UNIVERSITY OF CALIFORNIA

Los Angeles

Understanding the Persistence of Deleterious Variation Across Taxa

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

by

Jazlyn Ambry Mooney

2020

ABSTRACT OF THE DISSERTATION

Understanding the Persistence of Deleterious Variation Across Taxa

by

Jazlyn Ambry Mooney

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2020

Professor Kirk Lohmueller, Chair

The genetic architecture of complex traits in humans has received considerable attention in the past few years, especially with the advent of large biobank data sets, which provide valuable insights about complex traits in humans. However, there are still outstanding questions about the genetic variation that contributes to disease phenotypes. For my dissertation research, I am examining the joint impact of population history and natural selection in order to determine how these forces have allowed for deleterious variation to accumulate in populations. First, I will determine how recent demography affects patterns of deleterious variation in human population isolates. Second, I will test how long-term small population size impacts genetic diversity and the distribution of deleterious variation in Ethiopian wolves. Lastly, I will use dogs as a model system to test how recent demography and artificial selection affect the distribution of deleterious variation and architecture of complex traits in breed dogs. Taken together, my research will allow us to develop a more complete picture of how demography shapes patterns of deleterious variation.

The dissertation of Jazlyn Ambry Mooney is approved.

Rita Cantor

Sriram Sankararaman

Nelson Freimer

Kirk Lohmueller, Committee Chair

University of California, Los Angeles

2020

DEDICATION

This dissertation is dedicated to my family and friends. Their constant love and support has encouraged and nurtured me. As well as my dear friends Taylor Brown and Sergio Méndez-Aguirre, two scientists whose lights shined so brightly and were gone too soon.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I want to thank my personal support team: my PhD sisters Malika Freund and Chantle Edillor; my dear friends Andrew Lopez, Jesse Garcia, Jessica Ochoa, Adriana Arneson, Douglas Arneson, Ryan Cho, Matilde Miranda, Jessica Ong, and Hannah Anderson; my parents Sonya and Ror; my grandparents Dr. Gloria Madrid, Luis, Donna, and Bob, and my brothers Keegan and Caith. This work would not have been possible, or meaningful, if not for you.

Next, I would like to acknowledge my advisor Kirk Lohmueller, whose support, patient teaching, encouragement, and faith in me has pushed me forward every year, thank you. I would also like to thank Jeffrey Long, my undergraduate advisor who encouraged me to try population genetics, believed in me before I did and welcomed me into his lab, and gave me more valuable knowledge than I ever realized. The other members of my committee Nelson Freimer, Sriram Sankararaman, and Rita Cantor who have given me their insightful feedback and words of support. I have enjoyed working with them and look forward to continuing our relationship.

I next wish to acknowledge my collaborators, lab mates, and colleagues who have helped guide me through my research and this PhD. In particular, from the Lohmueller Lab past and present: Arun Durvasula, Clare Marsden, Annabel Beichman, Tanya Phung, Eduardo Amorim, Abigail Yohannes, Christian Huber, Jacqueline Robinson, Bernard Kim, Chris Kyriazis, Xinjun Zhang, Tina del Carpio, Ying Zhen, and Diego Ortega-Del Vecchyo; from UCLA: Janet Sinsheimer, Bogdan Pasanuic, and Susan Service; and beyond: Charleston Chiang and Nicholas Mancuso.

Last, but not least, I want to acknowledge Taylor Brown and Sergio Méndez-Aguirre. Thank you for sharing your lives with me. I would not be the scientist or person I am without you.

VITA

EDUCATION

2020	PhD Candidate, Human Genetics University of California, Los Angeles | Los Angeles, CA
2014	Bachelor of Science, Biology University of New Mexico | Albuquerque, NM
2014	Bachelor of Science, Anthropology University of New Mexico | Albuquerque, NM

PEER-REVIEWED PUBLICATIONS

Mooney JA*, Huber CD*, Service S, Sul JH, Marsden CD, Zhang Z, et al. Understanding the hidden complexity of Latin American population isolates. American Journal of Human Genetics. 2018;103(5):707–26. (* indicates equal contribution).

Sura, S. A., Smith, L. L., Ambrose, M. R., Amorim, C. E. G., Beichman, A. C., Gomez, A. C., … **Mooney, J.** (2019). Ten simple rules for giving an effective academic job talk. *PLoS Computational Biology*, *15*(7).

Chapter 1: Introduction

For many years, population geneticists have been interested in the persistence of deleterious variation, and how the demographic history of a population impacts its distribution and prevalence. Indeed, it is known that the demographic history of a population affects the efficiency of selection, which can allow some deleterious variants to avoid being eliminated from the genome immediately by natural selection. Since some deleterious variation may contribute to complex disease, it is imperative that I understand why and how deleterious variation remains within a population. For my dissertation, I will focus on three projects that aim to understand the relationship between deleterious variation and demographic history.

In **Chapter 2**, I focus on understanding how demography affects the distribution of deleterious variants in a young admixed population isolate. While most studies have focused on populations where the founders all came from the same ancestral population, I investigated genomic diversity of recently admixed population isolates from Costa Rica and Colombia and compared their diversity to a well-studied population isolate, the Finnish. I examine whole genome sequence data from 449 individuals, ascertained as families to build multi-generational pedigrees, with a mean sequencing depth of coverage of approximately 36X, alongside a handful of 1000 Genomes populations. I find that the genetic diversity and genomic background of population isolates varies immensely. Specifically, admixture increased genetic diversity of the Colombian and Costa Rican isolates; but I still observe an enrichment of IBD segments and a larger burden of ROH in the Latin American isolates relative to the Finnish, suggesting that there is no single genetic signature shared across all population isolates. Further, I determine that long runs of homozygosity are generated by recent consanguinity, and that recent consanguinity has led to an enrichment of derived deleterious variation in the Costa Rican and Colombian isolates. Thus, this work also provides a mechanism for how recent consanguinity can reduce fitness in natural populations. A version of this work has been published in *The American Journal of Human Genetics* as:

Mooney JA*****, Huber CD*, Service S, Sul JH, Marsden CD, Zhang Z, et al. Understanding the hidden complexity of Latin American population isolates. American Journal of Human Genetics. 2018;103(5):707–26. (* indicates equal contribution).

In **Chapter 3**, I evaluate genetic diversity and measure the genomic impact of long-term small population size in Ethiopian wolves. The Ethiopian wolf (EW) is the rarest canid in the world and has experienced a long-term population decline as well as very recent crashes due to disease. Thus, quantifying the current genetic variation in this population is crucial for conservation efforts and provides a model for assessing how deleterious variation accumulates in small populations. To this end, 10 Ethiopian wolves from the Bale Mountains were sequenced to high-coverage (40x). I compared the EW to four populations of breed dogs (N=41) and two populations of wolves (N=24). I observe low diversity across the genome in the EW compared to both dog and wolf populations. In addition to exhibiting low diversity, the EW carry more putatively deleterious variants relative to both dogs and wolves as well. I infer the demographic history of the EW, which suggests that this population has experienced a series of both ancient and recent contractions resulting in a current effective population size of approximately 100 individuals. Lastly, I find evidence of adaptation to their high-altitude environment through natural selection at CREBBP.

In **Chapter 4**, I test how recent demography and artificial selection affect the distribution of deleterious variation in breed dogs. I use dogs as a model to gain insight about the genetic architecture of shared traits in humans because dog demography has simplified the genetic architecture of complex traits and allowed deleterious variation to persist at appreciable levels in their genome. I compiled a data set that features genetic variation data and clinical and morphological trait information from over 4,300 dogs and wolves, to test various hypotheses. First, I test whether dog breeds that are closely related to each other, reflected by large amounts of runs of homozygosity (ROH) and identity-by-decent (IBD), have an increased incidence of disease relative to those breeds with less ROH and IBD segments. Second, I examine the relationship between ROH burden and case-control-status for eight traits. Lymphoma case-status is positively associated with the amount of the genome within an ROH and, on average, cases carry more ROH than controls. Conversely, ROH appear to show a protective effect against developing portosystemic vascular anomalies across multiple breed dogs. Third, I characterize the relationship between genetic disease alleles reported in the Online Mendelian Inheritance in Animals (OMIA) database and IBD and ROH. I find that the number of causal variants identified correlates with the popularity of that breed rather than the ROH or IBD burden, suggesting an ascertainment bias in OMIA. Lastly, I use the distribution of ROH to identify regions of the genome with depletions of ROH as potential hotspots for inbreeding depression and find multiple loci where ROH are never observed. My results suggest that inbreeding has played a large role in shaping genetic and phenotypic variation in dogs, and that there remains an excess of understudied breeds that have the potential to reveal new disease-causing variation.

In **Chapter 5**, I summarize all my results and synthesize them through the lens of demography, and how it can influence the prevalence of deleterious variation in the genome.

Chapter 2: Understanding the Hidden Complexity of Latin American Population Isolates

2.1 Introduction

The use of population isolates to map Mendelian and complex diseases has been a key feature of medical genomics. In addition to experiencing the bottleneck involved with the migration out of Africa, some populations underwent subsequent bottlenecks and remained in relative seclusion afterward. These populations formed present-day isolates[1]. The genomes of population isolates are thought to exhibit several hallmark features of genetic variation. Due to bottlenecks associated with their founding, it is thought that isolates should carry lower levels of genetic diversity and lower haplotype diversity than closely related non-isolated populations. Drift experienced by isolates is magnified by small population size, which generates more linkage disequilibrium (LD) than in non-isolated populations. In addition to increased LD, individuals from isolated populations tend to share more regions of the genome identical by descent (IBD) due to small population size. Further, due to the isolation after founding and recent mating practices, isolates may have larger regions of the genome found in runs of homozygosity (ROHs) due to recent inbreeding. Lastly, bottlenecks and inbreeding should impact patterns of deleterious variation[2–4]. Consequently, one would predict that individuals from isolates will have fewer segregating sites, and the remaining deleterious variants will be segregating at a higher frequency[5]. Indeed, genomic studies over the last decade have documented several of these signatures[6–8]. However, it is known that not all isolates share the same demographic history. Therefore, it is essential that we understand how the factors shaping genetic variation in a population are influenced by the unique demographic history of the population.

One archetypal human population isolate that has been extensively studied is the Finnish [7,9–11]. Finland was populated through two separate major migrations. Briefly, a small number of founders, relative isolation, serial bottlenecks, and recent expansion in Finland has allowed drift to play a large role in shaping the gene pool of this population[11]. The aforementioned demographic history of Finland has led to an increase in the prevalence of rare heritable Mendelian diseases, which has made this population particularly fruitful for identifying disease associated variants[10,12]. Most of the studies in Finland employed LD mapping in affected families and well curated genealogical records to identify causal and candidate variants[10]. More recently, it has been possible to apply population-based linkage analyses to identify disease associated variants as an alternative to genome-wide association studies (GWAS)[13] due to the availability of whole genome sequence data in conjunction with extensive electronic health records.

A number of studies have shown that power to detect causal variants can be improved by studying population isolates other than the Finnish[8,14–16]. For example, the Greenlandic Inuit experienced an extreme bottleneck which caused a depletion of rare variants and segregating sites in their genome[16].The remaining segregating variants are maintained at higher allele frequencies and a larger proportion of these SNPs are deleterious when compared to non-isolated populations. Another study of South Asian populations showed similar results. Specifically, South Asian populations have experienced more severe founder effects than the Finnish[15], thus creating an excess of rare alleles associated with recessive disease. A study of European population isolates compared the isolates with the closest non-isolated population from similar geographic regions[8], and found that the total number of segregating sites was depleted across all isolates relative to the comparison non-isolate. Of the sites that were segregating in isolates, between ~30,000- 122,000 sites existed at an appreciable frequency (minor allele frequency (MAF) > 5.6%), while remaining rare (MAF < 1.4%) in all the non-isolate population samples[8]. The authors surmised that these common and low-frequency variants could be useful in GWAS for novel associations, as they included SNPs that had been previously associated with cardio-metabolic traits[8,17].

While there have been many studies of genetic variation in population isolates, the studies described above have focused on populations where the founders all came from the same ancestral population. The founders of Latin American population isolates have come from distinct continental populations. We sampled individuals from mountainous regions of Costa Rica and Colombia where geographic barriers resulted in populations remaining isolated since their founding in the 16th and 17th centuries, until the mid-20th century[18]. Both groups share a similar demographic history, having originated primarily from admixture between a few hundred European males and Amerindian females, with a limited contribution from African founders. After the founding event, both populations experienced a subsequent bottleneck and then a recent expansion, within the last 300 years, the expansion increased the population size over 1000-fold since the initial founding event[18]. The effect that admixture has had on overall patterns of genetic variation in isolates remains elusive, and it remains unclear whether these populations share the typical genomic signatures seen in population isolates. While the small founding population size could reduce diversity, because the Costa Rican and Colombian isolates were founded from multiple diverse populations, they could potentially have increased in diversity relative to other population isolates. Lastly, the impact of admixture on deleterious variation also remains unclear.

To better understand patterns of genetic variation in admixed isolated populations, we compared the Colombian and Costa Rican population isolates to a benchmark isolate, the Finnish, as well as other 1000 Genomes Project populations[19]. We observe that relative to the Finnish, Latin American isolates have increased genetic diversity but an excess of IBD segments. Moreover, we detect an increase in the proportion of an individual's genome that falls within a long ROH in Latin American isolates relative to all other sampled populations and an enrichment of deleterious variation within these long ROH. Demographic simulations and analysis of extended pedigrees indicate that the enrichment of long ROH is primarily a consequence of recent inbreeding in Latin American isolates. Next, we examine the relationship between the proportion of European, Native American, and African ancestry and the amount of the genome within an ROH, as well as the relationship to an individual's pedigree inbreeding coefficient. Further, we examine demography across both recent and ancient timescales in these isolates. Our work sheds light on how the distinct demographic histories of population isolates affect both genetic diversity and the distribution of deleterious variation across the genome.

2.2 Methods and Material

Pedigree Data for Costa Rican and Colombian Individuals

Our study included 10 Costa Rican (CR) and 12 Colombian (CO) multi-generational pedigrees ascertained to include individuals affected by Bipolar Disorder 1. The sampled families are clumped geographically to some degree, and it is worth noting that the Central Valley of Costa Rica and Antioquia are population isolates but each population contains several million people. In Costa Rica there is only one psychiatric hospital, and the Antioquia Department of Colombia has few hospitals, thus most cases were originally identified in the largest hospital in a city of more than 3 million people. More extensive details about the curation of pedigree data and clinical assessments of diagnosis can be found in Fears et al.[20].

Identifying Unrelated Individuals

We defined unrelated individuals as those who are at most third-degree relatives. We chose this threshold of relatedness because the families from CR and CO are known to be cryptically related. We used KING[21] to identify 30 unrelated individuals from CR and CO. 24 of the 30 unrelated individuals in the CO are founders in the pedigree and 15 of the 30 unrelated individuals in the CR are founders, and each family sampled is represented by at least one individual, but some families had as many as seven individuals. The algorithm implemented in KING estimates familial relationships by modeling the genetic distance between a pair of individuals as a function of allele frequency and kinship coefficient, assuming that SNPs are in Hardy-Weinberg equilibrium.

Further, we also used PC-AiR[22] and PC-Relate[23] to estimate relatedness as these two methods are robust to population structure, cryptic relatedness, and admixture. We found that 28 of the 30 CO unrelated individuals and 26 of the 30 CR unrelated individuals were contained in the list of unrelated individuals from PC-AiR[22]. Complete overlap was not expected because we retained third-degree relatives when using KING to allow for cryptic relatedness of families sampled from Costa Rica and Colombia due to their demographic history.

Lastly, we used KING to identify 30 unrelated individuals from the following 1000 Genomes Project[19] populations: Yoruba (YRI), CEPH-European (CEU), Finnish (FIN), Colombian (CLM), Peruvian (PEL), Puerto Rican (PUR), and Mexican from Los Angeles (MXL). We used these 30 unrelated individuals per population for all analyses unless otherwise stated (**Figure S1)**.

Genotype Data Processing

We generated a joint variant call file (VCF) containing single nucleotide polymorphisms (SNPs) from two separate data sets. The first data set contained 210 whole genome sequences sampled from the aforementioned 1000 Genomes Project populations[19]. The second data set contained 449 whole genome sequences from Costa Rican and Colombian individuals. Variants in the second data set were called following the GATK best practices pipeline[24] with the HaplotypeCaller of GATK. All multi-allelic SNVs and variants that failed Variant Quality Score Recalibration were removed. Genotypes with genotype quality score ≤ 20 were set to missing. Further quality control on variants was performed using a logistic regression model that was trained to predict the probability of each variant having good or poor sequencing quality. Individuals with poor sequencing quality and possible sample mix-ups were removed, and all sequenced individuals had high genotype concordance rate between whole genome sequences and genotypes from microarray data. All sequenced individuals had consistency between the reported sex and sex determined from X chromosome as well as between empirical estimates of kinship and theoretical estimates. More information on sequencing and quality control procedures is discussed in Sul et al. 2018 (unpublished)[25].

We used the following protocol to merge these two datasets. First, we used guidelines from the 1000 Genomes Project strict mask to filter the Costa Rican and Colombian VCFs as well as the 1000 Genomes Project VCFs. Then, we used GATK to remove sites from both sets of VCFs that were not bi-allelic SNPs or monomorphic. Next, we merged the 1000 Genomes Project VCFs with the Costa Rican and Colombian VCFs into a single joint-VCF for each chromosome. We only used autosomes for our analyses. Lastly, we filtered the merged joint-VCF to only contain sites that were present in at least 90% of individuals. There were a total of 57,597,196 SNPs and 1,891,453,144 monomorphic sites in the final data set. We ensured that the merged data sets were comparable by examining the number of derived putatively neutral alleles across the 30 unrelated individuals in all sampled populations, and found few differences between populations, which is consistent with theory[5] (**Figure S2**).

Calculating Genetic Diversity

We computed two measures of genetic diversity from sites called across all 30 unrelated individuals from each population: pi (π) and Watterson's Theta (θ_w). The average number of pairwise differences per site (π) was calculated across the genome as:

$$\pi = \frac{n}{n-1}\frac{\sum_{i=1}^{L} 2p_i(1-p_i)}{L},$$

where *n* is the total number of chromosomes sampled, *p* is the frequency of a given allele, and *L* is the length in base pairs of the sampled region. θ_w was computed by counting the number of segregating sites and dividing by Watterson's constant, or the *n*-1 harmonic number[26].

Site Frequency Spectrum (SFS)

Site frequency spectra were generated using the 30 unrelated individuals from each population. SNPs with missing data were removed from these analyses. There was a total of 16 SNPs out of the 57,597,196 SNPs that were removed due to missing data.

Linkage Disequilibrium Decay

We calculated LD between pairs of SNPs for all unrelated individuals. First, we applied a filter to remove SNPs that were not at a frequency of at least 10% across all populations. Next, pairwise r^2 values were calculated using VCFTools[27]. SNP pairs were then binned according to physical distance (bp) between each other and r^2 was averaged within each bin.

Identifying Identity by Descent Segments

To detect regions of the genome that have shared IBD segments between pairs of individuals, we first removed singleton SNPs in each population since singletons are not informative about IBD. Then, we called IBD segments using IBDSeq[28]. IBDSeq is a likelihood-based method that is designed to detect IBD segments in unphased sequence data. We chose to use IBDSeq because other methods that require computational phasing could be biased when applied to Latin American population isolates, as they do not have a publicly available reference population to aid in phasing. We compared IBDSeq to two well-known methods Beagle[29] and GERMLINE[30] to determine whether it was feasible to use IBDSeq on an admixed population (**Figure S3**). Data for Beagle and GERMLINE was phased beforehand with SHAPEIT[31] (see Web Resources) using the 1000 Genomes as the reference panel. Beagle produced the shortest IBD segments while GERMLINE produced the longest IBD segments. IBDSeq produced segments with a length distribution similar to what we observed in Beagle, though the average segment length was slightly larger, which we expected given that IBDSeq was created to call longer segments that would have previously been broken up when using Beagle for phasing. We used the default parameters for IBDSeq.

Next, we filtered the pooled IBD segments to remove artifacts. First, we calculated the physical distance spanned by each IBD segment. Then, we totaled the number of SNPs that fell within each segment. We observed an appreciable number of IBD segments that were extremely long but sparsely covered by SNPs (**Figure S4**). IBD segments were removed if the proportion of the IBD segment covered by SNPs was not within one standard deviation (0.0043) of the mean proportion covered (0.0221) across all IBD segments (**Figure S4**). Strong deviations from the mean could indicate that the IBD segment spans a region of the genome with low mappability where we are only calling the SNPs at the outer ends of the segment. Therefore, the true segment length might be much shorter than what is being calculated by IBDSeq. Lastly, we converted from physical distance to genetic distance using the deCODE genetic map[32]. A file that contains all the IBD segments (unfiltered) alongside code used to filter can be found on GitHub (see Web Resources).

Enrichment analyses of IBD segments

To determine whether certain populations contain more IBD segments than others, we followed the IBD score procedure outlined by Nakatsuka and collegues[15]. A population's IBD score was calculated by computing the total length of all IBD segments between 3 and 20 cM. The score difference is the difference between the query population's IBD score and the Finnish IBD score. The score ratio is the ratio of each population's IBD score relative to the Finnish IBD score. The significance of enrichment relative to the Finnish was evaluated using a permutation test for each population, where IBD segment length was held fixed and labels of the two populations were permuted. We recalculated the score on a total of 10,000 permutations to generate a null-distribution of scores for each isolate. The code can be found on GitHub (see Web Resources).

Estimating Effective Population Size

We used the output files from IBDSeq to estimate the recent effective population size through time from the 30 unrelated individuals from each sampled population. We estimated effective population size by using the default settings in IBDNe[33]. We set the minimal IBD segment length equal to 2cM since that is the suggested setting when using sequence data.

Identifying Runs of Homozygosity

Runs of homozygosity were identified for each individual using VCFTools, which implements the procedure from Auton et al. 2009[34]. Next, we examined the number of callable sites that lie within each ROH. We found that there was a bi-modal distribution of coverage for ROH, where some ROH appeared to contain almost no callable sites, while others had much higher coverage. We only kept ROHs that were at least 2Mb in length, which we called long runs of homozygosity, and were at least 60% covered by callable sites. (**Figure S5)**. A file that contains the final ROHs can be found on GitHub (see Web Resources).

Calculating Inbreeding Coefficients

SNP-based inbreeding coefficients were calculated using VCFTools[27]. VCFTools calculates the inbreeding coefficient *F* per individual using the equation $F = \frac{O-E}{N-E}$, where *O* is the observed number of homozygotes, *E* is the expected number of homozygotes (given population allele frequency), and *N* is the total number of genotyped loci.

Pedigree-based inbreeding coefficients were computed using the R package kinship2[35].

Demographic Simulations

In order to investigate how aspects of the population history affect current day genetic diversity in Latin American isolated populations, we simulated genetic variation data using the forward simulation software SLiM 4.2.2[36]. We simulated a sequence length of 10Mb under uniform recombination rate of 1×10^{-8} crossing-over events per chromosome per base position per generation and under a mutation rate of 1.5×10^{-8} mutations per chromosome per base position per generation. Every simulation contained intergenic, intronic, and exonic regions, but only nonsynonymous new mutations experienced natural selection in accordance with the distribution of selection coefficients estimated in Kim et al. 2017[37]. Within coding sequences, we set nonsynonymous and synonymous mutations to occur at a ratio of 2.31:1[37,38]. The chromosomal structure of each simulation was randomly generated, following the specification in the SLiM 4.2.2 manual (7.3), which is modeled after the distribution of intron and exon lengths in Deutsch and Long[39].

We assumed an effective population size in the ancestral African population of 10,000 individuals, and a reduction in size to 2,000 individuals, starting 50,000 years ago, reflecting the colonization of the European, Asian, and American continents. The population then recovers to a size of 10,000 individuals 5,000 years ago. The colonization bottleneck is assumed to occur 500 years ago by an admixture event with a European population (70% admixture proportion) and is followed by an immediate reduction in population size to 1,000 individuals. The recent expansion in population size is modeled by an increase in population size to 10,000 individuals 200 years ago. We simulated data with recent inbreeding and without recent inbreeding. In the former case, inbreeding started at the time of the European colonization 500 years ago and continues until the present. Inbreeding is implemented with the “mateChoice” function in SLiM. Because SLiM’s pedigree track function is only valid for at most second-degree related individuals, 50% of the time, mating occurs randomly. However, in the remaining cases, mating occurs between close relatives with a relatedness coefficient bigger than 0.25. This produces levels of consanguinity similar to those seen empirically as measured by *F* (see Results). We also tested if such high observed values of *F* can be explained by random mating during an extreme bottleneck with a bottleneck to 100 individuals, and a bottleneck to 64 individuals, during colonization 500 to 200 years ago. To increase the speed of the simulations, we reduced mutation rate by a factor of 5, and verified the results of the simulations with theoretical predictions of the relationship between *F* and population size over time[40]. Finally, we sampled a total of 60 random individuals and calculated summary statistics on the sample data. The simulation script can be found on GitHub (see Web Resources).

Annotation of Variants

The ancestral allele was determined using the 6-primate EPO alignment (see Web Resources) and we restricted to only those sites called with the highest confidence. After filtering, 54,049,081 SNPs remained. Subsequently, exonic SNPs were annotated using the SeattleSeq Annotation website (see Web Resources). A total of 693,301 SNPs were annotated as either nonsynonymous or synonymous. We further classified these sites as either putatively neutral or deleterious using Genomic Evolutionary Rate Profiling (GERP) scores[41]. GERP scores are generated using a multiple-sequence alignment of the hg19 reference to 33 other mammalian species. When calculating the rejected substitutions (RS) score, which we will refer to as the GERP score, the hg19 reference genome is removed to eliminate confounding due to deleterious derived alleles. A GERP score less than two was considered as putatively neutral and a GERP score greater than 4 was considered as putatively deleterious for the 404,302 classified SNPs.

