UNIVERSITY OF CALIFORNIA

Los Angeles

Demography-aware inference of the strength of natural selection

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioinformatics

by

Vicente Diego Ortega Del Vecchyo

2016

ABSTRACT OF THE DISSERTATION

Demography-aware inference of the strength of natural selection

by

Vicente Diego Ortega Del Vecchyo

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2016

Professor John Novembre, Co-chair

Professor Kirk Lohmueller, Co-chair

Levels of genetic and possibly phenotypic variation are influenced by how natural selection acts to change the frequency of deleterious and advantageous mutations. However, the demographic history of a population influences the efficacy of natural selection to keep deleterious variants at low frequencies and to raise the frequency of advantageous mutations. I present three projects where we study how natural selection works in the context of different demographic histories. On the first project, I study the early demographic history of dogs and wolves since their divergence using genomic data from three dogs and three wolves. I inferred population bottlenecks in dogs and wolves and I found evidence for gene flow between dogs and wolves after their divergence. I develop a summary statistic to find the most plausible demographic model for dogs and wolves, where I found evidence for a demographic model stating that dogs evolved from one single location. This project laid the foundations to study how advantageous and deleterious variants work in the context of the bottlenecks found in dogs and wolves. On the second chapter, I study how population bottlenecks and inbreeding have influenced levels of deleterious genetic variation in dogs using 90 whole-genome sequences from breed dogs, village dogs and gray wolves. I used the ratio of heterozygosity at amino-acid changing variants over silent variants to show how bottlenecks associated with domestication and breed formation in dogs have affected

the efficacy of negative selection. I show multiple lines of evidence indicating that bottlenecks, and not inbreeding, are driving the patterns of deleterious genetic variation we observed in dogs. On the third project, I develop a novel likelihood-based method that uses the lengths of pairwise haplotype identity by state among rare-variant carrying haplotypes. The method conditions on the present-day frequency of the allele and is based on theory predicting that, under constant population sizes, the alleles under negative selection are on average younger than neutral alleles and should have higher average levels of haplotype identity among variant carriers. We developed a computational framework to obtain the probability distribution of the lengths of pairwise haplotype identity given a certain selection coefficient, demographic scenario and present-day allele frequency. Simulations indicate that our method provides unbiased estimates of selection under constant population sizes and realistic demographic scenarios. We show how our method can also be used to estimate the parameters that define the distribution of selective coefficients of a set of rare variants. We provide an example of how to apply this method to estimate the distribution of selective coefficients of a set of amino-acid changing variants in the UK10K, a large genomic dataset of British individuals.

Introduction

One of the most important problems in evolutionary biology is to understand what processes are maintaining the genetic variation observed within different species. A fundamental process affecting levels of genetic variation is the continuous input of new mutations every generation. These new mutations can be neutral, deleterious or advantageous. A neutral mutation does not affect the fitness of the individual. On the other hand, the fitness of an individual is decreased or increased by the action of deleterious and advantageous mutations, respectively. Levels of genetic diversity are affected by how natural selection is acting against these mutations. If mutations are advantageous, then natural selection will increase their frequency in the population. On the contrary, natural selection acts against deleterious mutations to decrease their frequency in the population.

A long-standing topic of interest has been to quantify the proportion of mutations that have a particular fitness effect value, going in a continuum from strongly advantageous to strongly deleterious. That information is commonly summarized in a continuous distribution called the distribution of fitness effects (DFE) of new mutations. Understanding the properties and shape of the DFE has been a topic of particularly fierce debate during the introduction of the neutral theory of molecular evolution. According to the neutral theory, the majority of new mutations are either deleterious or neutral, with advantageous mutations appearing rarely in the population. Since deleterious mutations are removed from the population and advantageous mutations do not appear often, the neutral theory holds that the majority of the genetic variation observed is due to neutral mutations. The importance and validity of the neutral theory

continues to be questioned nowadays. In the first section of the introduction, I will introduce the neutral theory of molecular evolution along with the related nearly neutral theory of molecular evolution. I will discuss evidence arguing against and in favor of those two theories.

In the second section, I will also review the importance of the interaction between past demographic history and the fitness effects of new mutations to define levels of genetic variation in a population.

Since past population history exerts an important influence on levels of genetic variation at neutral sites and under selection, it is crucial to infer it as accurately as possible. The recent availability of large-scale genomic datasets has fueled a lot of interest in the development of new methods to more accurately infer past demographic history. In the third section of my introduction, I will discuss these methods along with their strengths and disadvantages.

In the fourth part of this introduction, I will discuss modern approaches to estimate the distribution of fitness effects in a population. I will point out what are their main assumptions, advantages and disadvantages.

In the last part of this introduction, I will give a brief roadmap describing the content from each chapter of this dissertation.

1.1 Neutral theory and nearly neutral theory of molecular evolution

The neutral theory of molecular evolution was formally introduced in two papers (Kimura 1968; King & Jukes 1969). First, Motoo Kimura argues that the average amino-acid substitution rate in three proteins, one substitution every 2 years, was consistent with

the majority of the substitutions being ‘almost neutral’. He makes that conclusion based on an estimate made by Haldane that no more one substitution per 300 generations should be expected to not create an untolerable genetic load (Haldane 1957). Although King and Jukes challenge Motoo Kimura’s genetic load argument, they support the idea that the estimated substitution rates in proteins are fast and are more in line with substitutions being neutral. Based on estimates of the rate of appearance of slightly deleterious and recessive lethal mutations from (Mukai 1964; Whitfield et al. 1966), King and Jukes estimate that 80 to 90 percent of spontaneous mutations are mildly deleterious, 5 to 10 percent are lethal and 5 to 10 percent are selectively neutral. King and Jukes’s paper is also remarkable because it was the first one to note that there are functional constraints on different sites or proteins affect the substitution rate, where a higher functional constraint reduces the substitution rate (Ohta & Gillespie 1996).

The neutral theory of molecular evolution was heavily contested when it was presented (Ewens 2012). This was due to prevailing views on molecular evolution at the time holding that most of the amino acid substitutions must be under selection. Codon usage biases and a different rate of amino-acid substitutions between different amino-acids, based on the PAM amino-acid substitution matrix from Margaret Dayhoff, were mentioned as counter-evidence for the neutral theory by (Richmond 1970; Clarke 1970). A fair balance of the debate was presented in (Crow 1972), where he also points out some testable hypothesis from the neutral theory. He states that, under neutrality: 1) The mutation rate should be equal to the substitution rate; 2) Levels of heterozygosity should be related to the mutation rate and effective population size following this formula $H = {4Nu}/{1 + 4Nu}$ from (Kimura & Crow 1964) ; 3) That the frequencies of

different alleles should be dependent on the mutation rate and effective population size, and could be used to test neutrality employing Ewens sampling formula (Ewens 1972). An important modification from the nearly neutral was the nearly neutral developed by Tomoko Ohta (Ohta 1973). Ohta proposes that the neutral theory could be extended to include the fixation of slightly deleterious mutations. This class of mutations could become fixated in small populations and, therefore, we should expect to see a negative correlation between the substitution rate and the population sizes between species. The main differences between the proportion of mutations in different categories is illustrated in Figure 1.1 (taken from (Ohta 1992). The core of the neutral theory is that the vast majority of mutations we observe, both within and between species, are neutral. Deleterious mutations are rarely observed because selection acts against them. Advantageous mutations do not appear often in the population and, therefore, they are rarely observed. In the nearly neutral theory, mutations can be slightly beneficial or slightly deleterious and can contribute to the number of substitutions and polymorphisms observed.

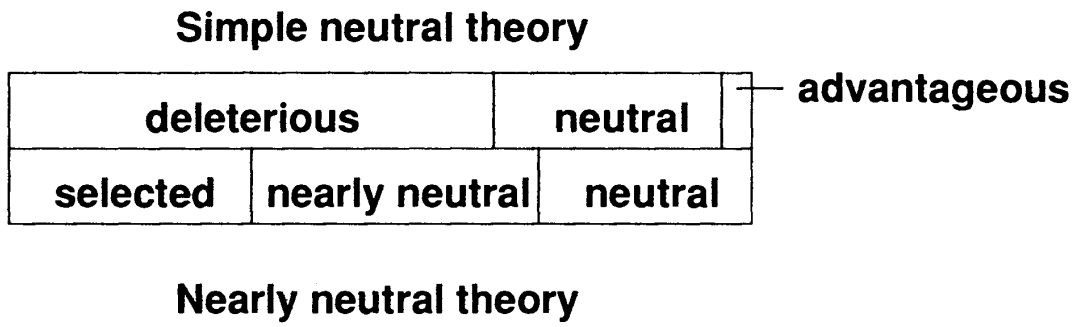

Figure 1.1.- Classification of new mutations in the simple neutral theory and nearly neutral theory. Figure taken from (Ohta 1992).

Later in the 1970s, Kimura and Ohta described more testable hypothesis from the neutral and nearly neutral theory (Kimura & Ohta 1974), including predictions stating that: 1) The rate of evolution for proteins should be constant, as long as the function of the protein is unaltered; 2) Less functional parts of a protein should have a higher substitution rate; 3) Mutant substitutions that disrupt less the existing structure and function of a molecule occur more frequently. Finally, they state that the neutral theory predicts that selective elimination of deleterious mutations and fixation of neutral or slightly deleterious mutations occur much more frequently than positive selection of advantageous mutations.

The neutral and nearly neutral theory of evolution was contested again due to the observation of differences in heterozygosity smaller than one order of magnitude in species with presumably large differences in population sizes (Lewontin 1974), inconsistent with the relationship between heterozygosity and population sizes in neutral sites found by (Kimura & Crow 1964). This problem has been revisited recently (Leffler et al. 2012). An important contribution to this subject was made by Maynard Smith and Haigh, who propose that the hitchhiking of neutral alleles with genomic regions experiencing selection could reconciliate the heterozygosity levels observed with the differences in population sizes (Maynard Smith & Haigh 1974).

The development of the neutral theory relied on information from protein sequences. This changed during the late 70's due to the development of new methods to sequence DNA and RNA (Sanger & Coulson 1975; Sanger et al. 1977; Maxam & Gilbert 1977). This provided further evidence for the neutral and nearly neutral theory. Synonymous

sites were found to have higher substitution rates compared to nonsynonymous sites in b-hemoglobins (Jukes 1978). This agrees with one principle from the neutral and nearly neutral theories stating that less functional parts of a protein have higher substitution rates. Linked to this principle, non-functional pseudogenes of the alpha-globin gene were found to have high substitution rates compared to functional forms of that same gene (Miyata & Yasunaga 1981).

The nearly neutral theory became widely accepted by the end of the 1980's. However, some of its assumptions continued to be questioned. One of the questioned assumptions is the constancy of selection coefficients over long periods of time (Ohta & Gillespie 1996). A strong criticism came from the estimated ratio of the variance of amino-acid substitution rate over the mean amino-acid substitution rate among lineages, which could be better explained by periods of changing substitution rates instead of the constant substitution rate predicted by the neutral theory (Gillespie 1984).

Another revolutionary advance done during the 1980's was done by Martin Kreitman's analysis of eleven copies of the Adh gene in Drosophila melanogaster (Kreitman 1983). This represents the first investigation using DNA trying to understand what evolutionary forces are driving the genetic variation observed inside a species. This advance also made possible to test for signals of selection using neutrality tests. One caveat of these neutrality tests, and any other neutrality test, is that they are affected by demography. One of the first neutrality test developed is the McDonald-Kreitman test, which tests for differences in the ratio of synonymous/nonsynonymous substitutions against the ratio of synonymous/nonsynonymous polymorphism as an evidence for neutrality (McDonald & Kreitman 1991). This test was shown later to be affected by changes in population size

(Eyre-Walker 2002). Fumio Tajima also developed the D statistic, which can test for an excess of low frequency variants as evidence for selection (F Tajima 1989). However, since a high proportion of low frequency variants can also be found due to a population expansion (F. Tajima 1989), care must be taken when interpreting the results of the D-statistic. We will return to the topic of how to take demography into account when studying selection in later sections of this introduction.

The neutral theory has continued to be criticized in subsequent decades. The problem with the lack of a consistent substitution rate across lineages, also called a constant molecular clock, has still been used as evidence against the neutral theory (Kreitman 1996). Another three additional observations done in the genus *Drosophila* not consistent with the neutral theory are: 1) The negative relationship between polymorphism and divergence; 2) The positive relationship between polymorphism and recombination and 3) Regional similarity in levels of polymorphism (Hahn 2008; Sella et al. 2009).

If the neutral or nearly neutral theory are true, we would expect to see a distribution of fitness effects of new mutations composed on its majority of neutral or nearly neutral mutations along with deleterious mutations. We would also expect to see that the distribution of fitness effects of segregating polymorphisms is mostly composed of neutral mutations. The problem of inferring the DFE of new mutations has been attacked before (more on that and the importance of the DFE of new mutations for other fundamental biological problems on section 1.4), while the problem of inferring the DFE of segregating mutations has been less studied (but see (Racimo & Schraiber 2014)). Chapter 4 of this dissertation in part will be devoted in part to develop methods to

estimate those two distributions. To infer those two distributions we must utilize information from genetic variation, where that information is also affected by the demographic history of a population. I will expand on this topic on the next section.

1.2 Interaction between selection and demography

When selection is acting on a mutation, the two types of selection that are more prevalent are positive selection and negative selection. Positive selection increases the frequency of an advantageous mutation. Negative selection decreases the frequency of a deleterious mutation. We can define the fitness of an individual containing the genotypes A1A1, A1A2 and A2A2 as 1, 1+s and 1+2s, respectively. Based on these definitions, if the selection coefficient $s > 0$ then the mutation is under positive selection while if $s < 0$ then the mutation is under negative selection. The efficiency of selection to act on the advantageous and deleterious mutations depends on the product of the effective population size N and the selection coefficient s. Due to this, many important summary statistics of genetic variation and other properties of an allele are a function of Ns. Exact formulas for some of these summary statistics and properties have been derived under a demographic model of a panmictic constant population size. Assuming additive effects, some properties that are a function of Ns are the fixation probability of a mutation u(p) (Kimura 1962), the age of an allele (Maruyama 1974), and the average time to fixation (Kimura 1980). The most important summary statistic of genetic variation, the site frequency spectrum, is also a function of Ns (Sawyer & Hartl 1992; Bustamante et al. 2001).

Sites under both positive and negative selection affect the variation on linked neutral and nearly neutral sites, and the magnitude of this effect is also dependent on past demographic history. When sites under positive selection, variation is reduced at linked neutral sites, a phenomenon called 'hitchhiking' (Maynared Smith & Haigh 1974; Gillespie 2000). The levels of diversity in the linked neutral sites are dependent on the population size when there is a single allele that raises in frequency fast in the population (Charlesworth & Charlesworth 2010), as well as in other models where there is recurrent selection on a particular site on different time points (Coop & Ralph 2012). Sites under negative selection also reduce the genetic variation at linked neutral sites, a phenomenon known as background selection. The reduction in genetic variation is also dependent on the effective population size. This has been calculated in a model without recombination events (Charlesworth et al. 1993) and in models with recombination events (Hudson & Kaplan 1995; Nordborg et al. 1996; Nordborg 1997). This points out once again the importance of demography in defining the efficiency of selection, both at the site under selection and at the linked neutral variation.

Analytical results to describe the genetic variation at sites under selection and at linked sites are only available for panmictic populations. However, most of the model species have non-equilibrium demographic scenarios. We must understand how genetic variation, both at sites under selection and linked neutral sites, is reduced in these demographic scenarios to be able to pursue two major objectives: Detect regions with genes under positive selection and to understand how deleterious variation changes due to different demographic events. Many efforts have been devoted recently to solve those two problems.

Efforts to understand how demography can change genome-wide diversity patterns and, therefore, our ability to detect regions with genes under positive selection have been carried out for different non-equilibrium scenarios. Those scenarios include bottlenecks and population subdivision (Pavlidis et al. 2010; Santiago & Caballero 2005; Jensen et al. 2005; Crisci et al. 2013; Barton 2000; Slatkin & Wiehe 1998). One of the projects where I was involved in found regions under positive selection in dogs and wolves, where we have both the effects of bottlenecks and population subdivision (Freedman et al. 2016). To do that project, we first needed to understand the demographic history of dogs and wolves. In Chapter 2, I describe the demographic analysis I performed to understand population size changes and migration rates in dogs and wolves.

The efficiency of selection to remove deleterious variants in non-equilibrium scenarios has received considerable attention recently (Brandvain & Wright 2016; Gravel 2016), particularly due to an interest in knowing whether recent demographic events involving reductions in population size can increase deleterious genetic variation. Some summary statistics lack power to measure if there has been an increase in deleterious genetic variation due to recent demographic events. Therefore, it is critical to carefully choose an appropriate summary statistic (Lohmueller 2014a). In Chapter 3, I will describe an investigation on patterns of deleterious genetic variation in dogs and wolves using a set of summary statistics that are sensitive to changes in deleterious genetic variation. Dogs and wolves are a great system to study patterns of genetic variation due to our knowledge on the timing and strength of bottlenecks in dogs, partly due to the work I describe in Chapter 2.

To be able to investigate patterns of genetic variation on sites under selection in non-equilibrium scenarios, we require a demographic model that is informed from genetic and historical data. In the next section of the data I will discuss recent approaches to infer a demographical model using genomic data.

1.3 Inference of demography using genome-wide scale data

Many ingenious approaches to infer past demographic history have been developed due to an increase in studies containing genome-wide data from many individuals. Some of the methods rely on the SMC (McVean & Cardin 2005) or the SMC' model (Marjoram & Wall 2006), which are highly accurate approximations to the ancestral recombination graph ARG, particularly the SMC' model (Wilton et al. 2015), that can facilitate the inference of demographic parameters of interest. The genealogical history across all sites in a chromosome is contained in the data structure known as the ARG. The main idea behind the SMC and SMC' models is that we:

1) Start with a random genealogy G_0 generated by the coalescent process at the left side of the chromosome with total branch length T_0.
2) Generate a random number that follows an exponential distribution with rate ρT_0. This is the distance to the next place where there has been a change in the local genealogy G_i.
3) Select a random place *x* uniformly in the genealogy.
4) Under the SMC coalescent: At the point x, create two branches (left and right) that will go towards the past. Delete the left branch, therefore eliminating that lineage. Then make the right side of the branch coalesce according to the usual coalescent probabilities anywhere up in the genealogy, including possibly coalescing farther away than the MRCA.

Under the SMC' coalescent: At the point x, create two branches (left and right) that will go towards the past. Make the right side of the branch coalesce according to the usual coalescent probabilities anywhere up in the genealogy, including possibly coalescing farther away than the MRCA. Then delete the left branch.

5) Go back to 2 and repeat until you reach the end of the sequence.

The main difference between the SMC and the SMC' model is that one more coalescent event is taken into account in the SMC' model, the possibility of the right branch coalescing with the left branch before deleting the left branch. An illustration with the main idea behind the SMC models is shown in Figure 1.2.

Figure 1.2.- A graphical illustration of the SMC algorithm taken from (McVean & Cardin 2005).

By providing an accurate approximation to the ARG, both the SMC and the SMC' model can be used to simulate sequence data that mimics patterns of genetic variation generated by the ARG (Chen et al. 2009) and to infer past demographic history. The first SMC model-based genome-wide scale method to infer past population sizes was PSMC (Li & Durbin 2011a). PSMC takes information from one single unphased genome and infers past population size changes under the assumption that all of its past history can be explained in terms of one panmictic population. Extensions developed later in the program MSMC use the SMC' model and take phased genomic data from many

individuals to estimate changes in population size and changes of the between and within coalescence rate (Schiffels & Durbin 2014). Additionally, information from the distribution of identity by state lengths combined with the SMC and SMC' models have been used to estimate past population size changes (Harris & Nielsen 2013).

Methods that rely on models different to the SMC and the SMC' have also been developed. Due to the development of tools that estimate the ARG in genomic data (Rasmussen et al. 2014), new demographic inference methods that employ the inferred ARG have been developed (Palacios et al. 2015). The sequentially Markov conditional sampling distribution SMCSD (Paul et al. 2011) has also been employed to infer past demographic history (Sheehan et al. 2013). Under the SMCSD, it is possible to calculate the probability of observing a certain haplotype given a set of other haplotypes and a particular demographic history. With this approach, it is possible to use a "leave-one-out" approach, where each of the haplotypes is left out in turn, and employ the SMCSD to calculate the full likelihood of the data given a particular demographic history.

Another popular model to infer past demographic history is the Poisson Random Field model PRF (Sawyer & Hartl 1992), where it is a assumed that mutations are independent from each other, always occur at a new site and each mutation follows an independent Wright-Fisher process. fastNeutrino (Bhaskar et al. 2015) infers past demographic events employing the PRF model, the coalescent model and information from the site frequency spectrum SFS, a summary statistic that displays the number or proportion of alleles at different frequencies in the population. Information from the SFS has also been employed to infer past demographic history employing a diffusion-based

composite likelihood approach (Gutenkunst et al. 2009) or a coalescent approach that estimates the composite likelihood of different models using simulations (Excoffier et al. 2013).

On the other hand, the program GPhoCS employs a standard coalescent model and a Bayesian method to infer the past demographic history employing genomic data from a set of individuals (Gronau et al. 2011a). GPhoCS uses information from a large number of genealogies from short neutral loci to get samples from the posterior distribution of demographic parameters that define past population sizes, migration rates and divergence times between a set of sampled genomes.

In Chapter 2 I explain my efforts to reconcile information from the two demographic approaches that were available at the time we published the paper based on that chapter, PSMC and GPhoCS, using genomic data from dogs and wolves. I develop a simple summary statistic that finds the best demographic model that explains the data. I also show that the demographic model is also concordant with patterns of gene flow detected using the D-statistic, which tests for asymmetries in the number of derived alleles between a source lineage (P3) and one of two other lineages (P1, P2) (Green et al. 2010; Durand et al. 2011).

1.4 The interaction between demography and the distribution of fitness effects

The distribution of fitness effects is one of the most important determinants of Evolution (Eyre-Walker & Keightley 2007). Apart of its importance to the neutral theory and to determine current levels of genetic variation, it is also relevant to understand current phenotypic variation, since the distribution of fitness effects can influence the evolution

of complex phenotypic traits (Lohmueller 2014b; Mancuso et al. 2015; Eyre-Walker 2010).

The Poisson Random Field model PRF is the framework currently employed to estimate the distribution of fitness effects. When the PRF model was proposed, it was used to estimate the strength of selection acting in the Adh gene using information from the number of substitutions and polymorphisms at synonymous and nonsynonymous sites (Sawyer & Hartl 1992). The use of polymorphism and divergence data coupled with inference under the PRF model allows the detection of even very weak selection (Akashi 1999) and has been successfully applied to infer selection in other species. Some examples of this are one study that found the distribution of selective coefficients on different genes in humans (Bustamante et al. 2005), and another one finding that, in a small set of genes, Arabidopsis tended to have a higher proportion of genes under negative selection compared to the higher proportion of genes under positive selection found in Drosophila (Bustamante et al. 2002).

In their foundational paper (Sawyer & Hartl 1992), Sawyer and Hartl also showed how the site frequency spectrum is affected by different strengths of selection under the PRF model. We show an example of how natural selection affects the site frequency spectrum in Figure 1.3, negative selection acts against deleterious alleles to keep them at low frequencies in the population. On the other hand, positive selection increases the frequency of advantageous alleles. The power of the site frequency spectrum to detect selection has been further analyzed, particularly when the assumption of having completely independent sites of the PRF model is violated (Bustamante et al. 2001). Violations to this assumption cause misleading estimates of selection. However, use of

the site frequency spectrum from many sites across the genome causes the sites to be more independent from each other and therefore provide a better estimate of selection, as seen in (Adam R Boyko et al. 2008a). Methods to estimate the distribution of selective coefficients using the site frequency spectrum and the PRF framework have been recently developed (Keightley & Eyre-Walker 2007; Adam R Boyko et al. 2008a; Loewe et al. 2006). The method from (Loewe et al. 2006) does not model past demographic events while the method of (Keightley & Eyre-Walker 2007) has the disadvantage that it only allows one population size change. On the other hand, (Adam R Boyko et al. 2008a) models more complicated demographic scenarios. The method first infers the demographic history using the site frequency spectrum at synonymous sites and then employs a likelihood-based method to find the parameters that better explain the site frequency spectrum at synonymous sites, based on the expected number of mutations $E(X_i|g(\gamma))$ at different frequencies given a distribution of selection coefficients:

$$E(X_i|g(\gamma)) = \frac{\theta}{2} \int_{-\infty}^{\infty} \int_{0}^{1} \binom{n}{i} x^i (1-x)^{n-i} f(x;\Theta,\gamma) g(\gamma) dx d\gamma$$

$$LL(g(\gamma)) = \prod_{i=1} E(X_i|g(\gamma))$$

Where θ is the genome wide mutation rate, n is the number of chromosomes, x is unknown frequency of the mutation, $f(x;\Theta,\gamma)$ is the probability of having a frequency x given a demographic history Θ and a selection coefficient γ, $g(\gamma)$ is the distribution of selection coefficients and $LL(g(\gamma))$ is the likelihood of the distribution of selection coefficients.

Current methods to infer the distribution of fitness effects of new mutations do not leverage linkage disequilibrium patterns in the data and only indirectly infer the distribution of fitness effects of segregating variants. In Chapter 4 I propose a method that takes into account the demographic history and the patterns of linkage disequilibrium to infer the distribution of fitness effects of variants at particular frequencies in the population. This method uses the information from the large number of variants at low observed frequency and can differentiate between negative and positive selection under certain realistic demographic scenarios. The application of this method is only feasible now thanks to the large-scale genomic dataset that are available now. Under the assumptions of the neutral theory, the distribution of segregating variants at low frequencies should be majorly neutral or nearly neutral.

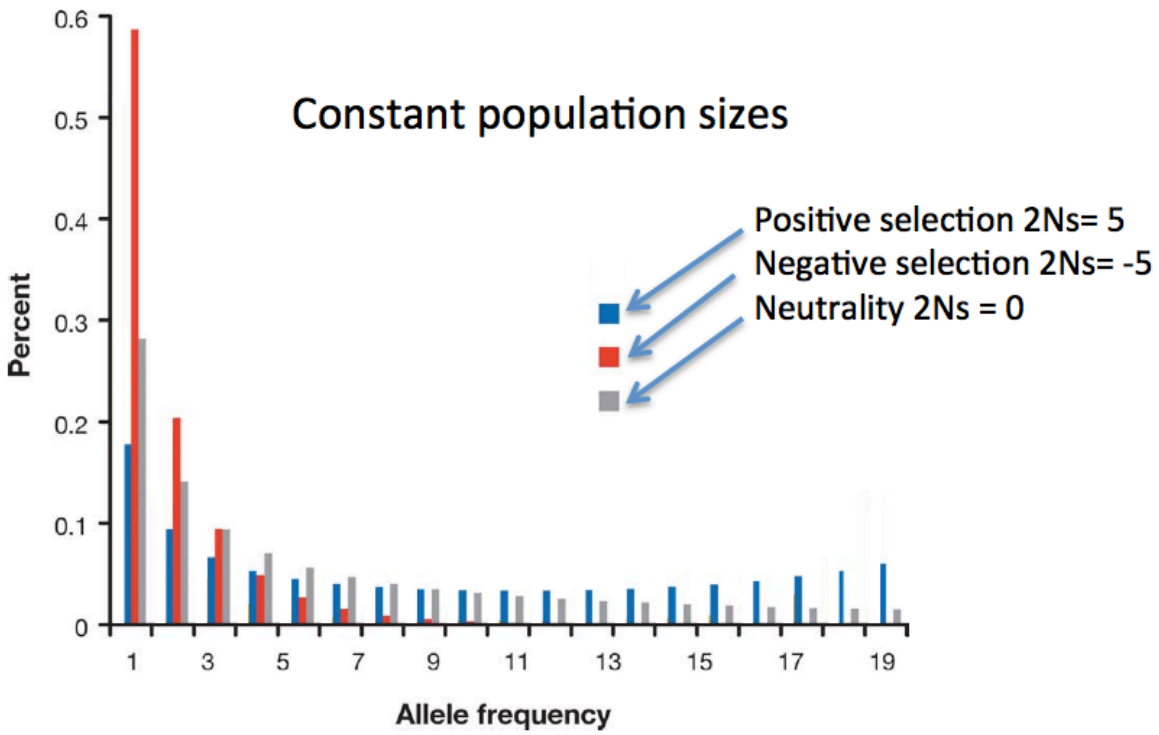

Figure 1.3.- Site frequency spectrum of alleles under selection. Notice how negative selection acts against deleterious alleles to increase the percent of alleles at low

frequency. On the other hand, positive selection increases the percent of alleles at higher frequency in the population.

1.5 Roadmap

This dissertation is divided into three different chapters, all unified into the common theme of understanding how past demographic history and natural selection act together to change levels of genetic variation.

In Chapter 2, I analyze patterns of gene flow and past population size changes in the early demographic history of dogs and wolves using genomic data from 3 wolves and 3 dogs. I inferred past population sizes and patterns of gene flow in both dogs and wolves. I show that the data is consistent with a demographic model that contains a single domestication event in dogs.

In Chapter 3, I present an analysis of how domestication bottlenecks have affected genetic variation on deleterious sites in dogs using genomic data from 71 dogs and 19 wolves. This investigation includes the development of the simulation software *PReFerSim* to analyze how past demographic history impact patterns of genetic variation at neutral and selected sites.

In Chapter 4, I introduce a new method that uses patterns of linkage disequilibrium to infer the distribution of fitness effects acting on a set of variants at a particular frequency in the population. To my knowledge, this is the first method that uses haplotypic information to infer the distribution of fitness effects. I show an application of this method to the UK10K dataset, which contains around 4,000 whole-genome sequences of individuals sampled from England.

Inferring the dynamic early history of dogs and wolves

This chapter explains my contributions to the following paper:

Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, Beale H, Ramirez O, Hormozdiari F, Alkan C, Vilà C, Squire K, Geffen E, Kusak J, Boyko AR, Parker HG, Lee C, Tadigotla V, Wilton A, Siepel A, Bustamante CD, Harkins TT, Nelson SF, Ostrander EA, Marques-Bonet T, Wayne RK, Novembre J. Genome sequencing highlights the dynamic early history of dogs. PLoS genetics. 2014;10(1):e1004016.

Introduction

To advance the understanding of dog origins and genetic changes early in dog domestication, we sequenced the genomes of six canid individuals, including three wolves (*Canis lupus*), an Australian Dingo, a Basenji and a golden jackal (*Canis aureus*). The three wolves sequenced were chosen to represent the broad regions of Eurasia where domestication is hypothesized to have taken place (Europe, the Middle East, and East/Southeast Asia) (Larson et al. 2012), and specifically, were sampled from Croatia, Israel, and China. Further, we sampled the Dingo and Basenji because relative to the reference Boxer genome, they are divergent lineages (Vonholdt et al. 2010) and maximize the odds to capture distinct alleles present in the earliest dogs. These lineages are also geographically distinct, with modern Basenjis tracing their history to hunting dogs of western Africa, while Dingoes are free-living semi-feral dogs

of Australia that arrived there at least 3500 years ago (Fig. 2.1) (Savolainen et al. 2004). As a result of their geographic isolation, the natural range of wolves has never extended as far south as the geographic sources for these two dog lineages (Larson et al. 2012), thus they are less likely to have overlapped with and admixed with wolves in the recent past. Sequencing the golden jackal allowed us to identify the ancestral state of variants arising in dogs and wolves.

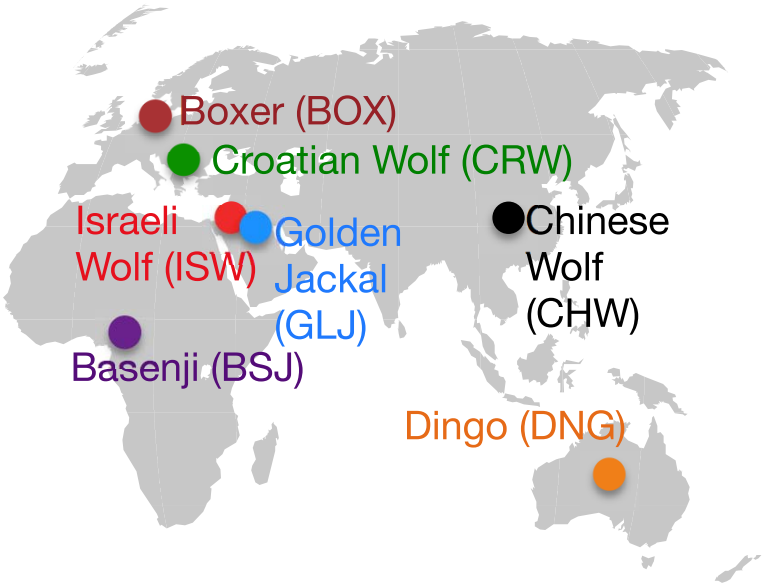

Figure 2.1.- Geographic distribution of sampled lineages

We chose to sequence a smaller number of individual genomes to high coverage, rather than larger numbers of individuals at low coverage, to take advantage of recently developed demography inference methods based on single high quality genomes (Durand et al. 2011; Gronau et al. 2011b; Li & Durbin 2011b). These methods allowed us to disentangle the effects of incomplete lineage sorting (ILS) and post-divergence gene flow, which pose a particular challenge in analysis of such recently diverged species as dogs and wolves (Larson & Burger 2013). Combining the results of

multiple complementary methods provided us with a synthetic, robust view of the shared history of dogs and wolves, including population divergence times, ancestral population sizes, and rates of gene flow. Using polymorphism data from 10 million single-nucleotide variant sites, we investigated: 1) the size of the ancestral wolf population at the time of wolf/dog divergence; 2) the geographic origins and timing of dog domestication; 3) post-divergence admixture between dogs and wolves.

