al.* 2013). Additional genes are discussed in Appendix 1-II.

Population-specific directional selection

The `BayeScan` algorithm identified 77 SNPs with a value of alpha above the FDR cut-off of 0.05 using the default prior odds of 10 (Figure 1-5). Forty-four of these SNPs were within 10kb of an annotated gene (Table 1-S5). Of the 27 annotated genes near SNPs with a positive alpha (indicating diversifying selection), GO analysis identified a single significantly enriched category of auditory receptor cell differentiation, and KEGG pathways related to oxytocin signaling and cardiac muscle contraction (Table 1-2; Table 1-S6**)**. When we used prior odds of 1000, a single SNP near *ANXA10* (discussed below) was significant, and when we used prior odds of 10 000 there were no significant SNPs.

The top candidate gene for positive selection from `BayeScan` was *ANXA10* (*Annexin A10*), a protein coding gene for which the function is not yet known (Table 1-S5). The only significantly enriched GO category, "auditory receptor cell differentiation," (Table 1-S6) contained two interesting candidate genes. The first gene, *PCDH15* (*Protocadherin-related 15*) plays a crucial role in upkeep of normal cochlear and retinal function (Le Guédard *et al.* 2007). The second candidate gene within that GO category was *CUX1* (*Cut-like homeobox 1*), which plays a broad role in mammalian development via regulation of morphogenesis (Lizarraga *et al.* 2002; Sansregret & Nepveu 2008).

Correlation between SNPs and environmental variables

Our samples of North American wolves showed extensive variation in environment (Figure 1-2). Using `Bayenv` we found multiple significant outlier SNPs for each of the 12 environmental variables (Figure 1-6; Figure 1-S9). Across all 12 sets of outlier loci, a single vegetation variable (normalized difference vegetation index) showed a significant enrichment of genic SNPs in the top 5% (Fisher's exact test; $P = 0.0326$) (Table 1-S7). Nonetheless, there were several significantly enriched GO categories for each of the 12 environmental variables we examined relating to hearing, morphology, pigmentation, smell, and organismal system (Table 1-3, Table 1-S8, Table 1-S9). For example, morphological categories such as "anatomical structure development" and "anatomical structure morphogenesis" were enriched in 11 and 10, respectively, of the temperature, precipitation, vegetation, and elevation variables (Table 1-3). Organismal system categories involved in "calcium ion binding" and "locomotion" were enriched in the majority of environmental variables, whereas "blood circulation" was enriched with mean annual temperature and all of the vegetation and elevation variables. Two categories related to hearing were enriched as well (Table 1-3). Of particular interest were GO categories related to pigmentation that were significantly enriched with annual mean temperature and vegetation variables.

Several top-ranking SNPs from `Bayenv` were located near genes implicated in energy regulation, metabolism, and water balance, and show high correlation with environmental variables. *LEPR* (*Leptin Receptor*) is a receptor for the adipocyte-specific hormone leptin, and is involved in obesity (Chua *et al.* 1996) and cold tolerance (Hancock *et al.* 2008). A SNP located less than 1kb upstream of the start codon of *LEPR* ranked above the 99.9th percentile for land cover classification (BF=106.8), and above the 95th percentile for minimum temperature and precipitation of the coldest month (Figure 1-7). Located less than 1kb downstream of *LIPG* (*Endothelial Lipase*) was a SNP above the 99.9th percentile (BF=71.1) for temperature seasonality. *LIPG* regulates lipid levels, specifically levels of HDL (Edmondson et al 2009; Tietjen et al 2012). Finally, an intronic SNP in *SLC14A2* (*Solute Carrier Family 14, Urea Transporter, Member 2*) ranked above the 99.5th percentile in elevation (BF=144.1). *SLC14A2* plays a major role in water and salt balance through the urinary concentration mechanism (reviewed in Smith & Fenton 2007).

A number of top-ranking SNPs from `Bayenv` were also located near candidate genes implicated in the Bone Morphogenetic Protein (BMP) pathway regulation of skeletal and eye development. For example, an intronic SNP in *SMOC1* (*SPARC Related Modular Calcium Binding 1*) ranked above the 99th percentile (BF=2.44) for maximum temperature of the warmest month. *SMOC1* is a member of the matricellular protein family that is crucial for eye and limb development in both mice and humans and may help modulate the BMP signaling pathway (Okada *et al.* 2011). Several additional members of the BMP pathway, including *BMP1*, *BMP4*, *BMP6*, *BMP7*, *BMP10*, *BMPER*, and *GDF5* (reviewed in Bragdon *et al.* 2011), were in the top 95th percentile for environmental variables related to temperature, precipitation, and elevation (e.g. Figure 1-4B). The SNPs near *BMPER* and *BMP10* are notable since their BFs were relatively high (BF=2.19, 99th and BF=2.49, 98th, respectively) (Figure 1-7). Additional SNPs above the top 99th percentile tagged two *FGF* (*Fibroblast growth factor*) genes, which are implicated in craniofacial skeletal formation in humans, dogs, and mice (e.g. Hünemeier *et al.*

2013). The first gene, *FGF3*, was tagged by a downstream SNP that was an outlier for percentage tree cover (BF=3.17), and the second gene, *FGF14*, was tagged by an intronic SNP highly ranked for mean diurnal temperature range (BF=5.46).

Finally, SNPs near genes within the pigmentation pathway were outliers. *TYR* (*Tyrosinase*) encodes an enzyme crucial to the conversion of tyrosine to melanin (Beermann *et al.* 2004). A SNP located in the intron of *TYR* ranked above the 99th percentile for annual mean temp (BF=2.5) and precipitation seasonality (BF=4.07) (Figure 1-7). *TYRP1* (*Tyrosinase-related protein 1*) was tagged by a SNP ranked above the 99th percentile (BF=1.32) for vegetation. *ASIP* (*Agouti Signaling Protein*) contained a SNP ranked above the 98th percentile for temperature diurnal range (BF=1.44) and elevation (BF=2.5) (Figure 1-7). *OCA2* was tagged by a SNP 7kb downstream that was an outlier for mean diurnal temperature range (BF=5.3, 99.7th). Finally, the ligand and receptor pair, *KITLG* and *KIT*, both were near to high ranking SNPs above the 95th percentile: KIT was tagged by a SNP (BF=3.05, 99.2th) for precipitation seasonality, and KITLG was tagged by a SNP (BF=1.17; 97.8th) for percentage tree cover.

"Meta-analysis" of candidate genes

GO enrichment of all genes within the top 0.5% of outliers from either only `Bayenv` and F_{ST}/XP-EHH, or `Bayenv`, F_{ST}/XP-EHH, and `BayeScan` (FDR≤0.05), identified significant enrichment in GO categories of "anatomical structure development", "locomotion", "sensory perception", "regulation of cation channel activity", and several human phenotype categories related to abnormal morphological development, increased body weight, and hair color (Table 1-S10, Table 1-S11). At the 5% level of candidate gene significance, there were 276 significantly enriched GO categories, of which at least 21 related to morphology (e.g. "limb development"), four related to movement (e.g. “locomotion”), nine related to sensory perception and stimulus (including “eye development”), 34 related to channel or transporter activity (e.g. “ion channel activity”), and nine were related to muscle (e.g. “muscle tissue development”; Table 1-S12). Consequently, 77 of 276 GO categories (28%) could be viewed as consistent with our hypotheses for local adaptation

At the 0.5% level, there was overlap between each pair of tests (except for between `BayeScan` and F_{ST}/XP-EHH), but no overlap among all three tests (Figure 1-S10). However, at the 5% level, there was overlap between each pair of tests, and 14 genes overlapped among all three tests (*ANK2*, *AZIN1*, *BCAS1*, *BTN1A1*, *CACNA2D3*, *CCDC33*, *FOXK1*, *KSR2*, *LOC100685844*, *LOC100855656*, *LOC100855681*, *LOC100856364*, *LRRC16A*, *MPPED2*).

Out of interest, we also determined the level of overlap between our candidate genes and those from multiple independent studies either focusing on wolves (i.e. Pilot *et al.* 2014; Zhang *et al.* 2014) or focusing on geographic variation in humans in North America (i.e. Hancock *et al.* 2008) (Figure 1-8). We reasoned that overlap of our genes with candidates from other studies may strengthen our case for selection acting on these genes. At the 5% level, the gene *CACNA2D3* (calcium channel, voltage-dependent, alpha 2/delta subunit 3) was common to the three selection tests applied here and both Pilot, *et al.* (2014) and Zhang, *et al.* (2014). *CACNA2D3* is involved in voltage-gated calcium channel activity and six different SNPs near *CACNA2D3* ranked above the 95th percentile in temperature, precipitation, and vegetation. The gene *AZIN1* (*antizyme inhibitor 1*) also overlapped between this study and that of Zhang, *et al.* (2014). Antizymes catalyze a rate-limiting step in polyamine biosynthesis and are crucial to cell development (Coffino 2001). A SNP near *AZIN1* was a 95th percentile only in the precipitation seasonality variable. Eight genes overlapped between our top 5% of `Bayenv` results and those of Hancock, *et al.* (2008), who examined whether tag SNPs in genes common to metabolic disorders were candidates for selection in relation to environment.

Phenotypic-genotypic association

For white coat color, we identified a SNP within an intron of *MITF* (*P*-value: 0.009791), a modulator of melanocyte-related genes such as *KIT* and *KITLG* (Goding 2000) and the gene implicated in white spotting in many dog breeds (Karlsson *et al.* 2007; Schmutz *et al.* 2009; Vaysse *et al.* 2011). All other SNPs near coat color genes were not significant following permutations. For black coat color, the most significant SNP tagging a pigmentation gene was within the intron of *TYR* (p-value: 0.02754). Both of these genes also were tagged by SNPs above the 95th percentile for at least one environmental variable in `Bayenv`.

Discussion

Population structure and genetic differentiation across populations

Our analysis of population structure of North American gray wolves revealed six major clusters that were associated with unique habitats (Figures 1, 2). These results were concordant with previous large-scale studies in wolves using microsatellites or SNPs (Geffen *et al.* 2004; Carmichael *et al.* 2007; vonHoldt *et al.* 2011). However, in contrast to previous studies, we found that mainland tundra wolves were highly admixed and contained genetic components of both Boreal Forest and Arctic subpopulations (Figure 1-1A, 1C). Additionally, the PCA did not provide any evidence of a mainland tundra subpopulation or two separate Boreal Forest subpopulations, as found by Carmichael *et al.* (2007) (Figure 1-1D). Our wide geographic sampling and thousands of SNP markers allowed us to make more subtle observations of structure than previously could be achieved (Carmichael *et al.* 2007). We also confirmed previous studies finding that British Columbia wolves are genetically and ecologically distinct (Muñoz-Fuentes *et al.* 2009). Our results highlight the differentiation of the British Columbia ecotype, which was one of the first to appear in STRUCTURE analysis as a separate group at increasing K values (K=4), and also the population separated on PC1 (Figure 1-S2). Using data from 12 environmental variables shown to be important in discriminating North American habitats (Harrigan *et al.* 2014), we distinguished six environmentally distinct populations using a Random Forest classification method (Figure 1-2). Precipitation was the climate variable that most strongly associated with the differences among ecotypes, which agrees with a previous analysis based on microsatellite loci and mtDNA (Geffen *et al.* 2004). Mean diurnal temperature range and maximum temperature of warmest month were also significant, which was a novel finding here. Random forest models had lower accuracy when assigning individuals to either West forest or Boreal forest, and to High Arctic or Arctic, which paralleled the moderate level of admixture identified from genetic data alone (Figure 1-1). Newer methods that incorporate environmental and genetic data without assuming a linear relationship may help further tease apart environmental differences among wolf ecotypes, especially those that are related to threshold responses in organisms (e.g. Fitzpatrick & Keller 2014). In summary, through the use of population structure and environmental classification methods, we demonstrated that environmental influences dominate population structure in wolves, with weaker trends according to distance, as might be expected for a highly mobile species.

Candidate genes for morphology

Both GO (Table 1-1; Table 1-3) and candidate gene analyses (Figure 1-6; Figure 1-7) suggested that selection on morphological pathways has occurred in North American wolves, as we predicted. Several genes within the bone morphogenetic protein pathway are top candidates in F_{ST}/XP-EHH and Bayenv. We found that *GDF5, BMP7*, and *NOTCH2* were located in candidate selective sweep regions in Boreal Forest wolves, British Columbia wolves, and West Forest wolves, respectively. Mutations within *GDF5* are associated with skeletal developmental disorders (Bragdon *et al.* 2011), functioning *BMP7* is necessary for normal cartilage and eye development (Bragdon *et al.* 2011), and mouse knockout experiments have shown that *NOTCH2* is critical for proper chondrocyte and bone development (Kohn *et al.* 2012). In other pathways for skeletal mineralization or limb development, we found top clusters for *ALPL* in Arctic wolves, *WNT5A* in British Columbia, and *WNT5B* in Atlantic Forest using F_{ST}/XP-EHH (Figure 1-S7). Mouse knockout experiments have shown that *WNT5A* and *WNT5B* are critical for chondrocyte proliferation and tooth development (Lin *et al.* 2011; Cai *et al.* 2011). SNPs either within the introns of these genes or in close proximity were also significant outliers in our environmental analyses of selection with Bayenv. For example, one of the top candidate sweep regions within the Boreal Forest wolves contained multiple SNPs at the 99.5th percentile near *GDF5* (Figure 1-4A) and was highly correlated with annual mean temperature (Figure 1-4B). If climate is influencing the prey type and availability, then wolves in differing environments may have evolved different skull morphologies as a result. Genes that are critical for tooth development, for example, may be under selection in response to diets consisting of smaller prey such as deer or fish, rather than elk or moose, which may require special dental adaptations or cranial bite force (Slater *et al.* 2009).

Some candidate genes were identified uniquely in Bayenv or F_{ST}/XP-EHH. An example of the former is an intronic SNP in *BMPER* that was above the 99th percentile for two environmental variables (Figure 1-7), but was not identified as a candidate from our F_{ST}/XP-EHH analysis. *BMPER* is a BMP-binding protein that may modulate BMP activity and play a role in endothelial cell differentiation (Moser *et al.* 2003). The frequency of this SNP had high correlation with annual precipitation and minimum temperature of coldest month (Figure 1-7). If *BMPER* affects bone development in wolves, then the association of an intronic SNP with precipitation may reflect large-scale differences in skull morphology due to precipitation (O'Keefe *et al.* 2013). On the other hand, if *BMPER* affects endothelial cell development, then its association with temperature of coldest month may reflect adaptations to modulate blood vessels in colder or warmer conditions. Some of our top candidate genes, specifically those within the BMP pathway, are also strongly associated with tooth formation (Lin *et al.* 2011; Cai *et al.* 2011). In general, we found that SNPs located near or within genes that are fundamental to bone, skeletal, and muscle development were highly correlated with both precipitation and temperature variables. It follows that a recent study that revisited skull measurement data collected on almost 300 wolves from all over North America (O'Keefe *et al.* 2013) found distinct trends in morphological variation, with higher mean body size at higher latitudes, and identified precipitation as a key factor driving the variation in cranial morphology.

Candidate genes for coloration

In the GO analyses of Bayenv results, we observed significant enrichment of categories related to pigmentation, melanin biosynthetic process, and melanosome membrane for environmental measures of temperature and vegetation (Table 1-3, Table 1-S8, Table 1-S9). The lack of similar GO categories in the F_{ST}/XP-EHH analysis may indicate that within a single ecotype multiple candidate pigment genes may not be under selection such that a GO analysis would be of limited use. Alternatively, if pigmentation is a result of polygenic selection and genes have not undergone a classic selective sweep, then XP-EHH would be unlikely to detect selection. Using F_{ST}/XP-EHH we did identify a candidate sweep region within High Arctic wolves that contained a single SNP tagging a single pigmentation gene, *KIT* (Figure 1-S6). As mentioned previously, *KIT* is a key component in the melanogenesis pathway, and given the low frequency of black wolves in the High Arctic (Musiani *et al.* 2007), may be involved in the higher frequency of light coat color. The ligand of *KIT*, *KITLG*, was also a top outlier, and its expression is correlated with the localization and migration of melanocytes (Wehrle-Haller 2003). In addition to *KIT* and *KITLG*, we also found that SNPs within the top 95th percentile of several environmental variables tagged other genes within the pigmentation pathway, including *TYR*, *TYRP1*, *ASIP*, *MYO5A*, and *OCA2*. Several of these genes have been associated with color polymorphisms within wild vertebrate populations (reviewed in Hubbard *et al.* 2010). For *ASIP*, *KITLG*, *OCA2*, and *TYR*, we observed relatively strong correlations with environmental variables (Figure 1-7). Within the pathway by which melanin pigment is produced, tyrosinase is the rate-limiting enzyme, and several mutations within *TYR*, the gene encoding tyrosinase, have been identified causing coat colors in mice along the spectrum of fully pigmented to albino (reviewed in Beermann *et al.* 2004). We suspect that similar mechanisms may occur in wolves, especially since we found significant association of a SNP in *MITF* and a second SNP in *TYR* with white and black coat color, respectively. The high number of pigmentation candidate genes warrants further study, perhaps through resequencing to identify functional variants, or measuring gene expression differences in wolves of known phenotype (e.g. Hoekstra *et al.* 2006; Linnen *et al.* 2013).

Candidate genes for metabolism, vision, and hearing

We identified high-ranking SNPs tagging genes that may affect metabolic and osmoregulatory performance, as well. A SNP located less than 1kb upstream of *LEPR* that was above the 95th percentile in vegetation, temperature, and precipitation variables. *LEPR* has been implicated in cold tolerance and cold adaptation, and here, the SNP tagging *LEPR* had a high correlation with the minimum temperature of the coldest month (Figure 1-7). An extremely high-ranking SNP also occurred upstream of *LIPG*, a gene that regulates lipid levels and in which loss-of-function mutations lead to increased levels of HDL (Edmondson *et al*. 2009; Tietjen *et al*. 2012). Wolves in especially cold environments may have evolved an increased ability to cope with cold stress by regulating fat metabolism via *LEPR* or *LIPG*. For example, pikas show a significant increase in the rate of non-synonymous substitutions in *LEPR* with lower temperatures (Yang *et al.* 2008), and studies in mice show that *LIPG* may aid in uptake to adipose tissue of free fatty acids (Kratky *et al*. 2003).

We also predicted that a second pathway by which North American wolves may adapt to environmental challenges such as heat or water stress is through osmoregulation. Indeed, two top-ranking tag SNPs from the environment association analysis of `Bayenv` were located within introns of *SLC14A2*, a key member of the urine concentrating mechanism in mammals (Smith & Fenton 2007). Previous research on cetaceans identified *SLC14A2* as a candidate gene for water and salt balance (Xu *et al.* 2013). GO category enrichment of genes involved in ion channel transport were some of the most prevalent in our results from Bayenv, accounting for 5% of categories (Table 1-3, Table 1-S8, Table 1-S9).

We found multiple genes and GO categories related to vision and hearing. One gene identified as a candidate in BayeScan and Bayenv was *PCDH15*, a member of the cadherin family of proteins that is highly expressed in the retina and cochlea (Alagramam *et al.* 2001). Mutations in this gene have been implicated in Usher Syndrome type 1F, a disease causing deafness (Le Guédard *et al.* 2007). BayeScan is subject to higher false positive rates under certain demographic scenarios (Lotterhos & Whitlock 2014), so it is possible that the SNP near *PCDH15* identified by BayeScan was a false positive. However, *PCDH15* was also an outlier in Bayenv for temperature, vegetation, and elevation variables, which strengthens the case for it being a true positive and demonstrates the utility of multiple selection tests with differing underlying models and assumptions. We also found other candidate genes for eye development and hearing, and several related significantly enriched GO categories from the union of all three selection tests (Table 1-S12), and in the Bayenv analysis (Table 1-3, Table 1-S8, Table 1-S9). Wolves inhabit a variety of terrain from open tundra habitats with strong seasonality in light to closed habitat temperate rainforests with more uniform light conditions. Such differences may exert divergent selection pressures on vision and hearing.

Comparison to previous wolf and vertebrate studies

To our knowledge, this study is the first large-scale genetic analysis of local adaptation in a non-human vertebrate across a substantial range of habitats. We found that precipitation and mean diurnal temperature range were some of the most influential environmental variables associated with SNP variation across the North American range of gray wolves (Figure 1-2).

This result is concordant with previous genetic analysis using microsatellite loci and mtDNA sequence variation suggesting that vegetation (Geffen *et al.* 2004) and habitat type (Carmichael *et al.* 2007; Muñoz-Fuentes *et al.* 2009) are the main drivers of wolf ecotype differentiation. Precipitation is also a significant correlate of morphological variation in wolves (O'Keefe *et al.* 2013). Consequently, local adaption in wolf ecotypes appears driven by strong environmental gradients, primarily in temperature and precipitation.

Our study provides an advance over previous research by identifying candidate genes in the context of environmental differences among genetically defined ecotypes. Notably, we confirm candidate genes that were outliers in sequencing and SNP genotypes studies of Old World gray wolves suggesting environmental difference may be driving local adaptation there as well. For example, at the 95th percentile cutoff, we observed 173 genes overlapping with a genome sequencing study on high altitude adaptation in Tibetan wolves (Zhang *et al.* 2014) and 14 genes overlapping with a SNP array-based study of demography and outlier SNPs tagging candidate genes in European wolves (Pilot *et al.* 2014) (Figure 1-8). The two genes common to all three sets, *CACNA2D3* and *AZIN1* are candidates for hypoxia in Zhang, *et al.* (2013). We speculate that *CACNA2D3* and *AZIN1* may serve this function in New and Old World wolves given wolf persistence at high and low altitude habitats (Figure 1-2A).

Functional interpretation of candidate genes under selection in our study was facilitated by a wide array of preexisting studies on pigmentation, disease, and other phenotypes in a variety of species (humans: reviewed in Sturm and Duffy 2012; lab mice: reviewed in Barsh 1996; *Peromyscus* mice: Manceau *et al*. 2011; sheep: Fariello, *et al*. 2014; cattle: Qanbari *et al*. 2014; Arctic skuas: Janssen *et al*. 2013). For example, eight of our candidate genes at the 95th percentile significance overlapped with environmentally correlated genes influencing the "metabolic syndrome" in humans (Hancock *et al.* 2008) (Figure 1-8). Interestingly, Hancock and colleagues chose to investigate these genes for their involvement in dyslipidemia (*CLOCK* and *PON1*), obesity (*LEPR*, *PPARGC1A*), hypertension (*EPHX2*), type II diabetes (*TCF7L2*), and a "metabolic syndrome" phenotype (*PTK2B*, *SCARB2*; see Hancock, *et al.* 2008 for details). The commonality with our study suggests the possibility of a general adaptation toolkit for environmental gradients, such as the *LEPR* gene for cold tolerance, which has also been implicated in cold tolerance and adipose tissue in Neanderthals and Denisovans (Sazzini *et al.* 2014), pikas (Yang *et al.* 2008), and mice (Chua *et al.* 1996). Similarly, we identified common mechanisms of pigmentation and morphology, especially major pathways of bone development such as BMP or WNT. Whereas in humans these genes have been implicated in diseases, selection on these genes in wolves may be a thermoregulatory response to large fluctuations in temperature, osmoregulatory response to differential water availability, or metabolic responses to varying diet and represent local adaptations resulting from divergent natural selection.

Our approach for identifying genes involved in adaptation was necessarily correlative and will require further study to confirm whether these candidate genes influence function or are false positives (Barrett & Hoekstra 2011). To determine if tag SNPs are actually associated with potential functional mutations in candidate genes, and if those mutations show evidence of selection (e.g. Domingues *et al.* 2012), new capture array approaches can be used to simultaneously capture exons from thousands of genes followed by high throughput sequencing (Hodges *et al.* 2007). Such verified candidate genes can then be subject to further functional inference or knockout studies to confirm function (e.g. Lewandoski 2001; Storz 2007; Kingsley *et al.* 2009; Manceau *et al.* 2011). Nonetheless, the validity of our approach is suggested by previous studies. For example, genes underlying traits shown to be under selection in humans, such as pigmentation, lactase tolerance, and hearing were initially identified as candidates using SNP genotyping (as we have done), and were verified with finer-scale resequencing (reviewed in Akey 2009). Genic SNPs with allele frequencies that follow environmental clines are especially convincing candidates for adaptation.

Two of the methods we used to infer selection (`Bayenv` and `BayeScan`) explicitly control for background demographic patterns, and we conservatively selected the very top few percent of outliers from F_{ST}/XP-EHH, which does not explicitly control for demography. However, future work incorporating empirically determined demographic models into selection scans and resequencing as discussed above may further clarify the level of false positives (e.g. Freedman *et al.*, in review). Furthermore, resequencing data, which is free from any ascertainment bias and which will more accurately describe variation within populations, is therefore a more sensitive approach to exploring selection. In fact, to further test our conclusions, we have resequenced exons and UTRs for over 1000 candidates genes from over 100 wolves from a similar geographic distribution, as well as 5Mb of non-genic "neutral" sequence. Extensive analyses of these data are described in a companion paper (Schweizer, *et al.*, submitted). We maintain that this general approach – first, using a genome-wide SNP array to identity candidate genes through the use of multiple statistical approaches including environmental data, and second, resequencing candidate genes by genome capture – provides an efficient paradigm for documenting and understanding local adaptation in a wide variety of non-model species.

Conclusions

Using a SNP genotyping array, we provided evidence for genetic subdivision in North American wolves that corresponds to distinct habitats, and consider these populations as ecotypes between which divergent natural selection may cause local adaptation despite gene flow. We demonstrated the utility of using multiple selection tests to build an extensive set of candidate genes that may have undergone selection among ecotypes and identify candidate genes for morphology, pigmentation, metabolism, vision and hearing in wolves. Many of these candidate genes also show evidence of local adaptation in Old World wolves and other species. These genes may define a genetic toolkit used by a wide variety of taxa to address climate and environmental variation as well as biotic factors such as food type availability. Our findings demonstrate that despite high mobility we can detect evidence of local adaptation through a moderately dense genomic scan. This result likely derives from high fidelity to natal habitats of dispersing wolves, strong ecological divergence among habitats, and relatively high levels of linkage in the wolf genome.

Appendix 1-I: Methods

Sample selection and genotyping

DNA was extracted using a QIAamp DNA mini kit (QIAGEN) following standard protocol and quantified using a Nanodrop 1000 (Thermo Scientific, Wilmington, Denver). Samples were prepared and genotyped using the Affymetrix "GeneChip® Mapping 500K Assay" protocol. Before the hybridization step, sample volumes were reduced to 35μL by heated evaporation at 30° C in order to allow the entire volume of each sample to be hybridized to a single array (vonHoldt *et al.* 2010; 2011).

A total of 61,585 SNPs were successfully called across 125 genotyped individuals, after seven individuals were removed from the data set due to poor call rates. After removing X-chromosome SNPs (n = 521), monomorphic SNPs, SNPs with a minor allele frequency <0.01, and SNPs with less than 95% call rate, a total of 42,587 SNPs were retained for analysis (henceforth referred to as 42K SNPs). To evaluate consistency of protocol and genotyping calls, five samples were fully processed and genotyped in duplicate. Called genotypes across duplicated samples differed at < 1.2% of SNP loci.

Because of the potential for linkage disequilibrium biasing our results, sliding windows of 50 SNPs within each chromosome were evaluated for high correlation ($r^2 \geq 0.2$) with `PLINK` v.1.06 (Purcell *et al.* 2007) using the complete sample across populations. If any pair of SNPs in any 50-SNP window was observed to have $r^2 \geq 0.2$, one SNP was randomly removed by `PLINK`, and the process was repeated in 5-SNP steps, as in vonHoldt *et al*. (2011). This pruning yielded a reduced data set of 22,084 SNPs (henceforth referred to as 22K LD-pruned) that are not in high LD due to physical proximity. This 22K LD-pruned dataset was used for population genetic analyses where indicated below, but the full 42K set was used for selection tests.

Phasing data with `fastPHASE`

The 111-wolf 42K SNP data set was run through `fastPHASE`, and preliminary runs using subsets of data indicated that no decrease in imputed genotype error rate was found by using > 20 haplotypic groups (results not shown; (Scheet & Stephens 2006)). Thus, phasing was performed with 20 haplotypic groups, with 20 random starts of the Expectation-Maximization algorithm, each running for 50 iterations. A cross validation procedure with 2000 SNPs and a 5% masking rate was applied 50 times for each K value. SNPs with complex ascertainments (n =

808; (vonHoldt *et al.* 2010) were removed from the phased data, and phased chromosomes were used as input for `XP-EHH`. Default parameters of `XP-EHH` were modified to allow for spacing of up to 1Mb between SNPs for calculating extended haplotype homozygosity (EHH) and up to 4Mb between SNPs for calculating integrated haplotype homozygosity (iHH) (vonHoldt et al. 2011).

Appendix 1-II: Results

Selective sweeps within ecotypes (*F_{ST}/XP-EHH)*

GO categories related to skeletal morphology were enriched in all ecotypes other than the High Arctic, whereas learning-related categories were only enriched in the High Arctic wolves. Metabolism categories were enriched in all but Arctic wolves. Categories related to musculature were only enriched in Arctic and British Columbia wolves. Two enriched KEGG pathways in British Columbia wolves were salivary secretion and arachidonic acid metabolism. These pathways contained five genes (out of 112 total at 0.5% significance) in common with our candidates.

A top candidate gene for morphology within the West Forest ecotype is *NOTCH2* (*Notch (Drosophila) Homolog 2*), which includes GO categories such as “positive regulator of the BMP signaling pathway” and “limb morphogenesis”. In humans, mutations within *NOTCH2* associate with Hajdu-Cheney syndrome, which is characterized by osteoporosis, facial anomaly, and premature loss of teeth (Isidor *et al.* 2011). Transgenic experiments within mice demonstrate the importance of proper *NOTCH2* function for chondrocyte and bone development (Kohn *et al.* 2012). Within the West Forest wolves, the cluster containing *NOTCH2* contains three SNPs above the 95th percentile, including one SNP with a joint percentile of 99.9%, and is an outlier using both reference populations. The function of other genes within the same cluster are either uncharacterized or related to vesicle trafficking.

A top candidate gene within the Boreal Forest ecotype was *GDF5* (*Growth Differentiation Factor 5*), which encodes a protein that is a member of the bone morphogenetic protein (BMP) family (Figure 4A). Proteins within the BMP pathway regulate cell growth and differentiation, and mutations within this gene are associated with several disorders related to skeletal development (reviewed in (Bragdon *et al.* 2011)). Two SNPs within the cluster containing *GDF5* together rank at the 99.5th percentile for joint F_{ST} and XPEHH. Another gene within this cluster is *LOC485853*, which has sequence similarity to *Otoferlin*. Mutations in *Otoferlin* have been associated with deafness in humans (Yasunaga *et al.* 1999), and cause deafness in mice models (Roux *et al.* 2006). A second interesting candidate gene in the Boreal Forest wolves is *MFAP2* (*Microfibrillar-associated protein*), which has been implicated in decreased ability to thermoregulate in mouse knockout mice (Craft *et al.* 2014). The sweep region containing this gene includes three SNPs and a joint Fst/XP-EHH percentile of 99.9%. Finally, Boreal Forest wolves also seem to have undergone a selective sweep in a candidate region containing *NEDD4L* (*Neural Precursor Cell Expressed, Developmentally Down-Regulated 4-Like E3 Ubiquitin Protein Ligase*). *NEDD4L* is the only gene within the sweep region (joint percentile 99.5%, 1 SNP), and is characterized by GO categories such as “response to salt stress” and “channel activity.” In humans, polymorphisms within *NEDD4L* are associated with increased salt sensitivity (Dahlberg *et al.* 2007).

One of the top sweep regions within the Arctic ecotype contained *GALNT5* (*Polypeptide N-Acetylgalactosaminyltransferase 5*). *GALNT5* is a member of a large family of genes involved in protein glycosylation (Bennett *et al.* 2012). The other gene within this sweep region is *ERMN* (*Ermin*), which functions in cellular development within the central nervous system (Brockschnieder 2006). Together these two genes are located in a region with two SNPs ranking at the 99.7th percentile. Another candidate sweep region, although lower in ranking, contains *ALPL* (*Alkaline Phosphatase, Liver/Bone/Kidney*), a gene implicated in skeletal development; in mice, for example, individuals without a functioning *ALPL* display abnormal craniofacial bone development starting as early as two weeks into development (Liu *et al.* 2014). Since lipid homeostasis may be crucial to wolves living in very cold environments, it is intriguing that another sweep candidate region within Arctic wolves contains the gene *ASXL1* (*Additional Sex Combs Like 1 (Drosophila)*). The *ASXL* family of genes is implicated in lipid homeostasis in humans and *ASXL1* inhibits apidogenesis in mice (Cristancho & Lazar 2011; Park *et al.* 2014).

The frequency of light coat color increases along a north-south gradient in wolf populations, with wolves of the High Arctic being predominantly white (Jolicoeur 1959; Musiani *et al.* 2007; Anderson *et al.* 2009). A high-ranking selective sweep candidate region within High Arctic wolves contains a single gene, *KIT* (v-*kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog*), which is an essential cell-surface receptor in the melanogenesis pathway. This sweep region contains a single SNP with a joint percentile of ~96% (Supplemental Figure 6). Two additional candidate regions in the High Arctic wolves contain genes that are important for vision, a function that may be heightened in response to lower levels of light during winter. A functioning version of *MAB21L1* (*Mab-21-Like 1*) is essential for lens placode development in mice (Yamada 2003) and a genetic variant within *CTNND2* (Catenin [Cadherin-Associated Protein], Delta 2) is strongly associated with high myopia in humans (Liu & Zhang 2014).

The top-ranking cluster within the British Columbia ecotype was identified both by using the West Forest ecotype as reference and by using all non-British Columbia ecotypes as reference. This region contains six SNPs with a maximum joint percentile of 99.7th. The region contains five genes, of which two are annotated. The first candidate is *LOC609648* (*similar to Protocadherin 8 isoform 2*), and may play a critical role in long-term memory (Frank & Kemler 2002). The second gene is *OLFM4* (*olfactomedin 4*), which encodes a glycoprotein expressed in the colon in humans that has been associated with childhood obesity in genome-wide association studies (Jonathan P Bradfield *et al.* 2012). British Columbia wolves, which are characterized by a small body and diet composed mainly of fish and deer, also show selective sweep regions containing genes related to morphology and dentition. A high-ranking sweep cluster contains *WNT5A* (*Wingless-Type MMTV Integration Site Family, Member 5A*), a gene which plays a critical role in determining tooth size during murine tooth development (Lin *et al.* 2011; Cai *et al.* 2011). *WNT5A* is the most suitable candidate gene within the sweep region, with the only other annotated gene being related to nerve cells (Supplemental Figure 7). The sweep region contains five SNPs, including one with a maximum joint percentile of 98.7%. Another promising candidate gene within British Columbia wolves is *BMP7* (*Bone Morphogenetic Protein 7*), one of several genes within the BMP pathway involved in bone growth. *BMP7* is thought to function in skeletal and eye development in humans, and mice with *BMP7* knockout mutations exhibit abnormal cartilage and eye morphology (reviewed in (Bragdon *et al.* 2011)). The region in British Columbia wolves that contains *BMP7* contains a single SNP. Measurements of British Columbia wolf skulls show that these wolves have shorter upper carnassial teeth than other wolves sampled (O'Keefe *et al.* 2013).

Two selective sweep regions within the Atlantic Forest wolves contain multiple candidate genes related to vision, oncogenesis, and lipid metabolism. The highest ranking cluster (4 SNPs, max. percentile product: 97.9%) contains *PLEKHB1* (*Pleckstrin Homology Domain Containing, Family B Member 1*), which is involved in retinal development in mice (Wan *et al.* 2011) A second gene within the same cluster, *MRPL48* (*Mitochondrial Ribosome Protein L48*), shows evidence in Antarctic icefish of gene duplications to increase mitochondrial function (Coppe *et al.* 2013). Another sweep cluster contains the following alluring candidates: *WNT5B* (*Wingless-type MMTB Integration Site Family, Member 5B*), a member of a well-known set of signaling proteins that coordinate chondrocyte development (Yang 2003); *ADIPOR2* (*Adiponectin Receptor 2*), a gene encoding a receptor of adiponectin, a key hormone involved in lipid metabolism (Yamauchi *et al.* 2007); and, *CACNA2D4* (*Calcium Channel, Voltage-Dependent, Alpha 2/Delta Subunit 4*), a member of the voltage-dependent calcium channel complex in which mutations have been identified that cause cone dystrophy (Wycisk *et al.* 2006).

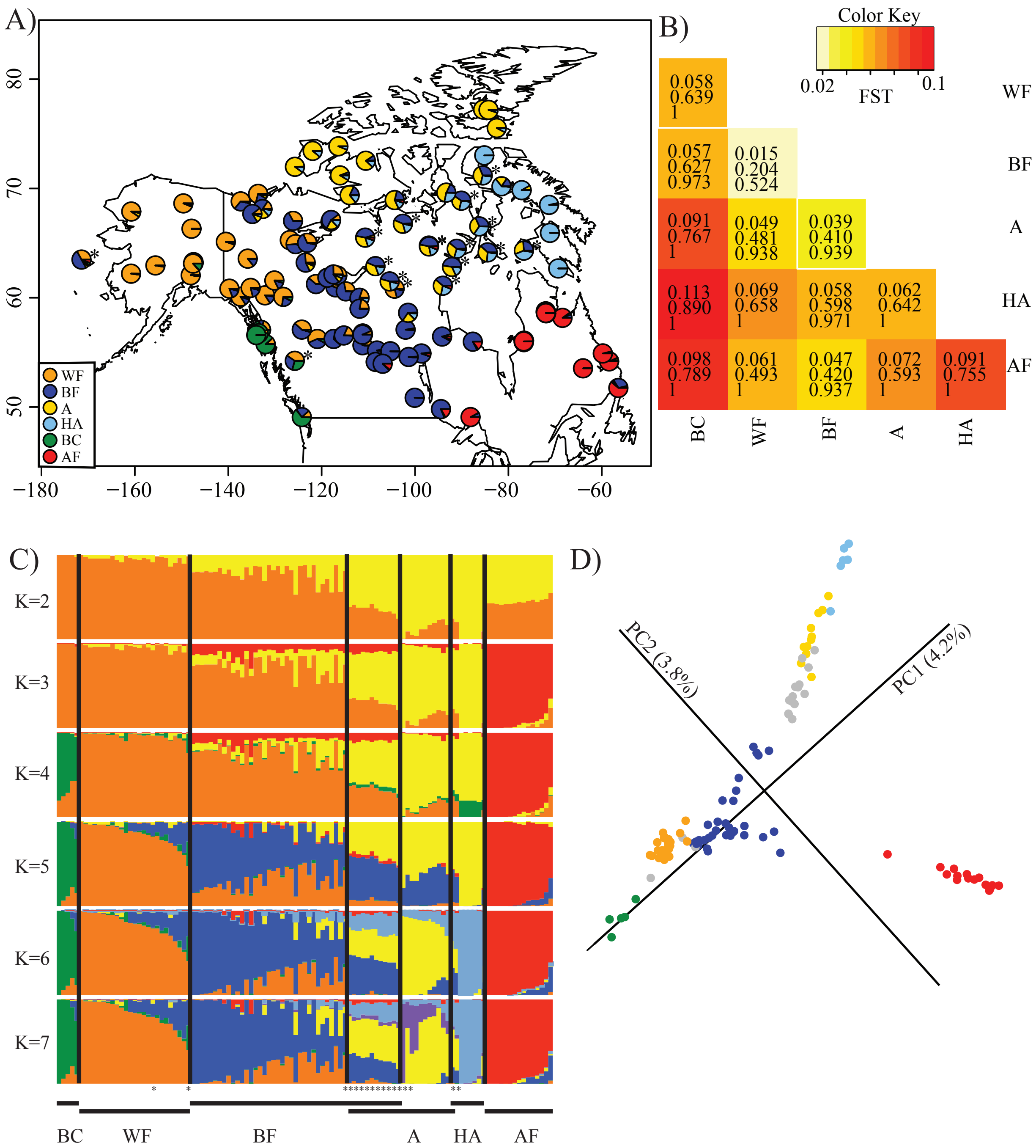


Figure 1-1. Genetic differentiation among wolf populations. A) Sampling locations of 111 wolves. Each wolf is symbolized by a pie chart with relative proportion of assignment at K=6 in STRUCTURE. Individuals with an asterisk were admixed at K=6 and were removed. B) Heatmap of pairwise Fst for 42K SNPs in 94 non-admixed wolves. The mean, 99th percentile, and maximum pairwise Fst are labeled. C) STRUCTURE plot for K=2 to K=7. Admixed individuals are marked with an '*' . D) PCA plot of 111 wolves for 22K LD-pruned data set, slightly rotated to match map in A). Colors represent STRUCTURE groupings, with admixed/removed individuals in gray. WF: West Forest, BF: Boreal Forest, A: Arctic, HA: High Arctic, BC: British Columbia, AF: Atlantic Forest

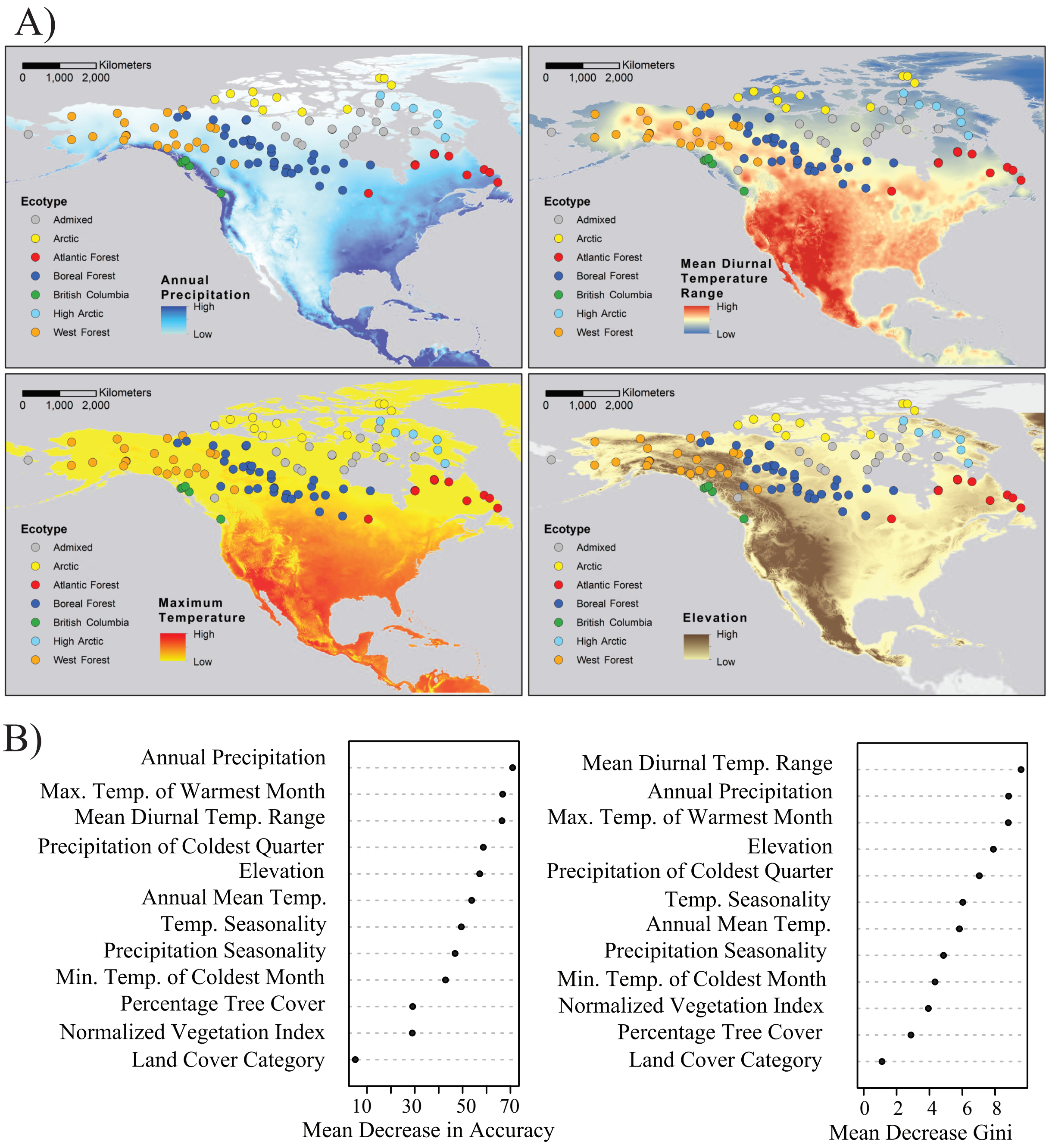


Figure 1-2. Environmental variation among wolves sampled in this study. A) Sampling location for wolves imposed on maps of variation for (clockwise, from top left) annual precipitation, mean diurnal temperature range, elevation, and maximum temperature. These variables were ranked as important from Random Forest analysis. B) Output from Random Forest analysis showing which environmental variables were most relevant in assigning individuals to their habitat. Environmental variables with higher mean decrease in accuracy (left) and higher mean decrease in Gini index (right) are shown. See Liaw and Wiener 2002 for details.

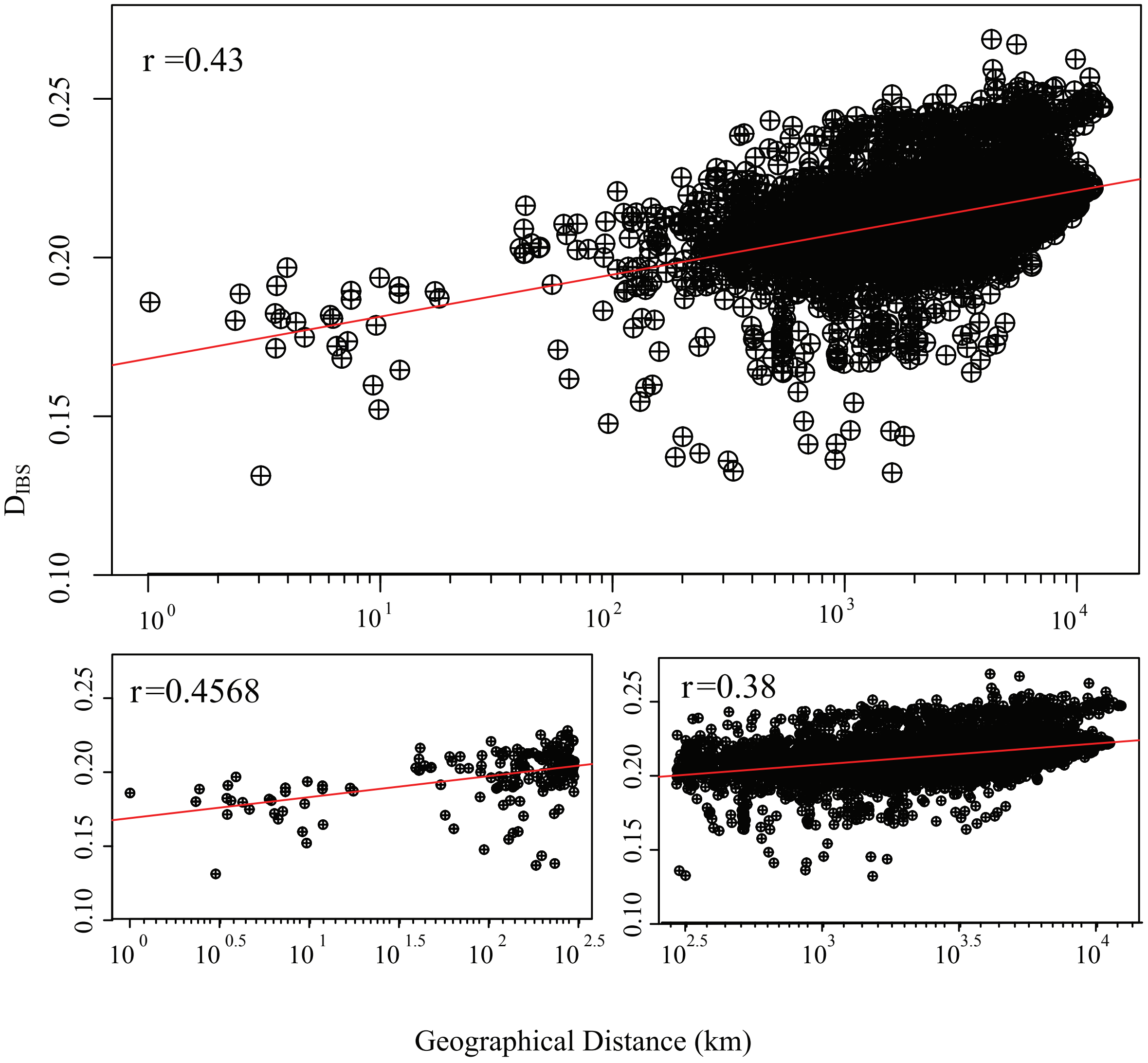


Figure 1-3. Isolation by distance for 111 North American wolves and the 42K SNP set. Pairwise Genetic distance (D_{IBS}) was calculated in PLINK and plotted against the log10 of geographic distance in kilometers for all samples (top panel), individuals closer than 300km (bottom left panel) and individuals further than 300km (bottom right panel).

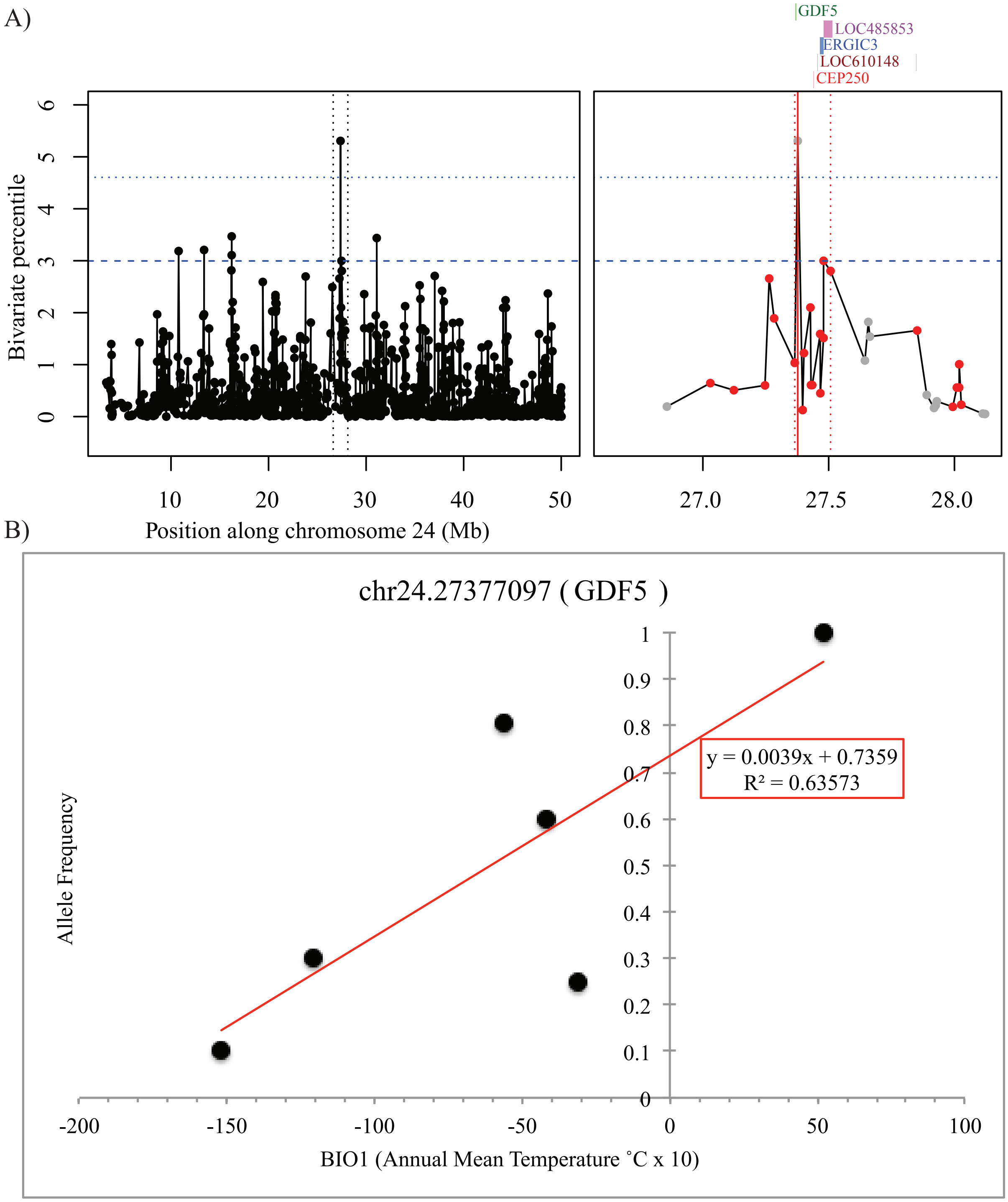


Figure 1-4. Signals of selection for Growth Differentiation Factor 5 (GDF5) in Boreal Forest wolves (A) and across all ecotypes (B). A) The bivariate percentile of Fst and XPEHH for SNPs along Chr 24 in Boreal Forest wolves, with a closer look at genes overlapping the cluster. B) The annual mean temperature for each ecotype is plotted along the x-axis, with the mean allele frequency of the SNP tagging GDF.

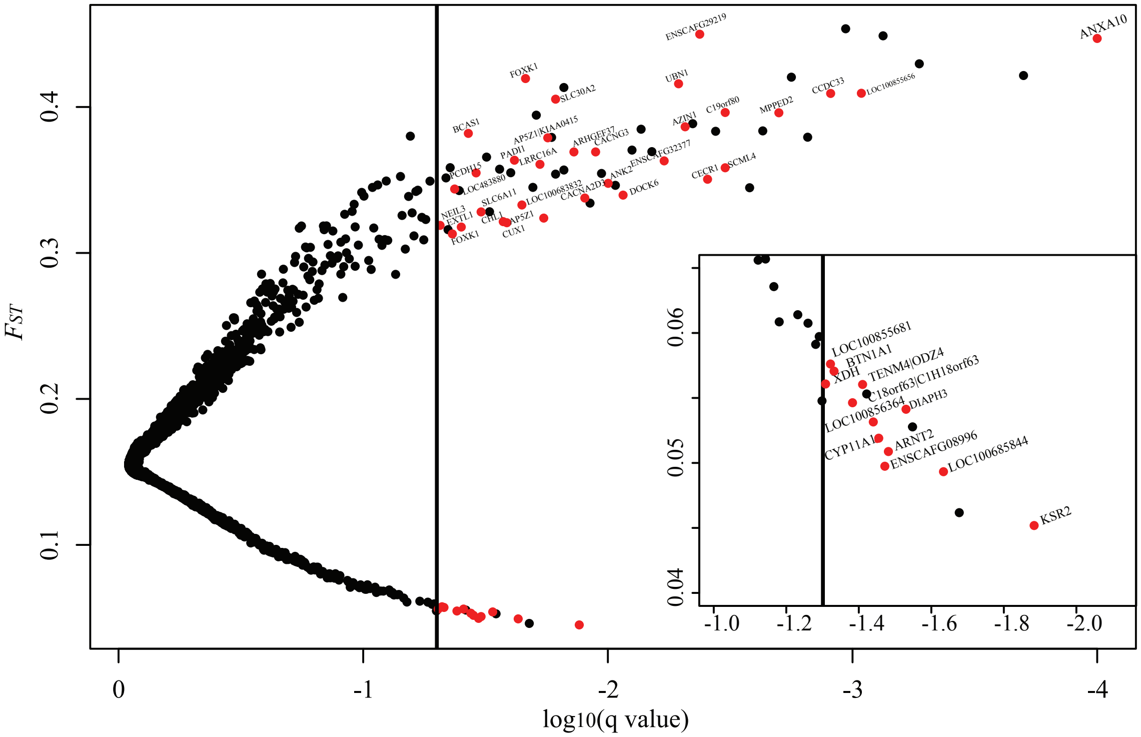


Figure 1-5. Signatures of selection in North American wolf ecotypes using BayeScan. The horizontal axis indicates the log10 of the q value (the FDR analog to the p-value) and the vertical axis is the mean Fst between each of six ecotypes. The vertical line indicates the log10 of FDR=0.05. SNPs tagging genes are highlighted in red with the gene name to the right. The inset shows a clearer picture of candidates with a low *FST*.

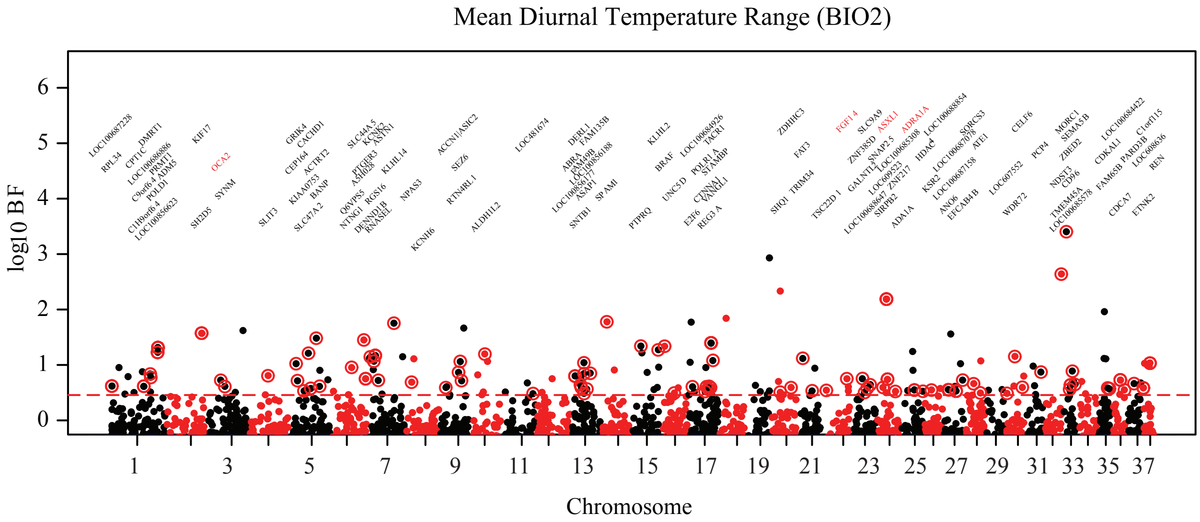


Figure 1-6. Manhattan plot of log10 Bayes Factors for Mean Diurnal Temperature Range (BIO2). The chromosomal locations of all SNPs with log10 BF >0 are plotted. SNPs within 10kb of a gene are circled in red, and the name of the gene closest to that SNP is labeled above. The empirical 99.5th is plotted (red dashed line). Genes mentioned in the Results and Discussion are highlighted in red.