Counting Deleterious Variants

We used three different statistics to count the number of deleterious mutations per individual. First, we tabulated the number of deleterious variants (the number of heterozygous plus the number homozygous derived genotypes). Second, we counted the total number of derived deleterious alleles (the number of heterozygous genotypes plus twice the number of homozygous derived genotypes). Third, we computed the total number of derived deleterious homozygous genotypes. A table that contains the counts of all deleterious and neutral variants can be found on GitHub (see Web Resources).

Testing for an enrichment of deleterious variation in ROHs

We were interested in whether there is an enrichment of nonsynonymous or loss-of-function mutations in ROH over non-ROH regions for the three different ways of counting deleterious variants outlined above. To account for differences in neutral variation, we standardized by synonymous variation, which is assumed to be neutral. Then, we calculated the ratio of nonsynonymous over synonymous variation in ROH regions divided by the ratio of nonsynonymous over synonymous variation outside of ROH. We computed significance using a permutation test, where the position of each SNP and its annotation as synonymous versus nonsynonymous was fixed and the positions of the vector of ROH annotations were randomly placed throughout the genome. Thus, the frequency distribution of synonymous and nonsynonymous SNPs, as well as the total amount of ROH and non-ROH annotations, is kept constant when compared to the unpermuted data. We recalculated the ratio for a total of 10,000 permutations to form a null-distribution of ratios and then computed significance.

Calculating Ancestry Proportions

We estimated genome-wide ancestry proportions in members of the CR and CO pedigrees using LAMP[42]. We generated ancestry estimates for all 838 pedigree members with SNP array genotype data, detailed information on the SNP array data can be found in Pagani et al.[43]. The ancestral reference populations were the CEU (*n*=112) and YRI (*n*=113) from HapMap[44,45], as well as 52 Native American samples from Central or South America. The Native American samples are the Chibchan-speaking subset of those used in Reich et al.[46], selected to originate from geographical regions relevant to CR/CO and to have virtually no European or African admixture (European and African ancestry < 0.00025). The allele frequencies were calculated for each reference population and were used as input files for LAMP alongside the following configuration parameters: offset=0.2, recombrate=1e^{-8}, generations=20, alpha=0.24,0.72,0.04, ldcutoff=0.1. Then, we computed global ancestry estimates from the LAMP output file.

Ancestry proportions in Table 1 for 1000 Genomes and Latin American populations were estimated using ADMIXTURE[47]. The analysis for the 1000 Genomes populations used 665,105 LD-pruned SNPs, an unsupervised learning model, and the number of source populations was set to K=3 (**Table 1**). The analysis for the Latin American isolates used a supervised learning model with K=3 source populations, composed of the European, African, and Native American populations mentioned above and 57,180 LD-pruned SNPs.

Accounting for Relatedness

We tested for correlations among several quantities computed for each individual in the Latin American population isolates. Because some of these individuals are closely related, the data points in our linear regression are no longer independent. We used the R-package GenABEL[48] to incorporate kinship when performing statistical tests for our correlations. We used the polygenic_hglm() function where the *formula* input was the equation for our linear model of interest and the *kinship.matrix* input was a kinship matrix computed from our pedigree computed using kinship2[35]. Our input took the following form: (F_{PED} ~ Length of genome in ROH, kin = kinshipMatrix, data = df). We also computed p-values from a genetic relatedness matrix (GRM) created using PC-AiR[22] and PC-Relate[23], both sets of p-values can be found in **Table S1**.

2.3 Results

Genetic Variation in Population Isolates

We first compared levels of genetic diversity in a sample of 30 unrelated individuals across the 1000 Genomes populations[19] and the CO and CR isolates. We split the genome into several different genomic regions and in each region summarized genetic variation using both the average number of pairwise differences (π) and Watterson's theta (θ_w) (**Figure 1A and B**). Overall, we found differences in diversity across the functional categories of sequence studied in all populations, with coding regions exhibiting the lowest diversity and intergenic regions the highest. These patterns are consistent with the role of purifying selection affecting coding diversity[37]. However, if we look genome-wide or focus on intronic regions, we see intermediate levels of diversity (**Table S2 and Table S3)**. We suspect that these categories are more strongly influenced by linked selection[49–51].

As we are interested in the role of demography in shaping genetic diversity, we focused on comparisons of intergenic levels of diversity as those are most likely to be neutrally evolving (**Figure 1A and B**). Overall, the YRI had the highest level of diversity ($\pi \approx 0.0010$; $\theta_w \approx 0.0012$) (**Table S2 and Table S3**). The European populations (CEU and FIN) had lower levels of diversity. The CEU and FIN had similar levels of π (approximately 0.0004), despite the FIN being considered an isolated population. However, the FIN had reduced numbers of SNPs as reflected by lower mean values of θ_w (CEU ≈ 0.00075 & FIN ≈ 0.00072). The CO and CR had levels of diversity comparable to that of several other Latin American populations in the 1000 Genomes Project (CLM and MXL). We found no clear pattern of the population isolates (FIN, CO, CR) having lower diversity than their most similar non-isolated population. Instead, diversity levels tended to be higher across all the sampled Latin American populations (CLM, CO, CR, MXL, and PUR) when compared to the European populations. One exception to this pattern is the PEL population, who had the lowest neutral levels of diversity ($\pi \approx 0.0007$; $\theta_w \approx 0.0007$).

Next, we examined the proportional site frequency spectrum (SFS; **Figure 1C & Figure S6**). Latin American populations had the highest proportion of singletons, as seen previously[52]. The CO and CR had similar proportions of singletons when compared to other 1000 Genomes Project Latin American populations. Conversely, the FIN had the lowest proportion of singletons in comparison to all sampled populations. The depletion of singletons relative to common variation supports the presence of a stronger founder effect during the FIN population history[11].

We also examined patterns of linkage disequilibrium (LD), since LD is affected by population size and recent bottlenecks[53,54]. **Figure 1D** shows the mean decay of r^2 with physical distance over 2Mb intervals across the genome in each population. We found that the YRI had the lowest levels of LD for each bin of physical distance, and the PEL formed the upper bound of the LD decay curves. The remaining Latin American populations (PUR, MXL, CLM, CO, CR) clustered together, close to the YRI, while the CEU and FIN are shifted toward higher values, like those seen in the PEL.

The FIN were previously shown to have more extensive haplotype blocks in their genome in comparison to the Latin American isolates[6]. In line with these findings, we observed faster LD decay in the Latin American isolates relative to the FIN. When considering pairs of SNPs 150kb or more apart, rates of LD decay become quite similar across all the sampled populations. Analogous to other diversity statistics, LD in the CO and CR closely resembled those of non-isolated Latin American populations. Once again, we found there is no clear pattern of having lower diversity or more LD that holds across all the population isolates (FIN, CO, CR) when compared to their most similar non-isolated population.

Latin American isolates carry more IBD segments than the Finnish

Next, we used IBD sharing between pairs of individuals to gain insight about more recent demographic events within populations (**Figure 2**). We quantified the amount of IBD within each population by computing an IBD score. Each population's IBD score was calculated by totaling the length of IBD segments between 3cM and 20cM. We expressed IBD scores for each population as the ratio of the IBD score for a given population relative to the IBD score in the FIN (**Figure 2A**). We also tabulated the total count of IBD segments for each population. The CEU showed the lowest number of both called IBD segments and the lowest IBD score relative to the FIN (p-value = 0.0001). Latin American populations formed the upper bounds of both total IBD segments called and IBD enrichment scores (**Figure 2A**). The PUR had the largest number of IBD segments (1402) and had a 2.1-fold increase in IBD score relative to the FIN (p-value < $1x10^{-4}$). The CO and CR isolates had a 1.8-fold and 2-fold increase in their IBD scores relative to the FIN (p-value < $1x10^{-4}$), as well as carrying more IBD segments than the FIN (**Figure 2B and 2C**). However, there were some Latin American populations that exhibited depletions in both IBD segments and IBD scores relative to the FIN. The MXL and PEL have the lowest number of IBD segments for the Latin American populations. Previous work has shown that a larger effective population size in admixed populations likely drove the depletion of IBD segments in these two Latin American populations[55].

Inferring the Demographic History of Latin American Isolates

We next leveraged the patterns of IBD described above to estimate the effective population size (N_e) through time using IBDNe[33] on the 30 unrelated individuals from each population (**Figure 3**). The use of only 30 unrelated individuals caused limitations for accurate estimation of N_e (see Discussion), but the general population size trajectory is likely to be robust to the number of individuals used. First, we found that recent demography differs vastly between the European populations (FIN and CEU). In general, CEU experienced population expansions over much of their demographic history. It was only in the most recent generations that they experienced a decrease in N_e. The FIN, on the other hand, have experienced a long population decline since their founding, approximately 4000 years ago, followed by a recent population expansion.

When analyzing the Latin American isolates, we detected a recent bottleneck, approximately 500 years ago (**Figure 3**). This bottleneck could correspond to the recorded bottleneck that followed the founding of these populations, and it appears to be much shorter and more severe than the bottleneck seen in the FIN. The strength and duration of bottlenecks varied across each of the Latin American populations. For example, we observed a more severe bottleneck in the CR, CO, CLM, and PUR than in PEL or MXL. However, we detected a subsequent period of growth across all populations following the bottleneck. The rate of growth differed across each population, and the PEL appeared to be growing at a much more rapid rate than any of the other Latin American populations.

Exploring Recent Consanguinity

Isolated populations may have experienced recent consanguinity. To test for this, we began by examining SNP-based inbreeding coefficients (F_{SNP}) (**Figure S7**). YRI individuals had the lowest median inbreeding coefficients (-0.0001) and the CO and CR isolates had the highest median inbreeding coefficients (0.0087 & 0.0086, respectively). Further, the CO and CR also had the highest maximum F_{SNP} values in the entire sample of unrelated individuals from any population (**Figure S7**). Median levels of F_{SNP} in the CEU (-0.0004) suggested that they are more homozygous than the FIN (-0.0007), which may be a result of how 1000 Genomes samples were selected. The PEL had the largest variance in F_{SNP} across any of the sampled populations.

Next, we examined patterns of long runs (>2Mb, see Methods) of homozygosity (ROH), since ROH have been linked to recent consanguinity[56–60]. The YRI and CEU had the lowest amount of their genome contained within an ROH (**Figure 4A**). The FIN had higher median (median = 11 Mb & s.d. = 6.3 Mb) amounts of their genome within an ROH in comparison to the CEU (median

= 2.4 Mb & s.d. = 2.1 Mb). Latin American isolates had the highest median amount of the genome contained within an ROH. Specifically, the CR had the highest median at 21.7 Mb (s.d. = 40.9 Mb). Further, the Latin American isolates also had the greatest variance in the amount of the genome contained within an ROH. For example, one of the CO individuals had approximately 230 Mb of their genome contained in long ROHs.

As expected, we found that the amount of the genome contained in a long ROH strongly correlated with an individual's F_{SNP} (CO: $R^2 = 0.8060$, p –value = 1.1×10^{-11}; CR: $R^2 = 0.7740$, p-value = 9.5×10^{-11}; FIN: $R^2 = 0.1288$, p-value = 0.03) (**Figure 4B-4D**). Indeed, individuals with higher values of F_{SNP} tended to have more of their genome within an ROH. Further, the individual with the highest F_{SNP} (0.133) also had the largest amount of their genome in long ROHs (230Mb).

The total number of ROH segments per individual followed a similar pattern as the total amount of genome within an ROH (**Figure S8**). For example, in populations with low values of F_{SNP}, ROH segments were not frequent. One YRI individual and three CEU individuals carried a ROH >4Mb, whereas more than 50% of CO and CR individuals carried an ROH >4Mb. Additionally, the longest ROHs identified (>20MB) only occurred in Latin American populations, who have the largest values of F_{SNP} (**Figure S8**). Importantly, the FIN individuals had significantly fewer ROH segments than the CO and CR, and most individuals had an F_{SNP} close to 0; while the Latin American isolates had the most ROH in comparison to any other sampled population, as well as the largest values of F_{SNP} (**Figure 4**).

Determining the Mechanisms that Generate Runs of Homozygosity

In principle, ROHs can be generated either by recent consanguinity over the last few generations, or by older historical processes, such as bottlenecks[56,58,60–63]. Based on both historical data[18] and inference from IBDNe analyses, Latin American population isolates show evidence of recent population bottlenecks. Therefore, we used two complementary strategies to test whether recent consanguinity or bottlenecks drove the observed increase in ROHs in the Latin American isolates. First, we used the extensive pedigree data for 449 sequenced individuals to calculate a pedigree inbreeding coefficient (F_{PED}). Most individuals had a F_{PED} of 0 (**Figure 5**). However, there were several individuals with values of F_{PED} as high as 0.07 in CR and 0.06 in CO. We observed a significant correlation between F_{SNP} and F_{PED} (R^2=0.1520 and p-value < 2 x10^{-16}), even after accounting for the non-independence of individuals based on their kinship (**Figure 5A**; see Methods). These correlations suggest that the recent consanguinity captured within the last few generations in the pedigree was a relevant factor to increase ROHs in the CO and CR populations. However, once we remove the four most influential individuals, the correlation between F_{SNP} and F_{PED} is no longer significant. These four individuals also account for approximately 7.5% of individuals with F_{PED} >0, so the reduction in sample size could also explain some component of the reduction in signal. F_{SNP} was a substantially better predictor of the amount of an individual's genome that falls within a ROH (R^2 = 0.7540 and p-value < 2 x 10^{-16}), than F_{PED} (R^2 = 0.2180 and p-value < 2 x 10^{-16}) (**Figure 5B and C**) likely due to the fact that F_{SNP} captured distant background relatedness within the population as well as the realized level of consanguinity, rather than the expected value[64]. Further, because the pedigrees were ascertained and analyzed separately, connections between pedigrees were not accounted for in F_{PED}, but were likely captured by $F_{SNP.}$

As a second approach to determine the mechanism driving the increase in ROHs in the CO and CR populations, we conducted forward in time demographic simulations. We simulated a 10Mb region under a demographic model that reflected changes in effective population size during the human expansion across the European, Asian and American continents, as well as the more recent bottleneck during the Spanish colonization about 500 years ago (**Figure 5D**; see Methods). Consanguineous nonrandom mating in the population was modeled to begin 500 years ago, leading to a mean value of F_{ROH} of about 0.075. This level of inbreeding matches the level of inbreeding in some of the CO and CR individuals, based on calculations using pedigree data.

Next, we investigated how severe the bottleneck caused by the Spanish colonization would have needed to be to generate such high levels of ROH, when assuming random mating instead of consanguineous mating. We found that a recent population bottleneck to 1000 individuals, as suggested by historical data for the Central Valley population of CR[65], is not capable of generating the large amounts of the genome within an ROH (>2Mb) that we observed for some of the individuals (**Figure 5D)**. We tested several more scenarios with severe bottlenecks where population size decreased to 100 and 64 individuals. A bottleneck to 100 individuals led to an F_{ROH} of only 0.003, which is considerably less than that estimated from the empirical data (**Figure 5D**). When we estimated F_{ROH} from simulation with 64 individuals, we observed the predicted value of 0.075 immediately following the bottleneck (i.e. 7.5% of the genome are in a ROH), and the value did not noticeably drop during the last 200 years even with the subsequent expansion of population size (**Figure 5D**). This matches theoretical predictions where the inbreeding coefficient, *F*, is related to the inbreeding effective population size (N_e) and number of generations[40] (*t*) according to the formula $F=1-(1-1/(2N_e))^t$.

Thus, bottlenecks or population structure would need to reduce inbreeding effective population size to approximately 60 individuals for multiple generations to generate ROH that are comparable to the empirical data. However, we believe this reduction the effective population size, to 64 individuals, is rather unlikely because such a low effective population size is not predicted by our estimates of N_e during the recent bottleneck (N_e >1000; see **Figure 3**), nor by the recent theoretical estimates of N_e in the Americas predicted by Browning and colleagues[66]. Further, historical data suggests that the lowest census population size for just Native Americans was 300 individuals in CO[18] and 1400 in CR[65], which is considerably more than 64 individuals, and does not include the unknown number of European and African American individuals. Since we observe considerable amounts of ROH even in the larger CR population, we conclude that recent consanguineous nonrandom mating was paramount for generating the long ROH that we observed in the Latin American isolates.

Global Ancestry

We looked at the relationship between intergenic π and proportion of ancestry per population (**Figure S9**). We saw that populations with the largest proportions of European and Native American ancestry tended to have lower diversity, and as we expected, populations with higher African ancestry had higher diversity (**Figure S9**).

Since the Latin American isolates originated from an admixture event between Native Americans, Africans, and Europeans, we tested for a correlation between different inbreeding metrics and the proportion of European, African, and Native American ancestry (**Figure S10**). We used the entire sequenced Costa Rican and Colombian data set (*n*=449) for the local ancestry analyses and accounted for relatedness of individuals in all the following reported p-values (see Methods). First, we examined the correlation between F_{PED} and global ancestry. We found that European ancestry was positively correlated with F_{PED} (R^2 = 0.0204; p-value = 0.0052) while Native American ancestry was negatively correlated with F_{PED} (R^2 = 0.0126; p-value = 0.0245). African ancestry was also negatively correlated with F_{PED} (R^2 = 0.0085; p-value = 0.0496).

Next, we examined the correlation between F_{SNP} and global ancestry. Similar to what we observed with F_{PED}, European ancestry was positively correlated with F_{SNP} (R^2 = 0.1120; p-value = 4.76×10^{-12}), Native American ancestry was negatively correlated with F_{SNP} (R^2 = 0.0705; p-value = 2.79×10^{-07}), and African ancestry was negatively correlated with F_{SNP} (R^2 = 0.0545; p-value = 3.49×10^{-08}). We expected that the correlation between F_{SNP} and global ancestry would be stronger than F_{PED} and global ancestry, since F_{SNP} captures the realized inbreeding coefficient rather than the expected inbreeding coefficient.

Lastly, we examined whether ancestry was correlated with the amount of the genome within an ROH (**Figure S10**). The correlation between ancestry and amount of the genome within an ROH followed the same trend as the correlation between ancestry, F_{PED}, and F_{SNP}. Native American ancestry and African ancestry were negatively correlated with the amount of the genome within a long ROH (R^2 = 0.1193; p-value = 9.04 x 10^{-12} and R^2 = 0.0467; p-value = 2.50 x 10^{-07}, respectively). European ancestry was positively correlated with the amount of an individual's genome within an ROH (R^2 = 0.1500; p-value 1.02 x 10^{-15}).

Recent Consanguinity is Correlated with an Increase of Deleterious Variation

It is well known that demography impacts patterns of deleterious variation in populations[2,5,52,56,67–71]. Thus, we compared patterns of putatively deleterious variation in the CO and CR to those in the FIN. Variants were classified as putatively deleterious or putatively neutral using GERP scores (see Methods). Recall that we consider three ways of counting deleterious variants in the genome of an individual: first, counting the number of heterozygous genotypes plus twice the number of homozygous derived genotypes (i.e. the total number of derived deleterious alleles), second, counting the number of heterozygous and homozygous derived genotypes (counting variants), and third, counting only the number of homozygote derived genotypes (counting homozygotes). The first quantity is most relevant if deleterious alleles are additive, while the third is most relevant if they are recessive. First, we looked at absolute counts of derived deleterious variation across isolates (**Figure S11**). Then, we used linear regression to test if there was a relationship between the amount of an individuals' genome in an ROH and the number of nonsynonymous sites in the genome for each counting method (**Figure 6**).

The FIN carried approximately 1% more derived deleterious nonsynonymous alleles per individual than CO and CR (p-value = 0.0007; p-value = 0.0013, Wilcoxon rank-sum test). However, there was no significant difference in the number of putatively neutral synonymous derived alleles per individual. These results suggest that the difference seen for putatively deleterious variants is not driven by data artifacts (**Figure S11**), and the FIN indeed have a slightly higher additive genetic load than the CO or CR. Turning to the number of variants per individual, FIN individuals carried significantly more deleterious nonsynonymous variants than the CR but not the CO (p-value = 0.0110). However, CO and CR did not differ significantly in the number of deleterious variants carried per individual (**Figure S11**). When we examined neutral synonymous variants, CO had significantly more variants than either FIN or CR (p-value = 8.56 x 10^{-06}; p-value = 0.0054, respectively). Finally, when counting the number of homozygous derived genotypes, we found that the FIN carried 3.3% more deleterious variants in the homozygous state per individual than CO but not the CR (p-value = 0.0003) (**Figure S11**). Additionally, the FIN carried significantly more neutral homozygous genotypes per individual than either population (CO p-value = 1.01 x 10^{-05}; CR p-value = 6.96 x 10^{-05}). The increased deleterious and neutral variation in homozygous form is an expected consequence of the long-term bottleneck that the FIN experienced during their founding.

We next tested whether the amount of the genome in an individual contained within a ROH was correlated with the number of nonsynonymous mutations carried by the individual. Counting nonsynonymous (NS) or synonymous (SYN) allele copies did not show any correlation with the amount of an individual's genome that falls within an ROH for the CR or FIN (**Figure 6A and D; Figures S12-15**). However, in the CO, as the amount of the genome within an ROH increased, individuals tended to carry more NS alleles, though this correlation was strongly driven by a single individual, who also had the highest F_{SNP} and F_{PED} (R^2 = 0.2393; p-value = 0.0036; **Figure S12**), and when this individual was removed the correlation no longer remained significant. Importantly, the number of SYN alleles per individual was not correlated with the amount of the genome in an ROH (p-value = 0.2261).

When counting variants per individual, we observed a significant negative correlation with the amount of an individuals' genome that falls within an ROH in the Latin American isolates (**Figure 6B and E; Figures S12-14**). The negative correlation is a result of heterozygous sites being lost when an ROH is formed due to inbreeding. Conversely, when counting homozygous genotypes per individual, we observed a significant positive correlation with the amount of an individual's genome that falls within an ROH in both the Latin American isolates and FIN (**Figure 6C and F; Figures S12-15**). Homozygous genotypes were the only statistic that correlated significantly with the amount of the genome in an ROH across all isolated populations for both SYN and NS sites. We observed a stronger correlation between the number of NS homozygous genotypes and the amount of an individual's genome within an ROH in the Latin American isolates (R^2 = 0.5000 (CO) & R^2 = 0.2165 (CR); p-value = 7.546^{-06}(CO) and p-value = 0.0059(CR)) compared to the FIN (R^2 = 0.1130 and p-value = 0.0389) (**Figures S12-15)**. This pattern exists because the majority of CO and CR individuals carried a larger proportion of their genome within an ROH, while the FIN individuals do not harbor many ROHs.

We next asked whether there was an enrichment or depletion of NS variants relative to SYN variants within versus outside of an ROH using a permutation test on the three different counting approaches (see Methods). When variants or allele copies were counted, none of the populations produced significant results (**Table 2**). When homozygous genotypes were counted, ROHs in the MXL and CR were enriched for homozygous NS genotypes relative to SYN homozygous genotypes (p-value = 0.0052 and p-value = 0.0169) (**Table 2**). Additionally, if we pooled the CR and CO populations, we also observed a significant enrichment of homozygous NS genotypes within an ROH compared to non-ROH regions of the genome (p-value = 0.0011). We tested whether F_{SNP} was correlated with the amount of deleterious variation per individual. We only used isolates for these regressions because we are particularly interested in how recent consanguinity affected deleterious variation in the genome. We observed the exact same pattern with F_{SNP} as with ROH (**Figure S16**). Briefly, counting NS or SYN allele copies did not show any correlation with F_{SNP} for the CR or FIN, but there was a significant correlation with NS allele copies in CO which was driven by a single outlier individual (**Figure S16; Figures S17-19**). The correlation with NS allele copies and F_{SNP} in CO did not remain once the outlier individual was removed. Counting NS and SYN variants per individual produced a significant negative correlation with F_{SNP} in the Latin American isolates (**Figure S14; Figures S17-19**). Counting the number of NS and SYN homozygous genotypes per individual was positively correlated with F_{SNP} in the both Latin American isolates and FIN (**Figure S16; Figures S17-19**). Again, counting homozygotes was the only method with significant results across all isolated populations for both SYN and NS variants. The ability to recapitulate the pattern we observed in ROH using F_{SNP} was reassuring and adds further support to the strong relationship between recent consanguinity and ROH.

Lastly, because we had multi-generational pedigrees for the Latin American isolates, we examined the correlation between putatively deleterious variation and recent consanguinity as measured by F_{PED}. All the following reported p-values account for kinship (see Methods). When we pooled the CO and CR individuals together, we did not observe any relationship between counting derived deleterious allele copies and F_{PED} after correcting for kinship (**Figure 7A**). Moreover, we observed a negative correlation between F_{PED} and the number of deleterious variants per individual ($R^2 = 0.0375$, p-value = 6.02×10^{-06}). The number of neutral variants per individual was also negatively correlated with F_{PED} (p-value = 2.26×10^{-10}) (**Figure 7B**). Finally, we observed a positive correlation between F_{PED} and derived deleterious homozygotes ($R^2 = 0.0575$, p-value = 1.0×10^{-06}) as well as between F_{PED} and the number of neutral derived homozygotes per individual. (p-value = 1.03×10^{-08}) (**Figure 7C**). These results suggest that recent consanguinity during the last few generations has increased the number of derived deleterious homozygous genotypes in these two populations.

2.4 Discussion

Here we present a comprehensive study of genetic diversity, demographic history, identity-by-descent, runs of homozygosity, and deleterious mutations in multiple admixed isolated populations. We show that admixture sufficiently increases genetic diversity of the Colombian and Costa Rican isolates, such that each isolate has diversity levels comparable to a non-isolated population. However, we still observe characteristics in the Latin American isolates that are hallmarks of an archetypal isolate, such as: an excess of IBD segments, cryptic relatedness within the population, and an enrichment of long ROH. Further, we demonstrate that long ROHs contain an enrichment of deleterious variants carried in the homozygous state, which has potential implications for fitness and disease risk.