Results

2.1 Genetic Distance Metrics

Individual genome sequences include valuable information about phylogenetic relationships between our samples. However, interpretation of these phylogenetic signals is challenging due to the possibility of post-divergence gene flow between dogs and wolves, as well as incomplete lineage sorting (ILS), which is a consequence of large ancestral population sizes. To provide accurate estimates of phylogeny given these demographic processes, we constructed a neighbor-joining (NJ) tree from a conservative estimator of genome-wide pairwise sequence distance for all pairs in our seven genomes, including the Boxer reference and using the golden jackal as an outgroup. The genome-wide pairwise distance metric employed between each of the 6 canid genomes and the reference Boxer sequence comes from (Gronau et al. 2011b) and is equal to:

$$d(X,Y) = \frac{1}{L}\sum_{i=1}^{L}[1 - \frac{1}{2}\max\left(\delta_{a_i c_i} + \delta_{b_i d_i}, \delta_{a_i d_i} + \delta_{b_i c_i}\right)] \quad \text{(Eq 2.1)}$$

where *X* and *Y* represent the two genomes being compared, *L* is the total number of sites utilized in the analysis, a_i and b_i are the two allele copies carried by individual *X*, c_i and d_i are the two allele copies carried by individual *Y* and δ_{jk} represents the Kronecker delta function (i.e. in this case equals one if allele *j* is identical to allele *k* and 0 otherwise). This measure represents a conservative estimate of the expected number of differences per site between individual chromosomes drawn.

We also computed the average number of nucleotide differences per site among a pair of randomly drawn alleles from each individual, using the following equation

$$d(X,Y) = \frac{1}{L}\sum_{i=1}^{L}[1 - \frac{1}{4}(\delta_{a_i c_i} + \delta_{b_i d_i} + \delta_{a_i d_i} + \delta_{b_i c_i})] \quad \text{(Eq 2.2)}$$

We took all of the sites across the genome that passed our quality filters to compute a matrix of pairwise distances between all canid genomes using E2.1 and E2.2 (Tables 2.1 and 2.2, respectively). The distances of all taxa to the golden jackal are very similar (approximately 0.0021) while the distances between dogs and wolves were about a half of that (0.0011). We used the matrix of pairwise distances generated by E2.1 and E2.2 to generate phylogenetic trees using the neighbor joining method as implemented on the program *neighbor* of the phylogenetic package *PHYLIP* (Felsenstein 1989).

In the neighbor-joining tree generated by using E2.1 (Figure 2.2A), all dogs were clustered into a single clade. Wolves also comprised a single clade, separated from other species by a branch of relatively short length. The Dingo was recovered as the outgroup to a clade comprised of Basenji and Boxer. Similarly, the Chinese wolf was

inferred as the outgroup to the clade formed by the Israeli and Croatian wolves. Thus, the phylogenetic tree supports the hypothesis that dogs and wolves are reciprocally monophyletic taxa.

The tree created using E2.2 (Figure 2.2B) differs from the previous tree in the position of the Chinese Wolf lineage. The Chinese Wolf appears as an outgroup to the clade comprised of the remaining dogs and wolves. However, the bootstrap support is low for both the branch that joins that lineage to the whole wolf-dog clade (54.2%) and the branch ancestral to the clade comprised of the Croatian and Israeli wolves 53.7%).

Table 2.1. Genome-wide pairwise sequence divergence, estimated using E2.1 using all the genomic sites that passed our genomic quality filters.

				Israeli	Croatian	Chinese	Golden
Boxer							
Basenji	0.00087						
Dingo	0.00094	0.00097					
Israeli wolf	0.00111	0.00105	0.00111				
Croatian wolf	0.00113	0.00110	0.00112	0.00101			
Chinese wolf	0.00114	0.00111	0.00111	0.00106	0.00105		
Golden jackal	0.00211	0.00211	0.00212	0.00209	0.00209	0.00210	

Table 2.2. Genome-wide pairwise sequence divergence, estimated using E2.2 using all the genomic sites that passed our genomic quality filters.

				Israeli	Croatian	Chinese	Golden
Boxer							
Basenji	0.00087						
Dingo	0.00094	0.00100					
Israeli wolf	0.00111	0.00112	0.00116				
Croatian wolf	0.00113	0.00117	0.00116	0.00115			
Chinese wolf	0.00114	0.00117	0.00115	0.00118	0.00115		
Golden jackal	0.00211	0.00214	0.00214	0.00214	0.00214	0.00214	

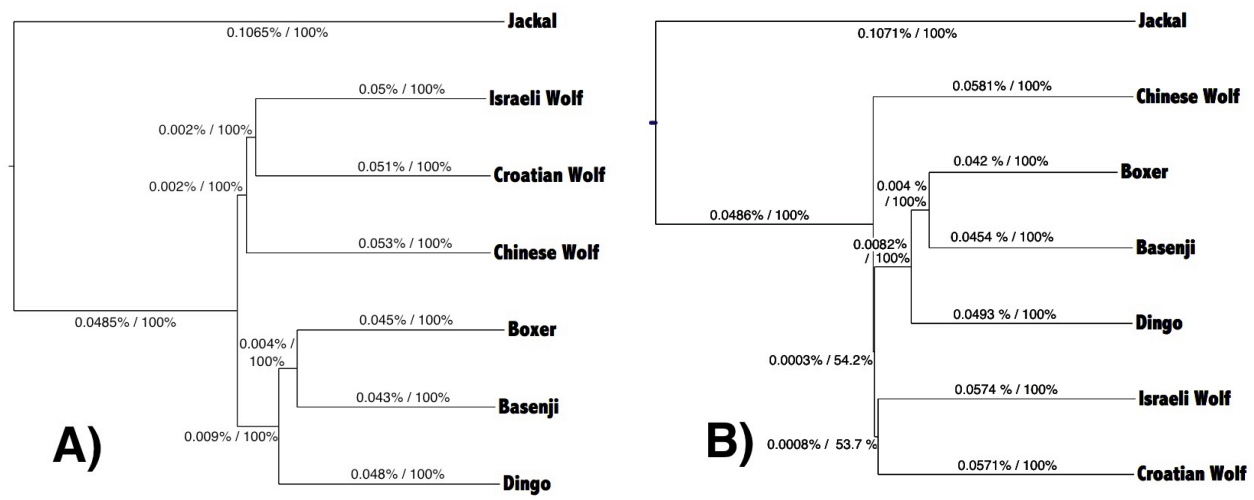

Figure 2.2 Neighbor-joining tree of canid samples plus the Boxer reference (CanFam3.0) for all positions passing our quality filters and for which there was no missing data for any sample. The distance metrics used were E2.1 and E2.2 for panel A) and B), respectively. For each branch, we report the genetic distance (left side of the slash) and the bootstrap support (right side of the slash). Bootstrap replicates were generated by dividing the genome of each species into windows of 500 kb based on the genomic coordinates of the Boxer reference, and then resampling with replacement from those windows until the bootstrapped genomes

for each species contain an equal or greater number of sites called as the true genomes.

2.2 Population Size Change Inference From Single Genome Sequences

We used the methods developed by (Li & Durbin 2011b) to infer the trajectory of population sizes across time for the six canid genome sequences. Briefly, the method uses the distribution of heterozygote sites across the genome and a pairwise sequentially Markovian coalescent (PSMC) model that defines a Hidden Markov Model, where the parameters are the mutation rate, recombination rate and the effective population sizes through time. The parameters are inferred through an Expectation-Maximization algorithm.

The genotypes for each diploid genome sample that passed the GF2 and SF filters were transformed into a sequence of ‘0’, ‘1’ and ‘.’, with one character for each 100bp, and where a ‘1’ was assigned if there were heterozygous sites in the window, 0 if there were none, and a ‘.’ was given if more than 90 positions were missing in the 100 bp window. Passing this data into the PSMC software, we ran 20 iterations of the Expectation-Maximization algorithm. The EM algorithm was run using an upper bound on the time to the most recent common ancestor equal to 10 in a $2N_0$ scale and an initial ρ/Θ set to the default value of 5. Following (Li & Durbin 2011b), the N_e was inferred across 64 different intervals for each dog genome, where the interval boundaries were set equal to:

$$t_i = 0.1 \exp\left[\frac{1}{n} \log\left(i + 100\right)\right] - 0.1$$

on a $2N_0$ scale ,where *i* takes values from 0 to 64. In a preliminary run we found that the number of recombination events inferred in the most recent time intervals by PSMC falls below 10. In such situations, the authors of PSMC recommend refraining from inferring a population size during such time intervals. Thus, we merged the first 6 intervals such that only a single N_e is inferred across them while the next 58 intervals were allowed to have interval-specific N_e values (in the Chinese wolf, the number of recombination events was higher and thus we continued to use all 64 intervals).

To translate from time units of generations to calendar years, we assume a generational time of 3 years for the wolves and the golden jackal. For the Dingo and the basenji, we used a generational time of 2 years from the present until the N_e interval that reached 10,000 years ago and for all N_e intervals further into the past, we used a generational time of 3 years. We found this scaling improved the concordance of the trajectories during the ancestral period where we expect them to be identical across lineages and is motivated by the known shorter generation time in domestic dogs. Following (Kerstin Lindblad-Toh et al. 2005), the mutation rate assumed was 1.0×10^{-8} per generation.

The full results including the golden jackal are shown in Figure 2.3. The golden jackal shows an apparent large increase in effective populations size around 80,000 years ago. We address interpretations of this signal in more detail in the results of our validation study (see below).

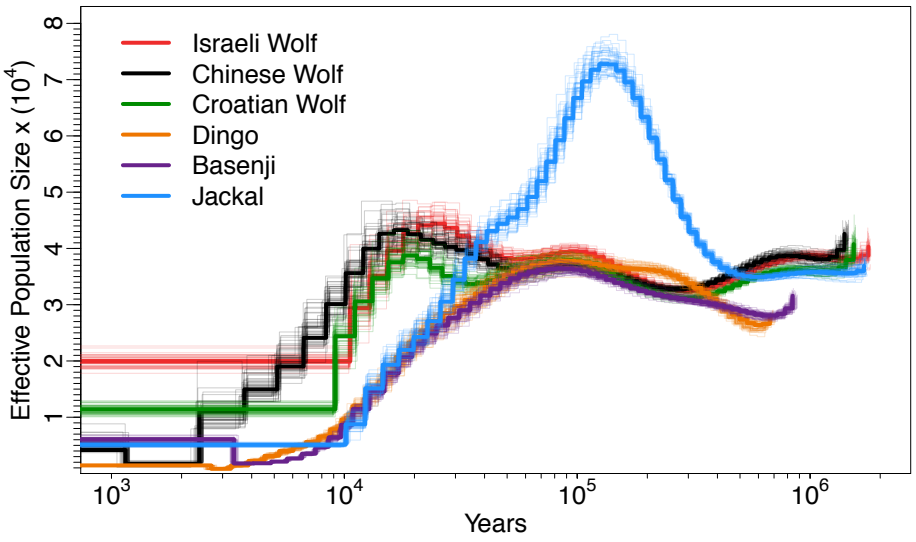

Figure 2.3. *Ne* trajectories of 6 canid lineages reconstructed using the PSMC method. Dark and light lines indicate whole genome based estimates and bootstrap estimates, respectively.

2.3 Quality control of population size change inferences

We assessed the confidence in our PSMC findings in three ways. First, to assess the certainty in the inferred *Ne* trajectories, we ran the PSMC method using the same settings for the initial estimations, assessing the variance in those estimates from 100 bootstrap replicates for each genome. To sample a bootstrap replicate, we divided the genome into segments of 5Mb, sampled with replacement from those segments until we obtained a sequence with approximately the same length as the original genome as defined by using the "-b" option in the PSMC software, and re-ran the EM-based *Ne*

estimation procedure. This analysis revealed a low variability among the *Ne* traces, comparable to what has been recovered in the analysis of human genome sequences (Figure 2.3) (Li & Durbin 2011b).

Second, we tested the sensitivity of the methods to long runs of homozygosity (RoH), as the Chinese wolf sample evidenced several runs. To test if long runs of homozygosity could bias the inference of Ne trajectories, we identified runs of homozygosity with the program PLINK (Purcell et al. 2007). As can be seen in Figure 2.4, the estimated trajectories are not affected by the removal of the RoH regions. This implies that the degree of inbreeding in the Chinese wolf is not large enough to bias the inference of ancestral demographic events estimated by the PSMC method.

Third, to investigate the sensitivity of PSMC to our choice of minimum acceptable genotype quality (GQ ≥20), we ran the PSMC analysis including the genotypes that passed our quality filters, but relaxing the GQ such that we included sites with GQ≥10 (as a contrast, Figure 2.3 use the genotypes that passed the GF2 and SF1 filters and had a GQ ≥ 20). Using this more liberal GQ threshold, values of *Ne* are lower by approximately 1,000 along the trajectory of all canids (Figure 2.5), however the *Ne* trajectories remain largely concordant. The effect is particularly strong in the golden jackal between 50,000 – 300,000 years ago, where using a lower GQ threshold reduces the estimates of *Ne* by 2,000. The difference between the dog and wolf *Ne* at earlier times (5,000-70,000 years) is more noticeable when using a higher GQ threshold. The reductions in *Ne* across the PSMC traces are consistent with expectations with respect to how confidence in genotype quality scales differently for homozygous versus heterozygous genotype calls. Homozygous sites can be called confidently with less data

that is of lower quality. Conversely, heterozygous calls will require more and higher quality data, such that genotype qualities at those sites will be higher. As a result, lowering the GQ threshold leads to the inclusion of disproportionately more homozygous genotypes than low quality heterozygous ones, reducing the observed heterozygosity within defined intervals, and as a result, the inferred N_e. Overall, although changes in GQ filtering does influence the estimates of the N_e trajectories, the magnitude of the changes are not large, and more importantly, the major patterns in the inferred trajectories are preserved.

Fourth, we simulated genome sequences arising from the demographic history inferred from a model inferred by the method *G-PhoCS* (more details on Text S9 from Freedman et al., 2014), a recently developed Bayesian demographic inference method, which assumes that wolves and dogs are reciprocally monophyletic taxa to determine if we could accurately reconstruct changes in N_e conditional on such a history. Specifically, for each species we simulated one hundred regions of 30Mb apiece using the program *MaCS* (Chen et al. 2009). We conducted these simulations under three different scenarios, varying the levels of gene flow between lineages. We used parameter values from the main results obtained with *G-PhoCS.* The scenarios tested used:

1) The full model inferred from *G-PhoCS* (Command Line 1 in Appendix, see command-line parameter listings below).

2) Our model inferred with *G-PhoCS* but with no gene flow between any species at any time (Command Line 2 in Appendix).

3) The model inferred by *G-PhoCS* but with only one form of gene flow, from golden jackal to the ancestor of dogs and wolves (Command Line 3 in Appendix).

4) The model inferred by *G-PhoCS* but with only one form of gene flow, from the ancestor of dogs and wolves to the golden jackal (Command Line 4 in Appendix).

5) The model inferred by *G-PhoCS* but only with gene flow from the Israeli wolf to the golden jackal (Command Line 5 in Appendix).

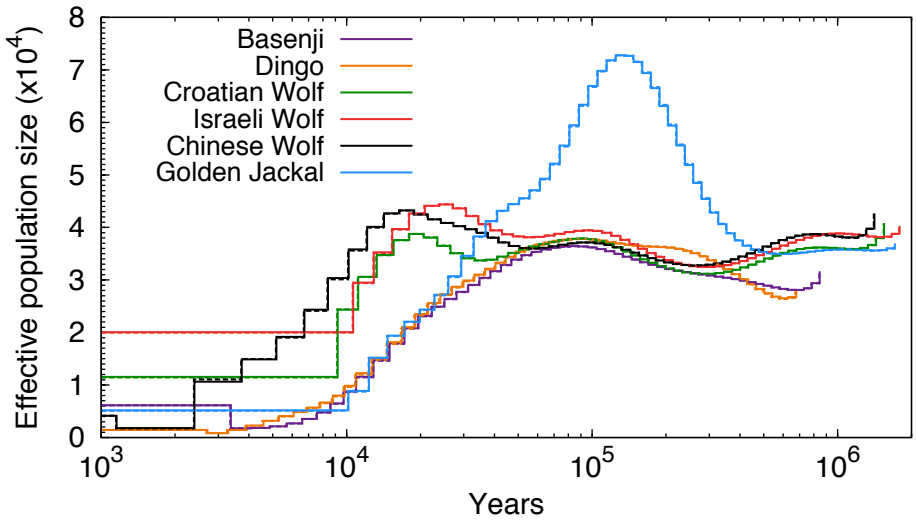

Figure 2.4. N_e trajectories of 6 canid lineages reconstructed with the PSMC method using all the genomic information that passed our quality filters (dashed lines) and excluding 43 regions with runs of homozygosity (solid lines).

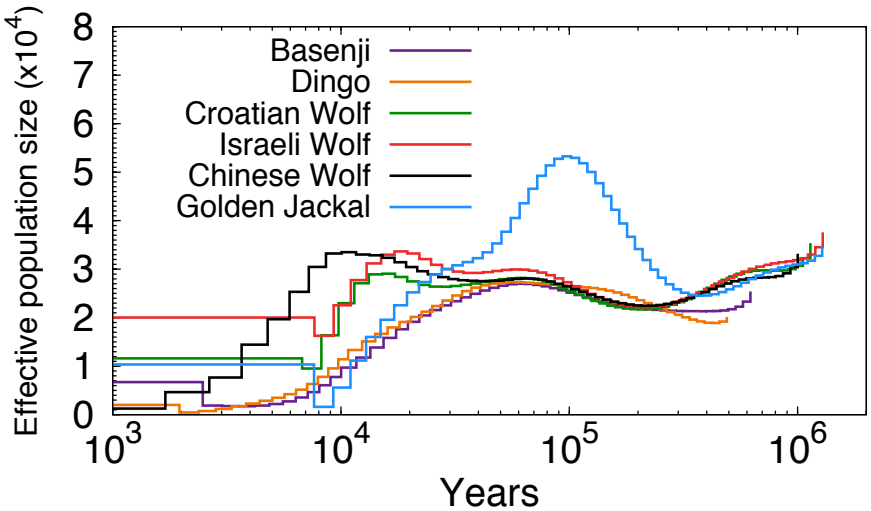

Figure 2.5 *Ne* trajectories of 6 canid lineages reconstructed with the PSMC method using the sites that had a GQ ≥ 10.

There are 7 different genomes being simulated in the command lines for each scenario. They are a haploid genome of the Boxer and diploid genomes for the Basenji, Dingo, Israeli wolf, Croatian wolf, Chinese wolf and Golden Jackal, respectively. Only the diploid genomes were used in this analysis. The output of *MaCS* was processed using perl scripts, so that each of the 30Mb regions was transformed into a binary sequence of '1' and '0', where each character was determined by the presence or absence of a heterozygote site in contiguous windows of 100bp. Then, for each lineage we used the 100 transformed binary sequences of 30Mb to run the PSMC method using the following command line:

./psmc -N20 -t10 -r5 -p "1*6+58*1" -o <Output file> <Input file>.

The recombination rate in all scenarios was assumed to be equal to 0.92 cM/Mb, a value that is equal to the mean recombination rate estimated in the dog genome in a

linkage map generated using microsatellites (Wong et al. 2010). In these simulations, we set the generational time to 3 years and mutation rate to 1×10^{-8} per bp per generation for all species.

We compared the N_e trajectories specified in the simulations with the estimations done by the PSMC method for each canid species. Scenarios 2 (Figure 2.6) and 3 (Figure 2.7) have remarkably similar and accurate trajectories inferred using the PSMC method for all species of canids. In scenarios 4 (Figure 2.8), 5 (Figure 2.9) and 1 (Figure 2.10), the N_e trajectories are also accurate for all species of canids but the golden jackal, where the estimate of N_e is inflated in the interval from 10,000 - 300,000 years ago, with a distinctive sharp peak between 100,000 and 300,000 years ago.

Admixture with wolves or the ancestor of dogs and wolves appears to generate the extreme upward bias in the inferred ancestral jackal N_e. In PSMC inferences from simulated jackal demographic histories the presence of jackal - dog/wolf ancestor and jackal - Israeli wolf migration bands (Figures 2.8 – 2.10) produced an artefactual spike in the jackal N_e trajectory. This sharp peak is similar to the one observed in the empirical data from the golden jackal, although in the N_e trajectory reconstructed from that data, the peak is slightly more recent. Overall, we conclude the peak in the N_e trajectory observed in the data is likely due to post- divergence gene flow between ancestors of contemporary golden jackals and Israeli wolves or the ancestor of dogs and wolves. Ongoing work has found evidence for multiple highly divergent jackal or jackal-like lineages in Africa and the Middle East (Koepfli et al. 2015).

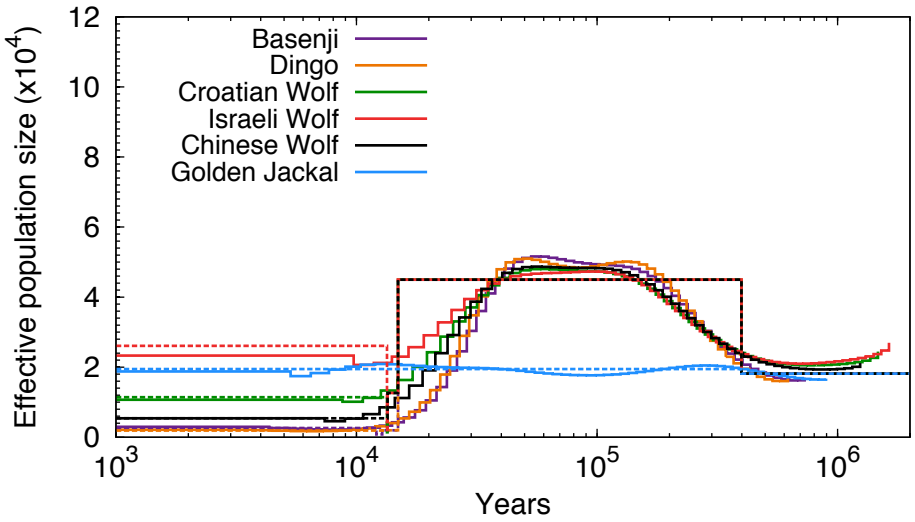

Figure 2.6 *Ne* trajectories of 6 canid lineages reconstructed using the PSMC method, for data simulated under the *G-PhoCS* inferred demographic history, excluding migration bands. The dotted lines show the actual *Ne* trajectories whereas the solid lines represent the inferred *Ne* trajectories.

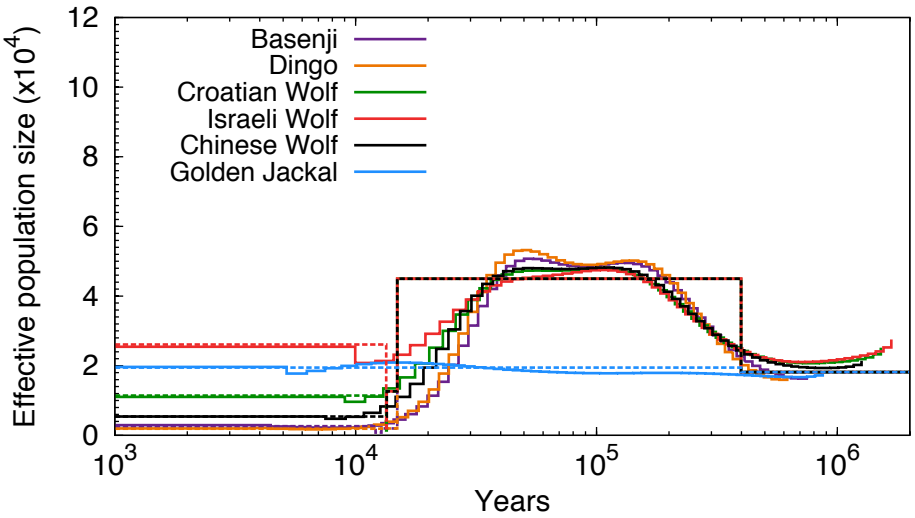

Figure 2.7. *Ne* trajectories of 6 canid lineages reconstructed using the PSMC method for data simulated under the *G-PhoCS* inferred demographic history, only including gene flow from the golden jackal to the ancestor of dogs and wolves.

Inferred *Ne* trajectories are shown with solid lines and the actual *Ne* trajectories are displayed with dotted lines.

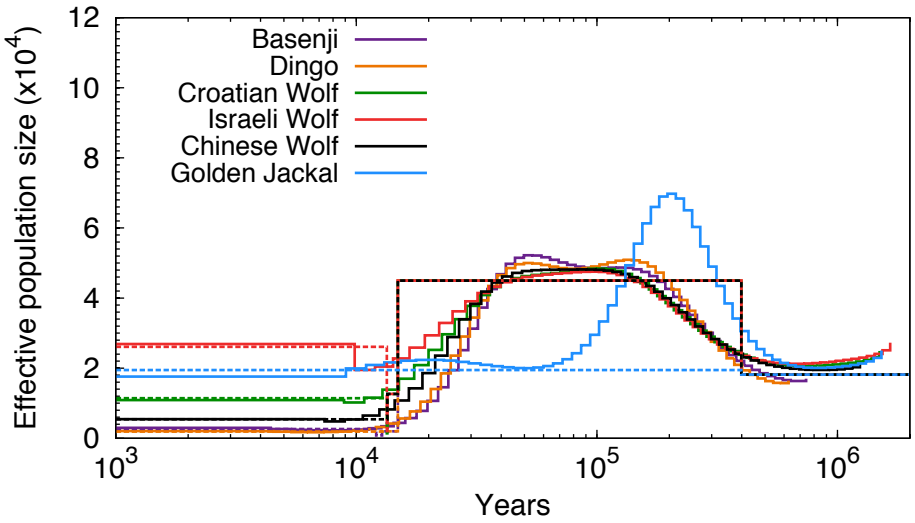

Figure 2.8. *Ne* trajectories of 6 canid lineages reconstructed using the PSMC method for data simulated under the *G-PhoCS* inferred demographic history, only including gene flow from the ancestor of dogs and wolves to golden jackal. Inferred *Ne* trajectories are shown with solid lines and the actual *Ne* trajectories are displayed with dotted lines.

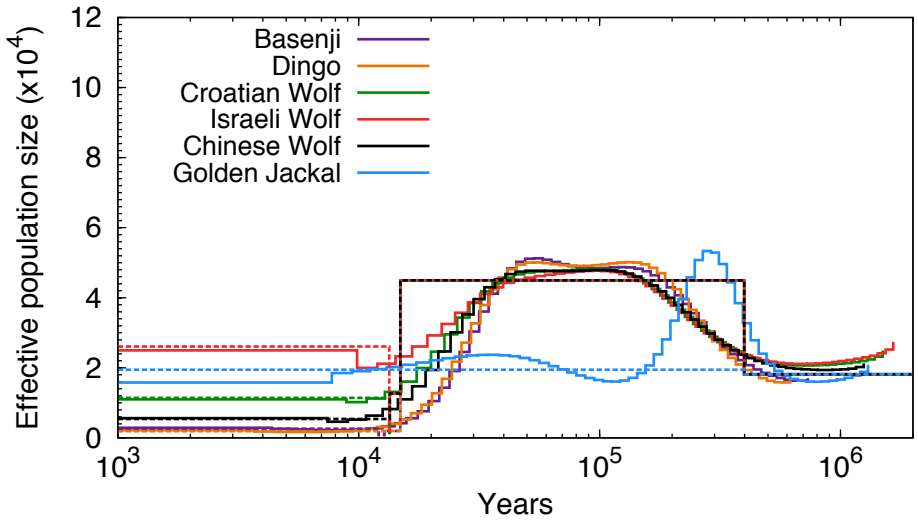

Figure 2.9. *Ne* trajectories of 6 canid lineages reconstructed using the PSMC method for data simulated under the *G-PhoCS* inferred demographic history, only including gene flow from Israeli wolf to golden jackal. Inferred *Ne* trajectories are shown with solid lines and the actual *Ne* trajectories are displayed with dotted lines.

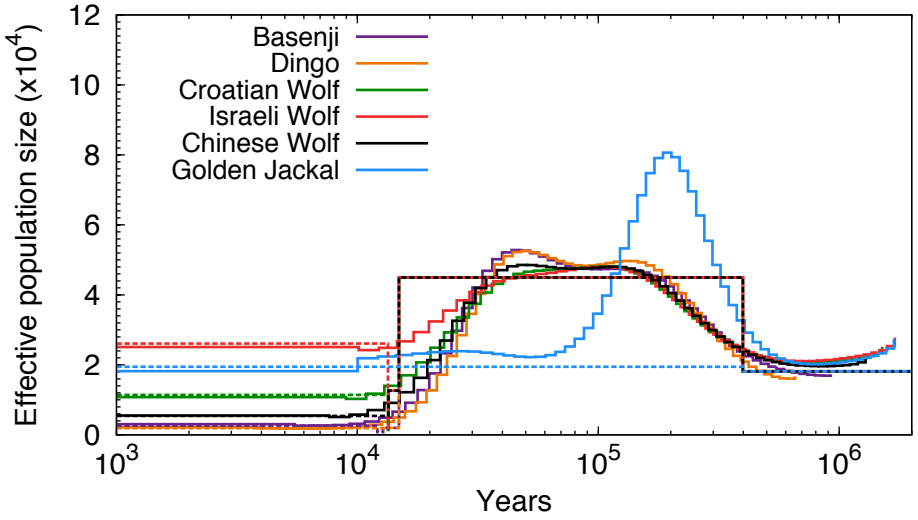

Figure 2.10. *Ne* trajectories of 6 canid lineages reconstructed using the PSMC method, for data simulated under the *G-PhoCS* inferred demographic history, including all detected gene flow. The actual *Ne* trajectories are shown as dotted lines whereas the inferred *Ne* trajectories are depicted by solid lines.

2.4 Post-Divergence Gene Flow

To investigate the extent of gene flow between wolves and dogs subsequent to their divergence, we employed a method recently developed by (Durand et al. 2011). This method tests for gene flow by testing for asymmetries in allele sharing between a source lineage (P3), and either of two receiving lineages (P1, P2). In this case, the ancestor of P1 and P2 is sister to the ancestor of P3. Given a site that is bi-allelic in (P1, P2, P3) where P3 is in state B and an outgroup (O) is in state A, there are two possible allelic configurations of P1-P2-P3-O that are informative with respect to gene flow between P3 and either P1 or P2: ABBA and BABA. In the absence of lineage- specific post-divergence gene flow and under selective neutrality, the genome-wide frequency of these configurations should be approximately equal. Thus, the null hypothesis is that there has not been gene flow between P3 and P1 or P2 after the divergence of P3 from P1 and P2. We defined an ABBA site as a site where P1 and the outgroup shared the same allele 'A' while P2 and P3 shared an alternative allele 'B'. A site was defined as a BABA site when the outgroup and P2 shared the allele 'A' and the alternative allele 'B' was shared between P1 and P3. The rejection of the null hypothesis indicates that there has been gene flow between P3 and either P1 or P2. Deviations from the null expectation were quantified using the D-statistic:

$$D = \frac{\sum_{i=1}^{n} C_{ABBA}(i) - \sum_{i=1}^{n} C_{BABA}(i)}{\sum_{i=1}^{n} C_{ABBA}(i) + \sum_{i=1}^{n} C_{BABA}(i)} \quad \text{E2.3}$$

where $C_{ABBA}(i)$ and $C_{BABA}(i)$ are indicator variables equal to 1 or 0 depending on the presence or absence of the ABBA and BABA sites at the ith site. To calculate the D statistic, we specified the golden jackal as our outgroup, and divided the reference genome into 422 segments of 5 Mb each, excluding the chromosome ends where the remaining segment is < 5Mb. Within these segments, we used stringent filtering criteria, excluding genomic positions with missing data, and sites that failed our set of quality filters. For each species at each site, with the exception of the haploid boxer reference, we randomly sampled one allele from the called genotype. We then calculated the D statistic from a total of *n* sites that met our quality control filters.

To be consistent with the evolutionary history reflected in the recovered neighbor-joining tree (see Figure 2.2), and to focus on gene flow most germane to evolutionary processes influencing wolf-dog divergence, we restricted testing to those cases where when one of the dog samples was P3, the other two (P1 and P2) were wolves, and vice versa (P3=wolf, P1 & P2 = dogs). Using these criteria, and including the boxer reference among the dogs, 18 tests were possible.

Following (Durand et al. 2011), the standard error of the statistic was calculated using a jackknife procedure (EFRON 1981). A Z-score was then obtained by dividing the value of the D statistic by its standard error. Z-scores with an absolute value ≥3 were considered significant. Rejection of the null hypothesis indicates that there has been gene flow between P3 and either P1 or P2 (Rasmussen et al. 2014). Negative significant Z scores indicate gene flow between P1 and P3 while positive significant Z scores indicate gene flow between P2 and P3.

We found evidence for post-divergence gene flow between three pairs of samples: basenji/Israeli wolf, boxer/Israeli wolf, and dingo/Chinese wolf (Table S8.4.1). The mean absolute value of *Z* was highest in basenji/Israeli wolf ($|\hat{Z}|$ = 9.27; range = 5.64 -12.11), compared to Chinese wolf/Dingo $|\hat{Z}|$ = 6.58; range = 3.58 – 10.14), and Israeli wolf/Boxer ($|\hat{Z}|$ = 6.15; range = 5.33 - 6.71).