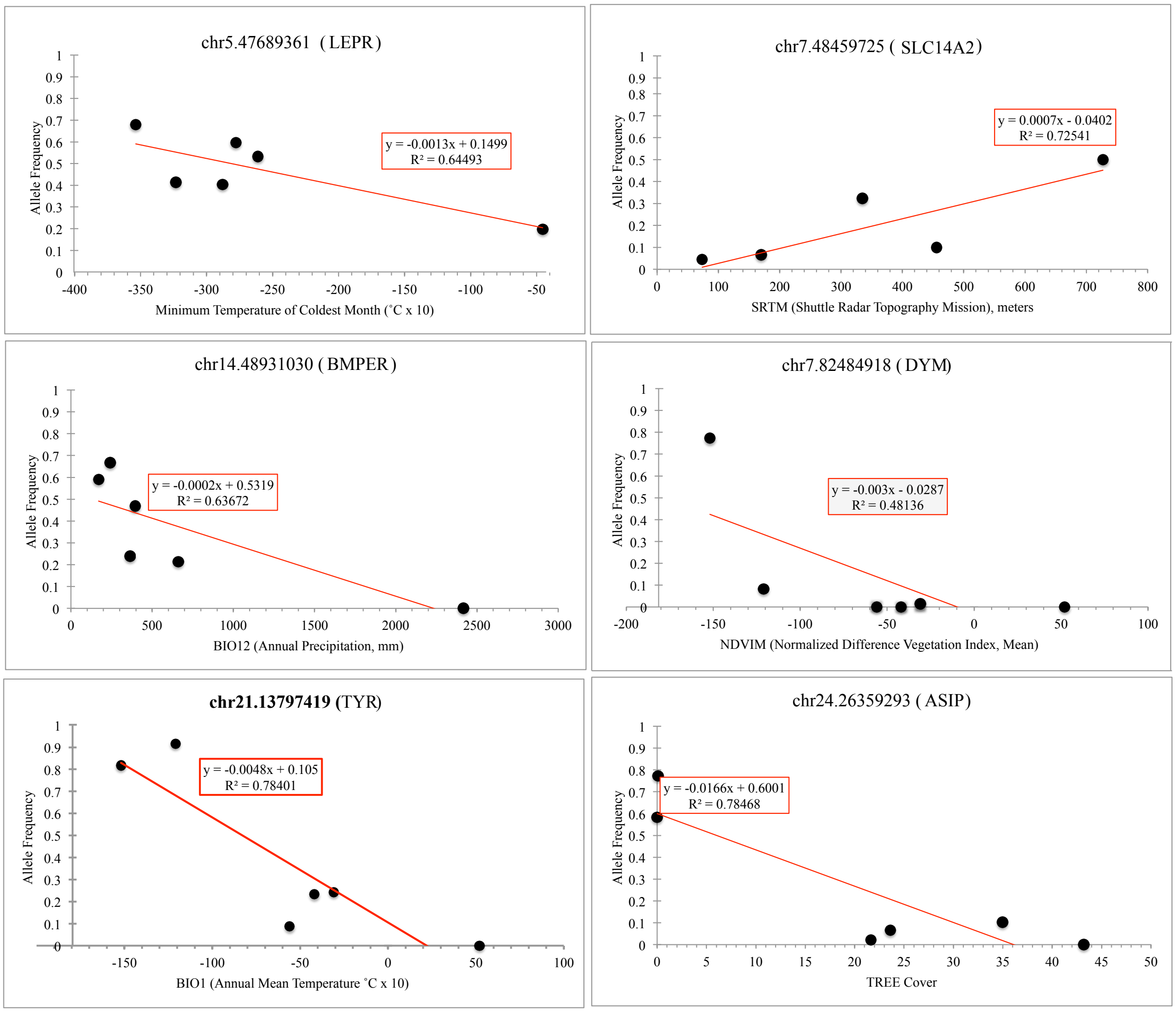


Figure 1-7. Examples of clinal variation of SNPs in metabolism, morphology, and pigmentation candidate genes. Population mean allele frequency for each SNP was plotted against the population mean of the environmental variable for which the SNP was a candidate outlier. Candidate genes are related to, metabolism (top panel, LEPR=Leptin Receptor, SLC14A2=Solute Carrier Family 14, Urea Transporter, Member 2), morphology (middle panel, BMPER=BMP Binding Endothelial Regulator, DYM=Dymeclin), and pigmentation (bottom panel, TYR=Tyrosinase, ASIP=Agouti Signaling Protein).

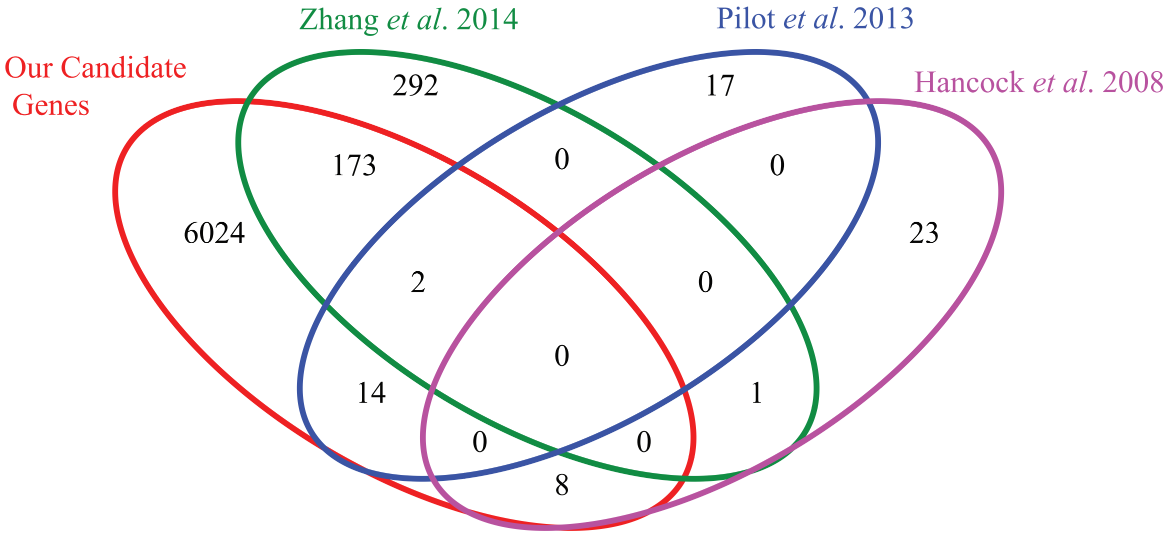


Figure 1-8. Venn diagram showing overlap between the union of our candidate genes from FST/XPEHH, BayeScan, and Bayenv at the 5% significance threshold, and three other studies. Zhang et al. (2014) and Pilot et al. (2014) used wolves as a study system, and Hancock et al. (2008) studied humans.

Table 1-1. Summary of Gene Ontology enrichment for F_{ST} /XP-EHH selection scan. For each ecotype, an example of a specific category related to each general category discussed in the main text is provided, with the significance of the specific category after Benjamini-Hochberg FDR correction. GO: gene ontology, HP: human phenotype, KEGG: Kyoto Encyclopedia of Genes and Genomes.

Ecotype	General Category	Example(s) of Specific Category	Significance of Specific Category	Type
West Forest	cardiovascular system	Abnormality of the cardiovascular system	2.61E-02	HP
	hearing	functional abnormality of the middle ear	4.01E-02	HP
	membranes	integral component of plasma membrane*	4.18E-02	GO
	metabolism	metabolic pathways	3.63E-02	KEGG
	organismal system	abnormality of the liver	3.67E-02	HP
	skeletal morphology	abnormality of the external nose	1.71E-02	HP
	vision	abnormality of the eye	2.18E-02	HP
Boreal Forest	immune response	immune system process*	2.85E-04	GO
	metabolism	lipid metabolic process*	1.37E-02	GO
	organismal system	tissue development*	2.31E-05	GO
	response to environment	response to external stimulus*	1.14E-04	GO
	skeletal morphology	ossification*	2.29E-04	GO
Arctic	immune response	positive regulation of lymphocyte mediated immunity	4.46E-02	GO
	musculature	abnormality of the musculature	5.00E-02	HP
	organismal system	functional abnormality of bladder	4.71E-02	HP
	skeletal morphology	abnormal bone ossification	4.38E-03	HP
High Arctic	brain function	learning or memory	1.08E-02	GO
	metabolism	histidine metabolism	5.00E-02	KEGG
British Columbia	dentition	misalignment of teeth	4.79E-02	HP
	diet	salivary secretion	4.22E-02	KEGG
	metabolism	Arachidonic acid metabolism	2.73E-03	KEGG
	musculature	Muscle hypertrophy	4.73E-02	HP
	organsismal system	protein transport*	9.31E-03	GO
	skeletal morphology	disproportionate short stature	4.86E-02	HP
	vision	aplasia/hypoplasia of the iris	3.93E-02	HP
Atlantic Forest	dentition	hypodontia	4.20E-02	HP
	metabolism	Glutathione metabolism	5.00E-02	KEGG
	organismal system	calcium ion transmembrane tranporter activity	5.00E-02	GO
	skeletal morphology	aplasia involving forearm bones	4.57E-02	HP

* Indicates category was enriched in top 0.5% candidate genes

Table 1-2. Summary of Gene Ontology enrichment for BayeScan selection scan. Significance of specific category after Benajamani-Hochberg FDR is provided, in addition to the type of category. GO: gene ontology, KEGG: Kyoto Encyclopedia of Genes and Genomes.

General Category	Example of Specific Category	Signifiance	Type
hearing	auditory receptor cell differentiation	5.00E-02	GO
organismal system	Oxytocin signaling pathway	1.02E-02	KEGG
cardiovascular system	Cardiac muscle contraction	2.11E-03	KEGG

Table 1-3. Summary of significant Gene Ontology enrichment for Bayenv selection scan. Within each general category, specific examples are provided. Shaded boxes indicate the GO category was enriched within genes at the 95th percentile or above for that environmental variable; an asterisk indicates significance at the 99.5th percentile candidate gene threshold. Gene lists are provided when the same set of genes was enriched across environmental variables, with bold font indicating genes discussed in the Results and/or Discussion. BIO1: annual mean temp., BIO2: mean diurnal temp. range, BIO4: temp. seasonality, BIO5: max. temp. of warmest month, BIO6: min. temp. of coldest month), BIO12: annual precipitation, BIO15: precipitation seasonality, BIO19: precipitation of coldest quarter, LC: land cover metric, NDVIM: normalized difference vegetation index, TREE: percentage tree cover, and SRTM: shuttle radar topography metric.

General Category	Specific GO Category	TEMPERATURE					PRECIPITATION			VEGETATION			ELEVATION	Genes
		BIO1	BIO2	BIO4	BIO5	BIO6	BIO12	BIO15	BIO19	LC	NDVIM	TREE	SRTM	
Hearing	ear development/inner ear development								x					CDH23,PRKRA,GLI2,GLI3,BCR,PAX8,DFNB31,**BMPER**,CHD7,LOC100856477,ECE1,DVL3,EYA1,GABRA5,TMC1,**USH2A**,TCAP,NTN1
	sensory perception of sound			x*									x	NAV2,Q1KQ08,**PCDH15**,ATP6V0A4,CDH23,SLC12A2,**USH2A**,LOXHD1,FBXO11,GPR98,ACCN1,CNTN5,FZD4,TUB
Morphology	anatomical structure development	x	x	x	x	x	x		x	x	x	x	x	varies
	anatomical structure morphogenesis	x		x	x	x	x		x	x	x	x	x	varies
	appendage development				x		x		x			x		varies
	appendage morphogenesis		x		x		x		x			x		varies
	growth						x							MBD5,ERBB4,MREG,CXCL16,LRP4,C10ORF90,SEMA7A,CTGF,CHD7,GLI2,COL9A1,PNPT1,DSCAM,ATRN,TCF7L2,FOXP1,GLI3,SEMA4D,LOC100856477,TBCE,ANKRD11,GAP43,GPD2,LOC100856261,SLIT3,BCL2,NRG1,CRIM1,ATP8A2,LPAR3,PPARD,PLXNA4,ESRRB,B6VCV6,ROS1,DHCR7,WISP2,PML,NPY1R,SMARCA2,POU1F1,INSR,PSEN2,BARHL2,LOC100855935
	limb morphogenesis		x		x		x		x			x		varies
	negative regulation of BMP signaling pathway				x									**BMPER**,SOST,DKK1,GREM1,SMAD6,TRIM33,TCF7L2
	positive regulation of ossification		x											FBN2,MIA3,**BMP4**,ZBTB16,OSR1,LRP6,**BMP6**
	regulation of growth					x	x		x					varies
	regulation of muscle tissue development/organ development/cell differentiation				x									HDAC9,TCF21,FLNB,COL19A1,FOXP2,HDAC2,USP19,TP63,NRG1,TGFBR3,Q8SPL8,Q38IV4,**PPARGC1A**,DKK1,Q6TYZ5,MYO18B,BOC,GREM1,MEGF10,SIN3B,EYA1,HDAC4,**TCF7L2**,NEBL,ZFAND5,PLCB1,LOC100856477,FKBP1A
Pigmentation	melanocyte differentiation*		x	x										**OCA2**,SOX10, GLI3,MREG
	melanin biosynthetic process	x									x	x		**MYO5A,TYR,ASIP,OCA2,TYRP1**
	melanin metabolic process	x										x		**MYO5A,TYR,ASIP,OCA2,TYRP1**
	melanosome membrane											x		**TYR,OCA2,TYRP1**
	pigmentation											x		**MYO5A,TYR,KITLG,ASIP,OCA2,**KIF13A,**TYRP1**,HPS3,SOX10,ATRN
Smell	olfactory lobe development	x												HTT,ROBO1,SLIT2,ERBB4,SALL1,SLIT1
System	blood circulation	x								x	x	x	x	varies
	calcium ion binding/calcium ion transmembrane transporter activity	x	x	x		x	x	x	x	x	x	x	x*	varies
	carbohydrate homeostasis						x		x					varies
	circulatory system process/development	x			x		x		x	x	x	x	x	varies
	developmental growth/process	x	x	x	x	x	x		x	x	x	x	x	varies
	fatty acid transport								x					SLC27A2,TNFRSF11A,B8XNP7,**PPARG**,PPARD,IRS2,FAM132B
	gated channel activity	x			x					x	x	x	x	varies
	lipoprotein particle receptor activity		x											VLDLR,LRP8,LDLR,LRP1B,LRP6
	locomotion	x	x	x	x	x	x		x	x	x	x	x	varies
	mesenchymal cell differentiation involved in kidney development/renal system		x											STAT1,**TCF21**,**BMP4**,OSR1
	multicellular organismal response to stress					x	x		x			x		varies
	negative regulation of ion transport						x							PTK2B,NOS1,BEST3,BCL2,IRS2,SLC9A3R1,NEDD4L
	passive transmembrane transporter activity		x		x		x			x		x	x	varies
	positive regulation of peptide (hormone) secretion					x	x		x					varies
	renal system development		x		x	x	x		x					varies
	positive regulation of lipid transport*			x										**LIPG**,TNFRSF11A
	response to stimulus	x	x	x	x	x	x	x	x	x	x	x	x	varies

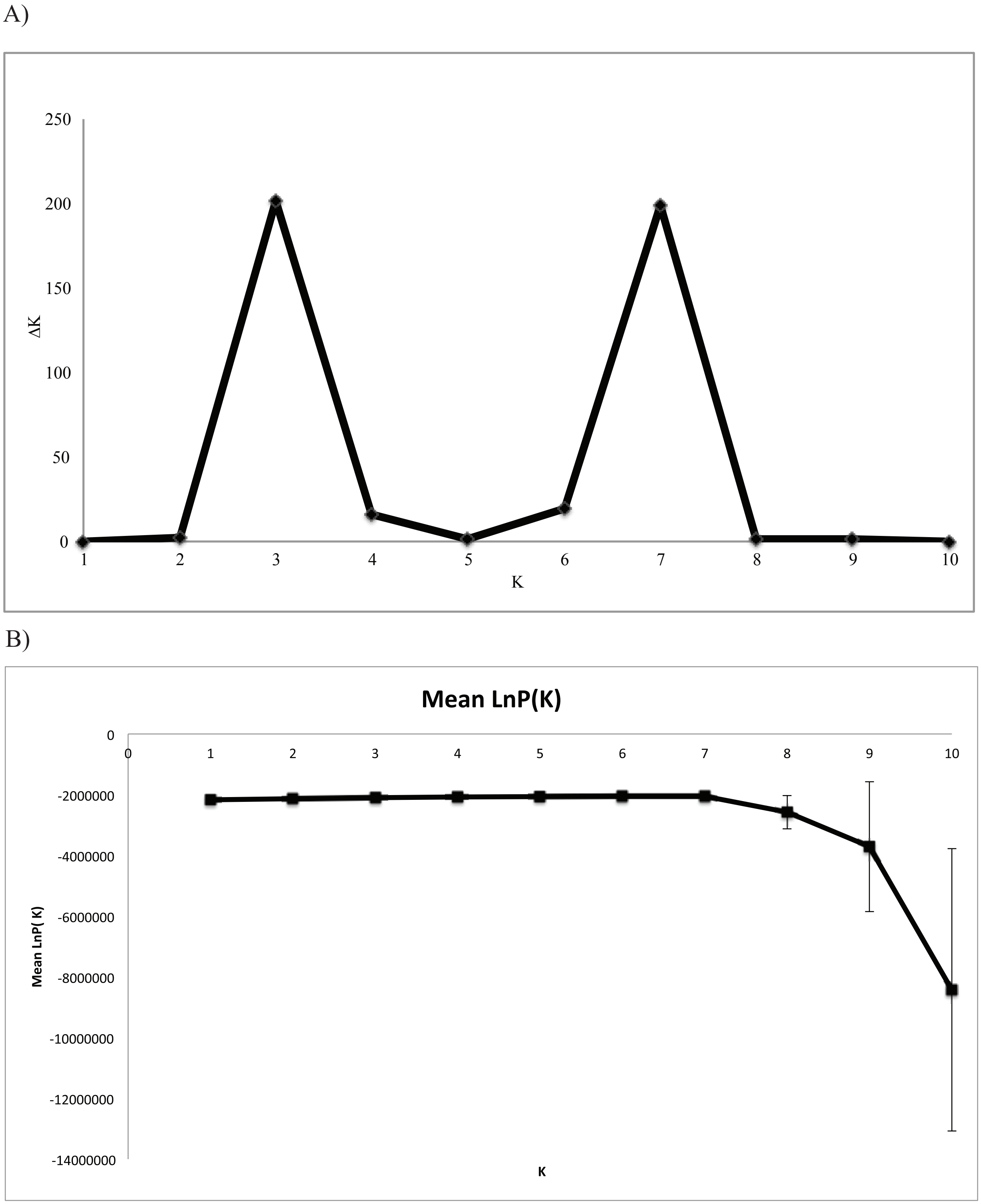


Figure 1-S1. Plot showing A) delta K and B) likelihood values for K=1 to K=10 from STRUCTURE run of 111 individuals and 22K LD-pruned SNPs. The standard deviation of mean likelihood values is shown at each value of K in B).

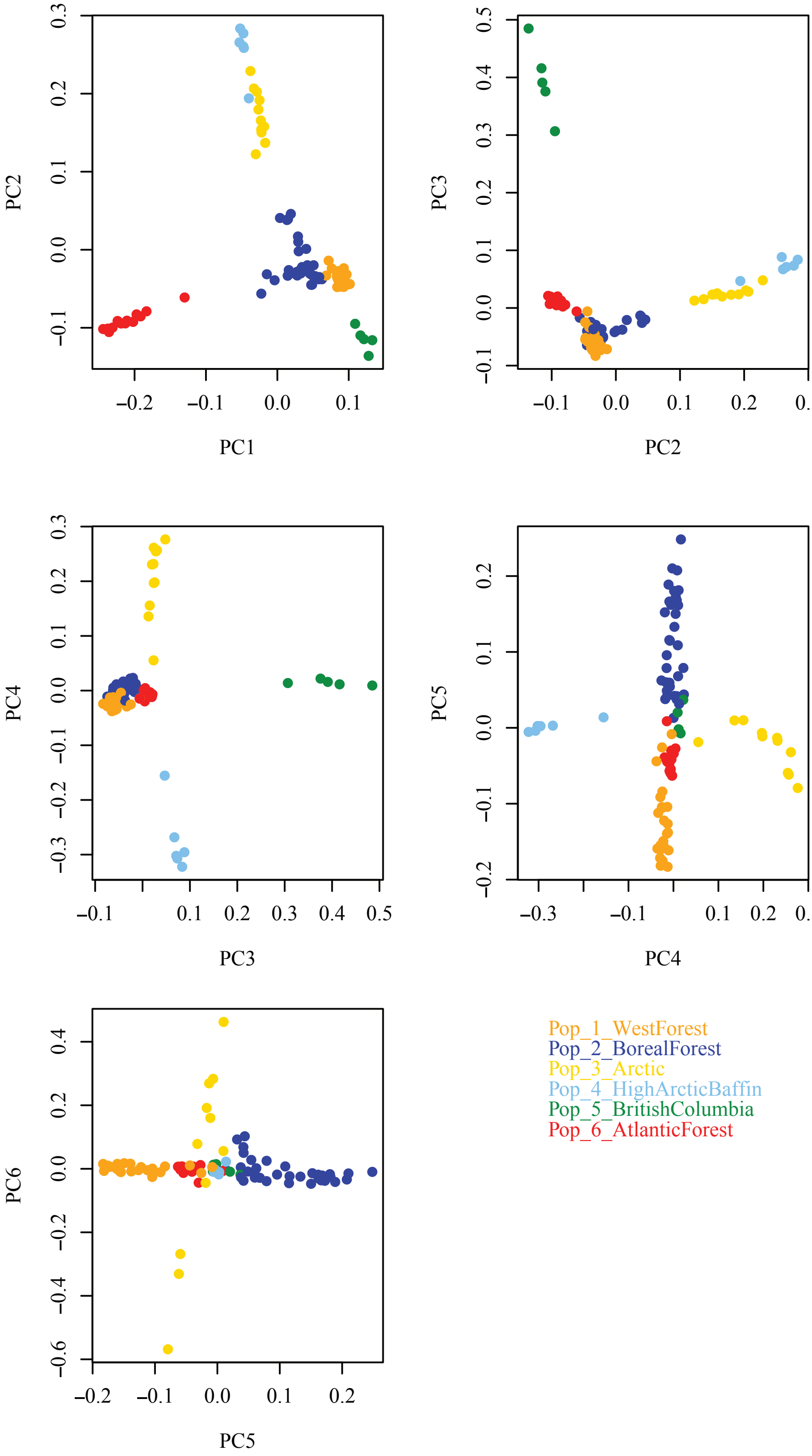


Figure 1-S2. First six principal components for 22K LD-pruned SNP set in 94 non-admixed individuals. Samples are color coded by their STRUCTURE grouping at K=6 as in the legend.

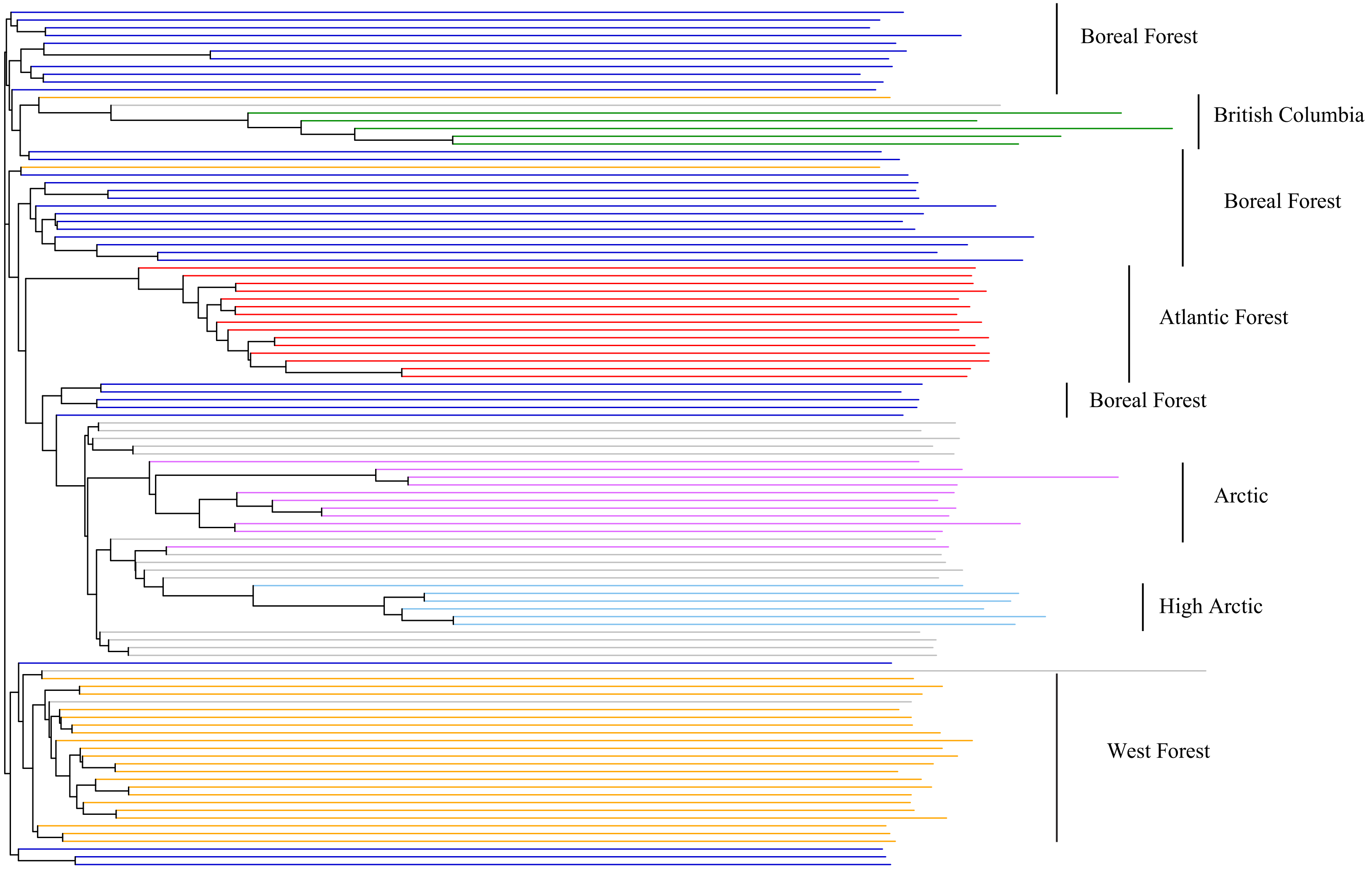


Figure 1-S3. Unrooted neighbor-joining tree based on 111 individuals genotyped at 22K LD-pruned SNPs. Branches are colored according to the STRUCTURE grouping at K=6, with samples that were excluded for selection tests in grey.

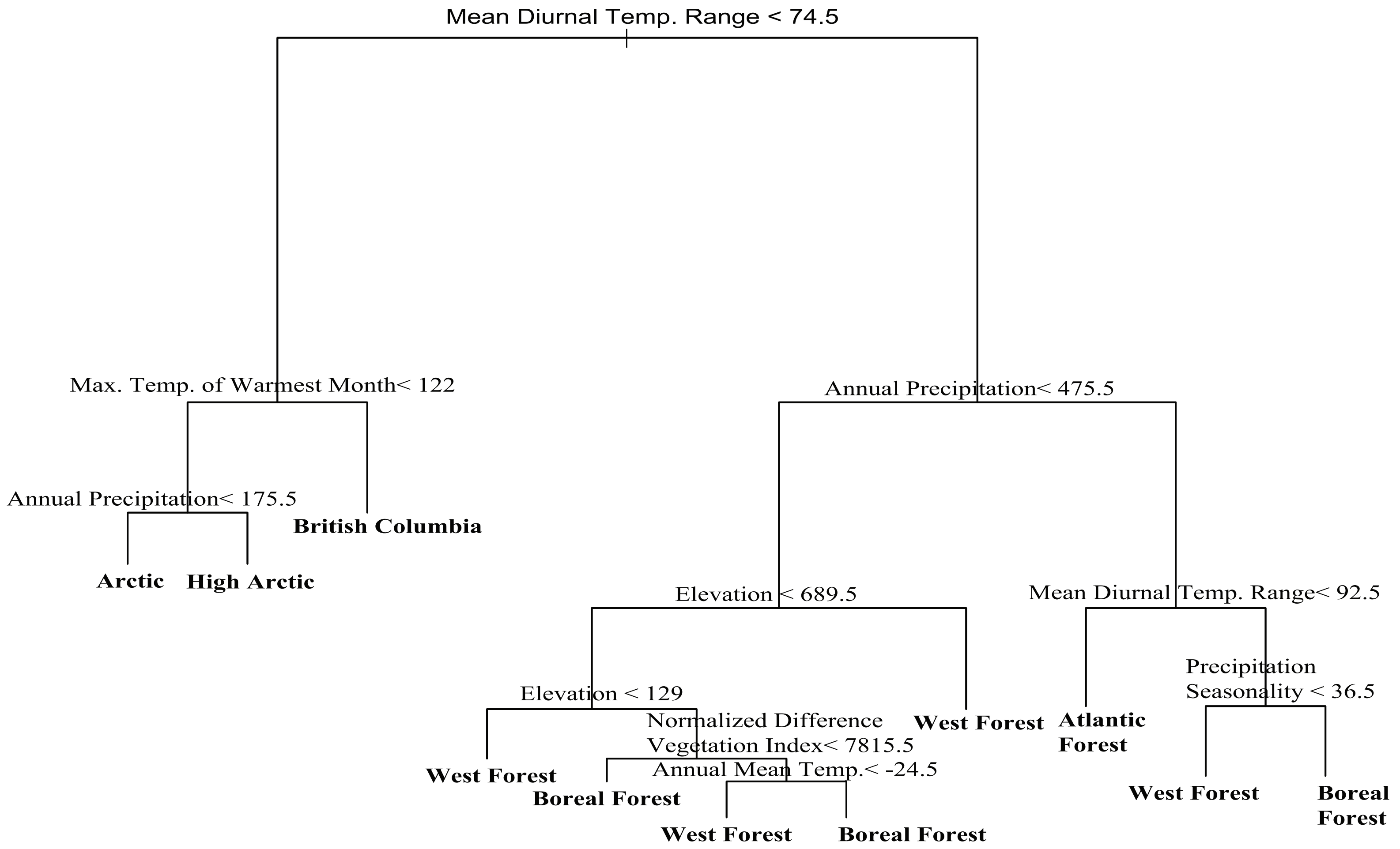


Figure 1-S4. Sample tree output from Random Forest classification scheme on environmental data. Numerical values at each node represent the specific threshold value at which samples were grouped for further classification. Precipitation variables are measured in milometers, temperature variables are measured in ˚C*10, elevation is measured in meters, and precipitation seasonality is the coefficient of variation.

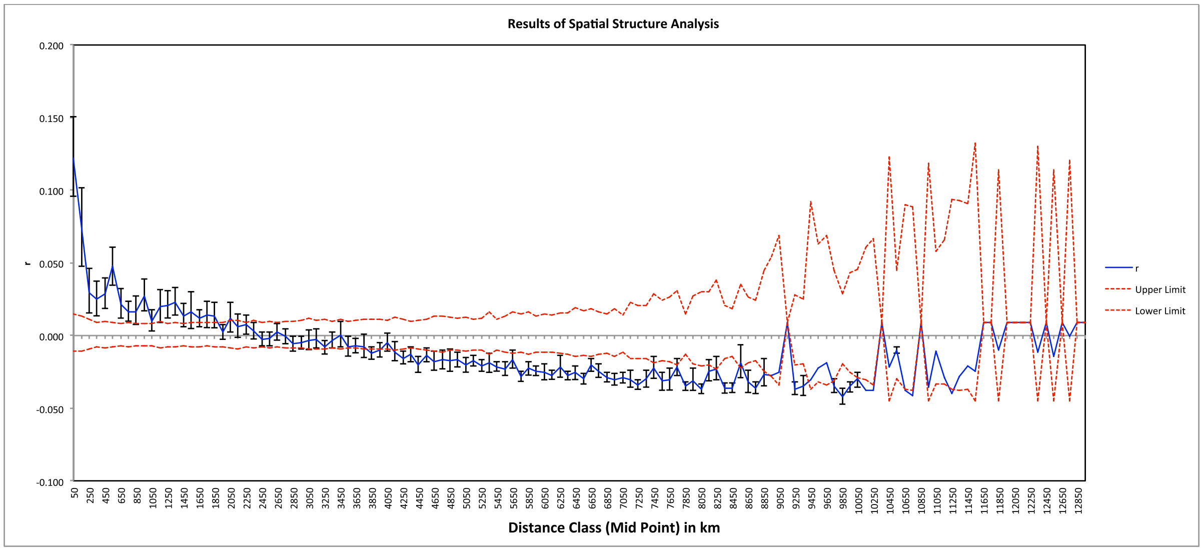


Figure 1-S5. Spatial autocorrelation measured within 130 even distance classes of 100km each. Significance was assessed using 9999 permutations, and the 95% confidence interval around r was determined using 9999 bootstraps.

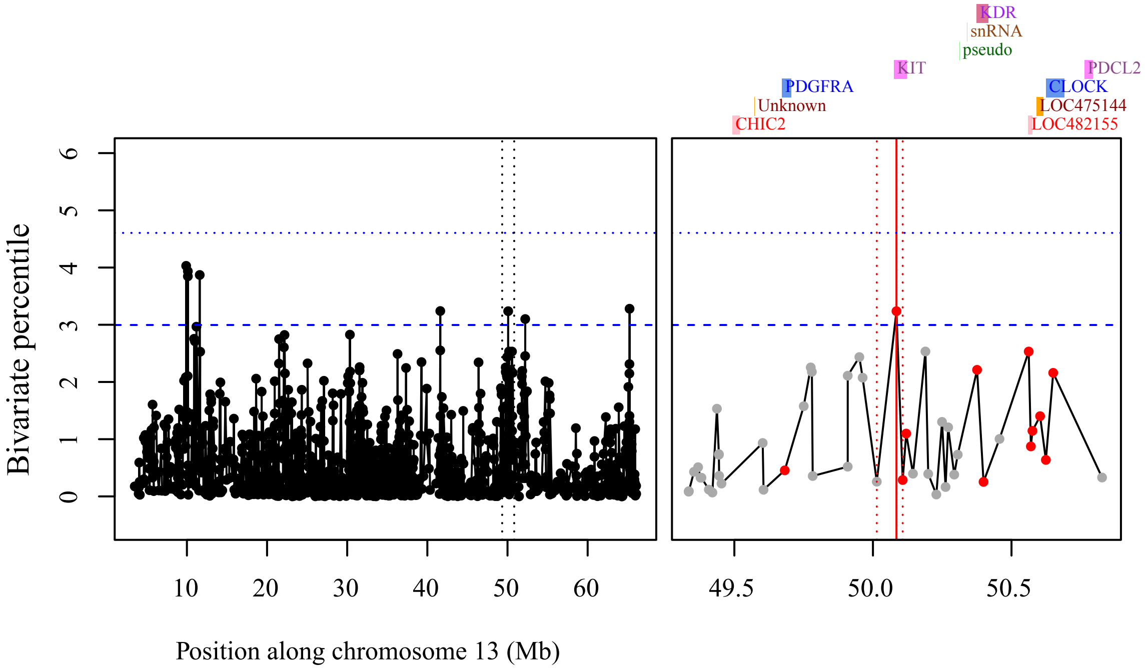


Figure 1-S6. The bivariate percentile of *FST* and *XP-EHH* for SNPs along chromosome 13 in High Arctic wolves, with a closer look at genes overlapping the cluster. The high-ranking cluster contains the KIT gene.

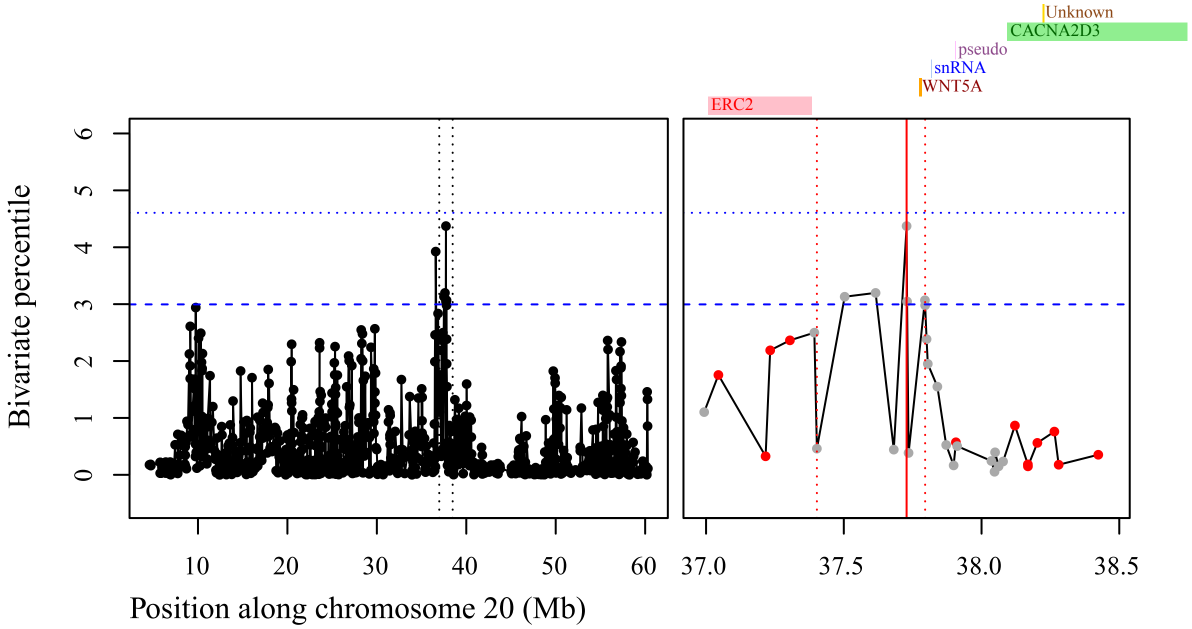


Figure 1-S7. The bivariate percentile of *FST* and *XP-EHH* for SNPs along chromosome 20 in British Columbia wolves, with a closer look at genes overlapping the cluster. The high-ranking cluster contains the WNT5A gene.

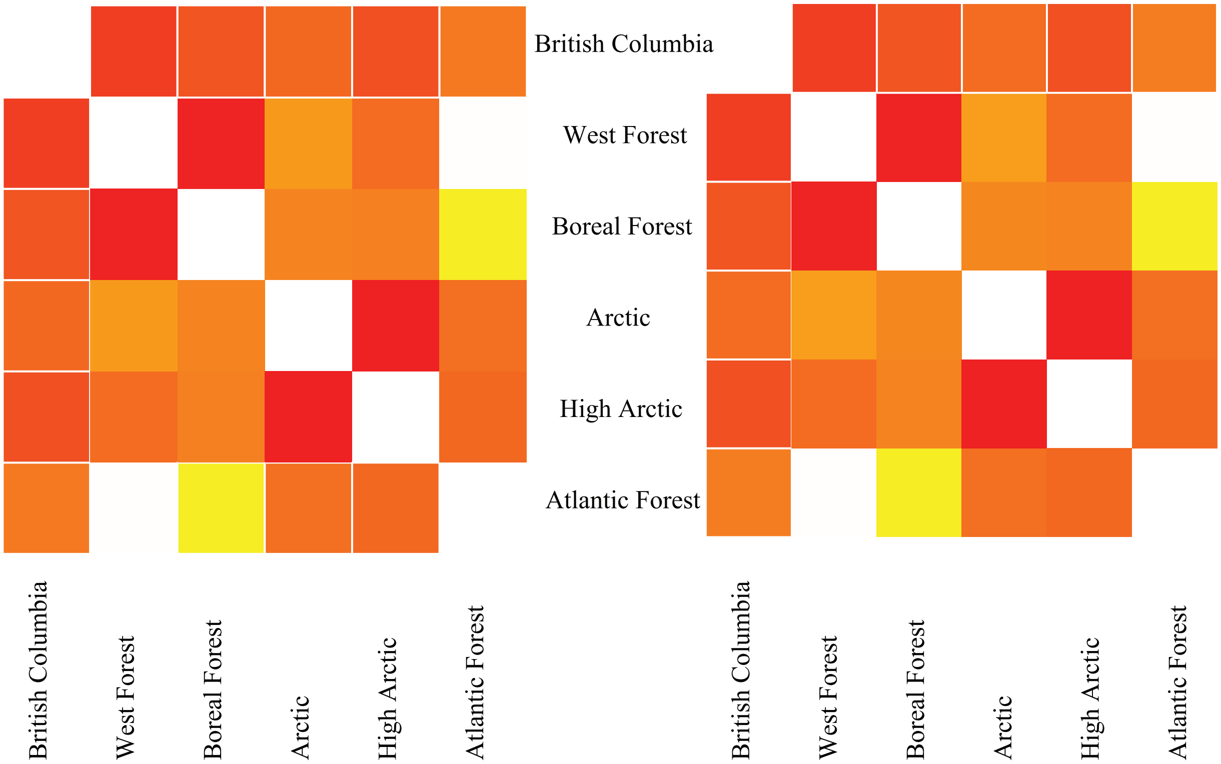


Figure 1-S8. Average covariance (left) and correlation (right) matrices after 500,000 iterations of the background matrix estimation in Bayenv using 10K random SNPs. Red heat colors correspond to more positive values and white/yellow heat colors correspond to more negative values.

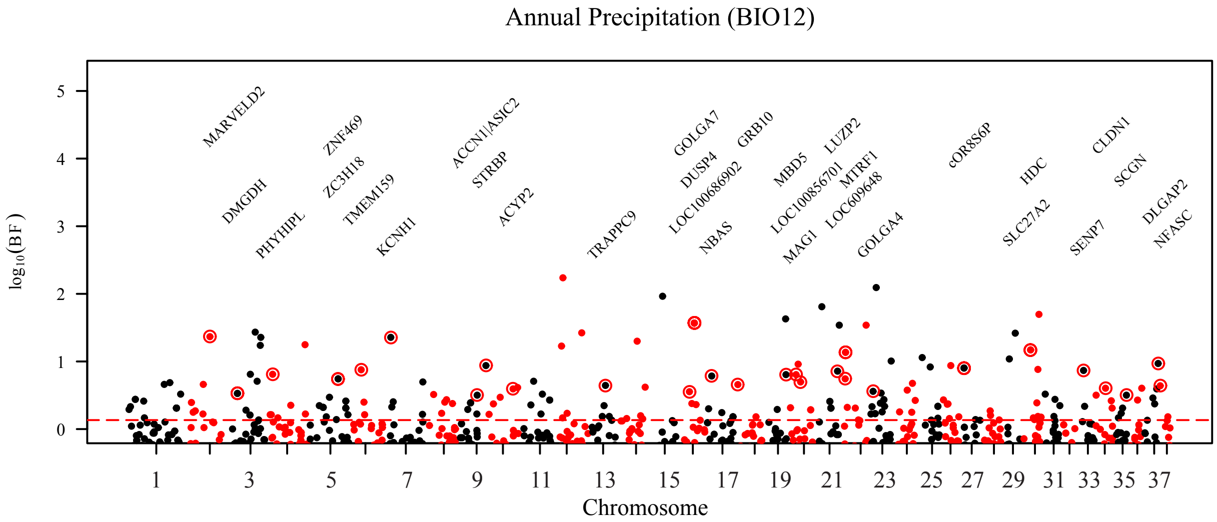
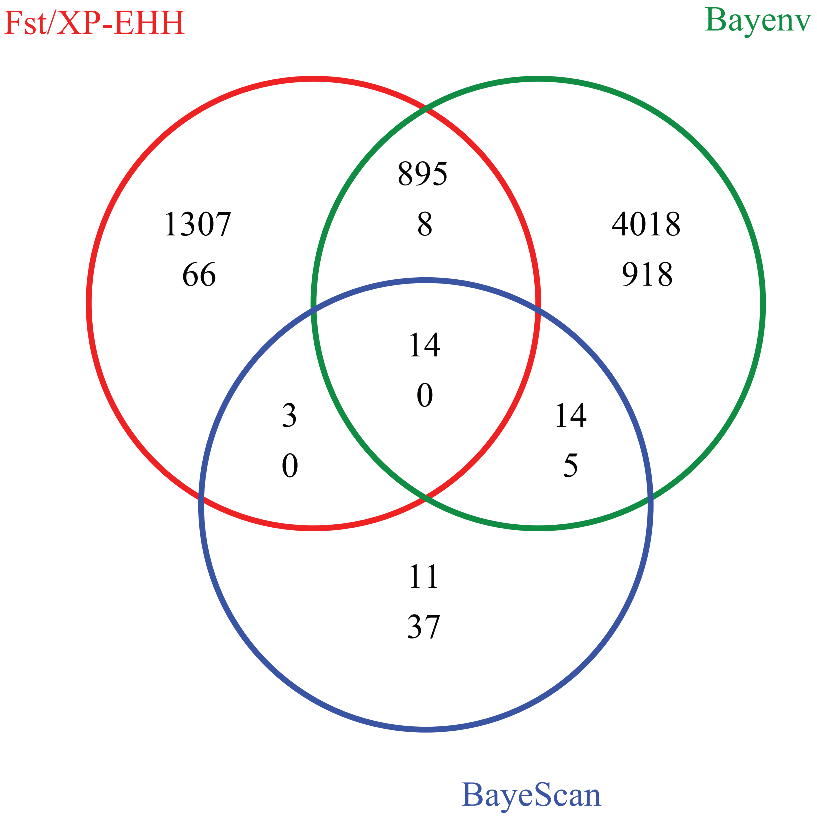


Figure 1-S9. Manhattan plot of log10 Bayes Factors for Annual Precipitation (BIO12). The chromosomal locations of all SNPs with log10 BF >0 are plotted. SNPs with BF>3 and within 10kb of a gene are circled in red, and the name of the gene closest to that SNP is labeled above. The empirical 99.5th and 95th percentile cutoffs are plotted.



Top: 95%; Bottom: 99.5%

Figure 1-S10. Venn diagram showing overlap between candidate genes from FST/XP-EHH, BayeScan, and Bayenv at the top 5% (top numbers) and top 0.5% (bottom numbers) significance level. For BayeScan, the same gene list was used for both comparisons.

Bibliography

Akey JM (2009) Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research*, **19**, 711–722.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Alagramam KN, Yuan H, Kuehn MH *et al.* (2001) Mutations in the novel protocadherin PCDH15 cause Usher syndrome type 1F. *Human Molecular Genetics*, **10**, 1709–1718.

Anderson TM, Candille SI, Musiani M *et al.* (2009) Molecular and evolutionary history of melanism in North American gray wolves. *Science*, **323**, 1339–1343.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Barsh G (1996) The genetics of pigmentation: from fancy genes to complex traits. *Trends in Genetics*, **12**, 299–305.

Beermann F, Orlow SJ, Lamoreux ML (2004) The Tyr (albino) locus of the laboratory mouse. *Mammalian Genome*, **15**, 749–758.

Bennett EP, Mandel U, Clausen H *et al.* (2012) Control of mucin-type O-glycosylation: A classification of the polypeptide GalNAc-transferase gene family. *Glycobiology*, **22**, 736–756.

Blair LM, Granka JM, Feldman MW (2014) On the stability of the Bayenv method in assessing human SNP-environment associations. *Human Genomics*, **8**, 1-13.

Boyko A, Quignon P, Li L, Schoenebeck J (2010) A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol*ogy, **19**: 795-803

Bradfield, H Rob Taal, Timpson NJ *et al.* (2012) A genome-wide association meta-analysis identifies new childhood obesity loci. *Nature Genetics*, **44**, 526–531.

Bragdon B, Moseychuk O, Saldanha S *et al.* (2011) Cellular Signalling. *Cellular Signalling*, **23**, 609–620.

Breiman L (2001) Random forests. *Machine learning*, **45**, 5–32.

Brockschnieder D (2006) Ermin, A Myelinating Oligodendrocyte-Specific Protein That Regulates Cell Morphology. *Journal of Neuroscience*, **26**, 757–762.

Cai J, Mutoh N, Shin J-O *et al.* (2011) Wnt5a plays a crucial role in determining tooth size during murine tooth development. *Cell and Tissue Research*, **345**, 367–377.

Carmichael LE, Krizan J, Nagy JA *et al.* (2007) Historical and ecological determinants of genetic structure in arctic canids. *Molecular Ecology*, **16**, 3466–3483.

Carneiro M, Rubin CJ, Di Palma F *et al.* (2014) Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science*, **345**, 1074–1079.

Chua SC, White DW, Wu-Peng XS *et al.* (1996) Phenotype of fatty due to Gln269Pro mutation in the leptin receptor (Lepr). *Diabetes*, **45**, 1141–1143.

Coffino P (2001) Regulation of cellular polyamines by antizyme. *Nature Reviews Molecular Cell Biology*, **2**, 188–194.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, **185**, 1411–1423.

Coppe A, Agostini C, Marino IAM *et al.* (2013) Genome Evolution in the Cold: Antarctic Icefish Muscle Transcriptome Reveals Selective Duplications Increasing Mitochondrial Function. *Genome Biology and Evolution*, **5**, 45–60.

Craft CS, Pietka TA, Schappe T *et al.* (2014) The Extracellular Matrix Protein MAGP1 Supports Thermogenesis and Protects Against Obesity and Diabetes Through Regulation of TGF. *Diabetes*, **63**, 1920–1932.

Cristancho AG, Lazar MA (2011) Forming functional fat:a growing understanding ofadipocyte differentiation. *Nature Publishing Group*, **12**, 722–734.

Dahlberg J, Nilsson L-O, Wowern von F, Melander O (2007) Polymorphism in NEDD4L Is Associated with Increased Salt Sensitivity, Reduced Levels of P-renin and Increased Levels of Nt-proANP (J Carr, Ed,). *PLoS ONE*, **2**, e432.

Darimont CT, Reimchen TE, Paquet PC (2003) Foraging behaviour by gray wolves on salmon streams in coastal British Columbia. *Canadian Journal of Zoology*, **81**, 349–353.

de Jong MA, Collins S, Beldade P, Brakefield PM, Zwaan BJ (2012) Footprints of selection in wild populations of Bicyclus anynanaalong a latitudinal cline. *Molecular Ecology*, **22**, 341–353.

Dobzhansky T (1948) Genetics of natural populations. XVI. Altitudinal and seasonal changes produced by natural selection in certain populations of Drosophila pseudoobscura and Drosophila persimilis. *Genetics*, **33**, 158-176.

Domingues VS, Poh Y-P, Peterson BK *et al.* (2012) Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution*, **66**, 1–15.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Edmondson AC, Brown RJ, Kathiresan S *et al.* (2009) Loss-of-function variants in endothelial lipase are a cause of elevated HDL cholesterol in humans. *Journal of Clinical Investigation*, **119**, 1042–1050.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 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.

Fariello M-I, Servin B, Tosser-Klopp G *et al.* (2014) Selection signatures in worldwide sheep populations. *PLoS ONE*, **9**, e103813.

Fitzpatrick MC, Keller SR (2014) Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, **18**, 1–16.

Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, **180**, 977–993.

Frank M, Kemler R (2002) Protocadherins. *Current Opinion in Cell Biology*, **14**, 557–562.

Freedman AH, Gronau I, Schweizer RM *et al.* (2014) Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genetics*, **10**, e1004016.

Geffen E, Anderson M, Wayne RK (2004) Climate and habitat barriers to dispersal in the highly mobile grey wolf. *Molecular Ecology*, **13**, 2481–2490.

Gipson P, Bangs E, Bailey T *et al.* (2002) Color patterns among wolves in western North America. *Wildlife Society Bulletin*, **30**, 821–830.

Goding CR (2000) Mitf from neural crest to melanoma: signal transduction and transcription in the melanocyte lineage. *Genes & Development*, **14**, 1712–1728.

Gray MM, Granka JM, Bustamante CD *et al.* (2009) Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, **181**, 1493–1505.

Gunther T, Coop G (2013) Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, **195**, 205-220.

Hancock A, Witonsky D, Ehler E *et al.* (2010) Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences*, **107**, 8924.

Hancock A, Witonsky D, Gordon A *et al.* (2008) Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, **4**, e32.

Harrigan RJ, Thomassen HA, Buermann W, Smith TB (2014) A continental risk assessment of West Nile virus under climate change. *Global Change Biology*, **20**, 2417–2425.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hodges E, Xuan Z, Balija V *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, **39**, 1522–1527.

Hoekstra HE, Hirschmann R, Bundey R, Insel P, Crossland J (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

Hubbard JK, Uy JAC, Hauber ME, Hoekstra HE, Safran RJ (2010) Vertebrate pigmentation: from underlying genes to adaptive function. *Trends in Genetics*, **26**, 231–239.

Hünemeier T, Gómez-Valdés J, De Azevedo S *et al.* (2013) FGFR1 signaling is associated with the magnitude of morphological integration in human head shape. *American Journal of Human Biology*, **26**, 164–175.

Isidor B, Lindenbaum P, Pichon O *et al.* (2011) Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. *Nature Publishing Group*, **43**, 306–308.

Janssen K, Mundy NI (2013) Molecular population genetics of the melanic plumage polymorphism in Arctic skuas (Stercorarius parasiticus): evidence for divergent selection on plumage colour. *Molecular Ecology*, **22**, 4634–4643.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Jolicoeur P (1959) Multivariate geographical variation in the wolf Canis lupus L. *Evolution*, **13**, 283–299.

Jones FC, Chan YF, Schmutz J *et al.* (2012) A Genome-wide SNP Genotyping Array Reveals Patterns of Global and Repeated Species-Pair Divergence in Sticklebacks. *Current Biology*, **22**, 83–90.

Karlsson EK, Baranowska I, Wade CM *et al.* (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature Genetics*, **39**, 1321–1328.

Koblmüller S, Nord M, Wayne RK, Leonard JA (2009) Origin and status of the Great Lakes wolf. *Molecular Ecology*, **18**, 2313–2326.

Kohn A, Dong Y, Mirando AJ *et al.* (2012) Cartilage-specific RBPj -dependent and -independent Notch signals regulate cartilage and bone development. *Development*, **139**, 1198–1212.

Kratky D, Zimmermann R, Wagner EM *et al.* (2005) Endothelial lipase provides an alternative pathway for FFA uptake in lipoprotein lipase–deficient mouse adipose tissue. *Journal of Clinical Investigation*, **115**, 161–167.

Le Guédard S, Faugère V, Malcolm S, Claustres M, Roux A-F (2007) Large genomic rearrangements within the PCDH15 gene are a significant cause of USH1F syndrome. *Molecular Vision*, **13**, 102–107.

Lewandoski M (2001) Conditional control of gene expression in the mouse. *Nature Publishing Group*, **2**, 743–755.

Liaw A, Wiener M (2002) Classification and regression by randomForest. *R news*, **2**, 18–22.

Lin M, Li L, Liu C *et al.* (2011) Wnt5a regulates growth, patterning, and odontoblast differentiation of developing mouse tooth. *Developmental Dynamics*, **240**, 432–440.

Liu J, Zhang H-X (2014) Polymorphism in the 11q24. 1 genomic region is associated with myopia: A comprehensive genetic study in Chinese and Japanese populations. *Molecular Vision*, **20**, 352.

Lindblad-Toh K, Wade CM, Mikkelsen TS *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819.

Linnen, C, Kingsley E, Jensen JD, Hoekstra HE (2009) On the Origin and Spread of an Adaptive Allele in Deer Mice. *Science,* ***325,*** 1095-1098.

Linnen CR, Poh YP, Peterson BK *et al.* (2013) Adaptive Evolution of Multiple Traits Through Multiple Mutations at a Single Gene. *Science*, **339**, 1312–1316.

Liu J, Nam HK, Campbell C *et al.* (2014) Tissue-nonspecific alkaline phosphatase deficiency causes abnormal craniofacial bone development in the Alpl−/− mouse model of infantile hypophosphatasia. *Bone*, **67**, 81–94.

Lizarraga G, Lichtler A, Upholt WB, Kosher RA (2002) Studies on the Role of Cux1 in Regulation of the Onset of Joint Formation in the Developing Limb. *Developmental Biology*, **243**, 44–54.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Manceau M, Domingues VS, Mallarino R, Hoekstra HE (2011) The Developmental Role of Agouti in Color Pattern Evolution. *Science*, **331**, 1062–1065.

Mech D, Boitani L (2003) Wolves: behavior, ecology, and conservation. (eds Mech D, Boitani L). The University of Chicago Press, Chicago, Illinois.

Moser M, Binder O, Wu Y *et al.* (2003) BMPER, a Novel Endothelial Cell Precursor-Derived Protein, Antagonizes Bone Morphogenetic Protein Signaling and Endothelial Cell Differentiation. *Molecular and Cellular Biology*, **23**, 5664–5679.

Mullen LM, Hoekstra HE (2008) Natural selection along an environmental gradient: a classic cline in mouse pigmentation. *Evolution*, **62**, 1555–1570.

Muñoz-Fuentes V, Darimont C, Wayne R, Paquet P, Leonard J (2009) Ecological factors drive differentiation in wolves from British Columbia. *Journal of Biogeography*, **36**, 1516–1531.

Musiani M, Leonard JA, Cluff HD *et al.* (2007) Differentiation of tundra/taiga and boreal coniferous forest wolves: genetics, coat colour and association with migratory caribou. *Molecular Ecology*, **16**, 4149–4170.

Nie X, Luukko K, Kettunen P (2006) BMP signalling in craniofacial development. *The International Journal of Developmental Biology*, **50**, 511-521.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Reviews Genetics*, **39**:197–218. Novembre J, Rienzo AD (2009) Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics*, **10**, 745–755.

O'Keefe FR, Meachen J, Fet EV, Brannick A (2013) Ecological determinants of clinal morphological variation in the cranium of the North American gray wolf. *Journal of Mammalogy*, **94**, 1223–1236.

Okada I, Hamanoue H, Terada K *et al.* (2011) SMOC1 Is Essential for Ocular and Limb Development in Humans and Mice. *American Journal of Human Genetics*, **88**, 30–41.

Oksanen J, Blanchet FG, Kindt R *et al.* (2013) Package 'vegan', *R Repository*.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)*, **20**, 289–290.

Park U-H, Seong M-R, Kim E-J *et al.* (2014) Biochemical and Biophysical Research Communications. *Biochemical and Biophysical Research Communications*, **443**, 489–494.

Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. *PLoS Genetics*, **2**, e190.

Peakall R, Smouse PE (2006) Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Pilot M, Greco C, vonHoldt BM *et al.* (2014) Genome-wide signatures of population bottlenecks and diversifying selection in European wolves. *Heredity*, **112,** 428–442.

Price A, Patterson N, Plenge R *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

Primmer CR, Papakostas S, Leder EH, Davis MJ, Ragan MA (2013) Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Molecular Ecology*, **22**, 3216–3241.

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

Pujolar JM, Jacobsen MW, Als TD *et al.* (2014) Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, **23**, 2514–2528.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Pyhäjärvi T, Hufford MB, Mezmouk S, Ross-Ibarra J (2013) Complex patterns of local adaptation in teosinte. *Genome Biology and Evolution*, **5**, 1594–1609.

Qanbari S, Pausch H, Jansen S *et al.* (2014) Classic Selective Sweeps Revealed by Massive Sequencing in Cattle (JK Pritchard, Ed,). *PLoS Genetics*, **10**, e1004148.

Reimand J, Arak T, Vilo J (2011) g:Profiler--a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, **39**, W307–W315.

Reimand J, Kull M, Peterson H, Hansen J, Vilo J (2007) g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, **35**, W193–W200.

Roux I, Safieddine S, Nouvian R *et al.* (2006) Otoferlin, Defective in a Human Deafness Form, Is Essential for Exocytosis at the Auditory Ribbon Synapse. *Cell*, **127**, 277–289.

Sabeti PC, Varilly P, Fry B *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Sansregret L, Nepveu A (2008) The multiple roles of CUX1: Insights from mouse models and cell-based assays. *Gene*, **412**, 84–94.