Taken together, our results support historical data which state that a recent admixture event, within the last 500 years, founded the Colombian and Costa Rican population isolates. After founding, a bottleneck corresponding to the Spanish Settlement occurred and each population has increased in size until the present day[18,72]. We see evidence of these processes in the inference of demography from IBD patterns. Importantly, the bottleneck experienced in the Latin American isolates was not as prolonged as that experienced by the Finnish. Further, the Finnish bottleneck occurred thousands of years ago. The difference in bottleneck timescales likely accounts for some portion of the higher genetic diversity observed in Latin American population isolates in comparison to the Finnish. In other words, the bottlenecks captured by IBDNe in the Latin Americans are too recent to markedly impact levels of heterozygosity. Further, the admixture process experienced by the Latin American isolates could increase levels of genetic diversity[52], especially because some individuals have appreciable levels of African ancestry[49].

We see little difference in patterns of genetic variation in the 1000 Genomes Colombian samples (CLM) and the isolated Colombian sample (CO) studied in this project. The CLM have similar levels of diversity and LD relative to the isolated CO. There is a modest increase in IBD segments and ROH in the isolated CO relative to CLM. The Latin American isolates occupy areas that were considered as being geographically isolated at the time of sampling -- the Central Valley of Costa Rica and the department of Antioquia in Colombia[18], while the 1000 Genomes sample was taken from Medellín which is included within the Antioquia region[73–76]. Thus, these results are a bit surprising as the CO samples studied in this project were from a more remote area and the individuals sampled in the 1000 Genomes Project were from a more cosmopolitan area. This finding likely indicates that the more ancient histories (prior to several hundred years ago) were likely more similar between these populations and have a greater influence on the patterns of genetic variation studied here.

Our results beg the question, what constitutes a population isolate? For example, is it a requirement that population isolates have low genetic diversity relative to the source population? Under this definition, the Latin American population isolates would not qualify as population isolates. The bottleneck in the Costa Ricans and Colombians seems to have had little effect on their genetic diversity, as their diversity levels are comparable to non-isolated Latin American populations. The Finnish, on the other hand, experienced a long-term bottleneck that has resulted in a depletion of segregating sites, and of the remaining segregating sites, there is an enrichment of deleterious variants relative to non-isolated populations[7,11], and would clearly qualify as an isolate. However, if one measures isolation based on IBD we see that there is an enrichment of IBD segments in the Latin American isolates relative to the Finnish. Further, looking at ROH, Latin American individuals from population isolates have a larger burden of ROH than Finnish, thus increasing the chances of identifying more shared genomic regions in the Latin American isolates than the Finnish. By this metric, the Latin American population isolates would qualify as population isolates. Thus, both the Costa Rican and Colombian populations and the Finnish are isolates but in different ways. For example, the Costa Ricans and Colombians are historical isolates, meaning these populations are not currently isolated but they exhibit many traits of an isolate, whereas the

Finnish are contemporary isolates, meaning the population is still isolated. Our work suggests that isolated populations have distinct demographic histories that impact genetic variation in different ways.

We find that Latin American isolates have the largest ROH burden in comparison to any other sampled population, which corroborates results from a recent review where authors state that populations with small N_e and recent consanguinity will harbor the largest amount of ROH[62]. Because previous research has shown a strong correlation between recent inbreeding, quantified by both F_{SNP} and F_{PED}, and long runs of homozygosity, we were particularly interested in the mechanism behind the generation of long ROH[56–61,77–79]. We used simulations to test which demographic scenarios could produce long ROH (**Figure 5**). These simulations and availability of extended pedigree data were crucial, because the F_{SNP} metric can also be influenced by a recent bottleneck. If small population size or admixture were responsible for generating the ROHs, these processes would not be reflected in F_{PED}. Thus, we would not expect to find a correlation between F_{PED} and the amount of the genome in ROHs. The observed correlation between F_{PED} and the amount of the genome in ROH suggests that recent consanguinity (as measured by F_{PED}) is related to the extent of long ROHs in the genome. Further, our simulations show that neither admixture nor a recent population bottleneck, unless unrealistically severe (see Results), could generate the high levels of long ROH that are observed in some individuals. It was only when we incorporated non-random mating into the simulation that levels of ROH comparable to what we observed in our data were produced.

Our results demonstrate that the Latin American population isolates have experienced more recent consanguinity than other population isolates, like the Finnish. Further, in Finland it has previously been shown that the frequency of consanguinity, due to first-cousin marriages, is quite low and the best predictors of these unions were socio-economic class and ethnicity, rather than geographic barriers or population density[80]. On the other hand, for the two Latin American isolates, consanguinity could be a consequence of increased geographic barriers preventing movement of individuals over more dispersed areas. It is also important to point out that it is unclear the extent to which ascertaining individuals from large pedigrees may impact the number of ROHs in our sample. Thus, the finding of an increase in ROHs may not be generalizable to Colombian and Costa Rican populations as a whole. However, we observed a similar pattern of increased ROH in the CLM, which suggests that the pedigree ascertainment of the CO and CR may not be generating the increase in ROHs.

We also tested how recent consanguinity affects deleterious variation in the genome. When counting homozygous derived deleterious genotypes, we found a positive correlation between the number of nonsynonymous homozygous genotypes and the amount of an individual's genome within an ROH (**Figure 6**). Further, we observed an enrichment for nonsynonymous homozygous derived genotypes relative to synonymous homozygous derived genotypes within ROHs versus the rest of the genome (**Table 2**). This enrichment could be a result of nonsynonymous mutations generally segregating at lower frequency and typically being carried as a single copy in an individual. When an ROH is formed, the chromosome that was carrying the mutation is copied, thus allowing the mutation to increase the number of homozygotes within the ROH[56,61]. Since long ROH are a product of recent consanguinity, and these populations have experienced recent consanguinity, we see a corresponding increase in the burden of deleterious variants in the genomes of Costa Rican and Colombian isolates. Because we are more likely to see deleterious variants in the homozygous form in areas of the genome that fall within an ROH, our work is particularly relevant for alleles associated with recessive diseases. Lastly, we provide a mechanism for how recent consanguinity can reduce fitness in natural populations[81–83]. Specifically, if gene-knockouts and deleterious mutations tend to be recessive[38,84–88], as suggested by several studies, then recent consanguinity will increase the number of homozygous derived deleterious variants carried by an individual in a long ROH, thus leading to an overall reduction of fitness in the sampled population[3].

Utilizing estimated ancestry proportions from across the genome, we tested for a correlation between an individual's ancestry and the amount of their genome that falls within an ROH, complementing the work of Szpiech et al. (unpublished)[89]. We found a positive correlation between the proportion of European ancestry and the amount of an individual's genome within a run of homozygosity. These results are consistent with the Latin American isolates originating from a small number of European founders, which would decrease genetic diversity and increase homozygosity for those areas of the genome containing European haplotypes. We observed a negative correlation between Native American ancestry and the amount of the genome contained within an ROH (**Figure S10**). This finding appears to be at odds with previous research[61,62] but largely agrees with conclusions drawn by Moreno and colleagues[90]. Thus, we believe that some of this difference may be due to distinct sampling strategies of the Native American source population in our study compared to previous work. The reference Native American population we used was composed of Chibchan-speaking individuals from Reich et al.[46]. Chibchan-speaking populations inherited their Native American ancestry from admixture between Southern and Northern American lineages[46]. Because our reference Native American population is admixed, and Native American populations tend to be small, it is likely that drift has affected different alleles in source populations[90] that formed the current Chibchan-speaking populations. The Chibchan-speaking populations may have more diversity, fewer fixed homozygous sites, than previously sampled Native American populations which could explain the negative correlation we observed between ancestry and ROH.

While we found evidence of recent bottlenecks and expansions within Latin American isolates using IBDNe[33] (**Figure 3**), our demographic inferences have some limitations. For example, the most current estimates of N_e are unrealistically large or small. The inaccurate estimates of N_e may be due to low sample size, since we only used 30 individuals and it has been suggested that IBDNe works best for larger sample sizes (>200 individuals)[33]. Indeed, the wide 95% confidence intervals around the most recent time points in the FIN suggests much uncertainty regarding the recent effective size of the last five generations and this estimate should not be taken literally. However, a recent study by the creators of IBDNe examined ancestry-specific effective population sizes through time by applying IBDNe to different ancestry segments[66]. Importantly, in that study, the overall genome-wide trajectories of $\mathrm{N_e}$ largely mirror those seen for the individual ancestry components[66]. Thus, we believe that it is appropriate to apply IBDNe to admixed populations. Further, we believe that the demographic patterns that we were able to detect in the Latin American populations (PUR, CO, CR, CLM, and MXL) are robust, as these patterns were recapitulated using a different larger dataset in the same paper[66].

In our study, the populations with the highest IBD scores were admixed (PUR, CO, CR, and CLM). Furthermore, because IBD segments may contain useful information for identifying regions of the genome that contain disease associated mutations, especially within individuals with the highest amounts of consanguinity, it may be useful to deconvolute ancestry for each segment when identifying disease associated mutations because disease prevalence may differ in each parental population. One population that may be of particular interest is the PUR, who demonstrated the largest enrichment of IBD segments while still exhibiting some of the highest levels of diversity. The PUR also stood out in several recent studies. Browning and colleagues found that the PUR had smaller founder sizes than other Latin American populations[66], while Belbin and colleagues used IBD segment mapping in Puerto Ricans sampled from BioMe biobank to identify a gene, COL27A1, that is involved in a common collagen disorder[91].

Population isolates have frequently been used for studying Mendelian[15,92–96] and complex diseases[14,17,97–101]. Our work shows that the genetic diversity and genomic background of population isolates varies immensely. Therefore, it is imperative that we understand the unique genetic diversity and demography belonging to each population isolate. When attempting to identify an isolate, one could use a composite test of with a number of features of interest: enrichment of IBD and/or ROH relative to an archetypal isolate, increase in shared IBD segments, enrichment of deleterious variation at intermediate allele frequencies, or small bottleneck effective population size. For example, if we knew beforehand that there was a history of consanguineous unions within the study population, then we would expect an enrichment of ROH in the composite test. Researchers could shape their study design to target the enrichment of ROH as a tool for disease mapping. This method has previously been used to identify human knockouts, discover novel loci associated with disease, and understand gene function[91,101–104]. Further, ROH could be particularly helpful to better understand disease architecture[105] since ROH may harbor more recessive mutations that do not have full penetrance. Thus, our work highlights the importance of understanding the demographic history of isolated populations, as differences in demographic history will greatly impact their patterns of genetic variation.

2.5 Figures

Figure 2.1 Patterns of genetic variation in the Colombian and Costa Rican populations compared to the 1000 Genomes populations.

(A) Diversity measured using the average pairwise differences between sequences, π. (B) Diversity measured using the number segregating sites, Watterson's theta (θ_W). (C) The site frequency spectrum (SFS) for each population (truncated at a SNP frequency of 15; full SFS **Figure S6**). (D) Average LD (r^2) between pairs of SNPs. All statistics were calculated using 30 unrelated individuals per population (see Methods). Box plots in (A) and (B) show the distribution over 22 autosomes. YRI: Yoruba 1000 Genomes; CEU: Ceph-European 1000 Genomes; FIN: Finnish 1000 Genomes; PEL: Peruvian 1000 Genomes; CLM: Colombian 1000 Genomes; CO: Colombia; CR: Costa Rica; MXL: Mexican from Los Angeles 1000 Genomes; and PUR: Puerto Rican 1000 Genomes.

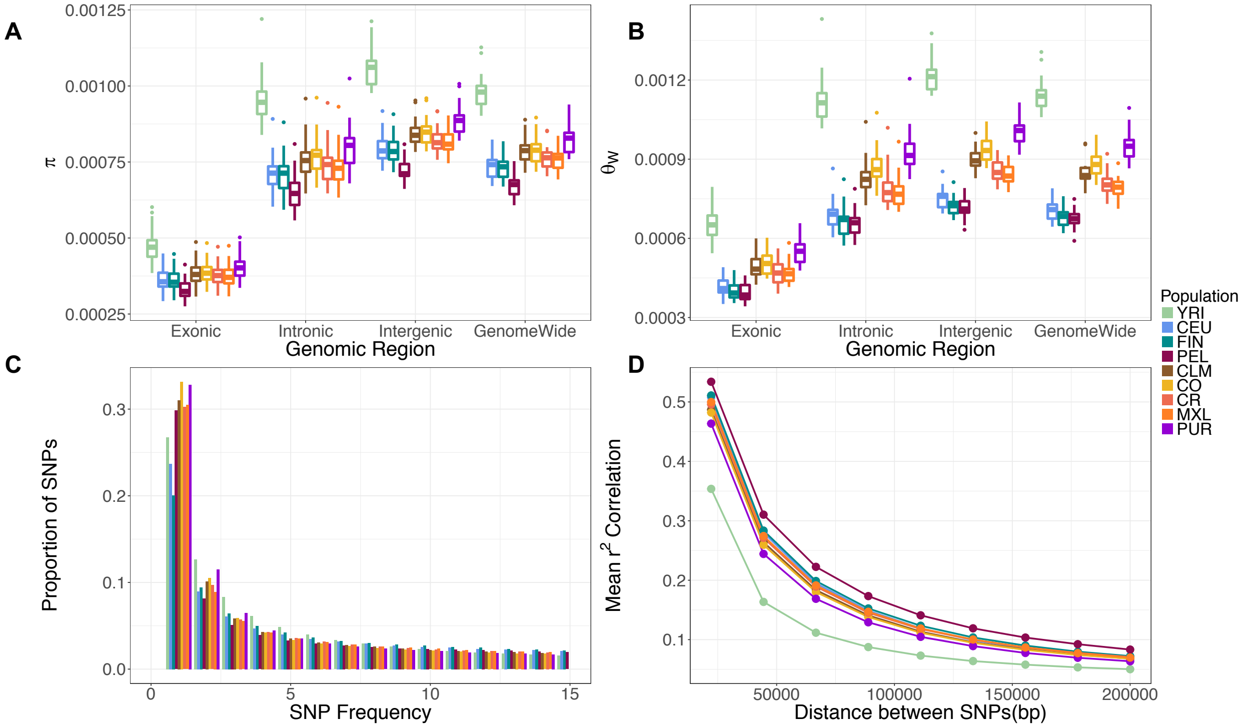


Figure 2.2 Latin American population isolates (CR and CO) have significantly more identity by descent (IBD) segments relative to the Finnish (FIN).

IBDSeq was used to generate IBD segments for the 30 unrelated individuals in each population. (A) Population score was calculated by summing all IBD segments between 3cM and 20cM for each population. Score difference is the population score minus the FIN population score. IBD enrichment for each population score is reported as relative to the FIN (i.e. FIN score is 1.0). (B&C) Histogram of 10,000 permutation tests of Colombia (p < 1.0 e-04) and Costa Rica (p < 1.0 e-04) population scores versus Finnish score. The observed score for each population is demarcated by the purple line. Population abbreviations are as in **Figure 1**.

A

Population	Population score	Score Difference	IBD segment counts per population	Relative to FIN	p-value
PUR	9339.79	4952.27	1402	2.13	< 0.0001
CR	9074.80	4687.28	1247	2.07	< 0.0001
CO	8314.22	3926.7	1177	1.89	< 0.0001
CLM	6702.69	2315.17	927	1.53	< 0.0001
FIN	4387.52	0	965	1.00	
MXL	529.71	-3857.81	105	0.12	0.0117
PEL	458.06	-3929.46	103	0.10	0.6911
YRI	339.97	-4047.55	65	0.08	0.0056
CEU	330.27	-4057.25	54	0.08	0.0001

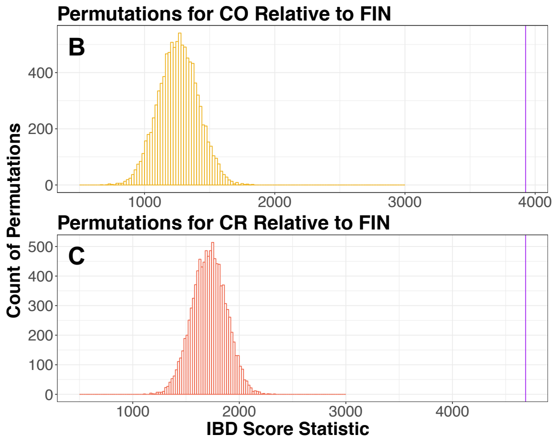


Figure 2.3 Recent effective population size differs across populations.

IBDNe[33] (see Methods) was used to infer effective population size (N_e) over the last 9000 years for each population. Shaded regions denote 95% confidence intervals. Note the FIN shows a long slow decline followed by recent growth. The CO and CR show sharp bottlenecks approximately 500 years ago followed by recent growth. While the overall trends in the population size trajectories appear to be robust to the use of smaller sample sizes in IBDNe, current estimates of Ne are likely inaccurate. Population abbreviations are as in **Figure 1**.

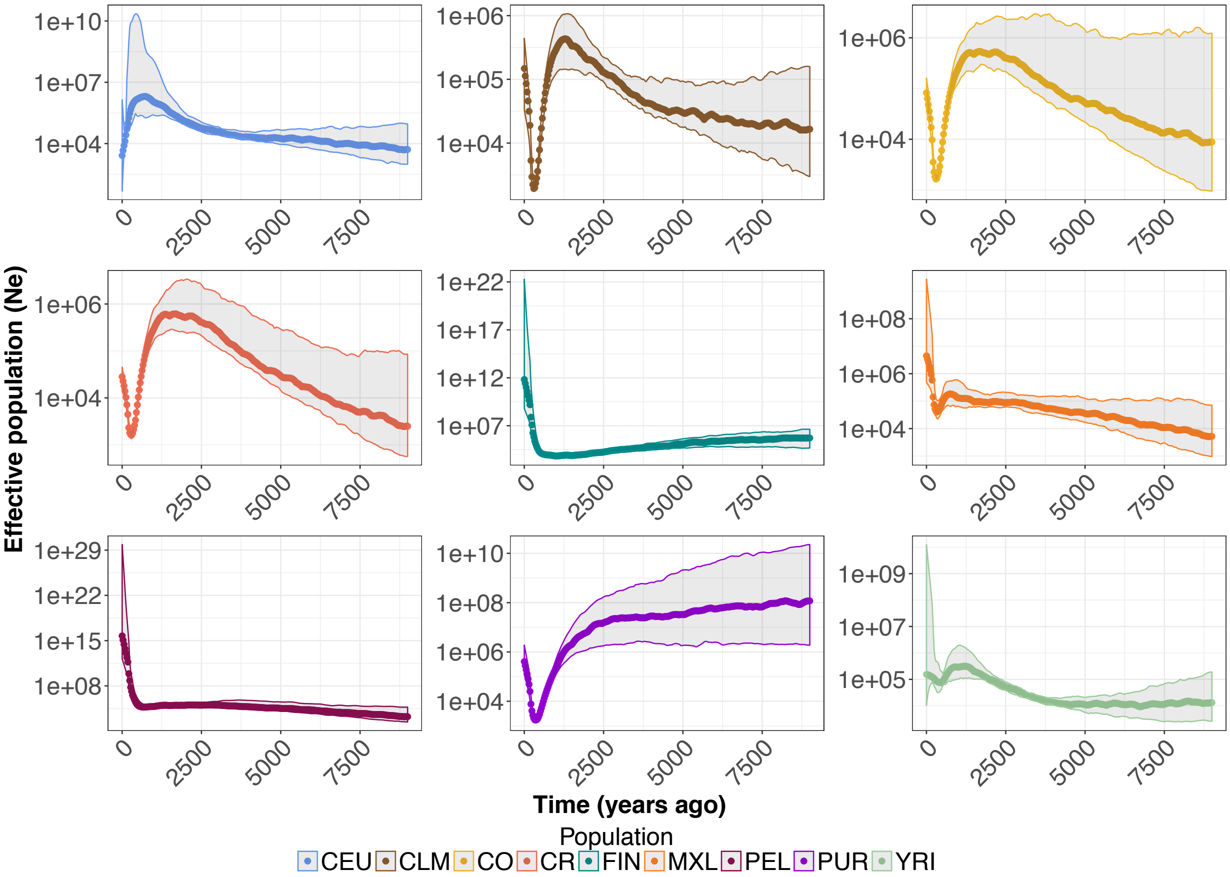


Figure 2.4 Length of the genome in a run of homozygosity (ROH) varies across populations and correlates with SNP inbreeding coefficient.

The length of the genome in an ROH was calculated for each unrelated individual (*n*=30 per population) by summing the physical distance (Mb) of each ROH >2Mb. (A) The length of the genome in an ROH varies by population. The black line within the violin marks the median. F_{SNP} for each individual was overlaid within the ROH violin plot. A blue hue indicates the lowest F_{SNP} and orange indicates the highest F_{SNP}. (B) Length of the genome in an ROH is strongly correlated with F_{SNP} in Colombians, (R^2 = 0.8060, p –value = 1.1 x 10^{-11}). (C) Length of the genome in an ROH is strongly correlated with F_{SNP} in Costa Ricans, (R^2 = 0.7740, p-value = 9.5 x 10^{-11}). (D) Length of the genome in an ROH is positively correlated with F_{SNP} in Finnish, (R^2 = 0.1288, p-value = 0.03). Population abbreviations are as in **Figure 1**.

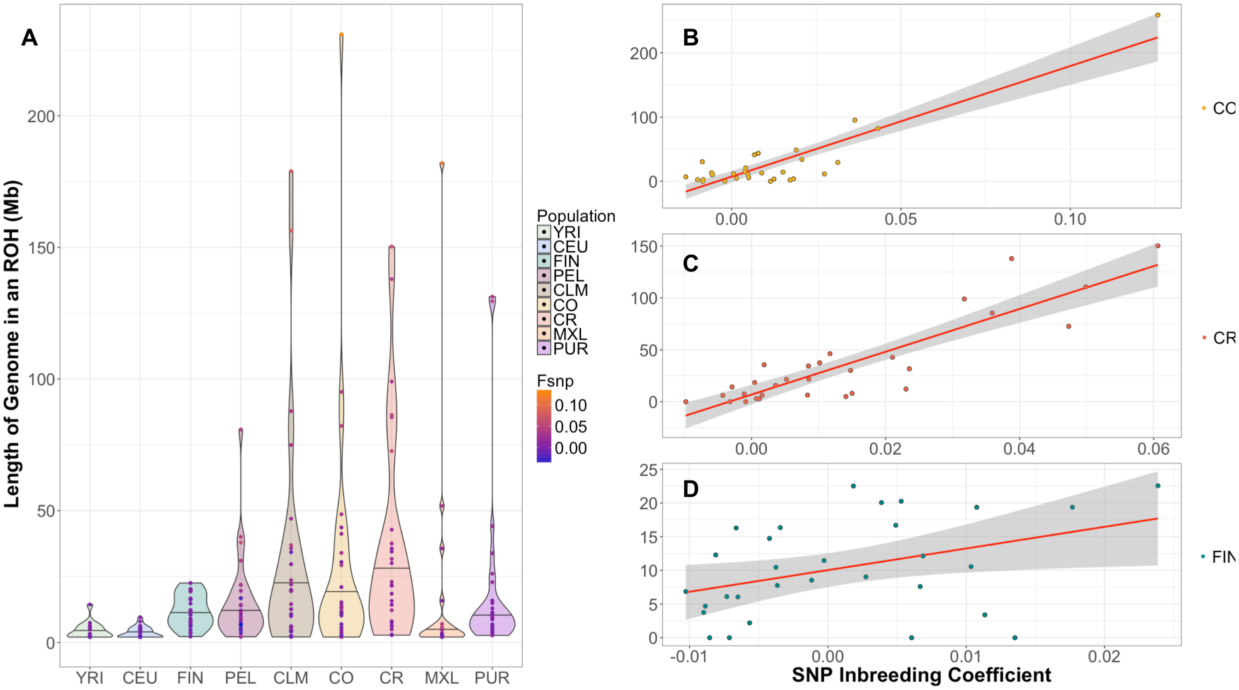


Figure 2.5 Recent consanguinity creates ROH in Costa Rica and Colombia. Similarity of Mendelian disorder gene sets.

Triangles represent the individuals that were sampled in the unrelated data set (n=30). (A) F_{SNP} is correlated with the pedigree inbreeding coefficient (F_{PED}; $R^2 = 0.1520$, p-value < 2 x10-16) in the full data. (B) The length of the genome in an ROH is correlated with F_{PED} ($R^2 = 0.2180$, p-value < 2 x 10-16). (C) The length of the genome in an ROH is correlated with F_{SNP} ($R^2 = 0.7540$, p-value < 2 x 10-16). (D) Forward simulations show that recent consanguinity during the last 500 years was important for the generation of ROHs in the Latin American isolates. Top panel shows the changes in population size used in the simulations. Bottom panel shows how the percent of the simulated in genome within an ROH changes over time for different demographic scenarios. Population abbreviations are as in **Figure 1**.