Because calculation of the D statistic does not account for the effects of gene flow between the outgroup and any of the three samples considered under a given test, it is possible that such gene flow could introduce bias. In particular, our analyses using *G-PhoCS* support gene flow between the jackal and the Israeli wolf and jackal and the ancestral wolf. Nevertheless, our ABBA/BABA results are not affected by this gene flow for the following reasons. First, only the gene flow with Israeli wolf could affect the calculation of the D statistic. Thus, this gene flow would not affect tests that did not include the Israeli wolf. Second, this gene flow would not affect tests with two dogs and one wolf (dog,dog,wolf,jackal = 1,2,3,4), as Israeli wolf –jackal gene flow would lead to an allelic configuration that is **AA or **BB and thus not evaluated in the test. It is possible that, in tests with two wolves (one of which is the Israeli wolf), jackal- Israeli wolf admixture could give appearance of gene flow between the dog in question and the other wolf in the test. For example, consider a test that includes Israeli wolf, Croatian wolf, Basenji, and Golden Jackal. If the 'B' allele resulted from a mutation that arose in the ancestor to dogs and wolves, the original configuration would be BBBA, but Israeli wolf –jackal admixture would convert it to ABBA, leading to an upwardly biased count of this configuration, which would contribute to a Croatian wolf-Basenji gene flow signal. Nevertheless, we found in all tests with two wolves and one dog that include the Israeli

wolf, the significant gene flow that is detected is between the Israeli wolf and the dog in question, the exact opposite of what would be expected if Israeli wolf –jackal gene flow were biasing the test statistic.

A complete view of the pairs of taxons for which we inferred post-divergence gene flow is shown in Figure 2.11.

Table 2.3. Estimation of post-divergence gene flow using the D Statistic (Durand et al. 2011). The outgroup in all comparisons is the golden jackal. Statistical significance is evaluated using a two-tailed Z test, with the additional requirement that that absolute value of the Z-score to be ≥3. Significant tests and sample pairs showing evidence for post-divergence gene flow are shown in bold.

P1	P2	P3	ABBA Sites	BABA Sites	D (%)	SE (%)	Z	p-value
Basenji	Dingo	Croatian wolf	164211	162364	0.57%	0.40%	1.42	0.16
Basenji	Dingo	**Israeli wolf**	158610	179656	-6.22%	0.51%	**-12.21**	**2.79×10^{-34}**
Boxer	Basenji	Croatian wolf	144942	146113	-0.40%	0.46%	-0.88	0.38
Boxer	**Basenji**	**Israeli wolf**	157007	147991	2.96%	0.52%	**5.64**	**1.67×10^{-8}**
Boxer	Dingo	Croatian wolf	177485	176031	0.41%	0.44%	0.94	0.35
Boxer	Dingo	**Israeli wolf**	176511	189294	-3.49%	0.52%	**-6.71**	**1.96×10^{-11}**
Croatian wolf	**Israeli wolf**	**Boxer**	226123	210897	3.48%	0.65%	**5.33**	**9.86×10^{-8}**
Croatian wolf	Israeli wolf	Dingo	213742	212876	0.20%	0.54%	0.38	0.71
Croatian wolf	**Israeli wolf**	**Basenji**	205695	182191	6.06%	0.62%	**9.74**	**1.99×10^{-22}**
Basenji	**Dingo**	**Chinese wolf**	173366	162030	3.38%	0.45%	**7.49**	**6.76×10^{-14}**
Boxer	Basenji	Chinese wolf	149172	147273	0.64%	0.41%	1.54	0.12
Boxer	**Dingo**	**Chinese wolf**	192400	175946	4.47%	0.44%	**10.14**	**3.77×10^{-24}**
Croatian wolf	Chinese wolf	Boxer	216145	219859	-0.85%	0.42%	-2.02	4.32×10^{-2}
Croatian wolf	**Chinese wolf**	**Dingo**	221737	212060	2.23%	0.44%	**5.10**	**3.48×10^{-7}**
Croatian wolf	Chinese wolf	Basenji	190706	191336	-0.16%	0.39%	-0.42	0.68
Chinese wolf	**Israeli wolf**	**Boxer**	242452	222327	4.33%	0.68%	**6.41**	**1.43×10^{-10}**
Chinese wolf	Israeli wolf	**Dingo**	223003	232071	-1.99%	0.56%	**-3.58**	**3.48×10^{-4}**
Chinese wolf	**Israeli wolf**	**Basenji**	216213	191475	6.07%	0.64%	**9.50**	**2.02×10^{-21}**

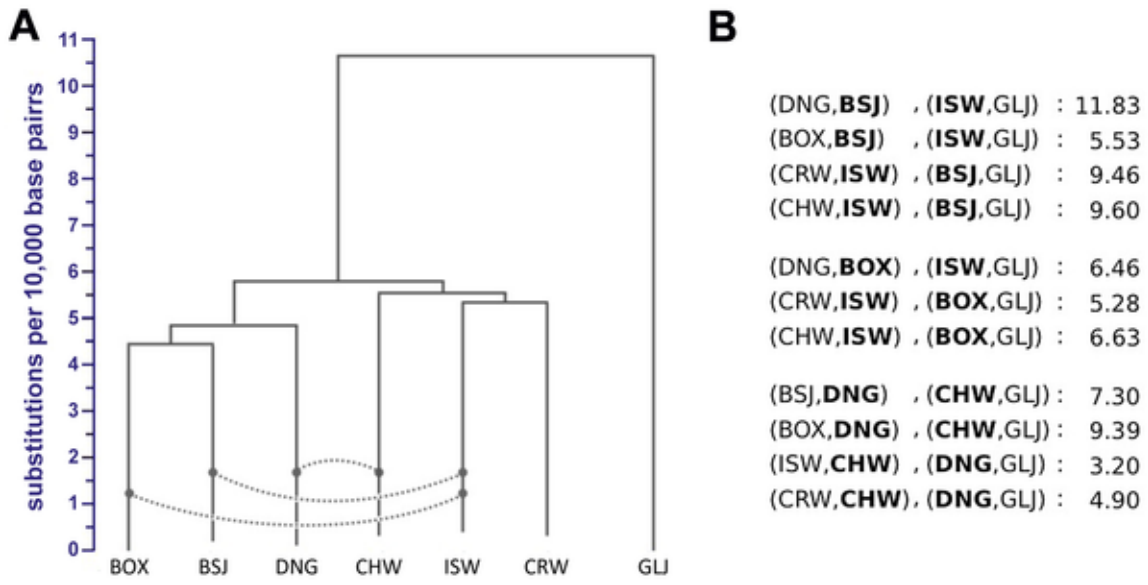

Figure 2.11 NJ tree constructed from genome-wide pairwise divergence, calculated using equation E8.1. All nodes have 100% bootstrap support. Dashed lines indicate admixture edges that were statistically significant in ABBA/BABA tests. (B) ABBA/BABA tests with significant Z-scores (values ≥3 are significant). All comparisons made are shown in Table 2.3. For each row, boldfaced labels indicate admixing lineages.

2.5 Model fit using the ABBA/BABA/BBAA configurations statistics

We tested the fit of the three models analyzed with *G-PhoCS* using the proportion of sites that contain alleles that are shared between two lineages but not the other two when comparing four species. The ABBA and BABA sites are defined following the notation seen in Section 2.4. On the other hand, a BBAA site is defined as one where the lineages P_1 and P_2 share one allele while the two other lineages P_3 and O share a different allele. The proportion of those three types of sites is reflective of the

genealogies contained in the data when comparing four lineages, where those genealogies are affected by gene flow and the divergence time between species. For a quartet of lineages P_1, P_2, P_3 and O we estimated the frequency of a site being ABBA, BABA or BBAA given that there are two alleles, each present in two of the four species as:

$$f(ABBA \mid two\ alleles, each\ in\ two\ species) = \frac{N(ABBA)}{N(ABBA) + N(BABA) + N(BBAA)}$$

$$f(BABA \mid two\ alleles, each\ in\ two\ species) = \frac{N(BABA)}{N(ABBA) + N(BABA) + N(BBAA)}$$

$$f(BBAA \mid two\ alleles, each\ in\ two\ species) = \frac{N(BBAA)}{N(ABBA) + N(BABA) + N(BBAA)}$$

(E2.4-6)

We refer to these estimates as relative frequencies of ABBA, BABA and BBAA sites, respectively. In the equations, N(ABBA), N(BABA) and N(BBAA) are the number of ABBA, BABA and BBAA sites.

The counts of ABBA, BABA and BBAA sites in the data were calculated using the 18 quartet configurations that are shown in Table 2.4 with two additional quartet configurations that contain either three dogs or three wolves. Those two additional configurations were added because they are informative about the actual phylogenetic relationships inside dogs and inside wolves, respectively. A demographic model would be more likely to be correct if it captures similar values for E2.4-2.6 as those seen in data. The estimates of the number of ABBA/BABA/BBAA sites in the data are shown in Table 2.4, along with the estimates of the relative frequency of those sites.

To mimic the empirical analysis (see above) we initially simulated 422 regions of 5Mb using the three models analyzed by *G-PhoCS* (more details on Text S9 from Freedman et al., 2014). However, because this produced an excess of ABBA/BABA/BBAA sites, to match the counts of these site classes seen in the data, we reduced our region size, instead simulating 422 regions of 2Mb. The simulations were performed using the following command lines:

1) Model where the dogs and wolves are each a separate clade (Command Line 7 in Appendix). This command line is identical to Command Line 1, with the only difference being the number of bases simulated.

2) Regional domestication model (Command Line 8 in Appendix).

3) Origin of dogs from the Israeli wolf (Command Line 9 in Appendix).

As a measure of the fit of each model to the data, we calculated the total difference between each model and the data in the relative frequencies of the ABBA/BABA/BBAA sites using the following equation:

$$\begin{aligned} Absolute\ Error &= \sum_{i=1}^{combinations} | f(ABBA \mid two\ alleles, each\ in\ two\ species)_{model} \\ &- f(ABBA \mid two\ alleles, each\ in\ two\ species)_{data} | \\ &+ | f(BABA \mid two\ alleles, each\ in\ two\ species)_{model} \\ &- f(BABA \mid two\ alleles, each\ in\ two\ species)_{data} | \\ &+ | f(BBAA \mid two\ alleles, each\ in\ two\ species)_{model} \\ &- f(BBAA \mid two\ alleles, each\ in\ two\ species)_{data} | \end{aligned} \quad \text{(E8.7)}$$

Overall, we found that the model which provided a better fit to the data, in terms of a smaller absolute error as estimated by E8.7, was the model which assumes that the dogs and wolves are each a separate clade whereas the model which provided the

worst fit was the one which assumes a regional domestication model (Table 2.5).

Using a threshold of 1.5% to look for important absolute differences between the data and the model in terms of relative frequencies, we found larger differences in the relative frequencies of BBAA sites in the data and the model that provided a better fit to the data in comparisons that included the Dingo, Chinese Wolf and another species of dog. We also found that the model which provided a better fit to the data incorrectly estimated the relative frequencies of ABBA sites in comparisons including the Chinese Wolf as P_1, Israeli wolf as P_2 and the Boxer or Basenji as P_3. Additionally, the number of BBAA sites in the quartet Boxer (P_1), Dingo (P_2) and Croatian Wolf (P_3) deviated substantially from those observed in the empirical data.

The regional domestication model overestimated the relative frequency of shared sites between Basenji and Dingo and underestimated the relative frequency of sites shared between (Dingo, Boxer) and (Boxer, Basenji) in comparisons that included the three dogs and the golden jackal. This shows that the phylogenetic relationships between dogs are more severely distorted under this model. This is also exemplified by the poor fit to the data in terms of the relative frequencies of ABBA/BABA/BBAA sites in the comparisons that include the Dingo, Boxer and another species of wolf. As in the model from Fig. 5A, the number of BBAA sites was also underestimated in the quartet Basenji (P_1), Dingo (P_2) and Chinese Wolf (P_3).

As with the best model, the model that posits the origin of dogs from the Israeli Wolf had poor fit to the data with respect to the relative frequency of BBAA sites in the comparisons of Boxer (P_1), Dingo (P_2) and Chinese Wolf (P_3). The latter model also had problems fitting the relative frequencies of the three types of sites we were

inspecting in comparisons that included the Israeli Wolf, Croatian Wolf and a dog. The relative frequency of BBAA sites in the comparison of Boxer, Dingo and Croatian Wolf was underestimated under this model.

Table 2.4. Estimates of the number of ABBA/BABA/BBAA sites in the six canid genomes. For each cell and each quartet comparison we report the number of ABBA/BABA/BBAA sites followed by the frequency of those three types of sites given that the site is bi-allelic with the two alleles found in two species each. The golden jackal was used as an outgroup in all comparisons.

			Data		
P1	**P2**	**P3**	**ABBA Sites**	**BABA Sites**	**BBAA Sites**
Basenji	Dingo	Croatian wolf	164211; 28.43%	162364; 28.11%	250958; 43.45%
Basenji	Dingo	Israeli wolf	158610; 27.18%	179656; 30.78%	245329; 42.04%
Boxer	Basenji	Croatian wolf	144942; 24.82%	146113; 25.02%	292896; 50.16%
Boxer	Basenji	Israeli wolf	157007; 26.71%	147991; 25.17%	282873; 48.12%
Boxer	Dingo	Croatian wolf	177485; 27.15%	176031; 26.93%	300095; 45.91%
Boxer	Dingo	Israeli wolf	176511; 26.50%	189294; 28.42%	300201; 45.07%
Croatian wolf	Israeli wolf	Boxer	226123; 34.16%	210897; 31.86%	224971; 33.98%
Croatian wolf	Israeli wolf	Dingo	213742; 32.78%	212876; 32.65%	225351; 34.56%
Croatian wolf	Israeli wolf	Basenji	205695; 35.29%	182191; 31.26%	194909; 33.44%
Basenji	Dingo	Chinese wolf	173366; 29.45%	162030; 27.52%	253270; 43.02%
Boxer	Basenji	Chinese wolf	149172; 24.91%	147273; 24.59%	302448; 50.50%
Boxer	Dingo	Chinese wolf	192400; 28.40%	175946; 25.97%	309223; 45.64%
Croatian wolf	Chinese wolf	Boxer	216145; 32.52%	219859; 33.08%	228675; 34.40%
Croatian wolf	Chinese wolf	Dingo	221737; 33.97%	212060; 32.49%	218959; 33.54%
Croatian wolf	Chinese wolf	Basenji	190706; 32.79%	191336; 32.90%	199502; 34.31%
Chinese wolf	Israeli wolf	Boxer	242452; 35.42%	222327; 32.48%	219803; 32.11%
Chinese wolf	Israeli wolf	Dingo	223003; 33.37%	232071; 34.73%	213209; 31.90%
Chinese wolf	Israeli wolf	Basenji	216213; 36.43%	191475; 32.26%	185855; 31.31%
Basenji	Dingo	Boxer	179362; 32.42%	216634; 39.16%	157265; 28.43%
Chinese Wolf	Croatian Wolf	Israeli Wolf	230181; 34.70%	208597; 31.44%	224601; 33.86%

Table 2.5. Estimates of the number of ABBA/BABA/BBAA sites in the three *G-PhoCS* models analyzed. For each cell and each quartet comparison we report: 1) The number of ABBA/BABA/BBAA sites; 2) The frequency of those three types of sites given that the site is bi-allelic with the two alleles found in two species each and 3) the difference of that frequency in the simulations minus what is estimated in the data (when this difference is bigger than 1.5%, we highlight the cell in bold). The lower row of the table indicates the fit of the model to the data as estimated by equation 8.7. The golden jackal was used as an outgroup in all comparisons.

			Fig. 5A model (Model where the dogs and wolves are each a			**Fig. 5B model (Regional**			**Fig. 5C model (Origin of dogs**		
			ABBA	**BABA**	**BBAA**	**ABBA**	**BABA**	**BBAA**	**ABBA**	**BABA**	**BBAA**
Basenji	Dingo	Croatian wolf	177596; 28.53%;	180202; 28.95%;	264624; 42.52%;	178773; 28.94%;	177186; 28.68%;	261870; 42.39%;	178434; 28.88%;	177152; 28.67%;	262289; 42.45%;
Basenji	Dingo	Israeli wolf	173506; 27.87%;	191296; 30.72%;	257817; 41.41%;	173256; 27.83%;	192556; 30.93%;	256705; 41.24%;	173222; 27.82%;	188792; 30.32%;	260580; 41.85%;
Boxer	Basenji	Croatian wolf	157926; 25.24%;	158158; 25.28%;	309616; 49.48%;	155013; 24.78%;	156346; 24.99%;	314275; 50.23%;	158543; 25.42%;	158872; 25.47%;	306268; 49.11%;
Boxer	Basenji	Israeli wolf	168735; 26.93%;	155221; 24.78%;	302524; 48.29%;	165943; 26.52%;	155130; 24.79%;	304670; 48.69%;	167349; 26.80%;	155402; 24.89%;	301725; 48.32%;
Boxer	Dingo	Croatian wolf	172541; 27.69%;	175379; 28.14%;	**275228; 44.17%;**	**148908; 23.69%;**	**148654; 23.64%;**	**331136; 52.67%;**	172536; 27.92%;	171583; 27.76%;	**273917; 44.32%;**
Boxer	Dingo	Israeli wolf	173388; 27.77%;	177664; 28.45%;	273358; 43.78%;	**147173; 23.27%;**	**155660; 24.61%;**	**329753; 52.13%;**	171562; 27.53%;	175185; 28.11%;	276446; 44.36%;
Croatian wolf	Israeli wolf	Boxer	205879; 33.27%;	201724; 32.60%;	211157; 34.13%;	208604; 33.71%;	200215; 32.36%;	209921; 33.93%;	208423; 33.80%;	**207350; 33.62%;**	200941; 32.58%;
Croatian wolf	Israeli wolf	Dingo	203877; 32.96%;	201160; 32.53%;	213431; 34.51%;	202216; 32.78%;	202568; 32.84%;	212020; 34.37%;	205800; 33.35%;	209303; 33.92%;	**201941; 32.73%;**
Croatian wolf	Israeli wolf	Basenji	215597; 34.74%;	197696; 31.85%;	207361; 33.41%;	216547; 34.95%;	196012; 31.63%;	207051; 33.42%;	216467; 35.11%;	**203118; 32.94%;**	197038; 31.95%;

Basenji	Dingo	Chinese wolf	188009; 30.16%;	177552; 28.49%;	**257728; 41.35%;**	188470; 30.47%;	174988; 28.29%;	**254996; 41.23%;**	185253; 29.98%;	173424; 28.06%;	259312; 41.96%;
Boxer	Basenji	Chinese wolf	160801; 25.64%;	158007; 25.20%;	308245; 49.16%;	156840; 25.13%;	155804; 24.97%;	311426; 49.90%;	157369; 25.29%;	159053; 25.56%;	305845; 49.15%;
Boxer	Dingo	Chinese wolf	184167; 29.48%;	170916; 27.36%;	**269545; 43.15%;**	**159174; 25.32%;**	**144656; 23.01%;**	**324831; 51.67%;**	178856; 28.94%;	168711; 27.30%;	**270441; 43.76%;**

Croatian wolf	Chinese wolf	Boxer	203311; 32.95%;	202091; 32.76%;	211562; 34.29%;	200348; 32.66%;	198041; 32.28%;	215078; 35.06%;	204468; 33.45%;	203864; 33.35%;	202947; 33.20%;
Croatian wolf	Chinese wolf	Dingo	213747; 34.53%;	196438; 31.74%;	208747; 33.73%;	209895; 34.22%;	193324; 31.52%;	210107; 34.26%;	210931; 34.50%;	201135; 32.90%;	199265; 32.60%;
Croatian wolf	Chinese wolf	Basenji	205710; 33.27%;	201464; 32.58%;	211167; 34.15%;	201556; 32.84%;	196880; 32.08%;	215250; 35.07%;	203801; 33.28%;	204552; 33.41%;	203964; 33.31%;
Chinese wolf	Israeli wolf	Boxer	**208018; 33.51%;**	205083; 33.04%;	207667; 33.45%;	210840; 34.03%;	204758; 33.05%;	203911; 32.91%;	210065; 34.00%;	209596; 33.92%;	198217; 32.08%;
Chinese wolf	Israeli wolf	Dingo	200720; 32.34%;	215312; 34.69%;	204645; 32.97%;	200301; 32.34%;	217224; 35.07%;	201859; 32.59%;	204194; 33.06%;	217493; 35.21%;	195969; 31.73%;
Chinese wolf	Israeli wolf	Basenji	**216436; 34.81%;**	202781; 32.61%;	202571; 32.58%;	217724; 35.14%;	201865; 32.58%;	199982; 32.28%;	218547; 35.38%;	204447; 33.10%;	194752; 31.53%;
Basenji	Dingo	Boxer	190695; 31.36%;	242304; 39.85%;	175036; 28.79%;	**244189; 40.10%;**	**219636; 36.07%;**	**145058; 23.82%;**	192265; 31.81%;	237327; 39.27%;	174739; 28.91%;
Chinese Wolf	Croatian Wolf	Israeli Wolf	208874; 33.63%;	203245; 32.73%;	208912; 33.64%;	206703; 33.38%;	198457; 32.05%;	214034; 34.57%;	204824; 33.27%;	200458; 32.56%;	210316; 34.16%;
		Absolute Error	0.4298			0.8219			0.4668		

Discussion

In this study, we generated high-quality individual canid genomes, and used them to uncover the history of dogs and gray wolves. Interpretation of the phylogenetic signals in these genomes was particularly challenging due to high levels of incomplete lineage sorting and post-divergence gene flow. We were able to disentangle the effects of these factors by using an array of recently developed statistical methods that together provided a detailed and robust inference of past demography for these canids. We used methods that rely on different aspects of this dataset: 1) whole-genome patterns of heterozygosity in single individuals (PSMC), 2) a subset of sites that are informative for post-divergence admixture (ABBA/BABA analyses) and 3) a set of neutral loci analyzed across all individuals jointly (*G-PhoCS*).

We found evidence of wolf-dog admixture in two divergent dog lineages (Basenji and Dingo). The fact that these lineages have been isolated from wolves geographically in the recent past suggests that this gene flow was ancestral and thus likely impacted multiple (if not most) dog lineages (Pickrell & Pritchard 2012; Vilà et al. 2005). Admixture has likely complicated previous inferences of dog origins. For instance, the presence of long shared haplotypes in Middle East wolves with several dog breeds (Vonholdt et al. 2010) may reflect historic admixture rather than recent divergence. Similarly, higher genetic diversity in East Asian dogs and affinities between East Asian village dogs and wolves (Pang et al. 2009; Savolainen et al. 2002; Wang et al. 2013) may be confounded by past admixture with wolves. In areas where village dogs (Boyko et al. 2009) roam freely and wolves have historically been in close proximity, admixture

may also be present and have non-trivial impact on patterns of genetic variation (Larson & Burger 2013).

Our inferences of ancestral population size from PSMC reveals an unexpected, roughly threefold population bottleneck in wolves. With PSMC, we detect the start of this bottleneck as early as 20 kya, while with *G-PhoCS* the bottleneck occurs at the timing of dog-wolf divergence, approximately 15kya. As our cross-validation between these two methods indicated that the timing of abrupt changes in N_e are overestimated by PSMC (Figure 2.6-2.10, more comparisons on Text S9 from Freedman et al. 2014), we place more confidence in the more recent date inferred with *G-PhoCS.* The bottleneck in wolves appears to have occurred before modern direct extermination campaigns by humans and within the timeframe of environmental and biotic changes associated with the ending of the Pleistocene. Although the specific cause of this bottleneck is unknown, it has important implications for understanding the process of dog domestication. Because of this bottleneck, we expect that at the onset of domestication, there was substantially more genetic diversity for selection to act on than observed in modern wolves. Direct comparisons of dog and wolf diversity (such as comparisons of heterozygosity) will not show as large a difference and thus previous studies that did not consider a wolf population decline (Kerstin Lindblad-Toh et al. 2005; Melissa M Gray et al. 2009) have underestimated the bottleneck associated with domestication. These previous studies estimated a two to fourfold reduction in dog $N_{e,}$, a far milder population contraction than the at least 16-fold reduction we infer here.

Overall, the genomes in this study reveal a dynamic and complex genetic history interrelating dogs and wolves. One question that remains unanswered has to do with

the geographic origins of dogs and the wolf lineage most closely related to them. Our analysis suggests that none of the sampled wolf populations is more closely related to dogs than any of the others and that dogs diverged from wolves at about the same time that the sampled wolf populations diverged from each other. One possible implication of this finding is that a more closely related wolf population exists today, but was not represented by our samples. We consider this unlikely, as we sampled the three major putative domestication regions, and previous SNP array studies have shown that wolf populations are only weakly differentiated, indicating that our sampled wolves should serve as good proxies for wolves in each broad geographic region (Vonholdt et al. 2010).

Another alternative is that the wolf population (or populations) from which dogs originated has gone extinct and the current wolf diversity from each region represents novel younger wolf lineages, as suggested by their recent divergence from each other. Our inference that wolves have gone through bottlenecks across Eurasia suggests a dynamic period for wolf populations over the last 20,000 years and that extinction of particular lineages is not inconceivable. Indeed, several external lines of evidence provide support for substantial turnover in wolf lineages. For example, ancient DNA, isotope, and morphologic evidence identify a divergent North American Late Pleistocene wolf (Leonard et al. 2007) and in Eurasia, similarly distinct wolves exist in the early archaeological record in Northern Europe and Russia, 15-36kya (Ovodov et al. 2011; Germonpré et al. 2012; Germonpré et al. 2009). Presumed changes in available prey (e.g. megafaunal extinctions) as habitats shrunk with the expansion of humans and agriculture also suggest the plausibility of wolf population declines and lineage turnover.

A remaining alternative for our inferred population phylogeny is that the basal lineage was absorbed into the three lineages sampled. Such a hypothesis is questionable though, as it requires there to be enough effective gene flow among the three wolf lineages such that no single lineage today serves best as a proxy for the basal lineage in our analysis. If true, the hypothesis that dogs were originally domesticated from a now-extinct wolf population suggests that ancient DNA studies will play a central role in advancing our understanding of the rapid transition from a large, aggressive carnivore to the omnivorous domestic companion that is a fixture of modern civilization.

Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs

The research from this chapter is contained in two papers:

** Denotes equal contributions*

Marsden CD*, Ortega-Del Vecchyo D*, O'Brien DP, Taylor JF, Ramirez O, Vilà C, Marques-Bonet T, Schnabel RD, Wayne RK, Lohmueller KE. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proceedings of the National Academy of Sciences of the United States of America. 2016; 113: 152-157.

Ortega-Del Vecchyo D, Marsden CD, Lohmueller KE. PReFerSim: Fast simulation of demography and selection under the Poisson Random Field model

Introduction

Many of the mutations that arise in genomes are weakly deleterious and reduce fitness but are not always eliminated from the population by purifying natural selection. Consequently, understanding the reasons why deleterious mutations persist in populations and the role of demographic history in this process is of considerable interest (Lohmueller 2014a; Lohmueller et al. 2008; Simons et al. 2014; Do et al. 2015; Fu et al. 2014; Henn et al. 2015; Gazave et al. 2013; Peischl et al. 2013; Schubert et al. 2014) . The radiation of domestic dogs offers a unique opportunity to address these

questions. Dogs were originally domesticated from ancestral gray wolf populations >15,000 years ago in a process involving one or more severe population bottlenecks (Freedman et al. 2014; Boyko 2011; vonHoldt et al. 2010). The more recent formation of modern dog breeds, which occurred over the last 300 years, involved additional population bottlenecks, intense artificial selection and inbreeding (K. Lindblad-Toh et al. 2005; Boyko 2011). Although this history is predicted to have resulted in the accumulation of deleterious variants, its specific effect on genome-wide patterns of deleterious variation remains unclear.

Here, we utilize complete genome sequencing data from 46 dogs representing 34 breeds, 25 village dogs, and 19 wolves to directly examine patterns of deleterious genetic variation across the dog genome. As over half of these data derive from our own sequencing efforts, this project represents the largest survey of dog genetic diversity based on genome sequences to date. Overall, we find that population bottlenecks associated with domestication have resulted in a proportional increase of amino acid changing variants in dogs relative to wolves and also have led to an increase in the additive genetic load in dogs relative to wolves. Our results indicate that the domestication process has dramatically re-shaped patterns of deleterious variation across the dog genome.

Results

3.1 Description of the data

Using a combination of in-house generated data (*n* = 52) and published sequences (*n* = 38); (Auton et al. 2013; Wang et al. 2013; Zhang et al. 2014), we collated a dataset of 90 canid whole genomes representing 46 breed dogs, 25 village dogs, 19 gray wolves as well as a single genome from a golden jackal to polarize ancestral and derived states. Our analyses focused on patterns of genetic diversity at putatively neutral sites far from genes, four-fold degenerate sites (non-amino acid changing coding variants) and zero-fold degenerate sites (amino acid changing coding variants).

We divided our dataset into two groups based on sequencing coverage. The first group contains the subset of genomes with high sequencing coverage (>15x) comprising 25 breed dogs and 10 wolves. For this dataset we called individual genotypes using GATK (DePristo et al. 2011). The second group consists of all 90 canid genomes. Many of these genomes have low sequence depth where genotype calls are less reliable. For these data, we estimated per individual heterozygosity (i.e. average pairwise differences between sequences) using a maximum likelihood approach based directly on the resampling of 4 sequence reads per site using the script FourSite (https://github.com/LohmuellerLab; SI from Marsden et al., 2016).

To assess the performance of this method, we compared our read-based estimates of heterozygosity to those from genotypes called using GATK (DePristo et al. 2011) on a subset of high-coverage genomes. We found the two estimates of heterozygosity to be highly concordant, suggesting that our estimator performs well (Figure 3.1). Importantly, because our read-based estimator was applied to subsamples of only four reads per individual, it is appropriate even for the lower-coverage genomes.

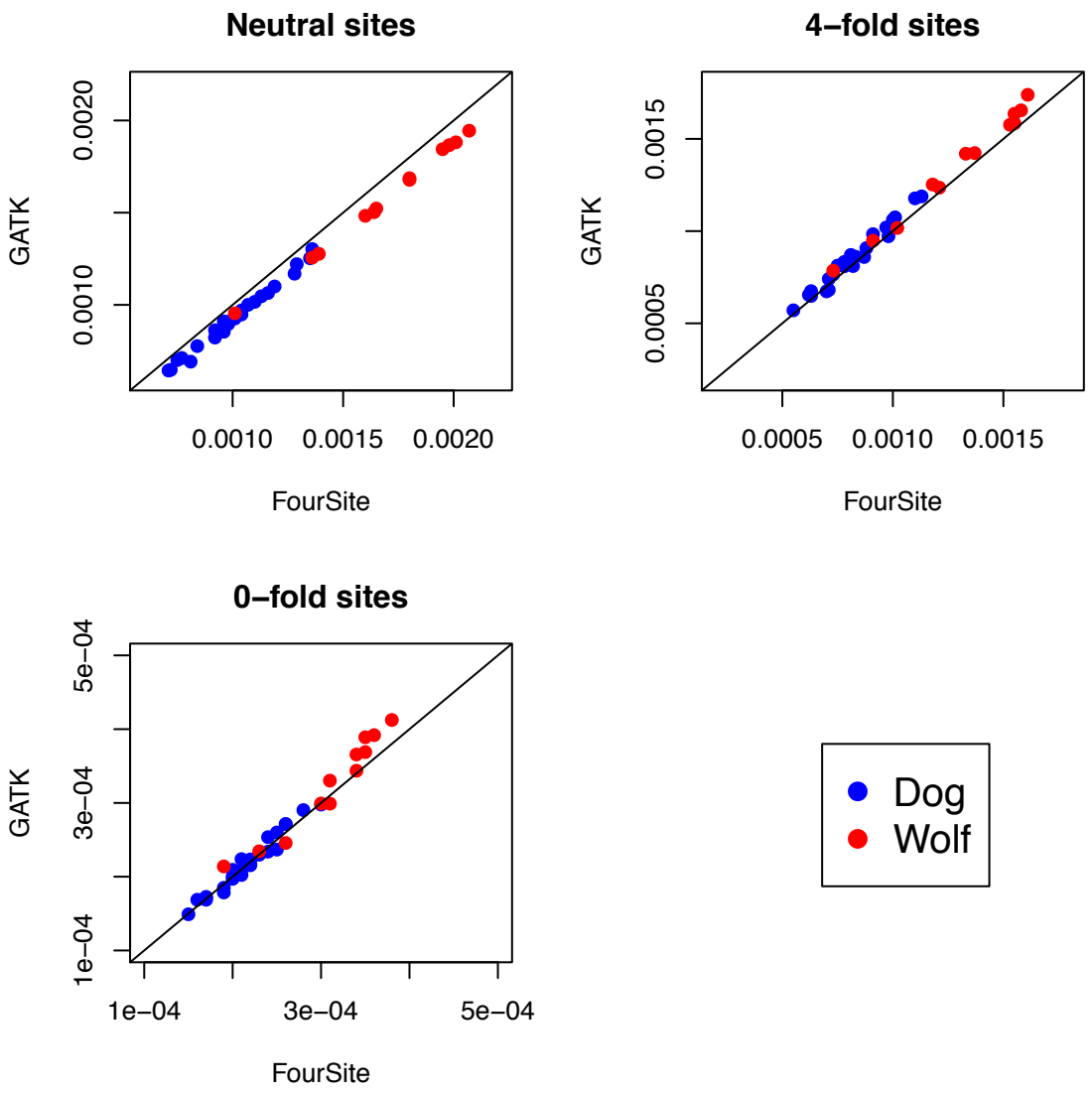

Fig. 3.1: Comparison of the read-based estimator of heterozygosity (FourSite) to the estimates based on GATK for high coverage individuals.