Sazzini M, Schiavo G, De Fanti S *et al.* (2014) Searching for signatures of cold adaptations in modern and archaic humans: hints from the brown adipose tissue genes. *Heredity*, **113**, 259–267.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, **78**, 629–644.

Schmutz SM, Berryere TG, Dreger DL (2009) MITF and White Spotting in Dogs: A Population Study. *Journal of Heredity*, **100**, S66–S74.

Slater GJ, Dumont ER, Van Valkenburgh B (2009) Implications of predatory specialization for cranial form and function in canids. *Journal of Zoology*, **278**, 181–188.

Smith CP, Fenton RA (2007) Genomic Organization of the Mammalian SLC14a2 Urea Transporter Genes. *Journal of Membrane Biology*, **212**, 109–117.

Staubach F, Lorenc A, Messer PW *et al.* (2012) Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus). *PLoS Genetics*, **8**, e1002891.

Storz JF (2007) Hemoglobin function and physiological adaptation to hypoxia in high-altitude mammals. *Journal of Mammalogy*, **88**, 24–31.

Sturm RA, Duffy DL (2012) Human pigmentation genes under environmental selection. *Genome Biology*, **13**, 248.

Tietjen I, Hovingh GK, Singaraja RR *et al.* (2012) Segregation of LIPG, CETP, and GALNT2 Mutations in Caucasian Families with Extremely High HDL Cholesterol (H Schunkert, Ed,). *PLoS ONE*, **7**, e37437.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E (2011) Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. *PLoS Genetics,* **7**, e1002316.

vonHoldt BM, Pollinger JP, Earl DA *et al.* (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Research*, **21**, 1-12.

vonHoldt BM, Pollinger JP, Lohmueller KE *et al.* (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, **464**, 898–902.

Wan J, Masuda T, Hackler L *et al.* (2011) Dynamic usage of alternative splicing exons during mouse retina development. *Nucleic Acids Research*, **39**, 7920–7930.

Wehrle-Haller B (2003) The role of Kit-ligand in melanocyte development and epidermal homeostasis. *Pigment cell research / sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society*, **16**, 287–296.

Weir B, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wright S (1951) Genetical structure of populations. *Annual Eugenics*, **166**, 247–249.

Wycisk KA, Zeitz C, Feil S *et al.* (2006) Mutation in the auxiliary calcium-channel subunit CACNA2D4 causes autosomal recessive cone dystrophy. *American Journal of Human Genetics*, **79**, 973–977.

Xu S, Yang Y, Zhou X *et al.* (2013) Adaptive evolution of the osmoregulation-related genes in cetaceans during secondary aquatic adaptation. *BMC Evolutionary Biology*, **13**, 1–9.

Yamada R (2003) Cell-autonomous involvement of Mab21l1 is essential for lens placode development. *Development*, **130**, 1759–1770.

Yamauchi T, Nio Y, Maki T *et al.* (2007) Targeted disruption of AdipoR1 and AdipoR2 causes abrogation of adiponectin binding and metabolic actions. *Nature Medicine*, **13**, 332–339.

Yang Y (2003) Wnt5a and Wnt5b exhibit distinct activities in coordinating chondrocyte proliferation and differentiation. *Development*, **130**, 1003–1015.

Yang J, Wang ZL, Zhao XQ *et al.* (2008) Natural selection and adaptive evolution of leptin in the ochotona family driven by the cold environmental stress. *PLoS ONE*, **3**, e1472.

Yasunaga S, Grati M, Cohen-Salmon M *et al.* (1999) A mutation in OTOF, encoding otoferlin, a FER-1-like protein, causes DFNB9, a nonsyndromic form of deafness. *Nature Genetics*, **21**, 363–369.

Yi X, Liang Y, Huerta-Sanchez E *et al.* (2010) Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science*, **329**, 75–78.

Zhang W, Fan Z, Han E *et al.* (2014) Hypoxia Adaptations in the Grey Wolf (Canis lupus chanco) from Qinghai-Tibet Plateau (JM Akey, Ed,). *PLoS Genetics*, **10**, e1004466.

Targeted capture and resequencing of 1040 genes reveal environmentally driven functional variation in gray wolves

ABSTRACT

In an era of ever-increasing amounts of genomic sequence data for individuals and populations, the utility of traditional single nucleotide polymorphisms (SNPs) array-based genome scans is uncertain. We previously performed a SNP array-based genome scan to identify candidate genes under selection among six distinct gray wolf (*Canis lupus*) ecotypes. Using this information, we designed a targeted capture of 1040 genes, including all exons and flanking regions, as well as 5000 1kb non-genic neutral regions and resequenced these regions in 107 wolves. Selection tests revealed striking patterns of variation within candidate genes relative to non-candidate regions and identified potentially functional variants related to local adaptation. We found 27% and 47% of candidate genes from the previous SNP array study had functional changes that were outliers in our analyses with `SweeD` and `Bayenv`, respectively. This result verifies the use of genome wide SNP surveys to tag genes that contain functional variants between populations. We highlight non-synonymous variants in *APOB*, *LIPG*, and *USH2A* that occur in functional domains of these proteins, and that demonstrate high correlation with precipitation seasonality and vegetation. We find Arctic and High Arctic wolf ecotypes have higher numbers of genes under selection, which highlight their conservation value and heightened threat due to climate change. This study demonstrates that combining genome wide genotyping arrays with large scale resequencing and environmental data provides a powerful approach to discern candidate functional variants in natural populations.

Introduction

The development of genome wide genotyping and sequencing technology has provided new high-resolution tools for exploring adaptation at the molecular level (reviewed in Perry 2014). Recent empirical and theoretical advancements have focused on signals of selection in multi-locus data sets, which are currently enabled by the availability of whole genome polymorphism data collected at relatively low cost (Li *et al.* 2014). Such “genome scans” are important for identifying regions of the genome that are tagged by divergent SNPs that may be located in or in linkage disequilibrium with genes under selection. However, these molecular studies of adaptation in natural populations may suffer from ascertainment bias if the SNP array was not designed for the same species. Furthermore, these studies often end with lists of candidate genes under selection that are not interrogated by resequencing (Scheinfeldt & Tishkoff 2013). As a result, potential functional variants are not identified that support the role of specific genes in adaptation. New DNA tools, such as the capture array (Hodges *et al.* 2007; Tewhey *et al.* 2009; Gnirke *et al.* 2009), allow for the enrichment in a DNA sample of specific gene regions for hundreds of candidate genes. This targeted enrichment, when followed by high-coverage next generation sequencing and careful quality control, can be used to confirm signals of selection (Burbano *et al.* 2010; Albert *et al.* 2011; Domingues *et al.* 2012) and pinpoint potential functional mutations (e.g. Ng *et al.* 2009; Bi *et al.* 2013). We aim to demonstrate that a genome scan followed by extensive resequencing is a synthetic approach to understanding local adaptation in non-model organisms.

The gray wolf (*Canis lupus*) was historically one of the most widespread mammals in North America (Leonard *et al*. 2005), with the ability to disperse large distances > 1000 km

(Wabakken *et al.* 2007). Despite their high mobility, wolves show remarkable morphologic and genetic differentiation at a local scale (Carmichael *et al.* 2007; Musiani *et al.* 2007; vonHoldt *et al.* 2011; O'Keefe *et al.* 2013; Schweizer et al, submitted). North American wolves are subjected to strong environmental gradients involving dramatic changes in temperature, precipitation, and vegetation between British Columbia and Arctic ecotypes (Schweizer *et al*, submitted). We previously analyzed 42,036 SNPs genotyped on the Affymetrix canid v2 SNP array and environmental data to explore local adaptation in wolf ecotypes (Schweizer *et al*, submitted). We identified six environmentally and genetically distinct wolf ecotypes: West Forest, Boreal Forest, Arctic, High Arctic, British Columbia, and Atlantic Forest. Based on results from three complementary selection tests and a review of the current literature, we identified candidate 1040 genes potentially under selection for confirmation using a capture array and resequencing approach.

We hypothesized that genes related to immunity, metabolism, morphology, pigmentation, and sensory functions have been preferentially selected in ecotypes. First, we expected immune challenges to vary with environment given the observed positive relationship between temperature and precipitation and pathogen persistence (Allen *et al.* 2002; Guernier *et al.* 2004; Dionne *et al.* 2007), perhaps due to higher metabolic rate, shorter life cycle, and increased density of parasites in warmer climates (Allen *et al.* 2002; Guernier *et al.* 2004). Second, we predicted that metabolic differences would be prevalent across ecotypes. Arctic species often have unique adaptations to survive in freezing temperatures, such as active sodium transport to maintain body heat (Stevens & Kido 1974), insulin resistance (Martin 2008; Odegaard & Chawla 2013), lipid metabolism or temperature sensitivity (Lynch *et al*. 2015). Additionally, changes in diet associated with different prey type availability have implications for increased lipid levels in the bloodstream and tolerance thereof (Akey *et al.* 2002; Liu *et al.* 2014; Clemente *et al.* 2014). A third prediction was that wolves would demonstrate local adaptation through morphology. Differences in terrain and prey type associated with migratory and non-migratory ecotypes (Musiani *et al*. 2007) may result in divergent selection for muscular/skeletal traits among ecotypes given differences in prey pursuit and acquisition in specific environmental contexts (MacNulty *et al.* 2009; Slater *et al.* 2009). Previous morphologic studies have shown that wolf skeletal features correspond with environmental and habitat differences (O'Keefe *et al.* 2013).

A fourth prediction concerned hair pigmentation, which varies geographically, with paler and whiter pelage more common in Northern regions (Gipson *et al.* 2002; Musiani *et al.* 2007; Anderson *et al.* 2009). A higher frequency of melanistic wolves is found in southern latitudes (Musiani *et al.* 2007; Anderson *et al.* 2009), and melanism is caused in some populations by a mutation in a beta defensin gene (*CBD103*) that also confers fitness advantages (Anderson *et al.* 2009; Coulson *et al.* 2011). A brownish tinge to gray wolves has been observed in coastal British Columbia wolves at a higher frequency than mainland gray wolves (Darimont & Paquet 2000), and may have some advantage for camouflage from prey (Jolicoeur 1959) or a secondary advantage due to epistasis, as may be the case for *CBD103* (Anderson *et al.* 2009). Evidence of local adaptation related to coat pigmentation is found in numerous other species (Hoekstra 2006; Hubbard *et al.* 2010). Our final prediction for local adaptation in wolf ecotypes was that genes affecting vision, hearing, and olfaction would be under selection. Wolves depend on their senses for socializing and prey capture (Mech 1970), and live and hunt in areas with differing light levels. The ability to find prey in dense or varied vegetation may require greater visual sensitivity, hearing and olfaction than in more open tundra environments.

In this study, we tested the utility of SNP based genome scans to tag genes under selection by resequencing candidates in wolves across an environmental gradient to confirm selection signals and pinpoint functional mutations. We supplemented this effort by sequencing additional candidate genes that were not previously tagged by SNPs in the genome scan, but which existing literature suggests may be functional in natural populations. We used a custom capture array to resequence 1040 candidate genes, including their exons and putative regulatory regions, in 107 wolves. With each of three selection tests, we used 5Mb of non-genic sequence to empirically control for genetic patterns due solely to background demography. We verified that up to 47% of candidate genes from the SNP array selection scan are outliers in the same or similar statistical tests using our sequence data and contain potentially functional mutations. We find significant clinal variation of missense SNPs that corresponds with environment. Using available protein databases, we highlight mutations in three genes (*APOB*, *LIPG*, *USH2A*) that appear to be under selection and in functional protein domains. We argue for more conservation focus on Arctic and High Arctic wolves because they demonstrate a high diversity of unique adaptations, yet are some of the most threatened ecotypes due to climate change.

Methods

Re-sequencing of candidate regions with capture array

Our capture array was designed to bind sequences from 1042 candidate genes. Of the total, 520 of the genes were outliers identified in previous SNP-based selection scans (Schweizer, *et al.* submitted), a total of 60 genes were *a priori* candidate genes based on a literature search, and the remainder were from previous iterations of the SNP genome scan that were modified after the capture array was designed. We chose genes implicated in function or disease that could conceivably be under selection in natural populations, such as those related to olfaction, immunity, thermoregulation, and morphology (Table 2-S1). The exons, plus 1000 bp upstream of each gene promoter, were targeted with unique 120bp RNA baits every 60 bp. Using the dog genome annotation, we also designed 5 000 1 kb regions (which we call neutral regions) with the aim to minimize effects of selection by maximizing the distance from annotated genes (>100kb), uniqueness within the genome, and several other genomic characteristics described previously (Freedman *et al.* 2014). These regions were intended to provide background neutral variation for downstream selection tests. Baits were designed by MYcroarray (Ann Arbor, Michigan) with a total targeted sequence length of ~8Mb.

Sample selection and library preps

North American wolves were previously sampled in a genome-wide selection scan based on SNP array genotyping (N=111, Schweizer *et al*. submitted). For this study, we re-extracted DNA from 78 of the same individuals for which blood or skin tissue sample remained, and selected an additional 39 individuals from similar geographic areas (Figure 2-1) (Carmichael *et al.* 2007; Musiani *et al*. 2007). All 117 of these individuals have known geographic locations (Schweizer et al, submitted), and 47 have coat color phenotype information ((Musiani *et al.* 2007); Denali National Park).

Genomic DNA was extracted using a Qiagen Mini Prep kit, then sheared using a Biorupter NGS Sonication System (Diagenode). Sequencing libraries were prepared following a with-bead library preparation protocol (Faircloth 2015), and samples were labeled with a unique 6bp index during adapter ligation to enable pooling of 24 individuals per lane (Faircloth & Glenn 2012). Libraries were target enriched and PCR amplified according to the MYbaits protocol (MYcroarray) after a 24-hour hybridization. Libraries were 100bp paired-end sequenced on a HiSeq 2000.

Sequence alignment and processing followed the general recommendations of the Broad Institute GATK v2.6-4 "Best Practices" pipeline (https://www.broadinstitute.org/gatk/guide/best-practices; see Supplemental Information for details). Reads were mapped to the reference dog genome (*Canis familiaris*; CanFam3.1) since previous work demonstrates that aligning wolf sequences to this reference produces high quality genotype calls and minimal reference bias due to the very short sequence divergence of wolves and dogs (~0.1%; Freedman, *et al*. 2014).

Variant filtering and final sample set

Variants were filtered with `GATK VariantFiltration` using 10 filter expressions, as recommended by the GATK "Best Practices" pipeline, as well as depth of coverage ≥10 and minimum genotype quality ≥ 30. Quality of sequence data was assessed using the `vcftools` package (Danecek *et al.* 2011), and we subsequently chose a minimum genotyping call rate of 95% for further analysis. Kinship among individuals was calculated using a linkage disequilibrium (LD) pruned set of neutral variants (using the --indep-pairwise 50 50 0.5 option in `PLINK` (Purcell *et al.* 2007), as in Schweizer et al, submitted) and `KING`, which accounts for population structure (Manichaikul *et al.* 2010). To remove related individuals, we used PRIMUS (Staples *et al.* 2012) and a maximum pairwise identity-by-descent of 0.1 (Fu *et al.* 2012).

Ecotype assignment of unrelated individuals was verified by both STRUCTURE v2.3.4 (Pritchard *et al.* 2000; Falush *et al.* 2003) and ADMIXTURE v1.23 (Alexander *et al.* 2009). We ran 10 independent runs of STRUCTURE, each with 20 000 burn-in and 50 000 sampling iterations for $K = 1$ through 10 with correlated allele frequencies under the admixture ancestry model, using a set of 28,195 LD-pruned neutral variants. The *greedy* algorithm within CLUMPP v 1.1.2 (Jakobsson & Rosenberg 2007) was used to control for variation in cluster labels across the 10 iterations of STRUCTURE. As an alternative clustering method, ADMIXTURE was run using the same data set. Ecotypes of these individuals were based on concordant STRUCTURE or ADMIXTURE assignments.

Variant annotation & gene annotation

Functional variants within genic regions were identified and annotated using the dog reference within Ensembl's Variant Effect Predictor (VEP) pipeline (release 78) (McLaren *et al.* 2010). The program, Sorting Intolerant From Tolerant (SIFT), which uses sequence alignment conservation across multiple species to identify the potential impact of mutations within coding regions (Ng & Henikoff 2003), was implemented through the VEP software. SIFT scores can be used to rank non-synonymous mutations as deleterious (score <0.05) or tolerated (score≥0.05) as an indication of potential functional impact. Given that non-synonymous mutations may have a functional impact even if they do not occur at highly conserved coding positions, we also used the Miyata score of biochemical similarity (Miyata *et* *al.* 1979). A Miyata score ≥ 1.85 means the amino acid substitution is significantly different in terms of biochemistry or size, and is an alternative to protein prediction algorithms such as `PolyPhen` that are designed for humans (Adzhubei *et al.* 2010; Marsden, *et al*, in prep).

Putative transcription factor binding sites (TFBS) were identified within genic regions using the profiles in the `JASPAR PHYLOFACTS` database (http://jaspar.genereg.net/). This database contains count matrices of conserved motifs in human, mouse, rat and dog originally identified by Xie and colleagues (2005). The motifs were converted to probability weight matrices and used with the motif finding program `FIMO` (Grant *et al.* 2011) (part of the `MEME` package v4.8.1: http://meme.sdsc.edu) to find matching occurrences in our sequence data.

Gene lists containing non-synonymous, deleterious non-synonymous, 5’ UTR, 3’UTR, or TFBS mutations were tested for enrichment of Gene Ontology (GO) categories by using the `R` 3.1.3 (http://www.R-project.org) package `gProfileR`, (Reimand *et al.* 2007; 2011), with “strong” hierarchical filtering and a Benjamani-Hochberg (BH) false discovery rate (FDR) correction to correct for multiple testing. A list of all genes sampled on the capture array was used as a statistical background for testing enrichment. GO category enrichment of outlier genes from `SweeD`, `BayeScan`, and `Bayenv` were also tested in this manner (see below).

Detection of regions of selective sweeps

We applied the site frequency spectrum (SFS)-based method of Nielsen et al. (2005), as implemented in the software `SweeD` (Pavlidis *et al.* 2013). This model detects selective sweeps from genomic SNP data using a composite likelihood ratio test to choose between neutral or selective sweep models and has the benefit that the null hypothesis is derived from the background pattern in the data itself. By comparing specific allele SFS to the empirical average, the location and magnitude of a selective sweep can be estimated (Nielsen *et al.* 2005).

For each ecotype, we ran SweeD on the data from neutral regions and genic regions separately, using a grid size of 10 000 and the option "strictPolymorphic." The P-value of each genic position likelihood score was determined by calculating the empirical percentile according to the distribution of likelihood values within the neutral regions using R, and P-value correction for multiple testing was achieved through a BH correction. A FDR threshold of 0.05 was used not as a strict threshold, but rather as a parameter to assign especially high support for outliers. Through this approach, we aimed to correct for neutral population demographic history without the assumptions of simulating data since there is no prior demographic model available.

In order to identify the genes nearest each grid position in the output from SweeD, we used BEDTools v2.21.0 (Quinlan & Hall 2010) to intersect the positions with the Ensembl annotation gene set (CanFam3.1, Ensembl release 79, March 2015), allowing a 6kb buffer on either side (N. Alachiotis, pers. comm.). Only genes that overlapped those queried by the capture array were annotated. We chose a P-value ≤0.01 for higher stringency, given some recent studies showing high false positive rate in SweeD under certain scenarios (Crisci *et al.* 2013). Using ANGSD (Nielsen, *et al*. 2012), we also generated unfolded SFS to verify patterns of allele frequency variation of the top 5% of genes from SweeD relative to neutral and genic regions (Appendix I).

Directional selection detection

To assess directional selection in each ecotype, we used the Bayesian method implemented in `BayeScan` v2.1 (Foll & Gaggiotti 2008). `BayeScan` tests whether the subpopulation-specific allele frequencies are significantly different from those within the common gene pool. Significance is assigned by a measure of support for a model in which selection explains allele frequency differences among populations versus a null model. In populations where isolation by distance is present, as is the case here (Schweizer et al, submitted), `BayeScan` can have high false positive rates unless a large set of neutral loci are used to generate empirical P-values (Lotterhos & Whitlock 2014). Therefore, we ran `BayeScan` separately for the neutral and genic regions, with prior odds of 10000 and 1000, respectively, and calculated empirical P-values of alpha, as we did for `SweeD`.

Environmentally correlated selection

To understand the effect of varying environments on clinal genetic variation in allele frequencies across North American wolf ecotypes, we implemented `Bayenv` (Coop *et al.* 2010). With `Bayenv`, we measured the support for a model in which SNPs covary linearly with an environmental variable over a model in which SNPs vary according to neutral expectation (Coop *et al.* 2010). We used 10 000 random neutral variants to generate a covariance matrix from the average of every 20 000th iteration over a total of 500 000 iterations. For each genic SNP, the selection mode of `Bayenv` was run with 100 000 iterations and 12 environmental variables previously shown to be influential in wolf ecotype differentiation (Schweizer, *et al*. submitted). The 12 variables were obtained from the WorldClim database as previously described (Hijmans *et al*. 2005; Schweizer *et al*. submitted), and measure temperature (annual mean temperature, mean diurnal temperature range, temperature seasonality, maximum temperature of warmest month, minimum temperature of coldest month), precipitation (annual precipitation, precipitation seasonality, precipitation of coldest quarter), vegetation (percentage tree cover, normalized difference vegetation index, and land cover category), and altitude. Each environmental variable was normalized as recommended (Coop *et al.* 2010). The final matrix of Bayes Factors (BF) was obtained by averaging each BF over a total of 10 independent runs in order to help control for sensitivity of MCMC sampling methods (Coop *et al.* 2010; Blair *et al.* 2014). The same procedure was done for a set of 15 000 random neutral variants in order to further control for background demographic patterns. We assigned an empirical P-value within `R` to the log10 BF of each genic variant using the neutral distribution. This approach has been shown to reduce falsely elevated BFs that may occur given the pure drift null model inherent within `Bayenv` (Hancock *et al.* 2010; Coop *et al.* 2010; Chen *et al.* 2012a; Lotterhos & Whitlock 2014). Variants with a BH-corrected FDR≤0.05 were highlighted as having additional support of being selection candidates.

Using the functional consequences of variants and SIFT scores annotated by `VEP`, we tested for significant excess of functional variants (missense, synonymous, or stop gained), regulatory variants (5'UTR, 3'UTR, or splice), or damaging variants (SIFT score <0.05) in each set of outlier loci. We performed a Fisher's exact test for count data in `R` using significance thresholds of P-value ≤0.05, P-value ≤0.005, and BF≥3 (Kass & Raftery 1995).

In order to assess how well we could assign individuals to an ecotype based solely on a subset of genotyped SNPs, we implemented a tree classification method by way of a random forest model (`randomForest` package in `R`; Liaw & Wiener 2002). The test uses a subset of individuals to train a model based on genotype data, then uses the remaining individuals to evaluate how well the model correctly assigns individuals to their population. This was repeated 5

000 times for significance. We used a set of both the significant missense SNP with BF>1 and a smaller set of SNPs located in genes related to known phenotypic traits (see Results).

Patterns of clinal variation in allele frequency

We also explored genetic and geographic patterns of variation in allele frequency for significant outliers from `Bayenv`. For variants with an uncorrected P-value ≤ 0.05, we plotted the average allele frequency of the reference nucleotide within each ecotype versus the average of the significant environmental variable within each ecotype. Best-fit linear model lines were plotted and the coefficient of correlation was calculated using the Pearson method. For the genic SNPs, we also identified the ancestral and derived alleles, when possible, by using allelic variation within Israeli wolf, Croatian wolf, Chinese wolf (all *Canis lupus* species), and Israeli golden jackal (*Canis aureus*) as an outgroup (Freedman *et al.* 2014).

Overlap assessment between outliers from capture array and SNP array

In order to gauge the utility of genome scans from SNP arrays in predicting candidates for selection, we examined the overlap of significant outliers between outliers on the capture and the Affymetrix SNP array (Schweizer et al, submitted), using only the genes that were assayed with both methods (n=739). We compared overlap within `Bayenv`, and `BayeScan`, and between `SweeD` and F_{ST}*/XP-EHH*. Although the approaches are different between `SweeD` (SFS-based) and F_{ST}*/XP-EHH* (haplotype-based) (Sabeti *et al.* 2007), both methods should identify regions containing genes that have been swept to high frequency as a result of selection.

Genotype association with coat color

Using coat color information from 47 individuals (8 black, 17 gray, 22 white) from among the West Forest, Boreal Forest, and Arctic ecotypes, we tested for significant associations between black or white coat color and each of the 13k genic SNPs using the variance component model within EMMAX (Kang *et al*. 2010). The set of LD-pruned neutral SNPs was used to calculate a Balding-Nichols kinship matrix, and genic SNPs were pruned for minor allele frequency ≥10% (Kang *et al*. 2010). Multiple testing P-value correction was performed within R.

Protein Models

For candidate genes that had publicly available protein structure information (see Results), we explored the effect of functional variants on structure. We extracted the coding sequence from the reference dog genome using custom Python scripts, translated the sequence to amino acids using phase information from Ensembl with *ExPASy* (Gasteiger 2003), then aligned the protein sequences to human annotated versions within Geneious *8.1.3* (Kearse *et al.* 2012) and determined where allelic variation occurred in our wolves. We also modeled functional impact on three-dimensional protein structure by using SWISS-MODEL (Arnold *et al.* 2006) to identify similar templates and model structures.

Results

Sequencing summary

The overall sequencing quality was high, with per-individual average unfiltered yield of 1,889.83 Mb ± 567.42 Mb, an average 88.91% ± 3.52% reads passing Illumina filters, and an average mean quality of 34.5 ± 0.88. After processing and removing low quality reads, an average of 89% ± 14% reads mapped to the dog reference genome and 86 % ± 6% of all reads mapped uniquely to the dog reference genome (i.e. after PCR duplicate removal). After genotyping and additional filtering, the mean depth of coverage over all regions on the capture array was 154.78X ± 64.45X, with mean neutral coverage of 181.65X ± 72.95X and mean genic coverage 89.61X ± 31.22X (Figure 2-S1). After filtering, we identified 4,918,729 neutral positions and 2,129,544 genic positions, of which 39,376 and 13,092 were variable, respectively (Table 2-1). The transition to transversion rate was 2.32 for neutral regions and 4.17 for exonic regions (Table 2-1), both of which are similar to values in wolves and humans (DePristo *et al.* 2011; Freedman *et al.* 2014; Zhang *et al.* 2014). Genotype quality was assessed by comparing genotypes for 198 SNP positions overlapping with the Affymetrix SNP array and the capture array target regions in 78 identical individuals (Schweizer *et al*. submitted). Genotyping concordance was higher than 99.5% (Table 2-S2).

After removal of two related individuals, ecotype assignment was confirmed for all individuals by concordant assignment in `Structure` and `Admixture`. Eight individuals were removed from further analyses since neither `Structure` nor `Admixture` could assign them to a single ecotype with >50% assignment. The remaining set of 107 individuals included 31 West Forest, 26 Boreal Forest, 30 Arctic, 6 High Arctic, 5 British Columbia, and 9 Atlantic Forest wolves (Figure 2-1).

Variant annotation & GO enrichment

`VEP` annotated a total of 13,092 variants (Table 2-1). GO enrichment analysis of genes containing functional variants (missense, deleterious missense, 5’UTR, 3’UTR, and TFBS) identified 80, 30, 50, 280, and 113 significantly enriched categories, respectively (BH-corrected

P-value ≤ 0.05). We focused on GO categories with a minimum of five genes at the highest hierarchical level, and found that four out of 31 categories overlapped between functional variant category types (Figure 2-S2). The most significantly enriched GO category was "detection of chemical stimulus involved in sensory perception" (P-value: 3.92e-06) in missense variants, with the next two most significant categories in related categories of "olfactory receptor activity" and "detection of stimulus involved in sensory perception". Within human phenotype categories, we identified 17 categories with a minimum of five genes, but no significant categories in 5'UTR variants (Figure 2-S3).

Candidate sweep regions

Using `SweeD`, we identified candidate selective sweep regions putatively under selection in each wolf ecotype (Figure 2-2; Figures 2-S4-S9). GO enrichment of significant outliers (P-value ≤ 0.01) with a minimum of two genes overlapping identified four significant categories at the highest hierarchy. "Defense response" was enriched in High Arctic wolves (P-value: 0.05), and cellular-related categories were enriched in Arctic, Atlantic Forest, and British Columbia wolves (Figure 2-S10). Human phenotype categories demonstrated enrichment of genes related to "round face" and "short neck" in Boreal Forest wolves, and "infantile onset" and "aplasia/hypoplasia involving the central nervous system" in Arctic wolves (Figure 2-S11). Arctic and High Arctic wolves had the highest numbers of candidate genes at this threshold and the highest number of significantly enriched GO-related categories (Figure 2-2A), as well as the highest numbers of unique candidate genes (Figure 2-2B). Furthermore, Arctic and High-Arctic ecotypes had the highest number of micro RNA categories (Figure 2-2A), and showed a high proportion of low-frequency and high-frequency derived alleles, relative to neutral regions (Figure 2-S12)

Within `SweeD` results, we focused on significant missense variant positions (P-value ≤0.05), since the functional effects are more directly interpretable, although many more variant types were identified in significant genes (Figures 2-S4-S9). We identified 25 genes (57 missense variants) in West Forest wolves, 29 genes (77 missense variants) in Boreal Forest wolves, 34 genes (112 missense variants) in Arctic wolves, 24 genes (78 missense variants) in High Arctic wolves, 25 genes (101 missense variants) in British Columbia wolves, and 34 genes (96 missense variants) in Atlantic Forest wolves.

There were several notable outlier genes from `SweeD`. *APOB* (*Apolipoprotein B*), which controls plasma cholesterol levels in a wide range of species (Farese *et al.* 1995), was an outlier in British Columbia wolves (Figure 2-S8). In Arctic wolves, a selective sweep region was centered on a candidate gene for hearing and vision, *PCDH15* (*protocadherin 15*) (Figure 2-S6). *PCDH15* has been implicated in a hearing disorder called Usher's Syndrome, which also has associated vision problems (Alagramam *et al.* 2001; Le Guédard *et al.* 2007). In both Arctic and Atlantic Forest wolves (Figure 2-S6 & S9), the olfactory receptor gene *OR6B1* was a significant outlier with missense mutations. Two canine beta-defensins, *CBD102* and *CBD1*, were highly ranked in West Forest and High Arctic wolves, respectively (Figure 2-S4 & 7). Canine beta-defensins represent a class of immunity genes that have also been recently characterized as ligands that are involved in the melanin pathway (Candille *et al.* 2007). The MHC class II gene, *DLA-DQA*, was also highly ranked in Atlantic Forest wolves (Figure 2-S9). The MHC complex, specifically those genes within class II cluster, are key to the genetic response to immune challenges (Wagner *et al.* 1996).

We observed three candidate genes for mammalian pigmentation that were significant outliers in SweeD. *TYR* (*Tyrosinase*), which encodes the rate-limiting enzyme that converts tyrosine to melanin within the pigmentation pathway, has been implicated in oculocutaneous albinism (light pigmentation of hair, eyes, and skin) in humans (Sturm & Duffy 2012), and was an outlier in Boreal Forest and Arctic wolves (Figure 2-S5 & 6). *TYRP1* (*Tyrosinase-related protein 1*), which was an outlier within British Columbia wolves (Figure 2-S8), can cause brown or white color in dogs and mice (Nakamura *et al.* 2002; Kaelin & Barsh 2013). *MLPH* (*melanophilin*) was an outlier in Atlantic Forest wolves, with a p-value<0.01 (Figure 2-S9), and mutations within *MLPH* have been associated with the “dilution” phenotype, in which eumelanin pigment appears diluted to silver or blue-like colors (Hume *et al.* 2006).

Directional selection with BayeScan

Our analysis with BayeScan identified 3 SNPs with a FDR≤0.05. One significant SNP causes a synonymous amino acid change in *UACA* (*Uveal Autoantigen With Coiled-Coil Domains And Ankyrin Repeats*), a gene that regulates apoptosis in response to stress, and has been implicated in multiple vision-related disorders (Yamada *et al.* 2001; Ohkura *et al.* 2004). Another significant SNP is located in an intron of *ATP10B* (*ATPase, Class V, Type 10B*), a gene involved in phospholipid translocating which is significantly associated with coronary artery disease and degree of atherosclerosis (Nolan *et al.* 2012). The remaining significant SNP was intergenic. GO enrichment of the two genes did not identify any significant categories with more than one gene overlap.

Environmentally correlated missense SNPs

Through the `Bayenv` method, we focused on the effect of 12 environmental variables on missense and TFBS variants. Only deleterious missense variants (i.e. with SIFT score ≤0.05 and p-value <0.005 or BF>3) were enriched in temperature seasonality and precipitation seasonality. We did not find significant enrichment of broader functional or regulatory mutations in any other environmental variables.

GO analysis of genes with significant variants (p-value ≤0.005) in `Bayenv` identified significant enrichment in multiple ecologically relevant top-level categories, including those related to vision, hearing, immunity, and homeostasis (Figure 2-3). There was overlap of GO categories among similar types of environmental variables (i.e. vegetation, precipitation, or temperature) (Figure 2-3). Human phenotype category enrichment of the same set of genes revealed categories for hearing, vision, and bone development (Figure 2-S13).

Several missense mutations were highly and significantly correlated with environmental variables (Figure 2-4, Table 2-2). Four genes within the olfactory receptor family had missense mutations that were significant outliers from `Bayenv` analysis (Figure 2-4, Table 2-2). Although to our knowledge there is no previously published data available for the function of these genes (*OR4S2, OR6B1, OR5B17, ENSCAFG00000012139*), the olfactory receptor gene family aids in the recognition by olfaction sensory neurons of vaporous odorant molecules.

Three well-known candidate genes for coat coloration also ranked highly in `Bayenv` (Table 2-2). Missense variants within *TYR* and *TYRP1* were both outliers. We did not find any missense mutations in *CBD103*, but this was not expected given previous studies showing that a 3bp deletion causes black coat color (Candille *et al.* 2007). We did find that an intron variant within 596 bp of the 3bp deletion was in perfect linkage (D'=1; Figure 2-4) and was significantly associated with maximum temperature of warmest month (BF=4.2; P-value=0.0037) and land cover type (BF=2; P-value=0.0076).

Two genes related to lipid metabolism also contain high-ranking missense mutations. *APOB* contained multiple missense mutations with a P-value≤0.05 (Figure 2-4, Table 2-2). *LIPG* (*Lipase, Endothelial*) is a second gene that regulates lipid levels and in which loss-of-function mutations lead to increased levels of HDL (Edmondson *et al.* 2009; Tietjen *et al.* 2012). The missense mutation in *LIPG* is most highly ranked in mean diurnal temperature range (Figure 2-4, Table 2-2).

Missense mutations in two genes implicated in vision and hearing were also highly ranked in `Bayenv`. Eight missense mutations occurred in *USH2A*, with four of them having BF >2 (Table 2-2). In humans, mutations in *USH2A* cause Usher Syndrome, which is characterized by hearing impairment and retinitis pigmentosa (also an outlier in SweeD above; Dreyer *et al.* 2000; Saihan *et al.* 2009). Here, we found significant correlation between a missense variant and the normalized vegetation difference index (Figure 2-4). We found three significant missense mutations within *PCDH15* as well (also an outlier in `SweeD` above; Figure 2-4, Table 2-2).

We identified multiple, high-ranking missense mutations within two immunity genes in the MHC complex (Figure 2-5, Table 2-2), including five within *DLA-DQA* and one within *DLA-DRB1* that were outliers for temperature, vegetation, and altitude variables. One missense mutation had a BF=164 for mean diurnal temperature range and was significant even after BH correction for multiple testing (p-value ≤ 10^{-5}).

We also examined whether any significant non-coding variants from `Bayenv` were located in putative TFBS, and found six variants (P-value ≤0.005) overlapping six genes (Table 2-3). Notably, we identified a high-ranking variant within a TFBS 567 bp upstream of *LEP* (*Leptin*), a gene that encodes a protein secreted from adipose tissue that is involved in obesity (Mammès *et al.* 1998). In humans, a 5' variant located 633bp upstream significantly associates with obesity (Li *et al.* 1999). A second potentially important variant was located in a putative TFBS for *FOXA3* (*Forkhead Box A3*) (Table 2-3). *FOXA3* is itself a transcription factor thought to control expression of multiple liver-related genes and differentiation of adipocytes (Xu *et al.* 2013) and glucose homeostasis (Shen 2001).

Finally, we determined whether these top-ranked SNPs could be used to correctly assign individuals to their ecotype, which may be informative for historical samples or those of unknown origin. We used two genotype data sets for this analysis. The first consisted of 121 missense SNPs with BF>1, and the second consisted of 31 SNPs (26 missense, 5 TFBS) that we discuss above based on their straightforward phenotypic relevance in wolves. Using the first data set, the random forest model made 22.43% errors in classifying individuals (Table 2-S3). Error in individual ecotype classification was low for West Forest and Arctic wolves. The model had more difficulty in classifying Boreal Forest and High Arctic wolves, but most often the incorrectly assigned individuals were classified as West Forest or Arctic, respectively, which are from similar types of habitats (Schweizer *et al*. submitted). Error rates were higher for British Columbia and Atlantic Forest wolves. For the second data set of 31 SNPs, the error rate from the random forest model was higher at 33.64%. Error rates were lowest for West Forest and Arctic ecotypes, although overall the model lost power to correctly assign individuals to their ecotype (Table 2-S4).

Patterns of clinal variation in allele frequency

Outlier genic SNPs from `Bayenv` showed large allele frequency differences across environmental variables (Figure 2-4, Figure 2-5, Table 2-2). Often, the High Arctic and British Columbia ecotypes were at opposite extremes of both the environmental variable and SNP allele frequency (Figure 2-4). Concordantly, we found that often the Boreal Forest and West Forest ecotypes have intermediate allele frequencies and environment (Figure 2-4).

For 19 of these outlier SNPs, we were able to infer the ancestral and derived alleles by comparing to previously sequenced wolf and golden jackal genomes (Table 2-2). For *LIPG*, *OR5B17*, *OR4S2*, *PCDH15*, and *TYR*, the Arctic and High Arctic ecotypes show an increase in derived allele frequency, with the greatest change occurring in *PCDH15* where Atlantic Forest wolves are almost fixed for the ancestral allele, and High Arctic wolves are fixed for the derived allele (Figure 2-4). In *APOB*, *DLA-DQA*, *OR4S2*, and *OR6B1*, we identified novel variants that were not previously observed in Old World wolves and a golden jackal (Freedman *et al.* 2014).

Selection test overlap assessment

We found relatively high overlap between significant genes with a P-value ≤0.05 for `SweeD`, `BayeScan`, and `Bayenv` (Figure 2-S14). Out of a total of 554 genes, 195 genes (35.4%) were common to two out of three methods and one gene was common to all three methods (*ATP10B*). For the former category, the majority (194/195) genes overlapped between `SweeD` and

Bayenv. Using a stricter threshold (SweeD P-value ≤0.01, BayeScan P-value ≤0.01, Bayenv BF ≥3), there were no genes common to all three methods (Figure 2-S14). There were, however, 28/233 genes (12.0%) common to Bayenv and SweeD, and 3/233 genes (1.3%) common to Bayenv and BayeScan. The 28 genes common to Bayenv and SweeD at this threshold included six candidate genes mentioned above: *AMOTL1*, *CBD1*, *CBD102*, *CBD103*, *MLPH*, *PCDH15*.

Capture array and SNP array overlap assessment

In order to assess how well our previous selection scan identified candidate genes (Schweizer, *et al*. submitted), we counted the overlap between candidate genes tagged by SNP array and sequenced by capture array. There were a total of 739 genes on the capture array that were within 10kb of a SNP on the Affymetrix dog SNP array. Bayenv performed the best, with 188/296 genes (47%) occurred in the top 5% rank in both platforms (Figure 2-S15). Selective sweep methods (SweeD and F_{ST}/XP-EHH) were also fairly concordant, with 73/270 (27%) genes overlapping, even though the analytical methods differed between the SNP array and the capture array (Figure 2-S15). No outlier genes from BayeScan on the SNP array were confirmed by gene sequencing on the capture array (out of six overlapping). We also observed cases in all three selection methods using resequencing data where the test identified significant genes that had been tagged by SNPs on the Affymetrix array but were not identified within the top 5% of genes (or FDR<0.05 for BayeScan) on the Affymetrix array.

Protein sequence models

To further explore the potential impact of selected variants on protein function, we chose three high-ranking genes, *APOB*, *LIPG*, and *USH2A*, for which protein domain structure and other relevant literature were readily available. For *APOB*, which is one of the most complex proteins in the genome with regard to exon structure, we focused on exon 26, which encodes the most important functional domains (Young 1990; Amrine-Madsen *et al.* 2003). Three missense mutations within *APOB* occurred in a region from AA 1425 to AA 1728 in humans (AA 1563 to AA 1866 in wolves) (Figure 2-S16A) which is crucial for forming buoyant triglyceride-rich LDL particles (Young 1990) and is a conserved OM channel domain (NCBI c121487). The mutations did not affect the type of side chain or size of amino acid. In *LIPG*, we identified a single missense mutation at position 420 causing an isoleucine (hydrophobic) to change to a threonine (polar). This mutation occurred within the PLAT domain of endothelial lipase (the protein encoded by *LIPG*) (Figure 2-S16B; Figure 2-6). Previous functional protein assays have demonstrated that endothelial lipase has a unique 23 AA region in the PLAT domain that is likely to be crucial to the unique capabilities of endothelial lipase to interface with HDL particles (Razzaghi *et al.* 2013), and found that our mutation occurred near the beginning of that 23 AA region (Figure 2-6). Finally, four highly ranked missense mutations occur within the longer isoform of *USH2A* (Figure 2-S16C). Two of these mutations, Ala2692Val and Asp2828Asn, occurred within a region of *USH2A* consisting of nine fibronectin type III domains (NCBI cd00063). We also identified a three base pair in-frame deletion (Ser1040del), predicted to be damaging by PROVEAN (Choi *et al.* 2012), within the functional laminin-type EGF-like motif domain (data not shown).

Genotype association with coat color

Using data from eight black, 17 gray, and 22 white wolves, we found significant associations with SNPs in pigmentation genes. In black wolves, eight SNPs had a corrected q-value≤0.05, and the most significant SNP in a pigmentation gene was an 3'UTR variant in *CBD103* (q-value: 0.02895). The other seven SNPs were within the selective sweep region for *CBD103* (Anderson *et al*. 2009). In white wolves, there were three significant SNPs (corrected q-value≤0.05), all within the 3'UTR region of *CBD103* (most significant q-value: 0.04467).

Discussion

Temperature-related variation in immune-related genes

We found missense variants within two MHC Class II genes, *DLA-DQA* and *DLA-DRB1*, that were significantly associated in frequency and heterozygosity with altitude, temperature, and percentage tree cover (Figure 2-5, Table 2-2). This followed our initial prediction that we would find variants in immunity genes among wolf ecotypes as a response to differences in pathogen prevalence at varying temperatures (Allen *et al.* 2002; Guernier *et al.* 2004; Dionne *et al.* 2007), and consequently we included multiple MHC and beta-defensin genes in the design of our capture array. Two beta-defensins, *CBD102* and *CBD103*, were also in sweep regions within Atlantic Forest and West Forest wolves, and an intron variant perfectly linked with the deletion causing black coat color was significantly associated with temperature and land cover variables. The deletion variant of *CBD103* had previously been highlighted in Yellowstone and Canadian wolves for its possible function in coat color and immunity (Anderson *et al.* 2009; Coulson *et al.* 2011), and we show that it is also found in Atlantic and West Forest populations from Denali National Park. Significant GO categories of "defense response" in both `Bayenv` and `SweeD` supported a role of immune response in these wolf ecotypes as well (Figure 2-3, Figure 2-S10).

MHC Class II genes encode cell surface immune receptors that respond to bacterial antigens in the extracellular environment, and variation within these genes is thought to improve the defense response to pathogens (reviewed in Bernatchez & Landry 2003). Temperature-related variation in immunity genes has been observed in salmon, where clinal variation reflects changing vector prevalence in streams (Dionne *et al.* 2007), and heterozygote advantage has been documented in direct response to zoonotics (Osborne *et al.* 2015) and associated with pathogen resistance (Bernatchez & Landry 2003). Likewise, we observed a correlation between heterozygosity for SNPs within MHC *DLA-DQA* and temperature variables (Figure 2-5), which may reflect selection for increased immunity in response to higher or lower levels of pathogen prevalence. In a previous study of MHC haplotype diversity in North American gray wolves (Kennedy *et al.* 2007), wolves of the boreal forest had the highest haplotype diversity at *DLA-DRB1*, *DLA-DQA1*, and *DLA-DBQ1*, and the authors hypothesized that this pattern may be due to habitat-based isolation or post-glacial recolonization history. Our results suggest that this pattern may also be due to temperature-related pathogen prevalence as we find that the frequency of the derived allele has increased from 0.25 in High Arctic wolves to 0.55 in Boreal Forest wolves, who also have the highest mean temperature of warmest month, relative to other populations (Figure 2-5A). We observe similar patterns of correlation with temperature for the *CBD103*-linked intron variant (Figure 2-4), with derived allele frequency increasing from 0.17 in High Arctic wolves to 0.57 in Boreal Forest wolves. Selective sweep regions containing immunity genes have also been identified in diverse species such as humans (Fagny *et al.* 2014), cattle (Qanbari *et al.* 2014), bank voles (White *et al.* 2013), and dogs (Akey *et al.* 2010).

Selection on vision, hearing, and olfaction genes

We identified multiple candidate genes and GO categories related to vision, hearing, and olfaction in wolves. Many of the genes in which we identified putatively selected missense mutations have been implicated in human vision and hearing disorders (i.e. *PCDH15* and *USH2A* in Usher syndrome) or have been well studied in multiple organisms (olfactory receptors genes). Based on habitat-related variation in light and vegetation, and given that wolves are visual hunters with pronounced olfactory sensitivity, it is not surprising that divergent selection for vision and hearing differences among ecotypes has occurred. Damaging mutations (SIFT score <0.05) within *PCDH15* and *USH2A* were outliers for multiple environmental variables, and *PCDH15* was within a selective sweep region for Boreal Forest and Arctic wolves (Supplemental Figures 2-5 & 2-6). For *PCDH15*, we observed an increase in derived allele frequency from near absence in Atlantic Forest wolves (0.06) to fixation (1) in High Arctic wolves, the latter of which experience the sharpest seasonal variation in light conditions. Multiple sensory-related GO categories were also enriched in `Bayenv` (Figure 2-3). Similarly, we predicted that differential ability to detect odorant molecules might be advantageous as a result of differing hunting conditions or intraspecific recognition factors across environments. We found multiple, damaging mutations within olfactory receptor genes, with allele frequency differences as much as 0.56 between Atlantic Forest and High Arctic wolves (Figure 2-4). Together these data imply local adaptation at the molecular level in different wolf ecotypes mediated by environmental factors.

The existing literature on disorders caused by mutations within *PCDH15* and *USH2A* is substantial (Dreyer *et al.* 2000; Alagramam *et al.* 2001; Le Guédard *et al.* 2007; Williams 2008; Yan & Liu 2010). In humans, non-synonymous mutations in *USH2A* are implicated in non-serious forms of deafness and ocultaneous albinism (Dreyer *et al.* 2000), while for *PCDH15*, large-scale genomic aberrations are more likely to cause similar disease symptoms of Usher Syndrome (Le Guédard *et al.* 2007). We found more damaging missense mutations in *USH2A* than in *PCDH15*, including a 3bp in-frame deletion occurring in the functional LE domain in the former that was characterized as putatively damaging. Given that our study design did not include characterizing large-scale indels, it is possible that we have not identified the actual location of selection within *PCDH15*, but rather have identified functional variants linked to deletions. Aside from the obvious functional and medical implications in humans, *PCDH15* has been identified as a candidate gene for selection related to echolocation in mammals (Parker *et al.* 2013), and as a gene within selective sweep regions in East Asian humans (Williamson *et al.* 2007; Grossman *et al.* 2010).

Olfactory receptor (OR) genes aid in sensing and distinguishing odorants in the environment and conspecifics from each other (reviewed in Ache & Young 2005) and are the most abundant gene class in canines with ~1100 genes known, or about 5% of the gene repertoire (Quignon *et al.* 2005). Because of their functional importance, OR genes have been implicated in selection in multiple organisms, including primates (Gilad *et al.* 2003), canids (Chen *et al.* 2012b), and cattle (Qanbari *et al.* 2014). In naturally occurring populations of *Drosophila*, OR genes show clinal variation and signals of selection (Reinhardt *et al.* 2014). In our previous selection scan (Schweizer *et al.* submitted), we detected significant outliers in `Bayenv` that tagged OR genes. None of those genes had functional variants once resequenced here, which suggests that regulatory variants upstream drove the signals on the SNP array. However, in this study we also found strong clines in a different set of OR genes. Given that each OR gene detects a distinct odorant, and specific variants within OR genes have been demonstrated to affect odor perception (Keller *et al.* 2007; Keller & Vosshall 2008), our results suggest that different OR genes may be selected in wolf ecotypes in response to varying habitats.

Metabolism

We found striking examples of selection on metabolic genes in wolf ecotypes. Extreme environmental differences between the most distinct ecotypes (British Columbia, Arctic, High Arctic) and associated diet differences are hypothesized to select for genetic variants influencing lipid levels and insulin regulation for cold tolerance and varying levels of dietary fat. For *LIPG*, the gene encoding endothelial lipase, we found a missense variant that significantly correlated with both mean diurnal temperature range and precipitation seasonality, with the derived allele frequency rising from 0 in British Columbia wolves to 0.83 in High Arctic wolves (Figure 2-4). This variant changes the amino acid from hydrophobic to polar within the functional PLAT domain (Figure 2-6; Razzaghi *et al.* 2013). Likewise, in *APOB*, we found three significant missense mutations within the highly functional 26th exon that may affect the formation of triglyceride-rich VLDL particles. Interestingly, we also found that Arctic and High Arctic ecotypes had a large proportion of their GO-related categories (i.e. GO, KEGG, Reactome, human phenotype, micro RNAs) represented by microRNA categories. MicroRNAs (miRNAs) are short segments of RNA that are involved in posttranscriptional regulation in many organisms and are increasingly implicated in adipocyte differentiation in humans and mice, and in response to environmental stress (Griffiths-Jones 2004; Zaragosi *et al*. 2011; Hilton *et al*. 2012; Wu *et al*. 2013; Lyons *et al*. 2013; Storey 2015).

To our knowledge *APOB* and *LIPG* have not previously been identified as selection candidates in wolves, other than in our initial SNP array-based selection scan (Schweizer *et al.*, submitted). Even so, both genes have been implicated in multiple diseases affecting humans and under selection in other organisms. For instance, in a genome-wide selection scan of polar bears and brown bears, *APOB* was one of the most statistically significant candidate genes, and contained mutations that may be functionally important for the high lipid diet of polar bears (Liu *et al.* 2014). Trout subjected to different fat content diets show differing expression levels of *LIPG* (Kolditz *et al.* 2008), and in humans, mutations in LIPG cause elevated HDL cholesterol (Edmondson *et al.* 2009; Razzaghi *et al.* 2013). *LIPG* and *APOB* are critical in the metabolism of HDL and LDL lipids, respectively, and are necessary for normal maintenance of lipid levels in the blood.

Pigmentation variation

We anticipated finding variants of genes involved in pigmentation pathways that may correspond to coat color variation in wolves, and that function in camouflage (Jolicoeur 1959) or have secondary effects on immunity and fitness (e.g. Anderson *et al.* 2009; Coulson *et al.* 2011). We did identify a selective sweep region within Boreal Forest and High Arctic wolves that included *CBD103*, and an intron variant within CBD103 in absolute linkage to the deletion haplotype that significantly varied with the maximum temperature of warmest month and land cover type. The frequency of the linked variant increases with increasing maximum temperature (Figure 2-4), and also with documented coat color frequencies (Gipson *et al.* 2002; Musiani *et al.* 2007; Anderson *et al.* 2009). Considering the observed clinal variation in wolf coat color, it is intriguing that we found a missense mutation within TYR that is a significant outlier for percentage tree cover in Bayenv and located in a sweep region within Boreal forest and Arctic wolves. The same mutation was not significant in a genotype-phenotype association for white coat color and was not found exclusively in white individuals (Figure 2-S17), suggesting it is one of several loci influencing color variation (Barsh 1996; Hoekstra 2006; Sturm & Duffy 2012).

Lack of strong evidence for morphological genes

Our evidence for selection on morphological variation was not as decisive as for other traits. We found this surprising given that size differences in wolves can facilitate more effective pursuit and capture of prey (MacNulty *et al.* 2009; Slater *et al.* 2009). We initially expected, given results from the SNP array-based genome scan (Schweizer *et al*. submitted), that we would find functional variants within genes implicated in skull morphology. However, we found no specific morphologically associated gene supported in the resequencing analysis. Conceivably, other genes or *cis*/*trans* factors affecting gene regulation may influence morphological variation in wolves and were not captured on the array. For example, we identified in our SNP array-based genome scan several genes within the BMP and WNT developmental pathways, but did not sequence them with our capture array due to design and space limitations~~.~~

Utility of study design

Overall, depending on the specific test, up to 47% of the candidate genes identified with the SNP array genome scan (Schweizer *et al*. submitted) were confirmed by resequencing as outliers with mutations that could affect function. Given that many of these regions were identified as outliers in the previous scan, their overlap with resequencing hits here is not necessarily an independent confirmation (Thornton & Jensen 2007). Nonetheless, resequencing reduced ascertainment bias in the genotype data since all variants, common and rare, were identified, thus enabling a higher resolution examination of diversity. Wolves and dogs share ~99% of polymorphisms; however, SNPs on the genotyping array were chosen to be common in a panel of dogs, which may provide a biased view of genetic differentiation in wolves. Our confirmation by resequencing suggests that the SNPs on the canine array are useful for identifying outlier regions despite ascertainment bias. We were able to computationally predict the effect of mutations through the use of SIFT and Miyata scores, and through detailed literature searches to identify functional domains within protein structures. Furthermore, the fact that many of our top candidates have been well studied in multiple organisms and identified in selection studies in those organisms supports the use of *a priori* candidates and candidate genes from other studies. In fact, 19 out of 60 *a priori* candidates contained variants that were significant in at least one of our selection methods. To further support our functional hypotheses based on coding or regulatory variation, future studies might utilize classical knock-in or knock-out experiments in mice (Lewandoski 2001), new methods such as CRISPR to target alleles (Cong *et al.* 2013) or proteomic approaches (e.g. Storz *et al.* 2010). Additionally, the use of tissue-specific wolf cell lines we have developed may allow allele specific patterns of gene expression to be assessed on a common genetic background (Johnston *et al*., unpublished data).

The use of extensive non-genic data (our “neutral regions” verified using the dog genome annotation, canFam3.1) as demographic controls offers an empirical approach to potentially reduce the false positive rate and remove potential ascertainment bias inherent to SNP genotyping arrays. Genome scans can suffer from high rates of false positives since multiple evolutionary forces can produce similar genetic patterns of variation (Nielsen *et al.* 2007). Inclusion of a demographic model in the analysis can potentially mitigate this problem, but demographic models have inherent simplifying assumptions that may not be realistic or are too difficult to infer with modeling (Lotterhos & Whitlock 2014). Our results suggest that the modeling approach embedded in `BayeScan` may be too conservative (Foll & Gaggiotti 2008) as only a few outlier regions were resolved and were largely not shared in common with our other two outlier approaches (Figure 2-S14, Figure 2-S15). In species with complex demographic histories, such as wolves, empirically based neutral controls may be preferable over explicit models that make specific demographic assumptions (Nielsen *et al.* 2005; Coop *et al.* 2010).

Finally, several important caveats should be noted about our experimental design. First, wolves have levels of LD allowing a moderately dense SNP array to tag genes within 10kb (the distance at which r^2=0.2 in outbred populations; see Gray *et al*. 2009). Secondly, the dog SNP array was enriched for genic regions, with over 60% of SNPs tagging genes within 10kb (Schweizer *et al*., submitted), which likely increased the efficacy of finding genes under selection, especially in comparison to random sequencing methods such as RAD-seq (Baird *et al.* 2008). The use of an exome or transcriptome capture array (e.g. Bi *et al.* 2012) is an alternative to our approach that would provide complete sequences for potentially all transcribed genes in a single experiment, but few of those genes are likely to be under selection. For example, our SNP genotyping array identified just over 1000 candidate genes from a total of about 12 000 tagged genes. Focusing capture on this reduced subset of genes allowed for higher coverage of each gene (>100x) and efficient use of sequencing resources (as many as 50 individuals per lane). Moreover, our uniquely designed capture array allowed extensive sampling of neutral regions as an empirical demographic control for selection tests, which most likely lowered the false positive rate (Lotterhos & Whitlock 2014). One limitation is the availability of genic SNP arrays for the study species, but technological improvements will likely reduce the cost of construction and application of such genotyping arrays in the near future.

Conservation implications

Our findings highlight local adaptation at the molecular level of wolf ecotypes in North America. Unfortunately, two of the ecotypes showing the greatest number of unique outlier genes are from the Arctic and High Arctic (Figure 2-2). These wolf ecotypes inhabit tundra environments that may disappear by the end of this century (Mech 2004; Gilg *et al.* 2012; Mahlstein & Knutti 2012), and are threatened by human impacts such as hunting (Musiani & Paquet 2004; Bryan *et al.* 2014). Most notably, we detect selection in Arctic and High Arctic wolves on genes influencing vision, immunity, pigmentation and metabolism (Figure 2-4). The high level of adaptive distinction found in these ecotypes might be expected given the extreme environment in which they live, but our molecular results provide a powerful mandate to enhance protection of these populations as they represent the most adaptively distinct North American wolves that we have sampled. The large number of GO-related category types in Arctic wolves demonstrates highly specific adaptations to their environment (Figure 2-2A). The large (>100) number of significantly enriched miRNA categories and the literature implicating miRNAs in adipocyte differentiation and extreme environment adaptation implies that Arctic wolves may have evolved regulatory responses to their environment (Figure 2-2A). Similarly, we find that British Columbia coastal wolves have a unique suite of molecular adaptations that support arguments for adaptive distinction (Muñoz-Fuentes *et al.* 2009). Differing sample sizes are unlikely to drive these patterns, as High Arctic and Arctic represent sample sizes at either extreme, but have similar numbers of genes and GO-related categories. The use of the relative number of genes and the top level GO-related categories under selection could potentially add to metrics for ranking conservation priorities based on the need for the preservation of adaptive diversity (Bonin *et al.* 2007; Gebremedhin *et al.* 2009). Specifically, the number of genes under selection provides a numerical ranking of adaptive diversity in each population akin to species diversity indices, whereas the GO categories represented by these genes are more similar to a higher order taxonomic grouping, such as genus or family. Therefore, those populations having the greatest number of unique genes and GO categories could be argued to deserve the greatest priority for conservation of adaptive diversity. Although GO categories are related and hierarchical, these simple indices are a possible alternative to other schemes for prioritizing the management of adaptive diversity (Fraser & Bernatchez 2001; Funk *et al.* 2012) and represent genome wide measures of adaptive divergence that can readily be incorporated into conservation schemes.