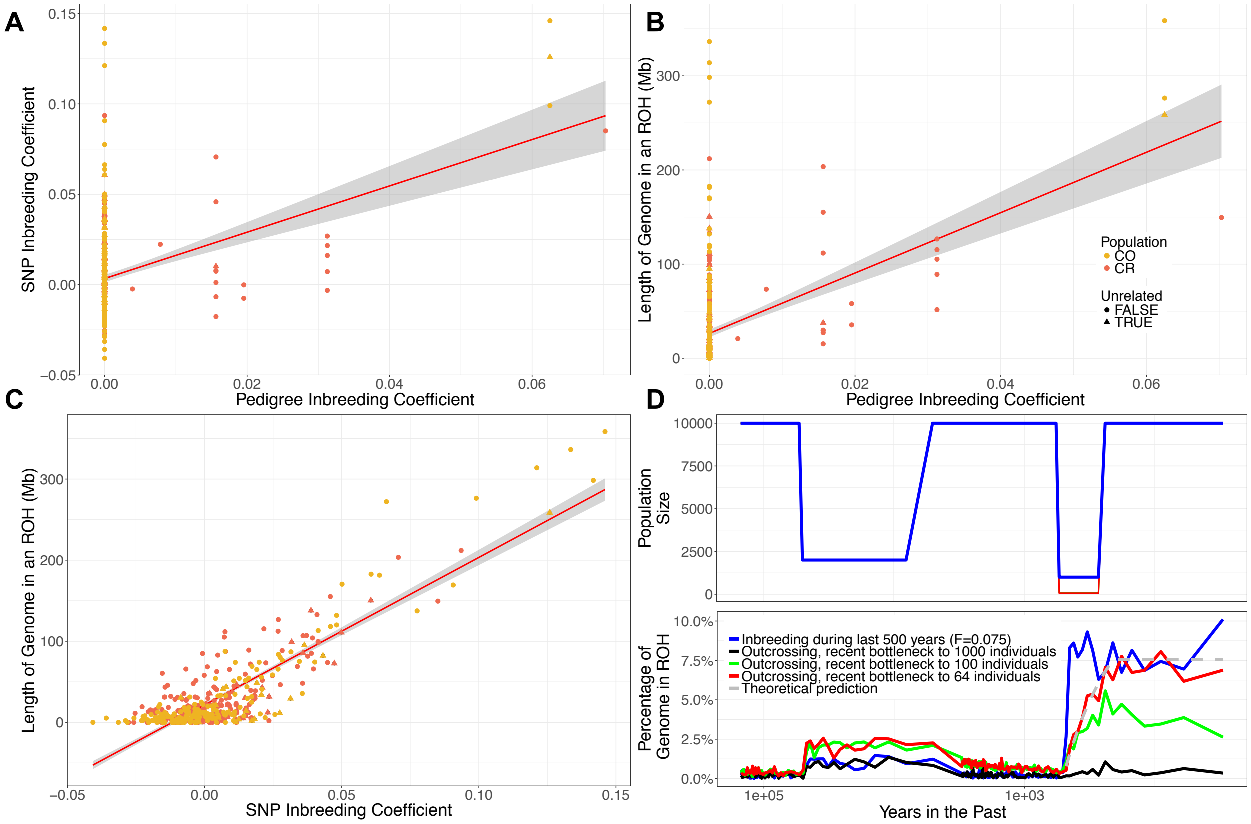


Figure 2.6 The correlation between ROH and nonsynonymous variation in the Colombian, Costa Rican, and Finnish samples.

The count of nonsynonymous and synonymous mutations per individual as a function of the length of the genome in an ROH in the Colombian (CO), Costa Rican (CR) and Finnish (FIN) populations: (A) Number of nonsynonymous alleles per individual. (B) Number of nonsynonymous variants per individual. (C) Number of homozygous nonsynonymous genotypes per individual. (D) The number of synonymous alleles per individual. (E) The number of synonymous variants per individual. (F) The number of homozygous synonymous genotypes per individual. Population abbreviations are as in **Figure 1**.

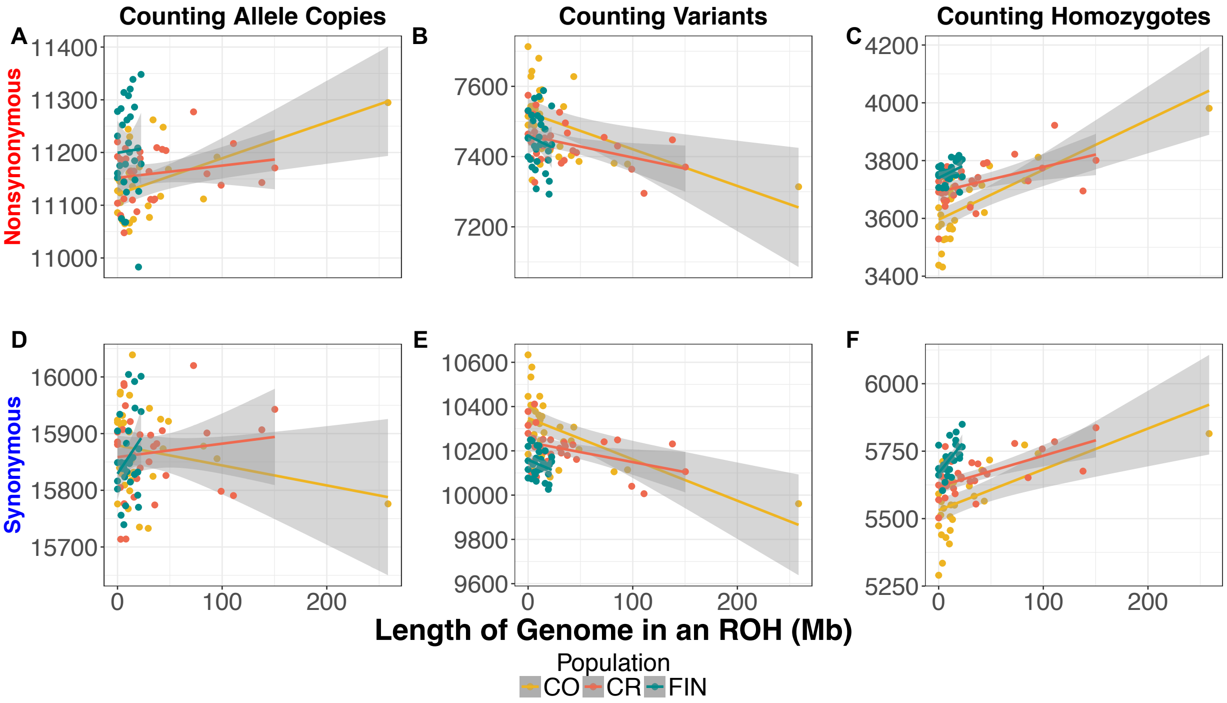
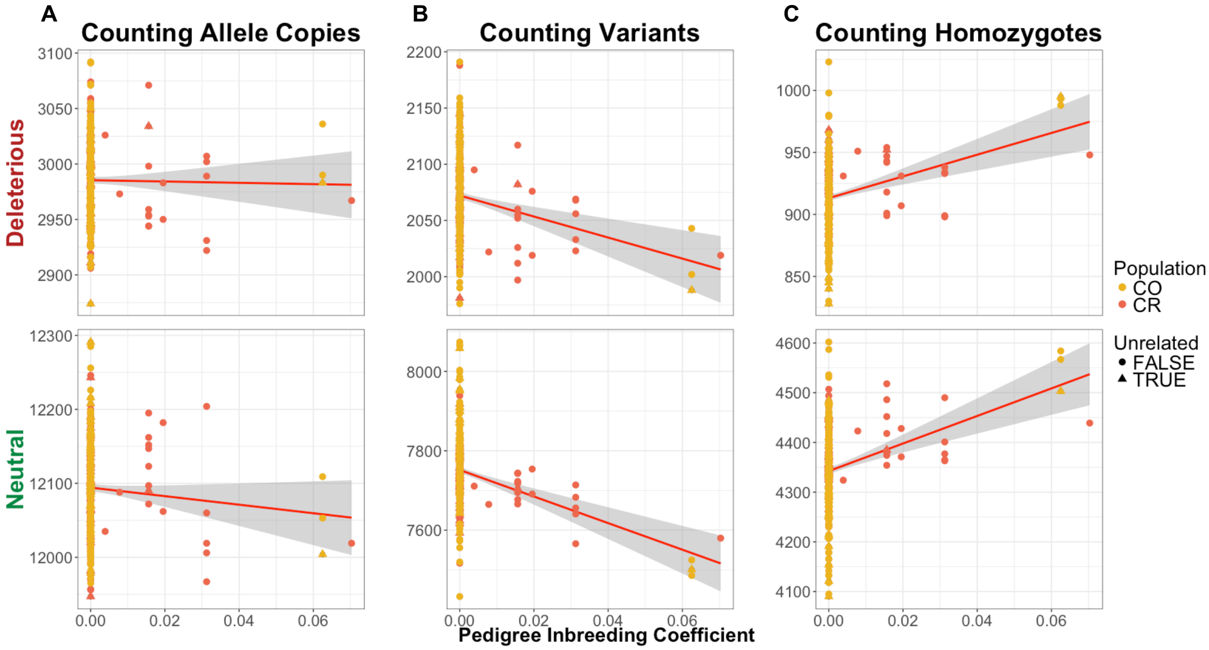


Figure 2.7 Pedigree inbreeding coefficient (F_{PED}) is correlated with deleterious variation.

Triangles represent the individuals that were sampled in the unrelated data set (n=30). Variants were predicted as either putatively deleterious (nonsynonymous) SNPs or putatively neutral (synonymous) SNPs using GERP[41]. Correlation between F_{PED} and the number of mutations per individual in Colombians and Costa Ricans. (A) Number of derived alleles per individual. (B) Number of variants per individual. (C) Number of homozygous derived genotypes per individual. The first row depicts the correlation between deleterious sites using each counting method and F_{PED} for sequenced individuals from Latin American isolates. The second row depicts the correlation between neutral sites using each counting method and F_{PED} in the same individuals. Population abbreviations are as in **Figure 1**.



2.6 Tables

Table 2.1 Ancestry proportions for each sampled population.

This table summarizes the average global ancestry percentages for each of the sampled populations found using ADMIXTURE[47]. Admixture proportions in CO and CR were estimated using supervised model with reference populations. Admixture proportions in other populations were inferred using an unsupervised model (See *Methods*). Population abbreviations are as in **Figure 1**.

Population	Native American	African	European
YRI	0.00	100.00	0.00
CEU	0.00	0.11	99.89
FIN	0.78	0.16	99.06
PEL	88.24	2.20	9.56
CLM	27.95	9.10	62.95
CO	20.43	6.64	72.93
CR	27.3	2.20	70.50
MXL	44.48	5.73	49.79
PUR	14.25	18.59	67.16

Table 2.2 Enrichment of nonsynonymous homozygous derived genotypes within ROHs

This table summarizes the results of our enrichment analyses for each population sampled as well as a combined super-population of Colombians and Costa Ricans (COCR). Odds ratios were calculated as the ratio nonsynonymous variants relative to synonymous variants within versus outside of an ROH for each counting method. Bolded text is used to indicate significant p-values after permutation test was conducted (see *Methods*). Population abbreviations are as in **Figure 1**.

Population	Allele Copies Odds Ratio	Allele Copies p-value	Variants Odds Ratio	Variants p-value	Homozygotes Odds Ratio	Homozygotes p-value
YRI	1.059	0.664	1.048	0.762	1.129	0.417
CEU	1.203	0.105	1.208	0.138	1.252	0.082
FIN	0.937	0.324	0.92	0.265	1.003	0.957
PEL	0.986	0.797	0.972	0.638	1.038	0.54
CLM	0.99	0.755	0.964	0.337	1.066	0.097
CO	1.008	0.828	0.985	0.714	1.074	0.097
CR	1.015	0.607	0.991	0.806	1.085	**0.0169**
CO & CR	1.025	0.283	1.002	0.936	1.088	**0.0011**
MXL	1.112	0.053	1.089	0.169	1.19	**0.005**
PUR	0.981	0.635	0.965	0.411	1.047	0.301

2.7 References

1. Peltonen, L., Palotie, A., and Lange, K. (2000). Use of population isolates for mapping complex traits. Nat. Rev. Genet. *1*, 182–190.

2. Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., and Nielsen, R. (2008). Proportionally more deleterious genetic variation in European than in African populations. Nature *451*, 994.

3. Charlesworth, D., and Willis, J.H. (2009). The genetics of inbreeding depression. Nat. Rev. Genet. *10*, 783.

4. Lohmueller, K.E. (2014). The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. PLOS Genet. *10*, e1004379.

5. Simons, Y.B., Turchin, M.C., Pritchard, J.K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. Nat. Genet. *46*, 220–224.

6. Service, S., DeYoung, J., Karayiorgou, M., Roos, J.L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J.A., et al. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat. Genet. *38*, 556–560.

7. Lim, E.T., Würtz, P., Havulinna, A.S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., et al. (2014). Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet *10*, e1004494.

8. Xue, Y., Mezzavilla, M., Haber, M., McCarthy, S., Chen, Y., Narasimhan, V., Gilly, A., Ayub, Q., Colonna, V., Southam, L., et al. (2017). Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. Nat. Commun. *8*, 15927.

9. Kittles, R.A., Perola, M., Peltonen, L., Bergen, A.W., Aragon, R.A., Virkkunen, M., Linnoila, M., Goldman, D., and Long, J.C. (1998). Dual origins of Finns revealed by Y chromosome haplotype variation. Am. J. Hum. Genet. *62*, 1171–1179.

10. Peltonen, L., Jalanko, A., and Varilo, T. (1999). Molecular genetics the Finnish disease heritage. Hum. Mol. Genet. *8*, 1913–1923.

11. Wang, S.R., Agarwala, V., Flannick, J., Chiang, C.W.K., Altshuler, D., Flannick, J., Manning, A., Hartl, C., Agarwala, V., Fontanillas, P., et al. (2014). Simulation of Finnish Population History, Guided by Empirical Genetic Data, to Assess Power of Rare-Variant Tests in Finland. Am. J. Hum. Genet. *94*, 710–720.

12. De La Chapelle, A., and Wright, F.A. (1998). Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc. Natl. Acad. Sci. *95*, 12416–12423.

13. Martin, A.R., Karczewski, K.J., Kerminen, S., Kurki, M.I., Sarin, A.-P., Artomov, M., Eriksson, J.G., Esko, T., Genovese, G., Havulinna, A.S., et al. (2018). Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. Am. J. Hum. Genet. *102*, 760–775.

14. Panoutsopoulou, K., Hatzikotoulas, K., Xifara, D.K., Colonna, V., Farmaki, A.-E., Ritchie, G.R.S., Southam, L., Gilly, A., Tachmazidou, I., Fatumo, S., et al. (2014). Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. Nat. Commun. *5*, 5345.

15. Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., Patterson, N., Bhavani, G.S., Girisha, K.M., Mustak, M.S., and Srinivasan, S. (2017). The promise of discovering population-specific disease-associated genes in South Asia. Nat. Genet. *49*, 1403.

16. Pedersen, C.-E.T., Lohmueller, K.E., Grarup, N., Bjerregaard, P., Hansen, T., Siegismund, H.R., Moltke, I., and Albrechtsen, A. (2017). The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. Genetics *205*, 787–801.

17. Tachmazidou, I., Dedoussis, G., Southam, L., Farmaki, A.-E., Ritchie, G.R., Xifara, D.K., Matchan, A., Hatzikotoulas, K., Rayner, N.W., and Chen, Y. (2013). A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. Nat. Commun. *4*, 2872.

18. Carvajal-Carmona, L.G., Ophoff, R., Hartiala, J., Molina, J., Leon, P., Ospina, J., Bedoya, G., Freimer, N., and Ruiz-Linares, A. (2003). Genetic demography of Antioquia (Colombia) and the central valley of Costa Rica. Hum. Genet. *112*, 534–541.

19. Consortium, 1000 Genomes Project (2015). A global reference for human genetic variation.

20. Fears, S.C., Kremeyer, B., Araya, C., Araya, X., Bejarano, J., Ramirez, M., Castrillón, G., Gomez-Franco, J., Lopez, M.C., and Montoya, G. (2014). Multisystem component

phenotypes of bipolar disorder for genetic investigations of extended pedigrees. JAMA Psychiatry *71*, 375–387.

21. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
22. Conomos, M.P., Miller, M.B., and Thornton, T.A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. *39*, 276–293.
23. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. *98*, 127–148.
24. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., and Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491.
25. Sul, J.H., Service, S.K., Huang, A.Y., Ramensky, V., Hwang, S.-G., Teshiba, T.M., Park, Y., Ori, A.P.S., Zhang, Z., Mullins, N., et al. (2018). Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates. BioRxiv 363267.
26. Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7*, 256–276.
27. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., and Sherry, S.T. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
28. Browning, B.L., and Browning, S.R. (2013). Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data. Am. J. Hum. Genet. *93*, 840–851.

29. Browning, B.L., and Browning, S.R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics *194*, 459–471.

30. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326.

31. Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179.

32. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., and Masson, G. (2002). A high-resolution recombination map of the human genome. Nat. Genet. *31*, 241.

33. Browning, S.R., and Browning, B.L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am. J. Hum. Genet. *97*, 404–418.

34. Auton, A., Bryc, K., Boyko, A.R., Lohmueller, K.E., Novembre, J., Reynolds, A., Indap, A., Wright, M.H., Degenhardt, J.D., Gutenkunst, R.N., et al. (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res. *19*, 795–803.

35. Sinnwell, J.P., Therneau, T.M., and Schaid, D.J. (2014). The kinship2 R package for pedigree data. Hum. Hered. *78*, 91–93.

36. Haller, B.C., and Messer, P.W. (2016). SLiM 2: flexible, interactive forward genetic simulations. Mol. Biol. Evol. *34*, 230–240.

37. Kim, B.Y., Huber, C.D., and Lohmueller, K.E. (2017). Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. Genetics *206*, 345–361.

38. Huber, C.D., Kim, B.Y., Marsden, C.D., and Lohmueller, K.E. (2017). Determining the factors driving selective effects of new nonsynonymous mutations. Proc. Natl. Acad. Sci. *114*, 4465–4470.

39. Long, M., and Deutsch, M. (1999). Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol. Biol. Evol. *16*, 1528–1534.

40. Kempthorne, O. (1957). An introduction to genetic statistics (John Wiley And Sons, Inc.; New York).

41. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.

42. Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. Am. J. Hum. Genet. *82*, 290–303.

43. Pagani, L., Clair, P.A.S., Teshiba, T.M., Fears, S.C., Araya, C., Araya, X., Bejarano, J., Ramirez, M., Castrillón, G., and Gomez-Makhinson, J. (2016). Genetic contributions to circadian activity rhythm and sleep pattern phenotypes in pedigrees segregating for severe bipolar disorder. Proc. Natl. Acad. Sci. *113*, E754–E761.

44. Consortium, I.H. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851.

45. Consortium, I.H. (2005). A haplotype map of the human genome. Nature *437*, 1299.

46. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., and Mesa, N. (2012). Reconstructing native American population history. Nature *488*, 370.

47. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

48. Aulchenko, Y.S., Ripke, S., Isaacs, A., and Van Duijn, C.M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics *23*, 1294–1296.

49. Lohmueller, K.E., Albrechtsen, A., Li, Y., Kim, S.Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A.F., and Grarup, N. (2011). Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. *7*, e1002326.

50. Cai, J.J., Macpherson, J.M., Sella, G., and Petrov, D.A. (2009). Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. *5*, e1000336.

51. Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., Sella, G., and Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. Science *331*, 920–924.

52. Kidd, J.M., Gravel, S., Byrnes, J., Moreno-Estrada, A., Musharoff, S., Bryc, K., Degenhardt, J.D., Brisbin, A., Sheth, V., and Chen, R. (2012). Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. Am. J. Hum. Genet. *91*, 660–671.

53. Stumpf, M.P., and Goldstein, D.B. (2003). Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr. Biol. *13*, 1–8.

54. Pritchard, J.K., and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. *69*, 1–14.

55. Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J.K., Muzzio, M., Rodriguez-Flores, J.L., Kenny, E.E., Gignoux, C.R., Maples, B.K., Guiblet, W., et al. (2013). Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. PLOS Genet. *9*, e1004023.

56. Pemberton, T.J., and Szpiech, Z.A. (2018). Relationship between Deleterious Variation, Genomic Autozygosity, and Disease Risk: Insights from The 1000 Genomes Project. Am. J. Hum. Genet. *0*,.

57. Kang, J.T., Goldberg, A., Edge, M.D., Behar, D.M., and Rosenberg, N.A. (2016). Consanguinity Rates Predict Long Runs of Homozygosity in Jewish Populations. Hum. Hered. *82*, 87–102.

58. Szpiech, Z.A., Xu, J., Pemberton, T.J., Peng, W., Zöllner, S., Rosenberg, N.A., and Li, J.Z. (2013). Long runs of homozygosity are enriched for deleterious variation. Am. J. Hum. Genet. *93*, 90–102.

59. McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., and Tenesa, A. (2008). Runs of homozygosity in European populations. Am. J. Hum. Genet. *83*, 359–372.

60. Kirin, M., McQuillan, R., Franklin, C.S., Campbell, H., McKeigue, P.M., and Wilson, J.F. (2010). Genomic runs of homozygosity record population history and consanguinity. PloS One *5*, e13996.

61. Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., and Li, J.Z. (2012). Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum. Genet. *91*, 275–292.

62. Ceballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M., and Wilson, J.F. (2018). Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet.

63. Lemes, R.B., Nunes, K., Carnavalli, J.E., Kimura, L., Mingroni-Netto, R.C., Meyer, D., and Otto, P.A. (2018). Inbreeding estimates in human populations: Applying new approaches to an admixed Brazilian isolate. PloS One *13*, e0196360.

64. Kardos, M., Taylor, H.R., Ellegren, H., Luikart, G., and Allendorf, F.W. (2016). Genomics advances the study of inbreeding depression in the wild. Evol. Appl. *9*, 1205–1218.

65. Escamilla, M.A., Spesny, M., Reus, V.I., Gallegos, A., Meza, L., Molina, J., Sandkuijl, L.A., Fournier, E., Leon, P.E., Smith, L.B., et al. (1996). Use of linkage disequilibrium approaches to map genes for bipolar disorder in the Costa Rican population. Am. J. Med. Genet. *67*, 244–253.

66. Browning, S.R., Browning, B.L., Daviglus, M.L., Durazo-Arvizu, R.A., Schneiderman, N., Kaplan, R.C., and Laurie, C.C. (2018). Ancestry-specific recent effective population size in the Americas. PLOS Genet. *14*, e1007385.

67. Kimura, M., Maruyama, T., and Crow, J.F. (1963). The mutation load in small populations. Genetics *48*, 1303–1312.

68. Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. Nature *246*, 96.

69. Hodgkinson, A., Casals, F., Idaghdour, Y., Grenier, J.-C., Hernandez, R.D., and Awadalla, P. (2013). Selective constraint, background selection, and mutation accumulation variability within and between human populations. BMC Genomics *14*, 495.

70. Peischl, S., Dupanloup, I., Kirkpatrick, M., and Excoffier, L. (2013). On the accumulation of deleterious mutations during range expansions. Mol. Ecol. *22*, 5972–5982.

71. Fu, W., Gittelman, R.M., Bamshad, M.J., and Akey, J.M. (2014). Characteristics of Neutral and Deleterious Protein-Coding Variation among Individuals and Populations. Am. J. Hum. Genet. *95*, 421–436.

72. Escamilla, M.A., Spesny, M., Reus, V.I., Gallegos, A., Meza, L., Molina, J., Sandkuijl, L.A., Fournier, E., Leon, P.E., Smith, L.B., et al. (1996). Use of linkage disequilibrium approaches to map genes for bipolar disorder in the Costa Rican population. Am. J. Med. Genet. *67*, 244–253.

73. Wang, S., Ray, N., Rojas, W., Parra, M.V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., and Hurtado, A.M. (2008). Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet. *4*, e1000037.

74. Bedoya, G., Montoya, P., García, J., Soto, I., Bourgeois, S., Carvajal, L., Labuda, D., Alvarez, V., Ospina, J., and Hedrick, P.W. (2006). Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proc. Natl. Acad. Sci. *103*, 7234–7239.

75. Safford, F., and Palacios, M. (2002). Colombia: Fragmented land, divided society (Oxford University Press, USA).

76. Carvajal-Carmona, L.G., Soto, I.D., Pineda, N., Ortíz-Barrientos, D., Duque, C., Ospina-Duque, J., McCarthy, M., Montoya, P., Alvarez, V.M., and Bedoya, G. (2000). Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. Am. J. Hum. Genet. *67*, 1287–1295.

77. Scott, E.M., Halees, A., Itan, Y., Spencer, E.G., He, Y., Azab, M.A., Gabriel, S.B., Belkadi, A., Boisson, B., Abel, L., et al. (2016). Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat. Genet.

78. Gaetano, C.D., Fiorito, G., Ortu, M.F., Rosa, F., Guarrera, S., Pardini, B., Cusi, D., Frau, F., Barlassina, C., Troffa, C., et al. (2014). Sardinians Genetic Background Explained by Runs of Homozygosity and Genomic Regions under Positive Selection. PLOS ONE *9*, e91237.

79. Li, L.-H., Ho, S.-F., Chen, C.-H., Wei, C.-Y., Wong, W.-C., Li, L.-Y., Hung, S.-I., Chung, W.-H., Pan, W.-H., Lee, M.-T.M., et al. (2006). Long contiguous stretches of homozygosity in the human genome. Hum. Mutat. *27*, 1115–1121.

80. Jorde, L.B., and Pitkänen, K.J. (1991). Inbreeding in Finland. Am. J. Phys. Anthropol. *84*, 127–139.

81. Wright, S. (1984). Evolution and the genetics of populations, volume 3: experimental results and evolutionary deductions (University of Chicago press).

82. Charlesworth, B., and Charlesworth, D. (1999). The genetic basis of inbreeding depression. Genet. Res. *74*, 329–340.

83. Wang, J., Hill, W.G., Charlesworth, D., and Charlesworth, B. (1999). Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. Genet. Res. *74*, 165–178.

84. Balick, D.J., Do, R., Cassa, C.A., Reich, D., and Sunyaev, S.R. (2015). Dominance of deleterious alleles controls the response to a population bottleneck. PLoS Genet. *11*, e1005436.

85. Mukai, T., Chigusa, S.I., Mettler, L.E., and Crow, J.F. (1972). Mutation rate and dominance of genes affecting viability in Drosophila melanogaster. Genetics *72*, 335–355.

86. Simmons, M.J., and Crow, J.F. (1977). Mutations affecting fitness in Drosophila populations. Annu. Rev. Genet. *11*, 49–78.

87. Phadnis, N., and Fry, J.D. (2005). Widespread correlations between dominance and homozygous effects of mutations: implications for theories of dominance. Genetics *171*, 385–392.

88. Agrawal, A.F., and Whitlock, M.C. (2011). Inferences about the distribution of dominance drawn from yeast gene knockout data. Genetics *187*, 553–566.

89. Szpiech, Z.A., Mak, A.C., White, M.J., Hu, D., Eng, C., Burchard, E.G., and Hernandez, R.D. (2018). Ancestry-dependent Enrichment of Deleterious Homozygotes in Runs of Homozygosity. BioRxiv 382721.