Lines denote the diagonal. Each blue point represents a breed dog. Each red point represents a wolf. Note the close correspondence between the estimates of heterozygosity obtained using FourSite to those from GATK. Importantly, only 4 reads per individual per site were used with FourSite while all the reads that passed our quality filters were used for calling genotypes with GATK.

3.2 Genome-wide patterns of deleterious variation

Because we typically have only 1-2 genomes per breed/population, we first focus on patterns of heterozygosity. To evaluate the role of population size in affecting deleterious variation, we calculate the ratio of 0-fold to 4-fold heterozygosity (Fay et al. 2001; Elyashiv et al. 2010; Akashi et al. 2012). This ratio is an estimate of the proportion of amino acid changing mutations that are not removed by selection. Assuming constant selection coefficients across populations, changes in this ratio indicate that demographic effects modulate the efficacy of selection. We chose this metric because it quantifies how demography affects selection without estimating parameters in complex demographic models for all populations (Akashi et al. 2012; Elyashiv et al. 2010).

In our data, the ratio of 0-fold heterozygosity to 4-fold heterozygosity shows a strong negative correlation with levels of neutral heterozygosity (Pearson's r = -0.534, $P<6 \times 10^{-8}$, Fig. 3.2B; Fig. 3.2C; Fig 3.3A and Table 3.1). Breed dogs have lower levels of neutral heterozygosity than wolves, consistent with their bottlenecked demographic history. However, they show disproportionately higher levels of amino acid (0-fold) heterozygosity (Fig. 3.2B). This result is concordant with previous estimates based on more limited data (a single boxer genome and mtDNA data (Cruz et al. 2008; Björnerfeldt et al. 2006)) and suggests that the proportional elevation in deleterious amino acid variation in dogs relative to wolves is seen across a wide range of breeds. Much of this pattern is driven by the difference between breed dogs and wolves. It diminishes when analyzing them separately (Fig. 3.3B), though statistical power also is

reduced. Patterns of neutral heterozygosity in the village dogs fall between those of breed dogs and wolves, consistent with their intermediate effective population size and variable levels of admixture between modern and ancient breeds (Shannon et al. 2015). However, the ratio of 0-fold to 4-fold heterozygosity in village dogs depends to some degree on the filters employed and is either similar to that in breed dogs or intermediate to that of dogs and wolves (Fig 3.3C). Interestingly, several wolf populations appear to show lower levels of neutral heterozygosity and higher ratios of 0-fold to 4-fold heterozygosity than breed dogs. These include the Tibetan wolves, which were previously shown to have very low genetic diversity (Zhang et al. 2014) and the Isle Royale wolf, which is a highly inbred island population derived from two founders in the 1950s ((Wayne et al. 1991). The negative correlation in Fig. 1 is unlikely to be driven by hypermutable CpG sites (Fig. 3.3C) or regions affected by selective sweeps (Fig. 3.3D), as it persists after removing these genomic features.

Table 3.1: Comparison of the regression parameter estimates across different data sets

Dataset	Intercept	SE	Lower CI	Upper CI	Slope	SE	Lower CI	Upper CI
High Coverage*	0.2759	0.0035	0.2691	0.2827	-21.8183	2.8901	-16.1537	-27.4828
Foursite on high coverage#	0.2921	0.0048	0.2827	0.3016	-28.5896	3.7965	-21.1484	-36.0308
FourSite	0.3003	0.0062	0.2878	0.3124	-28.4416	4.8208	-18.9929	-37.8903

low-

coverage

---

*Denotes the estimates from the 35 high coverage genomes where genotypes were called using GATK.

#Denotes the estimates from the 35 high coverage genomes that were treated analyzed as low-coverage genomes. We sampled four reads per site and estimated heterozygosity using FourSite.

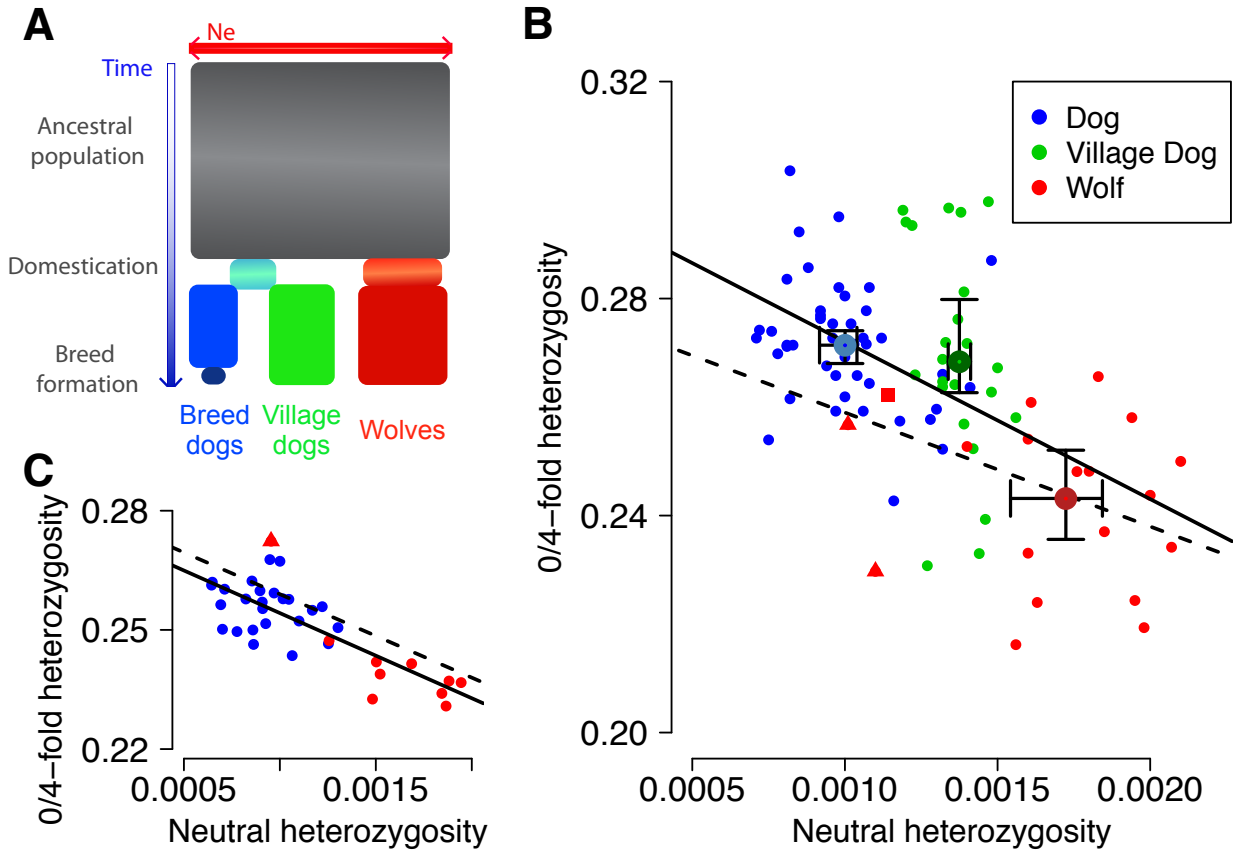

Fig. 3.2. Population history and deleterious genetic variation. (A) Conceptual model of dog domestication used in population genetic simulations. Box widths are proportional to estimated population sizes (Table 3.2). **(B)** The ratio of 0-fold to 4-fold heterozygosity vs. neutral genetic diversity. Observed heterozygosity is based on four reads per individual. The larger circles represent the trimmed median values for each population group and the error bars denote 95% confidence intervals on the trimmed median for each population group. Triangles denote the Tibetan wolves. A square denotes the Isle Royale wolf. The solid black line denotes the best-fit linear regression line (Intercept = 0.301, slope = -29.00, r = -0.534, $P < 6 \times 10^{-8}$). The dashed line denotes the best-fit linear regression line from forward simulations of demography and negative selection (SI Appendix, Tables S4 and S7). **(C)** The ratio of 0-fold to 4-fold heterozygosity vs. neutral genetic diversity in the 35 high-coverage genomes where genotypes were called using GATK. The solid black line denotes the best-fit linear regression line (Intercept = 0.276, slope = -21.40, r = -0.777, $P < 5 \times 10^{-8}$) and the dashed line is as described in (**B**).

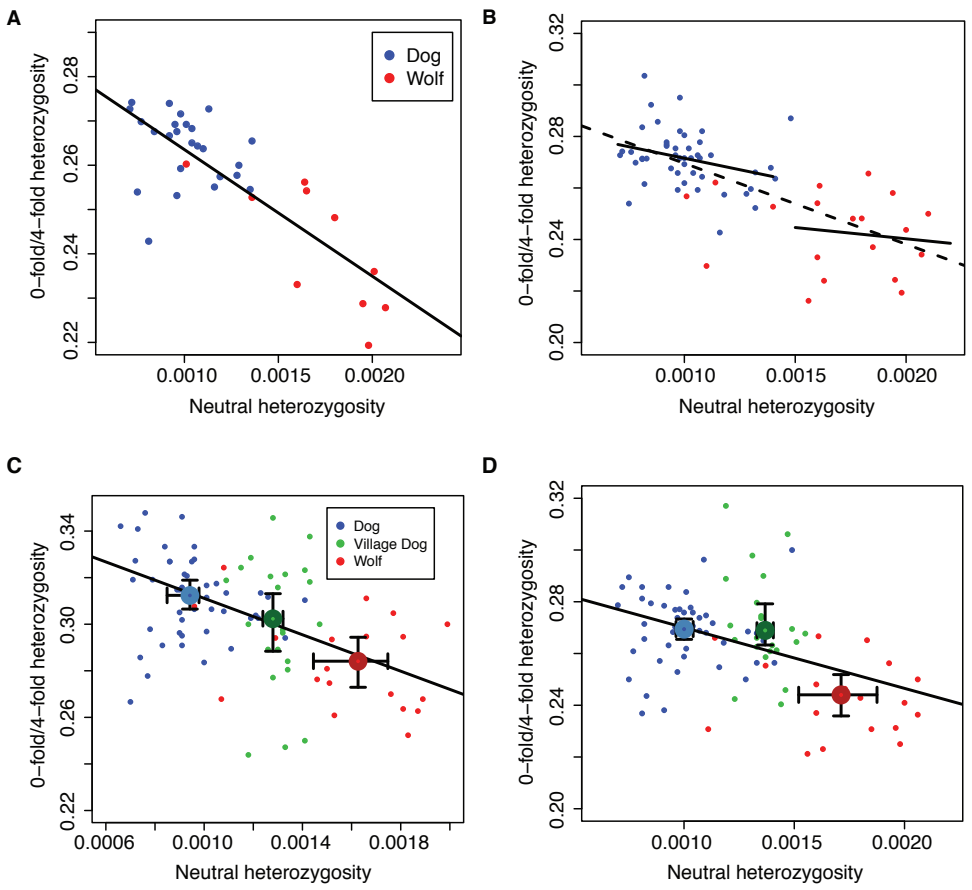

Fig. 3.3. The ratio of 0-fold to 4-fold heterozygosity is negatively correlated with neutral genetic diversity.

(A) This analysis uses 35 high coverage genomes treated as though they have lower-coverage. We sampled 4 reads per site and estimated heterozygosity using FourSite. The solid line denotes the best-fit linear regression line (Intercept = 0.276, slope = -21.4, r = 0.78, $P < 5 \times 10^{-8}$). **(B)** This analysis considers dogs (blue) and wolves (red) separately. Heterozygosity was computed using four reads per individual. The dashed line denotes the best-fit linear regression line (Intercept = 0.30, slope = -31.4, r = -0.674, $P < 6 \times 10^{-10}$) for both breed dogs and wolves together. Dogs show a slight negative

relationship (solid line over blue points; Intercept = 0.29, slope = -17.82, r = -0.30 P = 0.043) while wolves do not (solid line over red points; Intercept = 0.26, slope = -8.75, r = -0.185, P = 0.45). However, due to the limited sample size, statistical power is diminished within each group. **(C)** This analysis filters CpG sites. Heterozygosity was computed using four reads per individual. Error bars denote 95% confidence intervals on the trimmed median for each population group. The solid line denotes the best-fit linear regression line (Intercept = 0.350, slope = -39.09, r = -0.468, $P < 4 \times 10^{-6}$). **(D)** This analysis filters sites near a selective sweep. Heterozygosity was computed using four reads per individual. Error bars denote 95% confidence intervals on the trimmed median for each population group. The solid line denotes the best-fit linear regression line (Intercept = 0.294, slope = -23.50, r = -0.430, $P < 3 \times 10^{-5}$).

3.3 Forward-in-time simulations using PReFerSim

I modified a forward-in-time simulator previously designed by Kirk Lohmueller (Lohmueller 2014b) to make it capable of modeling demographic scenarios with many population size change events, different distributions of fitness effects, changes in the relaxation of selective constraints and in the inbreeding coefficients through time. Additionally, I added options to print nine different sets of summary statistics of genetic variation and the possibility to follow allele frequency trajectories conditioning on the present-day allele frequency. The interface of inputs and output files from the program was also modified to make it more user-friendly and to facilitate the use of the software in computing clusters. The program, PReFerSim, is available on:

https://github.com/LohmuellerLab/PReFerSim

PReFerSim performs simulations under the Poisson Random Field model, where it is assumed that the number of independent mutations arising each generation follows a Poisson distribution with a mean equal to $2N_iul$, where N_i is the effective population size in generation i, u is the mutation rate and l is the number of independent sites being simulated. Each simulation replicate was performed assuming a sequence length of 10 million independent sites, where each site can contain one or two alleles and only one mutation can take place in each site. Since its emergence, the frequency of a mutation p_{i+1} in successive generations follows a binomial distribution $Bin(N_{i+1},p')$, where:

$$p' = \frac{(1-s)(p^2+Fpq)+(1-hs)pq(1-F)}{q^2+pqF+(1-hs)2pq(1-F)+(1-s)(p^2+Fpq)},$$

and p is the frequency of the derived allele in generation N_i, q is the frequency of the ancestral allele, s is the selection coefficient, F is the inbreeding coefficient and h is the dominance coefficient. For neutral sites, the value of s is equal to 0. Synonymous sites were also assumed to be neutral and had a value of s equal to 0. Values of s for nonsynonymous mutations were drawn from a gamma distribution of selective effects assuming that, unless otherwise noted, all the mutations were additive (h = 0.5). Simulations were performed under a variety of models of population history. See below for further details on the mutations rates, demographic models, and distributions of selective effects used in the simulations.

At the end of the simulation, we sampled individual animals and computed their heterozygosity. Because some of our simulations included recent inbreeding, and inbreeding results in an increase in homozygosity relative to a randomly mating population, we included its effects in computing heterozygosity. Specifically, the

genotype for animal *i* at site *j* was drawn from a multinomial distribution with probabilities:

$$P(Genotype) = \begin{cases} p_j^2 + p_j q_j F & \text{Homozygous for the derived allele} \\ 2 p_j q_j (1-F) & \text{Heterozygous} \\ q_j^2 + p_j q_j F & \text{Homozygous for the ancestral allele} \end{cases},$$

where p_j is the frequency of the derived allele in the population at site *j*, q_j is the frequency of the ancestral allele at site *j* and *F* is the inbreeding coefficient used in the simulation.

Heterozygosity was computed by dividing the number of heterozygous sites over the number of sites simulated per simulation replicate (10 million). We also determined the proportion of heterozygous sites when sampling one allele from two different canids. To do this, we sampled two genotypes following the previous equation. Then, we sampled one allele from each of the two genotypes and determined that the site was heterozygous if the two sampled alleles were different. The total proportion of heterozygous sites in two dogs was obtained by dividing the number of heterozygous sites by the number of sites simulated.

We accounted for differences in mutation rates (μ) across 0-fold, 4-fold, and neutral sites using the following approach. First, we assume that CpG sites have a 10-fold higher mutation rate than do non-CpG sites. The mutation rates employed for the neutral, 0-fold and 4-fold sites used for the simulations were dependent on the proportion of CpG sites found within those three categories of sites in humans and our data. We then obtained estimates for the proportion of CpG sites in different functional categories from the literature. (Veeramah et al. 2014) found that in humans 5.75% of

the autosomal 0-fold sites were CpG sites while 9.1% of the autosomal 4-fold sites were CpG sites. To estimate the proportion of CpG sites in neutral variants, we observed that when we looked for all the mutations involving a 'CG' motif (i.e., any site after a C and any site before a G) from our data, where not necessarily all of those 'CG' motifs were CpG sites, we observed that 'CG' motifs were 9.1% more frequent in 0-fold sites than in neutral sites. This suggested that there should be around 9.1% more CpG sites in 0-fold sites as compared to neutral sites. Therefore, we reasoned that the proportion of CpG sites in neutral sites was equal to 5.27%. Using these numbers, we obtained the mutation rates of different categories of sites as:

$$\mu = P(CpG)(10)(B) + (1 - P(CpG))(B),$$

where *P*(CpG) is the proportion of sites that are CpGs in that category of sites and *B* is the background mutation rate for non-CpG sites. The factor of 10 indicates the assumed 10-fold increase in the mutation rate at CpG cites. We began with a neutral mutation rate of $\mu = 2 \times 10^{-8}$. Then, for neutral sites with P(CpG) = 5.27%, we obtain $B = 1.356 \times 10^{-8}$. That equation can also be used to obtain the mutation rate for 0-fold and 4-fold sites by using that same value of *B* and replacing *P*(CpG) by 5.75% and 9.1%, respectively. Using this procedure we obtain mutation rates of $\mu = 2.468 \times 10^{-8}$ for 4-fold sites and $\mu = 2.059 \times 10^{-8}$ for 0-fold sites. These mutation rates were used for the forward simulations.

We examined three different models of population history for canids. First, we used the demographic model for wolves and dogs inferred in Freedman et al. (Freedman et al. 2014) as a basis to simulate genetic variation that mimics the demographic history of wolves, village dogs and breed dogs (Table 3.2). Because we

assumed a per-base pair per-generation neutral mutation rate of 2 x10^{-8}, and Freedman et al. assumed a mutation rate of 1 x 10^{-8}, we rescaled the N_e and divergence times from the Freedman et al. study. Breed dogs were assumed to have been formed 100 generations ago to be consistent with the historical records and previous work (M. M. Gray et al. 2009; Boyko 2011). This breed formation was modeled as a decrease in population size. We explored different realistic effective population sizes for the most recent (around 2,500) generations in all populations to assess their effect on neutral heterozygosity and the ratio of nonsynonymous to neutral site heterozygosity. The values shown in SI Appendix, Table 3.2 show the final parameter values used in the simulations. Our second scenario also used the Freedman et al. (Freedman et al. 2014) demographic model as a backbone. But, here we increased the effective population size in the second epoch from 44,993 to 60,000 individuals. As expected, this model showed higher values of neutral heterozygosity. The parameters of this model are given in SI Appendix, Table 3.3. Finally, the third model that we considered was that fit to village dogs and wolves by Wang et al. (Wang et al. 2013) (Table 3.4). We made several simplifying assumptions and replaced exponential growth with piece-wise constant population sizes (Table 3.4).

Table 3.2: Forward simulation parameters based on the Freedman et al. demographic model

	Wolves		Village Dogs		Breed Dogs	
	Number of chromosomes ($2N_e$)	**Number of generations**	**Number of chromosomes ($2N_e$)**	**Number of generations**	**Number of chromosomes ($2N_e$)**	**Number of generations**
Epoch 1	18,169	145,352	18,169	145,352	18,169	145,352
Epoch 2	44,993	63,898	44,993	63,898	44,993	63,898
Epoch 3	24,000	237	1999	347	1999	347
Epoch 4	30,000	2243	15,000	2133	8000	2033
Epoch 5	-	-	-	-	1000	100

Epoch 1 denotes the ancestral population size. Epoch 4 denotes the current effective population size for Wolves and Village Dogs while Epoch 5 represents the current effective population size for Breed Dogs. This demographic model was used for the regression line presented in Fig. 3.2B.

Table 3.3: Forward simulation parameters based on the Freedman et al. model with larger ancestral population sizes

	Wolves		Village Dogs		Breed Dogs	
	Number of chromosomes ($2N_e$)	**Number of generations**	**Number of chromosomes ($2N_e$)**	**Number of generations**	**Number of chromosomes ($2N_e$)**	**Number of generations**

Epoch 1	18,169	145,352	18,169	145,352	18,169	145,352
Epoch 2	60,000	63,898	60,000	63,898	60,000	63,898
Epoch 3	2400	237	1999	347	1999	347
Epoch 4	30,000	2243	15,000	2133	8000	2033
Epoch 5	-	-	-	-	1000	100

Epoch 1 denotes the ancestral population size. Epoch 4 denotes the current effective population size for Wolves and Village Dogs while Epoch 5 represents the current effective population size for Breed Dogs.

Table 3.4: Forward simulation parameters based on the Wang et al. model

	Wolves		Village Dogs		Breed Dogs	
	Number of chromosomes ($2N_e$)	**Number of generations**	**Number of chromosomes ($2N_e$)**	**Number of generations**	**Number of chromosomes ($2N_e$)**	**Number of generations**
Epoch 1	35,000	280,000	35,000	280,000	35,000	280,000
Epoch 2	33,020	3556	5666	2556	5666	2,556
Epoch 3	-	-	11,332	1000	11,332	900
Epoch 4	-	-	-	-	200	100

Epoch 1 denotes the ancestral population size. Epoch 2 denotes the current effective population size for Wolves, Epoch 3 is the current effective population size for Village Dogs and Epoch 4 represents the current effective population size for Breed Dogs

Because the distribution of selective effects has not been estimated for new nonsynonymous mutations in dogs, the optimal parameters to use are not immediately clear. Thus, we examined different distributions of selective effects for new nonsynonymous mutations (Table 3.5). First we used estimates from other species. We fully acknowledge the distribution of selective effects may vary across species and these values may not be appropriate for dogs. However, our goal here is to determine whether plausible distributions of selective effects combined with demography can generate the qualitative patterns seen in our data, rather than perform a rigorous assessment of model fit. First, we used the gamma distribution that had been fit to human nonsynonymous SNP data by Boyko et al. (A. R. Boyko et al. 2008). Second, we

used a gamma distribution that had been fit to nonsynonymous SNPs in 10 *M. m. castaneus* individuals (D. L. Halligan et al. 2013). Importantly, because the ß (or scale) parameters of the gamma distribution are typically estimated as the population scaled selection coefficients (2*Ns*), we converted values of 2*Ns* drawn from the distribution into values of *s* by dividing by twice the relevant population (Table 3.5).

However, we found that both of these distributions of selective effects did not match the regression parameters relating the 0-fold/4-fold ratio and neutral heterozygosity for the observed data (Figure 3.4-3.5). In particular the Boyko et al. (A. R. Boyko et al. 2008) model from humans predicted a 0-fold/4-fold ratio that was too low compared to our data. This suggests that our data contains more nearly neutral ($s < 0.0001$) mutations than had been estimated from humans. Models including a few percent more mutations with $s < 0.0001$ better fit the observed data. The Halligan et al. mouse (D. L. Halligan et al. 2013) model (Table 3.5), which includes more mutations with $s < 0.0001$, better matches the observed 0-fold/4-fold ratio in dogs, but does not have a steep enough slope. This model contains too few moderately deleterious mutations ($0.0001 < s < 0.01$) that could be effectively removed by selection from the wolf population, but persist due to drift in dogs.

There are several possible reasons for this lack of fit of previous models to our data. First, the distribution of selective effects could be different in dogs than in humans and mice. The human and mouse distributions appear to differ from each other (Table 3.5), supporting the notion that this distribution may not be constant across species. Second, our simulations used to generate the relationship between the 0-fold/4-fold heterozygosity ratio and neutral heterozygosity assume that all variants are independent

of each other. If the real data includes substantial Hill-Robertson effects, then the data could differ from our simulations, even when the correct distribution of selective effects was used. Third, the distribution of selective effects may differ between dogs and wolves, perhaps because of domestication. If more new genetic variants in dogs became neutral after domestication, they may be able to drift to higher frequency, increasing the 0-fold/4-fold ratio. More detailed work on the distribution of selective effects in dogs and wolves is needed to distinguish among these possibilities.

Table 3.5: Parameters for the gamma distributions of selective effects on new mutations used in forward simulations of demography and selection

Model	**α**	**β**	***N***	**% mutations** $s<0.0001$	**% mutations** $0.0001<s<0.001$	**% mutations** $0.001<s<0.01$	**% mutations** $s>0.01$
Boyko (A. R. Boyko et al. 2008)	0.184	319.8626	1000	27.89	14.68	21.90	35.54
Mice (D. L. Halligan et al. 2013)	0.11	8,636,364	10^6	32.6	9.40	12.10	45.86
Gamma Test 1	0.25	250	10000	33.00	24.86	32.91	9.23
Gamma Test 2	0.3	100	10000	34.30	31.44	32.06	2.21

α denotes the shape parameter of the distribution of selective effects while β denotes the scale parameter. *N* refers to the population size that the beta parameter was scaled

by. Remaining columns provide the proportions of new mutations having different selection coefficients. The regression line from the Gamma Test 2 distribution is shown in Fig. 3.2B.

Because previously published distributions of selective effects appeared to not match the observed data, we explored several additional custom gamma distributions (Table 3.5). We found that models including a greater proportion of weakly deleterious (s < 0.001) and fewer strongly deleterious (s > 0.01) mutations provided a better fit to the data. In particular, a gamma distribution with a shape parameter of 0.3 and scale parameter of 0.05 (in terms of s) predicted regression coefficients intermediate between those seen in the low and high coverage datasets (Fig. 3.2B) under the Freedman et al. demographic model shown in Table 3.2. Under the Wang et al. demographic model (Table 3.4), a gamma distribution with a shape parameter of 0.25 and scale parameter of 0.125 reasonably predicts the observed regression parameters (Figure 3.4-3.5). While these distributions mimic the empirical patterns, other more complex distributions may be more biologically reasonable. As discussed above, further work on the distribution of selective effects is necessary to distinguish among these possibilities.

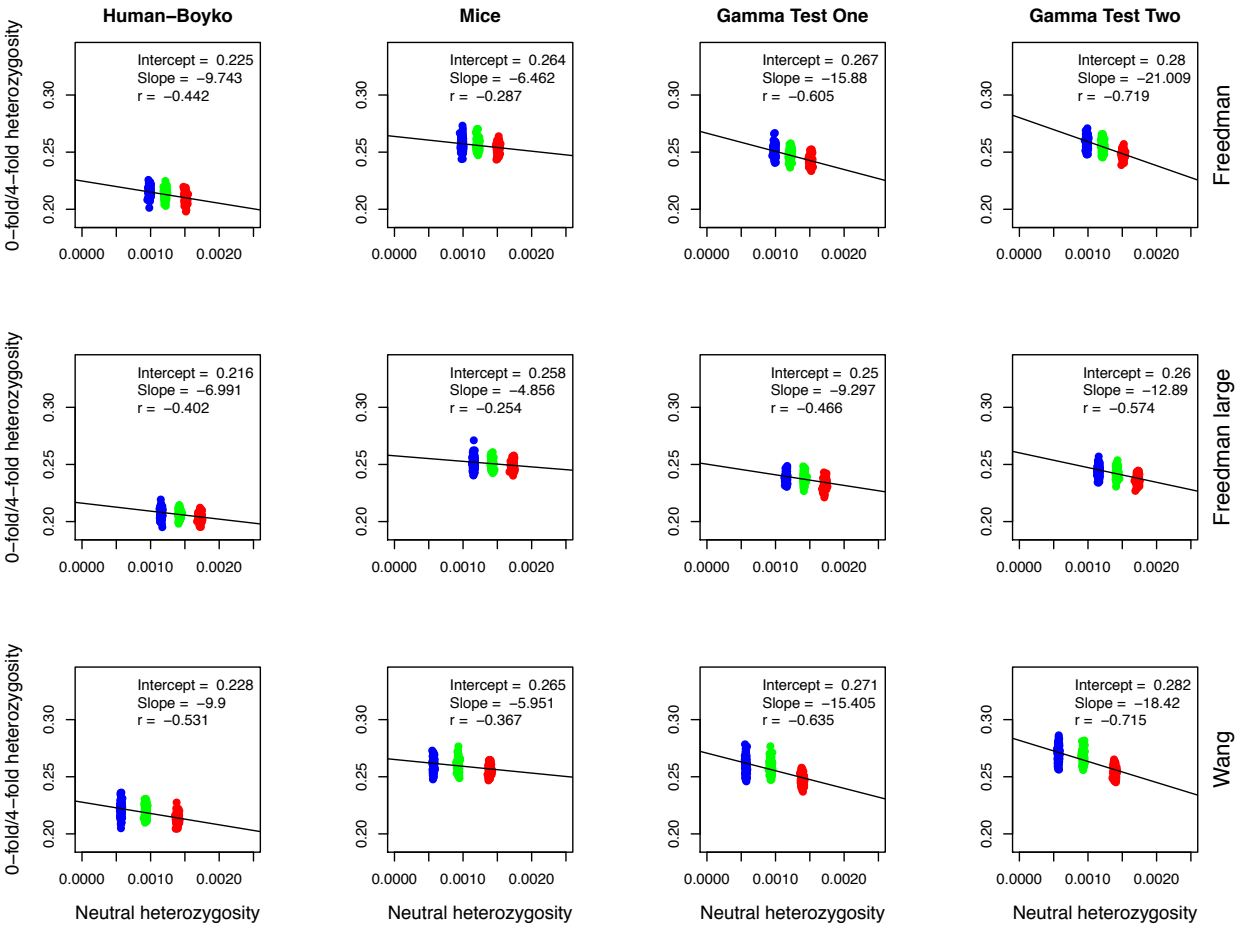

Fig. 3.4: Ratio of 0-fold to 4-fold heterozygosity vs. neutral heterozygosity from the forward simulations under different models of demography and selection.

Rows denote the different demographic models. “Freedman” refers to the Freedman et al. model (Table 3.2). “Freedman large” refers to the Freedman et al. model, but increasing the size of the ancient population size (Table 3.3). “Wang” denotes our implementation of the model fit in Wang et al. (Table 3.4). Columns denote different distributions of selective effects (Table 3.5). Lines are from the best-fit linear regression. Blue points denote breed dogs, green points denote village dogs, and red points represent wolves. In all cases, models of demography and selection predict a negative relationship between the ratio of 0-fold to 4-fold heterozygosity vs. neutral heterozygosity.

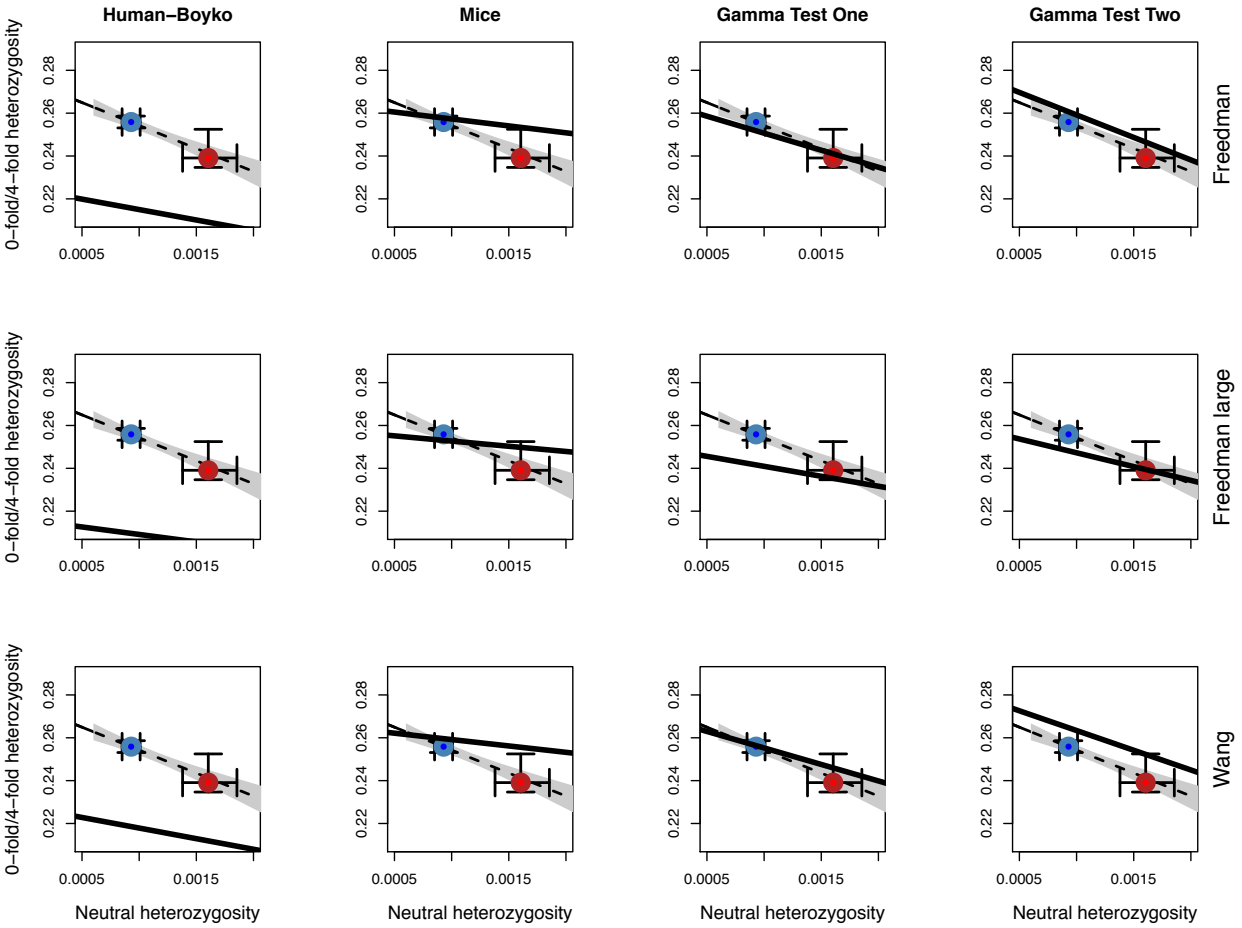

Fig. 3.5: Models of purifying selection and demography predict a similar negative relationship between 0-fold/4-fold heterozygosity and neutral heterozygosity as seen in the high quality genomes.