Appendix 2-I: Additional Methods

Sample selection

The quantity of DNA was assessed with the Qubit Fluorometer High Sensitivity Kit, and the quality of DNA was measured with a NanoDrop Spectrophotometer and visualization after electrophoresis on a 1.5% agarose gel. Samples were chosen for shearing if they consisted of at least 600ng-1000ng of dsDNA, and a molecular weight on agarose gel of greater than 1kb.

Sequence alignment and processing

Briefly, demultiplexed fastq reads passing the Illumina filter were trimmed for remaining adapter sequences and a minimum base quality of 20, using *fastq_illumina_filter 0.1* (http://cancan.cshl.edu/labmembers/gordon/fastq_illumina_filter/) and *trim_galore 0.3.1* (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Forward and reverse reads were aligned then mapped to the reference dog genome (CanFam3.1) using *bwa aln* with a seed length of 28 and *bwa sampe* with an insert size of 1000bp between paired ends (Li & Durbin 2010). Duplicates were removed using *samtools rmdup*, then local realignment was performed using *GATK 2.6-4* (DePristo *et al.* 2011). After fixing mate information with *picard tools* (http://picard.sourceforge.net), we generated a set of "known" variant sites for the *GATK Base Quality Score Recalibration (BQSR)* by calling individual genotypes with *samtools mpileup*, then intersecting variants that were observed at least twice on each strand within an individual with those variants that were observed in at least two individuals. The resulting set of variant positions was used with the *–knownSites* flag during *GATK BQSR.* The *GATK Unified Genotyper* algorithm was used to call SNPs over the capture array intervals with a padding of 1000 bp.

SweeD and site frequency spectra

Using newly developed methods for allele frequency estimation for next-generation sequence data (ANGSD; (Korneliussen *et al*. 2014; Nielsen *et al*. 2012)), we generated unfolded site frequency spectra (SFS) for individuals from the six ecotypes sequenced on the capture array in five categories: 1) ~5Mb of neutral regions, 2) all genic regions, 3) 0-fold and 4) 4-fold degenerate non-synonymous sites, and 5) genes overlapping sites with a p-val <0.05 in SweeD.

For the ancestral reference, we used a Kenyan golden jackal sequence generated with the following command options within ANGSD: -doFasta 2 -doCounts 1 -minMapQ 30 -minQ 20 -setMinDepth 6 -setMaxDepth 50 (Koepfli, *et al.*, accepted). We calculated the SFS separately for non-synonymous and 4-fold degenerate sites, since the non-synonymous sites should confirm a signature of purifying selection and the 4-fold degenerate sites should more closely mirror the signal from our neutral regions.

Appendix 2-II: Additional Results

Candidate genes in sweep regions

We identified multiple genes of relevance to wolf ecotype differences, specifically those involved in morphogenesis, lipid metabolism and other metabolic aspects, sensory perception, immunity, and pigmentation. A significant outlier in West Forest wolves (p-value<0.01) was *AMOTL1* (*Angiomotin Like 1*), a gene that functions in vascular development during embryogenesis (Zheng *et al.* 2009). In Boreal Forest wolves, a high-ranking gene was *DAAM2* (*Dishevelled Associated Activator Of Morphogenesis 2*), which is involved in dorsal patterning and spinal cord formation (Lee & Deneen 2012). Within Atlantic Forest wolves, *COL22A1* (*Collagen, Type XXII, Alpha 1*) encodes a collagen gene with expression at muscle ends, and in which knockdown studies in zebrafish induce muscular dystrophy (Charvet *et al.* 2013).

ACMSD (*Aminocarboxymuconate Semialdehyde Decarboxylase*), a gene that also regulates fatty acids in the diet (Egashira *et al.* 2004), was a significant outlier in Atlantic Forest wolves (Supplemental Figure 9). Two genes related to glucose transport and insulin were also outliers in West Forest and Arctic wolves, respectively (Supplemental Figure 4 & 6).

Polymorphisms near *EXOC4* (*Exocyst Complex Component 4*) are significantly associated with fasting glucose level and type II diabetes in humans (Laramie *et al*. 2008), and a polymorphism in *CACNA1E* (*Calcium Channel, Voltage-Dependent, R Type, Alpha 1E Subunit*) has been associated with type II diabetes and reduced insulin secretion (Holmkvist *et al.* 2007).

FOXN1 (*Forkhead box N1*), through mouse knockout studies, has been shown to be involved in normal development of body hair and peripheral T lymphocytes in the blood (Cunliffe *et al.* 2014). Mutations in *FOXN1* cause SCID and the "nude" or congenital alopecia (hair loss) phenotype in humans and mice (Adriani *et al.* 2004). This gene was an outlier in both Arctic and High Arctic wolves (P-value ≤0.05).

We examined the unfolded site frequency spectrum (SFS) from the capture array re-sequencing data. Compared to the SFS for neutral regions, the genic regions, 0-fold degenerate sites, and 4-fold degenerate sites showed an excess of low frequency, derived alleles for West Forest, Boreal Forest, Arctic, and High Arctic wolves (Supplemental Figure 12). This was not the case for British Columbia or Atlantic Forest wolves, where genic regions had slightly lower proportions of derived alleles than neutral regions. This may be due to either the absence of sweeps, low power due to the small sample sizes of these populations, or negative selection. For outlier regions in `SweeD` with a p-value <0.05, the proportion was higher than genic regions for low-frequency, derived alleles and lower than genic regions for high-frequency, derived alleles for West Forest, Arctic, High Arctic, and Atlantic Forest wolves, which is consistent with a signature of selective sweep. As above, the lack of consistent signal across all ecotypes may be due to small sample sizes for some populations.

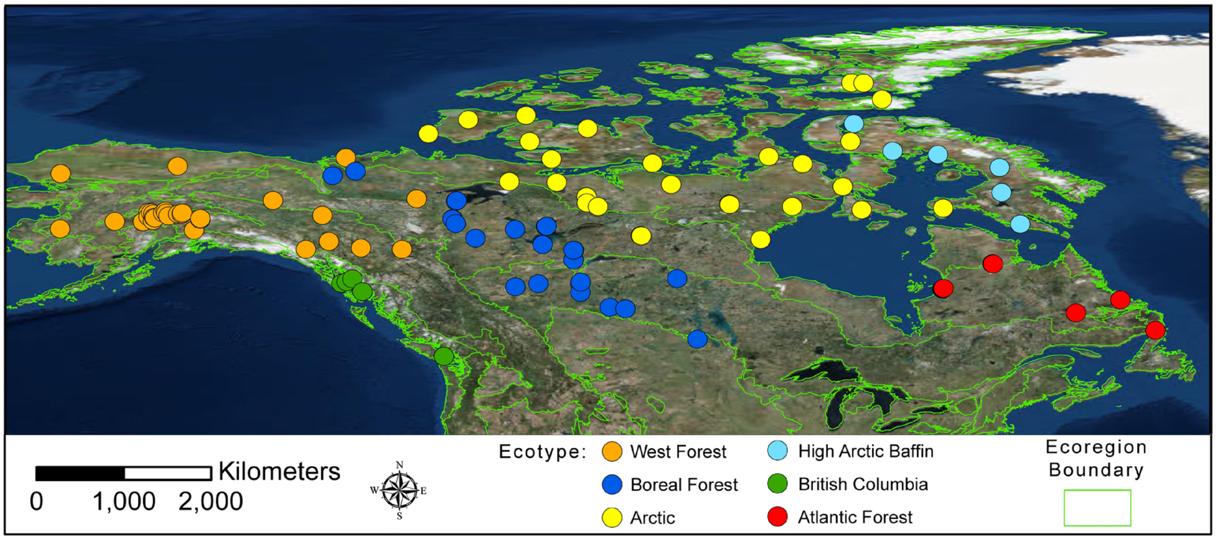


Figure 2-1. Sampling location of 107 wolves superimposed on a satellite image, with colored circles indicating the genetically and environmentally determined ecotypes of West Forest, Boreal Forest, Arctic, High Arctic Baffin, British Columbia, and Atlantic Forest (see legend). Green boundaries show major Environmental Protection Agency Ecoregions (http://www.epa.gov/naaujydh/pages/ecoregions.htm).

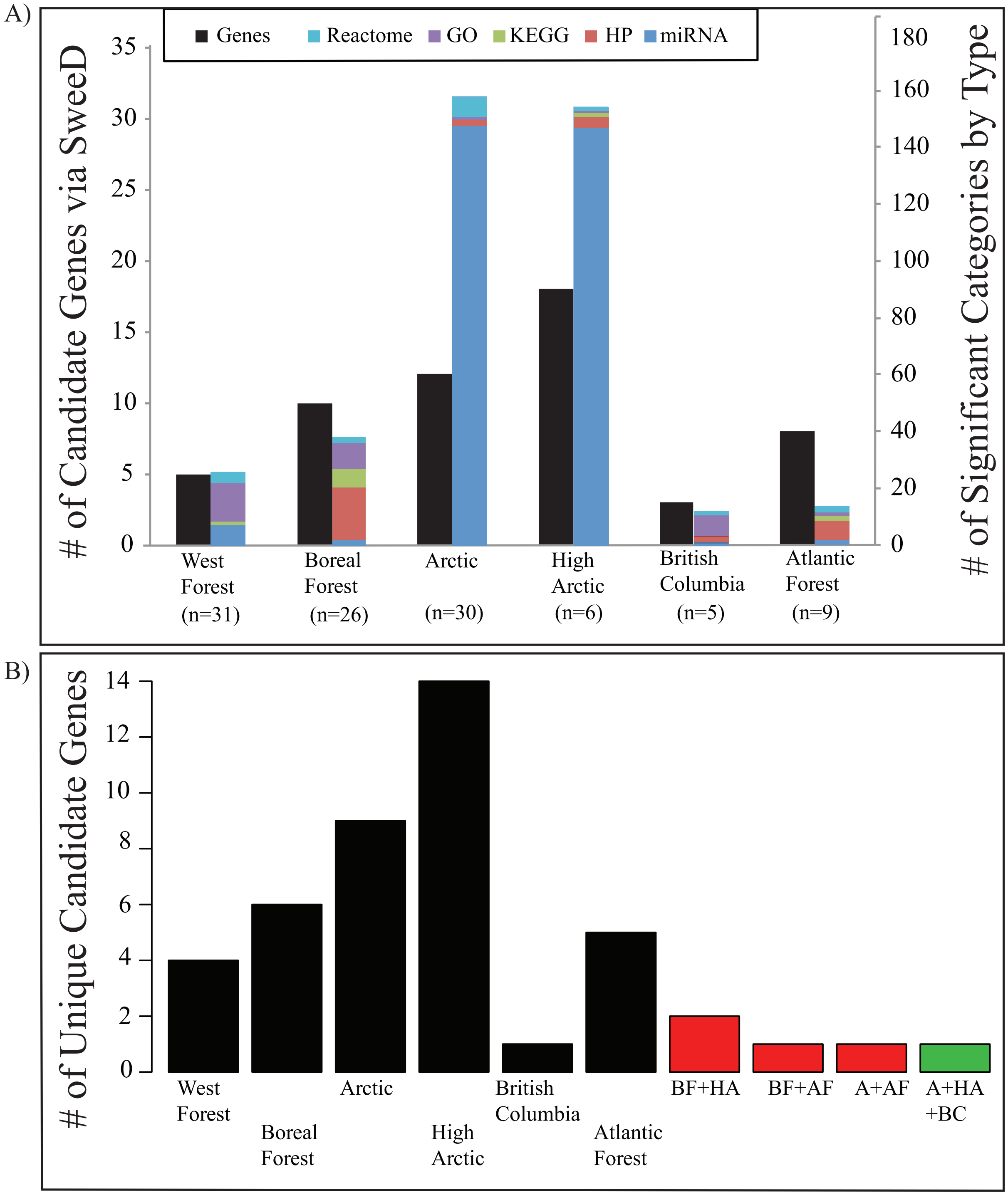


Figure 2-2. Counts of genes and GO-related categories from SweeD. A) The number of candidate genes and GO-related categories within wolf ecotypes (n: sample size), using a significance threshold of p≤0.01. Reactome: Reactome Biological Pathway; KEGG: Kyoto Encyclopedia of Genes and Genomes Pathway; GO: Gene Ontology; HP: Human Phenotype; miRNA: miRBase microRNAs. B) The number of unique candidate genes at the same significance threshold within each ecotype and within more than ecotype. Ecotypes are coded as follows: WF (West Forest), BF (Boreal Forest), A (Arctic), HA (High Arctic), BC (British Columbia), AF (Atlantic Forest).

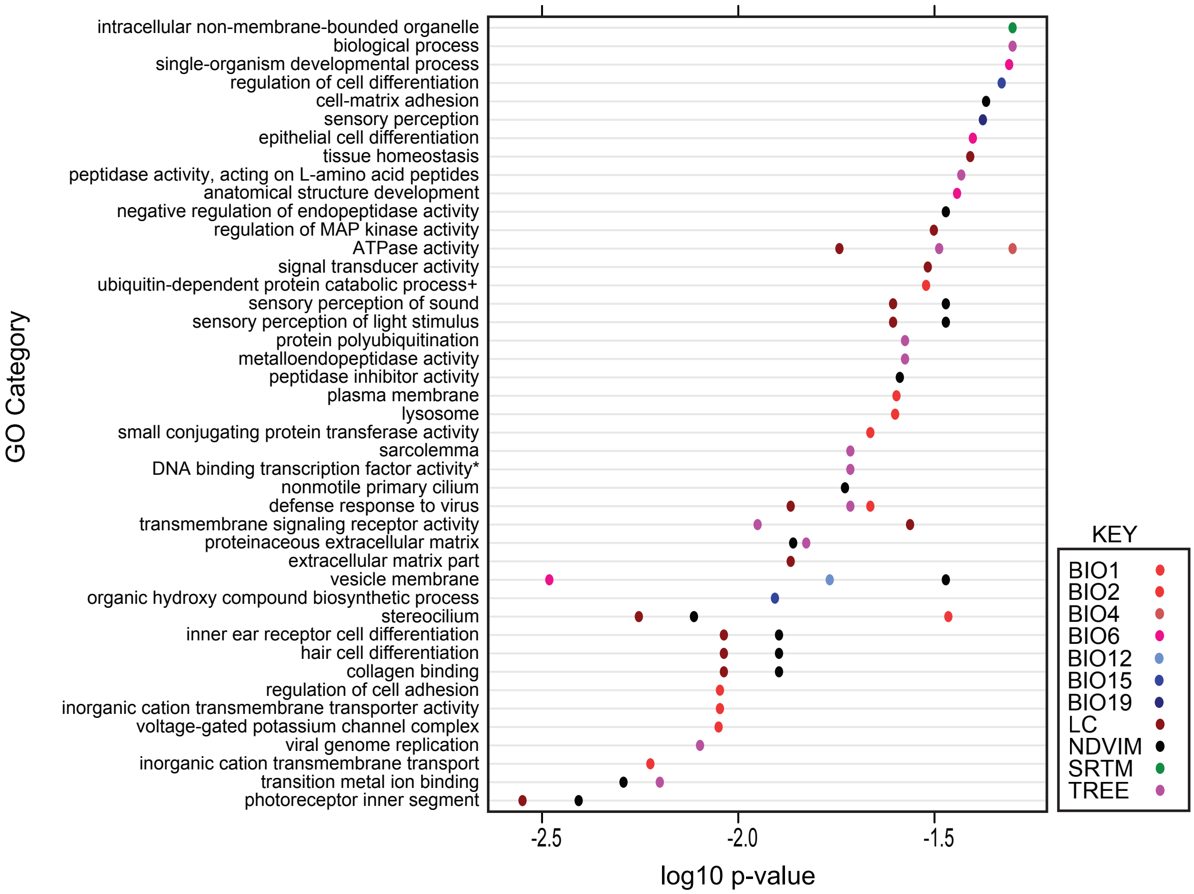


Figure 2-3. Significantly enriched GO categories containing genes with mutations significant in Bayenv (p<0.005). Only categories with a minimum of two genes are shown, with the log10 p-value as calculated by gProfiler and significant after multiple testing. Two categories shortened for space limitation are marked: "positive regulation of sequence-specific DNA binding transcription factor activity" (*) and "protein ubiquitination involved in ubiquitin-dependent protein catabolic process" (+). Environmental variables are related to temperature (red colors; BIO1: annual mean temp., BIO2: mean diurnal temp. range, BIO4: temp. seasonality, BIO5: max. temp. of warmest month, BIO6: min. temp. of coldest month), precipitation (blue colors; BIO12: annual precipitation, BIO15: precipitation seasonality, BIO19: precipitation of coldest quarter), vegetation (green colors; LC: land cover metric, NDVIM: normalized difference vegetation index, TREE: percentage tree cover) and elevation (black; SRTM: shuttle radar topography metric).

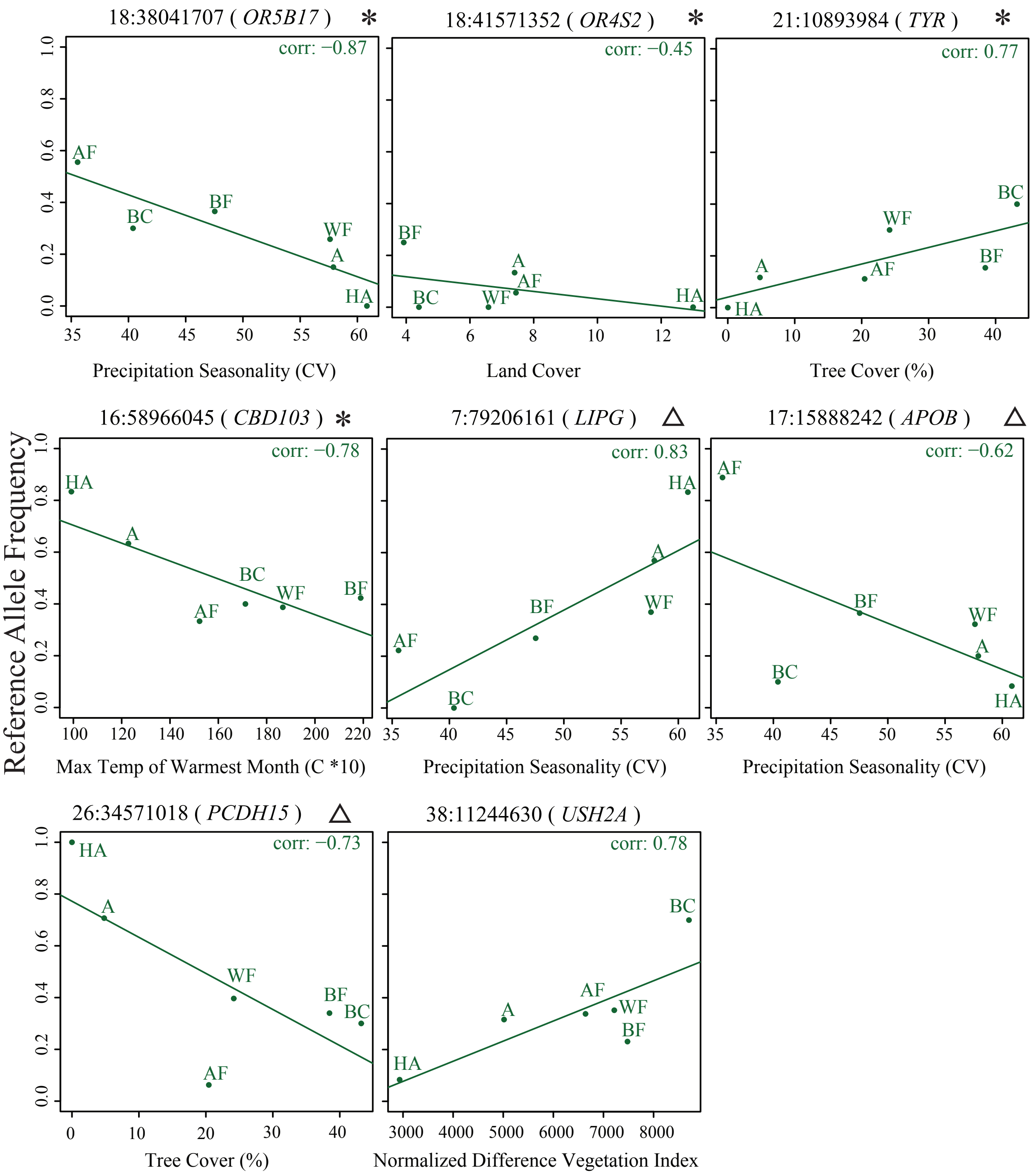


Figure 2-4. Clinal variation in allele frequency for eight significant variants from Bayenv. For each SNP, the reference allele frequency (y-axis) and environmental variable (x-axis) are plotted, with linear best fit lines and Pearson's correlation. The SNP location and gene name are provided, and an indication that the derived allele is reference (Δ) or non-reference (*), if known. Ecotypes are coded as follows: WF (West Forest), BF (Boreal Forest), A (Arctic), HA (High Arctic), BC (British Columbia), AF (Atlantic Forest).

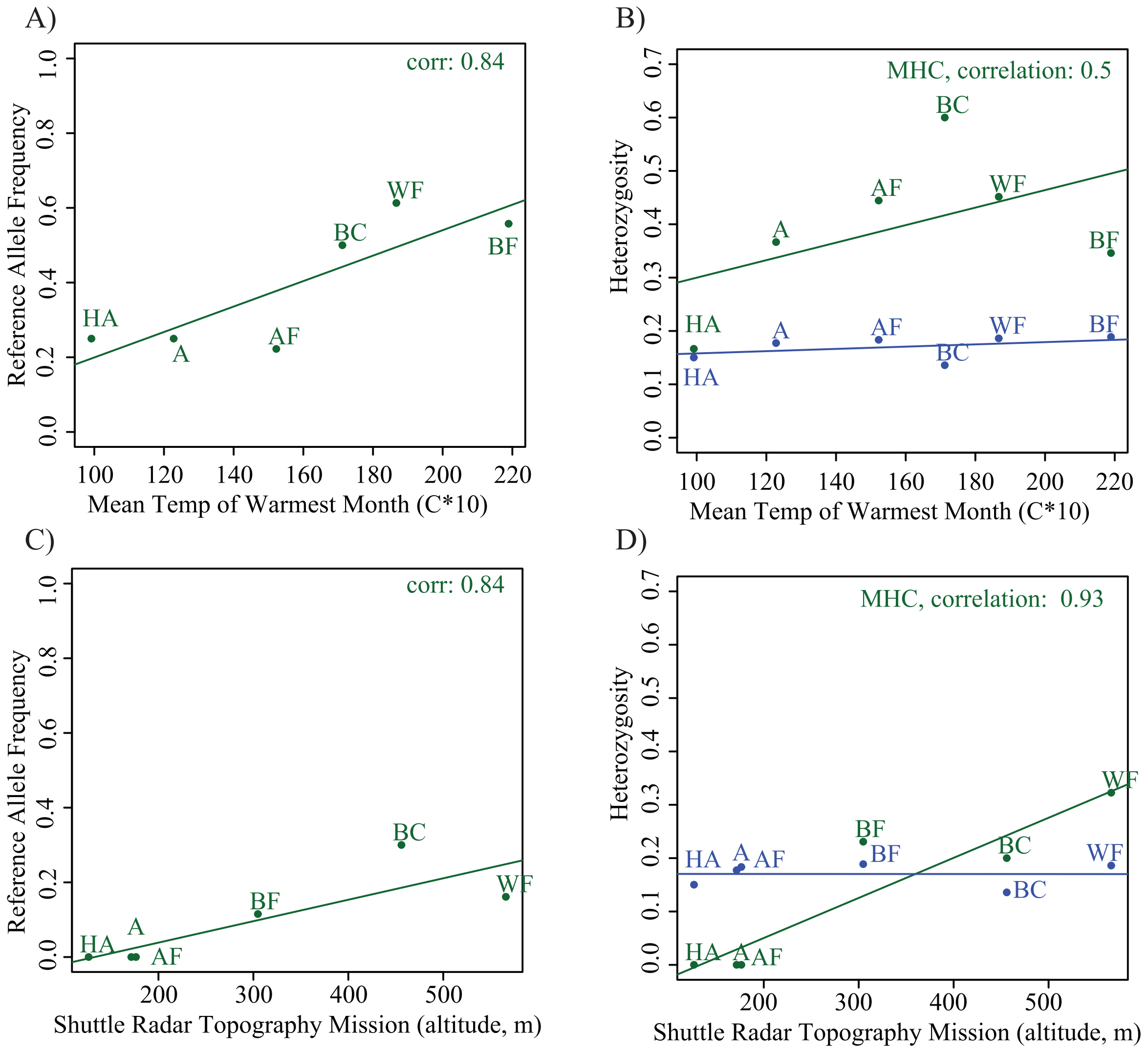


Figure 2-5. Allele frequency and heterozygosity in MHC Class II genes. A and B correspond to *DLA-DQA* missense mutation (chr12:2225338; derived allele is reference), C and D correspond to *DLA-DRB1* missense mutation (chr12:2164457; ancestral state unknown). The SNP location and gene name are provided. Heterozygosity of random neutral SNPs is provided (blue). Ecotypes are coded as follows: WF (West Forest), BF (Boreal Forest), A (Arctic), HA (High Arctic), BC (British Columbia), AF (Atlantic Forest).

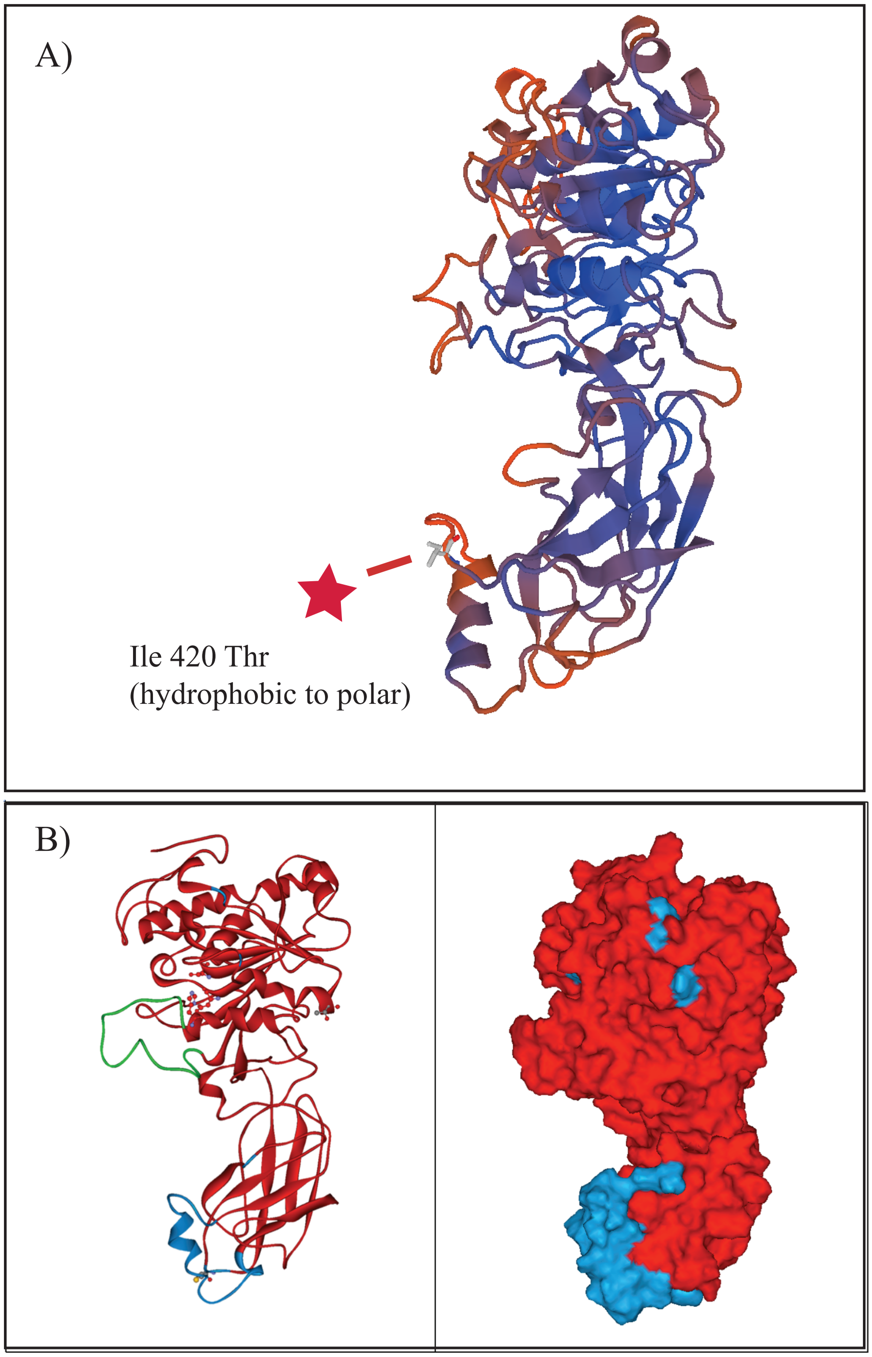


Figure 2-6. Location of missense mutation in *LIPG* within 3D protein structure. A) The mutation, which changes an Isoleucine (hydrophobic) to a threonine (polar) at amino acid (AA) position 420, occurs in a B) 23 AA structural motif (blue helix structure) within the PLAT domain of *LIPG* . The full 3D structure of *LIPG* is provided for reference. Previous work indicates this domain may enhance the lipid binding function of *LIPG* . Part B) is reproduced, with permission, from Razzaghi, et al (2013).

Table 2-1. Summary statistics and number of positions for genic and neutral regions. Categories are ordered from highest impact to lowest impact, as annotated by Ensembl Variant Effect Predictor. Variants were filtered to have a 95% genotyping call rate.

	Genic	**Neutral**
All Sites Pass Filter	2,129,544	4,918,729
Variable Sites Pass Filter	13,092	39,376
Transititions:Transversions	2.672	2.325
Exonic Ti/Tv	4.171	--
Overlapped Genes with Variants	717	--
Overlapped Transcripts with Variants	928	--
Variant Consequences (Most Severe)		
splice donor variant	1 (0.008%)	--
splice acceptor variant	3 (0.023%)	--
stop gained	6 (0.046%)	--
initiator codon variant	5 (0.038%)	--
missense variant	798 (6.095%)	--
deleterious missense	220 (1.680%)	--
splice region variant	158 (1.207%)	--
synonymous variant	1288 (9.838%)	--
5 prime UTR variant	139 (1.062%)	--
TFBS/5 prime UTR variant	27 (0.206%)	--
3 prime UTR variant	123 (0.940%)	--
TFBS/3 prime UTR variant	1 (0.008%)	--
non coding transcript exon variant	17 (0.130%)	--
intron variant	5884 (44.943%)	--
TFBS/intron variant	98 (0.749%)	--
upstream gene variant	321 (2.452%)	--
TFBS/upstream gene variant	302 (2.307%)	--
downstream gene variant	54 (0.412%)	--
TFBS/downstream gene variant	8 (0.061%)	--
intergenic variant	4295 (32.806%)	--
TFBS/intergenic variant	78 (0.596%)	--

Table 2-2. Summary table of significant non-synonymous SNPs identified in Bayenv (BF>1). SNPs with allele frequencies plotted against environmental variables are in **bold**. The genotype and ancestral allele (Freedman et al 2014) is provided when possible (otherwise indicated with '?'). For each SNP, SIFT scores ≤ 0.05 and Miyata scores ≥ 1.85 are in **bold.** See manuscript for details.

Gene	SNP	Environmental Variable		Nucleotide Mutation	Ancestral	Amino Acid Mutation	SIFT Score	Miyata Score	Bayes Factor	P-value
APOB (Lipid Metabolism)	17:15881156	BIO15	Precipitation Seasonality	C/T	?	Arg3812Lys	1	0.4	4.07	0.00293
	17:15881156	TREE	Percentage Tree Cover	C/T	?	Arg3812Lys	1	0.4	2.26	0.00853
	17:15881156	SRTM	Altitude	C/T	?	Arg3812Lys	1	0.4	1.89	0.02400
	17:15882467	SRTM	Altitude	C/T	C	Val3510Ile	1	0.85	1.23	0.04413
	17:15884825	BIO15	Precipitation Seasonality	C/T	C	Val2724Ile	0.89	0.85	2.56	0.00560
	17:15884825	BIO5	Max. Temp. of Warmest Month	C/T	C	Val2724Ile	0.89	0.85	1.57	0.01107
	17:15884825	BIO2	Mean Diurnal Temp. Range	C/T	C	Val2724Ile	0.89	0.85	1.15	0.01573
	17:15886439	SRTM	Altitude	G/T	G	Leu2186Ile	0.06	0.14	13.10	0.00267
	17:15888047	SRTM	Altitude	C/T	C	Gly1650Ser	0.45	0.85	2.27	0.01833
	17:15888047	BIO5	Max. Temp. of Warmest Month	C/T	C	Gly1650Ser	0.45	0.85	1.13	0.01813
	17:15888117	NDVIM	Vegetation Index	C/T	?	Met1626Ile	0.27	0.29	1.96	0.00793
	17:15888242	BIO15	Precipitation Seasonality	G/A	A	Leu1585Phe	0.32	0.63	7.86	0.00160
	17:15888242	BIO1	Annual Mean Temp.	G/A	A	Leu1585Phe	0.32	0.63	2.89	0.00360
DLA-DQA (Immunity)	12:2221262	TREE	Percentage Tree Cover	G/C	G	Glu25Asp	0.8	0.9	1.05	0.02167
	12:2225320	SRTM	Altitude	A/C	C	Met99Leu	0.64	0.41	2.77	0.01467
	12:2225320	BIO5	Max. Temp. of Warmest Month	A/C	C	Met99Leu	0.64	0.41	1.19	0.01660
	12:2225320	TREE	Percentage Tree Cover	A/C	C	Met99Leu	0.64	0.41	1.11	0.02033
	12:2225338	BIO2	Mean Diurnal Temp. Range	A/C	C	Lys105Gln	0.28	1.06	164.00	0.00000
	12:2225338	BIO5	Max. Temp. of Warmest Month	A/C	C	Lys105Gln	0.28	1.06	6.58	0.00213
	12:2225338	SRTM	Altitude	A/C	C	Lys105Gln	0.28	1.06	1.92	0.02307
DLA-DRB1 (Immunity)	**12:2164457**	SRTM	Altitude	G/A	?	Pro36Ser	0.72	0.56	2.31	0.01820
LIPG (Lipid Metabolism)	7:79206161	BIO2	Mean Diurnal Temp. Range	A/G	G	Ile420Thr	0.43	**2.14**	1.58	0.01027
	7:79206161	BIO15	Precipitation Seasonality	A/G	G	Ile420Thr	0.43	**2.14**	1.01	0.01767
OR4S2 (Olfaction)	18:41571000	BIO15	Precipitation Seasonality	T/C	T	Tyr82His	**0**	**2.27**	1.01	0.01767
	18:41571136	BIO4	Temp. Seasonality	G/A	G	Arg127His	1	0.82	1.19	0.01173
	18:41571261	SRTM	Altitude	C/A	C	Leu169Ile	0.13	0.14	1.30	0.04107
	18:41571352	LC	Land Cover Type	G/A	G	Ser199Asn	**0.01**	1.31	7.34	0.00173
	18:41571352	BIO5	Max. Temp. of Warmest Month	G/A	G	Ser199Asn	**0.01**	1.31	1.75	0.00987
	18:41571352	NDVIM	Vegetation Index	G/A	G	Ser199Asn	**0.01**	1.31	1.03	0.01873
OR5B17 (Olfaction)	**18:38041707**	BIO15	Precipitation Seasonality	C/T	C	Ala97Val	0.06	1.85	2.26	0.00593
	18:38041775	BIO5	Max. Temp. of Warmest Month	T/C	?	Cys120Arg	1	**3.06**	1.28	0.01480
OR6B1 (Olfaction)	16:5885672	NDVIM	Vegetation Index	C/G	C	Val48Leu	1	0.91	15.20	0.00073
PCDH15 (Vision and	**26:34571018**	TREE	Percentage Tree Cover	A/G	G	Asn1555Asp	0.08	0.65	1.04	0.02207
	26:34571630	TREE	Percentage Tree Cover	G/A	?	Glu1755Lys	1	1.14	1.07	0.02127
TYR (Pigmentation)	21:10893984	SRTM	Altitude	C/T	?	Val59Ile	0.31	0.85	1.39	0.03707
	21:10893984	TREE	Percentage Tree Cover	C/T	?	Val59Ile	0.31	0.85	1.06	0.02140
TYRP1 (Pigmentation)	11:33329087	BIO6	Min. Temp. of Coldest Month	G/A	G	Arg416Lys	0.39	0.4	2.81	0.00220
	11:33329087	BIO12	Annual Precipitation	G/A	G	Arg416Lys	0.39	0.4	1.18	0.00447
USH2A (Vision and Hearing)	**38:11244630**	NDVIM	Vegetation Index	C/T	?	Ala3218Thr	0.58	0.9	5.00	0.00273
	38:11244630	LC	Land Cover Type	C/T	?	Ala3218Thr	0.58	0.9	1.30	0.01207
	38:11244661	BIO19	Precipitation of Coldest Quarter	T/G	T	Gln3207His	0.33	0.32	3.25	0.00173
	38:11244661	BIO12	Annual Precipitation	T/G	T	Gln3207His	0.33	0.32	2.57	0.00213
	38:11244661	BIO4	Temp. Seasonality	T/G	T	Gln3207His	0.33	0.32	2.47	0.00527
	38:11244661	BIO6	Min. Temp. of Coldest Month	T/G	T	Gln3207His	0.33	0.32	1.55	0.00440
	38:11288551	SRTM	Altitude	C/T	C	Asp2828Asn	1	0.65	21.70	0.00167
	38:11297838	BIO5	Max. Temp. of Warmest Month	G/A	A	Ala2692Val	0.33	**1.85**	1.10	0.01867

Table 2-3. Summary table of significant SNPs from Bayenv that are located in transcription factor binding sites.

Gene	SNP	Environmental Variable		Consequence	Nucleotide Mutation	Bayes Factor	P-value
ATP6V1C2	17:7601977	BIO2	Mean Diurnal Temp. Range	5 prime UTR variant	C/A	15	0.000933
COBLL1	36:9958211	BIO6	Min. Temp. of Coldest Month	intron variant	C/T	6.26	0.001333
	36:9958211	BIO4	Temp. Seasonality	intron variant	C/T	4.56	0.002800
	36:9958211	BIO19	Precipitation of Coldest Quarter	intron variant	C/T	3.83	0.001533
	36:9958211	BIO12	Annual Precipitation	intron variant	C/T	3.28	0.001667
FOXA3	1:109757833	TREE	Percentage Tree Cover	upstream gene variant	T/A	4.29	0.004000
GPR116	12:15071633	BIO5	Max. Temp. of Warmest Month	intron variant	C/T	5.67	0.002400
KLF12	22:27986043	BIO15	Precipitation Seasonality	intron variant	G/A	3.14	0.004000
LEP	14:8121117	TREE	Percentage Tree Cover	intron variant	A/G	10.7	0.001400
	14:8121117	LC	Land Cover Type	intron variant	A/G	5.77	0.002267

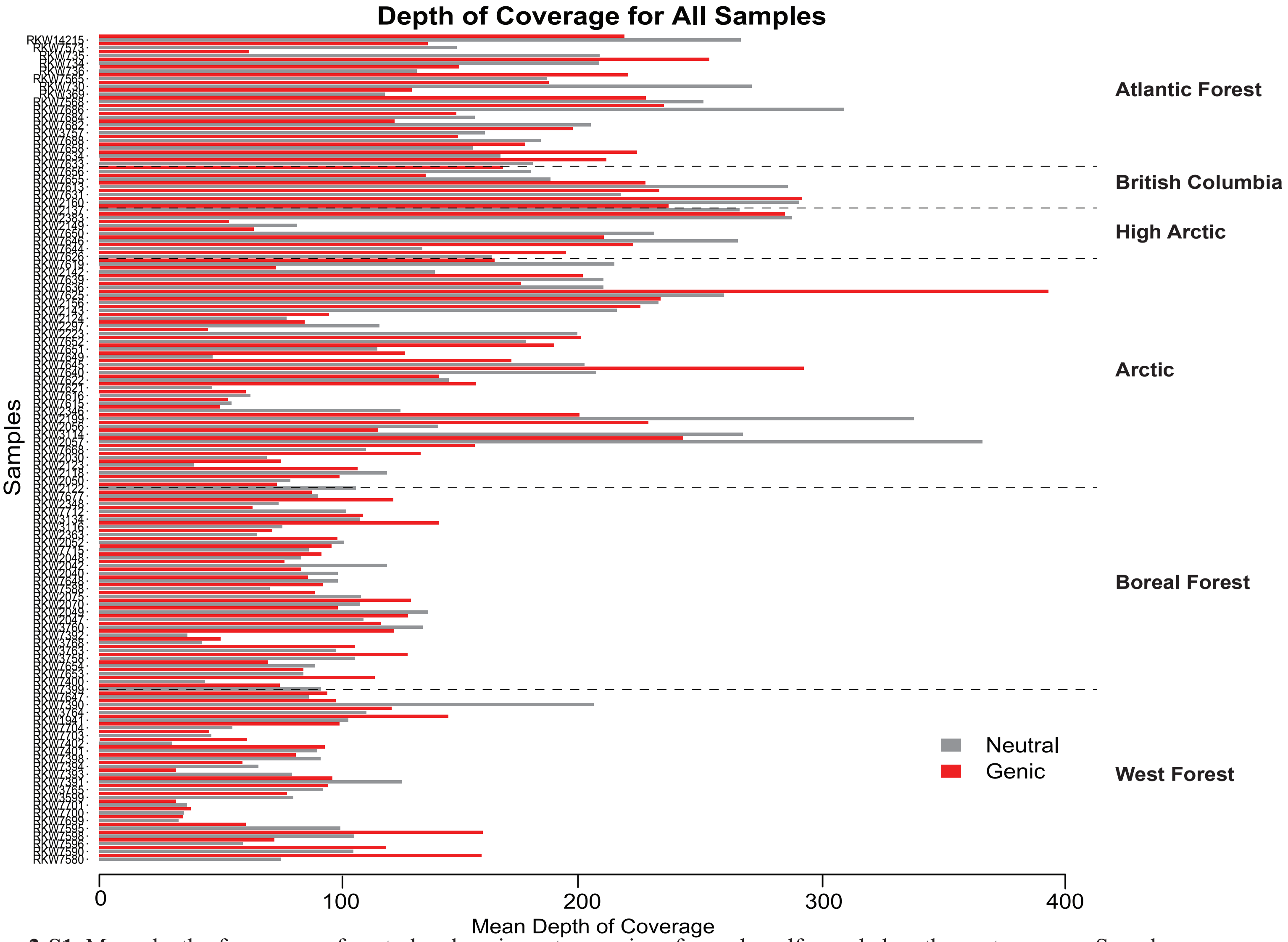


Figure 2-S1. Mean depth of coverage of neutral and genic capture regions for each wolf sampled on the capture array. Samples are grouped by the ecotype, indicated on the right, and each individual has a bar representing the mean depth of coverage over all neutral (gray) or genic (red) positions.

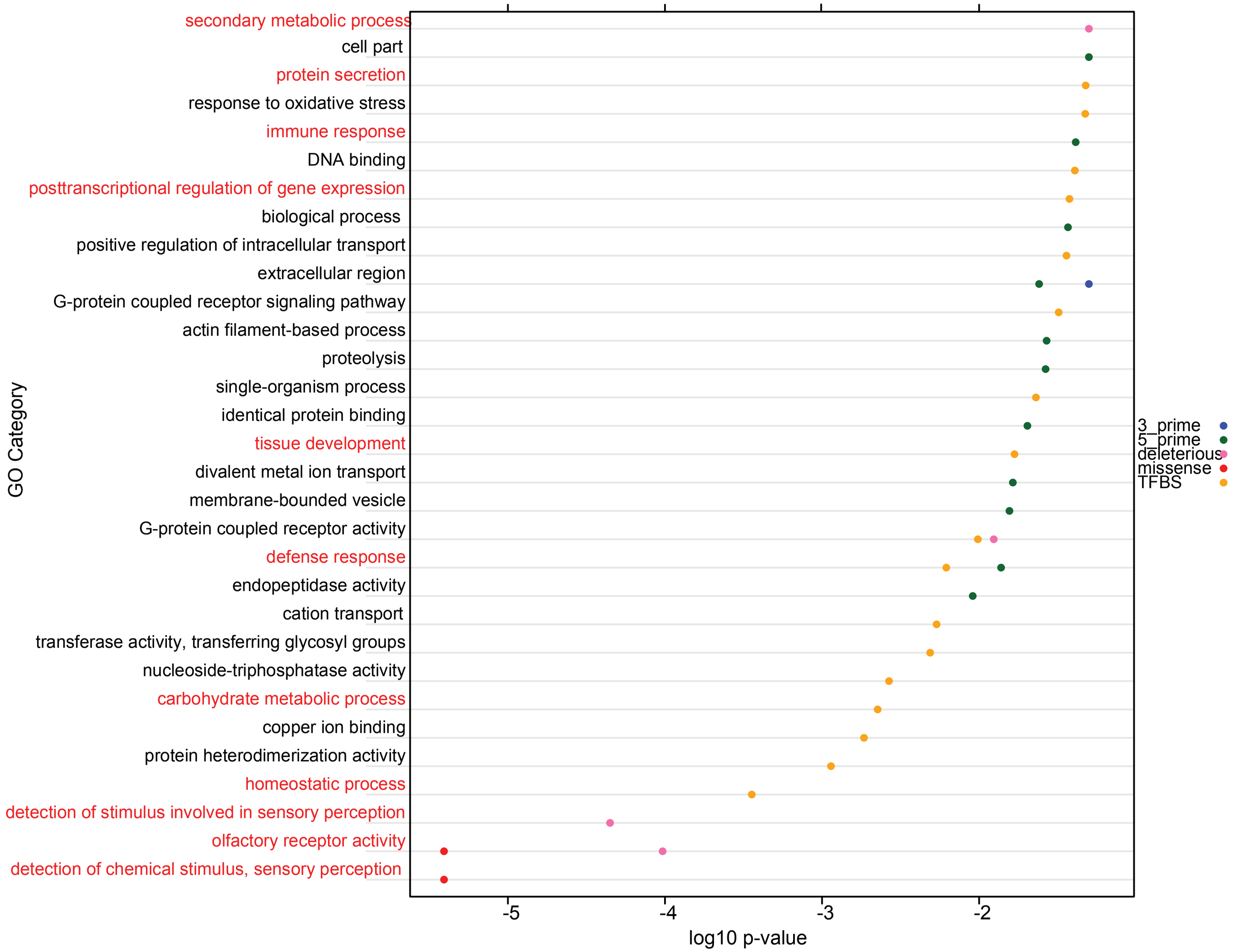


Figure 2-S2. Significantly enriched gene ontology (GO) categories containing genes with functional variants as annotated by VEP, SIFT, and our TFBS databases (see methods for details). Only categories with a minimum of five overlapping genes are shown, with the log10 p-value as calculated by gProfiler and significant after multiple testing.

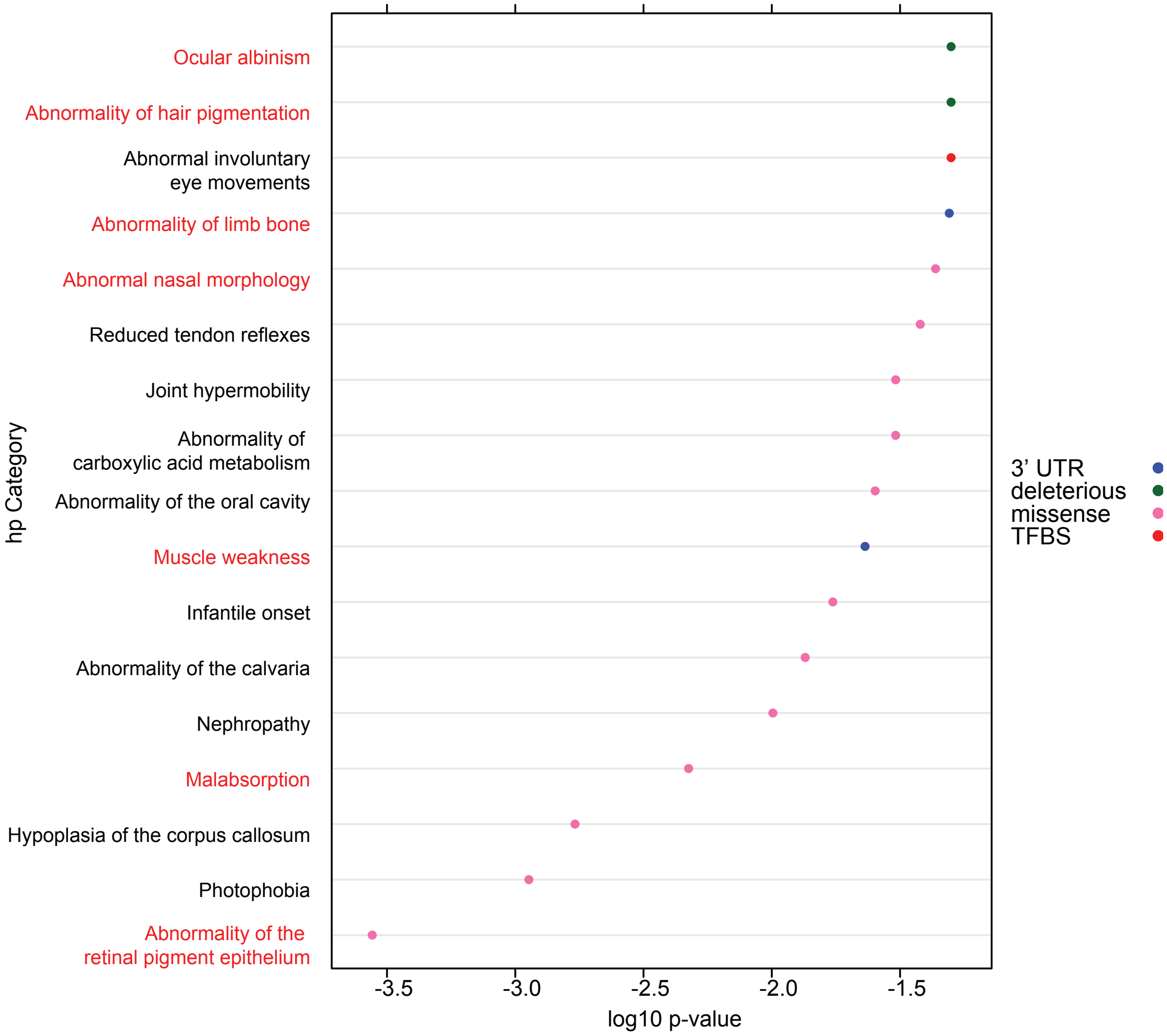


Figure 2-S3. Significantly enriched human phenotype (hp) categories containing genes with functional variants as annotated by Variant Effect Predictor, SIFT, and our TFBS databases (see methods for details). Only categories with a minimum of five overlapping genes are shown, with the log10 p-value as calculated by gProfiler and significant after multiple testing. There were no significant hp categories for 5 prime UTR sites with five or more genes.

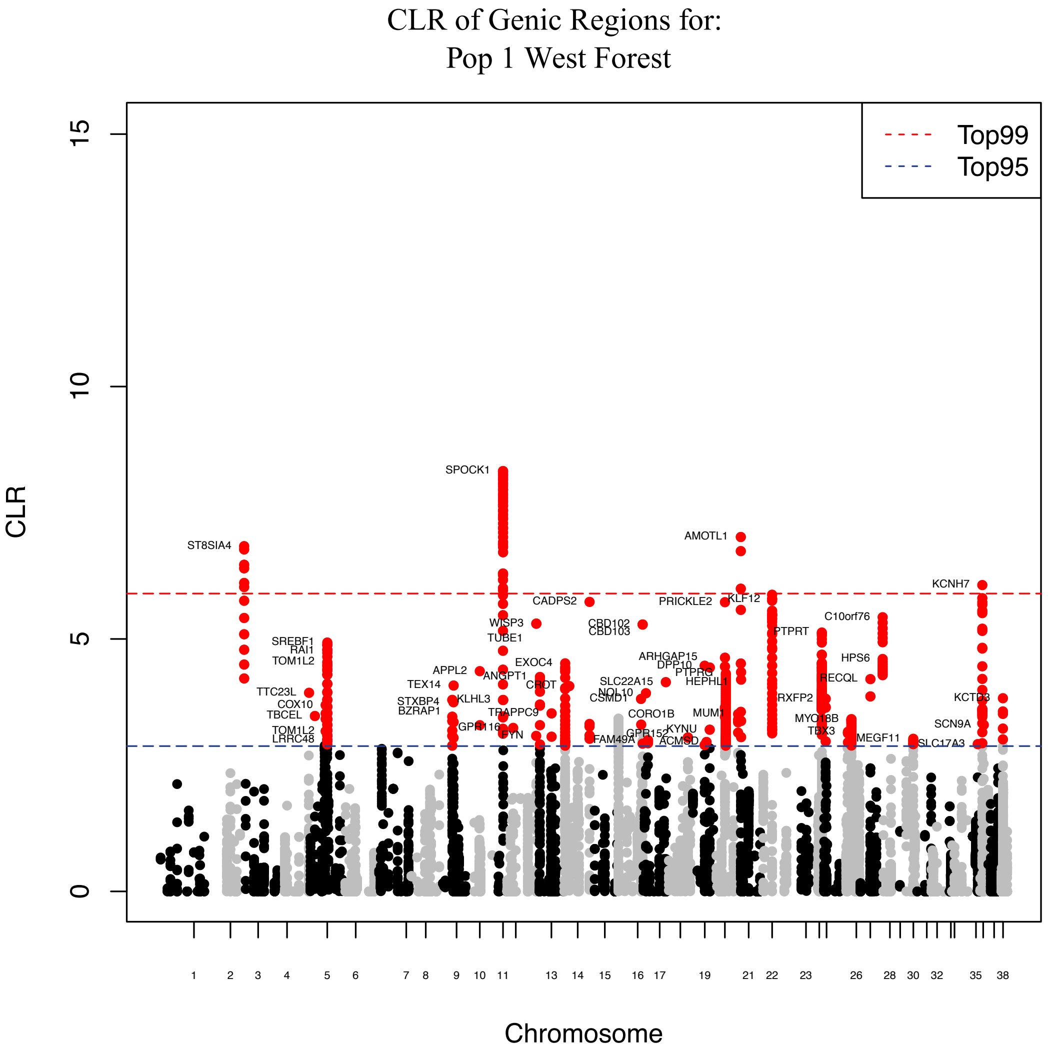


Figure 2-S4. Manhattan plot of the SweeD composite likelihood ratio (CLR) score for each genic position within West Forest wolves. Dashed lines indicate significance thresholds are shown (blue, p≤0.05; red, p≤0.01). Black and gray variants correspond to alternating chromosomes. Variants with a p-value ≤0.05 are shown in red, and variants with a q-value (the FDR equivalent of a p-value)≤0.05 are shown in orange. The variant with the highest CLR within each sweep region is labeled with the closest gene within 6kb.

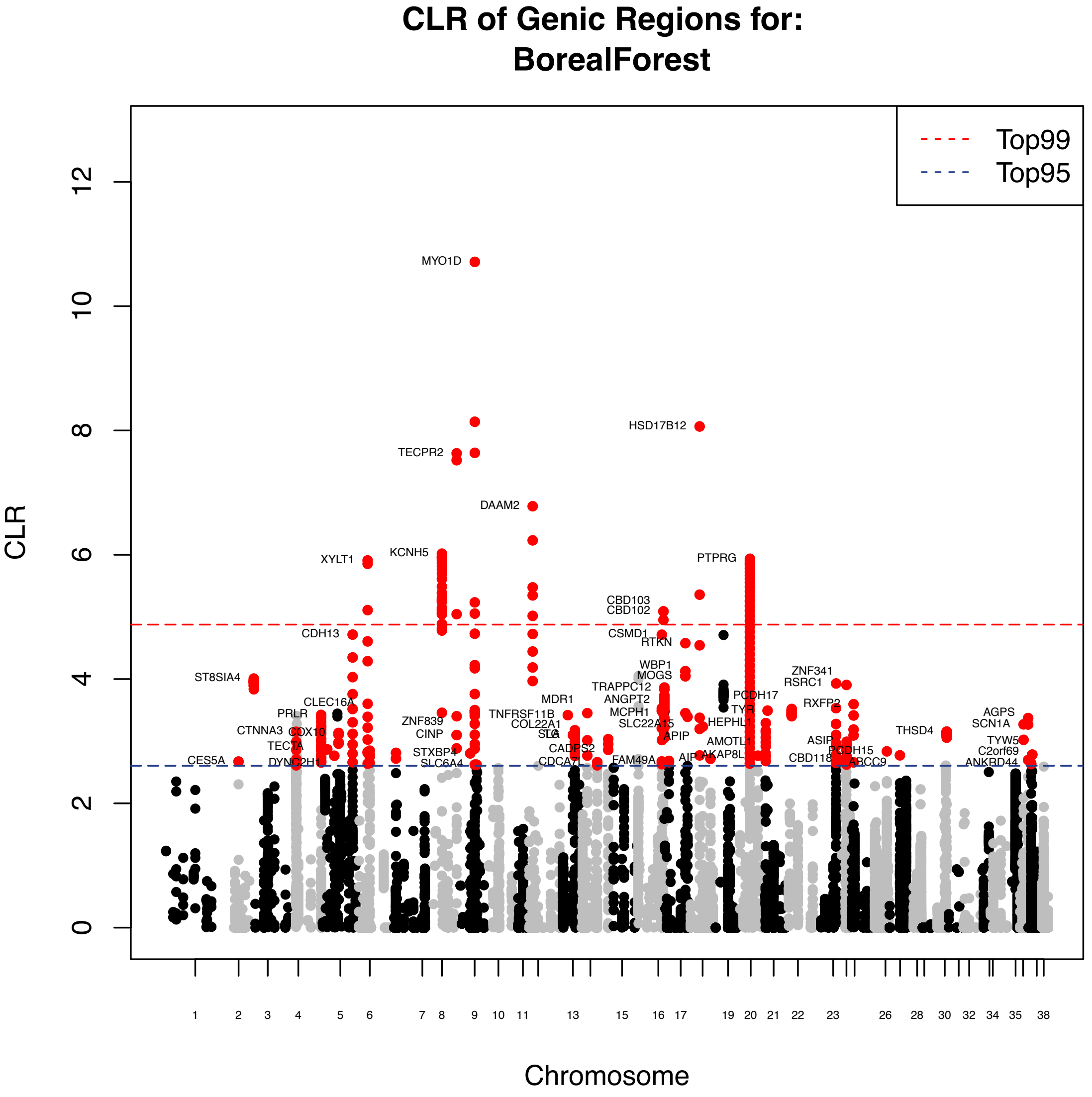


Figure 2-S5. Manhattan plot of the SweeD composite likelihood ratio (CLR) score for each genic position within Boreal Forest wolves. See caption of Figure 2-S4 for details.