90. Moreno-Estrada, A., Gignoux, C.R., Fernández-López, J.C., Zakharia, F., Sikora, M., Contreras, A.V., Acuña-Alonzo, V., Sandoval, K., Eng, C., and Romero-Hidalgo, S. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science *344*, 1280–1285.

91. Belbin, G.M., Odgis, J., Sorokin, E.P., Yee, M.-C., Kohli, S., Glicksberg, B.S., Gignoux, C.R., Wojcik, G.L., Van Vleck, T., and Jeff, J.M. (2017). Genetic identification of a common

collagen disease in Puerto Ricans via identity-by-descent mapping in a health system. Elife *6*, e25060.

92. Myerowitz, R., and Costigan, F.C. (1988). The major defect in Ashkenazi Jews with Tay-Sachs disease is an insertion in the gene for the alpha-chain of beta-hexosaminidase. J. Biol. Chem. *263*, 18587–18589.

93. Hästbacka, J., de la Chapelle, A., Mahtani, M.M., Clines, G., Reeve-Daly, M.P., Daly, M., Hamilton, B.A., Kusumi, K., Trivedi, B., and Weaver, A. (1994). The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell *78*, 1073–1087.

94. Ruiz-Perez, V.L., Ide, S.E., Strom, T.M., Lorenz, B., Wilson, D., Woods, K., King, L., Francomano, C., Freisinger, P., and Spranger, S. (2000). Mutations in a new gene in Ellis-van Creveld syndrome and Weyers acrodental dysostosis. Nat. Genet. *24*, 283.

95. Verhoeven, K., Villanova, M., Rossi, A., Malandrini, A., De Jonghe, P., and Timmerman, V. (2001). Localization of the gene for the intermediate form of Charcot-Marie-Tooth to chromosome 10q24. 1-q25. 1. Am. J. Hum. Genet. *69*, 889–894.

96. Valente, E.M., Bentivoglio, A.R., Dixon, P.H., Ferraris, A., Ialongo, T., Frontali, M., Albanese, A., and Wood, N.W. (2001). Localization of a novel locus for autosomal recessive early-onset parkinsonism, PARK6, on human chromosome 1p35-p36. Am. J. Hum. Genet. *68*, 895–900.

97. McInnes, L.A., Reus, V.I., Barnes, G., Charlat, O., Jawahar, S., Lewitzky, S., Yang, Q., Duong, Q., Spesny, M., and Araya, C. (2001). Fine-scale mapping of a locus for severe bipolar mood disorder on chromosome 18p11. 3 in the Costa Rican population. Proc. Natl. Acad. Sci. *98*, 11485–11490.

98. Ober, C., Tan, Z., Sun, Y., Possick, J.D., Pan, L., Nicolae, R., Radford, S., Parry, R.R., Heinzmann, A., and Deichmann, K.A. (2008). Effect of variation in CHI3L1 on serum YKL-40 level, risk of asthma, and lung function. N. Engl. J. Med. *358*, 1682–1691.

99. Stacey, S.N., Sulem, P., Jonasdottir, A., Masson, G., Gudmundsson, J., Gudbjartsson, D.F., Magnusson, O.T., Gudjonsson, S.A., Sigurgeirsson, B., and Thorisdottir, K. (2011). A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. Nat. Genet. *43*, 1098.

100. Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Masson, G., Agnarsson, B.A., Benediktsdottir, K.R., Sigurdsson, A., Magnusson, O.T., Gudjonsson, S.A., and Magnusdottir, D.N. (2012). A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat. Genet. *44*, 1326.

101. Saleheen, D., Natarajan, P., Armean, I.M., Zhao, W., Rasheed, A., Khetarpal, S.A., Won, H.-H., Karczewski, K.J., O'Donnell-Luria, A.H., and Samocha, K.E. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature *544*, 235.

102. Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. *33*, 228.

103. Lencz, T., Lambert, C., DeRosse, P., Burdick, K.E., Morgan, T.V., Kane, J.M., Kucherlapati, R., and Malhotra, A.K. (2007). Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc. Natl. Acad. Sci. *104*, 19942–19947.

104. Mezzavilla, M., Vozzi, D., Badii, R., Alkowari, M.K., Abdulhadi, K., Girotto, G., and Gasparini, P. (2015). Increased rate of deleterious variants in long runs of homozygosity of an inbred population from Qatar. Hum. Hered. *79*, 14–19.

105. Ku, C.S., Naidoo, N., Teo, S.M., and Pawitan, Y. (2011). Regions of homozygosity and their impact on complex diseases and traits. Hum. Genet. *129*, 1–15.

Chapter 3: Long-term small population size and the accumulation of deleterious variation, the tale of the Ethiopian wolf

3.1 Introduction

For many years, For many years, conservation biology has been concerned with the consequences of small effective population sizes in endangered species, the most severe consequence ultimately being extinction[1–5]. Time to extinction can be accelerated in small populations due to the accumulation of deleterious mutations, inbreeding depression, or loss of adaptive potential all of which result in a large reduction in fitness[6]. Since the demographic history of a population impacts the distribution and prevalence of deleterious variation[6–8] elucidating the interaction between demography and genetics is a key component of effective conservation. Further, recent research has shown that the small-population paradigm[1] may not fit all small populations[9], and that some small populations have the ability to avoid extinction presumably through the purging of strongly deleterious recessive alleles. One example of a species with long-term small population size that has yet to be examined is the Ethiopian wolf, which is currently the world's most endangered canid.

The Ethiopian wolf is endemic to the Ethiopian Highlands, where it is restricted to the Afroalpine habitat due to their specialized diet of burrowing mole rats[10]. The Ethiopian wolf population has presumably been declining since the last interglacial period (~10-15,000 years ago) when the Afroalpine habitat began to retract[11], and has experienced recent crashes caused by anthropogenic activity. The expansion of human agriculture into the Highlands has resulted in increased habitat fragmentation, range restriction, in addition to rabies and canine distemper virus outbreaks from increased proximity to domestic dogs[12–17]. Two populations have become locally extinct in the last 25 years[18] and today, there remain less than 500 individuals[19,20] sub-divided between six small isolated populations, which have between 15 and 200 wolves. These numbers are particularly concerning given the smaller proportion of mature wolves, and the social breeding system, whereby only a single dominant female and a few males breed per pack, which results in a very small effective population size[13,14,21]. The largest remaining population is found in the Bale Mountains, where approximately 250 wolves reside as 3 genetically differentiated sub-populations[22]. Gene-flow occurs between these subpopulations through both male and female dispersal, and it is believed these mating behaviors allow for inbreeding avoidance despite this small population size[22,23]. Previous surveys of genetic diversity have been conducted using either mitochondrial DNA[24] or microsatellites[22,23]. These studies have shown that diversity in the remaining Ethiopian wolf populations is low, and have suggested that long-term effective population sizes (N_e) are small across subpopulations from the Bale Mountains[22]. However, this research was limited to a handful of neutral loci, which can introduce biases into results, since most of the genome is not being sampled. We were able to use whole-genome sequence data from 10 Ethiopian wolves in order to assess genome wide diversity levels, examine the prevalence of deleterious variation, and infer the demographic history of these endangered canids.

Here we determine how both ancient and recent demography has affected the genetic diversity of Ethiopian wolves. We generated high-coverage (~40X) whole genome sequence data of 10 contemporary Ethiopian wolves from the Bale Mountains. To examine the effect of long term small population sizes on deleterious mutations, we compare the Ethiopian wolf with two gray wolf populations: 1) the Arctic wolf, a large extant gray wolf population and 2) the Isle Royale wolf, an isolated island population which was founded in the 1940's by 2-3 wolves. This population has since remained at a low population size averaging less than 25 wolves, and exhibited several features of inbreeding depression[25,26]. We also contrast these wild canids to four domestic dog breeds: the Labrador Retriever, Pug, Border Collie and Tibetan Mastiff.

We find that the Ethiopian wolf exhibits remarkably low diversity relative to both gray wolves and breed dogs, as well as an enrichment of derived putatively deleterious variation. The inferred demography of the Ethiopian wolf includes multiple bottlenecks and suggests that the current N_e is quite low. Despite the low N_e, the distribution of runs of homozygosity (ROH) in the Ethiopian wolf does not suggest that there has been recent inbreeding in the population, demonstrating that small populations can avoid inbreeding under specific social and mating structures. We also find evidence of adaptation to high-altitude through positive selection at the CREBBP gene.

3.2 Results

Genetic diversity of the Ethiopian wolf

We used multiple approaches to compare the genetic diversity of the Ethiopian wolf to gray wolves and dogs (Fig. 1). First, we assessed the phylogenetic relationships among canids, and performed hierarchical clustering on the shared identity-by-state (IBS) segments between individuals. We observed the Ethiopian wolf as the deepest divergence as it formed a separate clade from both wolves and dogs (Fig. 1A), which is concordant with recent phylogenetic work showing Ethiopian Wolves as being closer to the root of the genus *Canis*[27]. Indeed, when we compute pairwise-F_{ST} between the seven sampled groups, we found that the Ethiopian wolf is very divergent from the rest, with an average pairwise-F_{ST} of approximately 0.8 (Supplementary Fig. 1), compared to 0.38 for Arctic wolves and 0.35 for Tibetan Mastiffs.

To summarize patterns of genetic variation in the Ethiopian wolf, we compared the genome-wide average number of pairwise differences (π) to those of other canids (Fig. 1B). The Ethiopian wolf had the lowest levels of diversity compared to all the other canids (mean $\pi \approx 0.000269$) and the Arctic wolf had the highest level of diversity (mean $\pi \approx 0.00156$). The Isle Royale wolf had levels of diversity close to breed dogs (mean $\pi \approx 0.00113$). The Pug had the lowest diversity in breed dogs (mean $\pi \approx 0.000679$). To further assess this depletion of diversity in the Ethiopian wolf, we examined the proportional site frequency spectrum (SFS) across all wolf populations (Fig. 1C). Similar to what we observed with π, Arctic wolves had the most diversity and highest proportion of singletons. The Ethiopian wolves and the Isle Royale wolves had a markedly lower proportion of singletons. For the Ethiopian Wolves, along with the reduction of singletons, there was a flattening of the SFS across the frequency spectrum, suggesting a loss of rare alleles. This flattening in the Ethiopian wolf is likely a consequence of population contractions. The Isle Royale wolf SFS showed an enrichment of intermediate frequency variants (relative to any other population) as well as increased fixed variation, suggesting there was a recent contraction in size coupled with severe inbreeding[26] in this group (Supplementary Fig. 2).

Measures of inbreeding

Next, we examined spatial patterns of heterozygosity and runs of homozygosity (ROH) across the genome to gain intuition about the demographic factors affecting the Ethiopian wolf population (Fig. 2). The Ethiopian wolf showed signs of genomic-flatlining, whereby individuals had very low diversity genome-wide and a few peaks of heterozygosity, alongside an enrichment of short and intermediate sized ROH. The observed distribution of ROH and heterozygosity in the genome suggest a long-term small population size in the Ethiopian wolf[28,29]. Conversely, the Arctic wolf had high levels of heterozygosity across the genome and ROH that were infrequent. These patterns are consistent with a demographic history of large population size [30–32]. Whereas, the Isle Royale wolf had a saw-tooth pattern of heterozygosity, where we observed stretches of high heterozygosity intertwined with long ROH. This pattern is well explained by the recent population crash and extensive inbreeding in their population history[26,33].

We found that breed dogs generally had very long ROH interspersed between regions with high heterozygosity. Long ROH were prominent in the Pug and Tibetan Mastiff, who also had lower levels of heterozygosity relative to the two other breed dogs. These patterns are likely consequences of extremely small effective population sizes in the recent history of the Pug and

Tibetan Mastiff. Overall, it is apparent that breed dogs have an enrichment of long ROH compared to wolves. The enrichment of long ROH is a consequence of the domestication bottleneck (common to all dogs) and bottlenecks of varying intensity experienced during breed formation. In addition to the breed formation bottleneck, some breeds experienced subsequent sharp decreases in population size during their history with strong artificial selection for phenotypic traits of interest. This intensive selection has further limited the effective size of dogs and results in even more ROH.

Demographic history of the Ethiopian wolf

Our results are largely consistent with previous surveys of genetic diversity[22–24] and demographic analysis which suggested that the effective population size of Ethiopian wolves has been decreasing through time[11,22]. However, none of the previous studies reconstructed the demographic history of the Ethiopian wolf. Given multiple genomes and high coverage data, we were able to infer the demographic history of the Ethiopian wolf using approximate Bayesian computation (ABC). We inferred demographic parameters using the distribution of segregating sites and π for approximately 21000 neutral windows across the genome in a subset of four individuals (due to sample limitation). We fit a 3-epoch model (**Fig. 3**) with a series of instantaneous population contractions, chosen based on our analyses using PSMC on an independent sample (Supplementary Fig. 3). We inferred a large ancestral population size of 159,754 individuals and an ancient population contraction approximately 13,572 generations ago. The ancient contraction was severe and decreased population size to approximately 11,600 individuals, or seven percent of the original population size. This contraction was followed by an additional recent severe contraction, which decreased the current effective population size to approximately 100 individuals (**Fig. 3**). Our estimated current effective population size is on the same order of magnitude as estimates from field surveys, which place current census size as 197 mature individuals and decreasing (**https://www.iucnredlist.org/species/3748/10051312**).

Deleterious variation

Given that we observed exceptionally low diversity in the Ethiopian Wolves (**Figs. 1** and **2**) and inferred multiple population size contractions over the last ~13,500 generations, we wondered about the prevalence of deleterious variation in the Ethiopian wolf. To compare across canids, we first polarized alleles using the Wild dog (*Lycaon pictus*) as the reference genome then used GERP to annotate variation as putatively neutral or deleterious. We used multiple methods to count deleterious variants in the genome of an individual: 1) summing homozygous derived genotypes (counting homozygotes); 2) summing the total number of homozygous and heterozygous derived genotypes (counting variants); and 3) tallying twice the number of homozygous derived genotypes plus heterozygous genotypes (counting alleles). If deleterious alleles are recessive counting derived homozygotes is most relevant, whereas counting alleles is relevant if alleles are additive[34,35].

Overall, as expected, breed dogs carried slightly more deleterious variation than the Arctic wolf, and the Isle Royale wolf carried deleterious variation at levels closer to dogs[26,36] (Supplementary Fig. 4). Remarkably, we found that the Ethiopian wolf carried more neutral and deleterious derived homozygotes than dogs and wolves (Supplementary Fig. 4). On average, we observed a 1.56-fold (p-value = 3.655e^{-06}) and 1.76-fold (p-value = 1.872e^{-05}) enrichment of deleterious homozygotes in the Ethiopian wolf relative to dogs or wolves respectively. When counting variants, all three wolf populations carried comparable numbers of deleterious derived variants, and there was a slight enrichment of synonymous variants relative to dog (1.13-fold & p-value = 3.653e^{-06}). The Ethiopian wolf also appears to have a depletion of neutral variants, suggesting that most variants are fixed rather than segregating as heterozygotes. This result is concordant with what we observe in terms of diversity (**Fig. 1B**) and heterozygosity (**Fig. 2**). Lastly, when counting alleles, we found the Ethiopian wolf carried a comparable number of neutral alleles as dogs and wolves, and a 1.29-fold (p-value = 3.665e^{-06}) and 1.24-fold (p-value = 1.872e^{-05}) increase of deleterious genotypes relative to dogs and wolves respectively. In sum, we find elevated levels of predicted deleterious homozygous genotypes and derived alleles in the Ethiopian wolf compared to other wolf populations and breed dogs, implying an accumulation of weakly deleterious variation due to a continued small population size.

Adaptation to high-altitude

The Ethiopian wolf exclusively inhabits the high-altitude (3,000-4,500 meters) regions of the Bale Mountains and has specialized to hunt mole rats native to this area[10,37]. Given that the Ethiopian wolf has been restricted to high-altitude for thousands of generations, we hypothesized that they may have adapted to a hypoxic environment much like the Tibetan wolves or mastiffs[38–41]. Thus, we performed a genome-wide scan for signatures of positive selection. We used a combination of outlier approaches (π and F_{ST}), counts of derived homozygous genotypes, and Gene Ontology (GO) terms to search for evidence of adaptive loci in the Ethiopian wolf. In **Fig. 4**, we have summarized the top 5% of outliers from the F_{ST} analyses between the Ethiopian wolf and the Arctic wolf. There was a single gene, CREB-binding protein (CREBBP), that fell within the top 5% of all F_{ST} comparisons, regardless of which outgroup was used. We also observed low values of π (Supplementary Fig. 5) and an enrichment of fixed derived homozygous sites (Supplementary Fig. 6) in CREBBP relative to all other genes in the Ethiopian wolf genome.

CREB is a transcription factor that is critical to cellular processes, such as cell development, proliferation, differentiation, hypoxia-response, and circadian rhythm[42–45]. One of the primary protein-protein interactions of CREBBP is with hypoxia inducible factor (HIF)-1α (Fig.4) and has been validated with multiple methods[46]. The CREB transcriptional complex (CREB, CREBBP, and p300) has been shown to interact with multiple hypoxia-activated genes via HIF-1α in response to oxygen deprivation[43,44]. Since p300 directly interacts with DNA bound HIF-1, the CREB transcriptional complex can physically modulate the cellular response to hypoxia[43]. Cells can specifically adapt to hypoxic conditions via HIF-1-dependent induction of erythropoietin (Epo) which increases red blood cell production[43,44,47]. Further, previous work on pathways involved in response to hypoxia in human Ethiopian populations identified BHLHE41 as showing a strong signature of selection, and CREBBP directly interacts with BHLHE41[39].

Loss of PRDM9

PRDM9, a gene known for its involvement in specific hot spots associated with meiotic recombination in humans and other mammalian species has knock-out mutations in dogs, coyotes, and golden jackals[48–51]. Due to the loss of PRDM9 function, meiotic recombination is believed to occur by a different mechanisms in canid species than other mammals[50]. Since the Ethiopian wolf is an outgroup on the canid phylogeny compared to dogs, coyotes, and golden jackals, we reconstructed the gene tree of PRDM9 (Supplementary Fig. 7) using the domestic cat as the outgroup. As we expected, the human and macaque clustered together, and remained close to the other species that retained a functional copy of PRDM9, whereas canids formed a single divergent clade. We used GeneWise[52] to test whether PRDM9 was functional and observed multiple frameshift mutations and stop codons in the zinc-finger region of PRDM9 in all the canid species (Supplementary File). Thus, our results indicate that functionality of PRDM9 was lost prior to the speciation of the Ethiopian wolf, and likely has been lost across the entire grouping of wolf-like canids since the Dhole also retained a non-functional copy of PRDM9.

3.3 Discussion

Our study is the first to use whole-genome sequence data to investigate genetic diversity and infer the demographic history of the Ethiopian wolf, the world's most endangered canid. Despite showing molecular evidence for adaptation to its high-altitude environment, the species has been steadily declining for millennia and experienced recent subpopulation extinction. We have shown that despite having a long-term small effective population size, low genetic diversity, and an enrichment of deleterious variation relative to breed dogs and gray wolves, the Ethiopian wolf has persisted for about 13,500 generation since an ancient contraction (or ~ 40,500 years assuming a 3-year generation time[11]) . This enrichment of derived deleterious homozygotes seems plausible given the long-term small population size of the Ethiopian wolf. Despite accumulating a large amount of deleterious variation, there have not been any reported cases of apparent inbreeding depression[22,23,53], leading us to believe that purging has perhaps removed strongly deleterious recessive variants, much like what was observed in the Island fox[9]. Thus, the long-term persistence of the Ethiopian wolf could presumably be enabled through the purging strongly of deleterious recessive alleles and avoidance of inbreeding. Overall, our findings have illuminated some of the interactions between demography and genetics in the Ethiopian wolf, and can be used to enhance conservation efforts, which have previously focused on addressing disease, monitoring movements and social behavior, and the ranges of extant sub-populations.

3.4 Methods

Ethiopian wolf samples

The Ethiopian wolf samples for this study were selected from 85 previously collected tissue samples from the Bale Mountains population for which relatedness and microsatellite data were available[22]. The Bale Mountains population has previously been shown to consist of three sub-populations, Sanetti Plateau, Morebawa, and Web Valley [22]. These populations are genetically differentiated due to restricted gene flow between areas. Moreover, individual packs are primarily composed of close relatives. We used the microsatellite data to select individuals for sequencing that clustered with the Morebawa population and avoided sampling close relatives. All selected wolves had pairwise relatedness values < 0.11. All but one selected wolf had a STRUCTURE cluster assignment value for the Morebawa population of > 0.88. The one exception, T431, had an assignment value of 0.489 and appeared to be admixed between Morebawa and Sanetti Plateau. The IDs of the wolves that were selected for sequencing included T699, T294, T279, T47, T392, T437, T431, T577, T852, T30, their tissue samples were sent to MedGenome for extraction, library preparation and sequencing to approximately 40x coverage using Illumina HiSeq X machines. The section below describes how the raw sequence data was processed.

Sequence data processing

We merged the newly generated Ethiopian wolf sequence data with several existing canid whole-genome sequencing datasets. These included: Arctic wolf (N=15)[26] with approximately 39X coverage, Isle Royale wolf (N=10)[26] with approximately 24X coverage, Pug (N=15)[54] with approximately 47X coverage, Labrador Retriever (N=10)[55] with approximately 30X coverage, Tibetan Mastiff (N=10)[55] with approximately 15X coverage, Border Collie (N=10)[55] with approximately 24X coverage.

Raw whole genome sequences were processed following GaTK best practices and with steps 1-9 in the NGS pipeline from Phung et al. 2019 (https://github.com/tanyaphung/NGS_pipeline). In brief, fastq files were first aligned to the dog genome (CanFam3.1) with BWA[56], duplicate reads were marked with Picard tools (http://broadinstitute.github.io/picard/), poor reads were removed using samtools[57,58], and base quality scores were recalibrated with BQSR in GaTK v3.8[59].We performed joint genotyping with Haplotype Caller and emitted all sites (variant and invariant). To reduce bias in SNP calling accuracy between dogs, where we had many samples, and the wolves, where we had fewer samples, we conducted joint genotyping on each dog breed and each wild canid species separately. For example, joint genotyping was conducted on the Ethiopian Wolves as a group, and jointly on the Labradors as a group.

We then applied *post hoc* filtering to the seven group VCFs. Specifically, we applied GaTK filtering recommendations for variant sites in non-model species: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < - 12.5, ReadPosRankSum < -8.0, and a minimum genotype quality (GQ) of 20 and we removed clustered SNPs (> 3 SNPs within 10 base pairs). For invariant sites, where no best practices were available, we removed sites with QUAL < 30, and RGQ < 1. For both variant and invariant sites, we applied a minimum depth filter of 10 for each genotype, as previous work has found heterozygous calls are unreliable below this depth (Marsden et al. 2016), and a maximum depth filter of 2.5 times the average genomic coverage (specific for each sample). We also removed any sites where all individuals were heterozygous, where fewer than 80% of individuals in a group had a genotype call after post hoc filtering was applied, and any sites within the USCS repeat regions (http://hgdownload.soe.ucsc.edu/goldenPath/canFam3/database/rmsk.txt.gz).

As a final step, we merged all autosomal data into a single joint-VCF for our analyses that contained sites that were present in at least 90% of individuals. There was a total of 8,818,790 sites in the final merged data set.

Relatedness

Relatedness between individuals was computed using VCFTools[60], specifically the --relatedness2 option which incorporates the KING[61] algorithm for computing pairwise kinship. These results were cross-checked using PLINK[62] --genome command to estimate relatedness. First degree relatives were removed from each sampled group resulting in the following sample sizes used for subsequent analyses: Ethiopian wolves (N=10), Arctic wolf (N=14), Isle Royale wolf (N=10), Pug (N=15), Labrador Retriever (N=10), Tibetan Mastiff (N=9), Border Collie (N=7).

Variant Annotation

The ancestral allele at each site was determined using the African Wild Dog[63] (NCBI SRA: SAMN09924608) aligned to the CanFam3.1 reference. For sites where the Wild Dog was homozygous, we used the Wild Dog allele as the ancestral allele. For sites where the Wild Dog was heterozygous and one of the alleles matched the CanFam3.1 reference allele, we used that allele as the ancestral allele. Variant annotation was done using Ensembl VEP (version 94) with SIFT annotations enabled[64,65]. Genomic Evolutionary Rate Profiling (GERP) scores[66] were used alongside VEP to annotate variants as either neutral (synonymous with GERP Score less than 2) or deleterious (nonsynonymous with a GERP score greater than 4). For details about how GERP scores were generated, see work from Marsden et al. 2016[36]. Counts were compared using a Wilcoxon Rank Sum test.

Runs of homozygosity

ROH were called three ways: 1) using all populations as the input for VCFTools[60], 2) using BCFTools[58] with and without a genetic map, and 3) ROH were called separately for each population using VCFTools. Ultimately, we chose to use option three for our analyses. Then, we tallied the number of callable sites in each individual's genome that lie within each ROH. Our final set of ROH included those at least 10kb in length and where at least 40% of the run overlapped callable sites. F_{ROH} was calculated as the length of the genome within a ROH of at least 1Mb divided by the total length of the CanFam3.1 genome.

Genetic diversity

We calculated pi (π), the average number of pairwise differences per site, in a subsample of six individuals from each population. π was computed across the genome as:

$$\pi = \frac{n}{n-1} \frac{\sum_{i=1}^{L} 2p_i(1-p_i)}{L}$$

where n is the total number of chromosomes sampled, p is the frequency of a given allele, and L is the length in base pairs of the sampled region.

Population differentiation

F_{ST} between all individuals from each sampled population was computed using Weir and Cockerham's formula[67] as implemented in VCFTools[60].

We used the unsupervised learning model in ADMIXTURE[68] to determine the number of distinct genetic clusters within the data. 780,150 LD-pruned SNPs were generated using the suggested PLINK[62] command from the manual and the number of source populations varied between K=2 and K=10. The lowest cross-validation error was produced when K=6.