Rows denote the different demographic models (Table Columns denote different distributions of selective effects (Table 3.5). Dark solid black lines are from the best-fit linear regression of the simulations under the particular model. The gray shaded region denotes the 95% CI on the linear regression line calculated from the 35 high quality genomes (e.g., the data shown in Fig. 3.2B). The dark blue and red points represent the trimmed medians from the observed data from the breed dogs and wolves, respectively.

The whiskers denote 95% CIs on the trimmed medians. Note that the Gamma Test 2 distribution of selective effects best fits the observed relationship between 0-fold/4-fold heterozygosity and neutral heterozygosity under the Freedman demographic model. The Gamma Test 1 distribution also provides a good fit under the Wang demographic model.

We also performed a set of simulations where all mutations were recessive. Here we used the demographic model shown in Table 3.2. We used two different distributions of selective effects, the gamma distribution inferred in Boyko et al. as well as our Gamma Test 2 distribution. Overall, we found that the intercept of the regression of the 0-fold/4-fold heterozygosity ratio on neutral heterozygosity was higher with recessive effects than additive effects (Fig 3.4 with Fig 3.6). This finding is not surprising because, for the same distribution of *s*, recessive mutations are only selected against in the homozygous state and can thus drift up in frequency and persist in the population more easily than variants with additive effects. In contrast to the additive case, the slope of the regression was weakly positive when assuming fully recessive mutations. This result is in agreement with the recent theoretical findings of Balick et al. (Balick et al. 2015). Essentially, recessive alleles that survive during a bottleneck will have drifted to higher frequency and have a higher probability of being in the homozygous state compared to the same alleles in non-bottlenecked populations. When in the homozygous state, the recessive deleterious mutations can be removed by selection, leading to the decrease in the 0-fold/4-fold ratio in the bottlenecked population relative to the non-bottlenecked population. Because these simulations do not match the patterns seen in our data, and simulations including additive effects provide a better fit, we conclude that most

segregating amino acid changing variants in dogs and wolves are probably not fully recessive. They may be fully additive, however.

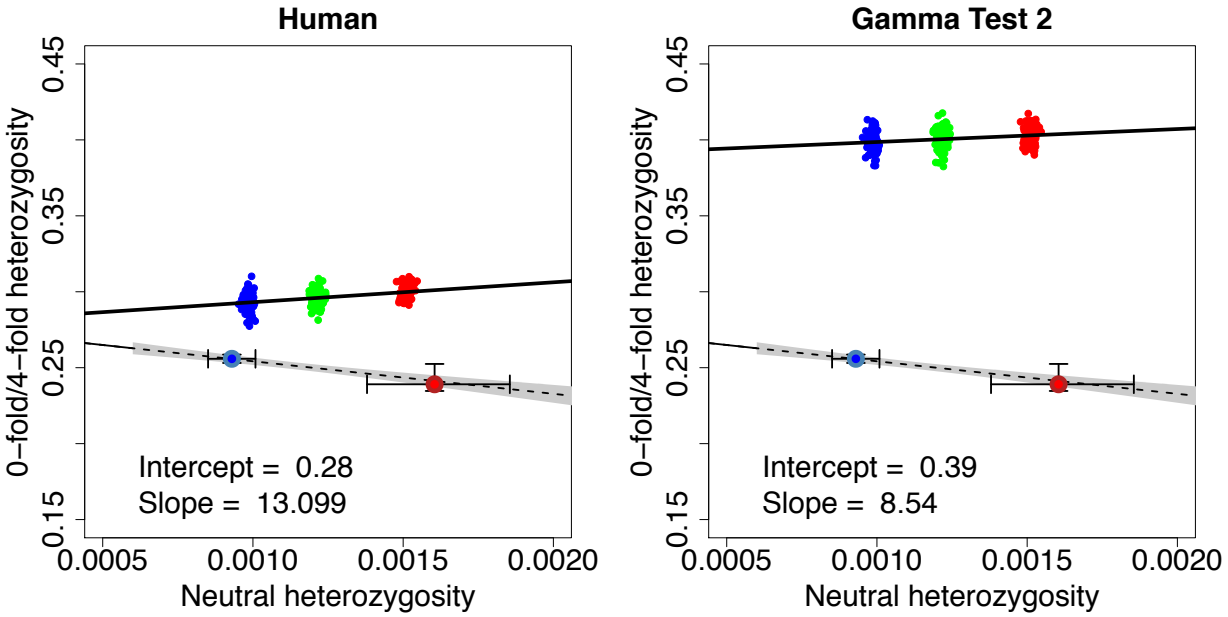

Fig. 3.6: Models with recessive effects predict a positive relationship between 0-fold/4-fold heterozygosity and neutral heterozygosity.

Breed dogs are in blue, village dogs in green, and wolves in red. All simulations assumed *h*=0 and the demographic parameters shown in Table 3.2. Columns denote different distributions of selective effects Table 3.5. The shaded gray lines denote the regression parameters from the simulations including additive effects. The clouds of blue, green, and red points denote the results of the simulations assuming recessive effects.

3.4 The role of recent inbreeding

Dogs from some breeds are homozygous for large (>1Mb) regions of the genome, suggesting recent mating among close relatives (i.e. inbreeding (Boyko et al. 2010), Fig 3.8). This inbreeding can reduce the effective population size, allowing deleterious alleles to drift higher in frequency and is a mechanism commonly assumed to account for the accumulation of deleterious mutations in dog genomes (McGreevy & Nicholas 1999) but has not been formally assessed. Based on three distinct analyses, we find that recent inbreeding is not driving the patterns shown in Fig. 3.2.

First, we conducted additional forward simulations including negative selection and recent inbreeding within breed dogs. Even strong inbreeding (F = 0.2) over the last 300 years, without the bottlenecks associated with domestication and breed formation, is insufficient to generate the observed negative relationship between the 0-fold/4-fold heterozygosity ratio and neutral heterozygosity (Fig. 3.7A). Second, we attempted to remove the effects of recent inbreeding on our analysis of heterozygosity. Because recent inbreeding increases the probability that two chromosomes within a given individual share a common ancestor with each other rather than with a chromosome from another individual (Fig 3.8A), it will reduce within-individual heterozygosity relative to between-individual heterozygosity (Wright 1951). Thus, we can obtain an estimate of heterozygosity removing the effects of inbreeding by sampling a single read from each individual at each site and determining whether the reads have different nucleotides. Forward simulations indicate that this approach removes the effects of recent inbreeding on heterozygosity (Fig 3.8B, Fig 3.8C). However, in contrast, in the actual data, neutral heterozygosity computed from two canids remains negatively correlated with the ratio of 0-fold to 4-fold heterozygosity (Fig. 3.8B), suggesting recent inbreeding

is not the cause of the association. Finally, when removing large runs of homozygosity (>2 MB) from our analyses, the negative relationship between neutral heterozygosity and the ratio of 0-fold to heterozygosity to 4-fold heterozygosity remained strong (Fig. 3.7C), indicating that it was not driven by patterns of variation within regions of the genome most affected by inbreeding. These unexpected findings imply that population bottlenecks, rather than recent inbreeding, are responsible for the proportional increase in amino-acid changing heterozygosity in breed dogs relative to wolves.

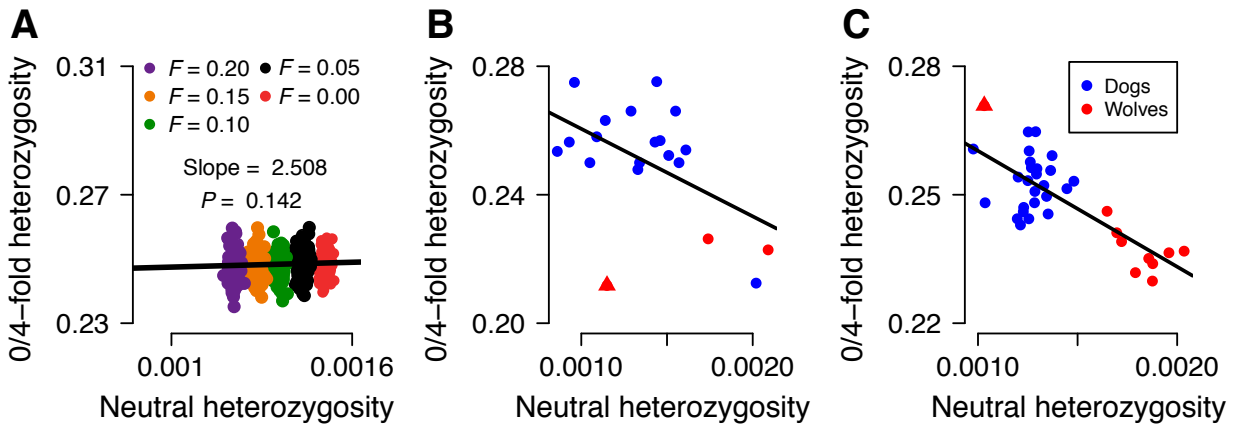

Fig. 3.7. Recent inbreeding does not drive the relationship between neutral heterozygosity and the 0-fold/4-fold heterozygosity ratio. (A) Forward simulations using a demographic model that includes inbreeding over the last 100 generations, but not bottlenecks associated with domestication or breed formation ("wolf" demographic model in Table 3.2). **(B)** Empirical results from computing heterozygosity using one read from each of two individuals per population. The solid line denotes the best-fit linear regression line (Intercept = 0.288, slope = -27.25, r = -0.502, P = 0.024). **(C)** The relationship between neutral polymorphism and the ratio of 0-fold to 4-fold heterozygosity persists when removing runs of homozygosity. The solid black line

denotes the best-fit linear regression line (Intercept = 0.276, slope = -21.40, r = -0.534, $P < 5 \times 10^{-8}$). This plot uses the same data as in Fig. 3.2C, but removing ROHs. Red triangles denote the Tibetan wolves.

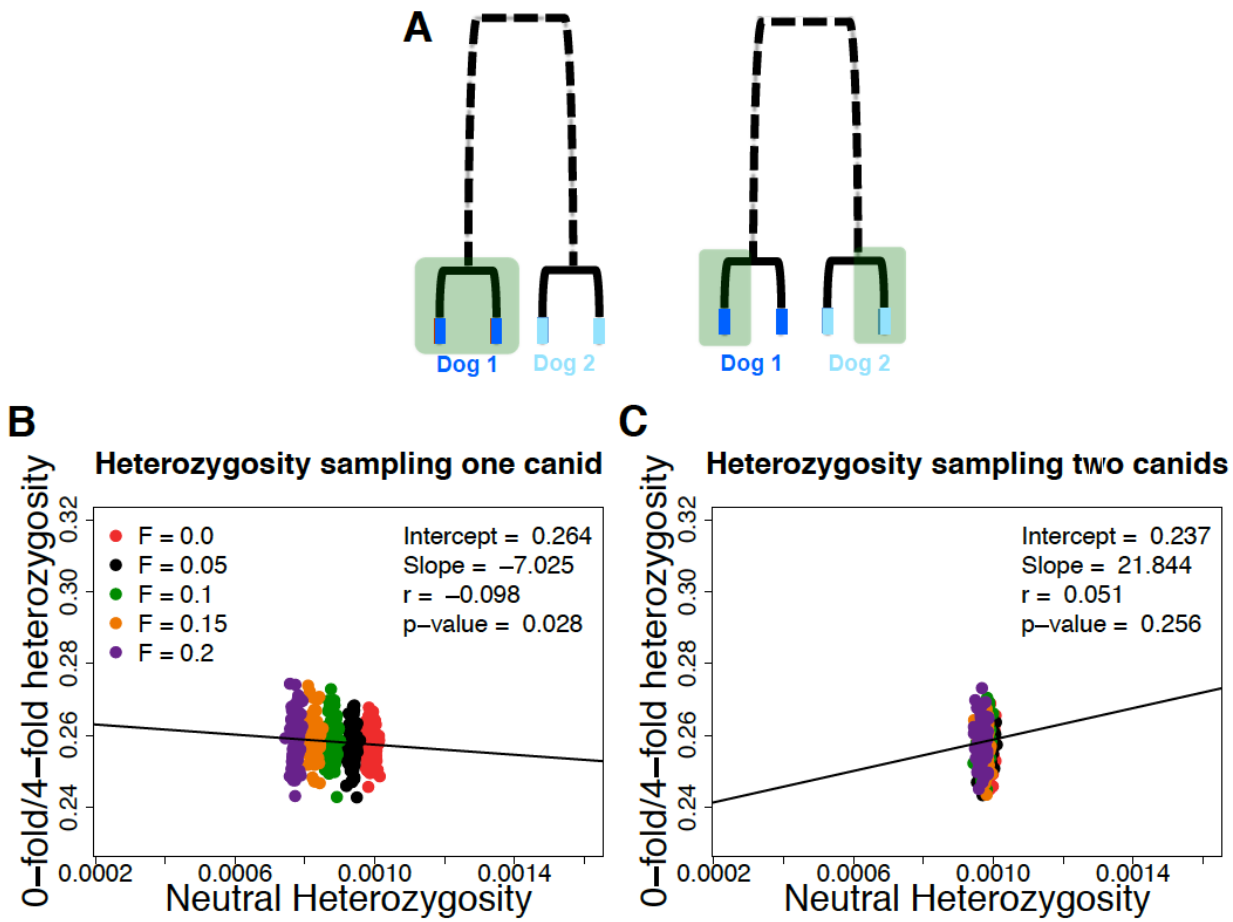

Fig. 3.8: Estimating heterozygosity using one chromosome from each of two individuals removes the effects of recent inbreeding.

(A) Inbreeding results in an increase in the probability that two chromosomes within an individual share a recent common ancestor with each other than with a chromosome in a different individual (i.e. chromosomes of the same color have a higher probability of coalescing with each other that with chromosomes of a different color). This will lead to

a reduction in heterozygosity (left panel). By computing heterozygosity from one read from each individual (i.e. from different colored chromosomes, right panel), we will remove this effect of inbreeding. **(B)** Forward simulations using the breed dog demographic model including population bottleneck along with 100 generations of inbreeding (model from Freedman et al., SI Appendix, Table S4) show a slight negative correlation between the ratio of 0–fold to 4-fold heterozygosity and neutral heterozygosity. This suggests recent inbreeding in certain dog breeds may slightly increase the 0-fold to 4-fold ratio. Lines denote the regression between the ratio of 0–fold to 4-fold heterozygosity vs. neutral heterozygosity. **(C)** Same simulations as in **(B)**, except here heterozygosity is computed using one chromosome from each of two dogs. Sampling from two dogs eliminates the reduction in heterozygosity due to recent inbreeding as well as the weak negative correlation seen in **(B)**.

Discussion

Our results show that the domestication process has dramatically affected patterns of deleterious variation across the dog genome. First, population history has had a genome-wide effect which increases the burden of deleterious variation in breed dogs as indicated by an elevated level of amino-acid changing variation relative to wolves where selection is more efficacious. Our demographic models suggest that repeated population bottlenecks and small effective population size have had a more profound effect on the accumulation of weakly deleterious variation than does recent inbreeding (i.e., mating between close relatives). Consequently, to minimize the accumulation of deleterious variation in the increasing number of species suffering from habitat loss and

fragmentation, conservation efforts should focus on maintaining sufficient population sizes in the wild and captivity, rather than focusing exclusively on inbreeding avoidance. Finally, our approach provides a comprehensive method for evaluating deleterious variation from genome data in the small isolated and threatened populations worldwide that can help prioritize their genetic management.

Inference of the distribution of fitness effects of segregating variants using haplotypic information

Introduction

The distribution of fitness effects is one of the most important determinants of Evolution (Eyre-Walker & Keightley 2007). Apart of its importance to the neutral theory and to determine current levels of genetic variation, it is also relevant to understand current phenotypic variation, since the distribution of fitness effects can influence the evolution of complex phenotypic traits (Lohmueller 2014b; Mancuso et al. 2015; Eyre-Walker 2010). Methods to estimate the distribution of selective coefficients using the site frequency spectrum and the PRF framework have been recently developed (Keightley & Eyre-Walker 2007; Adam R Boyko et al. 2008a; Loewe et al. 2006). These methods infer the distribution of fitness effects of new mutations and can only indirectly infer the distribution of fitness effects of observed mutations. Estimating the fitness effects of observed mutations is relevant to validate predictions of the nearly neutral theory, which predicts that most of the observed variation should be nearly neutral, and also is of interest in debates regarding the deleterious segregating variation observed in different populations.

Recent large sample size haplotype datasets have been recently generated and provide an important source of information to quantify the strength of selection acting on segregating variants. They are particularly important because they facilitate the finding of low-frequency variants, where we should found most of the deleterious genetic variation. An important source of information that we can extract from these datasets

comes from the linked variation around putatively functional low frequency variants. Using data from the Netherlands Genome Project, (Kiezun et al. 2013) found that, conditioning on the variants having a certain frequency in the population, nonsynonymous variants have a higher linkage with neighboring neutral variation compared to synonymous variants. This is in line with Takeo Maruyama's results showing that deleterious variants at a certain frequency have a younger age compared to neutral variants, implying that you should also expect to see less variation around deleterious variants.

Here we propose an approach to use patterns of linkage disequilibrium to infer the strength of natural selection acting on variants at a certain frequency in the population. Our approach uses information from the pairwise identity by state lengths L to infer the distribution of fitness effects acting on putatively functional variants at a certain frequency. This approach is the first one to estimate selection conditioning on the present-day frequency of the allele and can help us improve our understanding of how selection is impacting the vast amount of low-frequency variants present in a population. We discuss how our method can be used to distinguish between alleles under positive and negative selection in non-equilibrium demographic scenarios. Finally, we also present results to show how the distribution of fitness effects of alleles at a particular frequency can be applied to infer the distribution of fitness effects of new mutations.

Results

2.1 Inference of selection

Our question of interest is to infer what is the strength of selection acting on variants at an allele frequency *f* in the present. The allele frequency in the present can be defined either as the population allele frequency, defined as the actual frequency of the allele in the population, or the sample allele frequency, which is the allele frequency calculated after sampling a set of chromosomes from the population. We inferred the strength of natural selection using a linkage disequilibrium statistic that measures the length of pairwise haplotypic identity by state (*L)*. This statistic calculates the distance to the first difference between a pair of phased haplotypes that contain a particular allele *a*. *L* can be measured going both to the right and left side of this allele *a*. The values of *L* are defined with respect to a set of discrete non overlapping windows $\boldsymbol{W} = \{w_1, w_2, \ldots, w_n\}$ that extend from the physical position of the allele to the right or left side of *a*. If the first difference between the pair of phased haplotypes falls inside a window w_i, then the value of *L* is equal to w_i for this particular pair of haplotypes. Our inference method uses information of many values of $\boldsymbol{L} = \{L_1, L_2, L_3, \ldots L_n\}$ from pairs of haplotypes that contain alleles at that particular frequency *f*.

The likelihood of having a particular selection coefficient s conditioning on the allele frequency *f* using information from one length $L = w_i$ can be estimated as:

$$P(L = w_i|s, f) = \mathcal{L}(s, f) = \int P(L = w_i|H_i)P(H_i|s, f)dH$$

Where H_i is a particular allele frequency trajectory. We can compute $P(L = w_i|H_i)$ via Monte Carlo simulations done using *mssel* (Kindly provided by Richard Hudson), which assumes the structured coalescent model to simulate haplotypes containing a site whose frequency trajectory is determined by H_i. We used *mssel* to simulate many pairs of haplotypes (in the order of thousands or tens of thousands) given an allele frequency

trajectory H_i and we computed the L value for each pair of haplotypes. We can use that distribution of L values for a given allele frequency H_i to find the probability $P(L = w_i|H_i)$ that L falls in a certain window w_i. One important point about these Monte-Carlo simulations is that we can add any information we possess about the recombination rate variation present in a particular region to avoid biases in the values of L due to an incorrect modeling of the recombination rate in the simulations.

The likelihood $\mathcal{L}(s, f)$ is found by integrating over the space of allele frequency trajectories that end at a frequency f in the present and have a selection coefficient s. One possible way to perform that integration step is to perform many simulations under the assumptions of the Poisson Random Field framework (PRF) and utilize rejection sampling to only keep those trajectories that end at a frequency f in the present. Under the PRF model, the number of mutations that enter the population each generation have a Poisson distribution with mean $2N_imL = \Theta/2$, they are independent and the frequency of each mutation changes each generation following a Wright-Fisher model with selection. We could generate many allele frequency trajectories under this framework given a particular value of s and just keep those trajectories that end at a frequency of f. However, this is inefficient and computationally demanding, since a lot of allele frequency trajectories will not end at a frequency f in the present. And it is particularly more challenging if we wish to calculate $\mathcal{L}(s, f)$ for a grid of values of s. In the next two sections we show an alternative approach we took to perform an efficient integration over the space of allele frequency trajectories given s and f using importance sampling.

2.2 Importance sampling

Broadly, the idea behind importance sampling approaches is that we have a "target" distribution *f(x)* and we would like to take a large number of samples from that distribution. In this case, the "target" distribution f(x) is the set of trajectories that end at frequency *f* in the present. If we wanted to estimate expectations from that target distribution using a Monte Carlo method, we could take a set of samples (x_1, x_2,…,x_n) from the distribution f(x) and then use the following equation:

$$E[f(x)] = \sum_{i=1}^{n} f(x_i)$$

In our case, sampling from this "target" distribution is complicated because when we simulate allele frequency trajectories going forward in time, the vast majority of them do not end at a frequency f in the present. In the importance sampling framework, the idea is to choose a "proposal" distribution g(x) from which we can easily sample random values X. Then, making use of the equality $\frac{g(x)}{g(x)} = 1$, we can estimate the expected value of f(x) by using the following equation:

$$E[f(x)] = \sum_{i=1}^{n} \frac{g(x_i)}{g(x_i)} f(x_i) = \sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} g(x_i)$$

Then, we define a variable $\omega_i = \frac{f(x_i)}{g(x_i)}$ and:

$$E[f(x)] = \sum_{i=1}^{n} \omega_i g(x_i)$$

In cases where either f(x) or g(x) are missing a normalizing constant so that their area under the curve is equal to 1, we must employ a self-normalized importance sampling estimator (Robert & Casella 2010):

$$E[f(x)] = \frac{\sum_{i=1}^{n} \omega_i g(x_i)}{\sum_{i=1}^{n} g(x_i)}$$

The selection of the "proposal" distribution g(x) is critical in the importance sampling framework to accurately estimate the expected values of f(x). The goal overall is that the random variables simulated under g(x) could often be obtained by sampling under f(x). A useful metric in this regard is the effective sample size ESS, which is equal to (Robert & Casella 2010):

$$ESS = \frac{1}{\sum_{i=1}^{n} \underline{\omega_i}}$$

Where:

$$\underline{\omega_i} = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j}$$

The ESS tells you the sample size used in a Monte-Carlo evaluation of f(x) that is equivalent to the importance sampling approach estimate. The ESS takes values between 1 and n, where a higher value of the ESS indicates that more samples from g(x) are contributing to the estimate of the expected f(x). This is a necessary, but not sufficient, condition to obtain an accurate estimate of the expected value of f(x) when using an importance sampling approach. Values of ESS close to 1 indicate that few

replicates of g(x) are making a contribution of the expected value of f(x), and therefore, the estimated expected value of f(x) is probably not accurate.

2.3 Integration over the space of allele frequency trajectories using importance sampling

(TODO.- Need to add the beta-binomial step to go from sample frequency to the frequency in the first generation)

To find the likelihood $\mathcal{L}(s, f)$ over many different values of s, we performed an efficient integration over the space of allele frequency trajectories using the importance sampling approach developed by (Slatkin 2001) with a modification regarding the proposal distribution we use. Here, the "target" distribution $f(x) = P(H_i|s, f)$ are samples of allele frequency trajectories that end at a frequency of f and have a selection coefficient s.

Following Slatkin (2001), given a derived allele *a* we can define the fitness of the genotypes *AA*, *Aa* and *aa* as 1, 1+s and 1+2s, respectively. We can define the trajectory H_i of an allele as the number of copies of the allele a per generation since the allele appeared in the population. Therefore, $H_i = \{i_T, i_{T-1}, i_{T-2}, \dots, i_2, i_1, i_0\}$, where $i_T = 0$, $i_{T-1} = 1$ and $i_0 = f * N_0$. The effective population sizes at those times are $N = \{N_T, N_{T-1}, N_{T-2}, \dots, N_2, N_1, N_0\}$. The allele appears in generation T-1, where it has 1 copy in the population. In the present, the allele has a frequency f in the present and the number of copies is equal to f * N_0.

Under a Wright-Fisher model with selection, the probability of moving from i_t to i_{t-1} copies of the allele is equal to:

$$P(i_{t-1}|i_t) = p_{i_t,i_{t-1}} = \binom{2N_{t-1}}{i_{t-1}} x_t'^{i_{t-1}}(1-x_t')^{2N_{t-1}-i_{t-1}}$$

Where:

$$x_t' = x_t \frac{1+2sx_t+s(1-x_t)}{1+2sx_t^2+2sx_t(1-x_t)}$$

The frequency of the allele at generation t is $x_t = \frac{i_t}{2N_t}$. Finally, the probability of the whole allele frequency trajectory Hi is then equal to:

$$P(H_i|s,f) = f(x) = \prod_{t=T-2}^{0} p_{i_t,i_{t-1}}$$

As a "proposal" distribution g(x), we use a Wright-Fisher neutral model, with one modification, where the frequency of the allele is equal to f in the present. This allows us to make sure that the allele has the frequency f that we want in the present. Under this proposal distribution, we are going to move backwards in time assuming that the allele is neutral. Under this proposal distribution, if $i_{t-1} = 1$, then i_t can take any value from 0 to $2N_t$. If $i_{t-1} = 0$ or $2N_t$ then we stop the allele frequency trajectory. If i_{t-1} is bigger than 1 and smaller than $2N_t$, then i_t can take any value from 1 to $2N_t$. Those three rules are used together to make sure that each trajectory going forward in time goes from 0 to 1 copy of the allele always.

Under the proposal distribution we use, the transition probabilities of going from i_{t-1} alleles in generation t-1 to i_t alleles in generation i_t is:

$$P(i_t|i_{t-1}) = q_{i_{t-1},i_t} = \begin{cases} \dfrac{\binom{2N_t}{i_t} x_{t-1}^{i_t}(1-x_{t-1})^{2N_t-i_t}}{1-\binom{2N_t}{i_t} x_{t-1}^{0}(1-x_{t-1})^{2N_t}} & if\ i_{t-1} = (2,2N_t)\ and\ i_t > 0 \\ \binom{2N_t}{i_t} x_{t-1}^{i_t}(1-x_{t-1})^{2N_t-i_t} & if\ i_{t-1} = 1 \\ 0\ \ if\ 1)\ i_{t-1} = 0\ or\ 2N_t\ ;2)\ i_{t-1} = (2,2N_t)\ and\ i_t = 0 \end{cases}$$

Where $x_{t-1} = \frac{i_{t-1}}{2N_{t-1}}$. By generating an allele frequency trajectory with this proposal distribution, we can get the probability of any sample from this proposal distribution g(x):

$$g(x) = \prod_{t=0}^{T} q_{i_{t-1},i_t}$$

Now that we have defined how to sample allele frequency trajectories using our proposal distribution, we can compute the weight for every simulated allele frequency trajectory from g(x) as $\omega_i = \frac{f(x_i)}{g(x_i)}$. For some of the proposed trajectories under g(x), the trajectory will end up at a frequency of 1 going backwards into the past , instead of 0. The value of ω_i for those trajectories is 0.

The expected value that we wish to obtain with this problem is $P(L = w_i|s,f)$. Under the importance sampling framework, this would be equal to:

$$P(L = w_i|s,f) = \mathcal{L}(s,f) = \frac{\sum_{i=1}^{n} \omega_i P(L = w_i|H_i)}{\sum_{i=1}^{n} \omega_i}$$

Using this approach, we can estimate $P(L = w_i|s,f)$ for different values of s using the same set of allele frequency trajectories generated from our proposal distribution. This alleviates the need to simulate a different set of allele frequency trajectories for each

value of the selection coefficient s that we want to evaluate and follows the idea of a driving value (Fearnhead & Donnelly 2001). The only values that we need to change to evaluate $P(L = w_i|s, f)$ are the importance sampling weights ω_i, where we will change the value of $P(H_i|s, f) = f(x)$ depending on the value of the selection coefficient s evaluated.

Finally, given a set of values ***L***=*{L₁, L₂, L₃, … Lₙ}*, we can estimate the composite likelihood of having that set of ***L*** values as:

$$\mathcal{L}(s, f) = \prod_{i=1}^{n} P(L = w_i|s, f)$$

2.4 Estimation of selection in constant population sizes

We performed forward-in-time simulations under the Poisson Random Field (PRF) framework using PReFerSim to obtain a set of 10,000 allele frequency trajectories from alleles under a particular strength of selection that were sampled at a 1% frequency in the present in a sample of 4,000 chromosomes (see Methods). We ran many repetitions of PReFerSim using a value of Θ equal to 1,000 until we obtained 10,000 alleles frequency trajectories where $\hat{p}$=1% using 5 different values of selection (4Ns = 0, -50, -100, 50, 100). To do this, in each repetition we first ran PReFerSim to obtain a list of alleles where $\hat{p}$=1%, Then, we ran PReFerSim again using the same random seed and printing the allele frequency trajectory of the alleles were $\hat{p}$=1%.

Using the 10,000 recorded allele frequency trajectories for each selection value 4Ns, we calculated the mean allele frequency across many generations going backwards into the past to obtain an average frequency trajectory (Figure 4.1A). The average allele frequency trajectory for neutral alleles (4Ns = 0) is kept at higher values during more time going backwards in time compared to alleles under natural selection, indicating that 1% frequency neutral alleles can persist during a longer time compared to alleles under selection. Another interesting feature is that alleles under the same absolute strength of selection have a remarkably similar average allele frequency trajectory, regardless of whether the allele is under positive or negative selection. However, there are slight but significant differences in the average allele frequency trajectory between alleles under positive and negative selection that share the same absolute value of 4Ns. Across all generations, the mean allele frequency has significant differences in 7 out of the 10 most recent generations between alleles with a 4Ns = 50 versus alleles with 4Ns = -50 (two-tailed Welch's t-test, p-value < 0.05) with a mean difference in frequency of 1.5e-4 when there are significant differences between the mean allele frequencies. When we compare the average allele frequency trajectory of alleles with a 4Ns = -100 against alleles with a 4Ns = 100, we find significant differences only in 126 out of the 150 most recent generations (two-tailed Welch's t-test, p-value < 0.05) with a mean difference of 2e-4 when the differences in frequency are significant. The difference in allele frequency in those recent generations is due to the fact that the actual population allele frequency for alleles under negative selection in the present is lower in alleles sampled at a 1% frequency in the population as compared to alleles under positive selection. This implies that, in this demographic scenario, alleles under negative selection tend to

actually have a lower frequency in the population when we sample them at a frequency of 1% compared to alleles under positive selection that share the same absolute value of 4Ns. On the other hand, when we sample trajectories that end at a population allele frequency of 1%, in contrast to sampling trajectories that end at a sample allele frequency of 1%, we see no significant differences at any time in the average frequency trajectories of alleles under positive and negative selection that have the same absolute value of 4Ns (Figure 5.2B).