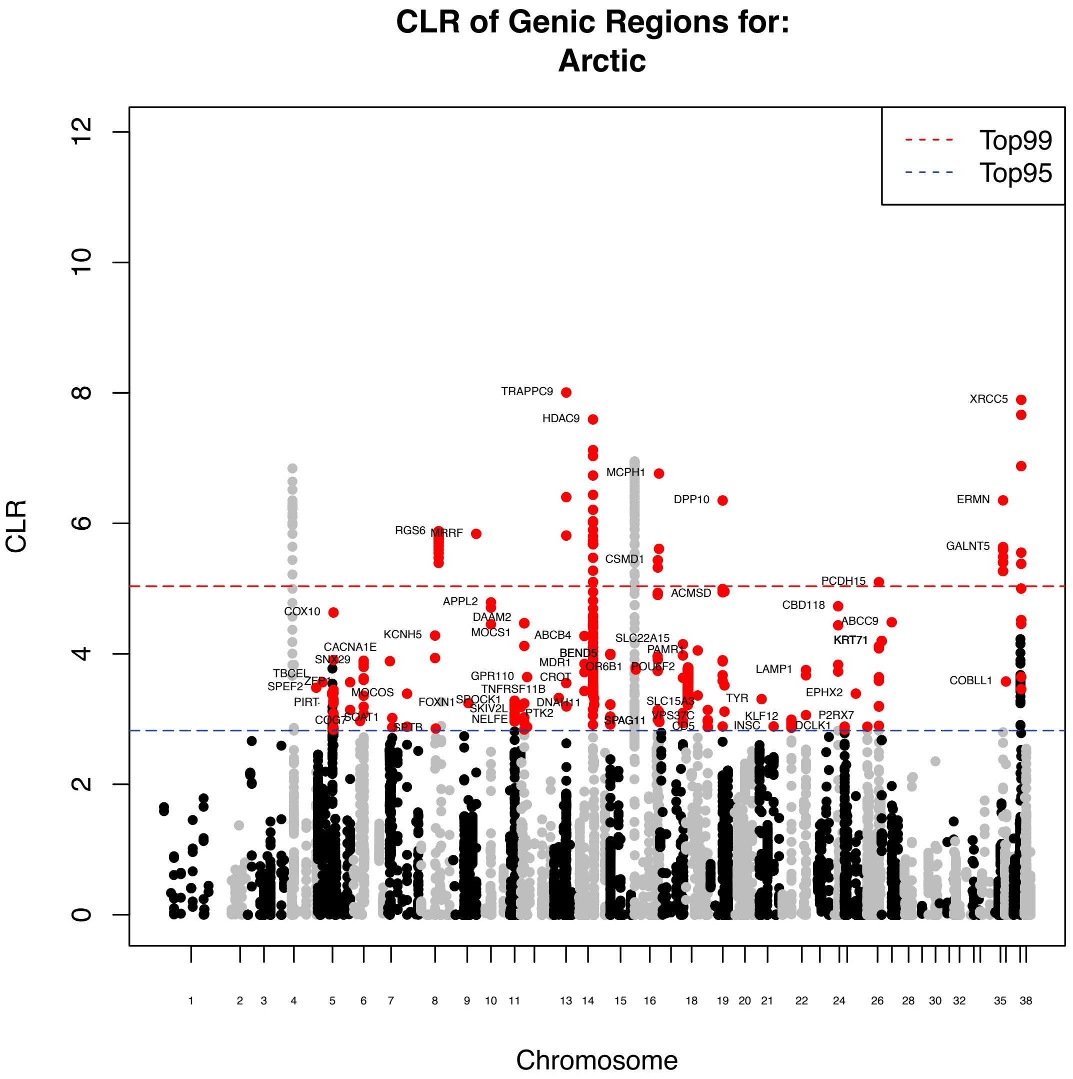


Figure 2-S6. Manhattan plot of the SweeD composite likelihood ratio (CLR) score for each genic position within Arctic wolves. See caption of Figure 2-S4 for details.

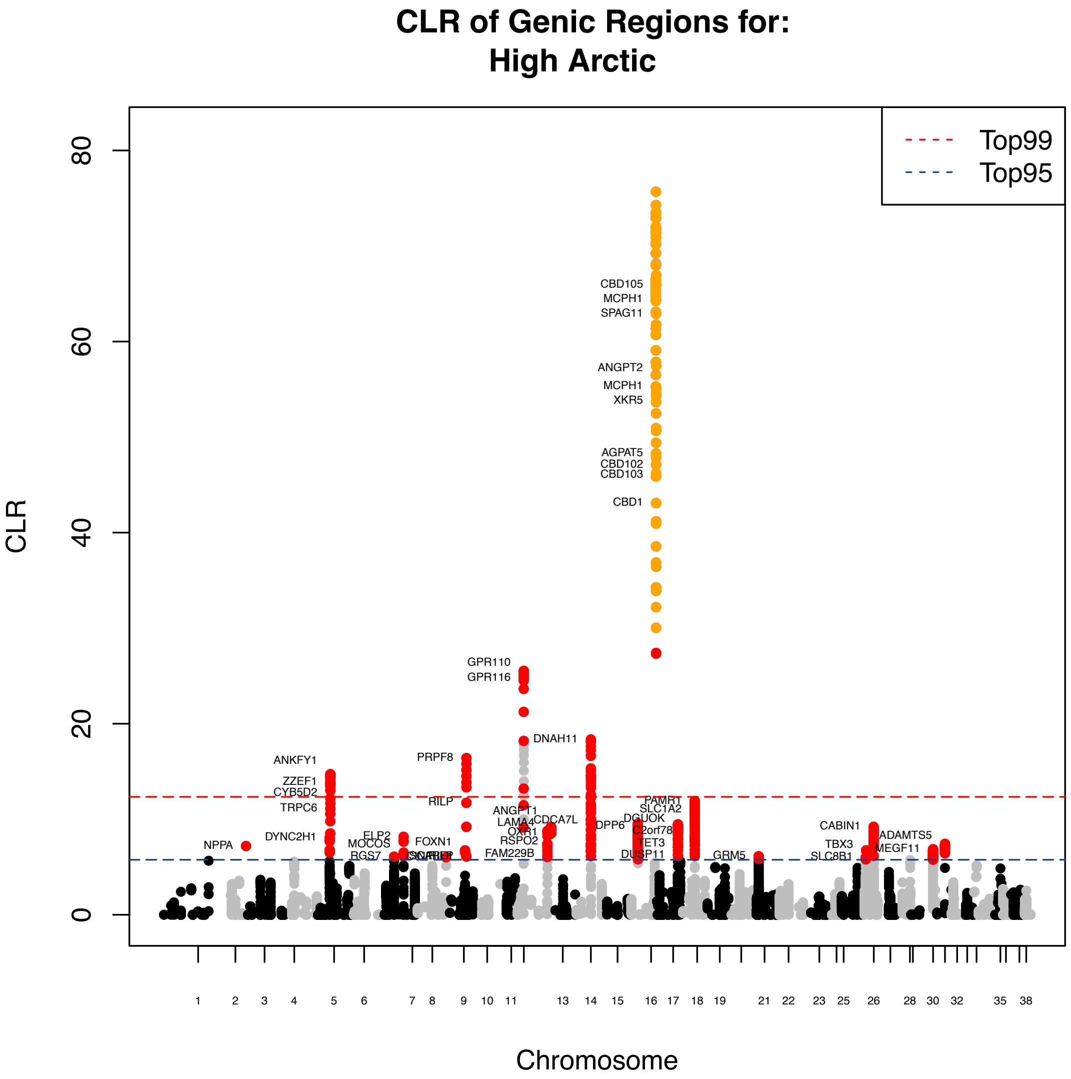


Figure 2-S7. Manhattan plot of the SweeD composite likelihood ratio (CLR) score for each genic position within High Arctic wolves. See caption of Figure 2-S4 for details.

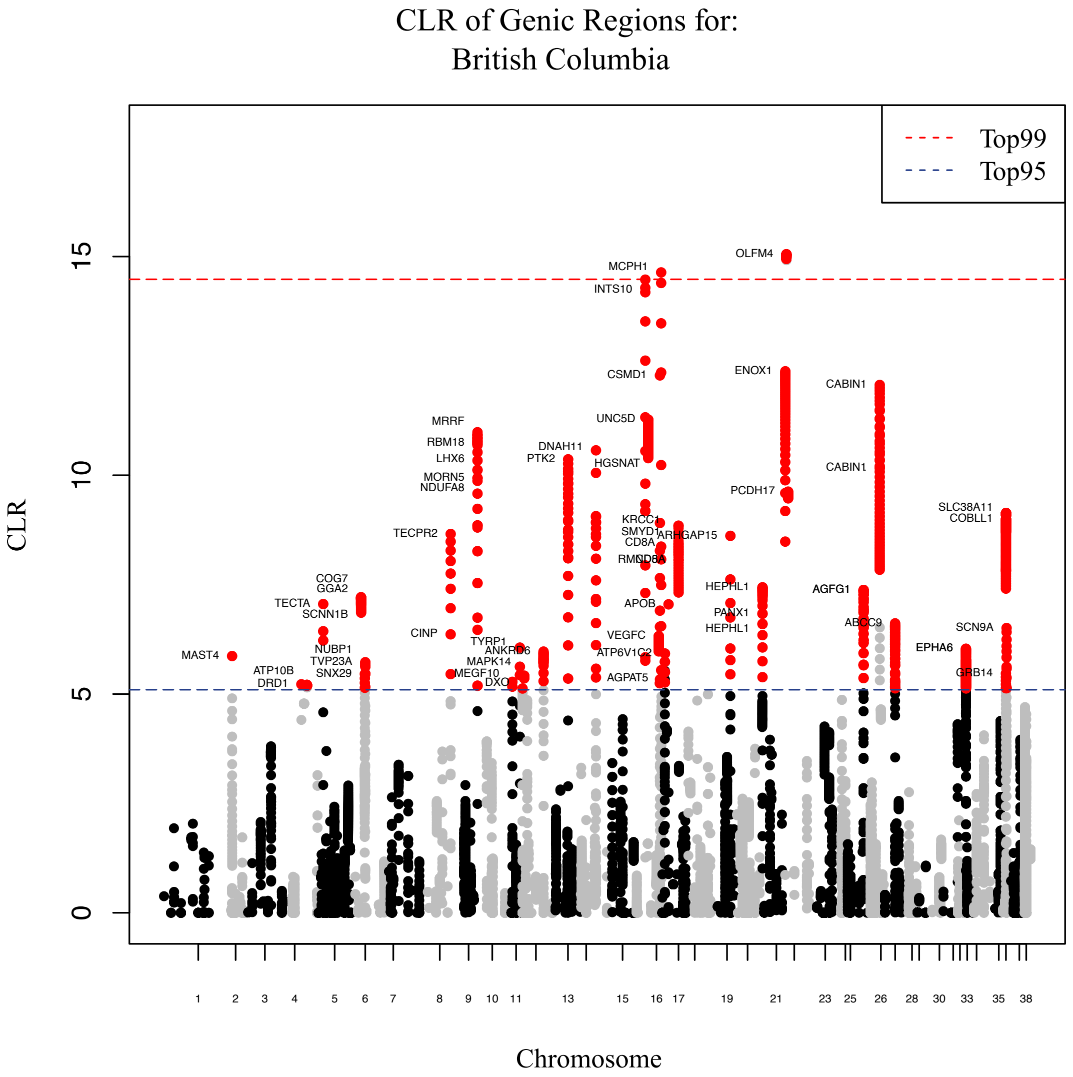


Figure 2-S8. Manhattan plot of the SweeD composite likelihood ratio (CLR) score for each genic position within British Columbia wolves. See caption of Figure 2-S4 for details.

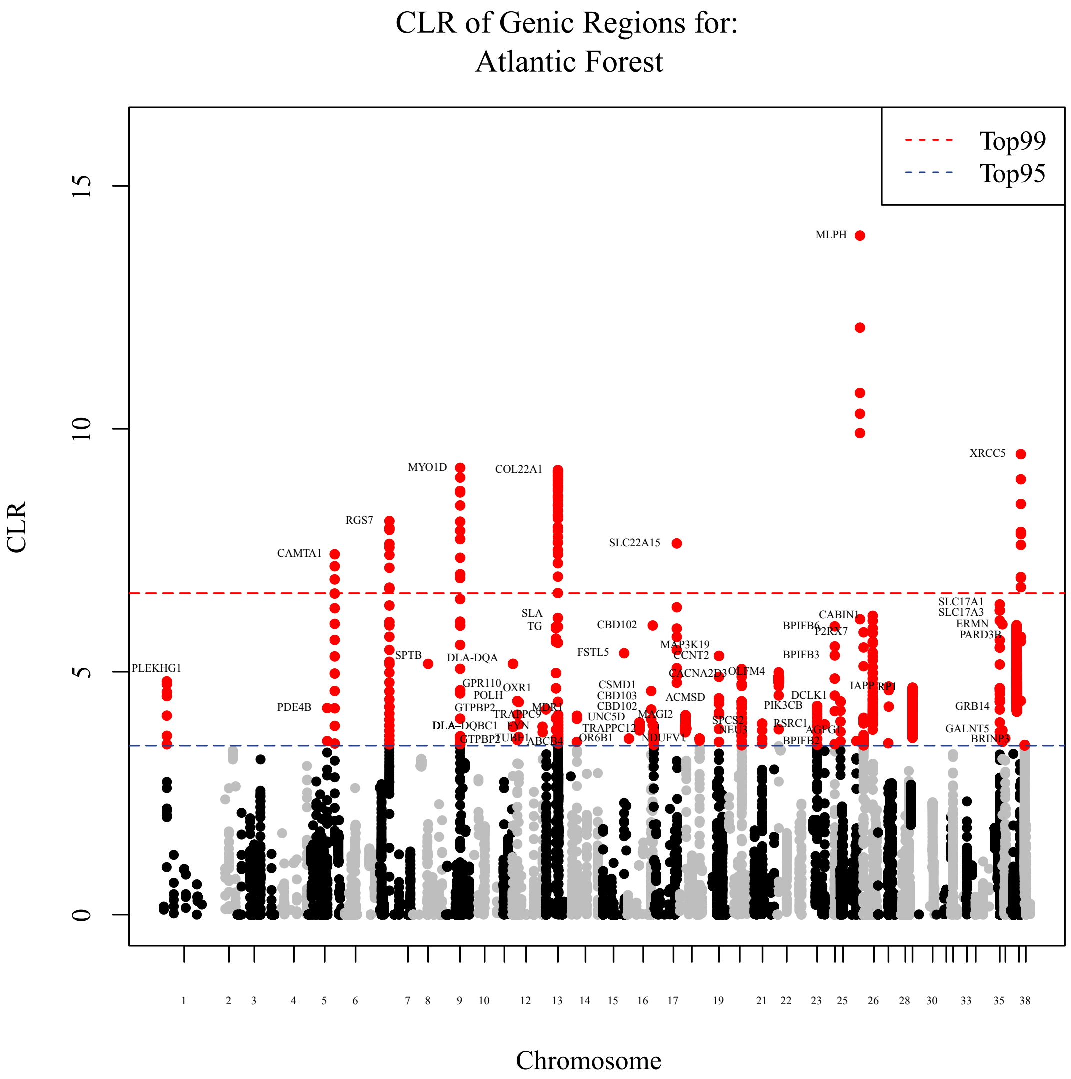


Figure 2-S9. Manhattan plot of the SweeD composite likelihood ratio (CLR) score for each genic position within Atlantic Forest wolves. See caption of Figure 2-S4 for details.

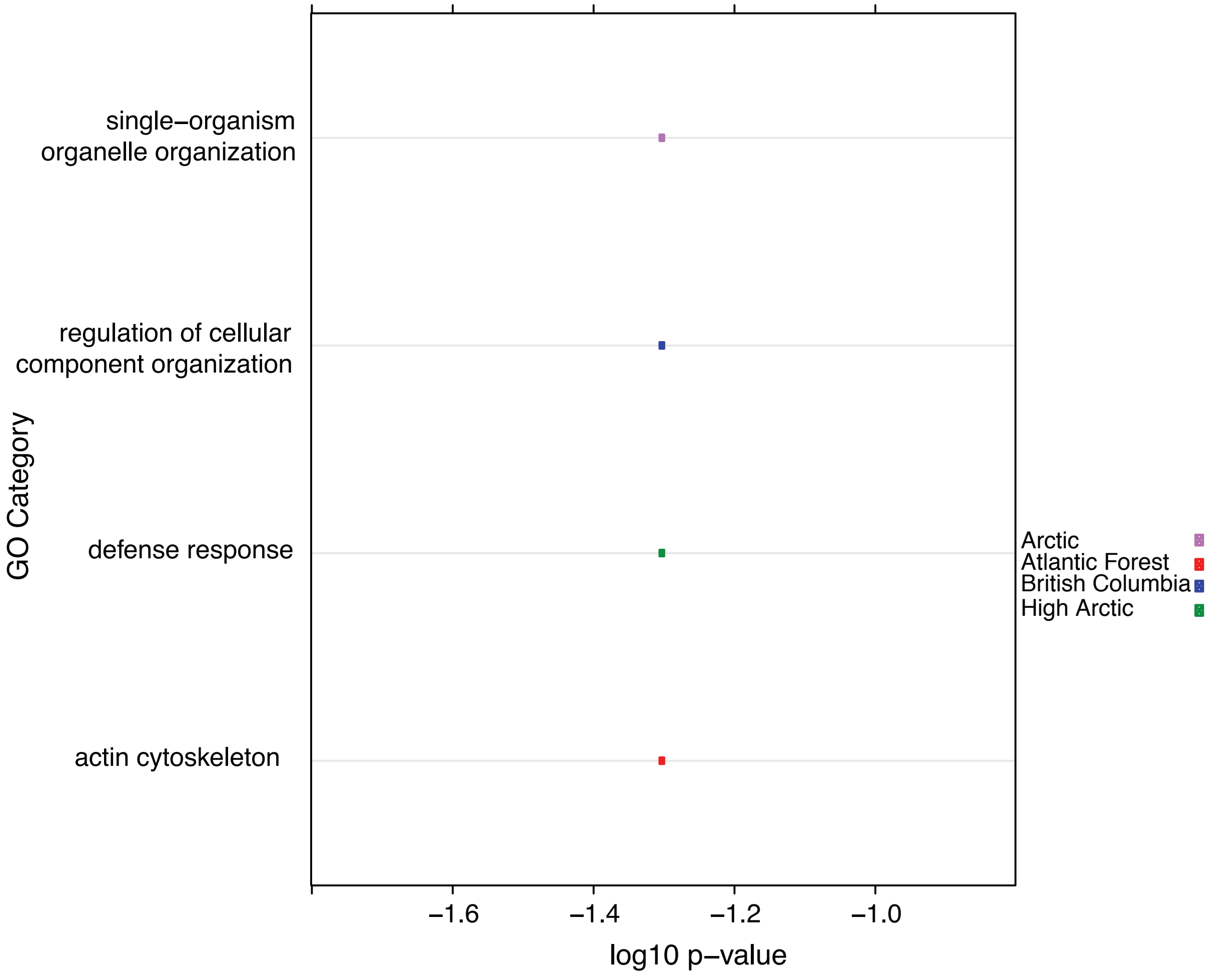


Figure 2-S10. Significantly enriched gene ontology (GO) categories containing genes with significant (p≤0.01) outliers from SweeD. Only categories with a minimum of two overlapping genes are shown, with the log10 p-value as calculated by gProfiler and significant after multiple testing.

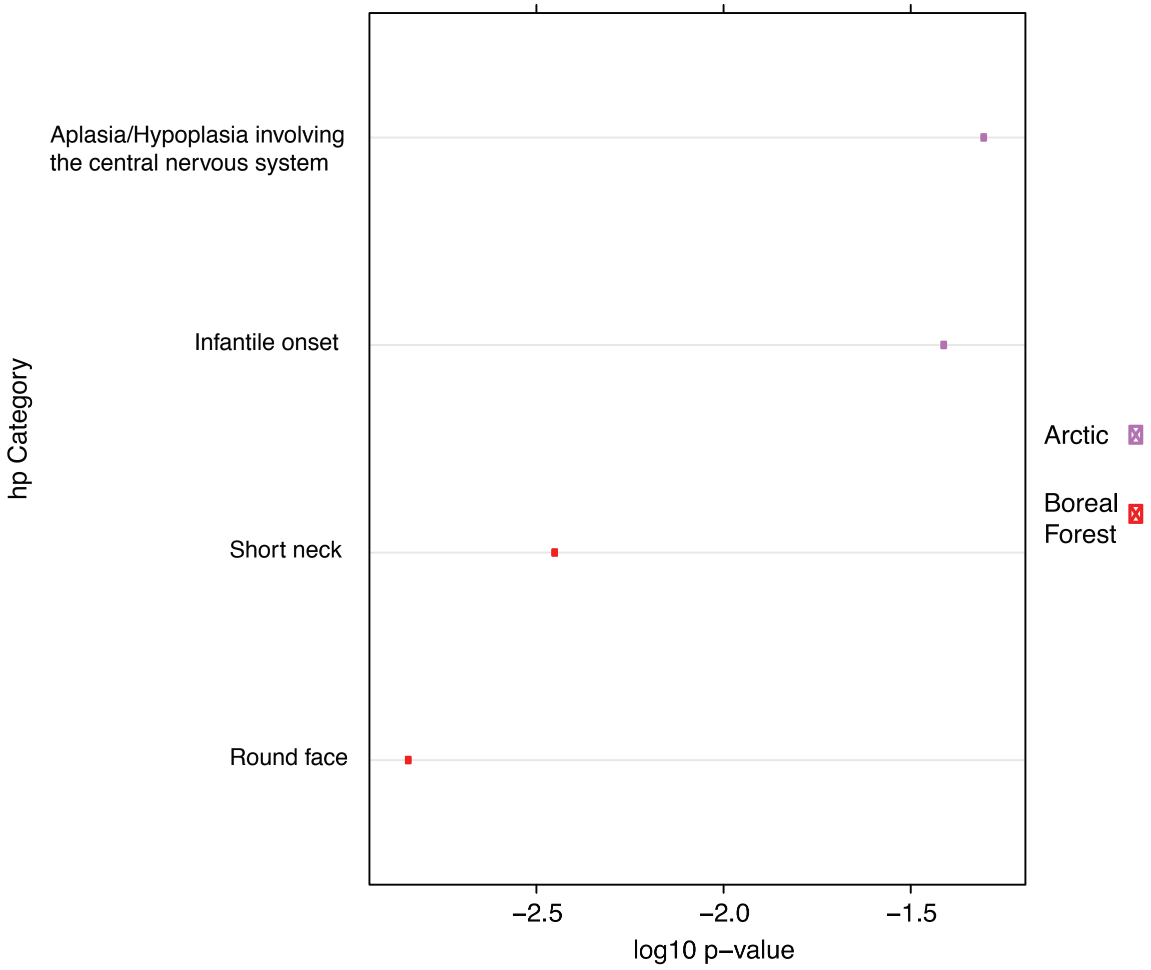


Figure 2-S11. Significantly enriched human phenotype (hp) categories containing genes with significant (p≤0.01) outliers from SweeD. Only categories with a minimum of two overlapping genes are shown, with the log10 p-value as calculated by gProfiler and significant after multiple testing.

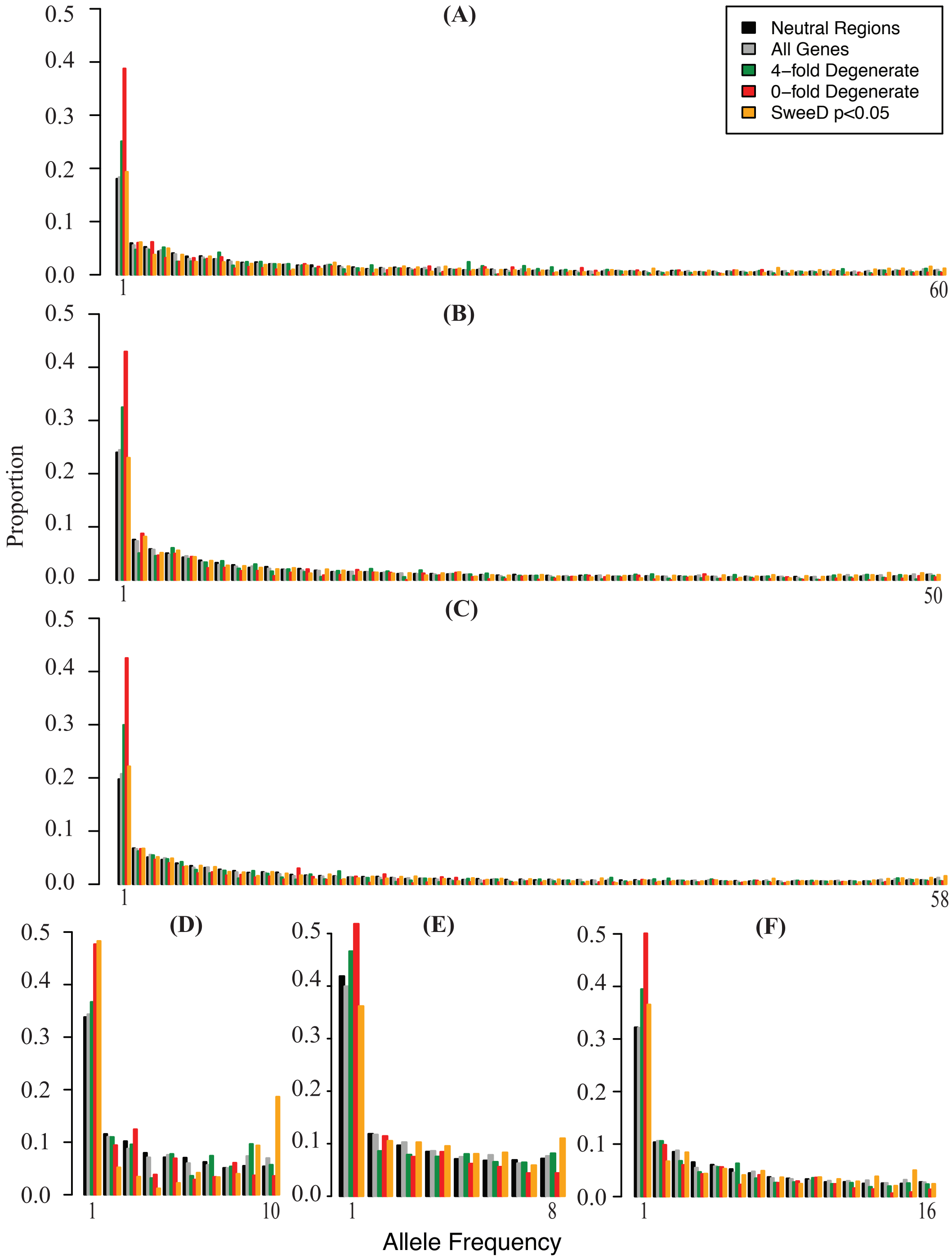


Figure 2-S12. Unfolded site frequency spectra (SFS) for neutral, genic, 0-fold degenerate, 4-fold degenerate, and selective sweep sites in A) West Forest, B) Boreal Forest, C) Arctic, D) High Arctic, E) British Columbia, and F) Atlantic Forest ecotypes. The proportion of sites at different allele frequencies is shown for neutral regions, all genic regions, 4-fold degenerate sites, 0-fold degenerate sites, and selective sweep regions (significant ($p \leq 0.05$) from Swee). The width of each SFS is relative to the total number of chromosomes sampled (i.e. the allele frequency).

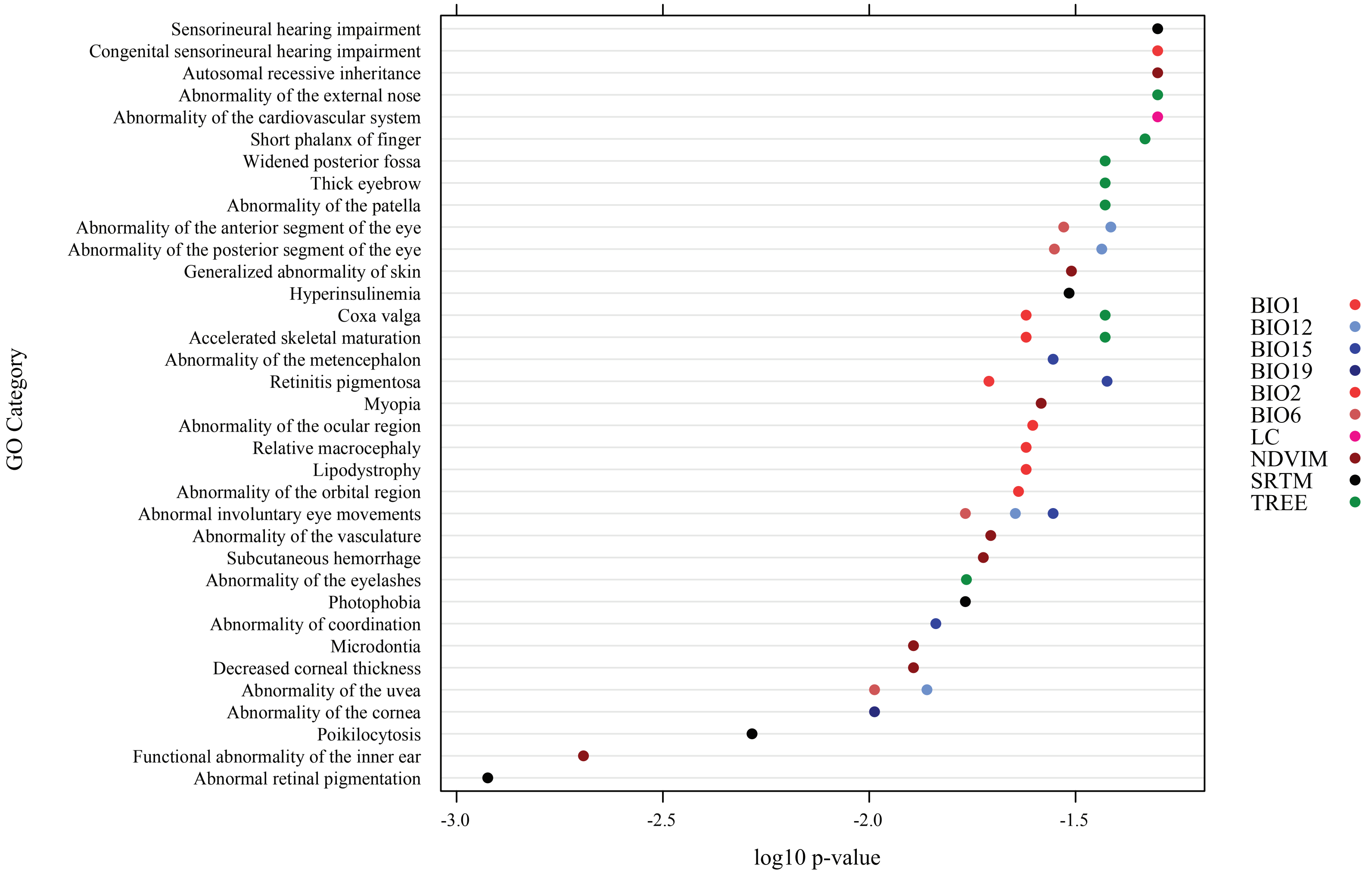


Figure 2-S13. Significantly enriched human phenotype (hp) categories containing genes with significant (p<0.005) outliers from Bayenv. Only categories with a minimum of two overlapping genes are shown, with the log10 p-value as calculated by gProfiler and significant after multiple testing.

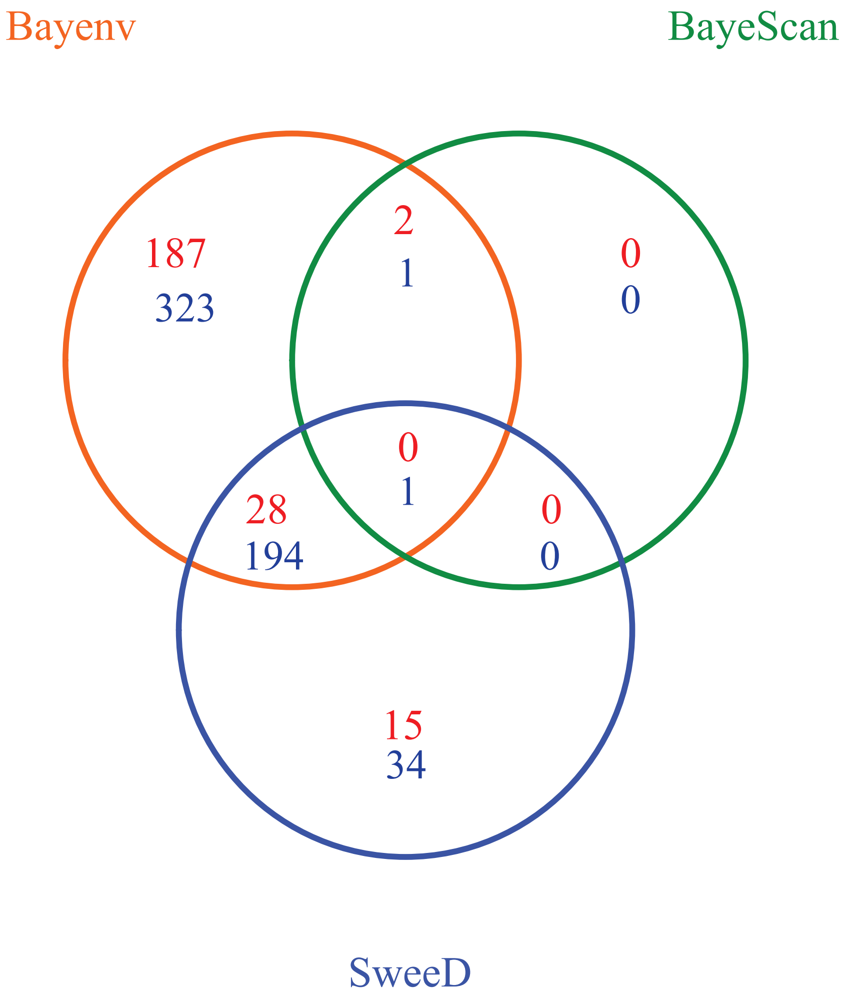


Figure 2-S14. Venn diagram showing overlap between candidate genes from SweeD, BayeScan, and Bayenv at two significance thresholds (top, red is p≤0.01; bottom, blue is p≤0.05).

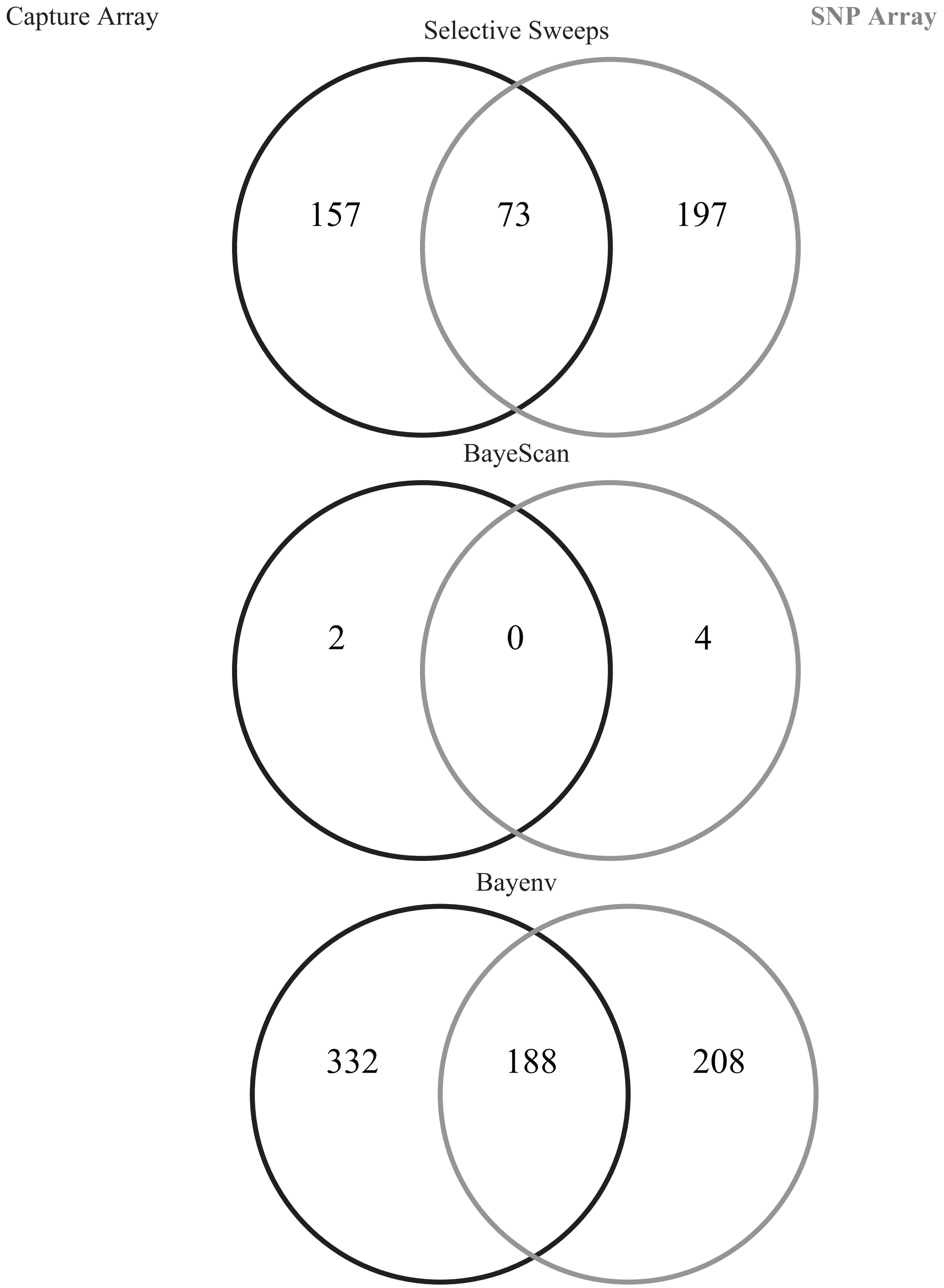


Figure 2-S15. Venn diagram showing overlap between candidate genes from SNP array and capture array for each selection test. The number of genes with significant variants identified at $p \leq 0.05$ from capture array data (left, black) and SNP array data (right, gray). The “Selective Sweeps” category refers to SweeD analysis with capture array data and Fst/XP-EHH analysis with SNP array data (see methods and Schweizer, et al. submitted for details).

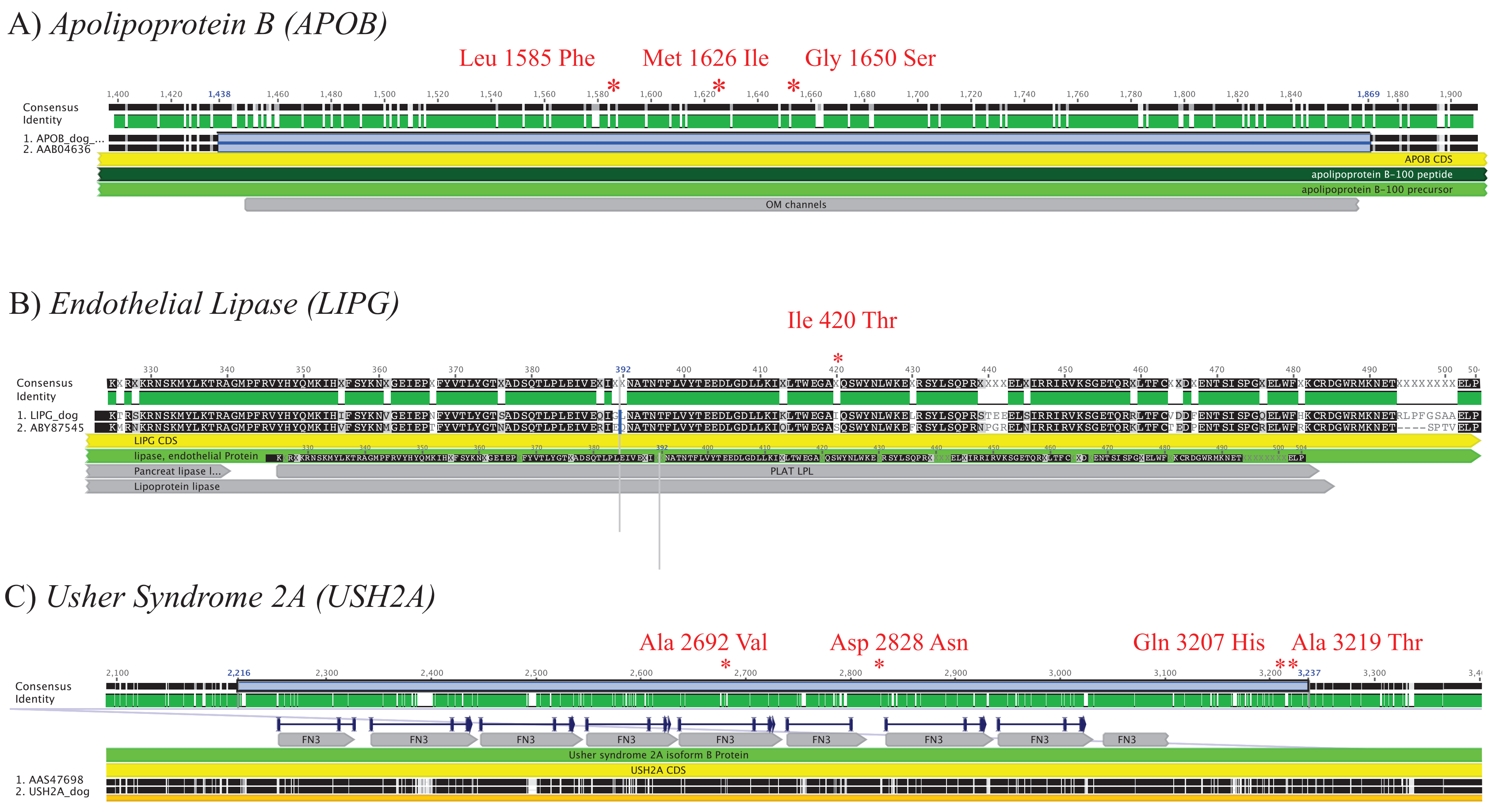


Figure 2-S16. Protein alignments between human and dog for A) APOB, B) LIPG, and C) USH2A. For each protein, a subset of the entire sequence is shown, with consensus and identity sequences generated within the Geneious program. The functional domains within each protein section (gray) are shown and the mutation and amino acid change within the protein sequence are noted in red.

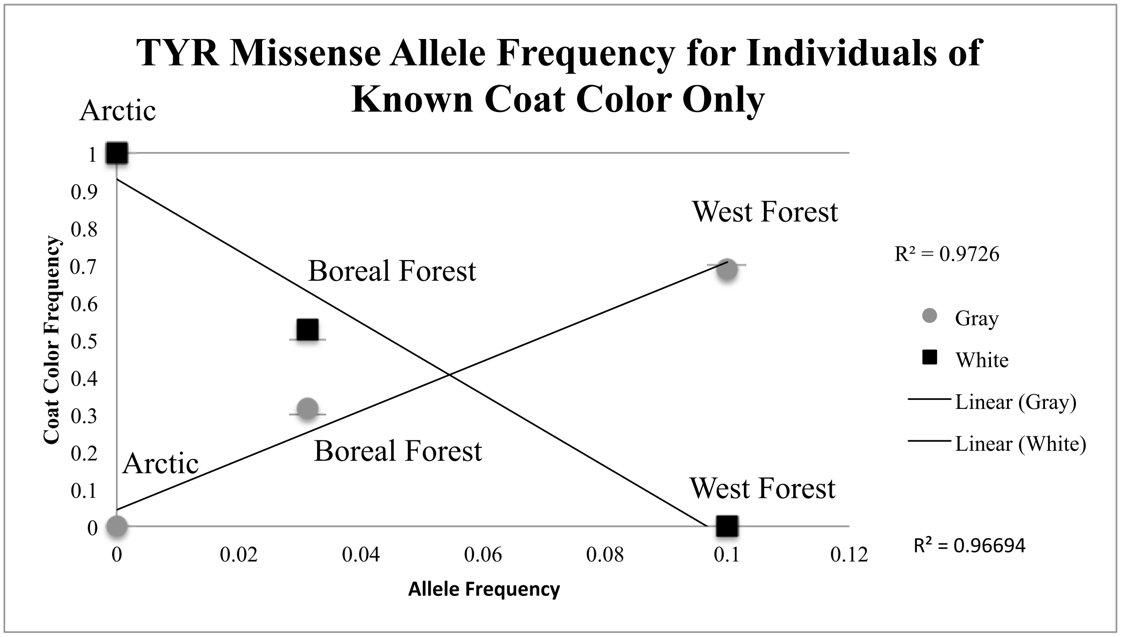


Figure 2-S17. Correlation of allele frequency of missense TYR with wolf coat color. Among populations with known phenotype (Arctic, Boreal Forest, and West Forest), the allele frequency of the non-reference variant is plotted against the frequency of white (black squares) or gray (gray circles) coat color.

Bibliography

Ache BW, Young JM (2005) Olfaction: Diverse Species, Conserved Principles. *Neuron*, **48**, 417–430.

Adriani M, Martinez-Mir A, Fusco F *et al.* (2004) Ancestral Founder Mutation of the Nude (FOXN1) Gene in Congenital Severe Combined Immunodeficiency Associated with Alopecia in Southern Italy Population. *Annals of Human Genetics*, **68**, 265–268.

Adzhubei IA, Schmidt S, Peshkin L *et al.* (2010) A method and server for predicting damaging missense mutations. *Nature Publishing Group*, **7**, 248–249.

Akey JM, Ruhe AL, Akey DT *et al.* (2010) Tracking footprints of artificial selection in the dog genome. *Proceedings of the National Academy of Sciences*, **107**, 1160–1165.

Akey J, Zhang G, Zhang K, Jin L, Shriver M (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Alagramam KN, Yuan H, Kuehn MH *et al.* (2001) Mutations in the novel protocadherin PCDH15 cause Usher syndrome type 1F. *Human Molecular Genetics*, **10**, 1709–1718.

Albert F, Hodges E, Jensen J, Besnier F (2011) Targeted resequencing of a genomic region influencing tameness and aggression reveals multiple signals of positive selection. *Heredity* **107,** 205–214.

Alexander D, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655.

Allen AP, Brown JH, Gillooly JF (2002) Global biodiversity, biochemical kinetics, and the energetic-equivalence rule. *Science*, **297**, 1545–1548.

Amrine-Madsen H, Koepfli K-P, Wayne RK, Springer MS (2003) A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. *Molecular phylogenetics and evolution*, **28**, 225–240.

Anderson TM, Candille SI, Musiani M *et al.* (2009) Molecular and evolutionary history of melanism in North American gray wolves. *Science*, **323**, 1339–1343.

Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, **22**, 195–201.

Baird N, Etter P, Atwood T *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Barsh G (1996) The genetics of pigmentation: from fancy genes to complex traits. *Trends in*

Genetics, **12**, 299–305.

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Bi K, Linderoth T, Vanderpool D *et al.* (2013) Unlocking the vault: next-generation museum population genomics. *Molecular Ecology*, **22**, 6018–6032.

Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC genomics*, **13**, 403–417.

Blair LM, Granka JM, Feldman MW (2014) On the stability of the Bayenv method in assessing human SNP-environment associations. *Human Genomics*, **8**.

Bonin A, NICOLE F, Pompanon F, MIAUD C, Taberlet P (2007) Population Adaptive Index: a New Method to Help Measure Intraspecific Genetic Diversity and Prioritize Populations for Conservation. *Conservation Biology*, **21**, 697–708.

Bryan HM, Smits JEG, Koren L *et al.* (2014) Heavily hunted wolves have higher stress and reproductive steroids than wolves with lower hunting pressure. *Functional Ecology*, **29**, 347–356.

Burbano HA, Hodges E, Green RE *et al.* (2010) Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science*, **328**, 723–725.

Candille SI, Kaelin CB, Cattanach BM *et al.* (2007) A -defensin mutation causes black coat color in domestic dogs. *Science*, **318**, 1418–1423.

Carmichael LE, Krizan J, Nagy JA *et al.* (2007) Historical and ecological determinants of genetic structure in arctic canids. *Molecular Ecology*, **16**, 3466–3483.

Charvet B, Guiraud A, Malbouyres M *et al.* (2013) Knockdown of col22a1 gene in zebrafish induces a muscular dystrophy by disruption of the myotendinous junction. *Development*, **140**, 4602–4613.

Chen J, Källman T, Ma X *et al.* (2012a) Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce (Picea abies). *Genetics*, **191**, 865–881.

Chen R, Irwin DM, Zhang Y-P (2012b) Differences in Selection Drive Olfactory Receptor Genes in Different Directions in Dogs and Wolf. *Molecular Biology and Evolution*, **29**, 3475–3484.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, **7**, e46688.

Clemente FJ, Cardona A, Inchley CE *et al.* (2014) REPOR TA Selective Sweep on a Deleterious Mutation in CPT1A in Arctic Populations. *American Journal of Human Genetics*, **95**, 584–589.

Cong L, Ran FA, Cox D *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, **185**, 1411–1423.

Coulson T, Macnulty DR, Stahler DR *et al.* (2011) Modeling Effects of Environmental Change on Wolf Population Dynamics, Trait Evolution, and Life History. *Science*, **334**, 1275–1278.

Crisci JL, Poh Y-P, Mahajan S, Jensen JD (2013) The impact of equilibrium assumptions on tests of selection. *Frontiers in genetics*, **4**, 1–7.

Cunliffe VT, Furley AJW, Keenan D (2014) Complete rescue of the nude mutant phenotype by a wild-type Foxn1 transgene. *Mammalian genome : official journal of the International Mammalian Genome Society*, **13**, 245–252.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, **27**, 2156–2158.

Darimont CT, Paquet PC (2000) The Gray Wolves (Canis lupus) of British Columbia's Coastal Rainforests: Findings from Year 2000 Pilot Study and Conservation Assessment. *Prepared for the Raincoast Conservation Society*, 1–72.

DePristo M, Banks E, Poplin R, Garimella K (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491-498.

Dionne M, Miller KM, Dodson JJ, Caron F, Bernatchez L (2007) Clinal variation in MHC diversity with temperature: Evidence for the role of host-pathogen interaction on local adaptation in Atlantic salmon. *Evolution*, **61**, 2154–2164.

Domingues VS, Poh Y-P, Peterson BK *et al.* (2012) Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution*, **66**, 1–15.

Dreyer B, Tranebjærg L, Rosenberg T *et al.* (2000) Identification of novel USH2A mutations: implications for the structure of USH2A protein. *European journal of human genetics : EJHG*, **8**, 500–506.

Edmondson AC, Brown RJ, Kathiresan S *et al.* (2009) Loss-of-function variants in endothelial lipase are a cause of elevated HDL cholesterol in humans. *Journal of Clinical Investigation*, **119**, 1042-1050.

Egashira Y, Murotani G, Tanabe A *et al.* (2004) Differential effects of dietary fatty acids on rat liver α-amino-β-carboxymuconate-ε-semialdehyde decarboxylase activity and gene expression. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, **1686**, 118–124.

Fagny M, Patin E, Enard D *et al.* (2014) Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets. *Molecular Biology and Evolution*, **31**, 1850–1868.

Faircloth, BC. 2015. Illumina TruSeq Library Prep for Target Enrichment. Available from http://ultraconserved.org (Accessed March 10, 2013).

Faircloth BC, Glenn TC (2012) Not All Sequence Tags Are Created Equal: Designing and Validating Sequence Identification Tags Robust to Indels. *PLoS ONE*, **7**, e42543.

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.

Farese RV, Ruland SL, Flynn LM, Stokowski RP, Young SG (1995) Knockout of the mouse apolipoprotein B gene results in embryonic lethality in homozygotes and protection against diet-induced hypercholesterolemia in heterozygotes. *PNAS*, **92**, 1774–1778.

Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, **180**, 977–993.

Fraser DJ, Bernatchez L (2001) Adaptive evolutionary conservation: towards a unified concept for defining conservation units. *Molecular Ecology*, **10**, 2741–2752.

Freedman AH, Gronau I, Schweizer RM *et al.* (2014) Genome Sequencing Highlights the Dynamic Early History of Dogs (L Andersson, Ed,). *PLoS Genetics*, **10**, e1004016.

Fu W, O'connor TD, Jun G *et al.* (2012) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, **27**, 489–496.

Gasteiger E (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, **31**, 3784–3788.

Gebremedhin B, Ficetola GF, Naderi S *et al.* (2009) Frontiers in identifying conservation units: from neutral markers to adaptive genetic variation. *Animal Conservation*, **12**, 107–109.

Gilad Y, Bustamante CD, Lancet D, Pääbo S (2003) Natural selection on the olfactory receptor gene family in humans and chimpanzees. *American Journal of Human Genetics*, **73**, 489–501.

Gilg O, Kovacs KM, Aars J *et al.* (2012) Climate change and the ecology and evolution of Arctic vertebrates. *Annals of the New York Academy of Sciences*, **1249**, 166–190.

Gipson P, Bangs E, Bailey T *et al.* (2002) Color patterns among wolves in western North America. *Wildlife Society Bulletin*, **30**, 821–830.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, **27**, 1017–1018.

Gray MM, Granka JM, Bustamante CD *et al.* (2009) Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, **181**, 1493–1505.

Grossman S, Shylakhter I, Karlsson E *et al.* (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, **327**, 879–883.

Guernier V, Hochberg ME, Guégan J-F (2004) Ecology Drives the Worldwide Distribution of Human Diseases. *PLoS Biology*, **2**, e141.

Hancock A, Witonsky D, Ehler E *et al.* (2010) Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences*, **107**, 8924.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hilton C, Neville MJ, Karpe F (2012) MicroRNAs in adipose tissue: their role in adipogenesis and obesity. *International Journal of Obesity*, **37**, 325–332.

Hodges E, Xuan Z, Balija V *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, **39**, 1522–1527.

Hoekstra HE (2006) Genetics, development and evolution of adaptive pigmentation in vertebrates. *Heredity*, **97**, 222–234.

Holmkvist J, Tojjar D, Almgren P *et al.* (2007) Polymorphisms in the gene encoding the voltage-dependent Ca2+ channel CaV2.3 (CACNA1E) are associated with type 2 diabetes and impaired insulin secretion. *Diabetologia*, **50**, 2467–2475.

Hubbard JK, Uy JAC, Hauber ME, Hoekstra HE, Safran RJ (2010) Vertebrate pigmentation: from underlying genes to adaptive function. *Trends in genetics*, **26**, 231–239.

Hume AN, Tarafder AK, Ramalho JS, Sviderskaya EV, Seabra MC (2006) A coiled-coil domain of melanophilin is essential for Myosin Va recruitment and melanosome transport in melanocytes. *Molecular biology of the cell*, **17**, 4720–4735.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics (Oxford, England)*, **23**, 1801–1806.

Jolicoeur P (1959) Multivariate geographical variation in the wolf Canis lupus L. *Evolution*, **13**, 283–299.

Jones SG (2004) The microRNA registry. *Nucleic Acids Research*, **32**, 109-111.

Kaelin CB, Barsh GS (2013) Genetics of Pigmentation in Dogs and Cats. *Annual Review of Animal Biosciences*, **1**, 125–156.

Kang HM, Sul JH, Service SK *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**, 348–354.

Kass RE, Raftery AE (1995) Bayes factors. *Journal of the american statistical association*, **90**, 773–795.

Kearse M, Moir R, Wilson A *et al.* (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics (Oxford, England)*, **28**, 1647–1649.

Keller A, Vosshall L (2008) Better smelling through genetics: mammalian odor perception. *Current opinion in neurobiology*, **18**, 364–369.

Keller A, Zhuang H, Chi Q, Vosshall LB, Matsunami H (2007) Genetic variation in a human odorant receptor alters odour perception. *Nature*, **449**, 468–472.

Kennedy LJ, Angles JM, Barnes A *et al.* (2007) DLA-DRB1, DQA1, and DQB1 Alleles and Haplotypes in North American Gray Wolves. *Journal of Heredity*, **98**, 491–499.

Kolditz CI, Paboeuf G, Borthaire M *et al.* (2008) Changes induced by dietary energy intake and divergent selection for muscle fat content in rainbow trout (Oncorhynchus mykiss), assessed by transcriptome and proteome analysis of the liver. *BMC genomics*, **9**, 506.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. *BMC bioinformatics*, **15**, 356.

Laramie JM, Wilk JB, Williamson SL *et al.* (2008) Polymorphisms near EXOC4 and LRGUK on chromosome 7q32 are associated with Type 2 Diabetes and fasting glucose; The NHLBI Family Heart Study. *BMC Medical Genetics*, **9**, 46.

Lee HK, Deneen B (2012) Daam2 Is Required for Dorsal Patterning via Modulation of Canonical

Wnt Signaling in the Developing Spinal Cord. *Developmental Cell*, **22**, 183–196.

Le Guédard S, Faugère V, Malcolm S, Claustres M, Roux A-F (2007) Large genomic rearrangements within the PCDH15 gene are a significant cause of USH1F syndrome. *Molecular Vision*, **13**, 102–107.

Leonard JA, Vilà C, Wayne RK (2005) Legacy lost: genetic variability and population size of extirpated US grey wolves (Canis lupus). *Molecular Ecology*, **14**, 9–17.

Lewandoski M (2001) Conditional control of gene expression in the mouse. *Nature Publishing Group*, **2**, 743–755.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **26**, 589.

Li H-D, Menon R, Omenn GS, Guan Y (2014) The emerging era of genomic data integration for analyzing splice isoform function. *Trends in genetics : TIG*, **30**, 340–347.

Li WD, Reed DR, Lee J-H *et al.* (1999) Sequence variants in the 5′ flanking region of the leptin gene are associated with obesity in women. *Annals of Human Genetics*, **63**, 227–234.

Liaw A, Wiener M (2002) Classification and regression by randomForest. *R news*, **2**, 18–22.

Liu S, Lorenzen ED, Fumagalli M *et al.* (2014) Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears. *Cell*, **157**, 785–794.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lynch VJ, Bedoya-Reina O, Ratan A *et al*. (2015) Elephantid genomes reveal the molecular bases of Woolly Mammoth adaptations to the arctic. *bioRxiv* doi:/10.1101/018366;

Lyons PJ, Lang-Ouellette D, Morin PJ (2013) Comparative Biochemistry and Physiology, Part D. *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, **8**, 358–364.

MacNulty DR, Smith DW, Mech LD, Eberly LE (2009) Body size and predatory performance in wolves: is bigger better? *The Journal of animal ecology*, **78**, 532–539.

Mahlstein I, Knutti R (2012) September Arctic sea ice predicted to disappear near 2°C global warming above present. *Journal of Geophysical Research*, **117**, D06104.

Mammès O, Betoulle D, Aubert R *et al.* (1998) Novel polymorphisms in the 5'region of the LEP gene: association with leptin levels and response to low-calorie diet in human obesity. *Diabetes*, **47**, 487–489.

Manichaikul A, Mychaleckyj JC, Rich SS *et al.* (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, **26**, 2867–2873.

Martin SL (2008) Mammalian hibernation: a naturally reversible model for insulin resistance in man? *Diabetes & vascular disease research : official journal of the International Society of Diabetes and Vascular Disease*, **5**, 76.

McLaren W, Pritchard B, Rios D *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)*, **26**, 2069–2070.

Mech LD (1970) The Wolf: The Ecology and Behavior of an Endangered Species. University of Minnesota Press, Minneapolis, Minnesota.

Mech LD (2004) Is climate change affecting wolf populations in the high arctic? *Climatic Change*, **67**, 87–93.

Miyata T, Miyazawa S, Yasunaga T (1979) Two types of amino acid substitutions in protein evolution. *Journal of molecular evolution*, **12**, 219–236.