PCA (Supplementary Fig. 8) of the genetic variation data was conducted in R using a combination of SNPRelate[69], PC-AiR[70], and PC-Relate[71]. We first generated a genetic relatedness matrix using SNPRelate's implementation of KING[61], to correct for ancestry, then PC-AiR and PC-Relate were used to perform PCA as these methods are robust to population structure, cryptic relatedness, and admixture. We also performed hierarchical clustering on the identity-by-state (IBS) matrix generated by SNPRelate's implementation KING using the default settings.

Demographic inference

We inferred the population history of the Ethiopian wolf using approximate Bayesian computation[72] (ABC). As natural selection can confound demographic inference[73,74], we selected windows along the genome where mutations are likely neutrally evolving. These windows were defined as being outside genes and conserved regions of the genome and greater than 0.4cM away from any gene. Then, the total number of segregating sites (S) in each window was tabulated using VCFTools --snpdensity option with a window size of 10bp, and π was computed across each window as well.

Following the approach taken in Robinson et al. 2016, we used PSMC[75] on a single individual to determine possible demographic history and inform priors for changes in population size and the bottleneck time. We considered a 3-epoch demographic model with serial contractions and priors for $N_{ANCIENT}$ ranging from to 40,000-200,000 individuals, $N_{INTERMIDIATE}$ from 60,000-1,000 individuals, and $N_{CURRENT}$ from 1,000-10 individuals. We drew parameter values from the aforementioned uniform prior distributions and performed 20,0000 coalescent simulations using ms[76] to simulate data for N=4 diploid individuals across approximately 21,000 neutral regions that were length-matched to our empirical data and at least 1 kilo-base in length, with a mutation rate[77] of $4.96e^{-09}$ and a generation time of 3 years [11].

For each draw from the prior, we then assessed the fit of the model to the data using a joint statistic of π and S. We calculated the following distance metric for each set of parameters from the prior to decide which simulations to accept and reject:

$\alpha = \sum_{i=0}^{M} |E_i - S_i|$, where (E_i) is the sum of π for a given interval and S in the empirical data, (S_i) is the sum of π and S in the simulated data, and M is the maximum number of bins of π and S this was set to 40. The top 200 simulations that minimized α were kept and composed the posterior distribution. **Supplementary Figure 9** compares bins of π and S from the top 200 simulations against the empirical data, shows the posterior and priors for each parameter, and **Supplementary Table 1** contains the 95% credible interval for each parameter point estimate.

Phylogenetic tree reconstruction and multiple sequence alignment in PRDM9

Given that Ethiopian wolves are an outgroup to most canids[27], we wanted to test whether they have a functional version of PRDM9. PRDM9 is involved in recombination and is responsible for positioning of recombination hotspots in the genome. We compared PRDM9 to another gene, GAPDH, which is a housekeeping gene in the glycolysis pathway. The PRDM9 and GAPDH coordinates of hg19, mm10, rheMac2, canFam2 were identified using the ECR browser (https://ecrbrowser.dcode.org/) and a fasta file was generated. The data for felCat was pulled from UCSC. These fasta files were subsequently merged with a second fasta file containing the homologous regions of a single Arctic wolf (AW15), a single Ethiopian wolf (EW7), a single Tibetan Mastiff (TM3), and a single Dhole[55] (Dhole01). Then, multiple sequence alignment was conducted using an online version of MAFFT (https://mafft.cbrc.jp/alignment/server/). Lastly, a neighbor-joining tree was constructed using the Jukes-Cantor substitution model (Jukes and Cantor 1969).

For PRDM9, GeneWise[52] was used to align the protein sequence of the Ethiopian wolf, Arctic wolf, and Tibetan Mastiff relative to hg19 DNA sequence. GeneWise allowed us to identify intronic and frameshift errors, which would affect the functionality of the gene.

Selection statistics to detect positive selection

We used a combination of outlier methods (F_{ST}, π, and derived allele counts) to identify potential regions of the genome undergoing natural selection in the Ethiopian wolf. Additionally, we compared the results from outlier methods across multiple breeds of dog (Border Collie and Tibetan) and the Arctic wolf. This is because the Tibetan Mastiff has been shown to have adapted to high-altitude[38–40,78]. As a first pass, we compared the average value of F_{ST} per gene, which was computed per site and between the Ethiopian wolf and each comparison population with VCFTools using the --weir-fst-pop command. We then intersected the top 5% F_{ST} outlier loci across all three comparisons and identified candidate loci for high-altitude adaptation that were either in the HIF-1α pathway, by using PANTHER (http://www.pantherdb.org/), or in a curated list of previously identified loci from high-altitude adaptation selection scans[39]. Lastly, we examined both π and derived allele count in the Ethiopian wolf for each locus.

3.5 Figures

Figure 3.1 Summaries of genetic variation in the Ethiopian wolf compared to other canids.

A. Hierarchical clustering based on shared IBD segments between individuals, the dendrogram was cut into 4 groups. Note that the Ethiopian wolves sit as an outgroup compared to other canids. **B.** Diversity in dogs and wolves measured using the average number of pairwise differences between sequences, π. Ethiopian wolves have exceptionally low genetic diversity. **C.** The folded site frequency spectrum (SFS) for each wolf population. The full-unfolded SFS can be found in the Supplement.

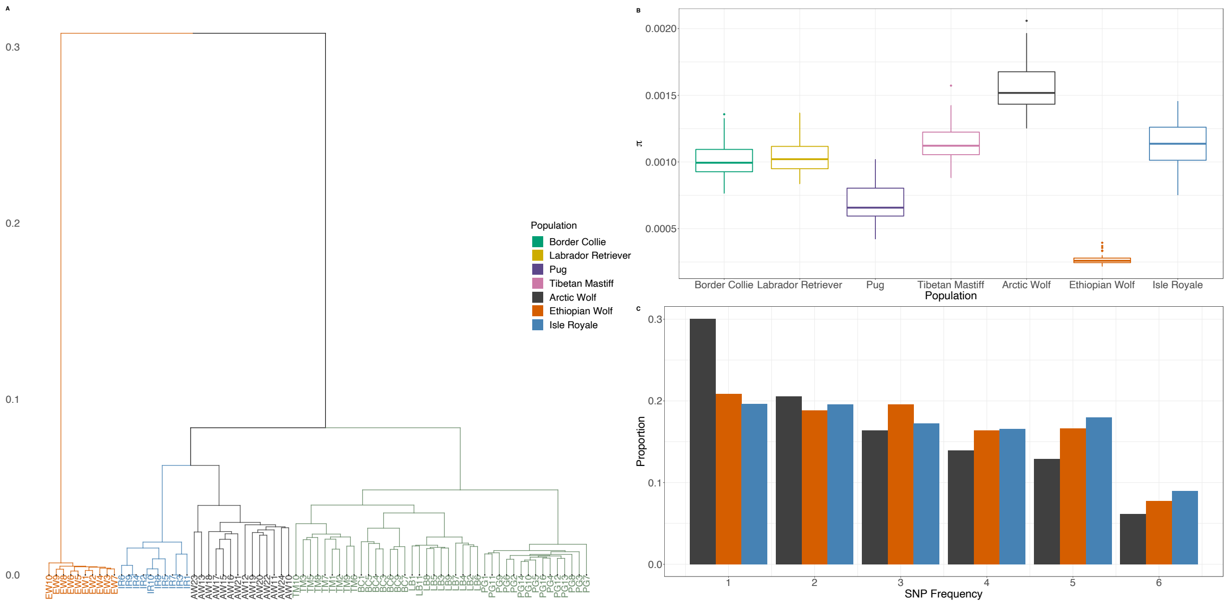


Figure 3.2 Distribution of homozygosity across the canid genome.

A. Sliding-windows of heterozygosity from a single representative sample in each population. Chromosomes are ordered from 1-38. Note the low heterozygosity across the entire Ethiopian wolf genome. **B.** The distribution of short [10Kb-1Mb), intermediate [1-10Mb), and long ROH [10Mb-63Mb) in dogs and wolves. Each row of the plot represents a single individual.

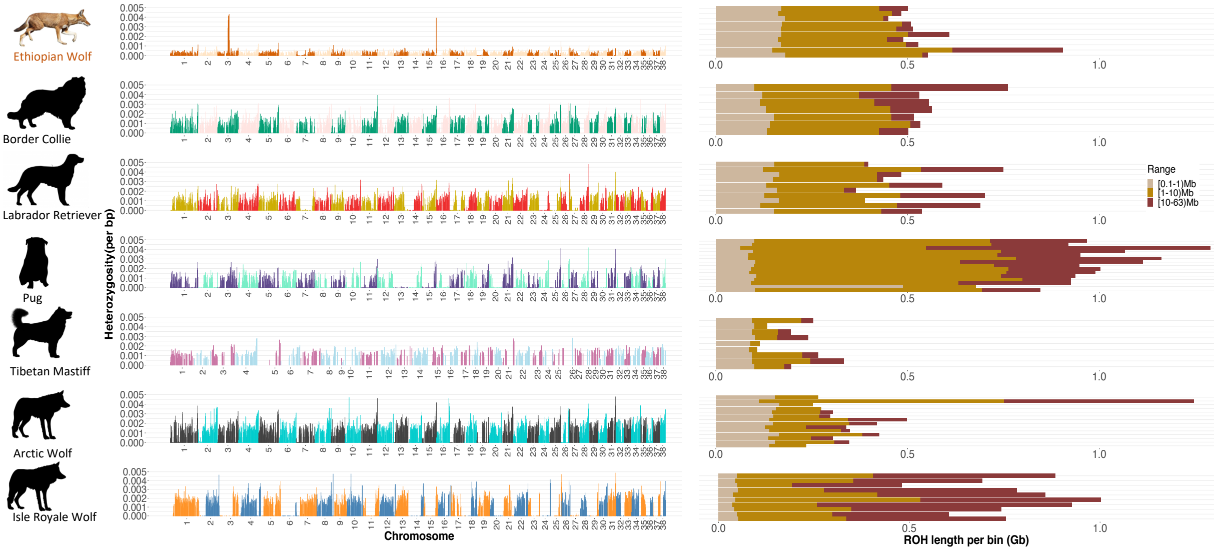


Figure 3.3 Ethiopian wolves have been at small size for an extended time.

The demographic history of the Ethiopian wolf was inferred using ABC and a three-epoch model of serial contractions. The inferred demographic history of the Ethiopian wolf, with point estimates (*maximum a posteriori* probability) of the population size and bottleneck times is shown below. **Supplementary Table 1** contains the credible intervals and point-estimates of each parameter.

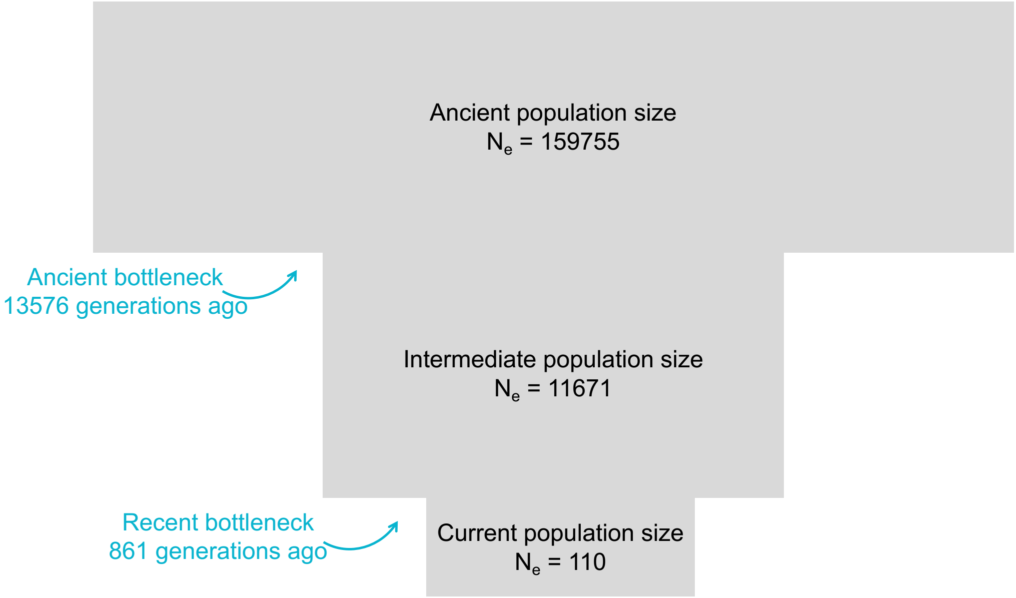
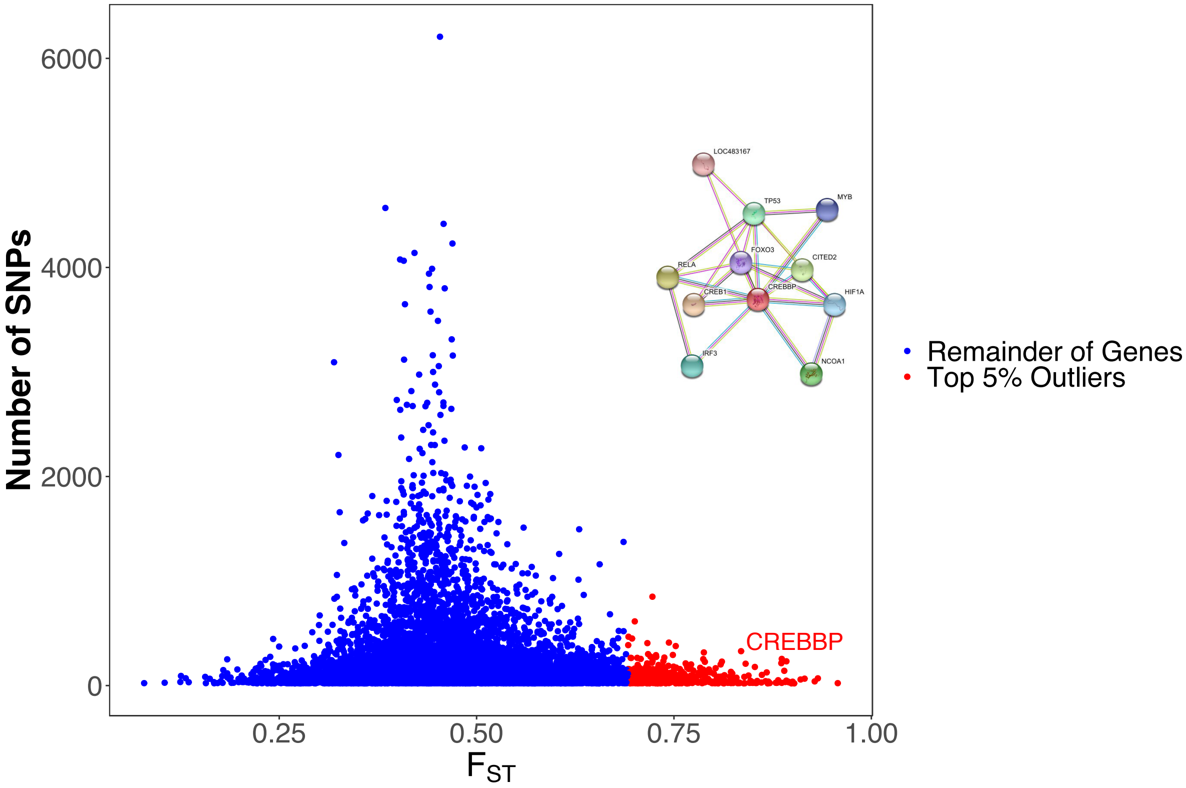


Figure 3.4 High-altitude adaptation.

F_{ST} between the Ethiopian wolf and Arctic wolf was computed for each gene, then normalized by SNP count. The top 5% F_{ST} outliers are shown in red and the remainder of the genome is blue. The primary protein-interaction network for CREBBP (represented by a red node) in *Canis lupus familiaris* is shown as an inset. Each node represents a protein-coding gene and the edges represent protein-protein associations that contribute to a shared function. The colors of each edge represent methods of validation of interactions (turquoise: curated database; purple: experimentally curated; green: text mining, black: co-expression).



3.6 References

1. Gilpin, M. E. Minimal viable populations: processes of species extinction. *Conservation biology: the science of scarcity and diversity* (1986).

2. Lande, R. & Barrowdough, G. Effective population size, genetic variation, and their use in population. *Viable populations for conservation* 87 (1987).

3. Lande, R. Risks of population extinction from demographic and environmental stochasticity and random catastrophes. *The American Naturalist* **142**, 911–927 (1993).

4. Caughley, G. Directions in conservation biology. *Journal of animal ecology* 215–244 (1994).

5. Frankham, R. Genetics and extinction. *Biological conservation* **126**, 131–140 (2005).

6. Lande, R. Genetics and demography in biological conservation. *Science* **241**, 1455–1460 (1988).

7. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96 (1973).

8. Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994 (2008).

9. Robinson, J. A., Brown, C., Kim, B. Y., Lohmueller, K. E. & Wayne, R. K. Purging of strongly deleterious mutations explains long-term persistence and absence of inbreeding depression in island foxes. *Current Biology* **28**, 3487–3494 (2018).

10. Sillero-Zubiri, C. & Gottelli, D. Diet and feeding behavior of Ethiopian wolves (Canis simensis). *Journal of Mammalogy* **76**, 531–541 (1995).

11. Gottelli, D., Marino, J., SILLERO-ZUBIRI, C. & Funk, S. M. The effect of the last glacial age on speciation and population genetic structure of the endangered Ethiopian wolf (Canis simensis). *Molecular Ecology* **13**, 2275–2286 (2004).

12. Whitby, J. E., Johnstone, P. & Sillero-Zubiri, C. Rabies virus in the decomposed brain of an Ethiopian wolf detected by nested reverse transcription-polymerase chain reaction. *Journal of wildlife diseases* **33**, 912–915 (1997).

13. Sillero-Zubiri, C., King, A. A. & Macdonald, D. W. Rabies and mortality in Ethiopian wolves (Canis simensis). *Journal of wildlife diseases* **32**, 80–86 (1996).

14. Randall, D. A. *et al.* Rabies in endangered Ethiopian wolves. *Emerging infectious diseases* **10**, 2214 (2004).

15. Randall, D. A. *et al.* An integrated disease management strategy for the control of rabies in Ethiopian wolves. *Biological Conservation* **131**, 151–162 (2006).

16. Laurenson, K. *et al.* Disease as a threat to endangered species: Ethiopian wolves, domestic dogs and canine pathogens. in *Animal Conservation forum* vol. 1 273–280 (Cambridge University Press, 1998).

17. Gordon, C. H. *et al.* Canine distemper in endangered Ethiopian wolves. *Emerging infectious diseases* **21**, 824 (2015).

18. Gottelli, D., Sillero-Zubiri, C., Marino, J., Funk, S. M. & Wang, J. Genetic structure and patterns of gene flow among populations of the endangered E thiopian wolf. *Animal Conservation* **16**, 234–247 (2013).

19. Marino, J., Sillero-Zubiri, C., Gottelli, D., Johnson, P. J. & Macdonald, D. W. The fall and rise of Ethiopian wolves: lessons for conservation of long-lived, social predators. *Animal Conservation* **16**, 621–632 (2013).

20. Group, I. C. S. *Strategic plan for Ethiopian wolf conservation*. (IUCN/SSC Canid Specialist Group Oxford, UK, 2011).

21. Marino, J. & Sillero-Zubiri, C. Canis simensis. *The IUCN Red List of Threatened Species* **10**, 2011–1 (2011).

22. Randall, D. A., Pollinger, J. P., Argaw, K., Macdonald, D. W. & Wayne, R. K. Fine-scale genetic structure in Ethiopian wolves imposed by sociality, migration, and population bottlenecks. *Conservation Genetics* **11**, 89–101 (2010).

23. Randall, D. A. *et al.* Inbreeding is reduced by female-biased dispersal and mating behavior in Ethiopian wolves. *Behavioral Ecology* **18**, 579–589 (2007).

24. Gottelli, D. *et al.* Molecular genetics of the most endangered canid: the Ethiopian wolf Canis simensis. *Molecular Ecology* **3**, 301–312 (1994).

25. Peterson, R. O., Vucetich, J. A., Bump, J. M. & Smith, D. W. Trophic cascades in a multicausal world: Isle Royale and Yellowstone. *Annual Review of Ecology, Evolution, and Systematics* **45**, 325–345 (2014).

26. Robinson, J. A. *et al.* Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Science Advances* **5**, eaau0757 (2019).

27. Gopalakrishnan, S. *et al.* Interspecific Gene Flow Shaped the Evolution of the Genus Canis. *Current Biology* **28**, 3441–3449 (2018).

28. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews Genetics* (2018).

29. Pemberton, T. J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *The American Journal of Human Genetics* **91**, 275–292 (2012).

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

31. Gray, M. M. *et al.* Linkage Disequilibrium and Demographic History of Wild and Domestic Canids. *Genetics* **181**, 1493–1505 (2009).

32. Pollinger, J. P. *et al.* A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome research* (2011).

33. Hedrick, P. W., Peterson, R. O., Vucetich, L. M., Adams, J. R. & Vucetich, J. A. Genetic rescue in Isle Royale wolves: genetic analysis and the collapse of the population. *Conserv Genet* **15**, 1111–1121 (2014).

34. Lohmueller, K. E. The distribution of deleterious genetic variation in human populations. *Current opinion in genetics & development* **29**, 139–146 (2014).

35. Simons, Y. B. & Sella, G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Current opinion in genetics & development* **41**, 150–158 (2016).

36. Marsden, C. D. *et al.* Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences* **113**, 152–157 (2016).

37. Gottelli, D. & Sillero-Zubiri, C. The Ethiopian wolf–an endangered endemic canid. *Oryx* **26**, 205–214 (1992).

38. Wang, M.-S. *et al.* Ancient Hybridization with Unknown Population Facilitated High Altitude Adaptation of Canids. *CURRENT-BIOLOGY-D-19-01554* (2019).

39. Witt, K. E. & Huerta-Sánchez, E. Convergent evolution in human and domesticate adaptation to high-altitude environments. *Philosophical Transactions of the Royal Society B* **374**, 20180235 (2019).

40. Miao, B., Wang, Z. & Li, Y. Genomic analysis reveals hypoxia adaptation in the Tibetan mastiff by introgression of the gray wolf from the Tibetan Plateau. *Molecular biology and evolution* **34**, 734–743 (2017).

41. Li, Y. *et al.* Population variation revealed high-altitude adaptation of Tibetan mastiffs. *Molecular biology and evolution* **31**, 1200–1205 (2014).

42. Shaywitz, A. J. & Greenberg, M. E. CREB: a stimulus-induced transcription factor activated by a diverse array of extracellular signals. *Annual review of biochemistry* **68**, 821–861 (1999).

43. Arany, Z. *et al.* An essential role for p300/CBP in the cellular response to hypoxia. *Proceedings of the National Academy of Sciences* **93**, 12969–12973 (1996).

44. Kallio, P. J. *et al.* Signal transduction in hypoxic cells: inducible nuclear translocation and recruitment of theCBP/p300 coactivator by the hypoxia-induciblefactor-1α. *The EMBO journal* **17**, 6573–6586 (1998).

45. Ginty, D. D. *et al.* Regulation of CREB phosphorylation in the suprachiasmatic nucleus by light and a circadian clock. *Science* **260**, 238–241 (1993).

46. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research* gkw937 (2016).

47. Scholz, H., Schurek, H.-J., Eckardt, K.-U. & Bauer, C. Role of erythropoietin in adaptation to hypoxia. *Experientia* **46**, 1197–1201 (1990).

48. Baker, Z. *et al.* Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *Elife* **6**, e24133 (2017).

49. Axelsson, E. *et al.* Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome research* **22**, 51–63 (2012).

50. Muñoz-Fuentes, V., Di Rienzo, A. & Vilà, C. Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. *PloS one* **6**, e25498 (2011).

51. Grey, C., Baudat, F. & de Massy, B. PRDM9, a driver of the genetic map. *PLoS genetics* **14**, (2018).

52. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome research* **14**, 988–995 (2004).

53. Sillero-Zubiri, C., Gottelli, D. & Macdonald, D. W. Male philopatry, extra-pack copulations and inbreeding avoidance in Ethiopian wolves (Canis simensis). *Behavioral Ecology and Sociobiology* **38**, 331–340 (1996).

54. Marchant, T. W. *et al.* Canine brachycephaly is associated with a retrotransposon-mediated missplicing of SMOC2. *Current Biology* **27**, 1573–1584 (2017).

55. Plassais, J. *et al.* Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nature communications* **10**, (2019).

56. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

57. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

58. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

59. Van der Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics* **43**, 11–10 (2013).

60. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

61. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

62. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

63. Chavez, D. E. *et al.* Comparative genomics provides new insights into the remarkable adaptations of the African wild dog (Lycaon pictus). *Scientific reports* **9**, 1–14 (2019).

64. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution* **35**, 1547–1549 (2018).

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

66. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology* **6**, (2010).

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

68. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655–1664 (2009).

69. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

70. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic epidemiology* **39**, 276–293 (2015).

71. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics* **98**, 127–148 (2016).

72. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).

73. Pouyet, F., Aeschbacher, S., Thiéry, A. & Excoffier, L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *Elife* **7**, e36317 (2018).

74. Schrider, D. R., Shanku, A. G. & Kern, A. D. Effects of linked selective sweeps on demographic inference and model selection. *Genetics* **204**, 1207–1223 (2016).

75. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics* **46**, 919 (2014).

76. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).

77. Phung, T. N., Wayne, R. K., Wilson, M. A. & Lohmueller, K. E. Complex patterns of sex-biased demography in canines. *Proceedings of the Royal Society B* **286**, 20181976 (2019).

78. Gou, X. *et al.* Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome research* **24**, 1308–1315 (2014).

Chapter 4: The impact of identity-by-descent on fitness and disease in natural and domesticated *canid* populations

4.1 Introduction

Identity-by-descent (IBD) segments are stretches of the genome that are inherited from a common ancestor and that are shared between at least two genomes in a population (Figure S1). Runs of homozygosity (ROH) form when an individual inherits the same segment of their genome identically by descent from both parents. Thus, ROH can be viewed as a special case of IBD, where IBD occurs within an individual rather than shared between individuals (1) (Figure S1). Recent consanguinity generates an increase in ROH, while decreases in population size can generate IBD regions.