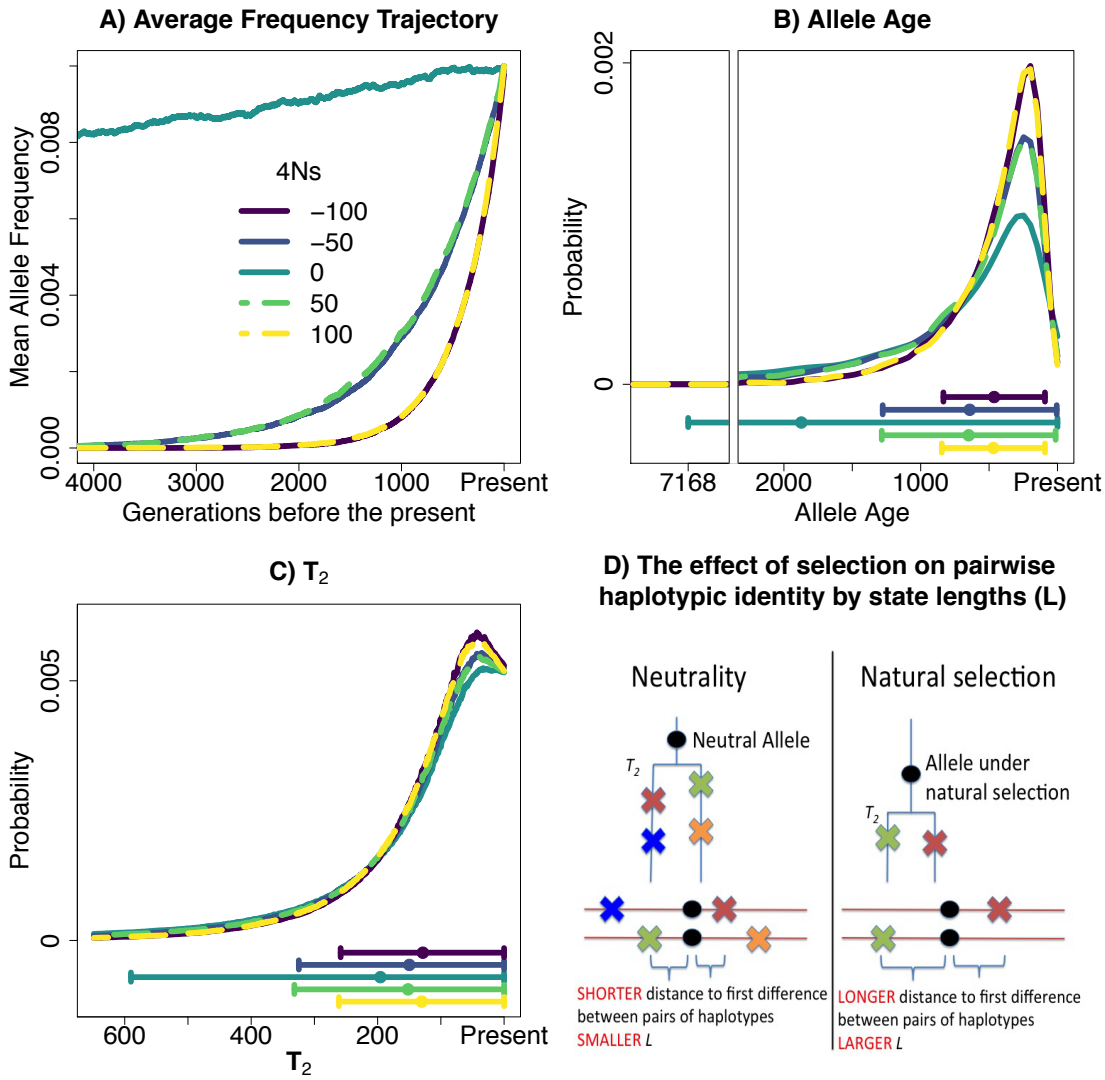

Figure 4.1.- **Properties of alleles sampled at a 1% frequency under different strengths of natural selection in a demographic scenario with a constant population size (N = 10,000).** Using forward-in-time simulations under the PRF model, we obtained 10,000 frequency trajectories for 1% frequency alleles under different strengths of selection. Using those frequency trajectories, we calculated: A) The mean allele frequency at different times in the past, in units of generations, to obtain an average frequency trajectory; B) The probability distribution of allele ages and C) The probability distribution of pairwise coalescent times T_2. Below B) and C), we show a dot

with two whiskers extending at both sides of the dot. The dot represents the mean value of the distribution and the two whiskers extend one s.d. below or above the mean. The whisker that extends one s.d. below the mean is constrained to extend until max(mean – s.d. ,0). In this demographic scenario, the alleles under a higher absolute strength of selection have younger ages and younger T_2 on average. The fact that alleles under higher strengths of selection have younger average T_2 values implies that those alleles tend to have larger L values as shown in D).

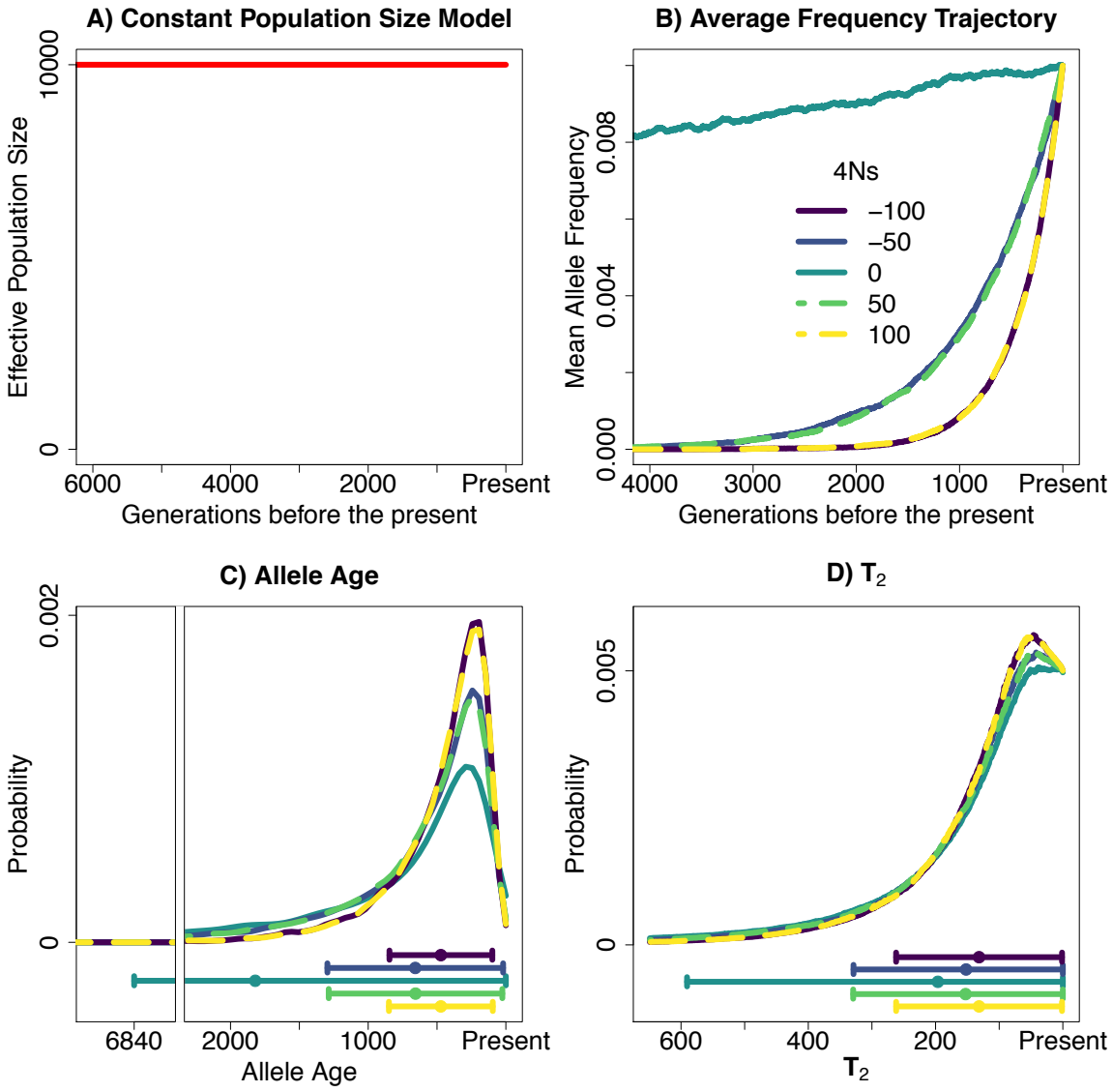

Figure 4.2.- **Properties of alleles sampled at a 1% population allele frequency under different strengths of natural selection in a constant population size scenario.** A) Population model analyzed. B) Mean allele frequency at different times in the past, in units of generations. C) Probability distribution of allele ages and D) Probability distribution of pairwise coalescent times T_2. The dot and whiskers below C) and D) represent the mean value of the distribution and the two whiskers extend at both sides of the mean until max(mean +- s.d. ,0).

We estimated the distribution of allele ages, defined as the number of generations ago when the allele emerged in the population, based on the 10,000 recorded allele frequency trajectories for each 4Ns value. We found that the distribution of ages on alleles under higher absolute values of 4Ns has a younger mean age and has a smaller standard deviation. The mean value of the ages are also is in close agreement with Maruyama's theoretical results, although they don't match exactly (Table 4.1). It is expected that the estimates from Maruyama's results and our simulations do not exactly match, since Maruyama's results assume that the population is in the diffusion limit, where the population size tends to infinity and 4Ns tends to a fixed value.

Table 4.1.- Comparison of estimates of the mean allele age of a 1% frequency variant in a demographic scenario of a constant population size (N = 10,000). The theoretical estimates were obtained by Maruyama (1974) and assume that the populations are in the diffusion limit, where N tends to infinity and the value of 4Ns tends to a fixed constant. We report two estimates of allele age from forward-in-time simulations using 10,000 forward-in-time allele frequency trajectories. One of the estimates uses alleles that were sampled at a 1% frequency based on a sample of 4,000 chromosomes. The other estimate is based on alleles that have a 1% population allele frequency. We also report the standard deviation of the allele ages, shown inside parenthesis, for the forward-in-time simulations.

4Ns	0	-50	50	-100	100
Maruyama's theoretical estimates	1840	632	632	452	452
Forward-in-time simulation estimates (based on the sample allele frequency)	1872.69 (5295.08)	641.49 (636.20)	649.00 (635.53)	462.68 (371.32)	468.42 (377.05)
Forward-in-	1820.33	658.03	656.10	471.87	472.17

time simulation estimates (based on the population allele frequency)	(5019.22)	(638.23)	(630.65)	(373.64)	(377.18)

We discretized and compressed each of the allele frequency trajectories according to a set of allele frequency boundaries to reduce the computing time needed to simulate haplotypes under the structured coalescent model with *mssel*. We then used each compressed allele frequency trajectory to estimate the distribution of pairwise coalescent times T_2 between a pair of haplotypes containing the allele changing in frequency. For each compressed allele frequency trajectory, we estimated the probability of coalescing at a time *t* as:

$$P(T_2 = t) = \left[\prod_{n=1}^{t-1}\left(1 - \frac{1}{Na_n}\right)\right]\frac{1}{Na_t}$$

Where Na_x denotes the number of individuals that have the derived allele a at generation x. Additionally, due to the way we compressed the allele frequency trajectories, where the allele frequency at the time that the allele emerges is not equal to 1/2N, the probability of T_2 in the generation *e* where the allele appears is equal to $1 - \sum_{e=1}^{e-1} P(T_2 = t)$. We averaged the probabilities of T_2 over all the simulated allele frequency trajectories given a particular value of *4Ns* to obtain the distribution of T_2

given a value of *4Ns*. Additionally, we also confirmed that this distribution of T_2 given T2 was correct by using an alternative method that employed simulations of 1,000 T_2 values in each allele frequency trajectory and then averaging over all the frequency trajectories obtained with a value of *4Ns* to obtain the distribution of T_2 given *4Ns* (Figure SX TO-DO).

The distribution of pairwise coalescent times T_2 across different values of *4Ns* shows that alleles under higher absolute values of *4Ns* have a more recent average value of T_2, and their distribution of T_2 has a smaller standard deviation (Figure 4.1C). As shown in Figure 4.1D, this implies that alleles under stronger absolute selection coefficients will have younger average ages and T_2. The fact that T_2 is younger in alleles under stronger selection coefficients, will lead to fewer mutations between haplotypes sharing the allele and higher average values of *L*, the statistic that we will employ to estimate the strength of selection acting in alleles sampled at a particular frequency in the population.

To infer selection, we took pairs of haplotypes that contained an allele at a 1% frequency in the population and then we divided the physical distance (in bp) surrounding the allele into 5 consecutive windows of 50 kb plus one extra window containing any distance bigger than 250 kb (Figure 4.3A). We resampled from the 10,000 simulated allele frequency trajectories given each *4Ns* value to perform 1 million simulations of pairs of haplotypes containing alleles at a 1% frequency under different strengths of selection. We estimated *L* in each of those pairs of haplotypes to create a probability distribution function of *L* given a set of 1% frequency alleles under different

strengths of selection. We found that alleles under the same absolute strength of selection have almost identical distributions of L (Figure 4.3B), a result in line with the similarities seen in the distribution of T_2 values for alleles under the same absolute strength of natural selection (Figure 4.1C). Since our inference method relies on differences in the distribution of *L* between alleles under different strengths of selection, this indicates that we should not have power to distinguish between alleles under the same absolute strength of selection, something that we corroborated in Figure 4.3C. We see that our method can accurately predict when an allele is neutral (4Ns = 0). However, in alleles under a strength of selection equal to -50 or 50, we see that the estimated values of selection tend to be similarly distributed around values of -50 or 50. The same trend is seen for alleles under a strength of selection equal to -100 or 100 (Figure 4.3C). When we display the estimated strength of selection in terms of absolute values of 4Ns, we see that our method produces an accurate estimate of the absolute strength of selection (Figure 4.3D). This indicates that our method can provide reasonable estimates of the absolute strength of natural selection in this demographic model, but cannot differentiate well between alleles under negative or positive selection. In constant population sizes, it is likely that no method would be able to distinguish between positively and negatively selected alleles at a frequency *f* due to the very slight differences in their mean allele frequency trajectories.

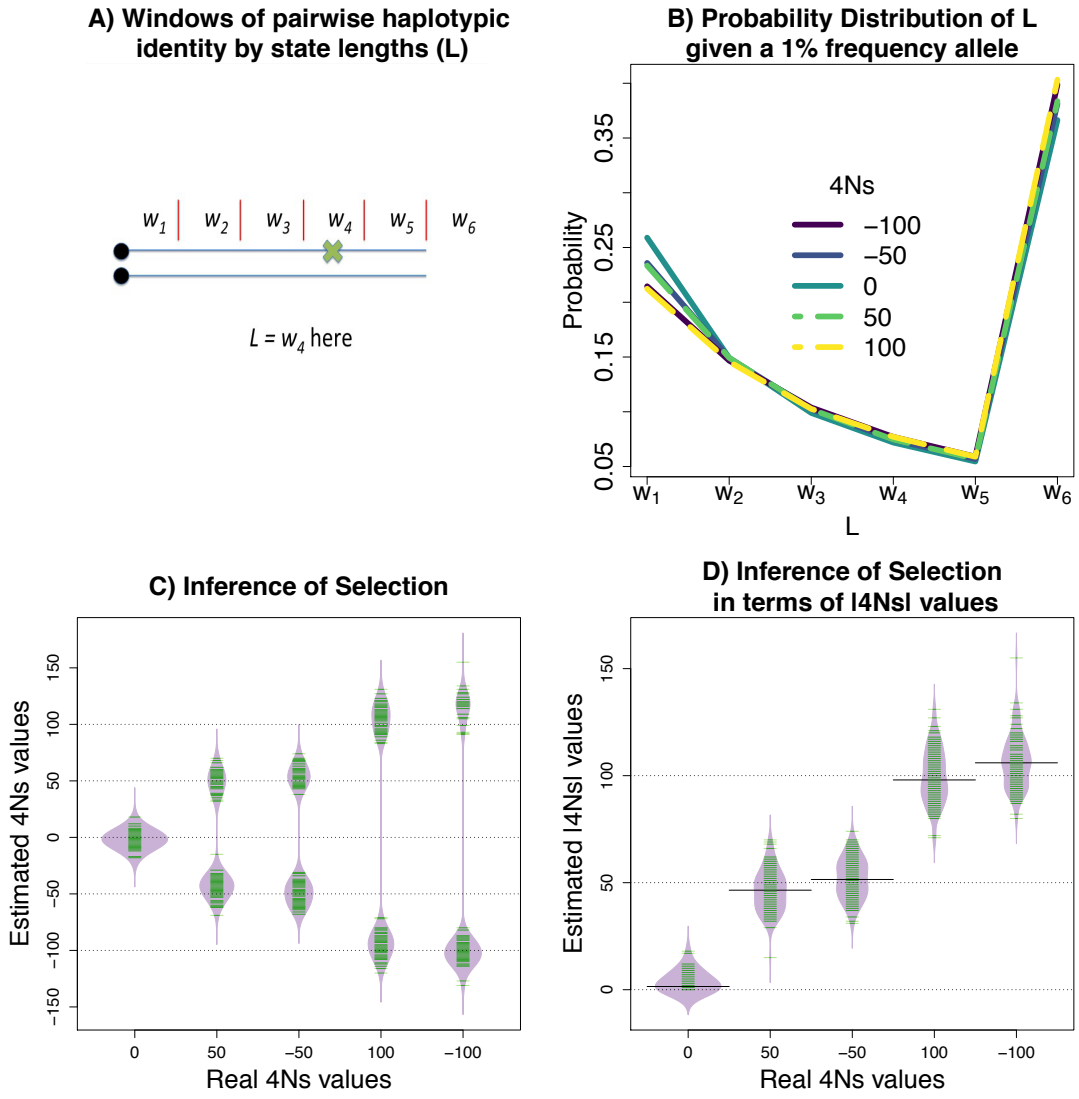

Figure 4.3.- **Estimation of the strength of natural selection in a constant population size model.** A) The physical distance in bp near an allele is divided into 5 non-overlapping equidistant windows of 50kb, with an extra window w6 indicating that there are no differences in the 250 kb next to the allele. B) The probability distribution of L given a 1% frequency allele and different values of 4Ns. C) Estimated selection values. D) Estimated selection values when we plot the estimated selection values as absolute 4Ns values.

2.5 Estimation of selection in non-equilibrium demographic scenarios

We used forward-in-time simulations to analyze the changes in frequency of alleles that have a 1% frequency in non-equilibrium demographic scenarios. First, we analyzed the shape of the average allele frequency trajectory in a population expansion scenario (Figure 4.4A) for alleles with different 4Ns values. In stark contrast with a constant population size scenario, we found that alleles under positive or negative selection show very distinct allele frequency trajectories. Alleles under positive selection keep increasing in frequency after the population expansion. On the other hand, alleles under negative selection increase in frequency before the expansion. After the expansion, the efficacy of selection increases due to the higher effective population size. This effect provokes that deleterious alleles decrease in frequency in the population until they reach an allele frequency value close to 1% (Figure 3B). The ages of alleles under the strongest absolute values of selection tend to be younger, although alleles under positive and negative selection that share the same |4Ns| value do not have the same mean allele age value and have a different standard deviation (Figure 3C). Another important distinction between the population expansion model and the constant population size model lies on the distribution of T_2, which is sensitive to differences in the sign of *4Ns* under our inspected population expansion model. In the values of *4Ns* that we inspected, we found that each of those values had a clearly different distribution of T_2. Alleles under the stronger positive selection had, on average, younger T_2 values. This is in agreement with their average allele frequency trajectory, which shows the

most rapid decrease in frequency. When we contrasted the T_2 distribution of the negatively selected alleles inspected (4Ns = -50, -100), we saw that their mean T2 value did not differ much, and their biggest difference relied on a slightly smaller standard deviation in the most deleterious allele (Figure 4.4D).

We used our method to infer the strength of selection in this demographic scenario and found that it provided accurate estimates of the strength of selection (Figure 4.5). In line with the differences found in the distribution of T_2 values between positively and negatively selected alleles sharing the same |4Ns| value, we found that our method was capable of inferring the strength of selection regardless the sign of the selection coefficient *s*. Therefore, under this non-equilibrium model it is possible to distinguish between alleles under positive and negative selection using haplotypic patterns. This does not mean we can differentiate between positive and negative selection in all non-equilibrium models. This will be dependent on the parameters of the non-equilibrium demography being studied. As an example, we show how in an ancient bottleneck there are not significant differences in the distribution of T_2 between alleles that have the same absolute strength of selection (Figure 4.6), pointing that we would not be able to differentiate between under positive or negative selection under this demographic model.

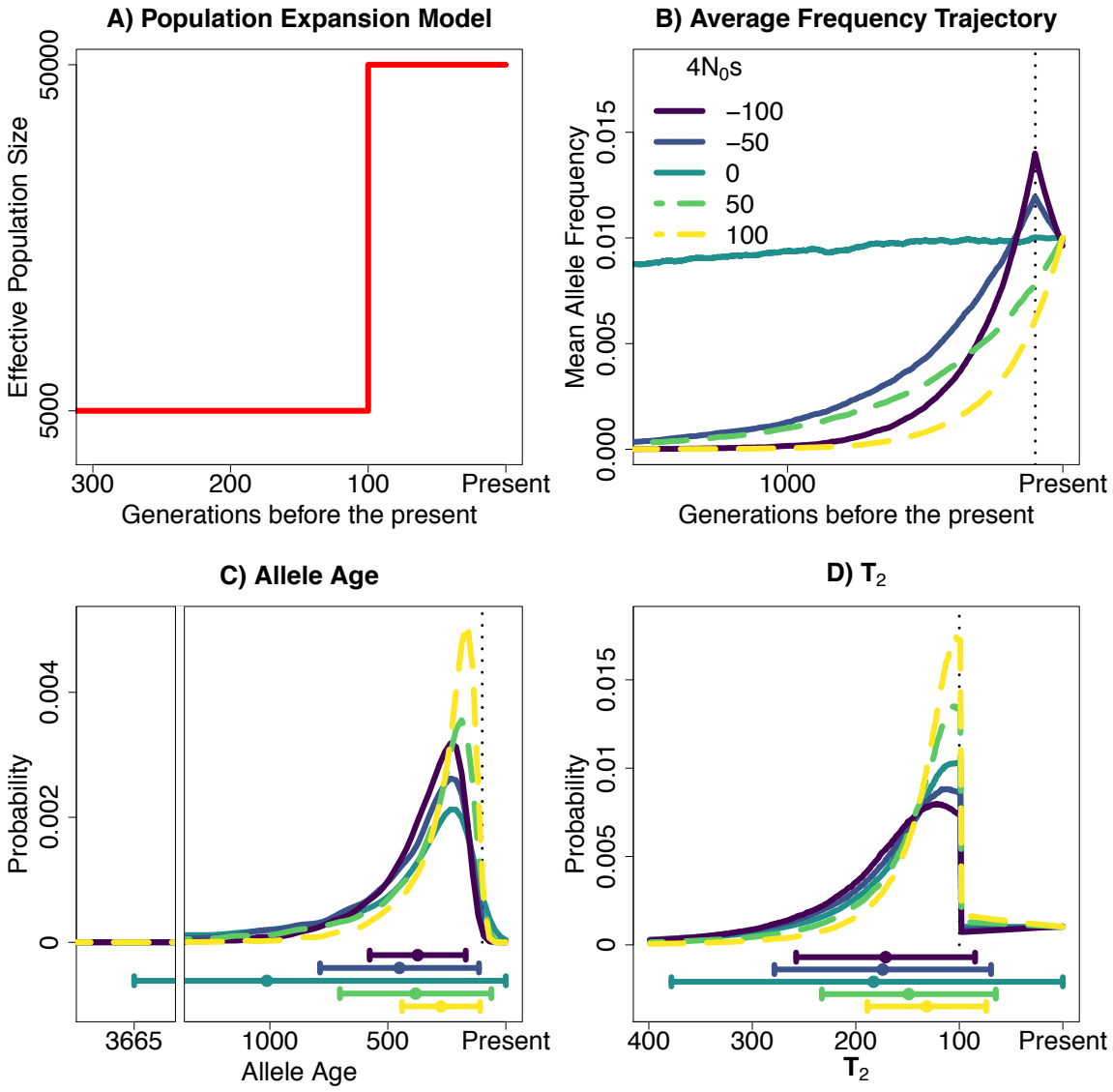

Figure 4.4.- **Properties of alleles sampled at a 1% frequency under different strengths of natural selection in a population expansion scenario.** A) Population expansion model analyzed. B) Mean allele frequency at different times in the past, in units of generations. Note that alleles under the same absolute strength of selection (4Ns) have a very different average allele frequency trajectory, in contrast to the constant population size scenario; C) Probability distribution of allele ages and D) Probability distribution of pairwise coalescent times T_2. The dot and whiskers below C)

and D) represent the mean value of the distribution and the two whiskers extend at both sides of the mean until max(mean +- s.d. ,0).

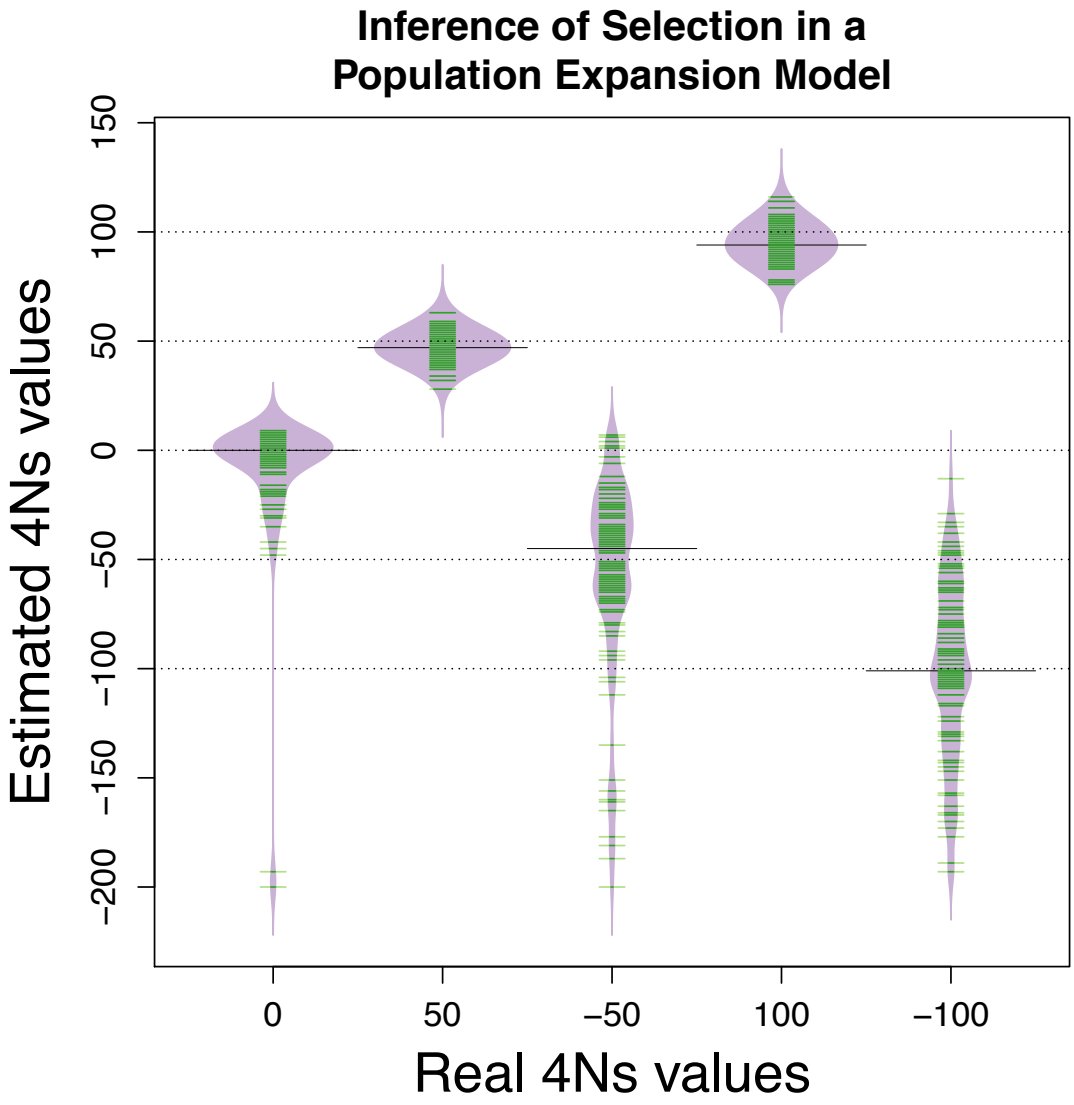

Figure 4.5.- **Estimation of the strength of natural selection in a population expansion model.**

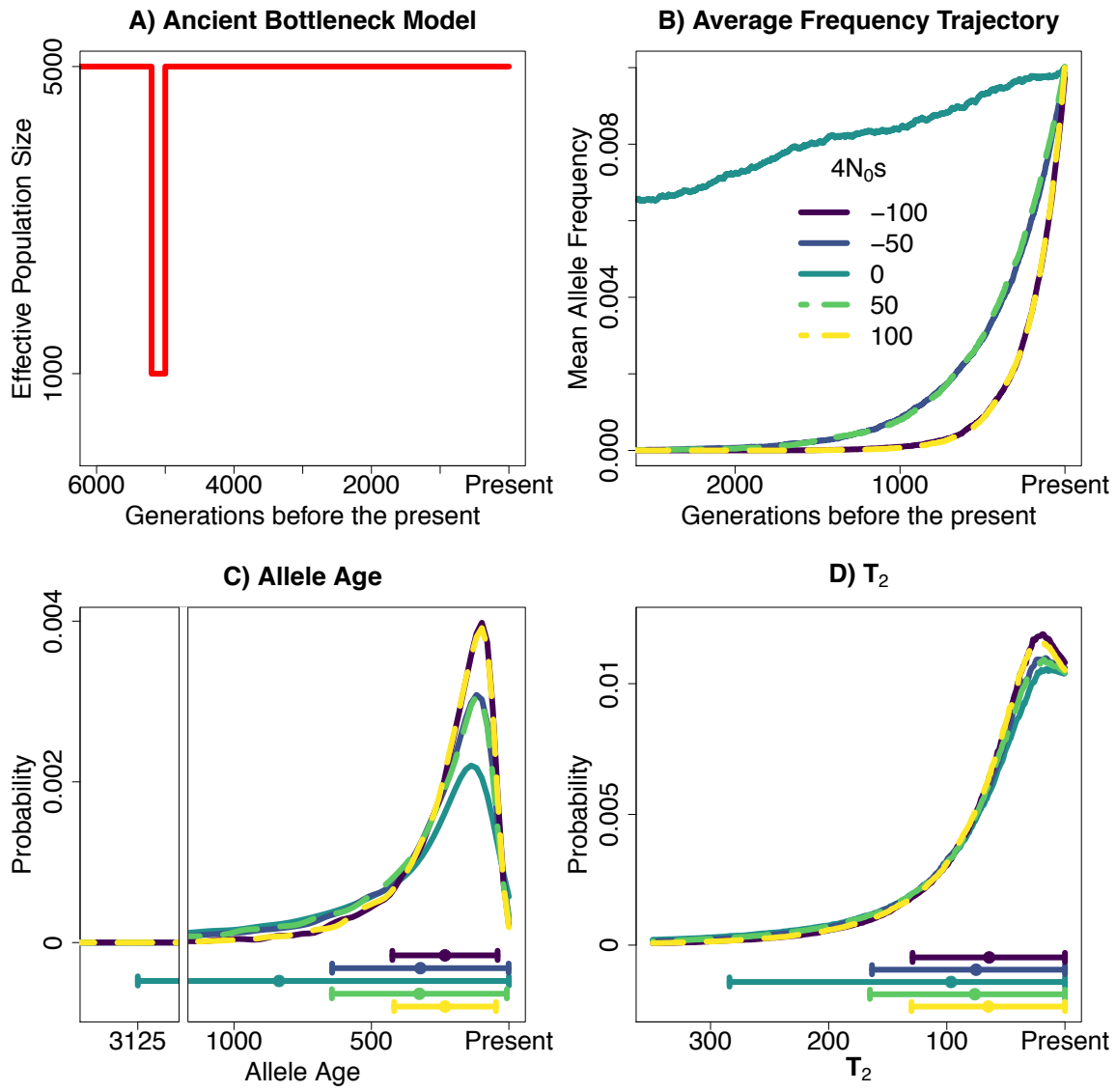

Figure 4.6.- **Properties of alleles sampled at a 1% frequency under different strengths of natural selection in an scenario with a bottleneck that took place 5,000 generations ago.** A) Population expansion model analyzed. B) Mean allele frequency at different times in the past, in units of generations. Note that alleles under the same absolute strength of selection (4Ns) have a very different average allele frequency trajectory, in contrast to the constant population size scenario; C) Probability distribution of allele ages and D) Probability distribution of pairwise coalescent times T_2.

The dot and whiskers below C) and D) represent the mean value of the distribution and the two whiskers extend at both sides of the mean until max(mean +- s.d. ,0).

2.6 Inference of the distribution of fitness effects of variants at a particular frequency

All of the variants present at a certain frequency in the population are likely to have different selection coefficients associated to them. We can also apply our composite likelihood framework to find the distribution of selection coefficients that better explain the distribution of *L* values seen in variants at a particular frequency. The probability of having a certain distribution of identity by state lengths L given a demographic scenario *D*, a set of variants at a frequency *f* and a distribution of selective coefficients $g(4Ns)$ is equal to:

$$P(L = w_i | g(4Ns), D, f) = \int_{\gamma=-\infty}^{\infty} P(L|4Ns, f, D) P(4Ns|g(4Ns)) d\gamma$$

Using the past equation, we tested if the distribution of haplotype lengths L can be used to estimate the parameters that define the distribution of fitness effects of variants at a particular frequency. We used distributions of L values obtained via simulations under the constant population size and population expansion demographic model from the past sections under two distributions of fitness effect of new mutations estimated in different species: one from humans (Adam R Boyko et al. 2008b) and another one from mice (Daniel L. Halligan et al. 2013).

We found that the estimated parameters of the DFE of 1% frequency variants do not cluster exactly around the actual parameters of the DFE of 1% frequency variants (Figure 4.7). However, they are responsive to changes in the DFE of new mutations employed to simulate the variants. This can be better seen by contrasting Figure 4.7A and Figure 4.7B, where the product of the estimated parameters of scale and shape in the Human DFE (Figure 4.7A) tends to be higher than that of the Mouse (Figure 4.7B). This is relevant because the result of that product gives the mean 2Ns values from the distribution. Therefore, we tend to have small variation in the estimated 4Ns values estimated with our approach in constant population sizes, as seen in Figure 4.8A and Figure 4.8B.