Muñoz Fuentes V, Darimont C, Wayne R, Paquet P, Leonard J (2009) Ecological factors drive differentiation in wolves from British Columbia. *Journal of Biogeography*, **36**, 1516–1531.

Musiani M, Leonard JA, Cluff HD *et al.* (2007) Differentiation of tundra/taiga and boreal coniferous forest wolves: genetics, coat colour and association with migratory caribou. *Molecular Ecology*, **16**, 4149–4170.

Nakamura M, Tobin DJ, Richards-Smith B, Sundberg JP, Paus R (2002) Mutant laboratory mice with abnormalities in pigmentation: annotated tables. *Journal of dermatological science*, **28**, 1–33.

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, **31**, 3812–3814.

Ng SB, Turner EH, Robertson PD *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, **8**, 857–868.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE*, **7**, e37558.

Nielsen R, Williamson SH, Kim Y *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Nolan DK, Sutton B, Haynes C *et al.* (2012) Fine mapping of a linkage peak with integrationof lipid traits identifies novel coronary arterydisease genes on chromosome 5. *BMC genetics*, **13**, 12.

O'Keefe FR, Meachen J, Fet EV, Brannick A (2013) Ecological determinants of clinal morphological variation in the cranium of the North American gray wolf. *Journal of Mammalogy*, **94**, 1223–1236.

Odegaard JI, Chawla A (2013) Pleiotropic actions of insulin resistance and inflammation in metabolic homeostasis. *Science*, **339**, 172–177.

Ohkura T, Taniguchi S-I, Yamada K *et al.* (2004) Detection of the novel autoantibody (anti-UACA antibody) in patients with Graves' disease. *Biochemical and Biophysical Research Communications*, **321**, 432–440.

Osborne AJ, Pearson J, Negro SS *et al.* (2015) Heterozygote advantage at MHC DRBmay influence response to infectious disease epizootics. *Molecular Ecology*, **24**, 1419–1432.

Parker J, Tsagkogeorga G, Cotton JA *et al.* (2013) Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, **502**, 228–231.

Pavlidis P, Zivkovic D, Stamatakis A, Alachiotis N (2013) SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. *Molecular Biology and Evolution*, **30**, 2224–2234.

Perry GH (2014) The Promise and Practicality of Population Genomics Research with Endangered Species. *International Journal of Primatology*, **35**, 55–70.

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 *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Qanbari S, Pausch H, Jansen S *et al.* (2014) Classic Selective Sweeps Revealed by Massive Sequencing in Cattle (JK Pritchard, Ed,). *PLoS Genetics*, **10**, e1004148.

Quignon P, Giraud M, Rimbault M *et al.* (2005) The dog and rat olfactory receptor repertoires. *Genome Biology*, **6**, R83.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, **26**, 841–842.

Razzaghi H, Tempczyk-Russell A, Haubold K *et al.* (2013) Genetic and Structure-Function Studies of Missense Mutations in Human Endothelial Lipase. *PLoS ONE*, **8**, e55716.

Reimand J, Arak T, Vilo J (2011) g:Profiler--a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, **39**, W307–W315.

Reimand J, Kull M, Peterson H, Hansen J, Vilo J (2007) g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, **35**, W193–W200.

Reinhardt JA, Kolaczkowski B, Jones CD, Begun DJ, Kern AD (2014) Parallel geographic variation in Drosophila melanogaster. *Genetics*, **197**, 361–373.

Sabeti PC, Varilly P, Fry B *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Saihan Z, Webster AR, Luxon L, Bitner-Glindzicz M (2009) Update on Usher syndrome. *Current Opinion in Neurology*, **22**, 19–27.

Scheinfeldt LB, Tishkoff SA (2013) Recent human adaptation: genomic approaches, interpretation and insights. *Nature Reviews Genetics*, **14**, 692–702.

Shen W (2001) Foxa3 (Hepatocyte Nuclear Factor 3gamma) Is Required for the Regulation of Hepatic GLUT2 Expression and the Maintenance of Glucose Homeostasis during a Prolonged Fast. *Journal of Biological Chemistry*, **276**, 42812–42817.

Slater GJ, Dumont ER, Van Valkenburgh B (2009) Implications of predatory specialization for cranial form and function in canids. *Journal of Zoology*, **278**, 181–188.

Staples J, Nickerson DA, Below JE (2012) Utilizing Graph Theory to Select the Largest Set of Unrelated Individuals for Genetic Analysis. *Genetic Epidemiology*, **37**, 136–141.

Stevens ED, Kido M (1974) Active sodium transport: a source of metabolic heat during cold adaptation in mammals. *Comparative Biochemistry and Physiology Part A: Physiology*, **47**, 395–397.

Storey KB (2015) Regulation of hypometabolism: insights into epigenetic controls. *The Journal of experimental biology,* **218**, 150-159.

Storz JF, Runck AM, Moriyama H, Weber RE, Fago A (2010) Genetic differences in hemoglobin function between highland and lowland deer mice. *The Journal of experimental biology*, **213**, 2565–2574.

Sturm RA, Duffy DL (2012) Human pigmentation genes under environmental selection. *Genome Biology*, **13**, 248.

Tewhey R, Nakano M, Wang X *et al.* (2009) Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biology*, **10**, R116.

Thornton KR, Jensen JD (2007) Controlling the false-positive rate in multilocus genome scans for selection. *Genetics*, **175**, 737.

Tietjen I, Hovingh GK, Singaraja RR *et al.* (2012) Segregation of LIPG, CETP, and GALNT2 Mutations in Caucasian Families with Extremely High HDL Cholesterol (H Schunkert, Ed,). *PLoS ONE*, **7**, e37437.

vonHoldt BM, Pollinger JP, Earl DA *et al.* (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Research*, **21**, 1-12.

Wabakken P, Sand H, Kojola I *et al.* (2007) Multistage, Long-Range Natal Dispersal by a Global Positioning System-Collared Scandinavian Wolf. *The Journal of wildlife management*, **71**, 1631–1634.

Wagner JL, Burnett RC, DeRose SA, Storb R (1996) Molecular analysis and polymorphism of DLA-DQA gene, *Tissue Antigens*, **48**, 199-204.

White TA, Perkins SE, Heckel G, Searle JB (2013) Adaptive evolution during an ongoing range expansion: the invasive bank vole (Myodes glareolus) in Ireland. *Molecular Ecology*, **22**, 2971–2985.

Williams DS (2008) Usher syndrome: Animal models, retinal function of Usher proteins, and prospects for gene therapy. *Vision Research*, **48**, 433–441.

Williamson SH, Hubisz MJ, Clark AG *et al.* (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genetics*, **3**, e90.

Wu C-W, Biggar KK, Storey KB (2013) Dehydration mediated microRNA response in the African clawed frog Xenopus laevis. *Gene*, **529**, 269–275.

Xie X, Lu J, Kulbokas EJ *et al.* (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

Xu L, Panel V, Ma X *et al.* (2013) The Winged Helix Transcription Factor Foxa3 Regulates Adipocyte Differentiation and Depot-Selective Fat Tissue Expansion. *Molecular and Cellular Biology*, **33**, 3392–3399.

Yamada K, Senju S, Nakatsura T *et al.* (2001) Identification of a Novel Autoantigen UACA in Patients with Panuveitis. *Biochemical and Biophysical Research Communications*, **280**, 1169–1176.

Yan D, Liu XZ (2010) Genetics and pathological mechanisms of Usher syndrome. *Journal of Human Genetics*, **55**, 327–335.

Yang J, Wang ZL, Zhao XQ *et al.* (2008) Natural selection and adaptive evolution of leptin in the ochotona family driven by the cold environmental stress. *PLoS ONE*, **3**, e1472.

Young SG (1990) Recent progress in understanding apolipoprotein B. *Circulation*, **82**, 1574–1594.

Zaragosi L-E, Wdziekonski B, Le Brigand K *et al.* (2011) Small RNA sequencing reveals miR-642a-3p as anovel adipocyte-specific microRNA and miR-30 as a key regulator of human adipogenesis. *Genome Biology*, **12**, R64.

Zheng Y, Vertuani S, Nystrom S *et al.* (2009) Angiomotin-Like Protein 1 Controls Endothelial Polarity and Junction Stability During Sprouting Angiogenesis. *Circulation Research*, **105**, 260–270.

Zhang W, Fan Z, Han E *et al.* (2014) Hypoxia Adaptations in the Grey Wolf (Canis lupus chanco) from Qinghai-Tibet Plateau. *PLoS Genetics*, **10**, e1004466.

Patterns of nucleotide and haplotype diversity of the K locus

Introduction

One of the most distinct phenotypes in nature is coloration, which is usually determined by the spatial distribution of pigmented hairs, fur, or scales across the body and pigment type (Protas & Patel 2008; Manceau *et al.* 2010). Coloration is often subject to positive selection (e.g. Hoekstra & Nachman 2003). In mammals, there are two main types of pigments that control coloration, eumelanin (black/brown) and pheomelanin (red/yellow) (Protas & Patel 2008). In many vertebrates, switching of these two pigment types is controlled by the *Agouti* (ligand) - *Mc1r* (receptor) pathway (Protas & Patel 2008). Although many melanistic phenotypes are caused by a mutation in *Agouti* or *Mc1r*, melanism in the gray wolf is caused by a mutation in the *CBD103* gene (also referred to as the K locus), which encodes an alternative ligand for *Mc1r* that outcompetes a functioning *Agouti* ligand if present (Candille *et al.* 2007; Anderson *et al.* 2009). The K^B allele contains a 3bp deletion that confers a dominantly inherited black (melanistic) coat color phenotype, whereas the wild-type K^y allele confers a gray (agouti) coat color in homozygotes (Anderson *et al.* 2009). In North American gray wolves, clinal variation in the frequency of the K^B allele corresponds with a transition from boreal coniferous forest (~50% black) to tundra/taiga (15% black) (Musiani *et al.* 2007; Anderson *et al.* 2009). Additionally, almost no polymorphism is observed within 60 Kb of K^B haplotypes, yet the wild-type K^y allele is highly polymorphic (Anderson *et al.* 2009). Together, these data support a molecular signature of positive selection for the K^B allele (Maynard-Smith & Haigh 1974). Previous SNP assays in wolves suggest that the K^B allele was introduced into the genome of North American wolves from the domestic dog via hybridization events that likely occurred between North American dogs and wolves after humans arrived in the New World around 13,000 years ago (Leonard *et al.* 2002; Anderson *et al.* 2009).

Coloration can serve several functions, including intraspecific and interspecific signaling and thermoregulation (Protas & Patel 2008). Much of the recent work exploring selection on coat color in mammals has involved mice (*Peromyscus* spp), where selective forces have favored mice that are better able to camouflage and avoid predation (Nachman *et al.* 2003; Mullen & Hoekstra 2008; Vignieri *et al.* 2010). In wolves, the selective advantage of light or dark coat color is not as clear. Wolf coat color varies from black to white, with tawny variations, with higher gray and white phenotype frequencies in the high Arctic where background vegetation is lighter (Musiani *et al.* 2007). However, gray and black individuals exist within the same populations, indicating that there may be fitness tradeoffs for different coat colors due to pleiotropic effects (Gipson *et al.* 2002; Musiani *et al.* 2007; Anderson *et al.* 2009; Coulson *et al.* 2011). In Yellowstone National Park (YNP), roughly 47% of wolves are black and 52% of wolves are gray, ranging from 35-76% (black) and 24-63% (gray) over the past 15 years (Coulson *et al.* 2011). An analysis of data from repeated observations of female wolves in YNP revealed that at the individual level, reproductive fitness is higher in gray wolves (i.e. larger litter size), yet black individuals have a higher overall survival rate (Coulson *et al.* 2011). Of interest in this regard is that the K locus is a member of the β-defensin family of antimicrobial peptides (Pazgier *et al.* 2006) and may be involved in adaptive immune response (Yang *et al.* 1999). Thus the apparent fitness benefit of larger litter sizes of gray female wolves may come at a cost to their immunity, as evident by a longer lifespan in black female wolves. Additionally, allele frequencies within *CBD103* are significantly associated with temperature variables across North American wolf ecotypes (Schweizer *et al*, submitted), and temperature and pathogen prevalence are positively related in a multitude of species (Allen *et al.* 2002; Guernier *et al.* 2004; Dionne *et al.* 2007), further supporting the notion that selection within *CBD103* likely involves immunity.

When a new beneficial mutation experiences positive selection, such as the mutation in the K locus, it will increase in frequency within a population along with linked neutral loci resulting in a “selective sweep” (Maynard-Smith & Haigh 1974). Neutral alleles close to the locus under positive selection experience a “hitchhiking effect” (Maynard-Smith & Haigh 1974) and when the favored allele goes to fixation, the selective sweep is considered complete. The effects of a selective sweep are strongest in the immediate vicinity of the locus under selection and will fade with increasing genetic distance. Recombination in the region can diminish the effects of a selective sweep by breaking up large haplotype blocks created by hitchhiking effects. Other factors, such as population demography and variation in recombination rate, can affect the ability to detect selection (Nielsen *et al.* 2007).

The previous study by Anderson *et al.* (2009) used a relatively small data set of approximately 50 SNPs genotyped in 47 Arctic and Yellowstone wolves from forest and tundra/taiga habitats, and was unable to address the details of the evolutionary history of the K locus due to limited genomic sequencing in the flanking regions of the locus. Thus, we designed a custom capture array to perform extensive re-sequencing of five megabases (Mb) surrounding the K locus core deletion in a larger sample of North American wolves from multiple areas to assess patterns of nucleotide and haplotype diversity, population-specific decay in linkage disequilibrium (LD), and hierarchical patterns of genetic divergence among populations. Additional genomic controls in the form of telomeric and non-telomeric regions provided an empirical background for diversity measured at similarly telomeric regions without selection.

These data provide insight into the evolutionary history of the K^B allele, and from which we infer that adaptive introgression most likely occurred first in the Northwest Territories or Yukon area of Canada, when native dogs and humans were co-existing in the Arctic. Furthermore, we find evidence for a strong, ongoing selective sweep in Yellowstone wolves that may be related to immunity and disease prevalence. These initial results form the basis for planned future analyses including extensive demographic modeling of the sweep, further assessment of the number, location and timing of introgression events, and estimating the strength and context of selection at the K locus in North American wolves. The history of selection at the K locus provides a striking example of a case of adaptive introgression, wherein a mutation originating in a domestic species was transferred and swept to high frequency in a wild progenitor.

Methods

Sampling

We selected wolf samples to maximize the following parameters: 1) the geographic distribution of samples across regions and ecotypes (Schweizer *et al*. submitted); 2) the number of samples with known coat color phenotype and life history data; and 3) the quality and amount of DNA. A total of 403 samples were used, and, after genotype and sample quality filtering, 381 samples were retained for analysis (Table 3-1). Samples from Yellowstone National Park (n=203) were chosen to maximize the number of known pedigree relationships (vonHoldt *et al.* 2008) and the life history data available for each individual (Stahler *et al.* 2012; Yellowstone Wolf Project, WY). The Yellowstone wolf population was founded in 1995 from two large Canadian source populations from Alberta and British Columbia, with additional wolves from

Northern Montana added the following year (Bangs & Fritz 1996), and has been closely studied since then (e.g. vonHoldt *et* al. 2008; Stahler *et al.* 2012). More North American wolves (n=130) were picked to represent a wide geographic distribution and multiple distinct ecotypes, and included many samples that were previously genotyped on the Affymetrix Dog SNP array (vonHoldt *et al.* 2010a; 2011; Schweizer *et al*. submitted). Wolves from Europe (n=12) and Asia (n=4) were chosen as "controls" for locations where no black wolves had been observed. Italian wolves (n=8) and known dog-wolf hybrids (n=3) were chosen that have black or gray phenotypes (M. Galaverni, unpublished data). Finally, purebred dog samples (n=21) from 20 dog breeds were chosen to represent different K allele genotypes (Candille *et al*. 2007) (Table 3-1).

Array Design

The capture array was designed with two main aims: 1) to reconstruct the origin, spread and ecological context of selection on the K^B allele; and 2) to estimate the ecotype-specific strength of selection on the K locus.

In order to address these aims, we extracted putatively "neutral" regions from the dog reference genome (CanFam3.1; Figure 3-1A) for background estimates of individual relatedness and population demography. Details of the design of these regions have been described elsewhere (Freedman *et al.* 2014; Schweizer *et al.* submitted) and follow guidelines set by previous studies in humans (Wall *et al.* 2008). Briefly, using the dog genome annotation (CanFam3.1) as a reference, we identified 1 Kb regions that were at least 100 Kb from any known or predicted genes, were not within highly repetitive regions of the dog genome, were within uniquely mapping regions of the genome as computed by TALLYMER (Kurtz *et al.*

2008), had PhastCons scores <0.5, and had GC content within two standard deviations of mean dog genome GC content. A total of 5073 autosomal 1 Kb regions were identified.

To specifically address the first aim, we designed an extensive resequencing approach for almost 5 Mb surrounding the 3bp causative mutation on chromosome 16. This included the 200 Kb "core" region partially sequenced previously (Anderson *et al.* 2009), plus 1 Kb segments spaced every 10 Kb, extending to the end of the chromosome ~560 Kb downstream of the mutation (the K locus is near the telomeres) and 4.2 Mb upstream of the mutation (henceforth called the "surrounding 5 Mb region"; Figure 3-1B). Furthermore, to test how the proximity of the K locus to the end of the chromosome impacts decay of LD and diversity statistics, we sequenced ten 1Kb fragments spaced on a 200Kb segment in non-telomeric regions for five larger and five smaller chromosomes, plus chromosome 16 (Figure 3-1B). Each non-telomeric region began at the mid-point of the chromosome. Finally, as an empirical background for selection at the K locus, we assessed decay of LD in other telomeric regions by similarly designing ten 1Kb fragments spaced on a 200 Kb segment in telomeric regions for five larger and five smaller chromosomes (Figure 3-1B). Each telomeric region began at the same relative distance from the end of the chromosome as the K locus mutation.

The array was also designed to capture a total of 1100 genes that were mostly used for selection studies in North American wolves (Schweizer *et al.* submitted). The coding region exons, plus 1Kb upstream of the transcript start site, were targeted for bait design. We used these gene sequence data here to assess the ranking of the K locus amongst other genes, in dogs versus wolves (see below).

In total, we designed regions to capture approximately 7 Mb of sequence from each individual. To do this, approximately 105,000 120bp RNA baits were designed by MYcroarray (Ann Arbor, Michigan) so as to maximize specificity of bait hybridization and unique mapping within the genome. These baits covered approximately 91% of the regions we aimed to capture.

Library Prep, Target Enrichment, and Sequencing

Samples were prepared as described previously (Schweizer *et al*. submitted). Briefly, we extracted genomic DNA from blood or tissue, and then sheared DNA of high quality and quantity using a Biorupter NGS Sonication System (Diagenode). Samples were sheared to approximately 300-450 bp fragment size, and randomized with respect to extraction, library prep and enrichment, and sequencing dates. Preparation of sequencing libraries followed the with-bead library preparation protocol of Faircloth *et al*. (2015), and each sample was barcoded with a unique 6 bp index sequence during adapter ligation so as to enable pooling of 24-25 individuals per lane (Faircloth & Glenn 2012). At the end of library preparation, samples were amplified with PCR as follows: 98˚C for 45 seconds, 16 cycles of 98˚C for 15 seconds, 60˚C for 30 seconds, 72˚C for 60 seconds, and a final step of 72˚C for 5 minutes. Concentration of amplified samples was measured with a Qubit dsDNA High Sensitivity kit, and samples with >500ng of library were dried down with a Speed Vac and reconstituted to 147 ng/µl. Subsequent to library preparation, samples were target enriched following the manufacturer's protocol (MYbaits by MYcroarray), with a 24 hour hybridization at 65˚C in an Eppendorf PCR machine. Samples were PCR amplified a second time using the above protocol.

In order to check that libraries were properly enriched, we performed enrichment qPCR on all samples prior to sequencing, following the recommendations for the Roche NimbleGen

Arrays (SeqCap EZ Library Prep Guide). For each sample, amplification of three ~100 bp on-target regions (an exon within the K locus gene, an exon within a second candidate gene, and a neutral region) and one off-target region (a single-copy control gene) was performed. Primers of 60-80 bp were designed in Primer3 (http://biotools.umassmed.edu/bioapps/primer3_www.cgi) using the CanFam3.1 genome sequence. Amplifications were done in duplicate, for 50 ng of genomic DNA, library, and enriched library, for each primer set and each sample. The qPCR Mastermix used the Roche High Resolution Melting Mix and standard qPCR cycle conditions. For a sample to be sufficiently enriched for sequencing the on-target regions had to be enriched by 20-200-fold and the off-target region had to show no enrichment. Enriched libraries were quantified, pooled in equimolar amounts, and then sequenced on a HiSeq 2000 with 100bp paired-end reads by the QB3 Vincent J. Coates Genomics Sequencing Laboratory (Berkeley, California, USA).

Sequence Alignment and Processing

Sequence alignment and processing followed the general recommendations of the Broad `Genome Analysis ToolKit` "Best Practices" pipeline (https://www.broadinstitute.org/gatk/guide/best-practices), with the specifics of our processing protocol published elsewhere (Schweizer *et al*. submitted). In short, demultiplexed fastq reads passing the Illumina filter were trimmed for remaining adapter sequences, then forward and reverse reads were aligned and mapped to the reference boxer genome (CanFam3.1) using `bwa aln` and `bwa sampe` with an insert size of 1000bp between paired ends (Li 2014). Previous work suggests that aligning wolf sequences to this reference produces high quality genotype calls and minimal reference bias due to the very short sequence divergence of wolves and dogs (~0.1%; Freedman *et al.* 2014). After duplicate removal with `samtools rmdup` (Li *et al*. 2009), local realignment with GATK, and fixing mate information with picard tools (http://broadinstitute.github.io/picard/), we ran GATK Base Quality Score Recalibration using a previously generated set of "known" variant sites (Schweizer *et al.* submitted). The GATK UnifiedGenotyper algorithm was used to call SNPs and indels (insertions and deletions) over the capture array intervals with a padding of 1000 bp.

Genotype Filtration

Variant positions identified by the Unified Genotyper were filtered with GATK VariantFiltration using ten filter expressions, as recommended by the Broad Best Practices pipeline. Using 5000 random variant positions from the vcf file, we generated histograms of the values for each of these annotations to justify the use of these filter values. In addition to the filters recommended by the Broad, we removed variant positions with genotype quality (GQ) < 30 and all positions with minimum depth (DP) ≤ 10.

Data Quality Control

Genotype concordance was assessed for 109 individuals and 204 sites that overlapped between the Affymetrix dog SNP array v2 and the capture array target intervals. Using the vcftools package (Danecek *et al.* 2011), we calculated the sequence-wide heterozygosity, transition/transversion ratio, site missingness and individual missingness for each set of regions.

The K locus indel genotype had previously been determined for 235 individuals using either Sanger sequence (Anderson *et al.* 2009) or high resolution melt curve analysis (Coulson *et al.* 2011; Schweizer & vonHoldt, unpublished data). We used these existing data to check quality control in our sequence samples.

We performed principal components analysis (PCA) within the `smartpca` package of `eigenstrat` (Price *et al.* 2006). First, we wanted to make sure that samples and various data set types (i.e. neutral) behaved according to what might be expected based on previous studies (vonHoldt *et al.* 2011; Pilot *et al.* 2013; Schweizer *et al*. submitted). Samples that did not group according to their expected population or species were dropped from further analysis. For this, we generated a set of LD-pruned SNPs using the "`--indep-pairwise 50 5 0.5`" option in `PLINK` (Purcell *et al.* 2007). We also generated a subset of unrelated individuals according to previous protocols (Schweizer *et al*. submitted).

Pedigree

A multi-generation pedigree of Yellowstone wolves was previously generated based on field observations and microsatellite genotyping (vonHoldt *et al.* 2008; 2010b). Within the present study, an accurate pedigree would be useful for determining the founder haplotypes of the K locus, for measuring the change in haplotype frequencies with each generation, and for calculating pedigree-based recombination and mutation rates within the K locus region. We sequenced 203 individuals from this pedigree to maximize the number of trios and duos, as well as other life history information collected previously (vonHoldt *et al.* 2008; Stahler *et al.* 2012). We double-checked familial relationships in two ways: 1) We calculated the pairwise relatedness in `Coancestry` (Wang 2010) using 1000 random LD-pruned SNPs from the neutral regions on the capture array and using the original 26 microsatellite loci data generated by vonHoldt and colleagues (2008; 2010); and 2) We used the `check` mode of `SHAPEITv2` (O'Connell *et al.* 2014) to count the occurrence of Mendelian inheritance errors. Given that some of these "errors" could be real *de novo* mutations, we verified the parentage for trios or duos with Mendelian inheritance error rates higher than 5%.

Phased haplotypes

Given the varying levels of relatedness among individuals sequenced here, we chose to use `SHAPEITv2` (O'Connell *et al.* 2014) to phase haplotypes. We filtered sites for a 100% call rate in all individuals, a DP ≥10 and variants with GQ ≥ 30. The remaining 7,761,114 sites were converted to `PLINK` format and phased with `SHAPEIT` by individual chromosomes with the following parameters: `--burn 10 --prune 10 --main 20 --states 200 --window 0.1 --rho 0.001 --effective-size 20000 --duohmm`. The pedigree information within Yellowstone wolves was used by `SHAPEIT` to improve the accuracy of phasing (O'Connell *et al.* 2014). Given that the K locus mutation is a deletion, we coded the indel genotype into the `PLINK` input file by coding individuals with a K^y as matching the reference genome, and individuals with a K^B allele as having an arbitrary 3bp genotype that did not match the reference genome. After phasing, we double-checked that individual indel genotypes were not affected. We also phased chromosome 16 specifically with a Kenyan golden jackal (*Canis aureus*) sequence (Koepfli *et al*. accepted) for downstream analyses that required the ancestral state of haplotype data.

Summary statistics on phased haplotypes

Using the phased haplotype data, we calculated a suite of summary statistics to investigate patterns of polymorphism and divergence within the K locus core and surrounding regions, neutral regions, genic regions, parallel telomeric, and non-telomeric regions. We predicted that statistics calculated using the neutral, telomeric, and non-telomeric regions might reflect what is known about population history and genome-wide patterns of variation, while statistics from the K locus regions might demonstrate evidence of selection. These statistics included the following: π, the average pairwise differences, Watterson's theta (θ_w), the average number of segregating sites (1975), haplotype diversity, the proportion of unique haplotypes out of all haplotypes, and Tajima's D, a measure of how well π and θ_w fit the infinite-sites model and an indicator of both selection and population history (Tajima 1989). These statistics were calculated for combinations of each sequence type, population, and K^B vs. K^y haplotype. We also calculated these statistics in sliding windows of 10 Kb with 1 Kb overlap. All calculations were performed within the `Python EggLib` package (De Mita & Siol 2012) using custom scripts.

Functional annotation

To investigate whether nearby mutations on the K^B haplotype have also been swept to high frequency that may be causing fitness differences between black heterozygote and black homozyote individuals, we annotated the functional effect of variants using Ensembl's Variant Effect Predictor and the implementation of `SIFT` (`Sorting Intolerant From Tolerant`) therein. `SIFT` uses sequence alignment conservation across multiple species to identify the potential impact of non-synonymous mutations within coding regions (Kumar *et al.* 2009) as deleterious (score <0.05) or tolerated (score≥0.05), as an indication of potential functional impact. Using these data we also calculated the ratio of non-synonymous to synonymous mutations within genes near the K locus.

Comparison of selection signals in K^B vs K^y

In addition to the summary statistics above, we used two methods to determine whether the selective sweep occurred in dogs or wolves, and whether the sweep occurred on the K^B or K^y containing haplotypes. First, we visualized large-scale patterns of variation in haplotypes by generating "haplotype structure" plots that showed the ancestral and derived state of each polymorphic position. These plots were also useful for visualizing regions where recombination events may have occurred.

Next, we implemented the extended haplotype homozygosity (EHH) test, which measures the relationship between the frequency of an allele of interest and the amount of LD surrounding it (Sabeti *et al.* 2002). Neutrally evolving alleles will take longer to reach high frequency and will have short-range LD because recombination causes decay in the LD, whereas positively selected alleles will rise in frequency quickly and will result in longer LD. Once a core mutation is identified (here, the K^B mutation), increasingly distant SNPs are used to measure the decay of LD from the core haplotype. This measure of decay is called the EHH and provides the probability that two randomly chosen chromosomes out of a population are identical between the core haplotype and the increasingly distant SNP.

Comparison of diversity among populations containing the K^B allele

Using geographic patterns of nucleotide and haplotype diversity, and decay of LD among populations, we sought to infer the geographic origins of the K^B allele in wolves (i.e. where and in which population the introgression may have occurred). The regions where introgression occurred will be identified based on the assumption that they have highest nucleotide and haplotype diversity, and the highest proportion of ancestral haplotypes (e.g. Tishkoff 2001; Gray *et al.* 2010). These origin populations should have had more time for both new mutations and recombination events to occur, leading to higher nucleotide diversity and a more rapid decay of LD. Conversely, non-origin populations should exhibit lower nucleotide diversity, slower decay of LD, and have haplotypes that are a subset of those in the source populations. Using the `EggLib` package, we calculated the LD decay by measuring the square Pearson’s correlation coefficient (r^2) between each polymorphic SNP, after filtering for a minimum allele frequency of 5%. Genomic distances were binned by 30 Kb for telomeric regions, and by exponentially increasing distances for the surrounding 5 Mb region, and the decay of r^2 was plotted in R 3.1.3 (http://www.R-project.org). Given that sample size can affect LD decay, we performed calculations using the entire set of samples for each population, plus a subset of up to eight haplotypes (or fewer if there were not eight). The K locus core and parallel telomeric regions were calculated separately, with the latter as a control for the effect of proximity to telomeres on the decay of LD.

Patterns of genetic divergence among populations

In order to uncover the introgression history at the K locus, we used neighbor-joining methods to infer the hierarchy between haplotypes from different geographic localities. Patterns using K locus data were compared to those based on the neutral data since the former should provide insight into the specific history at the selected K locus, and the latter should reflect patterns previously identified by population genetics using genome-wide data sets (vonHoldt *et al.* 2010a; 2011; Pilot *et al.* 2013; Schweizer *et al.* submitted). We first calculated the pairwise number of differences between each haplotype for the K locus core data set and for the neutral data set. We next constructed neighbor joining trees using the package ape 3.1-2 in R (Mech & Boitani 2003; Paradis *et al.* 2004; Musiani *et al.* 2007). Trees were generated with 1000 bootstraps, and then visualized within the ape package. Given that recombination events occurring within the 200 Kb region may have confounded the signal of ancestry, we calculated trees using surrounding 4 Kb, 10 Kb, 60 Kb, 100 Kb, and 200 Kb intervals around the 3bp indel.

Results

Success of targeted enrichment & sequencing summary

Overall, target enrichment and sequencing were both highly successful (Table 3-2, Figure 3-2). A total of 8.397 billion reads were generated across 20 lanes of sequencing on the HiSeq 2000, with an average of 82.34% ± 18.94% of reads per individual passing Illumina quality filters and uniquely mapping to the dog reference (canFam3.1) (Table 3-2). All individuals sequenced had approximately 60% or more of the target regions with at least 25X coverage, although there were many samples that had higher coverage, with mean sample coverage of 150X±61X (Figure 3-2). After genotype filtering and quality control, there remained 10,922,248 positions with 162,815 variants and 12,774 indels. A total of 24 wolves were removed from further analysis based on low sequencing coverage, low call rate, or discordant phenotype and genotype at the K locus (see below).

Quality control: Concordance with CanMap Samples

One hundred nine individuals sequenced for this experiment were previously genotyped on the Affymetrix Dog SNP array (vonHoldt *et al.* 2010a; 2011; Schweizer, *et al*. submitted), and 204 Affymetrix Dog SNP positions overlap with the targeted sequencing done for this study. Using these data, we calculated that genotype concordance was above 99.4% for all called genotypes (Table 3-3).

Quality control: Concordance with Known Coat Color and K locus Genotypes

Two-hundred fifteen wolves from Yellowstone NP and Denali NP and 20 dogs were previously Sanger-sequenced or HRM genotyped for the K locus indel (Candille *et al.* 2007;

Anderson *et al.* 2009; unpublished data). These individuals, plus an additional 43 wolves, were also of known coat color, from field observations or photographic evidence (Musiani *et al*. 2007; Stahler *et al.* 2012; Denali National Park Service). The K locus genotype based on the capture array data was concordant with HRM or Sanger data in 100% of individuals and all samples were concordant with known coat color except for two gray colored wolves that carried the K^B deletion. Given that wolf coat color can gray with age (Anderson *et al.* 2009), we did not remove these two samples. Only samples that met all of the applicable concordance checks and genotype quality filters were used for further analysis for a total of 378 samples (Table 3-1).

Pedigree

We confirmed parentage and familial relationships among 203 Yellowstone wolves, and after correcting a subset of the relationships, we identified 88 full trios and 69 paired duos (Figure 3-3). These relationships were further confirmed with data from resequencing SNPs and existing microsatellites, and with recorded behavioral observations made by the Wolf Project scientists at Yellowstone National Park (D. Stahler, E. Stahler, pers. comm.). This pedigree was used to phase haplotypes among all 378 samples (Table 3-4), to confirm the coat color and genotype of founder individuals within Yellowstone, and to observe which founder individuals (and their haplotype at the K locus) contributed to subsequent generations within Yellowstone.

Summary statistics for quality control

To further assess sequencing quality and to summarize genetic patterns, we split our sequence data into six types of genomic categories: 1) neutral; 2) the K locus core (200 Kb region); 3) the K locus core plus surrounding (the 200 Kb region plus intervals up to 5 Mb surrounding); 4) genic; 5) telomeric; and 6) non-telomeric regions (Table 3-4). We also measured variation within MHC regions sequenced on the array as an additional example of regions with high levels of variability (Bernatchez & Landry 2003). The transition/transversion ratio across all sites was 2.01 (Figure 3-4), and values within each category ranged from 1.17 in MHC to 3.65 in exonic regions. These values were consistent with expectations based on past research in humans, dogs, and wolves (DePristo *et al.* 2011; Freedman *et al.* 2014; Zhang *et al.* 2014; Schweizer *et al*. submitted). Additionally, heterozygosity values measured across the same regions (Figure 3-5) were in line with previous calculations based on SNP array data and expectations based on population histories (Gray *et al.* 2009; vonHoldt *et al.* 2010a; 2011) and references therein). For example, Mexican wolves are a distinct lineage that has undergone a recent bottleneck and as a result is largely inbred. Consistent with this history, the heterozygosity (H_o) values across all genomic categories were low ($H_{o,\ neutral}$: 0.00084). Similarly low values were observed in Indian ($H_{o,\ neutral}$: 0.00074) and Italian wolves ($H_{o,\ neutral}$: 0.00072), both of which have undergone recent bottlenecks (Lucchini *et al.* 2004; vonHoldt *et al.* 2011). Surprisingly, wolves from Sasketchawan ($H_{o,\ neutral}$: 0.00068) and coastal British Columbia ($H_{o,\ neutral}$: 0.00068) had even lower heterozygosity although a recent demographic decline has not been noted. Other North American wolf populations from Yukon ($H_{o,\ neutral}$: 0.00088), Newfoundland ($H_{o,\ neutral}$: 0.00088), Northwest Territories ($H_{o,\ neutral}$: 0.00090), Nunavat, ($H_{o,\ neutral}$: 0.00092) and Alaska ($H_{o,\ neutral}$: 0.00127) had higher heterozygosity values, which is consistent with them deriving from a large population (Mech & Boitani 2003). Additionally, comparisons of heterozygosity values by genomic categories were consistent with expectations based on molecular biology. For example, exonic regions are under stronger pressure from negative selection than genic regions and therefore have less heterozygosity (mean $H_{o,\ exonic}$: 0.00045; mean $H_{o,\ genic}$: 0.00125). Furthermore, heterozygosity was higher in telomeric regions than in non-telomeric regions (mean $H_{o,\text{ telomeric}}$: 0.00125; mean $H_{o,\text{ non-telomeric}}$: 0.00071), a finding that is consistent with telomeric regions in dogs having higher recombination rates and higher diversity (Auton *et al.* 2013). Finally, the heterozygosity within MHC regions was the highest of any region for all samples other than Italian and Indian wolves (mean $H_{o,\text{ MHC}}$: 0.00353), which might be expected given that high heterozygosity at MHC confers a selective advantage over non-heterozygotes across a wide variety of species (reviewed in Bernatchez & Landry 2003).

CBD103 diversity relative to other genes in dogs and wolves

By comparing genic coding and neutral region diversity between dogs and wolves, we found that *CBD103* was consistently more diverse in dogs than in wolves (Figure 3-6, 3-7, 3-8, 3-9). Measured in terms of haplotype diversity, *H*, *CBD103* showed no haplotype diversity in wolves (*H: 0)*, and higher diversity in dogs (*H*: 0.095), although in both groups the diversity of *CBD103* ranked on the low end of all genes (Figure 3-6). Haplotype diversity for neutral regions was equal to one for both dogs and wolves. In measures of nucleotide diversity, π, wolves showed no diversity in *CBD103*, while dogs had some diversity (π_{wolves}: 0; π_{dogs}: 0.00047; Figure 3-7). In dogs, the neutral region diversity was lower than that of *CBD103* in dogs ($\pi_{\text{neutral,dogs}}$: 0.00030) and in wolves the neutral region diversity was higher than that of *CBD103* in wolves ($\pi_{\text{neutral,wolves}}$: 0.0001). Values of Watterson's Theta, θ_w, showed a higher number of segregating sites within *CBD103* in dogs (θ_w: 0.00138) than in wolves (θ_w: 0) (Figure 3-8). Here, the values of theta of the neutral regions were very similar within dogs (θ_w: 0.000522) and wolves (θ_w: 0.000526), although this value was higher than *CBD103* in wolves, and lower than *CBD103* in dogs (Figure 3-8). A final comparison of measures of diversity, Tajima's *D*, was not applicable in wolves due to a lack diversity, but in dogs Tajima's *D* (*D:* -1.163) was below the value for neutral regions (*D*: -1.737) (Figure 3-9), suggesting an excess of rare alleles at low frequency.

Functional annotation

Of the 1040 genes sequenced on the array, six were annotated within the 200 Kb core region, all of which were canine beta-defensins (*CBD1*, *CBD102*, *CBD103*, *CBD105, SPAG11E*, *SPAG11B*). Within these six genes, VEP annotated 13 synonymous and 30 missense variants. Of the latter, 10 were deleterious and 20 were tolerated. The ratio of non-synonymous to synonymous mutations was above one for *CBD102* (d_N/d_S=3), *SPAG11E* (d_N/d_S =2.5), *SPAG11B* (only 2 non-synonymous), and *CBD105* (d_N/d_S =3). Interestingly, *CBD103* had a dn/ds of one suggesting weak or absent positive selection, although the sample sizes were too small to establish statistical significance. Deleterious mutations (`SIFT` score <0.05) occurred in *CBD1* (one), *SPAG11E* (one), *SPAG11B* (two), and *CBD105* (six).

Haplotype Diversity of K^B vs Ky in dogs vs wolves

As a first visualization of haplotype structure within the 200 Kb core region, we generated haplotype structure plots, in which each line represents a single haplotype, with ancestral and derived alleles colored differently (Figure 3-10). Within our samples, the K locus haplotypes containing K^B (above the light blue line in Figure 3-10) showed much less variability than haplotypes containing K^y (below the light blue line in Figure 3-10) across all wolf populations, and variability was diminished in K^B wolves relative to K^B dogs (Figure 3-10). Yellowstone wolves seemed to be largely composed of a single K^B haplotype, with very little variability. Italian wolf K^B haplotypes visually appeared much more similar to the dogs than to other wolf haplotypes, which might be expected given the observed recent admixture with dogs in that population (Verardi *et al.* 2006).

Next, we obtained four measurements of diversity (π, θ_w, H, and Tajima's D) to explore patterns among the five region types (K locus core, K locus core + 5 Mb, neutral, non-telomeric, and telomeric). For the K locus core and K locus core + 5 Mb regions, measurements of haplotype diversity were consistently lower in wolves than in dogs, and lower in K^B-containing haplotypes than in K^y-containing haplotypes (Table 3-5; Figure 3-11). Similarly, Tajima's D was only negative within wolf K^B haplotypes ($D_{\text{wolf,KB}}$: -2.19), which is an indication of an excess of rare alleles at low frequency, and suggests either a selective sweep or recent population expansion. For the other three region types, whose variability should be more indicative of population demographic history, the wolves were the same or had higher diversity than dogs (e.g. Neutral $\pi_{\text{dog,KB}}$: 0.00109, Neutral $\pi_{\text{dog,Ky}}$: 0.00113, Neutral $\pi_{\text{wolf,KB}}$: 0.00138, Neutral $\pi_{\text{dog,KB}}$: 0.00150; Table 3-5; Figure 3-11), suggesting similar demographic history. Within the neutral regions, for example, wolves had higher values of π, θ_w, and D (Figure 3-11). Telomeric regions had consistently higher values of π and θ_w, most likely because recombination rate increases towards telomeres and increases diversity in those regions as a result (Auton *et al.* 2013). In dog and wolf K^y haplotypes, diversity of the K locus regions was consistently higher than that of the telomeric regions (e.g. K locus core $\pi_{\text{wolf,Ky}}$: 0.00227; Telomeric $\pi_{\text{wolf,Ky}}$: 0.00173; K locus core $\pi_{\text{dog,Ky}}$: 0.00202; Telomeric $\pi_{\text{dog,Ky}}$: 0.00142; Table 3-5), whereas the K^B haplotypes were consistently lower than the telomeric regions, although this pattern was more pronounced in wolves than in dogs (e.g. K locus core $\pi_{\text{wolf,KB}}$: 0.00038; K locus core $\pi_{\text{dog,KB}}$: 0.00132; Table 3-5; Figure 3-11). These results taken together suggest that the low levels of diversity found within K^B haplotypes, especially in wolves, are not typical with regards to telomeric regions in general.

Three of the summary statistics (π, θ_w, and D) were calculated in sliding windows along each of four region types (K locus core, K locus core plus surrounding 5 Mb, telomeric, and non-telomeric) so that the genomic position underlying differences between wolves and dogs could be identified. Within the K locus region, both dog and wolf K^B haplotypes had lower nucleotide diversity (Figure 3-12), and, interestingly, both wolf and dog K^B haplotypes dipped to near zero values near the 3bp deletion ($\pi_{\text{wolf, KB}}$: 0.000049, $\pi_{\text{dog, KB}}$: 0.00017; Figure 3-12A,B). However, the extent of decreased diversity was across a much wider region in the wolf K^B haplotypes than in the K^B dogs. At the core allele, values of θ_w were similarly low for K^B haplotypes ($\theta_{\text{w wolf, KB}}$: 0.00046, $\theta_{\text{w dog, KB}}$: 0.00038; Figure 3-13A,B), and *D* was below negative two in K^B haplotypes ($D_{\text{wolf, KB}}$: -2.140, $D_{\text{dog, KB}}$: -2.701; Figure 3-14A,B). In wolf K^B haplotypes, D <-2 for 76 Kb of sequence (out of 100 Kb total sequenced within the 200 Kb core region). As in both π and θ_w, dog K^B haplotypes also dipped in diversity, but over a much narrower region than K^B wolves (Figure 3-13A,B). In contrast, diversity patterns across the telomeric and non-telomeric regions were largely consistent between K^B and K^y individuals, and between dogs and wolves (Figure 3-12C,D, Figure 3-13C,D, Figure 3-14C,D). Telomeric regions had slightly higher diversity than non-telomeric regions (mean $\pi_{\text{telomeric}}$: 0.00145±0.00037; (mean $\pi_{\text{non-telomeric}}$: 0.00113±0.00039; (mean $\theta_{\text{w,telomeric}}$: 0.00138±0.00033; (mean $\theta_{\text{w,non-telomeric}}$: 0.00104±0.00030), which is consistent with a higher recombination rate in the former. Again, diversity of the parallel telomeric regions, as measured with π and θ_w, was lower than for the K locus 200 Kb region ((mean π_{Klocus}: 0.001514±0.00091; (mean $\theta_{\text{w, Klocus}}$: 0.00152±0.00053). Across all telomeric and non-telomeric regions, Tajima's *D* was never lower than ~ -0.6 or ~ -1.2, respectively, in either dog or wolf K^B individuals (Figure 14C,D), whereas across the K locus 200 Kb region, the minimum Tajima's *D* fell below ~ -2.7 in wolf K^B haplotypes (Figure 14A,B). These results further suggest that the patterns of diversity measured at the K locus, especially with regards to the K^B haplotype, are abnormal, even in comparison to similarly positioned regions along similarly sized chromosomes.

The lower values of Tajima's *D* within the K locus region, in contrast to the telomeric and non-telomeric regions, are a measure of support for an excess of low frequency alleles that are consistent with a selective sweep within the K locus region.

Using the extended haplotype homozygosity (EHH) score, we found that the haplotype containing the derived K^B allele (blue in Figure 3-15A) had more extensive homozygosity in wolves than in dog, as shown by the slower decay of homozygosity in wolves. This pattern was visible in wolves up to 4 Mb upstream of the derived K^B allele (Figure 3-16A), whereas in dogs the homozygosity reached zero much closer to the mutation at around 185 Kb upstream of the derived K^B allele (Figure 3-16A). Similarly, haplotype bifurcation plots for both region sizes showed that most wolves have a single, common derived haplotype (the thick blue line in the right side of Figure 3-15B and Figure 3-16B). Haplotype bifurcation within dogs showed that all K^B haplotypes have a single, narrow core region around the 3bp deletion, but haplotypes quickly accumulate variability in close proximity to the mutation (multiple branching haplotypes near the vertical dashed blue line in Figure 3-15B and Figure 3-16B). This suggests a recent and dramatic selective sweep across wolf populations. For the ancestral K^y haplotype, haplotypes accumulated variants close to the core variant (dashed vertical blue line in Figure 3-15C and Figure 3-16C) and began branching quickly outwards along the chromosome in both dogs and wolves. Given that these patterns might be driven by uneven sampling among wolf populations (Table 3-1), especially those of the K^B haplotype where our sampling is dominated by Yellowstone wolves, we also explored patterns of EHH within different wolf populations.

Comparison of diversity among populations containing the K^B allele

Given that Alberta and Newfoundland each only had a single K^B haplotype, we focused on relative diversity among Alaska, the Northwest Territories, Yellowstone, and Yukon, and found that haplotypes from Yukon had the greatest diversity (π: 0.001196, *H*: 1, and θ_w: 0.001196), with the Northwest Territories slightly less diverse (π: 0.00094, *H*: 0.916, and θ_w: 0.00066) (Figure 3-17). Yellowstone wolves showed the lowest diversity (π: 0.00011, *H*: 0.749, and θ_w: 0.00032), despite including the largest number of sampled K^B haplotypes (n=97; Table 3-1). Importantly, the Yellowstone population is otherwise very diverse (Figure 3-5; mean π_{neutral}: 0.00135, mean H_{neutral}: 1, and mean $\theta_{\text{w, neutral}}$: 0.00083), having been founded from three distinct populations from Montana, Alberta and British Columbia. Even across the entire 5 Mb region, there were multiple individuals within YNP that had identical K^B haplotypes as indicated by $H = 0.958$ (Figure 3-17). Values of Tajima's *D* measured in Yellowstone wolves within the 200 Kb core region were below negative two for K^B haplotypes (*D*: -2.17) but not for K^y haplotypes (*D*: 1.29), which suggests an ongoing selective sweep in Yellowstone (Figure 3-17). Alaska K^B haplotypes also had a slightly negative Tajima's *D* (*D*: -0.537), whereas those in the Northwest Territories were above 2 (*D*: 2.23). Although the sample size was too low within Yukon wolves to calculate *D*, the combination of π, *H*, and θ_w statistics point to the highest K^B diversity in Northwest Territories or Yukon, with lowest diversity in Alaska and Yellowstone. The entire 5 Mb region had low diversity in Yellowstone wolves, with the most extreme reduction occurring within the 200 Kb region, as evidenced by low π, *H*, θ_w, and *D* values. Within the K^y haplotypes, the difference between the K locus core and the K locus core plus surrounding 5Mb (Δ) was much less drastic than in the K^B haplotypes (K^B $\Delta\theta_w$: 0.00091, K^y $\Delta\theta_w$: -0.00010; K^B ΔH: 0.082, K^y ΔH: -0.034; K^B $\Delta\pi$: 0.0010, K^y $\Delta\pi$: -0.0001) (Figure 3-17), and, for those populations that also had K^B haplotypes, Yellowstone 200 Kb K^y haplotypes had relatively lower diversity ($\pi_{\text{Yellowstone}}$: 0.00188, $H_{\text{Yellowstone}}$: 0.94, $\theta_{\text{w,Yellowstone}}$: 0.00132) than the 200 Kb K^y haplotypes from the other three populations (Alaska, Northwest Territories, Yukon), which were all more similar (π_{Alaska}: 0.00236, H_{Alaska}: 0.99, $\theta_{\text{w, Alaska}}$: 0.00188; π_{NWT}: 0.00221, H_{NWT}: 0.99, $\theta_{\text{w, NWT}}$: 0.00197; π_{Yukon}: 0.00167, H_{Yukon}: 0.92, $\theta_{\text{w, Yukon}}$: 0.00161).

Levels of LD were much higher within K^B haplotypes than K^y haplotypes for the K locus region (Figure 3-18), and the K locus region had overall higher levels of LD than those of the 10 parallel telomeric regions (Figure 3-19). Within the K^y haplotypes (Figure 3-18A), most geographic populations had LD below r^2=0.2 for the 5 Mb region at distance ranging from 20 Kb to 1 Mb from the 3 bp deletion (Northwest Territories, n=8, $r^2_{0.2}$≈20 Kb; Russia, n=8, $r^2_{0.2}$=20 Kb; Alberta, n=8, $r^2_{0.2}$≈20-30 Kb; Dogs, n=8, $r^2_{0.2}$≈ 40-50 Kb; Yukon, n=8, $r^2_{0.2}$≈100 Kb; Alaska, n=8, $r^2_{0.2}$≈125 Kb; Saskatchewan, n=8, $r^2_{0.2}$≈125 Kb; Nunavut, n=8, $r^2_{0.2}$≈200 Kb; Quebec, n=8, $r^2_{0.2}$≈800 Kb; Italy, n=8, $r^2_{0.2}$≈1 Mb). The exception is for Ukraine (n=4), Newfoundland (n=5), British Columbia (n=2), and Yellowstone (n=8), which may be due either to sample size or population history. Of note is that the Northwest Territories r^2≤0.2 with distances greater than ~20 Kb, suggesting that it was one of the most diverse populations that also contained K^B haplotypes (Figure 3-18A). For K^B haplotypes (Figure 3-18B), only dogs, and wolves from the Northwest Territories and Alaska ever reached r^2=0.2 (Northwest Territories, n=8, $r^2_{0.2}$≈100 Kb; Dogs, n=8, $r^2_{0.2}$≈150; Alaska, n=6, $r^2_{0.2}$≈4 Mb), whereas samples from Yellowstone (n=8), the Yukon (n=3), and Italy (n=2) did not. The two K^B haplotypes from Italian wolves were in complete linkage for the entire 5 Mb region, which may reflect an extremely recent introgression event. In contrast, LD decayed rapidly over the ~200 Kb parallel telomeric regions, with LD being only slightly higher in dogs than in wolves (dogs, n=16, $r^2_{30\text{ Kb}}$: 0.2-0.48; wolves, n= 118,

$r^2_{30\ \mathrm{Kb}}$: 0.07-0.23) (Figure 3-19). In general, smaller chromosomes had lower LD than larger chromosomes (Figure 3-19).

Using the EHH statistic, we explored the decay of homozygosity within each geographic population containing K^B wolves (Figure 3-20), and found that Alaska and Yellowstone had the most extensive EHH in the derived K^B haplotypes (blue lines in Figure 3-20A,C). In contrast, Northwest Territories (Figure 3-20B) and Yukon (Figure 3-20D) populations had EHH scores that dropped off more quickly. Interestingly, the Northwest Territories had an uneven degree of EHH decrease on either side of the deletion, with homozygosity extending more towards the 3' end of the region (Figure 3-20B), which may reflect fewer recombination events on that side of the K locus. By comparing the extent of EHH in the surrounding 5 Mb region, we found that most of the extended homozygosity we reported above in all K^B wolves (Figure 3-16A, right panel) was driven exclusively by the Yellowstone wolves since they were the only population that had that pattern once geographic localities were analyzed separately (Figure 3-20, right panel). Patterns of haplotype bifurcation within the 200 Kb core region among the four geographic localities showed that the Yellowstone wolf population seemed to be dominated by a single haplotype, with few variants and a block ~15 Kb upstream of the deletion showing no variation among samples (Figure 3-21C). This pattern was not similarly reflected within the K^y haplotypes in Yellowstone, in which variation was much more evenly spread among samples and within close proximity to the core allele. Both Alaska and the Northwest Territories showed regions surrounding the deletion with no variability among samples (Figure 3-21A,B), although the Northwest Territories displayed that pattern mostly downstream of the deletion (Figure 3-21B). Yukon haplotypes bifurcated very close to the deletion (Figure 3-21D).

To explore whether patterns of EHH within the reintroduced Yellowstone population were a result of limited diversity within the founders or a subsequent sweep, we measured EHH solely within 12 founders, of which five were K^B/K^y and seven were K^y/K^y. Decay of EHH surrounding the core mutation occurred more rapidly in Yellowstone founders (Figure 3-22A) than in the Yellowstone population as a whole (Figure 3-20C). In the derived K^B haplotypes (blue in Figure 3-22A), EHH was equal to one up to ~45 Kb upstream, but decreased to 0.6 immediately downstream of the mutation. This is in contrast to the current Yellowstone population (Figure 3-20C), in which EHH was ≥0.9 up to ~80 Kb upstream, and was ≥0.8 even 90 Kb downstream of the mutation. Over the 5 Mb region, EHH decayed to zero in K^B haplotypes ~ 1 Mb upstream (Figure 3-22A), whereas in the current Yellowstone population remained above zero even ~4 Mb upstream (Figure 3-20C). The most common K^B haplotype among the founders (Figure 3-22B) is similar in bifurcation to the most common K^B haplotype in the recent population (Figure 3-21C). Given that the Yellowstone population is not inbred (vonHoldt *et al.* 2008; 2010b), this suggests an advantageous haplotype rose in frequency subsequent to the founding generation.

Hierarchical patterns of divergence among populations

Using neighbor joining trees generated with pairwise nucleotide divergence from neutral regions, we found that major groupings were concordant with major geographic or species differences (Figure 3-23). For instance, dogs formed a single grouping near European and Asian wolves, which is concordant with likely origins of domestication (Freedman *et al.* 2014). Similarly, Yellowstone wolves formed a single cluster near wolves within territories from which the founders originated (e.g. Alberta; Bangs & Fritz 1996). Neighbor joining trees generated using pairwise distances calculated from K locus haplotypes were less definitive (Figure 3-24).

All K^B haplotypes were within a single cluster on the tree (blue in Figure 3-24), which was in contrast to the neutral tree, in which individuals with K^B and K^y alleles grouped according to geography rather than K locus allele. By focusing on the K locus section of the tree (Figure 3-25), we found that wolf and dog K^B haplotypes were mostly clustered separately from one another. One exception was a black Labrador K^B haplotype, which was sister to almost all of the K^B wolves (Yellowstone, Northwest Territories, Alberta). This suggests a possible origin of K^B haplotypes in wolves from introgression with black Labradors, or their common ancestor. Additional dog K^B haplotypes were also found clustered with dog K^y haplotypes and admixed Italian K^y haplotypes. Also, a cluster of four Northwest Territories K^B haplotypes grouped sister to a large group of K^y wolf haplotypes. These unusual groupings of K^B with K^y haplotypes may reflect recent recombination events that have placed the K^B allele onto a K^y haplotype.

Discussion

In general, our results support those of Anderson *et al*. (2009), who sequenced intervals within the 180 Kb region surrounding the K locus and demonstrated the following: 1) a 3bp deletion causes dominantly inherited black coat color in wolves; 2) the region surrounding the mutation shows signatures of a selective sweep within K^B haplotypes; and 3) the mutation arose in wolves as a result of introgression with dogs. Our results provide further support of these findings through a wider geographic sampling, more extensive sequencing, and a full characterization of variation in the Yellowstone wolf population. Using multiple summary statistics, LD decay, and EHH, we find clear evidence of a selective sweep within the K^B haplotypes. Diversity differences between wolves and dogs point to a selective sweep within wolves rather than dogs (Figure 3-11, Figure 3-12, Figure 3-15). We found that wolf populations from Yellowstone and Alaska displayed the highest homozygosity surrounding the core deletion in K^B, but not K^y, haplotypes, with homozygosity extending in K^B haplotypes up to 4 Mb downstream and >1 Mb upstream of the deletion (Figure 3-20, Figure 3-21). Populations in the Northwest Territories and Yukon had the highest K^B haplotype diversity, lowest LD, and the shortest extent of homozygosity. Given evidence of a selective sweep, future work will estimate the strength of selection and timing of introgression events in the framework of demographic models (Schweizer *et al*. in prep).

The insight provided from the results of this study would not have been possible without a focused, resequencing effort through the design of a custom capture array and the substantial yield capabilities of next generation sequencing technologies. Here, we were able to sequence over 8 billion reads in 403 wolves and dogs, and to target 7 Mb of specific regions. This approach has still not been well applied in non-model organisms, although the capabilities exist. Capture array resequencing has been applied to a study of pigmentation genes in beach mice (*Peromyscus* sp; Domingues *et al.* 2012) and a study resequencing quantitative trait loci for behavioral traits in wild rats (*Rattus* sp; Albert *et al.* 2011). A novelty within our approach was the use of internal, genomic controls (i.e. the telomeric and non-telomeric regions), which provide an empirical background for expectations of diversity within the same proximate regions as the K locus, but without the likely effects of a selective sweep. Furthermore, the regularly spaced sequence intervals up to 5 Mb around the deletion enabled us to measure impacts of selection without contiguous sequencing (i.e. we sequenced ~800 Kb rather than 5 Mb). Resequencing of functional loci such as the K locus is indispensable to determining the selective forces that influence genomic evolution.

Although the approach might be novel within wolves and other non-model organisms, the observation of adaptive introgression is increasingly common. New statistical tests and complete sequencing data from present and archaic humans have enabled the identification of multiple selective sweeps resulting from introduction of a new variant by introgression (reviewed in Racimo *et al.* 2015). For example, an adaptive allele within the gene *EPAS1*, which confers adaptation to high-altitude hypoxia in Tibetans (Beall *et al.* 2010), shows evidence of having introgressed into Tibetans or East Asians from archaic Denisovans, and subsequently increased in frequency due to selection (Beall *et al.* 2010; Huerta-Sanchez *et al.* 2014). In Tibetans, the mutations within *EPAS1* are in a 32 Kb window at ~80% frequency and are common only with Denisovans (Huerta-Sanchez *et al.* 2014). A second example concerns populations of house mice within Europe, which show evidence of adaptive introgression with wild Algerian mice in which a segment of ~10 Mb of DNA was introgressed from the latter that included adaptive alleles conferring warfarin-resistance (Song *et al.* 2011). Furthermore, this ~10 Mb region demonstrates signatures of selection, with a time of origin dating back to when anticoagulant rodenticides were in use (Song *et al.* 2011). Using simulations, the authors were able to estimate a selection coefficient, s=0.28-0.33.