Dogs provide an excellent model system for testing how ROHs and IBD patterns impact complex traits and reproductive fitness. The unique demographic and selective history of dogs includes a domestication bottleneck coupled with subsequent breed formation bottlenecks and strong artificial selection. This demography has allowed for the persistence of deleterious variation in their genome, simplified genetic architecture of complex traits, and an increase in both ROH and IBD segments within breeds (2–7). Specifically, the average F_{ROH} was approximately 0.3 in dogs (8), compared to 0.005 in humans (9). The large amount of the genome in ROHs in dogs, combined with a wealth of genetic variation and phenotypic data in dogs (3, 6, 8, 10–12) will allow us to test how these factors influence complex traits. Further, many of the deleterious alleles within dogs likely arose relatively recently within a breed, and dogs tend to share similar disease pathways and genes to humans (5, 13, 14), making our results relevant for complex traits in humans as well.

Despite IBD segments and ROHs being ubiquitous in genomes, the extent to which they affect the architecture of complex traits as well as reproductive fitness has remained elusive. Given that ROH are formed by inheritance of the same ancestral chromosome from both parents, there is a much higher probability of the individual to become homozygous for a deleterious recessive variant (9, 15) that was once carried in a heterozygous form, leading to a reduction in fitness. This prediction was verified in recent work in non-human mammals that has shown that populations suffering from inbreeding depression tend to have an increase in ROHs (16, 17). ROHs in human populations are enriched for deleterious variants, however the extent to which these variants impact phenotypes has not been demonstrated (9, 15, 18). Along these lines, several studies have associated an increase in ROHs with complex traits in humans (19–24), though some associations remain controversial (25–29). Determining how ROHs and IBD influence complex traits and fitness could provide a mechanism for complex trait architecture across populations that differ in IBD and ROH burdens.

Here, we use IBD segments and ROH from 4,741 breed dogs and 379 wolves to determine the recent demographic history of dogs and wolves and establish a connection between recent inbreeding and deleterious variation associated with both disease and inbreeding depression. This comprehensive dataset contains genotype data from 172 breeds of dog, village dogs from 30 countries, and gray wolves from British Colombia, North America, and Europe. We use IBD segments to infer the recent demographic history of these canids. In agreement with previous studies, we find that breed dogs experienced breed formation bottlenecks of varying degrees, and as expected, find no evidence of a breed formation bottleneck in village dogs or wolves (3). We test for an association with the burden of ROH and case-control status for a variety of complex traits. We find that an increase in ROH is associated with lymphoma, portosystemic vascular anomalies, and cranial cruciate ligament disease within breed dogs. Remarkably, we also find that the number of disease-associated causal variants identified in a breed is positively correlated with breed popularity rather than burden of IBD or ROH in the genome, suggesting ascertainment biases also exist in databases of dog disease mutations. Lastly, we identify multiple loci that may be associated with inbreeding depression by examining localized depletions of ROH across dog genomes.

4.2 Results

Global patterns of diversity across dogs and wolves

In order to examine diversity in dogs and wolves, we merged three previously published genotype array-based datasets (10–12). As an initial quality check, we used principal component analysis (PCA) to examine the relationship between domesticated dogs, village dogs, and wolves (Figure1A). We observed a split between dogs and wolves within the first PC. The dogs that fall closest to wolves trace their origins back to Australasia (Figure 1A), which has been previously shown to be the origin of some of the more ancient dog breeds (6, 30). When we performed a separate PCA with only dogs, dogs clustered by clade (Figure S2). We also observed separation based on the geographic location of wolf populations, which originate from populations from Europe or North America. A PCA with only wolves showed clear clustering by the location of where samples originated from either the United States, Mexico, or Europe (Figure S3).

Breed-specific bottlenecks are captured in IBD

We next turned our attention to IBD patterns and called IBD segments using IBDSeq (31). We then tested whether the recent demographic history of dogs, consisting of breed formation bottlenecks within the last few hundred years (7, 14, 32), had an impact on the IBD patterns. Although the demographic history of dogs has been well studied over the years, the majority of these works have focused on the origin of dogs thousands of years ago, and their geographic origin remains an ongoing point of contention (6, 30, 33–36). We inferred the recent demographic history of 10 standard breeds of dog, village dogs, and gray wolves from North America and Europe (Figure 1B) using patterns of IBD sharing between individuals (37). When conducting demographic inferences using IBDNe, we restricted our analyses to populations with at least 50 individuals (see Discussion). The IBDNe analyses show that all breed dogs have experienced a domestication bottleneck followed by another severe bottleneck approximately 200 years ago which corresponds with modern breed formation during the Victorian Era (1800s) (Figure 1B). Though the strength of breed formation bottleneck varied across breeds, and was less pronounced in mixed breeds, all bottlenecks were followed by a subsequent increase in population size. Notably, the Maltese and Rottweiler appeared to have undergone the most severe bottlenecks with the Golden retriever and German Shepherd dog close behind. The boxer also seemed to have experienced a severe bottleneck, but this bottleneck may be linked to reference bias (CanFam3.1 reference genome is from a boxer) as the confidence intervals are also largest for the boxer. As we expected, the village dog (feral street dogs) and gray wolves do not appear to have experienced the domestication bottleneck which is concordant with their history (3). Instead, the village dogs showed a much weaker prolonged bottleneck followed by an increase in population size (Figure 1B). The European gray wolves have a demographic trajectory similar to the village dogs, and the American gray wolves appear to have just experienced a prolonged decline in population size (Figure 1B), which matches recent ecological studies (10, 38, 39). Thus, the formation of dog breeds in the last 200 years has increased the amount of IBD within dogs compared to wolves and village dogs.

Long ROH are enriched in most breed dogs

Given that ROH reflect recent common ancestry between the two chromosomes carried by an individual (Figure S1) and the inter-breed differences in the strength of bottlenecks (Figure 1B) during the breed formation period, we next sought to examine the burden (total amount) of long ROH (greater than 2Mb) across breeds. We observed that F_{ROH}, the proportion of the genome within a long ROH, varies across breed and by extension clade (Figure S4). The majority of breed dogs contained a larger amount of their genome in ROH and thus a larger average value of F_{ROH}, a likely consequence of having experienced both a domestication and breed formation bottleneck as well as inbreeding (40). The Jack-Russell terrier was the exception and had reduced ROH relative to other breed dogs, which corresponds to previous works where it was found to be an outlier (6, 7). In village dogs, we observed that mean values of F_{ROH} fall much closer to what we observed in the wolves. This is expected since village dogs only experienced the domestication bottleneck and were left to reproduce without selective breeding. Lastly, we examined ROH among wolves, the European gray wolf has mean values of F_{ROH} that are comparable to the village dogs and markedly lower than the American gray wolves. The American gray wolf exhibited increased mean values of F_{ROH}, likely due to having experienced a recent bottleneck (Figure 1B) as a result of being pushed to near extinction (41).

Disease traits are associated with ROH burden

We hypothesize that the prevalence of ROH and identity-by-descent (IBD) segments could be associated with recessive genetic disease in each breed (Figure S1). ROH form when an individual inherits the same segment of their genome identically by descent from both parents (1), and the formation of ROH results in an increased probability of the individual to be homozygous for a deleterious recessive variant (15, 42). Thus, we predict that breeds with large amounts of ROH and IBD segments will have an increased incidence of recessive-disease. We tested this hypothesis using data from 4,342 dogs where we had case-control status for subsets of the data across 8 clinical and morphological phenotypes (Figure 3). For the majority of traits, there was not a significant association with ROH burden, even when stratified by breed, however we do note there are a larger number of associations than we would expect under the null (Figure 3). We observed a significant association between the burden of ROH and case-control status for five traits: portosystemic vascular anomalies (PSVA) in Yorkshire terriers ($\beta = -0.394$ & $p < 0.027$), lymphoma within both Labrador (β = -0.604 & $p < 0.0340$) and Golden retrievers (β = 0.913 & $p < 0.001$), cranial cruciate ligament disease (CLLD) in Labrador retrievers (β = -0.403 & $p < 0.003$), elbow dysplasia (ED) across all breeds (β = 0.238 & $p < 0.047$), and mast cell tumors (MCT) across all breeds (β = 0.286 & $p < 0.027$). For lymphoma in golden retrievers, case-status is positively associated with the amount of the genome within an ROH (OR = 2.491, Table S1) and on average cases carried more ROH than controls (Figure S5). Conversely, ROH appeared to show a protective effect against developing PSVA in Yorkshire terriers, or CLLD and lymphoma in Labrador retrievers.

Breed popularity, rather than ROH, correlates with Mendelian disease incidence. If dogs with more ROH and IBD segments have a higher incidence of genetic diseases, we hypothesize that dogs with the largest amount of ROH and/or IBD would carry the most disease associated variants. We tested this hypothesis using data from Online Mendelian Inheritance in Animals (OMIA) which included a count of causal variants identified in each breed. We observed that breeds with the lowest amounts of ROH/IBD have the most identified causal variants (Figure 4 B and C). Further, those breeds with the most ROH/IBD have no causal variants identified, apart from the Kerry Blue terrier that has been used in a single study (Figure 3 A and B). This finding was unexpected, and so we sought additional factors that might explain the variation in the number of OMIA variants per breed. We chose to examine the popularity of different dog breeds over time using data compiled by the American Kennel Club (AKC). We observed a strong positive correlation between the overall breed popularity with the number of causal variants identified in each breed ($R^2 = 0.168$ & $p = 1.145 \times 10^{-06}$) (Figure 3C). We find that the most popular breeds, such as the retrievers, have the most causal variants identified in genomic studies. Given that we do not observe a positive relationship between IBD or ROH and the total number of causal variants in a breed, and that breeds with excess amounts of IBD and/or ROH have almost no causal variants identified, our results indicate that there are large-scale ascertainment biases in

OMIA reported disease associated variants in dogs. More causal variants have been identified in the more popular breeds, containing fewer ROH, rather than those breeds with an increased prevalence of the associated Mendelian disease containing more ROH. Our results suggest that there are many understudied breeds that may be valuable for variant discover such as the Bearded collie, Belgian Sheep dog, Bedlington terrier, or Dogue de bordeaux, and these breeds are prone to serious health conditions according to the American Kennel Club (akc.org/dog-breeds).

ROH reveal genes with recessive lethal mutations

Given the relatively high values of F_{ROH} observed for breed dogs, much of the genome should be in in ROH in at least one individual. We hypothesized the genes not contained within a ROH in any individual or showing a deficit of ROH compared to the rest of the genome, contain recessive lethal variants, because individuals homozygous for these mutations would not be viable. Across 4,342 dogs, we observed 27 genes where at least one exon does not overlap a ROH in any individual. To test whether this is unusual, we permuted the locations of the ROHs within each individual and re-counted the number of genes with an exon not containing a ROH. We found that if ROHs were randomly distributed across the genome, we would expect to see ROH in all exons across genes (Figure S6). Thus, there are more genes not overlapping ROHs than expected by chance ($p < 0.0001$) suggesting the presence of segregating recessive lethal mutations across breed dogs.

We next intersected these 27 genes with the 90th percentile constrained coding regions (CCRs) identified in human populations (43). CCRs were found to be enriched for disease-causing variants, especially in dominant Mendelian disorders, and authors suggested the 95th percentile CCRs may be enriched for embryonic lethal mutations (43). We expect that genes containing recessive lethal mutations would be conserved across species. We tested whether the genes not overlapping ROH were enriched for CCRs. On-average, one would expect to see 18 genes above the 95th percentile CCRs. However, we observed that 23 out of 27 of our genes of interest fall above the 95th percentile of the CCR distribution (p = 0.025) (Figure 4). Additionally, we observed a 2.94-fold enrichment of non-ROH genes relative to ROH genes in CCRs (p = 0.041) (Figure 4). Taken together, these results suggest that more of the genes with an exon not in a ROH are the exons devoid of variation in humans. Thus, these genes may be targets of strongly deleterious mutations affecting viability.

We also tested whether our ROH analyses could be affected by low single nucleotide polymorphism (SNP) density, since the analyses thus far used only SNP genotype data. Because whole-genome sequence data would have an increased density of SNPs, we repeated our analyses using two sets of sequence data. The first dataset represents samples from four different breeds of dog: Pug (N=15) with approximately 47X coverage (44), Labrador retriever (N=10) with approximately 30X coverage (45), Tibetan mastiff (N=9) with approximately 15X coverage (45), Border collie (N=7) with approximately 24X coverage (45). The second dataset was previously published (see 16), and contains 220 samples from human populations. We find that of the 27 genes with at least one exon not overlapping an ROH in any dog, three genes ANKH, FYTTD1, and PRMT2 have exons not overlapping an ROH in all three data sets (Table S2). One of these genes, ANKH, has known Mendelian phenotypes that have been reported in Online Mendelian Inheritance in Man (OMIM) and is also a 95th percentile CCR (47). Another caveat to these results is the location of these three genes ANKH, FYTTD1, and PRMT2, which all reside towards the end of the chromosome in dogs (Figure S8). Nevertheless, the relative distribution of ROHs and these three genes not containing ROH were concordant across both VCFTools and PLINK (Figure S7).

4.3 Methods

Genomic data

Genotype data were aggregated from two published studies (10, 11), and all original data files are publicly available through Dryad. The Fitak et al. data (10) was lifted over to CanFam3.1 then merged with Hayward et al. data using PLINK (65). SNPRelate (66) was used to perform PCA (Figure 1), identify duplicate individuals, and unrelated individuals. Duplicate individuals and potential hybrid individuals were removed from the data set, and the final data set contained 4,741 breed dogs and 379 wolves. Code for merging data is available at https://github.com/jaam92/DogProject_Jaz/tree/master/MergeFitkakAndCornell.

American Kennel Club (AKC) data

We used AKC registration data from 1926 to 2005 to compute breed popularity. This data was curated from previous work (67) and contains information for approximately 150 recognized breeds. To compute popularity through time, we drop the first entry for each breed, as this number reflects older dogs and new litters, then use the remaining data as the total number of new registrants per year. The popularity score is defined as the integral from the second entry through 2005.

Online Mendelian Inheritance in Animals (OMIA) data

We downloaded all Likely Causal variants listed on OMIA. However, only causal variants associated with disease were used. The Likely Causal criteria is met if there is at least one publication to be listed where the variant is associated with a disorder. If a variant had been identified in multiple breeds, it was counted in each breed. The total number of causal variants per breed downloaded from OMIA is available here: https://github.com/jaam92/DogProject_Jaz/tree/master/LocalRscripts/OMIA.

Calling identity-by-descent (IBD) segments and inferring effective population size

To call IBD segments we used software IBDSeq (31) on its default settings. The IBD segments were input into software IBDNe (37) to infer effective population size. A pedigree based recombination map (68) was used as the input genetic map for IBDNe. We used the default settings in IBDNe but set the minimum IBD segment length to 4cM, as that is the suggested length to reliably call IBD segments in genotype data when using IBDSeq. We only inferred effective population sizes in populations with at least 50 unrelated individuals and assumed a conversion rate of 3 years per generation for visualizing results. We use at least 50 unrelated individuals because previous work has shown that demographic trajectories are robust to smaller sample sizes, though accurate estimates of effective population size (N_e) remain limited (9).

Calling runs of homozygosity (ROH)

VCFTools, which implements the procedure from Auton et al. 2009 (69), was used to call ROH in all individuals. Next, we performed quality control of the raw ROH. We only kept ROHs that contained at least 50 SNPs, were at least 100 kilo-bases long, and where SNP coverage was within one standard deviation of mean SNP coverage across all remaining ROH. A file that contains the final ROHs and scripts for running quality control can be found here: https://github.com/jaam92/DogProject_Jaz/tree/master/LocalRscripts/ROH. We also called ROH using PLINK, with the following parameters: --homozyg-window-het 0 --homozyg-snp 41 --homozyg-window-snp 41 --homozyg-window-missing 0 --homozyg-window-threshold 0.05 --homozyg-kb 500 --homozyg-density 5000 --homozyg-gap 1000 and then repeated the filtering listed above for VCFTools ROH.

Computing IBD and ROH scores

We computed each population's IBD and ROH scores using an approach similar to Nataksuka et al. (70). A population's IBD score was calculated by computing the total length of all IBD segments between 4 and 20 cM and normalizing by the sample size. A population's ROH score was computed using all ROH that passed quality control and normalizing by the sample size.

Association test and effect size estimates

For these analyses, we only used the subset of breed dog data from Hayward et al. where we had phenotype information (11). We computed the association between F_{ROH} and each trait using a generalized linear mixed model which is implemented in the R package GMMAT (71). Following the protocol from Hayward et al., we did not include covariates in the association test, and included the kinship matrix as a random effect in the model, to control for population stratification due to co-variation of the amount of ROH per breed with the incidence of the phenotype in the breed. P-values were determined using a Wald test with a significance threshold of p = 0.05. We do note that there is no correction for multiple testing here. For more details on clinical trait ascertainment see (11). We generated the kinship matrix two different ways: 1) Using R package PC-Relate (72) on the SNP genotype matrix and 2) By computing the total amount of the genome within a ROH that is shared between two individuals, shared ROH (SR).

$$\text{SR} = \sum_{j=1}^{i} X_{Gj}$$

Where $X_{Gj} \in \{0, 1\}$ is whether the genotype at the j^{th} base pair falls within a ROH shared by both individuals. SR was computed for each pair of individuals and bound between 0 (no sharing) and 1 (complete sharing with oneself). Results reported in the main text use the kinship matrix computed from ROH sharing. We also compared these results to those when not using a kinship or ROH matrix (Figure S10).

Identifying depletions of ROH

To find the number of genes expected to contain at least one exon without a ROH, ROH in each individual were permuted to a new location on the same chromosome using BEDTools shuffle. Next, we created a bed file containing the permuted ROH locations and intersected this file with the exon locations from CanFam3.1 and counted the number of genes with at least one exon where we did not observe any overlap with an ROH. We repeated our permutation test 10,000 times to create a null expectation. To examine the overlap between regions lacking ROH and constrained coding regions (CCRs) from Havrilla et al. (43) we used BEDTools (73) to intersect ROH with the top 10% of CCRs and exon ranges for CanFam3.1 (52) which came from Ensembl (74). Then, we tabulated the total number of genes where there was at least one exon where we never observed any overlap (including partial overlap) with a ROH (non-ROH genes) and the converse (ROH genes), as well as the count of whether these non-ROH and ROH genes fell within a CCR. Significance of the ratio of non-ROH genes relative to ROH genes within a CCR was assessed using a Fisher's Exact test. We also computed the expected number of non-ROH genes within the 90th percentile CCRs by randomly sampling an equal number of genes from the entire gene set and intersecting the randomly sampled genes with CCRs. We repeated this random sampling 100,000 times to build the null expectation.

4.4 Discussion

Here we show how the population history of dogs has increased the number of regions of the genome carried in ROHs and IBD segments, affecting phenotypes and fitness. Our work contributes to a burgeoning number of studies associating ROH burden with complex traits and is one of the first studies to directly show this association in dogs (8, 40, 48). Dogs provide an excellent model to examine the connection between ROH burden and disease due to their unique demographic history, as well as their simplified trait architecture due to artificial selection. Further, one can avoid many of the challenges encountered when using human data (need large sample size, correcting for confounding due to socio-economic status, religion, educational attainment etc.) (11, 40, 49). For example, religion has been shown to confound associations with ROH burden and major depressive disorder (50) and small sample size was shown to be a potential reason for replication failures when examining the association of ROH burden and Schizophrenia (25).

Dog breeds were initially formed through domestication of one or more ancestral wild populations, in a process involving population bottlenecks. Then, over the last two hundred years or so, modern dog breeds were formed (2, 4, 6, 11, 30, 51–53). This breed formation process resulted in additional population bottlenecks. Our results confirm that the severity of the breed formation bottleneck varies (Figure 1), which to our knowledge has only been examined one other time using genetic variation data (54). Further our results seem to match historical records quite well. For example, the Rottweiler appears to have experienced one of the most severe bottlenecks (Figure 1), and records document the Rottweiler almost disappearing in the early 1900's and subsequently being revived by a handful of individuals (55). Conversely in breeds with a less severe bottleneck shown by the IBDNe analysis, such as American cocker spaniels, there is a corresponding breed history that does not include a population crash, instead noting long-term breed popularity (55). Lastly, it is quite apparent that village dogs and wolves did not experience the domestication bottleneck, which we expected since they were not domesticated (5). Instead, we see much weaker recent bottlenecks in the wolves, which are likely connected to anthropogenic factors such as hunting and habitat fragmentation. While estimates of current N_e may be inaccurate due to low sample size, the shape trajectory of N_e has been shown to be more robust to the low sample size (9). Thus, our work demonstrates how recent demographic history has affected patterns of IBD and ROH within modern dog breeds.

We hypothesized that the increase in IBD and ROH in certain dog breeds would have led to an increase in the presence of recessive Mendelian diseases. Thus, we expected to observe a positive correlation between the IBD and ROH carried within a breed and the number of causal variants identified in OMIA, due to the increased probability of revealing fully or partially recessive mutations due to excess homozygosity. Instead, we found a negative association between IBD and ROH scores and the number of causal variants identified (Figure 3). Interestingly, many of the breeds with the largest amounts of IBD and ROH have had no causal variants identified through 2020. Instead, we found a positive correlation with breed popularity and the number of causal variants identified. This counterintuitive result could be caused by 1) increased numbers of popular breed dogs seen in veterinary offices, 2) increased funding and genomic studies of disease in popular breeds (through clubs or direct-to-consumer genomics) 3) a combination of both.

Given this ascertainment bias that we have observed in OMIA, researchers may want to shift their focus to some of these understudied breeds, as there may be more potential to discover new disease-associated variants. Ascertainment bias is not unique to OMIA and has been observed in human databases like OMIM (56). In the case of human data, authors found that OMIM contains an enrichment of diseases caused by high frequency recessive-alleles. They suggested that the bias is caused by the method through which these variants are identified. Many variants have been identified in isolated human populations, where there may be elevated levels of relatedness, which increases the probability of mapping higher frequency deleterious variants (56).

We find that, for some traits, increased homozygosity of low frequency variants can impact phenotype. Low-frequency variants harbored in ROH likely become more important in the context of inbreeding and their net-effect can lead to severe inbreeding depression. Because purebred dogs have had severe inbreeding, in a large sample of dogs (>4,000 individuals), ROHs are expected to cover the entire genome in at least one individual (Figure S7). By searching for regions of the genome devoid of ROHs across all individuals, we can potentially identify genes containing strongly deleterious mutations, possibly underlying inbreeding depression. These regions could be lacking ROH because strongly deleterious recessive mutations lurk as heterozygotes in the founders of the breed. Then, individuals that are homozygous for these regions are no longer viable and are not sampled in our study. Similar to what has been shown in Scandinavian wolves (17), we find that there are multiple exons spanning 27 different genes where there are no ROH across all 4,300 dogs. Further, we observe that 23 of these genes were within the top 90th percentile CCRs. Overall, CCRs tend to be enriched for pathogenic variants linked to clinical phenotypes (43). However, there are some CCRs that did not have any known pathogenic or likely pathogenic variants suggesting mutations in these exons could cause extreme developmental disorders or potentially be embryonic lethal (43). Therefore, studying mutations in these exons without ROH that overlap the CCR data could be helpful both for identifying new disease phenotypes and for identifying variants with large fitness effects that could be linked to inbreeding depression or embryonic lethality.

We examined the locations of the exons without ROHs. We find that ROH tend to not occur at all within the exons of some genes or occur at exons toward the end of the gene (Figure S9). Perhaps, the location of ROH depletions could be related to nonsense-mediated decay. For example, deleterious variants within these ROH depleted regions could have large effects on gene expression, through nonsense-mediated decay, if they were to become homozygous. Previous work has suggested there is a connection between nonsense-mediated decay and large-effect low frequency variants that disrupt splicing and potentially alter mRNA stability (57–59). Further, nonsense-mediated decay has been shown to play a role in disease through numerous mechanisms such as patterns of inheritance, modulating the disease phenotype, and causing different traits to manifest from mutations in the same gene (60).

Our findings have implications for understanding the architecture of complex traits in other species, such as humans. Specifically, the fact that we find a relationship between ROH and certain phenotypes (Figure 2), suggests that recessive mutations play a role in some traits. Much of the existing genome-wide association studies (GWAS) in humans have largely suggested that complex traits are highly polygenic with many additive effects (61–64). These differences across species likely reflect differences in genetic architecture driven by the demographic history of the populations combined with natural selection. Nevertheless, searching for recessive variants underlying complex traits in humans may be a fruitful avenue of research. Further, variation in the amount of the genomes in ROHs across human populations (9, 15, 18, 42), could lead to population-specific architectures for complex traits. For example, causal variants in populations with a higher burden of ROHs may be more recessive and less polygenic than in populations with fewer ROHs. Future research will be needed to fully elucidate which mutations are directly responsible for severe inbreeding depression and the functional impact of these deleterious mutations. Additional work could examine which models of trait architecture (deleterious variants segregating at a low frequency in the population, additive model, recessive model), and demography could generate the association with ROH burden that we detected. In conclusion, the joint analysis of IBD and ROH can provide considerable information about both demography and selection in the genome. This information is especially valuable in the context of fitness and disease and allows us to shed light on recent population history.

4.5 Figures

Figure 4.1 Genetic variation data reveals the recent history of canids.

A) PCA of breed dogs, village dogs, and wolves. The 350 village dogs were sampled from 32 different countries around the world and the 379 wolves were sampled from populations across Europe and North America. **B)** Effective population size (N_e) trajectories through time of breed dogs, village dogs, and wolves were inferred using IBDNe, when there were at least 50 unrelated samples per population (*Methods*). Shaded regions in the plots indicate the 95% confidence interval of the inferred population size at each time point.