The demographic scenario used affects the estimates of the DFE. Under a constant population size, the estimates of the DFE follow a decay that resembles a curve. Note that this curve decay causes the product of the scale and shape parameters to keep relatively similar values. Under a population expansion model, the estimates of the shape and scale show a wider variation (Figure 4.7C and Figure 4.7D). This is in line with our results shown in Figure 4.3D and 4.1D, where we show that there are not as much differences in the pairwise coalescent time distribution between variants with different negative selection coefficients in a population expansion scenario as compared with a constant population size scenario. Due to the bigger variation in the estimates of the parameters that define the DFE of variants at a 1% frequency, we also see a bigger variation in the mean 4Ns values estimated in a population expansion as compared to a constant population size (Figure 4.7).

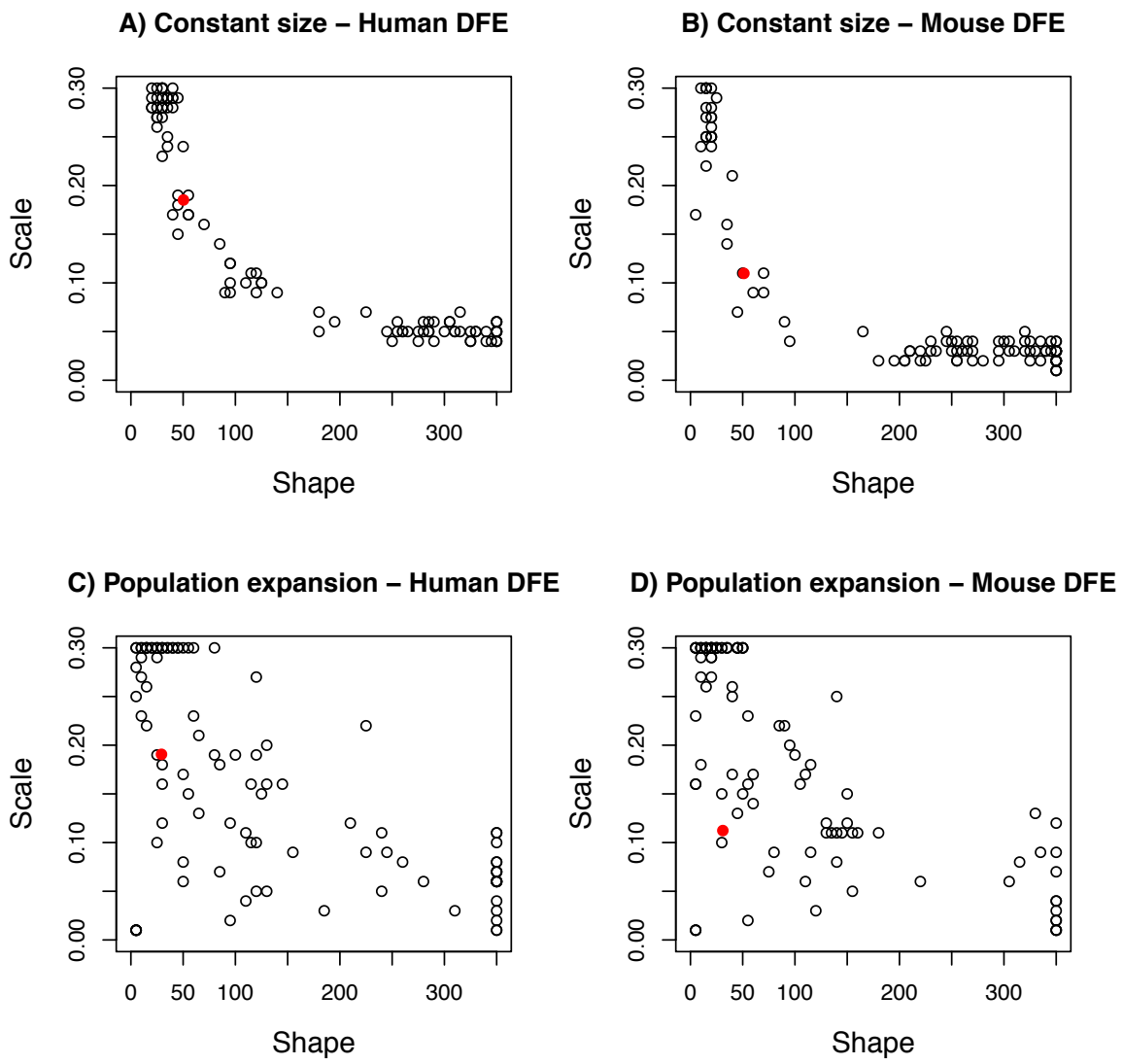

Figure 4.7.- **Estimates of the distribution of fitness effect of variants at a 1% frequency**. We tested if our method was capable of estimating the parameters of the DFE of variants at a particular frequency in two demographic models and two DFE's. The red dot indicates the scale and shape parameter from a gamma distribution that give a better adjustment to the selective coefficient of 50,000 variants at a 1% frequency from a particular DFE and demographic model. Each other dot contains an estimated value of the scale and shape parameter inferred from 100,000 L values.

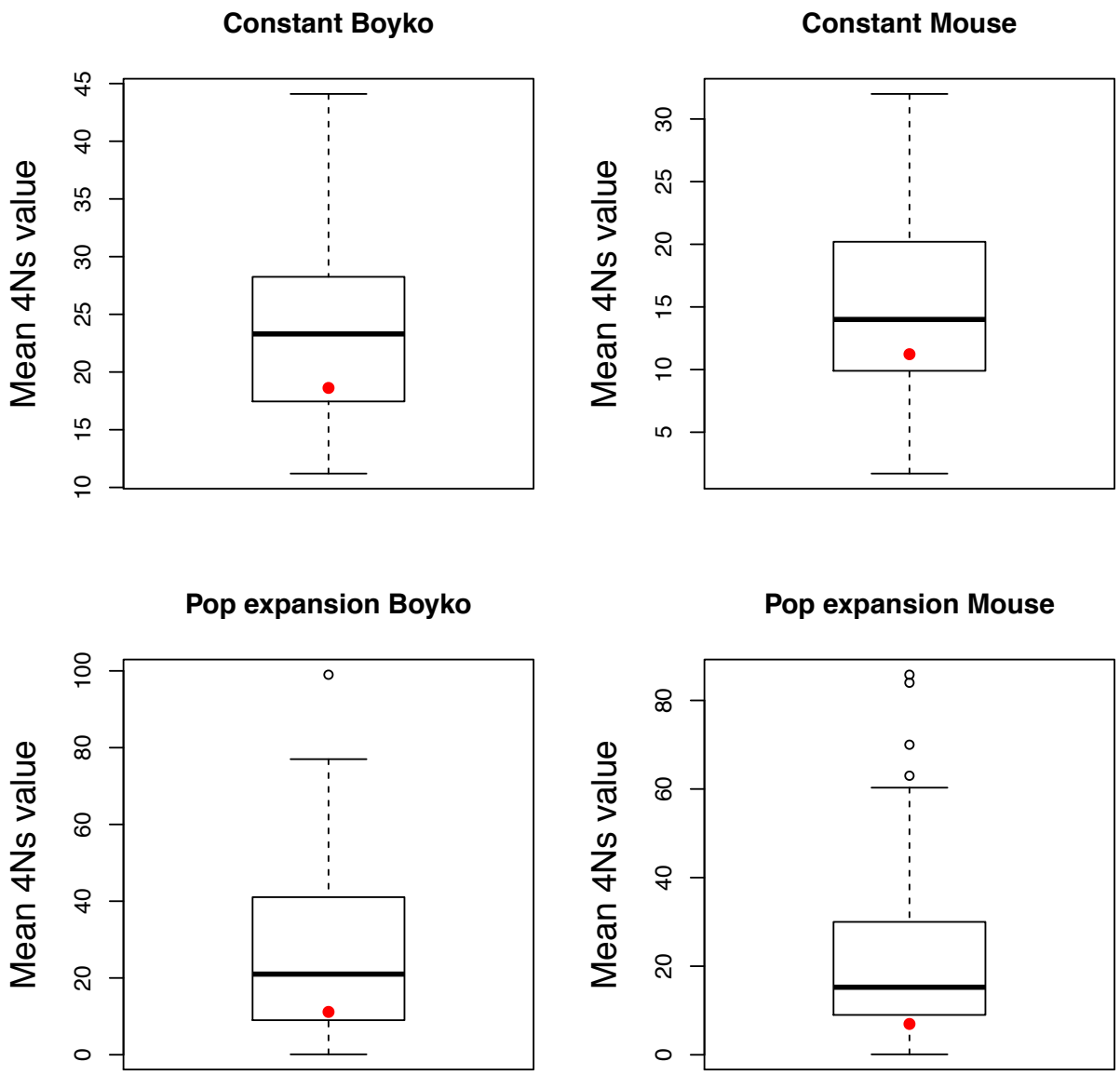

Figure 4.8.- **Estimation of the mean 4Ns values in 1% frequency variants.** The boxplots show the distribution of the estimated mean 4Ns values based on the estimates of the DFE parameters shown in Figure 5. The red dots show the actual 4Ns value from 50,000 1% frequency variants from each particular DFE and demographic model employed.

2.7 Connecting the distribution of fitness effects of variants at a particular frequency with the distribution of fitness effects of new variants

So far we have only been concerned about the distribution of fitness effects of variants at a particular frequency in the population. In this section, we want to show how we can relate the distribution of fitness effects of variants at a particular frequency in the population to the distribution of fitness effects of new variants. To do this, we define that the values of 2Ns are contained in an interval [S_1,S_2]. Then, we can make use of Bayes Theorem and define that:

$$P(B|C,D) = \frac{P(B|A,C,D)\,P(A|C,D)}{P(A|B,C,D)}$$

Where the events defined in that formula are:

A.- The allele has an x% frequency

B.- Allele has a selection coefficient 2Ns that falls in the interval [S_1,S_2].

C.- Distribution of fitness effects of new mutations.

D.- Demographic scenario

The information contained across all non-overlapping intervals of [S_1,S_2] from 0 to infinity of P (B |C, D) defines the distribution of fitness effects of new mutations. Since this information is independent of the demographic scenario, then P(B|C,D) = P(B|C). This is the information we would like to infer.

P (B|A,C,D) is the distribution of fitness effects of variants at a particular frequency given a certain distribution of fitness effects of new mutations and a particular demographic scenario. The method proposed in the past section of this paper estimates this probability.

P (A|C,D) is the probability that a variant has a certain frequency given a distribution of fitness effects of new mutations and a demographic scenario. This can be easily counted both in data and in simulations just by looking at the proportion of variants at a certain frequency.

P (A|B,C,D) is the probability that an allele has a certain frequency given that the allele has a selection coefficient contained in a certain interval, that there is a certain distribution of fitness effects of new mutations and a certain demographic scenario. To simplify the calculation of this probability, we can make the assumption that all the mutations in the interval $[S_1,S_2]$ have very similar selection coefficients, which is more likely to be true when the interval is not very big. Under that assumption P(A|B, C, D) = P(A| B, D). This probability can be found via forward-in-time simulations, where we simulate variants that have a selection coefficient contained in a certain interval $[S_1,S_2]$ in a particular demographic scenario. Then, the proportion of variants in that simulation that have a x% frequency is equal to P(A| B, C, D).

We applied the past equation to obtain an estimate of the distribution of fitness effects of new mutations in a population expansion scenario using the distribution of fitness effects of new mutations inferred in humans. Given that we have an estimate of P (B|A,C,D) by using the L distribution, that we can calculate P (A|C,D) from the data and

that we estimated P(A| B, D) via simulations, we can recover an estimate of the distribution of fitness effects of new mutations via the distribution P(B|C,D) (Figure 4.9).

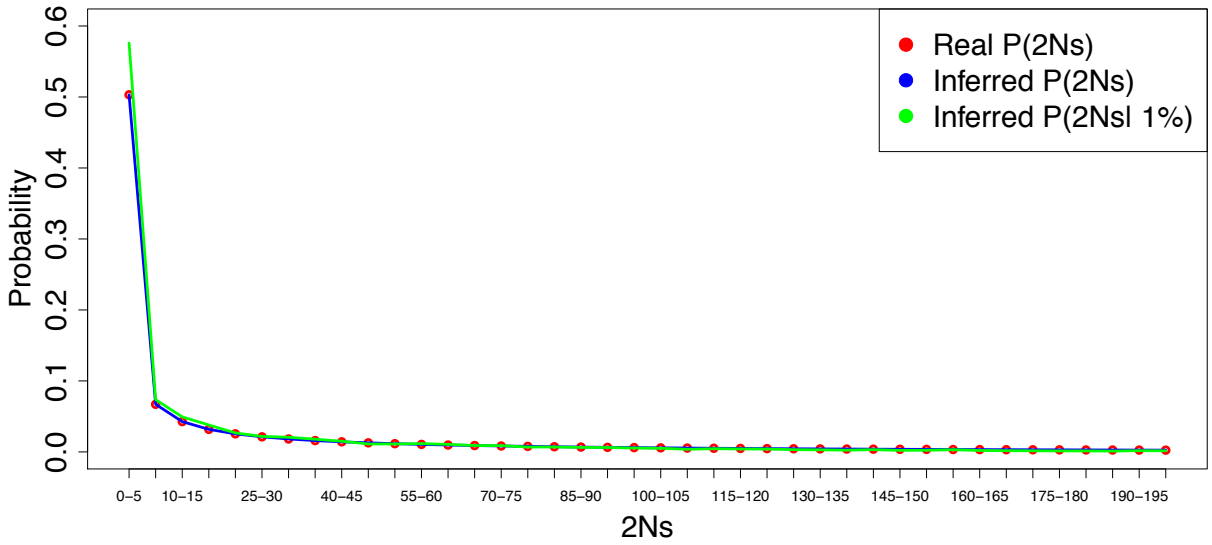

Figure 4.9**.- Inferring the distribution of fitness effects of new mutations from the distribution of fitness effects of variants at a certain frequency.** We use the Boyko distribution of fitness effects and a population expansion demographic scenario. Real P(2Ns) refers to the distribution of fitness effects of new mutations. Inferred P(2Ns | 1%) is one estimate of the distribution of fitness effects using the L distribution; this is P(B|A,C,D) from equation X. Inferred P(2Ns) is an estimate of the distribution of fitness effects using P(2Ns | 1%) and equation X.

2.8 Inference of the distribution of fitness effects of 1% frequency variants in the UK10K dataset

We used the UK10K dataset to infer the distribution of fitness effects of 1% frequency variants at nonsynonymous sites. First, we combined information from the ALSPAC and TwinsUK cohorts to get a total of 3,714 individuals. We estimated the past demographic history of the UK10K data using information from the site frequency spectrum at synonymous sites. To do this, we employed the program fastNeutrino (Bhaskar et al. 2015), where we specified that the demographic history for the ancient population history that took place more than 920 generations ago is equal to the demographic model from (Gravel et al. 2011a). For the most recent time, we specified that there are three different epochs of population size change with the timings of the population size changes left as free parameters. We used fastNeutrino to inferred 6 demographic parameters, the population sizes for the three most recent epochs and the three most recent times of population size change. The site frequency spectrum given the demographic history inferred by fastNeutrino and the site frequency spectrum in the data follow almost identical distributions (Figure 4.10), with the KL divergence between those two distributions being equal to 3.466e-05.

We investigated if it was possible to infer the strength of natural selection based on the inferred demographic history (Figure 4.11A). We found that under the inferred demographic model, the frequency trajectories of 1% frequency alleles were dissimilar between alleles under different strengths of natural selection (Figure 4.11B). In a similar way to the population expansion model (Figure 4.4B), we found that deleterious alleles tended to more quickly decrease in frequency when population sizes were larger. We found that both the distribution of allele ages (Figure 4.11C) and pairwise coalescent times T_2 (Figure 4.11D) were different for alleles under the same absolute strength of

selection. The fact that the distribution of pairwise coalescent times is different for alleles under the same |4Ns| value indicates that under this demographic model it is possible to distinguish between positive and negative selection using haplotype signatures. We performed simulations under 5 different strengths of selection and found that our method gave accurate estimates of the strength of natural selection (Figure 4.12).

We applied our method to the 1% frequency nonsynonymous variants from the UK10K dataset. We inferred that nonsynonymous variants have a selection coefficient 4Ns = 0. (MORE WORK TO COME IN THE REAL DATA).

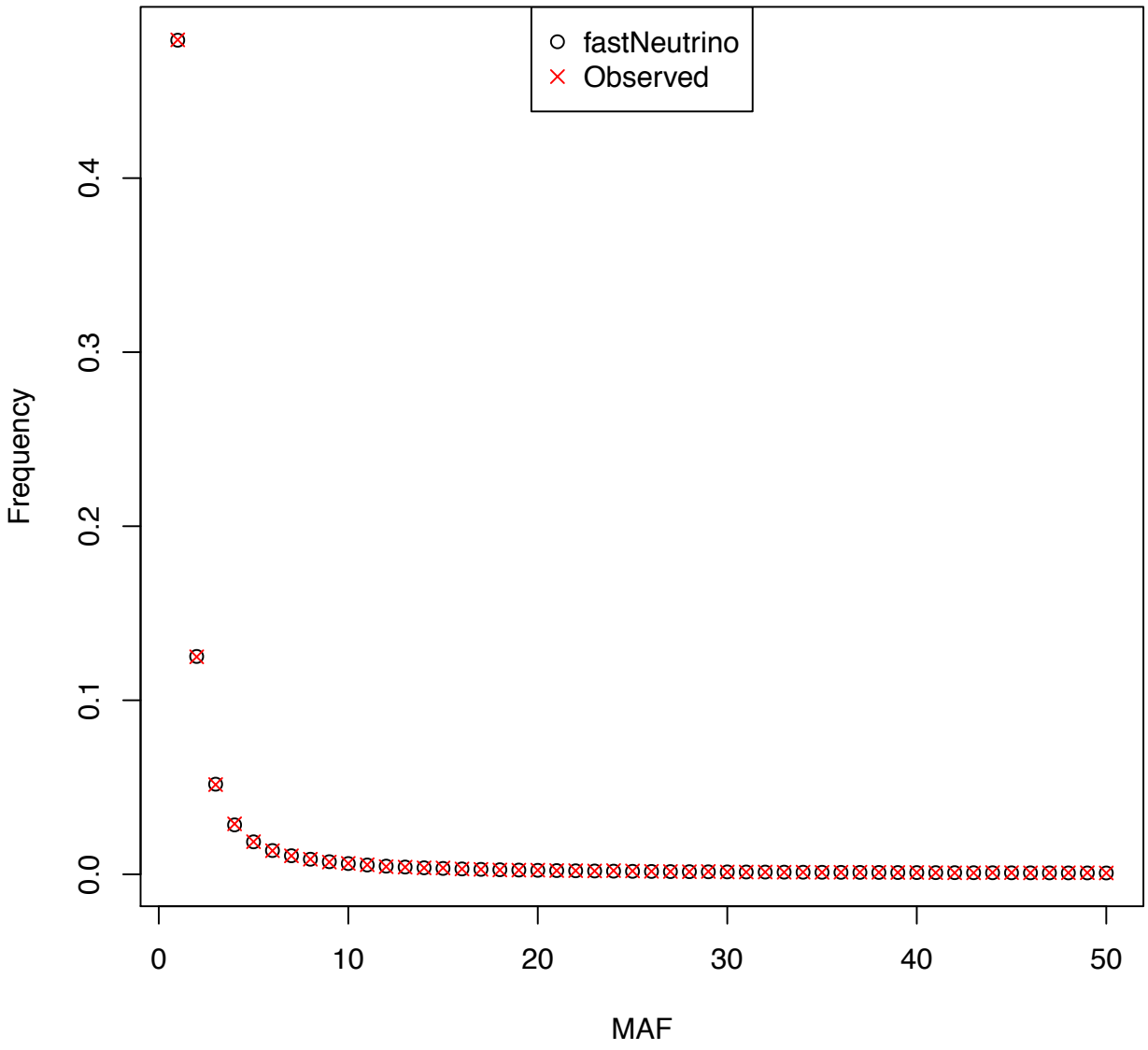

Figure 4.10.- Site frequency spectrum from the UK10K dataset compared with the SFS obtained using the demographic parameters inferred by fastNeutrino (Bhaskar et al. 2015) for the three most recent population size epochs and using Gravel's demographic history for the most ancient population sizes (Gravel et al. 2011b). The agreement between both distributions, based on the KL divergence is 0.00185.

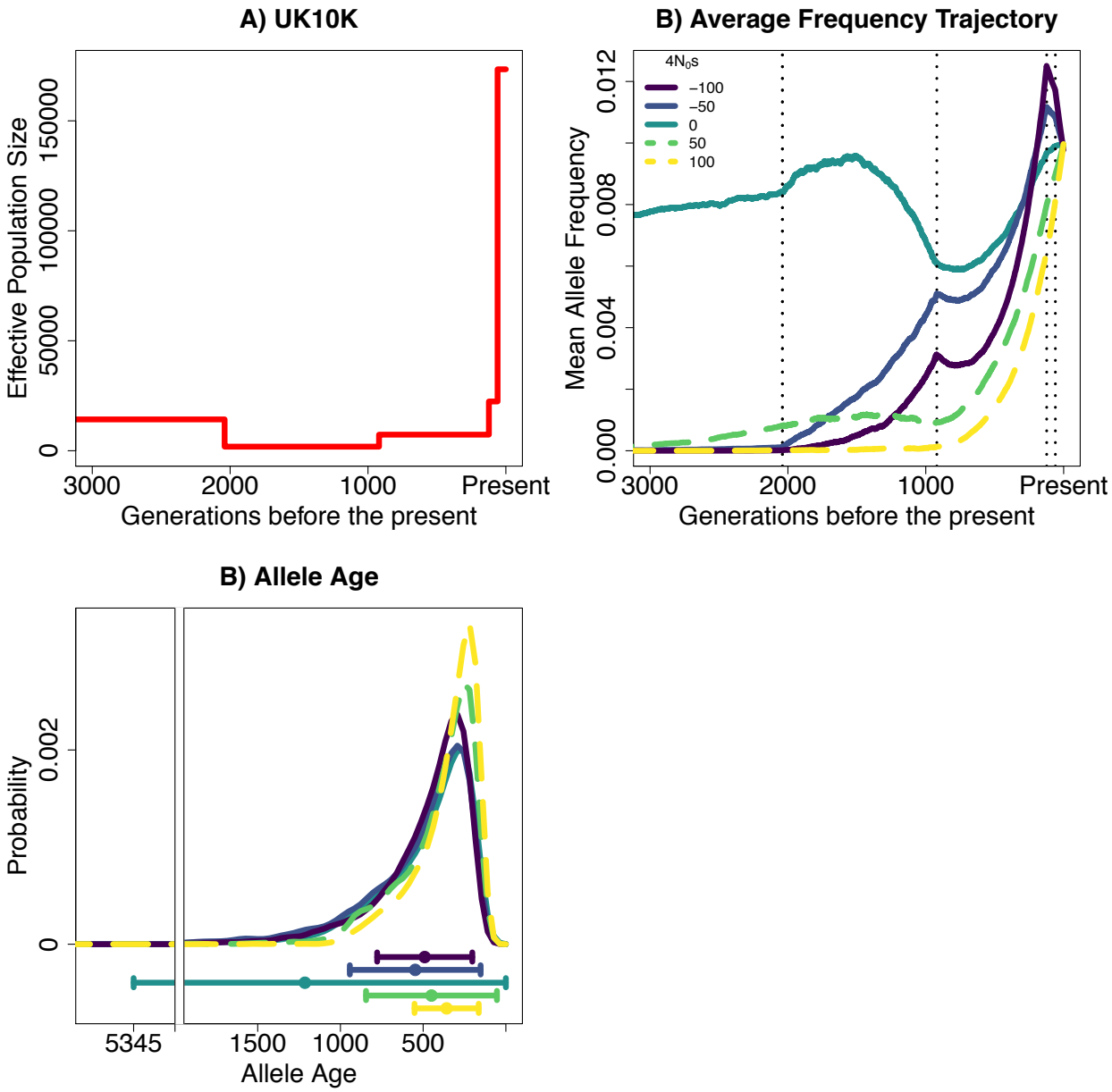

Figure 4.11.- **Properties of alleles sampled at a 1% frequency under different strengths of natural selection in the demographic scenario inferred in the UK10K data.** A) Population model inferred in the UK10K dataset. B) Mean allele frequency at different times in the past, in units of generations.; C) Probability distribution of allele ages and D) Probability distribution of pairwise coalescent times T_2. The dot and

whiskers below C) and D) represent the mean value of the distribution and the two whiskers extend at both sides of the mean until max(mean +- s.d. ,0).

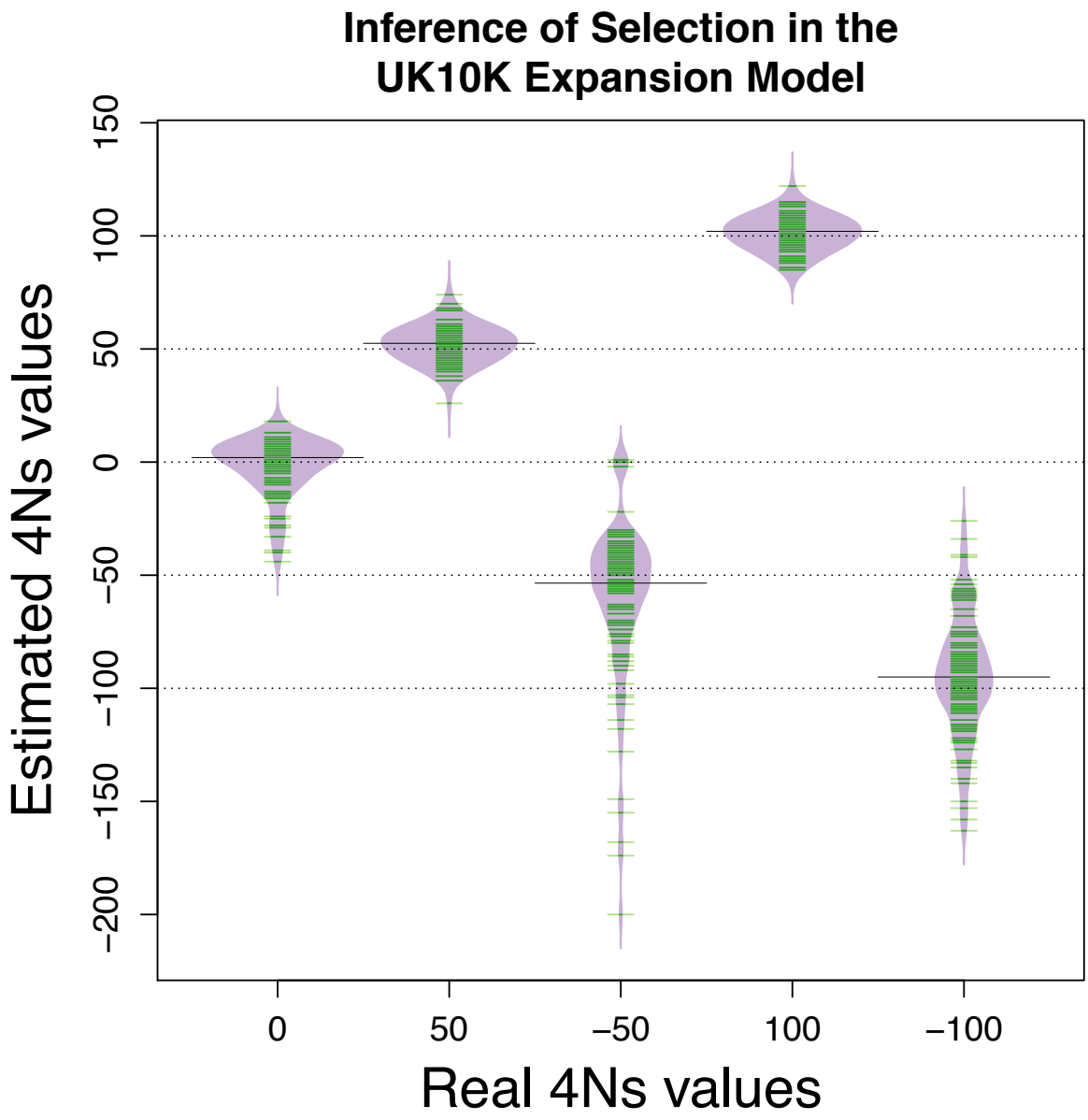

Figure 4.12.- **Estimation of the strength of natural selection in the demographic model inferred in the UK10K dataset.**

Discussion

We have developed a composite likelihood method to estimate the strength of natural selection acting on alleles at a certain frequency in the population. This method takes demography into account and uses the differences in the distribution of pairwise identity by state lengths L on alleles under different strengths of selection. We found that in a constant population size scenario, the distribution of L captures differences in the absolute strength of the selection coefficient 4Ns. However, since the distribution of L does not differ between advantageous and deleterious alleles under the same strength of selection, it is not possible to differentiate between positive and negative selection using that particular haplotype signature. More broadly, since the mean allele frequency trajectory is almost identical for deleterious and advantageous alleles under the same selective constraint, any statistic based on haplotype signatures will be insufficient in that scenario to distinguish between positive and negative selection in a model with only one selected allele with linked neutral variation surrounding it.

On the other hand, we found that the distribution of L is sufficient to differentiate between advantageous and deleterious alleles under some non-equilibrium demographic scenarios, including the demographic scenario inferred from the UK10K dataset. This is encouraging, since most of the natural species are very likely to have evolved a non-equilibrium demographic scenario and it is precisely in those scenarios where we would like to be able to differentiate between alleles under different strengths of selection.

We found a particularly interesting pattern of the mean allele frequency trajectories of deleterious alleles segregating at 1% frequency alleles when the population is expanding. These alleles tend to increase in frequency when the population size is low. However, they decrease in frequency when the population expands due to a higher efficacy of selection. This suggest that it is likely that, on average, deleterious alleles would tend to come from higher frequencies in the recent past on populations under expansion.

We propose a general strategy to infer the distribution of fitness effects of alleles at a particular frequency in the population. Here we used information from the site frequency spectrum at synonymous sites to infer a demographic model. However, it is possible to use information from other sources such as haplotypic data. After the demographic model is inferred, it is possible to calculate the distribution of L given selection and the demographic model to use the composite likelihood approach to estimate selection. We only explored demographic models where there is one single population changing its size in the past. One avenue of future exploration is to understand how the distribution of L changes in models where there are more demes and migration is allowed between the demes. Migration should not affect the distribution of L for lower frequency alleles, since they are more likely to be geographically restricted and gene flow would have a smaller effect on the patterns of variation in low frequency variant carrying haplotypes.

Our method assumes that every variant under putative selection does not appear more than once in the population. Recurrent mutations have been found in the population, particularly in large-sample size studies such as ExAc. However, recurrent mutations

are more unlikely for low frequency variants, since there is less time for many mutations to appear in the same site on those variants.

When we estimated parameters that define the DFE of segregating variants, we found that our method can provide reasonable estimates of the parameters that would lead to estimating an accurate value of the mean of the DFE. However, estimating the variance of the DFE accurately is more challenging, as can be seen from the variation of estimates of the DFE parameters.

Although here we analyzed the distribution of fitness effects of nonsynonymous variants at a certain frequency, it is possible to determine the distribution of fitness effects of variants with a different functional category. One possibility is to try to determine the strength of selection of alleles on variants that are predicted to be more deleterious based on the fitcons scores or the C-scores (Racimo & Schraiber 2014). This can help us to obtain genome-wide estimates of the selection coefficient of variants, based on their predicted functional category. This is of particular interest to genome-wide association studies, due to the interest in understanding the association between associated variants and their selection coefficients on different complex traits.