Often, admixture is viewed as a negative event that threatens local adaptations or leads to outbreeding depression (Whitney *et al.* 2006 and references therein). However, in our study, and those mentioned previously, we demonstrated that admixture could enhance adaptation. The object of selection on the K^B allele may not have been for the melanistic coat color, but rather for the immunological effects. The K locus is a member of the canine β-defensin family of antimicrobial peptides (Pazgier *et al.* 2006) and may be involved in adaptive immune response (Yang *et al.* 1999). In dogs, the K locus demonstrates antimicrobial activity against respiratory pathogens (Erles & Brownlie 2010), and black wolves have higher fitness than gray wolves (Coulson *et al.* 2011). In fact, black wolves of Yellowstone National Pack showed higher survivorship during the three documented distemper outbreaks (Stahler, pers. comm.). Given the dog populations are a reservoir for canine distemper, there is a possibility that dogs provide both the resistance genes and the pathogens to maintain those genes in wolf populations. Supporting this view is the observation that the populations with the fewest black in the high Arctic are not in close proximity to dogs.

Our extensive geographic sampling of wolves from 10 states and provinces within North America suggests that the admixture between dogs and wolves occurred within Northern Canada, most likely in the Northwest Territories. The most ancestral population with regards to K^B introgression is predicted to have had the most time for recombination to break down haplotype blocks and therefore greater diversity. Wolves within the Northwest Territories demonstrated some of the highest levels of diversity within the K^B haplotypes (Figure 3-17), the lowest LD (Figure 3-18), and lowest extent of haplotype homozygosity (Figure 3-20). Wolves from Yukon demonstrated similar patterns of diversity, and so are also candidates for K^B origin in wolves. However, the low number of K^B haplotypes within Yukon prohibited us from fully exploring this idea, especially with regards to LD decay. Samples from Alaska and Yellowstone lack genetic diversity at the K^B haplotype, have high levels of LD within K^B haplotypes, and have long tracts of extended homozygosity, which suggests the K^B allele is more recent in these populations.

Several recent studies exploring geographic origins of Native American dog breeds have suggested that Arctic dog breeds such as the Inuit sled dog, the Canadian Eskimo dog and the Greenland dog have archaic mitochondrial DNA haplotypes and show evidence of ancient admixture with wolves (Brown *et al.* 2013; van Asch *et al.* 2013). Furthermore, dogs within these populations represent the only living dog breeds with mitochondrial haplotypes that are unique to New World wolves. These results suggest that the original introgression event likely occurred in the High Arctic regions of Northern America, where dogs and native people first coexisted.

We also found genetic evidence for a strong ongoing selective sweep in Yellowstone wolves based on summary statistics. The Yellowstone population was originally founded in 1995 with 31 wolves from two discreet populations in Alberta and British Columbia, and 10 additional wolves from Montana added the following year. Given this history, we would expect an amplification of heterozygosity. However, the Yellowstone wolf K^B haplotypes demonstrate very low levels of diversity (Figure 3-10, Figure 3-17), high LD (Figure 3-18B), and an almost completely monomorphic core region (Figure 3-20C). This pattern might reflect lower diversity in the founding stock, but we found that in general the Yellowstone population has average levels of heterozygosity, although not as high as from more Northern populations. Furthermore, extensive studies of the Yellowstone wolf pedigree using microsatellite data demonstrate that wolves avoid inbreeding and currently have an inbreeding coefficient near zero (vonHoldt *et al.* 2008; 2010b). We also show that the limited number of Yellowstone founders that we sequenced had substantially higher diversity than the current population (Figure 3-22), and consequently, diversity within the K locus region has been lost in the current population. Thus, our findings imply differential selection of certain haplotypes within the Yellowstone population, and the increase in LD and homozygosity is a result suggests a strong, ongoing selective sweep. Furthermore, there is ample evidence for differential selection in Yellowstone wolves related to the genotype at the K locus (Coulson *et al.* 2011). Black heterozygote wolves have a higher overall survival rate than black homozygote wolves, which implies that the selective effects of a K^B allele may have to do with disease rather than pigmentation since both genotypes have the same coat color (Coulson *et al.* 2011). Yellowstone may be unique among North American wolf populations in that it surrounded by ranch land with free ranging dogs. As mentioned above, these dogs are potentially a reservoir for canine disease such as distemper which causes substantial mortality in the Yellowstone wolf population (Stahler *et al.* 2012; Stahler pers. comm.) and could be the cause underlying the ongoing selective sweep unique to Yellowstone wolves. Elsewhere, as represented by our wolf populations, the density of dogs is lower and the encounter probabilities less given the absence of cattle ranching and range lands.

The results presented here represent an initial exploration of patterns of nucleotide and haplotype diversity within the K locus. The sequencing data can be used to infer specific evolutionary parameters such as the effective population size of each population and locus-specific estimates of recombination rate, both of which will be necessary for an ecotype-specific estimation of the strength of selection on the K locus and timing of introgression events. Current analyses are underway to further explore patterns of selection at the K locus.

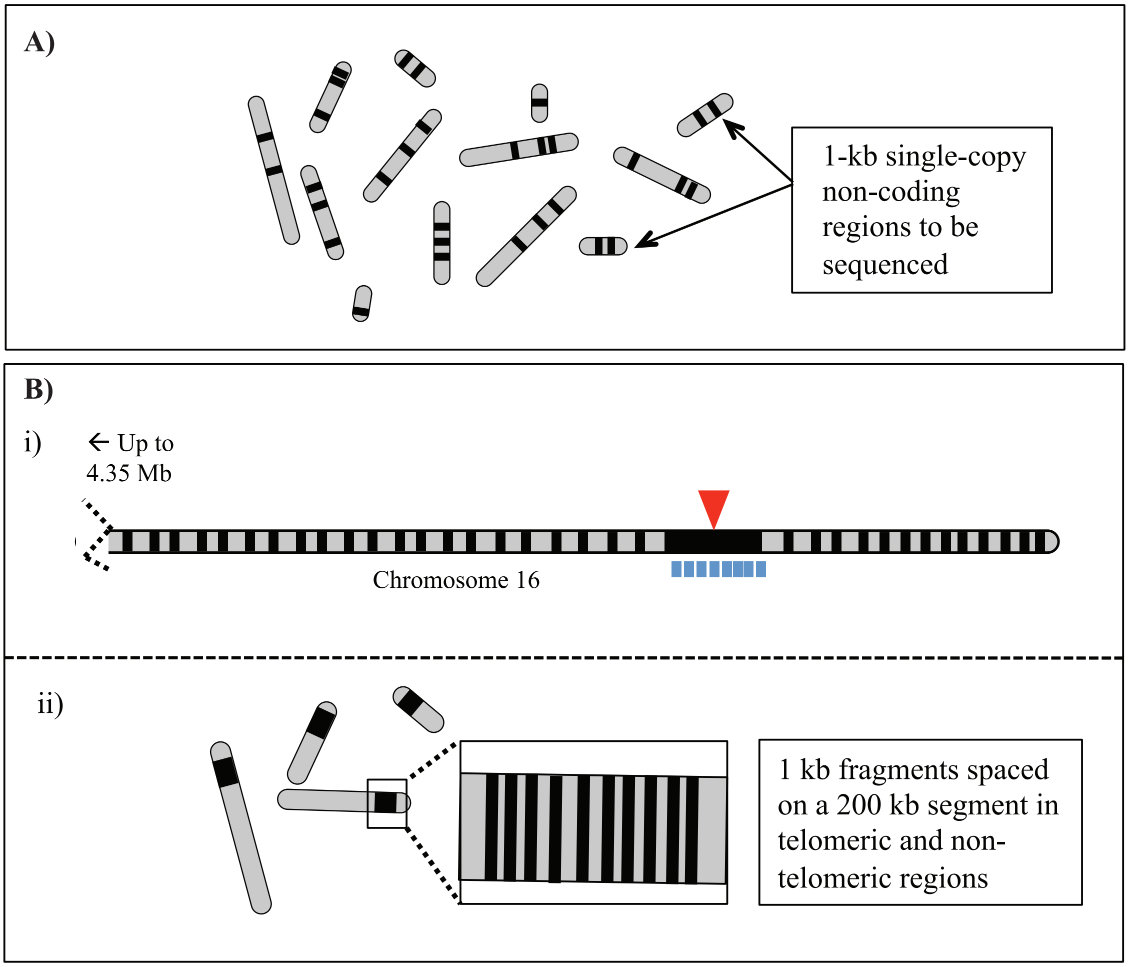

A)
1-kb single-copy non-coding regions to be sequenced
B)
i)
← Up to 4.35 Mb
Chromosome 16
ii)
1 kb fragments spaced on a 200 kb segment in telomeric and non-telomeric regions

Figure 3-1. Schematic of re-sequencing strategy and design for the capture array. A) Aim 1 includes capture of 5000 putatively neutral 1-Kb fragments spaced throughout the genome. B) Aim 2 includes capture of the K locus indel mutation (red triangle) and surrounding region. i) A 200 Kb core region (black bar below K^B indel) and 1-Kb fragments spaced every 10 Kb to capture variation up to 5 Mb surrounding the K locus, ii) 1kb segments sampling a similar core as in (i) and located in 20 telomeric and non-telomeric regions. Chromosome position on scale bar below. Region sequenced by Anderson *et al.* (2009) represented by blue fragments and arrows.

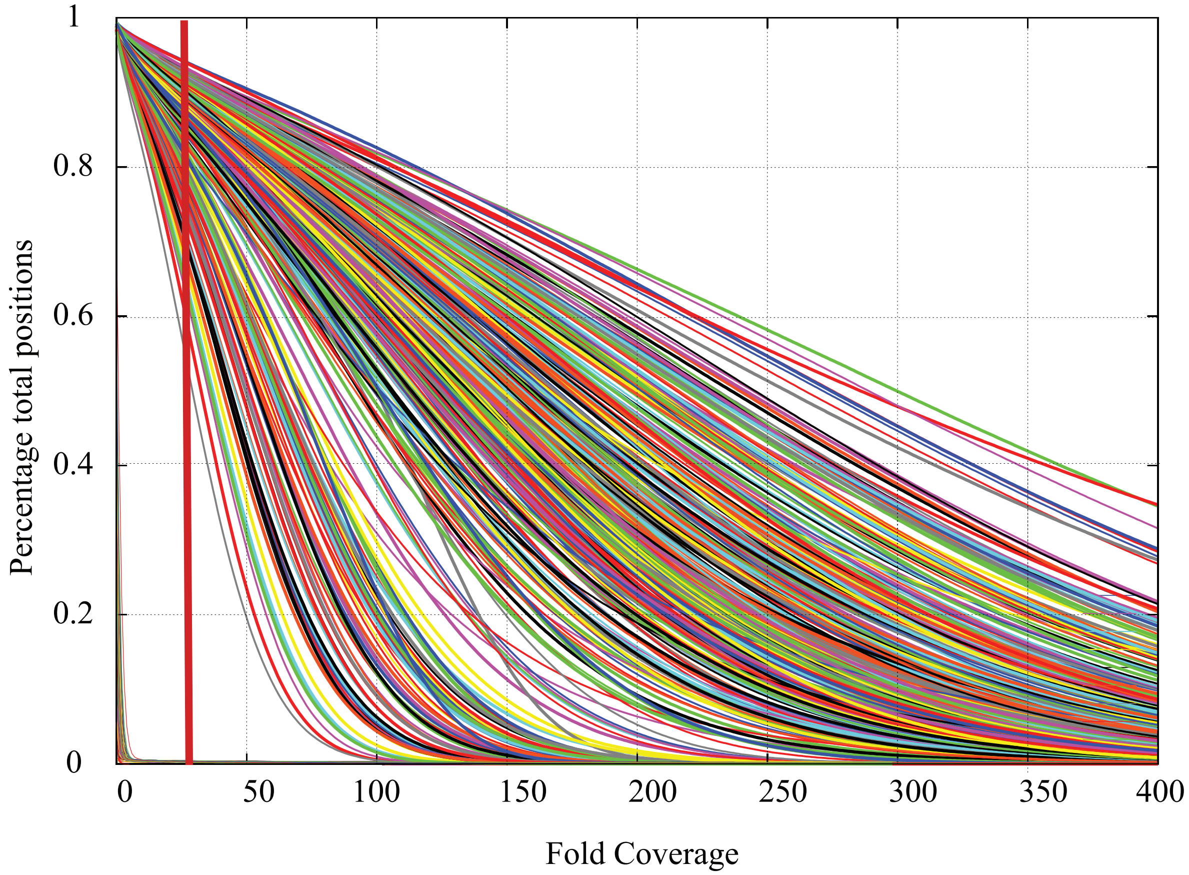


Figure 3-2. Fold-coverage and enrichment for 403 sequenced wolves and dogs. The percentage of total genomic positions covered at increasing coverage per base pair (fold-coverage) is plotted. Fold coverage over the capture array target regions (bold lines) and the entire reference genome (thin lines in bottom left corner) are shown. The low coverage of the genome but high coverage of the target regions demonstrates a successful capture and enrichment. The red vertical line indicates 25X fold coverage.

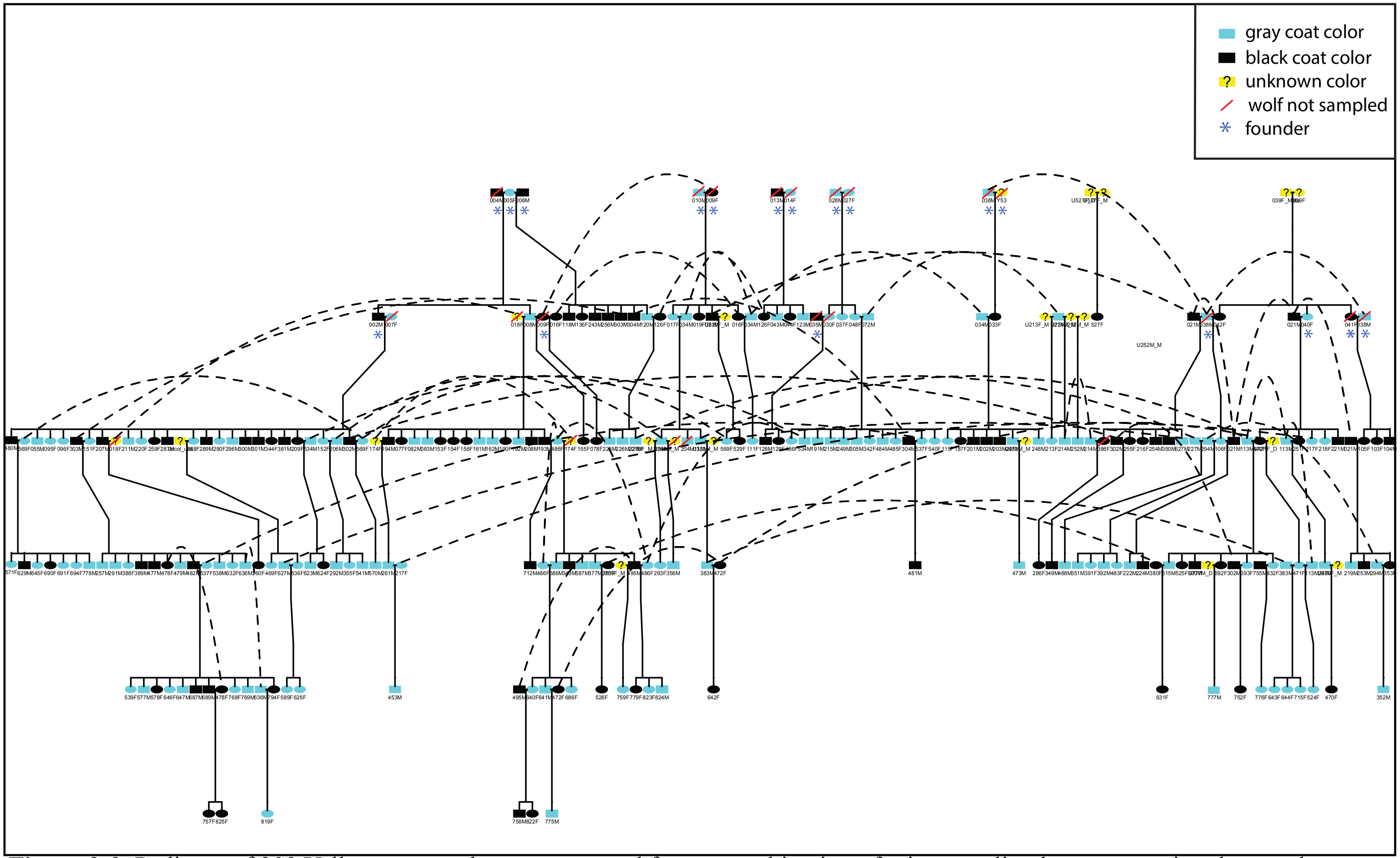


Figure 3-3. Pedigree of 203 Yellowstone wolves, constructed from a combination of microsattelite data, sequencing data, and field-based observations. Individuals with unknown coat color are yellow. Founder individuals (indicated with *) were used to explore changes in K locus diversity since foundation of the Yellowstone wolf population. Solid lines between generations represent parentage, while dashed lines represent the same individual placed elsewhere in the pedigree. Microsattelite data and field observations are from vonHoldt *et al* (2008; 2010).

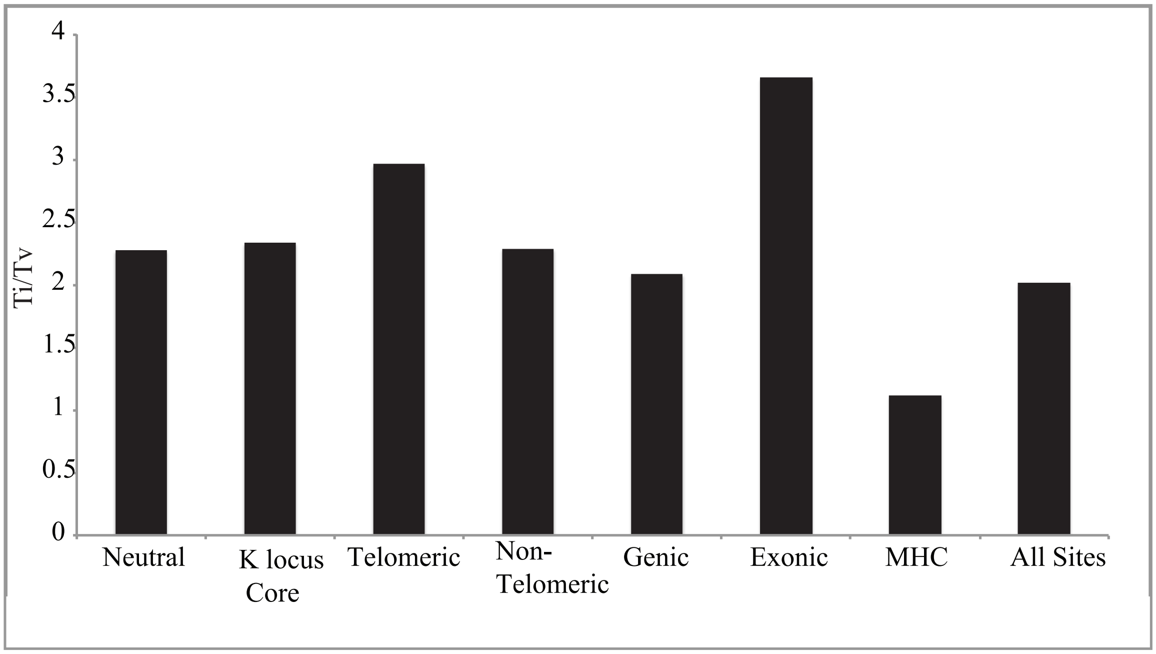


Figure 3-4. Transition/Transversion (Ti/Tv) ratios for multiple genomic regions sequenced on the capture array.

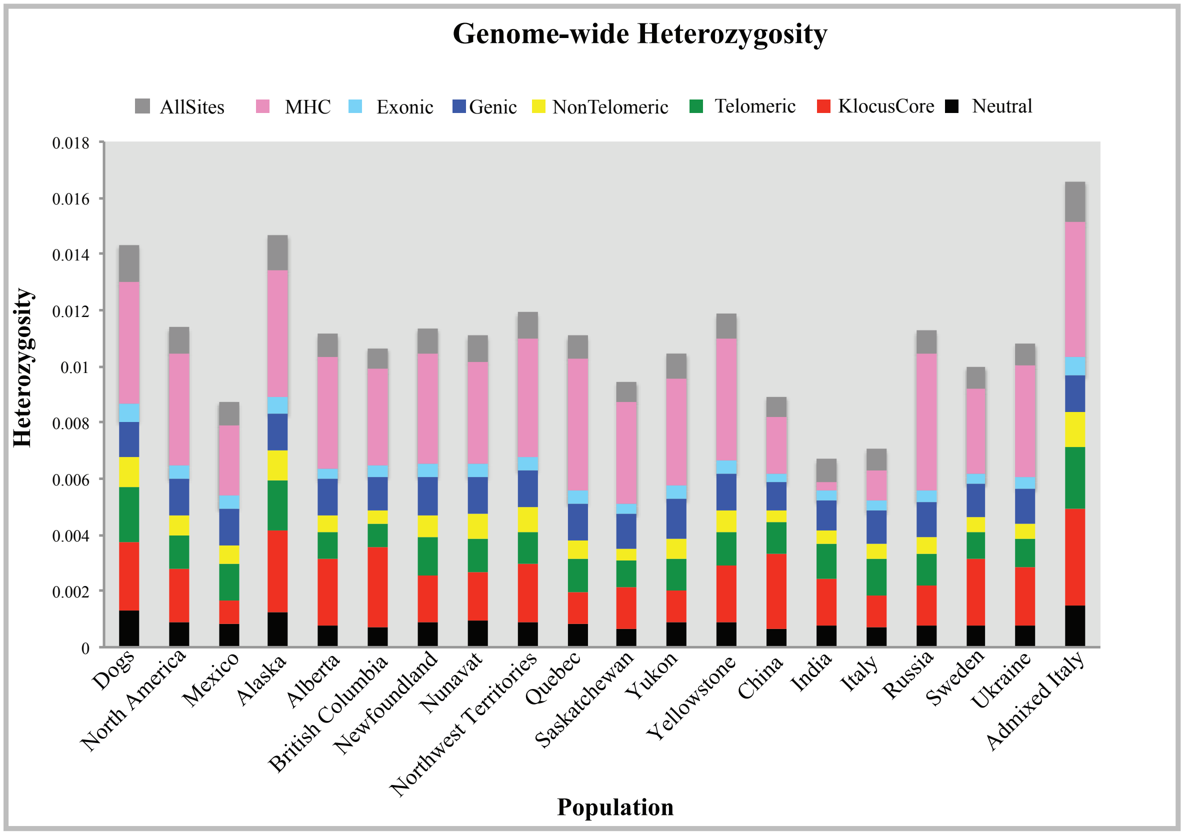


Figure 3-5. Genome-wide heterozygosity for filtered sites measured over eight types of genomic regions. All populations are gray wolves unless otherwise noted (i.e. "Dogs" and "Admixed Italy").

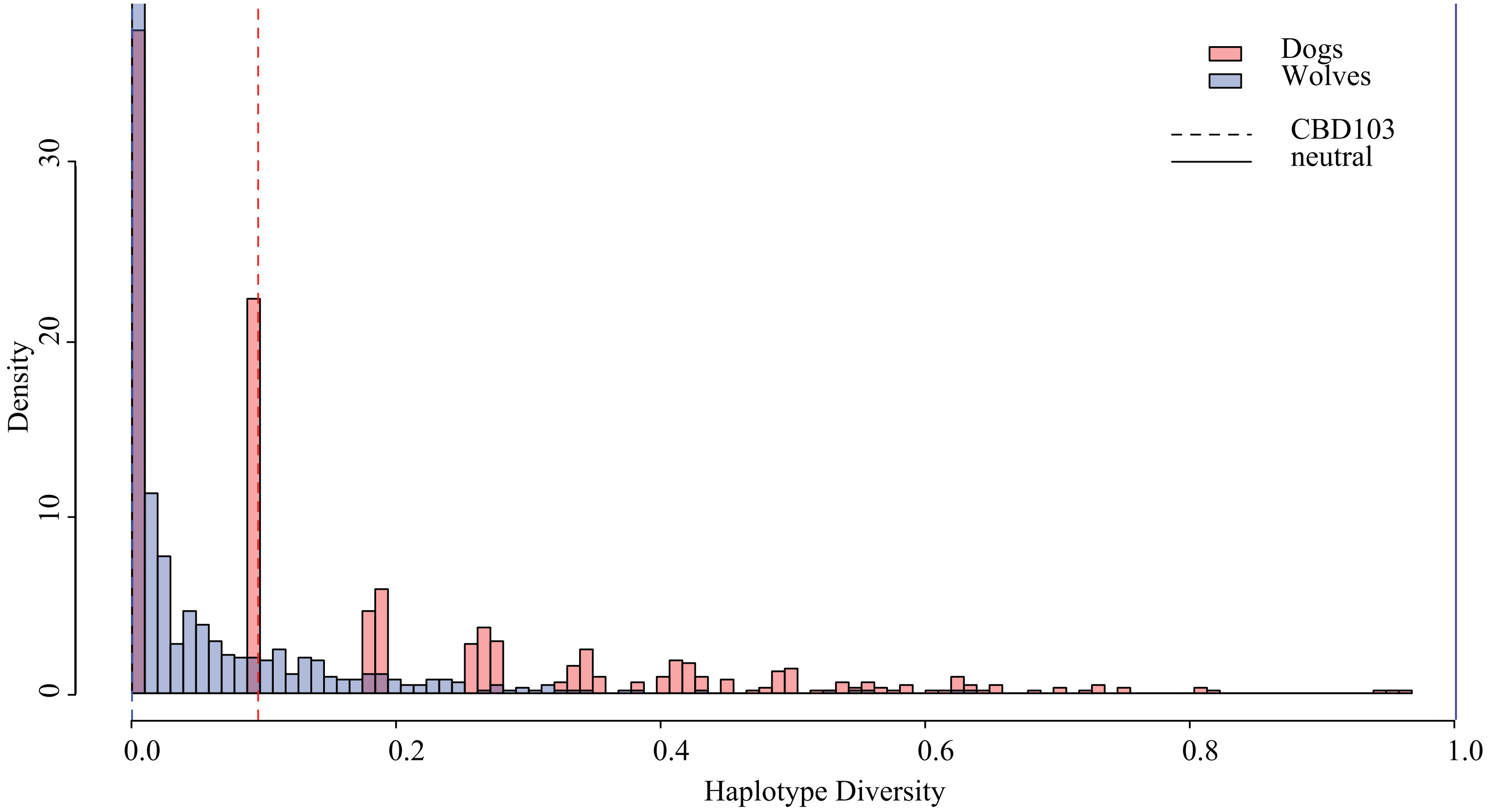


Figure 3-6. Histogram of haplotype diversity for a total of 1040 genes sequenced in dogs (red) and wolves (blue). The diversity of CBD103 is indicated with a vertical dashed line, and the diversity across 5Mb of neutral sequence data is indicated with a vertical solid line.

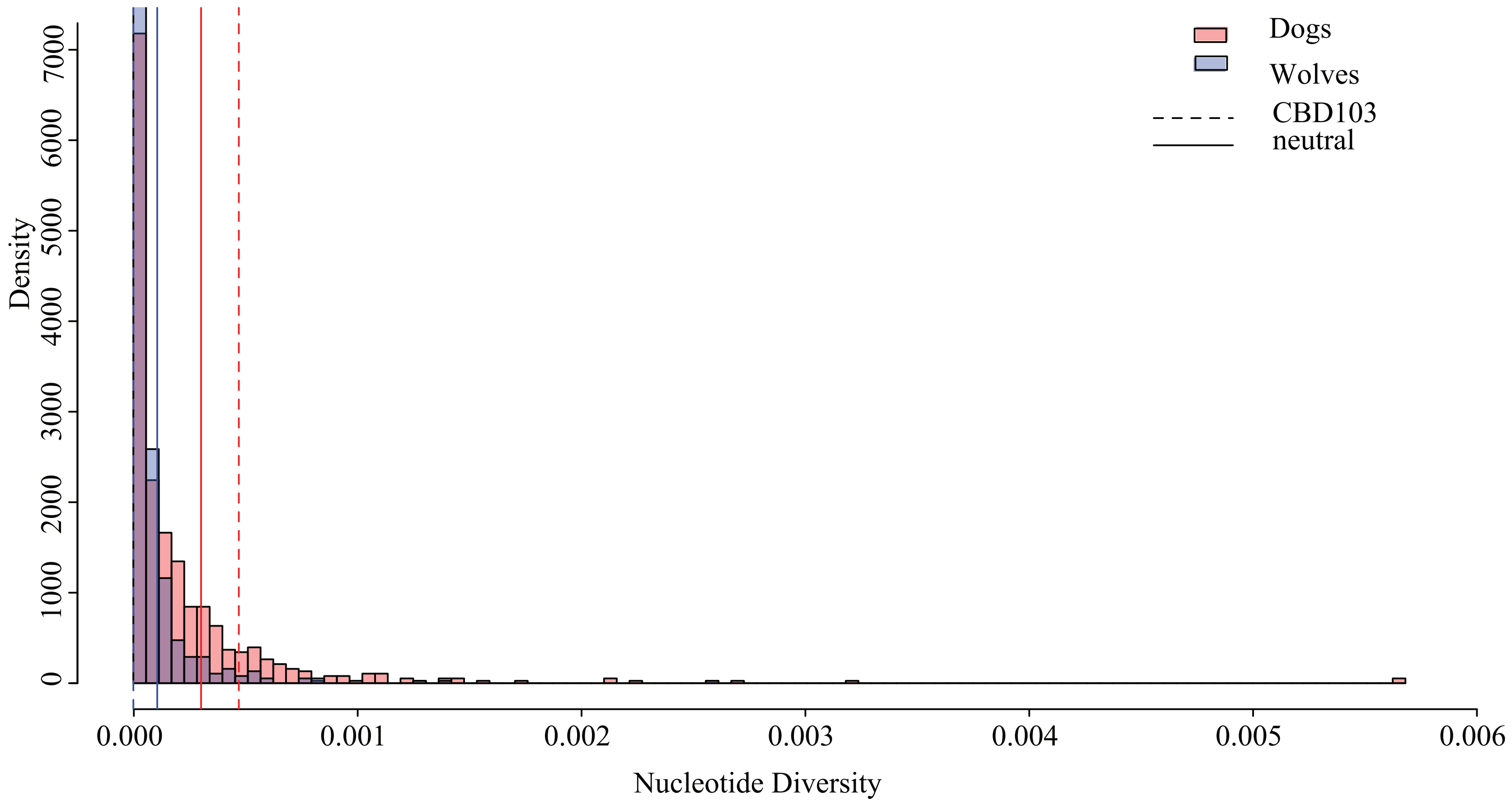


Figure 3-7. Histogram of haplotype diversity for a total of 1040 genes sequenced in dogs (red) and wolves (blue). The diversity of CBD103 is indicated with a vertical dashed line (visible at the zero line), and the diversity across 5Mb of neutral sequence data is indicated with a vertical solid line.

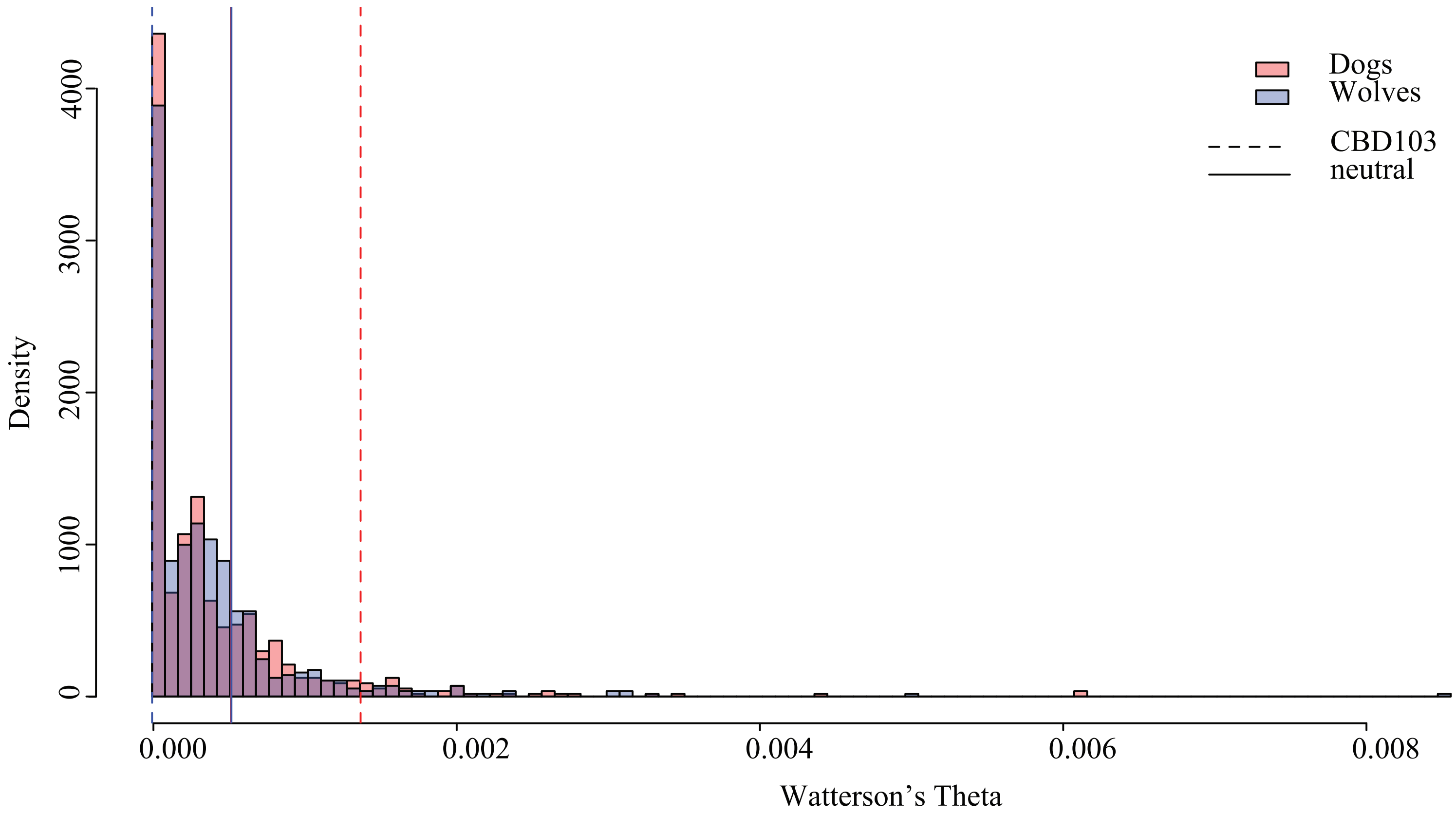


Figure 3-8. Histogram of Watterson's Theta diversity for a total of 1040 genes sequenced in dogs (red) and wolves (blue). The diversity of CBD103 is indicated with a vertical dashed line (in wolves, visible at theta=0), and the diversity across 5Mb of neutral sequence data is indicated with a vertical solid line.

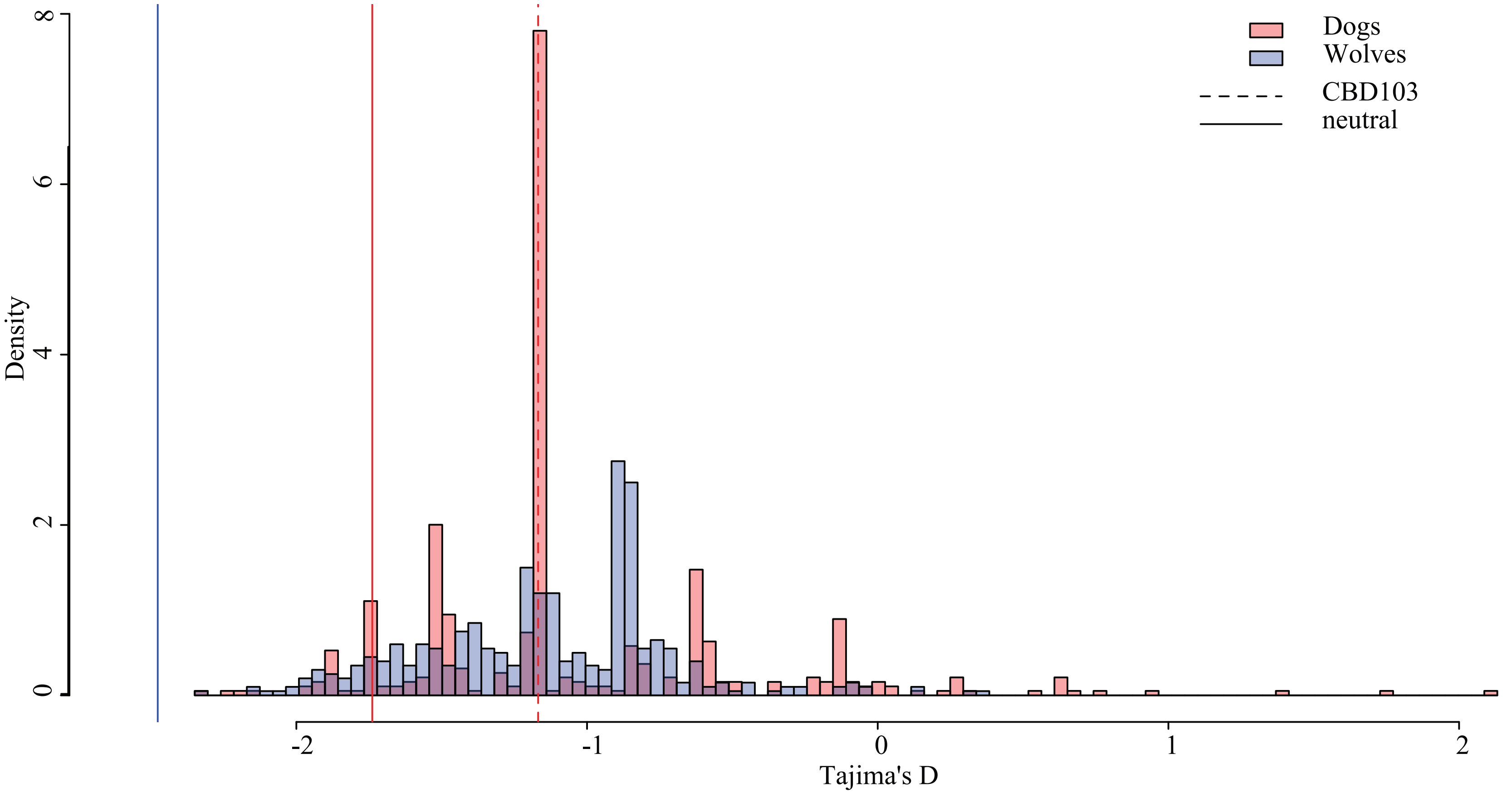


Figure 3-9. Histogram of Tajima’s *D* for a total of 1040 genes sequenced in dogs (red) and wolves (blue). The diversity of CBD103 is indicated with a vertical dashed line, and the diversity across 5Mb of neutral sequence data is indicated with a vertical solid line. Tajima’s D within wolves could not be calculated since both theta and pi were 0.

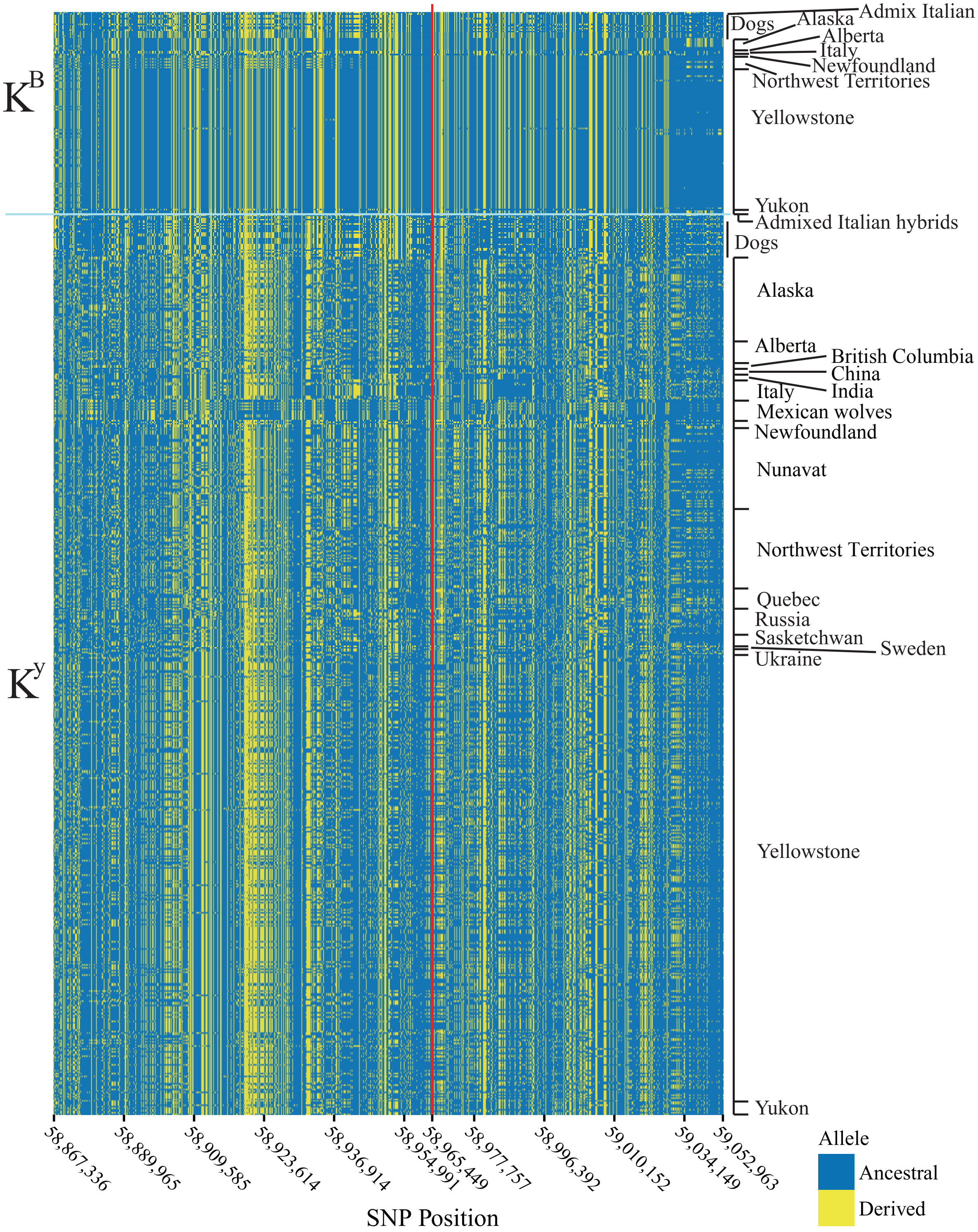


Figure 3-10. Haplotype structure plot for variable sites within the 200 Kb core region. Each row represents a single haplotype, colored for the ancestral (blue) or derived (yellow) allele variant surrounding the 3 bp deletion (vertical red line). Haplotypes are grouped according to geographic location, with K^B-containing haplotypes above the light blue horizontal line, and K^y-containing hapotypes below.

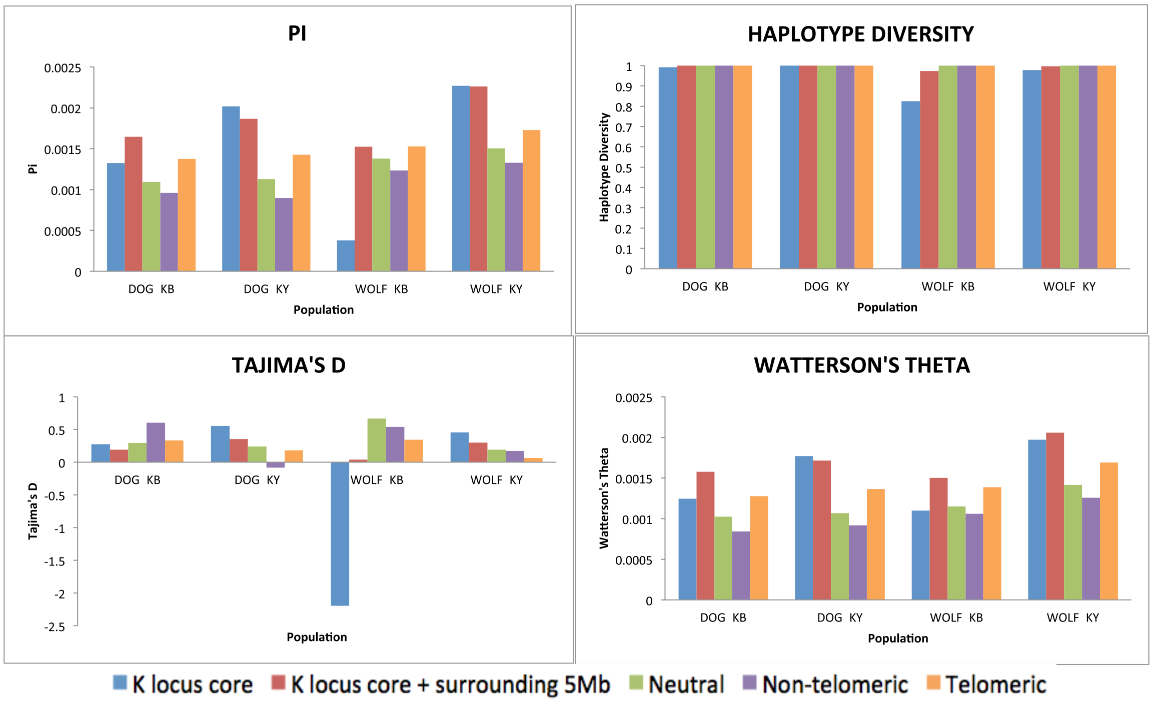


Figure 3-11. Measurements of pi, haplotype diversity, Tajima's D, and Watterson's theta in five different genomic regions. For neutral, non-telomeric, and telomeric regions, the same haplotypes that were K^B or K^y on chromosome 16 were used.

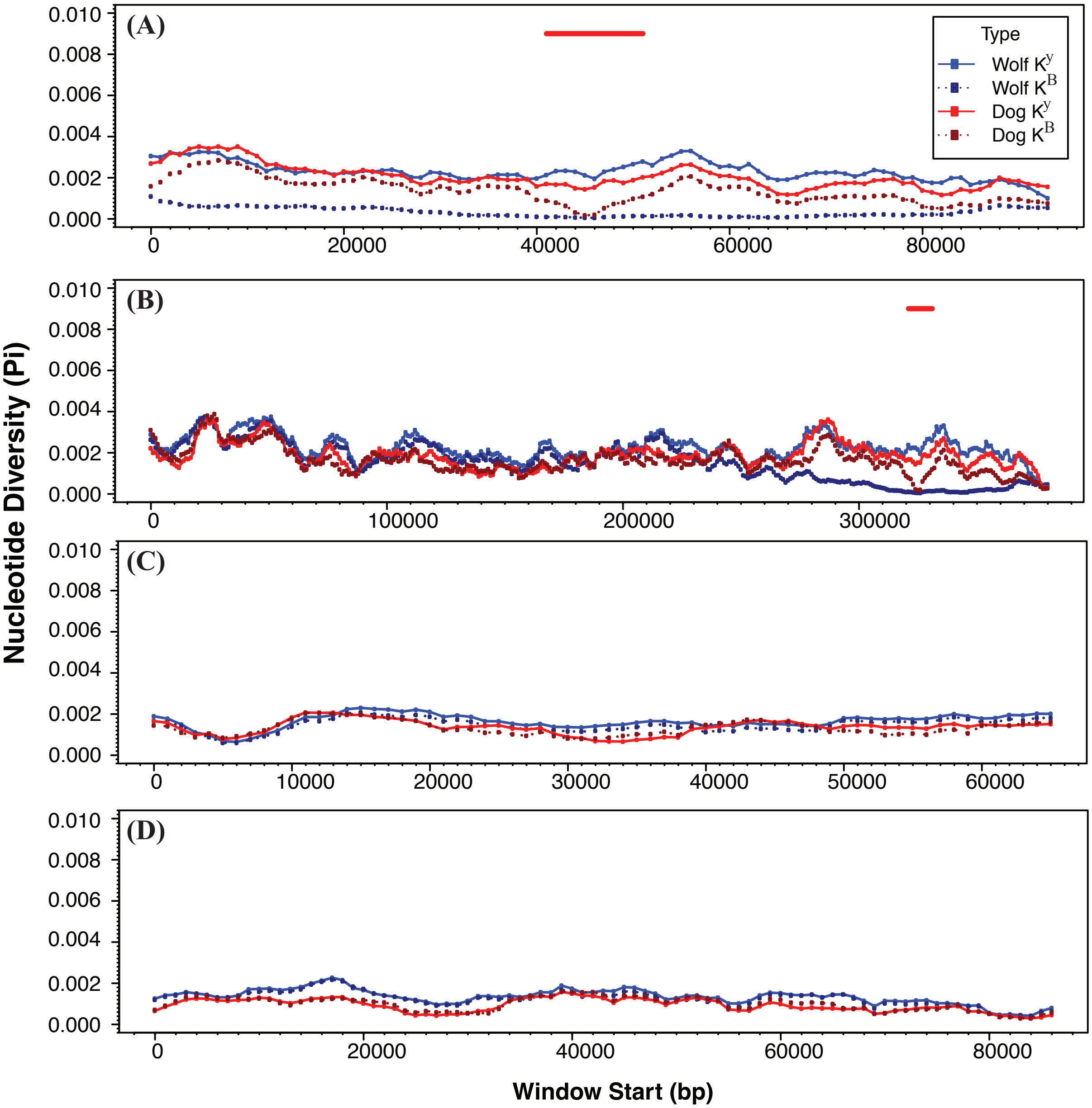


Figure 3-12. Nucleotide diversity in windows for dogs and wolves. Statistic was calculated in 10 Kb windows with 1 Kb step size for K^B and K^y haplotypes in wolves and dogs (see key in (A)) for: (A) K locus core, (B) K locus core plus surrounding 5 Mb, (C) Telomeric, and (D) Non-telomeric regions. Red horizontal lines in (A) and (B) indicate windows containing the 3 bp deletion. For each region, the phased sites were concatenated into a single sequence, with telomeric and non-telomeric regions ordered from smallest chromosome to largest chromosome.

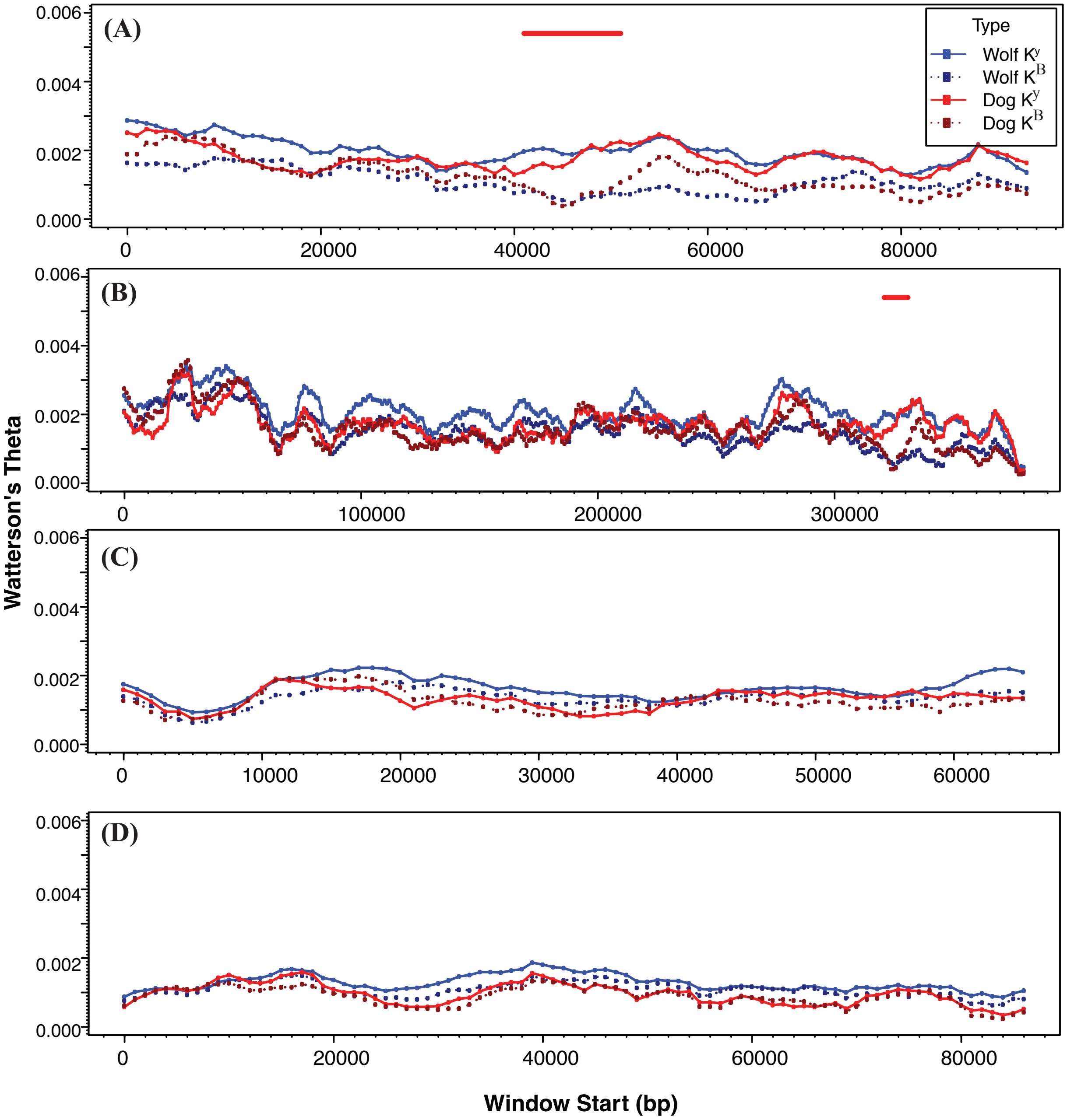


Figure 3-13. Watterson's Theta calculated in windows for dogs and wolves. Statistic was calculated in 10 Kb windows with 1kb step size for K^B and K^y haplotypes in wolves and dogs (see key in (A)) for: (A) K locus core, (B) K locus core plus surrounding 5 Mb, (C) Telomeric, and (D) Non-telomeric regions. Red horizontal lines in (A) and (B) indicate windows containing the 3 bp deletion. For each region, the phased sites were concatenated into a single sequence, with telomeric and non-telomeric regions ordered from smallest chromosome to largest chromosome.

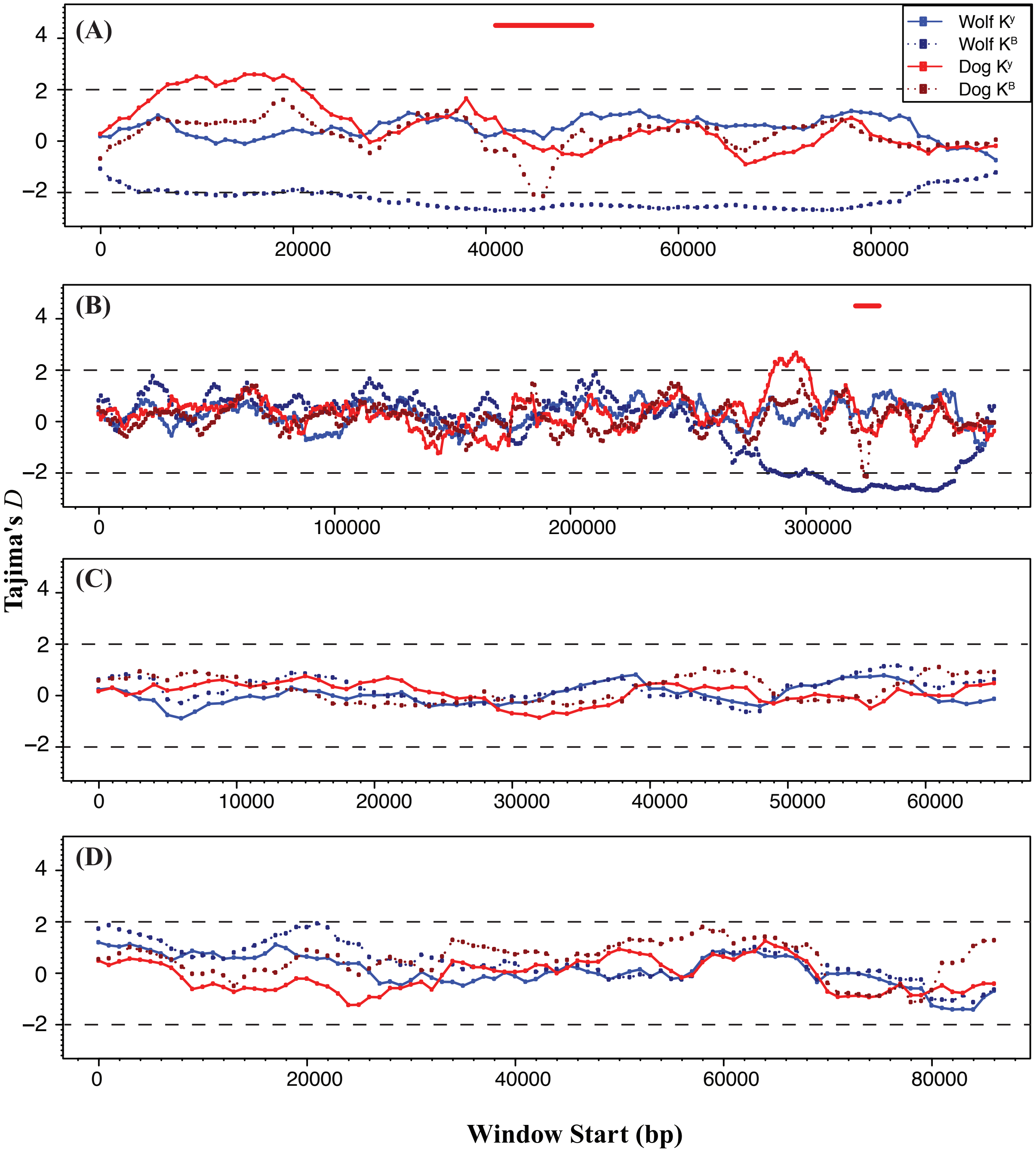


Figure 3-14. Tajima's *D* calculated in windows for dogs and wolves. Statistic was calculated in 10 Kb windows with 1 Kb step size for K^B and K^y haplotypes in wolves and dogs (see key in (A)) for: (A) K locus core, (B) K locus core plus surrounding 5 Mb, (C) Telomeric, and (D) Non-telomeric regions. Red horizontal lines in (A) and (B) indicate windows containing the 3 bp deletion. Telomeric and non-telomeric regions are concatented and ordered from smallest chromosome to largest.