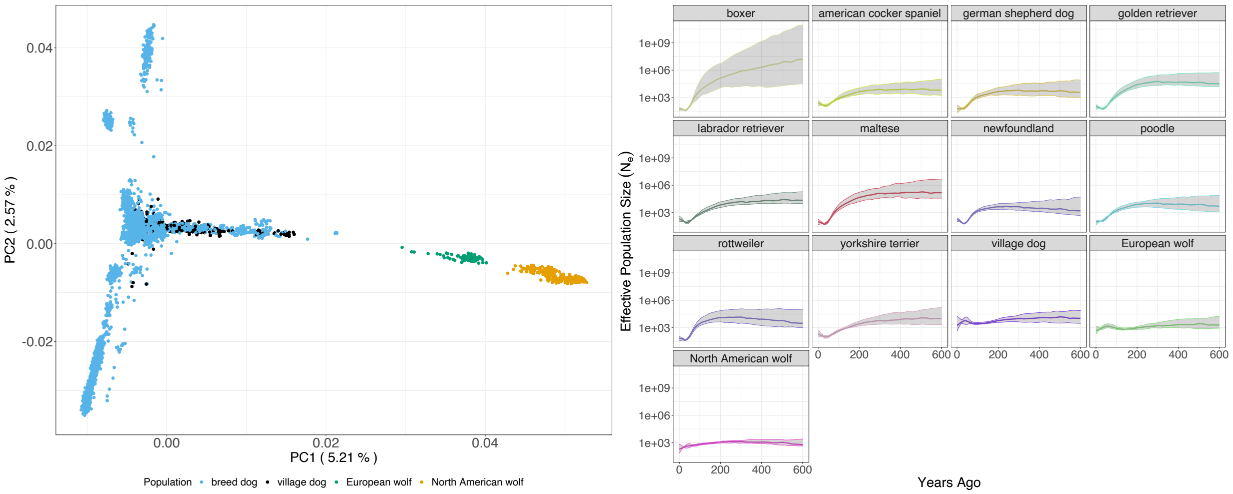


Figure 4.2 Association between runs of homozygosity (ROH) and phenotypes.

This figure shows the effect of ROH burden on eight quantitative traits, results are presented both stratified by breed and across all breeds. A significant effect of ROH burden on a trait ($p < 0.05$, note there is no correction for multiple testing here) is indicated with a red point. An effect size greater than 0 indicates an increase in ROH with the trait or disease status, and less than 0 represents the converse. Phenotype abbreviations: portosystemic vascular anomalies (PSVA); mitral valve degeneration (MVD); mast cell tumor (MCT); elbow dysplasia (ED); and cranial cruciate ligament disease (CCLD).

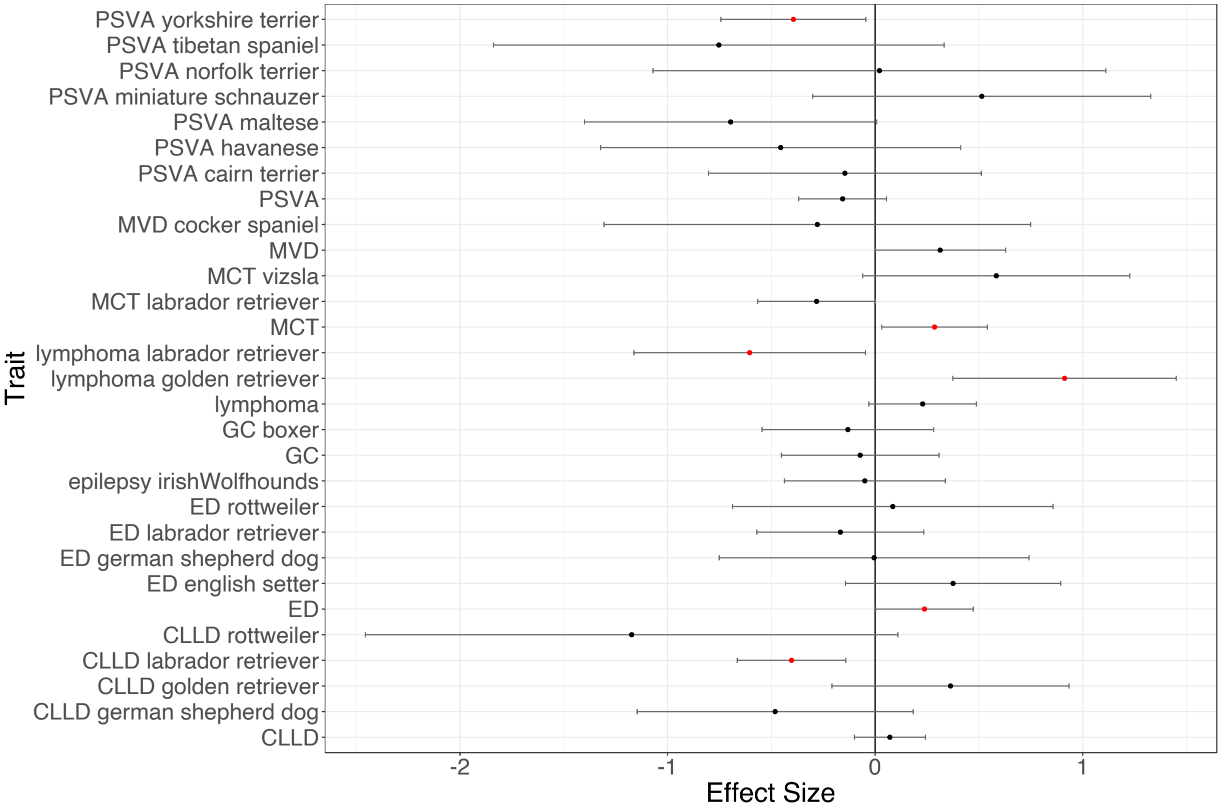


Figure 4.3 Breed popularity is positively correlated with number of causal variants identified in OMIA.

Here we show the correlation between causal variants identified in each breed that have been reported in OMIA and three different metrics. The shaded regions in each plot represent the confidence interval. **A)** The correlation between within-breed ROH and the total number of causal variants associated disease identified in the breed that have been reported in OMIA. **B)** The correlation between within-breed IBD and the total number of causal variants associated with disease identified in the breed that have been reported in OMIA. **C)** The correlation between breed popularity over time and the total number of causal variants associated with disease.

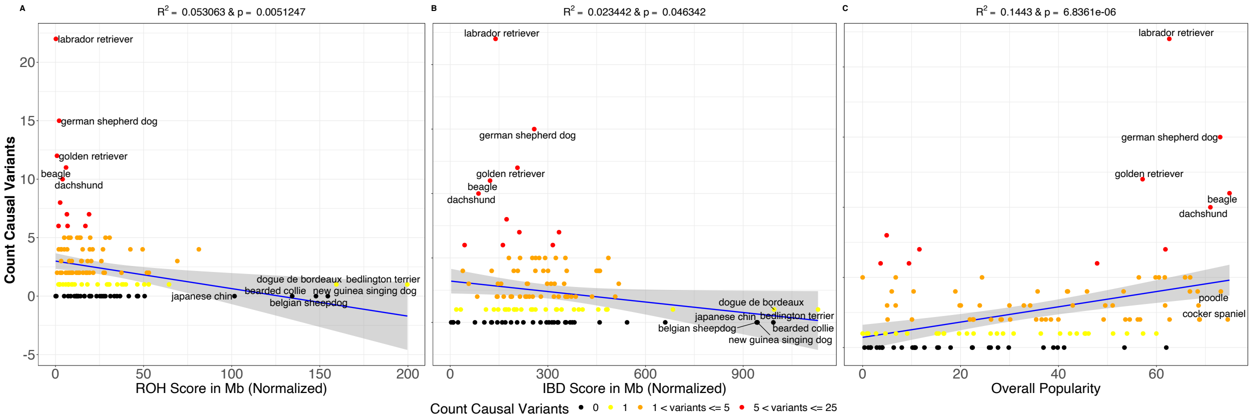
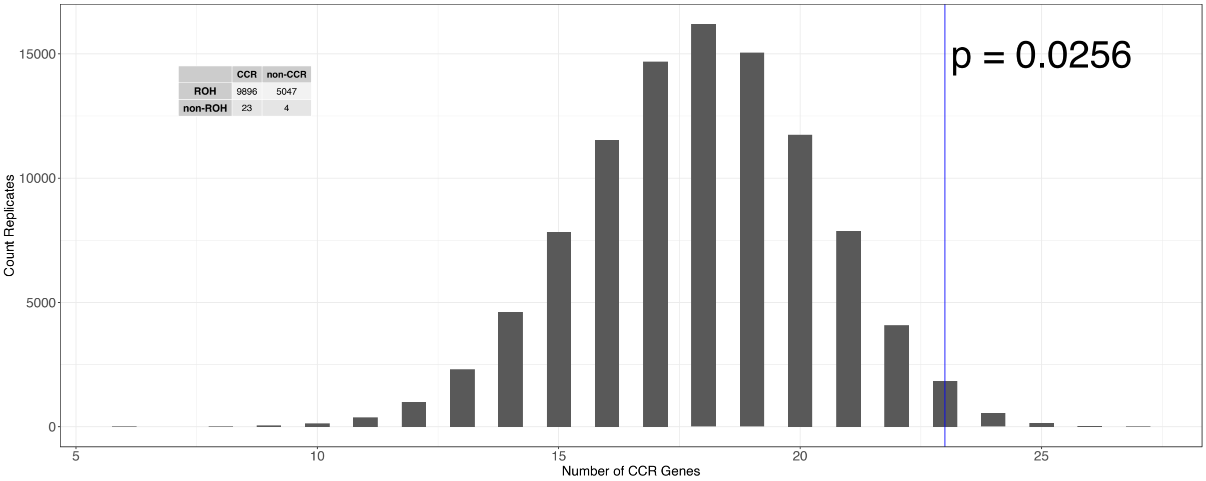


Figure 4.4 Histogram of the expected number of genes that fall into the top 10% constrained coding regions (CCRs) over 100,000 replicates, given a random sample equal to the number of genes of with a least one exon that does not contain an ROH.

The empirical data is demarcated by the blue line (p = 0.025). The contingency table shows the count of genes classified as either ROH or non-ROH and CCR or non-CCR. We observed a 2.94-fold enrichment of genes with at least one exon without an ROH, non-ROH genes, in CCRs (p = 0.041) relative to genes with an ROH.



	CCR	non-CCR
ROH	9896	5047
non-ROH	23	4

4.6 References

1. R. McQuillan, *et al.*, Runs of homozygosity in European populations. *The American Journal of Human Genetics* **83**, 359–372 (2008).

2. C. D. Marsden, *et al.*, Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences* **113**, 152–157 (2016).

3. A. R. Boyko, *et al.*, A simple genetic architecture underlies morphological variation in dogs. *PLoS biology* **8**, e1000451 (2010).

4. A. H. Freedman, K. E. Lohmueller, R. K. Wayne, Evolutionary history, selective sweeps, and deleterious variation in the dog. *Annual Review of Ecology, Evolution, and Systematics* **47**, 73–96 (2016).

5. A. R. Boyko, The domestic dog: man’s best friend in the genomic era. *Genome biology* **12**, 216 (2011).

6. B. M. vonHoldt, *et al.*, Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**, 898–902 (2010).

7. H. G. Parker, *et al.*, Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell reports* **19**, 697–708 (2017).

8. A. J. Sams, A. R. Boyko, Fine-Scale Resolution of Runs of Homozygosity Reveal Patterns of Inbreeding and Substantial Overlap with Recessive Disease Genotypes in Domestic Dogs. *G3: Genes, Genomes, Genetics* **9**, 117–123 (2019).

9. J. A. Mooney, *et al.*, Understanding the Hidden Complexity of Latin American Population Isolates. *The American Journal of Human Genetics* **103**, 707–726 (2018).

10. R. R. Fitak, S. E. Rinkevich, M. Culver, Genome-Wide Analysis of SNPs Is Consistent with No Domestic Dog Ancestry in the Endangered Mexican Wolf (Canis lupus baileyi). *Journal of Heredity* **109**, 372–383 (2018).

11. J. J. Hayward, *et al.*, Complex disease and phenotype mapping in the domestic dog. *Nature Communications* **7**, 10460 (2016).

12. A. V. Stronen, *et al.*, North-South Differentiation and a Region of High Diversity in European Wolves (Canis lupus). *PLOS ONE* **8**, e76454 (2013).

13. T. Awano, *et al.*, Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. *Proceedings of the National Academy of Sciences* **106**, 2794–2799 (2009).

14. A. L. Shearin, E. A. Ostrander, Leading the way: canine models of genomics and disease. *Disease Models & Mechanisms* **3**, 27–34 (2010).

15. Z. A. Szpiech, *et al.*, Long runs of homozygosity are enriched for deleterious variation. *The American Journal of Human Genetics* **93**, 90–102 (2013).

16. J. A. Robinson, *et al.*, Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Science Advances* **5**, eaau0757 (2019).

17. M. Kardos, *et al.*, Genomic consequences of intensive inbreeding in an isolated wolf population. *Nature ecology & evolution* **2**, 124 (2018).

18. Z. A. Szpiech, *et al.*, Ancestry-dependent enrichment of deleterious homozygotes in runs of homozygosity. *The American Journal of Human Genetics* (2019).

19. M. C. Keller, *et al.*, Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS genetics* **8** (2012).

20. G. Assié, T. LaFramboise, P. Platzer, C. Eng, Frequency of germline genomic homozygosity associated with cancer cases. *Jama* **299**, 1437–1445 (2008).

21. M. D. Bacolod, *et al.*, The signatures of autozygosity among patients with colorectal cancer. *Cancer research* **68**, 2610–2621 (2008).

22. T. Lencz, *et al.*, Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proceedings of the National Academy of Sciences* **104**, 19942–19947 (2007).

23. M. Ghani, *et al.*, Association of long runs of homozygosity with Alzheimer disease among African American individuals. *JAMA neurology* **72**, 1313–1323 (2015).

24. R. McQuillan, *et al.*, Evidence of inbreeding depression on human height. *PLoS genetics* **8** (2012).

25. E. C. Johnson, *et al.*, No reliable association between runs of homozygosity and schizophrenia in a well-powered replication study. *PLoS genetics* **12**, e1006343 (2016).

26. S. L. Spain, J.-B. Cazier, R. Houlston, L. Carvajal-Carmona, I. Tomlinson, Colorectal cancer risk is not associated with increased levels of homozygosity in a population from the United Kingdom. *Cancer research* **69**, 7422–7429 (2009).

27. V. Enciso-Mora, F. J. Hosking, R. S. Houlston, Risk of breast and prostate cancer is not associated with increased homozygosity in outbred populations. *European Journal of Human Genetics* **18**, 909–914 (2010).

28. A. K. Siraj, *et al.*, Colorectal cancer risk is not associated with increased levels of homozygosity in Saudi Arabia. *Genetics in medicine* **14**, 720–728 (2012).

29. F. J. Hosking, *et al.*, Genome-wide homozygosity signatures and childhood acute lymphoblastic leukemia risk. *Blood, The Journal of the American Society of Hematology* **115**, 4472–4477 (2010).

30. L. M. Shannon, *et al.*, Genetic structure in village dogs reveals a Central Asian domestication origin. *Proceedings of the National Academy of Sciences* **112**, 13639–13644 (2015).

31. B. L. Browning, S. R. Browning, Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data. *The American Journal of Human Genetics* **93**, 840–851 (2013).

32. F. C. Calboli, J. Sampson, N. Fretwell, D. J. Balding, Population structure and inbreeding from pedigree analysis of purebred dogs. *Genetics* **179**, 593–601 (2008).

33. C. Vilà, *et al.*, Multiple and ancient origins of the domestic dog. *Science* **276**, 1687–1689 (1997).

34. O. Thalmann, *et al.*, Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. *Science* **342**, 871–874 (2013).

35. P. Skoglund, E. Ersmark, E. Palkopoulou, L. Dalén, Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology* **25**, 1515–1519 (2015).

36. G.-D. Wang, *et al.*, Out of southern East Asia: the natural history of domestic dogs across the world. *Cell research* **26**, 21 (2016).

37. S. R. Browning, B. L. Browning, Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics* **97**, 404–418 (2015).

38. P. W. Hedrick, P. S. Miller, E. Geffen, R. Wayne, Genetic evaluation of the three captive Mexican wolf lineages. *Zoo Biology: Published in affiliation with the American Zoo and Aquarium Association* **16**, 47–69 (1997).

39. L. E. Harding, *et al.*, Genetic management and setting recovery goals for Mexican wolves (Canis lupus baileyi) in the wild. *Biological conservation* **203**, 151–159 (2016).

40. F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, J. F. Wilson, Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews Genetics* (2018).

41. J. A. Leonard, C. Vilà, R. K. Wayne, Legacy Lost: Genetic variability and population size of extirpated US gray wolves. *Molecular Ecology*, 198–206 (2005).

42. T. J. Pemberton, *et al.*, Genomic patterns of homozygosity in worldwide human populations. *The American Journal of Human Genetics* **91**, 275–292 (2012).

43. J. M. Havrilla, B. S. Pedersen, R. M. Layer, A. R. Quinlan, A map of constrained coding regions in the human genome. *Nature genetics* **51**, 88–95 (2019).

44. T. W. Marchant, *et al.*, Canine brachycephaly is associated with a retrotransposon-mediated missplicing of SMOC2. *Current Biology* **27**, 1573–1584 (2017).

45. J. Plassais, *et al.*, Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nature communications* **10** (2019).

46. T. N. Phung, R. K. Wayne, M. A. Wilson, K. E. Lohmueller, Complex patterns of sex-biased demography in canines. *Proceedings of the Royal Society B* **286**, 20181976 (2019).

47. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33**, D514–D517 (2005).

48. D. W. Clark, *et al.*, Associations of autozygosity with a broad range of human phenotypes. *Nature communications* **10**, 1–17 (2019).

49. M. C. Keller, P. M. Visscher, M. E. Goddard, Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* **189**, 237–249 (2011).

50. A. Abdellaoui, *et al.*, Association between autozygosity and major depression: Stratification due to religious assortment. *Behavior genetics* **43**, 455–467 (2013).

51. H. G. Parker, *et al.*, Genetic structure of the purebred domestic dog. *science* **304**, 1160–1164 (2004).

52. K. Lindblad-Toh, *et al.*, Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803 (2005).

53. E. A. Ostrander, R. K. Wayne, A. H. Freedman, B. W. Davis, Demographic history, selection and functional diversity of the canine genome. *Nature Reviews Genetics* **18**, 705 (2017).

54. D. L. Dreger, *et al.*, Whole-genome sequence, SNP chips and pedigree structure: building demographic profiles in domestic dog breeds to optimize genetic-trait mapping. *Disease models & mechanisms* **9**, 1445–1460 (2016).

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

56. C. E. G. Amorim, *et al.*, The population genetics of human disease: The case of recessive, lethal mutations. *PLoS genetics* **13** (2017).

57. T. Gte. Consortium, The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

58. R. Cheung, *et al.*, A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions. *Molecular cell* **73**, 183–194 (2019).

59. B. P. Lewis, R. E. Green, S. E. Brenner, Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences* **100**, 189–192 (2003).

60. J. N. Miller, D. A. Pearce, Nonsense-mediated decay in genetic disease: friend or foe? *Mutation Research/Reviews in Mutation Research* **762**, 52–64 (2014).

61. P. M. Visscher, *et al.*, 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**, 5–22 (2017).

62. T. A. Manolio, *et al.*, Finding the missing heritability of complex diseases. *Nature* **461**, 747 (2009).

63. J. Yang, *et al.*, Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42**, 565 (2010).

64. E. A. Boyle, Y. I. Li, J. K. Pritchard, An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

65. C. C. Chang, *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

66. X. Zheng, *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

67. S. Ghirlanda, A. Acerbi, H. Herzog, J. A. Serpell, Fashion vs. function in cultural evolution: The case of dog breed popularity. *PLoS One* **8**, e74770 (2013).

68. C. L. Campbell, C. Bhérer, B. E. Morrow, A. R. Boyko, A. Auton, A pedigree-based map of recombination in the domestic dog genome. *G3: Genes, Genomes, Genetics* **6**, 3517–3524 (2016).

69. A. Auton, *et al.*, Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome research* **19**, 795–803 (2009).

70. N. Nakatsuka, *et al.*, The promise of discovering population-specific disease-associated genes in South Asia. *Nature genetics* **49**, 1403 (2017).

71. H. Chen, *et al.*, Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics* **98**, 653–666 (2016).

72. M. P. Conomos, A. P. Reiner, B. S. Weir, T. A. Thornton, Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics* **98**, 127–148 (2016).

73. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

74. F. Cunningham, *et al.*, Ensembl 2019. *Nucleic acids research* **47**, D745–D751 (2019).

Chapter 5: Conclusion

5.1 Discussion

The interaction between demography and selection has long been of interest to population geneticists and is especially pertinent in the context of deleterious variation (1-3). This is because the demographic history of a population affects the efficiency of selection, thus allowing for some deleterious variants to avoid being eliminated from the genome immediately by natural selection (2-3). Since some portion of the deleterious variation that persists within the genome may contribute to complex disease, the goal of my research was to characterize how demography can influence this deleterious variation. To this end, I examined the distribution and abundance of deleterious variation across multiple taxa that have distinct demographic histories to better understand how demography impacts selection's ability to remove such variation.

In **Chapter 2**, I studied genetic variation in two Latin American population isolates and observed a marked increase in genetic diversity of the Colombian and Costa Rican isolates relative to my benchmark population isolate, the Finnish. Despite this increase in genetic diversity, the Latin American population isolates still had an enrichment of IBD segments and a larger burden of ROH relative to the Finnish. Further, I determined that these long runs of homozygosity are generated by recent consanguinity, and that recent consanguinity has led to an enrichment of derived deleterious variation in the Costa Rican and Colombian isolates. The aforementioned result shows how recent consanguinity can reduce fitness in natural populations, by increasing the burden of deleterious variation in the genomes of individuals who are inbred.

In **Chapter 3**, I took a similar approach of evaluating genetic diversity in a population isolate, but this time focusing on the rarest canid in the world, the Ethiopian wolf. Here, I was interested not only in inferring the demographic history of the Ethiopian wolf, but also measuring the genomic impact of long-term small population size on deleterious variation in this wolf-like canid. I found that the Ethiopian wolf has experienced long-lasting population decline since its origins in the Bale Mountains. I inferred the current estimated effective size to be approximately 100 individuals, which matches the census size quite well. Given these results, the Ethiopian wolf provides an excellent model for assessing how deleterious variation accumulates in small populations. As a first step, I examined overall levels of diversity in the Ethiopian wolf. These wolves exhibited the lowest levels of diversity when compared to either breed dogs or gray wolf populations. In addition to exhibiting low diversity, the Ethiopian wolf carried more putatively deleterious variants relative to both breed dogs and gray wolves. Given that the Ethiopian wolf shows no outward signs of inbreeding depression and few long runs of homozygosity, I believe that despite having a small population size there has not been recent inbreeding in the Bale Mountain population. Thus, this enrichment of deleterious variation is likely due to long-term small population size and perhaps those variants with large fitness effects have been purged from the population over time.

Finally, **in Chapter 4**, I tested how recent demography and artificial selection affected the distribution of deleterious variation in breed dogs. Dogs have experienced both an ancient domestication bottleneck and a recent severe breed formation bottleneck, which has simplified the genetic architecture of complex traits and allowed deleterious variation to persist at appreciable levels in their genome. I examined the relationship between ROH burden and case-control-status for eight clinical and morphological traits. I find that ROH burden is sometimes associated with increased disease risk and other times appears to be protective. One standout association was lymphoma in Golden retrievers, where I saw that case-status was positively associated with the amount of the genome within an ROH and, on average, cases carry more ROH than controls. I also identify several understudied dog breeds that could be of interest for future disease-association studies due to their large burden of ROH and/or IBD burden. Lastly, I used the distribution of ROH across the genome to identify potential hotspots for inbreeding depression and located 27 loci where ROH are never observed. I used constrained coding regions from humans, which contain an enrichment of disease-associated variants, to functionally validate inbreeding-depression candidate loci. I found that on average, one would expect to see 18 loci in these constrained coding regions, however in our data I observed 23 (p = 0.025). This enrichment of inbreeding-depression candidate loci in these constrained coding regions is promising, and may also suggest that these regions are relevant across multiple taxa. Further, one would expect the enrichment since homozygous recessive variants in these loci should have large fitness effects.

In sum, my research has shown that both ancient and recent demography can affect the distribution and frequency of deleterious variation in the genomes of multiple species. When inbreeding occurs, I observe an enrichment of deleterious variation in the genome alongside an enrichment of ROH. I also find that these ROH can be associated with increased risk of disease or used to identify regions of the genome that may be associated with inbreeding depression across taxa. Additionally, I show that long-term small populations have the ability to persist despite having relatively little diversity and accumulating large amounts of deleterious variation in their genomes. Perhaps this persistence is due to the ability of a small population to purge extremely deleterious recessive variants from their genome early on in their history. This purging of recessive variants is possible because of the population's small size, which allows for selection to act on recessive variants which would likely be masked in a heterozygous form in larger populations. The remaining deleterious variants detected in my study of the isolated Ethiopian wolves may be only weakly deleterious, and not severely impact fitness.

In the future, my work could be extended to gain a broader understanding of the interaction between demography and selection. Though it remains quite hard to dissect this interaction, one path that could be explored further is the use of long ROH to determine the fitness effects of deleterious variants. Due to the recent origins of long ROH, the impact of recombination and selection would presumably be negligible in these regions of the genome. My work may also be of interest to those who wish to use the demographic history of a population to build a set of expectations about what the distribution of deleterious variation, IBD, and ROH in the genome would be. In conclusion, one should always think critically about the role of demography and how it can shape variation in any population.

5.2 References

1. Ohta T. Slightly Deleterious Mutant Substitutions in Evolution. Nature. 1973;246(5428):96.

2. Lande R. Genetics and demography in biological conservation. Science. 1988;241(4872):1455–1460.

3. Lohmueller KE, Indap AR, Schmidt S, et al. Proportionally more deleterious genetic variation in European than in African populations. Nature. 2008;451(7181):994.