Appendix - Simulation Command Lines

Command Line 1. *G-PhoCS* model with the full set of migration bands inferred:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 –n 1 0.000010
-n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n 5 0.000457 -n 6
0.000217 -n 7 0.000778 -m 2 4 4505.0 -m 4 2 1840.0 -m 3 6 573.0
-m 6 3 942.0 -m 4 7 58.0 –m 7 4 1162.0 -ej 0.0000403 2 1 -en
0.0000403 1 0.000032 -em 0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0
–em 0.0000403 2 4 0.0 -em 0.0000403 4 2 0.0 -ej 0.0000427 3 1 -en
0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 –em 0.0000427 6 1 0.0
-em 0.0000427 3 6 0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -
en 0.0000446 4 0.000056 –em 0.0000446 1 4 0.0 -em 0.0000446 4 1
0.0 -em 0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6
4 –en 0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 –em 0.0000449 4
1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449 7 4 0.0 -ej 0.0000496
4 1 -en 0.0000496 1 0.001800 –em 0.0000496 1 4 0.0 -em 0.0000496
4 1 0.0 -em 0.0000496 4 7 0.0 -em 0.0000496 7 4 0.0 -em
0.0000496 1 7 17.0 -em 0.0000496 7 1 746.0 -ej 0.0013275 7 1 -en
0.0013275 1 0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 2. The model inferred from *G-PhoCS* but with no gene flow between any species at any time:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 –n 1 0.000010
-n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n 5 0.000457 -n 6
0.000217 -n 7 0.000778 -m 2 4 0.0 –m 4 2 0.0 -m 3 6 0.0 -m 6 3
0.0 -m 4 7 0.0 -m 7 4 0.0 -ej 0.0000403 2 1 -en 0.0000403 1
0.000032 –em 0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0 -em
0.0000403 2 4 0.0 -em 0.0000403 4 2 0.0 -ej 0.0000427 3 1 –en
0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 –em 0.0000427 6 1 0.0
-em 0.0000427 3 6 0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -
en 0.0000446 4 0.000056 –em 0.0000446 1 4 0.0 -em 0.0000446 4 1
0.0 -em 0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6
4 –en 0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 –em 0.0000449 4
1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449 7 4 0.0 -ej 0.0000496
4 1 -en 0.0000496 1 0.001800 –em 0.0000496 1 4 0.0 -em 0.0000496
4 1 0.0 -em 0.0000496 4 7 0.0 -em 0.0000496 7 4 0.0 -em
0.0000496 1 7 0.0 -em 0.0000496 7 1 0.0 -ej 0.0013275 7 1 -en
0.0013275 1 0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 3. The model inferred from *G-PhoCS* but with only one event of gene flow, from the golden jackal to the ancestor of dogs and wolves:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 –n 1 0.000010
-n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n 5 0.000457 -n 6
0.000217 -n 7 0.000778 -m 2 4 0.0 –m 4 2 0.0 -m 3 6 0.0 -m 6 3
0.0 -m 4 7 0.0 -m 7 4 0.0 -ej 0.0000403 2 1 -en 0.0000403 1
```

```
0.000032 –em 0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0 -em
0.0000403 2 4 0.0 -em 0.0000403 4 2 0.0 -ej 0.0000427 3 1 –en
0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 –em 0.0000427 6 1 0.0
-em 0.0000427 3 6 0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -
en 0.0000446 4 0.000056 –em 0.0000446 1 4 0.0 -em 0.0000446 4 1
0.0 -em 0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6
4 –en 0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 –em 0.0000449 4
1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449 7 4 0.0 -ej 0.0000496
4 1 -en 0.0000496 1 0.001800 –em 0.0000496 1 4 0.0 -em 0.0000496
4 1 0.0 -em 0.0000496 4 7 0.0 -em 0.0000496 7 4 0.0 -em
0.0000496 1 7 17.0 -em 0.0000496 7 1 0.0 -ej 0.0013275 7 1 -en
0.0013275 1 0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 4. The model inferred from *G-PhoCS* but with only one event of gene flow, from the ancestor of dogs and wolves to golden jackal:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 -n 1 0.000010
-n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n 5 0.000457 -n 6
0.000217 -n 7 0.000778 -m 2 4 0.0 -m 4 2 0.0 -m 3 6 0.0 -m 6 3
0.0 -m 4 7 0.0 -m 7 4 0.0 -ej 0.0000403 2 1 -en 0.0000403 1
0.000032 -em 0.0000403 1 4 0.0 –em 0.0000403 4 1 0.0 -em
0.0000403 2 4 0.0 -em 0.0000403 4 2 0.0 -ej 0.0000427 3 1 -en
0.0000427 1 0.000080 –em 0.0000427 1 6 0.0 -em 0.0000427 6 1 0.0
```

```
-em 0.0000427 3 6 0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -
en 0.0000446 4 0.000056 -em 0.0000446 1 4 0.0 -em 0.0000446 4 1
0.0 –em 0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6 4
-en 0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 -em 0.0000449 4 1
0.0 -em 0.0000449 4 7 0.0 -em 0.0000449 7 4 0.0 –ej 0.0000496 4 1
-en 0.0000496 1 0.001800 -em 0.0000496 1 4 0.0 -em 0.0000496 4 1
0.0 -em 0.0000496 4 7 0.0 –em 0.0000496 7 4 0.0 -em 0.0000496 1 7
0.0 -em 0.0000496 7 1 746.0 -ej 0.0013275 7 1 -en 0.0013275 1
0.000727 –em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 5. The model inferred from *G-PhoCS* but with only one event of gene flow, from Israeli wolf to golden jackal:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 -n 1 0.000010
-n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n 5 0.000457 -n 6
0.000217 -n 7 0.000778 -m 2 4 0.0 -m 4 2 0.0 -m 3 6 0.0 -m 6 3
0.0 -m 4 7 0.0 -m 7 4 1162.0 –ej 0.0000403 2 1 -en 0.0000403 1
0.000032 -em 0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0 -em
0.0000403 2 4 0.0 –em 0.0000403 4 2 0.0 -ej 0.0000427 3 1 -en
0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 -em 0.0000427 6 1 0.0
–em 0.0000427 3 6 0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -en
0.0000446 4 0.000056 -em 0.0000446 1 4 0.0 -em 0.0000446 4 1 0.0
-em 0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 –ej 0.0000449 6 4 -en
0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 -em 0.0000449 4 1 0.0
```

-em 0.0000449 4 7 0.0 –em 0.0000449 7 4 0.0 -ej 0.0000496 4 1 -en 0.0000496 1 0.001800 -em 0.0000496 1 4 0.0 -em 0.0000496 4 1 0.0 –em 0.0000496 4 7 0.0 -em 0.0000496 7 4 0.0 -em 0.0000496 1 7 0.0 -em 0.0000496 7 1 0.0 -ej 0.0013275 7 1 -en 0.0013275 1 0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0

Command Line 6. *ms* command line that uses the demographic history estimated from *G- PhoCS*.

./ms 7 1 -t 1000 -r 920 1000 -I 7 1 1 1 1 1 1 1 -n 1 0.000010 -n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n 5 0.000457 -n 6 0.000217 -n 7 0.000778 -m 2 4 4505.0 -m 4 2 1840.0 -m 3 6 573.0 -m 6 3 942.0 -m 4 7 58.0 -m 7 4 1162.0 -ej 0.0000403 2 1 -en 0.0000403 1 0.000032 -em 0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0 -em 0.0000403 2 4 0.0 –em 0.0000403 4 2 0.0 -ej 0.0000427 3 1 -en 0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 -em 0.0000427 6 1 0.0 –em 0.0000427 3 6 0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -en 0.0000446 4 0.000056 -em 0.0000446 1 4 0.0 -em 0.0000446 4 1 0.0 -em 0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 –ej 0.0000449 6 4 -en 0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 -em 0.0000449 4 1 0.0 -em 0.0000449 4 7 0.0 –em 0.0000449 7 4 0.0 -ej 0.0000496 4 1 -en 0.0000496 1 0.001800 -em 0.0000496 1 4 0.0 -em 0.0000496 4 1 0.0 –em 0.0000496 4 7 0.0 -em 0.0000496 7 4 0.0 -em 0.0000496 1 7 17.0 -em 0.0000496 7 1 746.0 -ej 0.0013275 7 1 –en 0.0013275 1

```
0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 7. Model where the dogs and wolves are each a separate clade, identical to Command Line 1, except for the simulation of smaller (2Mb) genomic regions.

```
./macs 13 2000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 -n 1 0.000010
-n 2 0.000106 -n 3 0.000077 -n 4 0.001044 –n 5 0.000457 -n 6
0.000217 -n 7 0.000778 -m 2 4 4505.0 -m 4 2 1840.0 -m 3 6 573.0
-m 6 3 942.0 -m 4 7 58.0 –m 7 4 1162.0 -ej 0.0000403 2 1 -en
0.0000403 1 0.000032 -em 0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0
–em 0.0000403 2 4 0.0 -em 0.0000403 4 2 0.0 -ej 0.0000427 3 1 -en
0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 –em 0.0000427 6 1 0.0
-em 0.0000427 3 6 0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -
en 0.0000446 4 0.000056 –em 0.0000446 1 4 0.0 -em 0.0000446 4 1
0.0 -em 0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6
4 –en 0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 –em 0.0000449 4
1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449 7 4 0.0 -ej 0.0000496
4 1 -en 0.0000496 1 0.001800 –em 0.0000496 1 4 0.0 -em 0.0000496
4 1 0.0 -em 0.0000496 4 7 0.0 -em 0.0000496 7 4 0.0 -em
0.0000496 1 7 17.0 - em 0.0000496 7 1 746.0 -ej 0.0013275 7 1 -
en 0.0013275 1 0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1
0.0
```

Command Line 8. Regional domestication model.

```
./macs 13 2000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 -n 1 0.000010
-n 2 0.000128 -n 3 0.000032 -n 4 0.000889 –n 5 0.000565 -n 6
0.000171 -n 7 0.000771 -m 1 2 20054 –m 2 1 59 -m 1 3 3459 -m 3 1
9560 -m 2 3 51 -m 3 2 7618 -m 4 5 5276 -m 5 4 48 -m 4 6 19 -m 6
4 4958 -m 5 6 26 -m 6 5 5312 -m 4 7 182.0 -m 7 4 1207.0 -ej
0.0000478 4 2 -en 0.0000478 2 0.000437 -em 0.0000478 1 2 0.0 –em
0.0000478 2 1 0.0 -em 0.0000478 2 3 0.0 -em 0.0000478 3 2 0.0 -
em 0.0000478 4 5 0.0 -em 0.0000478 5 4 0.0 -em 0.0000478 4 6 0.0
-em 0.0000478 6 4 0.0 –em 0.0000478 4 7 0.0 -em 0.0000478 7 4 0.0
     -ej 0.0000614 5 1 -en 0.0000614 1 0.000162 -em 0.0000478 1 2
0.0 –em 0.0000478 2 1 0.0 -em 0.0000478 1 3 0.0 -em 0.0000478 3 1
0.0 -em 0.0000478 4 5 0.0 -em 0.0000478 5 4 0.0 -em 0.0000478 5
6 0.0 -em 0.0000478 6 5 0.0 –ej 0.0000617 6 3 -en 0.0000617 3
0.000017 -em 0.0000478 3 2 0.0 -em 0.0000478 2 3 0.0 -em
0.0000478 1 3 0.0 –em 0.0000478 3 1 0.0 -em 0.0000478 6 5 0.0 -em
0.0000478 5 6 0.0 -em 0.0000478 4 6 0.0 -em 0.0000478 6 4 0.0 -
ej 0.0000618 2 1 -en 0.0000618 1 0.000252 –ej 0.0000626 3 1 -en
0.0000626 1 0.001790 -em 0.0000626 1 7 3.0 -em 0.0000626 7 1
782.0 -ej 0.0013859 7 1 –en 0.0013859 1 0.000682 -em 0.0013859 1
7 0.0 -em
      0.0013859 7 1 0.0
```

Command Line 9. Origin of dogs from the Israeli wolf.

```
./macs 13 2000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 -n 1 0.000010
-n 2 0.000103 -n 3 0.000076 -n 4 0.000894 –n 5 0.000445 -n 6
0.000221 -n 7 0.000765 -m 2 4 5032.0 -m 4 2 1196.0 -m 3 6 865.0
-m 6 3 524.0 -m 4 7 142.0 –m 7 4 1063.0 -ej 0.0000401 2 1 -en
0.0000401 1 0.000025 -em 0.0000401 1 4 0.0 -em 0.0000401 4 1 0.0
–em 0.0000401 2 4 0.0 -em 0.0000401 4 2 0.0 -ej 0.0000419 3 1 -en
0.0000419 1 0.000029 -em 0.0000419 1 6 0.0 –em 0.0000419 6 1 0.0
-em 0.0000419 3 6 0.0 -em 0.0000419 6 3 0.0 -ej 0.0000444 4 1 -
en 0.0000444 1 0.000186 –em 0.0000444 1 4 0.0 -em 0.0000444 4 1
0.0 -em 0.0000444 4 7 0.0 -em 0.0000444 7 4 0.0 -ej 0.0000447 5
1 –en 0.0000447 1 0.000229 -em 0.0000447 1 4 0.0 –em 0.0000447 4
1 0.0 -em 0.0000447 4 7 0.0 -em 0.0000447 7 4 0.0 -ej 0.0000450
6 1 -en 0.0000450 1 0.001801 –em 0.0000450 1 4 0.0 -em 0.0000450
4 1 0.0 -em 0.0000450 4 7 0.0 -em 0.0000450 7 4 0.0 -em
0.0000450 1 7 5.0 -em 0.0000450 7 1 778.0 -ej 0.0013954 7 1 -en
0.0013954 1 0.000663 -em 0.0013954 1 7 0.0 -em 0.0013954 7 1 0.0
```

Bibliography

Akashi, H., 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: Statistical power to detect directional selection under stationarity and free recombination. *Genetics*, 151(1), pp.221–238.

Akashi, H., Osada, N. & Ohta, T., 2012. Weak selection and protein evolution. *Genetics*, 192(1), pp.15–31.

Ash, E.C., 1927. *Dogs: Their History and Development*, London: E. Benn Limited.

Auton, A. et al., 2013. Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genet*, 9(12), p.e1003984.

Balick, D.J. et al., 2015. Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck. *PLoS Genet*, 11(8), p.e1005436.

Barton, N.H., 2000. Genetic hitchhiking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 355(1403), pp.1553–1562.

Bhaskar, A., Wang, Y.X.R. & Song, Y.S., 2015. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, (25), pp.268–279.

Björnerfeldt, S., Webster, M.T. & Vilà, C., 2006. Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Research*, 16(8), pp.990–994.

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

Boyko, A.R. et al., 2008a. Assessing the evolutionary impact of amino acid mutations in the human genome. *Plos Genetics*, 4(5), p.e1000083.

Boyko, A.R. et al., 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, 4(5), p.e1000083.

Boyko, A.R. et al., 2008b. Assessing the evolutionary impact of amino acid mutations in the human genome. *Plos Genetics*, 4(5), p.e1000083.

Boyko, A.R. et al., 2009. Complex population structure in African village dogs and its implications for inferring dog domestication history. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33), pp.13903–13908.

Boyko, A.R., 2011. The domestic dog: man's best friend in the genomic era. *Genome Biol*, 12(2), p.216.

Brandvain, Y. & Wright, S.I., 2016. The Limits of Natural Selection in a Nonequilibrium World. *Trends in genetics : TIG*, 32(4), pp.201–210. Available at: http://www.sciencedirect.com/science/article/pii/S0168952516000147.

Bustamante, C.D. et al., 2001. Directional selection and the site-frequency spectrum. *Genetics*, 159(4), pp.1779–1788.

Bustamante, C.D. et al., 2005. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062), pp.1153–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16237444 [Accessed March 9, 2012].

Bustamante, C.D. et al., 2002. The cost of inbreeding in Arabidopsis. *Nature*,

416(6880), pp.531–534. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11932744&retmode=ref&cmd=prlinks\npapers3://publication/doi/10.1038/416531a.

Charlesworth, B. & Charlesworth, D., 2010. *Elements of Evolutionary Genetics*, Roberts & Company Publishers.

Charlesworth, B., Morgan, M.T. & Charlesworth, D., 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4), pp.1289–303. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205596&tool=pmcentrez&rendertype=abstract.

Chen, G.K., Marjoram, P. & Wall, J.D., 2009. Fast and flexible simulation of DNA sequence data. *Genome research*, 19(1), pp.136–42. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2612967&tool=pmcentrez&rendertype=abstract [Accessed March 12, 2012].

Clarke, B., 1970. Darwinian Evolution of Proteins Author (s): Bryan Clarke Source : Science , New Series , Vol . 168 , No . 3934 (May 22 , 1970), pp . 1009-1011 Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/. *Science*, 168, pp.1009–1011.

Club, A.K., 1997. *The Comple Dog Book*, New York, NY: Howell Book House.

Coop, G. & Ralph, P., 2012. Patterns of neutral diversity under general models of selective sweeps. *Genetics*, 192(1), pp.205–224.

Crisci, J.L. et al., 2013. The impact of equilibrium assumptions on tests of selection.

Frontiers in Genetics, 4(NOV), pp.1–7.

Crow, J.F., 1972. The dilemma of nearly neutral mutations: how important are they for evolution and human welfare? *Journal of Heredity*, 63, pp.306–316. Available at: file:///Y:/532.pdf.

Cruz, F., Vila, C. & Webster, M.T., 2008. The legacy of domestication: accumulation of deleterious mutations in the dog genome. *Mol Biol Evol*, 25(11), pp.2331–6.

DePristo, M.A. et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5), pp.491–8.

Do, R. et al., 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genetics*, 47(2), pp.126–131. Available at: http://dx.doi.org/10.1038/ng.3186.

Durand, E.Y. et al., 2011. Testing for ancient admixture between closely related populations. *Mol Biol Evol*, 28(8), pp.2239–2252.

EFRON, B., 1981. NONPARAMETRIC ESTIMATES OF STANDARD ERROR - THE JACKKNIFE, THE BOOTSTRAP AND OTHER METHODS. *Biometrika*, 68(3), pp.589–599.

Elyashiv, E. et al., 2010. Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. *Genome research*, 20(11), pp.1558–73. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2963819&tool=pmcentrez&rendertype=abstract [Accessed March 28, 2012].

Ewens, W.J., 2012. James F. crow and the stochastic theory of population genetics.

Genetics, 190(2), pp.287–290.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1), pp.87–112.

Excoffier, L. et al., 2013. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, 9(10).

Eyre-Walker, A., 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics*, 162(4), pp.2017–2024.

Eyre-Walker, A., 2010. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences of the United States of America*, 107 Suppl, pp.1752–1756. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20133822&retmode=ref&cmd=prlinks\npapers3://publication/doi/10.1073/pnas.0906182107.

Eyre-Walker, A. & Keightley, P.D., 2007. The distribution of fitness effects of new mutations. *Nature reviews. Genetics*, 8(8), pp.610–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17637733 [Accessed March 8, 2012].

Fay, J.C., Wyckoff, G.J. & Wu, C.I., 2001. Positive and negative selection on the human genome. *Genetics*, 158(3), pp.1227–1234. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461725&tool=pmcentrez&rendertype=abstract.

Fearnhead, P. & Donnelly, P., 2001. Estimating recombination rates from population genetic data. *Genetics*, 159, pp.1299–1318.

Felsenstein, J., 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5, p.164.

Freedman, A.H. et al., 2016. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. *PLoS Genetics*, 12(3), pp.1–23.

Freedman, A.H. et al., 2014. Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genet*, 10(1), p.e1004016.

Fu, W. et al., 2014. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *American Journal of Human Genetics*, 95(4), pp.421–436. Available at: http://dx.doi.org/10.1016/j.ajhg.2014.09.006.

Gazave, E. et al., 2013. Population Growth Inflates the Per-Individual Number of Deleterious Mutations and Reduces Their Mean Effect. *Genetics*, 195(3), pp.969–978.

Germonpré, M. et al., 2009. Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. *Journal of Archaeological Science*, 36(2), pp.473–490. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0305440308002380 [Accessed March 11, 2012].

Germonpré, M., Lázničková-Galetová, M. & Sablin, M. V., 2012. Palaeolithic dog skulls at the Gravettian Předmostí site, the Czech Republic. *Journal of Archaeological Science*, 39(1), pp.184–202.

Gillespie, J.H., 2000. Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics*, 155(2), pp.909–919.

Gillespie, J.H., 1984. The molecular clock may be an episodic clock. *Proceedings of the National Academy of Sciences of the United States of America*, 81(24), pp.8009–13. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=392283&tool=pmcentrez&rendertype=abstract.

Gravel, S. et al., 2011a. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29), pp.11983–11988. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3142009&tool=pmcentrez&rendertype=abstract.

Gravel, S. et al., 2011b. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29), pp.11983–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3142009&tool=pmcentrez&rendertype=abstract [Accessed March 9, 2012].

Gravel, S., 2016. When is selection effective. *Genetics*, Early Onli(XXX), p.XXX.

Gray, M.M. et al., 2009. Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, 181(4), pp.1493–505.

Gray, M.M. et al., 2009. Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, 181(4), pp.1493–505. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2666515&tool=pmcentrez&rendertype=abstract [Accessed July 11, 2011].

Green, R.E. et al., 2010. A draft sequence of the Neandertal genome. *Science (New*

York, N.Y.), 328(5979), pp.710–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20448178 [Accessed July 19, 2011].

Gronau, I. et al., 2011a. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, 43(10), pp.1031–1034. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245873&tool=pmcentrez&rendertype=abstract.

Gronau, I. et al., 2011b. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, 43(10), pp.1031–4. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245873&tool=pmcentrez&rendertype=abstract [Accessed March 9, 2012].

Gutenkunst, R.N. et al., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*, 5(10), p.e1000695. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2760211&tool=pmcentrez&rendertype=abstract [Accessed March 1, 2012].

Hahn, M.W., 2008. Toward a selection theory of molecular evolution. *Evolution*, 62(2), pp.255–265.

Haldane, J.B.S., 1957. The cost of natural selection. *Journal of Genetics*, 55(3), pp.511–524.

Halligan, D.L. et al., 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet*, 9(12), p.e1003995.

Halligan, D.L. et al., 2013. Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. *PLoS Genetics*, 9(12).

Harris, K. & Nielsen, R., 2013. Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, 9(6).

Henn, B.M. et al., 2015. Estimating the mutation load in human genomes. *Nature Reviews Genetics*, 16(6), pp.333–343.

Hudson, R.R. & Kaplan, N.L., 1995. Deleterious background selection with recombination. *Genetics*, 141, pp.1605–1617.

Jensen, J.D. et al., 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics*, 170(3), pp.1401–1410.

Jukes, T.H., 1978. Neutral changes during divergent evolution of hemoglobins. *Journal of Molecular Evolution*, 11, pp.302–307.

Keightley, P.D. & Eyre-Walker, A., 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177(4), pp.2251–2261. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2219502&tool=pmcentrez&rendertype=abstract.

Kiezun, A. et al., 2013. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. *PLoS genetics*, 9(2), p.e1003301. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3585140&tool=pmcentrez&rendertype=abstract [Accessed January 10, 2014].

Kimura, M., 1980. Average Time until Fixation of a Mutant Allele in a Finite Ppopulation under Continued Mutation Pressure: Studies by Analytical, Numerical, and Pseudo-Sampling Methods. *Proceedings of the National Academy of Sciences of the United States of America*, 77(1), pp.522–526.

Kimura, M., 1968. Evolutionary Rate at the Molecular Level. *Nature*, 217, pp.624–626. Available at: http://linkinghub.elsevier.com/retrieve/pii/B0122270800004328.

Kimura, M., 1962. On the probability of fixation of mutant genes in a population. *Genetics*, 47(391), pp.713–719.

Kimura, M. & Crow, J.F., 1964. The Number of Alleles that can be maintained in a finite population. *Genetics*, 49, pp.725–738.

Kimura, M. & Ohta, T., 1974. On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 71(7), pp.2848–2852.

King, J.L. & Jukes, T.H., 1969. Non-Darwinian evolution. *Science*, 164(3881), pp.788–798.

Koepfli, K.P. et al., 2015. Genome-wide evidence reveals that African and Eurasian golden jackals are distinct species. *Current Biology*, 25(16), pp.2158–2165.

Kreitman, M., 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. *Nature*, 304(5925), pp.412–417. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6410283?dopt=abstract\npapers3://publication/uuid/7232C446-A003-482F-A7B7-BD4311A78E99.

Kreitman, M., 1996. The neutral theory is dead. Long live the neutral theory. *Bioessays*,

18(8), pp.678–83; discussion 683. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8760341\nhttp://onlinelibrary.wiley.com/doi/10.1002/bies.950180812/full.

Larson, G. et al., 2012. Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proc Natl Acad Sci USA*, 109(23), pp.8878–83. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3384140&tool=pmcentrez&rendertype=abstract.

Larson, G. & Burger, J., 2013. A population genetics view of animal domestication. *Trends in Genetics*, 29(4), pp.197–205. Available at: http://dx.doi.org/10.1016/j.tig.2013.01.003.

Leffler, E.M. et al., 2012. Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? *PLoS Biology*, 10(9).

Leonard, J.A. et al., 2007. Megafaunal Extinctions and the Disappearance of a Specialized Wolf Ecomorph. *Current Biology*, 17(13), pp.1146–1150.

Lewontin, R.C., 1974. Chapter 5 BT - The Genetic Basis of Evolutionary Change. *The Genetic Basis of Evolutionary Change*, pp.1–84. Available at: http://authors.library.caltech.edu/5456/1/hrst.mit.edu/hrs/evolution/public/papers/lewontin1974/lewontin1974_chap5.pdf\npapers2://publication/uuid/2E5A729E-E8A0-4C82-BF8E-4F742F491DFA.

Li, H. & Durbin, R., 2011a. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357), pp.493–496.

Li, H. & Durbin, R., 2011b. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357), pp.493–496.

Lindblad-Toh, K. et al., 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069), pp.803–19.

Lindblad-Toh, K. et al., 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069), pp.803–819. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16341006.

Loewe, L. et al., 2006. Estimating selection on nonsynonymous mutations. *Genetics*, 172(2), pp.1079–1092. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1456207&tool=pmcentrez&rendertype=abstract.

Lohmueller, K.E. et al., 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature*, 451(7181), pp.994–997.

Lohmueller, K.E., 2014a. The distribution of deleterious genetic variation in human populations. *Current Opinion in Genetics and Development*, 29, pp.139–146. Available at: http://dx.doi.org/10.1016/j.gde.2014.09.005.

Lohmueller, K.E., 2014b. The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. *PLoS Genetics*, 10(5).

Mancuso, N. et al., 2015. The contribution of rare variation to prostate cancer heritability. *In Submission*, 48(1), pp.30–35. Available at: http://dx.doi.org/10.1038/ng.3446.

Marjoram, P. & Wall, J.D., 2006. Fast “coalescent” simulation. *BMC Genetics*, 7(1),

p.16. Available at: http://www.biomedcentral.com/1471-2156/7/16/abstract\nhttp://www.biomedcentral.com/1471-2156/7/16\nhttp://www.biomedcentral.com/content/pdf/1471-2156-7-16.pdf.

Maruyama, T., 1974. The age of an allele in a finite population. *Genetical research*, 23(2), pp.137–143. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4417585.

Maxam, a M. & Gilbert, W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), pp.560–564.

Maynard Smith, J. & Haigh, J., 1974. The hitch-hiking effect of a favourable gene. *Genetical research*, 23(1), pp.23–35.

Maynared Smith, J. & Haigh, J., 1974. The hitch-hiking effect of a favourable gene. *Genetical research*, 23(1), pp.23–35.

McDonald, J.H. & Kreitman, M., 1991. Adaptive protein evolution at the ADH locus in Drosophila. *Nature*, 351, pp.652–654.

McGreevy, P.D. & Nicholas, F.W., 1999. Some Practical Solutions to Welfare Problems in Dog Breeding. *Animal Welfare*, 8(4), pp.329–341.

McVean, G. a T. & Cardin, N.J., 2005. Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1459), pp.1387–93. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1569517&tool=pmcentrez&rendertype=abstract [Accessed August 29, 2011].

Miyata, T. & Yasunaga, T., 1981. Rapidly evolving mouse alpha-globin-related pseudo

gene and its evolutionary history. *Proceedings of the National Academy of Sciences of the United States of America*, 78(1), pp.450–3. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=319071&tool=pmcentrez&rendertype=abstract.

Mukai, T., 1964. The genetic structure of natural populations of Drosophila melanogaster. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics*, 50(500), pp.1–19. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1210633&tool=pmcentrez&rendertype=abstract.

Nordborg, M., 1997. Structured coalescent processes on different time scales. *Genetics*, 146(4), pp.1501–1514. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1208092&tool=pmcentrez&rendertype=abstract.

Nordborg, M., Charlesworth, B. & Charlesworth, D., 1996. The effect of recombination on background selection. *Genetical research*, 67(2), pp.159–174. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8801188.

Ohta, T., 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428), pp.96–98.

Ohta, T., 1992. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*, 23, pp.263–286.

Ohta, T. & Gillespie, J., 1996. Development of Neutral and Nearly Neutral Theories. *Theoretical population biology*, 49(2), pp.128–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8813019.

Ovodov, N.D. et al., 2011. A 33,000-year-old incipient dog from the Altai Mountains of Siberia: evidence of the earliest domestication disrupted by the Last Glacial Maximum. *PloS one*, 6(7), p.e22821. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3145761&tool=pmcentrez&rendertype=abstract [Accessed March 17, 2012].

Palacios, J.A., Wakeley, J. & Ramachandran, S., 2015. Bayesian Nonparametric Inference of Population Size Changes from Sequential Genealogies. *Genetics*, 201(1), pp.281–304. Available at: http://www.genetics.org/content/201/1/281.abstract.

Pang, J.F. et al., 2009. MtDNA data indicate a single origin for dogs south of yangtze river, less than 16,300 years ago, from numerous wolves. *Molecular Biology and Evolution*, 26(12), pp.2849–2864.

Paul, J.S., Steinrücken, M. & Song, Y.S., 2011. An accurate sequentially markov conditional sampling distribution for the coalescent with recombination. *Genetics*, 187(4), pp.1115–1128.

Pavlidis, P., Jensen, J.D. & Stephan, W., 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics*, 185(3), pp.907–922.

Peischl, S. et al., 2013. On the accumulation of deleterious mutations during range expansions. *Molecular Ecology*, 22(24), pp.5972–5982.

Pickrell, J.K. & Pritchard, J.K., 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, 8(11).

Purcell, S. et al., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), pp.559–575.

Racimo, F. & Schraiber, J.G., 2014. Approximation to the Distribution of Fitness Effects across Functional Categories in Human Segregating Polymorphisms. *PLoS Genetics*, 10(11).

Rasmussen, M.D. et al., 2014. Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics*, 10(5).

Richmond, R.C., 1970. Non-Darwinian Evolution: A Critique. *Nature*, 225, pp.1025–1028.

Robert, C.P. & Casella, G., 2010. *Introducing Monte Carlo Methods with R*, Springer.

Sanger, F. & Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3), pp.441–448.

Sanger, F., Nicklen, S. & Coulson, a R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=431765&tool=pmcentrez&rendertype=abstract.

Santiago, E. & Caballero, A., 2005. Variation after a selective sweep in a subdivided population. *Genetics*, 169(1), pp.475–483.

Savolainen, P. et al., 2004. A detailed picture of the origin of the Australian dingo, obtained from the study of mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.*,

101(33), pp.12387–12390.

Savolainen, P. et al., 2002. Genetic evidence for an East Asian origin of domestic dogs. *Science*, 298(5598), pp.1610–3. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12446907.

Sawyer, S.A. & Hartl, D.L., 1992. Population Genetics of Polymorphism and Divergence. *Genetics*, 132, pp.1161–1176.

Schiffels, S. & Durbin, R., 2014. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8), pp.919–25. Available at: http://dx.doi.org/10.1038/ng.3015.

Schubert, M. et al., 2014. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences*, 111(52), pp.E5661–E5669.

Sella, G. et al., 2009. Pervasive natural selection in the Drosophila genome? *PLoS genetics*, 5(6), p.e1000495. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2684638&tool=pmcentrez&rendertype=abstract.

Shannon, L.M. et al., 2015. Genetic structure in village dogs reveals a Central Asian domestication origin. *Proceedings of the National Academy of Sciences*, p.201516215.

Sheehan, S., Harris, K. & Song, Y.S., 2013. Estimating Variable Effective Population Sizes from Multiple Genomes : A Sequentially Markov. , 194(July), pp.647–662.

Simons, Y.B. et al., 2014. The deleterious mutation load is insensitive to recent

population history. *Nat Genet*, 46(3), pp.220–4.

Slatkin, M., 2001. Simulating genealogies of selected alleles in a population of variable size. *Genetical research*, 78(1), pp.49–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11556137.

Slatkin, M. & Wiehe, T., 1998. Genetic hitch-hiking in a subdivided population. *Genetical Research*, 71(2), pp.155–160. Available at: <Go to ISI>://WOS:000075333100007.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), pp.585–95. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1203831&tool=pmcentrez&rendertype=abstract.

Tajima, F., 1989. The effect of change in population size on DNA polymorphism. *Genetics*, 123(3), pp.597–601.

Veeramah, K.R. et al., 2014. Evidence for Increased Levels of Positive and Negative Selection on the X Chromosome versus Autosomes in Humans. *Molecular Biology and Evolution*, p.msu166.

Vilà, C., Seddon, J. & Ellegren, H., 2005. Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends in Genetics*, 21(4), pp.214–218.

vonHoldt, B.M. et al., 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, 464(7290), pp.898–902.

Vonholdt, B.M. et al., 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, 464(7290), pp.898–902. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20237475 [Accessed July 19, 2011].

Wang, G.D. et al., 2013. The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat Commun*, 4(May), p.1860. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23673645.

Wayne, R.K. et al., 1991. Conservation genetics of the endangered Isle Royale gray wolf. *Conservation Biology*, 5(1), pp.44–51.

Whitfield, H.J., Martin, R.G. & Ames, B.N., 1966. Classification of aminotransferase (C gene) mutants in the histidine operon. *Journal of Molecular Biology*, 21(2), pp.335–355. Available at: http://www.sciencedirect.com/science/article/pii/0022283666901033.

Wilton, P.R., Carmi, S. & Hobolth, A., 2015. The SMC' Is a Highly Accurate Approximation to the Ancestral Recombionation Graph. *Genetics*, 200(May), pp.343–355.

Wong, A.K. et al., 2010. A comprehensive linkage map of the dog genome. *Genetics*, 184(2), pp.595–605. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2828735&tool=pmcentrez&rendertype=abstract [Accessed September 8, 2011].

Wright, S., 1951. The genetical structure of populations. *Annals of Eugenics*, 15(4), pp.323–354.

Zhang, W. et al., 2014. Hypoxia Adaptations in the Grey Wolf (Canis lupus chanco) from Qinghai-Tibet Plateau. *PLoS Genet*, 10(7), p.e1004466.