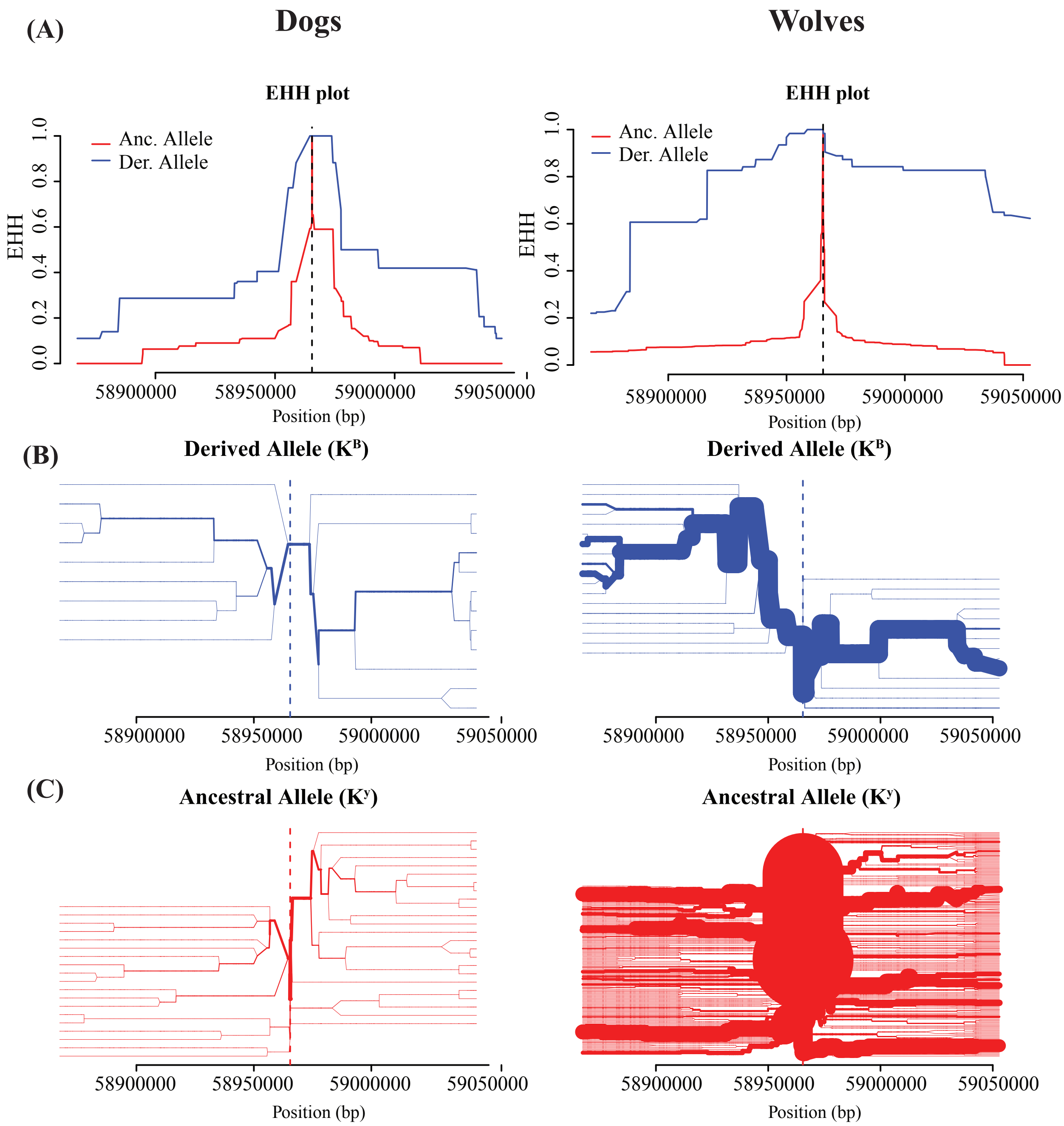


Figure 3-15. Extended haplotype homozygosity (EHH) decay and haplotype bifurcation plots for the 200 Kb core region in dogs (left) and wolves (right). (A) EHH scores along the 200 Kb K locus region show the decay of EHH with increasing distance from the core allele (vertical dashed line), for both ancestral K^y (red) and derived K^B (blue) haplotypes. (B) The haplotype bifurcation for derived K^B haplotypes within dogs (left) and wolves (right) regions. (C) Same as (B) but for ancestral K^y haplotypes.

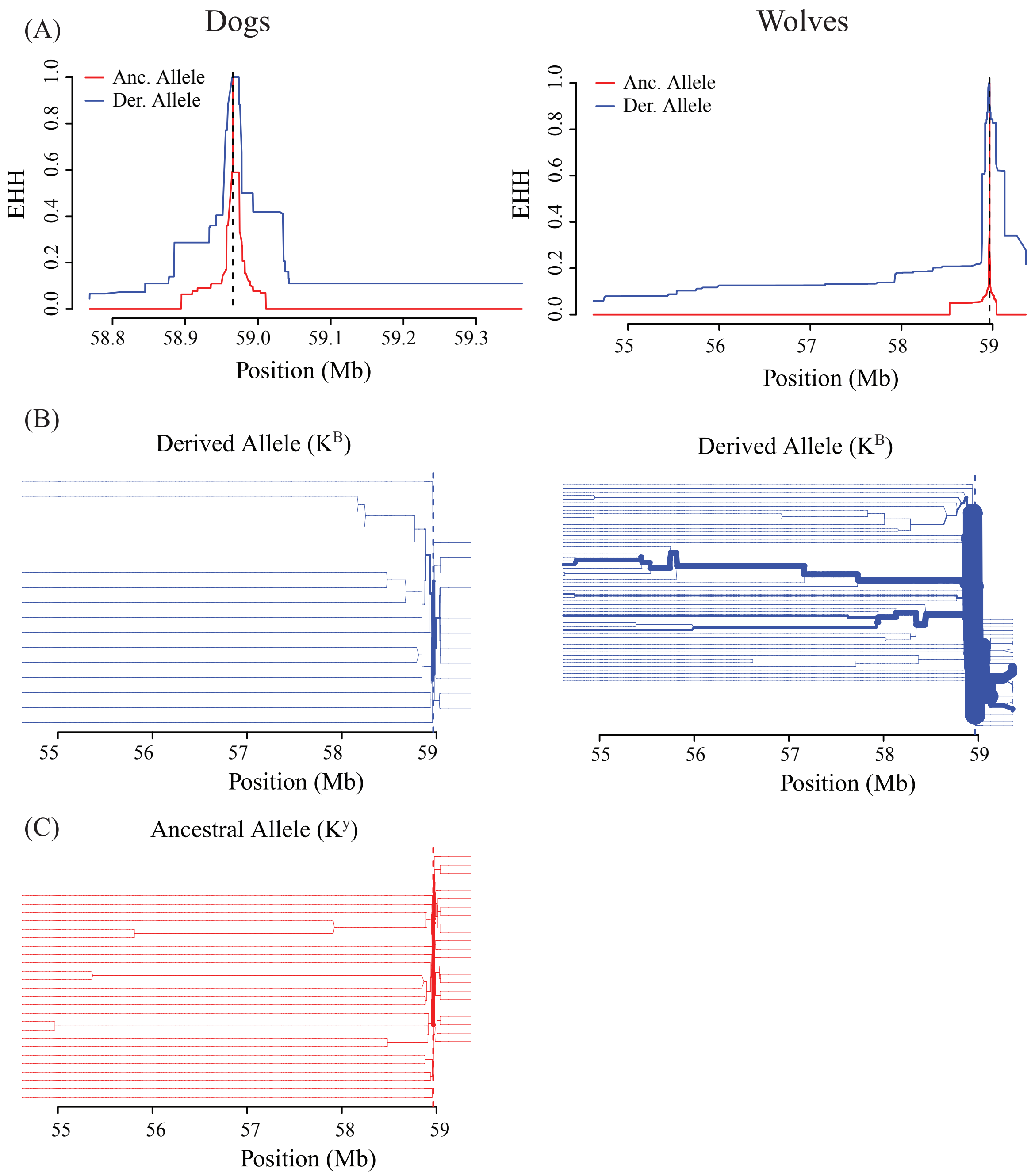


Figure 3-16. Extended haplotype homozygosity (EHH) decay and haplotype bifurcation plots for the 200 Kb core region plus surrounding 5 Mb in dogs (left) and wolves (right). (A) EHH scores along the 200 Kb K locus region show the decay of EHH with increasing distance from the core allele (vertical dashed line), for both ancestral K^y (red) and derived K^B (blue) haplotypes. Note the different extent of the scale between dogs and wolves. (B) The haplotype bifurcation for derived K^B haplotypes within dogs (left) and wolves (right) regions. (C) Same as (B) but for ancestral K^y haplotypes.

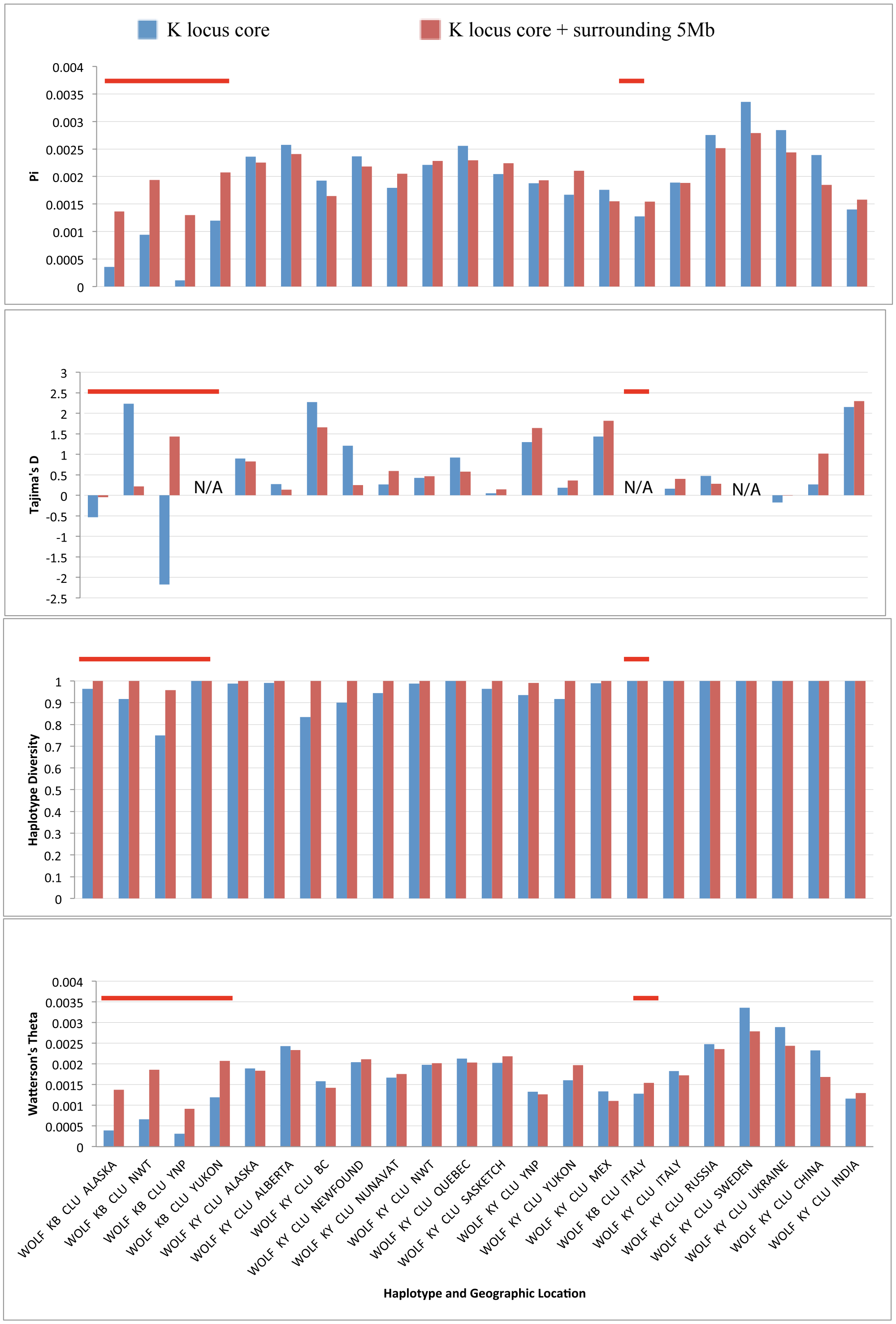


Figure 3-17. Nucleotide diversity, Tajima's *D*, haplotype diversity, and Watterson's Theta for K^B and K^y haplotypes separated by geographic location. Populations with the K^B allele are indicated by a horizontal red line.

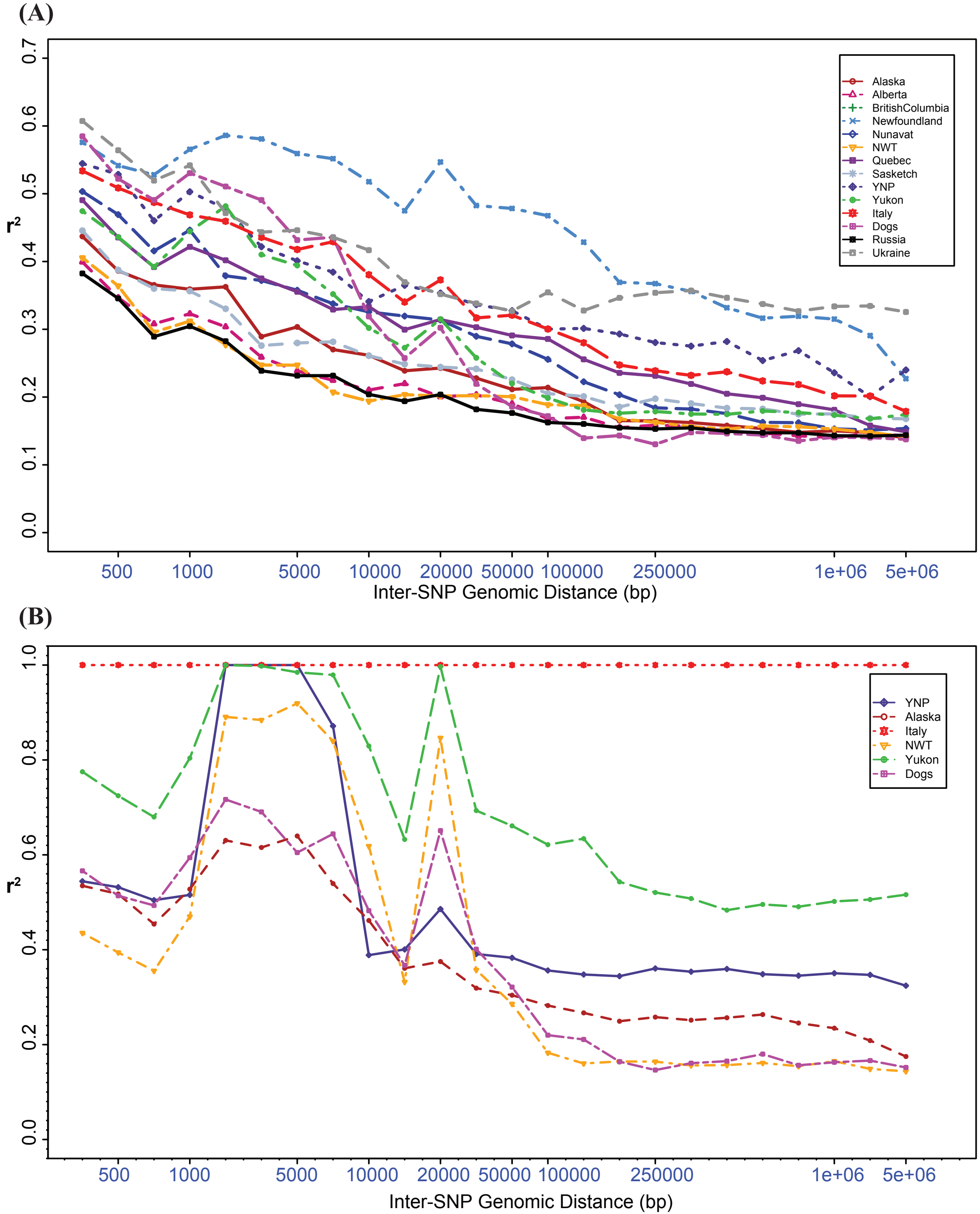


Figure 3-18. Decay of linkage disequilibrium (measured by r^2) for the K locus core plus surrounding 5Mb region, in (A) K^y-containing haplotypes and (B) K^B-containing haplotypes . Each population has been downsampled to a maximum of 8 haplotypes, and variants were filtered for a minor allele frequency greater than 0.05. Note the log scale and the slightly different y-axis scales.

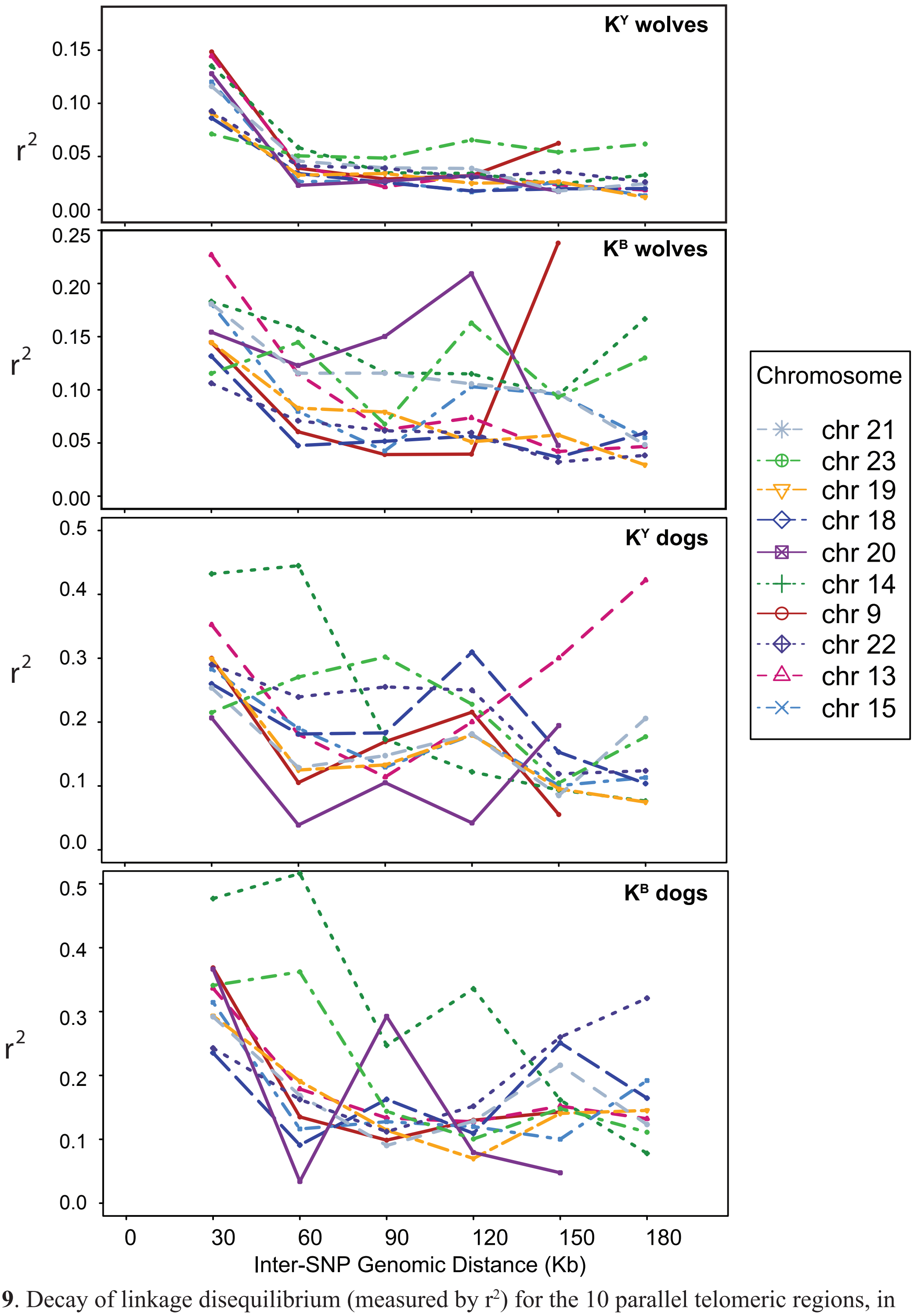


Figure 3-19. Decay of linkage disequilibrium (measured by r^2) for the 10 parallel telomeric regions, in wolves and dogs, for the same "K^y" and "K^B" individuals that were used for the K locus core. Each set has been downsampled to a maximum of 8 haplotypes, and variants were filtered for a minor allele frequency greater than 0.05. Note the log scale and the slightly different y-axis scales. Chromosomes in key are ordered from smallest to largest.

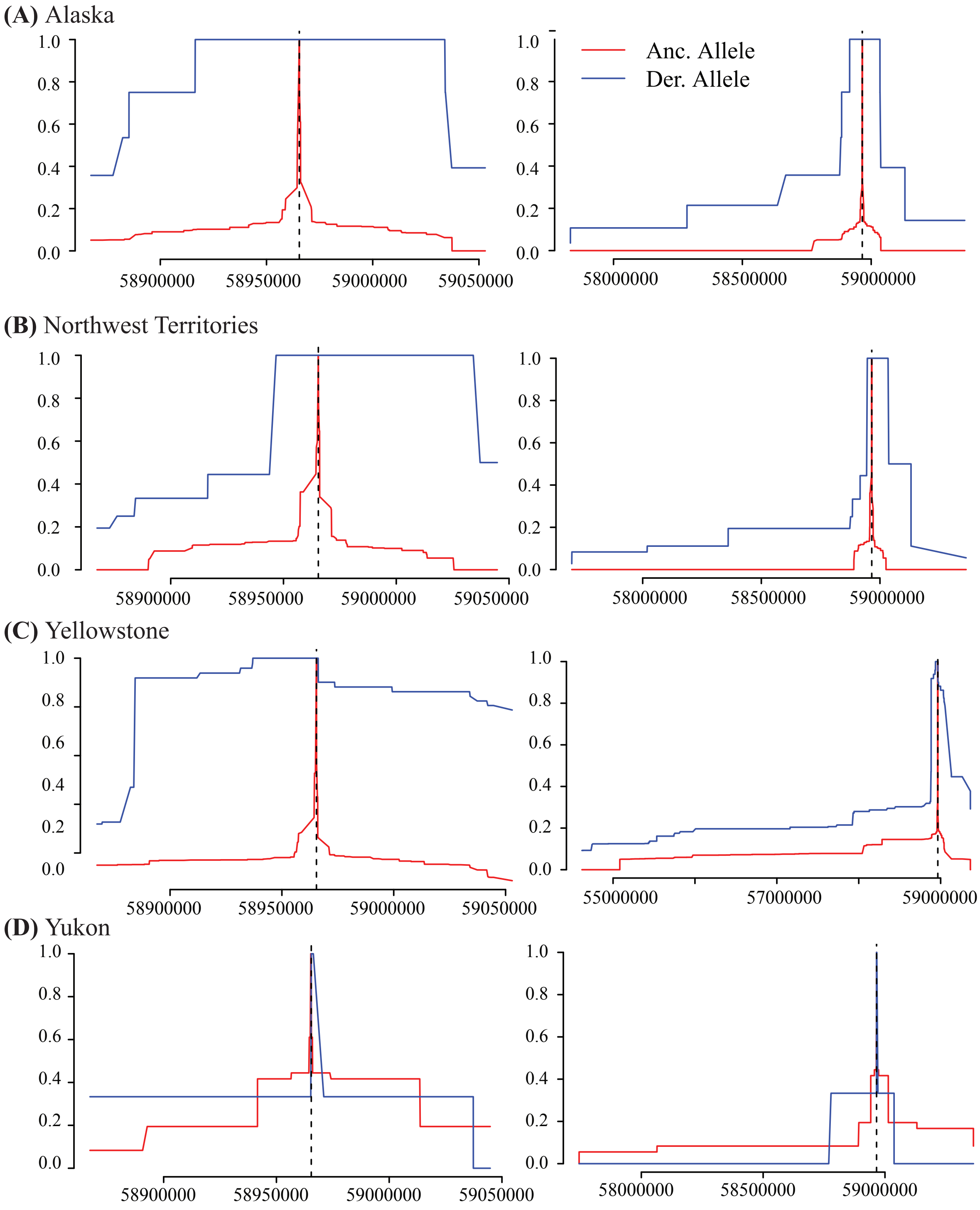


Figure 3-20. Decay of extended haplotype homozygosity (EHH) in ancestral K^y (red) and derived K^B (blue) haplotypes for North American populations containing more than a single K^B haplotype. The x-axis of each plot is the genomic position in bp along chromosome 16, and the y-axis is the EHH score. Plots on the left are of the 200 Kb core region, and plots on the right are of the 200 Kb core region plus surrounding 5Mb. Note the much more extensive decay for the surrounding 5Mb in Yellowstone wolves.

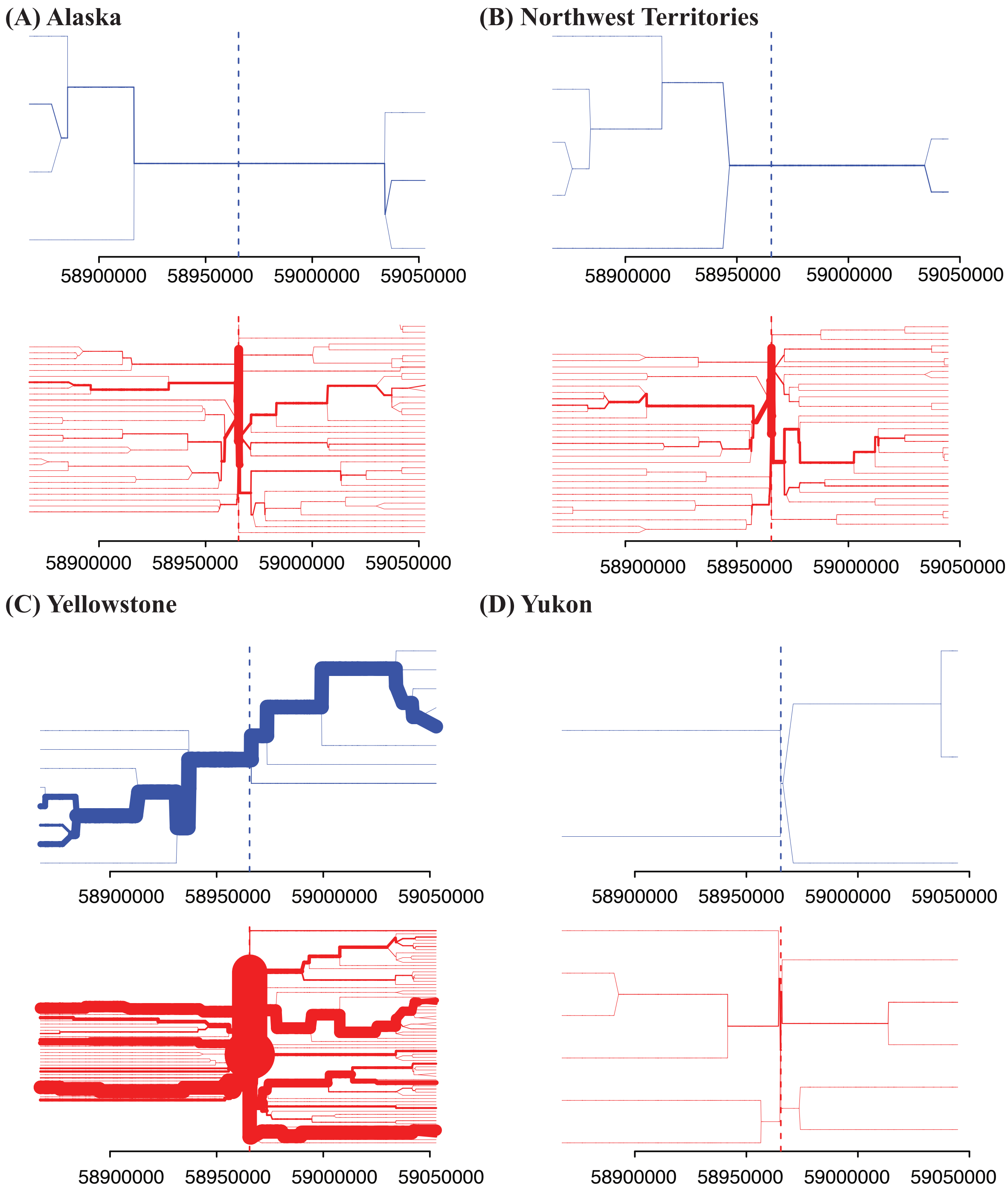


Figure 3-21. Patterns of haplotype bifurcation in four North American populations containing the K^B allele. The K^B allele is derived (blue), while the K^y allele is ancestral (red). The vertical dashed line indicates the location of the 3 bp deletion, branches represent haplotype bifucations, and the thickness of the line is proportional to the number of chromosomes. The position along chromosome 16, in bp, is provided at the bottom of each plot.

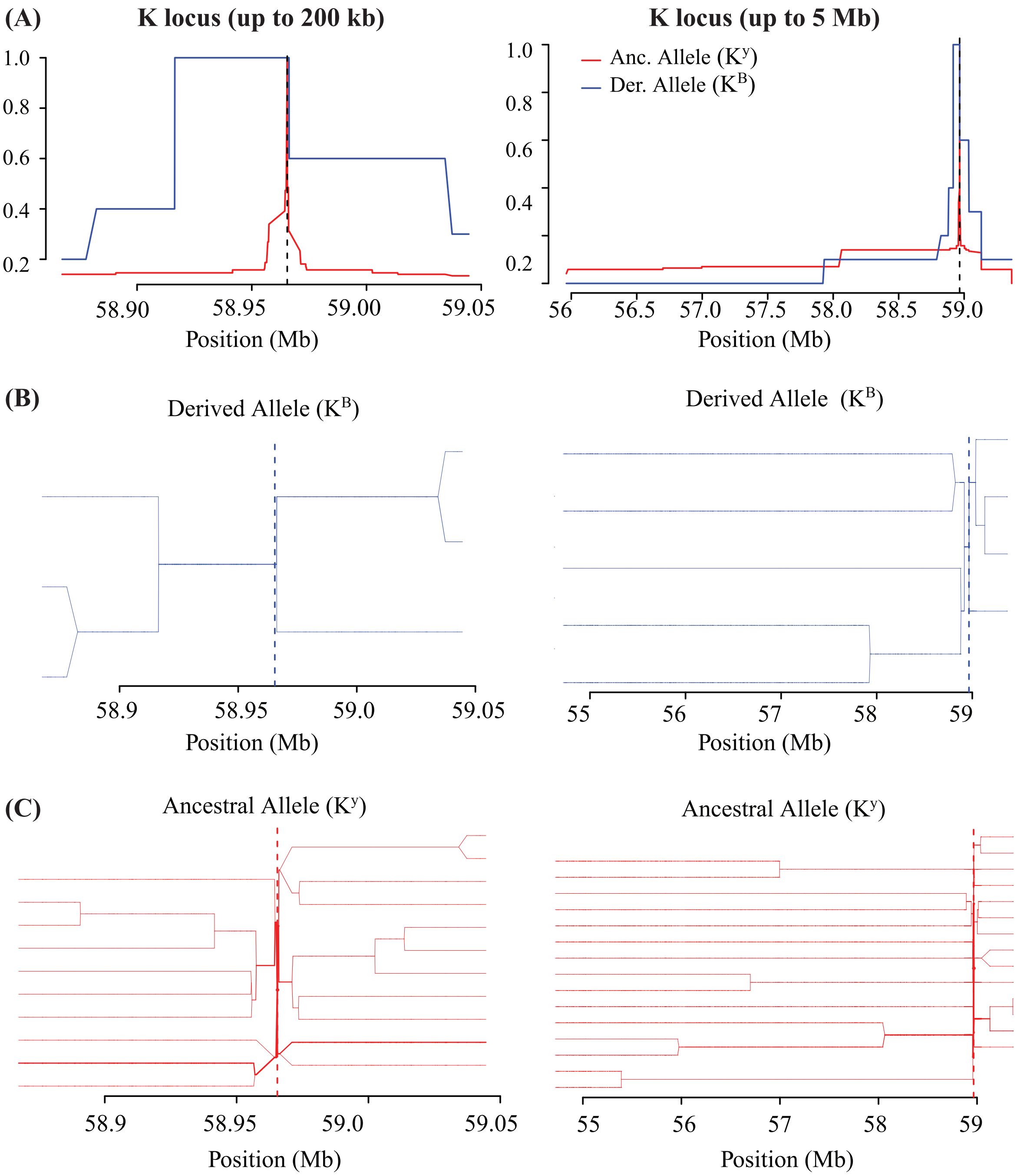


Figure 3-22. Extended haplotype homozygosity (EHH) and haplotype bifurcation within 12 Yellowstone founders. (A) EHH scores along the 200 Kb (left) and 5 Mb (right) K locus region show the decay of EHH with increasing distance from the core allele (vertical dashed line), for both ancestral K^y (red) and derived K^B (blue) haplotypes. Note that EHH does not extend as far upstream in these founders as in the current population (see Figure 3-20C). (B) The haplotype bifurcation for derived K^B haplotypes at 200 kb (left) and 5 Mb (right) regions. (C) Same as (B) but for ancestral K^y haplotypes.

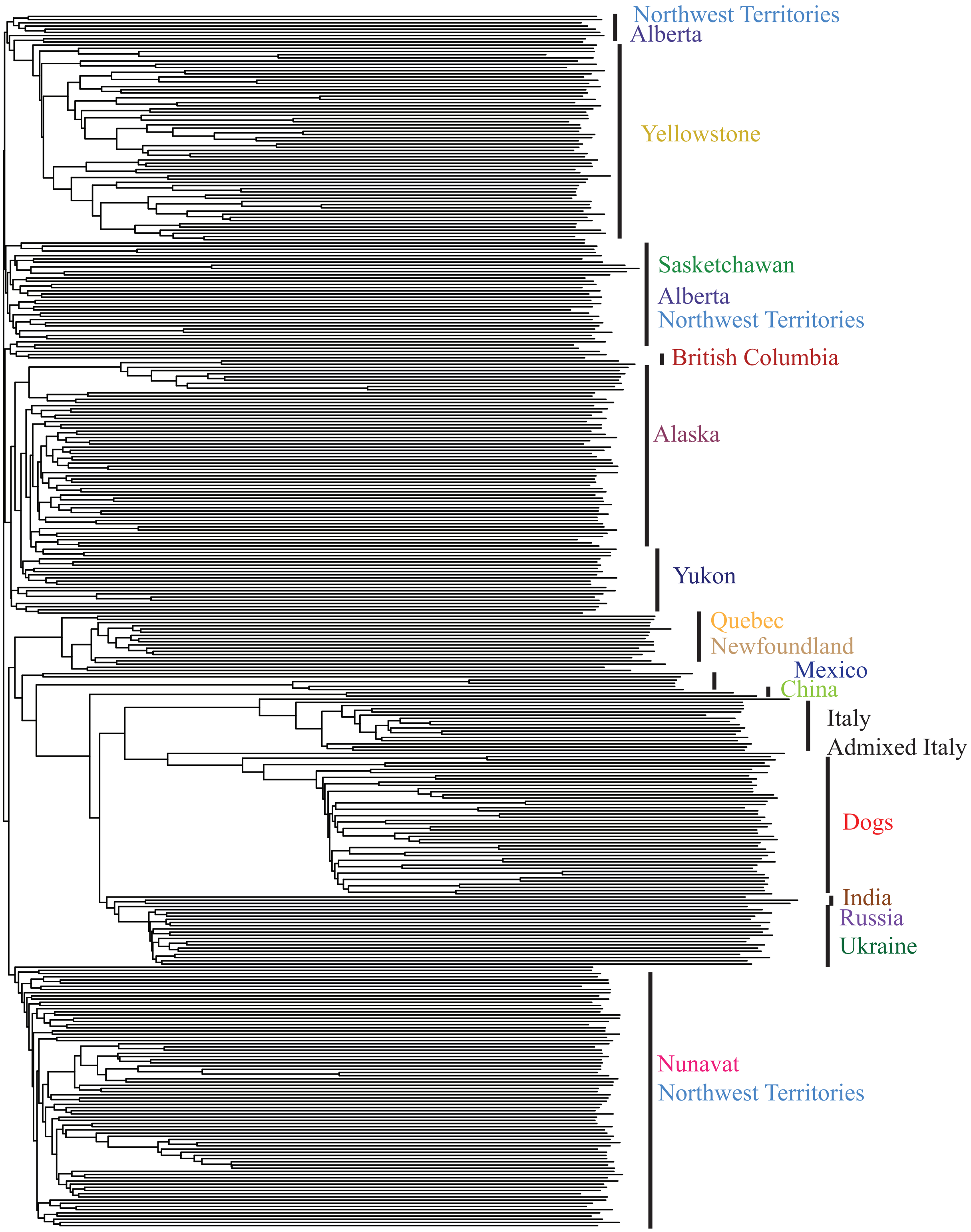


Figure 3-23. Neighbor joining tree based on pairwise differences between 52,872 variable neutral sites in 190 unrelated individuals. Groupings are concordant with geography rather than K^B genotype.

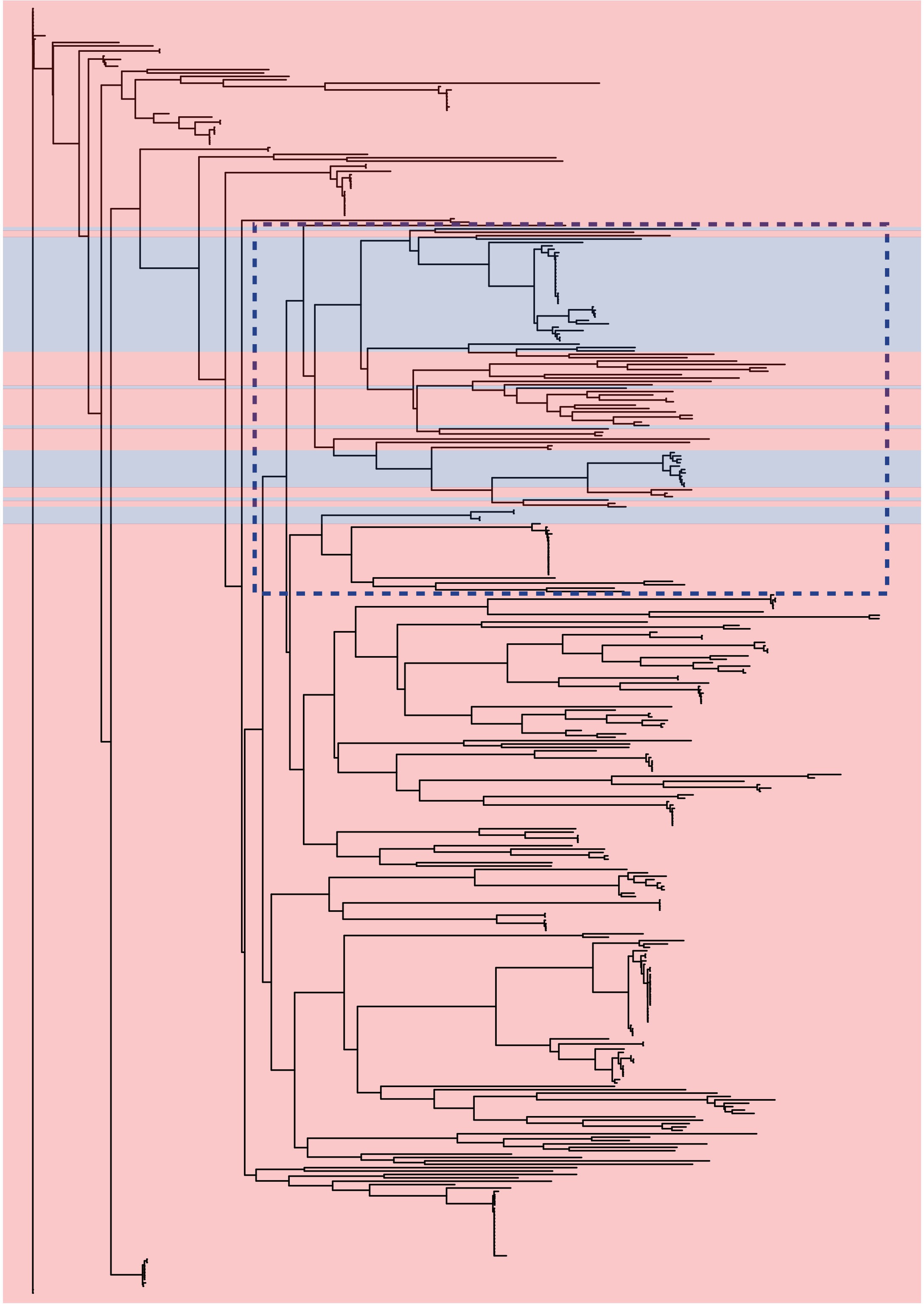


Figure 3-24. Neighbor-joining tree based on pairwise distance of sites within 200 Kb core region for 190 unrelated individuals, with K^B haplotypes (blue) and K^y (red) marked. See Figure 3-25 for zoom.

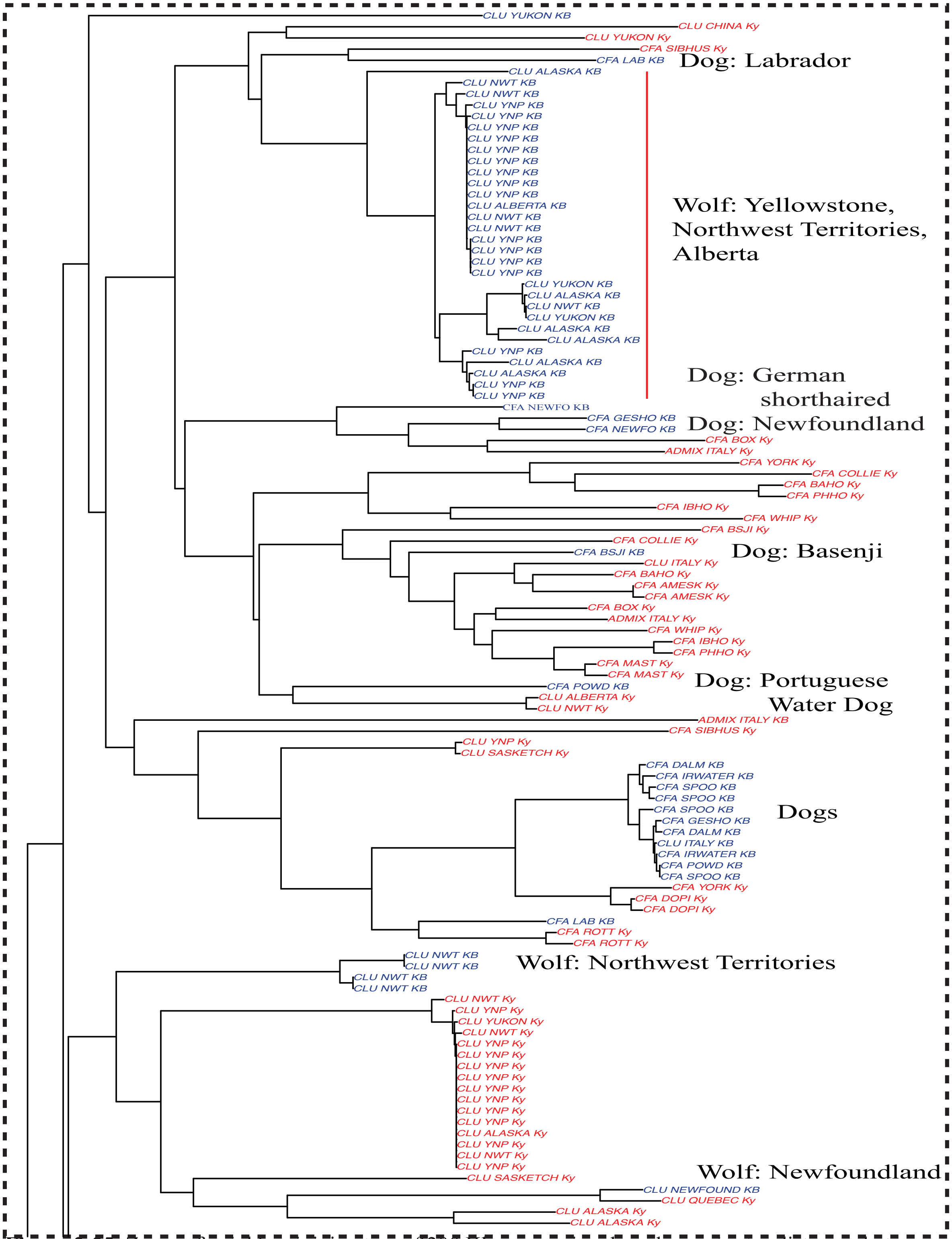


Figure 3-25. Zoom of neighbor-joining tree of 200 Kb core region, based on pairwise distances between haplotypes from 190 unrelated individuals. Iindividuals with K^B haplotypes are colored in blue. Dogs and wolf populations are labeled.

	# Sequenced	# Used for Analyses	K^B haplotypes	K^y haplotypes	Black	Gray	White	Unknown
Gray Wolf	**381**	**357**	**121**	**593**				
United States	253	243	105	381				
Alaska	34	33	8	58	6	16	0	11
Mexican Wolf	7	7	0	14	0	0	0	7
Yellowstone	212	203	97	309	90	113	0	0
Canada	101	90	14	166				
Alberta	8	8	1	15	1	3	3	1
British Columbia	2	2	0	4	0	0	0	2
Manitoba	4	0	0	0				
Newfoundland	4	3	1	5	0	0	0	3
Northwest Territories	32	32	9	55	4	4	13	11
Nunavut	29	28	0	56	0	0	9	19
Ontario	2	0	0	0				
Quebec	8	7	0	14	0	0	0	8
Saskawatchen	5	4	0	8	0	0	0	4
Yukon	7	6	3	9	0	0	0	6
Old World	27	24	2	46				
China	2	2	0	4	0	0	0	2
India	2	2	0	4	0	0	0	2
Italy	11	8	2	14	2	2	0	4
Russia	9	9	0	18	0	0	0	9
Sweden	1	1	0	2	0	0	0	1
Ukraine	2	2	0	4	0	0	0	2
Domestic dog	**21**	21	17	25				
Breeds fixed for K^B	8	8	16	0				
Breeds fixed for k^y	6	6	0	12				
Breeds variable for k^y and k^{br}	4	4	1	7				
Breeds with unknown K allel	3	3	0	6				
TOTAL	402	378	138	618				

Table 3-1 Sampling information for wolves and dogs sequenced on the capture array.

Breeds fixed for K^B: Dalmation, German Shorthaired Pointer, Irish Water Spaniel, Labrador Retriever, Newfoundland, Standard Poodle

Breeds fixed for K^y: Basset Hound, Collie, Doberman Pinscher, Rottweiler, Siberian Husky, Yorkshire Terrier

Breeds with unknown K alleles: American Eskimo Dog, Ibizan Hound, Pharaoh Hound

Breeds variable for Ky and Kbr: Boxer, Mastiff, Whippet, Basenji

	Yield (Mbases)	Reads PF	# Reads	Perfect Index Reads	≥ Q30 Bases (PF)	Mean Quality Score (PF)	Mapped Reads	Mapped to canfam3.1	Unique Mapped Reads	PCR Duplicates	Unique Mapped Reads
Average	1,956.92	89.29%	22,033,756	96.41%	86.74%	34.42	19,344,340	88.62%	18,086,557	7.71%	82.34%
Standard Deviation	550.45	3.36%	6,532,489	5.31%	3.01%	0.96	6,402,333	12.17%	6,228,979	5.33%	18.94%

Table 3-2. Summary of sequencing yields averaged over 403 individuals. PF: Pass Filter

		Sequencing Genotype			
		0/0	0/1	1/1	Number of Sites
Array Genotype	0/0	**99.490%**	0.510%	0.000%	6080
	0/1	0.179%	**99.702%**	0.119%	3351
	1/1	0.019%	0.459%	**99.522%**	5232

Table 3-3. Genotyping concordance for 109 individuals and 204 sites that overlapped between the Affymetrix dog SNP array v2 and the capture array target intervals. Homozygote reference: 0/0; Heterozygote: 0/1; Homozygote non-reference: 1/1

Region Type	Num. Sites Pass Filter	Num. Variable Sites Pass Filter	Num. Phased Sites	Num. Variable Phased Sites
All	10,922,248	162,815	7,761,114	80,309
Neutral	5,025,818	56,662	4,908,660	52,972
K locus core	134,897	3,252	102,529	1,542
K locus core + 5Mb surrounding	461,205	9,737	389,456	6,114
Genic	3,740,912	92,323	695,949	1,554
Telomeric	86,905	1,550	74,834	986
Non-telomeric	101,023	1,074	95,670	908

Table 3-4. Sites passing filter for multiple genomic regions. Filters applied to sequencing data were depth of coverage ≥ 10, genotype quality ≥ 30, and call rate ≥95%. Phased data had these same filters, but with a 100% call rate. Numerical differences between sum of region types and "All" cateogry are due to additional regions being sequenced as a result of genomic DNA overlap on capture baits.

Region Type	**Group**	**θw**	**π**	***D***	***H***
K locus Core	Dog K^B	0.00124	0.00132	0.27244	0.99265
	Dog K^Y	0.00177	0.00202	0.55465	1.00000
	Wolf K^B	0.00110	0.00038	-2.19223	0.82479
	Wolf K^Y	0.00197	0.00227	0.45498	0.97791
K locus Core + 5Mb	Dog K^B	0.00158	0.00165	0.18990	1.00000
	Dog K^Y	0.00171	0.00187	0.35472	1.00000
	Wolf K^B	0.00150	0.00152	0.04168	0.97314
	Wolf K^Y	0.00206	0.00226	0.29925	0.99748
Neutral	Dog K^B	0.00102	0.00109	0.29581	1.00000
	Dog K^Y	0.00107	0.00113	0.23887	1.00000
	Wolf K^B	0.00115	0.00138	0.66693	1.00000
	Wolf K^Y	0.00141	0.00150	0.18982	1.00000
Non-Telomeric	Dog K^B	0.00084	0.00096	0.60382	1.00000
	Dog K^Y	0.00092	0.00090	-0.08376	1.00000
	Wolf K^B	0.00106	0.00123	0.54102	0.99972
	Wolf K^Y	0.00126	0.00133	0.16952	0.99997
Telomeric	Dog K^B	0.00128	0.00137	0.33132	1.00000
	Dog K^Y	0.00136	0.00142	0.18242	1.00000
	Wolf K^B	0.00139	0.00153	0.34434	0.99972
	Wolf K^Y	0.00169	0.00173	0.06507	0.99998

Table 3-5. Summary statistics for K^B and K^y haplotypes within dogs and wolves, for each of five region types within the genome. Θw: Watterson's Theta, π: Nucleotide Diversity, *D*: Tajima's D, *H*: Haplotype Diversity

Bibliography

Albert F, Hodges E, Jensen J, Besnier F (2011) Targeted resequencing of a genomic region influencing tameness and aggression reveals multiple signals of positive selection. *Heredity*.

Allen AP, Brown JH, Gillooly JF (2002) Global biodiversity, biochemical kinetics, and the energetic-equivalence rule. *Science*, **297**, 1545–1548.

Anderson TM, Candille SI, Musiani M *et al.* (2009) Molecular and evolutionary history of melanism in North American gray wolves. *Science*, **323**, 1339–1343.

Auton A, Rui Li Y, Kidd J *et al.* (2013) Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs (I Henderson, Ed,). *PLoS Genetics*, **9**, e1003984.

Bangs, EE & Fritts, SH (1996) Reintroducing the gray wolf to central Idaho and Yellowstone National Park. Wildlife Society Bulletin, **24**, 402–413.

Beall CM, Cavalleri GL, Deng L (2010) Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Brown SK, Darwent CM, Sacks BN (2013) Ancient DNA evidence for genetic continuity in arctic dogs. *Journal of Archaeological Science*, **40**, 1279–1288.

Candille SI, Kaelin CB, Cattanach BM *et al.* (2007) A -defensin mutation causes black coat color in domestic dogs. *Science*, **318**, 1418–1423.

Coulson T, Macnulty DR, Stahler DR *et al.* (2011) Modeling Effects of Environmental Change on Wolf Population Dynamics, Trait Evolution, and Life History. *Science*, **334**, 1275–1278.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, **27**, 2156–2158.

De Mita S, Siol M (2012) EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC genetics*, **13**, 27.

DePristo M, Banks E, Poplin R, Garimella K (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*.

Dionne M, Miller KM, Dodson JJ, Caron F, Bernatchez L (2007) Clinal variation in MHC diversity with temperature: evidence for the role of host-pahtogen interaction on local adaptation in Atlantic salmon. *Evolution*, **61**, 2154–2164.

Domingues VS, Poh Y-P, Peterson BK *et al.* (2012) Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution*, **66**, 1–15.

Erles K, Brownlie J (2010) Expression of Î²-defensins in the canine respiratory tract and antimicrobial activity against Bordetella bronchiseptica. *Veterinary Immunology and Immunopathology*, **135**, 12–19.

Faircloth, BC. 2015. Illumina TruSeq Library Prep for Target Enrichment. Available from http://ultraconserved.org (Accessed March 10, 2013).

Faircloth BC, Glenn TC (2012) Not All Sequence Tags Are Created Equal: Designing and Validating Sequence Identification Tags Robust to Indels (S-H Shiu, Ed,). *PLoS ONE*, **7**, e42543.

Freedman AH, Gronau I, Schweizer RM *et al.* (2014) Genome Sequencing Highlights the Dynamic Early History of Dogs (L Andersson, Ed,). *PLoS Genetics*, **10**, e1004016.

Gipson P, Bangs E, Bailey T *et al.* (2002) Color patterns among wolves in western North America. *Wildlife Society Bulletin*, **30**, 821–830.

Gray MM, Granka JM, Bustamante CD *et al.* (2009) Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, **181**, 1493–1505.

Gray MM, Sutter NB, Ostrander EA, Wayne RK (2010) The IGF1 small dog haplotype is derived from Middle Eastern grey wolves. *BMC biology*, **8**, 16.

Guernier V, Hochberg ME, Guégan J-F (2004) Ecology Drives the Worldwide Distribution of Human Diseases. *PLoS Biology*, **2**, e141.

Hoekstra HE, Nachman M (2003) Different genes underlie adaptive melanism in different populations of rock pocket mice. *Molecular Ecology*, **12**, 1185–1194.

Huerta-Sanchez E, Jin X, Asan *et al.* (2014) Altitude adaptation in Tibetans caused byintrogression of Denisovan-like DNA. *Nature*, 1–17.

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, **4**, 1073–1081.

Kurtz S, Narechania A, Stein JC, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC genomics*, **9**, 517.

Leonard JA, Wayne RK, Wheeler J *et al.* (2002) Ancient DNA evidence for Old World origin of New World dogs. *Science*, **298**, 1613–1616.

Li H (2014) Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples. *arXiv.org*, **q-bio.GN**.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078–2079.

Lucchini V, Galov A, Randi E (2004) Evidence of genetic distinction and long-term population decline in wolves (Canis lupus) in the Italian Apennines. *Molecular Ecology*, **13**, 523–536.

Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE (2010) Convergence in pigmentation at multiple levels: mutations, genes and function. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*, **365**, 2439–2450.

Maynard-Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genet Res*, **23**, 23–25.

Mech D, Boitani L (2003) Wolves: behavior, ecology, and conservation. (eds Mech D, Boitani L). The University of Chicago Press, Chicago, Illinois.

Mullen LM, Hoekstra HE (2008) Natural selection along an environmental gradient: a classic cline in mouse pigmentation. *Evolution*, **62**, 1555–1570.

Musiani M, Leonard JA, Cluff HD *et al.* (2007) Differentiation of tundra/taiga and boreal coniferous forest wolves: genetics, coat colour and association with migratory caribou. *Molecular Ecology*, **16**, 4149–4170.

Nachman M, Hoekstra HE, D'Agostino S (2003) The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences*, **100**, 5268–5273.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, **8**, 857–868.

O'Connell J, Gurdasani D, Delaneau O *et al.* (2014) A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics*, **10**, e1004234–e1004234.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)*, **20**, 289–290.

Pazgier M, Hoover DM, Yang D, Lu W, Lubkowski J (2006) Human beta-defensins. *Cellular and molecular life sciences : CMLS*, **63**, 1294–1313.

Pilot M, Greco C, vonHoldt BM *et al.* (2013) Genome-wide signatures of population bottlenecks and diversifying selection in European wolves. *Heredity*.

Price A, Patterson N, Plenge R *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

Protas ME, Patel NH (2008) Evolution of coloration patterns. **24**, 425–446.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E (2015) Evidence for archaic adaptive introgression in humans. *Nature Publishing Group*, 1–13.

Sabeti PC, Reich DE, Higgins JM *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Song Y, Endepols S, Klemann N *et al.* (2011) Adaptive Introgression of Anticoagulant Rodent Poison Resistanceby Hybridization between Old World Mice. *Current Biology*, **21**, 1296–1301.

Stahler DR, MacNulty DR, Wayne RK, vonHoldt B, Smith DW (2012) The adaptive value of morphological, behavioural and life-history traits in reproductive female wolves (F Pelletier, Ed,). *The Journal of animal ecology*, **82**, 222–234.

Tajima, F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585-595.

Tishkoff SA (2001) Haplotype Diversity and Linkage Disequilibrium at Human G6PD: Recent Origin of Alleles That Confer Malarial Resistance. *Science*, **293**, 455–462.

van Asch B, Zhang A-B, Oskarsson MCR *et al.* (2013) Pre-Columbian origins of Native American dog breeds, with only limited replacement by European dogs, confirmed by mtDNA analysis. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20131142.

Verardi A, Lucchini V, Randi E (2006) Detecting introgressive hybridization between free-ranging domestic dogs and wild wolves (Canis lupus) by admixture linkage disequilibrium analysis. *Molecular Ecology*, **15**, 2845–2855.

Vignieri SN, Larson JG, Hoekstra HE (2010) The selective advantage of crypsis in mice. *Evolution*, **64**, 2153–2158.

vonHoldt BM, Pollinger JP, Earl DA *et al.* (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Research*, **21**.

vonHoldt BM, Pollinger JP, Lohmueller KE *et al.* (2010a) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, **464**, 898–902.

vonHoldt BM, Stahler DR, Bangs EE *et al.* (2010b) A novel assessment of population structure and gene flow in grey wolf populations of the Northern Rocky Mountains of the United States. *Molecular Ecology*, no–no.

vonHoldt BM, Stahler DR, Smith DW *et al.* (2008) The genealogy and genetic viability of reintroduced Yellowstone grey wolves. *Molecular Ecology*, **17**, 252–274.

Wall JD, Cox MP, Mendez FL *et al.* (2008) A novel DNA sequence database for analyzing human demographic history. *Genome Research*, **18**, 1354–1361.

Wang J (2010) coancestry: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, **11**, 141–145.

Watterson, GA (1975) On the number of segregating sites in genetical models without recombination. T*heoretical Population Biology,* **7**:256-276.

Whitney KD, Randell RA, Rieseberg LH (2006) Adaptive Introgression of Herbivore Resistance Traits in the Weedy Sunflower Helianthus annuus. *The American Naturalist*, **167**, 794–807.

Yang D, Chertov O, Bykovskaia S *et al.* (1999) β-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6. *Science*, **286**, 525.

Zhang W, Fan Z, Han E *et al.* (2014) Hypoxia Adaptations in the Grey Wolf (*Canis lupus chanco*) from Qinghai-Tibet Plateau. *PLoS Genetics*, **10**, e1004466.